

Data Mining and Cyber Threat Analysis – Five Trends

Robert Grossman
University of Illinois at Chicago
& Open Data Partners
February, 2003

Trend 1. Alert Management Systems

Focus in deployment is shifting from models to alerts, from data mining systems to alert management systems.

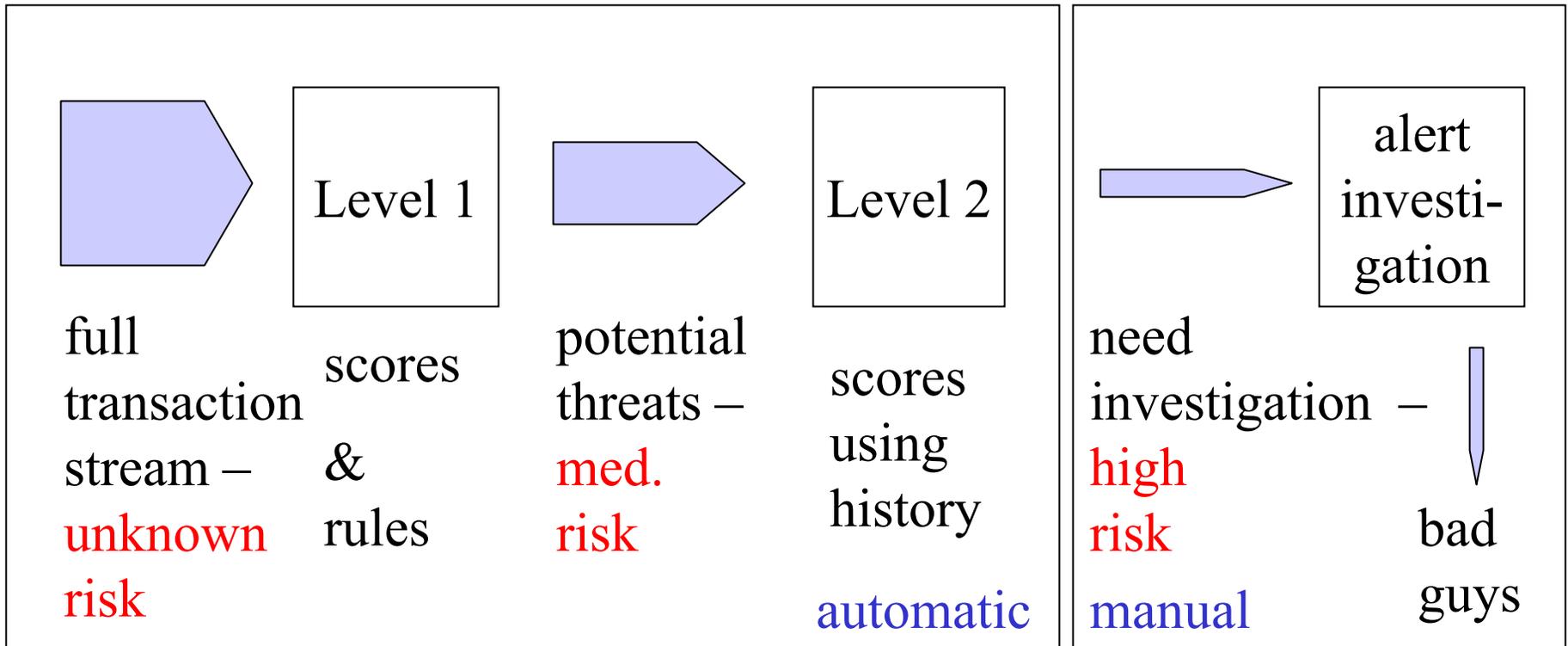
What is an Alert Management System?

- ❑ An Alert Management System (AMS) is a real time system which maintains *profiles* about individuals, threats, or other entities and in real time processes *events* and returns alerts about profiles and their risks.
- ❑ Examples: credit card fraud detection, threat assessment systems, intrusion detection systems, homeland defense, etc.

What are the Five Critical Functions of an AMS?

1. Scoring – compute risk scores for transactions, profiles, targets, etc.
2. Linking – social network analysis of targets
3. Matching – against watch lists, e.g. OFAC
4. Checking – regulations & policies
5. Routing – analysts have finite capacity

Alert Data Flow



1a. What is Scoring?

Mining data in motion—
assigning scores to data in real time.

Data Mining/Statistical Models

Summarization Models

Scores result from applying models to data.

- clustering
- associations

- tree-based methods
- neural nets
- k-nearest neighbors

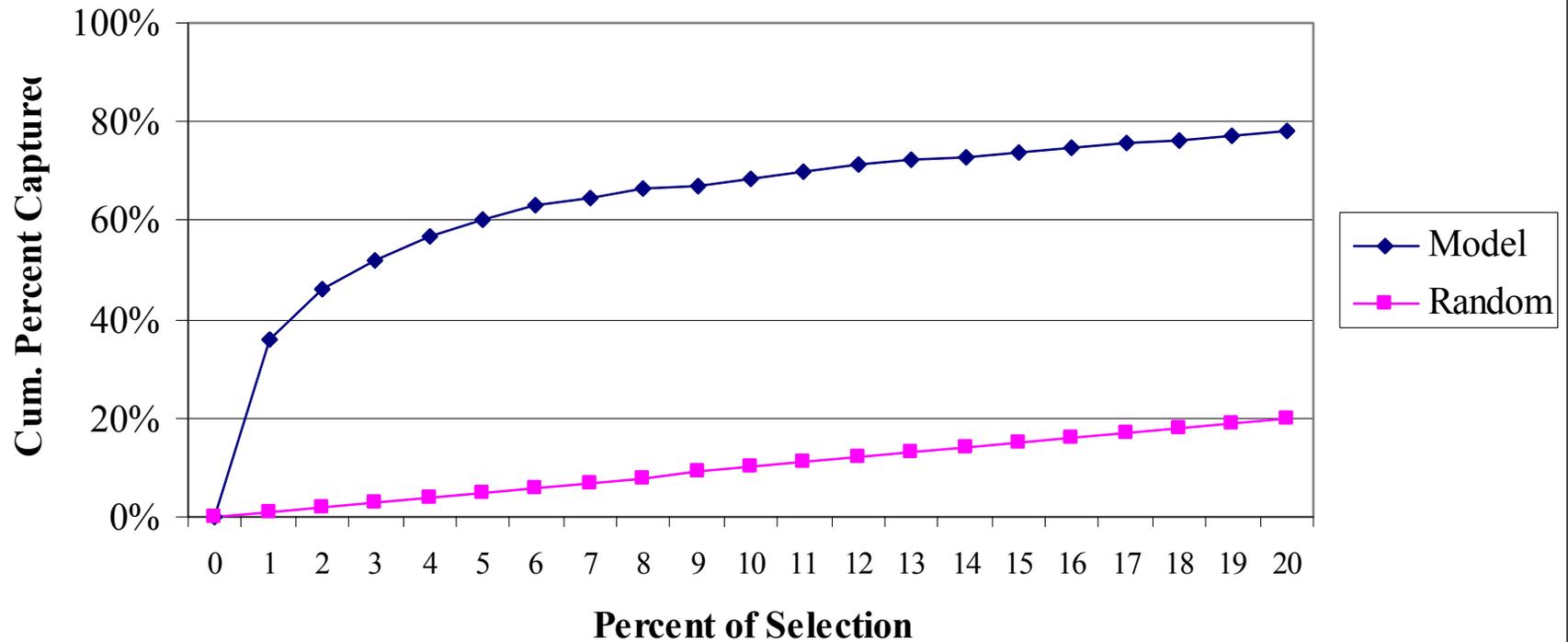
- contact chaining
- social network analysis

