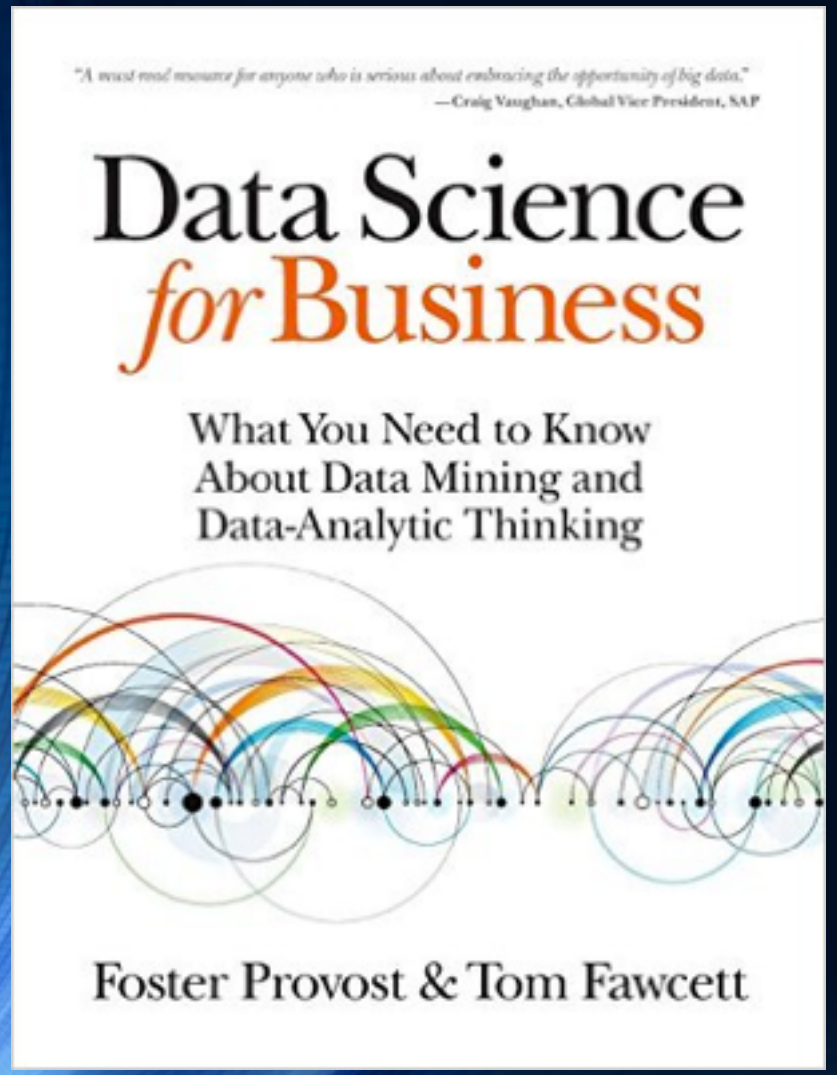
CSCI E-84
A Practical Approach to Data Science

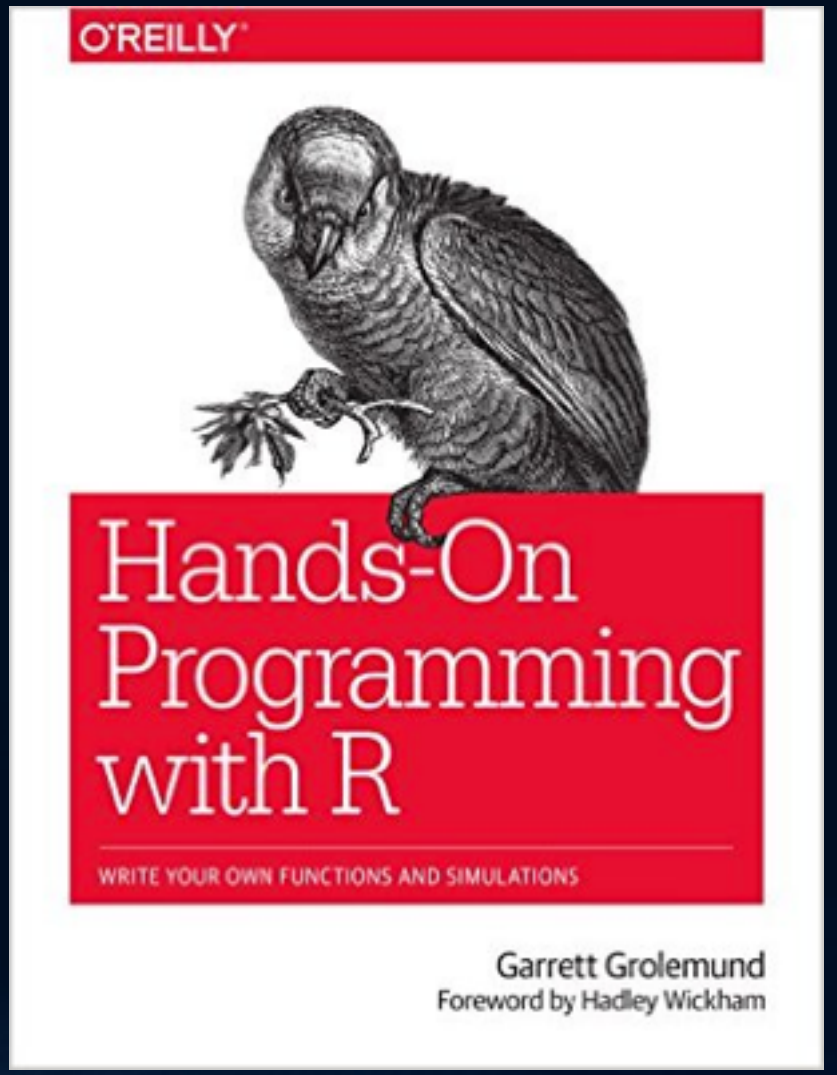
Ramon A. Mata-Toledo, Ph.D.
Professor of Computer Science
Harvard Extension School

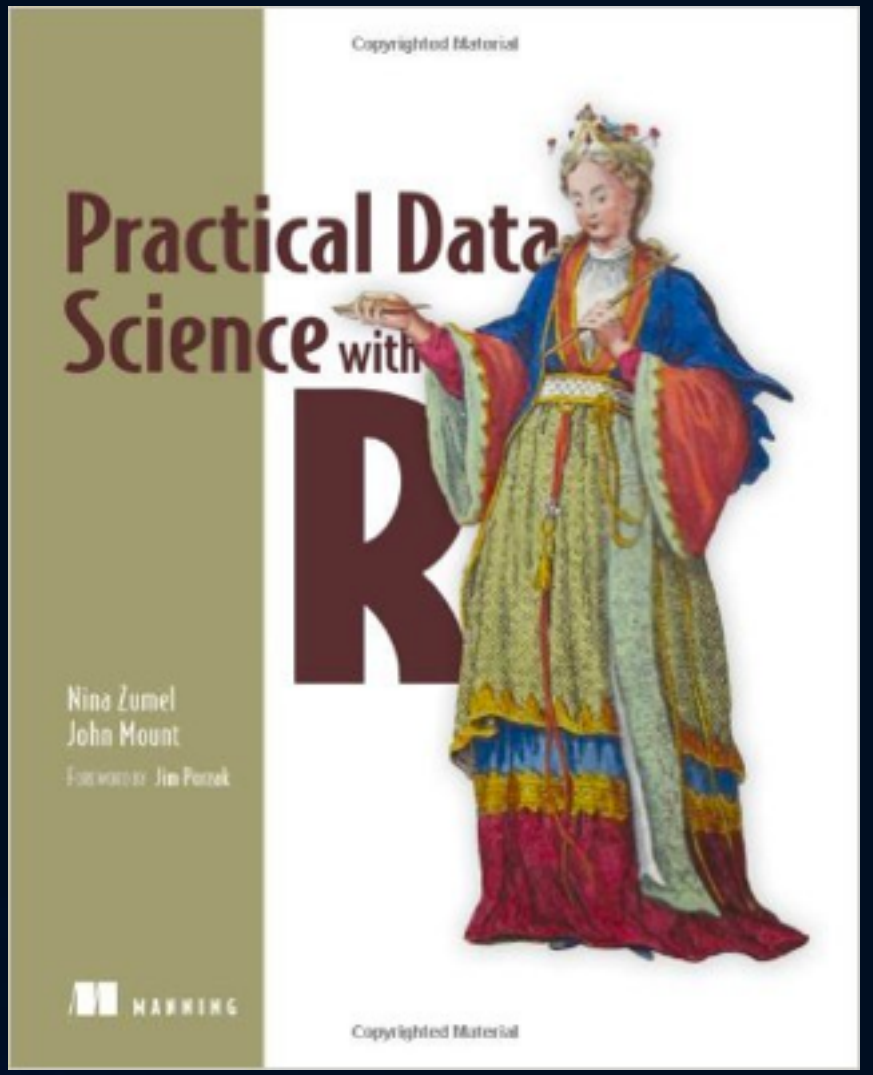
Unit 1 - Lecture 1

Wednesday, August 31, 2016

Required Textbooks







Main Objectives of the Course

This course is an introduction to the field of data science and its applicability to the business world.

