Lecture

Data Science

Prof. Dr. Steffen Staab

Christoph Kling

Consulting

Dr. Christoph Tempich
Head of Consulting

christoph.tempich@inovex.de
Christoph Tempich bei Google+
Christoph Tempich bei Twitter
Christoph Tempich bei XING

Organization

- Module number: 04IN2043

Target audience

- Master Web Science, Inf, CV, eGov, WI, IM
- Bachelor (Wahlpflicht): Inf, CV, WI

- http://www.uni-koblenz-landau.de/campus-koblenz/fb4/west/teaching/ss14/proseminar
- Klips-registration

Tutorials (Übungen)

- Programming with data
  - Statistics environment (e.g. R)
  - Cloud programming (e.g. Hadoop)
- Paper based excercises

Date: (tbd)

Will ask for:

- Knowledge, understanding and capability
  - Acquired from lecture
  - Acquired from pen-and-paper excercises
  - Acquired from programming

Subject to change

DATA SCIENCE: OVERVIEW

Acknowledgements

Most slides from

Hanspeter Pfister
pfister@seas.harvard.edu

Joe Blitzstein
blitzstein@stat.harvard.edu

https://drive.google.com/folderview?id=0BxYkKyLxfsNVd0xicUVDS1dIS0k

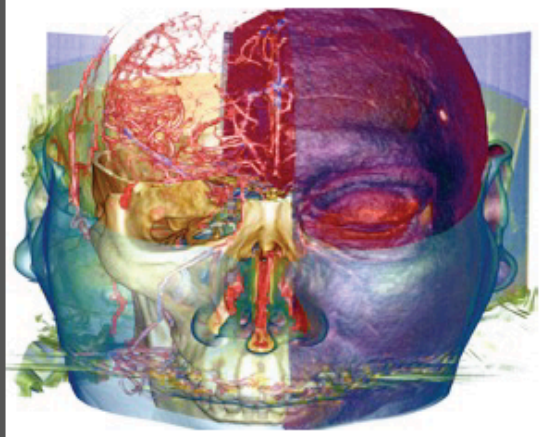
http://www.kdnuggets.com/2013/11/harvard-cs109-data-science-course-resources-free-online.html

Overview (subject to change)

1. History and background, from statistics to programming; examples
2. Process
3. Background in statistics
4. Hypothesis driven research
5. Programming paradigms
6. Visualization
7. Simple machine learning on large scale data
8. Example application domain: text
9. Privacy

Data Science

To gain insights into data through computation, statistics, and visualization



A Data Scientist Is...

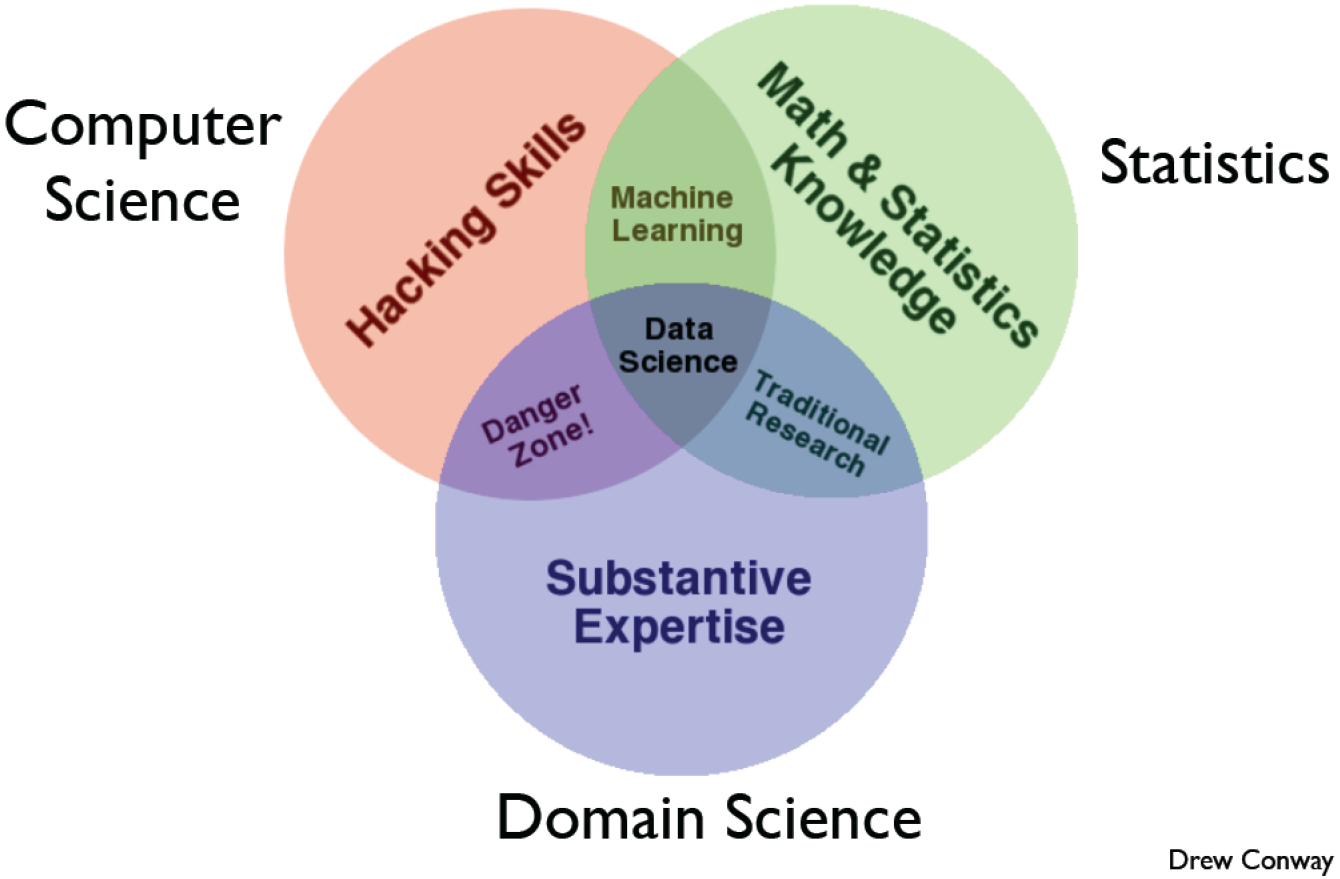
"A data scientist is someone who knows more statistics than a computer scientist and more computer science than a statistician."

- Josh Blumenstock

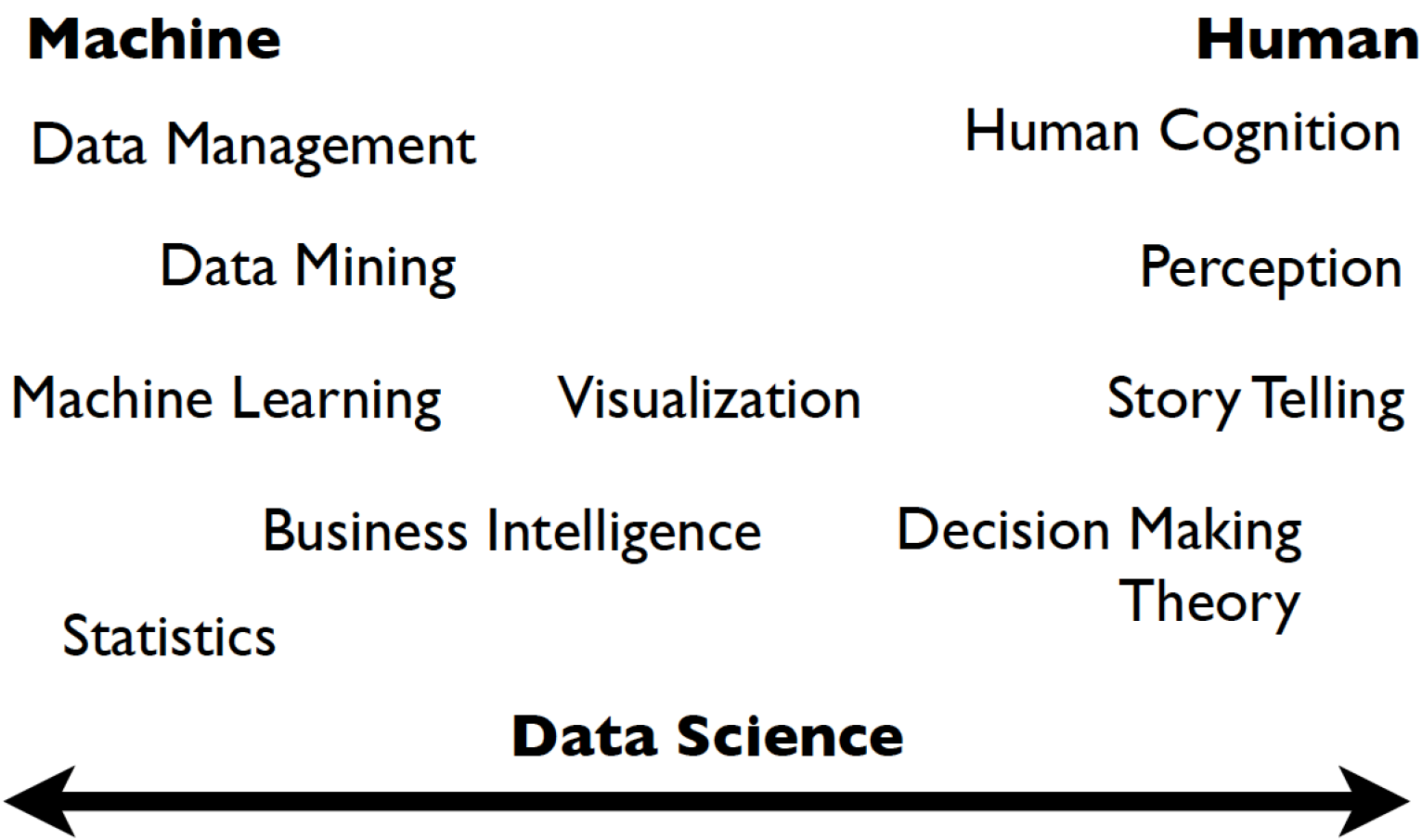
"Data Scientist = statistician + programmer + coach + storyteller + artist"

- Shlomo Aragmon

Data Science



Drew Conway



Inspired by Daniel Keim, "Visual Analytics: Definition, Process, and Challenges"

EXAMPLES

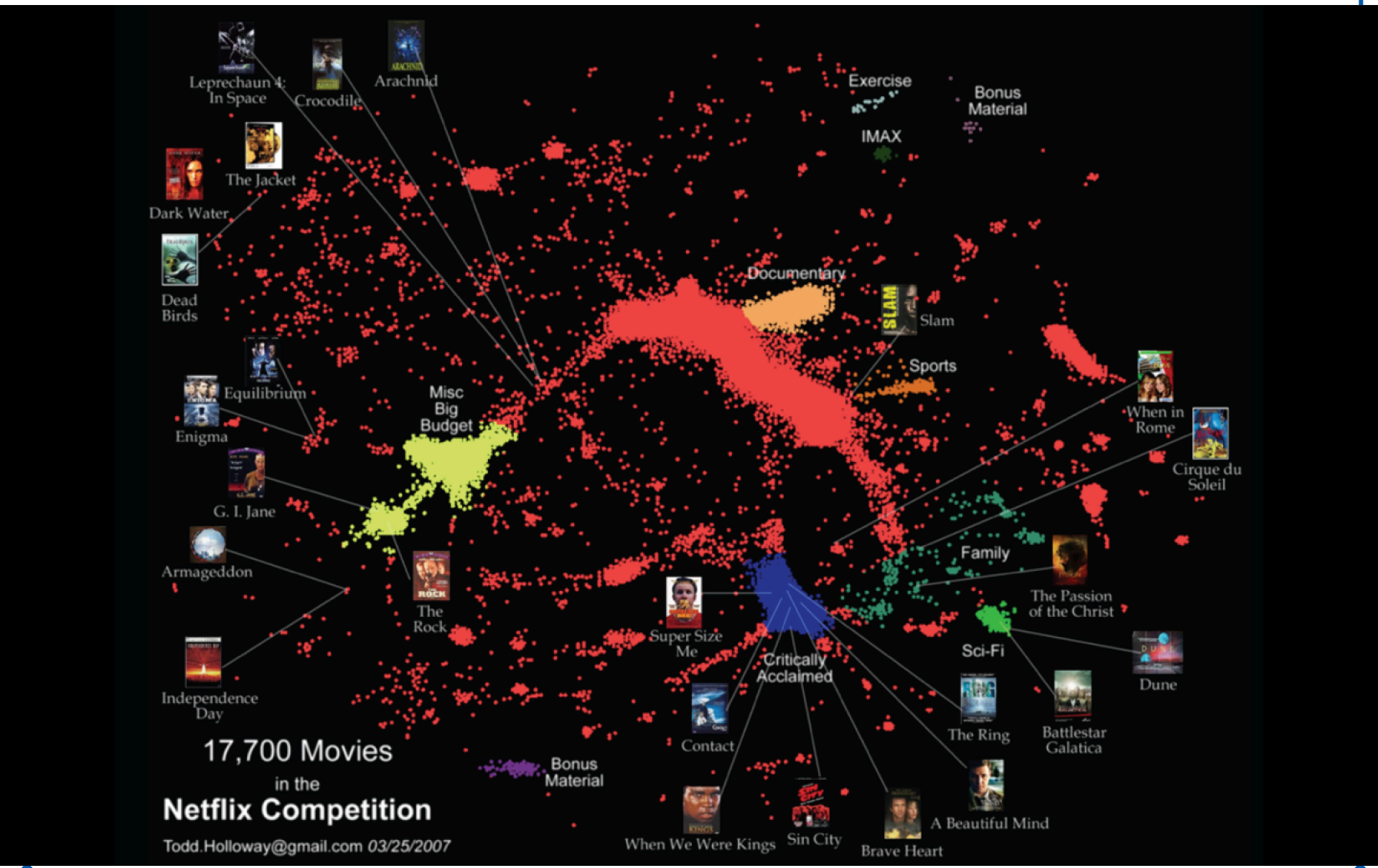
“Nate Silver won the election” – Harvard Business Review

Silver: „Pundits are no better than a coin toss.“

Some Key Principles

- use many data sources (the plural of anecdote is not data)
- understand how the data were collected (sampling is essential)
- weight the data thoughtfully (not all polls are equally good)
- use statistical models (not just hacking around in Excel)
- understand correlations (e.g., states that trend similarly)
- think like a Bayesian, check like a frequentist (reconciliation)
- have good communication skills
  (What does a 60% probability even mean? How can we visualize, validate, and understand the conclusions?)

