ECE20008-01 Project Practice 1

Big Data Analytics

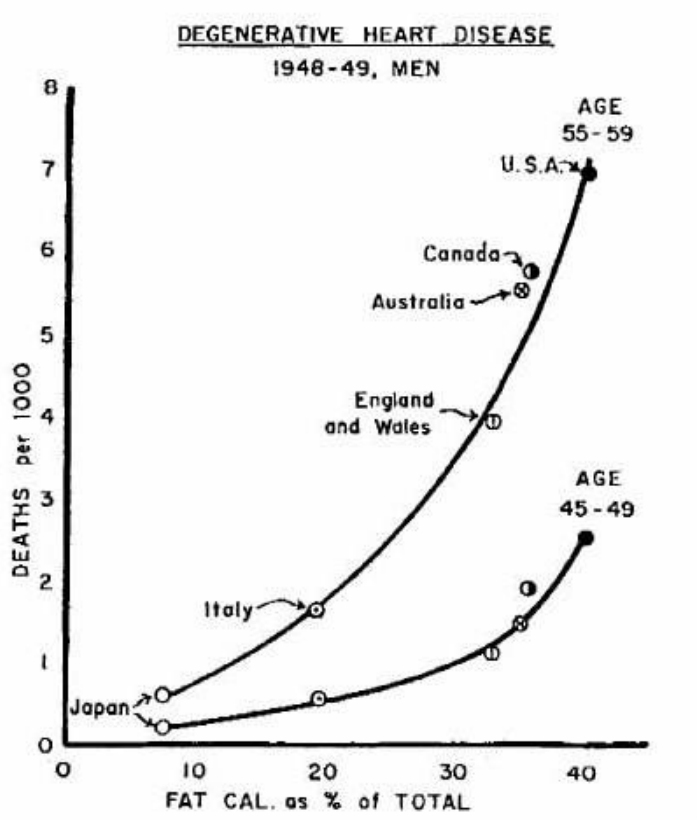
5 Sep, 2017

This material partly takes the contents/presentations from "What is Data Science" published by O'Reilly, John Canny's lecture note on "Introduction to Data Science" at UCB, Randy Paffenroth's slide on "Musing on Data Science and Student Experiencing Data Analytics"

The photo is taken at https://www.myswitzerland.com/de/goldwaschen-am-simplon.html

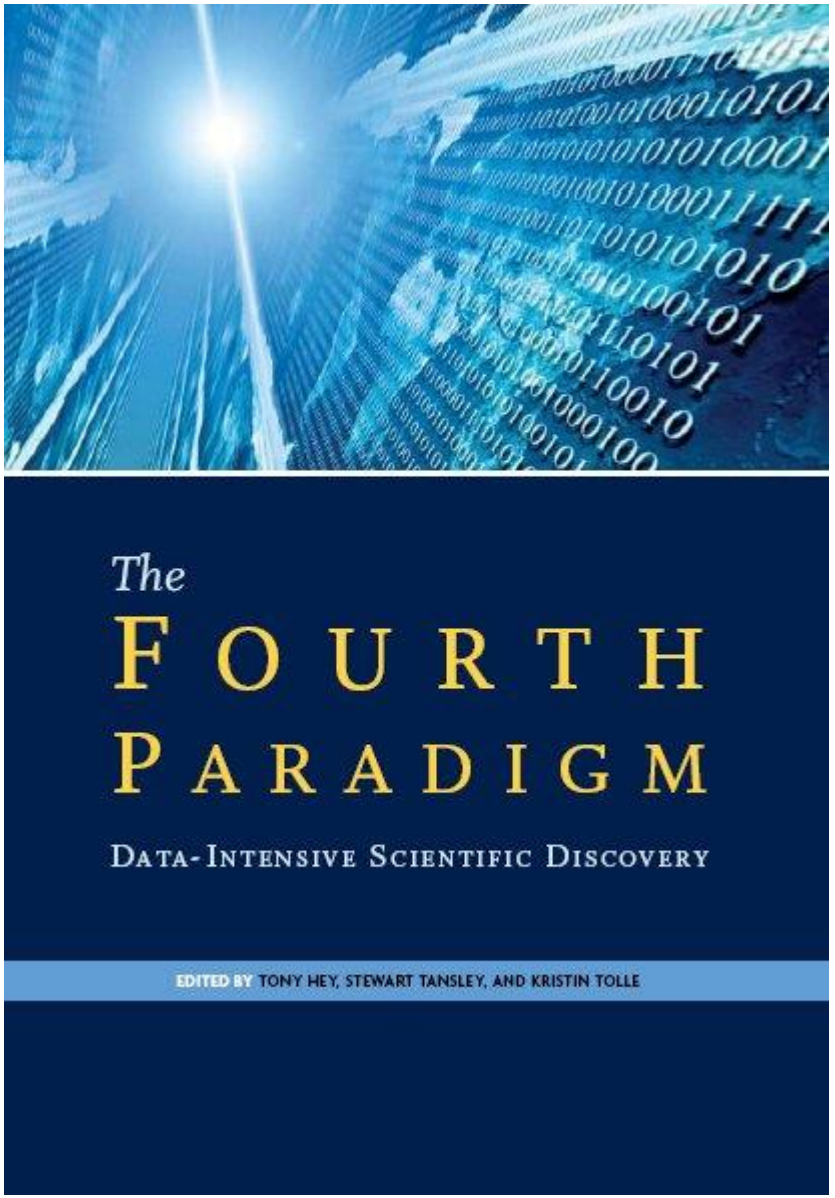
Data makes everything clearer

- Seven Countries Study (Ancel Keys, UCB 1925,28)
- 13,000 subjects total, 5-40 years follow-up.

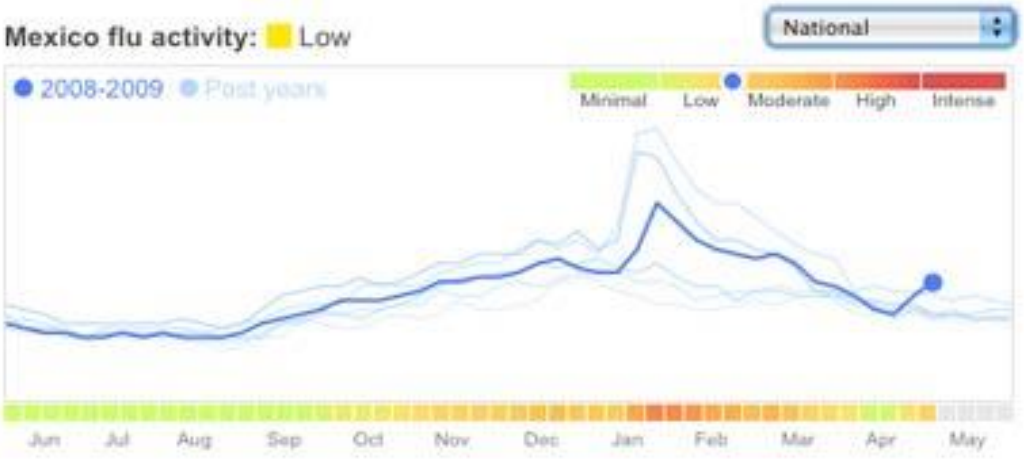


Fourth Paradigm of Science

- Thousands of years
  - Empirical
- Few hundreds of years
  - Theoretical
- Last fifty years
  - Computational
  - “Query the world”
- Last twenty years
  - eScience (**Data Science**)
  - “Download the world”



Data makes everything clearer



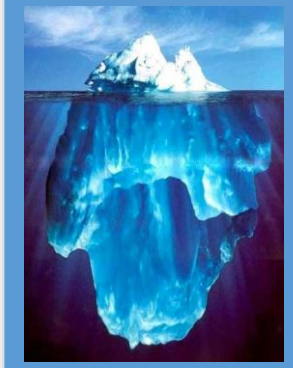
e.g.,
Google Flu Trends:

Detecting outbreaks
two weeks ahead
of CDC data

New models are estimating
which cities are most at risk
for spread of the Ebola virus.

