Local Terabytes, Remote Terabytes, and Distributed Terabytes: Four Case Studies in Data Mining

Robert Grossman
National Center for Data Mining
University of Illinois at Chicago
and
Open Data Partners

Introduction: What is data mining?

What is Data Mining?

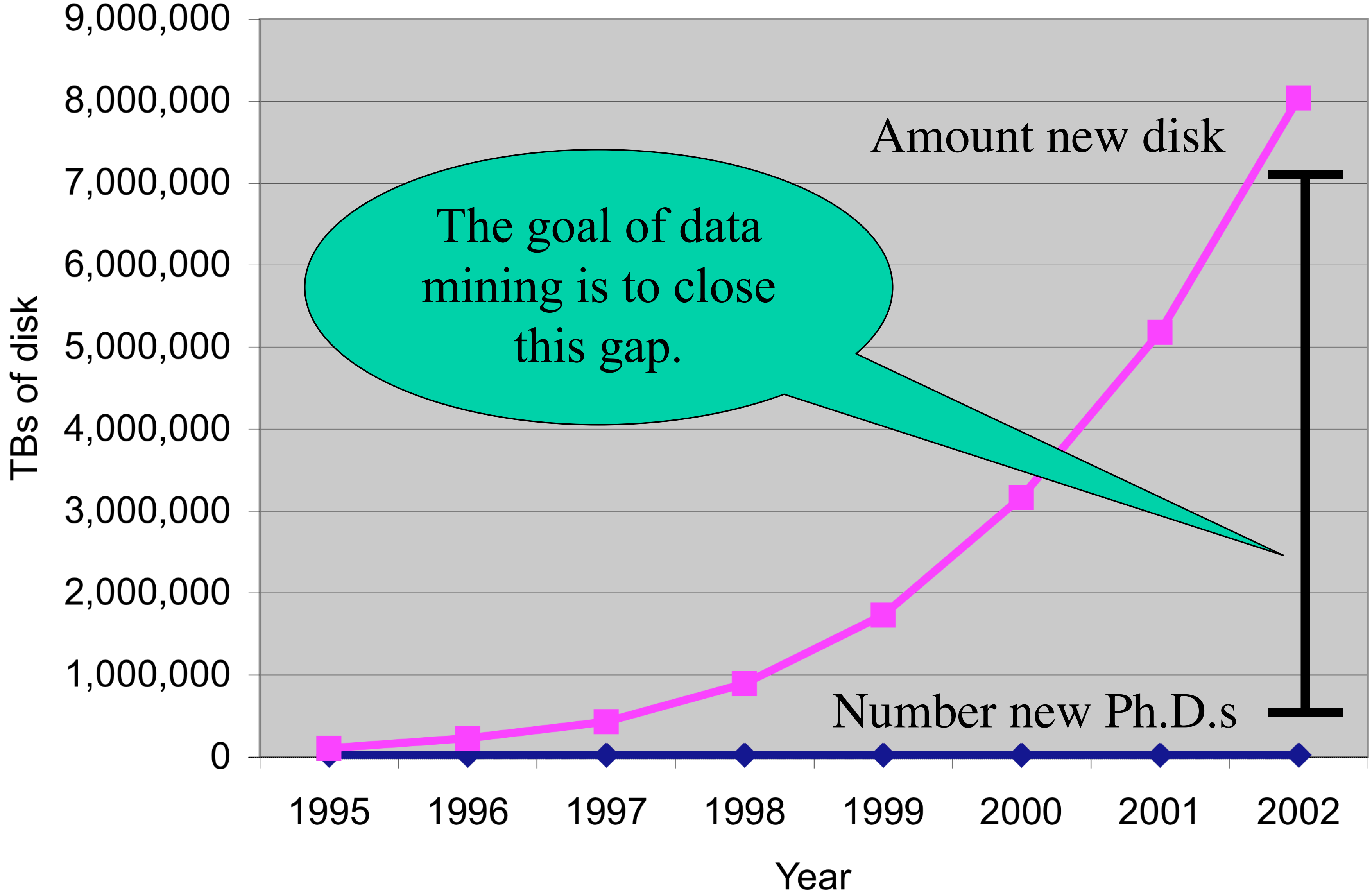
Short definition:

- Finding interesting structure in data. (Interesting implies actionable.)

Long definition:

- Semi-automatic discovery of patterns, correlations, changes, associations, anomalies, and other statistically significant structures in large data sets.

Why Bother: The Data Gap



Two Cultures: Data Science vs Decision Support

Goal	Gain understanding	Take an action
Data	collected & cleaned until conclusions **proven**	analyzed until results are **due**
Methodology	hypothesis testing	lift of model (measured e.g. by ROC)
Challenge	data analysis	data access & cleaning; implementation
Evaluation	results published	ROI, improvement over current decision or business process

Where is the Data?

Number Resources

Billions

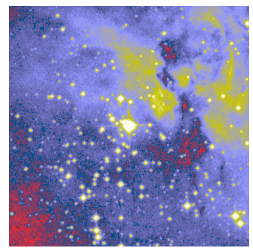
Millions

Thousands

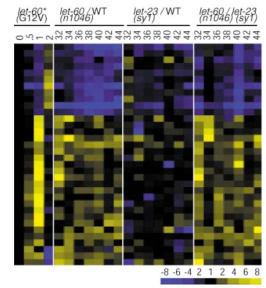
Units

Rigidity of Structure

Relational Databases



Grid-based Databases



Web-based Data

Device-based Data

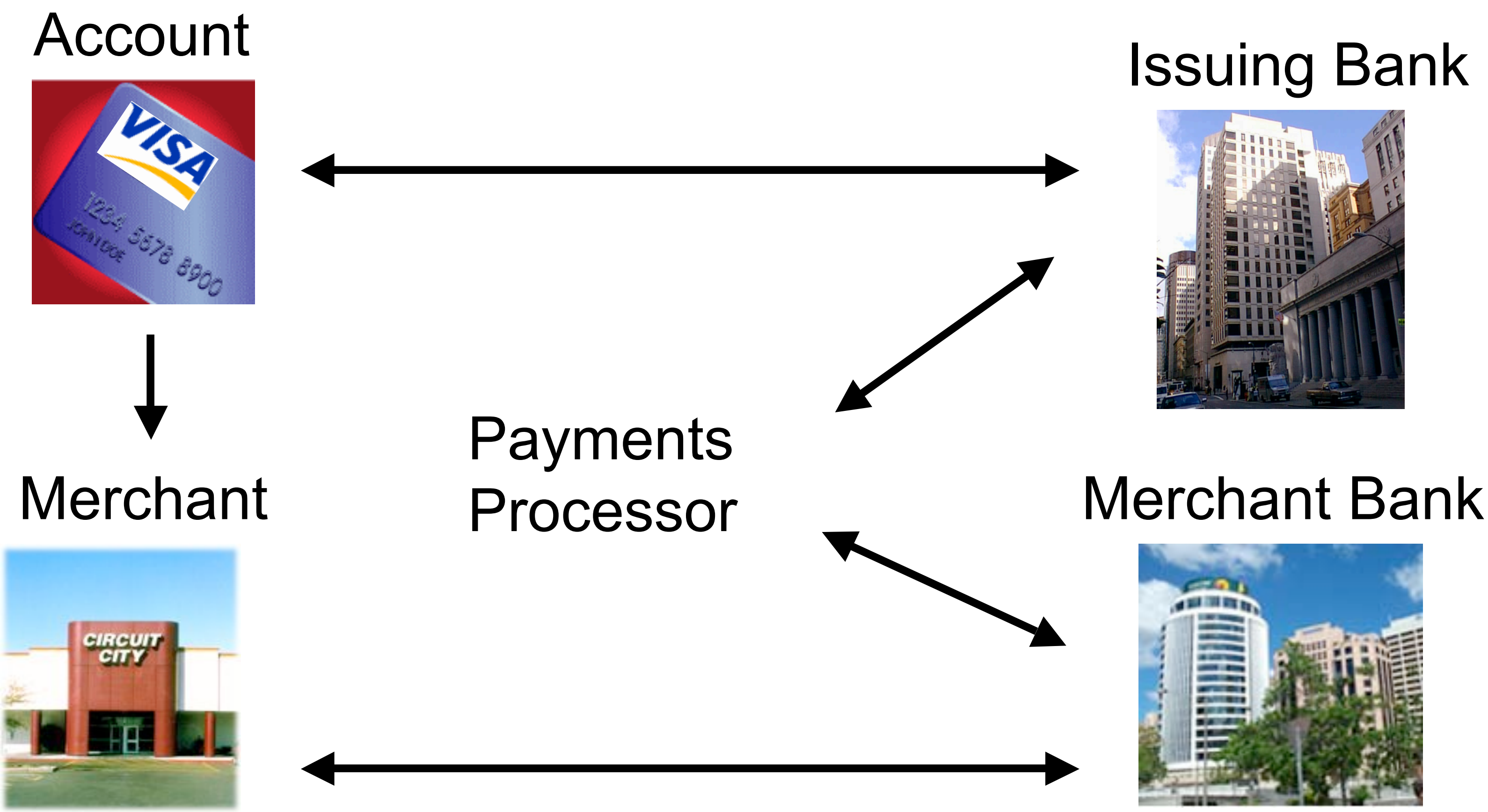
Time

Centralized Control

User Base

Organization Collaboration Community Consumers

Case Study 1:
Payments Card Fraud System



Working with your homogeneous terabytes.

