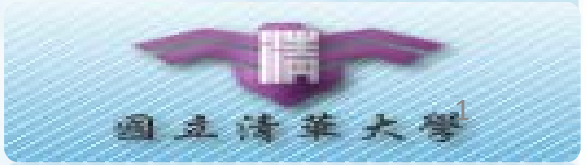
Cloud Programming: Lecture3 – Big Data

National Tsing-Hua University

2015, Spring Semester

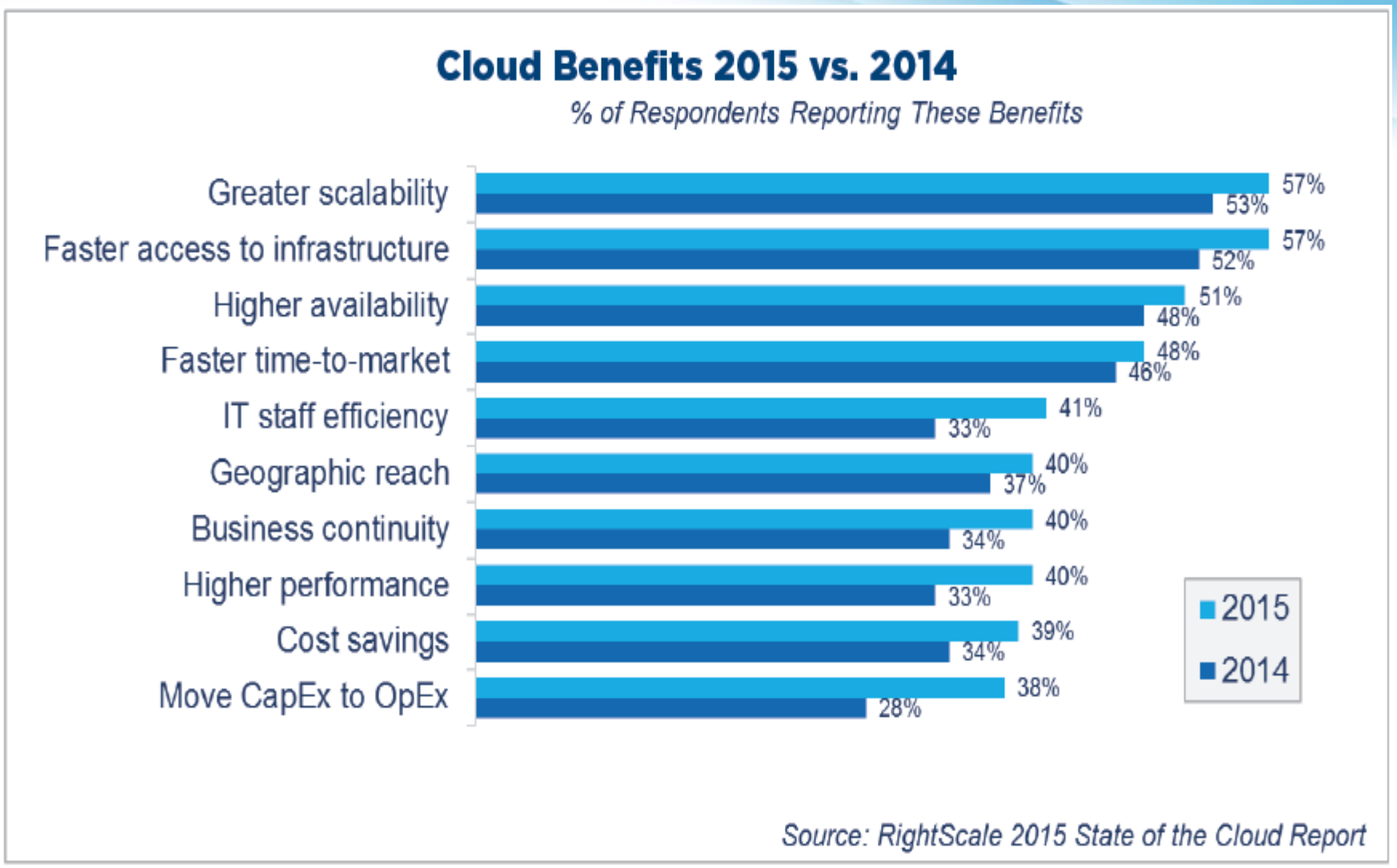


Outline

- Introduction
- Big data
- Data science
- Data processing tools
- Conclusion

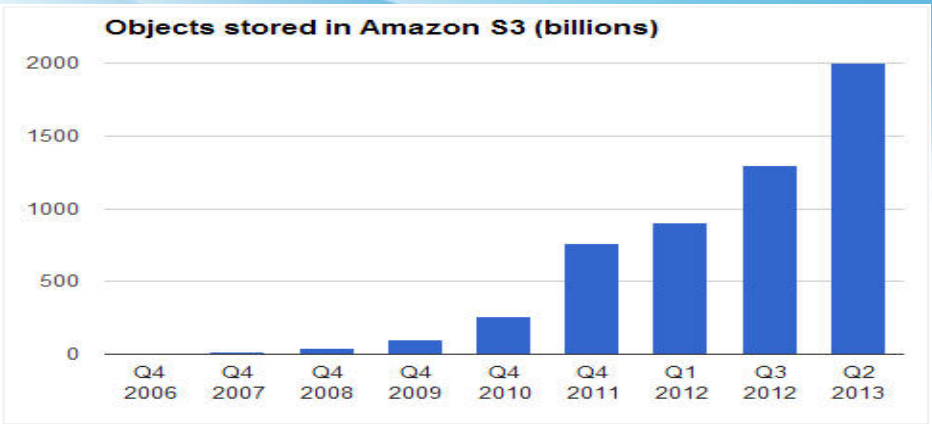
Technology Changes the World…

Cloud Benefits

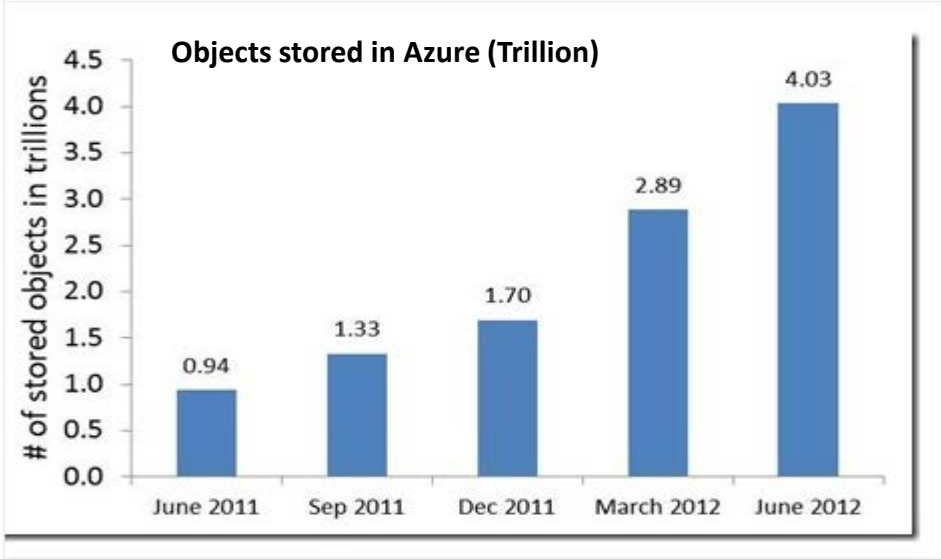


Data in Cloud

- Cloud computing has become a viable, mainstream solution for data processing, storage and distribution
- S3
  - Stores 2 trillion objects in 2013 = 250 objects for each person on Planet Earth
  - Regularly peaks at over 1.1M requests per second
- Azure
  - Stores 4 trillion objects in 2012
  - Regularly peaks at over 880K requests per second



www.theregister.co.uk/2013/04/18/amazon_2_trillion_s3/

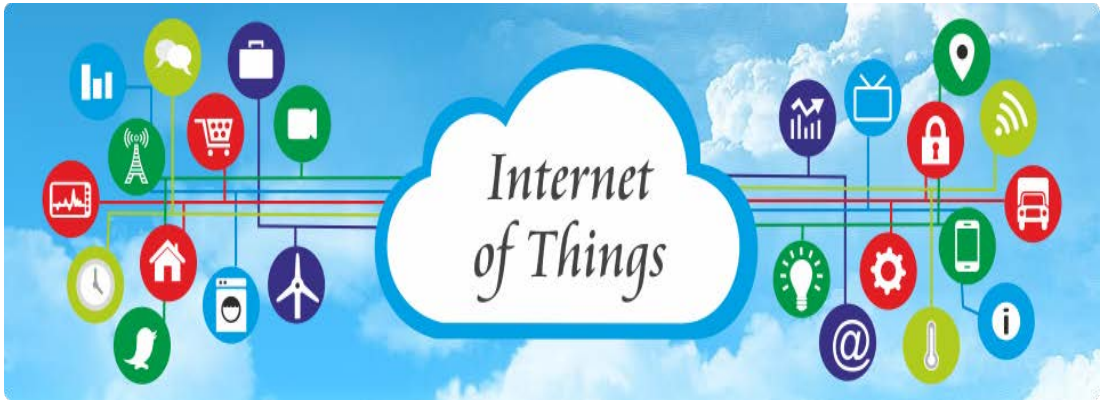


http://www.theregister.co.uk/2012/07/20/azure_four_trillion/

Killer App for the Public Cloud

- Big Data
  - Greater scalability
  - Higher performance
  - Cost savings
  - Faster time-to-market

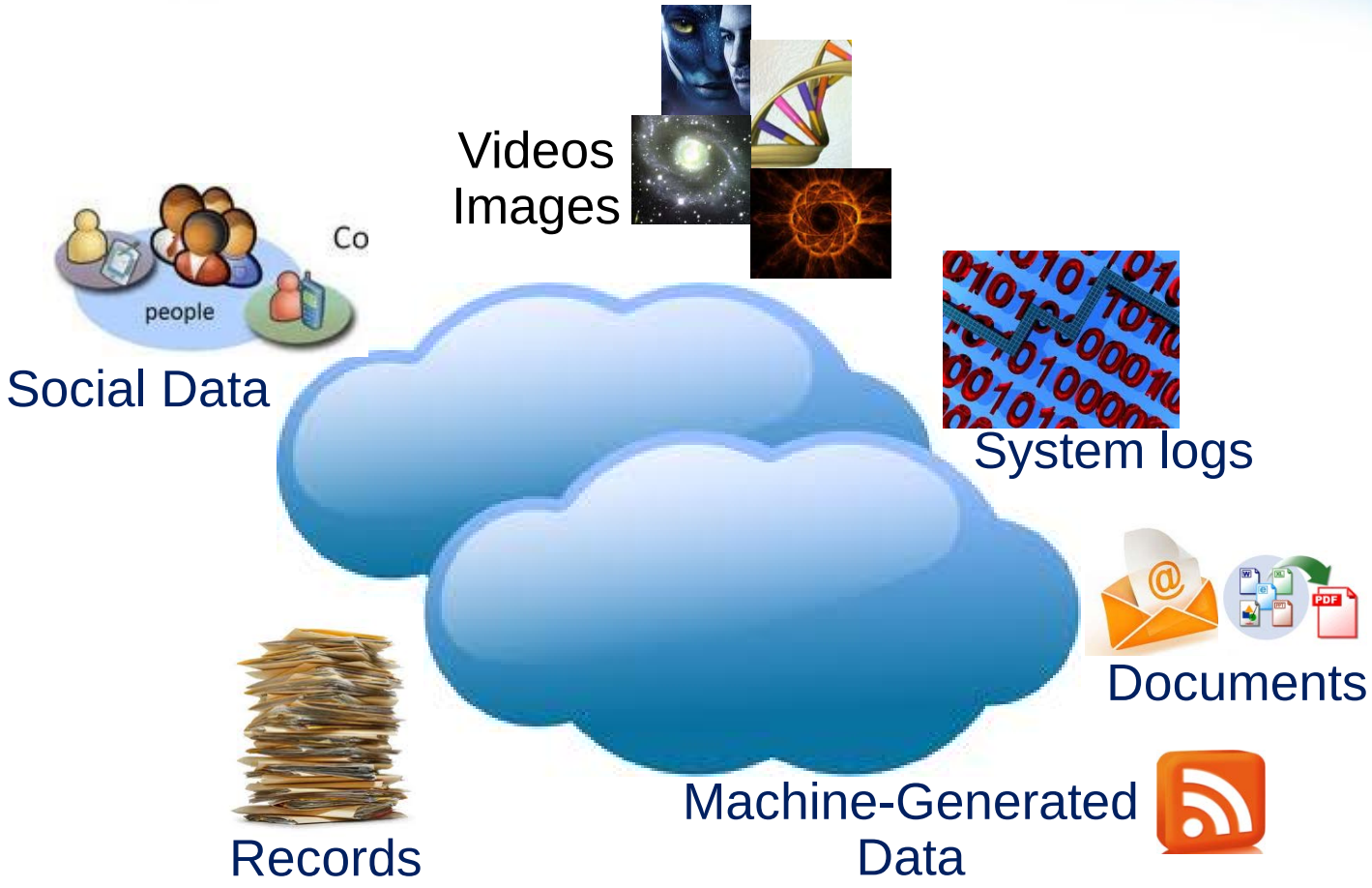
- IoT
  - Higher availability
  - Faster access
  - Geographic reach