Predictive Models

Network/Graph

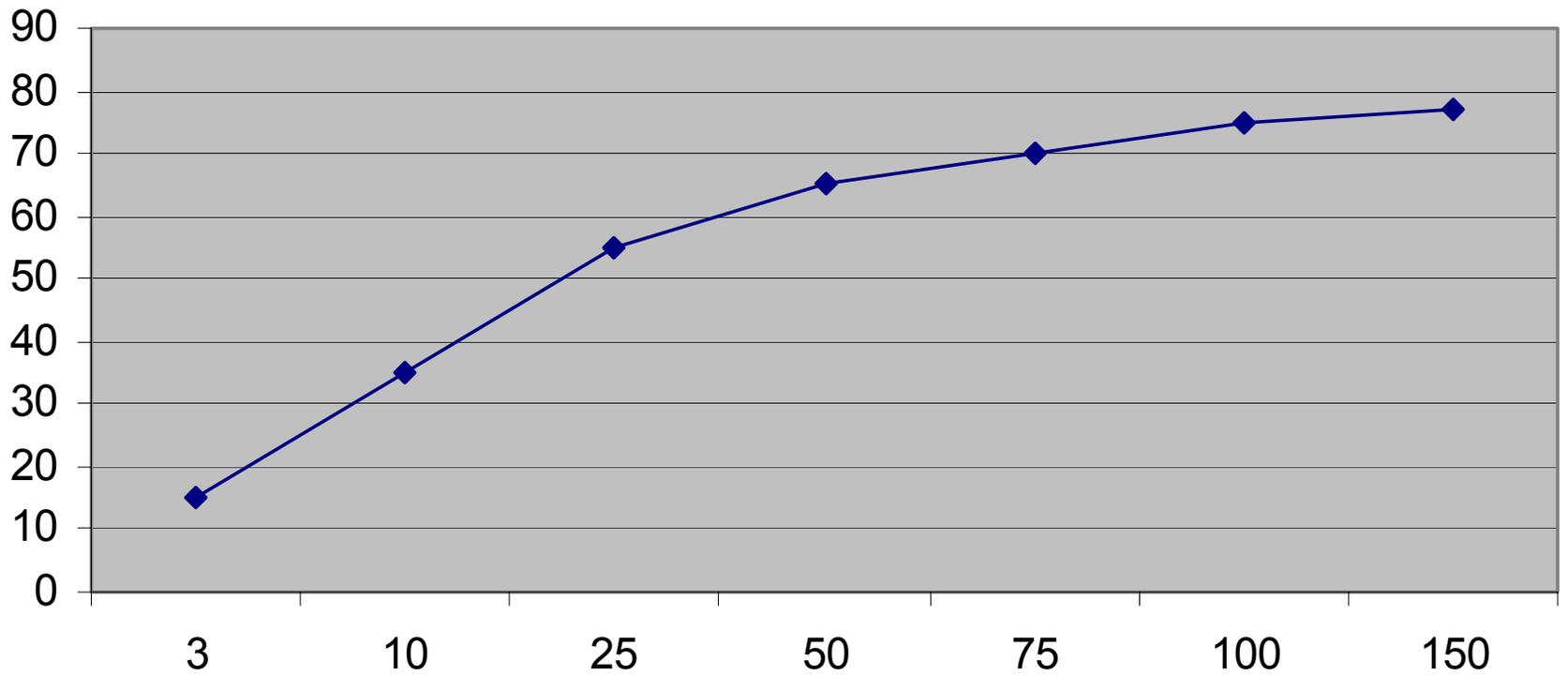
Detection Rates

Chart 2
Cumulative Percent Captured By Score Percentile (1st 20)



False Positive

False Positive vs. Detection Rate



Comparing Different Models

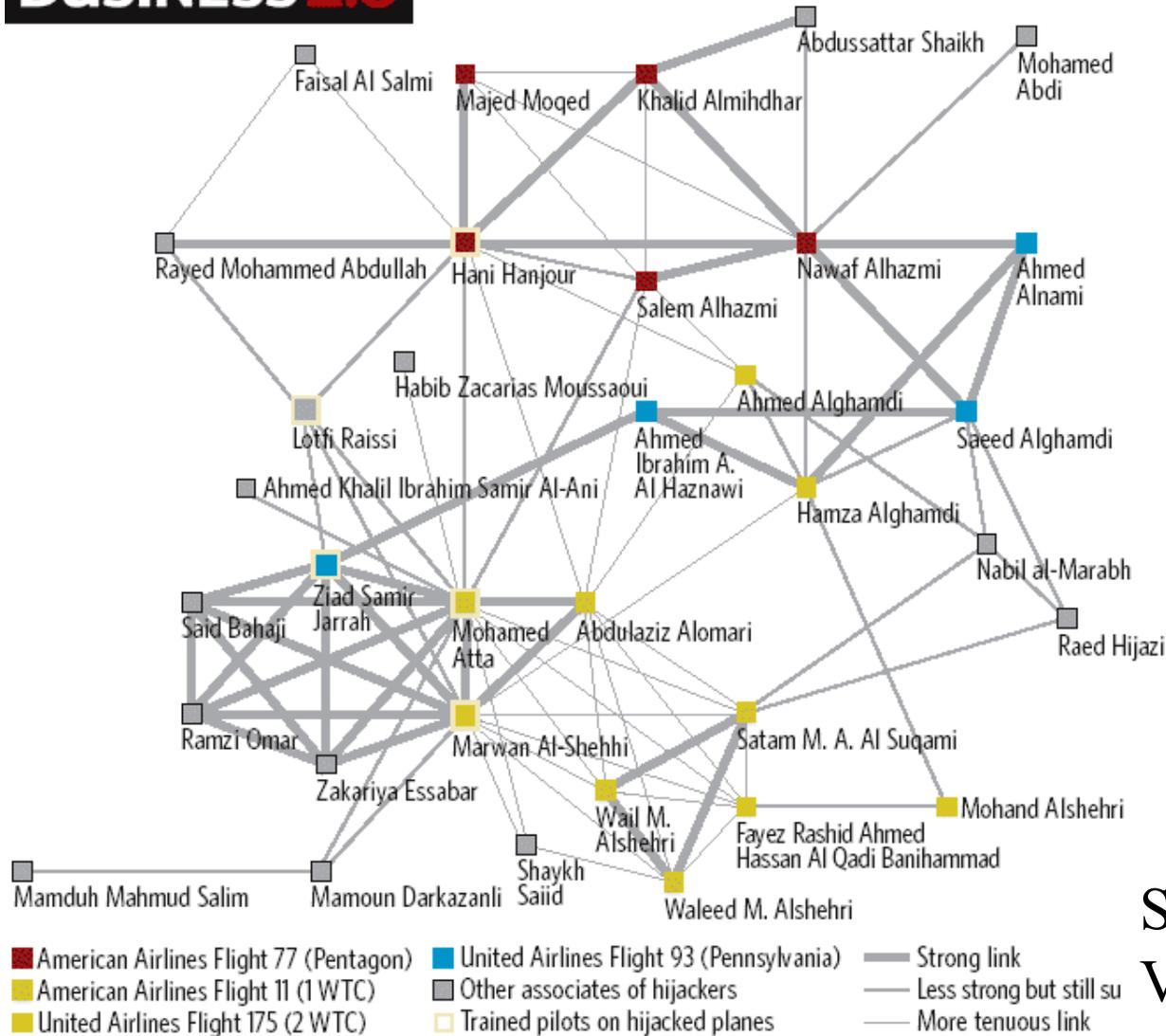
	Trees	Neural Networks	Rules
Accuracy	Yes	Yes	No
Easy to Maintain	Yes	No – hard to retrain	Yes – small No – large
Easy to interpret	Yes	No	Yes
Scalable to large data	Yes	No	No

1b. What is Linking?