At the end of the semester you should be able to:

- Understand how data science fits in an organization and be able to gather, select, and model large amounts of data.
- Identify the suitable data characteristics and the methods that will allow you to extract knowledge from it using the programming language R.

What This Course is NOT

An in-depth calculus based statistics research course.

What is this course all about?

It IS a practical "hands-on approach" to understanding the methodology and tools of Data Science from a business prospective. Some of the technical issues we will briefly consider are:

- Statistics
- Database Querying using SQL
- Regression Analysis
- Explanatory versus Predictive Modeling
- Data Warehousing

Brief History of Data Science/Big Data

The term data science can be considered the result of three main factors:

- Evolution of data processing methods
- Internet – particularly the Web 2.0
- Technological Advancements in computer processing speed/ storage and development of algorithms for extracting useful information and knowledge from big data

Brief History of Data Science/Big Data (continuation)

However, "…Already seventy years ago we encounter the first attempts to quantify the growth rate in the *volume of data* or what has popularly been known as the "information explosion" (a term first used in 1941, *Oxford English Dictionary*). "

Selected Milestones in the History of Data Science/Big Data

Theoretical/Academic:

(Source: http://www.forbes.com/sites/gilpress/2013/05/09/a-very-short-history-of-big-data/)

- November 1967 - B. A. Marron and P. A. D. de Maine publish "Automatic data compression" in the *Communications of the ACM*, stating that "*The 'information explosion' noted in recent years makes it essential that storage requirements for all information be kept to a minimum*."

Theoretical/Academic

- April 1980 - I.A. Tjomsland gives a talk titled "Where Do We Go From Here?" at the Fourth IEEE Symposium on Mass Storage Systems, in which he says "Those associated with storage devices long ago realized that Parkinson's First Law may be paraphrased to describe our industry—

Data expands to fill the space available'…. I believe that large amounts of data are being retained because users have no way of identifying obsolete data; the penalties for storing obsolete data are less apparent than are the penalties for discarding potentially useful data."

Theoretical/Academic

- July 1986 - Hal B. Becker publishes "Can users really absorb data at today's rates? Tomorrow's?" in *Data Communications*. Becker estimates that

"*the recoding density achieved by Gutenberg was approximately 500 symbols (characters) per cubic inch—500 times the density of [4,000 B.C. Sumerian] clay tablets. By the year 2000, semiconductor random access memory should be storing 1.25X10^11 bytes per cubic inch*."

Here we are making the basic assumption 1 character = 1 byte (8 binary digits)

Theoretical/Academic

- 1996 - *Digital storage becomes more cost-effective for storing data than paper* according to R.J.T. Morris and B.J. Truskowski, in "The Evolution of Storage Systems" *IBM Systems Journal*, July 1, 2003.

- 1997 - Michael Lesk publishes "How much information is there in the world?" Lesk concludes that "*There may be a few thousand petabytes of information all told; and the production of tape and disk will reach that level by the year 2000. So in only a few years, (a) we will be able [to] save everything–no information will have to be thrown out, and (b) the typical piece of information will never be looked at by a human being*."

Theoretical/Academic

August 1999 - Steve Bryson, David Kenwright, Michael Cox, David Ellsworth, and Robert Haimes publish "Visually exploring gigabyte data sets in real time" in the *Communications of the ACM*. It is the first CACM article to use the term "Big Data" (the title of one of the article's sections is "Big Data for Scientific Visualization"). The article opens with the following statement:

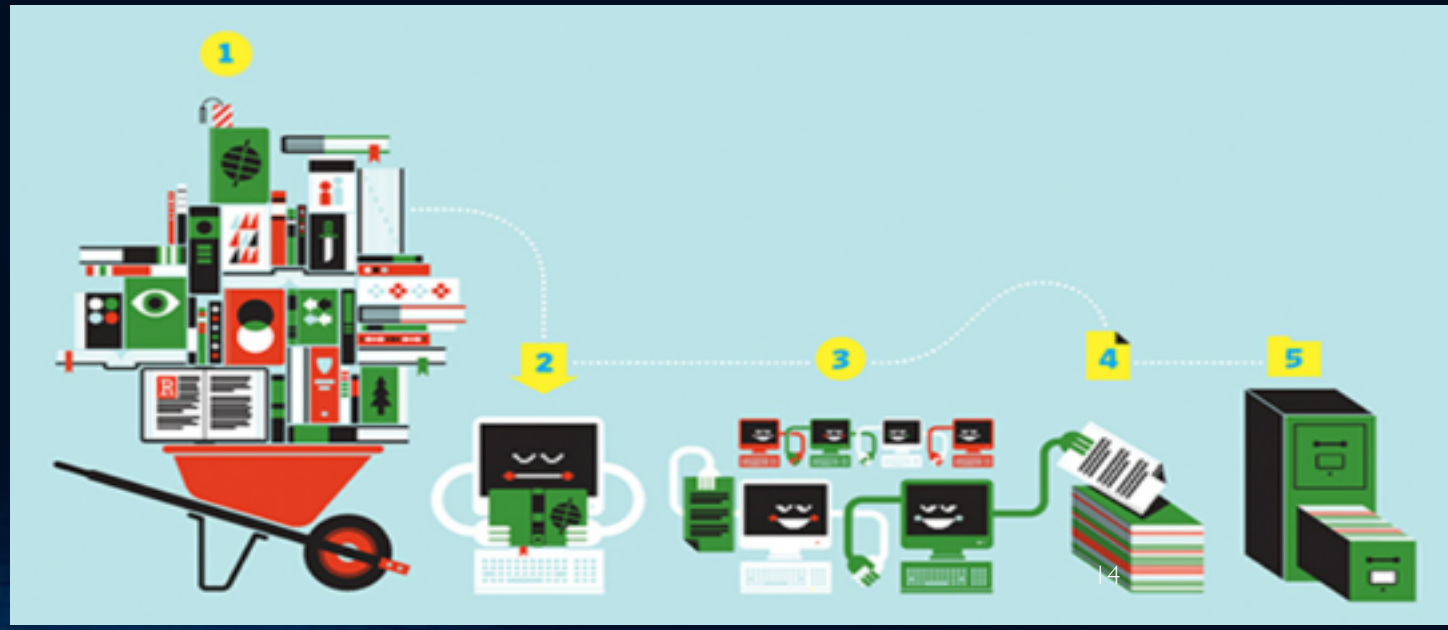
"*Very powerful computers are a blessing to many fields of inquiry. They are also a curse; fast computations spew out massive amounts of data. Where megabyte data sets were once considered large, we now find data sets from individual simulations in the 300GB range. But understanding the data resulting from high-end computations is a significant endeavor; …it is just plain difficult to look at all the numbers. And as Richard W. Hamming, mathematician and pioneer computer scientist, pointed out, the purpose of computing is insight, not numbers.*"

October 2000 - Peter Lyman and Hal R. Varian at UC Berkeley publish "How Much Information?"

It is the first comprehensive study to quantify, in computer storage terms, the total amount of new and original information (not counting copies) created in the world annually and stored in four physical media: paper, film, optical (CDs and DVDs), and magnetic.

The Google White Papers

- Foreshadowing of the data explosion
- Developed 2 white papers
  - Map Reduce (http://static.googleusercontent.com/media/research.google.com/en//archive/mapreduce-osdi04.pdf)
  - Google File System (GFS) (http://static.googleusercontent.com/media/research.google.com/en//archive/gfs-sosp2003.pdf)
- Basic precept was based on a distributed file system
  - Clusters of smaller, cheaper computers



Selected Software for Handling Big Data

SPSS is a widely used program for statistical analysis in social science. It is also used by industry and academic researchers.

The original SPSS manual (Nie, Bent & Hull, 1970) has been described as one of "sociology's most influential books" for allowing ordinary researchers to do their own statistical analysis.

In addition to statistical analysis, data management (case selection, file reshaping, creating derived data) and data documentation (a metadata dictionary was stored in the datafile) are features of the base software.

SPSS was acquired by IBM in 2009.

Selected Software for Handling Big Data

SAS was developed at North Carolina State University from 1966 until 1976, when SAS Institute was incorporated.

SAS was further developed in the 1980s and 1990s with the addition of new statistical procedures, additional components and the introduction of JMP.

A point-and-click interface was added in version 9 in 2004.

A social media analytics product was added in 2010.

Software Tools – (continuation)

R is a programming language and free software environment for statistical computing and graphics created, in 1997, by Ross Ihaka and Robert Gentleman at the University of Auckland, New Zealand.

R is named after the initials of the first names of its authors and as a play on the name of *S, the* programming language on which it is based.

R can be downloaded from https://cran.r-project.org/ (The Comprehensive R Archive Network)

Source:
(https://www.stat.auckland.ac.nz/~ihaka/downloads/Interface98.pdf).