“Big Data” Sources

It’s All Happening On-line



Every:
Click
Ad impression
Billing event
Fast Forward, pause,…
Server request
Transaction
Network message
Fault
…

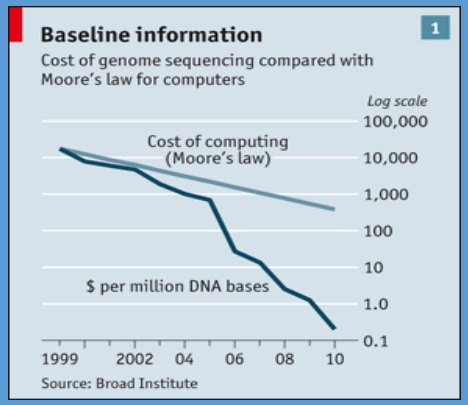
User Generated (Web & Mobile)




….
.


Internet of Things / M2M

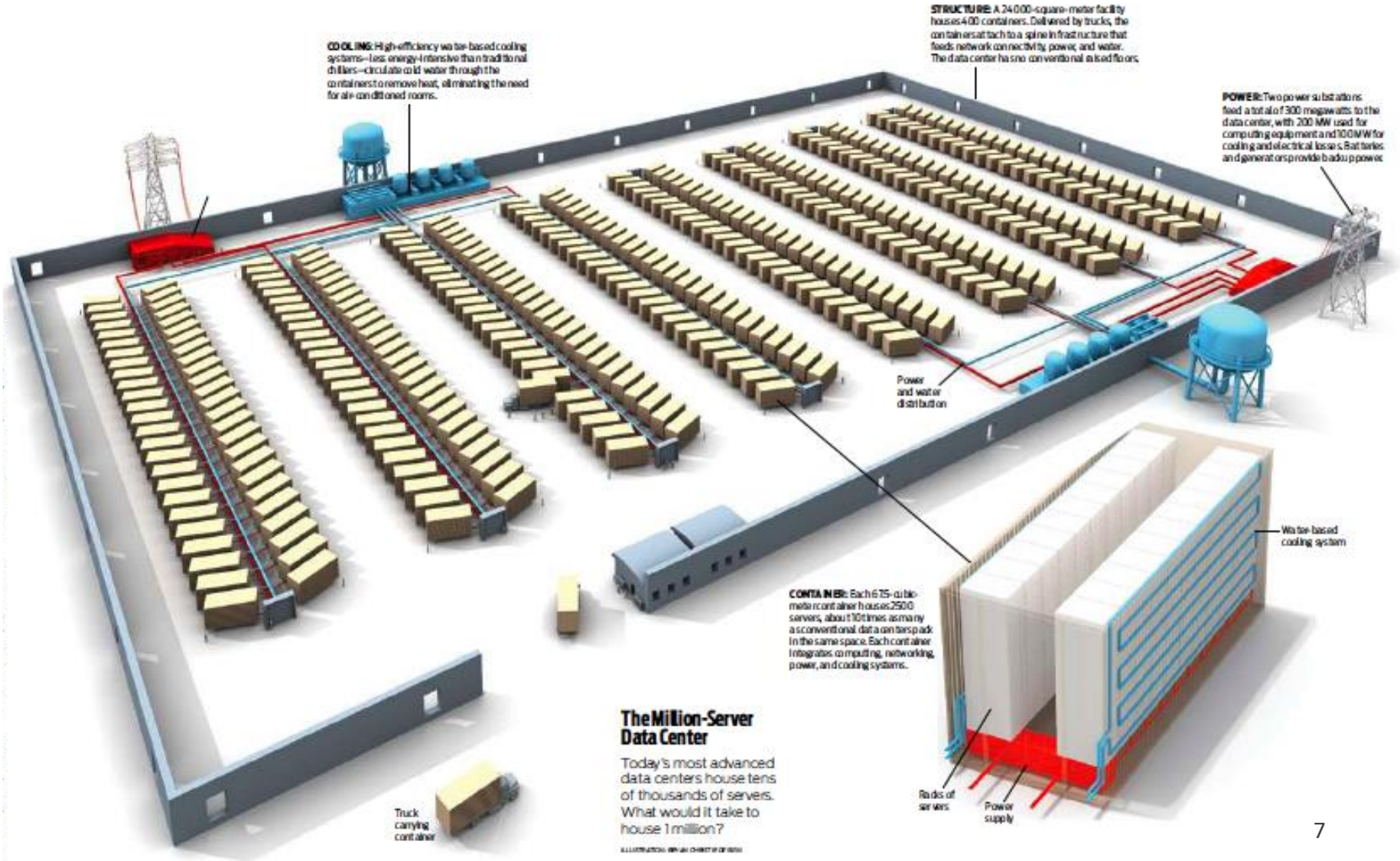
Health/Scientific Computing



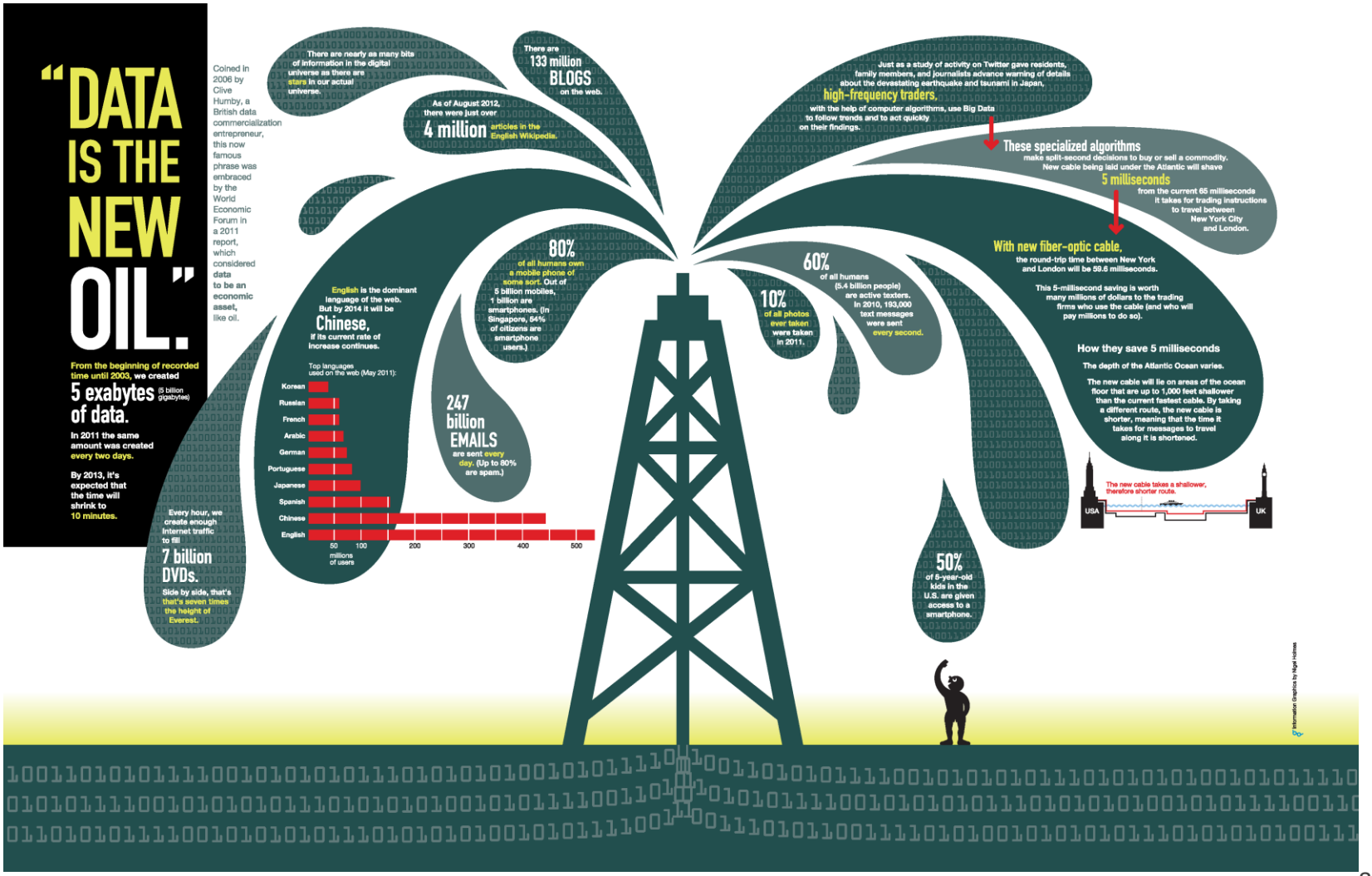
DB size = 50 billion sites

Google server farms
2 million machines (est)

