CS 277, Data Mining

Introduction

Padhraic Smyth
Department of Computer Science
Bren School of Information and Computer Sciences
University of California, Irvine

Today's Lecture

- Discuss class structure and organization
  - Assignment 1
  - Projects
- What is data mining?
- Examples of large data sets
- Brief history of data mining
- Data mining tasks
  - Descriptive versus Predictive data mining
- The "dark side" of data mining

Goals for this class

- Learn how to extract useful information from data in a systematic way

- Emphasis on the process of data mining and how to analyze data
  - understanding specific algorithms and methods is important
  - but also…emphasize the "big picture" of why, not just how
  - less emphasis on mathematics than in 273A, 274A, etc

- Specific topics will include
  - Text classification, document clustering, topic models
  - Web data analysis, e.g., clickstream analysis
  - Recommender systems
  - Social network analysis

- Builds on knowledge from CS 273A, 274A

Logistics

- Grading
  - 20% homeworks
    - 2 assignments over first 3 weeks
    - Assignment 1 due Wednesday next week
  - 80% class project
    - Will discuss in later lecture
- Web page
  - www.ics.uci.edu/~smyth/courses/cs277
- Reading
  - Weekly reading plus reference material
  - No required textbook
- Prerequisites
  - Either ICS 273 or 274 or equivalent

Assignment 1

http://www.ics.uci.edu/~smyth/courses/cs277/assignment1.xht

Due: Wednesday Jan 15th

Outline

- Census income data
- Exploratory Data Analysis tasks
- Software
  - R/Rminer, Matlab, and scikit-learn

Software Options

- General programming languages
  - Python, Java, C++, etc
  - Strengths: flexibility, speed
  - Weaknesses: lack of built-in functionality for data analysis (e.g., for plotting), learning curve
  - Ideal for final implementation/operational use

- Scientific programming environments
  - R, Matlab, Python (NumPy, SciPy, scikit-learn)
    - http://scikit-learn.org/stable/index.html
  - Strengths: considerable flexibility, many built-in library functions and public toolboxes
  - Weaknesses: slower than general programming languages, learning curve
  - Ideal for prototyping and development

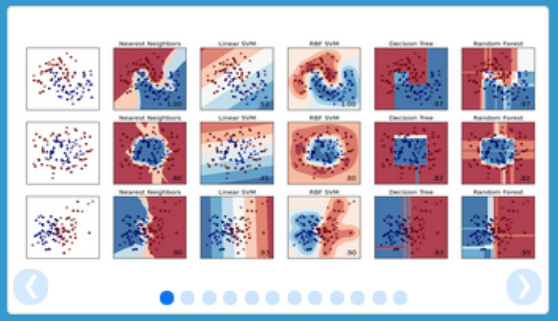
- Data analysis packages
  - Weka, SAS, etc
  - Strengths: little or no learning curve, ideal for analysts with limited programming skills
  - Weaknesses: not flexible, limited programmability, may be unable to handle large data sets
  - Useful for exploratory data analysis and for analysts with limited programming skills

scikit learn

Home Installation Documentation Examples

Google™ Custom Search Search

Fork me on GitHub



scikit-learn

Machine Learning in Python

- Simple and efficient tools for data mining and data analysis
- Accessible to everybody, and reusable in various contexts
- Built on NumPy, SciPy, and matplotlib
- Open source, commercially usable - BSD license

Classification

Identifying to which set of categories a new observation belong to.

Applications: Spam detection, Image recognition.
Algorithms: *SVM*, *nearest neighbors*, *random forest*, ... — *Examples*

Regression

Predicting a continuous value for a new example.

Applications: Drug response, Stock prices.
Algorithms: *SVR*, *ridge regression*, *Lasso*, ... — *Examples*

Clustering

Automatic grouping of similar objects into sets.

Applications: Customer segmentation, Grouping experiment outcomes
Algorithms: *k-Means*, *spectral clustering*, *mean-shift*, ... — *Examples*

Dimensionality reduction

Reducing the number of random variables to consider.

Applications: Visualization, Increased efficiency
Algorithms: *PCA*, *Isomap*, *non-negative matrix factorization*. — *Examples*

Model selection

Comparing, validating and choosing parameters and models.

Goal: Improved accuracy via parameter tuning
Modules: *grid search*, *cross validation*, *metrics*. — *Examples*

Preprocessing

Feature extraction and normalization.

Application: Transforming input data such as text for use with machine learning algorithms.
Modules: *preprocessing*, *feature extraction*. — *Examples*

News

On-going development: What's new (changelog)

Community

Questions? See stackoverflow # scikit-learn

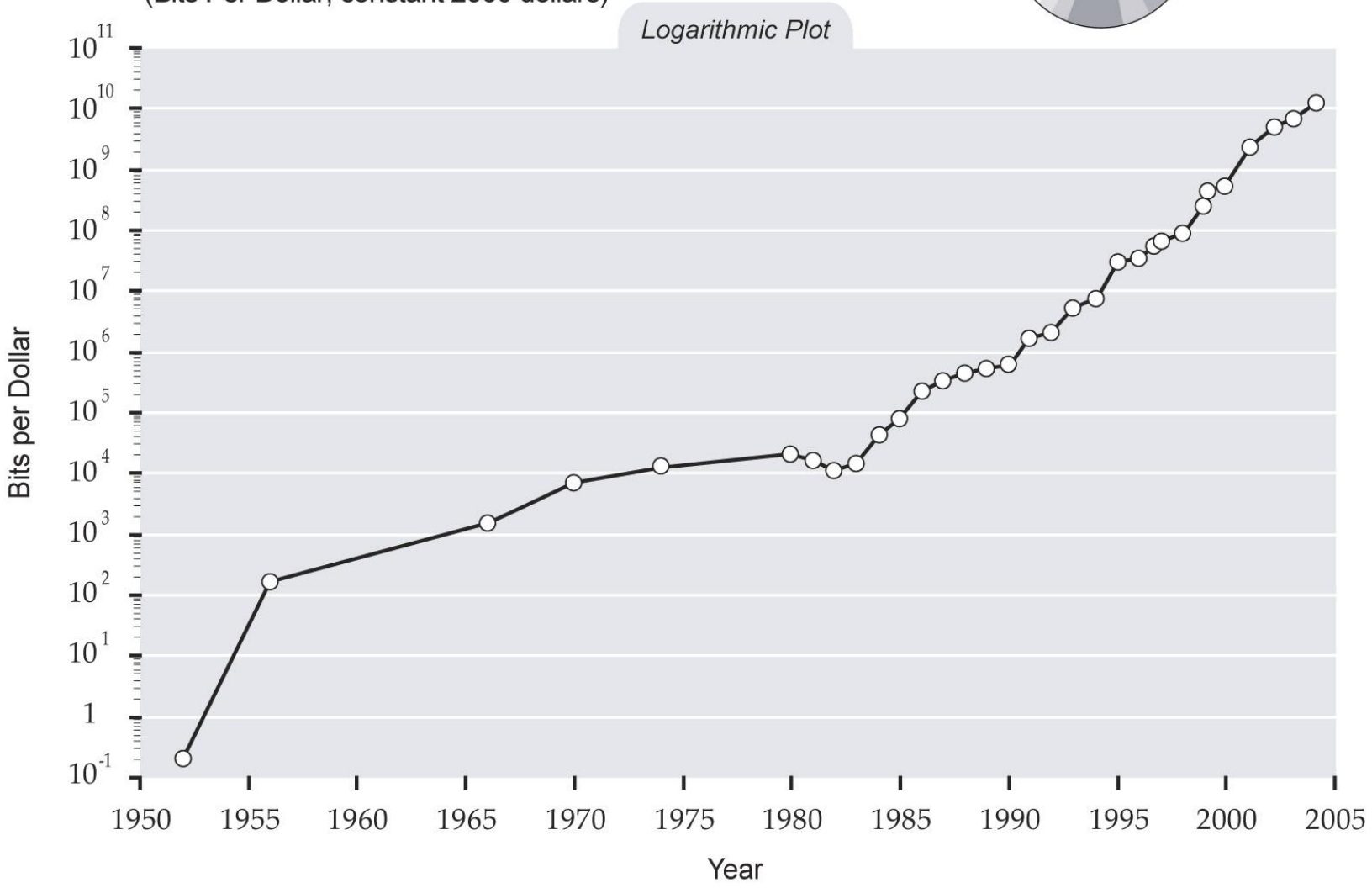
Mailing list: scikit-learn-

Who uses scikit-learn?

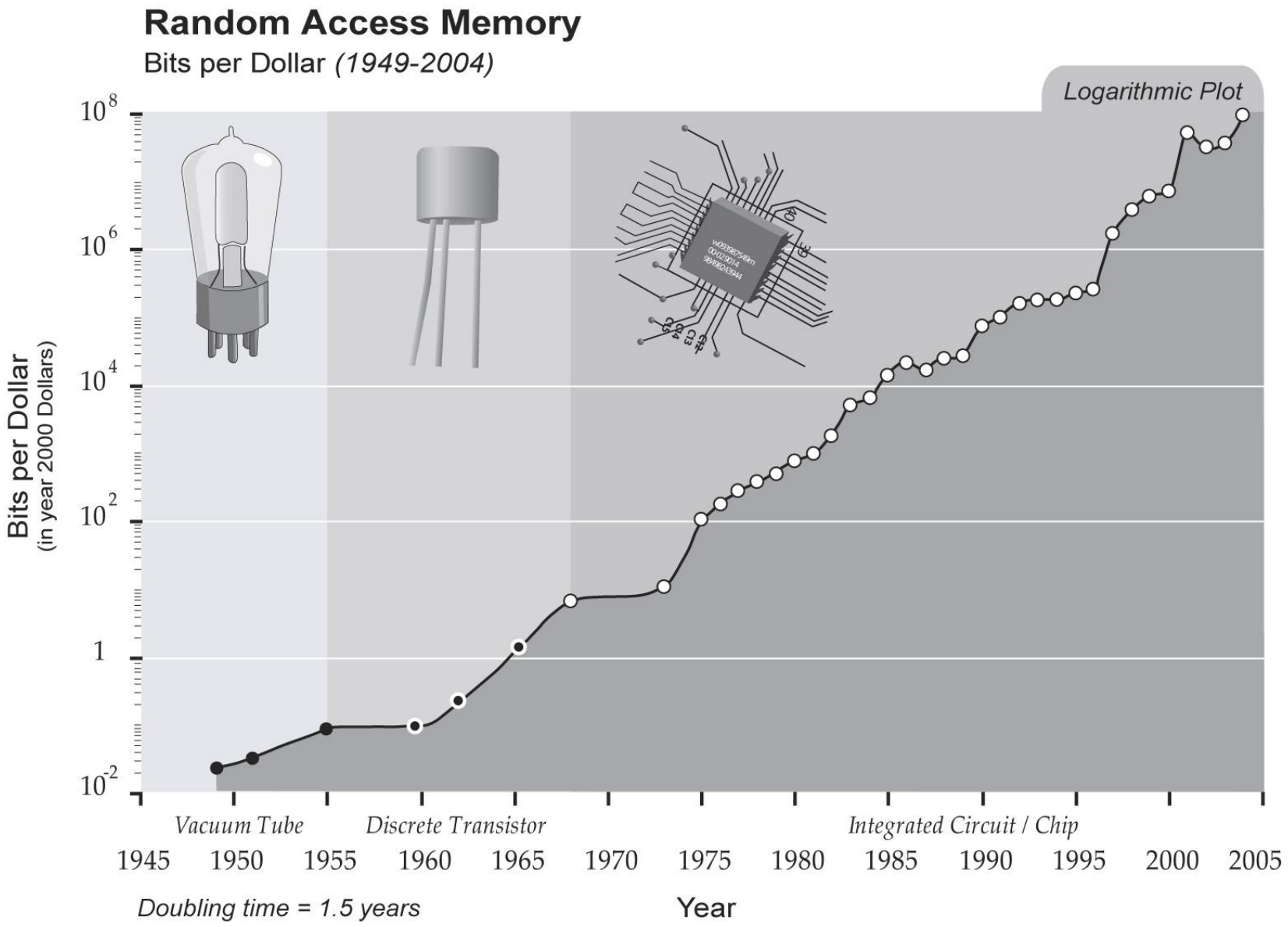
A SHORT HISTORY OF DATA MINING

Technological Driving Factors

- Larger, cheaper memory
  - Moore's law for magnetic disk density
    "capacity doubles every 18 months" (Jim Gray, Microsoft)
  - storage cost per byte falling rapidly
- Faster, cheaper processors
  - the supercomputer of 20 years ago is now on your desk
- Success of relational database technology
  - everybody is a "data owner"
- Advances in computational statistics/machine learning
  - significant advances in theory and methodologies

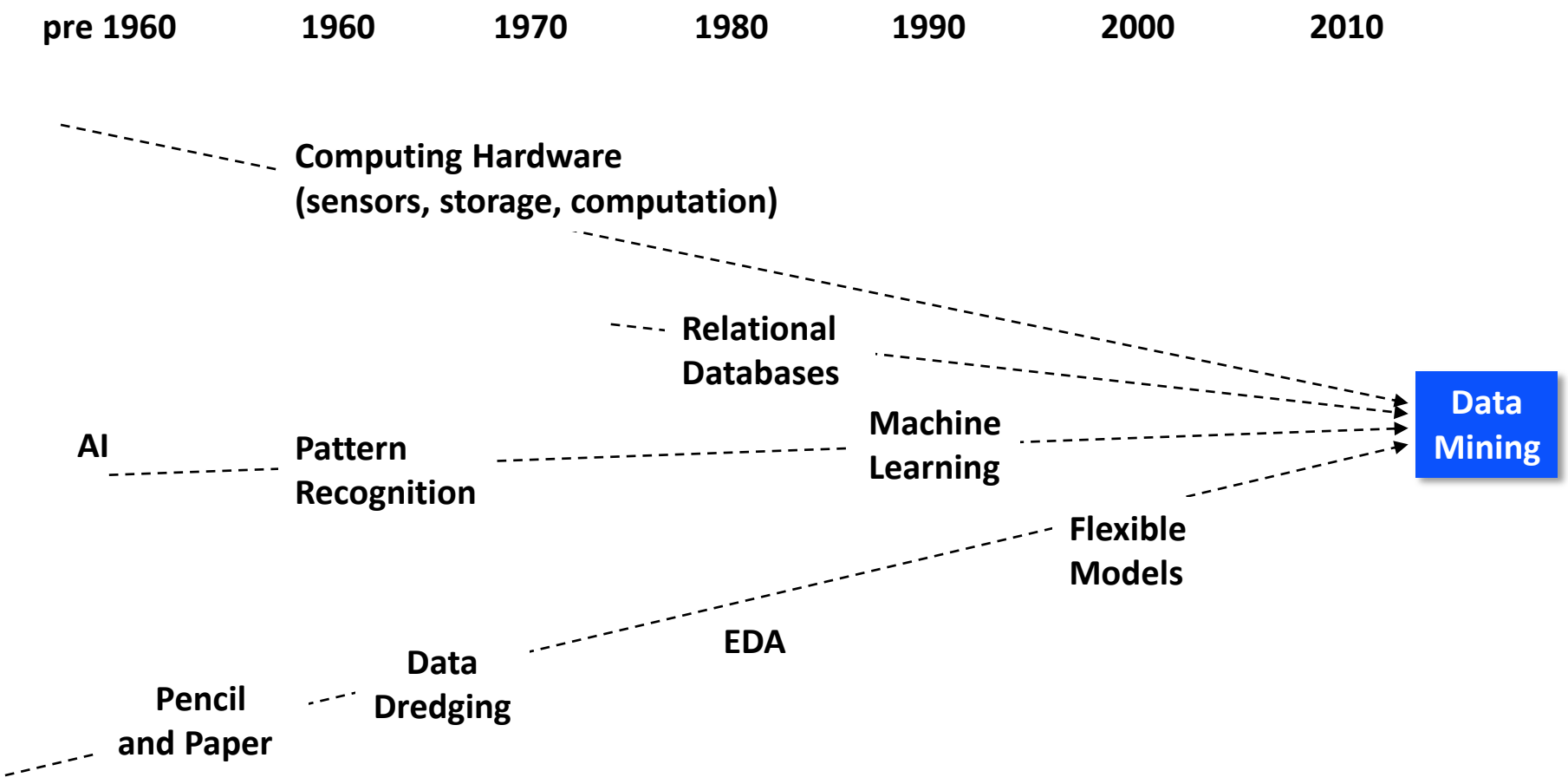


From Ray Kurzweil, singularity.com



From Ray Kurzweil, singularity.com

Origins of Data Mining

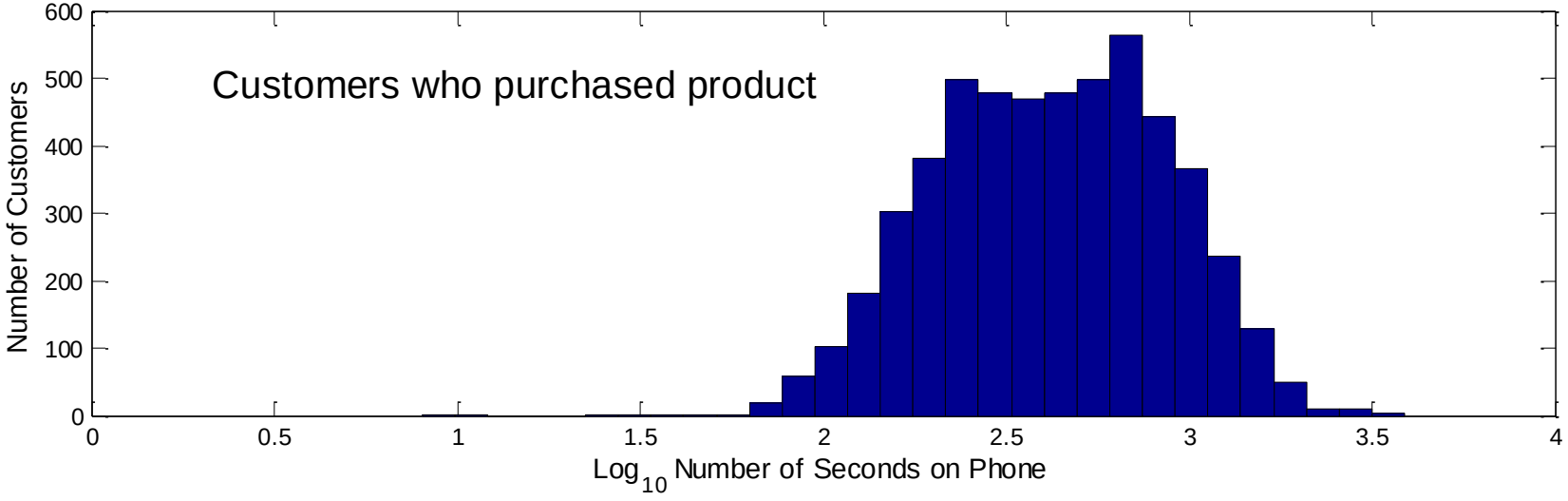


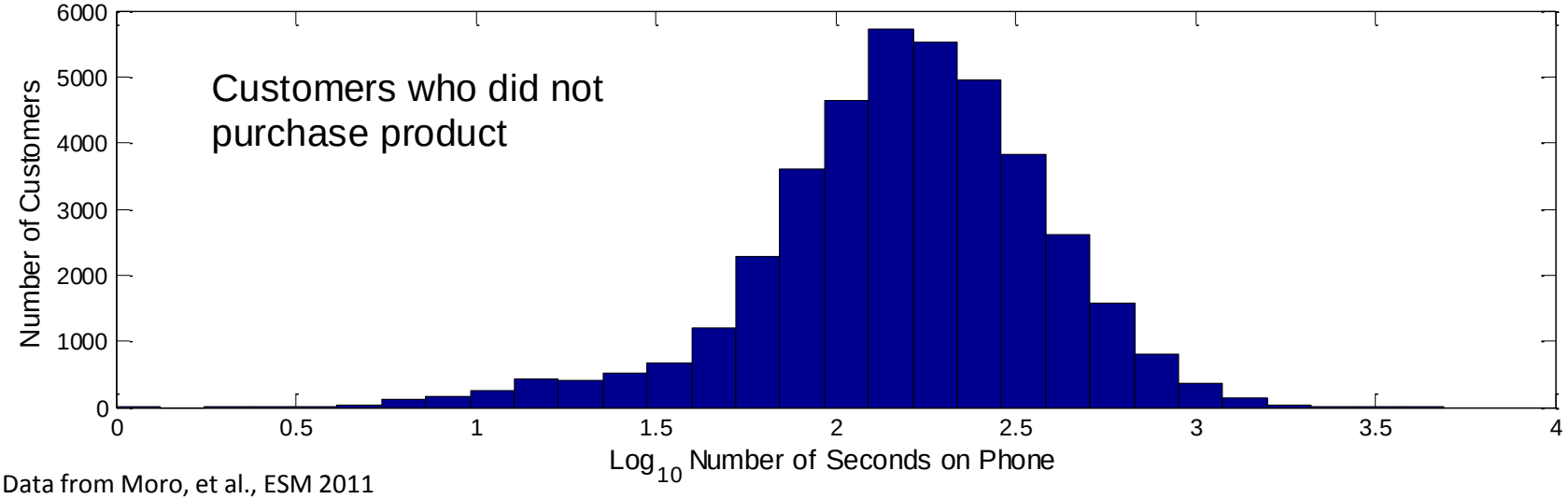
Two Types of Data

- Experimental Data
  - Hypothesis H
  - design an experiment to test H
  - collect data, infer how likely it is that H is true
  - e.g., clinical trials in medicine
  - analyzing such data is primarily the domain of statistics

- Observational or Retrospective or Secondary Data
  - massive non-experimental data sets
    - e.g., logs of Web searches, credit card transactions, Twitter posts, etc
  - Often recorded for other purposes, cheap to obtain
  - assumptions of experimental design no longer valid
  - often little or no "strong theory" for such data
  - this type of data is where data mining is most often used….

A Predictive Variable for Banking Data





Data from Moro, et al., ESM 2011

What are the main Data Mining Techniques?

- Descriptive Methods
  - Exploratory Data Analysis, Visualization
  - Dimension reduction (principal components, factor models, topic models)
  - Clustering
  - Pattern and Anomaly Detection
  - ….and more

- Predictive Modeling
  - Classification
  - Ranking
  - Regression
  - Matrix completion (recommender systems)
  - …and more

Different Terminology

- Statistics
- Machine Learning
- Data Mining
- Predictive Analytics
- “Big Data”
- And more…..