Outline

- Introduction
- Big data
- Data science
- Data processing tools
- Conclusion

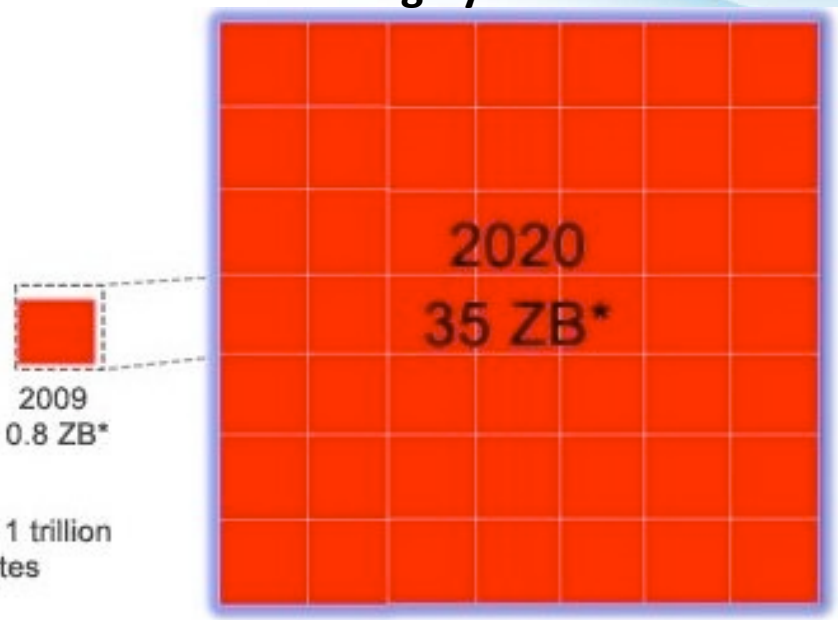
A World Full of Data



How Many Data

- Amount of data in the world
  - 800 Terabytes, 2000
  - 160 Exabytes, 2006
  - 0.8 Zettabytes, 2009
  - 2.7 Zettabytes, 2012
  - 35 Zettabytes by 2020
- Data generated in ONE day
  - 7 TB, Twitter
  - 10 TB, Facebook

Digital Universe 2009-2020, Growing by a factor of 44

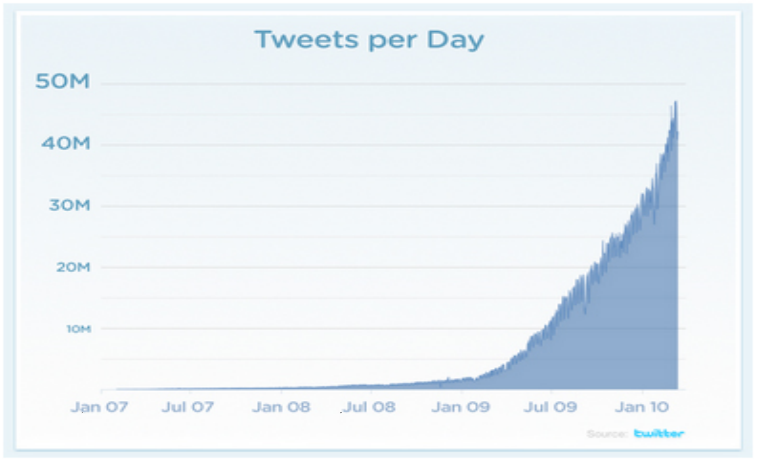


35ZB = enough data to fill a stack of DVDs reaching half way to Mars

http://www.storagenewsletter.com/news/miscellaneous/digital-universe-decade-emc-idc

The Explosion of Data

- A increased number and variety of **data sources** that generate large quantities of data
  - Sensors(e.g. measurements)
  - Mobile devices(e.g. phone)
  - Social Network (e.g. twitter, wikis)
  - OLTP (e.g. bank transactions)



Mobile device

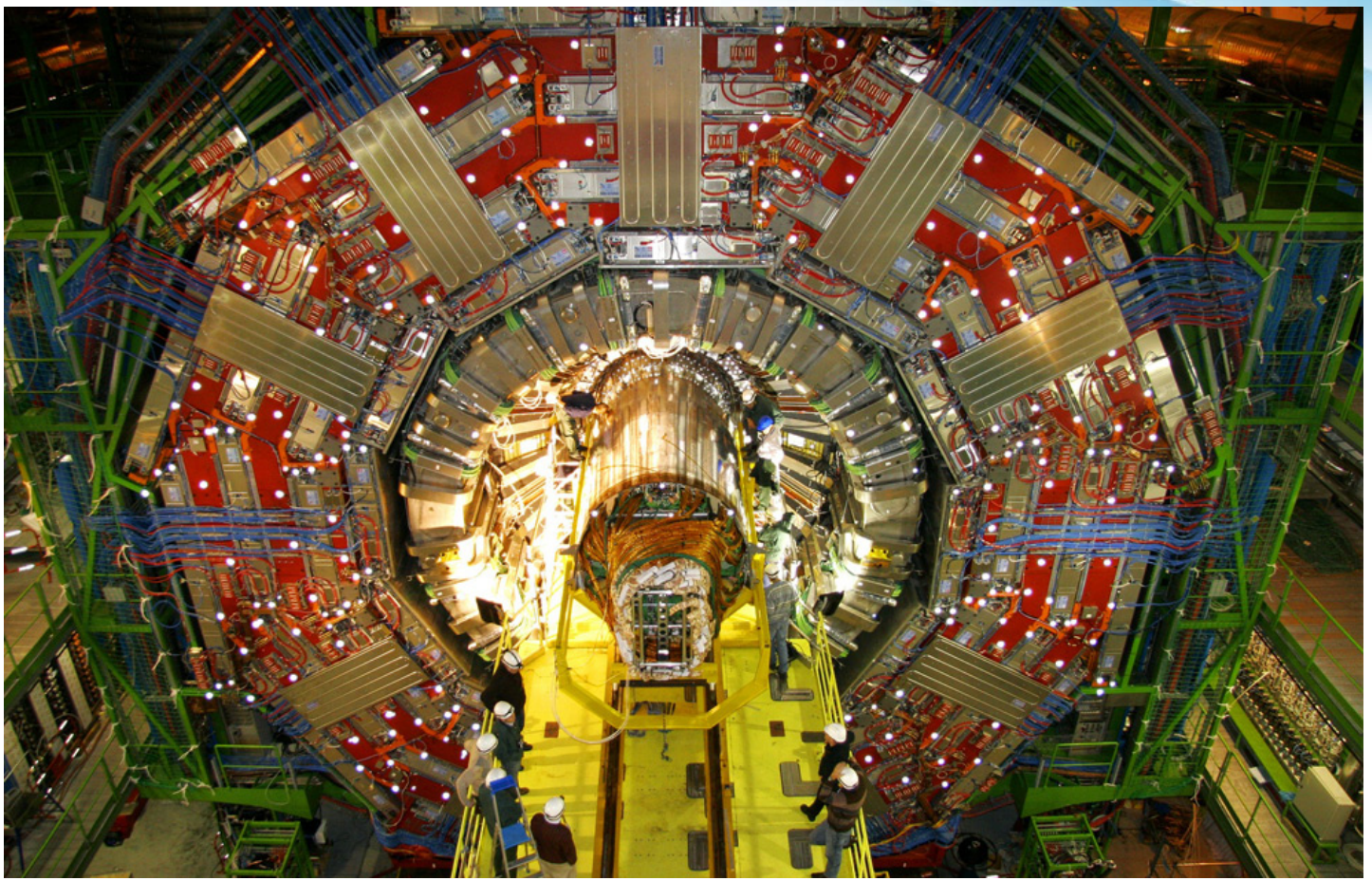
Sensors

OLTP

Social Networks

Scientific Devices

The Explosion of Data

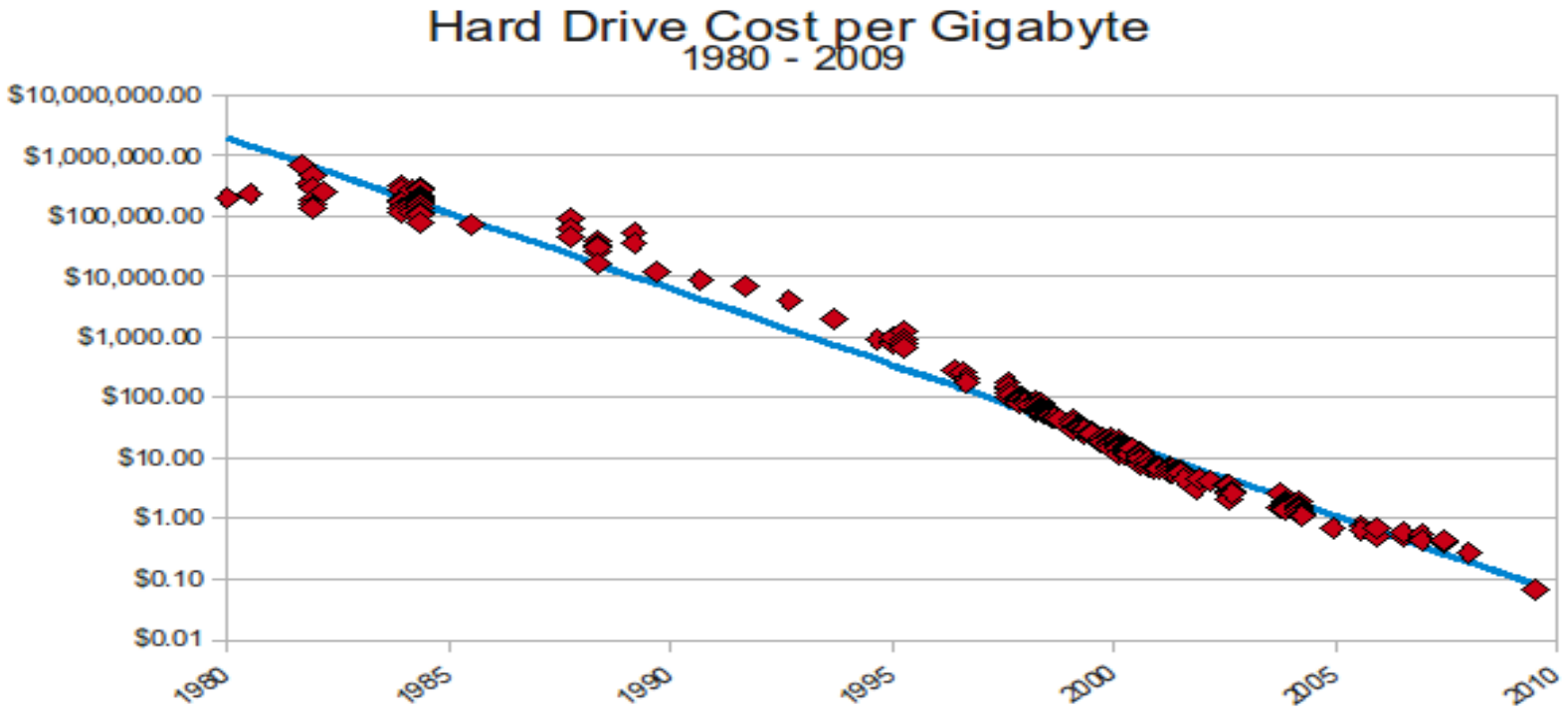


- CERN's Large Hydron Collider (LHC) generates 15 PB a year

The Explosion of Data

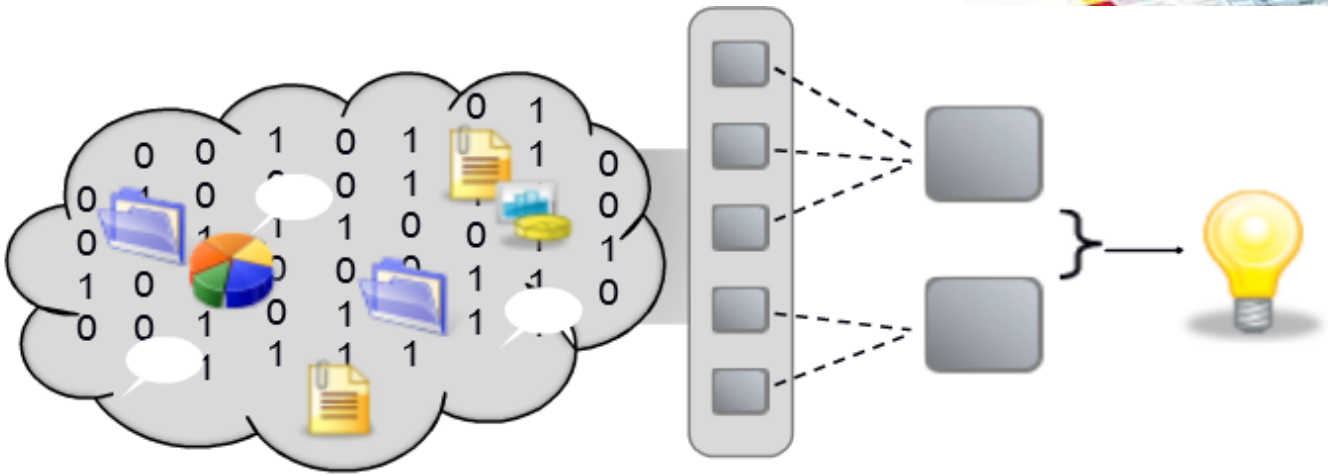
- Dramatic decline in the **cost of HW**, especially storage
  - The cost reduction is in the order of about 40-45% per year – which means it becomes half in 2 years

➔ It is FREE for storage: $1+1/2+1/4+\ldots = 2 \neq \infty$



The Explosion of Data

- Realize data is “too valuable” to delete
  - Diagnose system
  - Understand user behavior
  - Evaluate merchandise & products
  - Make business decision



The promise of Big Data

- Data contains information of great business value
- If you can extract those insights you can make far better decisions
- ...but is data really that valuable?

NATURE | NEWS

Drug data reveal sneaky side effects

Mining of surveillance data highlights thousands of previously unknown consequences when drugs are taken together.