Challenges

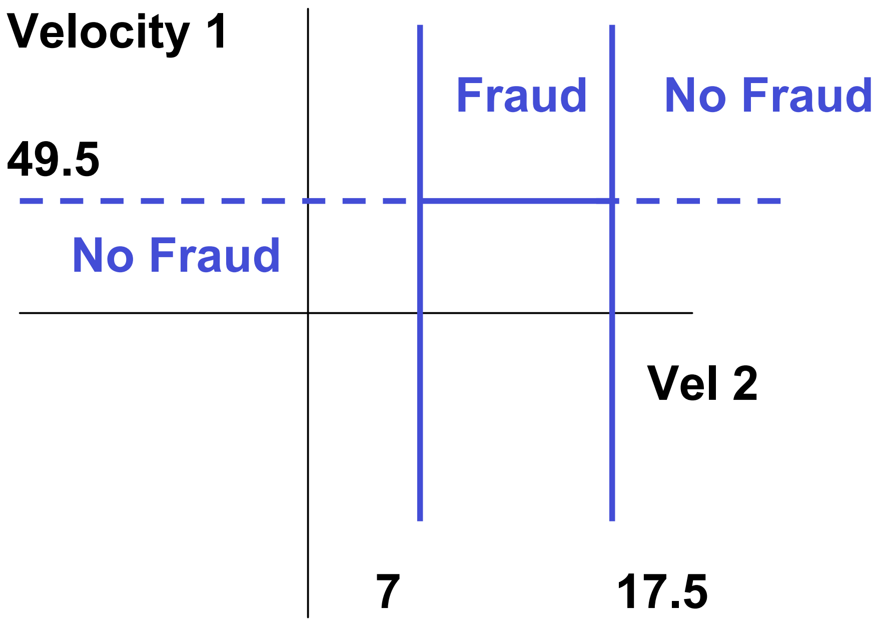
- Technical
  - Develop algorithms that scale to out of memory data.
  - Develop algorithms that scale to high dimensional data.
- Practical
  - Develop algorithms that can be quickly deployed into operational systems.

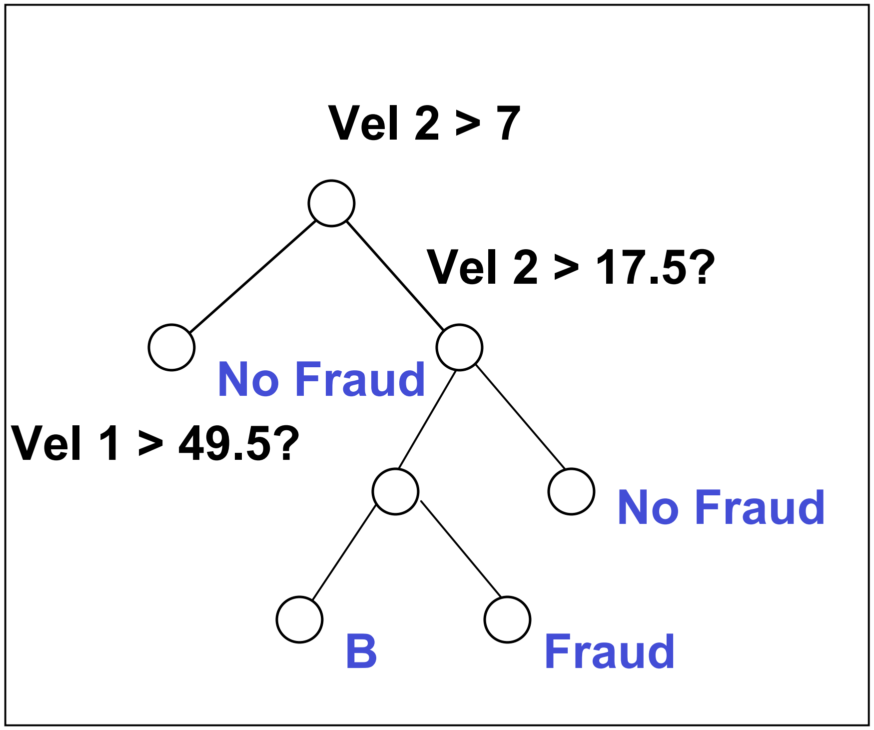
Classification Trees

Vel 1	Vel 2	MCC	Ind 1	Fraud
02	14	33	0	0
24	56	31	0	0
23	51	31	1	1
13	45	28	0	0

- Want a function y = f(x), which predicts the red variable Y using one or more of the blue variables x = (Vel 1, Vel 2, MCC, Ind 1)
- Assume each row is classified 0 or 1

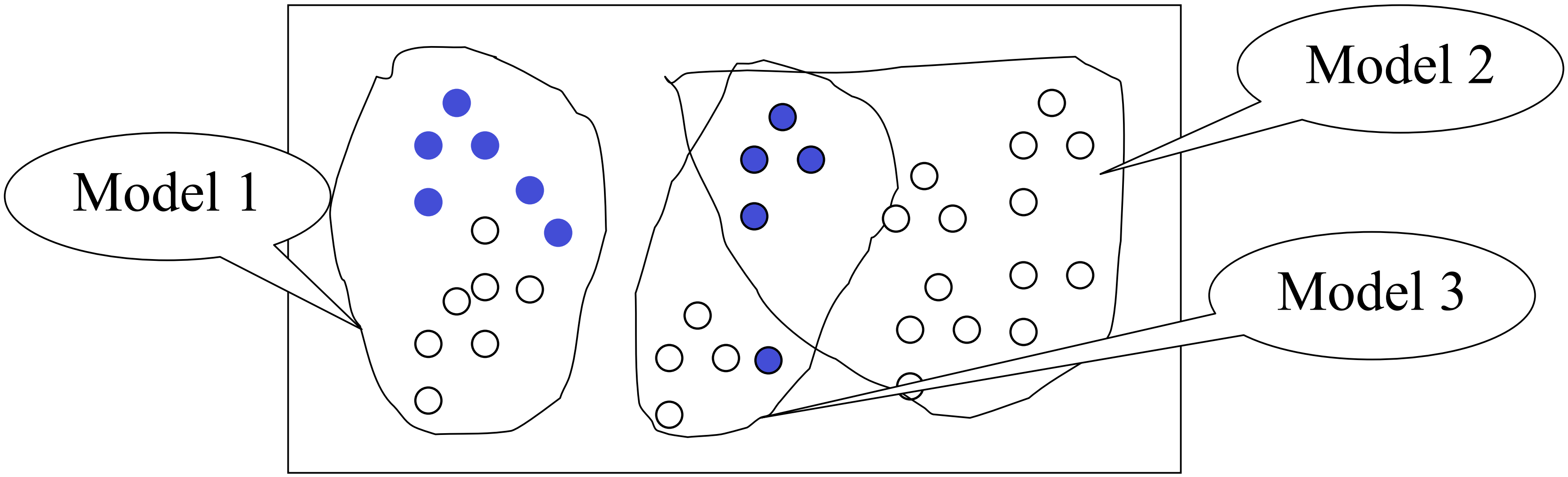
Trees Partition Feature Space





- ❑ Trees partition the feature space into regions by asking whether an attribute is less than a threshold.

Key Idea: Combine Weak Learners



- It is often better to build several models, and then to average them, rather than build one complex model.
- Work in the algebra generated by the multiple classifiers $f_1(x)$, $f_2(x)$, $f_3(x)$, etc.