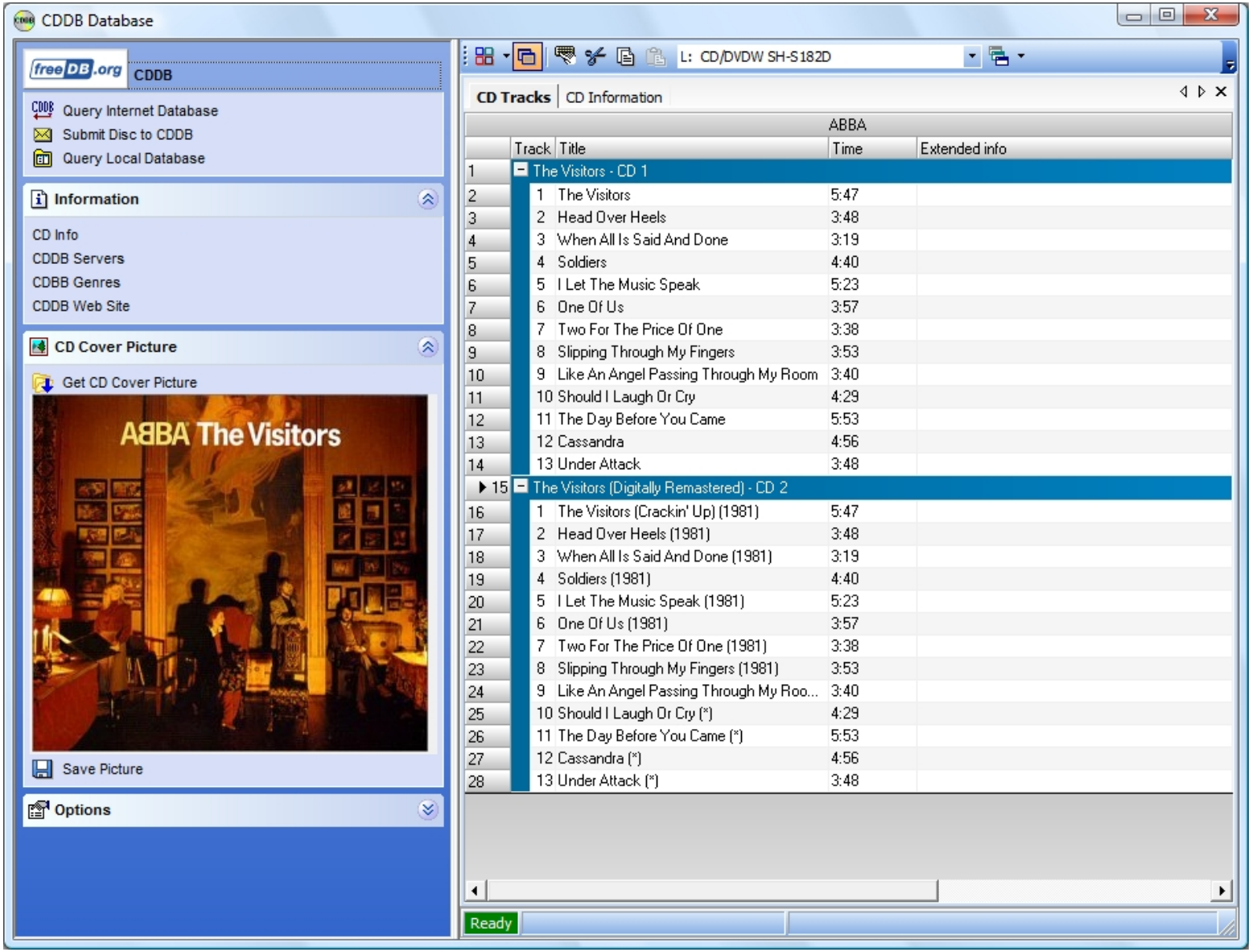
The Million-Server Data Center



“Data is the New Oil” – World Economic Forum 2011



Early Data-driven Applications



Why the all the Excitement?

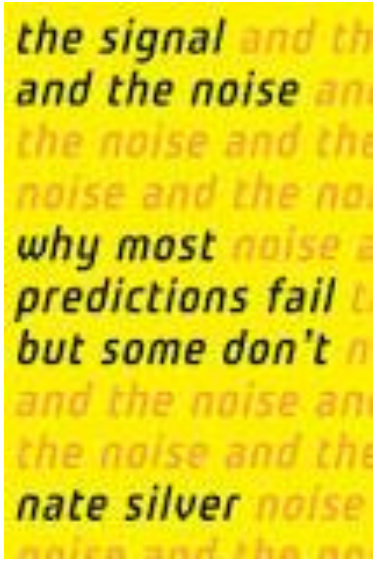
elections2012

Live results | President | Senate | House | Governor | Choose your

Numbers nerd Nate Silver's forecasts prove all right on election night

FiveThirtyEight blogger predicted the outcome in all 50 states, assuming Barack Obama's Florida victory is confirmed

Luke Harding
guardian.co.uk, Wednesday 7 November 2012 10.45 EST



Data and Election 2012 (cont.)

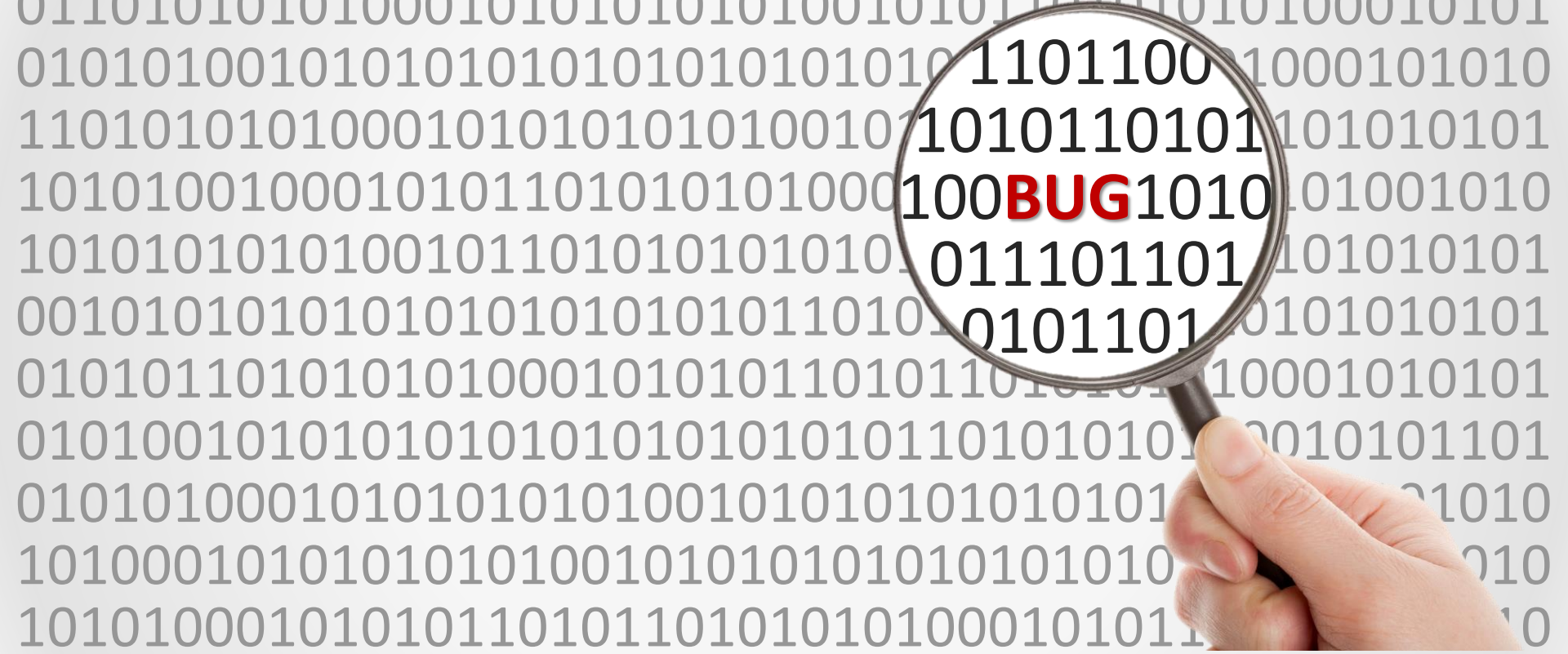
- …that was just one of several ways that Mr. Obama's camp aign operations, some unnoticed by Mr. Romney's aides in Boston, helped save the president's candidacy. In Chicago, the campaign recruited a team of behavioral scientists to b uild an extraordinarily sophisticated database

  …that allowed the Obama campaign not only to alter the v ery nature of the electorate, making it younger and less wh ite, but also to create a portrait of shifting voter allegiances . The power of this operation stunned Mr. Romney's aides on election night, as they saw voters they never even knew existed turn out in places like Osceola County, Fla.

