Sciometrics joins NIST's contactless fingerprint capture devices research program

Jun 8, 2016 | Stephen Mayhew

CATEGORIES Biometric R&D | Biometrics News | Fingerprint Recognition

Sciometrics has revealed a partnership with the National Institute of Standards and Technology (NIST) to develop standards for image quality for contactless fingerprint capture devices such as smartphones and other mobile devices.

The purpose of the partnership, a Cooperative Research and Development Agreement (CRADA), is to develop new methods and standards that can support evaluation of these devices by the U.S. Government for certification and use by federal agencies. Certification by the Federal Bureau of Investigation (FBI) for submission of fingerprints for background checks is an important topic in the CRADA program.

"We are pleased to contribute to the science and standards in an initiative that will greatly expand mobile identity

ID for All – Biometrics and digital ID initiatives in developing nations

MOST READ THIS WEEK

Fingerprint biometric attendance systems suspended in India as impacts of coronavirus considered

Biometrics industry lobbies for responsible facial recognition regulation

Biometric passports, digital ID initiatives in Africa and contactless tech top this week's biometrics news

ICAO publishes Master List of digital certificates for biometric passport authentication

management while dramatically lowering the costs," said Mark Walch, CEO of Sciometrics. "As we announced earlier this month, our mobile SlapShot app, the subject of the CRADA, is already in beta testing with government agencies. It has also generated interest from the private sector ranging from app security providers to smartphone manufacturers who see an opportunity to replace fingerprint sensors."

NIST, working with the FBI's Biometric Center of Excellence, started its Contactless Fingerprint Capture Device Measurement Research Program in 2015. Through the CRADA, Sciometrics will provide its SlapShot mobile app for blind testing and join NIST's other CRADA partners, including 3M, MorphoTrack and Hoyos Laboratories. The CRADA program requires a close working relationship between NIST and each CRADA participant to ensure the performance metrics that emerge from the collaboration will be suitable for a broad range of devices. Sciometrics will contribute to the development of standards while working to improve its products.

The SlapShot app, when in "administered" mode, captures all ten fingers plus a face in less than a minute using only a phone's camera. And a "fingerprint selfie" version will allow users to sign into apps or their phones using between one and four fingers. The Android app obtains images already proven to match against non-customized third party AFIS systems. An iOS app has also been prototyped and will be added later this year.

Article Topics

biometrics | data collection | fingerprint recognition | identity verification | mobile identity credentialing | National Institute of Standards and Technology | NIST | Sciometrics | touchless

Comments

Idemia successfully pilots contactless biometric access technology at stadium in Japan

Biometrics and digital ID in Africa this week: increasing CCTV surveillance, new ID document contracts

Aadhaar's architect discusses what went into world's biggest biometric repository

Infineon reference design for Qualcomm Snapdragon 865 to ease mobile 3D facial biometric payments

TSA deploys new credential authentication tech for checkpoint screening

Biometric vending machines draw $20M investment as retail app launches facial recognition

Daily Biometrics News »»

More Biometrics Companies »»

BIOMETRICS RESEARCH

Consumer biometrics, speech recognition, 3D sensors and digital ID markets all to grow by at least 17 percent CAGR

6 Replies to “Sciometrics joins NIST’s contactless fingerprint capture devices research program”

1. **BiometricUpdate** says:
   June 8, 2016 at 8:19 pm

   Sciometrics joins NIST’s contactless fingerprint capture devices research program https://t.co/TFmXQglOpk
   Reply

2. **Techscram** says:
   June 8, 2016 at 8:29 pm

   Sciometrics joins NIST’s contactless fingerprint capture devices research program https://t.co/JnrQq0JRav
   Reply

3. **BiometricJustin** says:
   June 9, 2016 at 3:05 am

   RT BiometricUpdate: @@Sciometrics joins NIST’s contactless fingerprint capture devices research program https://t.co/1xAQpZf1HW @usnistgov
   Reply

4. **BiometricAlli** says:
   June 9, 2016 at 6:05 am

   RT BiometricUpdate: @Sciometrics joins NIST’s contactless fingerprint capture devices research program https://t.co/xDcOEH4FJL @usnistgov
   Reply

Biometrics-as-a-service monetization strategies to drive border control market to $3.5B by 2025

Auto biometrics market projected for massive growth as industry players demo new technology

More Biometrics Research »»

BIOMETRICS WHITE PAPERS

Biometrics in Government: addressing vendor lock-in in ABIS systems

Biometrics best practices: Optimizing security in fingerprint technology deployments

How Latin American banks use mobile biometrics and liveness detection to prevent fraud

More White Papers »»

BIOMETRICS EVENTS

connect:ID
Washington D.C.: Mar 11 - Mar 12, 2020

ISC West (postponed until July)
Las Vegas, NV: Mar 17 - Mar 18, 2020

Biometrics Institute US Conference 2020
Washington, DC: Mar 24 - Mar 25, 2020

More Biometrics Events »»

5. **BiometricAmy** says:
June 9, 2016 at 9:05 am

RT BiometricUpdate: @Sciometrics joins NIST's contactless fingerprint capture devices research program https://t.co/DRRe99oqrs @usnistgov
Reply

6. **rawlsonking2** says:
June 9, 2016 at 12:05 pm

RT BiometricUpdate: @Sciometrics joins NIST's contactless fingerprint capture devices research program https://t.co/0M0aHQXjhG @usnistgov
Reply

Leave a Reply

Enter your comment here...

This site uses Akismet to reduce spam. Learn how your comment data is processed.

EXPLAINING BIOMETRICS

- Explainer: Electronic ID (eID)
- Explainer: Facial Recognition
- Explainer: Retinal Scan Technology
- U.K. Law Enforcement Agencies Using High-End Fingerprint Readers
- Explainer: Finger Vein Recognition

More Biometrics Explainers »

Sciometrics

SCIOMETRICS

The Science of Assurance

Sciometrics specializes in patented object recognition technologies, primarily in support of the global military, intelligence, and law enforcement communities. The technology evolved out of optical character recognition technology and has since resulted in advancements within multilingual handwriting biometrics, latent print matching, tool mark analysis, and other related pattern matching capabilities.

Over the last decade, Sciometrics has used its technologies to perform a number of pattern recognition research and development projects for intelligence and law enforcement agencies. Commercialization of the algorithms for more general distribution has taken place over the last several years, resulting in our four current products: *LatentSleuth* our advanced latent print analysis product, *Slapshot*, our smartphone fingerprint capture app, *Digital Lift*, our smartphone latent lift capture app, and *FLASH ID*, our industry leading handwriting biometric product.

Our products and technologies are in use by the Federal Bureau of Investigation, the Department of Defense, and other state, local, and international agencies.

Our Products

LatentSleuth

LatentSleuth

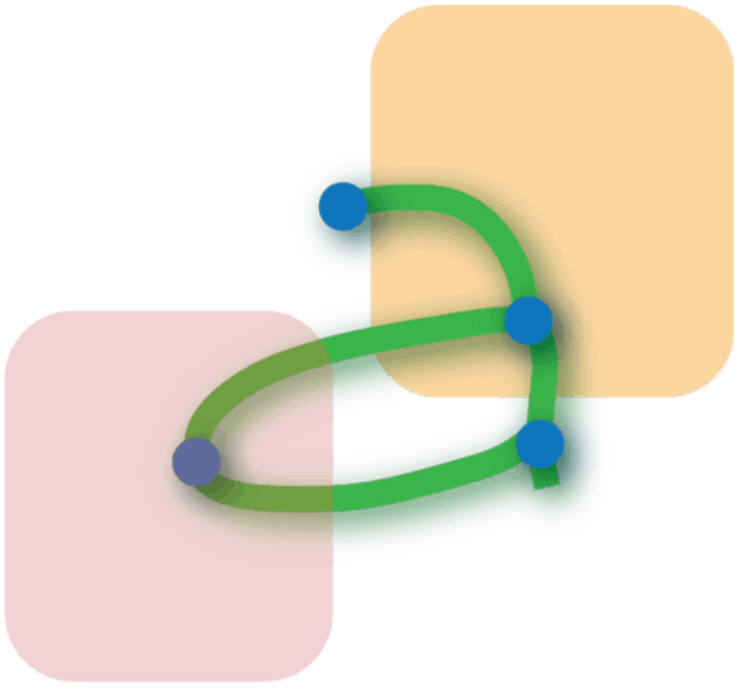
LatentSleuth is a fingerprint analysis, visualization and matching tool designed to assist Latent Print Examiners in the analysis of latent prints against known suspects and match candidates returned from AFIS fingerprint matching systems.

SLAPSHOT

Slapshot

SlapShot is an Android app that turns a commercial Android Smartphone into a biometric capture and matching device. It is designed for ease of use, allowing every officer, soldier, and agent to be a biometic sensor.

Digital Lift



Flash ID

Digital Lift

Digital Lift is an Android app that turns a commercial Android Smartphone into a latent lift capture device. Lift cards, gellifters, and undeveloped latents can be captured and quickly submitted for analysis.

FLASH ID

FLASHID application is a unique handwriting analysis, matching and visualization tool designed to assist law enforcement, military and intelligence professionals in the identification of suspects, criminals, and terrorists via their handwriting.

Sciometrics Technologies

Over the last decade, Sciometrics has used its graphing techniques to perform a number of pattern recognition research and development projects for intelligence and law enforcement agencies.

These include:

Biometrics:
Latent Fingerprint Matching: using fragments - as small as 6x6 millimeters – which did not contain enough minutiae for traditional matching techniques.
Mobile Biometrics: imaging, ingestion and matching of biometrics using standard mobile sensors, i.e., an Android smartphone.
Handwriting Analysis: handwriting biometrics in multiple languages.
Face: using graphing techniques to correct for gaze angles in surveillance photos.
Voice: matching voice snippets.
Biometric Fusion: combining face, small handwriting samples, and small latent fingerprints.

Forensics
Pollen Identification: analysis of 3D pollen images to determine geographic origin of suspect materials.
Tool Mark identification: analysis and identification of tool marks related to crime scenes.
Object recognition and extraction: automated recognition of objects in large image galleries and separating them from backgrounds scenes.

Our Leadership

Mark Walch - President

Richard Smith - VP Products & Services

Robyn Felix - Contract & Operations

Daniel Gantz - Senior Software Engineer

Donald Gantz, Ph.D. - Advisor

CONTACT FORM

Name

Email

Phone

Message

SEND FORM

Address

Sciometrics, LLC
14150 Parkeast Circle, Suite 140
Chantilly, VA 20151

Contacts

Email:
info@sciometrics.com
Phone: +1 (703) 793-2282
Fax: +1 (703) 793-2280

Donald Gantz, Ph.D. A Full Professor of Statistics, Dr. Gantz was the founding Chair of the AIT Department and now serves as the Director of the Document Forensics Laboratory. He has also as served as Interim Associate Dean for Undergraduate Studies in the Volgenau School at George Mason University (GMU). Dr. Gantz is currently applying his cutting edge analysis of handwriting to multi-language document exploitation and biometric identification. He has been active in the research and application of geographic information systems, modeling systems, and decision support systems to transportation demand management and traffic mitigation. Other areas of research include analysis of latent fingerprints, the relationship between TB incidence and socioeconomic factors, surveillance systems to track infection, computer performance evaluation, flight design, and litigation related analyses.

Dr. Gantz earned his Ph.D. and M.A. in Mathematics at the University of Rochester, and his B.A. in Mathematics at Fordham University.

The Volgenau School of Engineering's Department of Applied Information Technology (AIT) recognized and celebrated Dr. Donald Gantz's 10 years of service as department chair and his 40 years as a member of the Mason community at the Johnson Center Bistro on Wednesday, May 14, 2014.

Associate Chair for Undergraduate Studies, Dr. Irene Bruno emceed the event and speakers included Dean Emeritus Dr. Lloyd Griffiths, and Associate Dean for Undergraduate Studies, Dr. Sharon Caraballo. The speakers thanked Dr. Gantz and acknowledged his many contributions to the university, the school, and the department. Dr. Griffiths presented an AIT recognition award to Dr. Gantz.

Dr. Gantz was hired by Mason in 1974 as an assistant professor in the Mathematics Department. When the engineering school was founded, and as it developed, Dr. Gantz continued to serve as faculty and assumed a variety of administrative duties. He was Interim Associate Dean of Undergraduate Studies, Assistant Dean for IT Programs, Associate Dean of Undergraduate Studies, Assistant Dean for Enrollment Management, Assistant Dean for IT Programs, and then AIT Chair. During his tenure as chair, some of Dr. Gantz's major accomplishments included promoting recognition at the school level for outstanding faculty and students and assembling an efficient and collegial departmental staff.

Towards the end of the dinner, faculty members Mr. Peter Farrell, Dr. Aditya Johri, Mr. Mike Lyons, and Mr. Ken Santucci gave a farewell toast. Mr. Farrell said, "While I will miss Don Gantz as Chair of the Applied IT Department, I will miss Don Gantz much more as my friend."

The event culminated with a farewell speech by Dr. Gantz where he thanked and acknowledged the faculty and staff for their contributions during his years of service. The dinner ended with closing remarks by Dr. Bruno.

VB Profiles

- Dashboard
- Market Profiles
- Events
-

- IOT Revolution LandscapeConnected Vehicles LandscapeAPI LandscapeAutonomous Vehicles LandscapeIOT Incubators LandscapeIntelligent Assistance and Bot Landscape

VB|Profiles

e.g. Search profiles and marl

Login Sign up

Login Sign up

Edit

Person

Donald Gantz

Emeritus Professor at Volgenau School of Engineering, George Mason University
Location: United States

Add to My Lists Claim

Share

Donald T. Gantz, PhD, is an Emeritus Professor, Founding Chair of the Department of Applied Information Technology, and past Dean of the Volgenau School of Engineering at George Mason University. Dr. Gantz has developed cutting edge methodologies for the quantification and analysis of handwriting and has applied these methodologies to multi-language document exploitation and biometric identification, including matching software and scoring algorithms for fingerprint identification. He helped pioneer the use of statistical and geographic information systems to analyze the relationship between disease and socioeconomic factors, as well as transportation demand management and traffic mitigation. He has lectured on methods for surveillance systems to detect levels of infection due to a natural epidemic or bioterrorism threat, and is one of the top experts used by FBI in validating fingerprint analysis.

2
Companies in Career
N/A
Related Markets
N/A
Colleagues
N/A
Related Investments

Alias

N/A

Industry

Educational Services

Tags

N/A

Topics of Influence

N/A
N/A

- Career
- Achievements
- Investments
- Related People

Edit
View all

Donald GantzCareer (2)

Intensity Analytics

Advisor

Volgenau School of Engineering, George Mason University

Emeritus Professor

Competencies

- Other

Edit

Donald GantzEducation

Add Education

No education has been recorded for Donald Gantz

Edit

Donald GantzAchievements and Recognitions

Add Milestone

No milestones has been recorded for Donald Gantz

Cisco, kloudtrack® and George Mason University Team to Deliver Innovation Sandbox for Government and Education Sectors

Sandbox Centers Create 'Proving Grounds' Where Government and Commercial Organizations Can Test Innovation Viability for Real-World Use-Cases and Bottom-Line Results

#GovFest, WASHINGTON DC, (September 25, 2017) – kloudtrack® a leader in cloud computing technologies and solution services for sensitive data, process management and Governance, Risk and Compliance (GRC) today announced a teaming relationship with George Mason University (GMU) and Cisco Systems (NASDAQ: CSCO) to establish the first *Innovation Sandbox*™ exchange catering to the innovation and technology road-mapping needs of government/public sector and education organizations.

Dr. Don Gantz (Center), Chair of the Applied IT Dept. of GMU's Volgenau School of Engineering, joins kloudtrack CEO | president Michael Binko (right) and GMU's Brett Berlin in announcing the first Innovation Sandbox for Public Sector/Government and Education communities at #GovFest. More than 1400 government IT decision makers and industry representatives attended the first of its king festival-type event in Washington DC.

This *Innovation Sandbox*, which will be located on GMU's Fairfax, VA campus just outside Washington DC, is being developed jointly by kloudtrack, Cisco and GMU's Department of Applied IT which is part of the Volgenau School of Engineering. This *Innovation Sandbox* will allow organizations to test a variety of innovative technologies and/or methodologies in a "sandbox" environment.

As with all *Innovation Sandbox* environments, the GMU sandbox will be provisioned in-tandem with selected partners that offer technologies and solutions in a variety of innovation categories (examples follow):

- Cloud Computing | Cloud Migration | Cloud 2.0™

Case Studies
Press Releases
Recent Articles

- Innovation Roadmapping
- Cyber | InfoSecurity
- Data Driven Innovation™ (Data Analytics | Big Data | Reporting)
- Mobility (BYOD | Cross-platform | Device Independence)
- Platform Integration | Legacy Data Migration | Portability
- Hosting (IaaS, SaaS, PaaS | Co-Location | Virtualization)
- Secure Collaboration | Safe Sharing
- Business Process Management | Workflow
- Solution Services (IT Assessments, ROI Metrics, Benchmarking, Cloud Broker, Certifications)

"Organizations in public sector, education as well as commercial ranks have become very risk-minded with regards to IT strategies and budgets," stated Donald T. Gantz, PhD and Chair of the GMU Applied IT Department. "No matter if the risk is mission-oriented or a result of the sequested procurement environment the results are the same. IT decision makers are becoming very selective and justifiably taking a use-case approach to purchases. The Innovation Sandbox concept allows these communities to bring their mission needs to a team of vendors and project analyst specialists with a goal to assess, identify, prioritize and roadmap innovations that meet critical objectives," added Gantz.

Now, more than ever, the driving factors behind IT deployments have become more economic than technical. The advent of innovations such as cloud, virtualization and mobility mean the vendor landscape is much more diverse.

"With standards such as the NIST Cloud Reference Architecture and more recently FedRAMP driving easier portability, organizations can now consider best-of-breed offerings when considering and procuring IT solutions," stated Jan M. Levine, founder | CTO for kloudtrack®. "The *Innovation Sandbox* environments are designed to give organizations the opportunity to test waters with a *Start Small … and Scale*™ approach with keen sensitivity to limited resources and measurable results," added Levine.

Deployments are expected to run on sandbox systems for a period of weeks or months at small costs. During these test-bed deployments teams will evaluate the viability and return on investment (ROI) for selected innovations. Then, decisions can be made about contracting and scaling the deployments with a solid base of results-oriented information.

"Organizations will be able to utilize the *Innovation Sandbox* to test sample-sets of data, users, processes and measure value metrics against real-world use case scenarios," stated Dan Kent, CTO- US Public Sector, Cisco Systems. "This concept and approach is consistent with Cisco's efforts to cultivate vendor ecosystems. With tightened budgets and a much more diverse environment of innovations available, CIOs have a rare opportunity to test the waters. The challenges we've heard have centered around finding vendors/technologies that work well together and are able to be deployed quickly on a small-scale with limited resources. These *Innovation Sandbox* environments offer an on-ramp to next-generation technologies that could have a dramatic impact on bottom-line results," added Kent.

The *Innovation Sandbox* approach works in a wide variety of industries. As announced previously, Cisco and kloudtrack have identified the following sectors for initial *Innovation Sandbox* efforts:

- Public Sector | Government | Education (federal, state, municipal, local)
- Financial Services (securities-insurance, banking, capital markets, deal-flow)
- Cyber Security
- Healthcare | Medical | Life Sciences | Pharma
- Manufacturing (3-D printing, rapid prototyping, operational efficiency, logistics)
- Energy | Utilities | Telecom
- Property | Asset Management
- Retail | Consumer

By the end of 2013, three *Innovation Sandbox* environments are expected to be operational with a broad community of partner organizations serving diverse customer needs. Invited partner organizations will include academia (universities, training centers), technology/solution vendors, consulting/professional service providers, system integrators, incubators/accelerators, economic development agencies (EDAs) as well as individual thought-leaders.

At this time, kloudtrack and Cisco are seeking innovation partners to jointly map sophisticated innovations and offerings to existing missions, task orders and new procurements in the industry sectors and innovation categories highlighted above. Interested parties should contact Nicholas Mumma at 240-499-3844 or nmumma@kloudtrack.com.

###

About kloudtrack®

Headquartered in Annapolis, MD kloudtrack® is a Software-as-a-Service and cloud computing (SaaS|Cloud) company that integrates best-practice management with cybersecurity and audit-trails that place a premium on monitoring users, transparency and collaborations as they relate to sensitive data assets and processes.

kloudtrack® is one of the first technology and solution providers to address specific international, federal, state and industry-imposed regulatory compliance standards in SaaS|Cloud environments across multiple vertical industry sectors such as fin-serv (SEC/FINRA/Basel Accords), health-medical (HIPAA/CCHIT/REMS) and government (FIPS/FISMA/FedRAMP) among others. This understanding and integration of GRC factors and certifications set kloudtrack® technologies, solutions and IP apart in the market. Learn more at www.kloudtrack.com.

About Cisco

Cisco (NASDAQ: CSCO) is the worldwide leader in IT that helps companies seize the opportunities of tomorrow by proving that amazing things can happen when you connect the previously unconnected. For ongoing news, please go to http://thenetwork.cisco.com.

About George Mason University (Where Innovation is Tradition)

Located in Northern Virginia near Washington, D.C., George Mason University is an innovative, entrepreneurial institution with global distinction in a range of academic fields. Within Virginia's largest university, the GMU Volgenau School of Engineering is home to more than 5,000 undergraduate and graduate students and consists of seven departments: Applied Information Technology; Bioengineering; Sid and Reva Dewberry Department of Civil, Environmental and Infrastructure Engineering; Computer Science; Electrical and Computer Engineering; Statistics; and Systems Engineering and Operations Research.

On October 10, 2017, **posted in:** Press Releases by mbinko

Comments are closed.

ACM DIGITAL LIBRARY

Donald T Gantz

Search for Donald T Gantz's work

Donald T Gantz

George Mason University

Author's Profile

Searched The ACM Guide to Computing Literature (3,201,853 records) |
Limit your search to The ACM Full-Text Collection (785,017 records)

RESULTS

Showing 1 - 3 of 3 Results

Feedback

ORT BY: LATEST

An approach to a capstone curriculum

Robert T. Quinn, Donald T. Gantz

SIGITE '09: Proceedings of the 10th ACM conference on SIG-information technology education

• October 2009,pp 150–154 • https://doi.org/10.1145/1631728.1631768

In this paper, we describe the required Senior Capstone Sequence for the George Mason University Bachelor of Science in Information Technology. The mission of the Bachelor of Science in Information Technology program is to prepare students to employ ...

RESEARCH-ARTICLE

Combining academic studies with it certifications: becoming a cisco regional academy

Louis R. D'Alessandro, Donald T. Gantz

SIGITE '09: Proceedings of the 10th ACM conference on SIG-information technology education

• October 2009,pp 209–214 • https://doi.org/10.1145/1631728.1631796

Influenced by unprecedented growth in the IT industry, especially in the Washington, DC metropolitan

ACM DIGITAL LIBRARY

Pictographic matching: a graph-based approach towards a language independent document exploitation platform

Mark A. Walch, Donald T. Gantz

HDP '04: Proceedings of the 1st ACM workshop on Hardcopy document processing
• November 2004,pp 53–62 • https://doi.org/10.1145/1031442.1031451

In this paper, we introduce the concept of Pictographic Matching as a tool for document exploitation across multiple languages. The primary technology supporting Pictographic Matching uses graph-based pattern matching to detect the signature of ...

1

Get Access

Mark Walch is the President of Sciometrics LLC, where he brings more than 25 years of technical and managerial experience. In this role, Mr. Walch oversees development of a variety of technologies related to the efficient and effective capture of data from images.

Mr. Walch is the principal architect of Pictographic pattern-matching methods, and has developed innovative Optical Character Recognition techniques that have been used to read hand print and cursive script for the U.S. Postal Service. Derivatives of these techniques have been used successfully to register independent vector data sets for upgrading map data on behalf of the U.S. Census Bureau.

Mr. Walch has also developed several automated techniques for accurate and cost-effective data capture from handwritten and printed forms, and pioneered the concept of “directed workflow”—a method for streamlining the way human operators review large quantities of data.

Mr. Walch holds advanced degrees from the University of Michigan and Yale University.

Mark Walch

President at Sciometrics LLC

Washington D.C. Metro Area · 181 connections · Contact info

Sciometric

Yale Unive

About

I am the President of Sciometrics LLC (formerly the Gannon Technologies Group) and the principal ar
Sciometrics' Graph-based Pattern Matching methods that can be used to extract very hard-to-get co
documents, photographs and signal data. Applications using Graph-based Pattern Matching have be

Experience

President

Sciometrics LLC

Dec 2013 – Present · 6 yrs 4 mos

Herndon, Virginia

AFI Sleuth AFIS

Products

President
Gannon Technologies
2000 – Present · 20 yrs
Herndon, Virginia

SCIOMETRICS FLASH ID

FLASH ID: Writer Identification

Chairman/CTO
Nortec LLC
1994 – 1999 · 5 yrs

Partner
Michael Fieldman & Partners
1982 – 1995 · 13 yrs

Specialized in Hospit and Health Facility...

Education

Yale University
MPH, Health Services Administration
1975 – 1977

University of Michigan
MArch, Architecture

ne of your email addresses. Please **update or confirm** your email. **More info**

highly skilled at this

Proposal Writing · 8

Endorsed by **Paul Maguire, who is highly skille** leag
this

Intelligence · 6

Les Gay and 5 connections have given endorsements

S

Recommendations

Received (0) Given (1)

Terry Pabst
Principal at Pabst Law Offices
January 5, 2011, Mark reported directly to Terry

Terry cam rnat
recently a y in
transform le" I
support sl tion
position r chn

Interests

Yale University Careers
155,673 followers

Bloomberg News
3,022,930 followers

University of Michigan
484,116 followers

University of Detroit Je

Messaging 1

Daniel Gantz is a senior software engineer at Sciometrics LLC. Daniel has worked for this group since graduating The University of Virginia's school of engineering in 2008 with a bachelor of science in computer science.

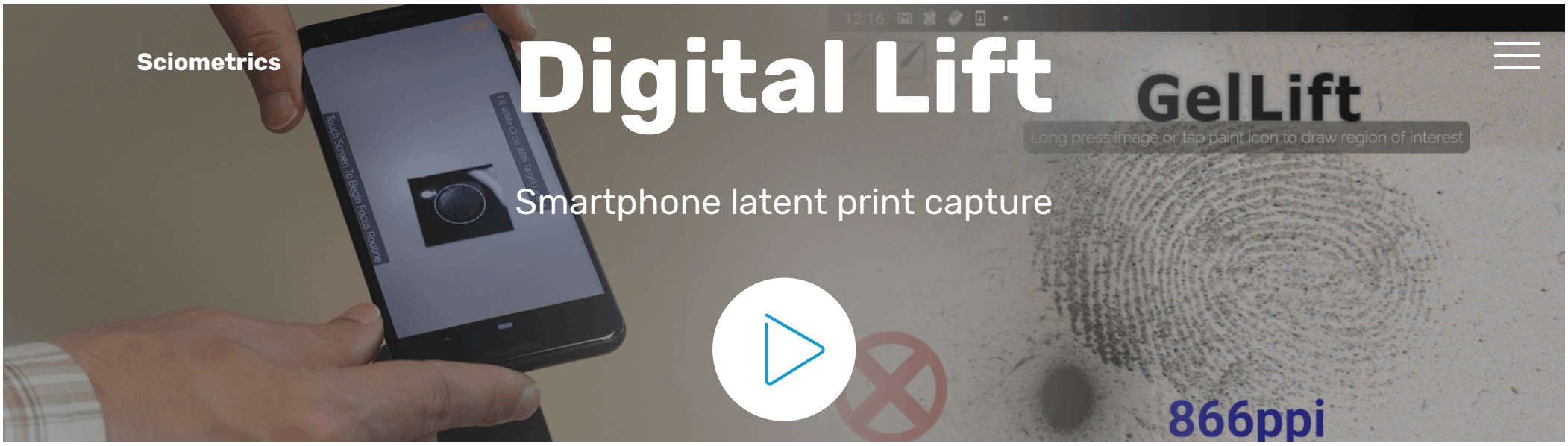
Daniel has helped as well as lead development for a number of different biometric systems from their R&D inception to their evolution into industrial level applications. The nature of these biometric systems include latent fingerprint, tool-mark and handwriting identification.