Mining data at rest –
bad guys tend to hang out with other bad guys.

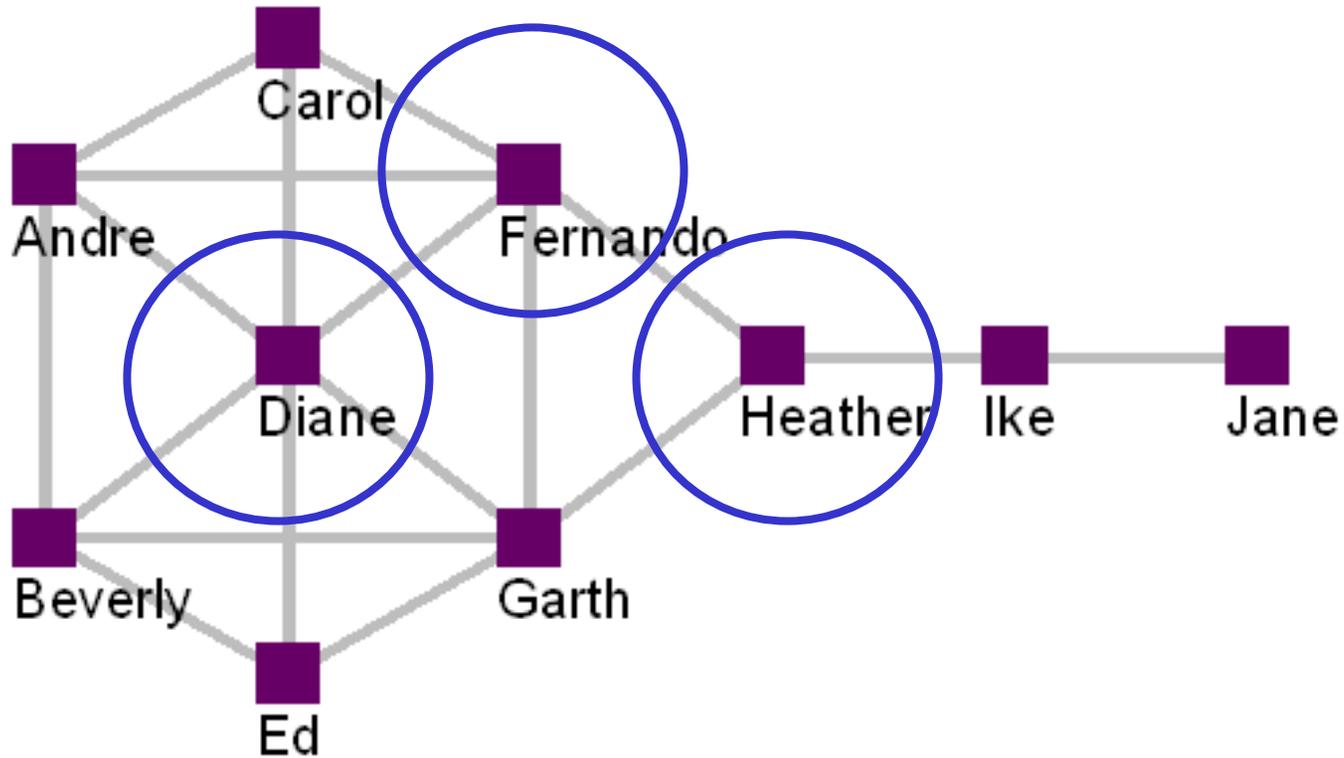
BUSINESS 2.0

Social Networks



Source:
Valdis Krebs

Social Networks



degree, betweenness, & closeness

1c. What is Matching?

Living with watch lists and other lists of good
and bad guys

Matching

Wall Street Journal
May 6, 2002.

The screenshot shows the Wall Street Journal website in a Microsoft Internet Explorer browser window. The address bar displays the URL: <http://online.wsj.com/article/0,,S8102063823287846600,00.html?mod=Page+One>. The page features the WSJ logo and navigation menus. The main article is titled "Crackdown on Terrorism Financing Ties Hands of Businessman in Sweden" by Christopher Cooper. The article text discusses the U.N. sanctions program and the U.S. Treasury Department's actions against Mr. Aden.

WSJ.com - Major Business News - Microsoft Internet Explorer

File Edit View Favorites Tools Help

Address <http://online.wsj.com/article/0,,S8102063823287846600,00.html?mod=Page+One> Go Links

WSJ.com THE WALL STREET JOURNAL ONLINE Search Quotes & Research

Other Journal Sites As of 11:28 p.m. EDT Sunday, May 5, 2002

News **Technology** **Markets** **Your Money** **Opinion** **At Leisure**

In Today's Paper Columnists Portfolio Setup Center Discussions Site Map Help Contact Us Log Out

THE WALL STREET JOURNAL Print Editions Customer Service Today In: BARRONS Online

WSJ.com THE WALL STREET JOURNAL ONLINE

PAGE ONE

Crackdown on Terrorism Financing Ties Hands of Businessman in Sweden

By **CHRISTOPHER COOPER**
Staff Reporter of THE WALL STREET JOURNAL

STOCKHOLM -- At the central train station, a clerk pushes a subway pass and a small pile of change toward Abdirisak Aden -- breaking international law as he does so.

The reason: Mr. Aden, a Somali-born former school principal and partner in a money-transfer office, is on the United Nations' list of those subject to economic sanctions. Citizens of all countries that honor the U.N. list are forbidden to conduct financial transactions with Mr. Aden unless the Security Council specifically approves it. No one can hire him, his bank account is frozen, and the mortgage on his two-bedroom condo is five months in arrears.

The U.N. devised its sanctions program as a way "to apply pressure on a state or entity ... without the use of force," it says. Over the years, the U.N. has targeted 13 governments, including Iraq, Libya and Sierra Leone, and a few prominent leaders such as Slobodan Milosevic of Yugoslavia. But after Sept. 11, the sanctions mechanism was pressed to service in a different task: helping the U.S. effort to block financing for terrorism. In the process, it ensnared for the first time some private individuals with no established links to rogue states or violent groups.

The U.S. Treasury Department added Mr. Aden's name to a list of groups and individuals it said had a role in funding terrorism. It faxed the list to the U.N. Security Council, which appended it without fanfare to its own economic-sanctions list.