New York TImes, Wed Nov 7, 2012

“Big Data” is so 2012

- “… the sexy job in the next 10 years will be statisticians,” Hal Varian, Google Chief Economist
- the U.S. will need 140,000-190,000 predictive analysts a nd 1.5 million managers/analysts by 2018. McKinsey Globa l Institute’s June 2011
- New Data Science institutes being created or repurpose d – NYU, Columbia, Washington, UCB,...
- New degree programs, courses, boot-camps:
  - e.g., at Berkeley: Stats, I-School, CS, Astronomy…
  - One proposal (elsewhere) for an MS in “Big Data Science ”



Data-driven Software Debugging : A Mutation Approach

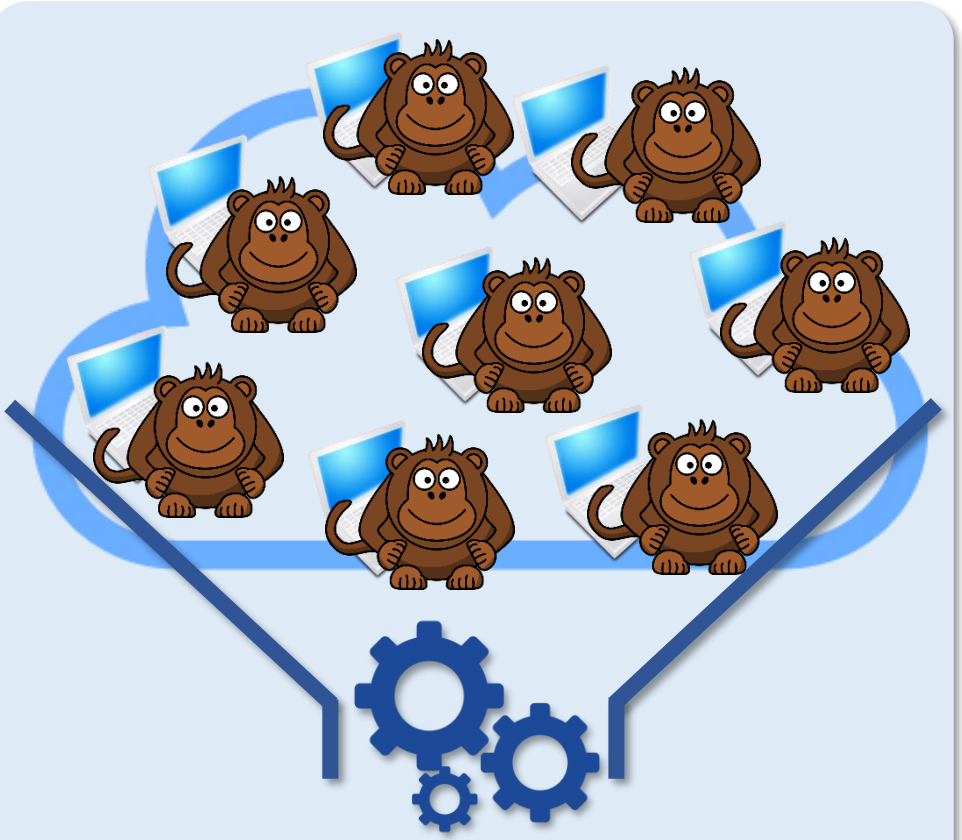
Shin Hong School of CSEE, Handong Global University

`printf`

+

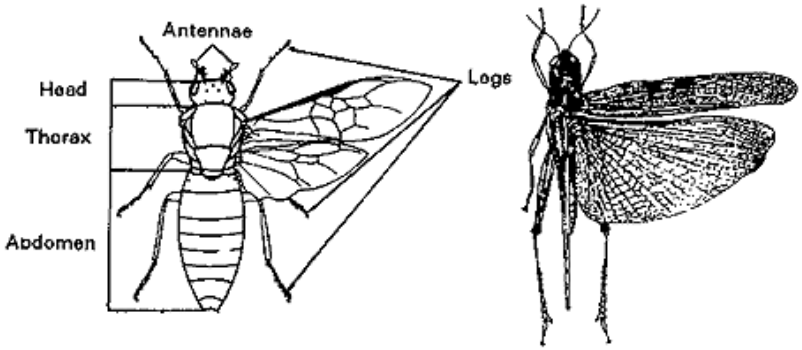
"Debugging [...] remains as **labor-intensive** and **painful as it was five decades ago**"