Netflix Prize


Leprechaun 4: In Space
Crocodile
Arachnid
Exercise
Bonus Material
IMAX
The Jacket
Dark Water
Dead Birds
Documentary
Slam
Sports
Equilibrium
Misc Big Budget
Enigma
When in Rome
Cirque du Soleil
G. I. Jane
Armageddon
The Rock
Family
The Passion of the Christ
Super Size Me
Critically Acclaimed
Sci-Fi
Dune
Independence Day
Contact
The Ring
Battlestar Galatica
17,700 Movies
in the
Netflix Competition
Todd.Holloway@gmail.com 03/25/2007
Bonus Material
When We Were Kings
Sin City
Brave Heart
A Beautiful Mind

Netflix Prize Progress

http://blogs.hbr.org/2012/10/big-data-hype-and-reality/

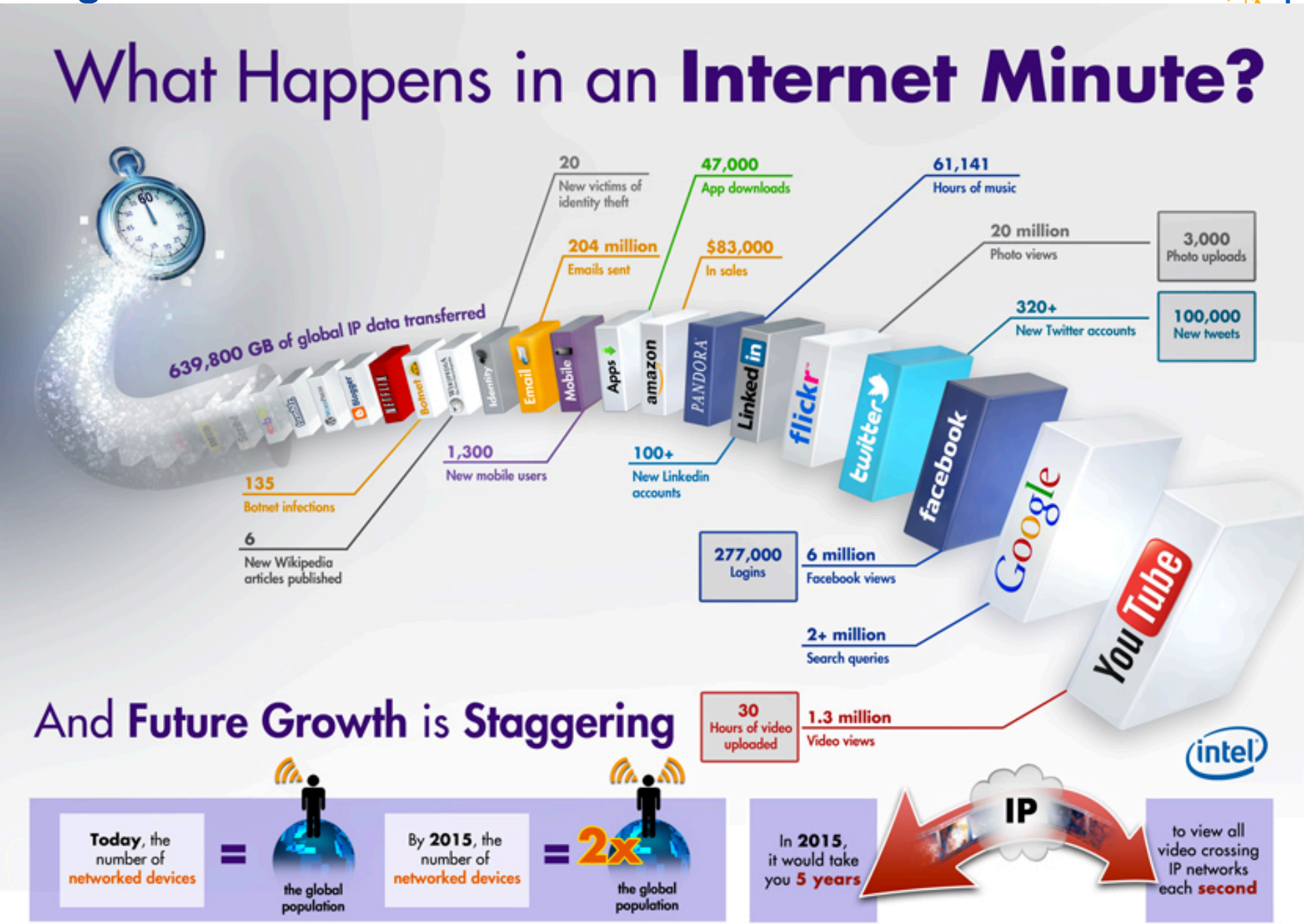
Some Challenges

- *massive data* (500k users, 20k movies, 100m ratings)
- *curse of dimensionality* (very high-dimensional problem)
- *missing data* (99% of data missing; not missing at random)
- *extremely complicated set of factors that affect people's ratings of movies* (actors, directors, genre, ...)
- need to avoid *overfitting* (test data vs. training data)

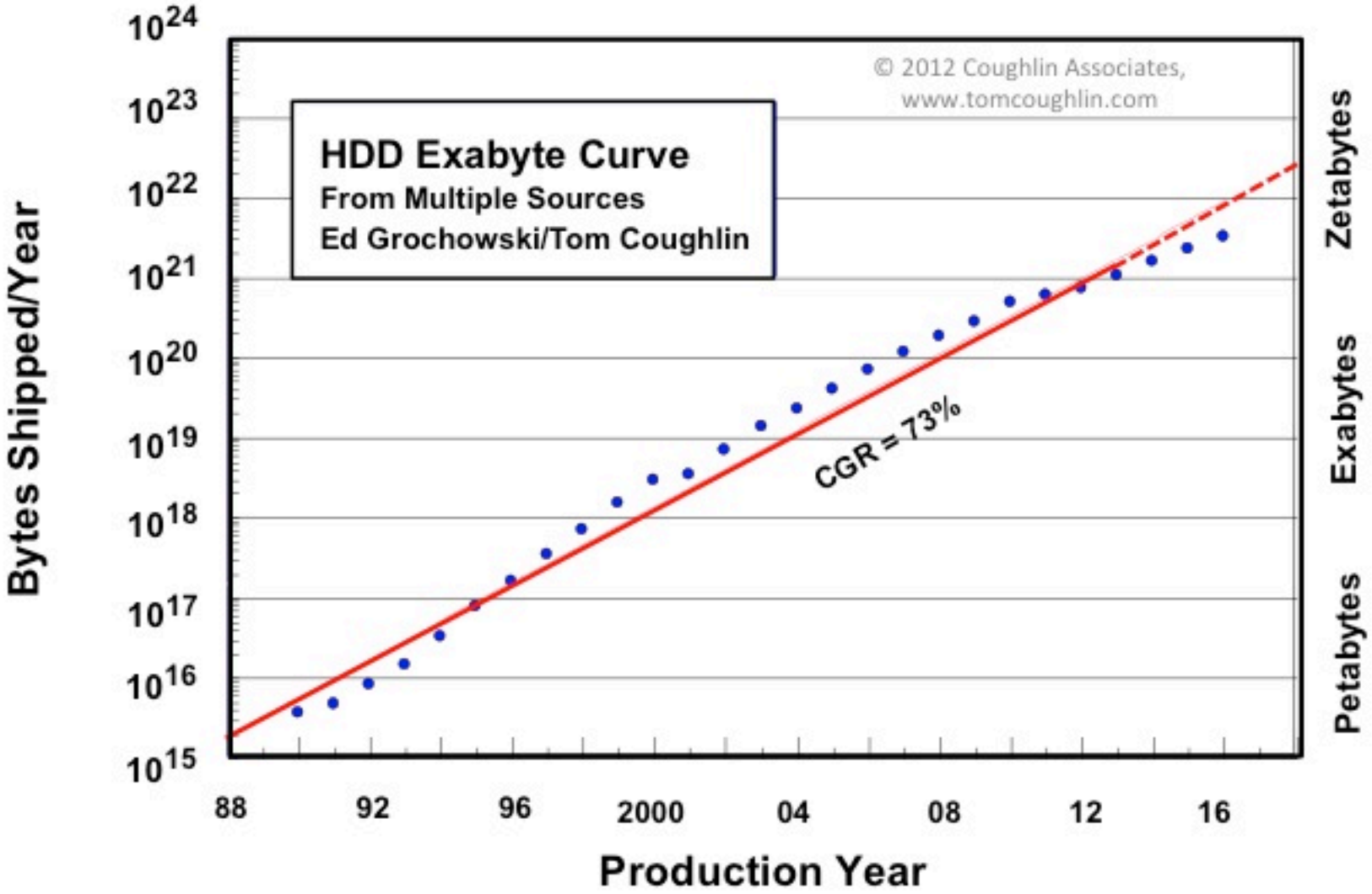
No silver bullet: Netflix prize only semi-successful

http://www.techdirt.com/blog/innovation/articles/20120409/03412518422/why-netflix-never-implemented-algorithm-that-won-netflix-1-million-challenge.shtml

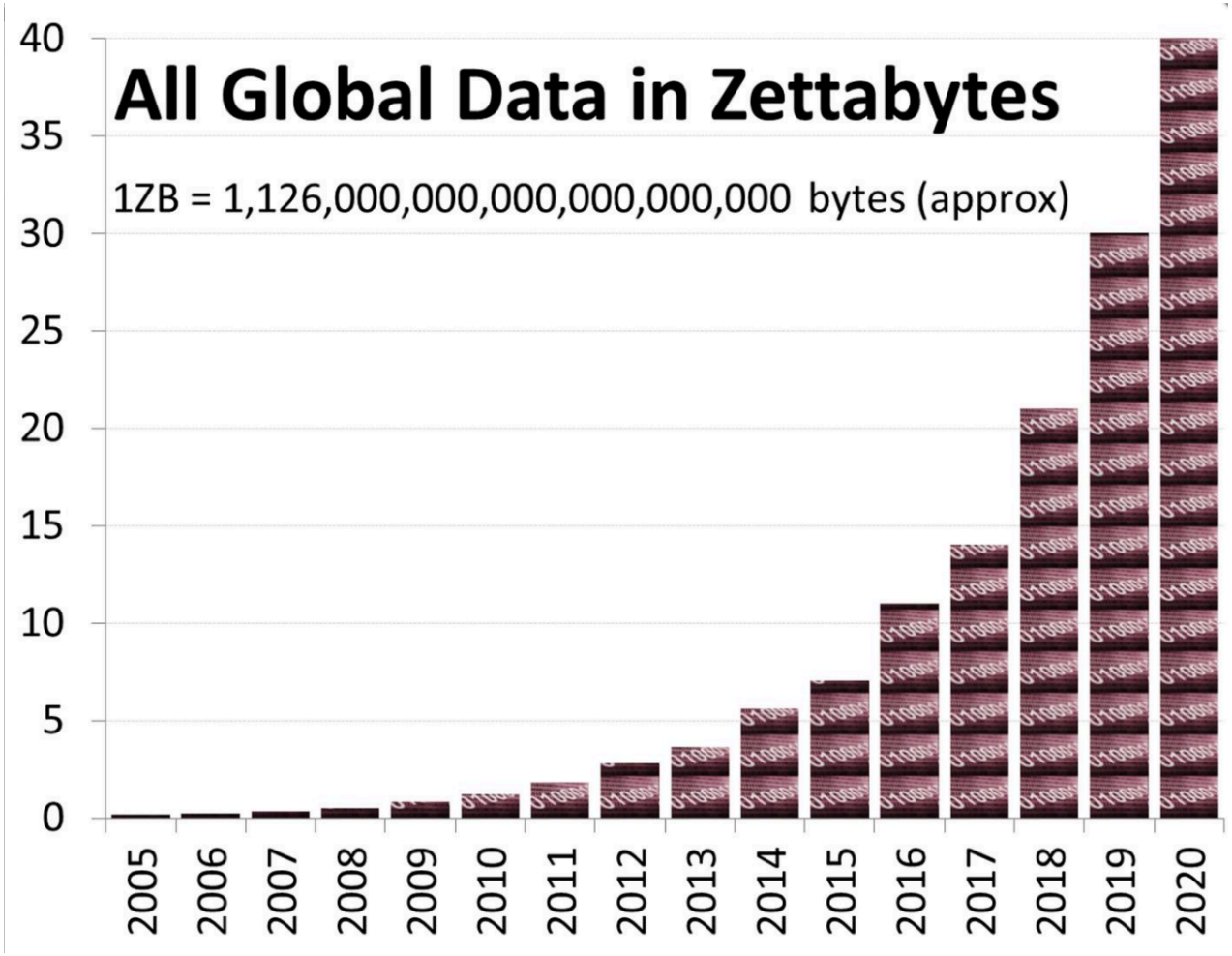
OPPORTUNITY: BIG DATA


What Happens in an Internet Minute?
639,800 GB of global IP data transferred
20
New victims of identity theft
204 million
Emails sent
47,000
App downloads
$83,000
In sales
61,141
Hours of music
20 million
Photo views
3,000
Photo uploads
320+
New Twitter accounts
100,000
New tweets
1,300
New mobile users
100+
New Linkedin accounts
135
Botnet infections
6
New Wikipedia articles published
277,000
Logins
6 million
Facebook views
2+ million
Search queries
30
Hours of video uploaded
1.3 million
Video views
And Future Growth is Staggering
Today, the number of networked devices
=
the global population
By 2015, the number of networked devices
=
2x
the global population
In 2015, it would take you 5 years
IP
to view all video crossing IP networks each second
intel