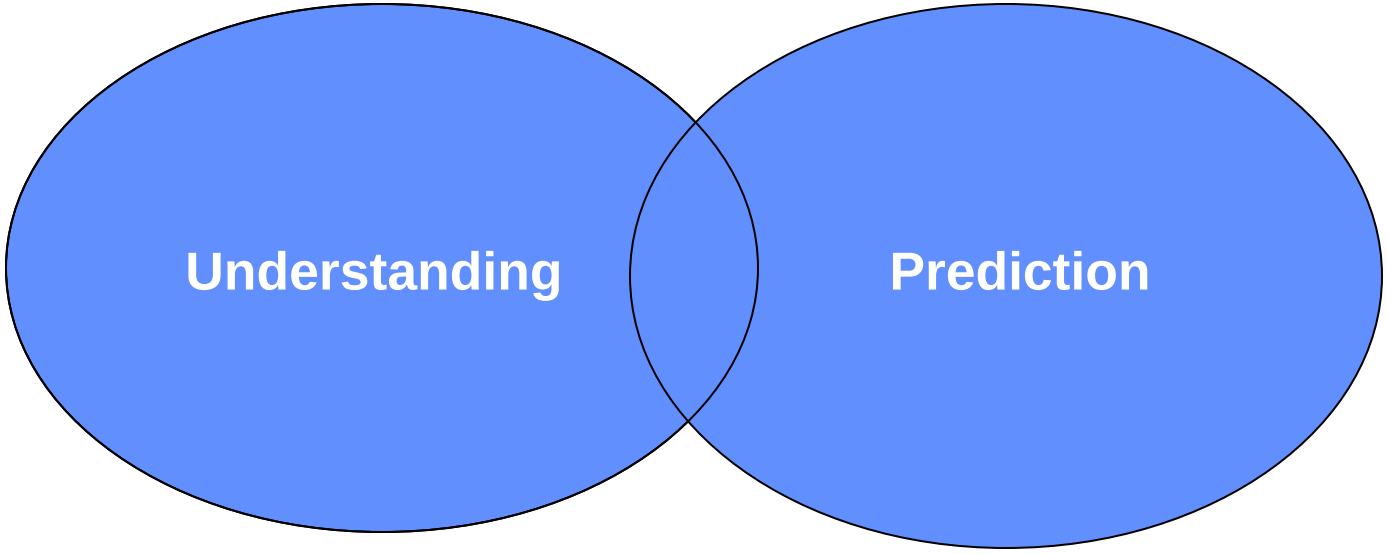
Differences between Data Mining and Statistics?

- Traditional statistics
  - first hypothesize, then collect data, then analyze
  - often model-oriented (strong parametric models)

- Data mining:
  - few if any a priori hypotheses
  - data is usually already collected a priori
  - analysis is typically data-driven not hypothesis-driven
  - Often algorithm-oriented rather than model-oriented

- Different?
  - Yes, in terms of culture, motivation: however…..
  - statistical ideas are very useful in data mining, e.g., in validating whether discovered knowledge is useful
  - Increasing overlap at the boundary of statistics and DM
    e.g., exploratory data analysis (work of John Tukey in the 1960's)

Two Complementary Goals in Data Mining



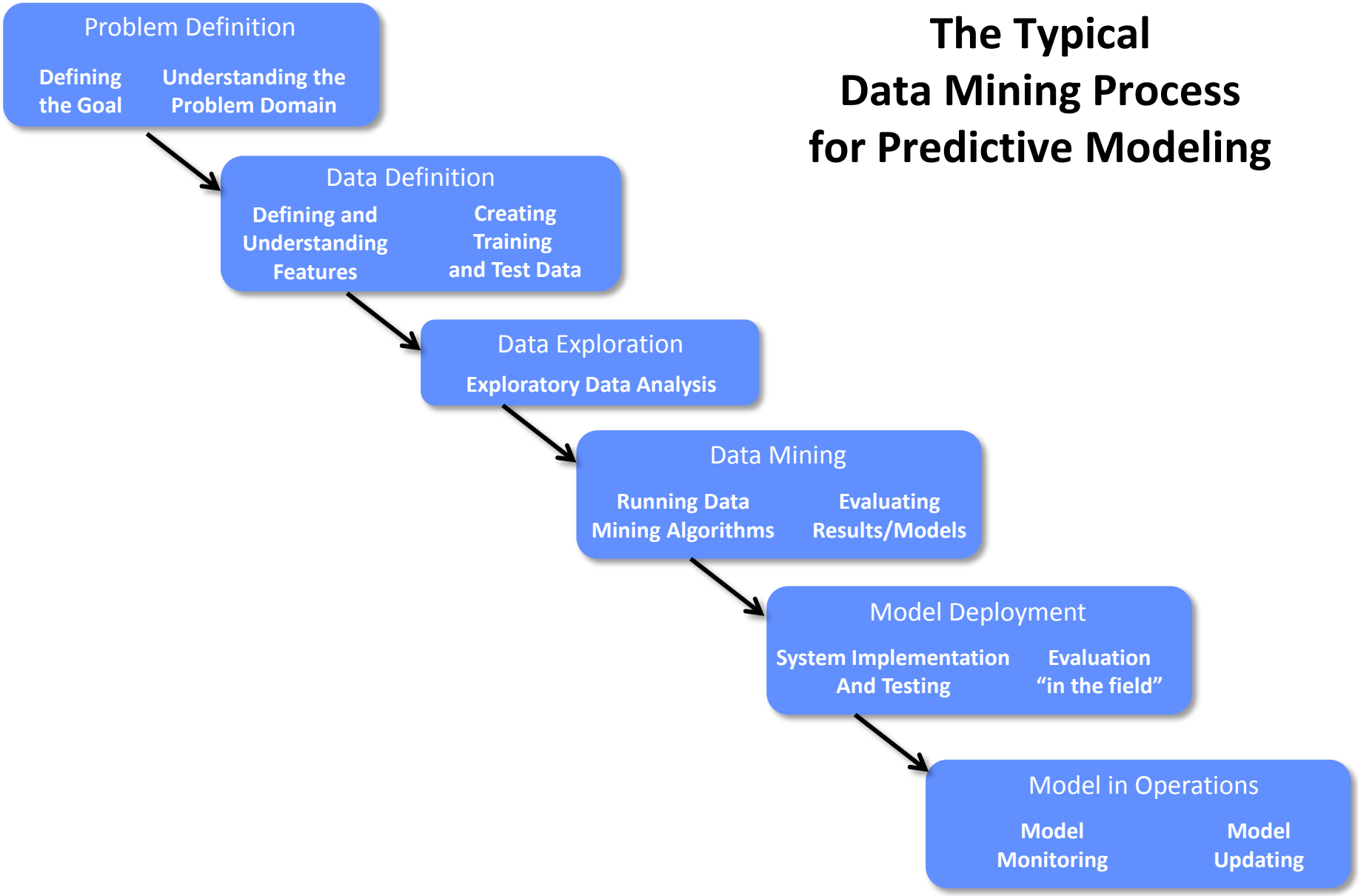
Differences between Data Mining and Machine Learning?

- Very similar in most respects……
  - Data mining relies heavily on ideas from machine learning (and from statistics)

- Some differences between DM and ML:
  - ML tends to focus more on mathematical/statistical "core" algorithms
  - DM is somewhat more applications-oriented
  - But distinctions are small…
  - Can effectively treat them as being much the same….

- Conferences
  - Data Mining: ACM SIGKDD, ICDM, SIAM DM
  - Machine Learning: NIPS, ICML
  - Text/Web oriented: WWW, WSDM, SIGIR, EMNLP, ACL, etc
  - Database oriented: SIGMOD, VLDB, ICDE
  - Similar problems and goals, different emphases

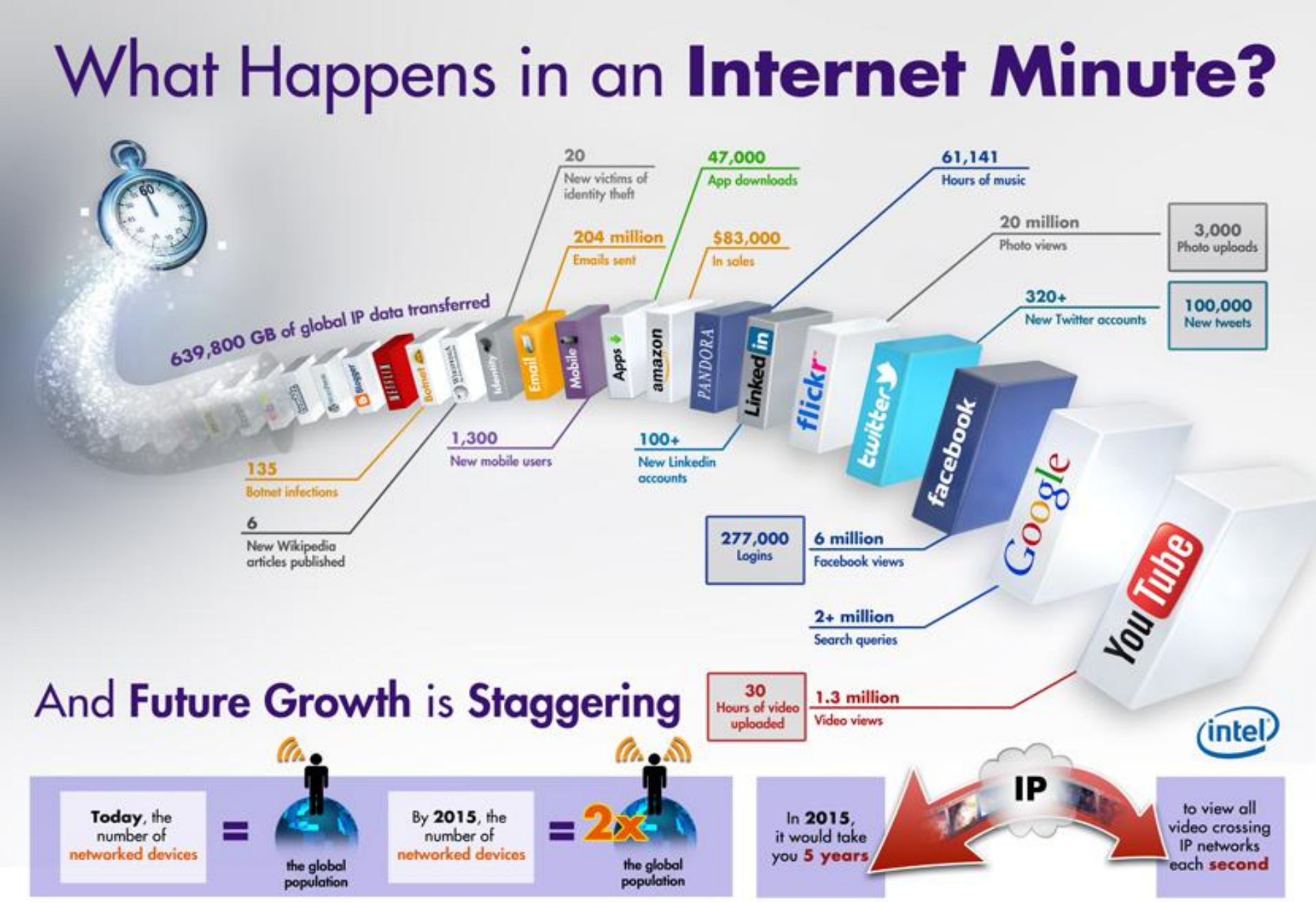
The Typical Data Mining Process for Predictive Modeling



EXAMPLES OF BIG DATA

Examples of "Big Data"

- The Web
  - Over 10^{12} (trillion) Web pages, 3 billion Google searches per day
- Text
  - PubMed/Medline: all published biomedical literature, 20 million articles
  - 4.5 million articles in the English Wikipedia
- Particle Physics
  - Large Hadron Collider at CERN: 600 million particle collisions/second, Gbytes/second, 31 million Gbytes (petabytes) per year
- Retail Transaction Data
  - eBay, Amazon, Walmart, Visa, etc : >100 million transactions per day

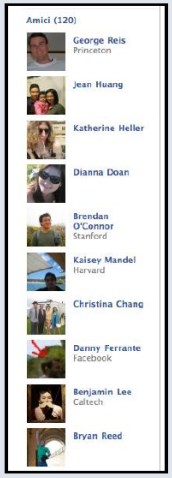


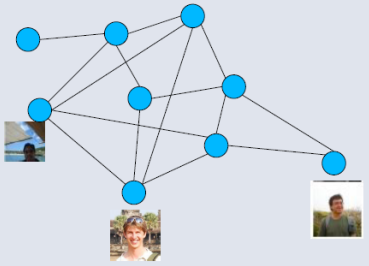
Graphic from http://www.intel.com/content/www/us/en/communications/internet-minute-infographic.html, downloaded in 2011

Graphics from Lars Backstrom, ESWC 2011

500 million 30-day active users

The Friendship graph



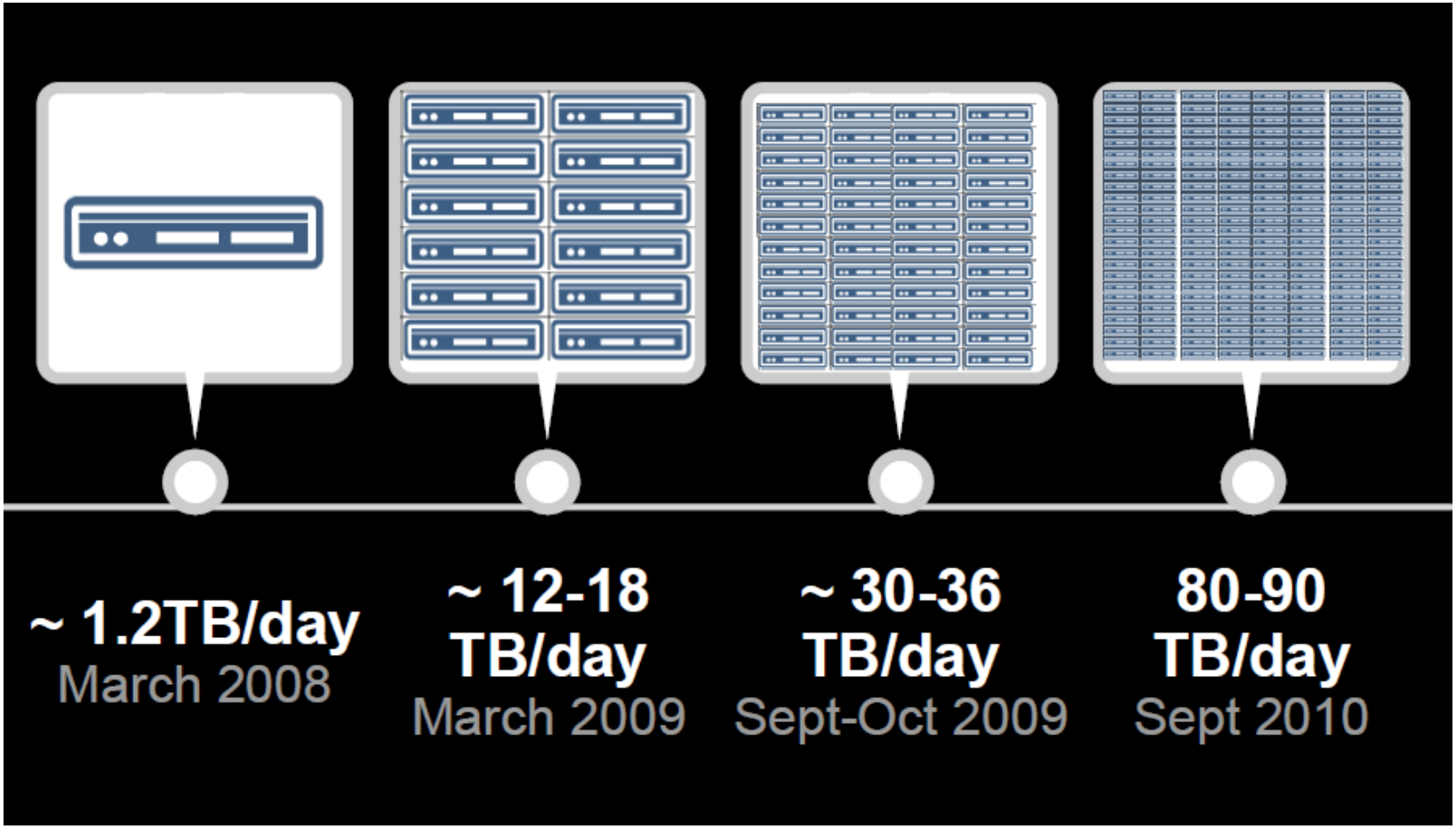


500M users each connect to an average of 130 other users = ~ 60 Billion Edges

Over 30 billion pieces of content shared every month

Over **3 billion** photos uploaded each month

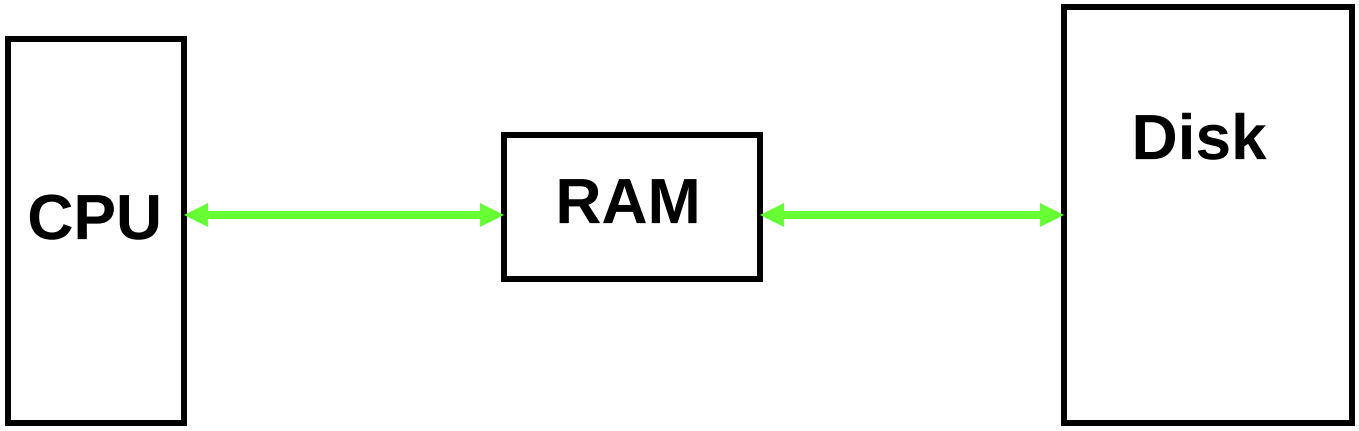
Facebook Data Storage



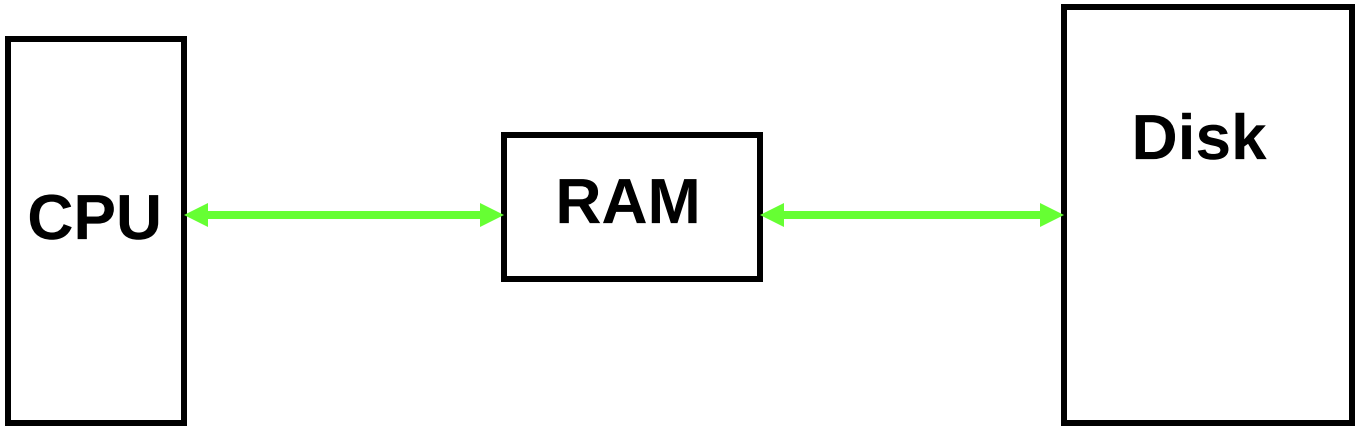
facebook

Graphics from Lars Backstrom, ESWC 2011

Computer Architecture 101



How Far Away are the Data?

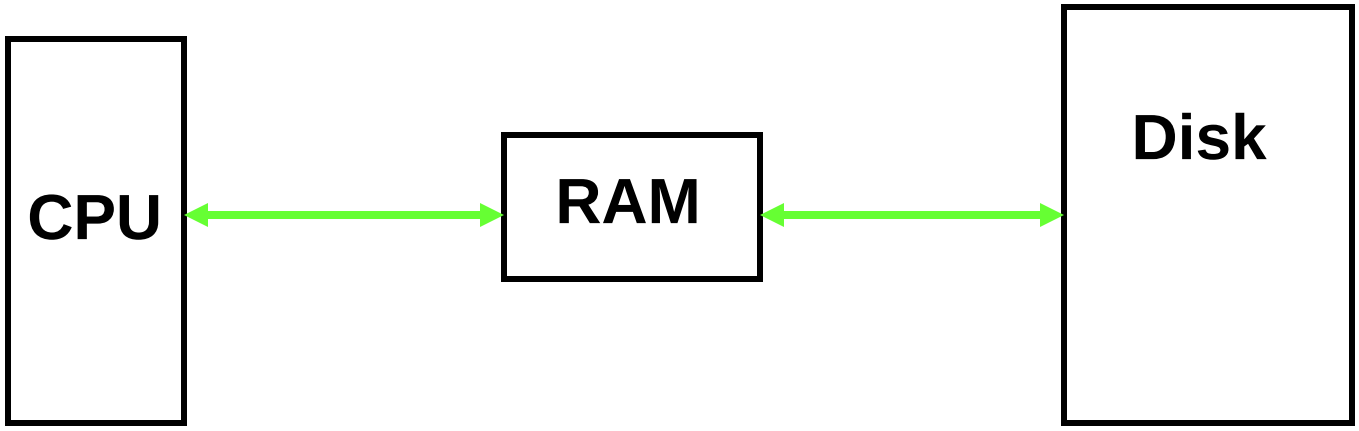


10^{-8} seconds

10^{-3} seconds

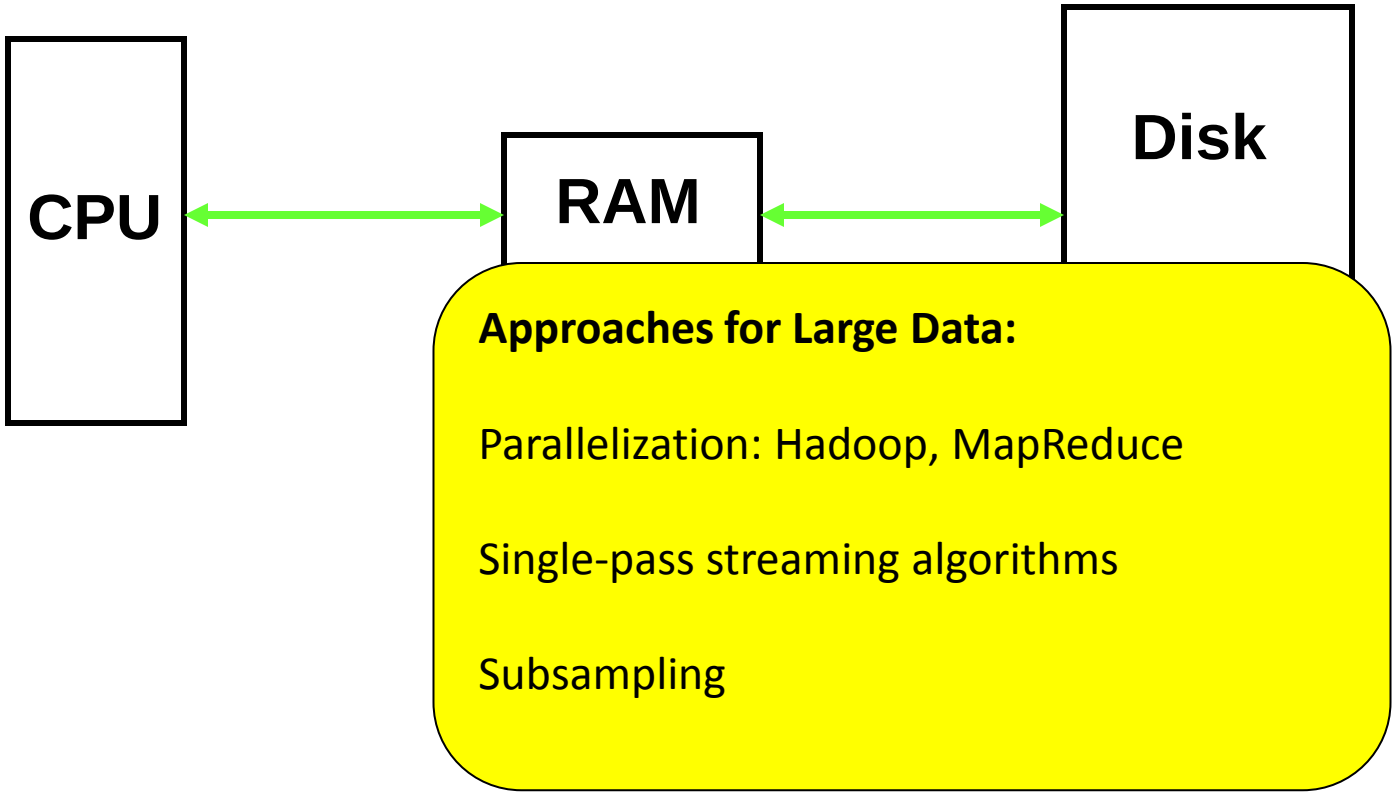
Random Access Times

How Far Away are the Data?