- Andreas Zeller, 2001



Automated, parallel generation of debugging information using mutation analysis

Taxonomy of Software Bug

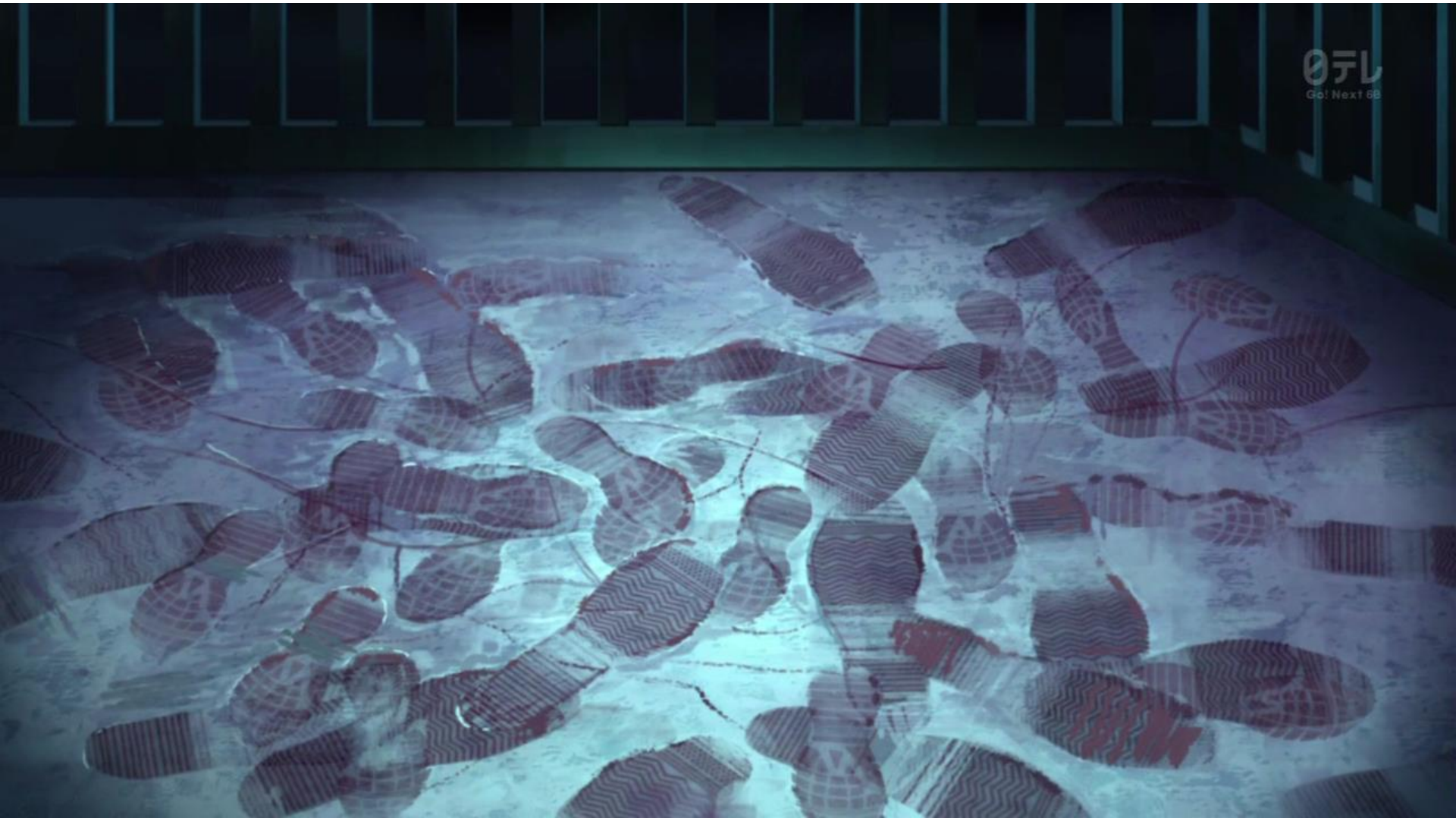


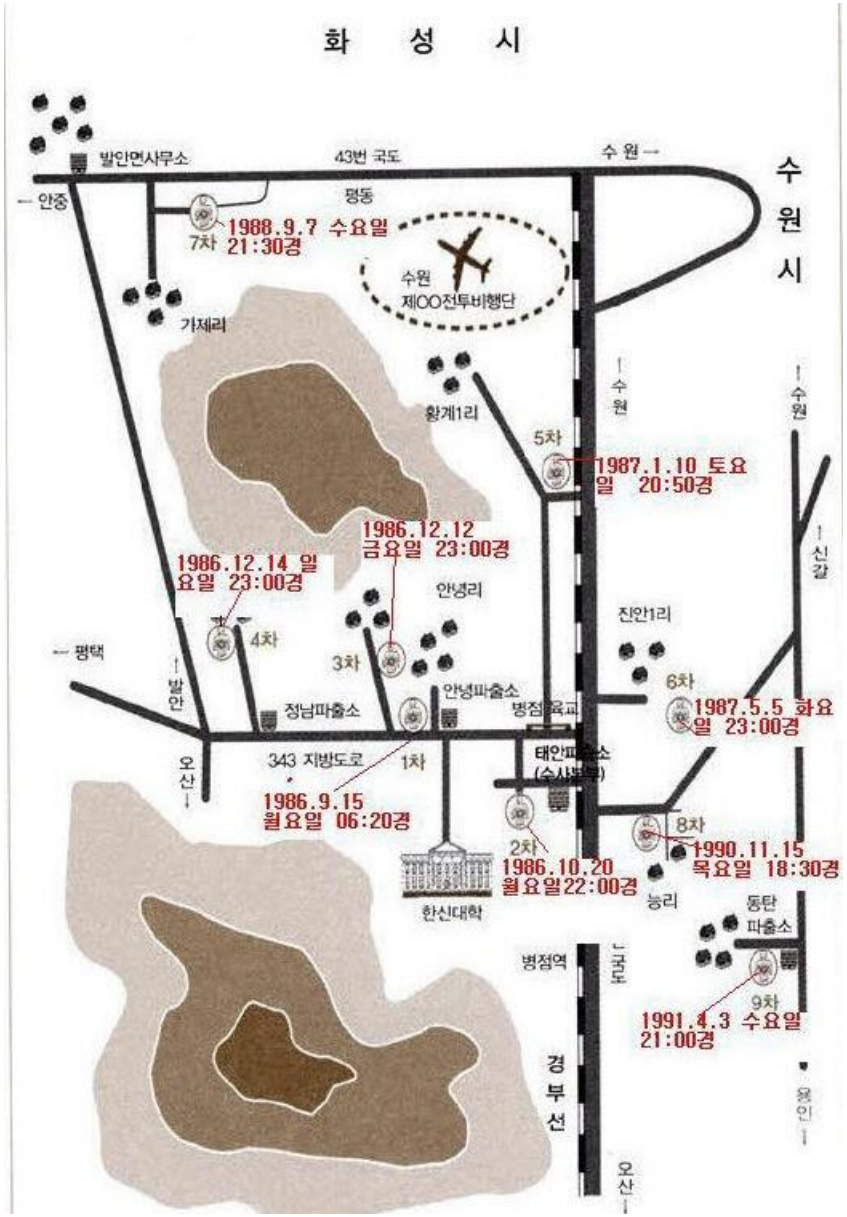
- Error (failure)
  : the execution with unexpected results

- Fault (bug, defect)
  : a flaw in code that is responsible for making an execution fail

- Execution-Infection-Propagation (PIE): how a fault causes an error
  - **Execution**: execution flow reaches to a fault
  - **Infection**: the fault produces an invalid state
  - **Propagation**: the invalid state leads to a symptom


221B


日テレ
Go! Next 8B


화 성 시
수 원 시
발안면사무소
43번 국도
수 원
안중
평동
1988.9.7 수요일
7차 21:30경
수원
제OO전투비행단
가제리
환계1리
5차
1987.1.10 토요일 20:50경
1986.12.12
금요일 23:00경
1986.12.14 일
요일 23:00경
안녕리
4차
3차
진안1리
6차
1987.5.5 화요일 23:00경
평택
발안
정남파출소
안녕파출소
병점초교
343 지방도로
1차
오산
1986.9.15
월요일 06:20경
2차
1986.10.20
월요일22:00경
8차
1990.11.15
목요일 18:30경
한신대학
능리
동탄
파출소
병점역
9차
1991.4.3 수요일
21:00경
경부선
신갈
용인
오산

Example

```
   mid(x, y, z){
01  m = y;
02  if (m < z)
03   if (m < x)
04    m = x;
05    if (z < m)
06     m = z;
07  else
08   if(x > y)
09    m = z; // m = y;
10   else if (x > z)
11    m = x;
12  print m ;
   }
```

mid(3,3,5)	**mid(4,3,3)**	**mid(4,3,2)**
01 m = 3	01 m = 3	m = 3
02 (m < 5)	02 !(m < 3)	!(m < 3)
03 !(m < 3)		
	07 else	07 else
	08(4 > 3)	08 (4 > 3)
	09 **m = 3**	**09 m = 2**
12 print 3	12 **print 3**	**12 print 2**

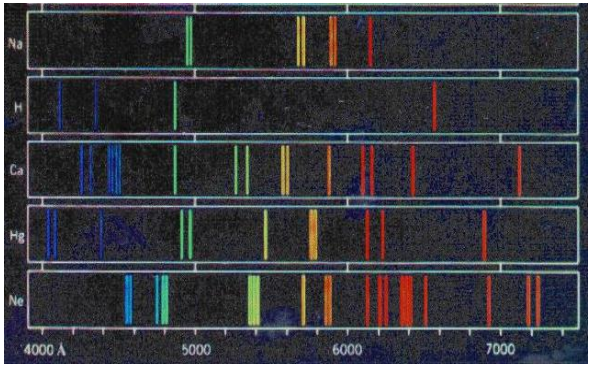
Detective Debugging Process

1. Observe erroneous executions of a program — **Test generation**
2. Make the program fail deterministically — **Error replay**
3. Devise a hypothesis about first invalid state — **Fault localization**
   - "A variable has a wrong value at a location"
4. Test the hypothesis by running the program — **Runtime Monitoring (interactive debug)**
   - Monitor the suspected variable
5. Change the program code to resolve the error — **Patch generation**

Fault Localization

- An activity of searching for **code entities responsible for producing an error**
- **The most critical part in debugging**
  - The most difficult step for human developers [Vessey *et al.*, '85]
  - The later steps rely on the accuracy of FL results
- Automated FLs: search-space reduction or prioritization
  1. Program slicing
  2. Spectrum-based fault localization
  3. Mutation-based fault localization

Spectrum-based Fault Localization



- A large test suite generates large execution information from which we can inductively reason about the cause of errors
  - Use code coverage as features (spectrum) of program behaviors

- Spectrum-based fault localization finds a code entity **suspicious** if an errors likely occurs when the code entity was executed.
  - Define a **suspicious score** of a code entity as a **correlations between the error and the code entity executions**
  - Example: Tarantulas [ASE'05]

$$suspScore(s) = \frac{\frac{fail(s)}{totalFail}}{\frac{pass(s)}{totalPass} + \frac{fail(s)}{totalFail}}$$