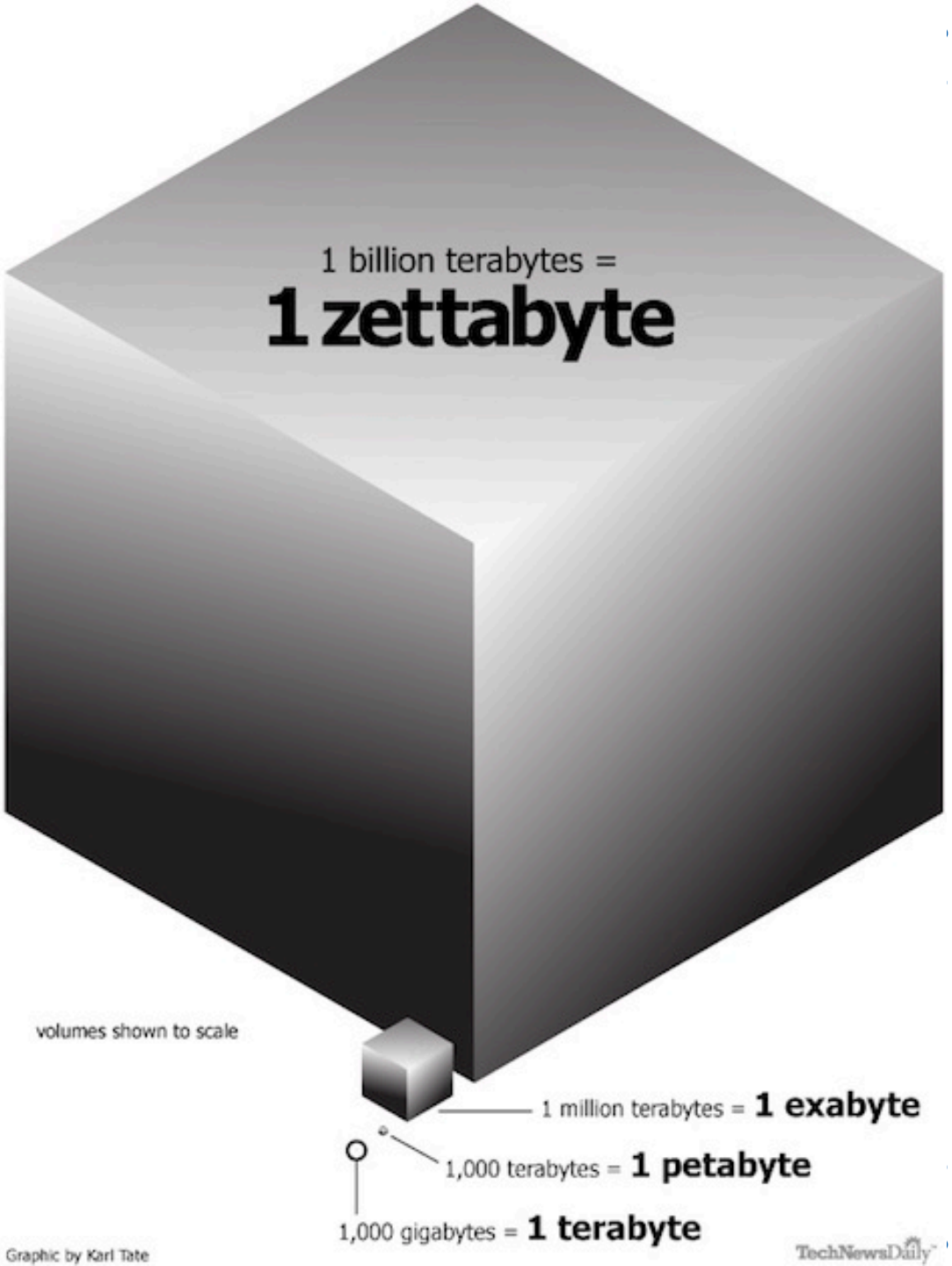
WW HDD CAPACITY SHIPMENTS (1999 THROUGH 2016)



http://www.forbes.com/sites/tomcoughlin/2012/10/03/have-hard-disk-drives-peaked/


All Global Data in Zettabytes
1ZB = 1,126,000,000,000,000,000,000 bytes (approx)
40
35
30
25
20
15
10
5
0
2005
2006
2007
2008
2009
2010
2011
2012
2013
2014
2015
2016
2017
2018
2019
2020

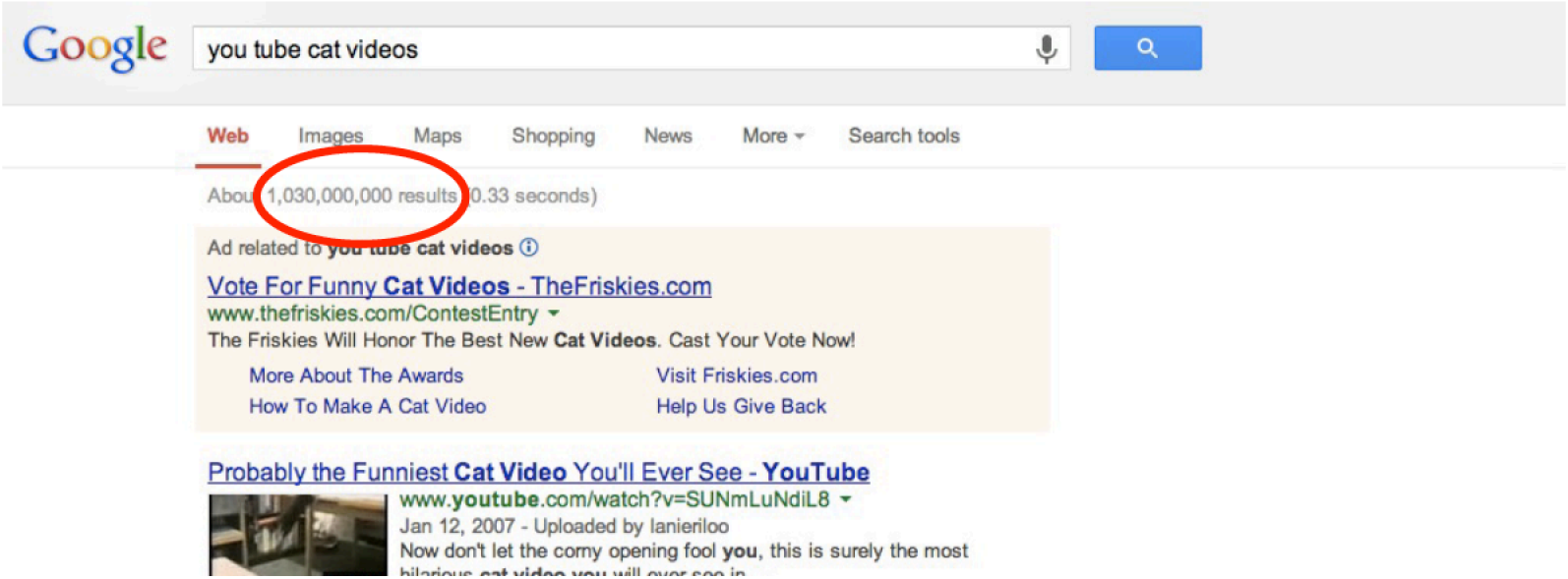
Big Data – what it means



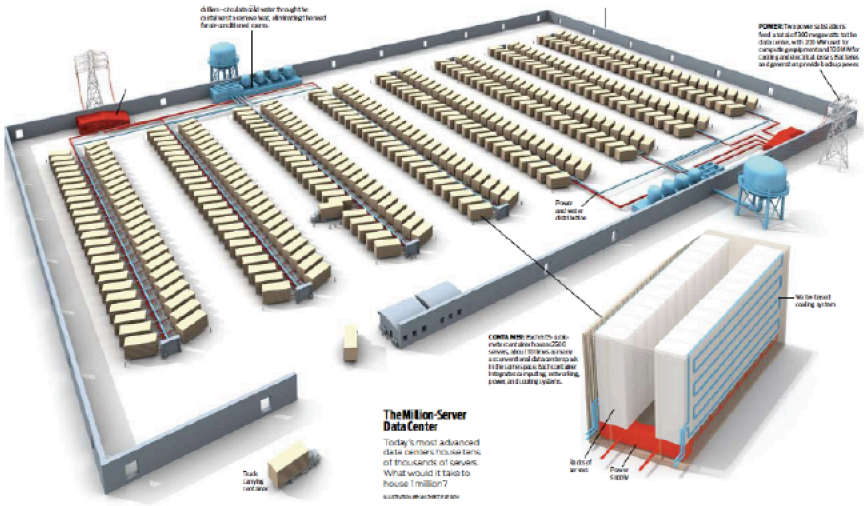
WeST
Steffen Staab
staab@uni-koblenz.de

"Between the dawn of civilization and 2003, we only created five exabytes of information; now we're creating that amount every two days."

Eric Schmidt, Google (and others)



Commodity Computing



Michael Franklin, UC Berkeley

Smarter Devices

GarageBand '08

Michael Franklin, UC Berkeley

Ubiquitous Connectivity

skype

Michael Franklin, UC Berkeley

Big Vs

- Volume
- Velocity
- Veracity
- Variety
- priVacy

CHE Survey

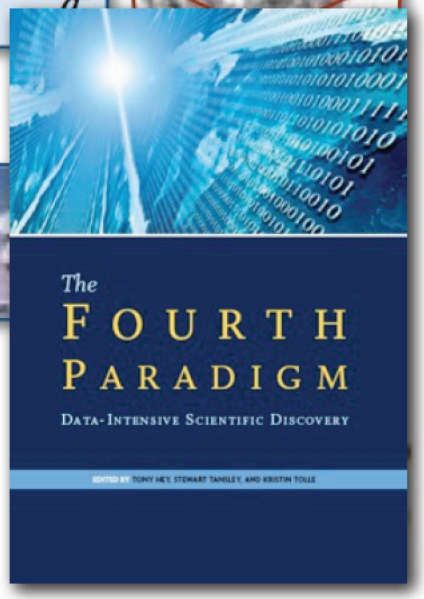
- To students of
  - Master in Informatik
  - Master in CV

- Please respond!