Combining Weak Learners

1 Classifier	3 Classifiers	5 Classifiers
55%	57.40%	59.30%
60%	64.0%	68.20%
65%	71.00%	76.50%

1

1 1

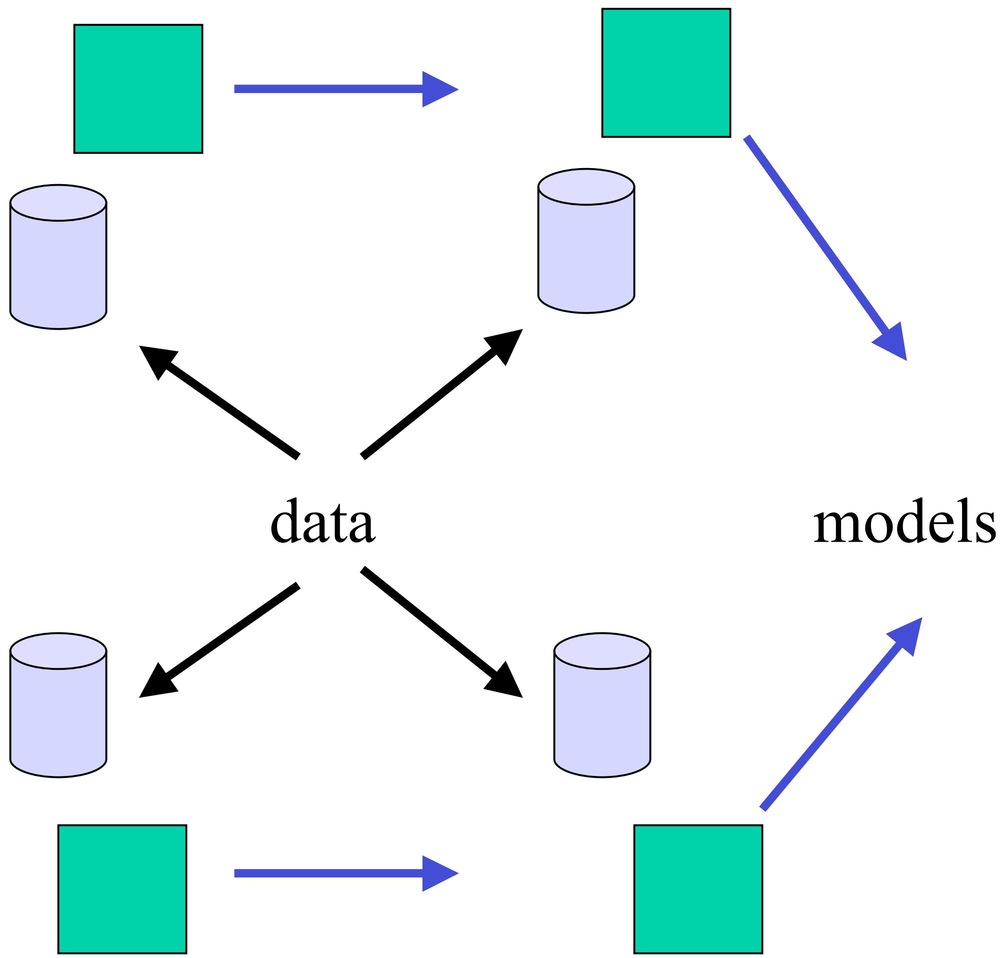
1 2 1

1 3 3 1

1 4 6 4 1

1 5 10 10 5 1

Building Models over Clusters

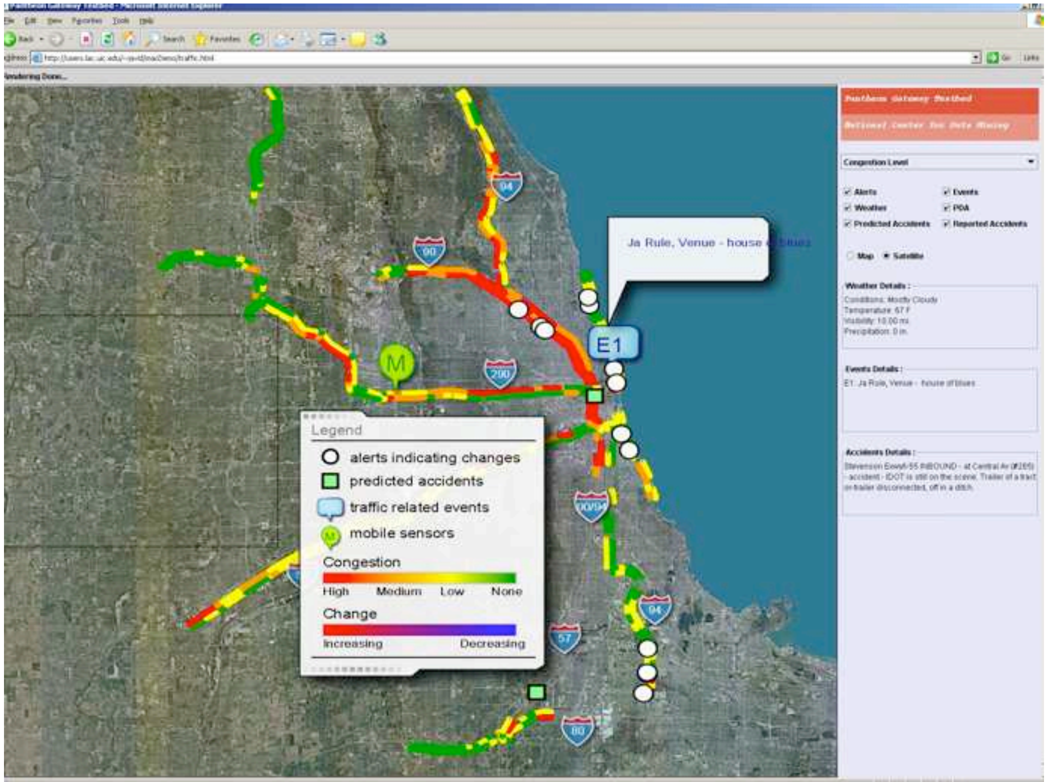


- Scatter data
- Build models (e.g. tree-based model)
- Gather models into **ensemble** (e.g. majority vote for classification & averaging for regression)

Lessons Learned

- Use tree based classifiers to deal with large number of attributes.
- Use ensembles of trees to deal with large amount of data. Used ensembles with 80+ trees. Implemented using clusters.
- Use column-wise warehouses to speed up statistical operations on large data ets.
- Big win from using standards-based scoring engines to deploy models in 24x7x365 systems so that no custom code is required when updating analytics
- Reduced deployment time from months to weeks week. Important for problems like fraud in which target responds and adapts

Case Study 2: Highway Traffic Data



- 833 road sensors
- weather data (images, xml)
- text data about special events

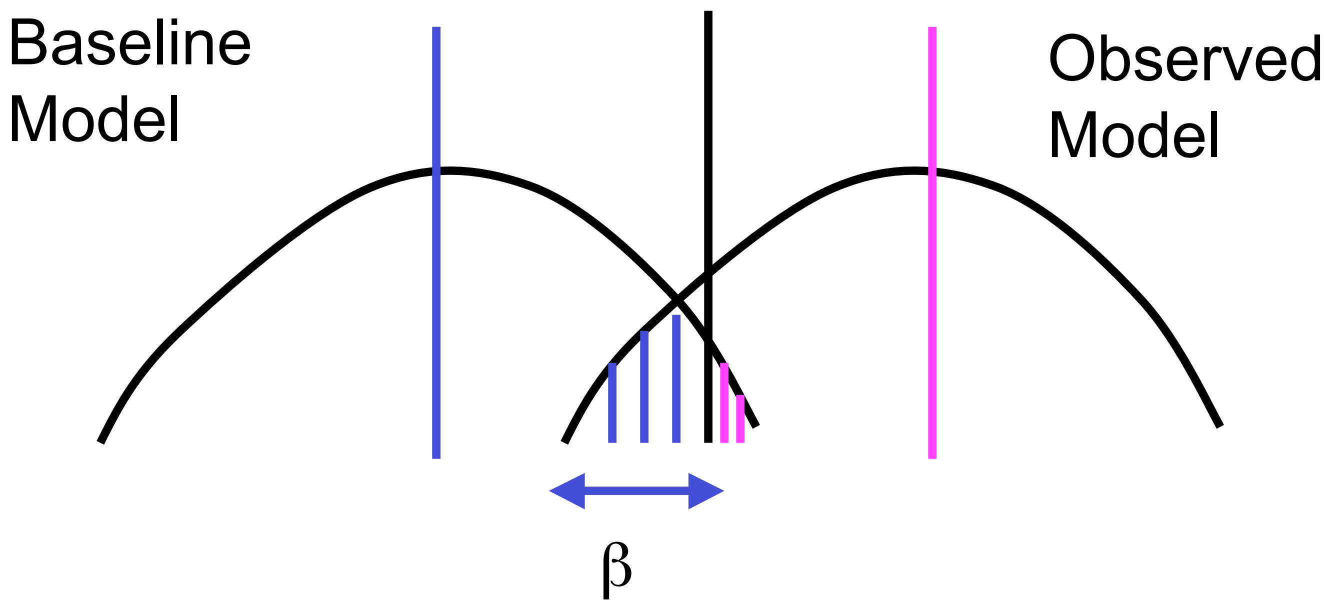
- ❑ Is the traffic speed and volume today (Tuesday, Nov. 15, 3 pm, convention event, no rain) **different** than the baseline?
- ❑ If so, send an alert to a PDA.

Working with your heterogeneous terabytes.