Rich Smith is Vice President of Products & Services for Sciometrics, LLC. In this role, Mr. Smith oversees the development of all software applications and delivery of services to Sciometrics's clients. Mr. Smith joined the company in January 2011 and brings with him over 25 years of experience developing language-based software applications. Most recently, Mr. Smith was Vice President of Engineering for Janya, Inc. Mr. Smith holds a B.S. in Computer Science from Canisius College.

Robyn Felix oversees the company’s financial and general operations, including human resources and facilities. Ms. Felix is a key contributor to both short- and long-term organizational development of the company, and also serves as the company Facilities Security Officer. Ms. Felix has over 12 years of experience in Government Contracting in the Intelligence Community, including working in Iraq and Afghanistan supporting the Global War on Terrorism. Her previous experience includes Deputy Director of Logistics and Deployment Operations supporting various INSCOM Language and Counter Intelligence contacts, as well as managing a CI contract for the 310th Military Intelligence Battalion at the 902nd . Ms. Felix has also worked as a Senior Management Analyst supporting the Missile Defense Agency’s Base Realignment and Closure (BRAC) relocation efforts, Logistics Officer for the MASINT Special Airborne Projects division at the Defense Intelligence Agency, and also managing the day to day operations at the Washington Planning Center; the US Navy's Cruise Missile Projects programs office testing facility for PMA-281 - Tomahawk mission planning.

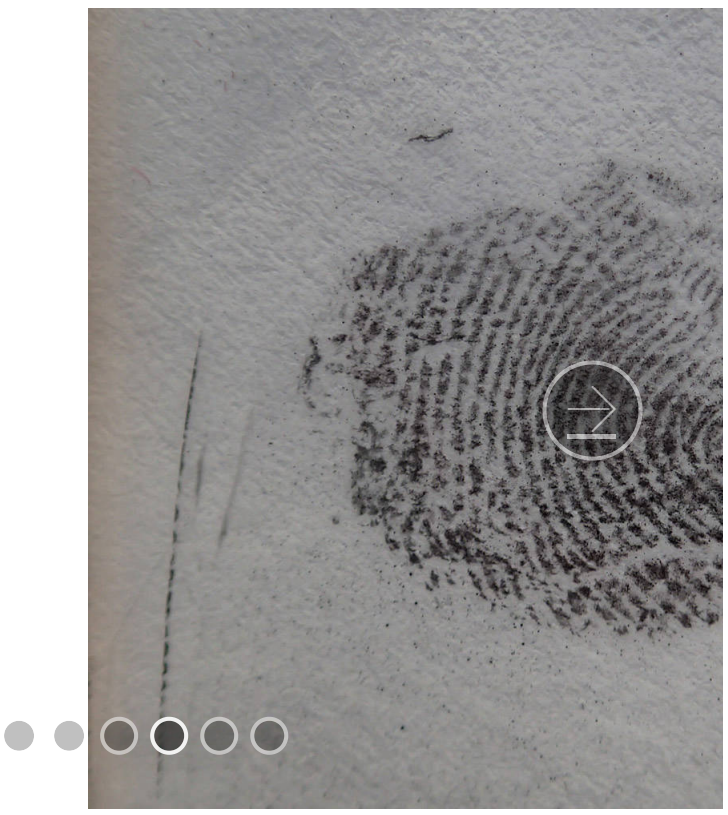
Past experience also includes living and working in Brazil for the US State Department.

Ms. Felix attended the University of Memphis.



Digital Lift is a mobile app that empowers smartphones to become latent print capture devices. Digital Lift leverages the incredible cameras found on modern smartphones. Digital Lift can capture images from traditional lift cards, gellifters, and, in some cases, undeveloped latents.

- **Easy to use** - Digital Liftuses a simple user interface with large icons and intuitive operation. Minimal training is required to use Digital Lift.
- **Triage from the crime scene**- Digital Lift allows an investigator to submit latents for evaluation immediately. The app records the time, location, and identity of the investigator.
- **Convenient** - Digital Lift runs on a device that is likely already being carried. No additional device to carry and no added weight.



CONTACT FORM

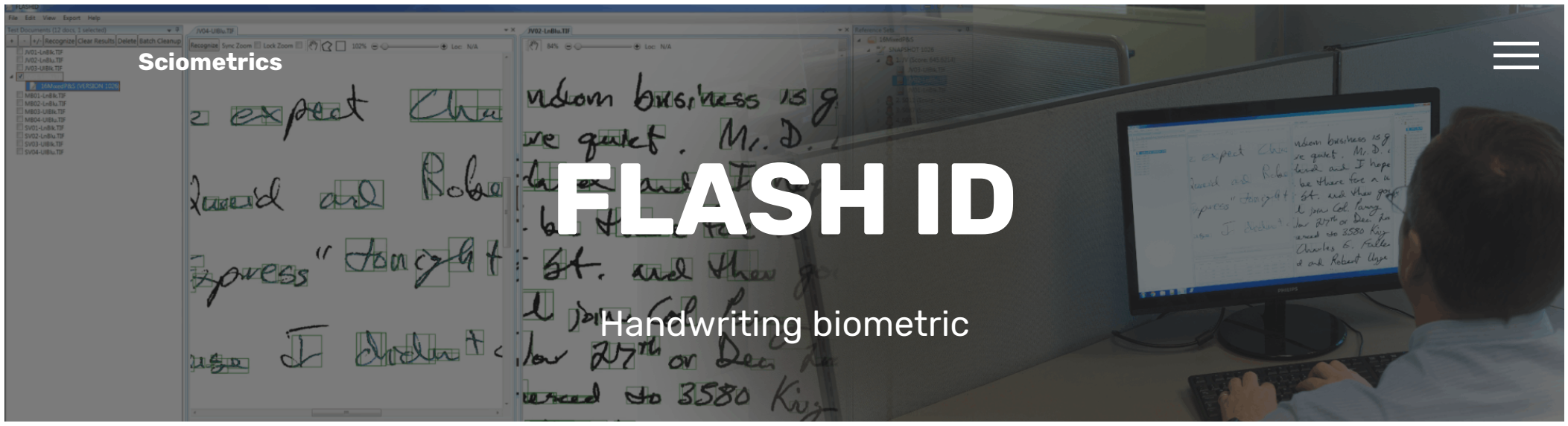
Contact us for more information or to request a demonstration.

Name

Email

Phone

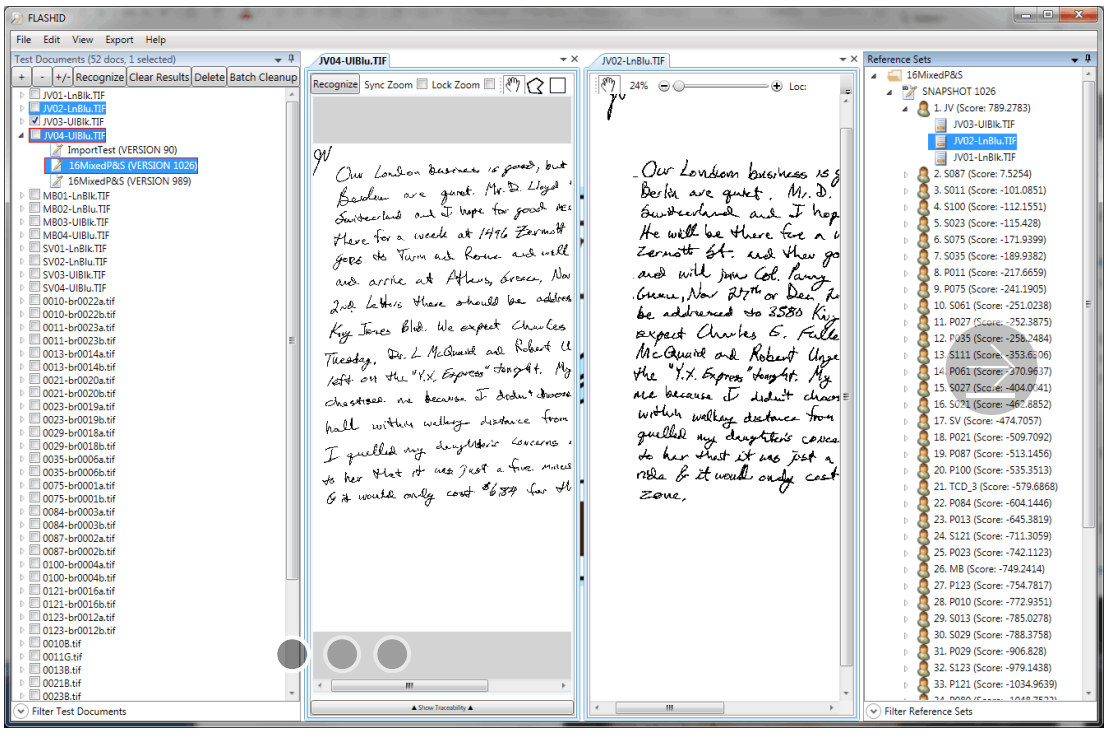
Message

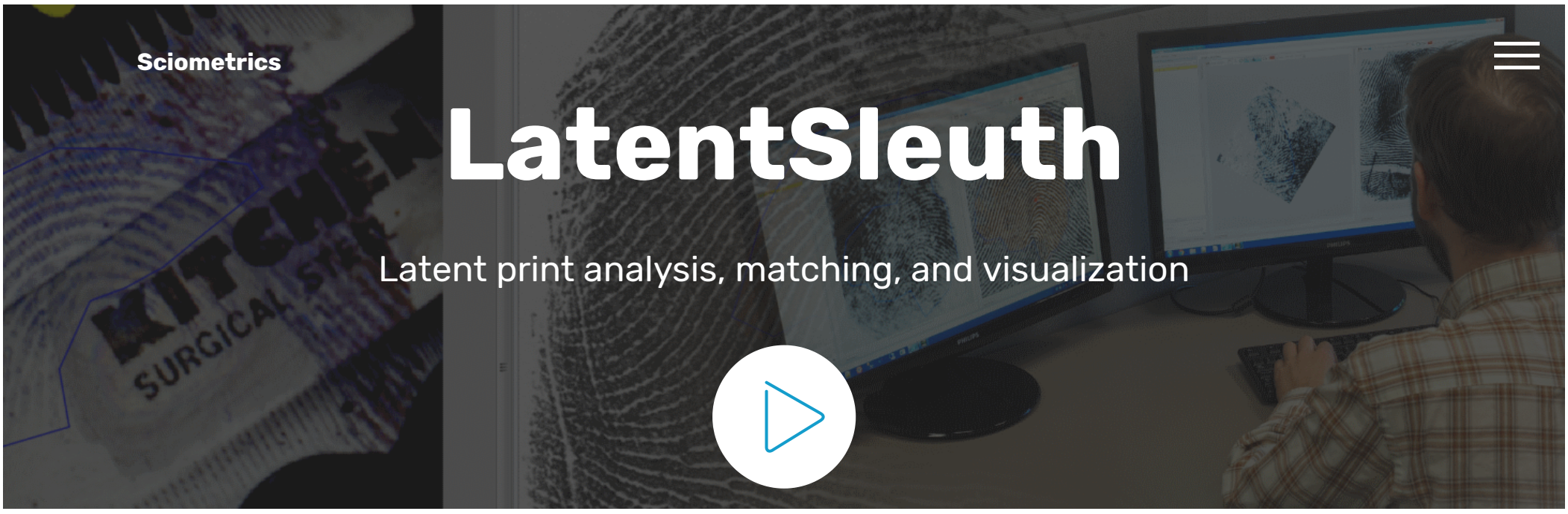


FLASH ID is a unique handwriting analysis, matching and visualization tool designed to assist law enforcement, military and intelligence professionals in the identification of suspects, criminals, and terrorists. This powerful tool for document comparison and evaluation is designed to assist examiners in comparing questioned documents against known writer document samples, and previously received questioned documents.

- **Easy to use** - FLASH ID does not require any manual markup or tracing of documents. Simply scan and compare.
- **Sophisticated algorithms** - Graph matching algorithms at the core of the technology compare the graphemes, first, by their topology and second, by their geometric features. Geometric features measure the shapes of curves, angles and distances among graph components. FLASH ID computes nearly 200 measurements of the sample writing.
- **Works on limited data** - FLASH ID is able to make comparisons with questioned documents as small as ¼ page of writing. Reference samples can be effectively matched with as little as one page of writing..
- **Visualization tools** - FLASH ID offers a comprehensive suite of visualization tools to assist the examiner in making their determination, including side-by-side viewing and an innovative “heat map” of writing similarity.
- **Image cleanup tools** - FLASH ID includes sophisticated image cleanup tools, including deskewing, noise removal, line removal, blob removal, form removal, color removal, contrast enhancement, smoothing, blurring, and B&W conversion.

Click here for more information.

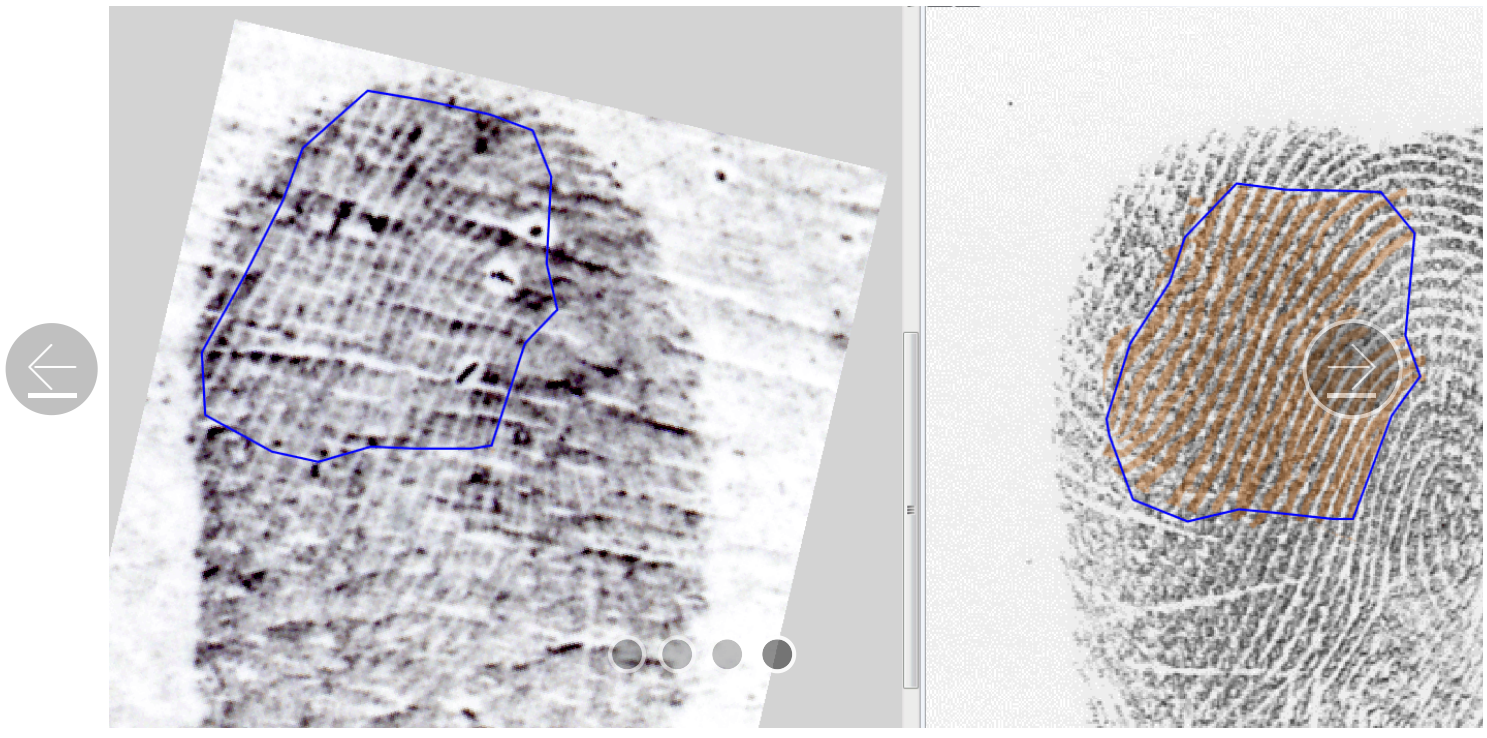




LatentSleuth uses a novel method for latent-to-reference matching that fully leverages ridge skeletons, is rotationally independent, and effectively handles distortion caused by “elasticity of skin” and other factors. The end result is a mapping between the latent region of interest (ROI) to a region in each reference print being compared, which allows for allows for comparisons based on a quantitative evaluation of the similarity of the latent and reference prints .

- **Easy to use** - LatentSleuth provides step-by-step guidance and a complete suite of tools for latent print prep, reference print ingestion, and annotation.
- **Match more latents** - LatentSleuth's powerful ridge skeleton approach significantly expands available comparison information for even minimal minutiae point and small size latent prints. Effective matching of smaller and/or distorted prints (including perspective distortion) , prints lacking cores and deltas, prints lacking known orientation, prints with ambiguous anatomical origin.
- **Search all types of reference prints** - Flats, rolls, major case impressions, ad hoc reference prints, etc.
- **Enhanced Visualization** - LatentSleuth produces very accurate overlays of the latent print onto each reference print, providing precise placement information for the LPE. These overlays can be exported for use in case documentation.
- **Save time** - The information provided by the overlay eliminates much of the time spent by the LPE searching for the correct orientation and placement of the latent print during the comparison phase of the ACE-V process.

Click here to learn more.



What our customers say

I was skeptical about the software after seeing my first demonstration. I knew that computerized latent print matching algorithms are limited by the availability of encodable minutiae. That was before I discovered Sciometrics' LatentSleuth. For the first time I'm seeing a software program match fingerprints using ridge tracings in a fashion similar to a human examiner. Actually in some ways Sciometrics is superior to the human examiner. The software doesn't need to know the orientation of the latent print or the anatomical source. It methodically searches all friction ridge detail for a match (and it doesn't get tired).

***Eric Sahota**, Las Vegas Metro PD Forensics Lab*

CONTACT FORM

Contact us for more information or to request a demonstration or free trial.

Name

Email

Phone

Message

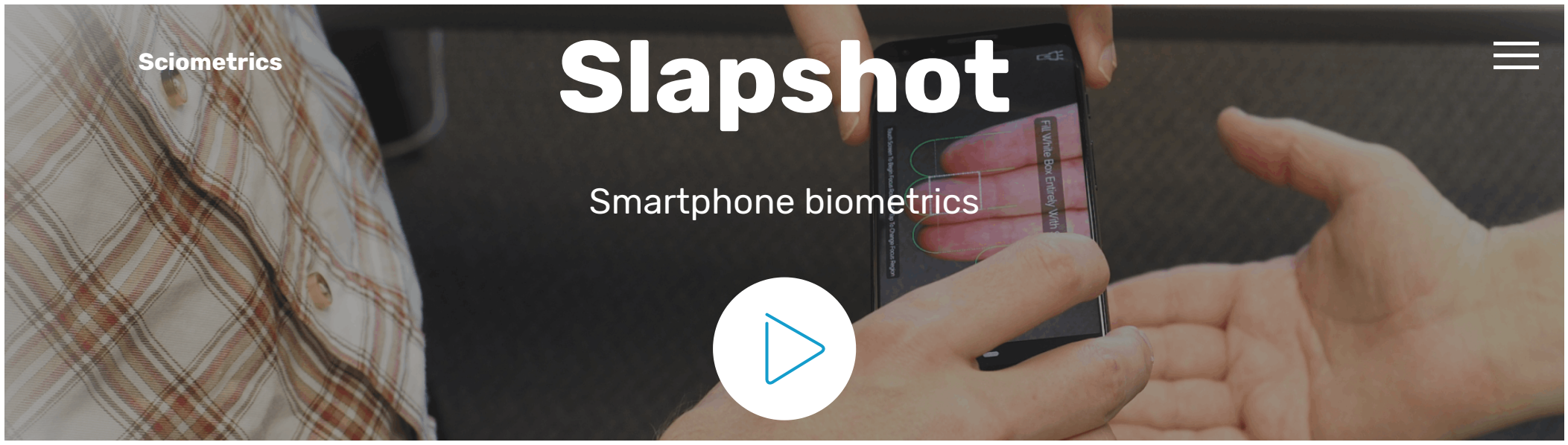
SEND FORM

Address

Sciometrics, LLC
14150 Parkeast Circle, Suite 140
Chantilly, VA 20151

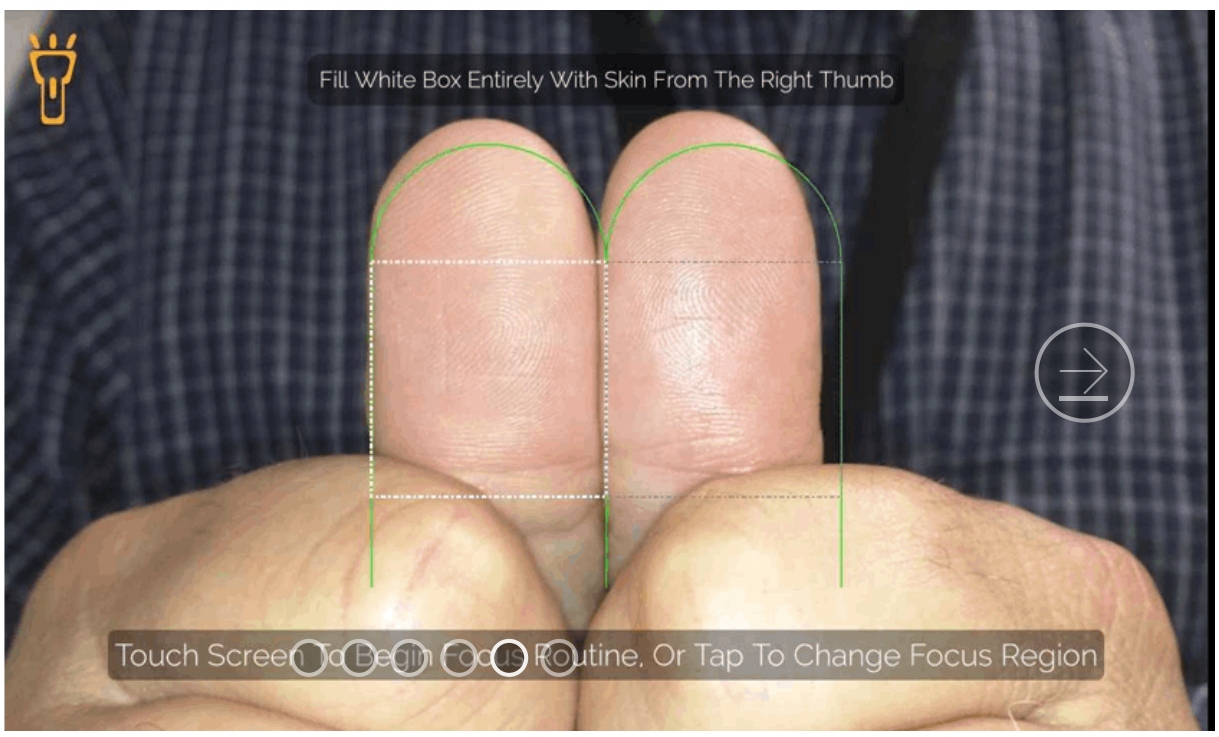
Contacts

Email:
info@sciometrics.com
Phone: +1 (703) 793-2282
Fax: +1 (703) 793-2280



Slapshot is a patented mobile app that empowers smartphones to become biometric capture and detection devices. Slapshot leverages commercial off-the-shelf hardware without special modifications or peripheral devices. Slapshot captures fingerprint biometric information by photographing the surfaces of the fingers and thumbs. Ten finger capture can be accomplished in less than a minute.

- **Easy to use** - Slapshot uses a simple user interface with large icons and intuitive operation. Minimal training is required to use Slapshot.
- **Multi-modal biometrics** - Slapshot supports fingerprint and facial recognition using state of the art matching algorithms.
- **Everyone is a sensor** - Slapshot runs on a device that is likely already being carried. No additional device to carry and no added weight.
- **Multiple modes of operation** - Slapshot supports on device matching, or submission to an AFIS system for matching against larger databases.
- **Capture additional information** - Slapshot can also capture additional information, including voice recording, scars, marks, and tattoos.



CONTACT FORM

Contact us for more information or to request a demonstration or free trial.

Name

Email

Phone

Hoyos Labs Joins National Research Program to Inform Creation of Contactless Fingerprint Technologies for Federal Agencies

NEWS PROVIDED BY
Hoyos Labs →
Feb 16, 2016, 10:00 ET

NEW YORK, Feb. 16, 2016 /PRNewswire/ -- Hoyos Labs, Inc., the global leader in biometric authentication, today announces a partnership with the National Institute of Standards and Technology (NIST) to develop new methodologies for measuring the image fidelity of contactless fingerprint capture devices, to support evaluation of these devices for future inclusion on the U.S. government's Certified Product Lists.

From local police stations to government agencies, today's law enforcement organizations utilize fingerprinting for a variety of screening and identification purposes. The most commonly used scanning devices digitally capture a fingerprint pressed on a glass surface with live-scan technology.

However, the disadvantages to this process include added time – particularly when capturing rolled prints – as well as hygienic concerns. Law enforcement officials are now exploring contactless fingerprint readers as a more effective and efficient alternative.

Following up on earlier programs of the Department of Homeland Security (DHS) and National Institute of Justice (NIJ) to evaluate contactless fingerprint capture technology, NIST – under sponsorship of the Federal Bureau of Investigation's (FBI) Biometric Center of Excellence – initiated its Contactless Fingerprint Capture Device Measurement Research Program. The NIST program requires direct interaction with emerging technologies to address the following industry challenges: data format standards and best practices, development of methods for certification testing, as well as interoperability assessment with legacy contact-based devices.

Through a Cooperative Research and Development Agreement (CRADA) with NIST, Hoyos Labs has offered up the company's touchless mobile biometric application, called 4F, for blind pilot testing. Hoyos Labs joins NIST's other CRADA partners, including 3M Company and MorphoTrack, in contributing its touchless fingerprint acquisition technology to ensure the performance metrics emerging from the collaboration will be suitable for a broad range of devices.

"Our participation in NIST's research program is mutually beneficial," said Hector Hoyos, founder and CEO of Hoyos Labs. "Not only are we playing a critical role in building a worldwide standard for testing contactless fingerprint scanners, but NIST researchers are also providing us with new scenarios that we have been previously unable to test using human subjects – to help us better understand any limitations to our product and make the necessary improvements."

Hoyos Labs' 4F identification technology simultaneously captures all four fingerprints by leveraging high-resolution camera capability and flash lighting mode on modern smartphones. Hoyos Labs conducted tests in accordance with the FBI's 2D image quality standards, which demonstrated 4F is capable of producing images that exhibit outstanding geometric fidelity, sharpness, detail rendition and grayscale dynamic range with low noise characteristics.

4F eliminates many of the distorting variables present with today's hardware scanners, such as elastic deformation when pressing a finger on the scanner surface, and delivers a biometric template that is a precise copy of the captured finger impression. For the protection sector, in particular, 4F presents law enforcement agents with the opportunity to cross-reference an individual's fingerprints against existing federal databases and confirm a person's identity within minutes.

Under the CRADA, results of the program's tests will be shared with Hoyos Labs scientists to be considered in future 4F technology research and development.

For more information on Hoyos Labs and its portfolio of biometric authentication solutions, please visit: http://www.hoyoslabs.com/.

About Hoyos Labs

Hoyos Labs is the leading biometrics-based end-to-end authentication infrastructure company, and the only company that can guarantee identity. The goal of Hoyos Labs is to develop and deploy enterprise and consumer identity assertion technology platforms that will conveniently and securely address the identity assertion and authentication challenges we face today. Hoyos Labs currently has offices in New York, Boston, San Juan (Puerto Rico), Oxford (United Kingdom), Bucharest (Romania), Timisoara (Romania) and Beijing (China).