Science Paradigms

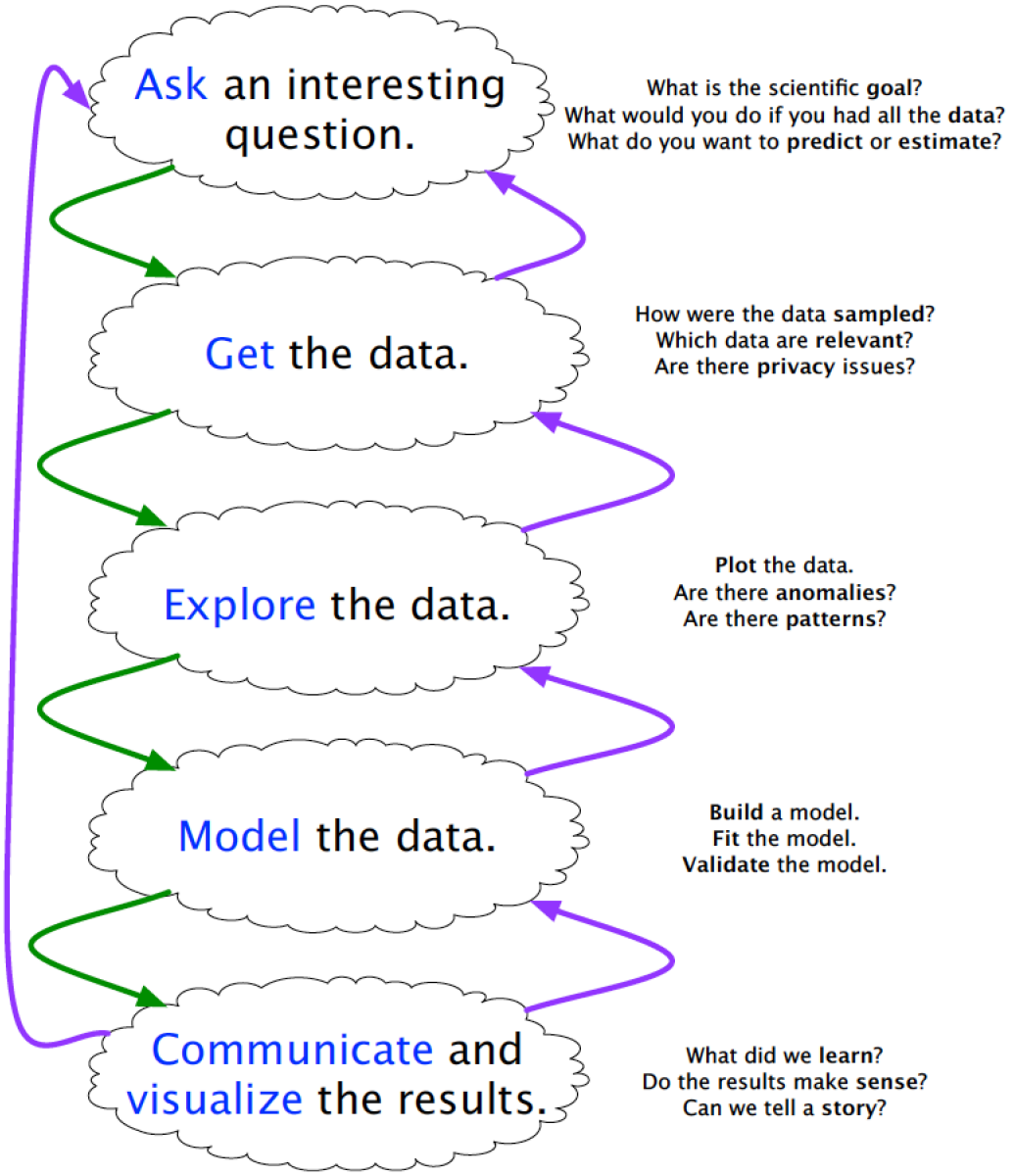
- Thousand years ago:
  science was **empirical**
  describing natural phenomena
- Last few hundred years:
  theoretical branch
  using models, generalizations
- Last few decades:
  a **computational** branch
  simulating complex phenomena
- Today: **data exploration** (eScience)
  unify theory, experiment, and simulation
  - Data captured by instruments
    or generated by simulator
  - Processed by software
  - Information/knowledge stored in computer
  - Scientist analyzes database/files
    using data management and statistics

$$\left(\frac{\dot{a}}{a}\right)^2 = \frac{4\pi G p}{3} - K\frac{c^2}{a^2}$$



Jim Gray, Microsoft

Data Science Process



Tasks

- *data munging/scraping/sampling/cleaning* in order to get an informative, manageable data set;
- *data storage and management* in order to be able to access data - especially big data - quickly and reliably during subsequent analysis;
- *exploratory data analysis* to generate hypotheses and intuition about the data;
- *prediction* based on statistical tools such as regression, classification, and clustering; and
- *communication* of results through visualization, stories, and interpretable summaries.

What do analysts do?

Enterprise Data Analysis and Visualization: An Interview Study

Sean Kandel, Andreas Paepcke, Joseph M. Hellerstein, and Jeffrey Heer

Abstract—Organizations rely on data analysts to model customer engagement, streamline operations, improve production, inform business decisions, and combat fraud. Though numerous analysis and visualization tools have been built to improve the scale and efficiency at which analysts can work, there has been little research on how analysis takes place within the social and organizational context of companies. To better understand the enterprise analysts' ecosystem, we conducted semi-structured interviews with 35 data analysts from 25 organizations across a variety of sectors, including healthcare, retail, marketing and finance. Based on our interview data, we characterize the process of industrial data analysis and document how organizational features of an enterprise impact it. We describe recurring pain points, outstanding challenges, and barriers to adoption for visual analytic tools. Finally, we discuss design implications and opportunities for visual analysis research.

Index Terms—Data, analysis, visualization, enterprise.

1 Introduction

Organizations gather increasingly large and complex data sets each year. These organizations rely on data analysis to model customer engagement, streamline operations, improve production, inform sales and business decisions, and combat fraud. Within organizations, an increasing number of individuals — with varied titles such as "business analyst", "data analyst" and "data scientist" — perform such analyses. These analysts constitute an important and rapidly growing user population for analysis and visualization tools.

Enterprise analysts perform their work within the context of a larger organization. Analysts often work as a part of an analysis team or business unit. Little research has observed how existing infrastructure, available data and tools, and administrative and social conventions within an organization impact the analysis process within the enterprise. Understanding how these issues shape analytic workflows can inform the design of future tools.

ery and wrangling, often the most tedious and time-consuming aspects of an analysis, are underserved by existing visualization and analysis tools. We discuss recurring pain points within each task as well as difficulties in managing workflows across these tasks. Example pain points include integrating data from distributed data sources, visualizing data at scale and operationalizing workflows. These challenges are typically more acute within large organizations with a diverse and distributed set of data sources.

We conclude with a discussion of future trends and the implications of our interviews for future visualization and analysis tools. We argue that future visual analysis tools should leverage existing infrastructures for data processing to enable scale and limit data migration. One avenue for achieving better interoperability is through systems that specify analysis or data processing operations in a high-level language, enabling retargeting across tools or platforms. We also note

What do analysts do?

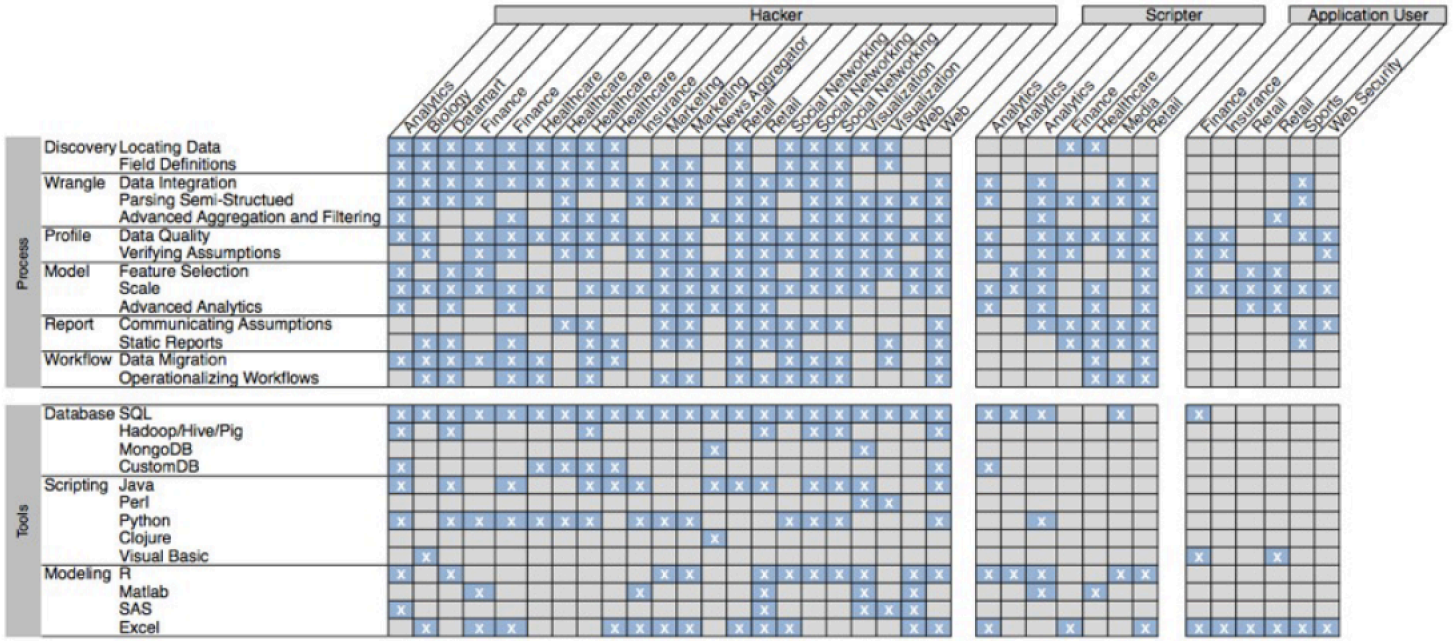


Fig. 1. Respondents, Challenges and Tools. The matrix displays interviewees (grouped by archetype and sector) and their corresponding challenges and tools. *Hackers* faced the most diverse set of challenges, corresponding to the diversity of their workflows and toolset. *Application users* and *scripters* typically relied on the IT team to perform certain tasks and therefore did not perceive them as challenges.

What do analysts do?

I spend more than half of my time integrating, cleansing and transforming data without doing any actual analysis. Most of the time I'm lucky if I get to do any analysis. Most of the time once you transform the data you just do an average... the insights can be scarily obvious. It's fun when you get to do something somewhat analytical

BE CAREFUL WITH STATISTICS

Ascombe's Quartet

Same mean, variance, correlation, and linear regression line

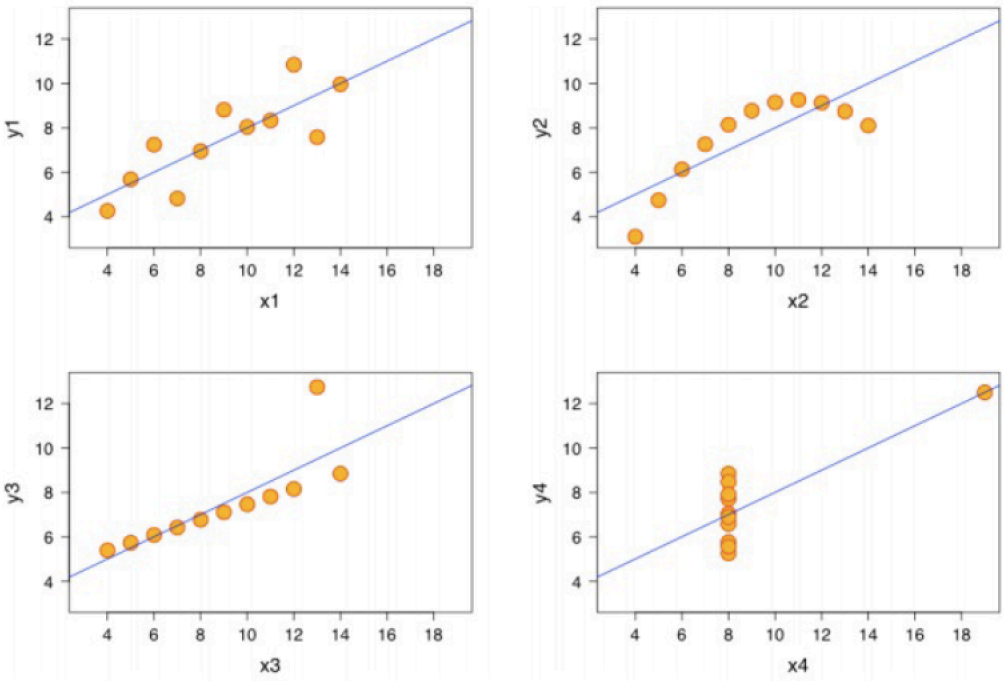
Anscombe's Quartet: Raw Data

	I		II		III		IV	
	x	y	x	y	x	y	x	y
	10.0	8.04	10.0	9.14	10.0	7.46	8.0	6.58
	8.0	6.95	8.0	8.14	8.0	6.77	8.0	5.76
	13.0	7.58	13.0	8.74	13.0	12.74	8.0	7.71
	9.0	8.81	9.0	8.77	9.0	7.11	8.0	8.84
	11.0	8.33	11.0	9.26	11.0	7.81	8.0	8.47
	14.0	9.96	14.0	8.10	14.0	8.84	8.0	7.04
	6.0	7.24	6.0	6.13	6.0	6.08	8.0	5.25
	4.0	4.26	4.0	3.10	4.0	5.39	19.0	12.50
	12.0	10.84	12.0	9.13	12.0	8.15	8.0	5.56
	7.0	4.82	7.0	7.26	7.0	6.42	8.0	7.91
	5.0	5.68	5.0	4.74	5.0	5.73	8.0	6.89
mean	9.0	7.5	9.0	7.5	9.0	7.5	9.0	7.5
var.	10.0	3.75	10.0	3.75	10.0	3.75	10.0	3.75
corr.	0.816		0.816		0.816		0.816	