AL BARAKAAT IN THE NEWS

- Though Its Corporate Face Is Gone, Barakaat Remains Active in Europe 04/22/02
- U.S. Estimates \$20 Million in Funds for Terrorists Have Been Cut Off 01/29/02
- Shutdown of Al Barakaat Severs Lifeline for Many Somalia Residents 12/04/01
- Network Suspected of Funding Terrorists Used Major Banks for Money Transfers 11/09/01

advertisement

free 3-year warranty
on hp multifunction printers

Done Internet

OFAC Entry: ADEN, Abdirisak, Skaftingebacken 8,
Spanga 163 67, Sweden; DOB 01 Jun 68
(individual) [SDGT]

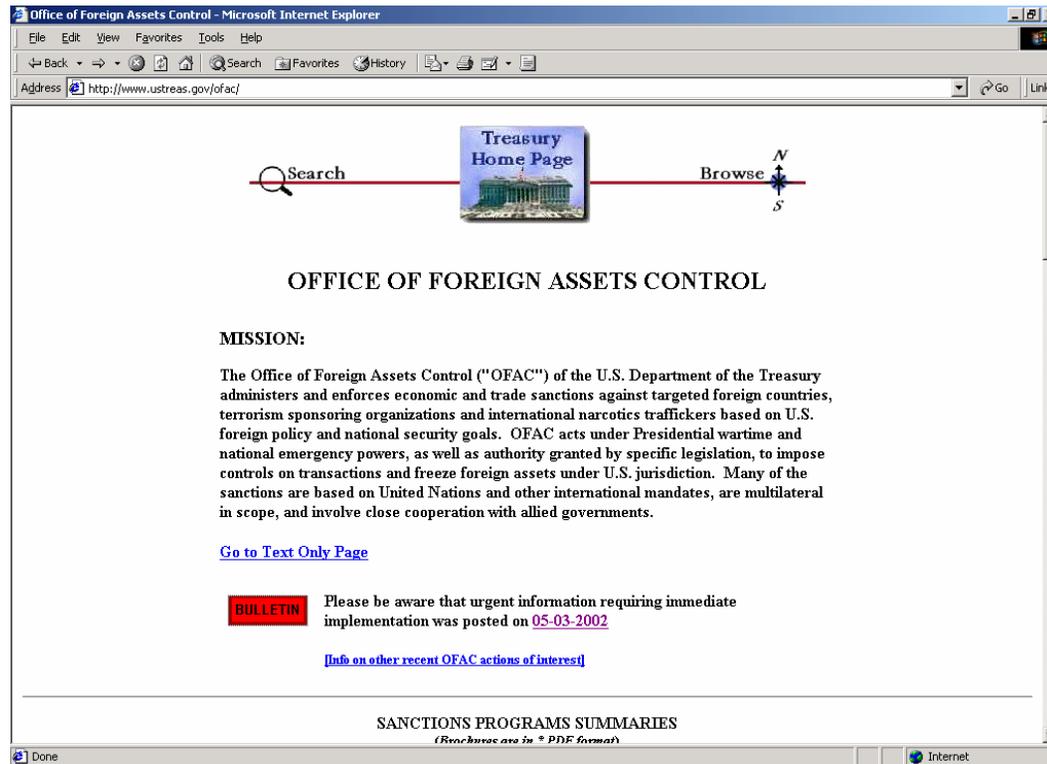
Similarity Search Is an Important Component of a Matching System

AHMED, Ahmed (a.k.a. ALI, Ahmed Mohammed Hamed; a.k.a. ABDUREHMAN, Ahmed Mohammed; a.k.a. ABU FATIMA; a.k.a. ABU ISLAM; a.k.a. ABU KHADIJAH; a.k.a. AHMED HAMED; a.k.a. Ahmed The Egyptian; a.k.a. AL-MASRI, Ahmad; a.k.a. AL-SURIR, Abu Islam; a.k.a. ALI, Ahmed Mohammed; a.k.a. ALI, Hamed; a.k.a. HEMED, Ahmed; a.k.a. SHIEB, Ahmed; a.k.a. SHUAIB), Afghanistan; DOB 1965; POB Egypt; citizen Egypt (individual) [SDGT]

1d. What is Checking?

How to stop worrying and learn to live with regulations.

There will be more and more regulations about what data can be used and how...



1e. What is Routing?

0.1% of 30,000 transactions/second
= 30/second at 10
minutes/investigation vs. 100
analysts and 8 hours per day.

Routing



- ❑ Routing is about getting the right information to the right person at the right time.

Trend 2. Real Time Data Mining

Exploiting Events and Profiles.

What is an Event?



login



email



message



scan



phone calls



credit card
transaction



cell phone call

- ❑ An **event** is real time information about an entity, eg. person, place, event, threat, opportunity, etc.

What is a Profile?



trips



calls



trips, #cc

summarized
information



0.26	0.86	0.94	...	0.70
------	------	------	-----	------

profiles

- A **profile** is the summarized data and attributes about an entity.

Trend 3.

In the past, we have built models and scored data at rest. In the future, more and more data will be streaming at faster and faster rates.

Premises

- ❑ Some data sets will be accessible via OC 12, GigE, 2.5 GigE, 10 GigE, etc. wide area networks – Photonic networks.
- ❑ There will be data mining services for high performance networks, *as well as for* commodity networks.
- ❑ Many applications will trade accuracy for speed in order to keep up with line speed
- ❑ Call these Photonic Data Services (PDS)