Heidi Ledford

14 March 2012

An algorithm designed by US scientists to trawl through a plethora of drug interactions has yielded thousands of previously unknown side effects caused by taking drugs in combination.

The work, published today in *Science Translational Medicine*[1], provides a way to sort through the hundreds of thousands of 'adverse events' reported to the US Food and Drug Administration (FDA) each year. "It's a step in the direction of a complete catalogue of drug–drug interactions," says the study's lead author, Russ Altman, a bioengineer at Stanford University in California.

Although clinical trials are often designed to assess the safety of a drug in addition to how well it works, the size of the trials needed to detect the full range of drug interactions would surpass even the large, late-stage clinical trials sometimes required for drug approval. Furthermore, clinical trials are often done in controlled settings, using carefully defined criteria to determine which patients are eligible for enrolment — including other conditions they might have and which medicines they can take alongside the trial drug.

A program predicts the potential side-effects of mixing different pills.

DWIMAGES/ALAMY

How Companies Learn Your Secrets

Antonio Bolfo/Reportage for The New York Times

By CHARLES DUHIGG
Published: February 16, 2012 | 570 Comments

Andrew Pole had just started working as a statistician for Target in 2002, when two colleagues from the marketing department stopped by his desk to ask an odd question: "If we wanted to figure out if a customer is pregnant, even if she didn't want us to know, can you do that? "

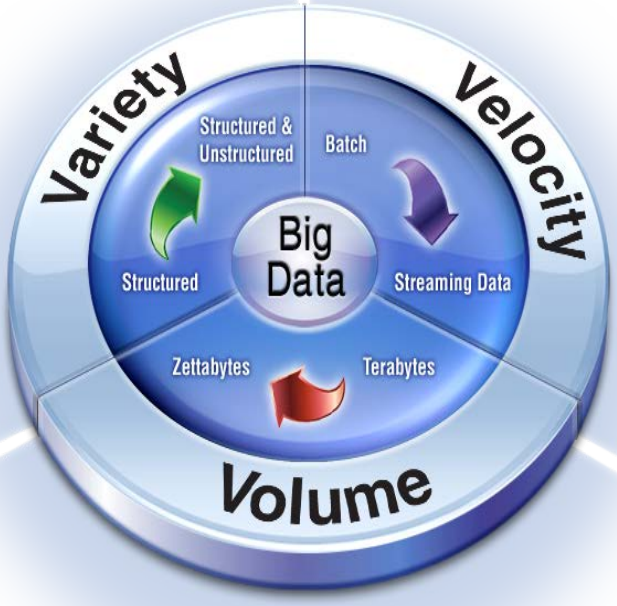
What Makes it Big Data?

*Extracting **values**(insight) from an immense volume, variety and velocity of data, in context, beyond what was previously possible*

Volume: Scale from Terabytes to Petabytes (1K TBs) to Zetabytes (1B TBs)

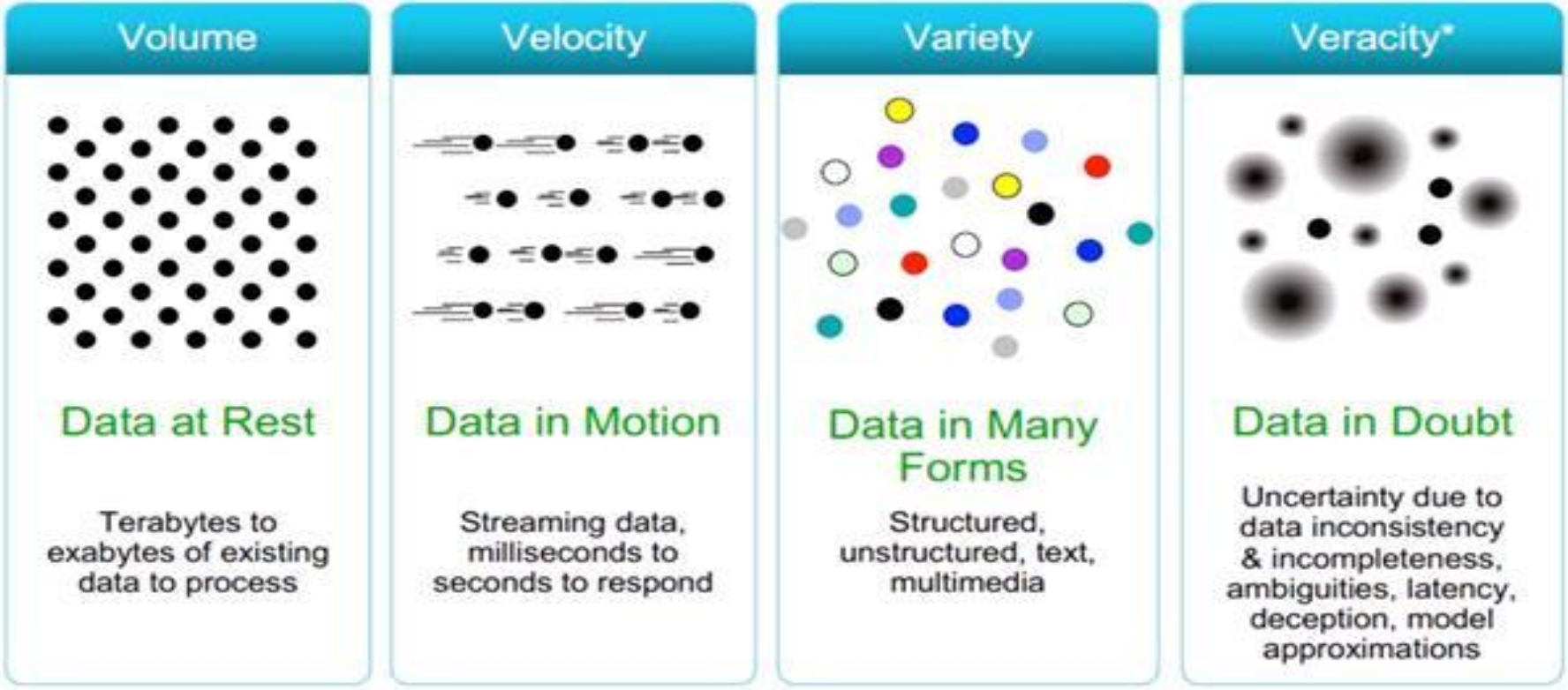
Variety: Manage the complexity of data in many different structures, ranging from relational, to logs, to raw text

Velocity: Streaming data and large volume data movement. How fast to process the data.



What Makes it Big Data?

- Some people even consider it as 5 Vs…



Big Data in Action

Tapping into diverse data sets

Finding and monetizing unknown relationships

Data driven business decisions

Big Data Examples

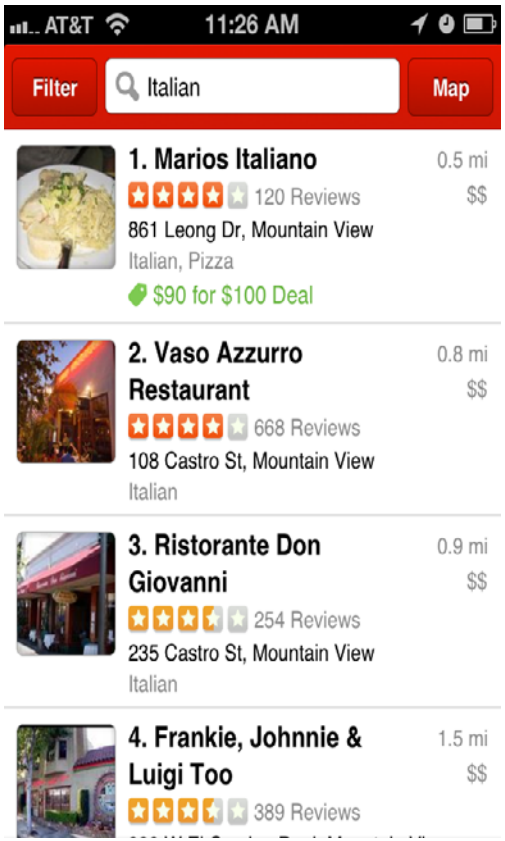
- Traditional Ad on TV
  - Like it or not,

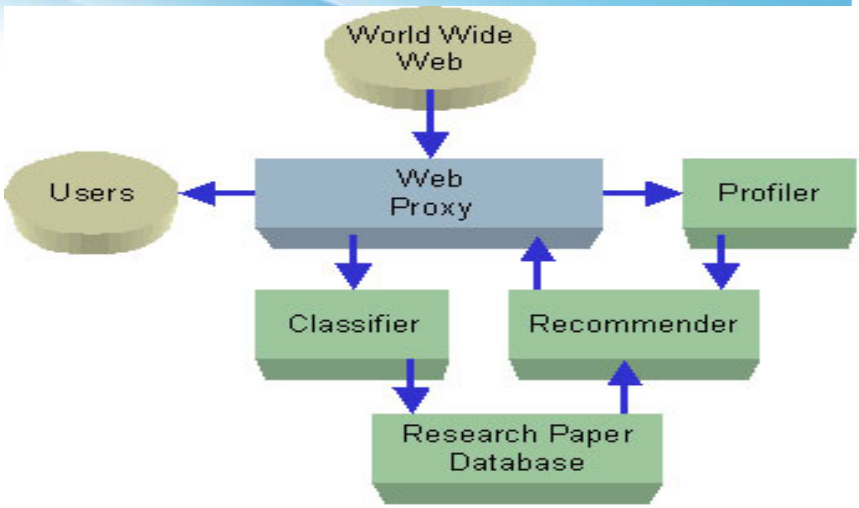
    everyone sees the SAME!

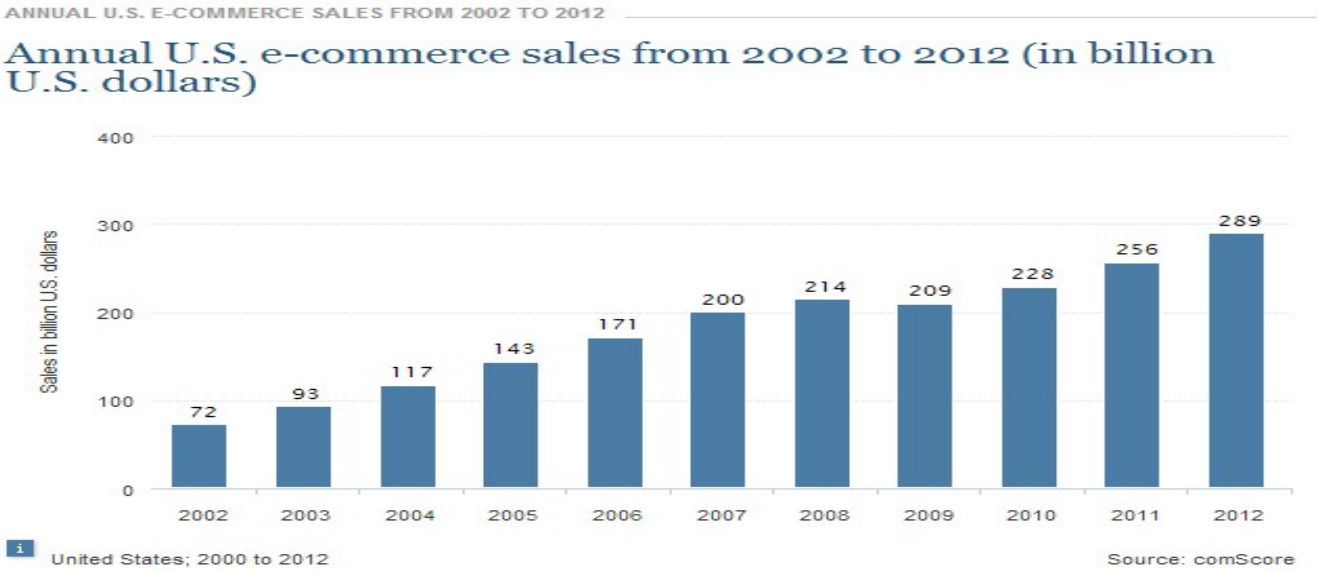
- Today's Ad on Internet
  - Collect user data
  - **Analyze** user behavior
  - Display Ad to match individual user