SOURCE Hoyos Labs

Related Links

http://www.hoyoslabs.com

FINOVATE BLOG

Tracking fintech, banking & financial services innovations since 1994

Back to Blog (https://finovate.com//blog/)

Hoyos Labs Rebrands as Veridium, Launches New Biometric Authentication Solution

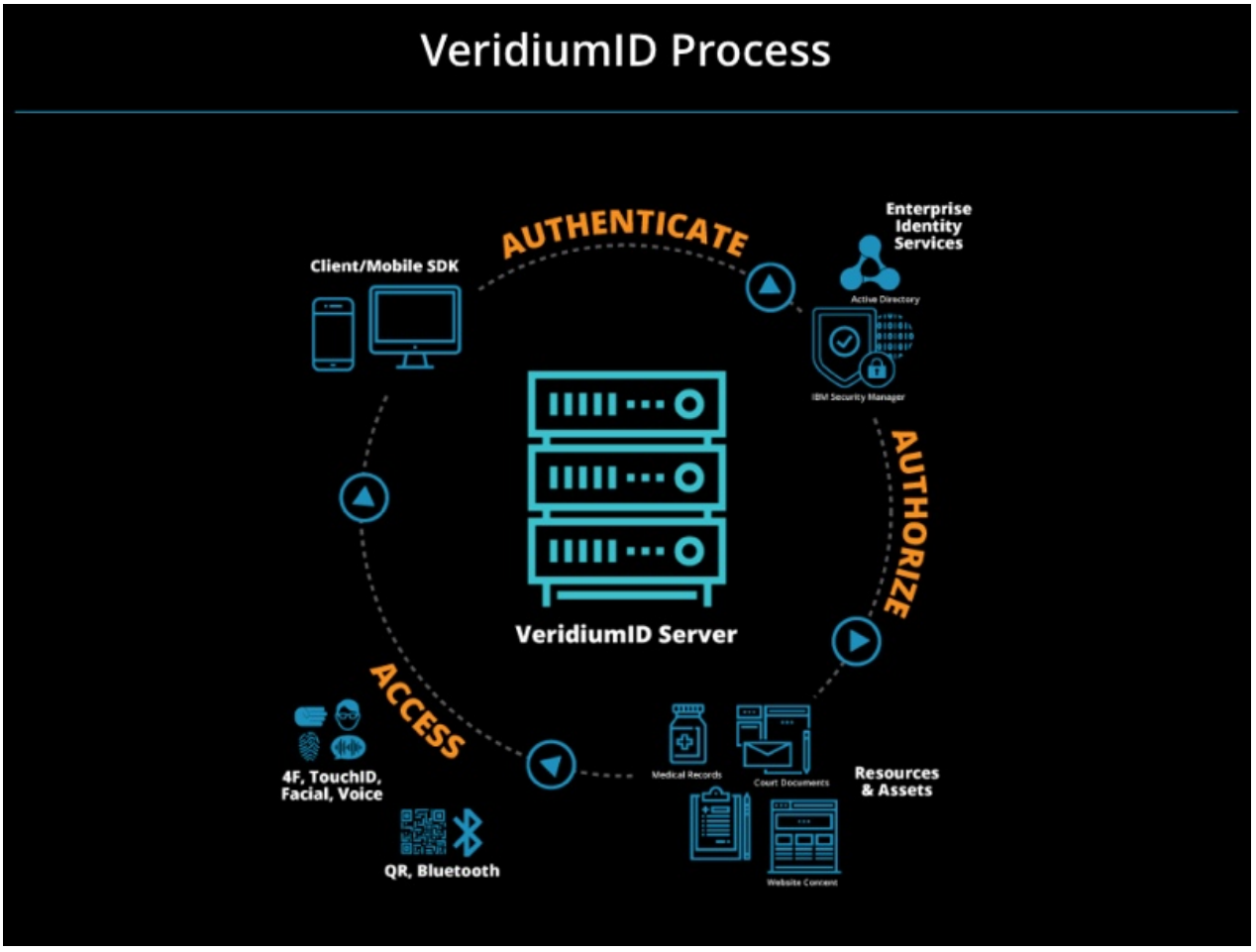
By David Penn (https://finovate.com/author/david/) Posted on September 19, 2016 (https://finovate.com/hoyos-labs-rebrands/)

Share:

(http://www.veridiumid.com)

With the launch of its new, end-to-end biometric authentication solution—*VeridiumID*—the security innovator that demoed its technology at FinovateFall (http://finovate.com/videos/finovatefall-2014-hoyos-labs/) 2014 as **Hoyos Labs** is now **Veridium** (http://www.veridiumid.com).

Company COO Todd Shollenbarger said, “We’re going to replace what you know (passwords) with what you are (biometrics) to safeguard enterprises from fraud and data breaches.” Pointing to the issues that typically inhibit broader adoption of biometric technology, such as high cost and technical complexity, Shollenbarger added that Veridium’s use of SaaS delivery for the backend—and the user’s smartphone to handle the image-capture for authentication—helps reduce both cost and complexity. “We’re making biometric authentication a solution of today, not tomorrow,” he said.



VeridiumID gives businesses the ability to leave password-based authentication and instead incorporate biometric methods using face, voice, or fingerprint. The combination of biometrics and multiple authentication factors in its new solution gives customers of financial services companies the ability to access their accounts and transact without cards and PIN codes, eliminates the need for insurance cards as an authentication tool in healthcare, and helps governments manage physical access in a more efficient and less cumbersome and costly way.

A server-based solution working with a mobile app and a front-end, mobile SDK, the technology uses virtual cryptography to break biometric data into pieces. Storing and distributing the data this way makes it harder for hackers to steal or compromise an entire biometric piece of data, such as a fingerprint. VeridiumID uses liveness detection, accommodates a number of plug-and-play biometric libraries, and provides support for an Active Directory or a FIDO authenticator. Deployable in the cloud or on-site, VeridiumID requires no additional hardware.

The goal of the rebrand from Hoyos Labs to Veridium is to focus the company's efforts in four key areas: financial services, healthcare, government, and the Global 2000. Veridium says it seeks to leverage its eight patents (with another 34 pending), as well as numerous R&D breakthroughs, to build an open standard for authentication that will defend the enterprise against data breaches today, while remaining flexible enough to "ensure continual iteration and enable the solutions to evolve at the forefront of the industry."

Founded in 2013 and headquartered in New York City, Veridium demonstrated the mobile app of its identity-assertion platform at FinovateFall 2014 (http://finovate.com/videos/finovatefall-2014-hoyos-labs/). The company also has offices in the U.K. (Oxford and London), as well as in Boston, Massachusetts, and Timisoara, Romania.

Categories: Hoyos Labs (https://finovate.com/category/hoyos-labs/), Veridium (https://finovate.com/category/veridium/)

Search

Upcoming Events

Subscribe by Email

Please enter your email address

Want us to cover your news?

Contact research@finovate.com (mailto:research@finovate.com)

Follow us on Twitter

(https://twitter.com/intent/follow?screen_name=finovate)

44829 Followers

Popular Posts

6 Banks Making Saving as Easy as Spending (https://finovate.com/6-banks-making-saving-as-easy-as-spending/)

Are U.S. Banks Leaning Toward "Closed Banking?" (https://finovate.com/are-u-s-banks-leaning-toward-closed-banking/)

FinovateEurope Best of Show Winners Announced (https://finovate.com/finovateeurope-best-of-show-winners-announced-2/)

Popular Demos

FinovateEurope 2020 – Fidel (https://finovate.com/videos/finovateeurope-2020-fidel/)

/ FinovateEurope, 2020

FinovateEurope 2020 – Altilia (https://finovate.com/videos/finovateeurope-2020-altilia/)

/ FinovateEurope, 2020

FinovateFall 2017 – Nvstr (https://finovate.com/videos/finovatefall-2017-nvstr/)

/ FinovateFall, 2017

(https://finovate.com/feed)
Latest Finovate blog posts (https://finovate.com/)

Nancy Giordano, Big Shifts, and How Exponential Technologies Will Change All (https://finovate.com/nancy-giordano-big-shifts-and-how-exponential-technologies-will-change-all/)

Airwallex Integrates with Xero to Help SMEs Reconcile Cross-Border Payments (https://finovate.com/airwallex-integrates-with-xero-to-help-smes-reconcile-cross-border-payments/)

Grab to Use Wirecard for Payment Processing (https://finovate.com/grab-to-use-wirecard-for-payment-processing/)

About

Contact Us (/contact/)

Who We Are (/who-we-are/)

Sponsor (/sponsor/)

Apply to Demo (/apply-to-demo/)

(https://twitter.com/finovate)

(https://www.facebook.com/Finovate-Conference-Series-1291423074220235/)

Finovate

FinovateFall (https://informaconnect.com/finovatefall/?vip_code=FKV2373FINOVATEREF&utm_source=FinovateWebReferral&utm_medium=web_listing&utm_campaFinovateWebReferral-

(https://www.linkedin.com/company/finovate-conference-series)

web_listing&utm_term=FinovateWebReferral&utm_content=FKV2373FINOVATEREF&tracker_id=FKV2373FINO

FinovateAwards (https://finance.knect365.com/finovate-industry-awards/)

FinovateAsia (https://informaconnect.com/finovateasia/?vip_code=FKV2374FINOVATEREF&utm_source=FinovateWebReferral&utm_medium=web_listing&utm_campaFinovateWebReferral-web_listing&utm_term=FinovateWebReferral&utm_content=FKV2374FINOVATEREF&tracker_id=FKV2374FINO

Demo Videos

Videos (https://finovate.com/videos/)

FinovateMiddleEast (https://finance.knect365.com/finovate-middle-east/)

FinovateEurope (https://informaconnect.com/finovateeurope/?vip_code=FKV2368FINOVATEREF&utm_source=FinovateWebReferral&utm_medium=web_listing&utm_campaFinovateWebReferral-web_listing&utm_term=FinovateWebReferral&utm_content=FKV2368FINOVATEREF&tracker_id=FKV2368FINO

FinovateSpring (https://informaconnect.com/finovatespring/?vip_code=FKV2369FINOVATEREF&utm_source=FinovateWebReferral&utm_medium=web_listing&utm_campaFinovateWebReferral-web_listing&utm_term=FinovateWebReferral&utm_content=FKV2369FINOVATEREF&tracker_id=FKV2369FINO

(https://informaconnect.com/)

Access Control and Time & Attendance

We promise secure access control to a variety of premises.

IDEMIA's innovative biometric terminals guarantee approved individuals a frictionless access to secured areas. Whether used alone or in combination with contactless cards, standalone or IP-networked, indoor or outdoor, our biometric access solutions are designed to be modular. IDEMIA's biometric time clocks provide accurate clock-in and clock-out records enabling organizations to increase workforce management efficiency.

GO TO BIOMETRIC DEVICES & AUTOMOTIVE MARKET

PRODUCT & SOLUTION

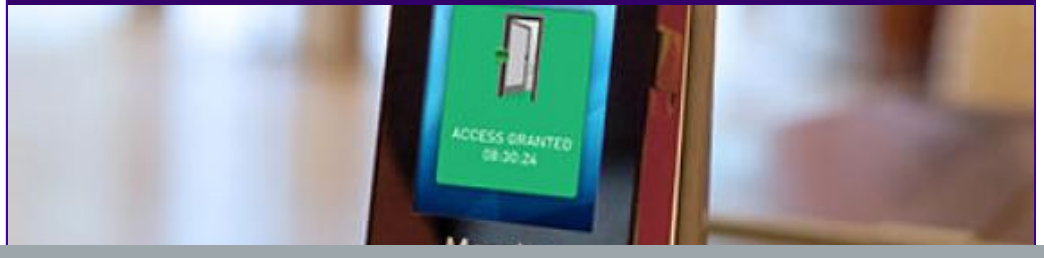
MORPHOWAVE COMPACT

Delivering frictionless access everywhere

PRODUCT & SOLUTION

MORPHOACCESS SIGMA LITE SERIES

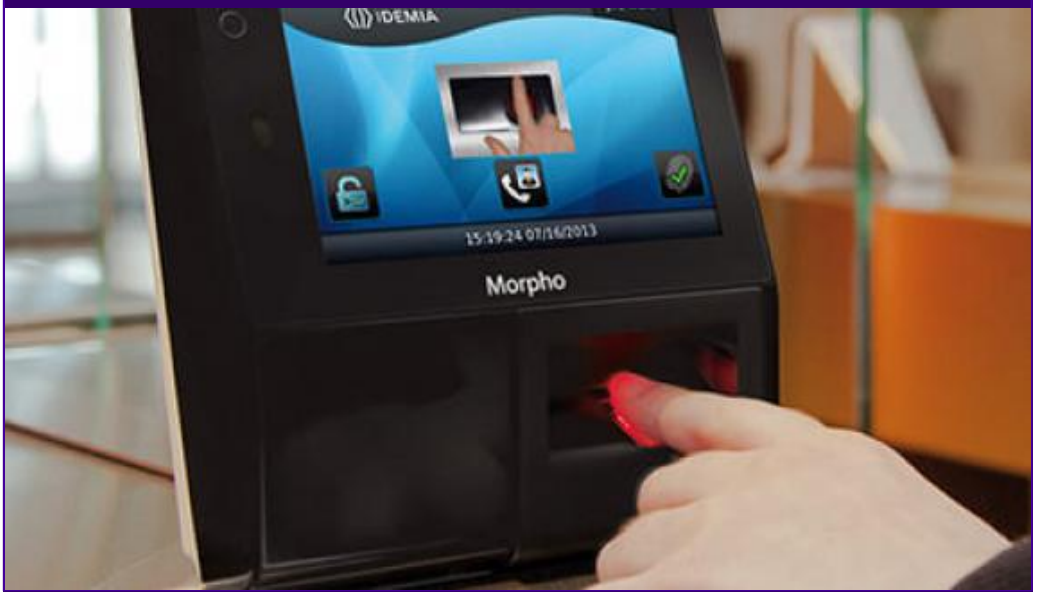
Providing world's #1 fingerprint identification in a slim design



PRODUCT & SOLUTION

MORPHOACCESS SIGMA SERIES

Configuring access & time transactions to your needs with versatile fingerprint terminals



PRODUCT & SOLUTION

MORPHOACCESS SIGMA EXTREME SERIES

Controlling access in the harshest outdoors conditions

PRODUCT & SOLUTION

MORPHOACCESS VP

Enhancing access control with hybrid finger vein and fingerprint technology

PRODUCT & SOLUTION

MORPHOTABLET 2

Managing mobile workforces with accurate biometric identification on the spot

PRODUCT & SOLUTION

MORPHOMANAGER

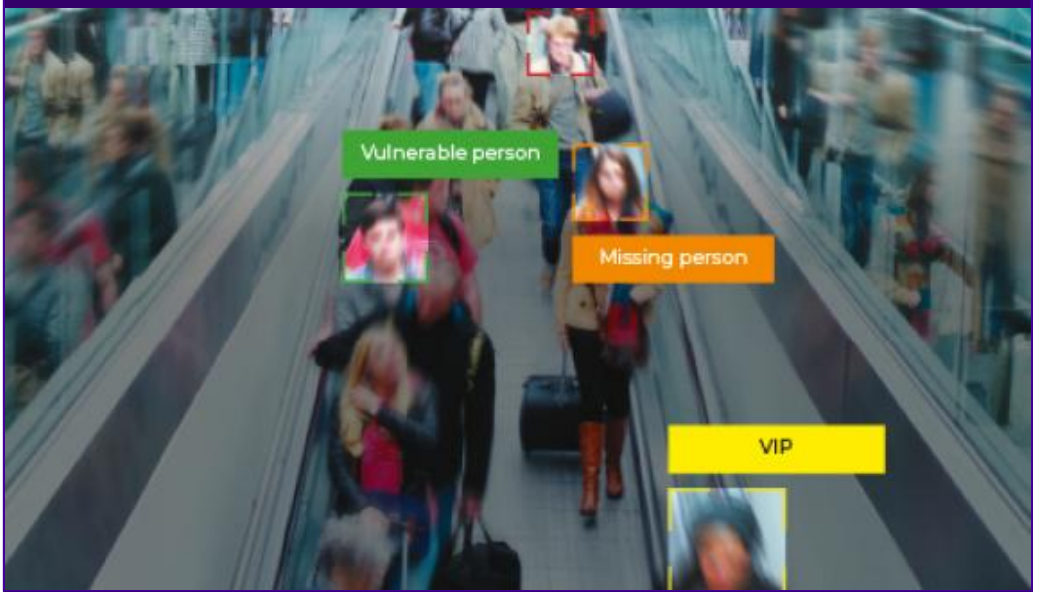
Centralizing user database and biometric terminal management

EN

PRODUCT & SOLUTION

AUGMENTED VISION

Enhancing video management infrastructure



GO TO THE TOP

SUPPLIERS

INVESTORS

MEDIA LIBRARY

EN

MANAGEMENT REPORTING FOR CAPACITY PLANNING AND COMPUTER PERFORMANCE EVALUATION

Donald T. Gantz

Center for Computational Statistics
George Mason University
Fairfax, Virginia 22030

ABSTRACT

Performance analysts and capacity planners keep management informed about how much work the system is doing, how well that work is being done, what problems exist in the system, and who has the problems. The information provided to management is backed up by a large volume of measurement data. Graphic displays are generally valuable for putting life into otherwise dull data and statistics. For computer performance evaluation (CPE) and capacity planning professionals, graphics are the critical media for communication to management. Good graphics speak for themselves. Graphics that present patterns that are readily interpretable by management also buy credibility for the analyst. The primary skill required for successful descriptive analysis and presentation is common sense, backed up by a good toolbox. SAS software is already the basic toolbox for CPE professionals. Many effective graphic techniques are easily implemented through descriptive SAS procedures. Other techniques are accomplished by using SAS procedures to create summary data sets and then to formulate reports by using SAS report writing and SAS/GRAPH capabilities. This paper presents graphic reporting methods that have been successful in the author's capacity planning work.

1. INTRODUCTION

The primary challenge for performance analysts and capacity planners is effective communication of their analyses to management. The challenge is met through succinct understandable reporting. Usually managers want a quick look at system performance or capacity through a one page summary [ART-89]. Further, most analysts have learned that technical language can often impede rather than facilitate communication. This is particularly true when the technical language is statistics. The visual display of information through graphics can provide the succinctness and clarity demanded of management reporting. Both analyst and manager need to *see* the data. Descriptive displays should be an *honest* picture of what has been measured. Patterns in a such a graphic display cry out for interpretation. Often a quick manager will provide an interpretation before the analyst opens his mouth. Such a scenario is often essential when a manager is being asked to base a decision on the analysis.

Management wants to know how much work the system is delivering and how well the work is being delivered. It is important to determine how close the system is to capacity. Capacity is the level of work at which service begins to degenerate. Addressing management's questions requires measurement, analysis and reporting. Measurement units need to be defined for workload and performance. Appropriate summaries for these units must be selected, and reports must be designed. Analysts can take the data summaries and reports that *come for free* in a performance database management system. Or they can *roll their own*. The author's experience is primarily in the latter case. Since most systems produce performance databases (PDB) in SAS, the natural toolbox for analysis and reporting is the SAS system.

A variety of papers discuss the issues of workload and performance measures, statistical summaries, and reporting formats [COM-87]. Workload measures tell what the system is doing. These include CPU, I/O and memory resource units as well as an aggregate of all these units called system resource units. Also, CPU utilization and transaction rates are measures of the amount of work being done. Capacity planners focus on the business which is generating the work for the system, and they often define business-based units [ROS-88, SNY-89]. Response time is what the user is experiencing and it is the usual measure of performance.

Statistical summaries include averages, sums, ranges, deviations and percentiles. The most ubiquitous analysis and reporting summary is the *average*. However, averages are skewed by outlying and bad data values. Additionally, averages do not relate to user experience as well as other measures. Who actually experiences the *average* measurement? Percentiles, on the other hand, are resistant to influence from outlier data and are measures of actual user experience. For instance, 95 percent of response times are lower than the $95th$ percentile measurement. Service level agreements and thresholds are set in terms of response time percentiles.

Reports generally expected by management include daily workload and performance profiles, weekly and monthly summaries, and trend analyses. In the following paragraphs, the author presents those data summaries and report formats which he has used in his capacity planning work.

2. BOX & WHISKER DIAGRAMS

A statistical toolbox is a handy asset for performance analysts and capacity planners. The requirement is met by a knowledge of applied statistical techniques together with the SAS system for implementation. However, many of the modern tools of applied statistics are unknown to capacity planners. In particular, applied statisticians utilize some very powerful descriptive techniques for exploratory data analysis [HOA-83, MOS-77, TUK-77]. Two of the techniques which are illustrated in the example reports in this paper are *Stem & Leaf Charts* and *Box & Whisker Diagrams*. A primary selling point for these tools is their ability to display a large amount of readily understandable information. These techniques have been heavily used by the author in his capacity planning work [GAN-87, GAN-88].

Stem & Leaf charts are bar charts which can be produced with

the SAS PROC CHART procedure. Box & Whisker diagrams are a succinct graphical technique for displaying percentile information [TUK-77]. Figure 1 presents the basic structure of a Box & Whisker diagram as utilized in this paper for reporting. In this definition, the diagram presents five percentiles of a data set: $5th$, $25th$, $50th$ (median), $75th$ and $95th$ percentiles. These percentile selections reflect the relative importance of outlying values in either direction. The choice of percentiles depends on the nature of the data to be displayed, and the diagrams can be based on other groups of percentiles as well. The $25th$ and $75th$ percentiles are represented by the left and right ends of the box, respectively. The $5th$ and $95th$ percentiles are represented by the ends of the whiskers which extend from the box. The $50th$ percentile (median) is represented by the letter 'M' inside the box. The mean is represented by the letter 'A.' When the median and the mean are the same value, our convention is to display only the 'A' in the diagram. Figure 2 presents the Box & Whisker diagram for a particular set of data. Note that for skewed distributions the mean can fall outside of the box.

The width of the box in a Box & Whisker diagram is called the *interquartile range*. The interquartile range is a measure of spread in the data as is the standard deviation of the data. The interquartile range has advantages over the standard deviation as a descriptive statistic. The interquartile range is readily interpretable and it is resistant to influence from unusually large data values (outliers). This resistance is demonstrated in some of the reports presented in the following paragraphs. These advantages are strong arguments for the use of percentile based descriptives, in general.

3. DAILY WORKLOAD PROFILE REPORTS

The amount of work performed by the system in a single day is profiled in Figure 3. This report is for a 3090/200 running a commercial workload of CICS transactions. Work is measured by aggregate resource units defined by Type 72 RMF records. Performance groups are used to identify CICS, database and other systems. This report is constructed using Stem & Leaf diagrams. Stems are the bars whose length summarizes all Type 72 RMF measurements in an hour for performance groups in a system. The leaf structure in the charts consists of the characters used to draw the bars. The characters used represent a particular system; for instance, 'C' represents CICS. Several aspects of the workload profile are immediate. These include absolute resource volume information, an obvious busy period profile, and a comparison of resources delivered to the separate systems. The readily understood and fully reasonable visual information in this report earn credibility for the reporter from management.

This report was produced with the SAS PROC CHART procedure. First, SAS code was used to read data from the PDB, manipulate the observation and variable structure, and relate performance groups to systems. Type 70 and Type 72 RMF records in the PDB are summarized into smaller and more efficiently stored data sets in the reporting data base. These summaries are then available to form this and other reports. For this report the PROC CHART statements are:

```
PROC CHART; BY DATE;
HBAR HOUR / DISCRETE TYPE=SUM
            GROUP=WORK
            SUBGROUP=PERFGRP
            SUM SUMVAR=SERVICE ;
```

Figure 4 presents a daily profile of CPU time consumed. Type 72 records measure CPU time delivered to specific performance groups while Type 70 records measure all CPU time consumed. By combining information from both Type 72 and Type 70 records, the report in Figure 4 includes the additional category of *overhead* CPU time. The author produces the daily profile reports just presented together with profiles of I/O, memory, TCB CPU time and SRB CPU time from a single SAS program. SAS macros and macro language greatly limit the amount of code required.

The same type of charts are used to report daily profiles for channel busy percentages (from Type 73 records), page rates (from Type 71 records), and other RMF information.

4. DAILY PERFORMANCE PROFILE REPORTS

System performance is determined by user response times. Response times for CICS transactions as measured by Type 110 records is response time within the computer, that is, internal response time. A report of response times should provide information concerning how well the system is meeting service level objectives. Typically, such objectives reference the percentage of transactions which experience response times of a predetermined magnitude. CICS response times for a single day's prime shift are summarized in Figure 5 using Box & Whisker diagrams. Box & Whisker diagrams are used to present percentiles of response times in each hour. The selection of the $10th$ and $95th$ percentiles for the whisker values of the diagrams reflects the relative importance of outlying values in either direction. This report immediately tells management the fraction of transactions experiencing various response times. Box & Whisker diagrams are of particular value for illustrating exactly what service level agreements would have been met by the response times. For instance, the diagrams tell what is the shortest reponse time threshold consistently met by 95 percent of transactions. In addition, the flow of the Box & Whisker diagrams indicates the complexities of response time variation over the day. Note that average response times are frequently greater than the $75th$ percentile response times; that is, the average is to the right of the box. This reiterates the caution of many CPE reporting experts that people do not experience averages, whereas, in contrast, n% of transactions have response times as low as the nth-percentile response time. Means (averages) are sensitive to, and hence skewed by, outlying data (unusually large or small values); however, percentiles are resistent to the influence of outliers. Our experience has been that the use of such a graphic report as that in Figure 5 which clearly presents complex structures has built credibility for our measurements with management.

Figure 5 also presents additional information on transaction counts and percentages that supplement the Box & Whisker description of response times. Such enhancement of a graphic display helps to maximize the amount of information available to management in a single report. There is a considerable literature on the enhancement of graphics [WAI-88].

The first step in producing the report in Figure 5 is to write the SAS statements for selecting the normal production transactions from the CICS Type 110 records part of the PDB. Then PROC UNIVARIATE is used to get the percentile variables which summarize internal response time. These summaries are saved in a reporting data base. Finally SAS report writing code is required to produce the report from the reporting data base file.

5. WEEKLY AND MONTHLY WORKLOAD REPORTS

The daily workload profile in Figure 4 is a snapshot of one day's service delivery. We now consider reports which present the variability in workload among days. We face the requirement of

aggregating workload charts like that in Figure 4 for a number of days. Again, the appropriate tool for this aggregation is Box & Whisker diagrams. Figure 6 presents an aggregation of 6 months of daily workload profiles of the type presented in Figure 4. Note that the robust daily profile shape showing morning and afternoon peaks carries over from Figure 4 to Figure 6. However, Figure 6 also illustrates the variation in daily workload. In particular, for any service delivery level, the diagrams provide information concerning the percentage of days in which that level was not exceeded. Figure 6 illustrates absolute workload levels, patterns of workload over the day, and daily variation in workload.

The daily workload profile reports like the report in Figure 4 were generated from *summaries* of the PDB Type 70 and Type 72 records. To produce the report in Figure 6, these summaries were run through the SAS PROC UNIVARIATE procedure to get a report file with the percentile variables which summarize workloads for a week, month, etc. SAS report writing code was used to produce the report from the reporting data base file.

Once the percentile summary data used to produce the report in Figure 6 is available in a reporting data base, selected parts of the summarized data can be used for additional reports. Also, with slight modifications to the progression of programs used to produce the report in Figure 6, other reports illustrating aspects of the variation in daily workload can be designed. For instance, Figure 7 uses Box & Whisker diagrams to study the variation in service delivery across the five day workweek. Several months of 2PM data is used in this report. Two PM is typically the busiest period in the particular data center whose data is being reported, so it is the most reasonable choice of a single hour to analyze. Note in this report that the entire week's data is summarized in a single Box & Whisker diagram and also broken down into daily Box & Whisker diagrams. The report in Figure 7 illustrates the resistance of percentiles to bad data. There were problems with the RMF data on a couple of Mondays, and the result was incorrect low measurements. Note that the mean (average) is skewed low by the bad data, but only the left whisker of the Box & Whisker diagram is affected. This illustrates the utility of these Box & Whisker based reports in that they can be generated to provide satisfactory results without major preliminary data cleaning.