The Data Stack – Replacing Apps over Operating Systems

6. Data Mining Applications

5a. Storage
Services

5b. Data Web
Services

5b. Data Grid
Services

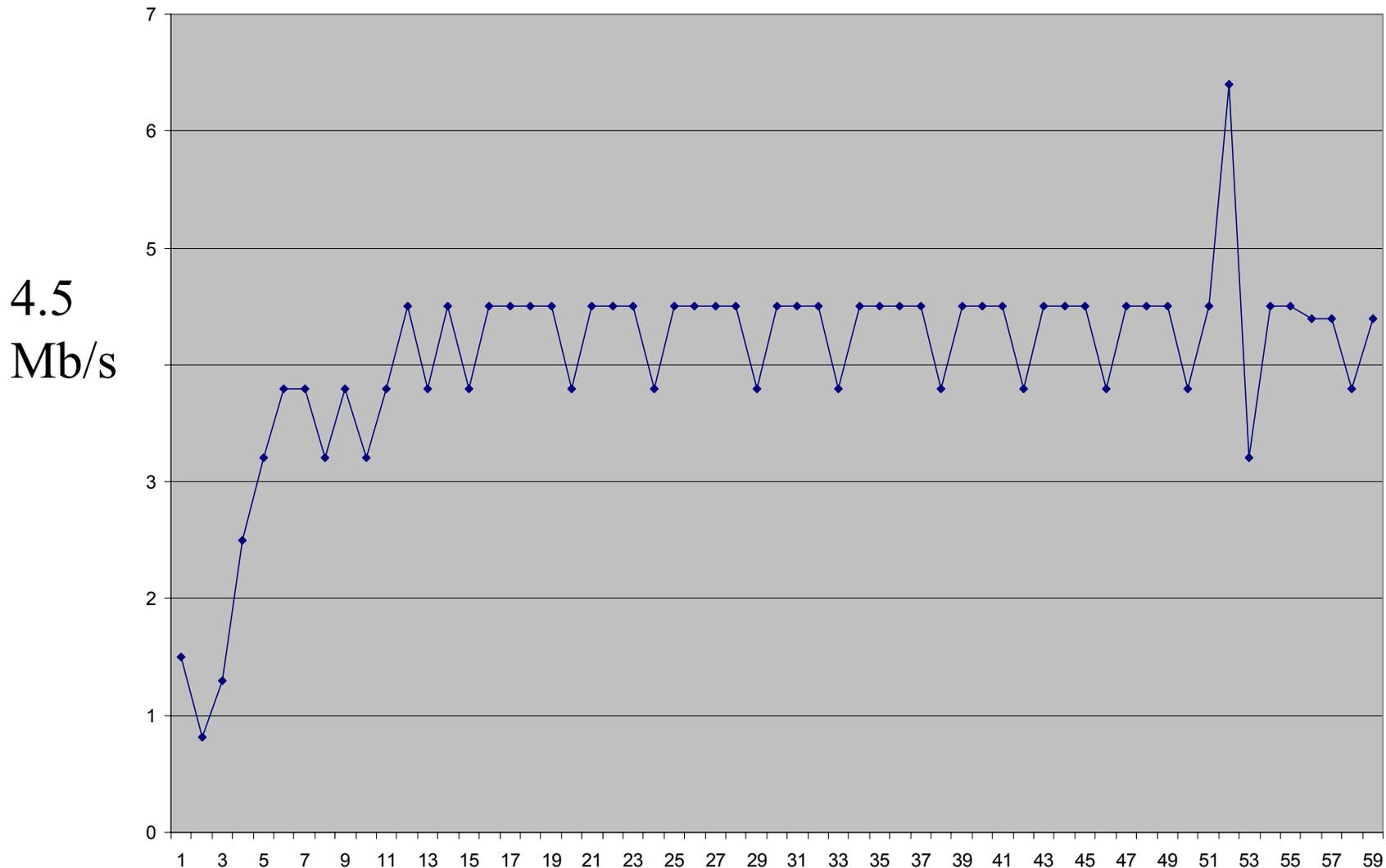
4. Transport – TCP, UDP, Reliable UDP

3. IP

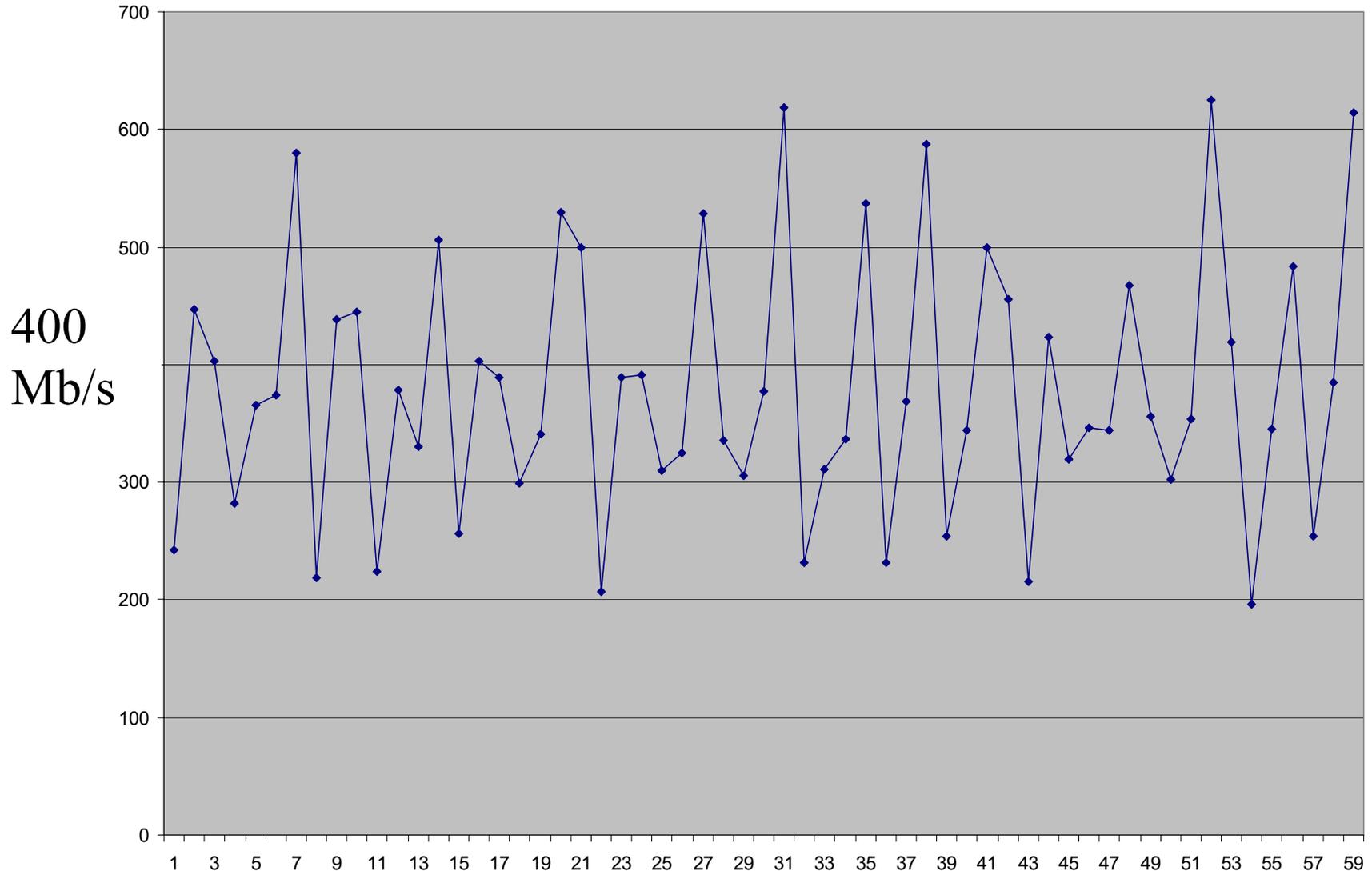
2. Photonic Path Services

1. Physical

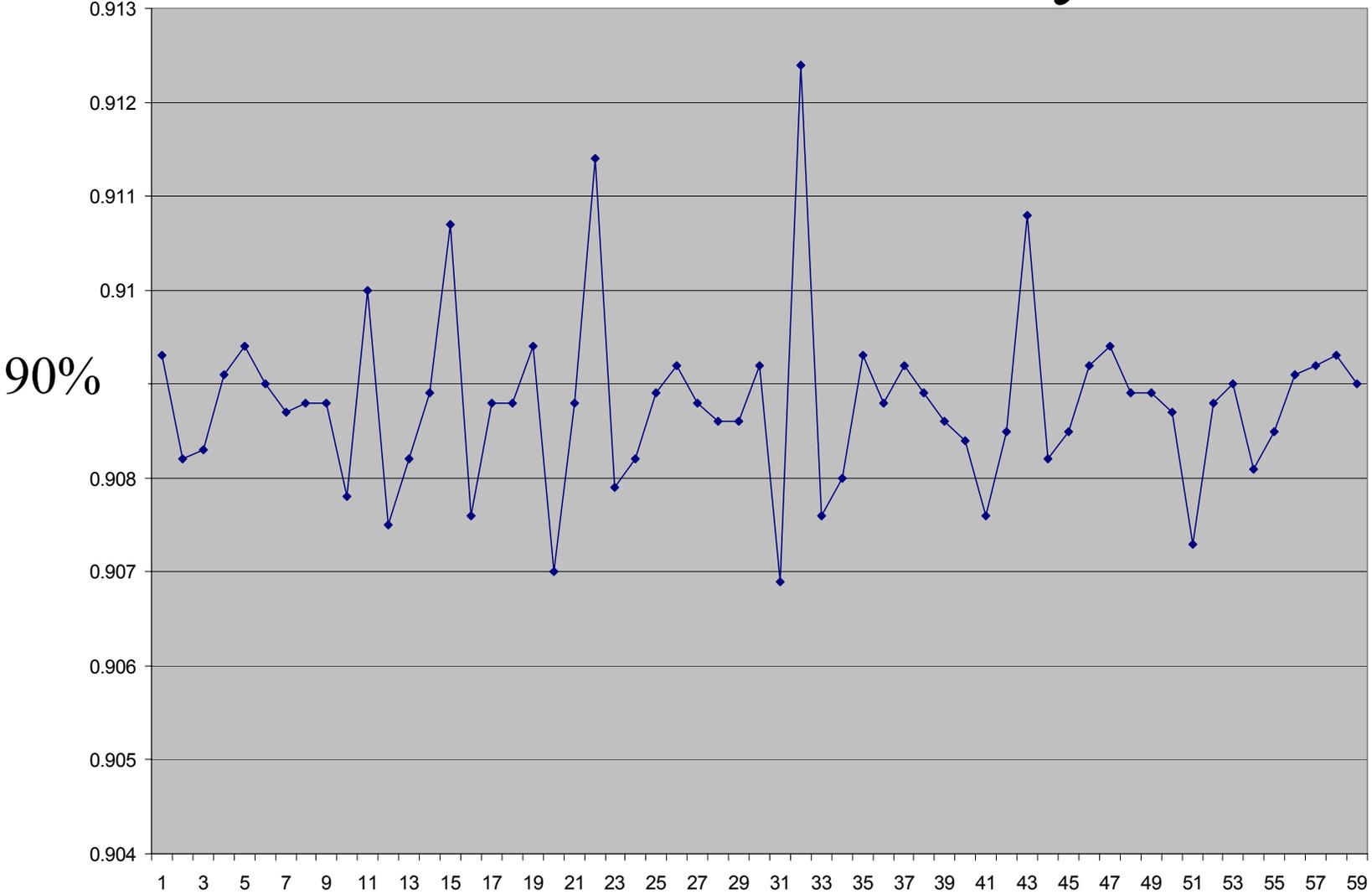