Big Data Examples

- Recommendation System







Some more examples

- Sports
  - basketball increasingly driven by data analytics
  - soccer beginning to follow
- Entertainment
  - House of Cards designed based on data analysis
  - increasing use of similar tools in Hollywood
- "Visa Says Big Data Identifies Billions of Dollars in Fraud"
  - new Big Data analytics platform on Hadoop
- "Facebook is about to launch Big Data play"
  - starting to connect Facebook with real life

https://delicious.com/larsbot/big-data

Applications for Big Data Analytics

Smarter Healthcare

Multi-channel

Finance

Log Analysis

Homeland Security

Traffic Control

Telecom

Search Quality

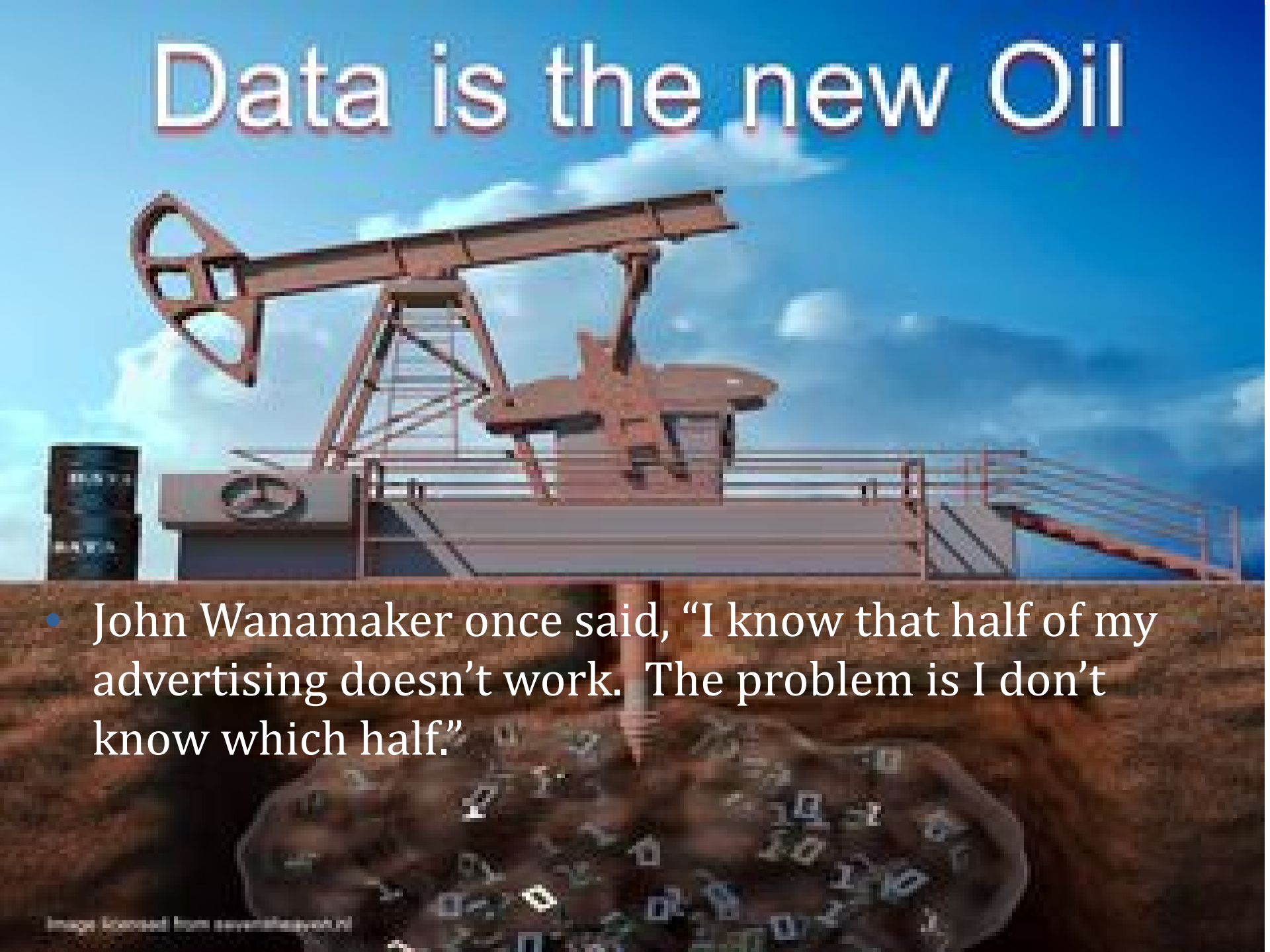
Manufacturing

Trading Analytics

Fraud and Risk

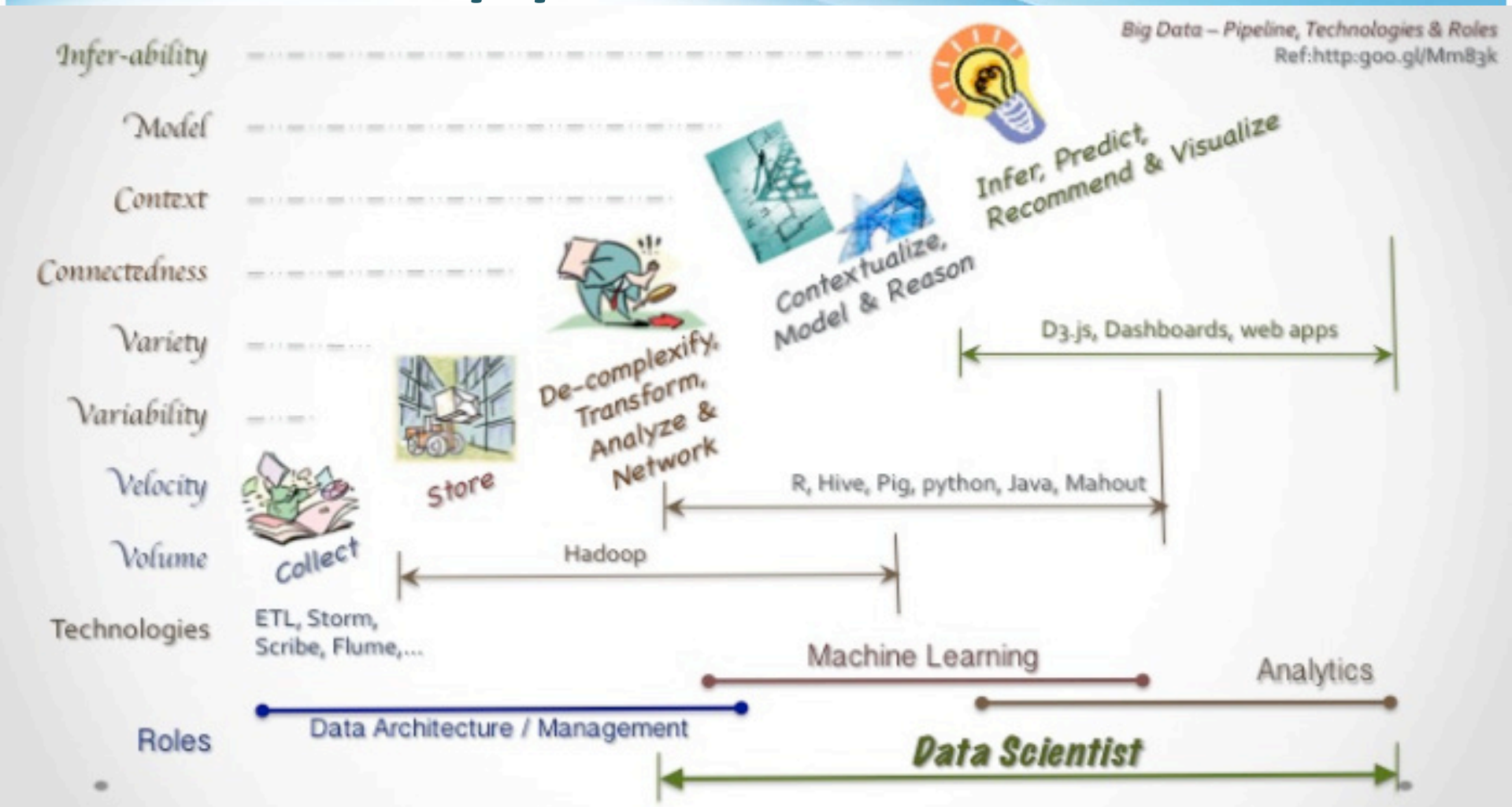
Retail: Churn, NBO

Data is the new Oil



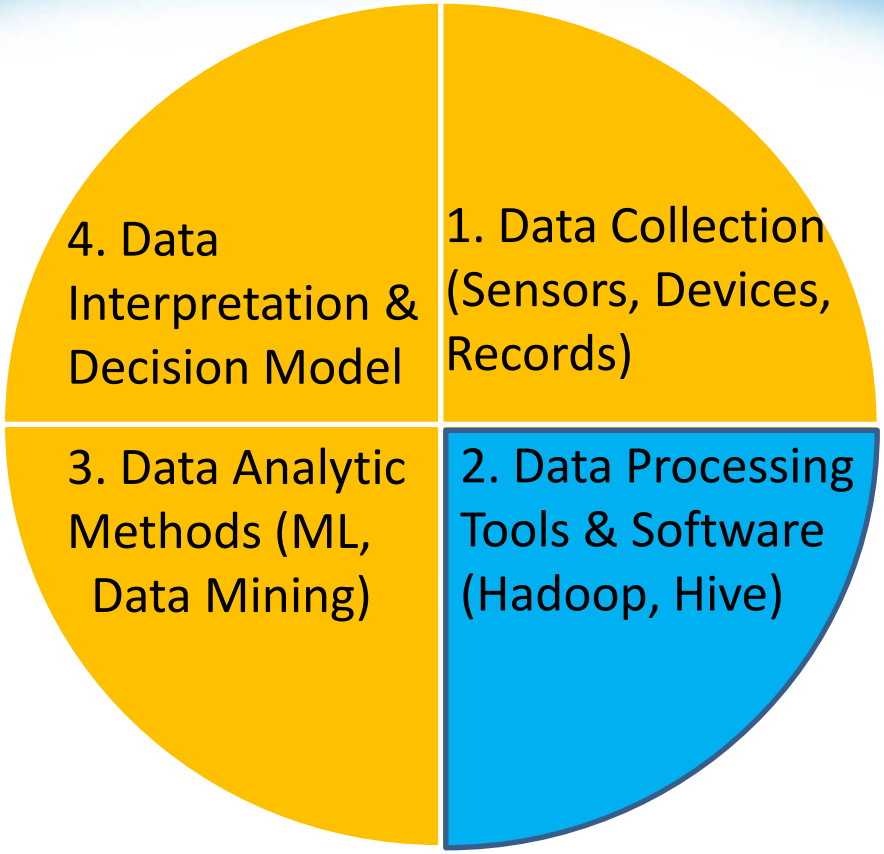
- John Wanamaker once said, “I know that half of my advertising doesn’t work. The problem is I don’t know which half.”

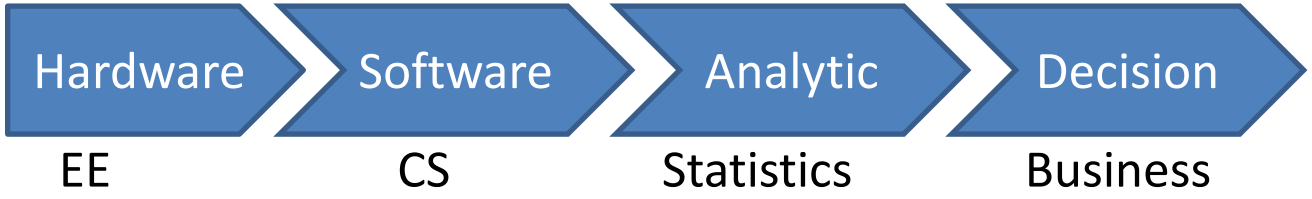
Moving from Big Data to Smart Data is a multistep process.



- "The issue is not about the volume of data but the ability to analyze and act on data in real time." http://tinyurl.com/atcanjw