The same type of charts are used to report daily profiles for channel busy percentages (from Type 73 records), page rates (from Type 71 records), and other RMF information. A daily profile for channel busy percents for a single channel across one day's prime shift is presented using the SAS PROC CHART PROCEDURE in the report in Figure 8. A daily profile for channel busy percents for *all* channels across one day's prime shift is presented using Box & Whisker diagrams in the report in Figure 9. Note that the Box & Whisker diagrams in Figure 9 are not as structured as the ones in our earlier reports. However, they provide a very useful summary of channel activity in the system. The report provides a quick comparison of channel activity. The Box & Whisker diagram for channel 2 in the report in Figure 9 is a summary of the entire bar chart in Figure 8.

6. WORKLOAD TREND REPORTS

There are a number of motivations for looking at trends in computer measurements. Each motivation prompts its own analysis and reporting approach. Figure 10 presents the flow of weekly service unit delivery over 22 weeks. Box & Whisker diagrams in Figure 10 summarize the data that is plotted in Figure 11. Each Box & Whisker diagram is a *strip* summary of data points for 30 data points (6 busy period hours in each of five weekdays). Figure 11 contains the data for weeks 22 through 29. Note that measurement problems on monday of week 29 result in six incorrectly low hourly measurements. However, only the average and the lower part of the Box & Whisker diagram is skewed although six of thirty values are *bad*.

Figure 10 illustrates some problems we face in defining trends. Note the smooth flow of box & whisker diagrams broken by a *jump* at week 22. Week 22 was the first week after a major hardware upgrade. The jump is almost entirely due to a jump in memory units with the new system. There is a *correction* in weeks subsequent to week 22 as the system is tuned. The analyst needs to use judgment in discarding some weeks of data from a trend analysis. Somewhere around week 25 the system appears to have settled and trends can again be tracked. There is a clear periodic (sine curve) component to the Box & Whisker flow. A straightforward approach to trend analysis is to separate a straight line growth component from the purely periodic component. The straight line component is the long term growth and the periodic component shows the seasonal variation. For most analyses, there is probably no need to use anything fancier than *ad hoc* techniques to specify these components. In the rare cases where sophisticated statistics are called for, the proper tools are found in time series analysis. In particular, taking first order differences will eliminate the jump (at week 14) as well as the straight line trend. A second order autoregressive model usually fits a sinusoidal component well; the resulting nonstationary second order autoregressive model in turn can be used as a forecasting tool.

For capacity planning, the Figure 10 style analysis ought to be applied separately for each business element. Then knowledge of the historic growth of that business element in purely business terms is linked to the measurement analysis to find the effect of the business element on data systems.

Figure 12 presents a very concentrated summary by month of the workload growth. The summary *smooths* the workload data presentation we saw in Figure 10. The smoothing cuts out seasonal variation in the data and allows the real growth to be more visible.

Figure 13 is similar in style to Figure 10, and it presents the trend in CPU consumption over the 22 weeks. There is no jump in this graph at week 22 like there is in the service unit graph. Otherwise the trend analysis problems in the two graphs are equivalent.

7. PERFORMANCE TREND REPORTS

We now discuss an approach to the task of aggregating Figure 5 daily performance profile information for a group of days. The variation across days that is to be captured in this instance is a variation in response time distributions as summarized by Box & Whisker diagrams. Focusing on just the 2PM busy hour, Figure 14 presents a stream of daily response time Box & Whisker summaries. One workload measurement, number of transactions in the hour, is included in the report. Part of Figure 14, 3 April 87 through 17 April 87, shows a pattern of improving response times in that both the box and the upper whisker are shrinking. In contrast, another part of Figure 14, 18 February 87 through 26 February 87, shows degenerating response times. These very definite patterns of changing response times provide management with the users' view of changes being implemented in the data center.

Figure 14 points out an underlying problem with doing statistical analysis on performance data. That is, the data is collected under a very changing environment. Both good and bad response times appear at each workload level. An analyst cannot

simply throw all data into a hopper to find a robust relationship between workload volume and response time. Differences in the system, other than workload, during the measurements may account for much of the response time variation. This data problem makes it more imperative that the analyst present to management clear and honest pictures of data rather than results of statistical analysis in technical jargon.

A close relationship should exist between the analyst generating these reports, the capacity planners familiar with business changes, and the data center performance people. Together, these experts should be able to explain the patterns that are apparent in the reports. The reports can serve to sharpen these experts' judgement concerning the effects that changes in their respective areas will have on performance.

Figure 15 presents a report which adds more workload information to the report in Figure 14. This new report has a *wings* look. The right *wings* consist of the response time Box & Whisker summary from Figure 14. The left *wings* consist of a Stem & Leaf summary of service units delivered during the hour. The stem is an aggregate of the three 2PM bars from the chart in Figure 3. The leaf information refers to the subsystem consuming the service: 'C' for CICS, 'D' for data base, and 'O' for overhead. The focus of data center evaluation is the busy period when the system is dealing with its peak workload. This report provides a very complete summary over time of the busy period. The value of this report depends very much on the appearance of performance patterns such as those occuring in Figure 15. Changing performance demands explanation. Is the system in need of tuning or approaching capacity? Reports such as this one help management to approach data center operations with concrete questions and criteria for evaluation.

8. CAPACITY

Capacity is the workload level at which performance begins to degenerate. The style of capacity graph produced from *analytic modeling* is presented in Figure 16. This is truely a ubiquitous graph in the capacity planning literature [ART-89, GRA-88,OLC-87,WIC-88]. Qualitatively, this curve warns the analyst that at some workload level performance will degenerate. Quantitatively, the graph plots the theoretical average response time at various workload levels. Workload is measured as the hourly transaction rate during the busy period.

The theme of this paper has been that other statistical summaries provide better user related performance information than averages. We will now look at a version of the capacity plot using Box & Whisker summaries. Figure 17 presents a scatter plot of data that has the characteristics of the average plot in Figure 16. Figure 18 presents a Box & Whisker diagram summary of this data. This graph shows how the percentiles of response time change as the rate of transactions increases. It is more important to know when the *95th* percentile response time crosses the service level threshold than when the average response time crosses the threshold. In fact, for this data, the *95th* percentile response time is at 15 seconds when the average response time is just at the threshold of five seconds.

The theoretical plot in Figure 16 and the data in Figure17 are appropriate for an unchanging system. However, none of us ever really measure an unchanging system. As pointed out in Paragraph 7 above, the system is generally under constant alteration either in its workload characteristics or its configuration and tuning. This fact emphasizes the value of exploratory type data display techniques to find patterns in data for the analyst and to report these patterns to management.

REFERENCES

[ART-89] Artis, H.P., "Capacity Surfaces," *Mainframe Journal*, February, 1989

[CHA-83] Chambers, J., Cleveland, W., Kleiner, B. and Tukey, P., *Graphical Methods for Data Analysis*, Duxbury Press, 1983.

[CLE-85] Cleveland, W., *The Elements of Graphing Data*, Wadsworth, 1985.

[COM-87] Comstock, C., "CICS Performance and Workload Reports," *CMG Transactions*, Summer, 1987, Issue Fifty-Seven.

[GAN-88] Gantz, D., "Tutorial on Statistical Methods for Analysis and Reporting for Capacity Planning," *Proceedings of the CMG'88, Dallas*, 1988.

[GAN-88] Gantz, D. and Wegman, E., "Invited Discussion of a Paper by Drs. Cleveland & McGill, *Journal of the Royal Statistical Society (A)*, Volume 10, Part 3, 1987.

[GAN-87] Gantz, D., Gonzalez, C. and Moore, M., "Capacity Planning for a Multi-Processor Environment Supporting CICS Transactions," *Proceedings of the CMG'87*, Orlando, 1987.

[GRA-88] Gray, W., "The Use of Analytic Modeling to Compare Various I/O Configurations," *CMG Transactions*, Fall, 1988, Issue Sixty-Two.

[HOA-83] Hoaglin, D., Editor, *Understanding Robust and Exploratory Data Analysis*, Wiley, 1983.

[MOS-77] Mosteller, F. and Tukey, J.W., *Data Analysis and Regression*, Addison-Wesley, 1977.

[OLC-87] Olcott, R., "I/O Tuning in the MVS/XA Environment," *CMG Transactions*, Summer, 1987, Issue Fifty-Seven.

[SAS-85] SAS Institute Inc., *SAS User's Guide: Basics*, Version 5, 1985.

[SNY-89] Snyder, M., "The Capacity Planning Database," *Mainframe Journal*, March, 1989.

[TUK-77] Tukey,J.W., *Exploratory Data Analysis*, Addison-Wesley, 1977.

[WAI-88] Wainer, H. and Thissen, D., "Plotting in the Modern World: Statistics Packages and Good Graphics," *CHANCE, New Directions for Statistics and Computing*, Vol. 1, No. 1, 1988.

[WIC-88] Wicks, R., *Balanced Systems and Capacity Planning*, Washington Systems Center Technical Bulletin, 1988.

BOX & WHISKER DIAGRAMS

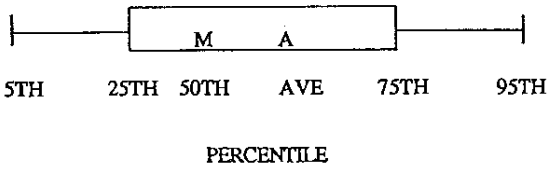


Figure 1

BOX & WHISKER DIAGRAM

Figure 2

COMPUTER RESOURCES DELIVERED
COMPUTER='3090'

COEFFICIENTS FOR SYSTEM RESOURCE UNIT CALCULATION
CPU - 10 SRB - 10 I/O - 10 MEMORY - 3
SRUS = CPUCOEFF * CPU TCB SECONDS * HARDWARE CONSTANT
+ SRBCOEFF * CPU SRB SECONDS * HARDWARE CONSTANT
+ I/O COEFF * UNITS
+ MEMCOEFF * MEMORY (PAGE SECONDS) UNITS

DATE=TUESDAY, MARCH 24, 1987

BAR CHART OF SUMS

WORK	HOUR		SERVICE SUM
		!	
CICS	8	!CCCCCCCCCCCCCCCCCCCC	103,941,452
	9	!CCCCCCCCCCCCCCCCCCCCCCCCCCCC	144,840,441
	10	!CCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCC	173,444,329
	11	!CCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCC	192,211,272
	12	!CCCCCCCCCCCCCCCCCCCCCCCCCCCC	140,438,479
	13	!CCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCC	179,126,138
	14	!CCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCC	195,555,436
	15	!CCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCC	159,053,815
	16	!CCCCCCCCCCCCCCCCCCCCCCCC	122,382,475
	17	!CCCCCCCCCCCCCCCCCC	93,483,954
	18	!CCCCCCCCCCCC	57,964,907
	19	!CCCCCC	32,056,258
		!	
DATA BASE	8	!DDDDDDDDDDD	53,174,617
	9	!DDDDDDDDDDDDD	72,017,917
	10	!DDDDDDDDDDDDDDDDDDDDD	102,898,928
	11	!DDDDDDDDDDDDDDDDDDDDD	105,089,152
	12	!DDDDDDDDDDDDDDDDDDDDD	104,869,696
	13	!DDDDDDDDDDDDDDDDDDDD	102,491,920
	14	!DDDDDDDDDDDDDDDDDDDDD	106,479,056
	15	!DDDDDDDDDDDDDDDDDD	94,896,240
	16	!DDDDDDDDDDDDDDDDD	87,071,760
	17	!DDDDDDDDDDDDD	66,033,975
	18	!DDDDDDDDD	43,443,752
	19	!DDDDDDDDD	43,072,900
		!	
OTHER	8	!OO	8,296,490
	9	!O	7,739,571
	10	!OO	9,849,176
	11	!OO	9,788,881
	12	!OO	10,868,967
	13	!OO	9,348,511
	14	!OO	9,470,224
	15	!OO	10,656,418
	16	!OO	10,452,706
	17	!OO	9,277,668
	18	!O	4,721,214
	19	!O	6,233,000
		!	
		----------+----------+----------+----	
		60,000,000 120,000,000 180,000,000	
		SRU'S	

Figure 3

COMPUTER RESOURCES DELIVERED
COMPUTER='3090'

COEFFICIENTS FOR SYSTEM RESOURCE UNIT CALCULATION
CPU - 10 SRB - 10 I/O - 10 MEMORY - 3

TOTUNITS = CPUCOEFF * CPU TCB SECONDS * HARDWARE CONSTANT
+ SRBCOEFF * CPU SRB SECONDS * HARDWARE CONSTANT

DATE=TUESDAY, MARCH 24, 1987

BAR CHART OF SUMS

WORK	HOUR		TOTUNITS SUM
		!	
CICS	8	!CCCCCCCCCCCCCCCCC	9,249,122
	9	!CCCCCCCCCCCCCCCCCCCCC	11,675,074
	10	!CCCCCCCCCCCCCCCCCCCCCCCCCCCCC	15,218,365
	11	!CCCCCCCCCCCCCCCCCCCCCCCCCCCCCCC	16,128,268
	12	!CCCCCCCCCCCCCCCCCCCCC	11,455,087
	13	!CCCCCCCCCCCCCCCCCCCCCCCCCCCCC	15,566,435
	14	!CCCCCCCCCCCCCCCCCCCCCCCCCCCCCCC	16,318,178
	15	!CCCCCCCCCCCCCCCCCCCCCCCCC	13,728,905
	16	!CCCCCCCCCCCCCCCCCCC	10,707,285
	17	!CCCCCCCCCCCCCCC	7,888,620
	18	!CCCCCCCCCC	5,104,939
	19	!CCCCC	2,701,804
		!	
DATA BASE	8	!DDDDDDDDDDD	6,112,536
	9	!DDDDDDDDDDDDDDD	8,078,075
	10	!DDDDDDDDDDDDDDDDDDDDDDD	11,554,361
	11	!DDDDDDDDDDDDDDDDDDDDDDD	11,633,116
	12	!DDDDDDDDDDDDDDDDDDDDDDD	11,623,865
	13	!DDDDDDDDDDDDDDDDDDDDDDD	11,418,950
	14	!DDDDDDDDDDDDDDDDDDDDDDD	11,725,664
	15	!DDDDDDDDDDDDDDDDDDDDD	10,496,005
	16	!DDDDDDDDDDDDDDDDDD	9,651,208
	17	!DDDDDDDDDDDDDD	7,371,732
	18	!DDDDDDDDDD	4,886,460
	19	!DDDDDDDDDD	4,892,854
		!	
OTHER	8	!OOOOO	2,939,324
	9	!OOOOO	2,622,827
	10	!OOOOOOA	3,412,724
	11	!OOOOOOA	3,488,076
	12	!OOOOOOOA	3,979,686
	13	!OOOOOA	3,480,639
	14	!OOOOOA	3,321,654
	15	!OOOOOOA	3,628,447
	16	!OOOOOOA	3,700,491
	17	!OOOOOA	3,226,833
	18	!OOOA	1,772,069
	19	!OOOOA	2,245,956
		!	
OVERHEAD	8	!SSSSSSSSSS	4,849,669
	9	!SSSSSSSSS	5,149,013
	10	!SSSSSSSSSSSSSS	7,570,660
	11	!SSSSSSSSSSSSSSSSS	8,981,901
	12	!SSSSSSSSSSS	5,852,826
	13	!SSSSSSSSSSSS	6,488,237
	14	!SSSSSSSSSSSS	6,656,347
	15	!SSSSSSSSSSSS	6,580,825
	16	!SSSSSSSSSS	5,884,266
	17	!SSSSSSSSS	4,916,073
	18	!SSSSSS	3,164,025
	19	!SSSSSSSSSS	5,435,279
		!	
		----------+----------+----------+----	
		6,000,000 12,000,000 18,000,000	