Software Tools (Continuation)

MongoDB is a cross-platform document-oriented database developed by MongoDB Inc., in 2007.

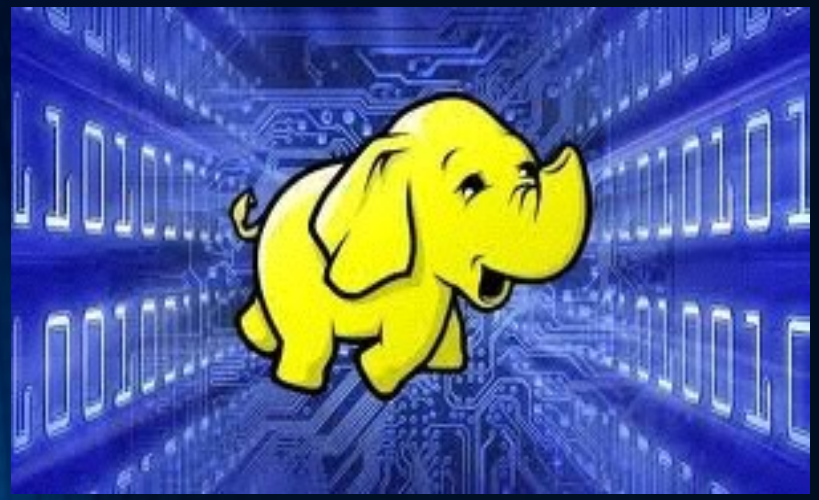
Classified as a NoSQL database, MongoDB eschews the traditional table-based relational database structure in favor of JSON-like documents with dynamic schemas (MongoDB calls the format BSON), making the integration of data in certain types of applications easier and faster.

MongoDB is free and open-source software (https://www.mongodb.com/)

Software Tools – (Continuation)

Hadoop

- Open source from Apache [http://hadoop.apache.org/]
- Implementation of the Google White Papers using the programming language Java
- Hadoop is now associated with a collection of technologies

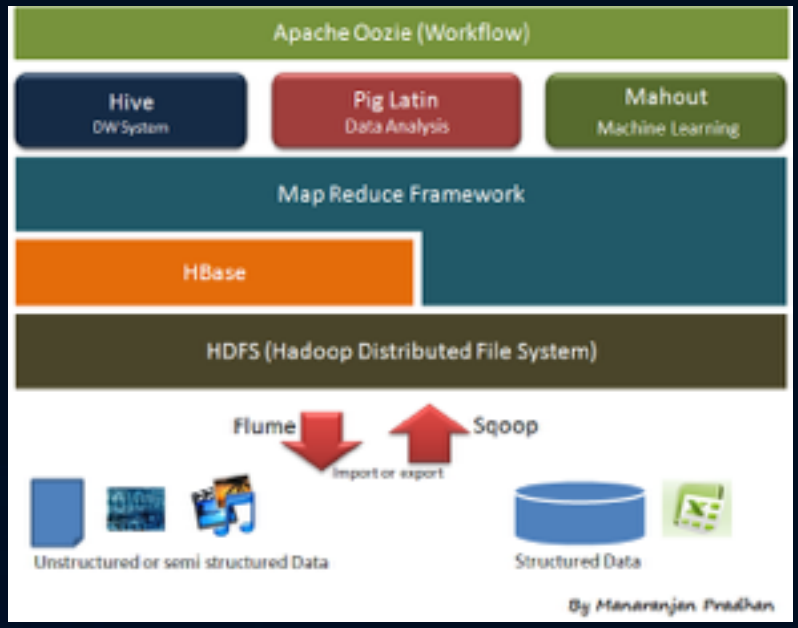


Design Choice

- Hadoop Example
  - Very good at storing files
  - Optimizes use of 'cheap resources' - no RAID needed here
  - Provides data redundancy
  - Good at sequential reads
  - Not so good at high speed random reads
  - Cannot update a file – must replace

Hadoop Ecosystem – other Technologies

- Yet Another Resource Negotiator (YARN)
- Scoop – imports tables from RDBMS
- Flume – deals with event data like web logs
- Hive – imposes metadata over flat data so SQL code can be used
- Impala – high speed analytic using distributed queries
- Hbase – NOSQL db for Hadoop stored data
- Mahout – machine learning algorithms
- Oozie – workflow manager
- Zookeeper – configuration management



Big Data

The processing of massive data sets that facilitate real-time data driven decision-making

Digital data grows by 2.5 quintillion (10^{18}) bytes every day

In practical terms BIG DATA is data that exceed the processing capacity of conventional database systems because it is :

- too big,
- it moves too fast
- does not conform to the structural requirement of traditional databases.

Units of Information storage

The smallest addressable unit in a computer's memory is the byte. A byte is equal to 8 consecutive binary digits or bits.

In the physical sciences a Kilo (K) stand for 1000 units. For example: 1 Km = 1000 meters (m)

1 Kg = 1000 grams (g)

However, in the computer field, when we refer to Kilobyte or Kb the K = 1024 bytes.

Units of Information storage
(continuation)

Larger information units and their names are shown in the figure below.

Source (https://en.wikipedia.org/wiki/Units_of_information)

Symbol	Prefix	SI Meaning	Binary meaning	Size difference
k	kilo	$10^3 = 1000^1$	$2^{10} = 1024^1$	2.40%
M	mega	$10^6 = 1000^2$	$2^{20} = 1024^2$	4.86%
G	giga	$10^9 = 1000^3$	$2^{30} = 1024^3$	7.37%
T	tera	$10^{12} = 1000^4$	$2^{40} = 1024^4$	9.95%
P	peta	$10^{15} = 1000^5$	$2^{50} = 1024^5$	12.59%
E	exa	$10^{18} = 1000^6$	$2^{60} = 1024^6$	15.29%
Z	zetta	$10^{21} = 1000^7$	$2^{70} = 1024^7$	18.06%
Y	yotta	$10^{24} = 1000^8$	$2^{80} = 1024^8$	20.89%

1 ZB is equivalent to 1 followed by 21 zeroes.

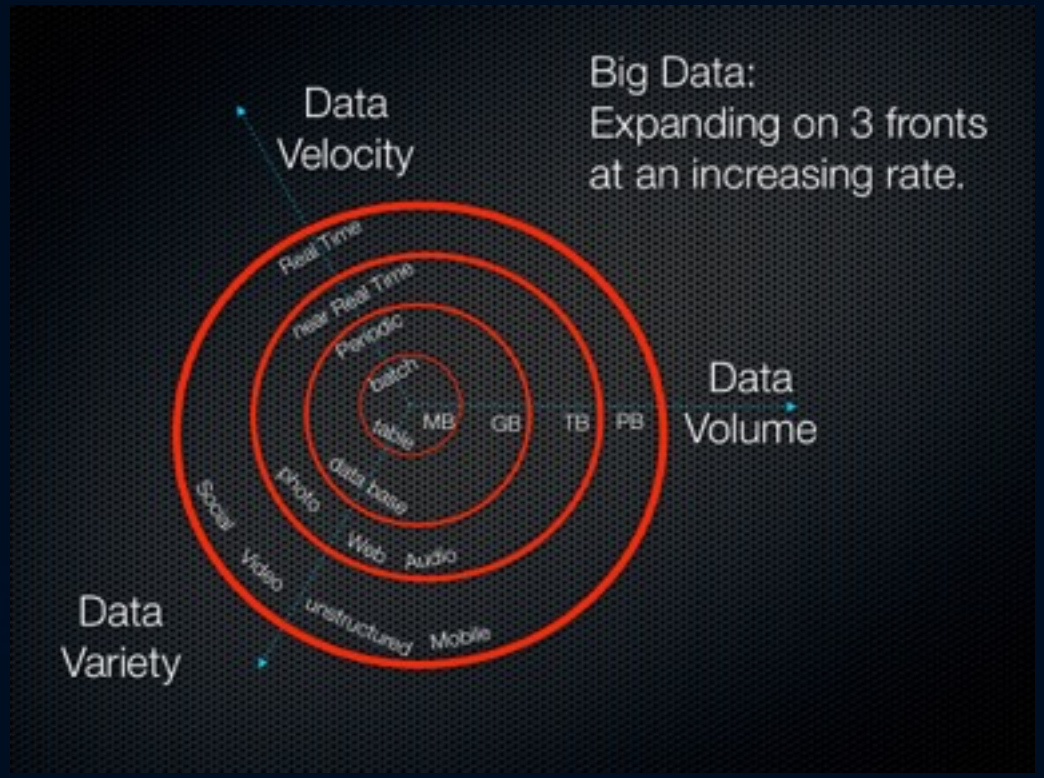
1 YB is equivalent to 1 followed by 24 zeroes.

The problem with data…..

over abundance, fragmentation, erroneous, heterogeneity, duplication, untrustworthy, unstructured.

What makes it Big Data?

