My Collaborations with Paul Kantor: From the Paper Neither of Us Can Access to Fish

Fred Roberts
DIMACS and CCICADA
Rutgers University

Happy 75th Birthday Paul!

Paul Kantor: A Search Through Google

Paul Kantor: A Search Through Google

Those who know Paul know he is strong
Well, he has strong opinions

Paul Kantor: A Search Through Google

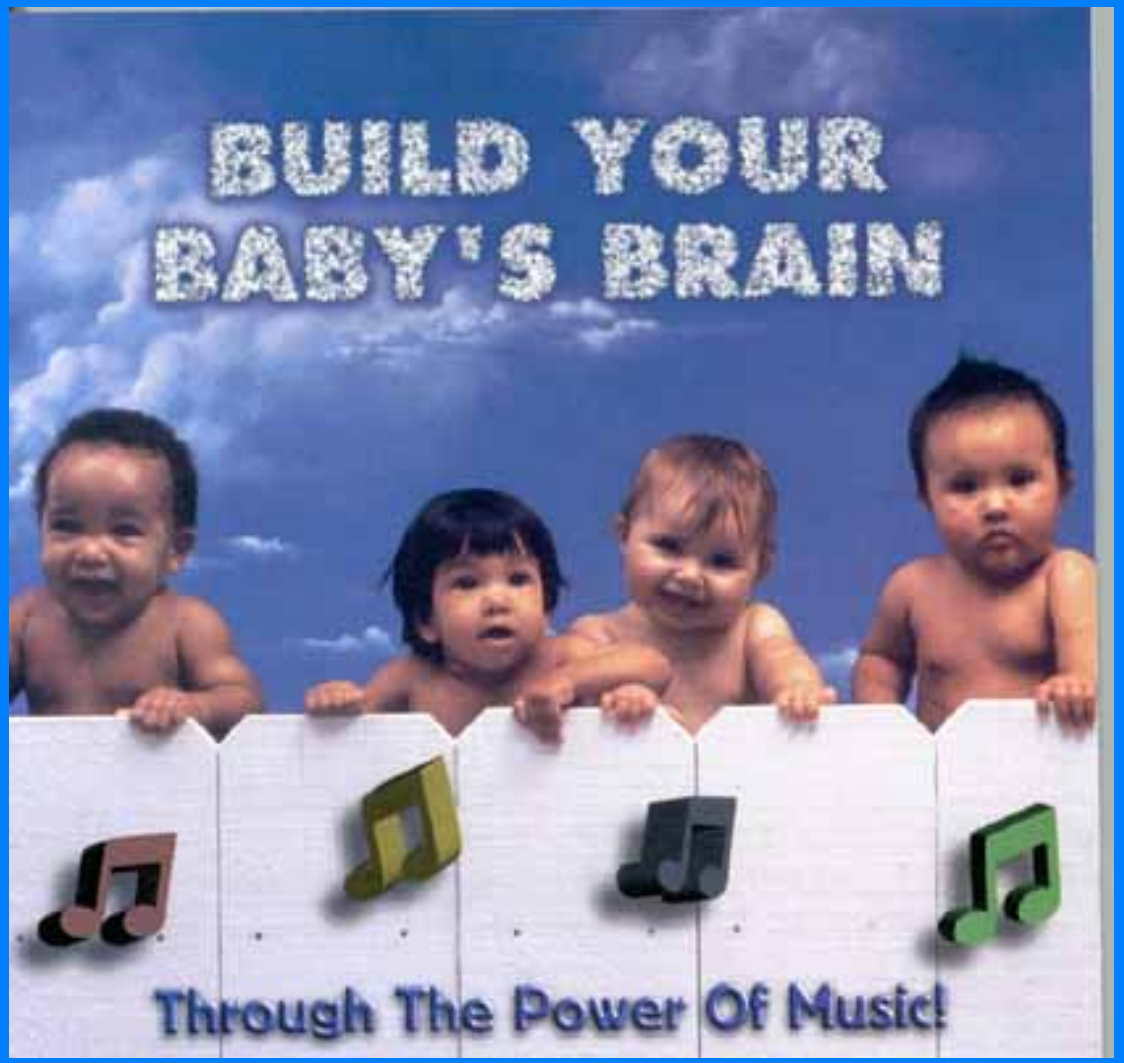
We have been known to debate different points of view.