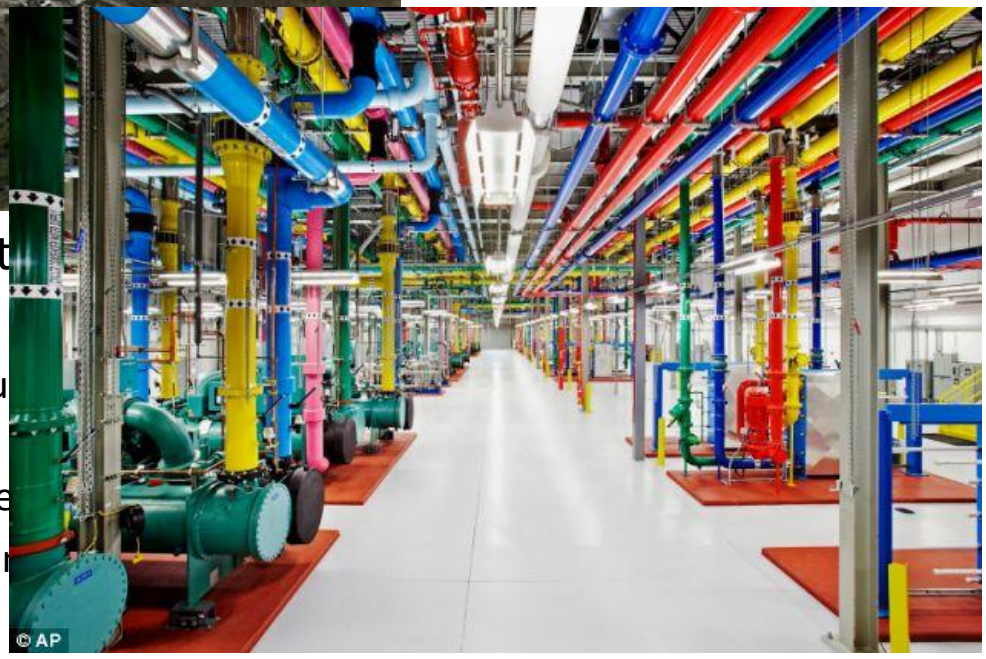


Effective Distances

How Far Away are the Data?



Engineering at Web Scale

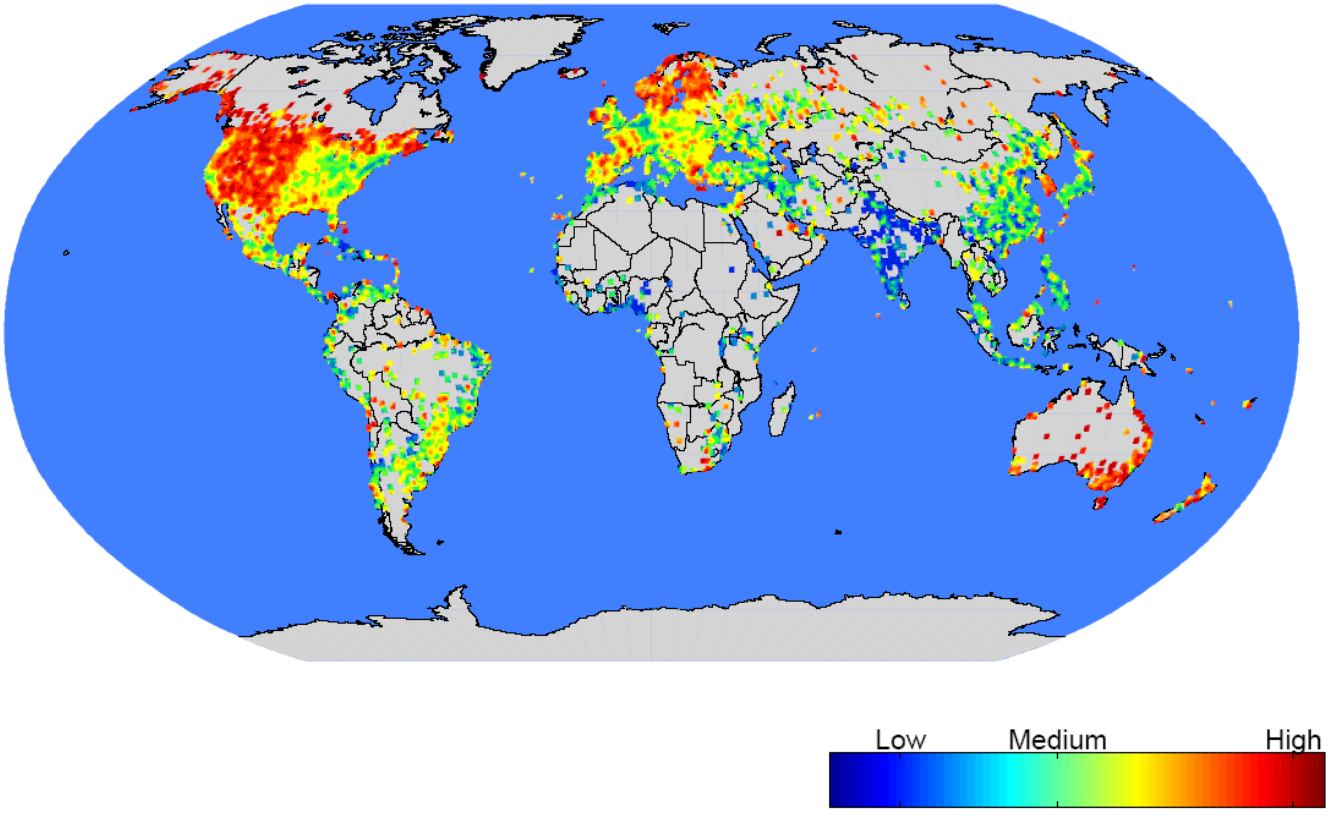


- How can a system have low lat
  pieces of content?
  - Massively parallel systems – requ
    and algorithms
  - Each piece of software in pipeline
  - Highly specialized, non-trivial to i

Instant Messenger Data

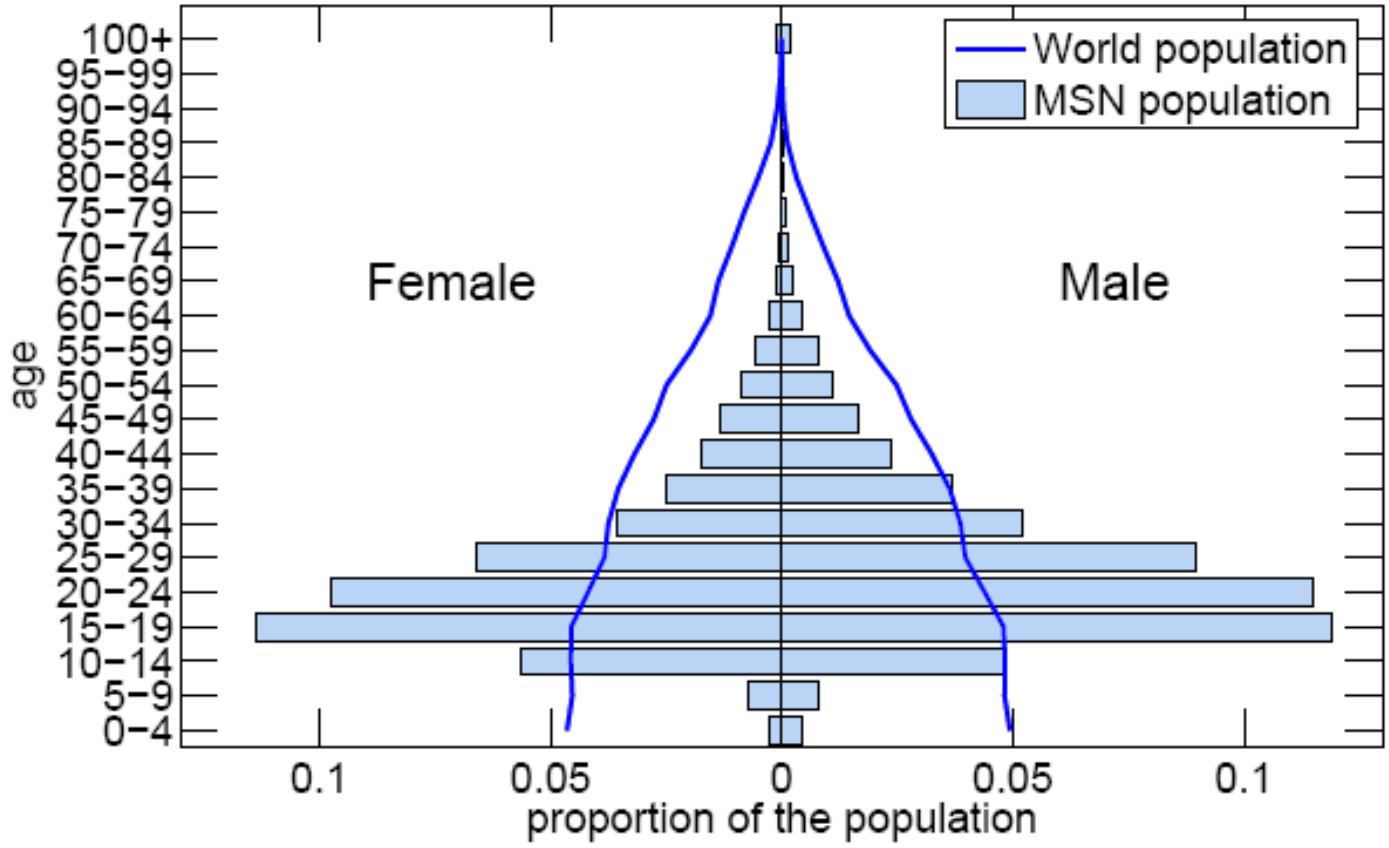
Jure Leskovec and Eric Horvitz, 2007

240 users over 1 month
1.3 billion edges in the graph



Linking Demograpics with IM Usage

Leskovec and Horvitz, 2007



The Internet Archive

Non-profit organization:
crawls and archives the Web for historical preservation

Currently:

- 374 billion Web pages archived since 1996
- over 50 billion unique documents
- Over 5 million digitized books
- 5.8 petabytes of data (5.8 x 10^{15} bytes)

Source: www.archive.org

Benefits of Combining Data, e.g, Property Data

99% of US Properties

+

2.6 Million Active Property Listings

+

Tax history on 125 M parcels

+

90 Million

+

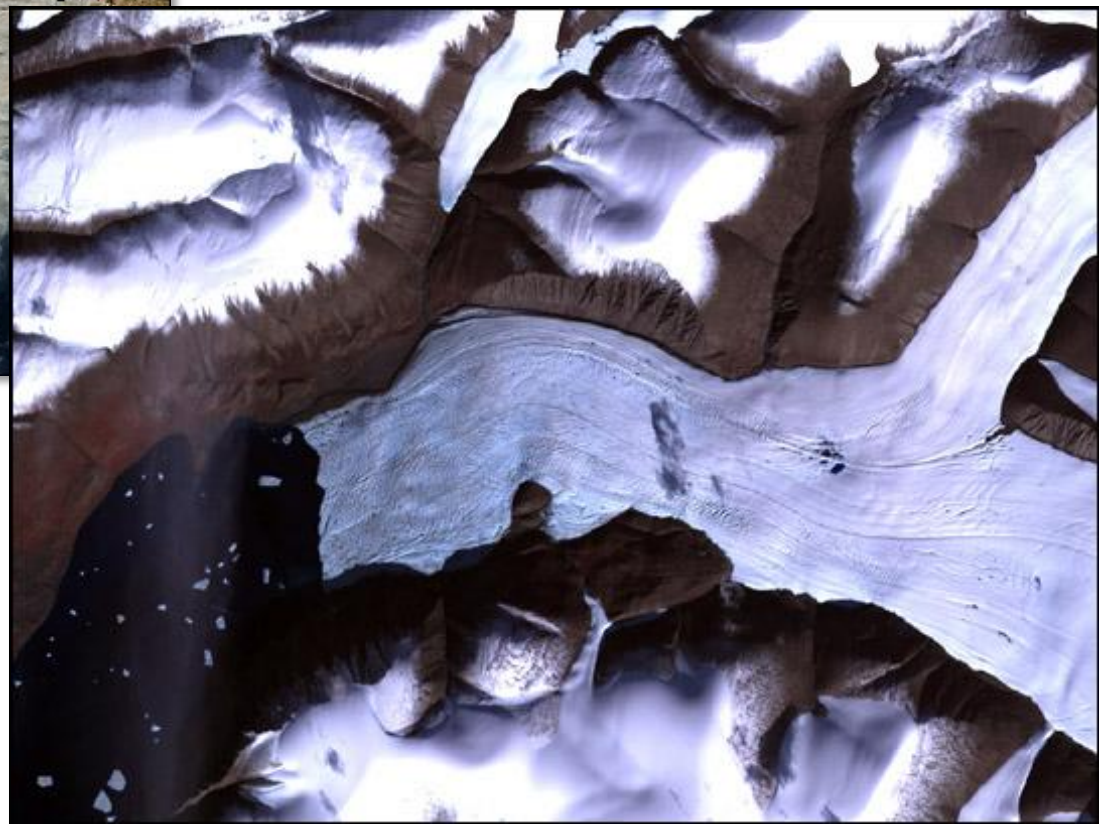
50 million active loans

+

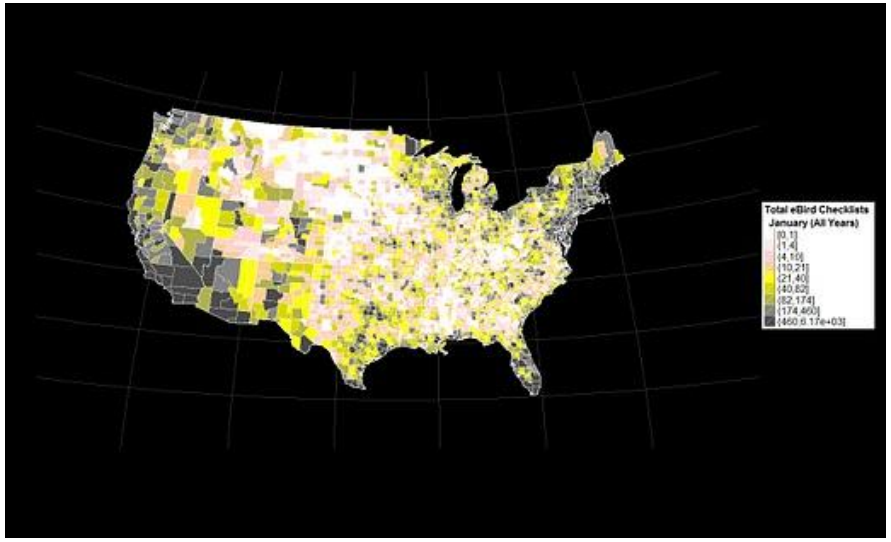
27 Million

NASA's MODIS satellite

entire planet
250m resolution
37 spectral bands
every 2 days



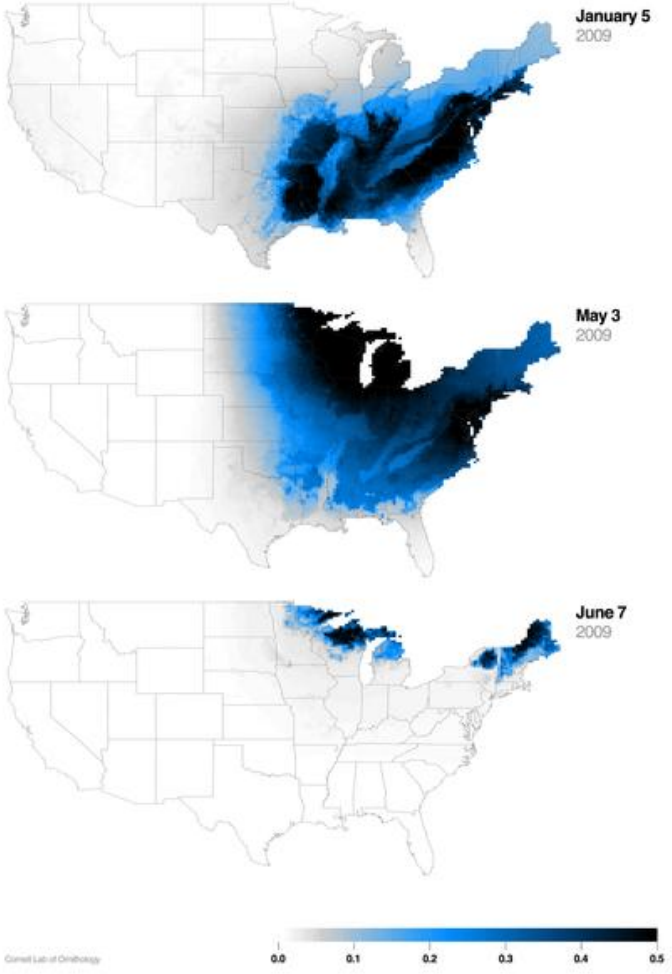
Ebird.org



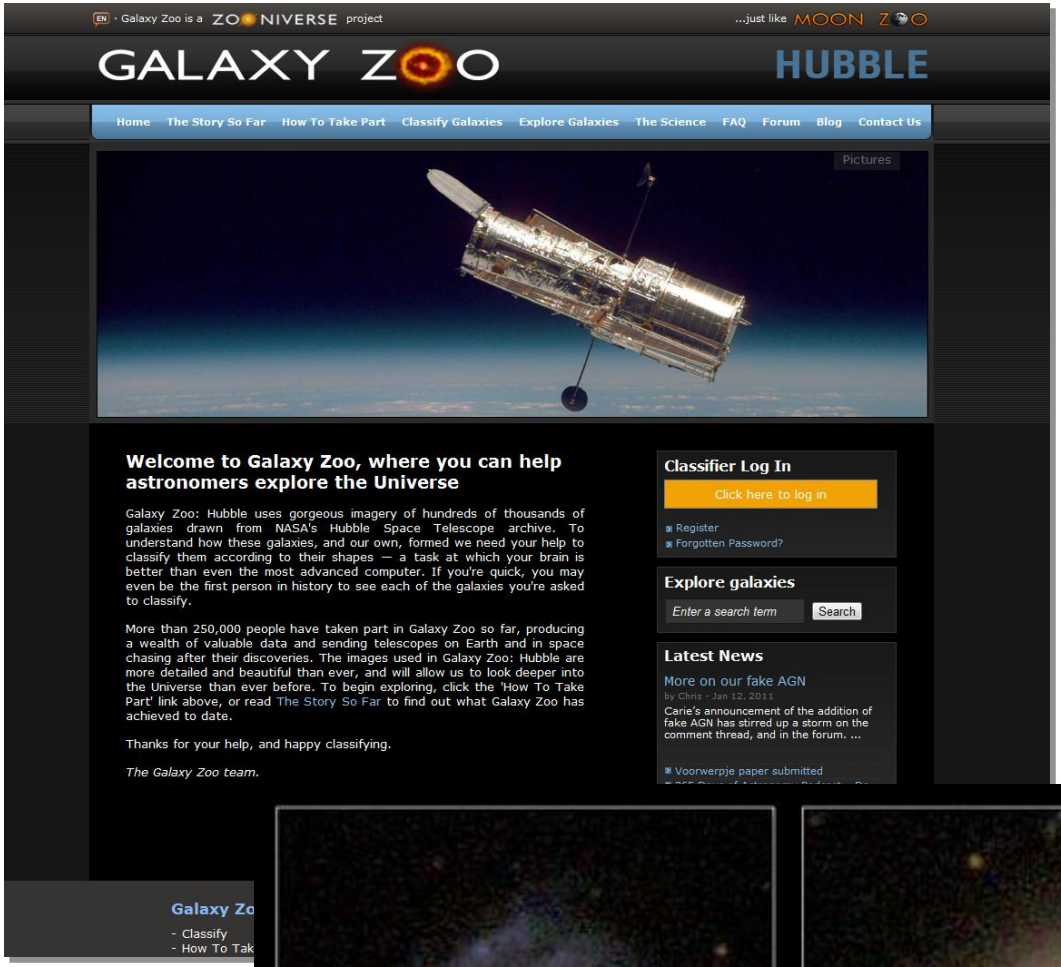
Over 1.5 million submissions per month

From Wood et al, PLOS Biology, 2011

White-throated sparrow distribution



From Raddick et al, Astronomy Education Review, 2009



200,000 citizen scientists
100 million galaxies classified

Quantified Self
self knowledge through numbers

MEETUPS • FOLLOW: • ABOUT Search

Numbers From Around the Web: Round 5

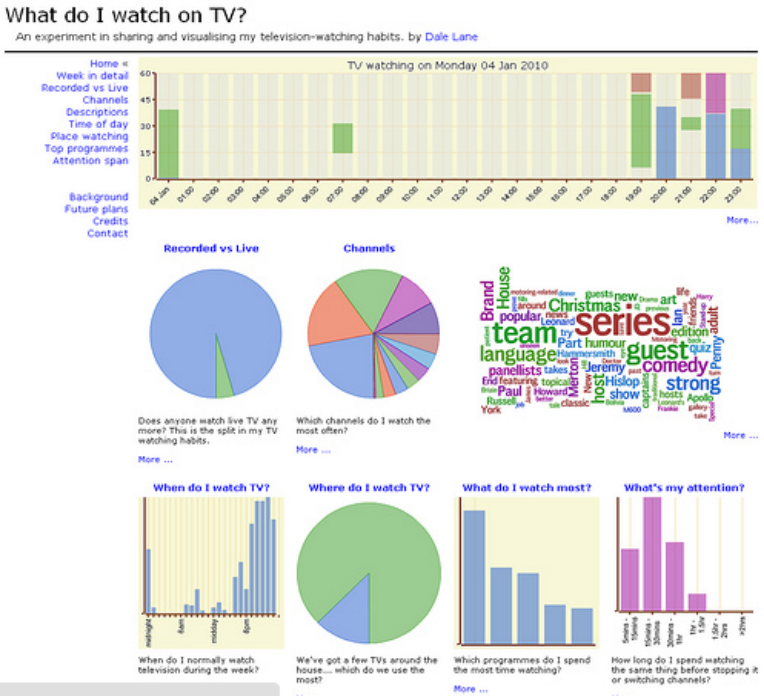
Posted on April 19, 2012 by Ernesto Ramirez

Today's NFATW post comes from Martin Sona, a QS friend and organizer for the QS Aachen/Maastricht meetup group, who pointed out this fascinating project on the QS Facebook group.