- Volume
- Velocity
- Variety
- Veracity
- Validity
- Value
- Volatility



Diya Soubra (2012), "The 3Vs that Define Big Data"

Volume > 2.8 (10^{21}) ZB

- Digital universe will reach 40 zettabytes (ZB) by 2020
- **About 373** times all grains of sand on the beaches (http://www.answers.com/Q/How_many_grains_of_sand_are_there_in_the_world)
- = Estimated weight **of 424** Nimitz aircraft carriers (maybe measured in long tons. Weight is expressed in displacement; could not find actual weight in any search. Classified?)

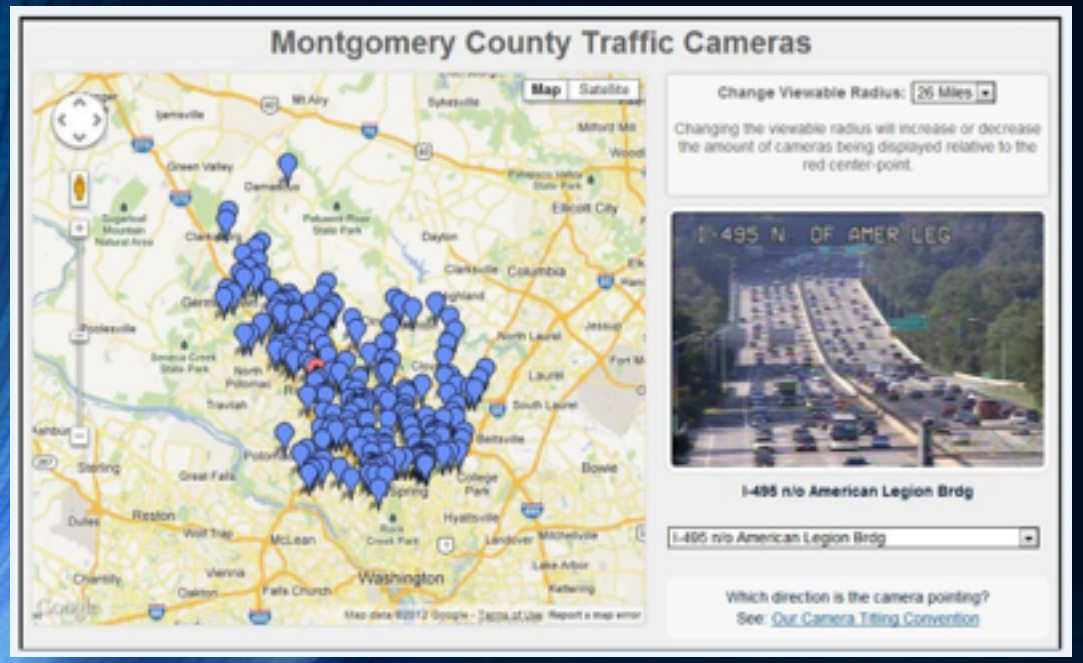
- Facebook (**2012**) processed **2.5** million pieces of content each day (= **500+** terabytes of data daily)

- Amazon sold **958,333** items every hour (on Cyber Monday) in 2016. (If you do the math, that's how about 23 millions items in a day)

Velocity – Data in Motion

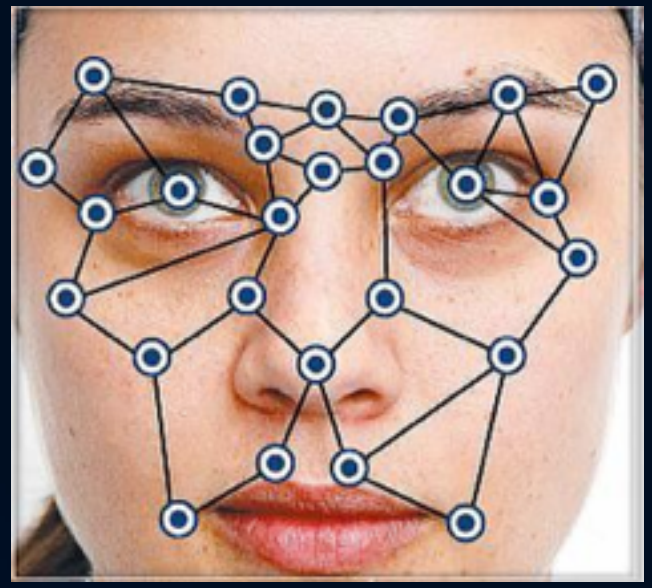
- Data Delivery – Streaming Data
- Results delivery – Real time, Near-real time, Periodic
- Measured in data volume per unit of time (Ex. 20 GB/sec)

Increasing rate of data flow into the organization. It considers incoming and outgoing flow of data. The faster we can capture, analyze, and transform data into useful information the better (tight loop).



Traffic

Disease Outbreak

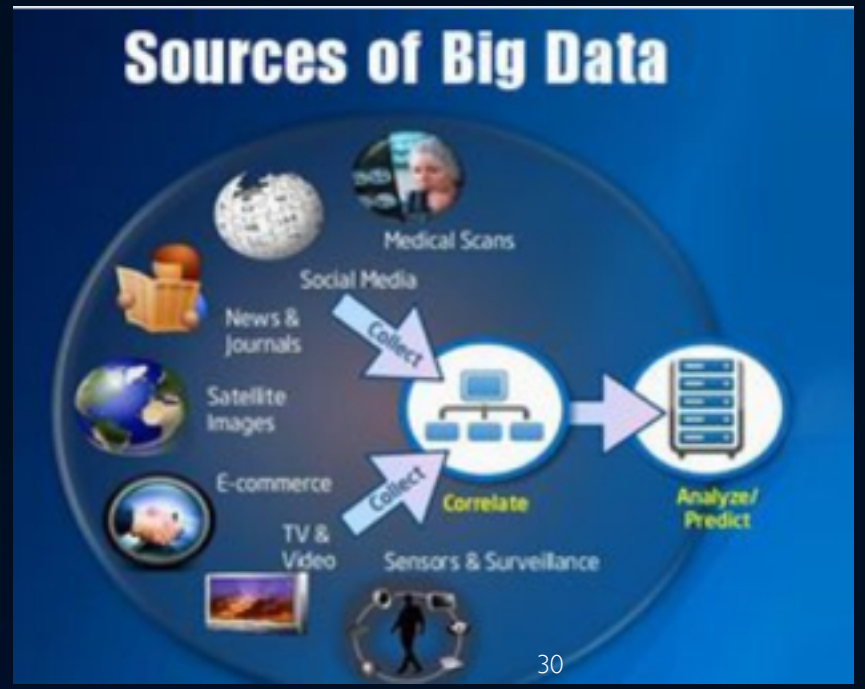


Facial Recognition

Variety – ALL types of data

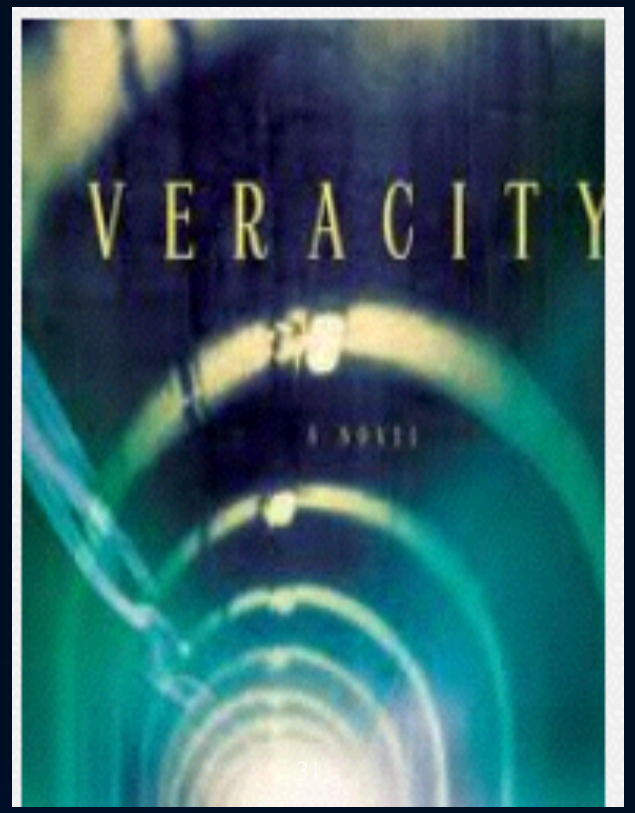
- Big data comes from EVERYWHERE

It is not a relational database with billions of structured rows – it's a mix of structured and multi-structured data.