Ingredients to Success

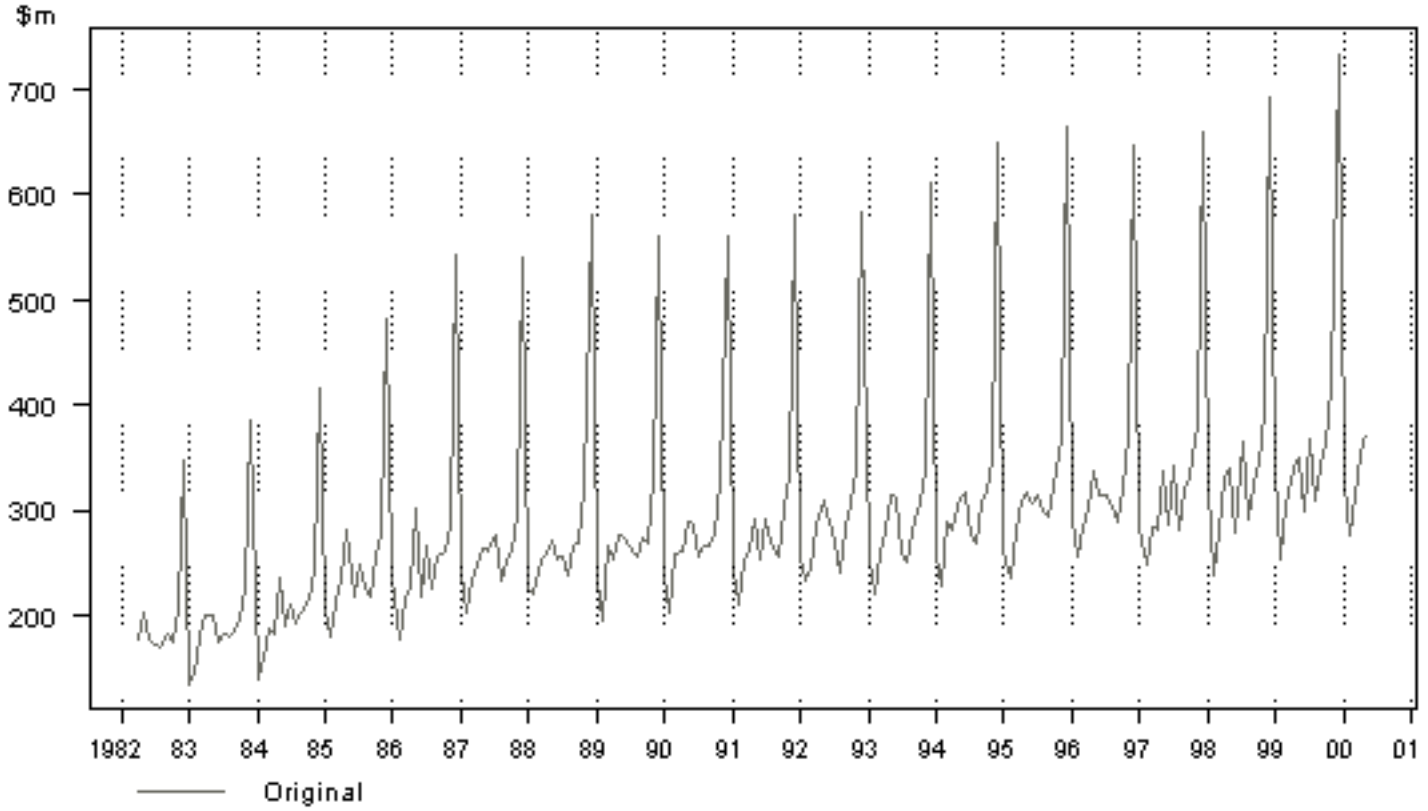




Outline

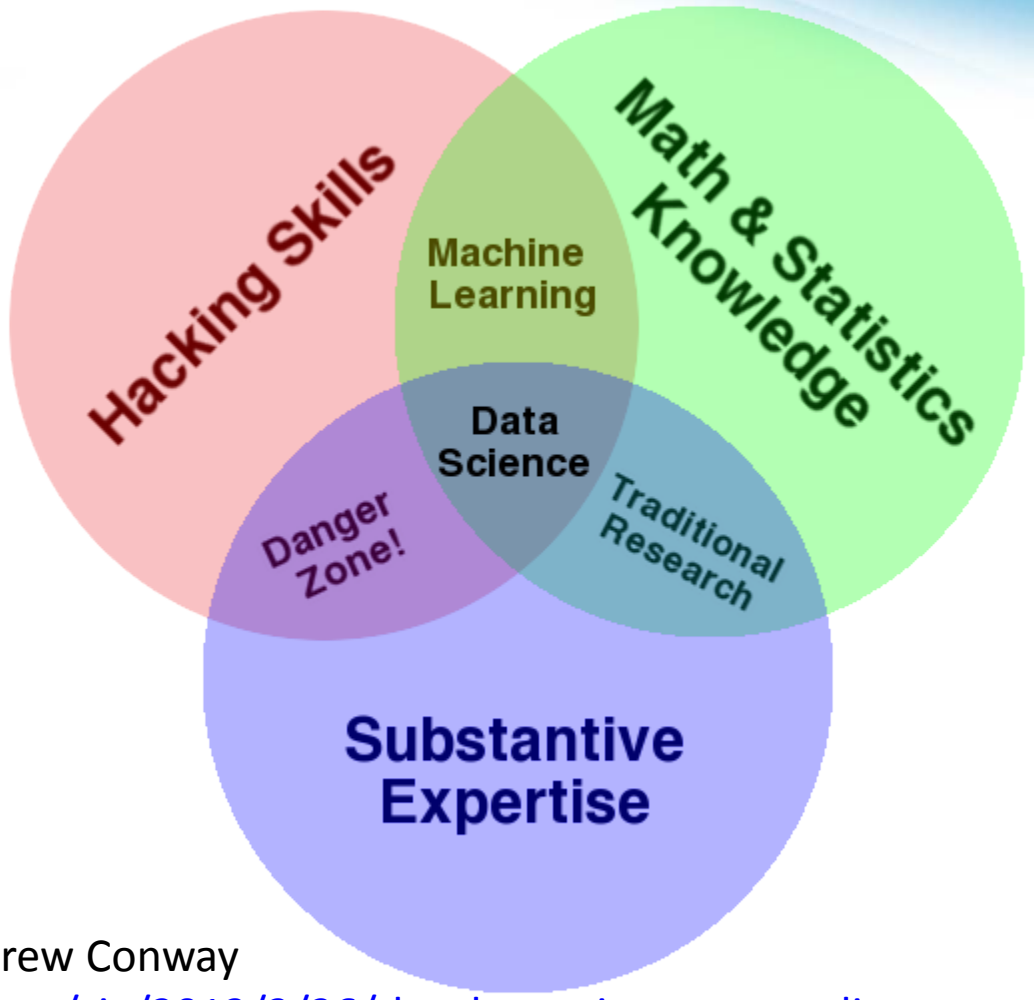
- Introduction
- Big data
- Data science
- Data processing tools
- Conclusion

How to extract insight from data?



Monthly Retail Sales in New South Wales (NSW) Retail Department Stores

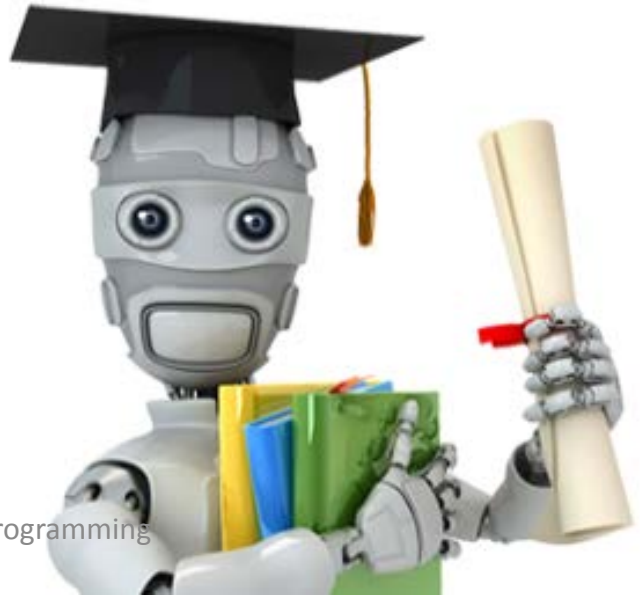
Data Science?



From data scientist Drew Conway
http://drewconway.com/zia/2013/3/26/the-data-science-venn-diagram

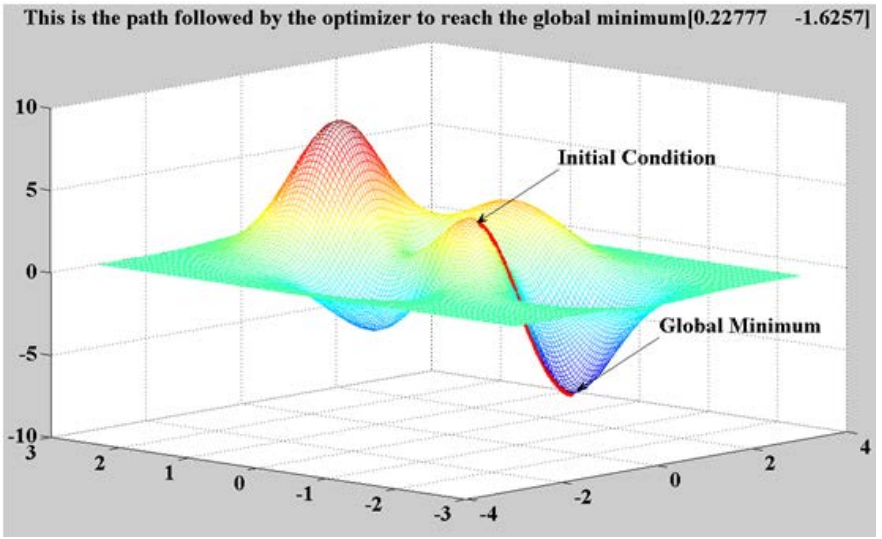
Two orthogonal aspects

- Analytics / machine learning
  - learning insights from data
- Big Data:
  - handling massive data volumes
- Can be combined, or used separately



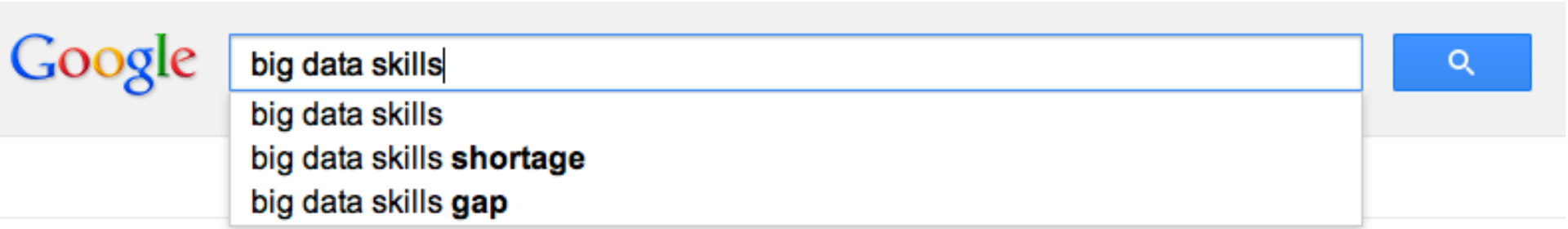
Types of algorithms

- Clustering
- Association learning
- Parameter estimation
- Recommendation engines
- Classification
- Similarity matching
- Neural networks
- Bayesian networks
- Genetic algorithms



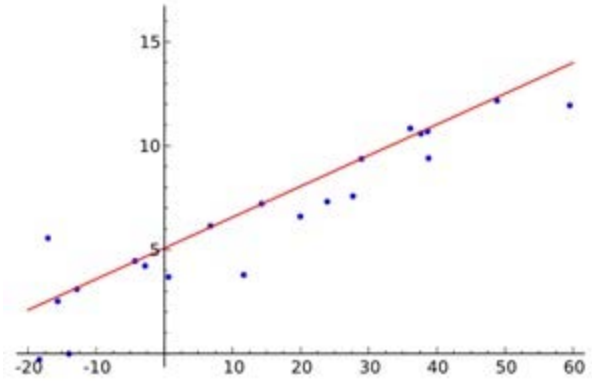
Big data skills gap

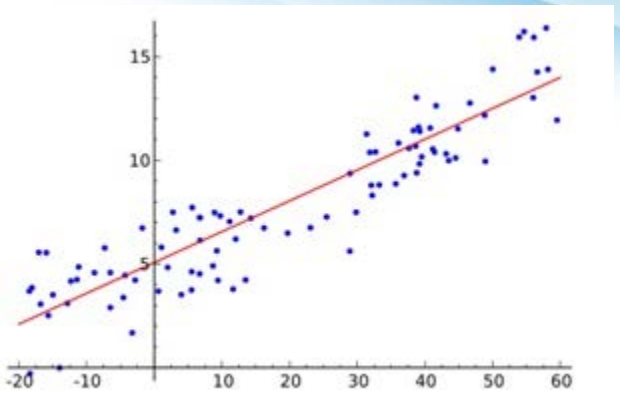
- Hardly anyone knows this stuff
- It's a big field, with lots and lots of theory
- And it's all maths, so it's tricky to learn

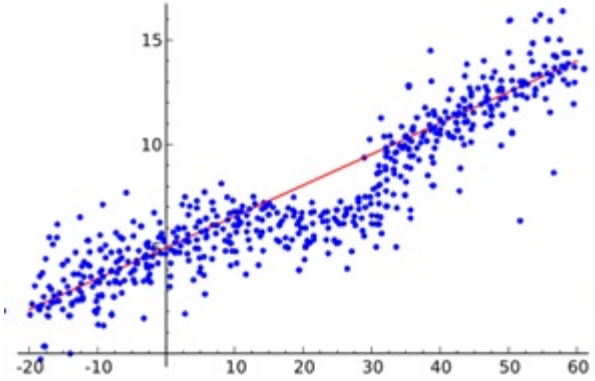


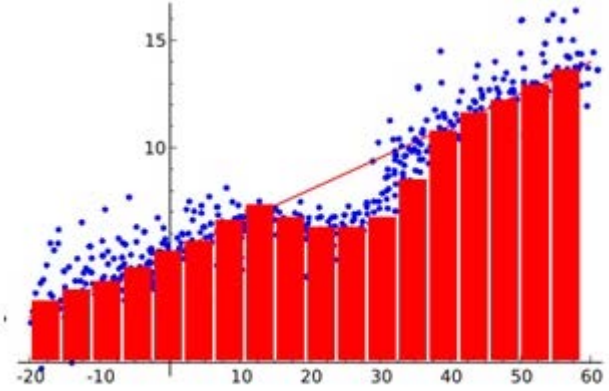
http://www.ibmbigdatahub.com/blog/addressing-big-data-skills-gap
http://wikibon.org/wiki/v/Big_Data:_Hadoop,_Business_Analytics_and_Beyond#The_Big_Data_Skills_Gap

More data reveals non-linear relationship in the dataset



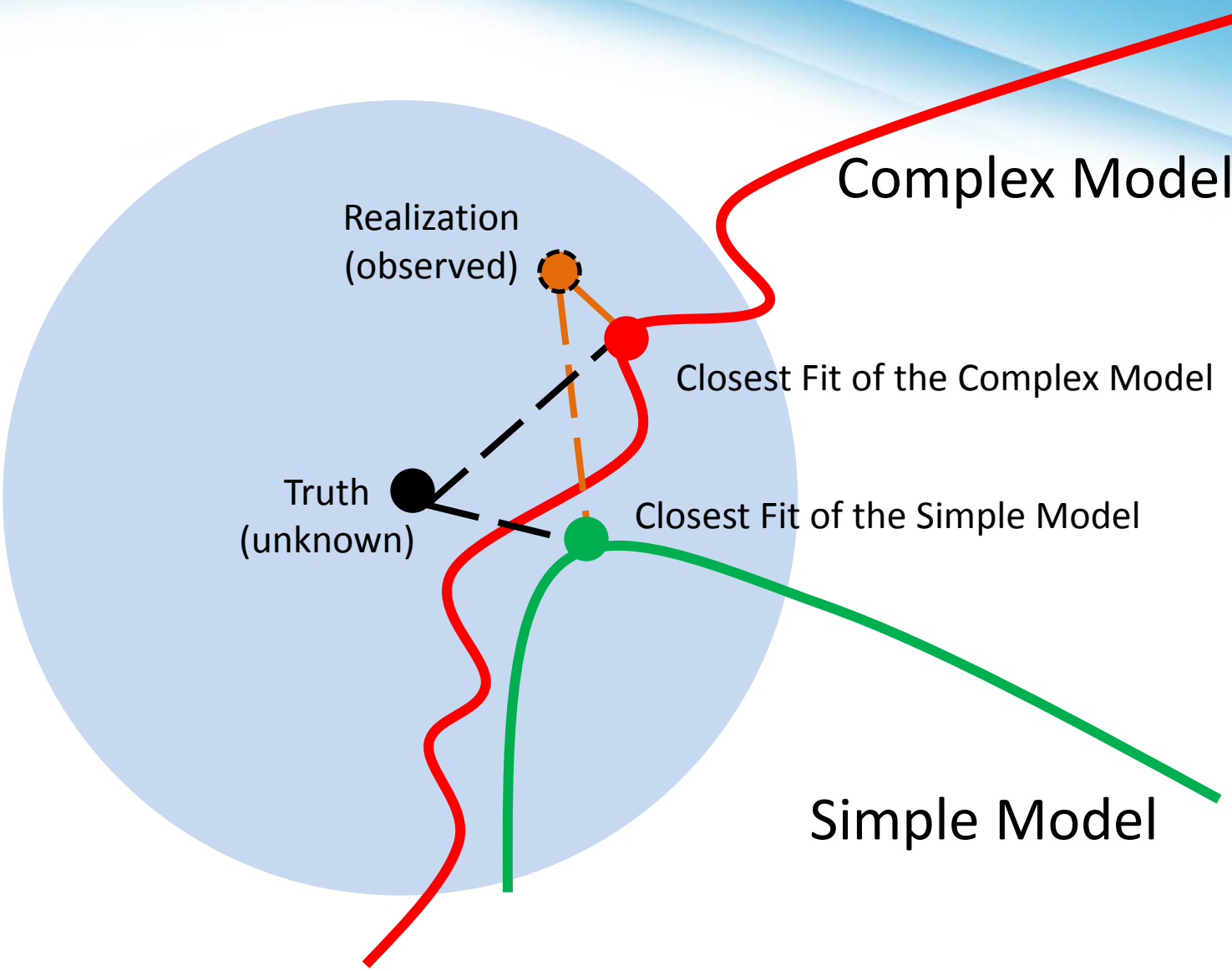






http://goo.gl/nZ5vZl

Why big data is helpful for data mining?