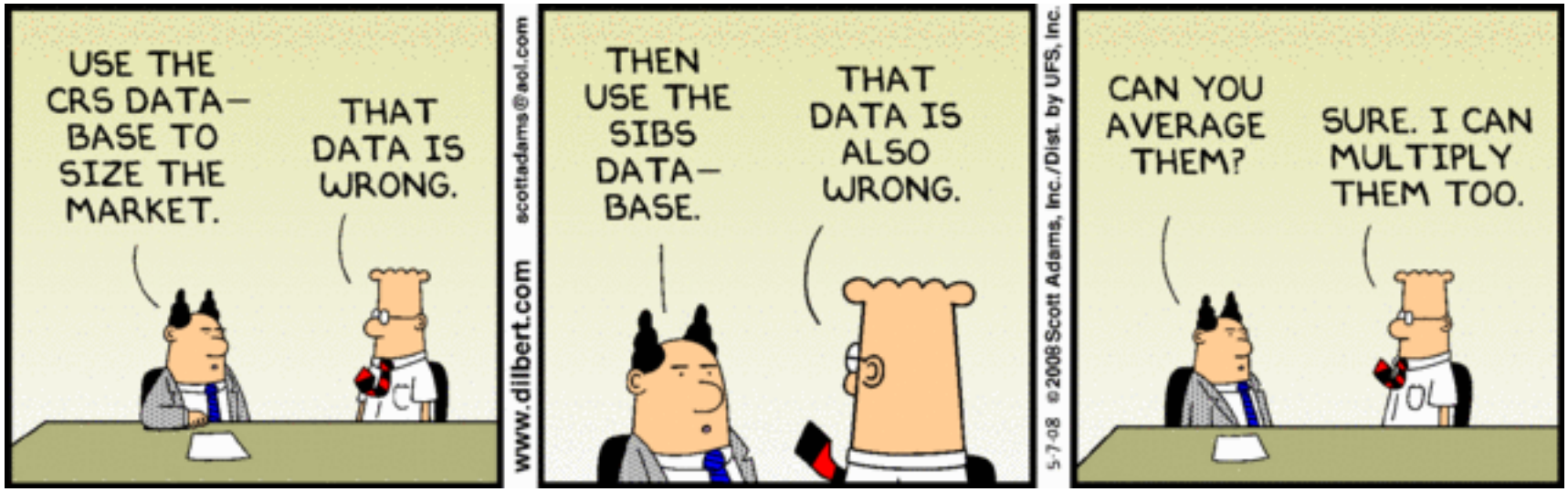
Anscombe '73

Ascombe's Quartet

Same mean, variance, correlation, and linear regression line



Anscombe '73


USE THE CRS DATA-BASE TO SIZE THE MARKET.
THAT DATA IS WRONG.
www.dilbert.com scottadams@aol.com
THEN USE THE SIBS DATA-BASE.
THAT DATA IS ALSO WRONG.
5-7-08 ©2008 Scott Adams, Inc./Dist. by UFS, Inc.
CAN YOU AVERAGE THEM?
SURE. I CAN MULTIPLY THEM TOO.

Example: Antibiotics (Will Burtin, 1951)

TELLING A STORY

Effectiveness of Antibiotics

Data & Questions

What are the data types?
What are possible questions?

Bacteria	Penicillin	Antibiotic Streptomycin	Neomycin	Gram stain
Aerobacter aerogenes	870	1	1.6	–
Brucella abortus	1	2	0.02	–
Bacillus anthracis	0.001	0.01	0.007	+
Diplococcus pneumoniae	0.005	11	10	+
Escherichia coli	100	0.4	0.1	–
Klebsiella pneumoniae	850	1.2	1	–
Mycobacterium tuberculosis	800	5	2	–
Proteus vulgaris	3	0.1	0.1	–
Pseudomonas aeruginosa	850	2	0.4	–
Salmonella (Eberthella) typhosa	1	0.4	0.008	–
Salmonella schottmuelleri	10	0.8	0.09	–
Staphylococcus albus	0.007	0.1	0.001	+
Staphylococcus aureus	0.03	0.03	0.001	+
Streptococcus fecalis	1	1	0.1	+
Streptococcus hemolyticus	0.001	14	10	+
Streptococcus viridans	0.005	10	40	+

https://www.americanscientist.org/issues/pub/2009/4/thats-funny

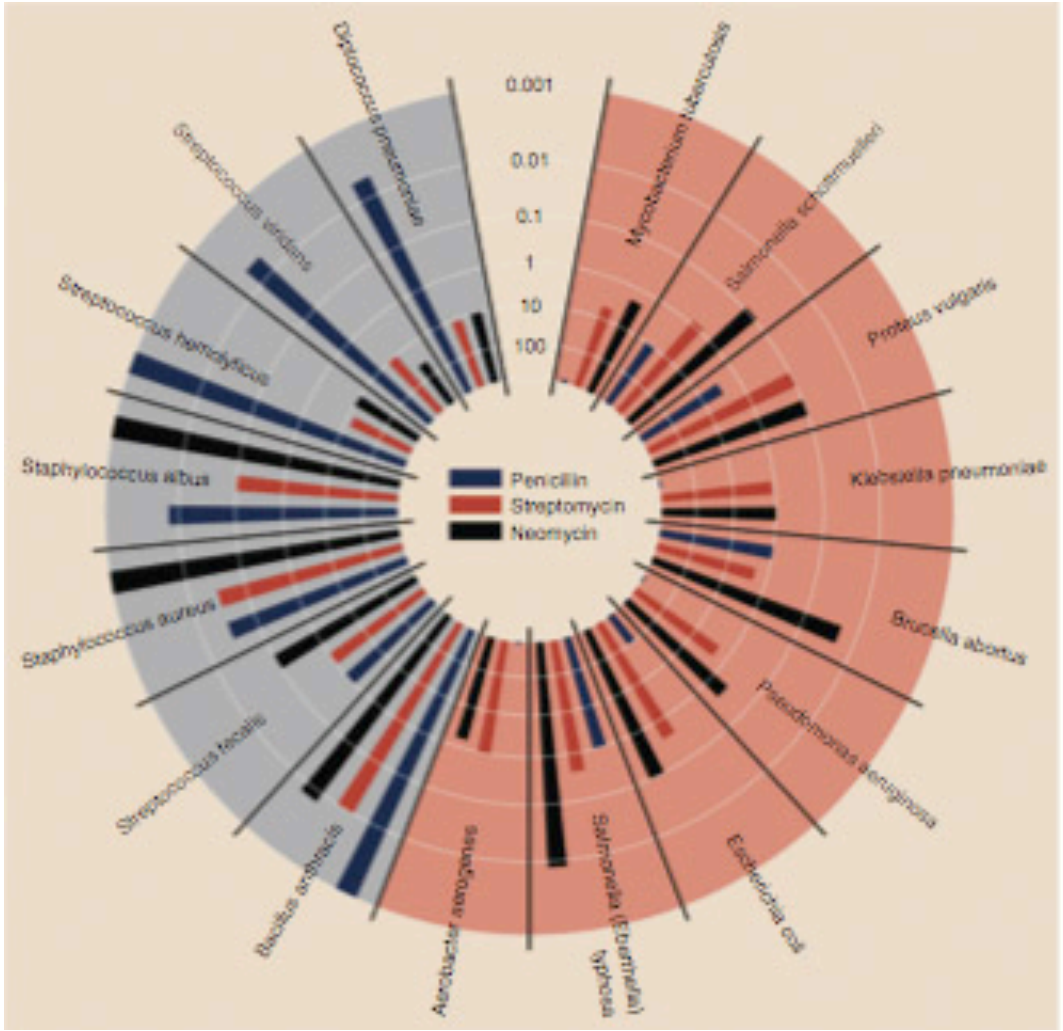
Data

- Genus & species of bacteria [string]
- Antibiotic name [string]
- Gram staining? [pos/neg]
- Minimum inhibitory concentration (mg/ml) [float] (lower == more effective)

Bacteria	Penicillin	Antibiotic Streptomycin	Neomycin	Gram stain
Aerobacter aerogenes	870	1	1.6	–
Brucella abortus	1	2	0.02	–
Bacillus anthracis	0.001	0.01	0.007	+
Diplococcus pneumoniae	0.005	11	10	+
Escherichia coli	100	0.4	0.1	–
Klebsiella pneumoniae	850	1.2	1	–
Mycobacterium tuberculosis	800	5	2	–
Proteus vulgaris	3	0.1	0.1	–
Pseudomonas aeruginosa	850	2	0.4	–
Salmonella (Eberthella) typhosa	1	0.4	0.008	–
Salmonella schottmuelleri	10	0.8	0.09	–
Staphylococcus albus	0.007	0.1	0.001	+
Staphylococcus aureus	0.03	0.03	0.001	+
Streptococcus fecalis	1	1	0.1	+
Streptococcus hemolyticus	0.001	14	10	+
Streptococcus viridans	0.005	10	40	+

Steffen Staab
staab@uni-koblenz.de

How effective are the drugs?

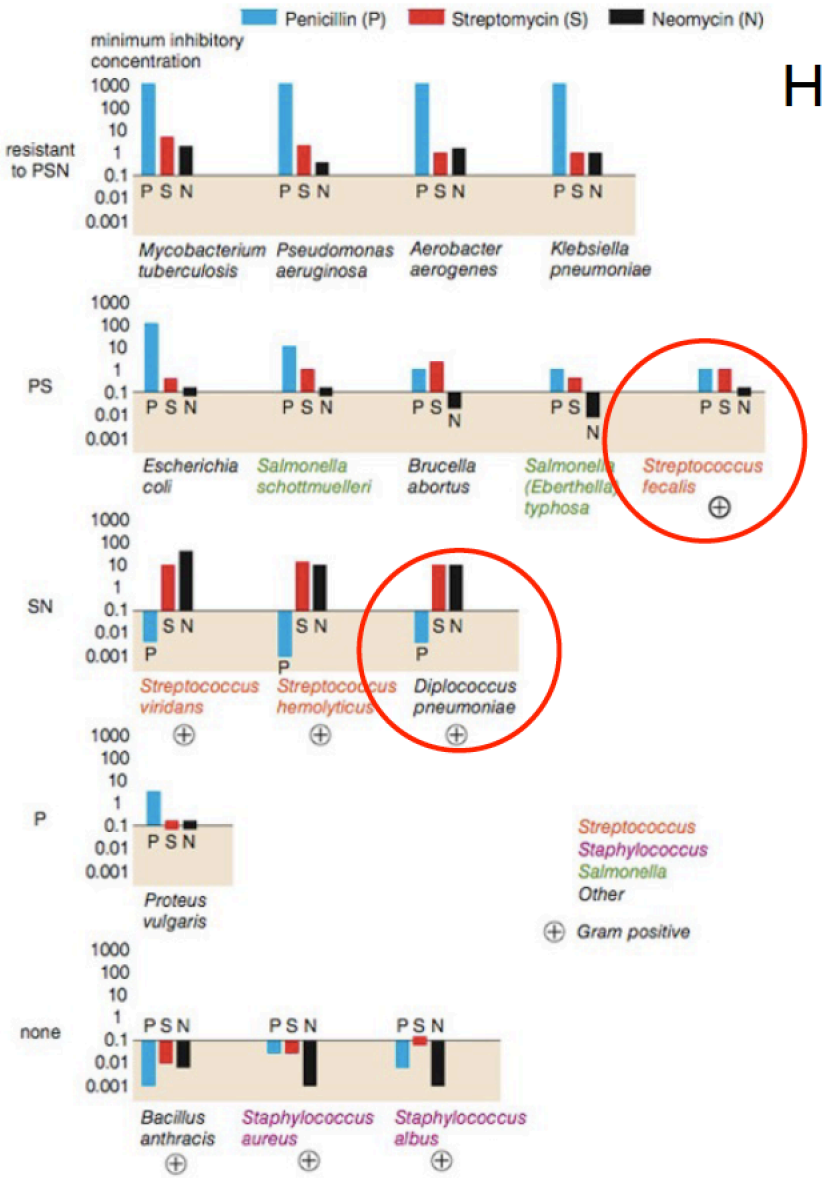


Not answered:

Which bacteria behave likewise?

https://www.americanscientist.org/issues/pub/2009/4/thats-funny

How do the bacteria compare?