While we are talking about Data…

- Veracity – the 4th V of Big Data
- Dirty data is a tough issue
- Uncertain data can't be cleaned
- Errors can snowball fast with big data
- Examples:
  - GPS signals bouncing around buildings in NYC
  - Weather conditions – sun, clouds, rain icon



Validity

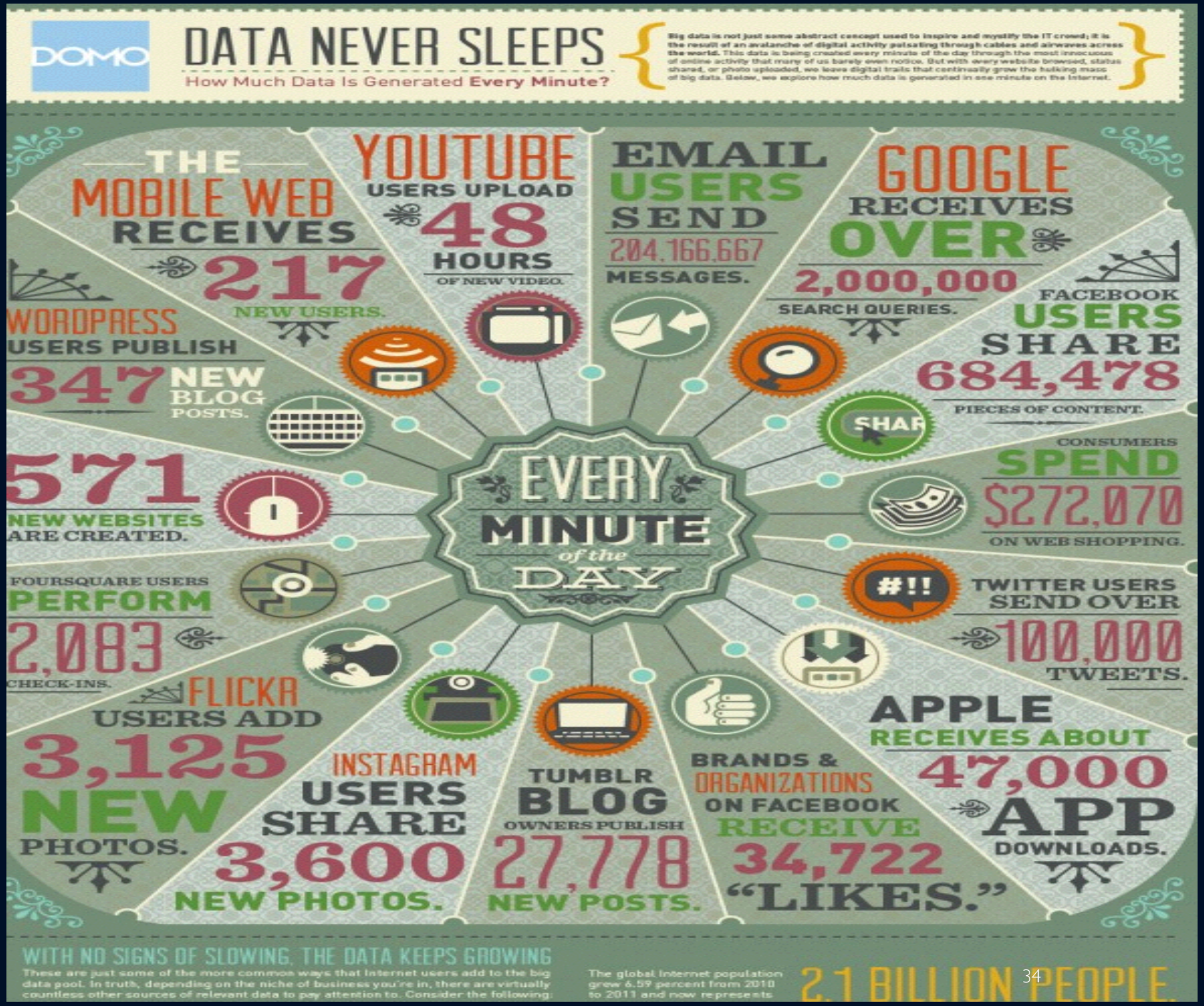
- Validity refers to the issue that the data being stored and mined is clean and meaningful to the problems or decisions that need to be made.

Value and Volatility

- Value refers usefulness and relevancy of the information extracted from the data as opposed to the practice of collecting data for archival or regulatory purposes.

- Volatility addresses the issue of for how long is the data valid and for how long should it be stored. Some analysis is required to determine when, in the future, the data is no longer relevant.

In Summary

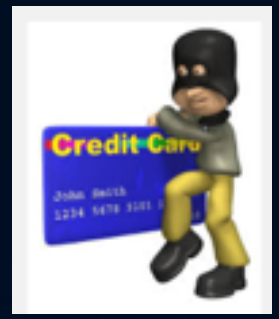

DOMO
DATA NEVER SLEEPS
How Much Data Is Generated Every Minute?
Big data is not just some abstract concept used to inspire and mystify the IT crowd; it is the result of an avalanche of digital activity pulsating through cables and airwaves across the world. This data is being created every minute of the day through the most innocuous of online activity that many of us barely even notice. But with every website browsed, status shared, or photo uploaded, we leave digital trails that continually grow the hulking mass of big data. Below, we explore how much data is generated in one minute on the Internet.
THE MOBILE WEB RECEIVES 217 NEW USERS.
YOUTUBE USERS UPLOAD 48 HOURS OF NEW VIDEO.
EMAIL USERS SEND 204,166,667 MESSAGES.
GOOGLE RECEIVES OVER 2,000,000 SEARCH QUERIES.
FACEBOOK USERS SHARE 684,478 PIECES OF CONTENT.
WORDPRESS USERS PUBLISH 347 NEW BLOG POSTS.
571 NEW WEBSITES ARE CREATED.
EVERY MINUTE of the DAY
CONSUMERS SPEND $272,070 ON WEB SHOPPING.
FOURSQUARE USERS PERFORM 2,083 CHECK-INS.
TWITTER USERS SEND OVER 100,000 TWEETS.
FLICKR USERS ADD 3,125 NEW PHOTOS.
INSTAGRAM USERS SHARE 3,600 NEW PHOTOS.
TUMBLR BLOG OWNERS PUBLISH 27,778 NEW POSTS.
BRANDS & ORGANIZATIONS ON FACEBOOK RECEIVE 34,722 "LIKES."
APPLE RECEIVES ABOUT 47,000 APP DOWNLOADS.
WITH NO SIGNS OF SLOWING, THE DATA KEEPS GROWING
These are just some of the more common ways that Internet users add to the big data pool. In truth, depending on the niche of business you're in, there are virtually countless other sources of relevant data to pay attention to. Consider the following:
The global Internet population grew 6.59 percent from 2010 to 2011 and now represents
2.1 BILLION PEOPLE.

Uses of Big Data - Current

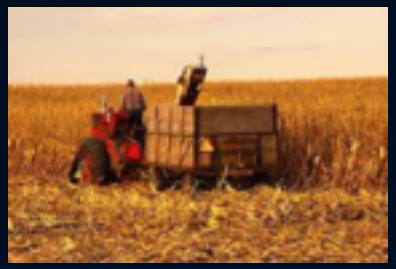
- "Customer-Centric"
  - Personalize the experience
- Predictive Analytics
  - Manage the risk