Dale Lane is a software developer for IBM living and working in Hampshire and he has been developing neat personal tools for his self tracking for the last few years. Let's take a look at a few of them.

Tracking TV Watching

Inspired by the background data collection offered by last.fm designed to capture music listening habits Dale set out to create his own "scrobbler" to better understand his TV viewing habits. What he came up with is amazing:



From quantifiedself.com

Welcome to Quantified Self

A place for people interested in self-tracking to gather, share knowledge and experiences, and discover resources. Learn more

QS 2012 Conference Register Now!

Get Started Here...

- QS Show&Tell Videos
- Guide to 400+ QS Tools
- QS Forums
- Ernesto's QS 101 Posts
- Raj's Toolmaker Talks

Global QS Event Calendar*

Thursday, April 19
7:00pm Berkeley QS meetup
Friday, April 20
9:00am Aachen/Maastricht (
Tuesday, April 24
12:30pm Bay Area QS Discus
Wednesday, April 25
3:00pm Pittsburgh QS meetu
Thursday, May 3
6:00pm Silicon Valley QS m

Google Calendar

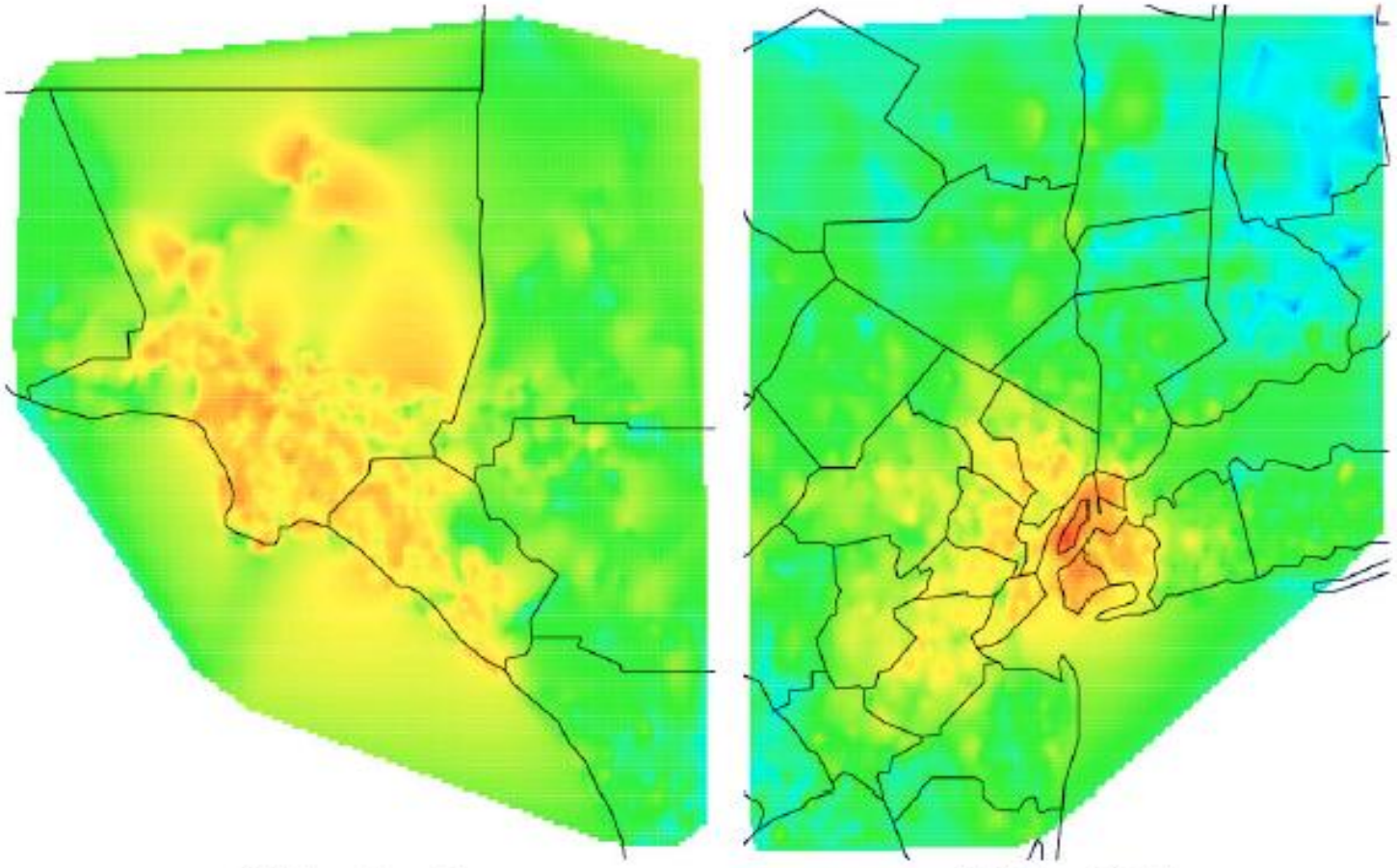
*Note: All times in PST!

QS Meetup Groups

USA - WEST
San Francisco
Silicon Valley
San Diego

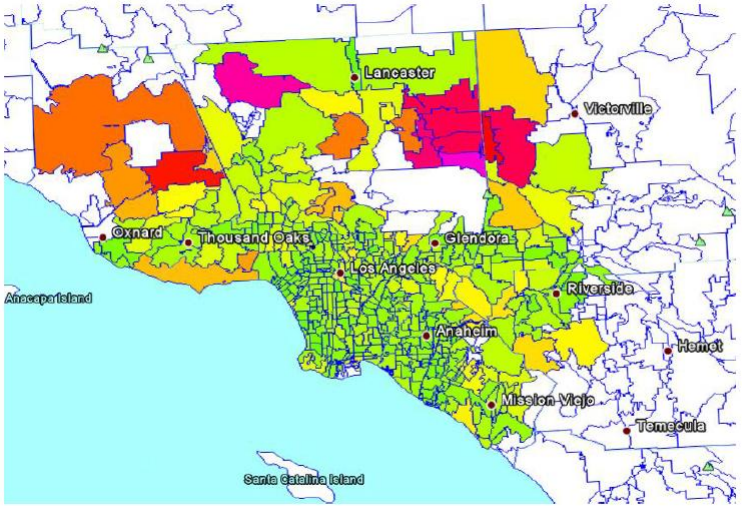
CANADA
Toronto
Vancouver
Montreal

Heatmaps of call densities from AT&T cellphones from 7pm to 8pm on weekdays over 3 months

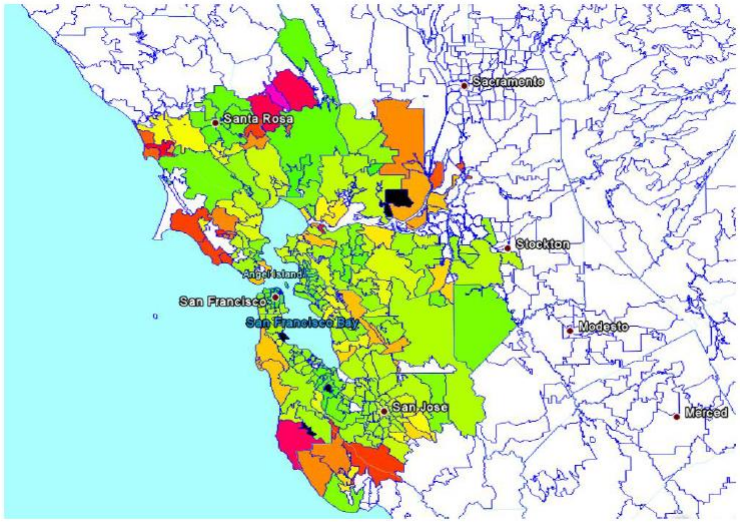


(a) Los Angeles (b) New York

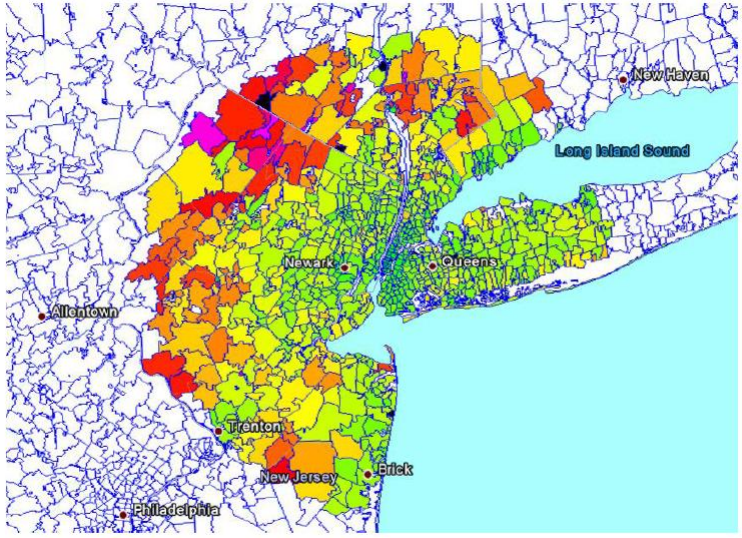
From Isaacman et al, Human mobility modeling at metropolitan scales
MobiSys 2012 Conference



(a) Los Angeles



(b) San Francisco

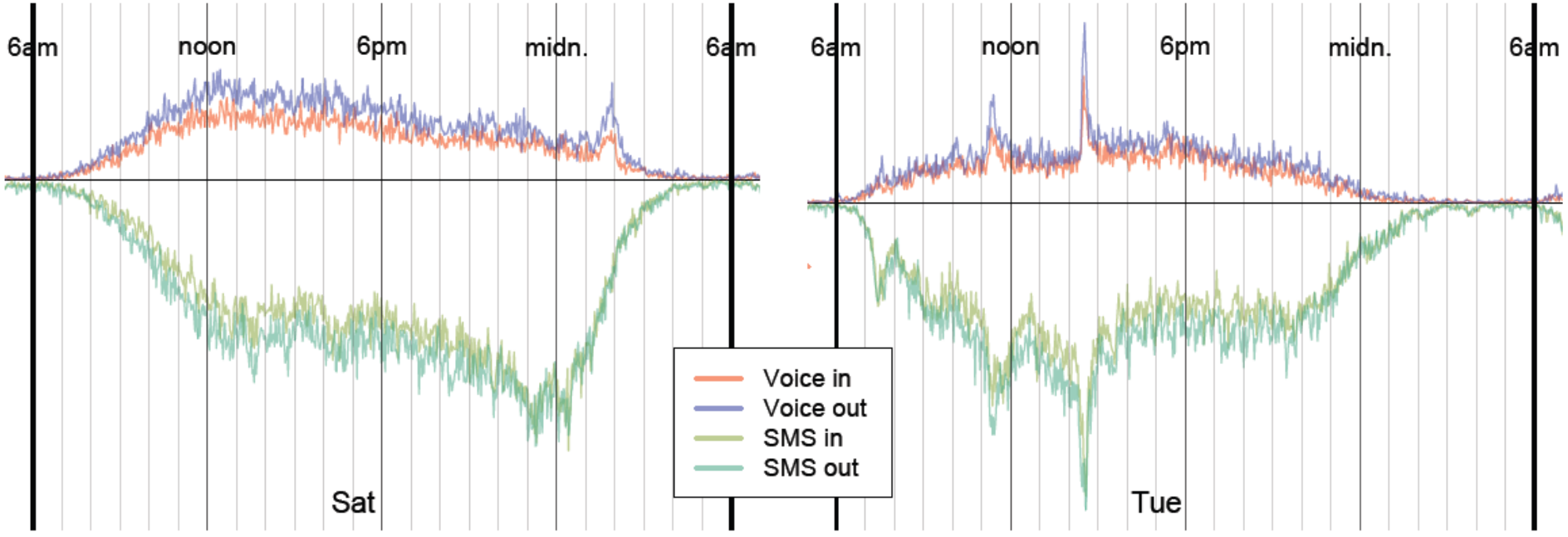


(c) New York

Inferred carbon footprints from commuting patterns using calling locations from cellphones

From Becker et al, Human mobility characterization from cellular network data, *Communications of the ACM,* in press

Voice and SMS Signatures for City Center and School Location

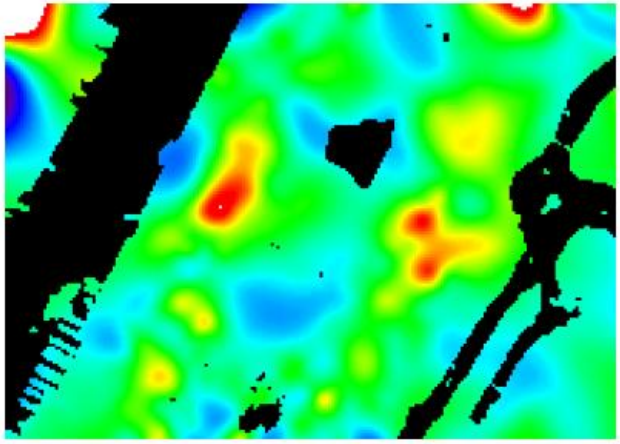


(a) Downtown Antenna

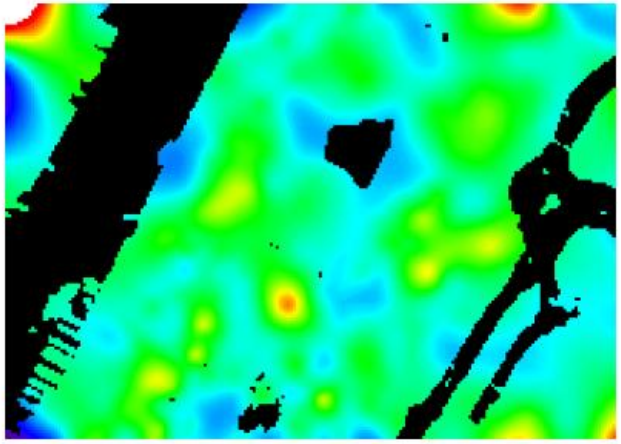
(b) High School Antenna

From Becker et al, A tale of one city: Using cellular network data for urban planning *Pervasive Computing Conference,* 2011

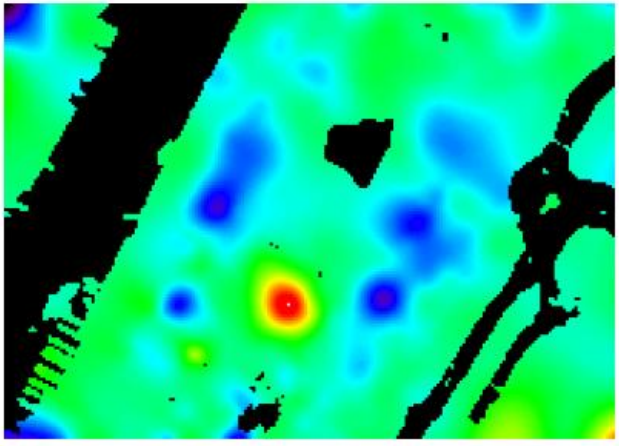
Change in Greenspace Usage from Winter to Summer



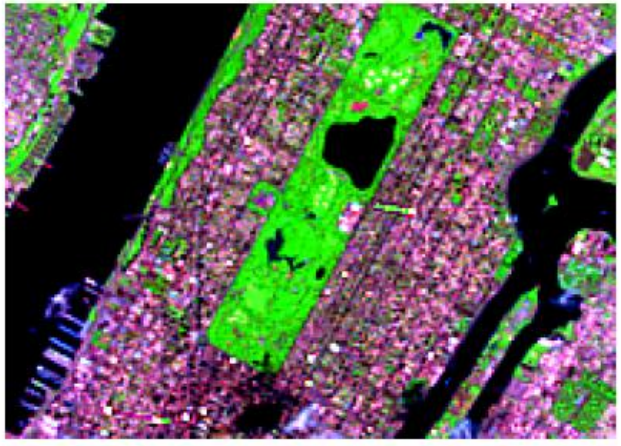
(a) Density: February 12, 2011, 2-3pm



(b) Density: July 9, 2011, 2-3pm



(c) Change in density: July minus February



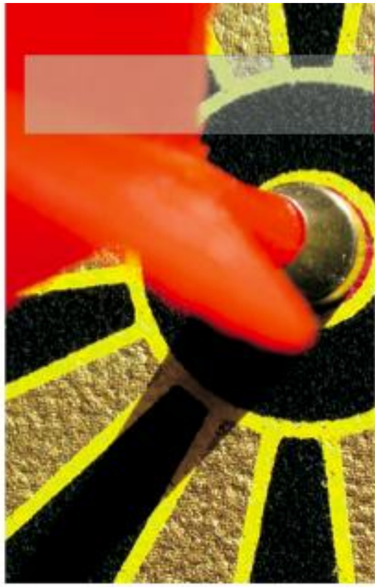
(d) Landsat 5: July 7, 2011, 11am

From Caceres et al, Exploring the use of urban greenspace through cellular network activity
Proceedings of 2nd PURBA Workshop, 2012

Whole new fields are emerging….

- Web and Text Mining
- Computational Advertising
- Computational Science
  - Bioinformatics
  - Cheminformatics
  - Climate Informatics
  - ….and so on
- Computational Social Science, Digital Humanities

….all driven primarily by data rather than theory



EXPERT OPINION

Contact Editor: **Brian Brannon,** bbrannon@computer.org

from IEEE Intelligent Systems, 2009

The Unreasonable Effectiveness of Data

Alon Halevy, Peter Norvig, and Fernando Pereira, *Google*

Eugene Wigner's article "The Unreasonable Effectiveness of Mathematics in the Natural Sciences"[1] examines why so much of physics can be neatly explained with simple mathematical formulas

such as $f = ma$ or $e = mc^2$. Meanwhile, sciences that involve human beings rather than elementary par-

behavior. So, this corpus could serve as the basis of a complete model for certain tasks—if only we knew how to extract the model from the data.

Learning from Text at Web Scale

The biggest successes in natural-language-related machine learning have been statistical speech recognition and statistical machine translation. The

TYPES OF DATA

A Data Set in Matrix Form

Patient ID	Zipcode	Age	….	Test Score	Diagnosis
18261	92697	55		83	1
42356	92697	19		-99	1
00219	90001	35		77	0
83726	24351	0		65	0
…..					
…..					
12837	92697	40		70	1

Terminology:

Columns may be called “measurements”, “variables”, “features”, “attributes”, “fields”, etc

Rows may be individuals, entities, objects, samples, etc

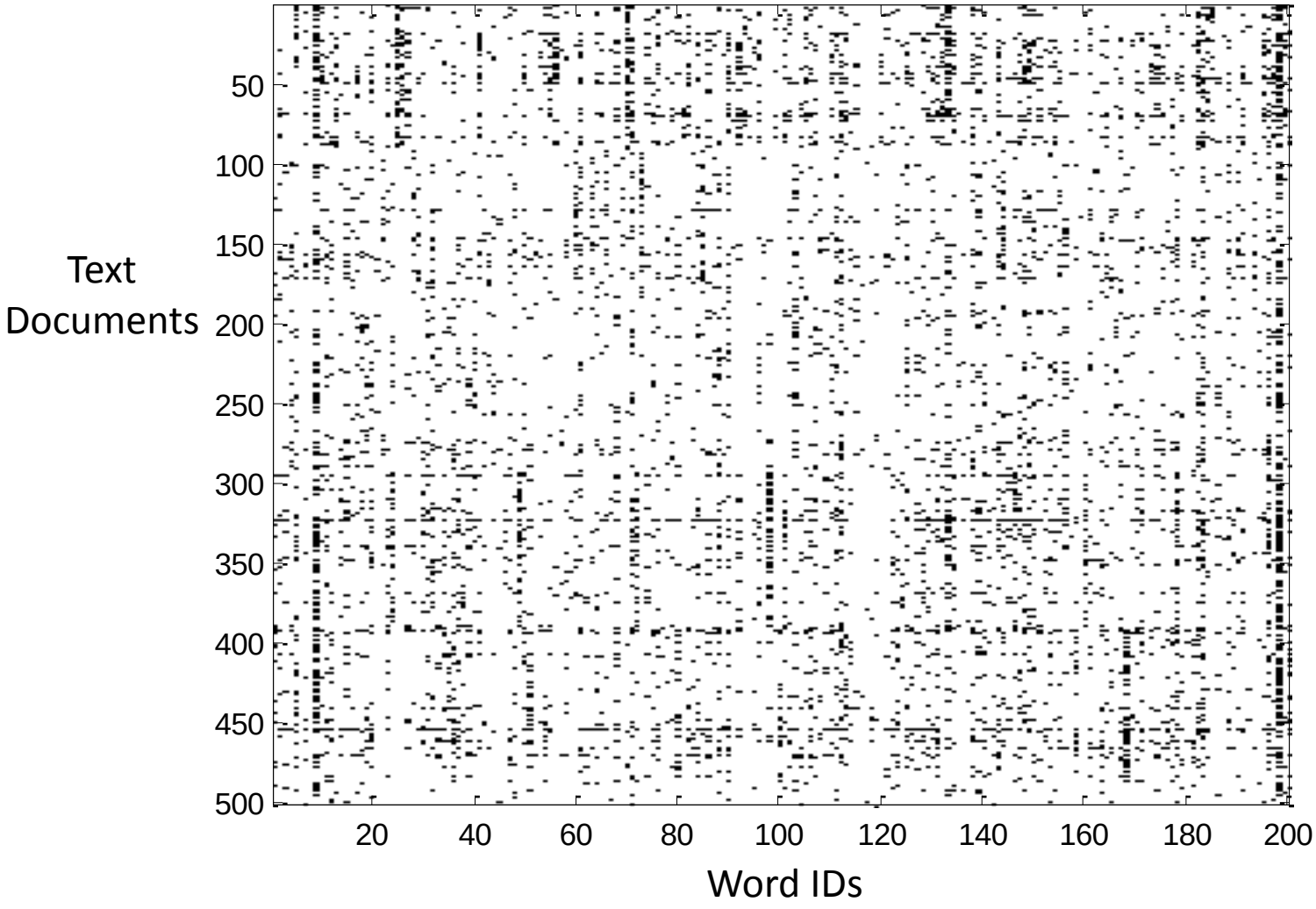
Prediction

Patient ID	Zipcode	Age	….	Test Score	Diagnosis
18261	92697	55		83	1
42356	92697	19		-99	1
00219	90001	35		77	0
83726	24351	0		65	0
…..					
…..					
12837	92697	40		70	1