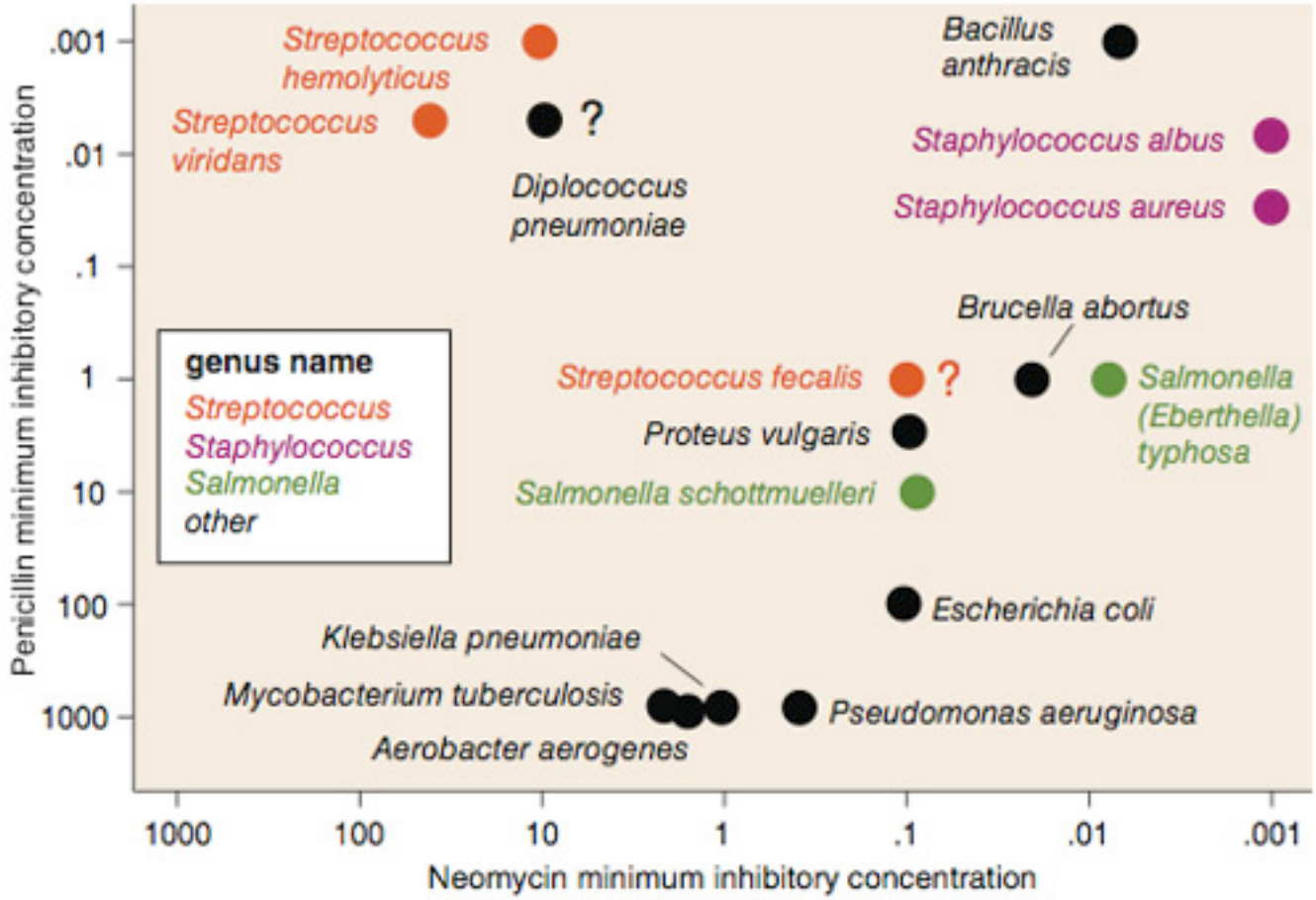


How do the bacteria compare?

Not a streptococcus!
(realized ~30 years later)

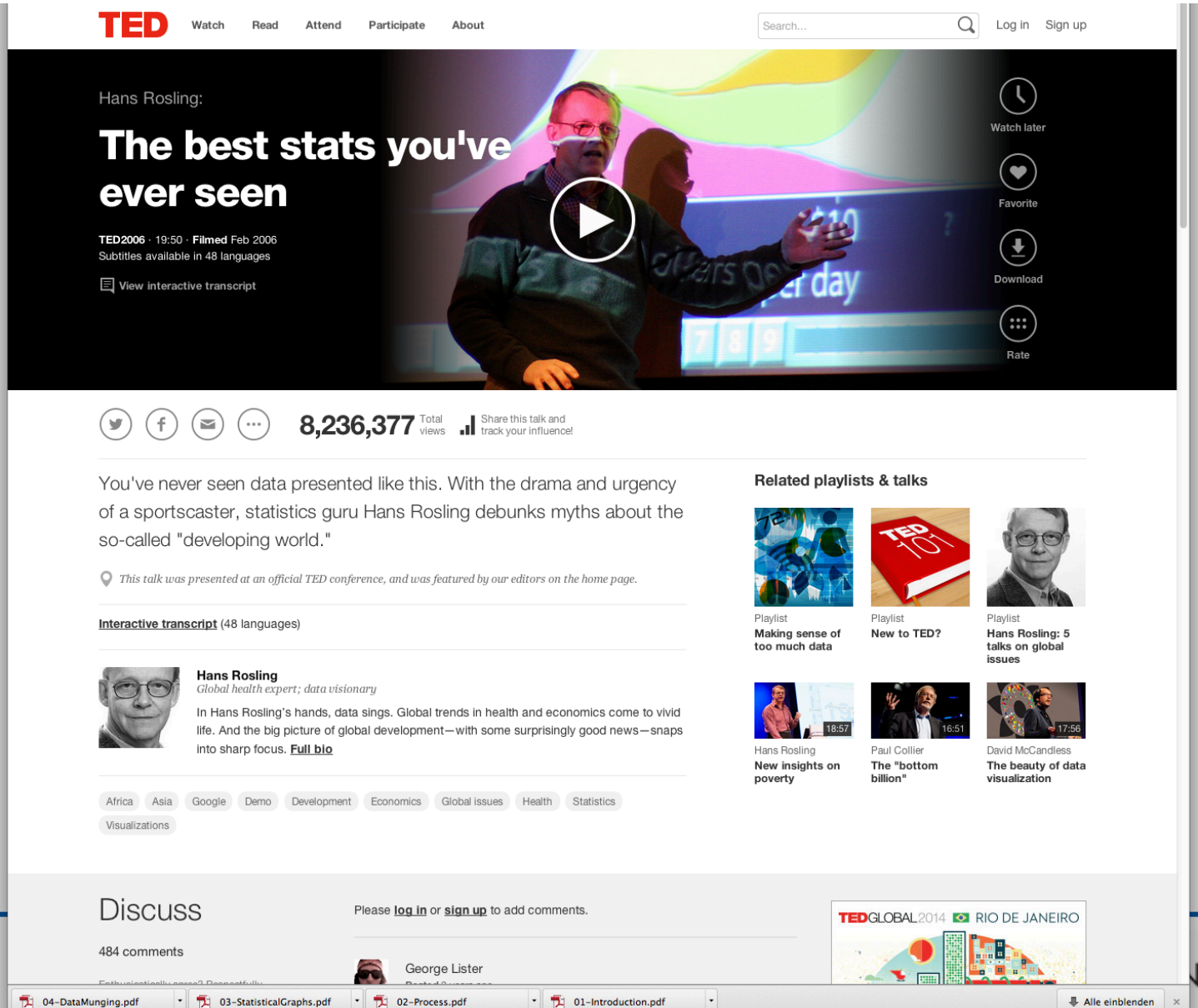
Really a streptococcus!
(realized ~20 years later)

Wainer & Lysen, "That's funny..."
American Scientist, 2009
Adapted from Brian Schmotzer


Penicillin minimum inhibitory concentration
.001
.01
.1
1
10
100
1000
Streptococcus hemolyticus
Streptococcus viridans
?
Diplococcus pneumoniae
Bacillus anthracis
Staphylococcus albus
Staphylococcus aureus
Brucella abortus
Streptococcus fecalis
?
Proteus vulgaris
Salmonella schottmuelleri
Salmonella (Eberthella) typhosa
genus name
Streptococcus
Staphylococcus
Salmonella
other
Escherichia coli
Klebsiella pneumoniae
Mycobacterium tuberculosis
Aerobacter aerogenes
Pseudomonas aeruginosa
1000
100
10
1
.1
.01
.001
Neomycin minimum inhibitory concentration

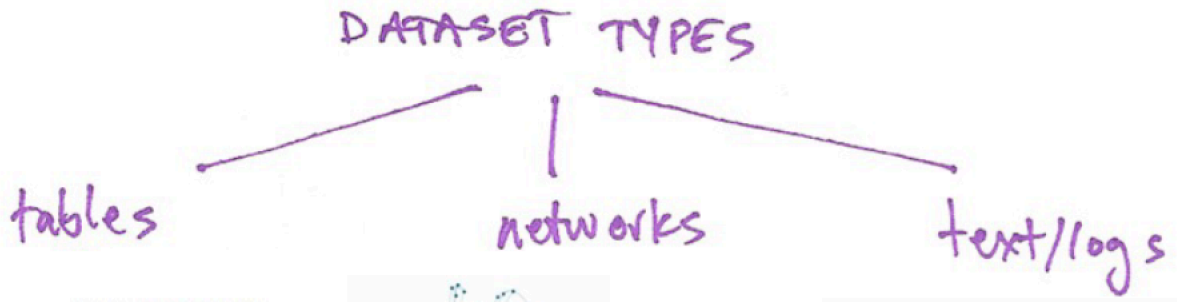
Ted Talks by Rosling

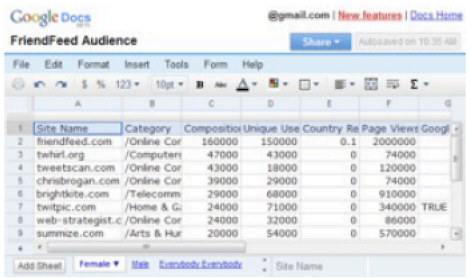
http://www.ted.com/talks/hans_rosling_shows_the_best_stats_you_ve_ever_seen

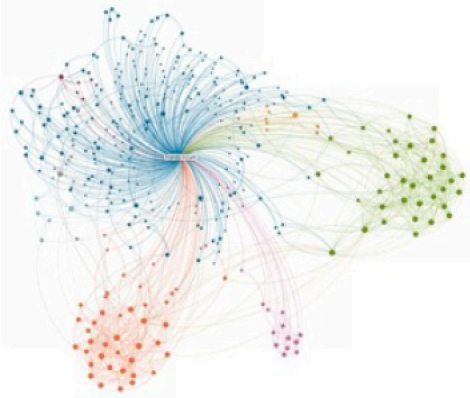


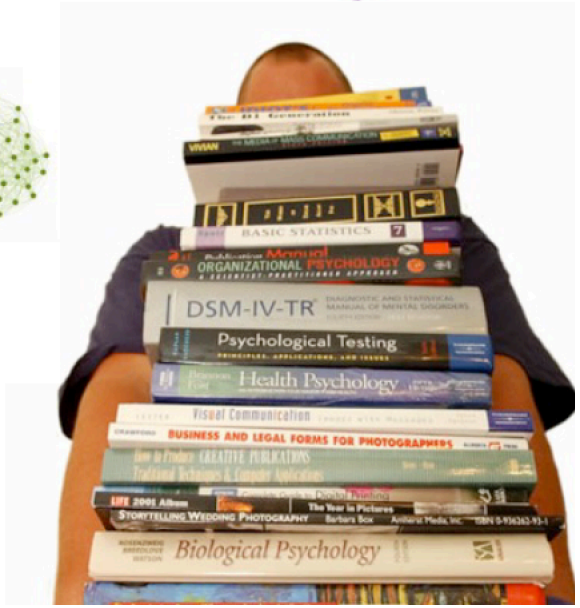
LOOKING AT DATA

Tamara Munzner, 2013









Semantics vs. Types

- **Data Semantics**: The real-world meaning
  - e.g., company name, day of the month, person height, etc.

- Data Type: Interpretation in terms of scales of measurements
  - e.g., quantity or category, sensible mathematical operations, data structure, etc

Data Types

- Nominal (Categorical) (N)
  - Are = or ≠ to other values
  - Apples, Oranges, Bananas,...
- Ordinal (O)
  - Obey a < relationship
  - Small, medium, large
- Quantitative (Q)
  - Can do arithmetic on them
  - 10 inches, 23 inches, etc.

On the theory of scales and measurements [S. Stevens, 46]

Steffen Staab
staab@uni-koblenz.de

Data Types

- Q - Interval (location of zero arbitrary)

  Dates: Jan 19; Location: (Lat, Long)

  Like a geometric point. Cannot compare directly.

  Only differences (i.e., intervals) can be compared

- Q - Ratio (zero fixed)

  Measurements: Length, Mass, Temp, ...