Challenges

- Technical
  - High volume, complex, multi-modal, distributed streaming data
  - Data highly heterogeneous
- Pragmatic
  - Real time alerts to PDA
  - Effectively providing awareness of changes

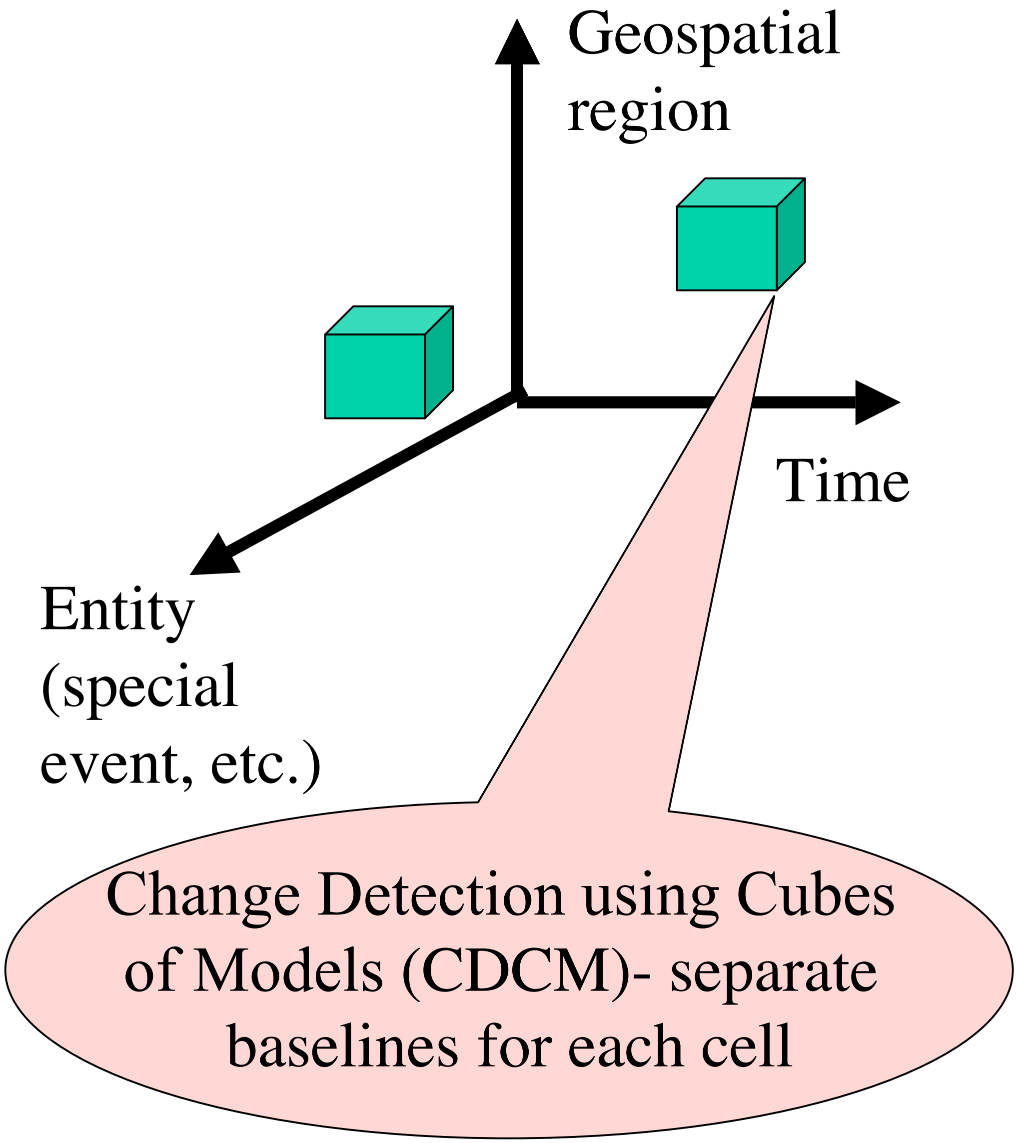
Change Detection Algorithms



- Sequence of events x[1], x[2], x[3], …
- Question: is the observed distribution different than the baseline distribution?
- Used CUSUM & Generalized Likelihood Ratio (GLR) tests

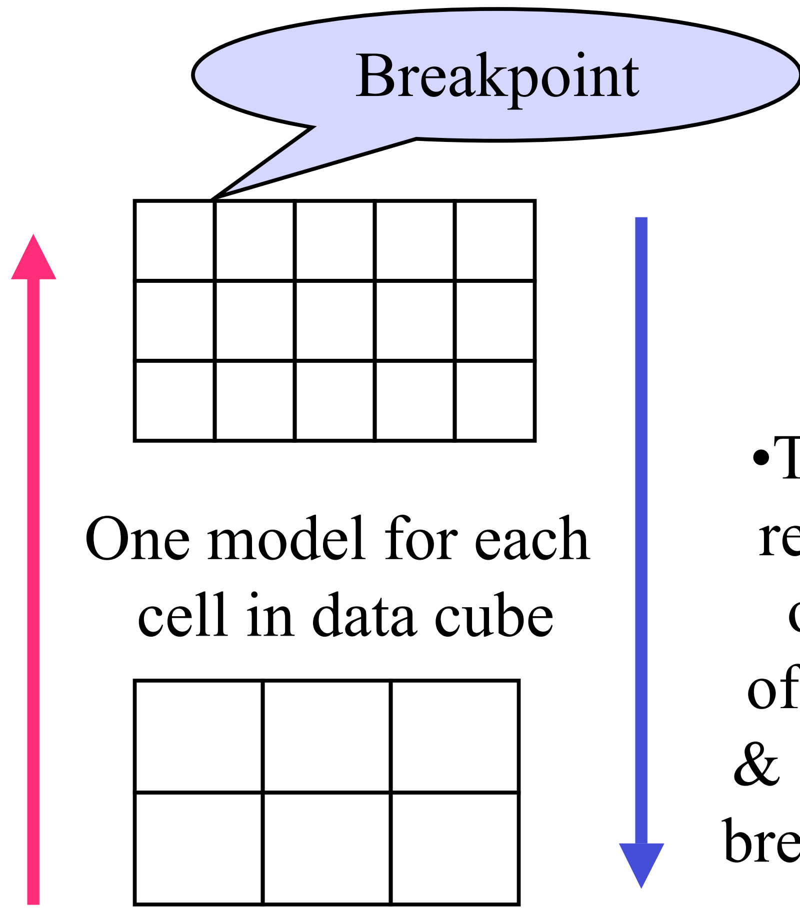
Key Idea 1: Build 10^4+ Models

1. Divide & conquer data (segment) using multidimensional data cubes
2. For each distinct cube, estimate parameters for separate statistical model
3. Detect changes from baselines and send alerts in real time



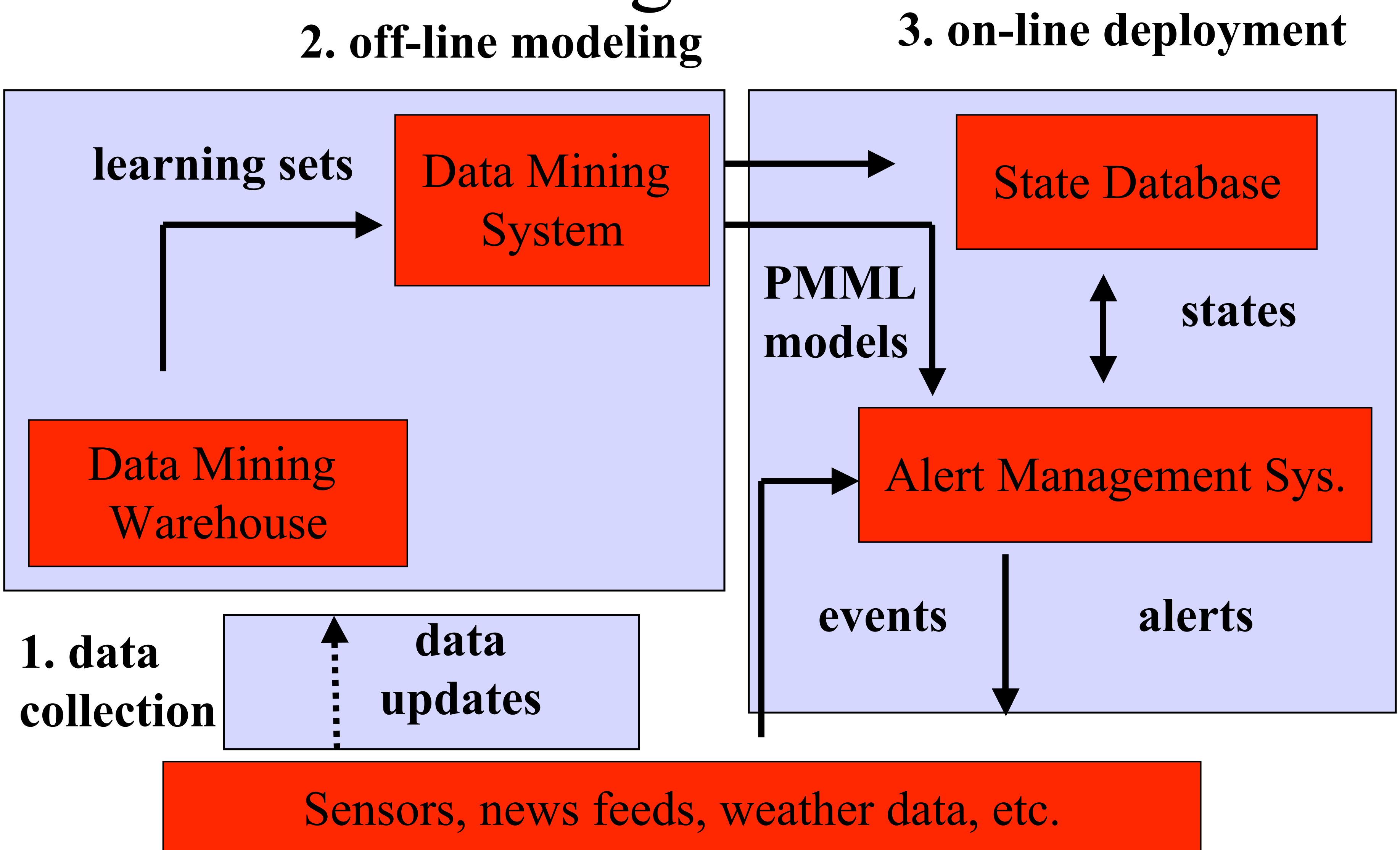
Greedy Meaningful/Manageable Balancing (GMMB) Algorithm