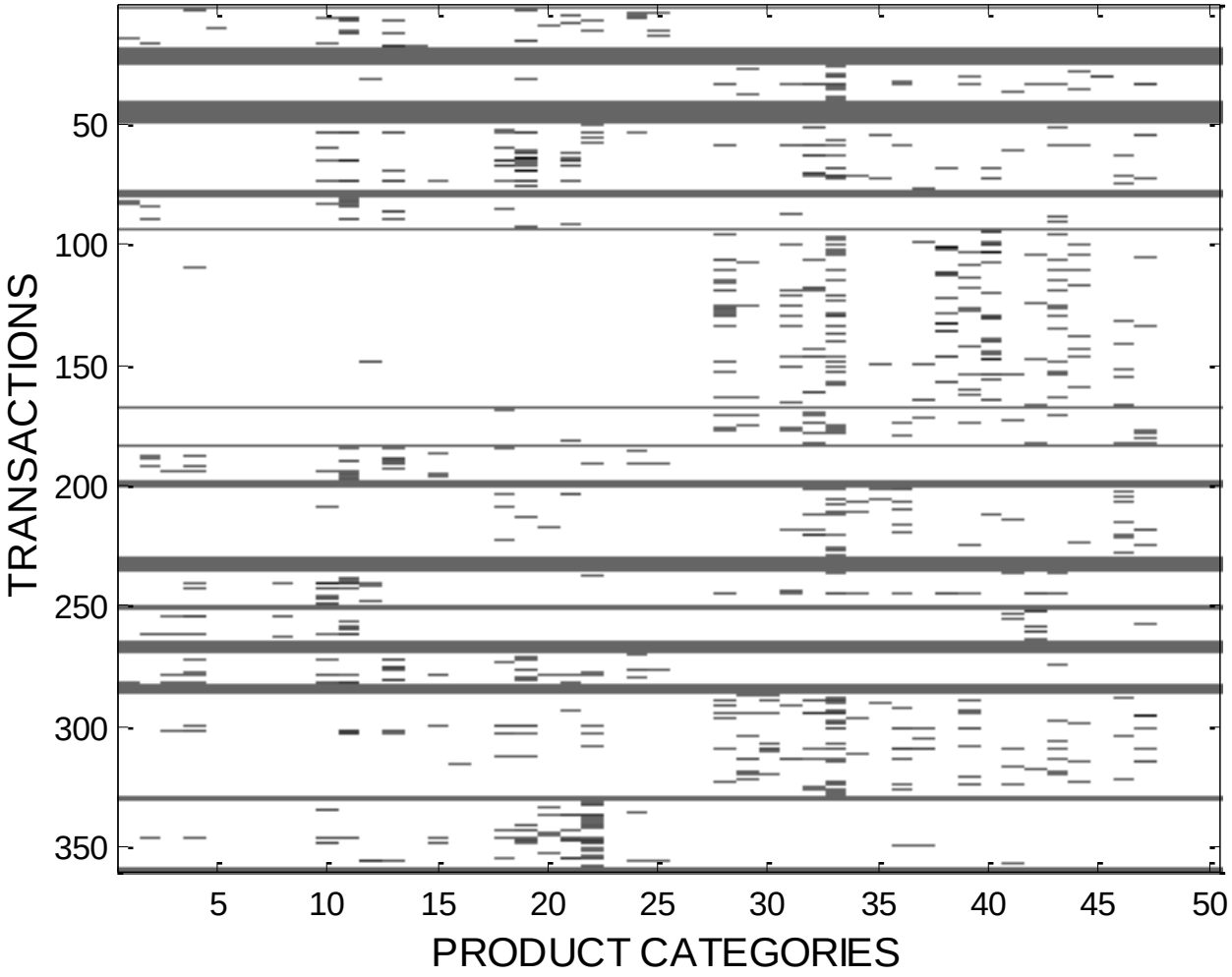
Clustering

Patient ID	Zipcode	Age	….	Test Score	Cluster
18261	92697	55		83	?
42356	92697	19		-99	?
00219	90001	35		77	?
83726	24351	0		65	?
…..					
…..					
12837	92697	40		70	?

Sparse Matrix (Text) Data



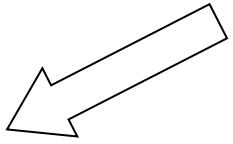
"Market Basket" Data



Sequential Data (Web Clickstreams)

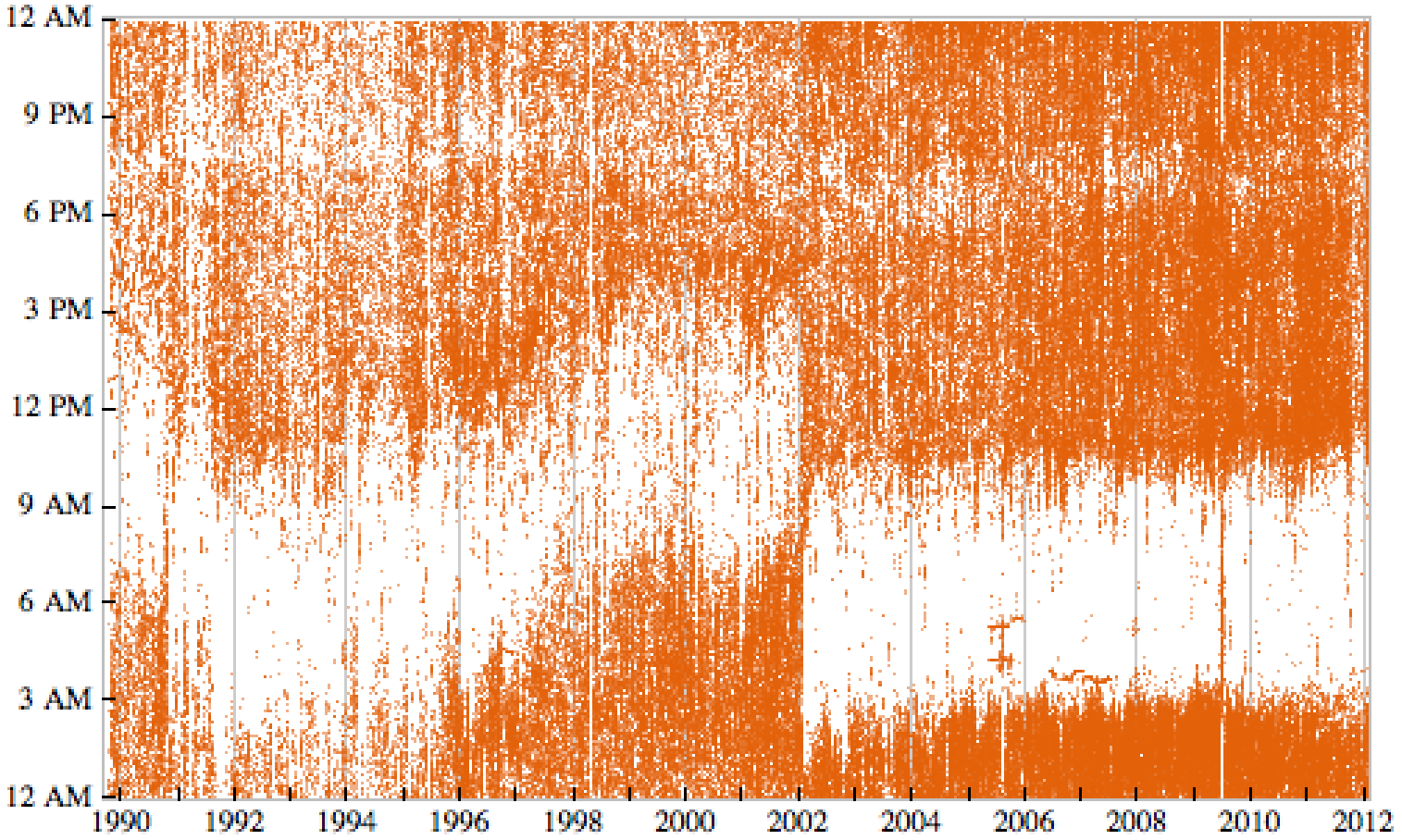
128.195.36.195, -, 3/22/00, 10:35:11, W3SVC, SRVR1, 128.200.39.181, 781, 363, 875, 200, 0, GET, /top.html, -,
128.195.36.195, -, 3/22/00, 10:35:16, W3SVC, SRVR1, 128.200.39.181, 5288, 524, 414, 200, 0, POST, /spt/main.html, -,
128.195.36.195, -, 3/22/00, 10:35:17, W3SVC, SRVR1, 128.200.39.181, 30, 280, 111, 404, 3, GET, /spt/images/bk1.jpg, -,
128.195.36.101, -, 3/22/00, 16:18:50, W3SVC, SRVR1, 128.200.39.181, 60, 425, 72, 304, 0, GET, /top.html, -,
128.195.36.101, -, 3/22/00, 16:18:58, W3SVC, SRVR1, 128.200.39.181, 8322, 527, 414, 200, 0, POST, /spt/main.html, -,
128.195.36.101, -, 3/22/00, 16:18:59, W3SVC, SRVR1, 128.200.39.181, 0, 280, 111, 404, 3, GET, /spt/images/bk1.jpg, -,
128.200.39.17, -, 3/22/00, 20:54:37, W3SVC, SRVR1, 128.200.39.181, 140, 199, 875, 200, 0, GET, /top.html, -,
128.200.39.17, -, 3/22/00, 20:54:55, W3SVC, SRVR1, 128.200.39.181, 17766, 365, 414, 200, 0, POST, /spt/main.html, -,
128.200.39.17, -, 3/22/00, 20:54:55, W3SVC, SRVR1, 128.200.39.181, 0, 258, 111, 404, 3, GET, /spt/images/bk1.jpg, -,
128.200.39.17, -, 3/22/00, 20:55:07, W3SVC, SRVR1, 128.200.39.181, 0, 258, 111, 404, 3, GET, /spt/images/bk1.jpg, -,
128.200.39.17, -, 3/22/00, 20:55:36, W3SVC, SRVR1, 128.200.39.181, 1061, 382, 414, 200, 0, POST, /spt/main.html, -,
128.200.39.17, -, 3/22/00, 20:55:36, W3SVC, SRVR1, 128.200.39.181, 0, 258, 111, 404, 3, GET, /spt/images/bk1.jpg, -,
128.200.39.17, -, 3/22/00, 20:55:39, W3SVC, SRVR1, 128.200.39.181, 0, 258, 111, 404, 3, GET, /spt/images/bk1.jpg, -,
128.200.39.17, -, 3/22/00, 20:56:03, W3SVC, SRVR1, 128.200.39.181, 1081, 382, 414, 200, 0, POST, /spt/main.html, -,
128.200.39.17, -, 3/22/00, 20:56:04, W3SVC, SRVR1, 128.200.39.181, 0, 258, 111, 404, 3, GET, /spt/images/bk1.jpg, -,
128.200.39.17, -, 3/22/00, 20:56:33, W3SVC, SRVR1, 128.200.39.181, 0, 262, 72, 304, 0, GET, /top.html, -,
128.200.39.17, -, 3/22/00, 20:56:52, W3SVC, SRVR1, 128.200.39.181, 19598, 382, 414, 200, 0, POST, /spt/main.html, -,

User 1	2 3 2 2 3 3 3 1 1 1 3 1 3 3 3 3
User 2	3 3 3 1 1 1
User 3	7 7 7 7 7 7 7 7
User 4	1 5 1 1 1 5 1 5 1 1 1 1 1 1
User 5	5 1 1 5
…	…

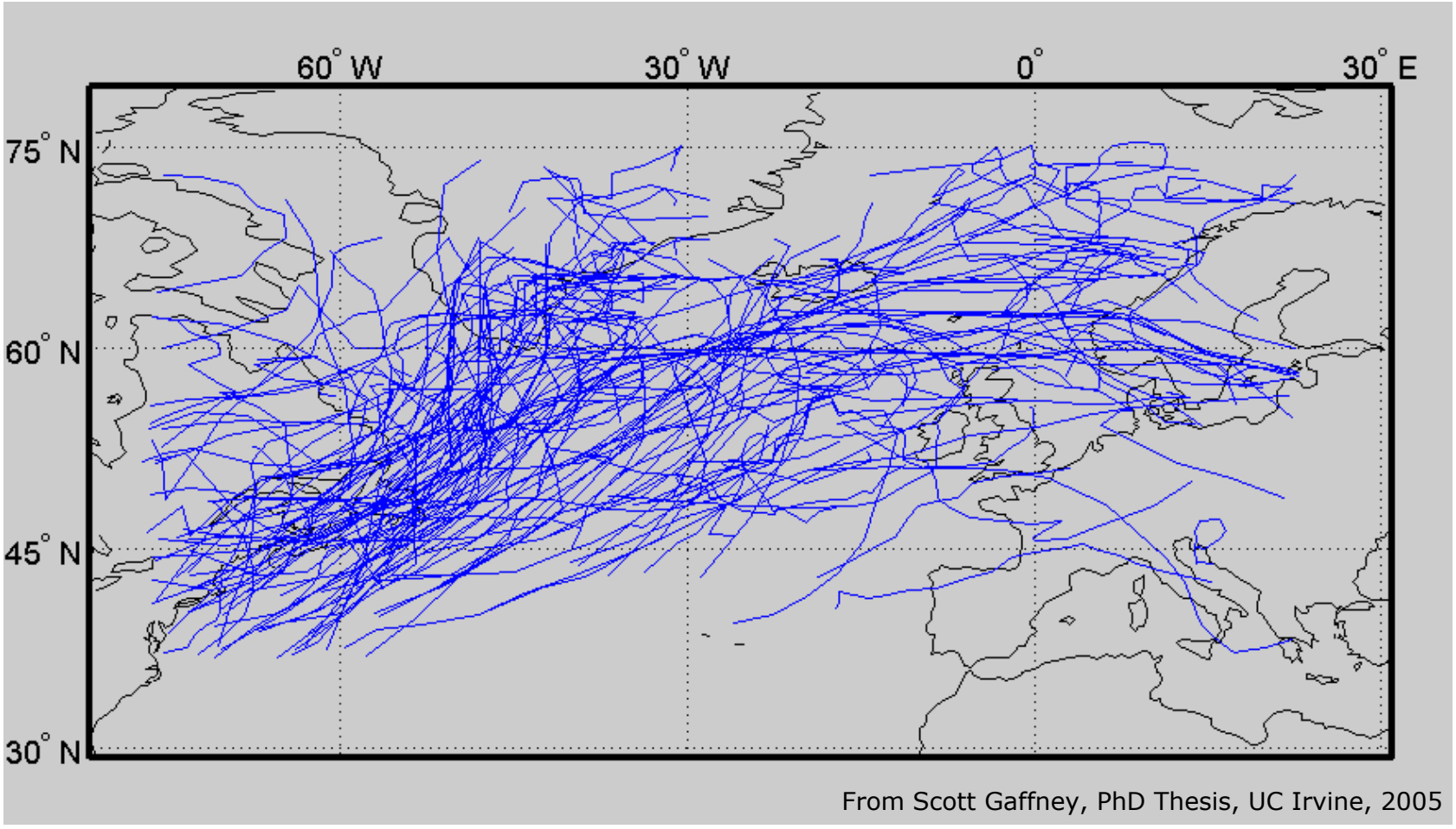


Emails over Time

Record of 300,000 emails sent since 1989
From: blog.stephenwolfram.com
***The Personal Analytics of My Life*, March 8th 2012**