TOTAL CPU UNITS

SYMBOL	PERFGRP	SYMBOL	PERFGRP	SYMBOL	PERFGRP
U	UNASSIGNED	O	OTHER	C	CICS
D	DATA BASE	A		S	SYSTEM OVERHEAD

Figure 4

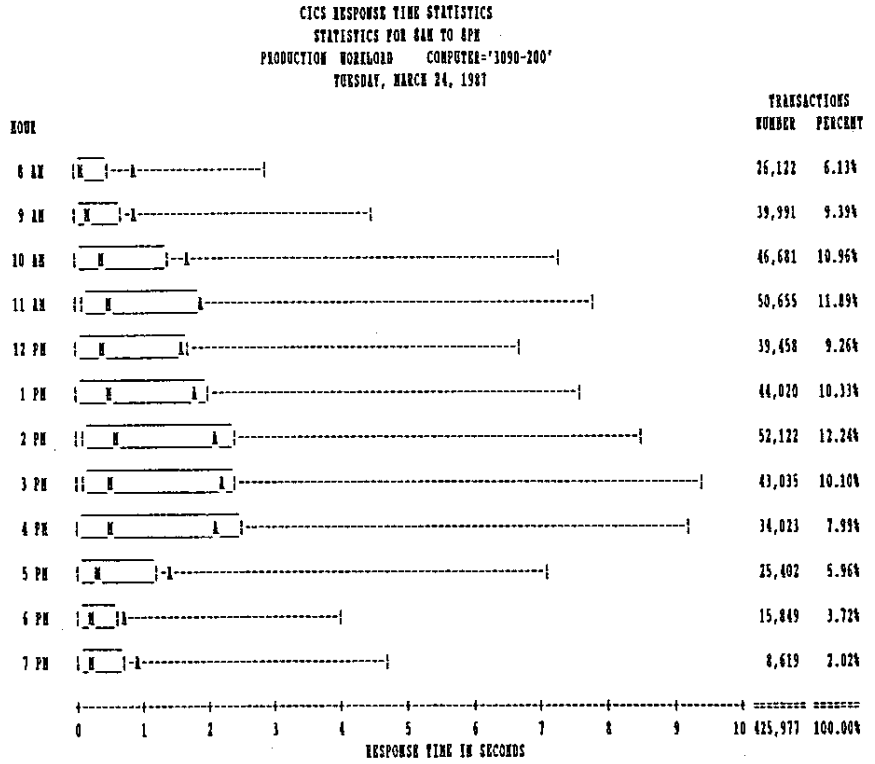


Figure 5

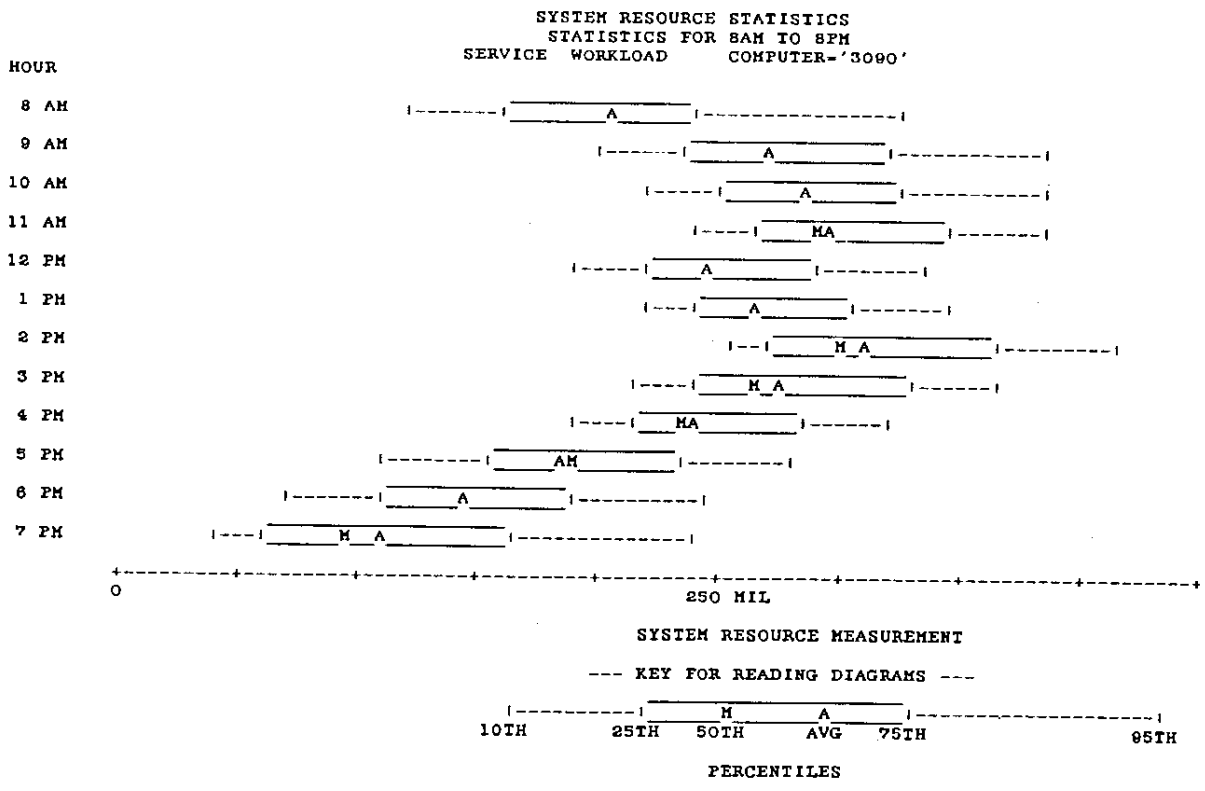


Figure 6

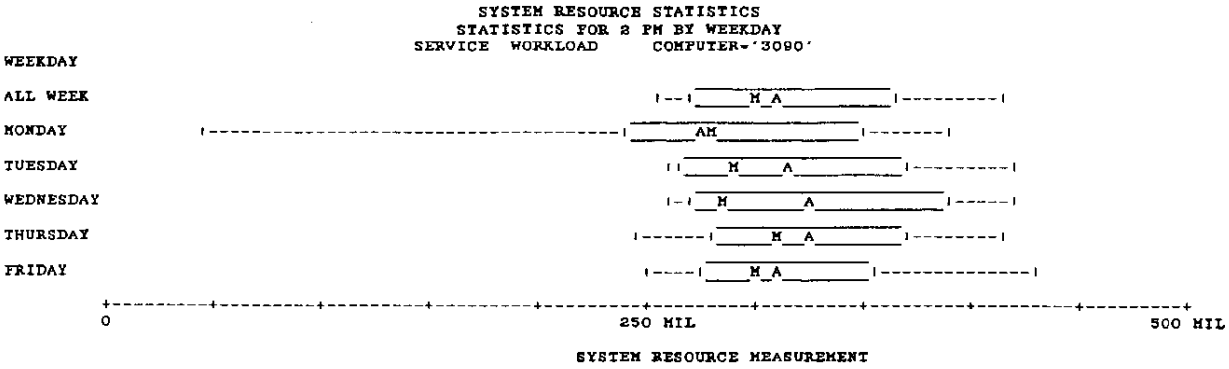


Figure 7

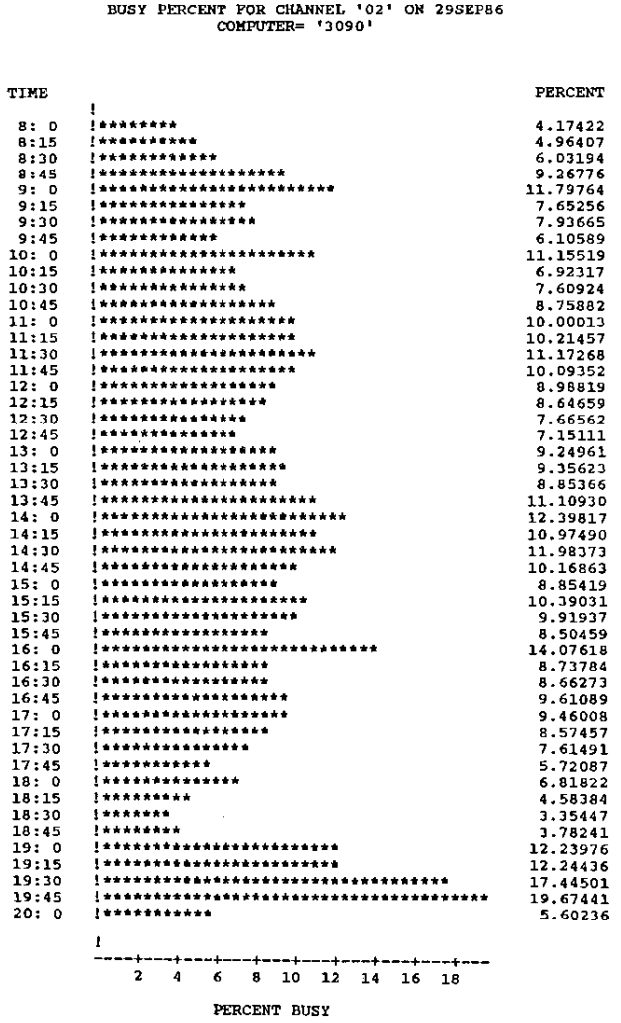


Figure 8

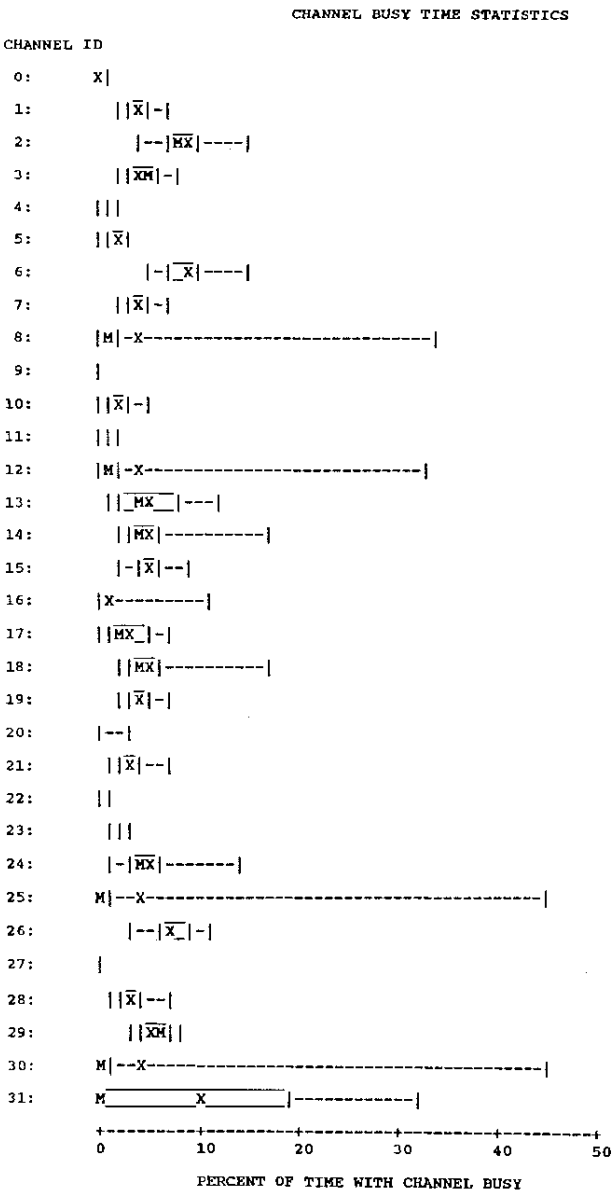


Figure 9

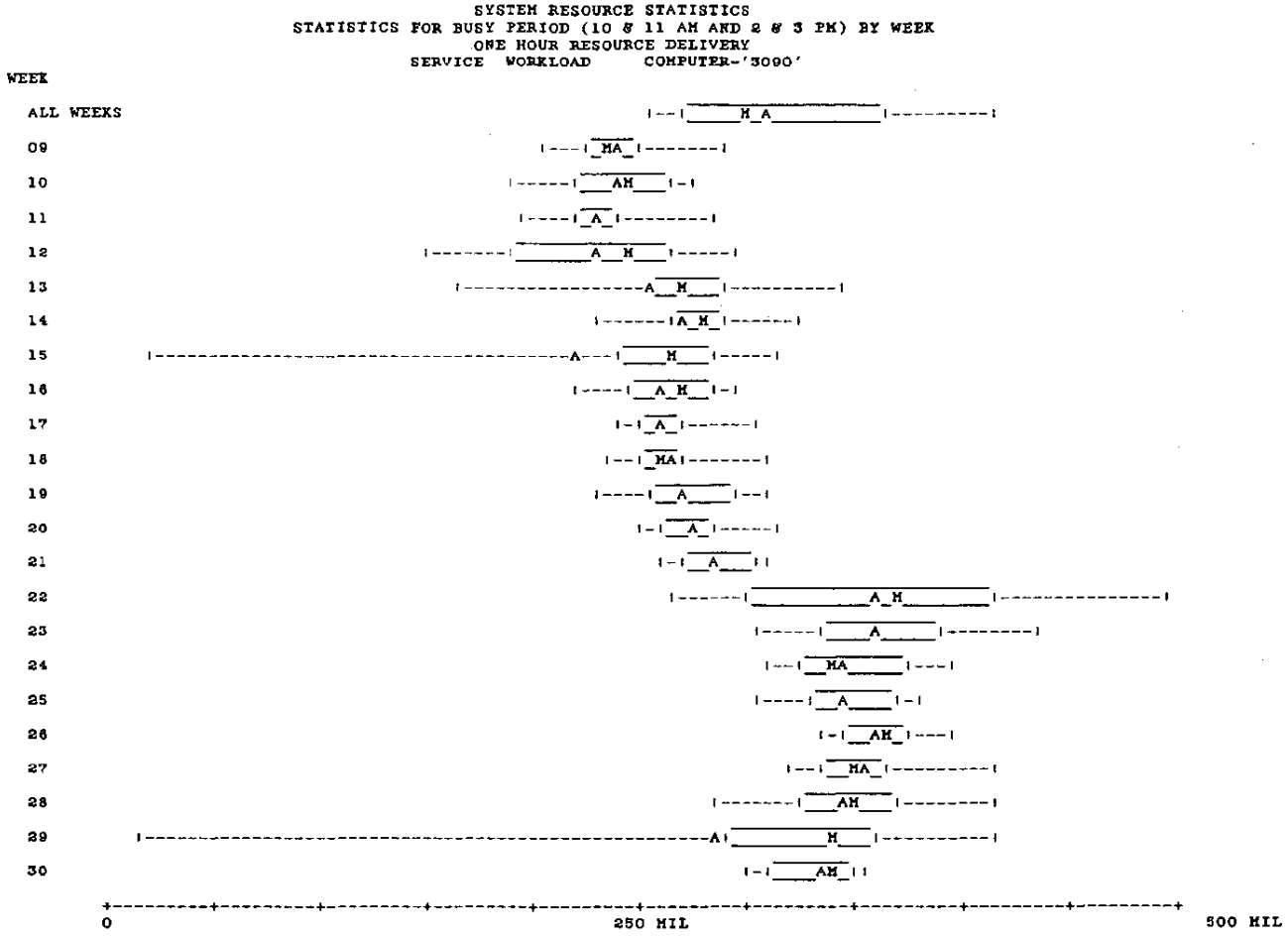


Figure 10

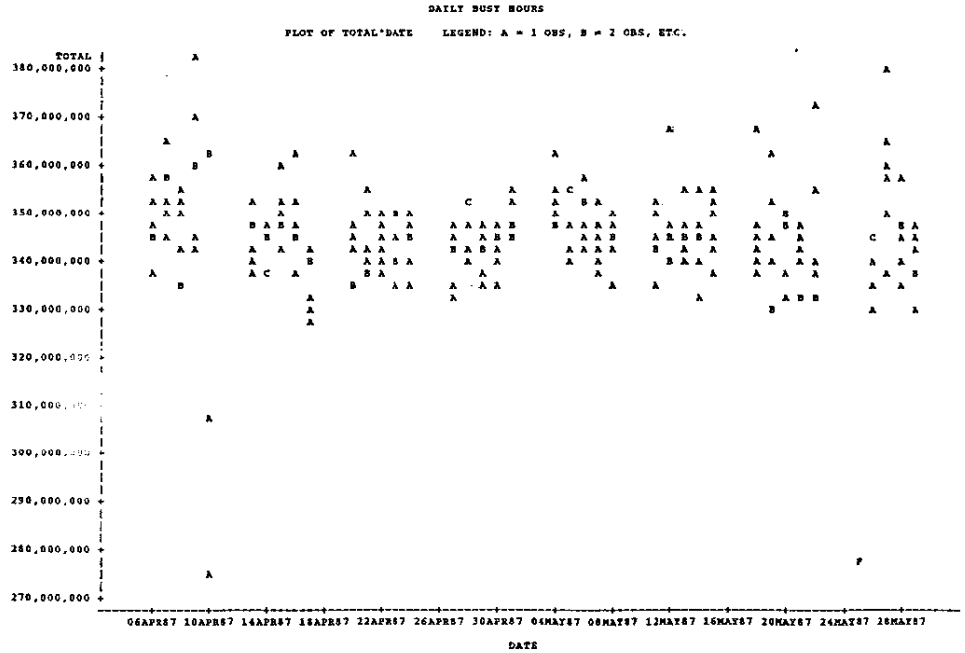


Figure 11

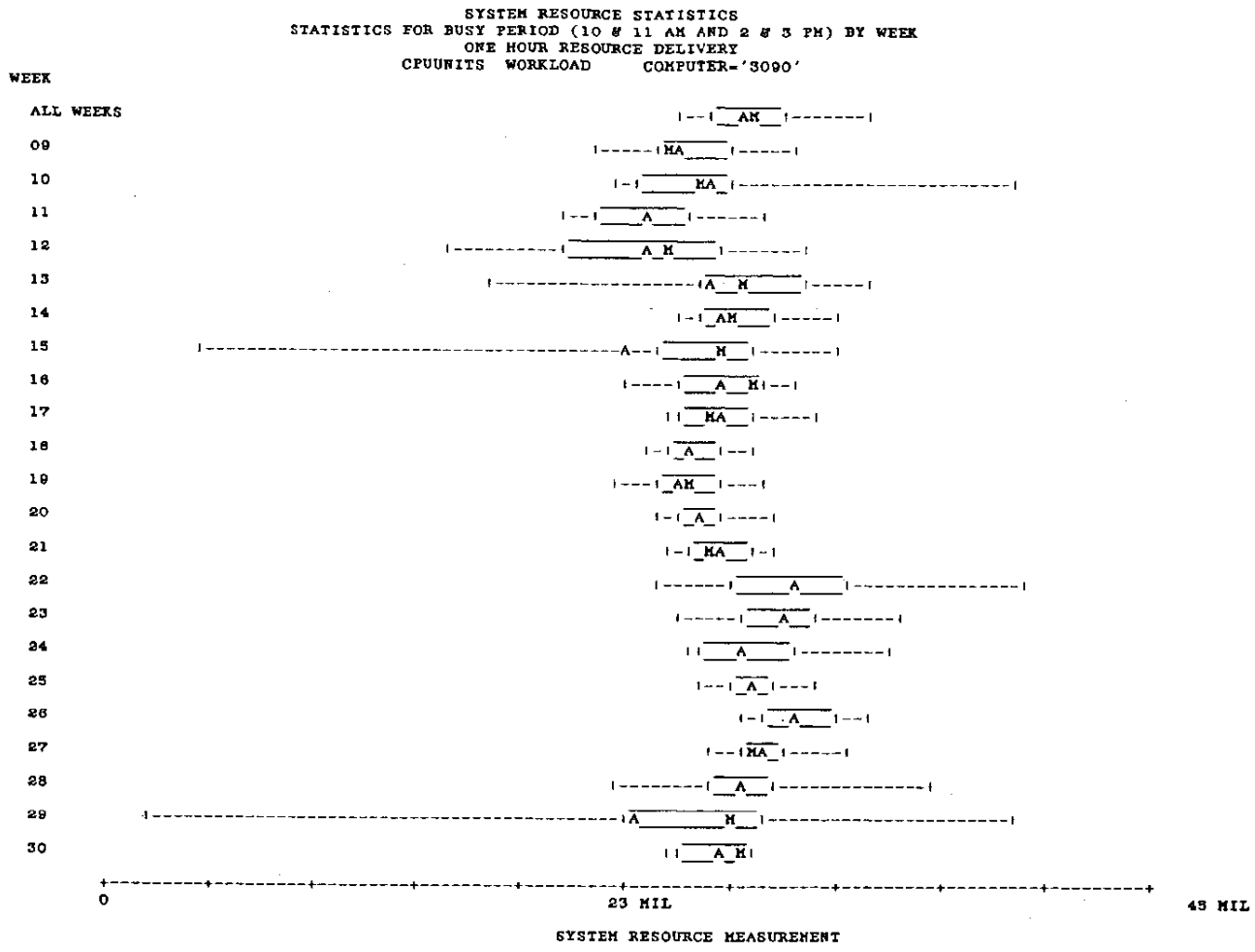


Figure 12

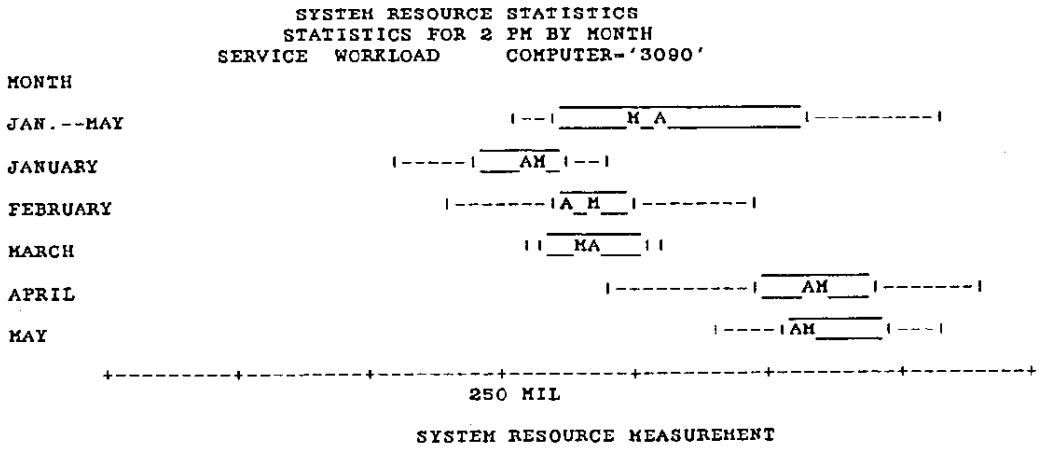


Figure 13

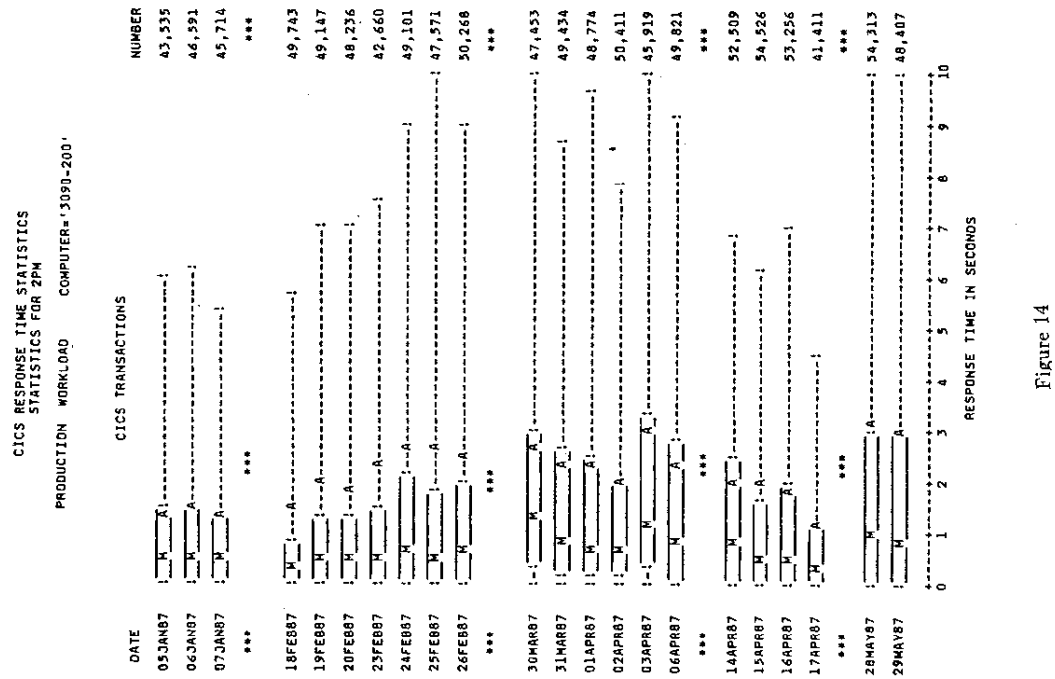


Figure 14

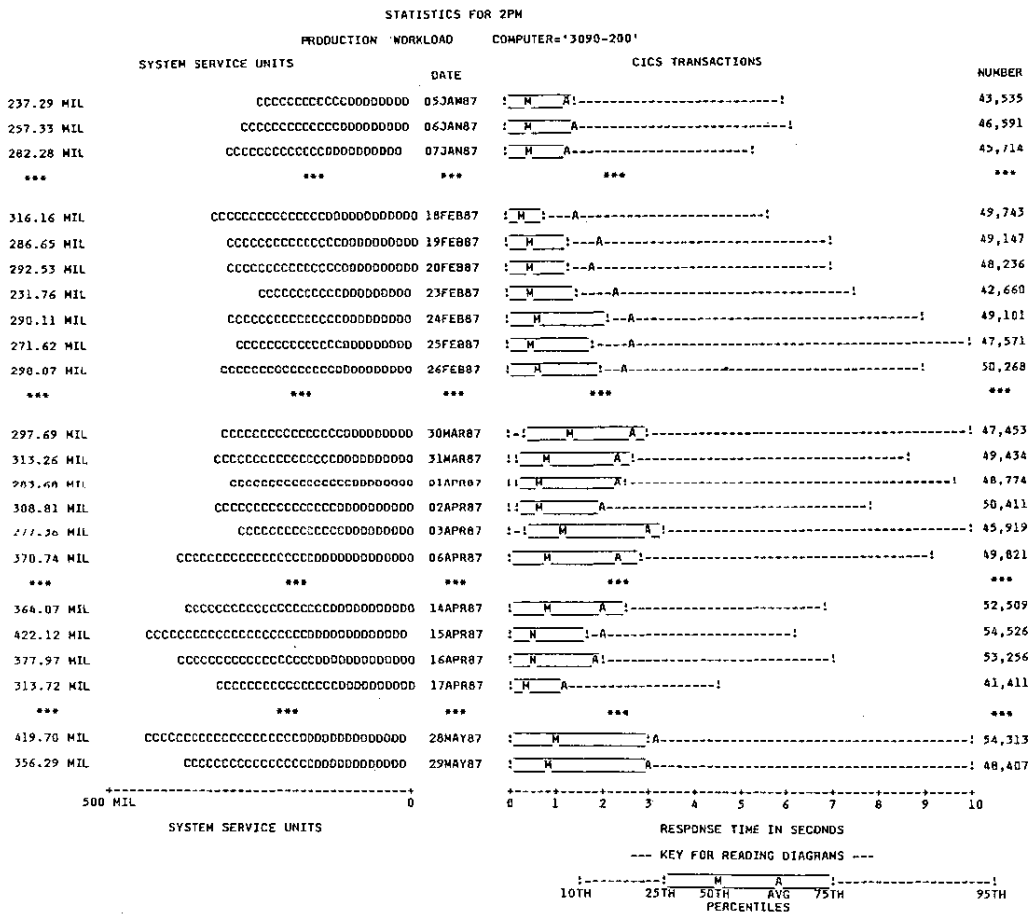


Figure 15

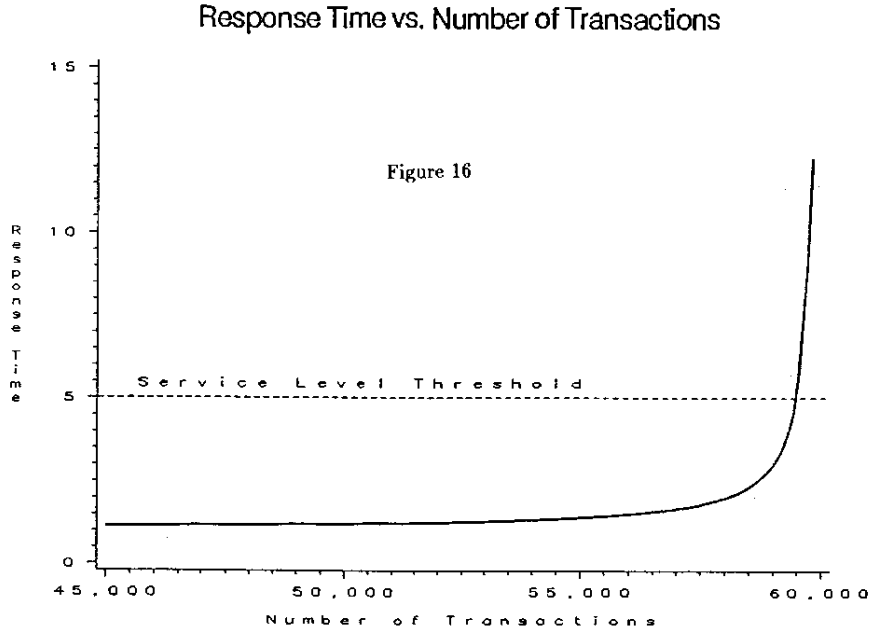


Figure 16

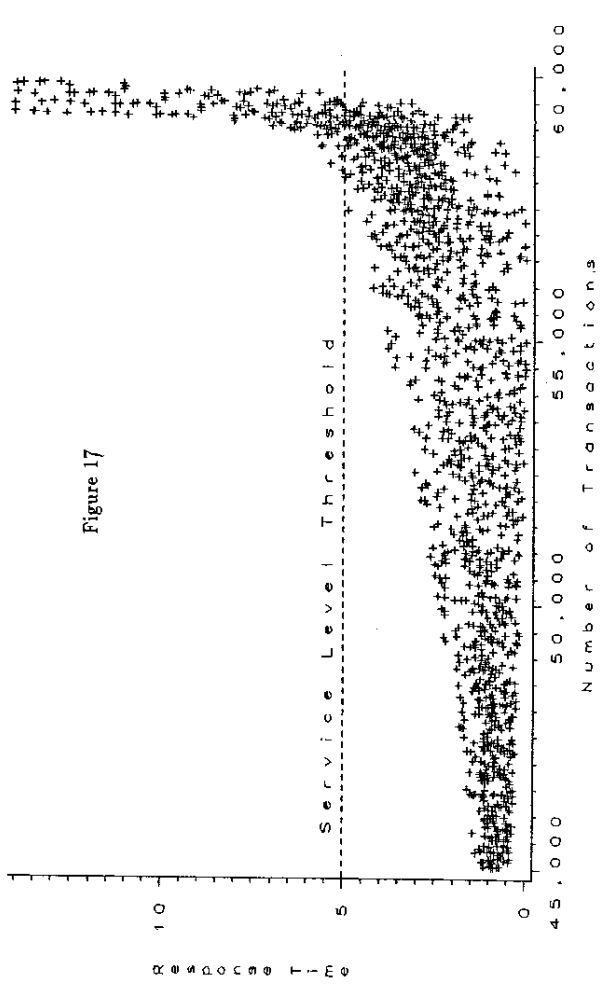


Figure 17

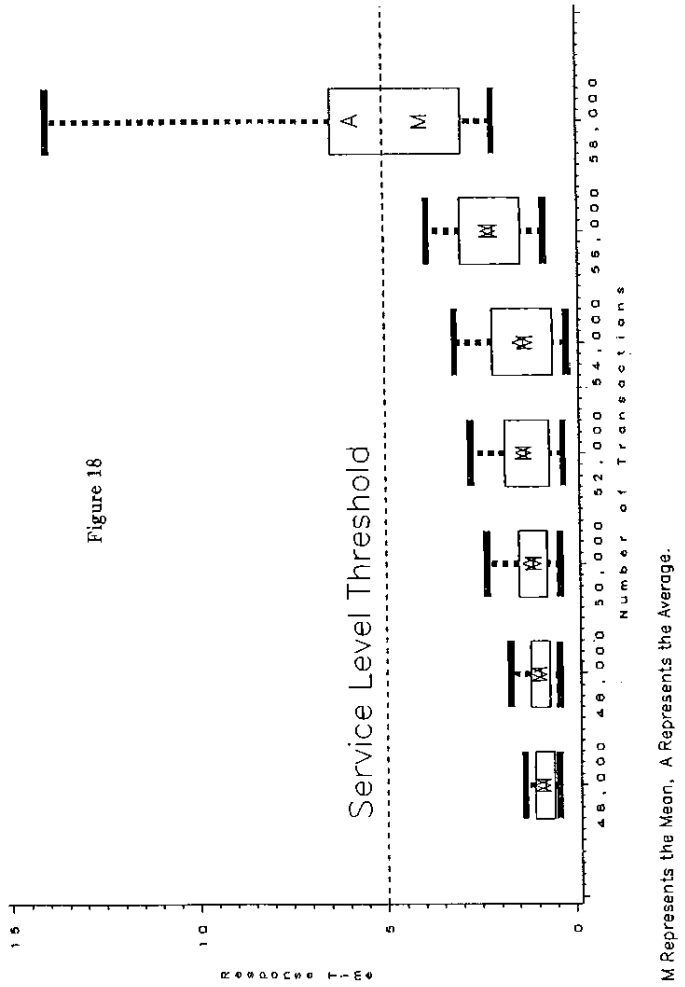


Figure 18

M Represents the Mean, A Represents the Average.

Journal of Forensic Sciences / Volume 62, Issue 3

Paper

A Set of Handwriting Features for Use in Automated Writer Identification[†,‡]

John J. Miller Ph.D., Robert Bradley Patterson Ph.D., Donald T. Gantz Ph.D., Christopher P. Saunders Ph.D., Mark A. Walch M.P.H., M.Arch., JoAnn Buscaglia Ph.D.

First published:05 January 2017
https://doi.org/10.1111/1556-4029.13345
Citations: 6

† This is publication number 14–27 of the FBI Laboratory Division. Names of commercial manufacturers are provided for identification purposes only, and inclusion does not imply endorsement of the manufacturer or its products or services by the FBI. The views expressed are those of the authors and do not necessarily reflect the official policy or position of the FBI or the U.S. government.

‡ Funded under a contract award to The Gannon Technologies Group by the FBI Laboratory Division.

Abstract

A writer's biometric identity can be characterized through the distribution of physical feature measurements (“writer's profile”); a graph-based system that facilitates the quantification of these features is described. To accomplish this quantification, handwriting is segmented into basic graphical forms (“graphemes”), which are “skeletonized” to yield the graphical topology of the handwritten segment. The graph-based matching algorithm compares the graphemes first by their graphical topology and then by their geometric features. Graphs derived from known writers can be compared against graphs extracted from unknown writings. The process is computationally intensive and relies heavily upon statistical pattern recognition algorithms. This article focuses on the quantification of these physical features and the construction of the associated pattern recognition methods for using the features to discriminate among writers. The graph-based system described in this article has been implemented in a highly accurate and

> approximately language-independent biometric recognition system of writers of cursive documents.

Citing Literature

[+] [−] Donald T. Gantz

[−] Person information

- *affiliation:* George Mason University, Fairfax VA, USA

[−] 2000 – 2009

2009

[c5] Robert T. Quinn, Donald T. Gantz: **An approach to a capstone curriculum.** SIGITE Conference 2009: 150-154

[c4] Louis R. D'Alessandro, Donald T. Gantz: **Combining academic studies with it certifications: becoming a cisco regional academy.** SIGITE Conference 2009: 209-214

[−] 1990 – 1999

1990

[c3] Donald T. Gantz, James R. Mekemson: **Flow profile comparison of a microscopic car-following model and a macroscopic platoon dispersion model for traffic simulation.** Winter

[−] Refine list

showing all 6 records

refine by search term

refine by type
- Conference and Workshop Papers (only)
- Editorship (only)

select all | deselect all

refine by coauthor
Susan L. Solomon (1)
Louis R. D'Alessandro (1)
James R. Mekemson (1)
Melvin L. Moore (1)
Carlos M. Gonzalez (1)
Gerard C. Blais (1)
Robert T. Quinn (1)

refine by venue
Int. CMG Conference (2)
SIGITE Conference (2)
Winter Simulation Conference (2)

Simulation Conference 1990:
770-774

[–] **1980 – 1989**

1988

[c2] Donald T. Gantz:
Tutorial On Statistical Methods For Analysis And Reporting. Int. CMG Conference 1988: 871-880

1987

[c1] Donald T. Gantz, Carlos M. Gonzalez, Melvin L. Moore:
Capacity Planning for a Multi-Processor Environment Supporting CICS Transactions. Int. CMG Conference 1987: 405-414

1985

[e1] Donald T. Gantz, Gerard C. Blais, Susan L. Solomon:
Proceedings of the 17th conference on Winter simulation, WSC 1985, San Francisco, CA, USA, December 11-13, 1985. ACM 1985, ISBN 0-911801-07-3 [contents]

[+] **Coauthor Index**

DB2™ PERFORMANCE MANAGEMENT

Donald T. Gantz, George Mason University
James M. Thompson, Federal Home Loan Mortgage Corporation

ABSTRACT

This paper describes analysis done on measurements of a DB2 production workload at a Fortune 500 company. DB2 is IBM's strategic database management system. Several sections of the company's computer and information systems staff have a need for DB2 performance information. The authors have developed and piloted a technique for relating the design structure of DB2 plans to their performance. SAS® Institute's products play a major supporting role in executing this technique. SAS programming, procedures, report writing, and graphing are used to augment and exploit a computationally intense classification tree software product.

INTRODUCTION

Database Administrators (DBAs) involved in technical support for application development provide expertise in strategic products such as DB2. These DBAs have the experience to translate design alternatives into resource requirements. In particular, they estimate the logical and physical I/O's that a design will require. The technique developed for this paper is aimed at developing products to help these DBAs predict the performance, in terms of CPU seconds and elapsed times, of alternative DB2 designs. In the spectrum of estimation accuracy – from gross to precise – these DBAs require the rough cut estimates that are provided by rules of thumb (ROTs). Early in the life cycle of plan development cheap, ballpark estimation is appropriate. Later in the life cycle, when the design is more fully specified, more expensive and accurate modeling techniques are appropriate. The technique outlined in this paper provides a framework for the development of ROTs for DB2 performance prediction. Our continuing work along these lines will develop and validate working ROTs.

In addition, we hope that the technique described in this paper will be applicable to problems facing production control and capacity planning staff. We will describe how products of the analysis can be used to classify and track the performance of DB2 plans as well as to define and monitor service level agreements with clients.

THE DB2 PRODUCTION ENVIRONMENT

The workload environment consisted of a high end 370/ESA class processor under MVS/ESA™ with approximately 500 GBytes of on-line storage. The Information Management System (IMS) subsystem was at version 2.2 and DB2 at version 2.1. This IMS subsystem, connected to DB2 for database services, processed 45,000 to 53,000 transactions per workday, 8 a.m. to 4 p.m..

The data for our analysis came from the MICS DB2 Component. The specific file was the DB2 User Activity file. At the detail level this file contained one record for each plan execution, from thread create to termination. Among the information presented were user identification; SQL interaction presented by the plan, SELECTs, FETCHs, UPDATEs, INSERTs, and DELETEs; some internal DB2 activities, locks and buffer access; and time statistics, CPU and elapsed time.

THE DB2 ANALYSIS PROBLEM

While needs vary among our several clients with DB2 performance analysis requirements, we believe that the techniques presented in this paper will play a fundamental role in satisfying each client. The basic contribution in this paper is a technique to define performance classes for DB2 plans. For the development DBA, the performance classes will provide performance predictions for structural design alternatives during application development. For the capacity planner, the performance classes provide a structure for defining service level agreements, tracking plan performance and projecting resource requirements.

There is broad testimony concerning the inherent unpredictability of DB2 performance. However, a company with a large strategic DB2 workload necessarily has created an environment for DB2 application development and production which delivers performance within specifications. This is, of course, both complex and expensive to accomplish. The DBAs and capacity planners charged with shaping this environment need all the help they can get from ROTs, tools, etc. The entire production environment is constrained by the strategic importance of DB2. In order to guarantee system performance, plan complexity is severely limited.