- More alerts
- Alerts more ***meaningful***
- To increase alerts, add breakpoint to split cubes, order by number of new alerts, & select one or more new breakpoints



- Fewer alerts
- Alerts more ***manageable***
- To decrease alerts, remove breakpoint, order by number of decreased alerts, & select one or more breakpoints to remove

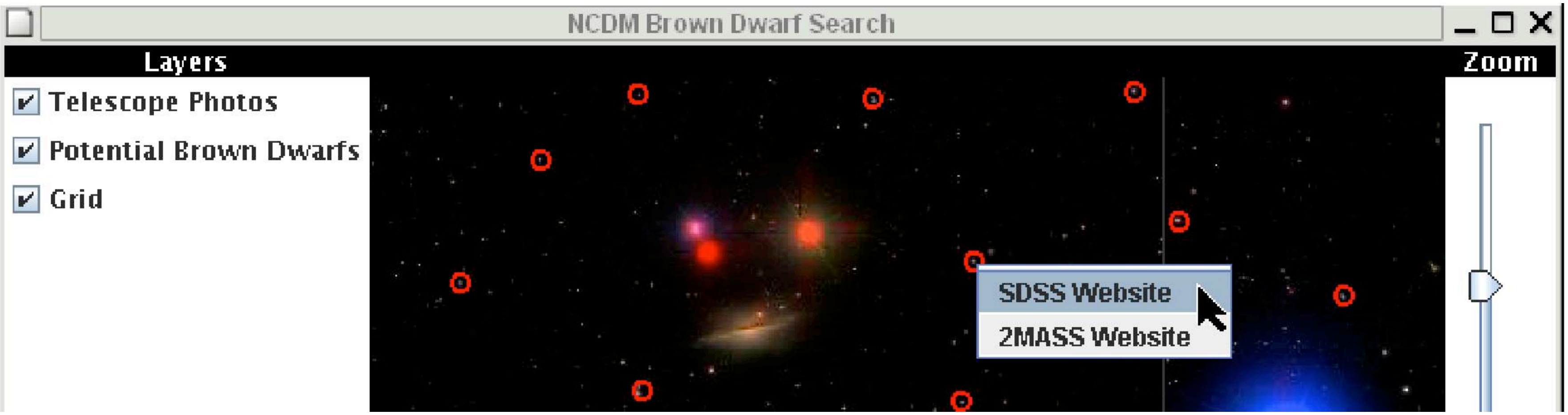
Key Idea 2: Event Based Data Mining Architecture



Lesson Learned

- Change Detection using Cubes of Models (CDCM) is an effective methodology for detecting changes in highly heterogeneous data
- The Greedy Meaningful/Manageable Balancing (GMMB) Algorithm is critical to building a functional system
- An architecture based upon Predictive Model Markup Language (PMML), specifically PMML-producers and PMML-consumers and a few basic segmentation techniques can effectively manage thousands to millions of individual statistical models

Case Study 3 - Integrating Streaming Data



Working with your friends' terabytes….

Finding Candidate Brown Dwarfs

- Sloan Digital Sky Survey (SDSS)
  - 82 million stars
  - Visible spectrum
- Two Micro All Sky Survey (2MASS)
  - 208 million stars
  - Infrared spectrum
- Two separate locations - Query at SC 05 in Seattle
  - SDSS in Tokyo & 2MASS in Chicago
- Found 289,283 Candidate Brown dwarfs
  - Common index structure for each cell in sky (metadata)
  - Object in both locations; infrared value is 2 degree brighter

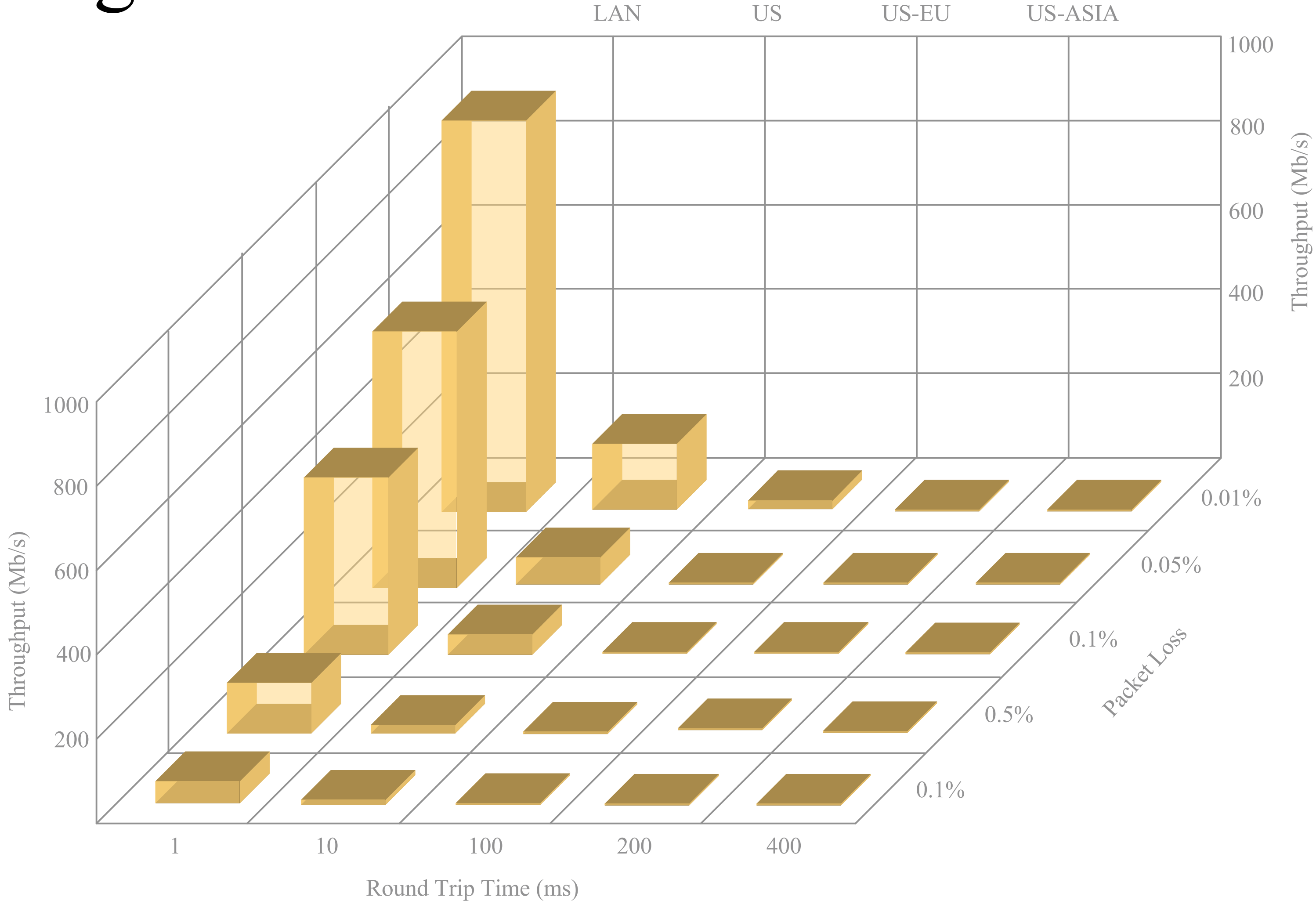
Challenges

- Technical - ***Streaming joins*** not well understood
- Practical - Accessing distributed terabytes of data over ***high bandwidth delay product networks*** is still a problem in practice