TCP Data Transport Chicago to Amsterdam over 622 Mb/s Link



Best Effort Distributed Merge Over PDS - Bandwidth



Best Effort Distributed Merge over PDS - Accuracy



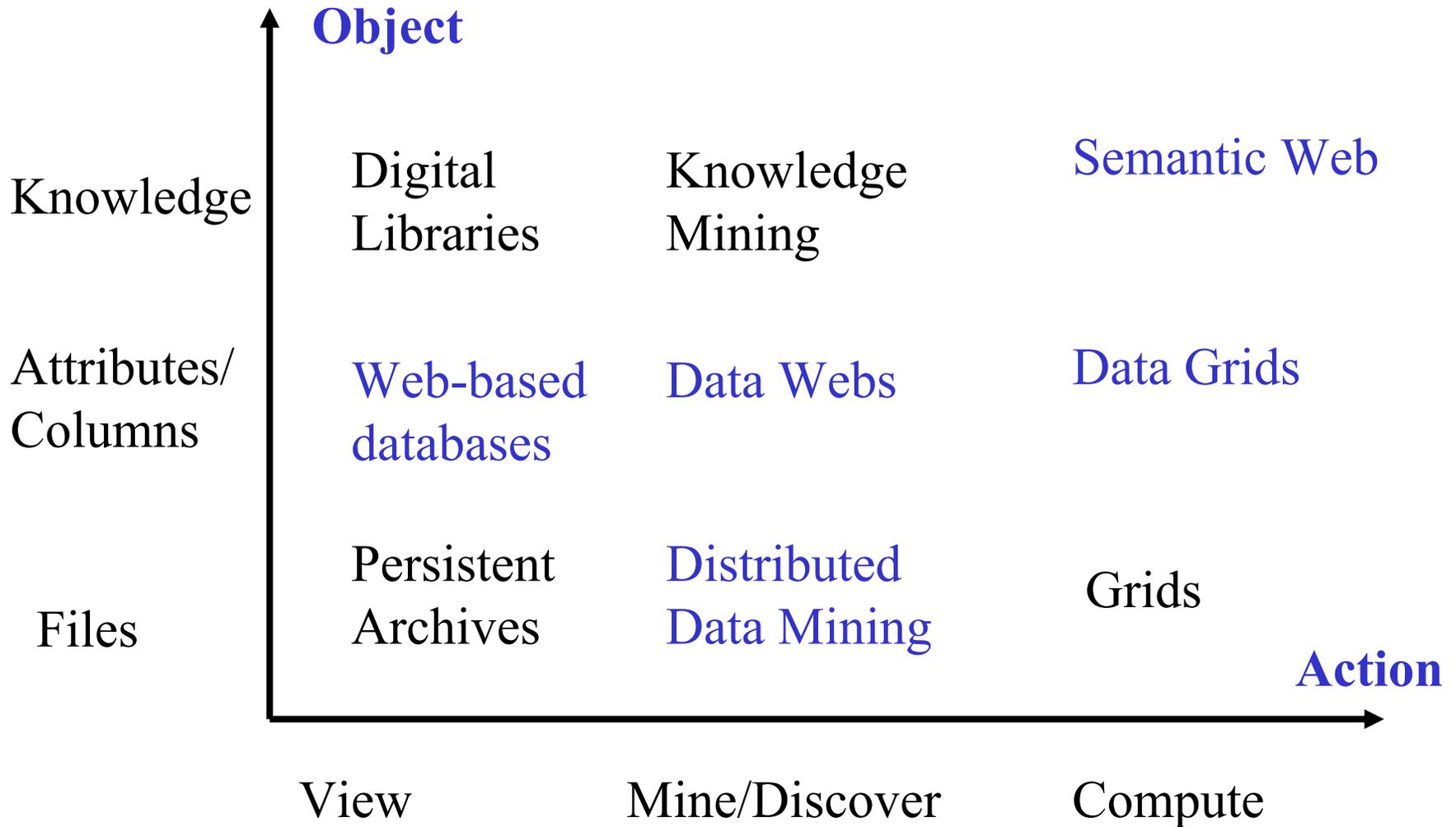
Trend 4. Data Webs for Data Exploration

We have developed some good data mining algorithms & systems, we need better algorithms and systems for data *exploration*.