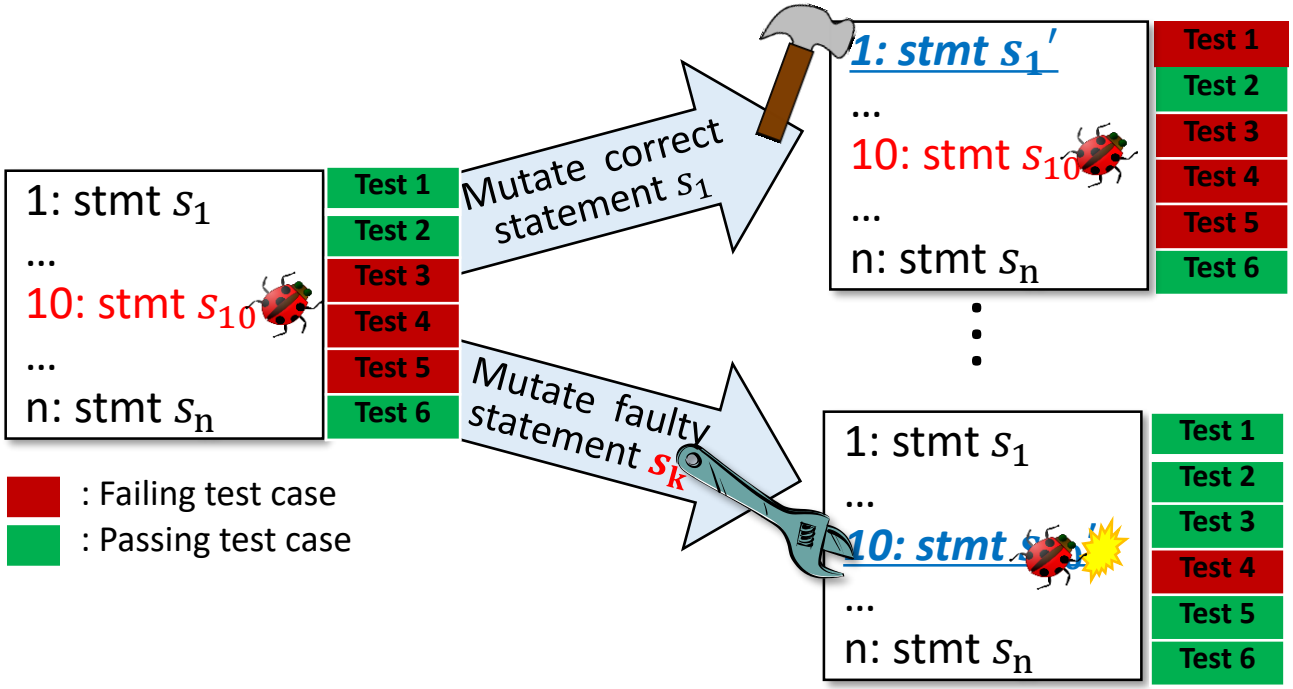
Example

	mid (2,2,1)	mid (1,2,3)	mid (5,5,5)	mid (5,3,4)	mid (2,1,3)	mid (3,2,1)	Susp. score	Rank
01 m = y ;	●	●	●	●	●	●	0.50	6
02 if (m < z)	●	●	●	●	●	●	0.50	6
03 if (m < x)		●		●	●		0.00	12
04 m = x ;				●	●		0.00	12
05 if (z < m)				●	●		0.00	12
06 m = z ;				●			0.00	12
07 else	●		●			●	0.71	3
08 if (x > y)	●		●			●	0.71	3
09 **m = z ;//Bug**						●	1.00	1
10 else if (x > z)	●						0.00	12
11 m = x ;	●						0.00	12
12 print m ; }	●	●	●	●	●	●	0.05	6
	Pass	Pass	Pass	Pass	Pass	Fail		

Mutation-Based Fault Localization (MBFL)

- Infer the suspiciousness of a code line by checking **how testing results change if the code line is mutated**.
- Conjectures

1. more mutants at *s* make passing TCs to failing (breaking)
2. more mutants at *s* make failing TCs to passing (partial fix)



MUSE: MUtant-baSEd Fault Localization

$$suspScore(s) = \sum_{m \in M(s)} \sum_{t \in Fail} F2P(m,t) - \sum_{m \in M(s)} \sum_{t \in Pass} P2F(m,t)$$

P

```
1: if(x > y){
2:     z = x + y;
       ...
```

TC1	TC2	TC3	TC4
Pass	Pass	Fail	Fail

M1

```
1: if(x < y){
2:     z = x + y;
       ...
```

TC1	TC2	TC3	TC4
Pass	Pass	Pass	Fail

M2

```
1: if(x > y+1){
2:     z = x + y;
       ...
```

TC1	TC2	TC3	TC4
Pass	Pass	Fail	Pass

M3

```
1: if(x > y){
2:     z = x - y;
       ...
```

TC1	TC2	TC3	TC4
Fail	Pass	Fail	Fail

M4

```
1: if(x > y){
2:     z = x * y;
       ...
```

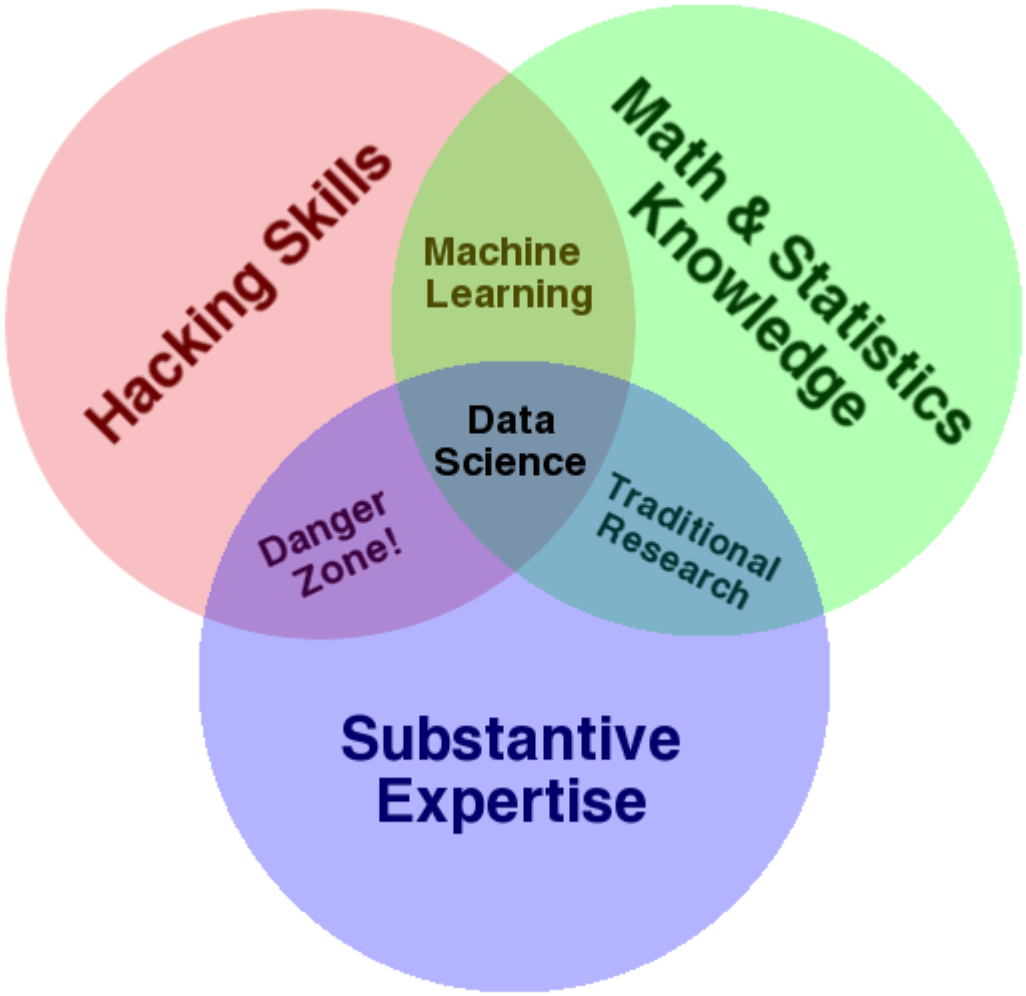
TC1	TC2	TC3	TC4
Fail	Pass	Fail	Fail

Data Science

"… on any given day, a team member could author a multistage processing pipeline in Python, design a hypothesis test, perform a regression analysis over data samples with R, design and implement an algorithm for some data-intensive product or service in Hadoop, or communicate the results of [those] analyses to other members of the organization"

- Jeff Hammerbacher (Cloudera, formerly Facebook)

Data Science – A Visual Definition



5 Vs of Big Data

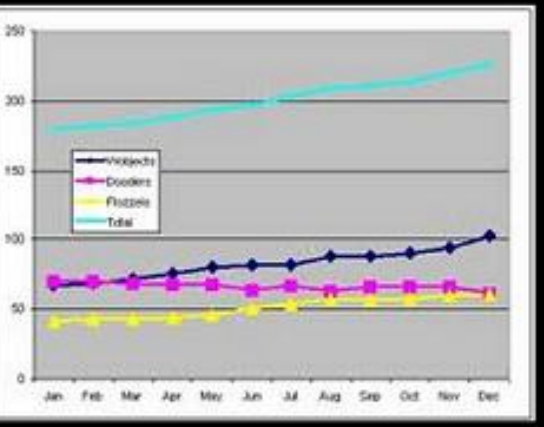
- Raw Data: Volume
- Change over time: Velocity
- Data types: Variety
- Data Quality: Veracity
- Information for Decision Making: Value