$$\text{Throughput} < \frac{\text{MSS}}{\text{RTT} \times \sqrt{\text{Loss}}}$$

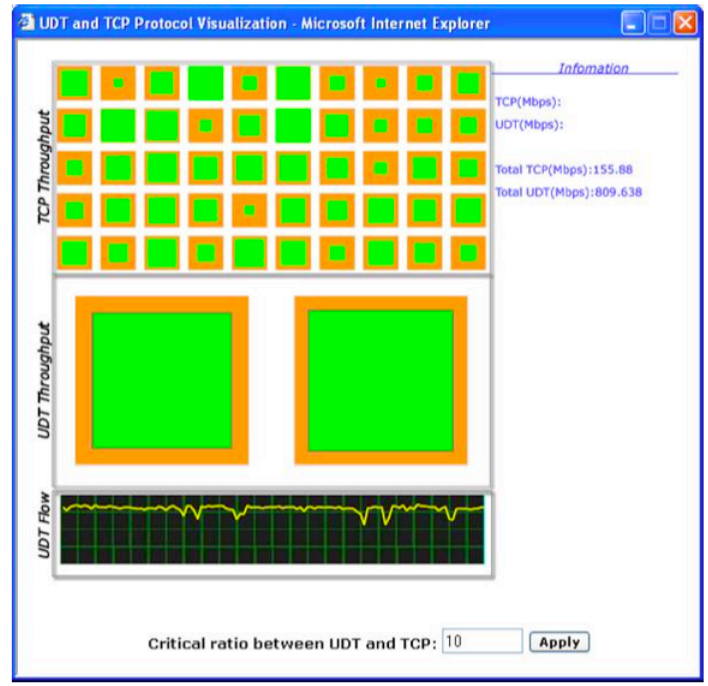
Mathis Equation

Current Protocols (TCP) Don’t Work Over High Bandwidth Wide Area Networks



Key Idea: Network Protocols Matter (Factors of 10x, 100x, 1000x)

1. Goal: Exploit available bandwidth of wide area 10 Gbps networks for distributed data mining.
2. Developed new application level network protocol - UDT
3. UDT is fair to other high volume data flows
4. UDT is friendly to commodity TCP flows.
5. UDT is easy to deploy since application level.



We developed streaming joins and data mining primitives over UDT

UDT Introduced AIMD with Decreasing Increases

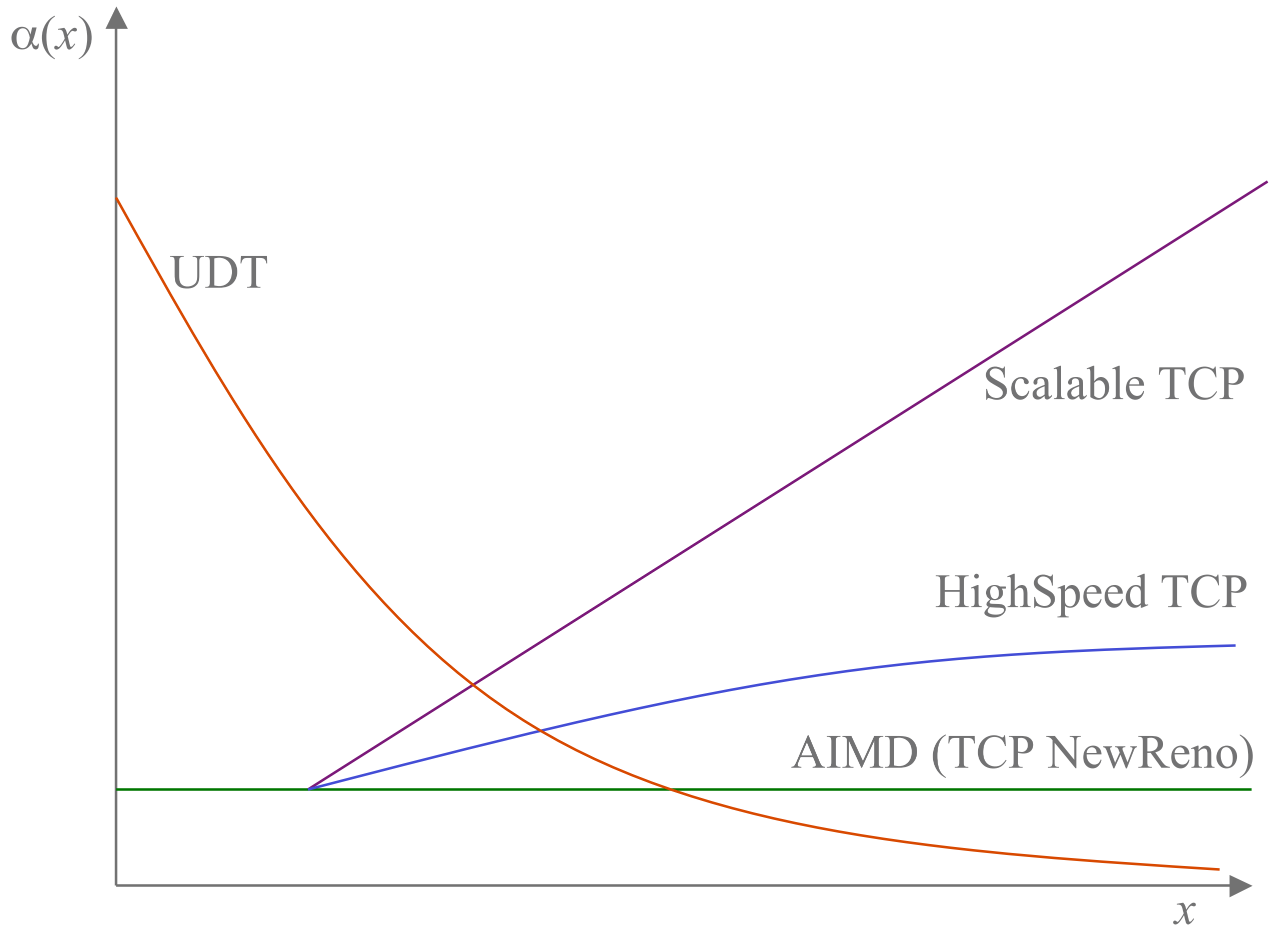
- AIMD (Additive Increases, Multiplicative Decreases)
  - $x = x + \alpha(x)$, for every constant interval (e.g., RTT)
  - $x = (1 - \beta)\, x$, when there is a packet loss event

  where x is the packet sending rate.

- TCP
  - $\alpha(x) \equiv 1$, and the increase interval is RTT.
  - $\beta = 0.5$

- AIMD with Decreasing Increase
  - $\alpha(x)$ is non-increasing, and $\lim_{x \to +\infty} \alpha(x) = 0$.

AIMD with Decreasing Increases



Case Study 4 - Integrating Proteomics Data

NORTHWESTERN UNIVERSITY

Proteomics Grid

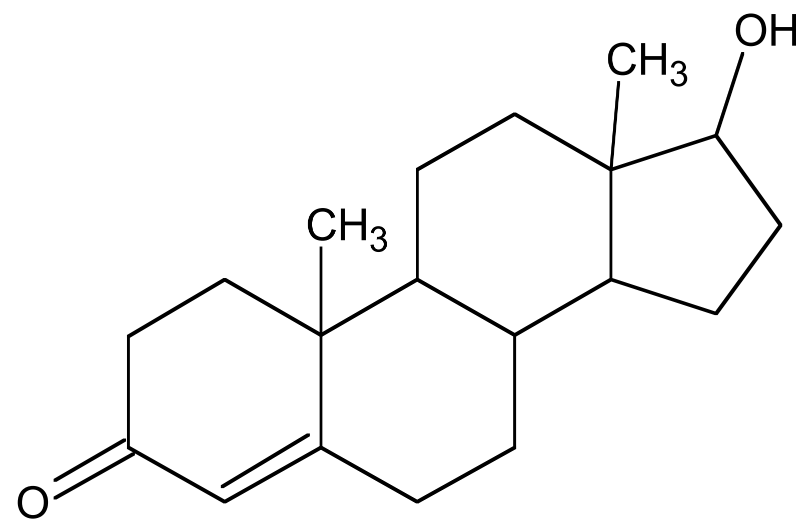
THE UNIVERSITY OF CHICAGO

UIC

Stranger's Gigabytes and Terabytes

What is a Chemical Key?