  Origin is meaningful, can measure ratios & proportions

  Like a geometric vector, origin is meaningful

On the theory of scales and measurements [S. Stevens, 46]

Data Types

- N - Nominal (labels)

  Operations: =, ≠
- O - Ordinal (ordered)

  Operations: =, ≠, >, <
- Q - Interval (location of zero arbitrary)

  Operations: =, ≠, >, <, +, − (distance)
- Q - Ratio (zero fixed)

  Operations: =, ≠, >, <, +, −, ×, ÷ (proportions)

On the theory of scales and measurements [S. Stevens, 46]

Examples

Nominal

Ordinal

Interval

Ration

Operations on

- Nominals: x=y, x≠y
- Ordinals: x=y, x≠y, x<y, x>y
- Interval scales: Ration scales (also):

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

$$s_x^2 = \frac{1}{n}\sum_{i=1}^{n} (x_i - \bar{x})^2$$

$$\Delta x_t = x_t - x_{t-1}$$

$$\frac{\Delta y}{\Delta_x} = \frac{y_{11} - y_{21}}{x_{11} - x_{21}}$$

Scale types with their properties according to Stanley Smith Stevens				
	Nominal scale	**Ordinal** scale	**Interval** scale	**Ratio** scale
Logical/ math operations × ÷	✗	✗	✗	✓
Logical/ math operations + −	✗	✗	✓	✓
Logical/ math operations < >	✗	✓	✓	✓
Logical/ math operations = ≠	✓	✓	✓	✓
Examples: *Dichotomous and non-dichotomous* Variable name (data values)	*Dichotomous:* Gender (male vs. female) *Non-dichotomous:* Nationality (American/Chinese/etc)	*Dichotomous:* Health (healthy vs. sick), Truth (true vs. false), Beauty (beautiful vs.ugly) *Non-dichotomous:* Opinion ('completely agree'/ 'mostly agree'/ 'mostly disagree'/ 'completely disagree')	Date (from 1457 BC to AD 2013) Latitude (from +90° to −90°) Temperature in Celsius degrees (from 10°C to 20°C)	Age (from 0 to 99 years) Temperature in Kelvin (from 10K to 20K)
Measure of central tendency	Mode	Median Mode	Arithmetic Mean Median Mode	Geometric Mean Arithmetic Mean Median Mode
Qualitative or Quantitative	Qualitative	Qualitative	Quantitative	Quantitative

https://en.wikipedia.org/wiki/Level_of_measurement

A	B	C	S	T	U
Order ID	Order Date	Order Priority	Product Container	Product Base Margin	Ship Date
3	10/14/06	5-Low	Large Box	0.8	10/21/06
6	2/21/08	4-Not Specified	Small Pack		2/22/08
32	7/16/07	2-High	Small Pack		7/17/07
32	7/16/07	2-High	Jumbo Box		7/17/07
32	7/16/07	2-High	Medium Box	0.6	7/18/07
32	7/16/07	2-High	Medium Box	0.65	7/18/07
35	10/23/07	4-Not Specified	Wrap Bag	0.52	10/24/07
35	10/23/07	4-Not Specified	Small Box	0.58	10/25/07
36	11/3/07	1-Urgent	Small Box	0.55	11/3/07
65	3/18/07	1-Urgent	Small Pack	0.49	3/19/07
66	1/20/05	5-Low	Wrap Bag	0.56	1/20/05
69	6/4/05	4-Not Specified	Small Pack	0.44	6/6/05
69	6/4/05	4-Not Specified	Wrap Bag	0.6	6/6/05
70	12/18/06	5-Low	Small Box	0.59	12/23/06
70	12/18/06	5-Low	Wrap Bag	0.82	12/23/06
96	4/17/05	2-High	Small Box	0.55	4/19/05
97	1/29/06	3-Medium	Small Box	0.38	1/30/06
129	11/19/08	5-Low	Small Box	0.37	11/28/08
130	5/8/08	2-High	Small Box	0.37	5/9/08
130	5/8/08	2-High	Medium Box	0.38	5/10/08
130	5/8/08	2-High	Small Box	0.6	5/11/08
132	6/11/06	3-Medium	Medium Box	0.6	6/12/06
132	6/11/06	3-Medium	Jumbo Box	0.69	6/14/06
134	5/1/08	4-Not Specified	Large Box	0.82	5/3/08
135	10/21/07	4-Not Specified	Small Pack	0.64	10/23/07
166	9/12/07	2-High	Small Box	0.55	9/14/07
193	8/8/06	1-Urgent	Medium Box	0.57	8/10/06
194	4/5/08	3-Medium	Wrap Bag	0.42	4/7/08

Semantics

A	B	C	S	T	U
Order ID	Order Date	Order Priority	Product Container	Product Base Margin	Ship Date
3	10/14/06	5-Low	Large Box	0.8	10/21/06
6	2/21/08	4-Not Specified	Small Pack	0.55	2/22/08
32	7/16/07	2-High	Small Pack	0.79	7/17/07
32	7/16/07	2-High	Jumbo Box		7/17/07
32	7/16/07	2-High	Medium Box		7/18/07
32	7/16/07	2-High	Medium Box	0.65	7/18/07
35	10/23/07	4-Not Specified	Wrap Bag	0.52	10/24/07
35	10/23/07	4-Not Specified	Small Box	0.58	10/25/07
36	11/3/07	1-Urgent	Small Box	0.55	11/3/07
65	3/18/07	1-Urgent	Small Pack	0.49	3/19/07
66	1/20/05	5-Low	Wrap Bag	0.56	1/20/05
69	6/4/05	4-Not Specified	Small Pack	0.44	6/6/05
69	6/4/05	4-Not Specified	Wrap Bag	0.6	6/6/05
70	12/18/06	5-Low	Small Box	0.59	12/23/06
70	12/18/06	5-Low	Wrap Bag	0.82	12/23/06
96	4/17/05	2-High	Small Box	0.55	4/19/05
97	1/29/06	3-Medium	Small Box	0.38	1/30/06
129	11/19/08	5-Low	Small Box	0.37	11/28/08
130	5/8/08	2-High	Small Box	0.37	5/9/08
130	5/8/08	2-High	Medium Box	0.38	5/10/08
130	5/8/08	2-High	Small Box	0.6	5/11/08
132	6/11/06	3-Medium	Medium Box	0.6	6/12/06
132	6/11/06	3-Medium	Jumbo Box	0.69	6/14/06
134	5/1/08	4-Not Specified	Large Box	0.82	5/3/08
135	10/21/07	4-Not Specified	Small Pack	0.64	10/23/07
166	9/12/07	2-High	Small Box	0.55	9/14/07
193	8/8/06	1-Urgent	Medium Box	0.57	8/10/06
194	4/5/08	3-Medium	Wrap Bag	0.42	4/7/08

Item

A	B	C	S	T	U
Order ID	Order Date	Order Priority	Product Container	Product Base Margin	Ship Date
3	10/14/06	5-Low	Large Box	0.8	10/21/06
6	2/21/08	4-Not Specified	Small Pack		2/22/08
32	7/16/07	2-High	Small Pack		7/17/07
32	7/16/07	2-High	Jumbo Box		7/17/07
32	7/16/07	2-High	Medium Box		7/18/07
32	7/16/07	2-High	Medium Box		7/18/07
35	10/23/07	4-Not Specified	Wrap Bag		10/24/07
35	10/23/07	4-Not Specified	Small Box	0.58	10/25/07
36	11/3/07	1-Urgent	Small Box	0.55	11/3/07
65	3/18/07	1-Urgent	Small Pack	0.49	3/19/07
66	1/20/05	5-Low	Wrap Bag	0.56	1/20/05
69	6/4/05	4-Not Specified	Small Pack	0.44	6/6/05
69	6/4/05	4-Not Specified	Wrap Bag	0.6	6/6/05
70	12/18/06	5-Low	Small Box	0.59	12/23/06
70	12/18/06	5-Low	Wrap Bag	0.82	12/23/06
96	4/17/05	2-High	Small Box	0.55	4/19/05
97	1/29/06	3-Medium	Small Box	0.38	1/30/06
129	11/19/08	5-Low	Small Box	0.37	11/28/08
130	5/8/08	2-High	Small Box	0.37	5/9/08
130	5/8/08	2-High	Medium Box	0.38	5/10/08
130	5/8/08	2-High	Small Box	0.6	5/11/08
132	6/11/06	3-Medium	Medium Box	0.6	6/12/06
132	6/11/06	3-Medium	Jumbo Box	0.69	6/14/06
134	5/1/08	4-Not Specified	Large Box	0.82	5/3/08
135	10/21/07	4-Not Specified	Small Pack	0.64	10/23/07
166	9/12/07	2-High	Small Box	0.55	9/14/07
193	8/8/06	1-Urgent	Medium Box	0.57	8/10/06
194	4/5/08	3-Medium	Wrap Bag	0.42	4/7/08

Attribute aka Feature

A	B	C	S	T	U
Order ID	Order Date	Order Priority	Product Container	Product Base Margin	Ship Date
3	10/14/06	5-Low	Large Box	0.8	10/21/06
6	2/21/08	4-Not Specified	Small Pack	0.55	2/22/08
32	7/16/07	2-High	Small Pack	0.79	7/17/07
32	7/16/07	2-High	Jumbo Box	0.72	7/17/07
32	7/16/07	2-High	Medium Box	0.6	7/18/07
32	7/16/07	2-High	Medium Box	0.65	7/18/07
35	10/23/07	4-Not Specified	Wrap Bag	0.52	10/24/07
35	10/23/07	4-Not Specified	Small Box	0.58	10/25/07
36	11/3/07	1-Urgent	Small Box	0.55	11/3/07
65	3/18/07	1-Urgent	Small Pack	0.49	3/19/07
66	1/20/05	5-Low	Wrap Bag	0.56	1/20/05
69	6/4/05	4-Not Specified	Small Pack	0.44	6/6/05
69	6/4/05	4-Not Speci		0.6	6/6/05
70	12/18/06	5-Low		0.59	12/23/06
70	12/18/06	5-Low		0.82	12/23/06
96	4/17/05	2-High		0.55	4/19/05
97	1/29/06	3-Medium		0.38	1/30/06
129	11/19/08	5-Low	Small Box	0.37	11/28/08
130	5/8/08	2-High	Small Box	0.37	5/9/08
130	5/8/08	2-High	Medium Box	0.38	5/10/08
130	5/8/08	2-High	Small Box	0.6	5/11/08
132	6/11/06	3-Medium	Medium Box	0.6	6/12/06
132	6/11/06	3-Medium	Jumbo Box	0.69	6/14/06
134	5/1/08	4-Not Specified	Large Box	0.82	5/3/08
135	10/21/07	4-Not Specified	Small Pack	0.64	10/23/07
166	9/12/07	2-High	Small Box	0.55	9/14/07
193	8/8/06	1-Urgent	Medium Box	0.57	8/10/06
194	4/5/08	3-Medium	Wrap Bag	0.42	4/7/08

1 = Quantitative
2 = Nominal
3 = Ordinal

A	B	C	S	T	U
Order ID	Order Date	Order Priority	Product Container	Product Base Margin	Ship Date
3	10/14/06	5-Low	Large Box	0.8	10/21/06
6	2/21/08	4-Not Specified	Small Pack	0.55	2/22/08
32	7/16/07	2-High	Small Pack	0.79	7/17/07
32	7/16/07	2-High	Jumbo Box	0.72	7/17/07
32	7/16/07	2-High	Medium Box	0.6	7/18/07
32	7/16/07	2-High	Medium Box	0.65	7/18/07
35	10/23/07	4-Not Specified	Wrap Bag	0.52	10/24/07
35	10/23/07	4-Not Specified	Small Box	0.58	10/25/07
36	11/3/07	1-Urgent	Small Box	0.55	11/3/07
65	3/18/07	1-Urgent	Small Pack	0.49	3/19/07
66	1/20/05	5-Low	Wrap Bag	0.56	1/20/05
69	6/4/05	4-Not Specified	Small Pack	0.44	6/6/05
69	6/4/05	4-Not Speci		0.6	6/6/05
70	12/18/06	5-Low		0.59	12/23/06
70	12/18/06	5-Low		0.82	12/23/06
96	4/17/05	2-High		0.55	4/19/05
97	1/29/06	3-Medium		0.38	1/30/06
129	11/19/08	5-Low	Small Box	0.37	11/28/08
130	5/8/08	2-High	Small Box	0.37	5/9/08
130	5/8/08	2-High	Medium Box	0.38	5/10/08
130	5/8/08	2-High	Small Box	0.6	5/11/08
132	6/11/06	3-Medium	Medium Box	0.6	6/12/06
132	6/11/06	3-Medium	Jumbo Box	0.69	6/14/06
134	5/1/08	4-Not Specified	Large Box	0.82	5/3/08
135	10/21/07	4-Not Specified	Small Pack	0.64	10/23/07
166	9/12/07	2-High	Small Box	0.55	9/14/07
193	8/8/06	1-Urgent	Medium Box	0.57	8/10/06
194	4/5/08	3-Medium	Wrap Bag	0.42	4/7/08

1 = Quantitative
2 = Nominal
3 = Ordinal

A	B	C	S	T	U
Order ID	Order Date	Order Priority	Product Container	Product Base Margin	Ship Date
3	10/14/06	5-Low	Large Box	0.8	10/21/06
6	2/21/08	4-Not Specified	Small Pack	0.55	2/22/08
32	7/16/07	2-High	Small Pack	0.79	7/17/07
32	7/16/07	2-High	Jumbo Box	0.72	7/17/07
32	7/16/07	2-High	Medium Box	0.6	7/18/07
32	7/16/07	2-High	Medium Box	0.65	7/18/07
35	10/23/07	4-Not Specified	Wrap Bag	0.52	10/24/07
35	10/23/07	4-Not Specified	Small Box	0.58	10/25/07
36	11/3/07	1-Urgent	Small Box	0.55	11/3/07
65	3/18/07	1-Urgent	Small Pack	0.49	3/19/07
66	1/20/05	5-Low	Wrap Bag	0.56	1/20/05
69	6/4/05			44	6/6/05
69	6/4/05			.6	6/6/05
70	12/18/06			59	12/23/06
70	12/18/06			82	12/23/06
96	4/17/05			55	4/19/05
97	1/29/06			38	1/30/06
129	11/19/08			37	11/28/08
130	5/8/08	2-High	Small Box	0.37	5/9/08
130	5/8/08	2-High	Medium Box	0.38	5/10/08
130	5/8/08	2-High	Small Box	0.6	5/11/08
132	6/11/06	3-Medium	Medium Box	0.6	6/12/06
132	6/11/06	3-Medium	Jumbo Box	0.69	6/14/06
134	5/1/08	4-Not Specified	Large Box	0.82	5/3/08
135	10/21/07	4-Not Specified	Small Pack	0.64	10/23/07
166	9/12/07	2-High	Small Box	0.55	9/14/07
193	8/8/06	1-Urgent	Medium Box	0.57	8/10/06
194	4/5/08	3-Medium	Wrap Bag	0.42	4/7/08
194	4/5/08	3-Medium	Wrap Bag	0.84	4/7/08

Nominal /Ordinal = Dimensions

Describe the data, independent variables

Quantitative = Measures

Numbers to be analyzed, dependent variables

Data vs. Conceptual Model

- Data Model: Low-level description of the data

  Set with operations, e.g., floats with +, -, /, *

- Conceptual Model: Mental construction

  Includes semantics, supports reasoning

Data	Conceptual
1D floats	temperature
3D vector of floats	space

Data vs. Conceptual Model

- From data model...