Paul Kantor: A Search Through Google

The google page says master taxidermist, Paul Kantor.

Taxidermy = art of preparing, stuffing, and mounting skins of animals. Paul prepares, stuffs, and mounts what??

Our first real collaboration arose when Paul educated me about Machine Learning



The story starts with infectious diseases – in fact, anthrax

Epidemiology to MMS

DIMACS
Center for Discrete Mathematics and Theoretical Computer Science
A National Science Foundation Science and Technology Center

- Sept. 11 attacks
- Subsequent anthrax attacks
- Concern about bioterrorism
- The DIMACS Center had sizeable NSF grant for mathematical and computational epidemiology
- KDD group contacted researchers with relevant large NSF funding to see what they might do for homeland security.
- We were invited because of bioterrorism

Epidemiology to MMS

- Somehow, Paul influenced me to come to the meeting with several research ideas.
- We never did do epidemiology
- But KDD group liked an idea involving machine learning.
- ***Paul coached me in my presentation***
- (You don't want to know what other machine learning people said about the presentation).
- Somehow, it was funded.
- ***The "Monitoring Message Streams" project was initiated.***

Monitoring Message Streams: Algorithmic Methods for Automatic Processing of Messages

MMS: Goal

Monitor huge communication streams, in particular, streams of textualized communication to automatically detect pattern changes and "significant" events

Motivation: monitoring
- email traffic, news, communiques, faxes, voice intercepts (with speech recognition)

MMS: Highlights of Achievements From 9/06 Presentation

- ***Nearest Neighbor (kNN) text classification:***
  - Reduced memory usage up to 10-fold
  - Increased execution speed up to 100-fold
  - Demonstrated scaling to tens of thousands of classes
- ***Logistic Regression:***
  - Sped up application of tens of thousands of logistic regression classifiers
  - Reduced size of these classifiers by 1000-fold while retaining state of the art effectiveness

MMS: Highlights of Achievements From 9/06 Presentation

- ***Vastly extended applications of logistic regression*** by developing algorithms for:
  - Learning logistic regression models online
  - Using domain texts/knowledge bases to greatly reduce need for labeled training data
  - Combining disparate sources of training examples through hierarchical priors
  - Automatically tuning regularization parameters
  - Using bootstrapping and other techniques to assess the uncertainty of a classifier's predictions.

MMS: Highlights of Achievements From 9/06 Presentation

- ***Our Bayesian Binary Regression (BBR) and Bayesian Multinomial Regression (BMR) packages***:
  - Downloaded hundreds of times
  - Increasingly used and cited
  - Constitute most efficient software in the world for ultra-high dimensional logistic regression
- ***Our Methods Yielded***:
  - Top performances on classification tasks in TREC2004 and 2005 evaluations
  - Top three overall results in Entity Resolution 1b Task of 2005 KDD Challenge

MMS: Highlights of Achievements

- ***Requirement: Once a year, submit a paper to Journal of the Intelligence Community Research and Development***
- ***P.B. Kantor*** *and F.S. Roberts* "Monitoring message streams: Algorithmic methods for automatic processing of messages," JICRD, Feb. 2007

MMS PROJECT TEAM:

Paul Kantor, Rutgers Communic., Info.& Library Studies
Dave Lewis, Consultant
Michael Littman, Rutgers CS
David Madigan, Rutgers Statistics
S. Muthukrishnan, Rutgers CS
Rafail Ostrovsky, Telcordia/UCLA
Fred Roberts, Rutgers DIMACS/Math
Martin Strauss, AT&T Labs/U. Michigan)
Wen-Hua Ju, Avaya Labs (collaborator)
Andrei Anghelescu, Graduate Student
Suhrid Balakrishnan, Graduate Student
Aynur Dayanik, Graduate Student
Dmitry Fradkin, Graduate Student
Peng Song, Graduate Student
Graham Cormode, postdoc
Alex Genkin, software developer
Vladimir Menkov, software developer

Epidemiology Also Led to Our Next Collaboration

- Strong epidemiological modeling group at Los Alamos
- Visit to Los Alamos – discussions with Kevin Saeger and Phil Stroud
- But not about epidemiology – about container inspection at ports

Epidemiology Also Led to Our Next Collaboration

- Led to joint NSF project
- "A Decision Logic Approach to the Port of Entry Inspection Problem"
- Work continued through funding from ONR and from Dept. of Homeland Security

Port of Entry Inspection Algorithms

Thanks to Capt. David Scott, US Coast Guard Captain of Port, Sector Delaware Bay, for taking us out on a tour of the port of Philadelphia

Sequential Decision Making Problem

- Stream of containers arrives at a port
  - Similar analysis for inspection prior to departure
- **The Decision Maker's Problem:**
  - Which to inspect?
  - Which inspections next based on previous results?
- **Approach:**
  - ***"decision logics"***
  - ***combinatorial optimization methods***
  - Builds on ideas of Stroud and Saeger at LANL
  - Need for new models and methods

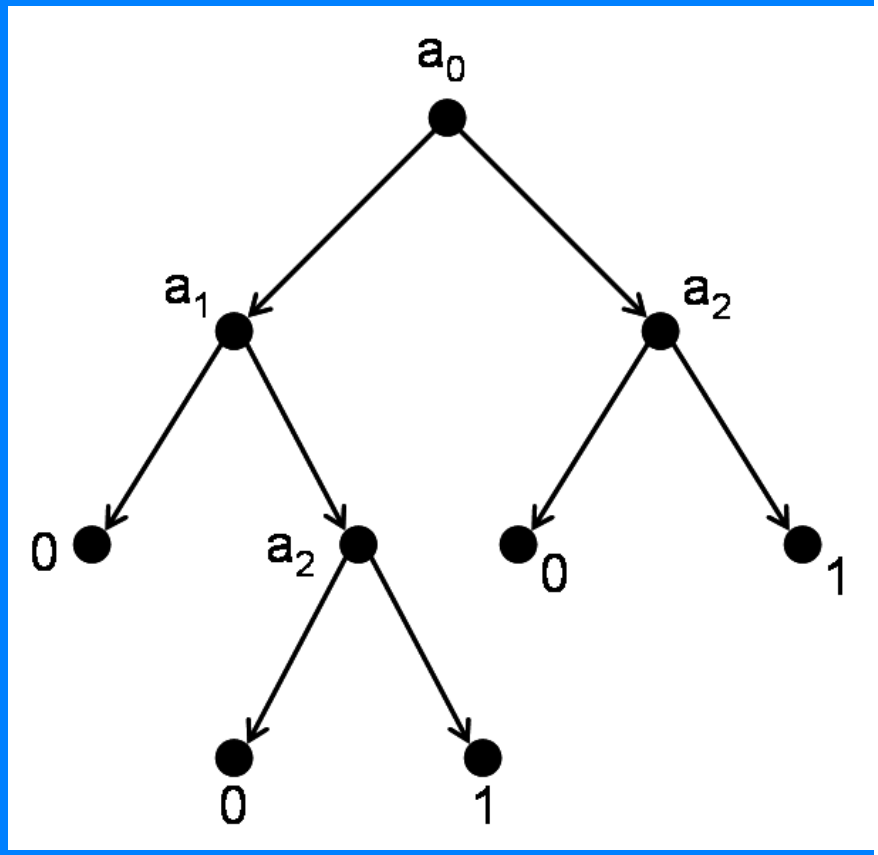
Binary Decision Tree Approach

- ***Sensors*** (or other "tests") measure presence/ absence of attributes: so 0 or 1
- Use two ***outcome categories***: 0, 1 (safe or suspicious)