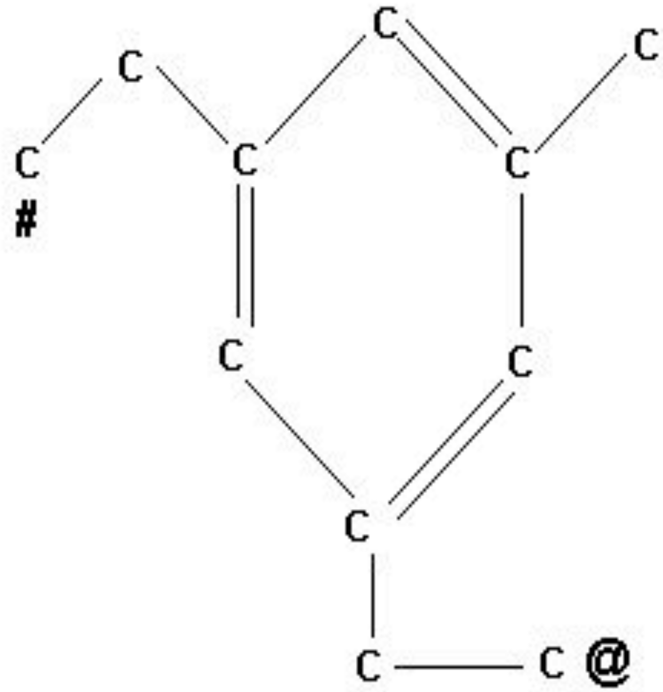
- Testosterone, $C_{19}H_{28}O_2$
- NSC id 9700
- CAS id 58-22-0
- 17-hydroxyandrost-4-en-3-one
- Androlin
- Cristerona T
- Homosteron



A Chemical key is a globally unique key or ID associated with a chemical compound.

Example 1

Name : 3,5-diethyl toluene



Two different Unique SMILES :

1) CCC1=CC(=CC(=C1)C)CC (started at #)

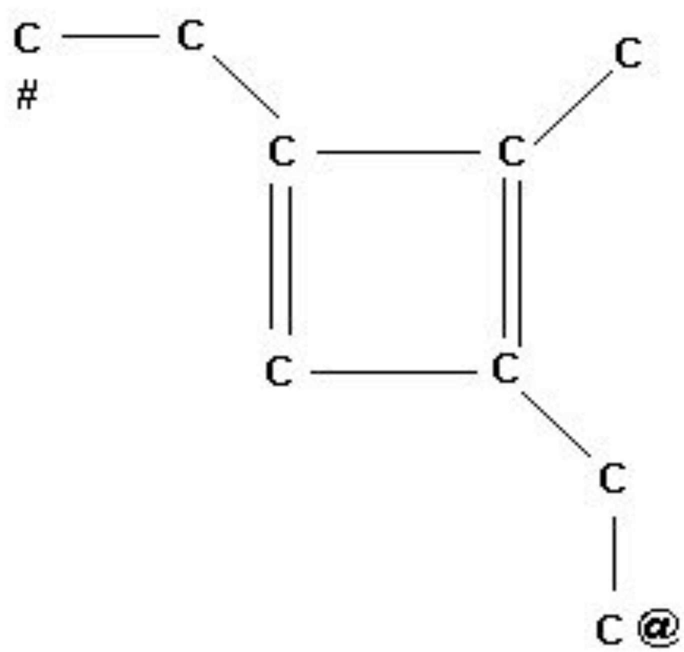
2) CCC1=CC(=CC(=C1)CC)C (started at @)

Universal Chemical Key (UCK)

85C7DC186897FD83D8ECB6B167D988BE

Example 2

Name : 1,3-diethyl-2-methylcyclobuta-1,3-diene



Two different Unique SMILES :

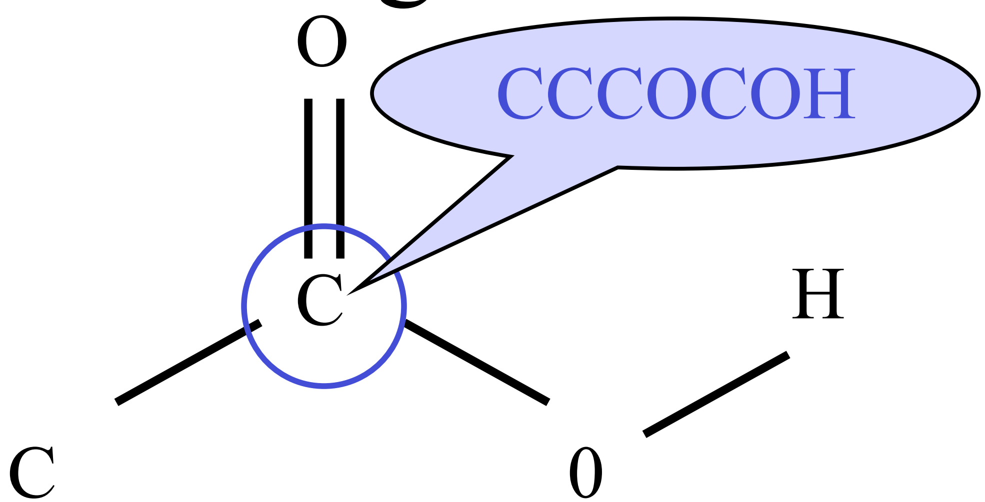
1) CCC1=CC(=C1C)CC (started at #)

2) CCC1=C(C)C(=C1)CC (started at @)

Universal Chemical Key (UCK)

DF0C98C94F6D95226C8FD00028F8F1CB

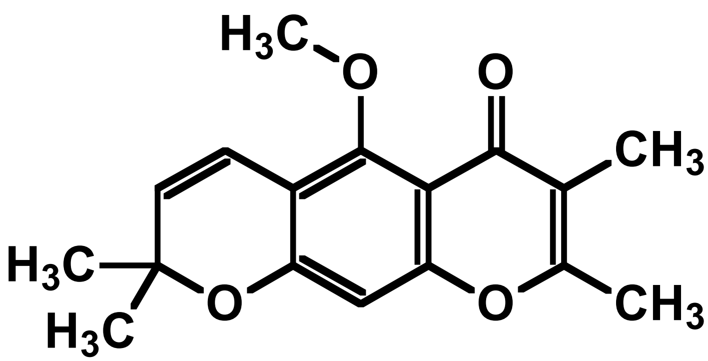
Unique Chemical Key (UCK) Algorithm - Path Labels



- The set of paths is naturally defined.
- Paths can be lex ordered.

1. Set of paths of length less or equal to 2 originating from C: **{CO, CC, COH}.**
2. Lexigraphically order: **[CC, CO, COH].**
3. Concatenate: **CCCOCOH** (path label)

Universal Chemical Keys (UCKs) - Graph Labels



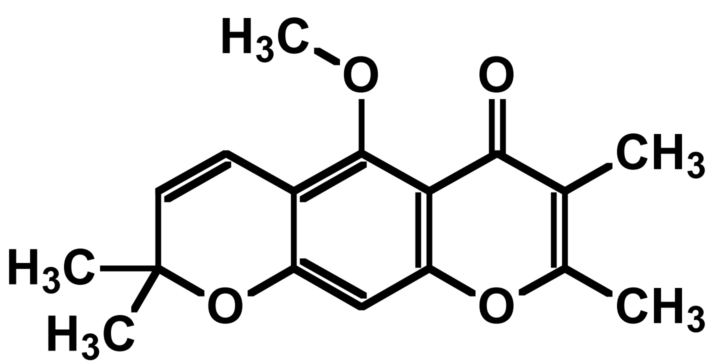
682322

Loop over all pairs of nodes u and v and form “natural labels”

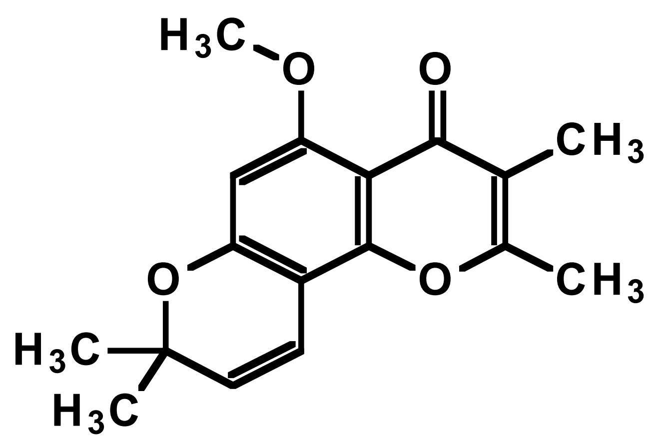
1. Fix depth d.
   Compute path labels $\lambda(u)$, for nodes u.
2. Loop over all pairs of nodes u and v, compute length of shortest path n and form $\lambda(u)$ n $\lambda(v)$.
3. Lex order.
4. Concatenate.
5. Hash.