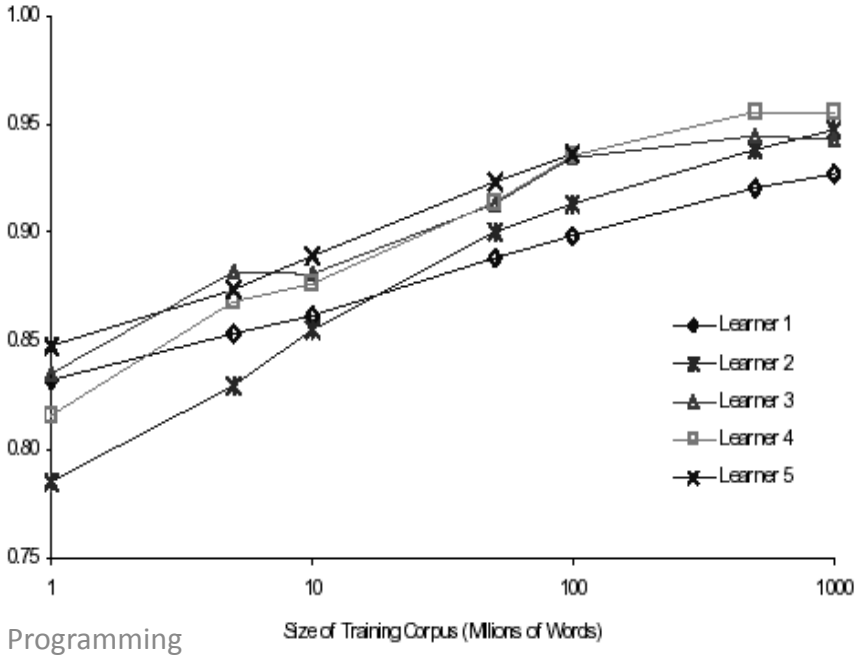


Modified from Fig 7.2 of "The elements of statistical learning" by T. Hastie et al.

A better algorithm, or more data?

- Task: Confusion set disambiguation:
  {you're | your}, {to | too | two}, {its | it's}
- 5 Algorithms: n-gram table, winnow, perceptron, transformation-based learning, decision trees

- Training: $10^6 \rightarrow 10^9$ words
- Lessons:
  - All methods improved to almost same point
  - Simple method can end above complex one



Banko and Brill, HLT 2001
Hovy 2011

Algorithm vs data

Big Data paradigm:

- When you are data poor
  - Not much you can do unless you have a good theory
    - Throw data away, rely only on the algorithm or the model
- When you are data rich
  - Let the data speak for itself

Language check using n-grams

- 全台大停電
  - 全台/大/停電
  - 全/台大/停電
- nowisthetimeforallgoodmentocometothe
  - Using dictionary and grammar for parsing
  - Let the data speak for itself
    - Probability of a segmentation

    = P(first word) x P(rest | first) ≈ P(first word) x P(rest)
    - Best segmentation = one with highest probability
    - Best segmentation of (“nowisthetimeforall…”)
    - P(“n”) x P(“owisthetimeforall…”)
    - P(“no”) x P(wisthebesttimeforall…”)
    - P(“now”) x P(isthebesttimeforall…”)
    - P(“nowi”) x P(“sthebesttimeforall…”)

Know-hows vs Know-whys

- Walmart found that, on Friday afternoons, young American males who buy diapers also have a predisposition to buy beer

- Why?
  - Multiple different explanations
    - Many new fathers like beer, but they don't go to bars because they need to take care of babies
    - New fathers buy beer to celebrate a newborn with friends
    - Other explanations?

Danger of Big Data

1. Big data is very good at detecting correlations, but **it never tells us which correlations are meaningful**.
   - Example: from 2006 to 2011 the United States murder rate was well correlated with the market share of Internet Explorer: Both went down sharply. But it's hard to imagine there is any causal relationship between the two.

http://www.nytimes.com/2014/04/07/opinion/eight-no-nine-problems-with-big-data.html

Danger of Big Data

2. Big data is at its best when analyzing things that are extremely common, but often falls short when analyzing things that are less common.
   - For example: no existing body of data will ever be large enough to include all the trigrams that people might use, because of the continuing inventiveness of language.

3. Statistics vs. Specifies
   - Big data is all about statistics: divining patterns and trends from large data sets.
   - Real challenge is information logistics**: how to get exactly the right information to, and from, the right people in the right formats at the right time.**

Danger of Big Data

4. Big data can only capture the past
   - Without theory, we cannot predict into a changing future

5. Data may never be big enough
   - The data you collect is useful. But it has holes which you need to fill. Which leads you to collect more data. But this data still has holes. So you need to collect even more data. *Ad infinitum*.

Can big data discover $E=mc^2$?

- Essentially, all models are wrong, but some are useful

-- George E. P. Box

Climbing to the Moon

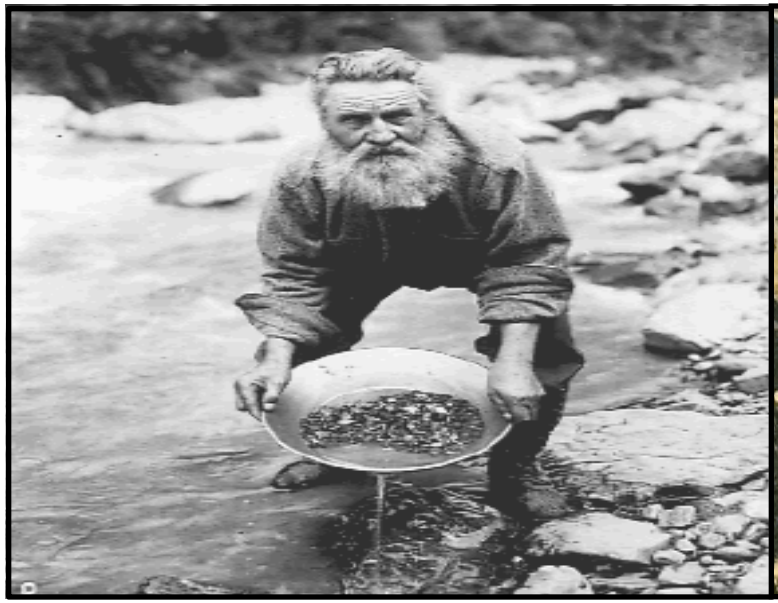
- The Big Data bandwagon is supposedly driven by 'evidence' . Trouble is, it's a **misleading sort of evidence**.
  - Imagine a world where every inch you got closer to the moon, you got rewarded. First, you climb the highest mountain. Then you start building skyscrapers, shouting all the while 'see, we're getting closer the moon. We have demonstrable evidence of success. Look at the rewards we are getting!'
  - But if you really want to get to the moon, you don't climb mountains and build skyscrapers. You build rockets. A completely different activity.

https://www.ctrl-shift.co.uk/index.php/news/2012/01/17/big-data-big-dead-end/

Outline

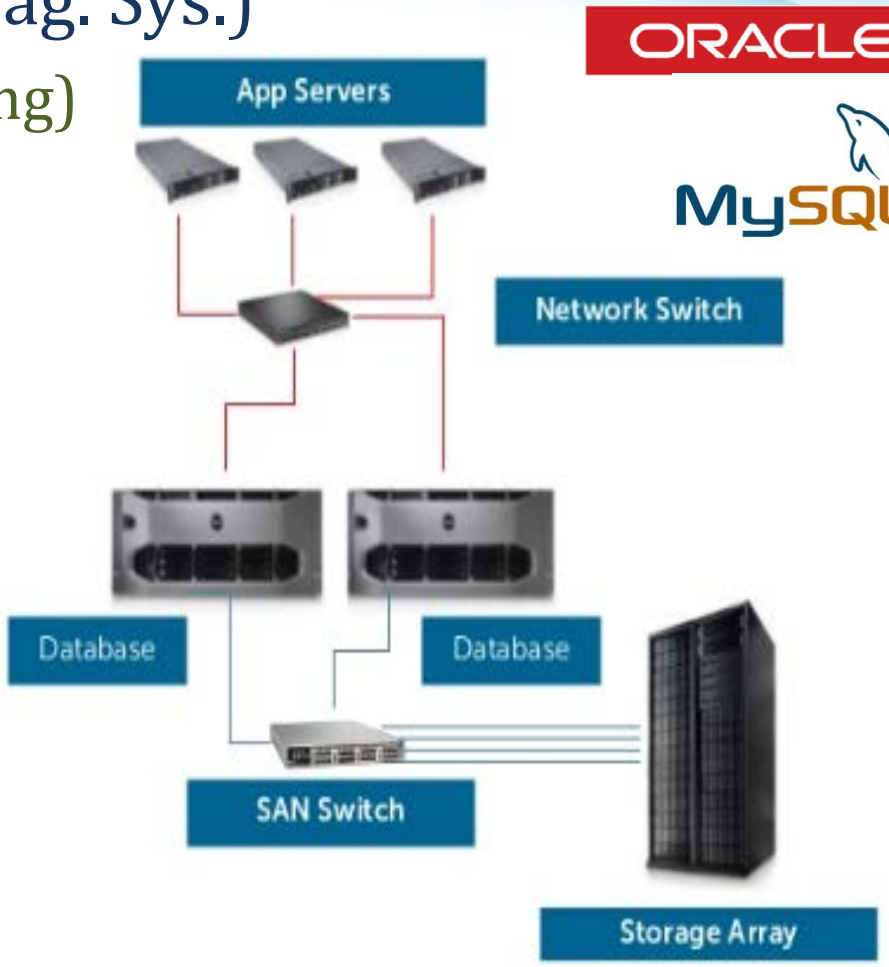
- Role of Cloud
- Big data
- Data science
- Data processing tools
- Conclusion

Why Not All of Big Data Before: Didn't have the Tools?



Conventional Solution

- RDBMS (Relational DB Mag. Sys.)
  - High availability (Clustering)
  - Standard Database
  - Business Intelligent Data Warehousing
  - Structure Data