Customer Recommendations
Streaming Routing

Online ad targeting

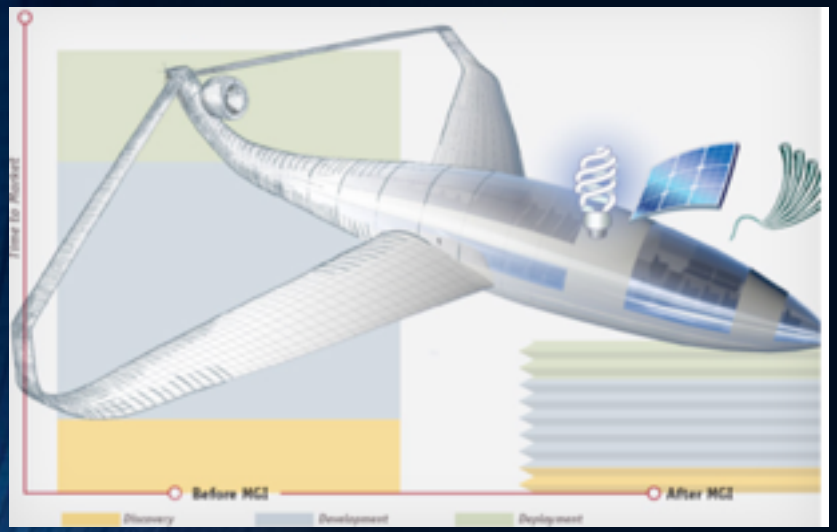


Credit Card Fraud



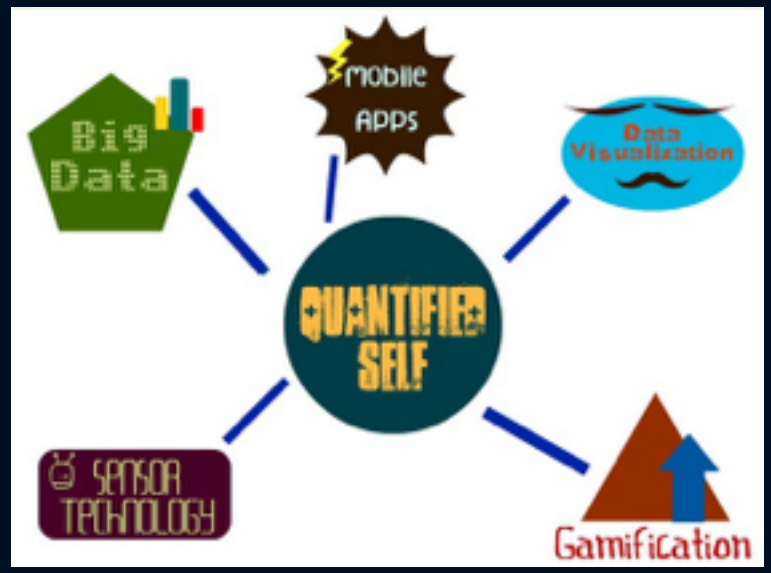
Crop Forecasts

Uses of Big Data - Future

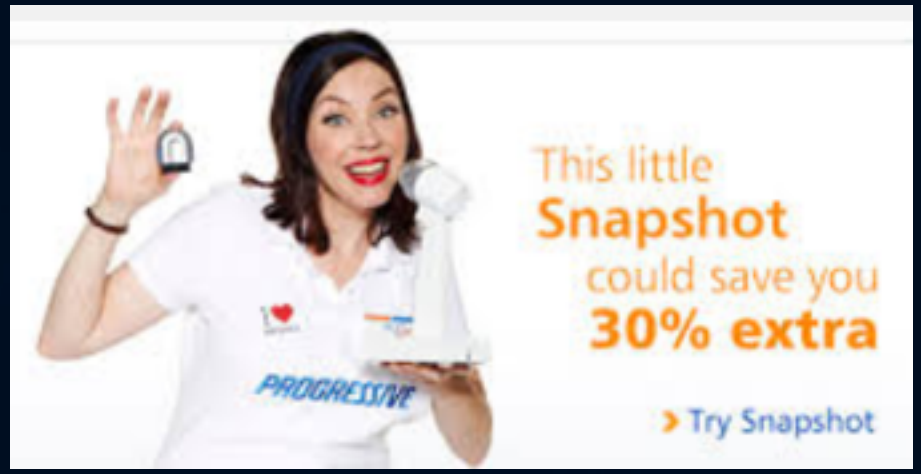


New material development

(Materials Genome Project)

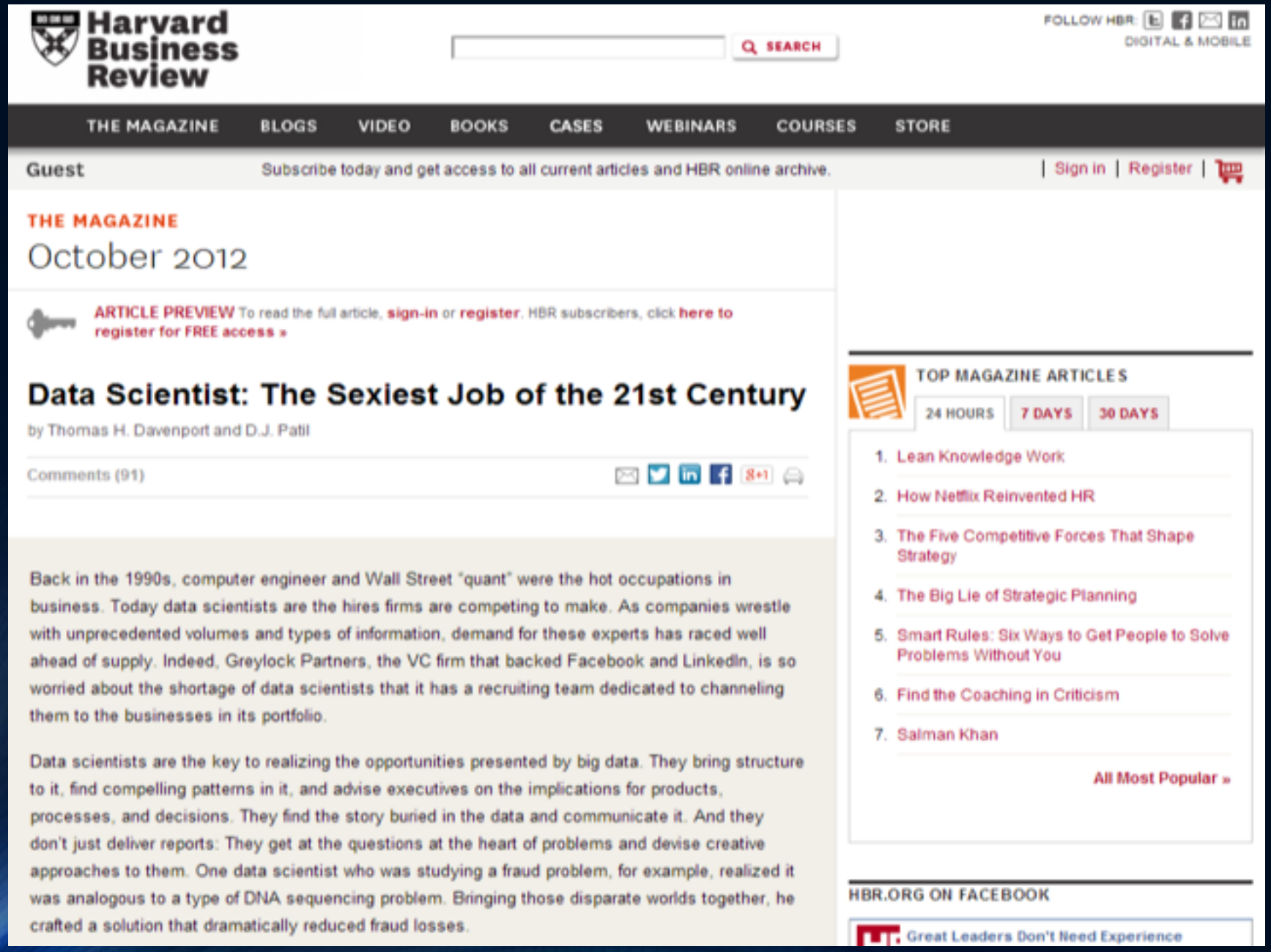


Self Quantification

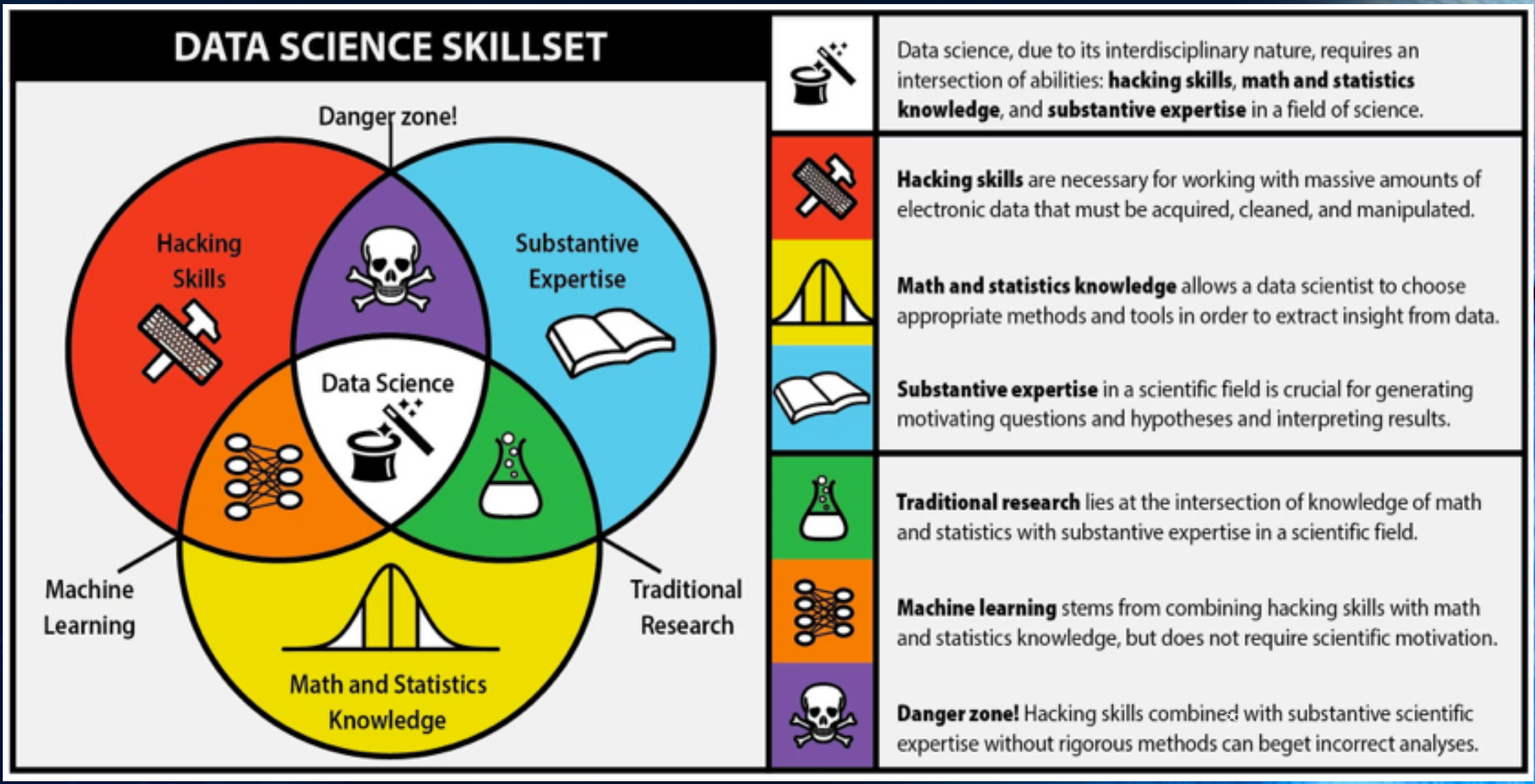


Pay as You Go

What is Data Science?