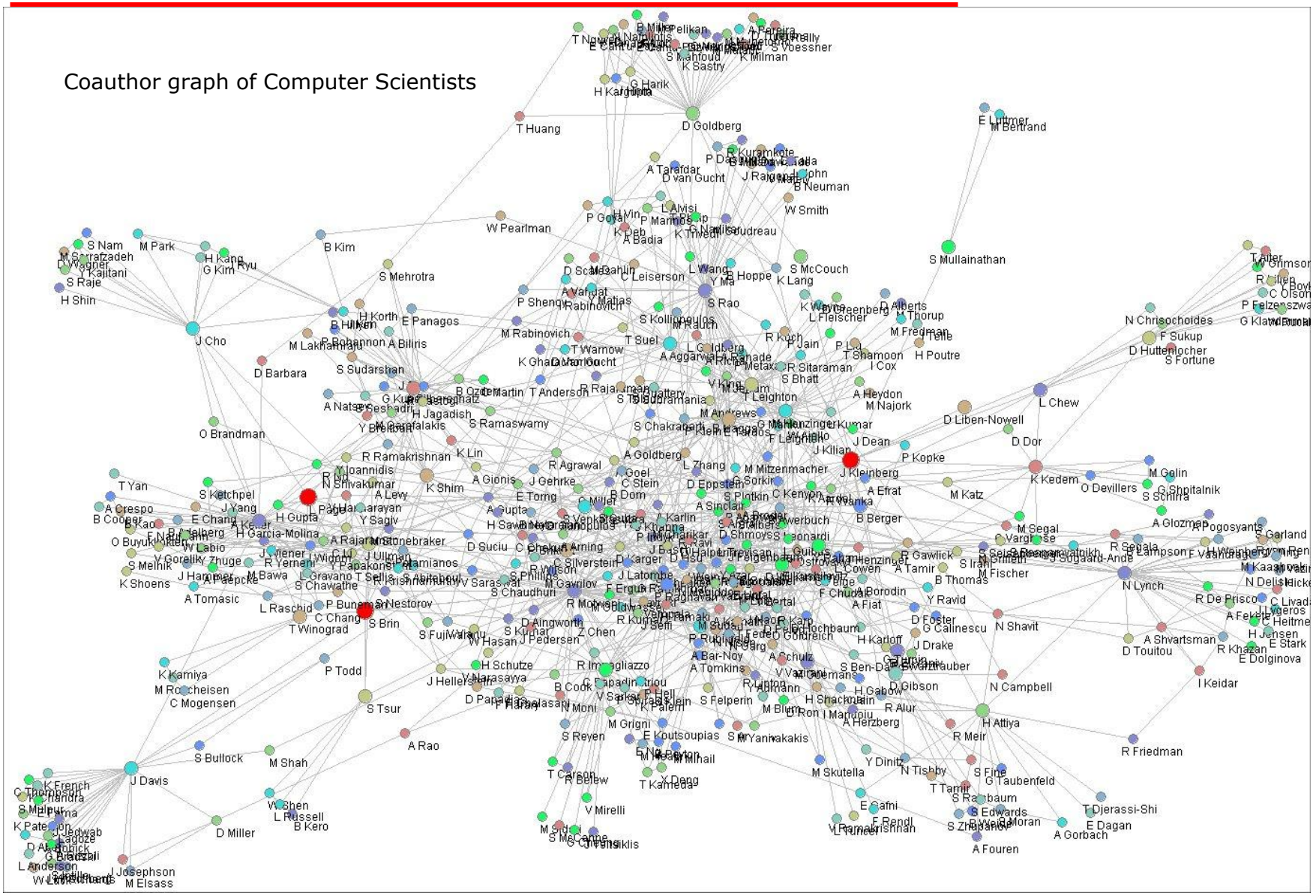


Spatio-temporal data

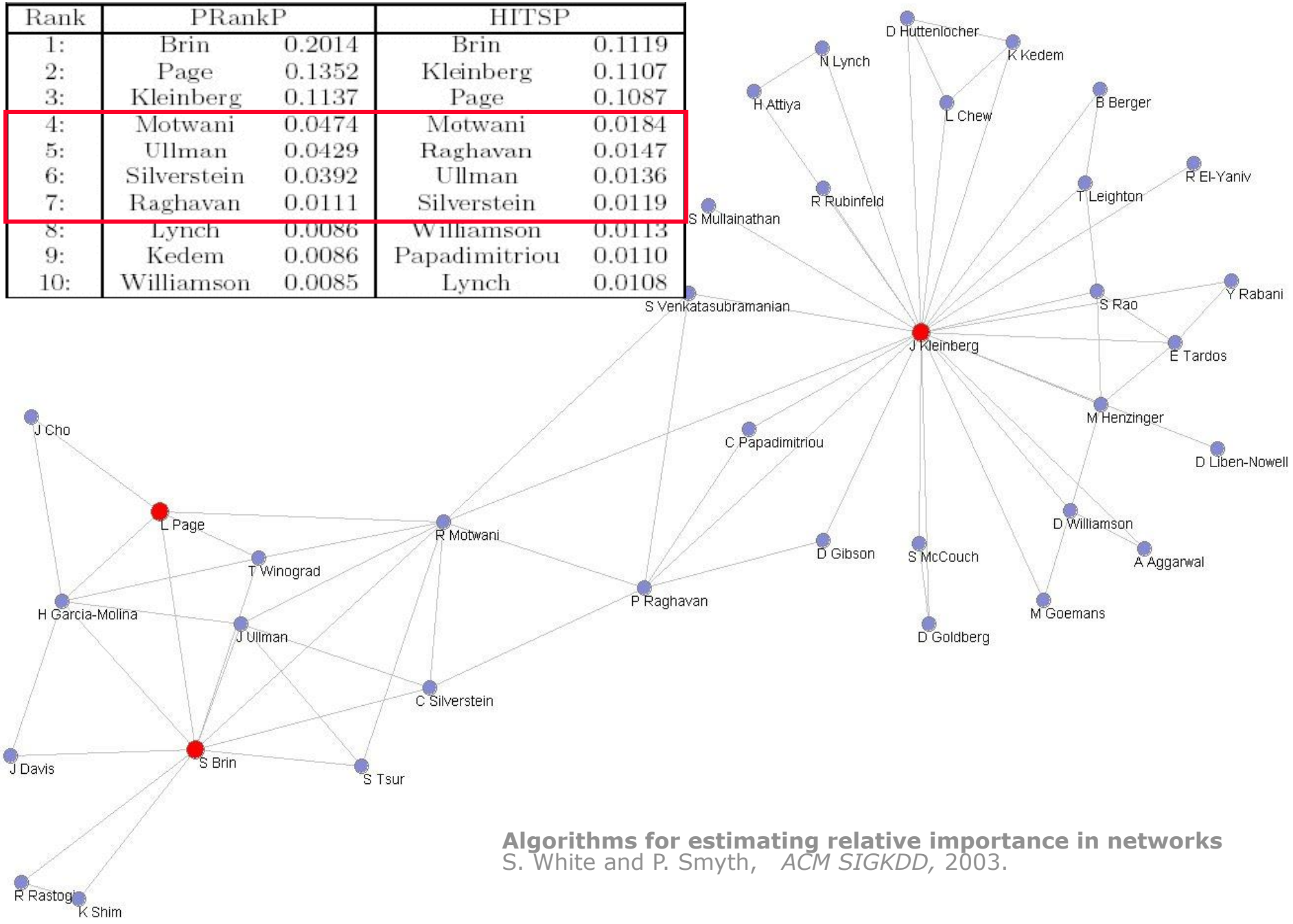


From Scott Gaffney, PhD Thesis, UC Irvine, 2005

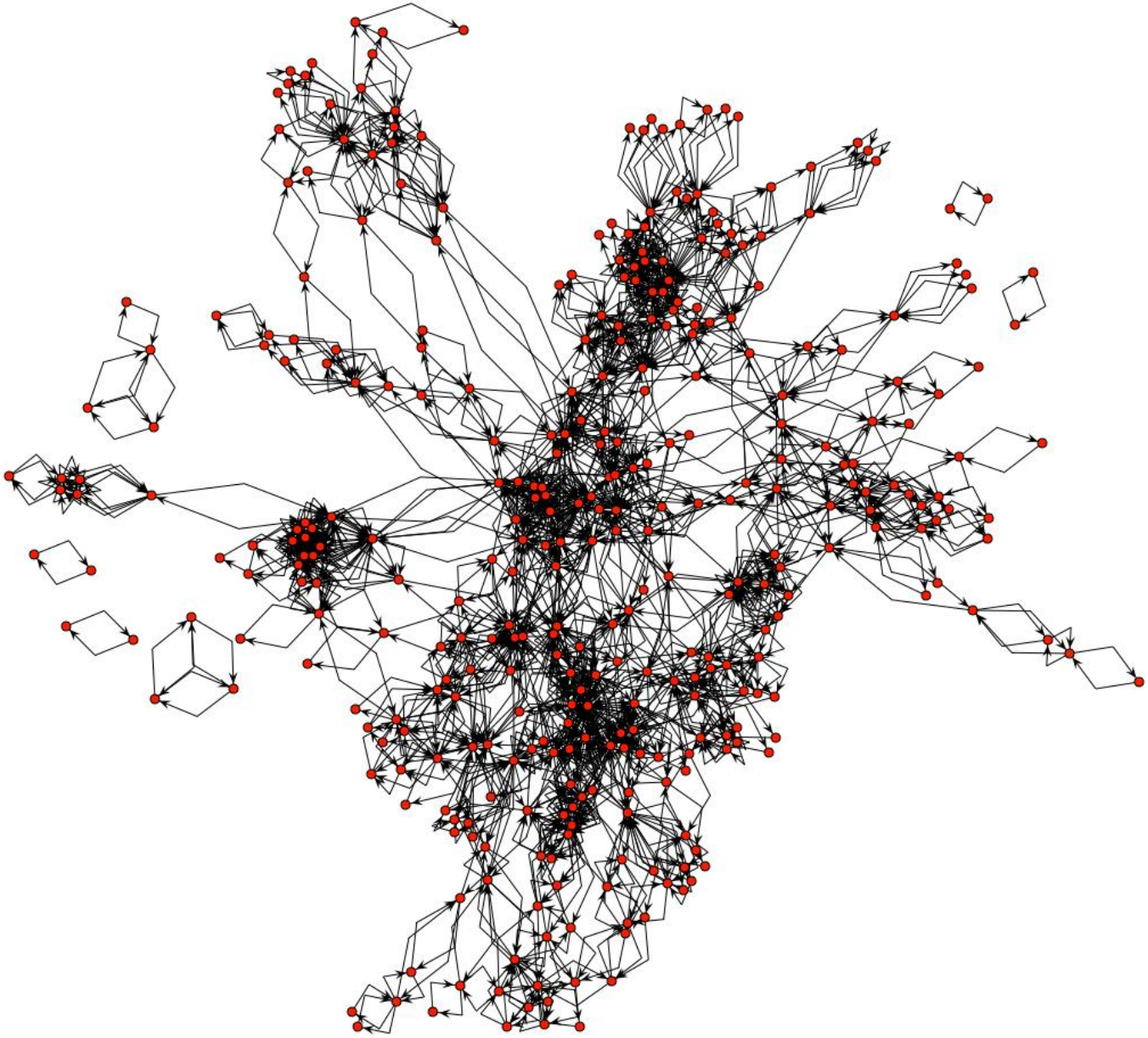
Relational Data

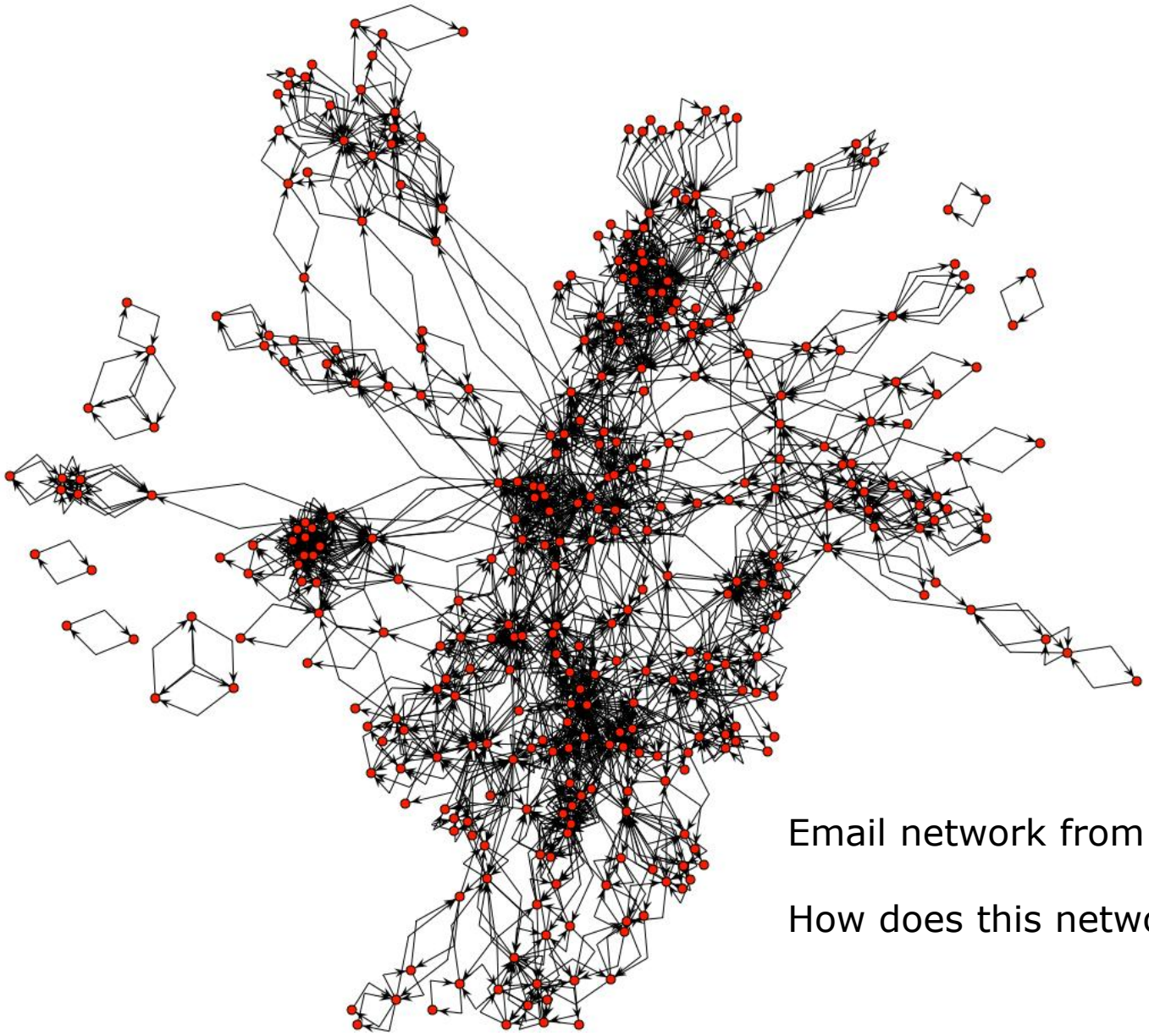


Rank	PRankP		HITSP	
1:	Brin	0.2014	Brin	0.1119
2:	Page	0.1352	Kleinberg	0.1107
3:	Kleinberg	0.1137	Page	0.1087
4:	Motwani	0.0474	Motwani	0.0184
5:	Ullman	0.0429	Raghavan	0.0147
6:	Silverstein	0.0392	Ullman	0.0136
7:	Raghavan	0.0111	Silverstein	0.0119
8:	Lynch	0.0086	Williamson	0.0113
9:	Kedem	0.0086	Papadimitriou	0.0110
10:	Williamson	0.0085	Lynch	0.0108



Algorithms for estimating relative importance in networks
S. White and P. Smyth, *ACM SIGKDD,* 2003.



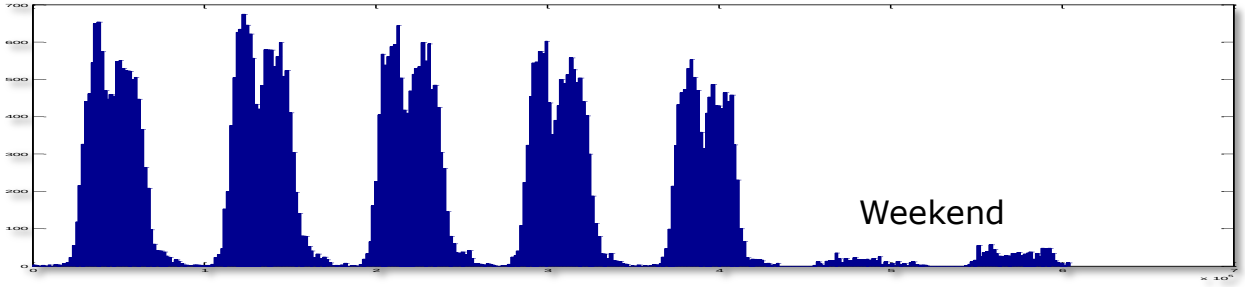


Email network from 500 HP researchers

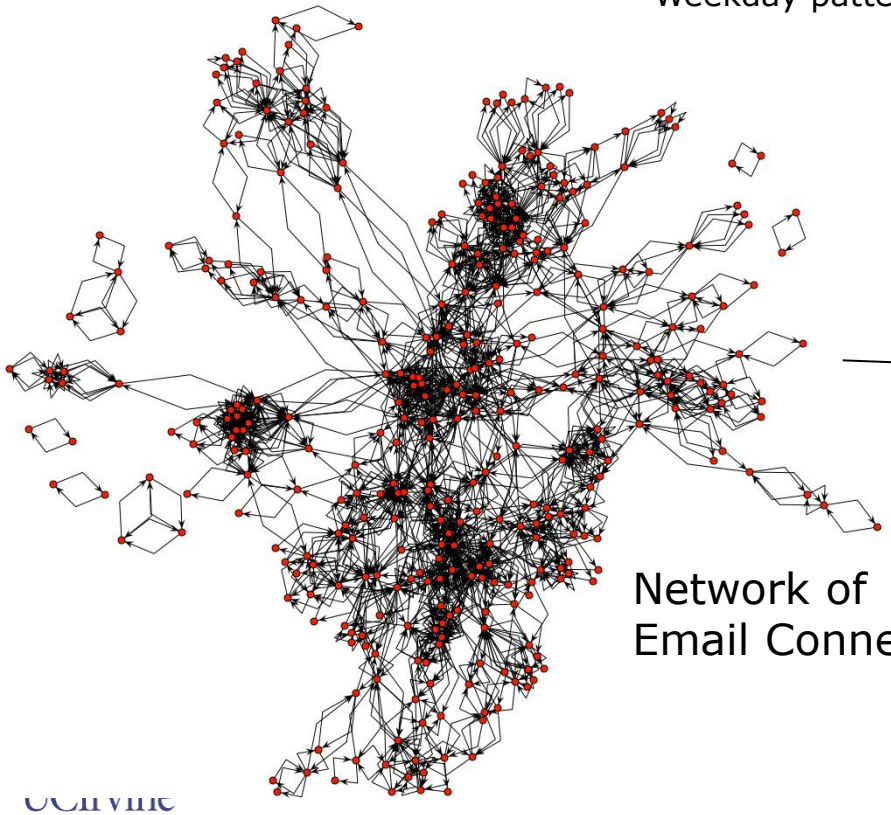
How does this network evolve over time?

Time Series Patterns in Email Data

One week of email volume versus time



Weekday patterns (note dip at lunch)



Network of
Email Connectivity

Research goal:
better understand and predict human behavior in these networks over time

DATA MINING TASKS

Different Data Mining Tasks

- Descriptive Methods
  - Exploratory Data Analysis, Visualization
  - Dimension reduction (principal components, factor models, topic models)
  - Clustering
  - Pattern and Anomaly Detection

- Predictive Modeling
  - Classification
  - Ranking
  - Regression
  - Matrix completion (recommender systems)

- Description and Prediction are intimately related
  - Good descriptions should have predictive power
  - We would ideally like to be able to describe/understand why good prediction models work

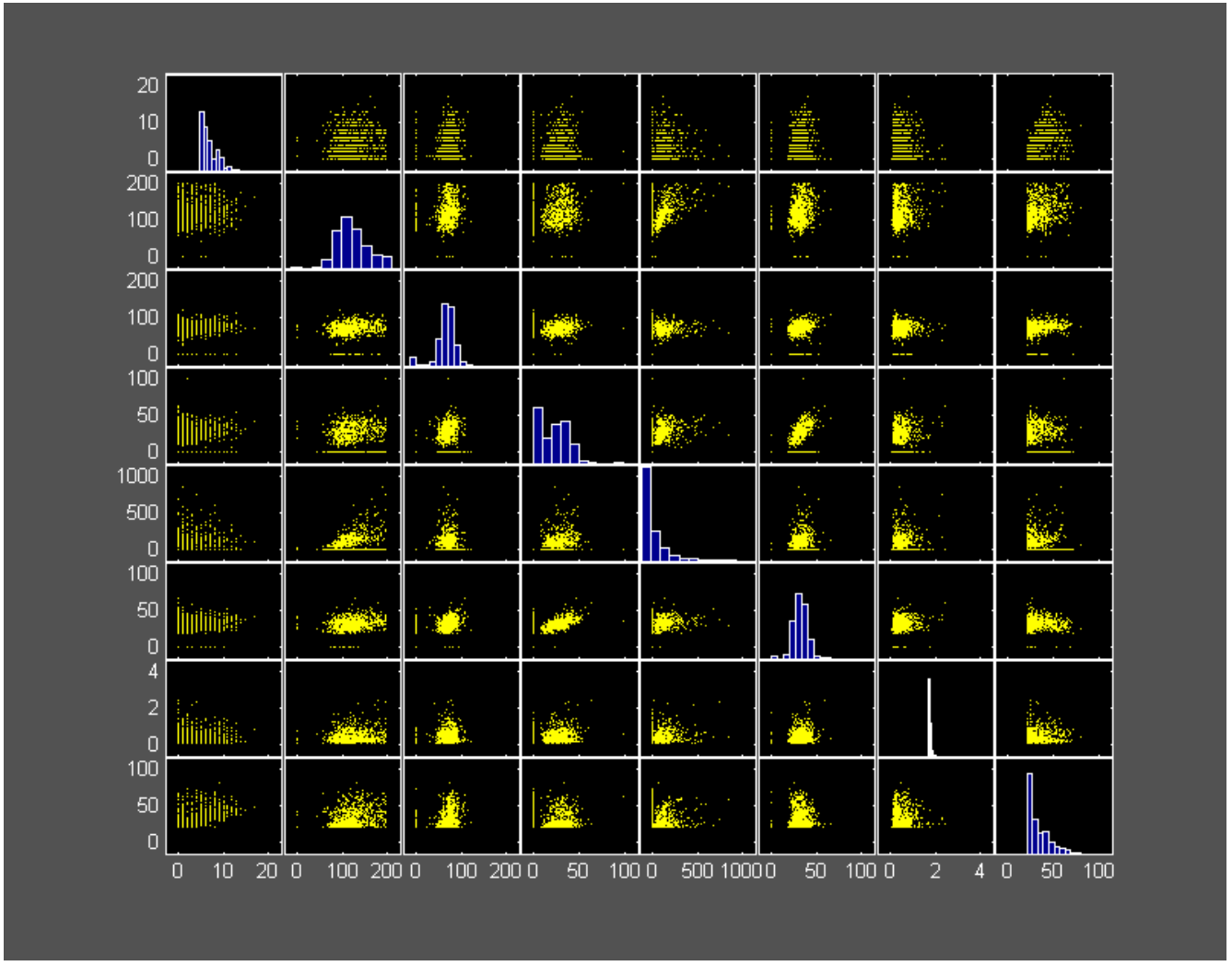
Descriptive Modeling

- Goal is to find an interpretable summarization of the data
  - Detect interesting and interpretable patterns
  - May lead to useful insights about the problem domain
  - May be useful before applying predictive models
    - e.g., discovering that there are 2 very different subpopulations in the data
  - Popular in the sciences, e.g., in biology, climate data analysis, astronomy, etc

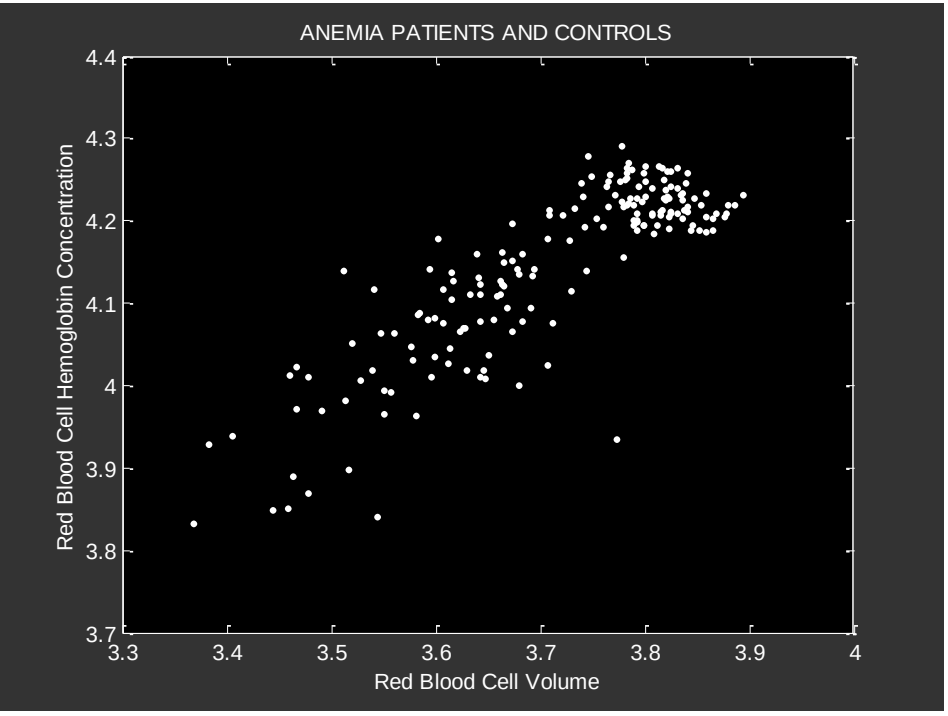
- Examples:
  - Summary statistics
  - Visualization
  - Cluster analysis:
    - Find natural groups in the data
  - Dependency models among the p variables
    - Learning a Bayesian network for the data
  - Dimension reduction
    - Project the high-dimensional data into 2 dimensions to reveal structure

Example of Descriptive Methods: Scatter Matrices

Pima Indians Diabetes data

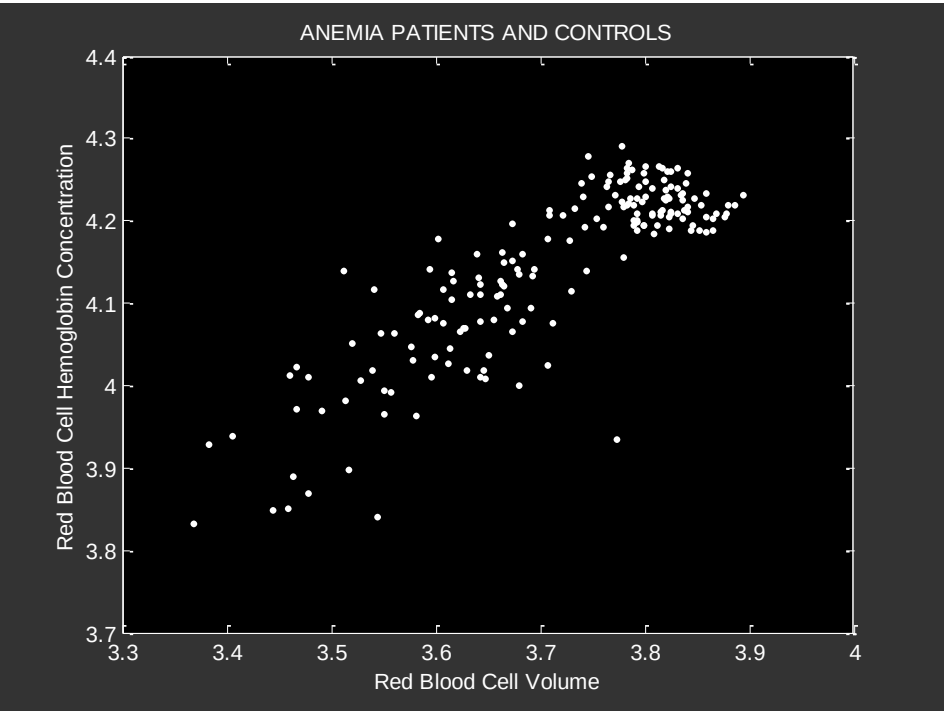


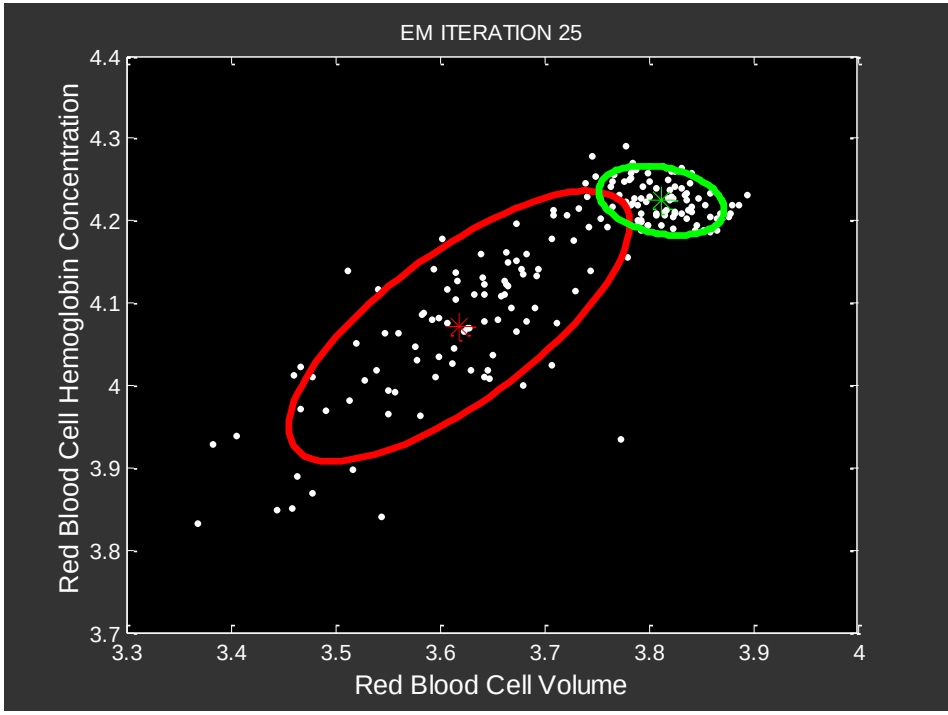
Example of Descriptive Modeling: Clustering Medical Patients