Example

NSC	Formula	UCK
682322	$C_{17}H_{18}O_4$	132020 …
682323	$C_{17}H_{18}O_4$	098900 …



682322



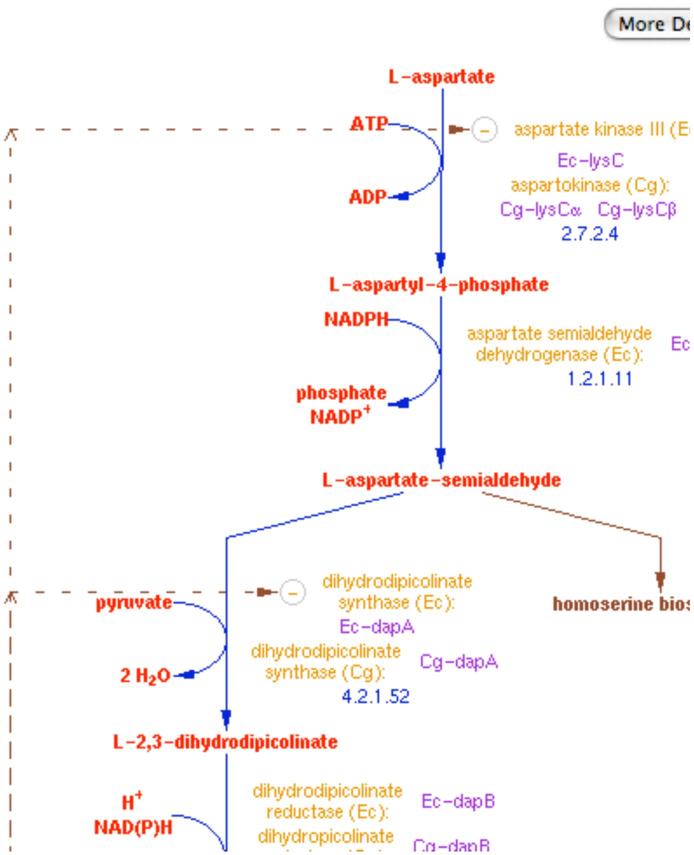
682323

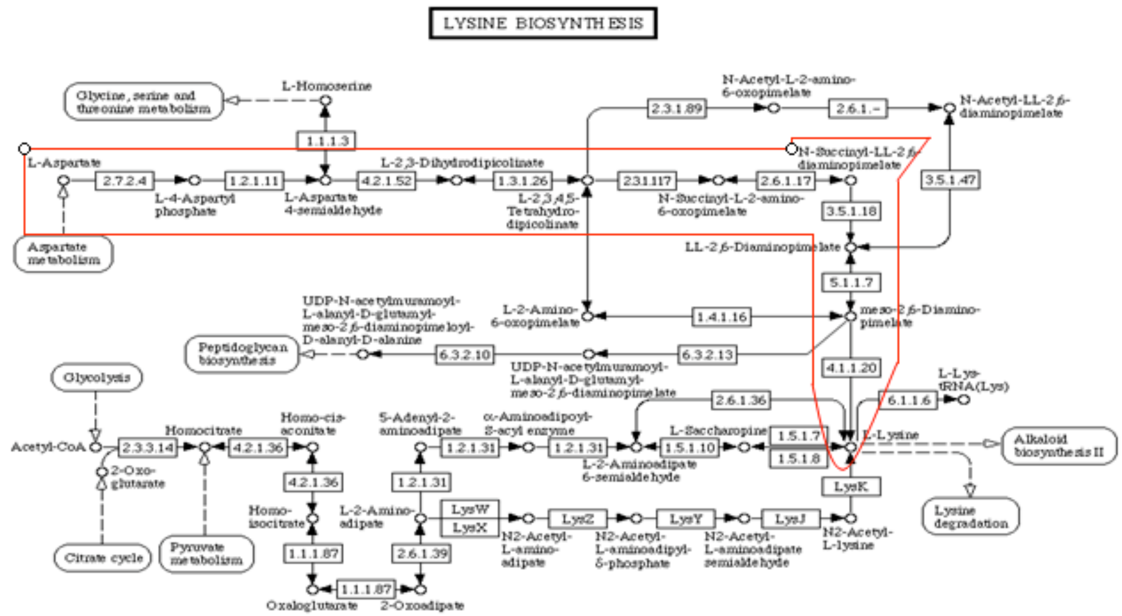
Application 1: Keys for Chemical Compunds (Analysis of NCI Database)

Description	Number	Remark
Total number of chemical compounds	236,917	Some compounds have duplicate entries
Number of chem. comp. with single entry	202,384	All gave unique UCK
Number chem. comp. 2 or more entries	33,533	UCK gave same key to same compounds

Application 2: Keys for Metabolic Pathways

MetaCyc Pathway: lysine biosynthesis I





KEGG database : Lysine biosynthesis

Conclusion

Three Trends for the Next Five Years

1. Forget data mining, the real pay-off is data integration, especially for distributed data
2. For many problems, streaming algorithms will be the only choice available, whether we like it or not
3. Analytic algorithms for working with more complex data, e.g. graphs, semi-structured data, etc. will become more and more important.

References (1 of 3)

1. Payments Card Fraud

R. L. Grossman, H. Bodek, D. Northcutt, and H. V. Poor, Data Mining and Tree-based Optimization, Proceedings of the Second International Conference on Knowledge Discovery and Data Mining, E. Simoudis, J. Han and U. Fayyad, editors, AAAI Press, Menlo Park, California, 1996, pp 323-326.

Robert L. Grossman, Alert Management Systems: A Quick Introduction, in Managing Cyber Threats: Issues, Approaches and Challenges, edited by Vipin Kumar, Jaideep Srivastava and Aleksandar Lazarevic, Springer Science+Business Media, Inc., New York, pages 281-291, ISBN 0-387-24226-0.

2. Change Detection for Highway Traffic

Robert L. Grossman, Michal Sabala, Javid Alimohideen, Anushka Aanand, John Chaves, John Dillenburg, Steve Eick, Jason Leigh, Peter Nelson, Mike Papka, Doug Rorem, Rick Stevens, Steve Vejcik, Leland Wilkinson, and Pei Zhang, Real Time Change Detection and Alerts from Highway Traffic Data, ACM/IEEE SC 2005 Conference (SC'05).

L. Wilkinson, A. Anand and R. Grossman. High-dimensional Visual Analytics: Interactive Exploration Guided by Pairwise Views of Point Distribution, IEEE Transactions on Visualization and Computer Graphics, 2006.

Rajmonda Sulo, Anushka Anand, Leland Wilkinson, Robert Grossman, and Stephen Eick, Topographically-Based Real-Time Traffic Anomaly Detection in a Metropolitan Highway System, submitted for publication.

References (2 of 3)

3. Sloan Distribution and Continuous Discovery

Yunhong Gu, Xinwei Hong, and Robert Grossman, Experiences in Design and Implementation of a High Performance Transport Protocol, SC 04

Yunhong Gu and Robert Grossman, Supporting Configurable Congestion Control in Data Transport Services, ACM/IEEE SC 2005 Conference (SC'05).

Robert L. Grossman, Yunhong Gu, David Handley, and Michal Sabala Joe Mambretti, Alex Szalay and Ani Thakar, Kazumi Kumazoe and Oie Yuji, Minsun Lee, Yoonjoo Kwon, and Woojin Seok, Data Mining Middleware for Wide Area High Performance Networks, Journal of Future Generation Computer Systems (FGCS), 2006.

Robert L. Grossman, Yunhong Gu, Xinwei Hong, Antony Antony, Johan Blom, Freek Dijkstra, and Cees de Laat, Teraflows over Gigabit WANs with UDT, Journal of Future Computer Systems, Elsevier Press, Volume 21, Number 4, 2005, pages 501-513.

Marco Mazzucco, Asvin Ananthanarayan, Robert L. Grossman, Jorge Levera, and Gokulnath Bhagavantha Rao, Merging Multiple Data Streams on Common Keys over High Performance Networks, Proceedings of the IEEE/ACM SC2002 Conference, 2002, IEEE Computer Society, page 67.

References (3 of 3)

4. Data Integration and Universal Chemical Keys (UCKs)

Greeshma Neglur and Robert L. Grossman, Assigning Unique Keys to Chemical Compounds for Data Integration: Some Interesting Counter Examples, 2nd International Workshop on Data Integration in the Life Sciences (DILS 2005), La Jolla, July 20-22, 2005.

Robert L. Grossman, Pavan Kasturi, Donald Hamelberg, Bing Liu, An Empirical Study of the Universal Chemical Key Algorithm for Assigning Unique Keys to Chemical Compounds, Journal of Bioinformatics and Computational Biology, 2004, Volume 2, Number 1, 2004, pages 155-171.

5. Data Mining and Data Integration Architectures

Robert Grossman, Mark Hornick, and Gregor Meyer, Data Mining Standards Initiatives, Communications of the ACM, Volume 45-8, 2002, pages 59-61.

Robert Grossman, and Marco Mazzucco, DataSpace - A Web Infrastructure for the Exploratory Analysis and Mining of Data, IEEE Computing in Science and Engineering, July/August, 2002, pages 44-51.

Thank you.

For more information: www.ncdm.uic.edu