Harvard Business Review
SEARCH
FOLLOW HBR
DIGITAL & MOBILE
THE MAGAZINE
BLOGS
VIDEO
BOOKS
CASES
WEBINARS
COURSES
STORE
Guest
Subscribe today and get access to all current articles and HBR online archive.
Sign in | Register
THE MAGAZINE
October 2012
ARTICLE PREVIEW To read the full article, sign-in or register. HBR subscribers, click here to register for FREE access »
Data Scientist: The Sexiest Job of the 21st Century
by Thomas H. Davenport and D.J. Patil
Comments (91)
Back in the 1990s, computer engineer and Wall Street "quant" were the hot occupations in business. Today data scientists are the hires firms are competing to make. As companies wrestle with unprecedented volumes and types of information, demand for these experts has raced well ahead of supply. Indeed, Greylock Partners, the VC firm that backed Facebook and LinkedIn, is so worried about the shortage of data scientists that it has a recruiting team dedicated to channeling them to the businesses in its portfolio.
Data scientists are the key to realizing the opportunities presented by big data. They bring structure to it, find compelling patterns in it, and advise executives on the implications for products, processes, and decisions. They find the story buried in the data and communicate it. And they don't just deliver reports: They get at the questions at the heart of problems and devise creative approaches to them. One data scientist who was studying a fraud problem, for example, realized it was analogous to a type of DNA sequencing problem. Bringing those disparate worlds together, he crafted a solution that dramatically reduced fraud losses.
TOP MAGAZINE ARTICLES
24 HOURS
7 DAYS
30 DAYS
1. Lean Knowledge Work
2. How Netflix Reinvented HR
3. The Five Competitive Forces That Shape Strategy
4. The Big Lie of Strategic Planning
5. Smart Rules: Six Ways to Get People to Solve Problems Without You
6. Find the Coaching in Criticism
7. Salman Khan
All Most Popular »
HBR.ORG ON FACEBOOK
Great Leaders Don't Need Experience

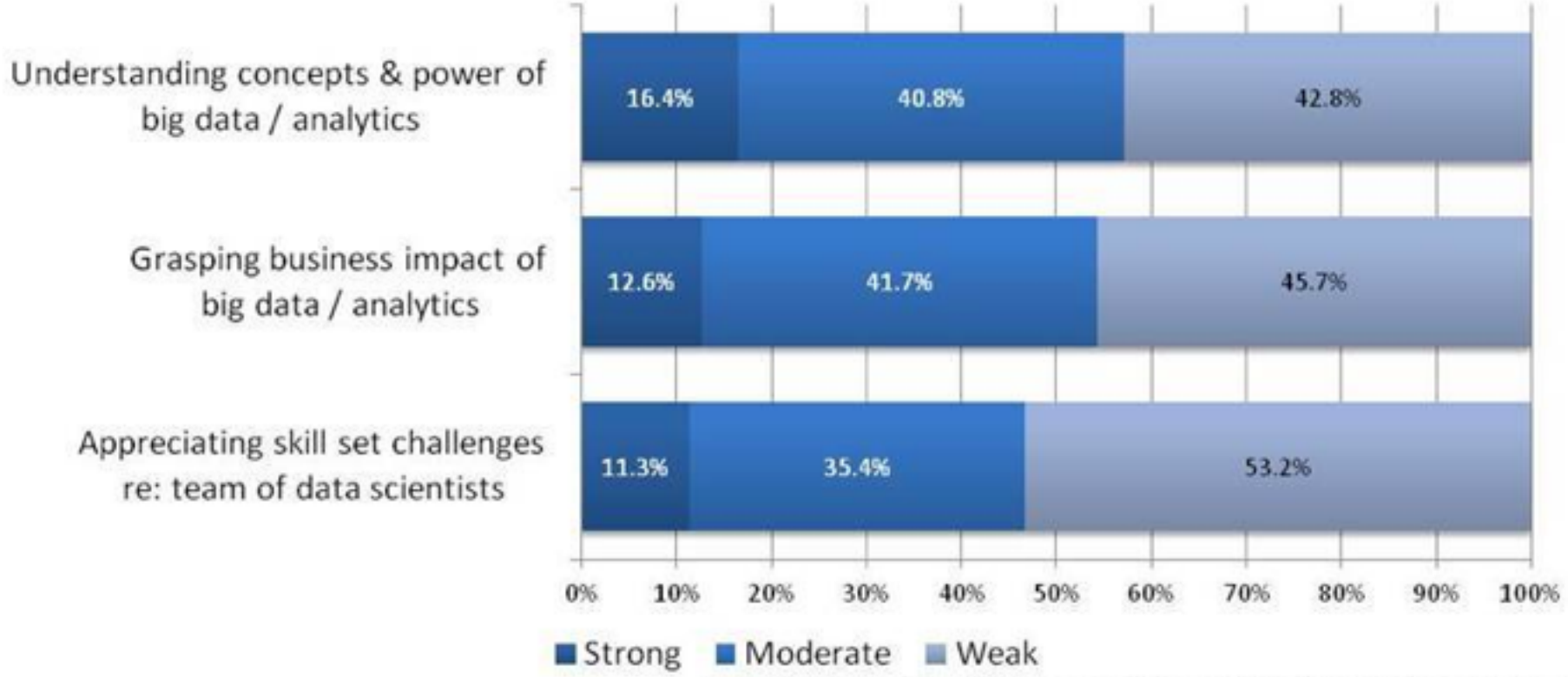
What does it take?



Is there anything else?

Data Science and Big Data Analytics

Management Skills Gap – Input From Over 1,000 Professionals



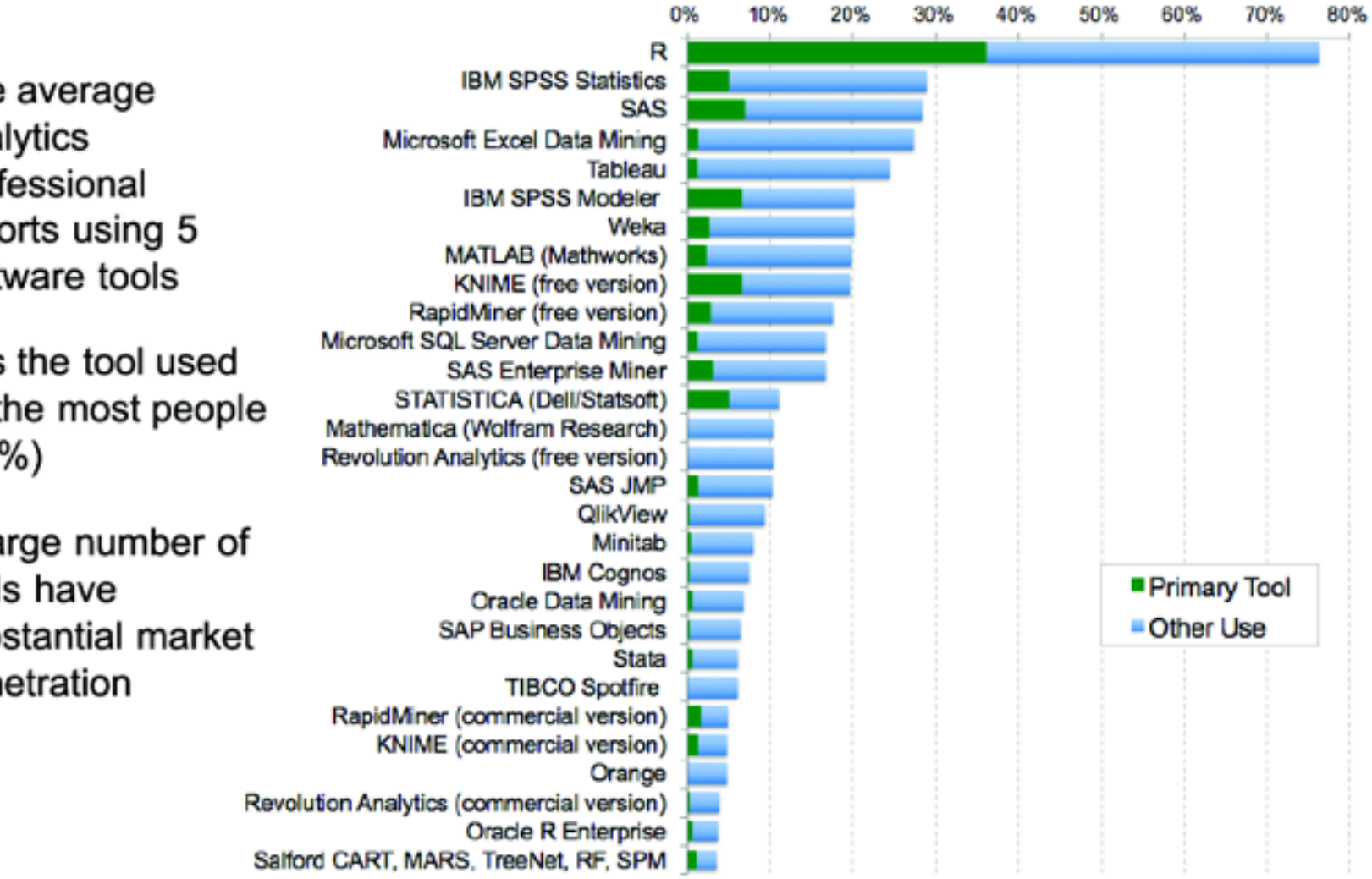
Source: Managing Information Storage: Trends, Challenges and Options – 2013-14 Study – To be Published May 7,2013.