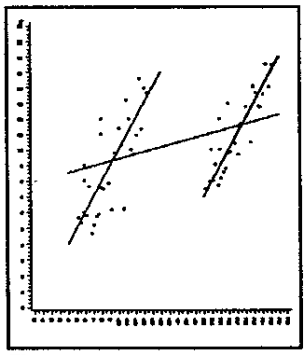


Figure 1

This fact has had a fundamental impact on our selection of the analysis approach outlined in this paper. Only certain design structures for plans can in fact exist within the constraints imposed by a particular production environment. The implication for the analyst is structural constraints on measurement data. Much of the data structure required by classic statistical techniques will not exist in actual measurement data. For example, multidimensional data may structurally degenerate to fall within lower dimensional hyperplanes or may be divided into separated clusters. Trusting the results of 'black box' statistical methods in these instances can lead to invalid conclusions. For instance, consider the linear regression analysis problem of fitting a straight line to the clustered data in Figure #1. When the clustered are fitted jointly, a fictitious line is drawn through the two clusters; the resulting fit is quite different than the proper fit obtained by modeling each cluster separately. The structural properties of the data then require caution in the selection of analysis techniques. The analyst must select from the techniques which will control for or even exploit and interpret the structures within the measurement data.

Specifically, the DB2 production measurement data analyzed for this paper includes executions of a large number of DB2 plans. The plan designs have been carefully developed and tuned to perform according to user requirements and within system specifications. These plans all execute daily with some intensity.

Performance of these plans is subject to the usual sources of variation, namely, concurrent competing workloads, current status of the databases, and specific characteristics of single plan executions. We undertook the analysis with the belief that there does exist 'typical' structural and performance patterns for each DB2 plan. One product of the analysis was anticipated to be a profiling of 'typical' DB2 plan design structures with associated performance profiles or predictions. This type of product is intended for development DBAs who require 'rules of thumb' to trade off between structural design alternatives during DB2 plan development. Capacity planners can use a similar product to classify production DB2 plans into structural design classes and associated performance guarantees. Performance guarantees in terms of response times and costs will depend on the structural class recorded for the particular execution of the plan. That is, the analysis products presented in these papers can be used to define and monitor 'service level agreements' with DB2 users. Trends concerning a DB2 plan breakdown into these performance classes can be a basis for feedback and discussion between production managers and DBAs.

Executions of a well designed DB2 plan should present a homogeneous set or cluster of measurement data. The analysis of this paper was done under the hypothesis that each production DB2 plan has a definable structural profile. Of course, multiple plans can have similar profiles. But the variety of DB2 plan structural profiles identified from measurement data will represent the possible structural design alternatives known to actually exist and operate within specifications for the production environment. These might, in fact, represent the only DB2 plan structural design alternatives which are feasible for the particular production environment. The bottom line fact, regardless, is that these DB2 plan structural designs do exist and operate within system constraints and thus present viable design alternatives to development DBAs.

CLASSIFICATION AND REGRESSION TREES (CART)™

We had to select a classification technique whose results were both accurate and practical. The caveats on data structures enumerated above cast suspicions on the potential accuracy of classical statistical methods. Also, the results of those methods are frequently too intricate to be practically interpreted and applied. A modern computational technique called classification tree analysis was selected. This technique is one of the new computationally intense statistical methods which are made possible through modern computing. The technique was implemented through CART (Classification and Regression Tree) software [BRE-84]. Ten current production DB2 plans were selected for analysis. The problem posed to CART was to discriminate between the ten plans using measurements on structural variables. Measurement data was from the MICS DB2 information area. For purposes of a pilot analysis to proof this approach and for a not too complex example for this paper, we reduced the number of structural variables used in the analysis. The reduction was accomplished through preliminary analyses and was guided by the kinds of measurement information which would relate to design level decisions. The following table summarizes the structural variables presented to CART and CART's assessment of the relative value of each variable in discriminating between the plans.

VARIABLE IMPORTANCE

MICS Variable	Relative Importance
DSUALOCK - IRLM lock requests	100
DSUPGPR0 - DB2 GET PAGE requests	97
DSUAUNLK - IRLM unlock requests	94
DSUQUERY - sum of FETCH and SELECT statements	77
DSUSPRU0 - system page updates	40
DSUSRIO0 - synchronous read requests	35
DSUUPDTE - number of UPDATE statements	27
DSUINSRT - number of INSERT statements	23
DSUNTNUS - authorization identifier changes	17
DSUNTDEL - normal program termination	10
DSUREASN - reason for the accounting record	10
DSUPRFH0 - number of prefetch read requests	4
DSUABORT - number of times work backed out	2
DSUDELET - number of DELETE statements	1
DSUSUSP - suspend count for the user	1
DSUDESC - number of DESCRIBE statements	0
DSUPREP - number of PREPARE statements	0
DSUGRANT - number of GRANT statements	0

CART defined the binary tree presented in Figure #2 in order to classify the DB2 plans through consideration of these variables. We selected a complexity level for the tree which caused it to have 15 terminal nodes. These terminal nodes are defined through 14 splitting nodes. Each splitting node specifies a variable and a splitting value. The splitting variables and values are referenced at the nodes in the Figure. Observations entering that splitting node are split to the left branch if their variable value is less than or equal to the splitting value. Otherwise the observation is split to the right branch. Each of the terminal nodes is associated with one of the ten plans that are being discriminated. The association is presented in the table below.

DB2 PLAN	ASSOCIATED TERMINAL NODES
K	5, 10
L	None
M	3
N	1, 12
O	2
P	8, 13, 15
Q	4, 14
R	6
S	7
T	9, 11

Figure #3 reports how well this tree does at discriminating between the 10 DB2 plans. CART divides the data into a 'learning' sample and a 'test' sample. The test sample is used to validate the modeling done with the learning sample. Accordingly, Figure #3 presents two tables, a learning sample table and a test sample table.

CART has built this tree using four MICS variables (DSUALOCK, DSUGPGR0, DSUUPDTE and DSUQUERY) for splitting. We view locking as a spurious effect rather than a structural design variable. With this view, we will prune the tree to obtain a tree which only splits on the variables DSUGPGR0, DSUUPDTE and DSUQUERY. Towards this end, Figure #4 presents the information available in CART output for each splitting node. Splitting is done in order to reduce a cost parameter associated with the tree. The variable selected for splitting minimizes cost for the resulting tree. The output tells us which other variables were competitors for splitting at that node, together with the cost improvement that such a choice would bring to the tree. In

CLASSIFICATION TREE DIAGRAM
15 TERMINAL NODE TREE INCLUDING SPLITS ON LOCKS

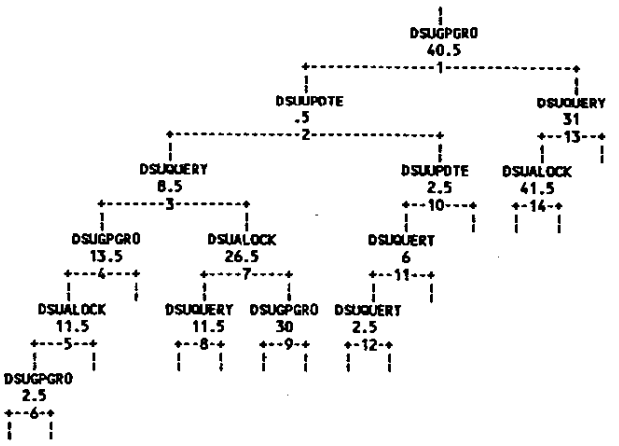


Figure 2

TEST SAMPLE CLASSIFICATION PROBABILITY MATRIX

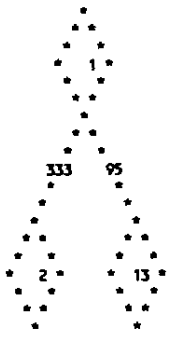
			TRUE DB2 PLAN								
			K	L	M	N	O	P	Q	R	S	T
PDB2	K		.83	.17				.05		.27		
	L											
	M		.07		.83	.08				.07		.07
	N					.23			.05			
	O		.10			.23	.83	.03	.05		.25	.12
	P				.17	.15	.17	.38	.18	.07		
	Q			.67		.23		.47	.73	.27	.13	.05
	R									.20	.25	.02
	S									.13	.38	.02
	T			.17		.08		.08				.72

LEARNING SAMPLE CLASSIFICATION PROBABILITY MATRIX

			TRUE DB2 PLAN								
			K	L	M	N	O	P	Q	R	S	T
PDB2	K		.81	.54				.07		.21	.03	.01
	L											
	M		.05		.70	.07		.01	.03		.03	.04
	N					.40		.01	.02			
	O		.05	.08	.05	.27	.87	.06		.14	.38	.09
	P		.03		.15			.47	.11	.07		
	Q			.38	.10	.10		.27	.85	.24		.01
	R		.03			.10	.13	.04		.28	.09	.01
	S									.07	.47	
	T		.04			.07		.06				.83

Figure 3

NODE INFORMATION



Node 1 was split on variable dsugpgr0
A case goes left if variable dsugpgr0 .le. 4.05e+01
Improvement = 7.1e-02 (C. T. = 1.1e-01)

Node	Cases	Class	Cost
1	428	1	0.82
2	333	1	0.77
13	95	7	0.49

Class	Number Of Cases			Within Node Prob.		
	Top	Left	Right	Top	Left	Right
1	77	76	1	0.18	0.23	0.01
2	13	13	0	0.03	0.04	0.
3	20	16	4	0.05	0.05	0.04
4	30	28	2	0.07	0.08	0.02
5	15	15	0	0.04	0.05	0.
6	70	32	38	0.16	0.10	0.40
7	66	18	48	0.15	0.05	0.51
8	29	27	2	0.07	0.08	0.02
9	32	32	0	0.07	0.10	0.
10	76	76	0	0.18	0.23	0.

	Surrogate	Split	Assoc.	Improve.
1	dsualock s	3.75e+01	0.71	6.1e-02

	Competitor	Split	Improve.
1	dsuaunlk	2.75e+01	6.1e-02
2	dsualock	3.75e+01	6.1e-02
3	dsuupdte	5.00e-01	4.3e-02
4	dsuspru0	5.00e-01	3.9e-02
5	dsuquerY	2.50e+00	3.7e-02

Figure 4

CLASSIFICATION TREE DIAGRAM
13 TERMINAL NODE TREE ELIMINATING SPLITS ON LOCKS

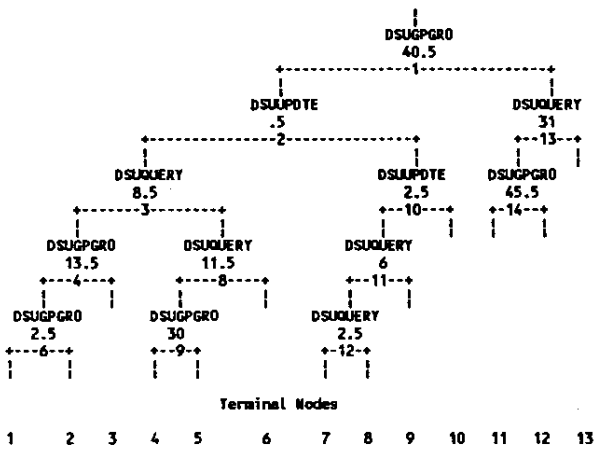


Figure 5

Node 14 was split on variable dsualock
A case goes left if variable dsualock .le. 4.15e+01
Improvement = 1.9e-02 (C. T. = 2.3e-02)

Node	Cases	Class	Cost
14	80	7	0.43
-13	23	6	0.39
-14	57	7	0.26

Figure 6

	Surrogate	Split	Assoc.	Improve.
1	dsuaunlk s	2.95e+01	0.56	1.6e-02

	Competitor	Split	Improve.
1	dsuaunlk	2.95e+01	1.6e-02
2	dsugpgr0	4.55e+01	1.4e-02
3	dsuinsrt	8.50e+00	7.9e-03
4	dsuspru0	2.45e+01	7.9e-03
5	dsuquerY	4.00e+00	6.6e-03

particular, a surrogate splitting variable associated with the selected splitting variable is identified. This is a variable that would accomplish a split of observations close to that provided by the actual splitting variable. The strength of association between the surrogate variable and the splitting variable is indicated. Through analysis of the node splitting information in the CART output, we trimmed the original tree to get the tree presented in Figure #5. For instance, from Figure #6, which presents splitting node 14 information from CART output, we learn that a competitive split can be made using a value of 45.5 for the variable DSUGPGRO. The new tree has 13 terminal nodes. These 13 terminal nodes define 13 structural classes for DB2 plan design.

ANALYSIS OF THE DB2 STRUCTURAL CLASSES

A graphic picture of the 13 structural classes is presented in Figure #7. This graphic is produced using the SAS/GRAPH® G3D Procedure. In order to provide a better visualization of these 13 classes, Figures #8, #9, #10 and #11 present separate pictures of groups of these classes. A scatter plot of measurement observation values is overlaid onto the 13 class regions in Figure #12 to give a further picture of the character of the data. Note that a large amount of the data falls into the plane where DSUUPDTE is 0.

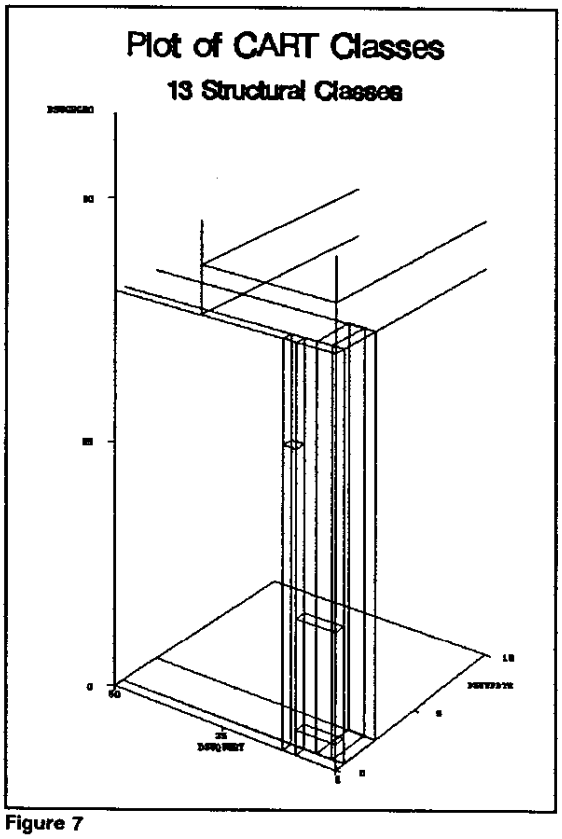


Figure 7

Now that structural classes for DB2 plans have been defined, we investigate the performance characteristics of these classes. Performance is quantified as response time and CPU time. A statistical graphic, called a 'Box & Whisker Diagram', is used in many of the descriptive graphics presented in the rest of this paper. This graphic presents a quick visual summary of percentile statistics for data. A box covers the interquartile range (from the 25th through the 75th percentiles) and whiskers extend out to the minimum value on the left and the maximum value on the right. Figure #13 demonstrates the association between this summary and a set of data. Box & whisker diagrams are used extensively for management reporting [GAN-87, GAN-88, GAN-90]. Figure #14 uses box & whisker diagrams to show the pattern of Class 2 CPU time across the 13 structural classes. This pattern establishes a performance hierarchy among the 13 classes. Other than Class 10, the graphic in Figure #14 sorts the classes by median Class 2 CPU time. Class 10 has relatively few data points and its statistical summarization is not as well defined as most of the other classes. Figures #15 and #16 present similar performance profiles in terms of Class 1 CPU time and elapsed time, respectively. Observe that the performance hierarchy established in Figure #14 is reiterated in each of the other Figures.

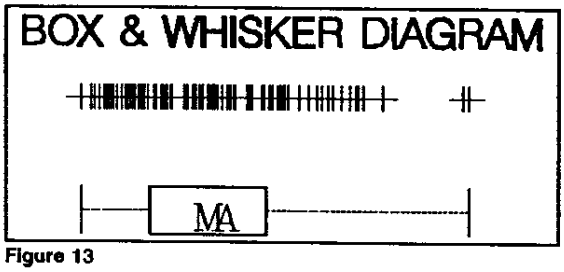


Figure 13

The performance hierarchy imposed on the 13 structural classes for DB2 plans shows the thresholds for increases in queries, updates and get page requests at which performance changes. These thresholds are meaningful since they are based on measurements from existing plans executing in the production environment. Figures #17 and #18 use the box & whisker diagrams to present 'profiles' of classes 6 and 11, respectively. The profiles include summaries of performance measurements, of structural variable measurements used for classification, and of additional structural variable measurements. These profiles present a contrast between performance predicted for two different design structures. The profile presents a summary of the levels of DB2 design complexity and performance for a class of DB2 plans.

Measurements for a particular DB2 plan might result in observations across a number of classes. One reason is that a single plan can be used to execute one of a variety of functions. Each observation in our analysis is a single execution of a plan, i.e., a DB2 thread. For instance, Figures #19, #20 and #21 show CPU time and elapsed time summaries for Plan P across several classes. Plan P, in fact, is associated with about ten different functions. A time series of measurement summaries for this plan would determine the robustness of the plans execution across these classes. A DBA could use this type of aid in determining when a particular plan was deviating from an established baseline, and perhaps even determine reasons for the deviation. Further, since the structural measurements of a plan's execution are directly related to performance, service level agreements for performance parameters can be framed within the context of established structural classes. Figure #22 presents a profile of Plan P measurements restricted to Class #12 alone. This could be considered the 'typical' structure and performance profile of Plan P.

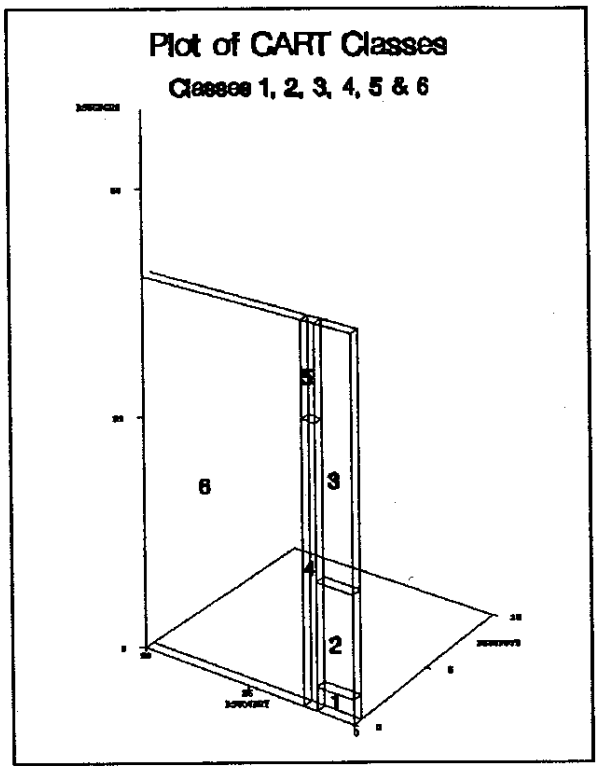


Figure 8

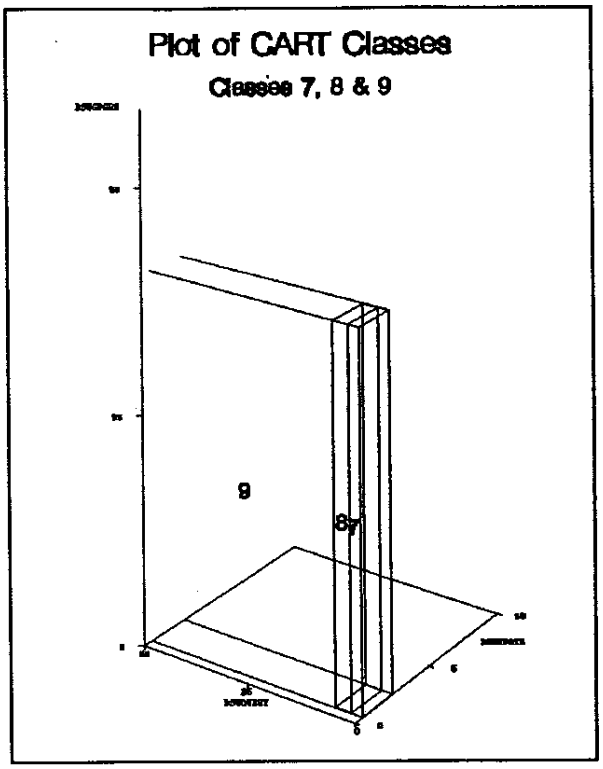


Figure 9

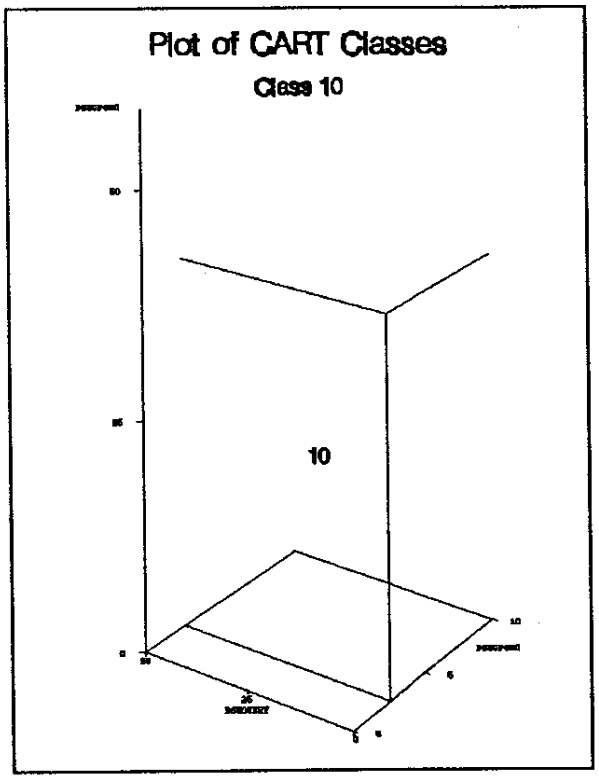


Figure 10

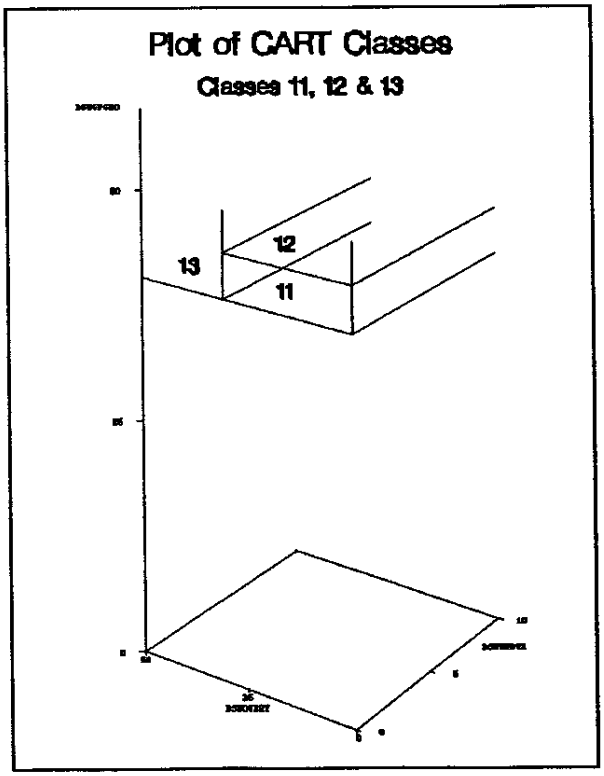


Figure 11

USING SAS TO ANALYZE CART OUTPUT

Figures #4 and #6 presented the splitting node information found in typical CART output. A long CART output describing the splitting logic for a large tree is an unwieldy object to digest. Additionally, an analyst who wants to put CART terminal node information into a SAS data set faces writing exhaustive logic statements. We wrote SAS programs that read the CART output and formed a SAS data set of the output information. We could then efficiently analyze the CART output information to understand and prune trees. For example, Figure #23, which gives surrogate splitting variable information for the tree in Figure #2, was generated from such a SAS data set.

Additionally, two SAS programs were written to use either the SAS data set of CART output or a very simple tree describing raw data file and produce SAS programs which would put CART terminal node information onto the original SAS data set of DB2 measurement data.

CONCLUSION

In closing, our work has achieved two important objectives. The first is that we have shown the utility of a modern classification tree software tool extended with the use of SAS Institute's products in identifying and modelling DB2 Plan structures. SAS software is a very powerful language with excellent analysis and reporting capabilities. This work exploited SAS Institute's product capabilities to provide additional insight into the results of another analysis tool.

Second, and more importantly, the combination of the two tools helped identify the performance hierarchy inherent in a DB2 workload. This hierarchy could be employed in developing ROTs that can be utilized early in development to size a new application system.

REFERENCES

[BRE-84] Breiman, L., et al, *Classification and Regression Trees*, Wadsworth, 1984.

[GAN-87] Gantz, D., Gonzalez, C., and Moore M., "Capacity Planning for a Multi-Processor Environment Supporting CICS Transactions," *Proceedings of the CMG'87*, Orlando, 1987.

[GAN-88] Gantz, D., "Tutorial on Statistical Methods for Analysis and Reporting for Capacity Planning," *Proceedings of the CMG'88*, Dallas, 1988.

[GAN-90] Gantz, D., "Management Reporting for Capacity Planning and Computer Performance Evaluation," *SUGI'15 SAS Users Group International 15th Annual Conference Proceedings*, Nashville, 1990.