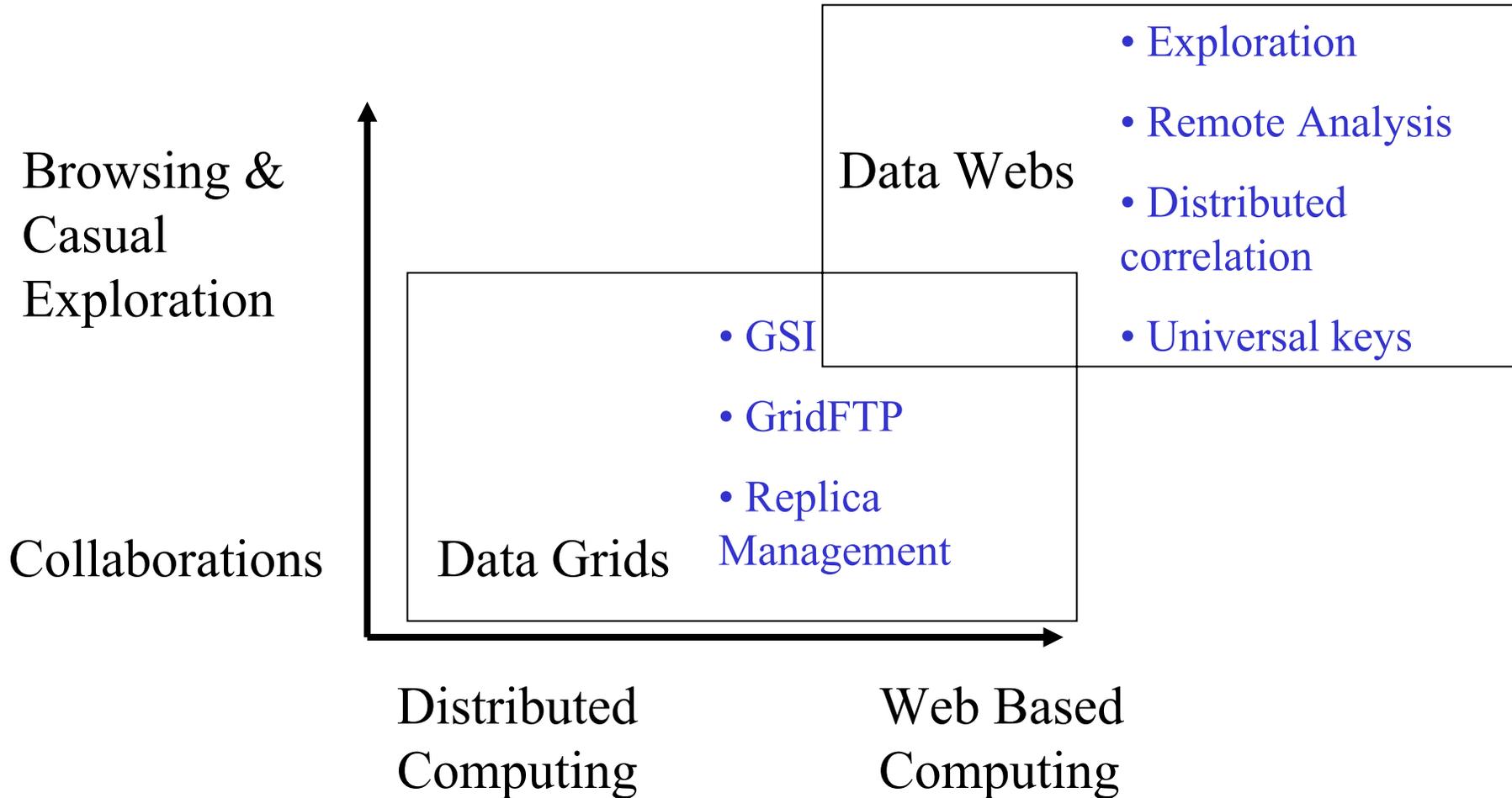
Paradigm Today

1. Learn about a new data source from a friend or over the web
2. FTP the data, federal express, or courier the data
3. A consultant or contractor spends a 1-3 months and then tells you whether or not to start a project to build a centralized data warehouse

Technologies for Global Data



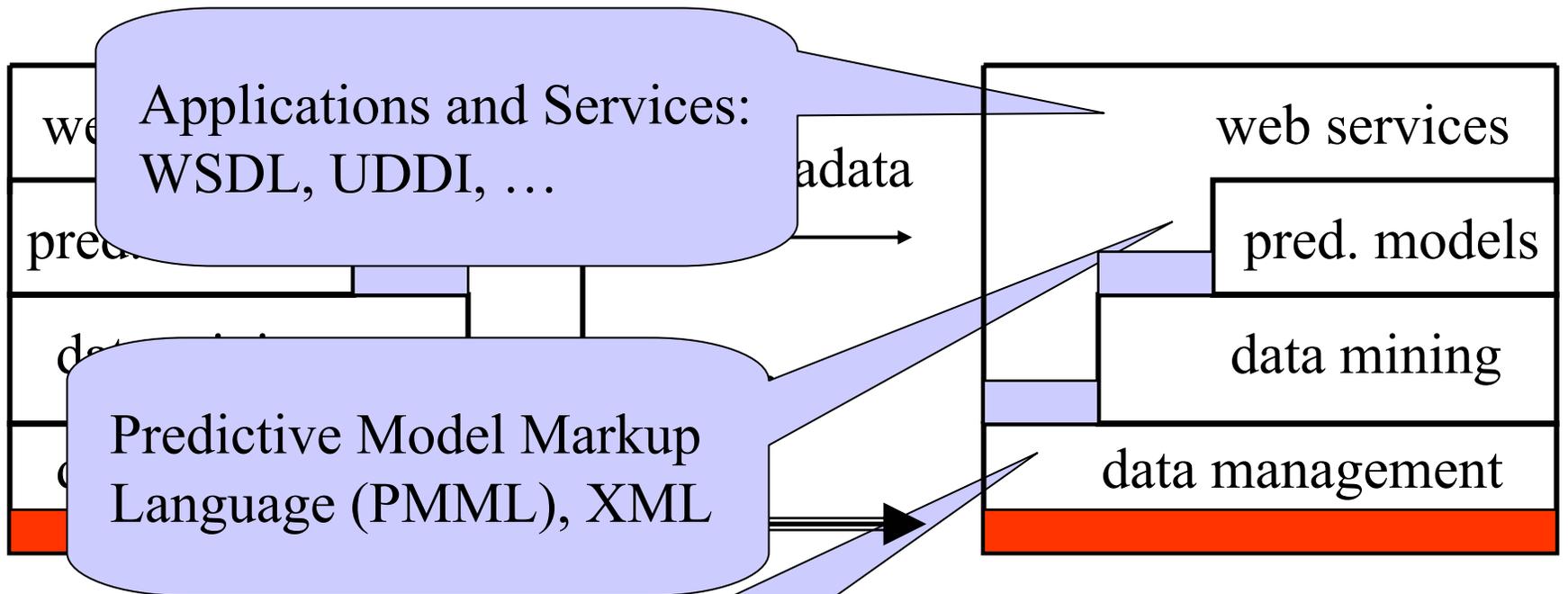
Data Grids vs. Data Webs



Trend 5.

Standards are maturing.

Maturing Standards



- Web services: WSDL, UDDI, ... layer in an open infrastructure.
- Database SQL, JDBC, ODBC, ... standard for data mining.