Data courtesy of Professor Christine McLaren,
Department of Epidemiology, UC Irvine

Example of Descriptive Modeling: Clustering Medical Patients



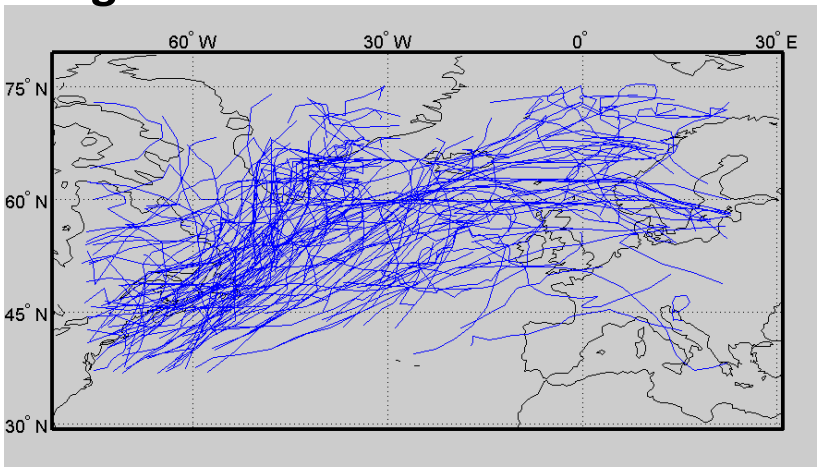


Data courtesy of Professor Christine McLaren,
Department of Epidemiology, UC Irvine

Example of Descriptive Modeling: Clusters of Storm Trajectories

From Gaffney et al., Climate Dynamics, 2007

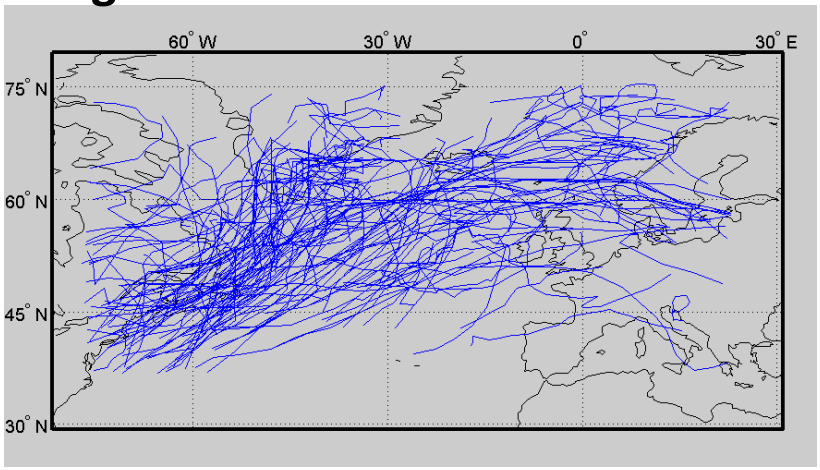
Original Data



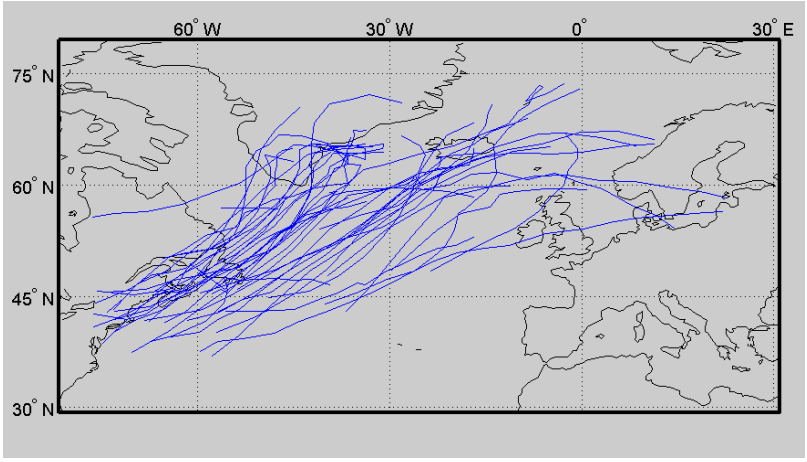
Example of Descriptive Modeling: Clusters of Storm Trajectories

From Gaffney et al., Climate Dynamics, 2007

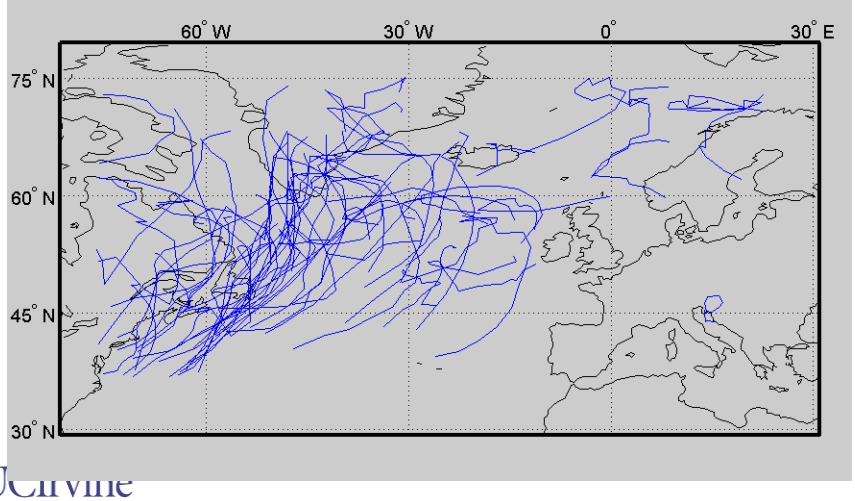
Original Data



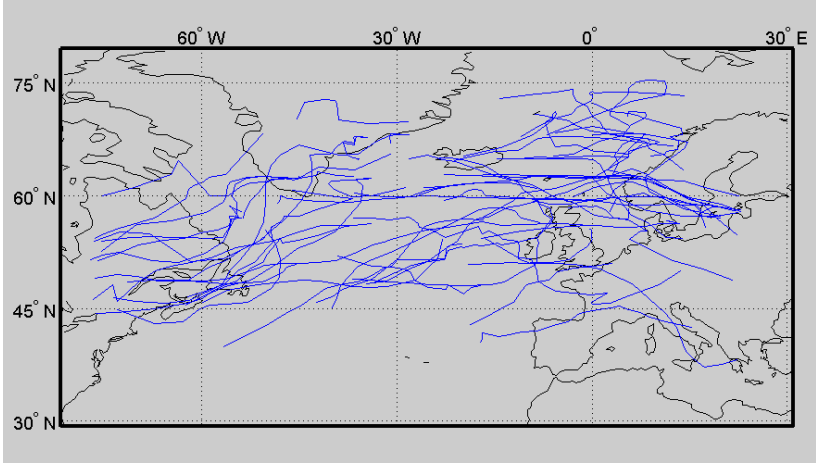
Iceland Cluster



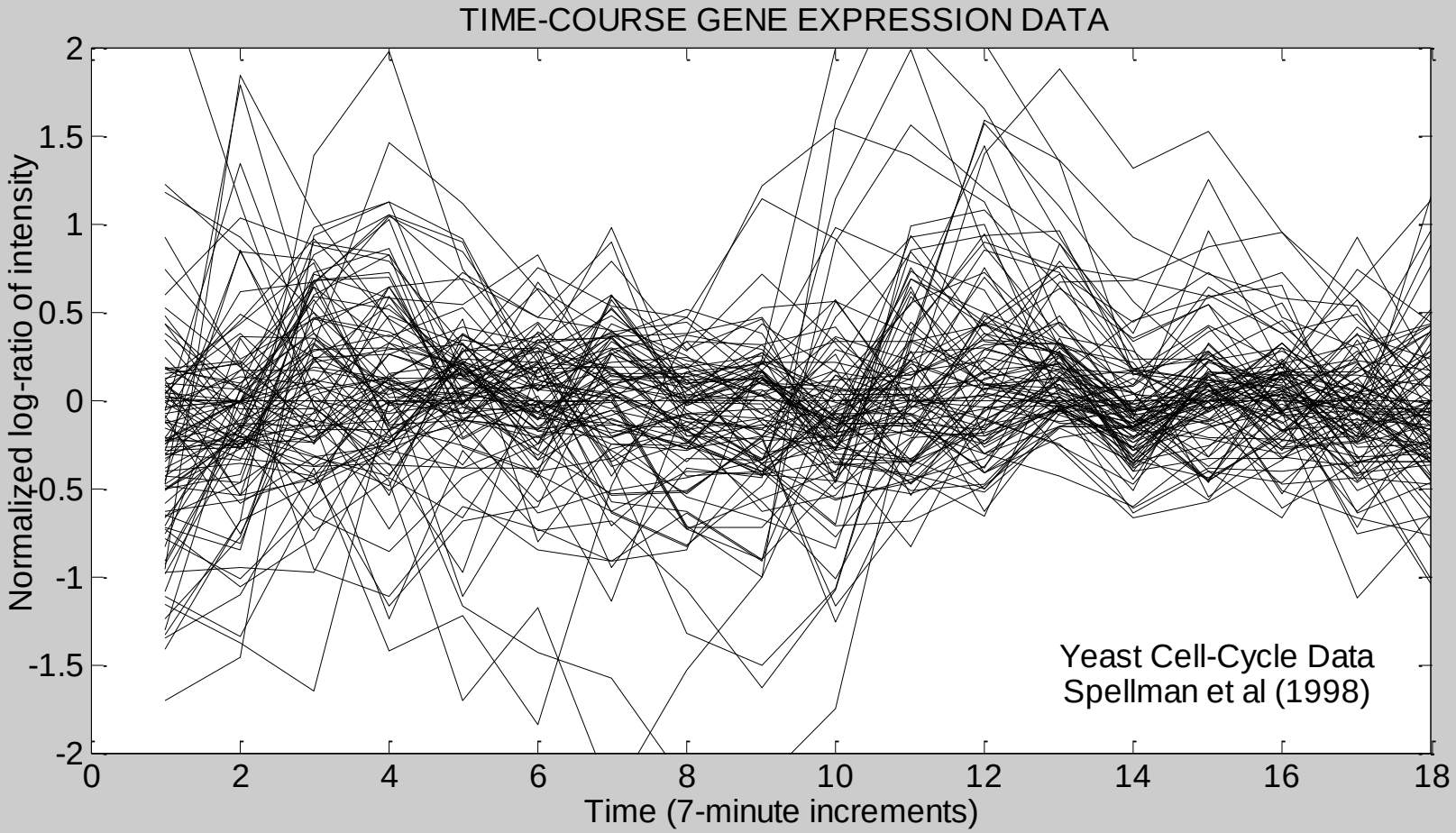
Greenland Cluster



Horizontal Cluster



Descriptive Modeling: Time-Course Gene Expression Data



Pattern Discovery

- Goal is to discover interesting "local" patterns in the data rather than to characterize the data globally

- given market basket data we might discover that
  - If customers buy wine and bread then they buy cheese with probability 0.9
  - These are known as "association rules"

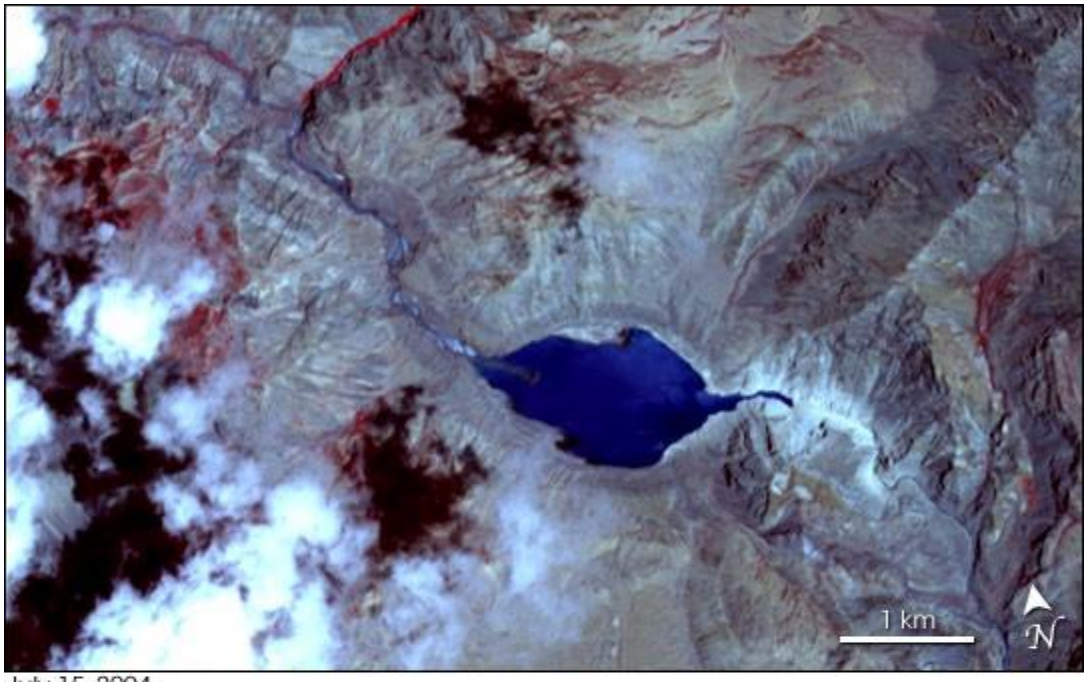
- Given multivariate data on astronomical objects
  - We might find a small group of previously undiscovered objects that are very self-similar in our feature space, but are very far away in feature space from all other objects

Example of Pattern Discovery

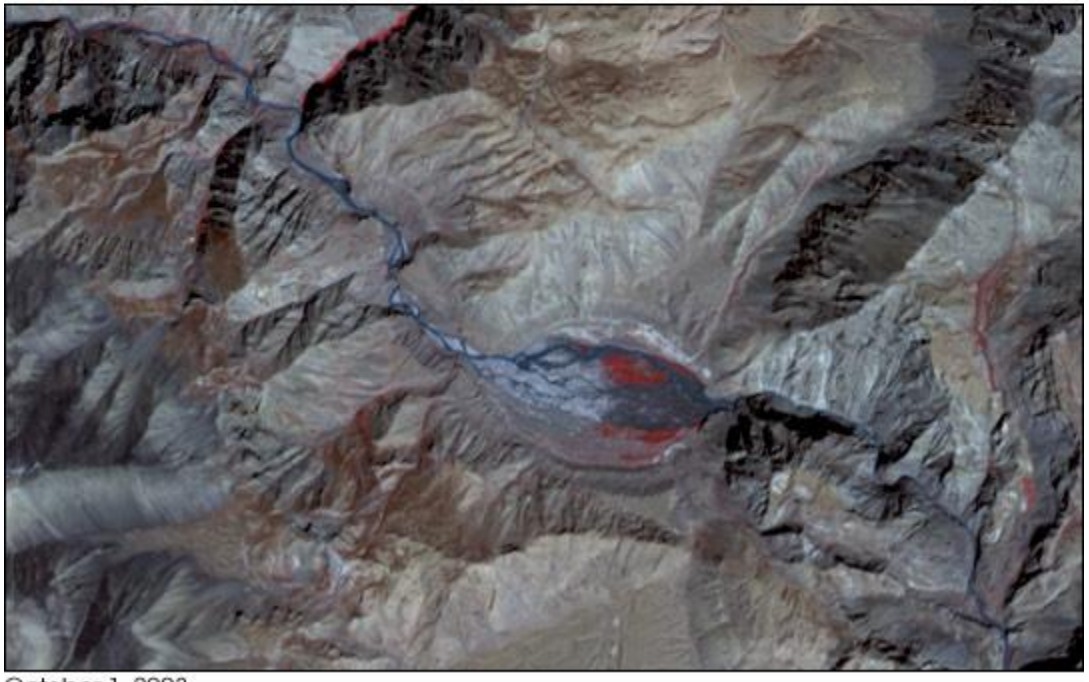
ADACABDABAABBDDBCADDDDBCDDBCCBBCCDADADAADABDBBDABABBCDD
DCDDABDCBBDBDBCBBABBBCBBABCBBACBBDBAACCADDADBDBBCBBCCBB
BDCABDDBBADDBBBBCCACDABBABDDCDDBBABDBDDBDDBCACDBBCCBBAC
DCADCBACCADCCCACCDDADCBCADADBAACCDDDCBDBDCCCCACACACCDAB
DDBCADADBCBDDADABCCABDAACABCABACBDDDCBADCBDADDDDCDDCADC
CBBADABBAAADAAABCCBCABDBAADCBCDACBCABABCCBACBDABDDDADAA
BADCDCCDBBCDBDADDCCBBCDBAADADBCAAAADBDCADBDBBBCDCCBCCCD
CCADAADACABDABAABBDDBCADDDDBCDDBCCBBCCDADADACCCDABAABBC
BDBDBADBBBBCDADABABBDACDCDDDBBCDBBCBBCCDABCADDADBACBBBC
CDBAAADDDBDDCABACBCADCDCBAAADCADDADAABBACCBB

Example of Pattern Discovery

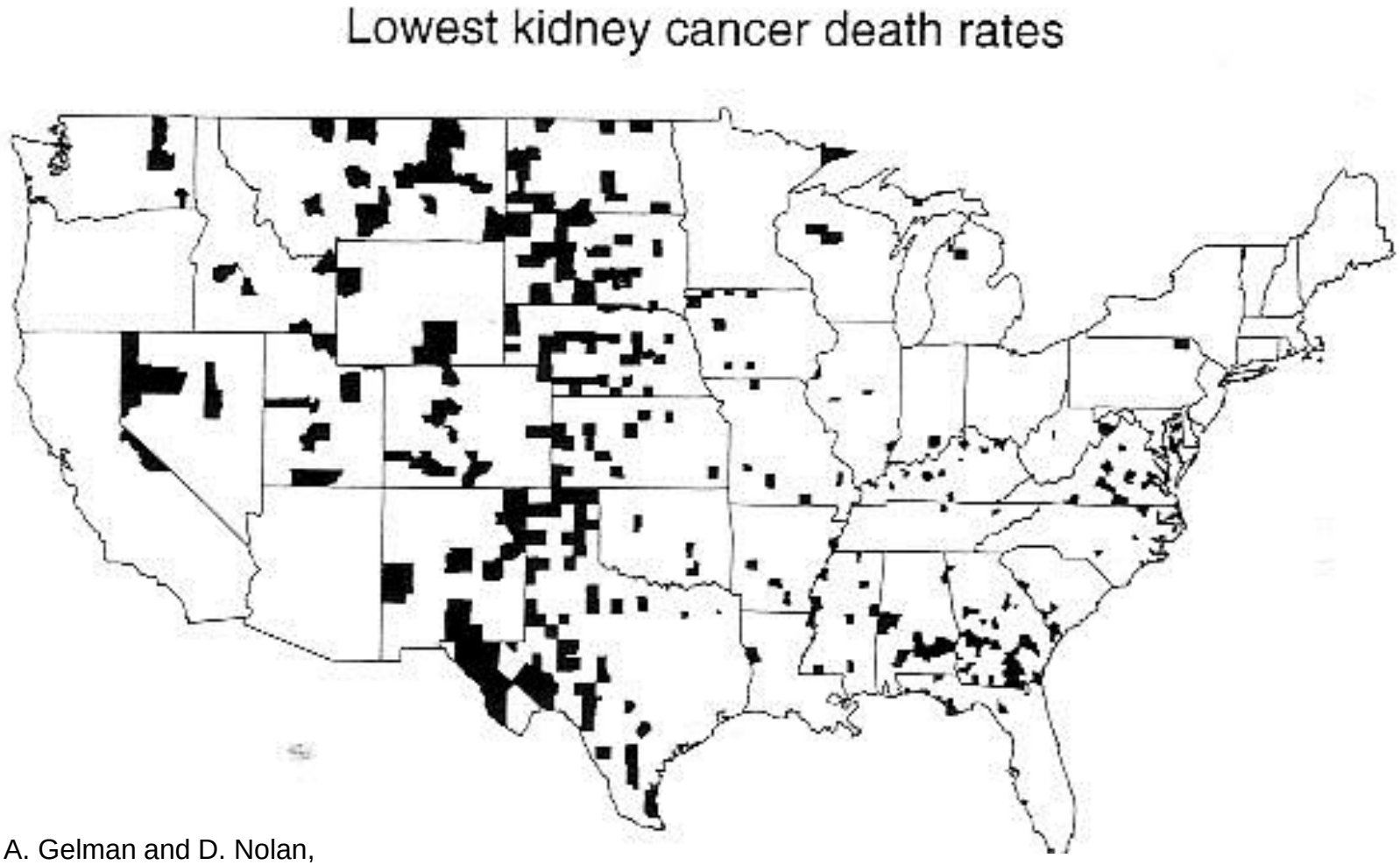
ADACABDABAABBDDBCADDDDBCDDBCCBBCCDADADAADABDBBDABABBCDD
DCDDABDCBBDBDBCBBABBBCBBABCBBACBBDBAACCADDADBDBBCBBCCBB
BDCABDDBBADDBBBBCCACDABBABDDCDDBBABDBDDBDDBCACDBBCCBBAC
DCADCBACCADCCCACCDDADCBCADADBAACCDDDCBDBDCCCCACACACCDAB
DDBCADADBCBDDADABCCABDAACABCABACBDDDCBADCBDADDDDCDDCADC
CBBADABBAAADAAABCCBCABDBAADCBCDACBCABABCCBACBDABDDDADAA
BADCDCCDBBCDBDADDCCBBCDBAADADBCAAAADBDCADBDBBBCDCCBCCCD
CCADAADACABDABAABBDDBCADDDDBCDDBCCBBCCDADADACCCDABAABBC
BDBDBADBBBBCDADABABBDACDCDDDBBCDBBCBBCCDABCADDADBACBBBC
CDBAAADDDBDDCABACBCADCDCBAAADCADDADAABBACCBB