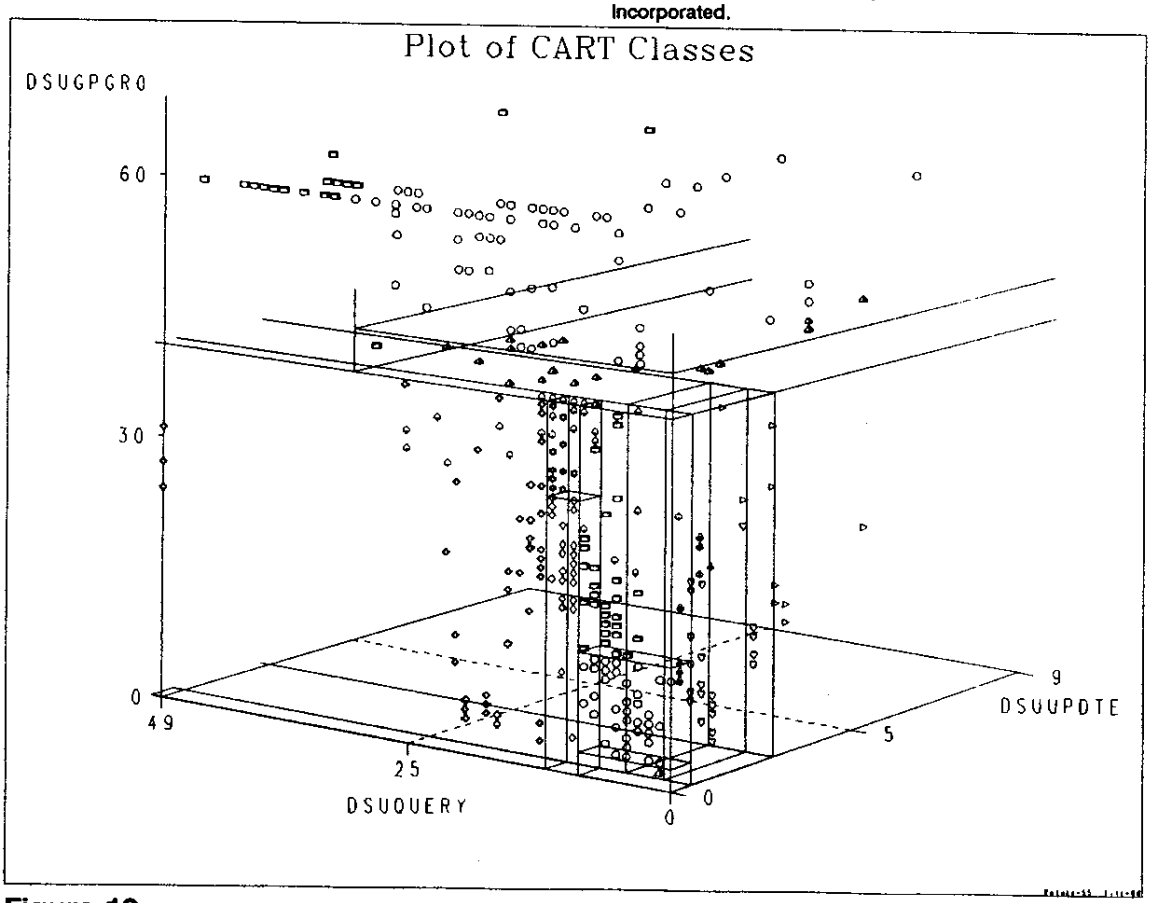


Figure 12

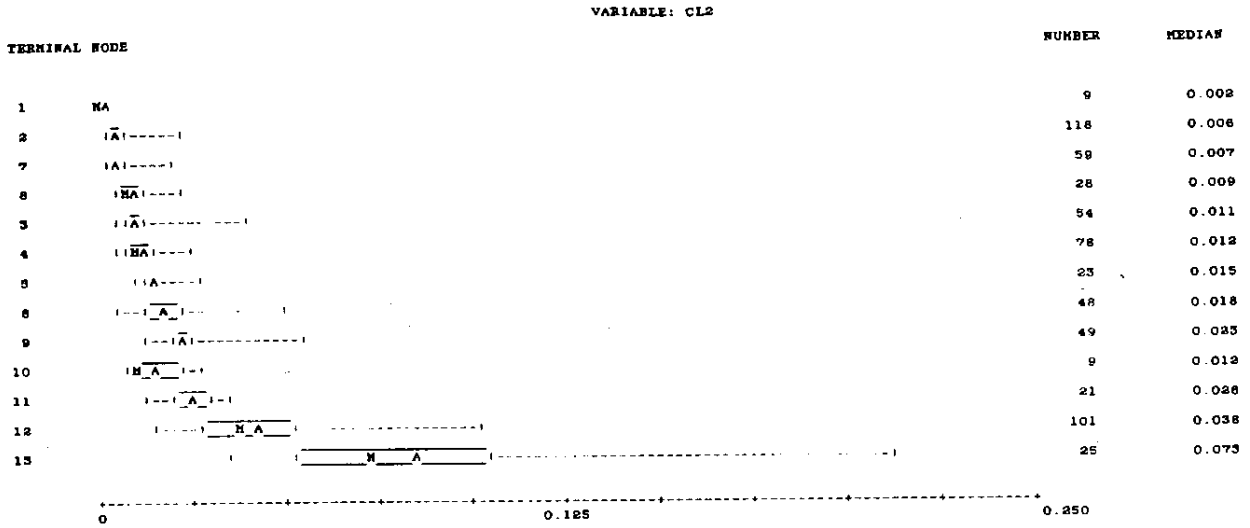


Figure 14

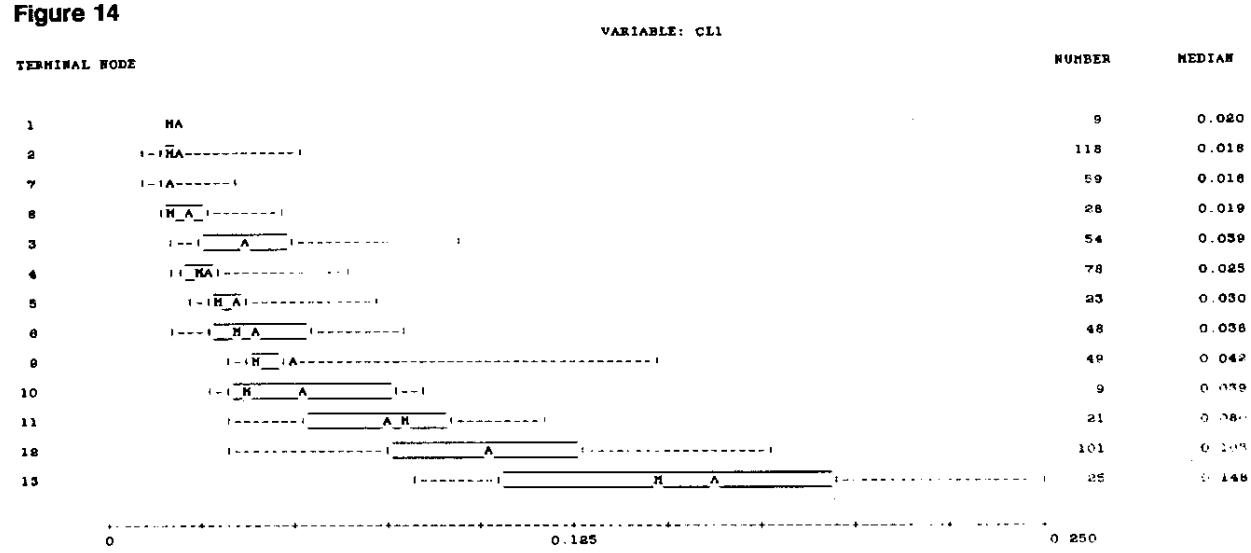


Figure 15

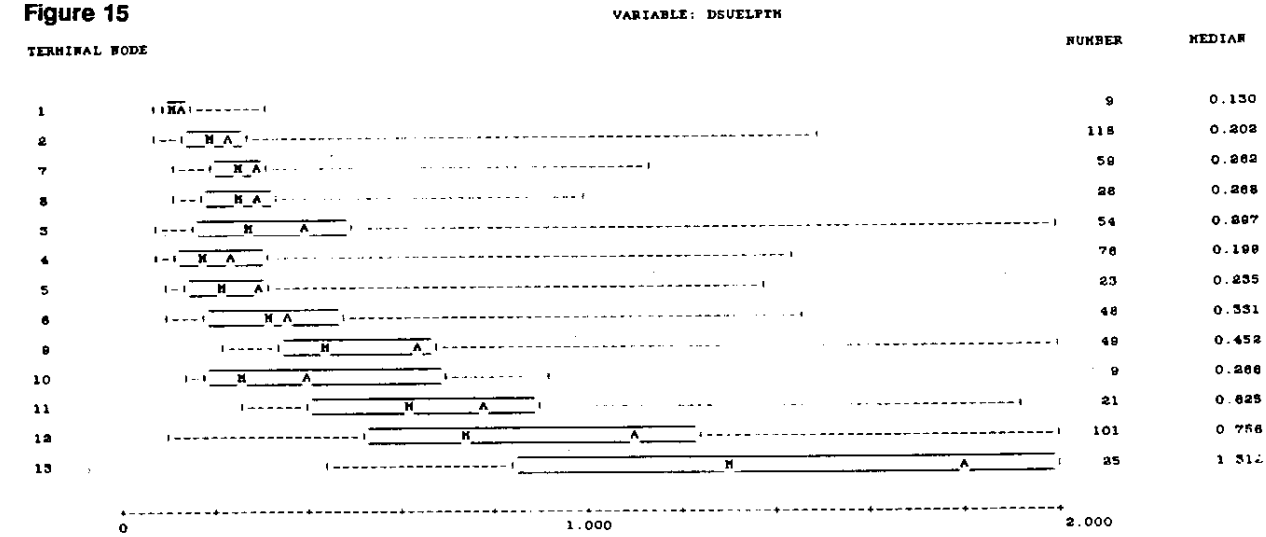


Figure 16

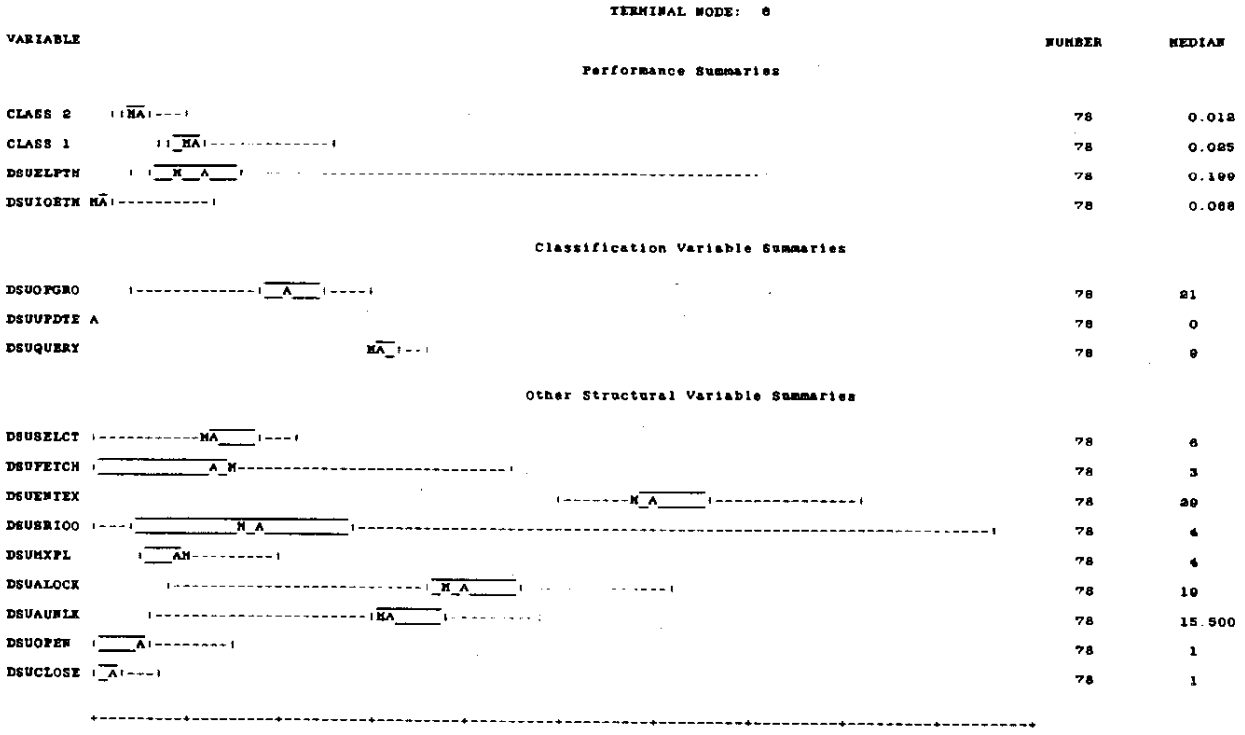


Figure 17

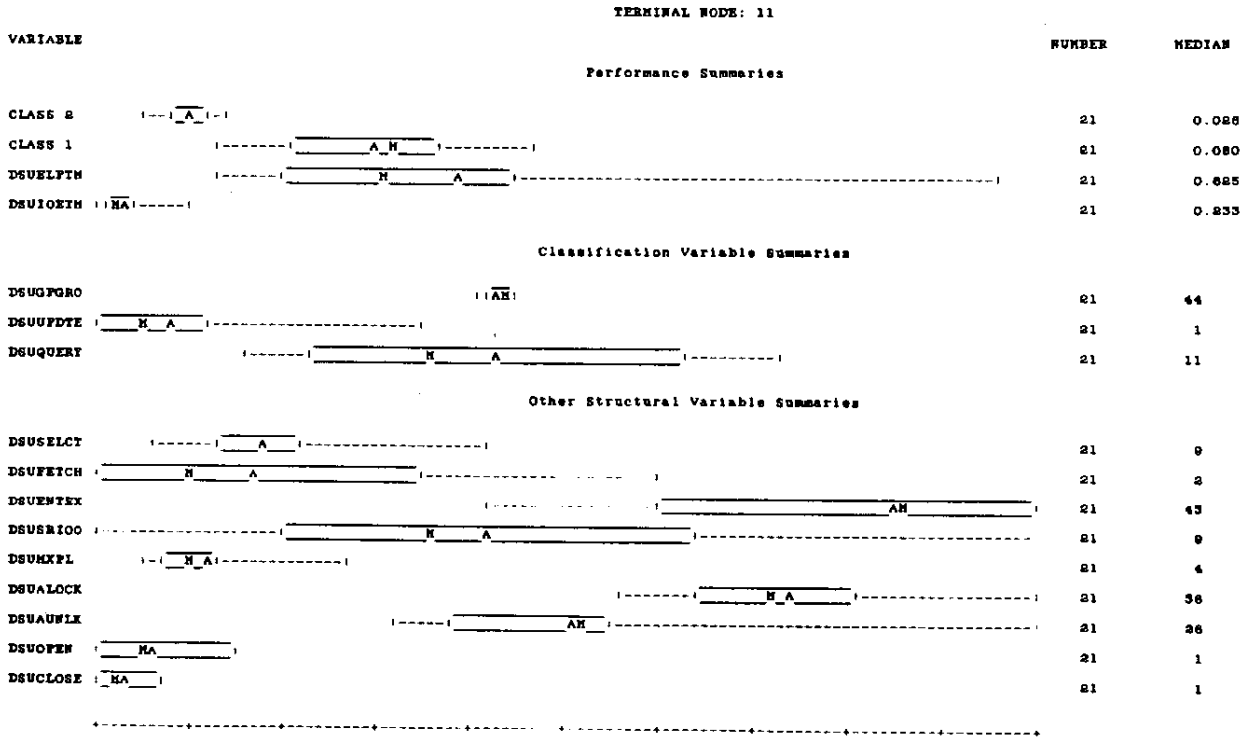


Figure 18

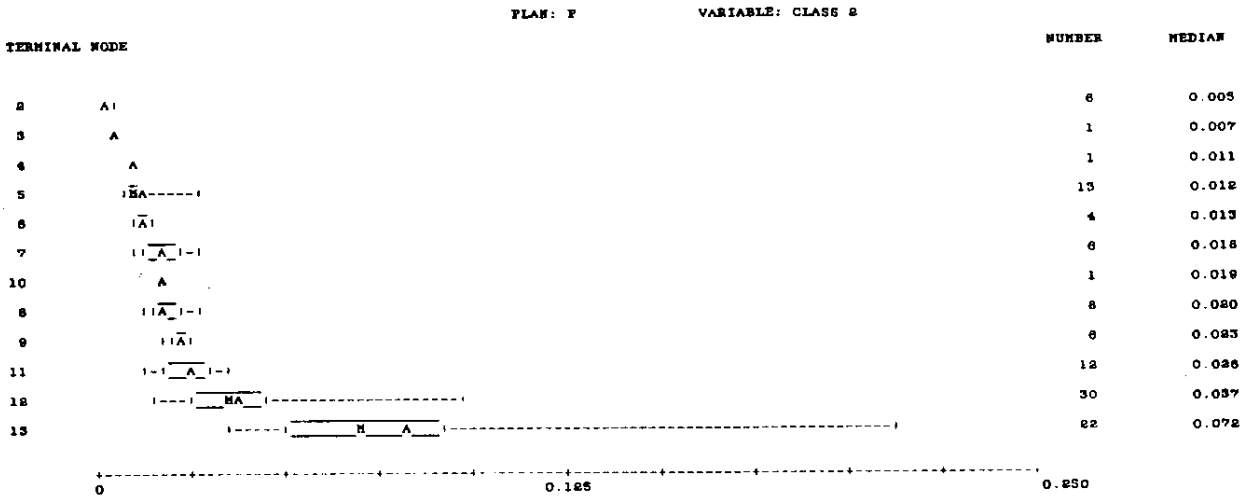


Figure 19

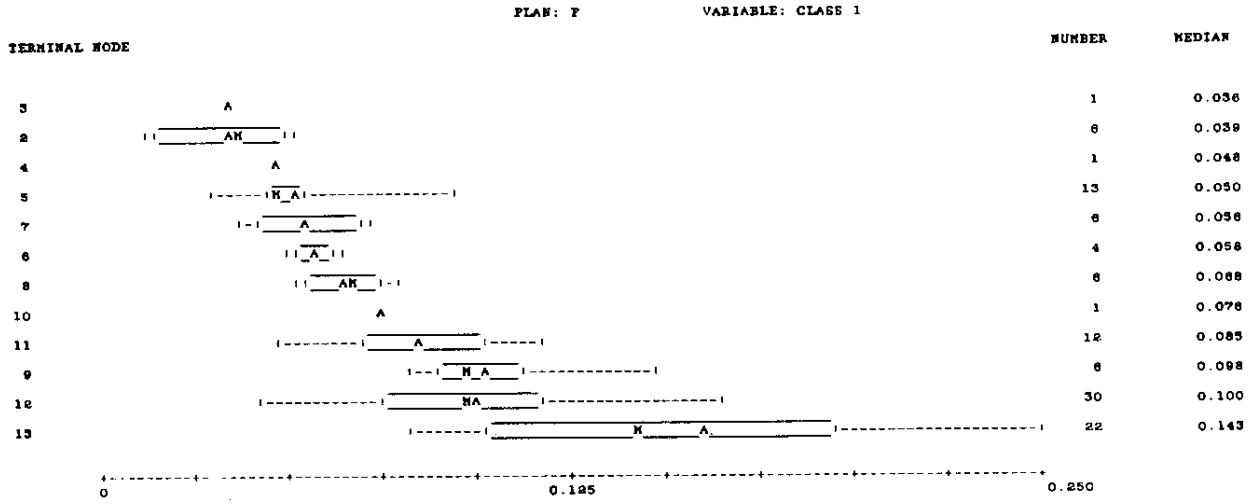


Figure 20

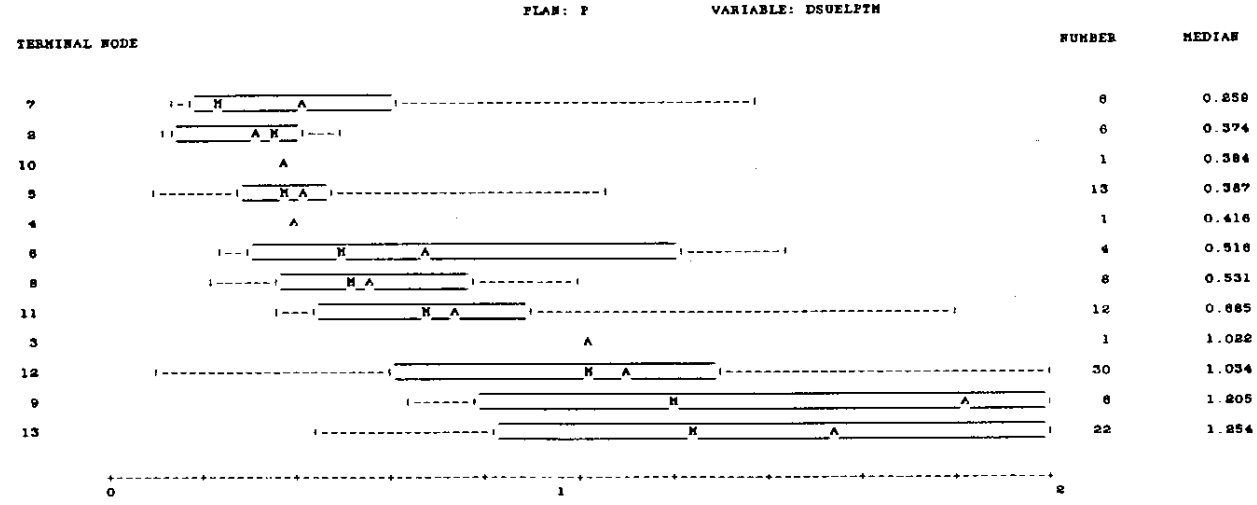


Figure 21

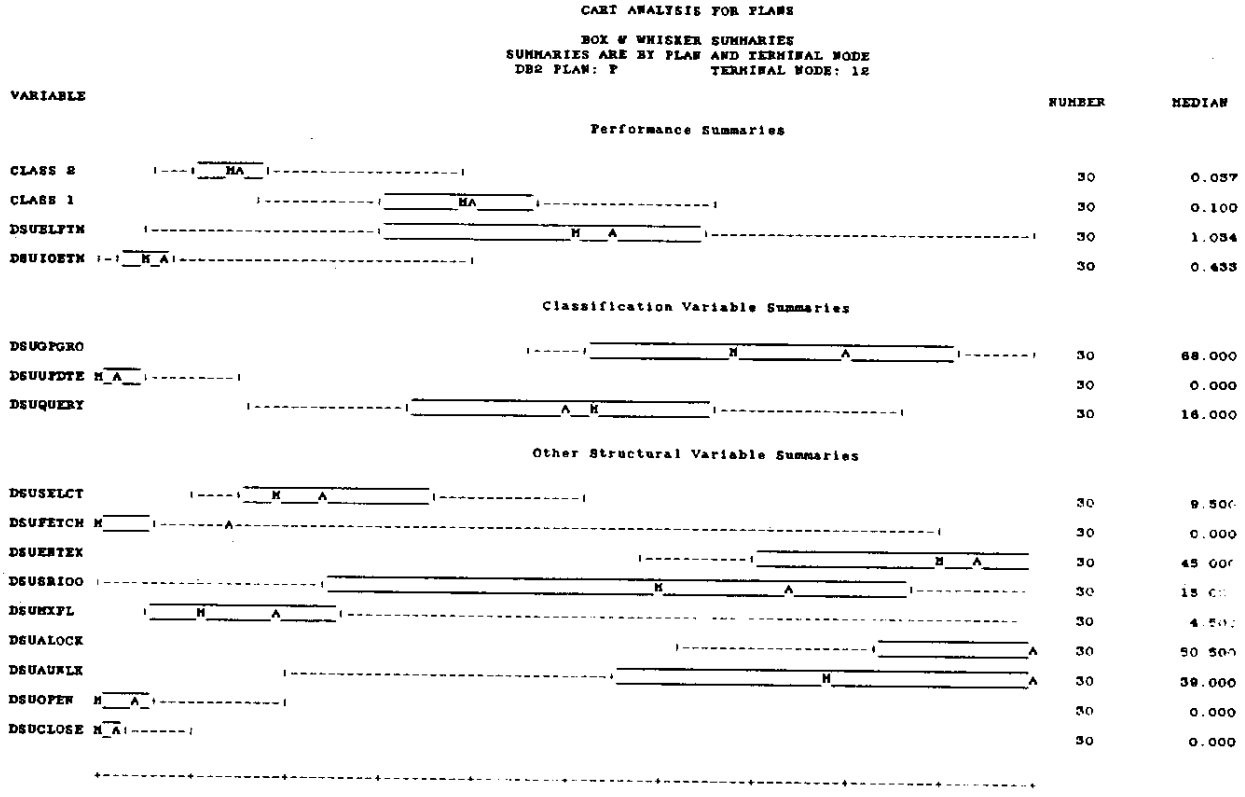


Figure 22

NODE SUMMARIES

CART COMPETITIVE SPLITS

15 TERMINAL NODE CLASSIFICATION TREE

Splitting MICS Variable=dsualock

Surrogate Splitting Variable	Cost Improvement Ratio	Association between Variables	Splitting Node
dsuaunlk	1.00000	0.96	5
dsuaunlk	0.84211	0.56	14
dsuquery	0.89655	0.42	7

Splitting MICS Variable=dsugpgr0

Surrogate Splitting Variable	Cost Improvement Ratio	Association between Variables	Splitting Node
dsualock	1.00000	0.88	4
dsualock	0.83333	0.85	6
dsualock	0.76471	0.66	9
dsuaunlk	0.85915	0.71	1

Splitting MICS Variable=dsuquery

Surrogate Splitting Variable	Cost Improvement Ratio	Association between Variables	Splitting Node
dsualock	0.86667	0.40	13
dsuaunlk	1.00000	0.94	11
dsuaunlk	0.82857	0.91	12
dsuaunlk	0.47826	0.41	8
dsugpgr0	0.87097	0.67	3

Splitting MICS Variable=dsuupdte

Surrogate Splitting Variable	Cost Improvement Ratio	Association between Variables	Splitting Node
dsuquery	0.933333	.	10
dsuspru0	0.904762	0.92	2

Figure 23

Advisory Board

Our Advisory Board is composed of a network of accomplished experts and distinguished thought leaders from a variety of backgrounds, who have generously agreed to provide ongoing help to Intensity Analytics in different areas. Whether as a whole, or on an individual basis, their consultation and guidance is invaluable to the company and our strategy.

Board Members

Steve Francis, President & Chairman, RAF Technology

Steve is a serial entrepreneur and gifted product strategist with a 30-year career ranging from hands-on software engineer to public company CEO. After graduating from Stanford University in Electrical Engineering/Computers with both Tau Beta Pi and Phi Beta Kappa honors, Steve helped pioneer the field of omnifont optical character recognition in a variety of roles at Calera Recognition Systems. He co-founded Pixel Translations, developers of ISIS (Image and Scanner Interface Specification) international standard for document scanner control. More recently, he was CEO of Avtrex, leading its software licensing business before the acquisition of Avtrex by Intel.

Dr. Donald T. Gantz, Emeritus Professor

Donald T. Gantz, PhD, is an Emeritus Professor, Founding Chair of the Department of Applied Information Technology, and past Dean of the Volgenau School of Engineering at George Mason University. Dr. Gantz has developed cutting edge methodologies for the quantification and analysis of handwriting and has applied these methodologies to multi-language document exploitation and biometric identification, including matching software and scoring algorithms for fingerprint identification. He helped pioneer the use of statistical and geographic information systems to analyze the relationship between disease and socioeconomic factors, as well as transportation demand management and traffic mitigation. He has lectured on methods for surveillance systems to detect levels of infection due to a natural epidemic or bioterrorism threat, and is one of the top experts used by FBI in validating fingerprint analysis.

Honorable Myron Greenberg

Minnesota District Court Judge (ret). After nearly 30 years on the bench, including eight years on the Minnesota Supreme Court Committee on Alternative Dispute Resolution, Mickey is now a distinguished mediator in complex legal cases. He has also taught and presented programs on settlement techniques for the Minnesota Judicial College, the American Bar Association Section on Alternative Dispute Resolution, the Wisconsin Bar Association, the International Academy of Mediators and the Litigation Management College at Northwestern University. He has extensive international experience working on behalf of the United States, having served as a Trial Judge and Appellate Judge on the Court of Bosnia and Herzegovina handling organized crime and corruption cases, and training judges and teachers in Kabul, Afghanistan. Along with IA's CEO, he lead IA's Seed Round funding – 40 investors; $2.9M raised.

Brian Moore, VP, PJM Advisors

Mr. Moore is Vice President of PJM Advisors, a family investment office specializing in the lower middle market. Brian began his career as a management strategy consultant with IBM Global Business Services specializing in finance transformation and corporate strategy project. Spending time in the automotive, industrial, and financial services sectors, Brian managed integration and operations teams both in the U.S. and abroad while delivering value to clients in various situations including bankruptcy and reorganizations, accelerated growth and transactions. He currently sits on the international board of Boys Hope Girls Hope and Aerovalve.

H. John Oechsle, President and CEO at Swiftpage

John is an international business executive and seasoned speaker whose strategic insights about market creation and product innovation have helped transform businesses around the world. His track record of building highly profitable and sustainable revenue growth for emerging companies and established global leaders includes key leadership positions in the U.S., Europe, Asia and Latin America with Johnson & Johnson, Kellogg Company, LandAmerica Financial, IHS, and DigitalGlobe. In addition, John created the vision, strategy and operational plans that helped triple revenue growth for two companies and resulted in their successful IPOs. John is an advocate for technology and education in

many industry awards, including the CTA's Bob Newman Lifetime Achievement Award for Outstanding Contribution to the Community.

Reginald Steer

Mr. Steer is a well-known and highly respected litigator who has tried some of the largest cases in the country, including those with billions of dollars at stake. He is an emeritus partner in Akin Gump Strauss Hauer & Feld, LLP, a prestigious international law firm, where he practiced complex commercial litigation for 13 years involving claims of antitrust violations, contract disputes, fraudulent business practices, unfair competition and intellectual property disputes. Before joining Akin Gump, Mr. Steer was a partner in Pillsbury Madison & Sutro for over twenty years, where he was a member of the executive committee, head of the technology committee and co-chair of the firm-wide patent practice in charge of litigation. He is a fellow of the American College of Trial Lawyers, which is limited to the top 1% of attorneys. Reg has taught in trial skills programs and on the subject of Robinson-Patman law, and was a contributor to the ABA Antitrust Section's Rule of Reason Monograph. He formerly served in the U.S. Army Judge Advocate General's Corps.