Data Mining Group

- ❑ Products shipping with PMML Version 1.1
- ❑ PMML Working Group Full Members
 - IBM, Magnify, Microsoft, MineIt, NCR, Oracle, Salford Systems, SAS, SPSS, xChange, University of Illinois at Chicago (over 20 vendors)
- ❑ PMML Working Group Supporting Members
 - Angoss, Insightful, KXEN, Microsoft, SGI ...
- ❑ Part of Source Forge

Problems with Current Techniques

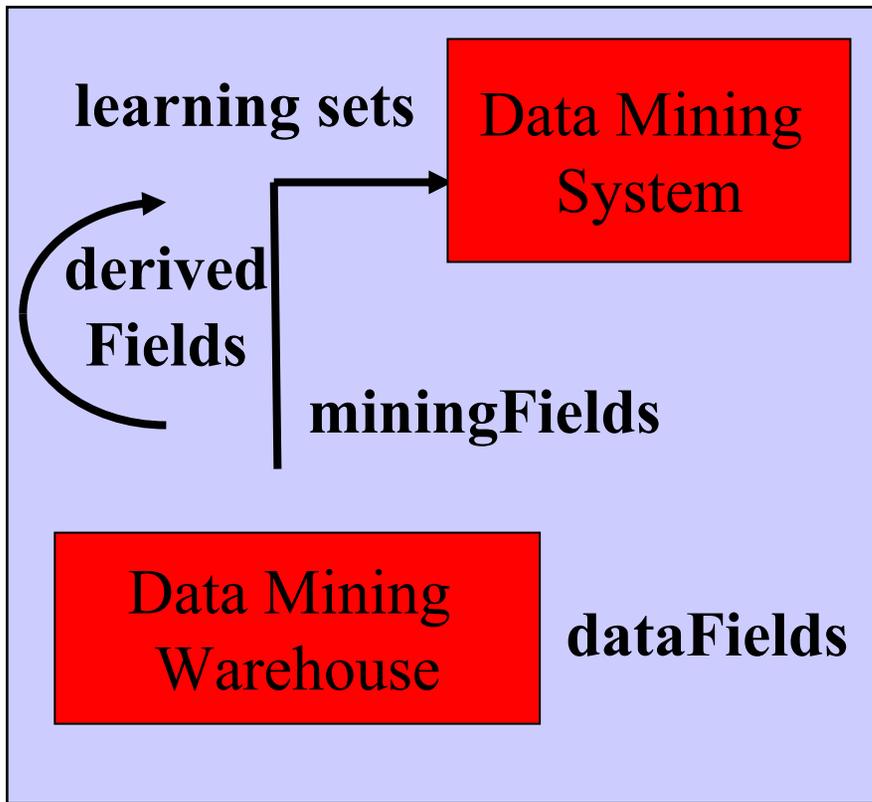
- ❑ Models are deployed in proprietary formats
- ❑ Models are application dependent
- ❑ Models are system dependent
- ❑ Models are architecture dependant
- ❑ Time required to integrate models with other applications can be long.

Predictive Model Markup Language (PMML)

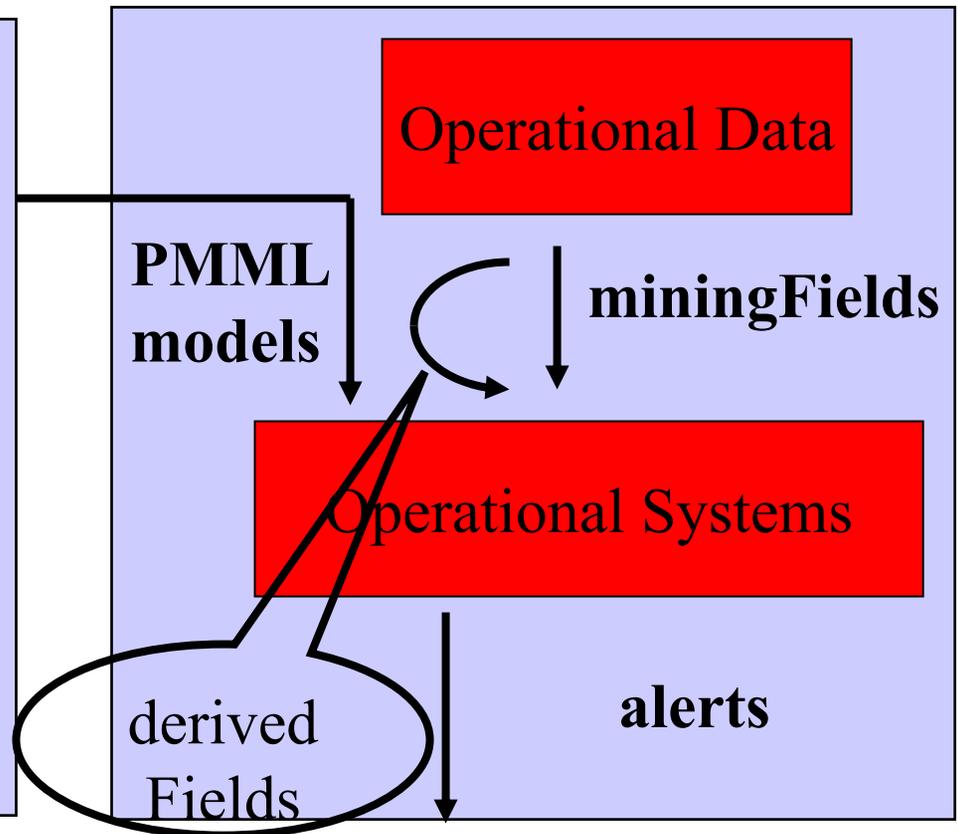
- ❑ Based on XML
- ❑ Benefits of PMML
 - Open standard for Data Mining & Statistical Models
 - Not concerned with the process of creating a model
 - Provides independence from application, platform, and operating system
 - Simplifies use of data mining models by other applications (consumers of data mining models)

PMML Producers, Consumers, & Data Flow

PMML Producers



PMML Consumers



Closely Related Standards



Object model
for representing
data mining metadata:
models, model results
(UML/DTD/XML)



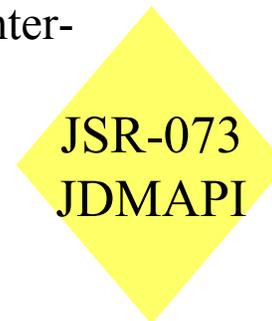
SQL objects for defining,
creating, and applying
data mining models, and
obtaining their results
(SQL)



Representation of data
mining models for inter-
vendor exchange
(DTD/XML)



SQL-like interface
for data mining
operations
(OLE DB/SQL)



Java API for defining,
creating, and applying
data mining models, and
obtaining their results
(Java)

Summary: Cyber Threat Analysis

1. Deployment is more about *alert management* than which algorithm.
2. Events and Profiles enable event driven applications.
3. There is a fundamental need to design algorithms for *high bandwidth* data streams, at 1 Gb/s and higher.
4. The best way to improve a model is to join new from a new source. Data web and *data exploration* systems are designed to make this easier.
5. Standards for data mining are maturing.

For More Information

Robert Grossman

grossman at uic.edu or rlg at opendata.biz
www.lac.uic.edu, www.opendata.biz,
www.rgrossman.com,

Standards

www.dmg.org (PMML, DWTP, etc.)

Data Webs

www.dataspaceweb.net or info at ac.uic.edu

Testbed

Terra Wide Data Mining Testbed (TWDM)
Terabyte Challenge Testbed
www.ncdm.uic.edu