- ***Binary Decision Tree***:
  - –Nodes are sensors or categories
  - –Two arcs exit from each sensor node, labeled left and right.
  - ***–Take the right arc when sensor says the attribute is present, left arc otherwise***

Binary Decision Tree Approach

- Reach category 1 from the root by:
  a_0 L to a_1 R a_2 R 1 or
  a_0 R a_2 R1
- ***Container classified in category 1 iff it has a_1 and a_2 and not a_0 or a_0 and a_2 and possibly a_1.***



Binary Decision Tree Approach

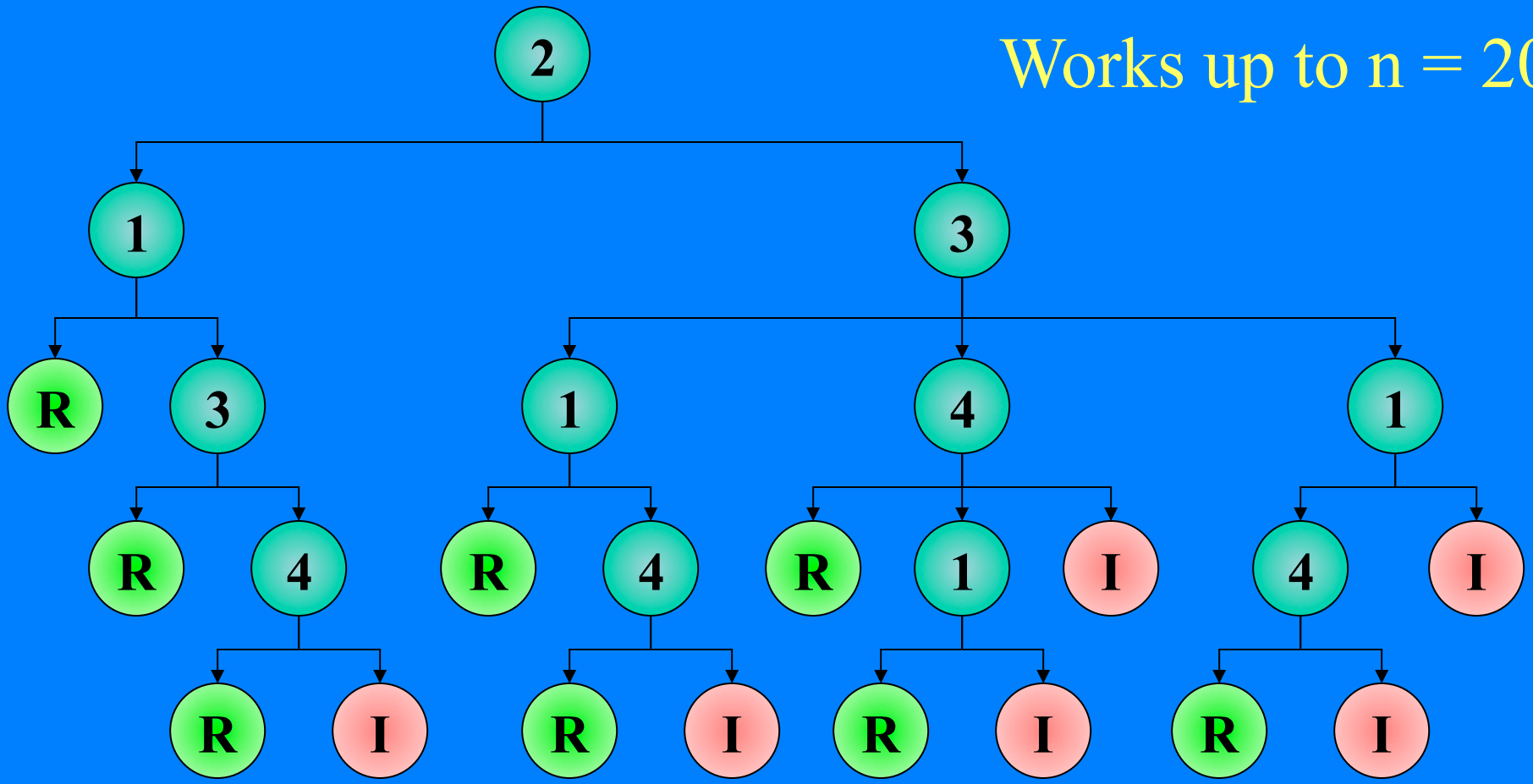
- How do we find a ***low-cost*** or ***least-cost*** binary decision tree corresponding to a Boolean function?
- Costs :
  - Inspection costs (use of tree nodes)
  - Delay costs
  - Fixed equipment costs
  - False positive, false negative
- n = no. of attributes