Alex Wellins, Co-Founder, The Blueshirt Group

As Co-Founder and Managing Director of The Blueshirt Group, Alex brings 30 years of investor relations, financial media relations and capital markets experience to the Intensity Analytics advisory board. After earning his B.A. from the University of Michigan, he spent his first years on the west coast providing marketing communications counsel to emerging Silicon Valley software startups. Alex then spent five years at Morgen-Walke Associates focusing on financial media relations then transitioned to manage investor relations programs for the largest group of technology clients at the firm. Alex has been a strategic advisor on the IPOs of many Internet, software and digital media companies and has participated in numerous high-profile M&A transactions in the technology sector.

In Memoriam

Intensity Analytics owes a great deal to two individuals who provided key assistance during our start-up phase, and were instrumental in the company's early days. While these two beloved friends and mentors are no longer with us, they remain close in our hearts.

Richard (Dick) Lemons

Dick had over 40 years experience in information technology and always with the goal of eliminating the mountains of paper in favor of emerging digital technologies. For over 12 years, he was Director of the NASA Scientific and Technical Information Facility (NASA-STIF) where the first ever online retrieval system was built. He had a breadth of experience building large business operations, such as GE's GEISCO unit, and a division of Informatics with over 1,500 people, as well as delivering major systems to NLM (MEDLINE and TOXLINE), and EPA (ENVIRON). Dick provided creative management, angel investment capital, and consulting services to start-up companies for more than 22 years. He is a Founding Member of the Westlake Chapter of Keiretsu Forum, and was an accomplished Bluegrass musician.

Jerry Marterella

In his long career, Jerry held senior manager and executive sales positions at Northrup Grumman, Integic, Computer Science Corporation (CSC), Digital Equipment Corporation, and Xerox. He was able to increase revenue at an unprecedented rate, in each of his positions. Jerry was able to grow the consulting division of CSC from $400 million to $850 million in annual revenue over 2 years without acquisitions. Recognizing the need for senior executives to grow professionally, he co-founded the Potomac Region Executive Leadership Breakfast, an organization that has grown to over 500 members. Jerry was also an award-winning wine maker, having won numerous Virginia Governor Cup Awards, with a tasting room that was often voted one of the best in Virginia.

Innovative Surveillance Methods for Rapid Detection of Disease Outbreaks and Bioterrorism: Results of an Interagency Workshop on Health Indicator Surveillance

Julie A. Pavlin, Farzad Mostashari, Mark G. Kortepeter, Noreen A. Hynes, Rashid A. Chotani, Yves B. Mikol, Margaret A. K. Ryan, James S. Neville, Donald T. Gantz, James V. Writer, Jared E. Florance, Randall C. Culpepper, Fred M. Henretig, Patrick W. Kelley

Am J Public Health. 2003 Aug; 93(8): 1230–1235.

PMCID: PMC1447943

Article PubReader PDF–186K Citation

Is Cited by the Following 12 Articles in this Archive:

A Practitioner-Driven Research Agenda for Syndromic Surveillance

Richard S. Hopkins, Catherine C. Tong, Howard S. Burkom, Judy E. Akkina, John Berezowski, Mika Shigematsu, Patrick D. Finley, Ian Painter, Roland Gamache, Victor J. Del Rio Vilas, Laura C. Streichert

Public Health Rep. 2017 Jul-Aug; 132(1 Suppl): 116S–126S. Published online 2017 Jul 10. doi: 10.1177/0033354917709784

PMCID: PMC5676517

Article PubReader PDF–320K Citation

Infectious Disease Surveillance in the Big Data Era: Towards Faster and Locally Relevant Systems

Lone Simonsen, Julia R. Gog, Don Olson, Cécile Viboud

J Infect Dis. 2016 Dec 1; 214(Suppl 4): S380–S385. Published online 2016 Nov 14. doi: 10.1093/infdis/jiw376

PMCID: PMC5144901

Article PubReader PDF–253K Citation

Using Undergraduate Researchers to Build Vector and West Nile Virus Surveillance Capacity

Grant Hokit, Sam Alvey, Jennifer M. O. Geiger, Gregory D. Johnson, Marni G. Rolston, Daniel T. Kinsey, Neva Tall Bear

Int J Environ Res Public Health. 2013 Aug; 10(8): 3192–3202. Published online 2013 Jul 31. doi: 10.3390/ijerph10083192

PMCID: PMC3774432

Article PubReader PDF–300K Citation

Using Chief Complaints for Syndromic Surveillance: A Review of Chief Complaint Based Classifiers in North America

Mike Conway, John N. Dowling, Wendy W. Chapman

J Biomed Inform. Author manuscript; available in PMC 2014 Aug 1.
Published in final edited form as: J Biomed Inform. 2013 Aug; 46(4): 734–743. Published online 2013 Apr 17.
doi: 10.1016/j.jbi.2013.04.003

PMCID: PMC3741452

Article PubReader PDF–695K Citation

Comparison: Flu Prescription Sales Data from a Retail Pharmacy in the US with Google Flu Trends and US ILINet (CDC) Data as Flu Activity Indicator

Avinash Patwardhan, Robert Bilkovski

PLoS One. 2012; 7(8): e43611. Published online 2012 Aug 30. doi: 10.1371/journal.pone.0043611

PMCID: PMC3431370

Article PubReader PDF–184K Citation

Telehealth Ontario Detection of Gastrointestinal Illness Outbreaks

Jaelyn M. Caudle, Adam van Dijk, Elizabeth Rolland, Kieran M. Moore

Can J Public Health. 2009 Jul; 100(4): 253–257. Published online 2009 Jul 1. doi: 10.1007/BF03403942

PMCID: PMC6973921

Article PubReader PDF–293K Citation

A multi-data source surveillance system to detect a bioterrorism attack during the G8 Summit in Scotland

N. MEYER, J. McMENAMIN, C. ROBERTSON, M. DONAGHY, G. ALLARDICE, D. COOPER

Epidemiol Infect. 2008 Jul; 136(7): 876–885. Published online 2007 Aug 3. doi: 10.1017/S0950268807009132

PMCID: PMC2870879

Article PubReader PDF–163K Citation

Syndromic surveillance: is it a useful tool for local outbreak detection?

Kirsty Hope, David N Durrheim, Edouard Tursan d'Espaignet, Craig Dalton

J Epidemiol Community Health. 2006 May; 60(5): 374–375. doi: 10.1136/jech.2005.035337
PMCID: PMC2563979
Article PubReader PDF–114K Citation

Using Ontario's "Telehealth" health telephone helpline as an early-warning system: a study protocol
Elizabeth Rolland, Kieran M Moore, Victoria A Robinson, Don McGuinness
BMC Health Serv Res. 2006; 6: 10. Published online 2006 Feb 15. doi: 10.1186/1472-6963-6-10
PMCID: PMC1431529
Article PubReader PDF–262K Citation

Public Health and Terrorism Preparedness: Cross-Border Issues
Debra Olson, Aggie Leitheiser, Christopher Atchison, Susan Larson, Cassandra Homzik
Public Health Rep. 2005; 120(Suppl 1): 76–83. doi: 10.1177/00333549051200S115
PMCID: PMC2569992
Article PubReader PDF–77K Citation

SARS Alert Applicability
Emerg Infect Dis. 2004 Aug; 10(8): 1512–1514. doi: 10.3201/eid1008.040221
PMCID: PMC3320393
Article PubReader PDF–84K Citation

Respiratory medical societies and the threat of bioterrorism
T O'Riordan, G Smaldone
Thorax. 2004 Mar; 59(3): 265–267. doi: 10.1136/thorax.2003.015321
PMCID: PMC1746971
Summary PDF–91K Citation

LIST OF PRESENTATIONS

[Presentations received have been highlighted]

SESSION 4

SESSION 5

SESSION 6

SESSION 7

DEMONSTRATION AND POSTER ABSTRACTS

GOVERNMENT PANEL

"Current State of Affairs," *Dan Lopresti, Lehigh University, Bethlehem, PA* - (Lopresti-SACH06-Panel.pdf)

"Language Translation Technologies: How ARL is Meeting the Challenges," *Luis Hernandez, Steve LaRocca, Ph.D., U.S. Army Research Laboratory, USA* - (LaRocca-SACH06-Panel.pdf)

"Philosophy," *Sargur N. Srihari, CEDAR, University at Buffalo, Amherst, New York* -- (Srihari-SACH06-Panel.pdf)

On Parametric Models for Pairwise Comparisons

with Applications to Estimation of Random Match Probabilities

Donald Gantz, Department of Applied IT, George Mason University
John Miller, Department of Statistics, George Mason University
Chris Saunders, Department of Applied IT, George Mason University

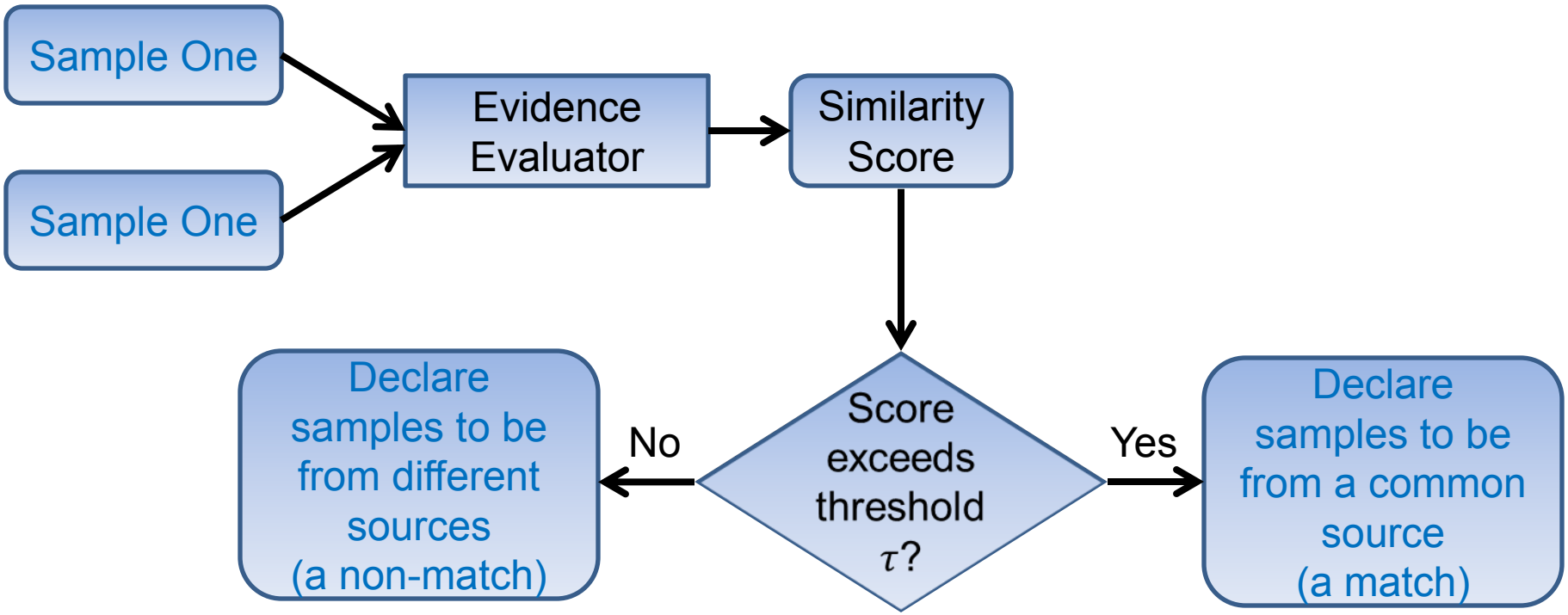
August 11, 2011

Trace Evidence Symposium
Kansas City, MO

Acknowledgement and Disclaimer

The research detailed in this presentation was supported in part by Award No. 2009-DN-BX-K234 awarded by the National Institute of Justice, Office of Justice Programs, US Department of Justice. The opinions, findings, and conclusions or recommendations expressed in this publication are those of the authors and do not necessarily reflect those of the Department of Justice.

A Common Situation in Forensics



The similarity score is a numerical value measuring association of the samples, with higher values suggesting that the samples come from a common source.

A Sampling Model

Suppose we assume that we have some population of objects which can lead to samples like those on the preceding slide.

A very important consideration of the evidence evaluation method is to know the probability that two randomly sampled objects from different sources would lead (erroneously) to a conclusion of a match. This quantity is called the random match probability.

(The goal of this presentation is to obtain a valid upper confidence limit for the random match probability based on a sample.)

One way to assess the random match probability is to obtain a random sample of n objects known to have each come from a different source and calculate the similarity score for each pair of objects. This is called taking all pairwise comparisons.

Let us call the score resulting from comparing sample i to sample j

$$s_{ij}.$$

There are $N = n(n-1)/2$ possible comparisons*.

* We are assuming that $s_{ij} = s_{ji}$. That is the similarity of object i to object j is the same as the similarity of object j to object i.

A Plausible Model for the Correlation Structure of the s_{ij}

The correlation between s_{ij} and s_{kl} should be:

- One if there are two subscripts in common
- A positive quantity if there is one subscript in common
- Zero if there are no subscripts in common

An Important Quantity

The mean of the scores from all possible pairs of different source objects in the population is a very important quantity in calculations of the random match probability based on scores. We will call this parameter θ.

A plausible estimate for this quantity is the mean of the N scores from the pairwise comparisons in the sample.

It will be important to have proper estimates of the variability of this mean in forming upper confidence limits for the random match probability.

Issues

Some researchers simply ignore the correlation structure and proceed as if there is a sample of N independent scores.

Other researchers believe that the correlation structure forces one to use only uncorrelated pairs (such as s_{12}, s_{34}, s_{56}, etc.).

We will show that it is possible to account for the correlation structure to create a confidence limit for the random match probability.

A Mathematical Model for a Score

We assume that

$$s_{ij} = \theta + a_i + a_j + e_{ij},$$

where θ is an unknown parameter, the a_i are i.i.d. $N(0, {\sigma_a}^2)$ $i = 1, \ldots, n$, and the e_{ij} are i.i.d. $N(0, {\sigma_e}^2)$, $i = 1, \ldots n-1; j = i+1, \ldots n$.

Our goal is to use this model to form a valid upper confidence limit for the random match probability.

The Model in Matrix Terms

We can rewrite this model in matrix terms by writing the scores (s_{ij}) in lexicographic order as a vector $\mathbf{y}$. The errors (e_{ij}) are listed in the same order and the a's are listed in order of their subscripts. There is a design matrix $\mathbf{P}$ (for pairwise) which has N rows and n columns. $\mathbf{P}$ is mostly composed of zeroes but has a one in the ith and jth columns for the row corresponding to s_{ij}.

Thus our model becomes

$$\mathbf{y} = \theta \mathbf{1}_N + \mathbf{Pa} + \mathbf{e},$$

where $\mathbf{y}$ and $\mathbf{e}$ are as described above, $\mathbf{a}$ is the vector of the a_i, and $\mathbf{1}_N$ is an N by 1 vector of ones.

The Expected Value and Covariance Matrix of the Vector of Scores

The N by 1 expected value of the vector of scores is just $\theta \mathbf{1}_N$.

The N by N covariance matrix of the vector of scores is

$$\boldsymbol{\Sigma} = {\sigma_e}^2 \mathbf{I}_N + {\sigma_a}^2 \mathbf{P}\mathbf{P}'.$$

We can convert the covariance matrix to a correlation matrix if we wish. The resulting correlation matrix contains mostly zeroes but has non-zero correlations of $r = {\sigma_a}^2/({\sigma_e}^2 + 2{\sigma_a}^2)$ in some positions.

Our Model Expressed Using Vectors and Matrices for $n = 8$

$$y = \begin{Bmatrix} S12 \\ s13 \\ s14 \\ s15 \\ s16 \\ s17 \\ s18 \\ s23 \\ s24 \\ s25 \\ s26 \\ s27 \\ s28 \\ s34 \\ s35 \\ s36 \\ s37 \\ s38 \\ s45 \\ s46 \\ s47 \\ s48 \\ s56 \\ s57 \\ s58 \\ s67 \\ s68 \\ s78 \end{Bmatrix} \quad P = \begin{Bmatrix} 1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 1 & 0 & 1 & 0 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 1 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 1 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 & 1 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 & 0 & 1 & 0 \\ 1 & 0 & 0 & 0 & 0 & 0 & 0 & 1 \\ 0 & 1 & 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 1 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 1 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 & 1 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 & 0 & 1 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 & 0 & 1 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 & 1 \end{Bmatrix}$$

Correlation Matrix for n = 8

$$r = \sigma_a{}^2/(\sigma_e{}^2 + 2\sigma_a{}^2)$$

1	r	r	r	r	r	r	r	r	r	r	r	0	0	0	0	0	0	0	0	0	0	0	0	0	0
r	1	r	r	r	r	r	r	0	0	0	0	0	r	r	r	r	r	0	0	0	0	0	0	0	0	0
r	r	1	r	r	r	r	0	r	0	0	0	0	r	0	0	0	0	r	r	r	r	0	0	0	0	0
r	r	r	1	r	r	r	0	0	r	0	0	0	0	r	0	0	0	r	0	0	0	r	r	r	0	0	0
r	r	r	r	1	r	r	0	0	0	r	0	0	0	0	r	0	0	0	r	0	0	r	0	0	r	r	0
r	r	r	r	r	1	r	0	0	0	0	r	0	0	0	0	r	0	0	0	r	0	0	r	0	r	0	r
r	r	r	r	r	r	1	0	0	0	0	0	r	0	0	0	0	r	0	0	0	r	0	0	r	0	r	r
r	r	0	0	0	0	0	1	r	r	r	r	r	r	r	r	r	r	0	0	0	0	0	0	0	0	0	0
r	0	r	0	0	0	0	r	1	r	r	r	r	r	0	0	0	0	r	r	r	r	0	0	0	0	0	0
r	0	0	r	0	0	0	r	r	1	r	r	r	0	r	0	0	0	r	0	0	0	r	r	r	0	0	0
r	0	0	0	r	0	0	r	r	r	1	r	r	0	0	r	0	0	0	r	0	0	r	0	0	r	r	0
r	0	0	0	0	r	0	r	r	r	r	1	r	0	0	0	r	0	0	0	r	0	0	r	0	r	0	r
r	0	0	0	0	0	r	r	r	r	r	r	1	0	0	0	0	r	0	0	0	r	0	0	r	0	r	r
0	r	r	0	0	0	0	r	r	0	0	0	0	1	r	r	r	r	r	r	r	r	0	0	0	0	0	0
0	r	0	r	0	0	0	r	0	r	0	0	0	r	1	r	r	r	r	0	0	0	r	r	r	0	0	0
0	r	0	0	r	0	0	r	0	0	r	0	0	r	r	1	r	r	0	r	0	0	r	0	0	r	r	0
0	r	0	0	0	r	0	r	0	0	0	r	0	r	r	r	1	r	0	0	r	0	0	r	0	r	0	r
0	0	r	0	0	0	r	r	0	0	0	0	r	r	r	r	r	1	0	0	0	r	0	0	r	0	r	r
0	0	r	r	0	0	0	0	r	r	0	0	0	r	r	0	0	0	1	r	r	r	r	r	r	0	0	0
0	0	r	0	r	0	0	0	r	0	r	0	0	r	0	r	0	0	r	1	r	r	r	0	0	r	r	0
0	0	r	0	0	r	0	0	r	0	0	r	0	r	0	0	r	0	r	r	1	r	0	r	0	r	0	r
0	0	r	0	0	0	r	0	r	0	0	0	r	r	0	0	0	r	r	r	r	1	0	0	r	0	r	r
0	0	0	r	r	0	0	0	0	r	r	0	0	0	r	r	0	0	r	r	0	0	1	r	r	r	r	0
0	0	0	r	0	r	0	0	0	r	0	r	0	0	r	0	r	0	r	0	r	0	r	1	r	r	0	r
0	0	0	r	0	0	r	0	0	r	0	0	r	0	r	0	0	r	r	0	0	r	r	r	1	0	r	r
0	0	0	0	r	r	0	0	0	0	r	r	0	0	0	r	r	0	0	r	r	0	r	r	0	1	r	r
0	0	0	0	r	0	r	0	0	0	r	0	r	0	0	r	0	r	0	r	0	r	r	0	r	r	1	r
0	0	0	0	0	r	r	0	0	0	0	r	r	0	0	0	r	r	0	0	r	r	0	r	r	r	r	1

The Eigenstructure of Σ

There are N eigenvectors of Σ:

- 1 eigenvector ($\mathbf{v}_1 = \mathbf{1}_N/\sqrt{N}$)
  with eigenvalue $\lambda_1 = {\sigma_e}^2 + 2(n-1){\sigma_a}^2$

- $(n-1)$ eigenvectors ($\mathbf{v}_2$ to $\mathbf{v}_n$)
  with eigenvalue $\lambda_2 = {\sigma_e}^2 + (n-2){\sigma_a}^2$

- $(N-n)$ eigenvectors ($\mathbf{v}_{n+1}$ to $\mathbf{v}_N$)
  with eigenvalue $\lambda_3 = {\sigma_e}^2$

Because eigenvectors are orthogonal, we have $\mathbf{v}_k'\mathbf{1}_N = 0$ for all $k > 1$.

The Likelihood for the Vector of Scores

The log-likelihood can be written as

$$-2lnL = ln(2\pi) + ln|\boldsymbol{\Sigma}| + (\mathbf{y} - \theta\mathbf{1}_N)'\boldsymbol{\Sigma}^{-1}(\mathbf{y} - \theta\mathbf{1}_N)$$

$$= ln(2\pi) + ln\lambda_1 + (n-1)ln\lambda_2 + (N-n)ln\lambda_3$$
$$+ \frac{N(\bar{y} - \theta)^2}{\lambda_1} + \frac{\mathbf{y}'(\sum_{k=2}^{n} \mathbf{v}_k \mathbf{v}_k')\mathbf{y}}{\lambda_2} + \frac{\mathbf{y}'(\sum_{l=n+1}^{N} \mathbf{v}_l \mathbf{v}_l')\mathbf{y}}{\lambda_3}$$

$$= ln(2\pi) + ln\lambda_1 + (n-1)ln\lambda_2 + (N-n)ln\lambda_3$$

$$+ \frac{N(\bar{y} - \theta)^2}{\lambda_1} + \frac{SS_a}{\lambda_2} + \frac{SS_e}{\lambda_3}$$

Unbiased Estimates of the Parameters of the Model

We can find unbiased estimators for all parameters in our model: First let $MS_a = SS_a/(n-1)$ and $MS_e = SS_e/(N-n)$.

$$\hat{\theta} = \bar{y}$$

$$\hat{\sigma}_a{}^2 = \frac{MS_a - MS_e}{n-2}$$

$$\hat{\sigma}_e{}^2 = MS_e$$

These estimates are closely related to REML estimates. We have derived closed form versions of all three of these estimates.

(For future reference, let $SS_t = SS_a + SS_e$ and $MS_t = SS_t/(N-1)$.)

Some More Important Quantities

The variance of a randomly selected score (the similarity score for two randomly selected objects) is given by

$$\sigma_s{}^2 = \sigma_e{}^2 + 2\sigma_a{}^2.$$

The variance of the mean of all scores in the sample ($\bar{y}$) is given by

$$\sigma_{\bar{y}}{}^2 = \frac{\sigma_e{}^2}{N} + \frac{4\sigma_a{}^2}{n}.$$

We can obtain unbiased estimates for each of these quantities by plugging in the unbiased estimates of the variance components. These will be designated by "hats".

Some Other Related Quantities

The expected value of MS_t is given by

$$\sigma_e{}^2 + 2\frac{n-1}{n+1}\sigma_a{}^2$$

which is almost the same as $\sigma_s{}^2$.

The expected value of MS_t/N is given by

$$\frac{\sigma_e{}^2}{N} + \left(\frac{1}{n+1}\right)\frac{4\sigma_a{}^2}{n}$$

which is *NOT AT ALL* the same as $\sigma_{\bar{y}}{}^2$!

The Random Match Probability Based on Our Model

For a given cutoff τ, the random match probability is the probability that a randomly selected s_{ij} will exceed τ. That is

$$RMP = P\{s_{ij} > \tau\}.$$

For our model

$$P\{s_{ij} > \tau\} = 1 - \Phi\left(\frac{\tau - \theta}{\sigma_s}\right) \equiv \pi$$

where Φ is the standard normal CDF.

Some Equivalent Mathematical Statements

The following statements are equivalent for any value of B (either random or not random).

$$\pi < B \Leftrightarrow 1 - \Phi\left(\frac{\tau - \theta}{\sigma_s}\right) < B \Leftrightarrow \Phi\left(\frac{\tau - \theta}{\sigma_s}\right) > 1 - B$$

$$\Leftrightarrow \frac{\tau - \theta}{\sigma_s} > \Phi^{-1}(1 - B) \equiv L$$

Thus we require a random quantity L such that

$$P\left\{\frac{\tau - \theta}{\sigma_s} > L\right\} = 1 - \alpha$$

It is straightforward to convert such an interval to an upper confidence bound for π.

A Hypothetical

Suppose for a moment that we knew σ_e and σ_a but that we estimated θ by $\bar{y}$. Then

$$P\left\{\frac{\tau-\bar{y}}{\sigma_{\bar{y}}}-\frac{\tau-\theta}{\sigma_{\bar{y}}}<z_\alpha\right\}=1-\alpha$$

BUT

$$\frac{\tau-\bar{y}}{\sigma_{\bar{y}}}-\frac{\tau-\theta}{\sigma_{\bar{y}}}<z_\alpha \Leftrightarrow \frac{\tau-\theta}{\sigma_s}>\frac{\tau-\bar{y}}{\sigma_s}-z_\alpha\frac{\sigma_{\bar{y}}}{\sigma_s}$$

We could use this to get our upper confidence bound for π. The resulting interval would have exactly correct coverage probability.

What would happen if we ignored the correlation structure?

If we ignored the correlation structure in this model, we could use the same formula from the previous slide but substituting the expected value of MS_t/N for $\sigma_{\bar{y}}$. This upper confidence bound has some very bad properties.

Actual Coverage Probability for Hypothetical Method

(Nominal Coverage Probability 0.95, True RMP = 0.000001)

n	ρ =.1	ρ =.5	ρ =1	ρ =2	ρ =10
10	.855	.758	.730	.712	.695
50	.708	.627	.611	.602	.593
100	.653	.592	.580	.573	.567
500	.571	.541	.536	.533	.530
1000	.551	.529	.525	.523	.521

$\rho = {\sigma_a}^2/{\sigma_e}^2$

A Confidence Interval Based on Fieller's Theorem

Results for New Method Based on Fieller's Theorem

(Nominal Coverage Probability 0.95; True RMP=0.001; 1,000,000 Simulations per Cell)

n	ρ =.1	ρ =.5	ρ =1	ρ =2	ρ =10
50	.9465 .0005 .0022	.9433 .0002 .0042	.9442 .0002 .0055	.9457 .0001 .0068	.9481 .0001 .0085
100	.9474 .0006 .0017	.9467 .0004 .0027	.9469 .0003 .0034	.9479 .0002 .0040	.9495 .0002 .0048
500	.9497 .0008 .0012	.9490 .0006 .0016	.9489 .0006 .0017	.9495 .0005 .0019	.9498 .0005 .0021
1000	.9501 .0009 .0012	.9497 .0007 .0014	.9495 .0007 .0015	.9497 .0006 .0016	.9497 .0006 .0017

Coverage Probability
Average Lower Bound
Average Upper Bound

Conclusions

- We have found a method which yields an approximately correct confidence interval for the random match probability based on pairwise comparisons.

- This method can be used in any situation where the scores can be (monotonically) transformed to approximate normality.

- As long as the model is valid, an estimate of the random match probability can be made even if no matches are observed in the sample.

- We are continuing research to obtain an upper bound (rather than a two-sided interval) for the random match probability.