  32.5, 54.0, -17.3, … (floats)

- using conceptual model...

  Temperature

- to data type

  Continuous to 4 significant figures (Q)

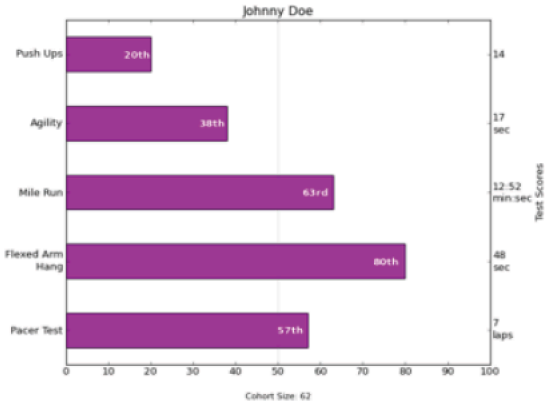
  Hot, warm, cold (O)

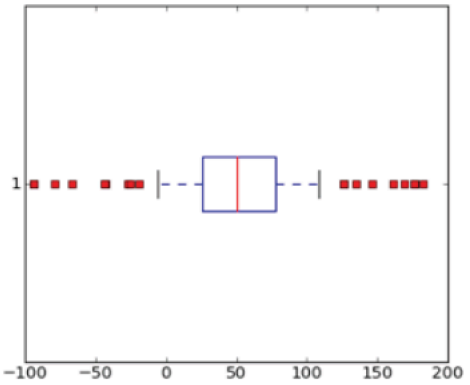
  Burned vs. Not burned (N)

Based on slide from Munzner

DATA DIMENSIONS

Univariate Data

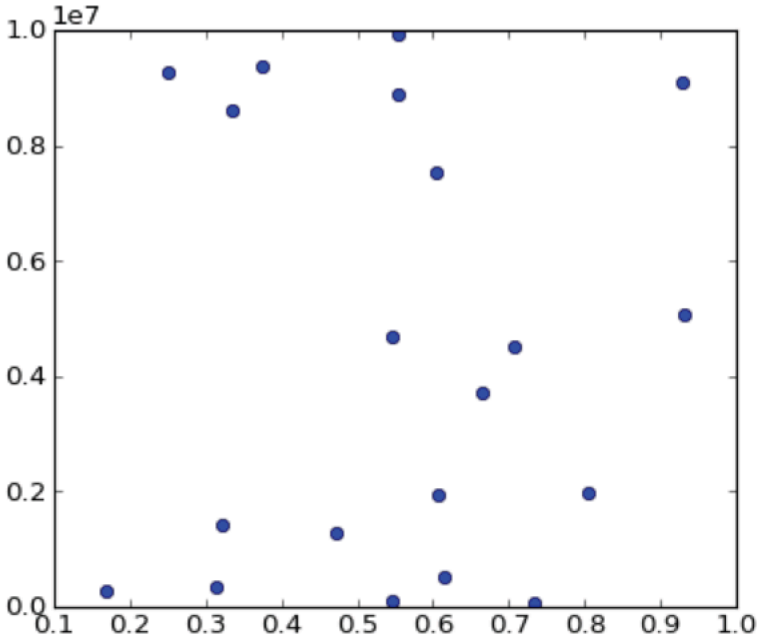




Based on slide from M. Agrawala

Bivariate Data

Scatterplot is common

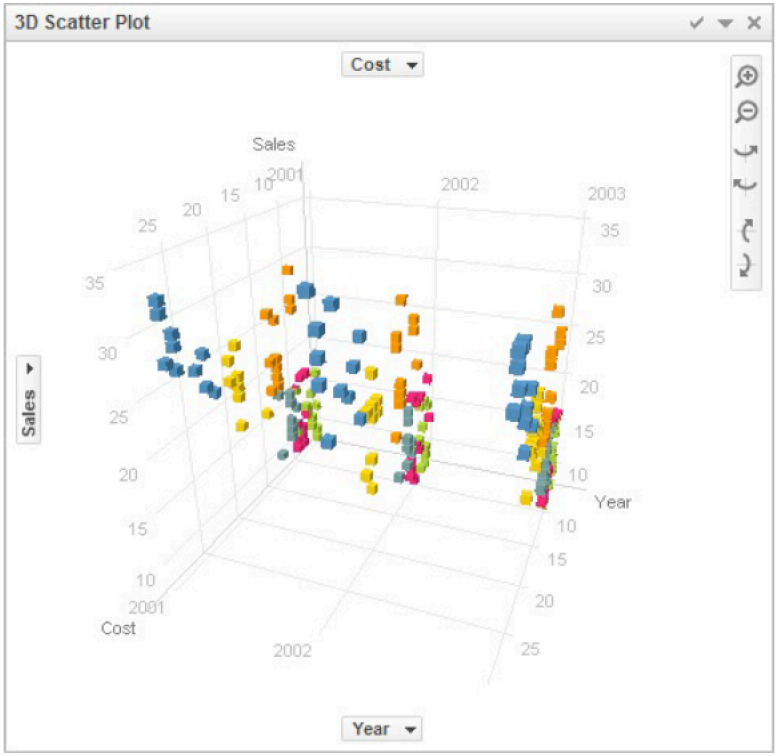


Not a function in general!

Based on slide from M. Agrawala

Trivariate Data

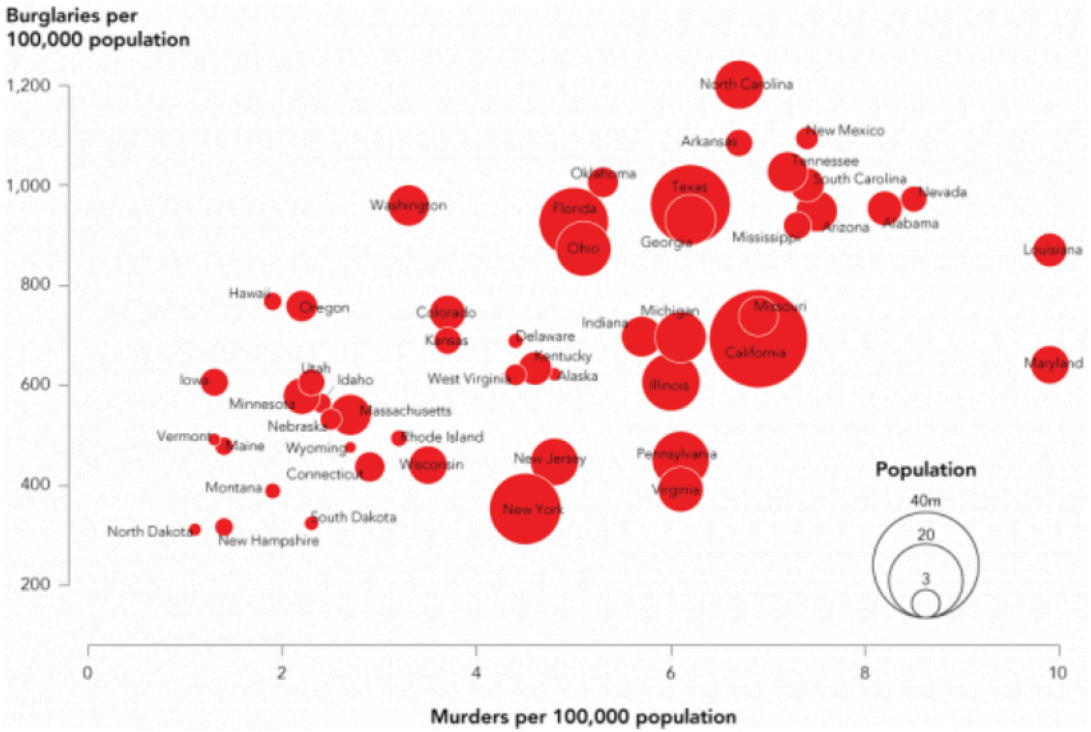
Do NOT use 3D scatterplots!



Based on slide from M.Agrawala

Trivariate Data

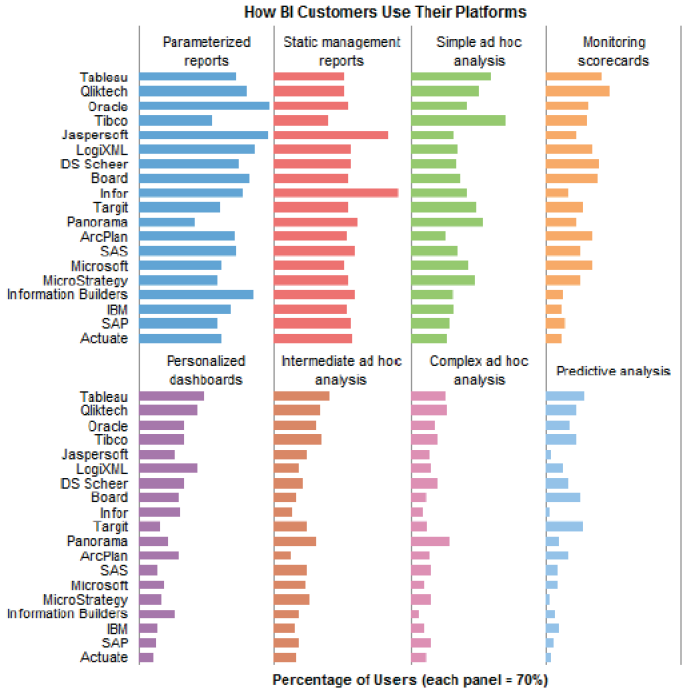
Map the third dimension to some other visual attribute



Based on slide from M.Agrawala

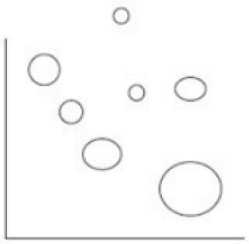
Multivariate Data

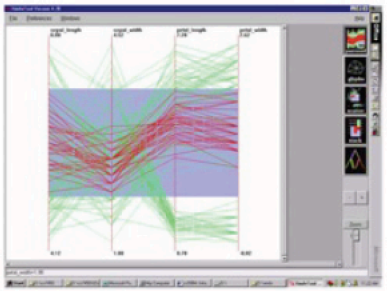
Give each attribute its own display (small multiples)

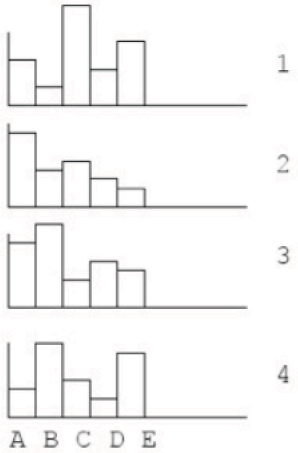


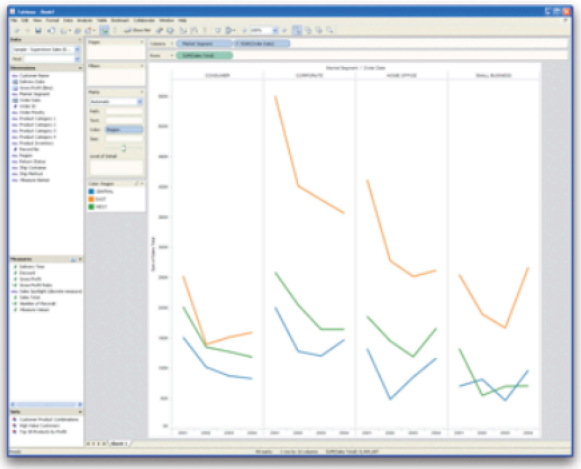
Based on slide from M. Agrawala

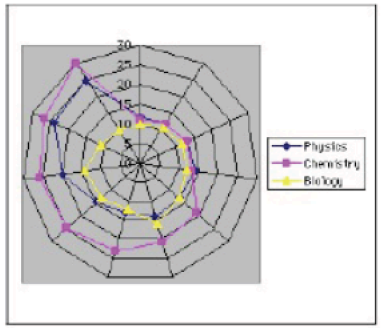
Multivariate Data Representations











Data Reduction

- Filtering: Eliminate some items or attributes
  - e.g., select range of interest, zoom in, remove outliers, etc.
- Aggregation: Represent a group of elements by a new derived element
  - e.g., take average, min, max, count, sum
  - Attribute aggregation a.k.a. dimensionality reduction