- Stroud and Saeger tools worked up to n =3
- n = 4 if specialize the decision function
- n = 4 at Port of Long Beach – Los Angeles
- Our methods work up to n = 10

Binary Decision Tree Approach

- Eventually led Boros and Kantor to develop their very clever ***SNSRtree algorithm***

Works up to n = 20



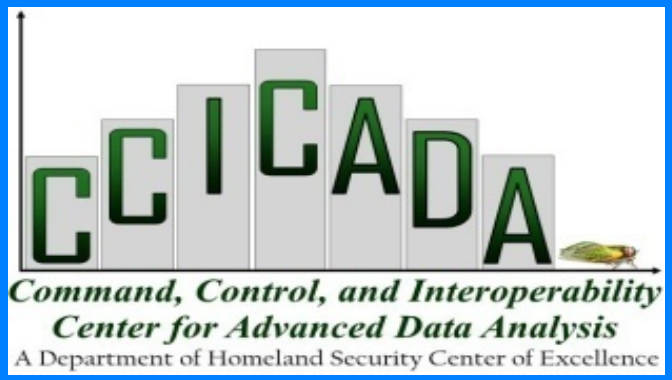
Detection = 81.5 27%, Cost = 0.1977826 units = $11.867 (< $13+)

Binary Decision Tree Approach

•E. Boros, E. Elsayed, ***P. Kantor***, F. Roberts, M. Xie, "Optimization problems for port-of-entry detection systems," in *Intelligence and Security Informatics: Techniques and Applications*, H. Chen and C. C. Yang (eds), Springer, 2008, 319-335.

Homeland Security

- Container inspection work morphed into work through Dept. of Homeland Security Center of Excellence:
CCICADA – Command, Control, and Interoperability Center for Advanced Data Analysis
- ***Paul as CCICADA Director of Research***
- Many collaborations



One Sample CCICADA Collaboration

- Subject: Fish

CCICADA
Command, Control, and Interoperability Center for Advanced Data Analysis
A Department of Homeland Security Center of Excellence

Estimating Violation Risk for Fisheries Regulations

Fisheries Rules

- The United States sets rules for fishing with the goal of maintaining healthy fish populations.
- Rules depend on specific species and include
  - Allowable locations to fish
  - Allowable seasons to fish
  - Catch quotas
- Violations of the rules leads to fines – sometimes quite large

Endangered Atlantic Cod

Fisheries Law Enforcement

- The US Coast Guard District 1 (based in Boston) uses a ***scoring system called OPTIDE to determine which commercial fishing vessels to board to look for violations.***
- The OPTIDE rule was built based on expert judgment and intuition.
- USCG asked us if their success rate in finding violations by boarding could be improved by use of sophisticated methods of data analysis.
- Goal: refine the ability to determine the risk profile of vessels.