Car

CarKey	MakeKey	ModelKey	ColorKey	Year
1	1	1	2	2003
2	2	1	3	2005
3	2	1	2	2005

Color

ColorKey	Color
1	Red
2	Green
3	Blue

MakeModel

ModelKey	MakeKey	Model
1	1	Pathfinder
1	2	Bluebird
2	1	Civic

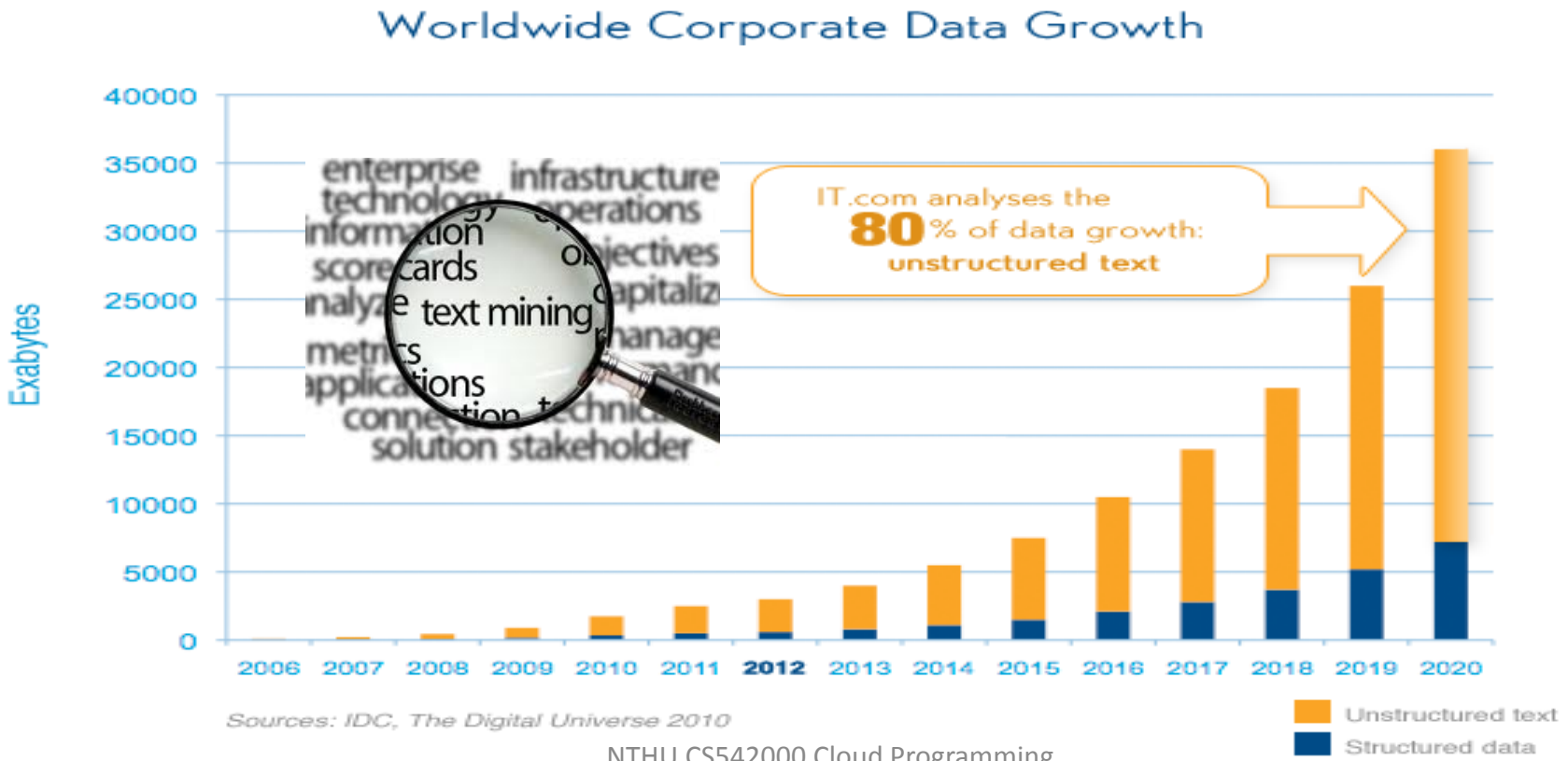
Make

MakeKey	Make
1	Nissan
2	Honda

Example of a Typical Relational Data Model

1. Growth of Unstructured Data

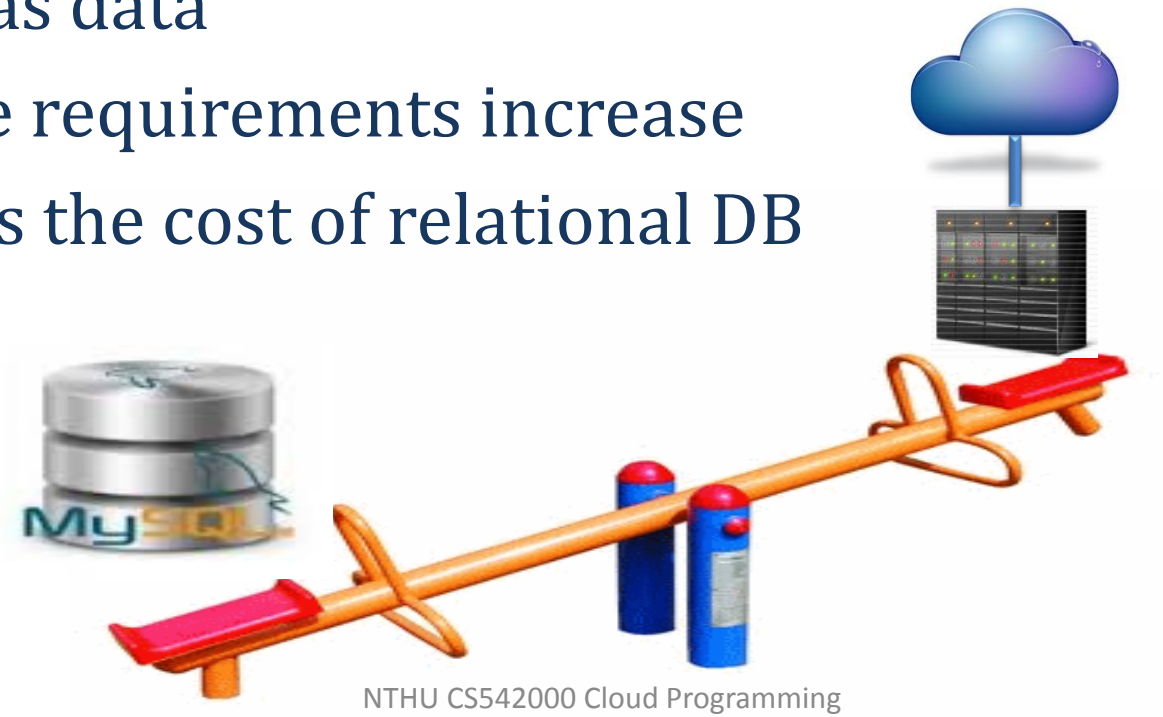
- Data can be of any type (Variety)
  - Not necessarily following any format or sequence
  - Not follow any rules, so is not predictable



2. Cost Expense

- The software license and expertise required to implement a relational database
- The ongoing costs of system maintenance and staff as data
- Space requirements increase

  raises the cost of relational DB



3. Scalability

- RDBMS scale well on a single node, but difficulty scaling out with more commodity hardware in parallel
  - Strong integrity requirement
    - ACID property: Atomicity, Consistency, Isolation, Durability
  - More expensive hardware (less fault tolerant)
    - High availability, reliability device (e.g. RAID)
  - Limited by legacy software architecture that was designed since 1970's~1980's

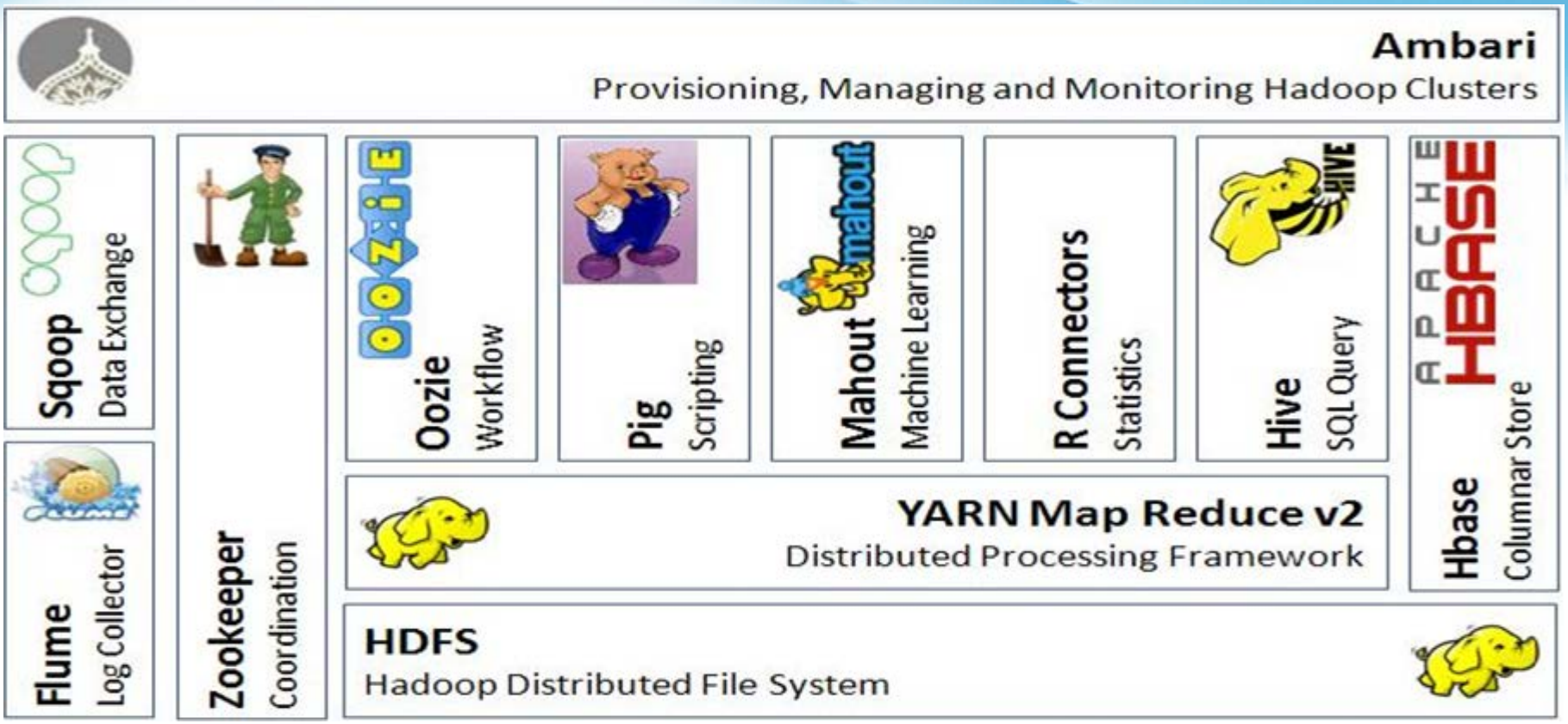
New data processing tools

- Platform & infrastructure
  - Public cloud services
- Distribute file systems
  - GFS, HDFS
- Parallel data processing
  - Batch: MapReduce, Hadoop, Spark
  - Streaming: STORM
  - Graph: GPS, GraphLab, Xgraph
- Database
  - NoSQL: 75 or so vendors giving up both SQL and ACID
  - NewSQL: Retain SQL and ACID but go fast with a new architecture

Google Data Processing Stack

- Google published the designs of **web-search engine**
  - SOSP 2003
    - The **G**oogle **F**ile **S**ystem
  - OSDI 2004
    - **MapReduce** : Simplified Data Processing on Large Cluster
  - OSDI 2006
    - **Bigtable**: A Distributed Storage System for Structured Data

Apache Hadoop Ecosystem

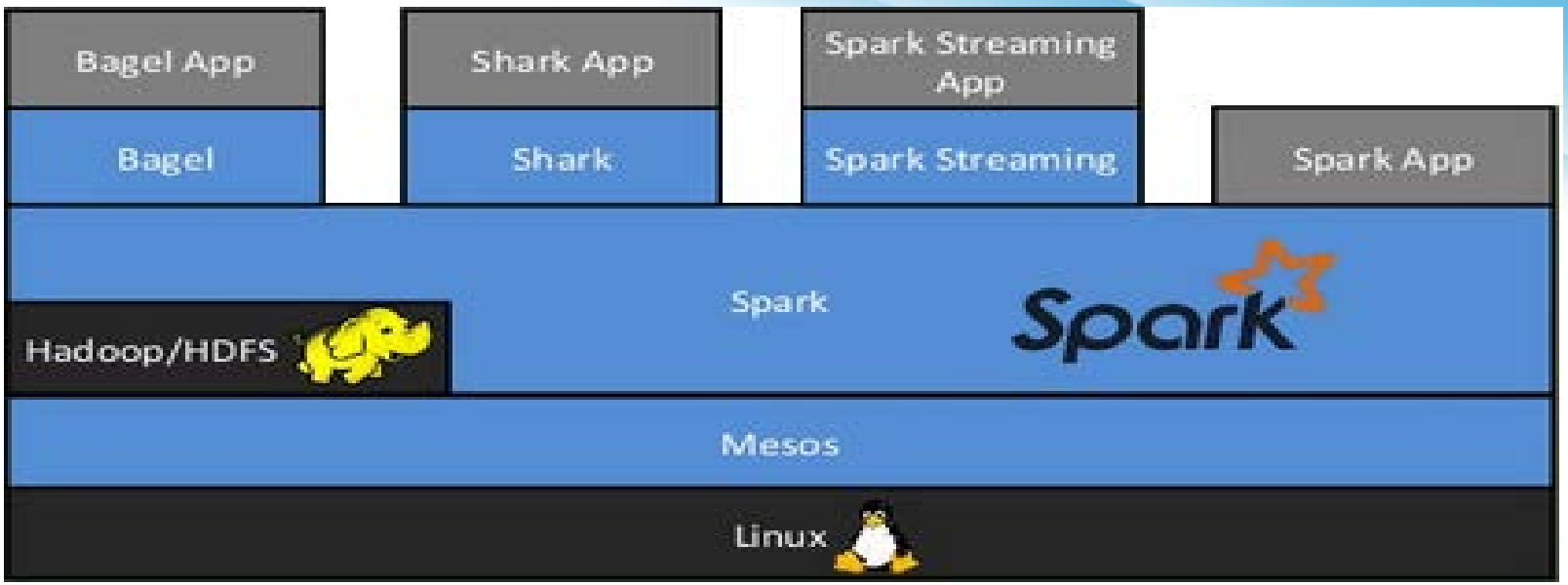


- **HDFS**: Hadoop Distributed *File System*
- **MapReduce**: A parallel *data processing framework*
- **Hbase**: Key-value *structured database*
- **Hive**: A high-level *SQL-like language*

Facebook's Hadoop Warehouse (2011)

- 2700 nodes
  - each with 8CPU, 32-48GB memory, 12 disks (1TB or 2TB)
- 19 PB of data in HDFS
  - 50.4 PB with replication
- 15 TB (compressed) data added daily
  - 40 TB new data
  - 110 TB of derived tables
- 150K jobs processed daily
  - Only 500 are MapReduce jobs. Rest in **Hive**

BDAS (Berkeley Data Analytics Stack)



- **Spark**: in-memory computing engine achieves 10x speedup over hadoop
- **Spark streaming**: small batch processing
- **Bagel**: Pregal-like graph programing interface
- **Shark**: SQL-like query interface