Data Scientist

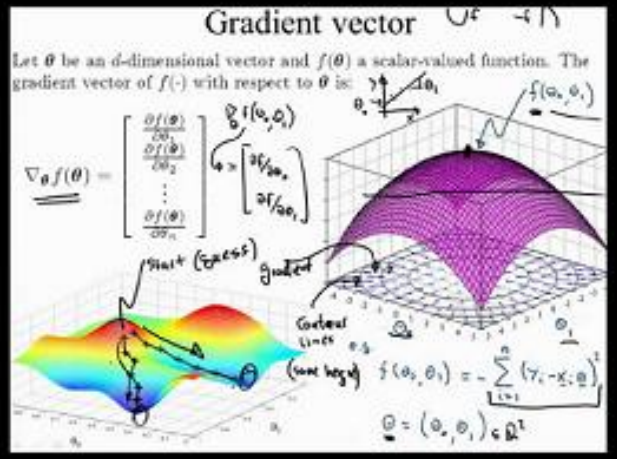
What my friends think I do

What my mom thinks I do

What society thinks I do



What my boss thinks I do

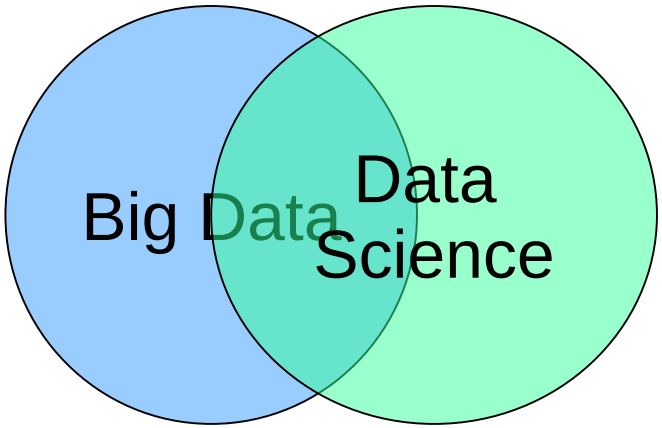


What I think I do

What I actually do

Is Big Data the same as Data Science?

- Are Big Data and Data Science the same thing?
  - I wouldn't say so...
  - Data Science can be done on small data sets.
  - And not everything done using Big Data would necessarily be called Data Science.
  - But there certainly is a substantial overlap!



Contrast: Databases

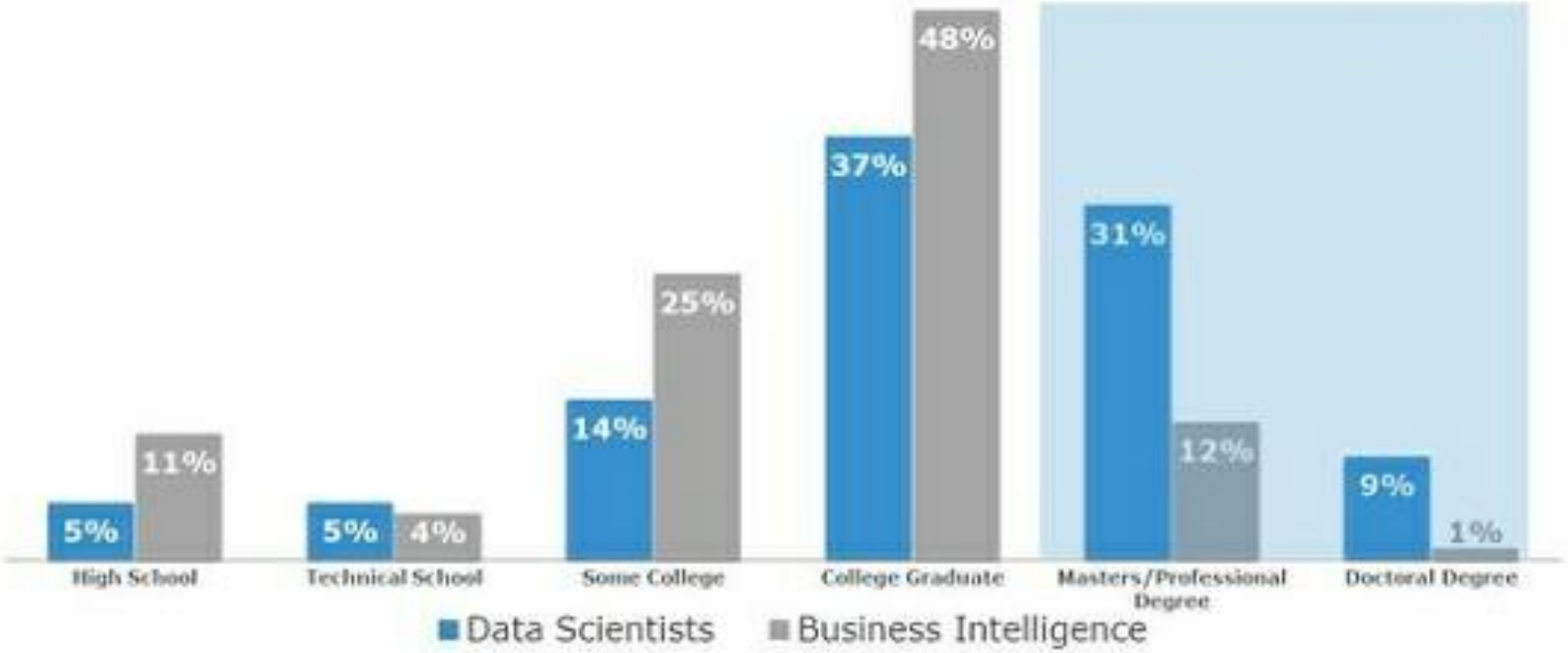
	Databases	Data Science
Data Value	"Precious"	"Cheap"
Data Volume	Modest	Massive
Examples	Bank records, Personnel records, Census, Medical records	Online clicks, GPS logs, Tweets, Building sensor readings
Priorities	Consistency, Error recovery, Auditability	Speed, Availability, Query richness
Structured	Strongly (Schema)	Weakly or none (Text)
Properties	Transactions, ACID*	CAP* theorem (2/3), eventual consistency
Realizations	SQL	NoSQL: Riak, Memcached, Apache River, MongoDB, CouchDB, Hbase, Cassandra,…

ACID = Atomicity, Consistency, Isolation and Durability CAP = Consistency, Availability, Partition Tolerance

Contrast: Machine Learning

Machine Learning	Data Science
Develop new (individual) models	Explore many models, build and tune hybrids
Prove mathematical properties of models	Understand empirical properties of models
Improve/validate on a few, relativ ely clean, small datasets	Develop/use tools that can handle massive datasets
Publish a paper	Take action!

Data science requires greater education

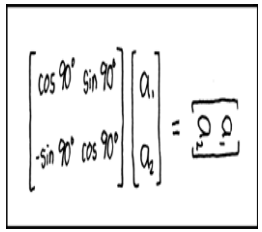


40% of data science professionals have an advanced degree – and nearly one in ten have a doctorate. In contrast, less than 1% of BI professionals have a PhD.