Many Goals of Fisheries Law Enforcement

- The project started with the following definition of the goal: ***Find a decision rule for deciding whether or not to board that leads to as large a percentage of times as possible in which boarding leads to finding a violation.***

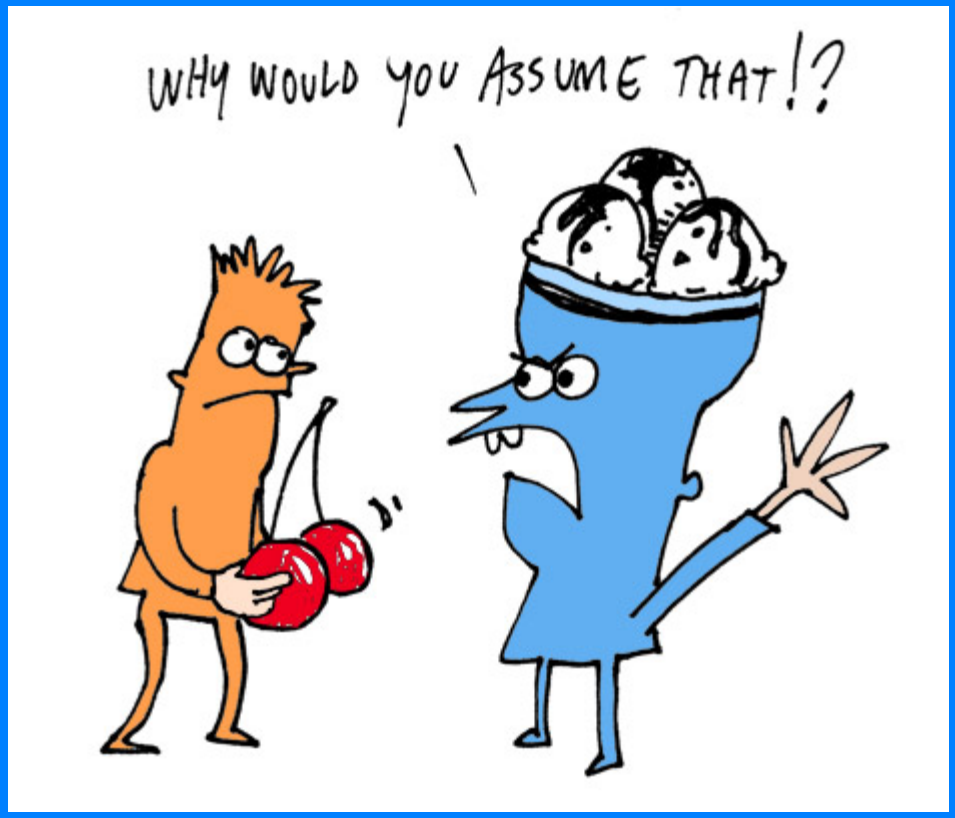
Fish Conclusions

- Project used machine learning to develop a new scoring rule: RIPTIDE (Rule Induction OPTIDE)
- Project used logistic regression to develop a new scoring rule: DE-OPTIDE (Data-Enhanced OPTIDE)
- OPTIDE, though based mostly on intuition, does quite well based on the features it uses.
- Both RIPTIDE and DE-OPTIDE improve over OPTIDE, but may require changes in number of Coast Guard vessels patrolling
- Many alternative approaches are needed to formalize all the multitude of goals in fisheries law enforcement: More research for me and Paul to do!

Fish

H. Chalupsky, R. DeMarco, F. Roberts, E. Hovy, ***P. Kantor***, A. Matlin, P. Mitra, B. Ozbas, J. Wojtowicz, M. Xie, "Estimating violation risk for fisheries regulations," in P. Perny, M. Pirlot, and A. Tsoukias (eds), *Proceedings of International Conference on Algorithmic Decision Theory III*, Lecture Notes in Computer Science, LNAI 8176, Springer, 2013, 297-308.

Thanks Paul
For many stimulating discussions over the years.



Thanks Paul for Many Memorable Talks

Congratulations Paul!!!

Many more successes!!!