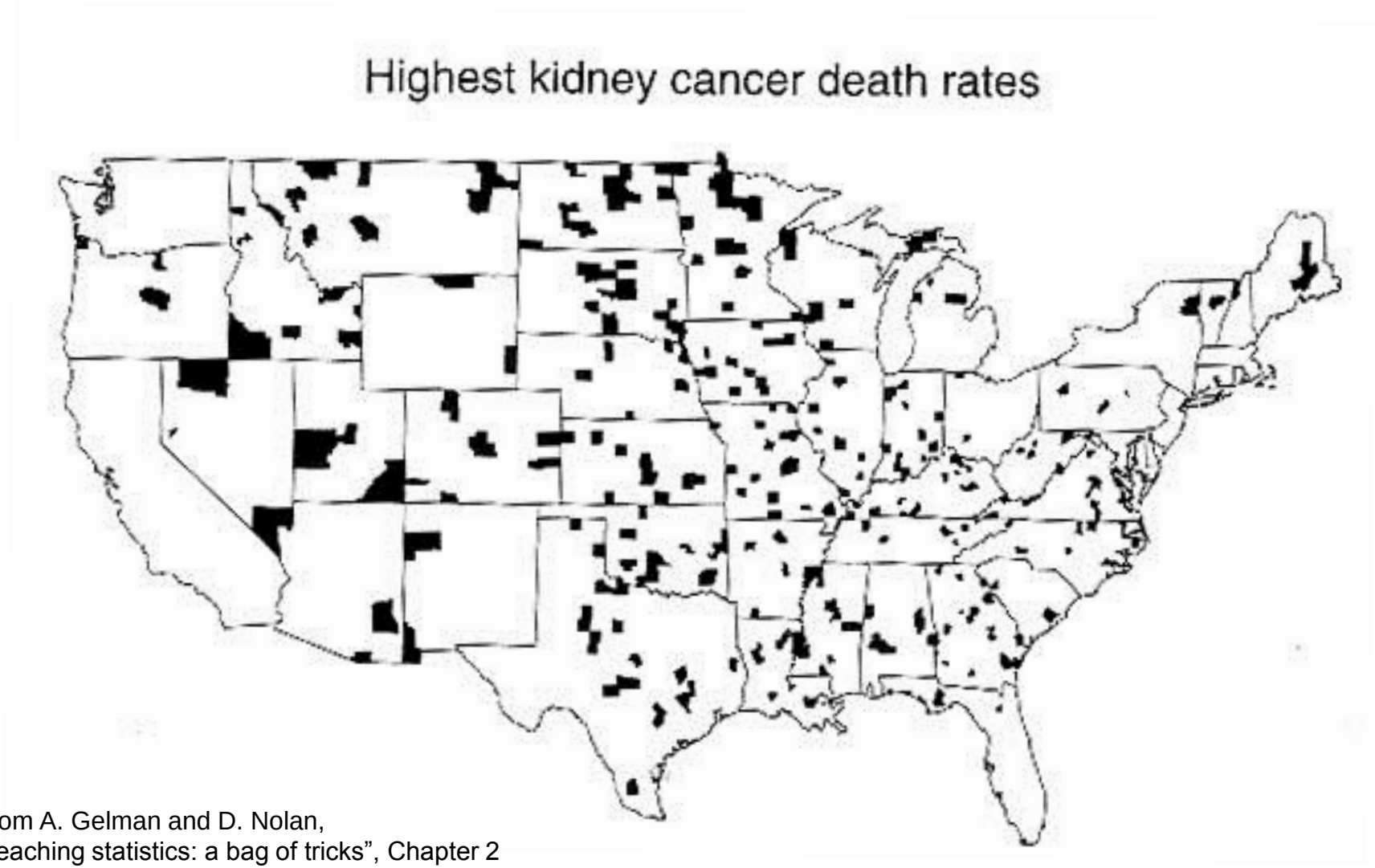
July 15, 2004



October 1, 2003



From A. Gelman and D. Nolan,
"Teaching statistics: a bag of tricks", Chapter 2
Oxford University Press, 2002



From A. Gelman and D. Nolan,
"Teaching statistics: a bag of tricks", Chapter 2
Oxford University Press, 2002

Different Data Mining Tasks

- Descriptive Methods
  - Exploratory Data Analysis, Visualization
  - Dimension reduction (principal components, factor models, topic models)
  - Clustering
  - Pattern and Anomaly Detection

- Predictive Modeling
  - Classification
  - Ranking
  - Regression
  - Matrix completion (recommender systems)

Predictive Modeling

- Predict one variable Y given a set of other variables $\underline{X}$
  - Here $\underline{X}$ could be a p-dimensional vector
  - Classification: Y is categorical
  - Regression: Y is real-valued

- In effect this is function approximation, learning the relationship between Y and $\underline{X}$

- Many, many algorithms for predictive modeling in statistics and machine learning

- Often the emphasis is on predictive accuracy, less emphasis on understanding the model

Predictive Modeling: Fraud Detection

- Credit card fraud detection
  - Credit card losses in the US are over 1 billion $ per year
  - Roughly 1 in 50k transactions are fraudulent

- Approach
  - For each transaction compute P(fraudulent | transaction)
  - Models like logistic regression are widely used to estimate P
  - Inputs to the model are features based on a customer's behavior and demographics
  - Model is built on historical data of known fraud/non-fraud
  - High probability transactions investigated by fraud police

- Example:
  - Fair-Isaac/HNC's fraud detection software based on neural networks, led to reported fraud decreases of 30 to 50%

- Issues
  - Significant feature engineering/preprocessing
  - false alarm rate vs missed detection – what is the tradeoff?

Predictive Modeling: Customer Scoring

- Example: a bank has a database of 1 million past customers, 10% of whom took out mortgages
- Use machine learning to rank new customers as a function of P(defaults on mortgage | customer data)
- Customer data
  - History of transactions with the bank
  - Other credit data (obtained from Experian, etc)
  - Demographic data on the customer or where they live
- Techniques
  - Binary classification: logistic regression, decision trees, etc
  - Many, many applications of this nature

THE DARK SIDE OF DATA MINING

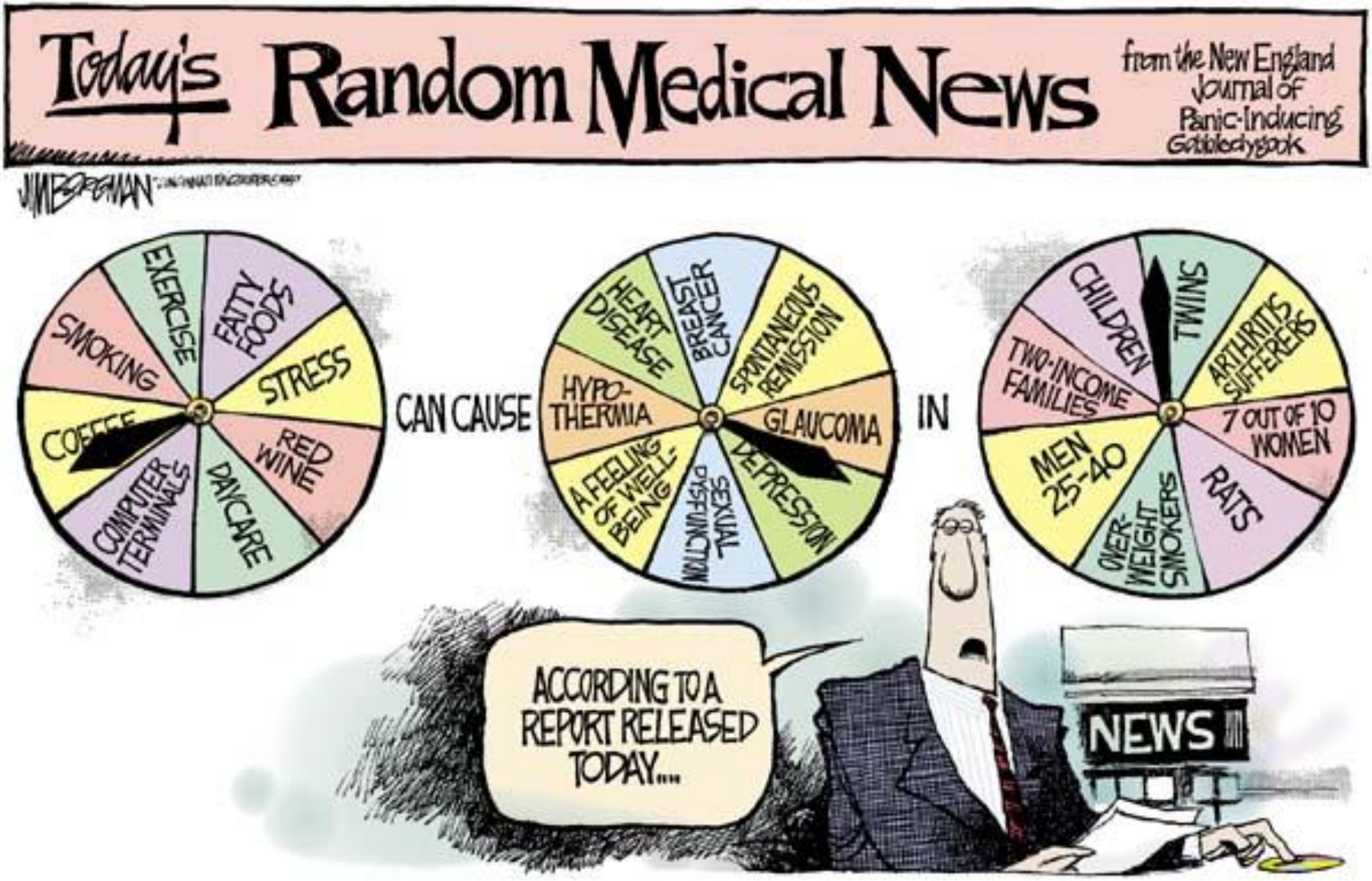
Data Mining: the Dark Side

- Hype

- Data dredging, snooping and fishing
  - Finding spurious structure in data that is not real
  - Correlation does not imply causation

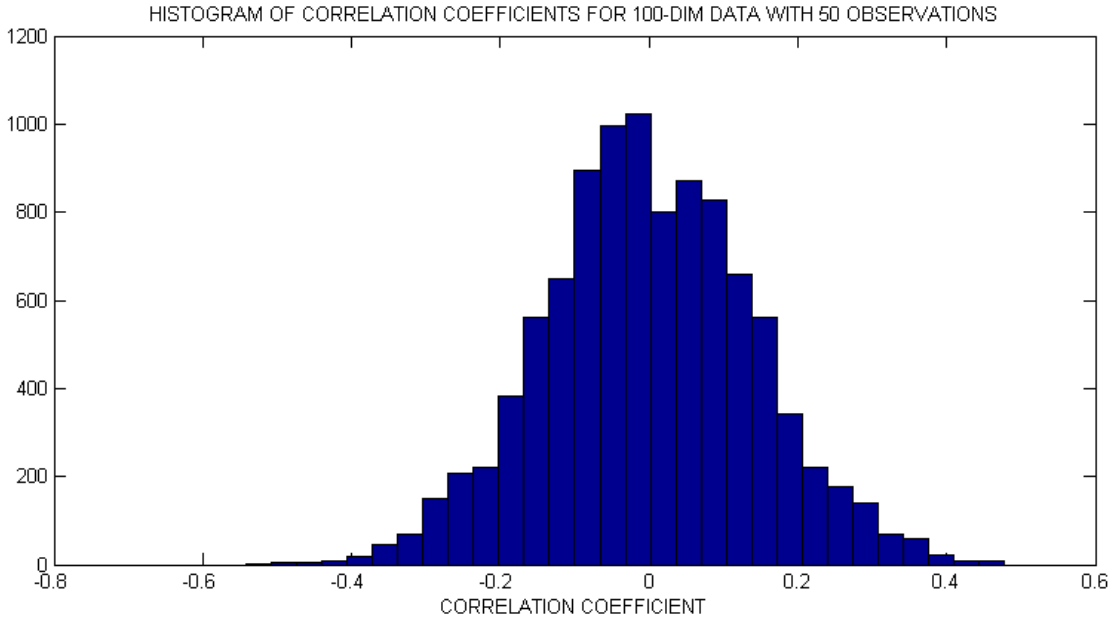
- historically, 'data mining' was a derogatory term in the statistics community
  - making inferences from small samples
    - e.g., the Super Bowl fallacy

- The challenges of being interdisciplinary
  - computer science, statistics, domain discipline


Today's Random Medical News
from the New England Journal of Panic-Inducing Gobbledygook
JIM BORGMAN
SMOKING
EXERCISE
FATTY FOODS
STRESS
COFFEE
RED WINE
COMPUTER TERMINALS
DAYCARE
CAN CAUSE
HEART DISEASE
BREAST CANCER
SPONTANEOUS REMISSION
HYPO-THERMIA
GLAUCOMA
A FEELING OF WELL-BEING
SEXUAL DYSFUNCTION
DEPRESSION
IN
CHILDREN
TWINS
ARTHRITIS SUFFERERS
TWO-INCOME FAMILIES
7 OUT OF 10 WOMEN
MEN 25-40
OVER-WEIGHT SMOKERS
RATS
ACCORDING TO A REPORT RELEASED TODAY...
NEWS

Example of "data fishing"

- Example: data set with
  - 50 data vectors
  - 100 variables
  - Even if data are entirely random (no dependence) there is a very high probability some variables will appear dependent just by chance.



Tom Swift and His Electric Factor Analysis Machine

J. SCOTT ARMSTRONG
Massachusetts Institute of Technology

Problems in the use of factor analysis for deriving theory are illustrated by means of an example in which the underlying factors are known. The actual underlying model is simple and it provides a perfect explanation of the data. While the factor analysis "explains" a large proportion of the total variance, it fails to identify the known factors in the model. The illustration is used to emphasize that factor analysis, by itself, may be misleading as far as the development of theory is concerned. The use of a comprehensive and explicit a priori analysis is proposed so that there will be independent criteria for the evaluation of the factor analytic results.

* * * * *

It has not been uncommon for social scientists to draw upon analogies from the physical sciences in their discussions of scientific methods. They look with envy at some of the mathematical advances in the physical sciences and one gets the impression that the social sciences are currently on the verge of some major mathematical advances. Perhaps they are—but there are many social scientists who would disagree. Their position is that we really don't know enough about what goes into our mathematical models in order to expect results which are meaningfully related to anything in the "real world."

In other words, the complaint is not that the models are no good or that they don't really give us optimum results; rather it is that the assumptions on which the model is based do not provide a realistic representation of the world as it exists. And it is in this area where the social sciences differ from the physical sciences.

But now, thanks to recent advances in computer technology, and to refinements in mathematics, social scientists can analyze masses of data and determine just what the world is like. Armchair theorizing has lost some of its respectability. The computer provides us with objective results.

Despite the above advances, there is still a great deal of controversy over the relevant roles of theorizing and of empirical analysis. We should note that the problem extends beyond one of scientific methodology; it is also an emotional problem with scientists. There is probably no one reading this paper who is not aware of the proper relationship between theorizing and empirical analysis. On the other hand, we all know of *others* who do not understand the problem. We are willing to label others as either theorists or empiricists; and we note that these people argue over the relative merits of each approach.

17

American Statistician, v.17, 1967

ALCHEMY IN THE BEHAVIORAL SCIENCES*

BY HILLEL J. EINHORN

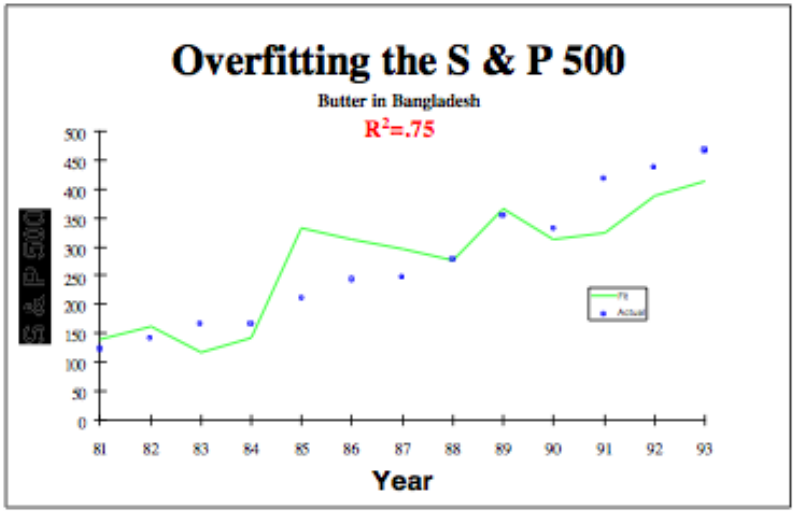
Access to powerful new computers has encouraged routine use of highly complex analytic techniques, often in the absence of any theory, hypotheses, or model to guide the researcher's expectations of results. The author examines the potential of such techniques for generating spurious results, and urges that in exploratory work the outcome be subjected to a more rigorous criterion than the usual tests of statistical significance.

Hillel Einhorn is Assistant Professor of Behavioral Science, Graduate School of Business, University of Chicago.

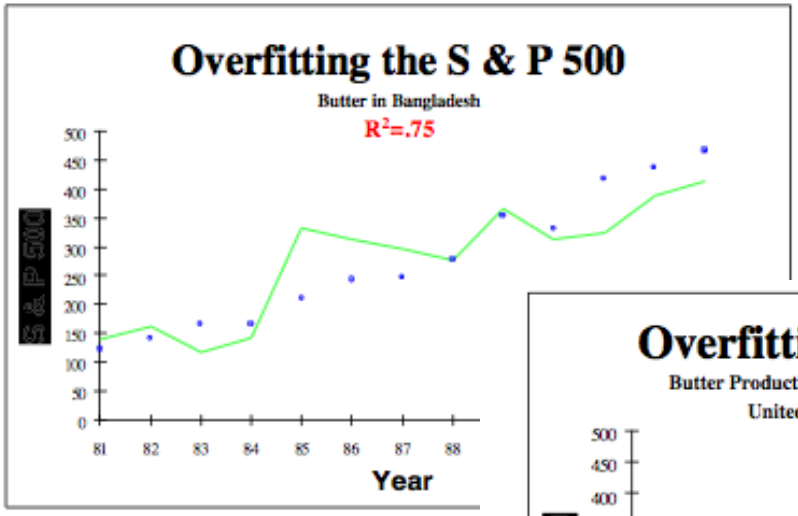
WITH THE LARGE-SCALE use of electronic computers, powerful new methods for data analysis have become quite prevalent. Although this development may be viewed with considerable enthusiasm by some,[1] others may view the gains to be derived from increased ability to handle large amounts of data with increasingly sophicated tools with more than a certain degree of skepticism. Such skepticism is based on the observation that as methods and techniques get more complicated, the role of theory in research is being dangerously ignored in favor of purely empirical work that proceeds without so much as a hypothesis. Like Pirandello's characters in search of an author, many of today's researchers seem to have an assortment of techniques in search of a substantive problem.

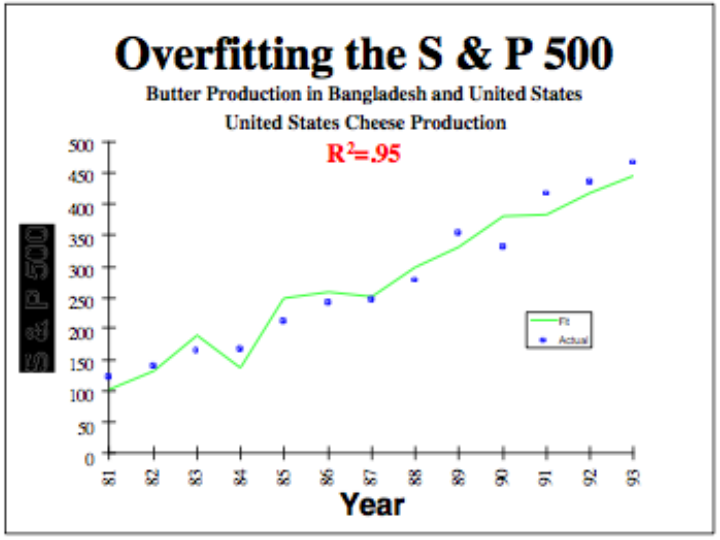
The general question of proceeding inductively or deductively in science is not easily answered. As Armstrong has put it,

Public Opinion Quarterly, 1972



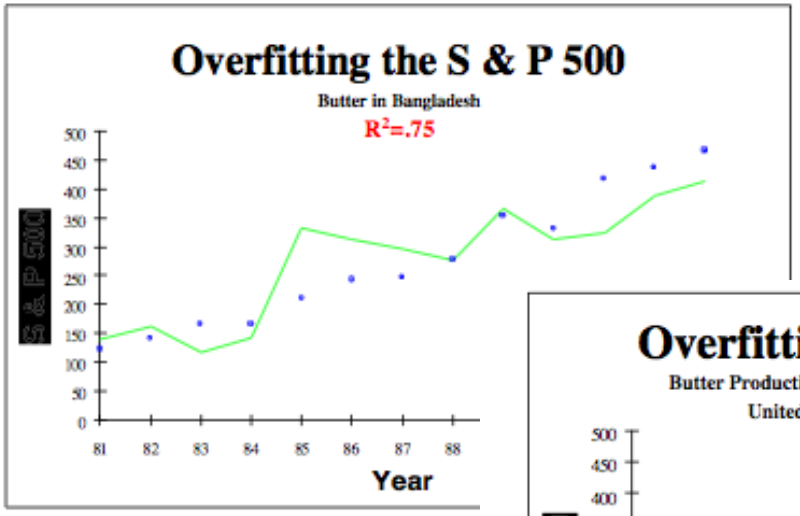
Stupid Data Miner Tricks: Overfitting the S&P 500 David Leinweber…

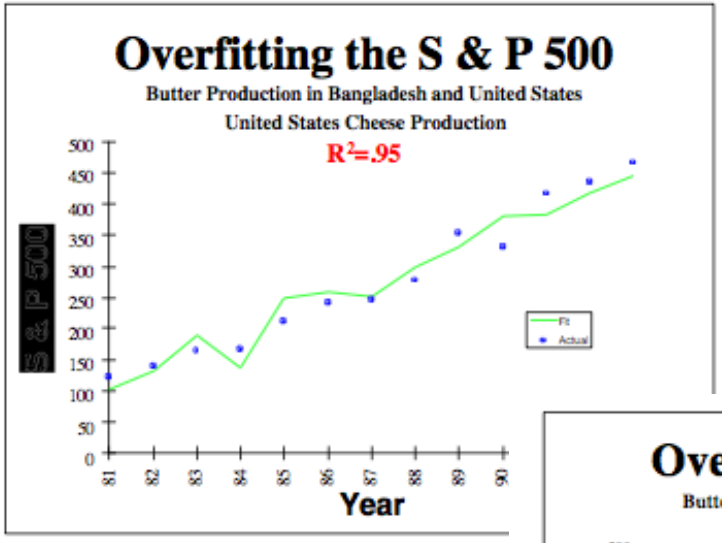


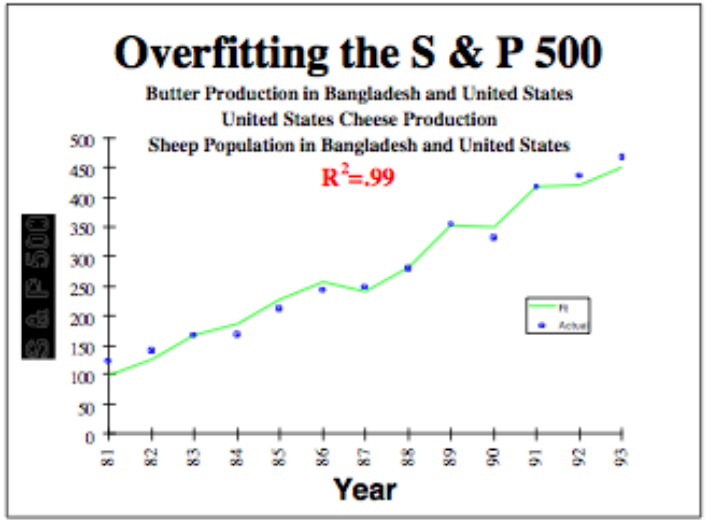


Stupid Data Miner Tricks: Overfitting the S&P 500
David Leinweber…

Stupid Data Miner Tricks: Overfitting the S&P 500
David Leinweber…







Assignment 1

http://www.ics.uci.edu/~smyth/courses/cs277/assignment1.xht

Due: Wednesday Jan 15th

Outline

- Census income data
- Exploratory Data Analysis tasks
- Software
  - R/Rminer, Matlab, and scikit-learn