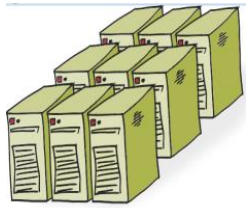
Data Scientist's Practice

Digging Around in Data

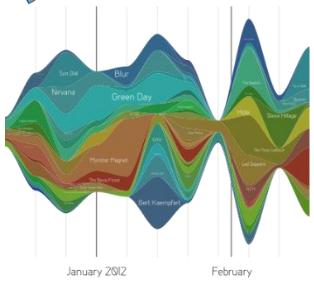
Clean, prep



Hypothesize Model



Large Scale Exploitation

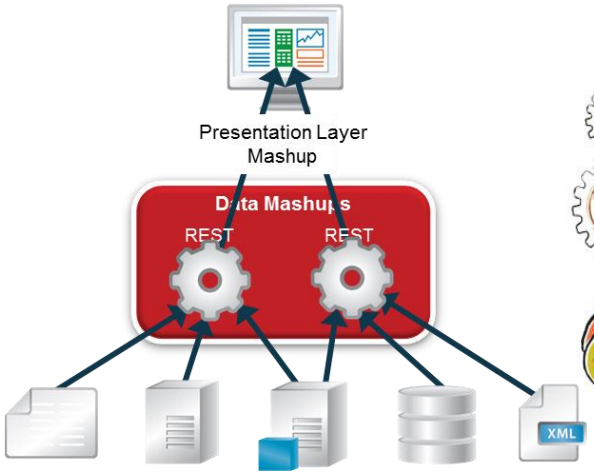


Evaluate Interpret

Data Gathering & Conditioning

- Getting data into a state where the data is usable
  - turning unformatted data in formatted one
  - handling missing, inconsistent and/or incomplete data

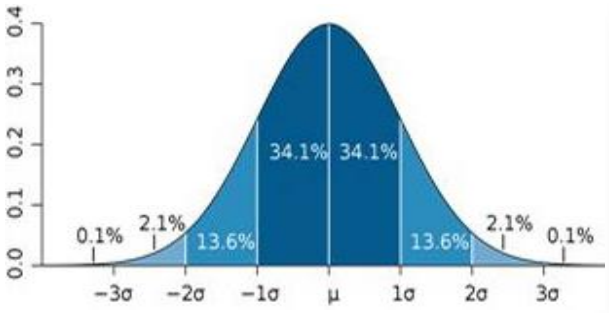
- Use parsers, text-editing tools, web scrappers, natural language processing libraries, or a tool witih human-in-the-loop etc.
  - E.g., clean up messy HTML with web scraping tools

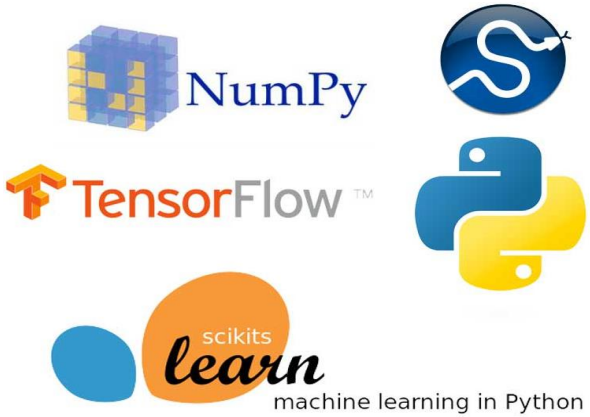


Ryan Mitchell

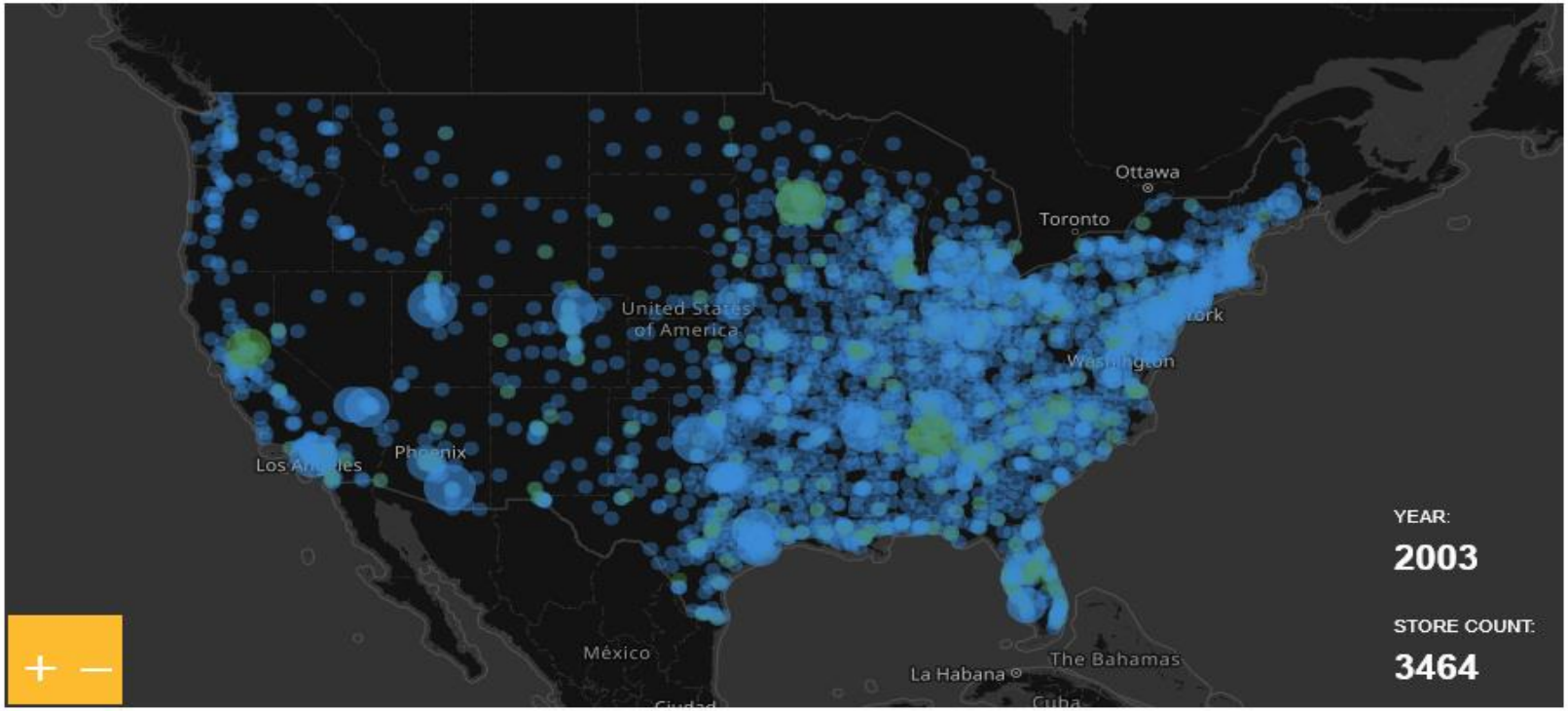
Statistics and Machine Learning

- Create statistical models to represent hypothesis on data
  - make data speak coherently
- Use machines to find patterns/models that explain large-scale data mechanically
  - to test hypotheses, to discover hidden patterns





Data Visualization



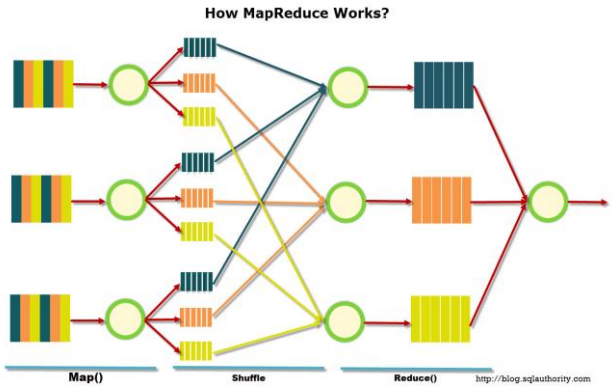
- Generate figures from a set of numbers for human to intuitively understand patterns/tendencies/messages in data
  - E.g., visualization is a key step at data conditioning and analysis design
- Tools: GnuPlot, R, IBM's Many Eyes
- Example: the growth of Walmart and Sam's Club (http://projects.flowingdata.com/walmart/)

Large Scale Query Processing

- "Big data" is when the size of the data itself becomes part of the problem
  - Traditional DBMS's do not scale at processing "Big data"
  - Traditional DBMS's are not effective to support flexible schema

- Operate NoSQL databases to store huge database systems
  - e.g., BigTable @ Google, Dynamo @ Amazon, Cassandra @ Facebook

- Construct data analysis applications to query on massive database efficiently
  - e.g., Hadoop with MapReduce, HDFS, HBase, Pig, etc.



Syllabus

- Topic: build movie recommendation system with collaborative filtering and classification technique

- Schedules
  - Understand collaborative filtering with Market-Basket analysis (9/26)
  - Handle movie rating dataset with Java libraries (9/26)
  - Implement a movie recommendation system (9/26)

  - Use a NoSQL Redis to manage the movie rating data (10/10)
  - Improve the recommendation system by using Redis (10/10)

  - Understanding the Random Forest method (10/17)
  - Practice RF classification with Weka (10/17)
  - Homework (10/17)

Term Project Topic

- Build a recommendation system with Amazon's fine food review dataset
  - use 500,000 food reviews opened at Kaggle
  - explore data to discover interesting findings