Outline

- Role of Cloud
- Big data
- Data science
- Data processing tools
- Conclusion

Conclusion

- The Original Big Data
  - Data with the property of 3 Vs: volume, velocity, variety, veracity.
- Big Data as Technology
  - New tools and technology developed for processing big data, such as Hadoop, NoSQL.
- Big Data as Value
  - Making better-informed decisions, discovering hidden insights and automating business processes.
- Big Data as Opportunity
  - An opportunity to make new discovery by analyzing **dark data** that was previously ignored because of technology limitations

The 15 hottest skills of 2014 on LinkedIn

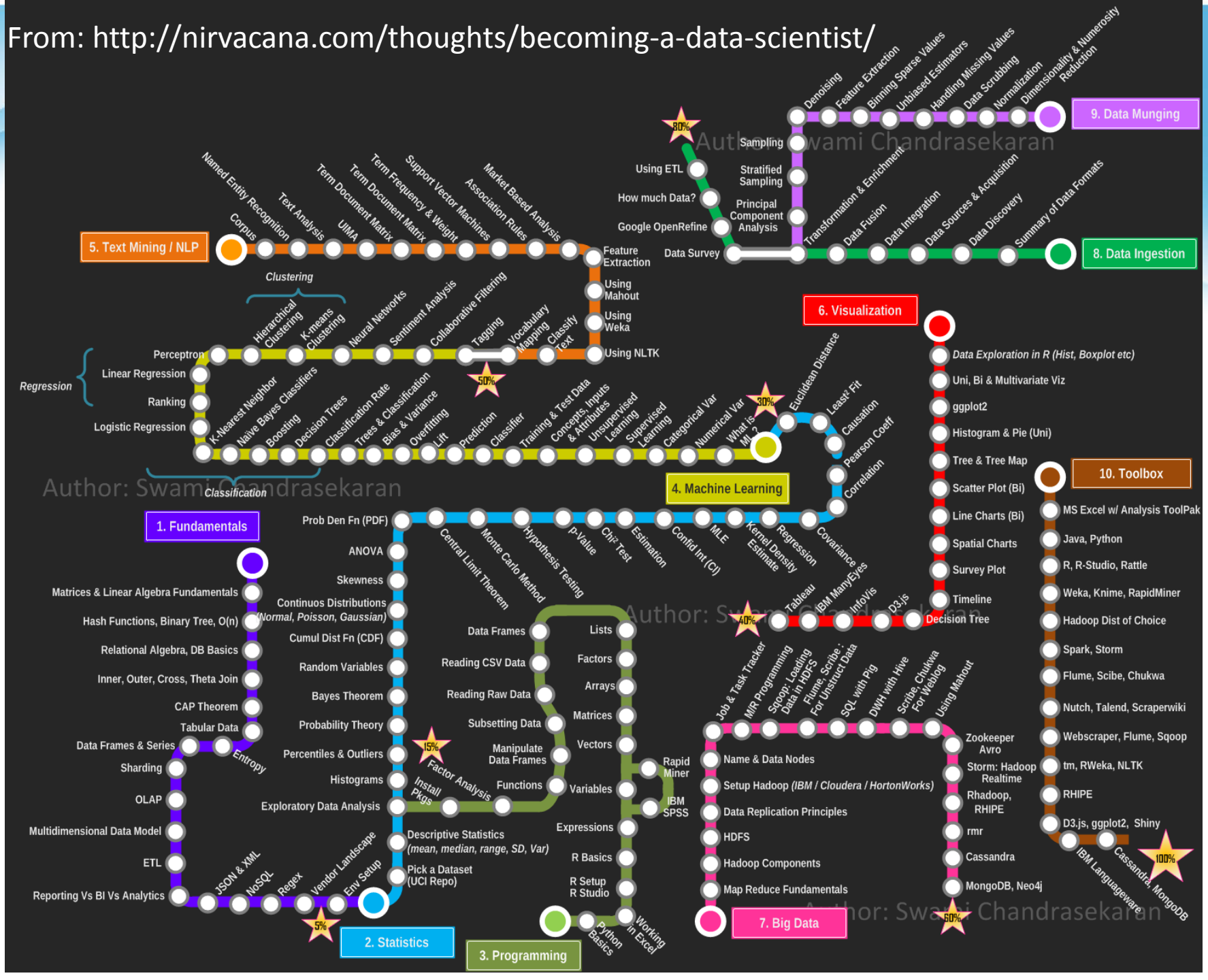
Global

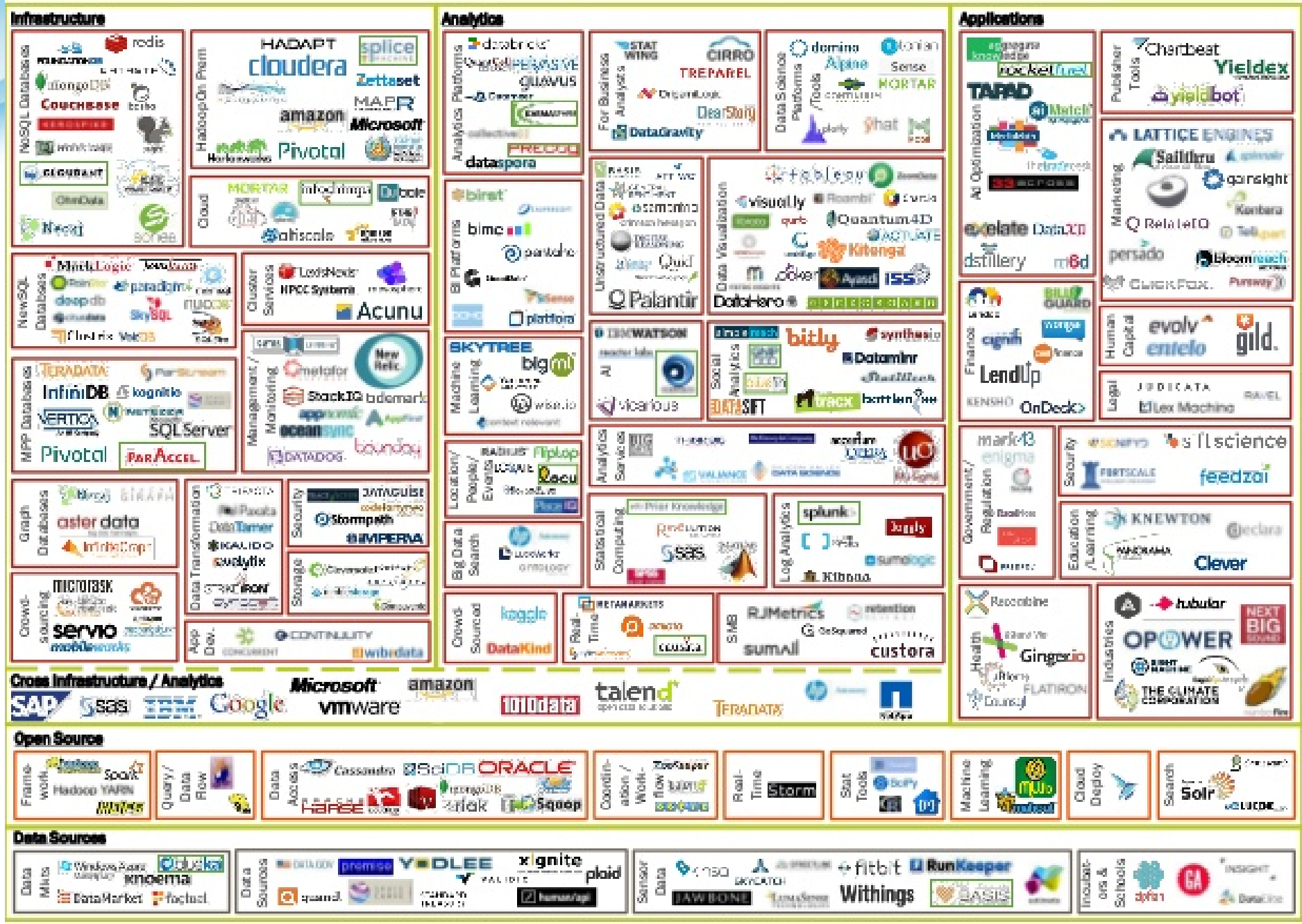
1. Statistical Analysis and Data Mining
2. Middleware and Integration Software
3. Storage Systems and Management
4. Network and Information Security
5. SEO/SEM Marketing
6. Business Intelligence
7. Mobile Development
8. Web Architecture and Development Framework
9. Algorithm Design
10. Perl/Python/Ruby
11. Data Engineering and Data Warehousing
12. Marketing Campaign Management
13. Mac, Linux and Unix Systems
14. User Interface Design
15. Recruiting

United States

1. Cloud and Distributed Computing
2. Statistical Analysis and Data Mining
3. Middleware and Integration Software
4. Network and Information Security
5. Mobile Development
6. Foreign Language Translation
7. Storage Systems and Management
8. Mac, Linux and Unix Systems
9. Java Development
10. Perl/Python/Ruby
11. Algorithm Design
12. Digital and Online Marketing
13. Computer Graphics and Animation
14. Data Engineering and Data Warehousing
15. Software QA and User Testing

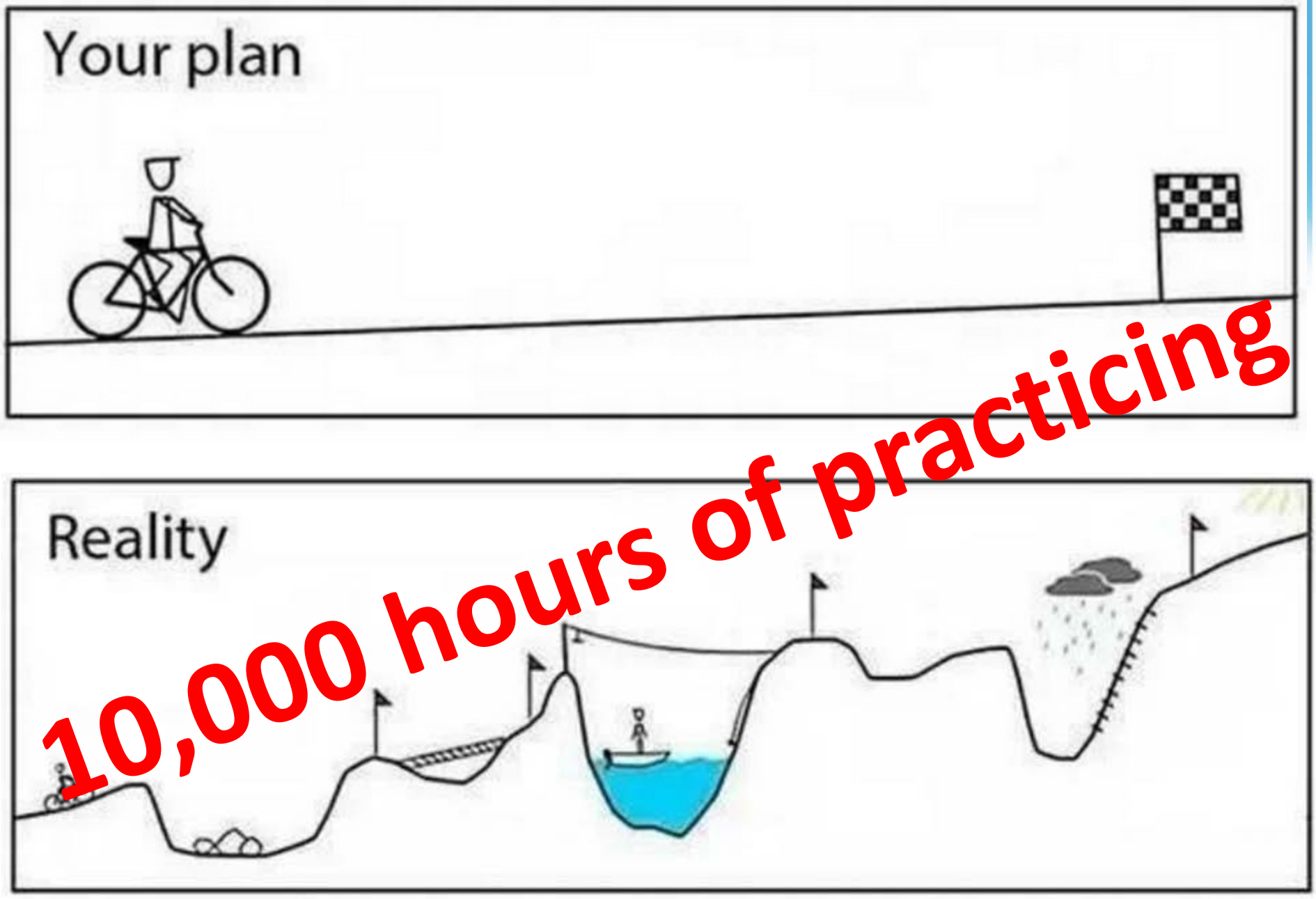
From: http://nirvacana.com/thoughts/becoming-a-data-scientist/





© Matt Turck (@mattturck), Sutian Dong (@sutiandong) & FirstMark Capital (@firstmarkcap)

http://www.slideshare.net/mjft01


Your plan
Reality
10,000 hours of practicing

Reference

- Lars Marius Garshol: Introduction to Machine Learning
- McKinsey Global Institute: Big data: The next frontier for innovation, competition, and productivity,
- Stavros Harizopoulos, Daniel J. Abadi, Samuel Madden, Michael Stonebraker: OLTP through the looking glass, and what we found there. SIGMOD Conference 2008: 981-992
- Hung-Hsuan Chen: The theory and practice on studying scholarly big data
- https://www.ctrl-shift.co.uk/index.php/news/2012/01/17/big-data-big-dead-end/
- http://www.nytimes.com/2014/04/07/opinion/eight-no-nine-problems-with-big-data.html