Tools in Data Science

The Tools We're Using

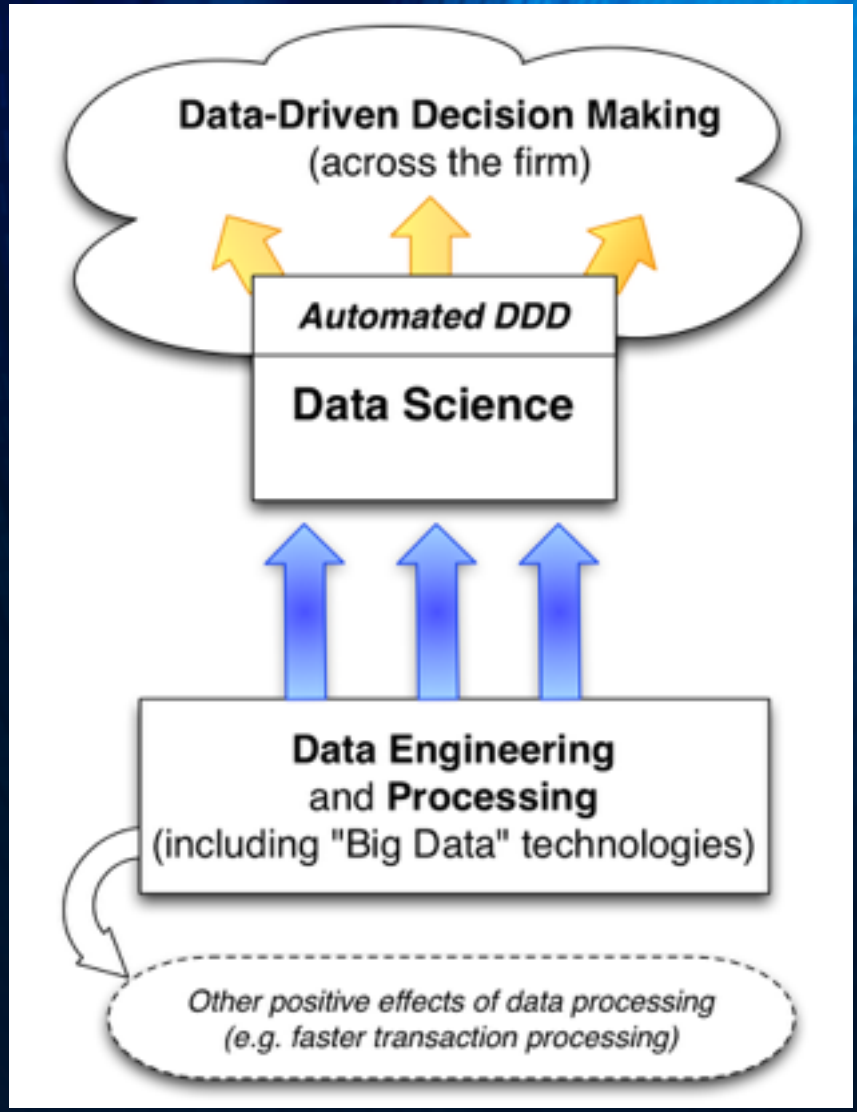
Vendors are excluded from tool-use analyses

- The average analytics professional reports using 5 software tools
- R is the tool used by the most people (76%)
- A large number of tools have substantial market penetration



Question: What Data mining / analytic tools did you use in the past year?
Question: What one data mining / analytic software package do you use most frequently in the past year?

Some Order in the Buzz…



Data in its raw form has low value. The objective of data engineering and data science is to increase the value of data for the entire organization.

Turning raw data into "actionable insight" is the objective of the data scientist in a data-driven decision making process. To do this data scientist use *DATA ANALYTICS*.

There are four types of Data Analytics:

Descriptive: based on historical data, descriptive analytics tries to answer questions such as "what happened?"

Diagnostic: based on historical data tries to answer questions such as "why did this happen?"

Predictive: based on historical and current data and through the use of model tries to discover trends or future events.

Prescriptive: based on the predictive results tries to optimize the process or structures of the organization.

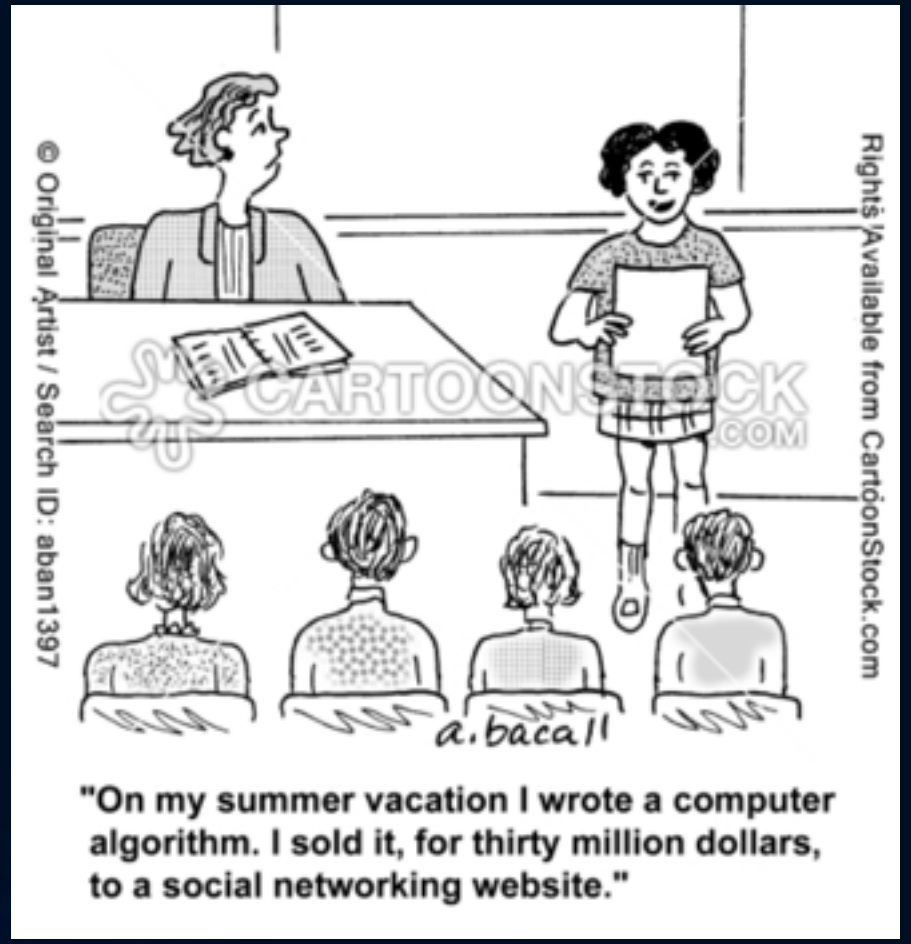
I/O The Real Issue

Processors aren't the problem, it's getting data off the hard disk

… and for those hopeful that Solid State drives would be a solution – they won't be replacing hard drives anytime soon

Algorithms

- What it really is all about
  - In Big Data, common association is with Predictive Algorithms
  - Room for new ideas but some believe we have enough of these for now…
  - Others say what you need is more data, simpler algorithms



What types of algorithms?

- Sorting
- Searching
- Streaming
- Filtering

- Classifying
- Predictive Learning via Rule Fit Ensembles
- Deterministic behavior algorithms

Summary

- From *scientific* discovery to business intelligence, "*Big Data*" is changing our world
- Big Data permeates most (all?) areas of computer science
- Opening the doors to lots of opportunities in the computing sciences

It's not just for Data Scientists….

Links to more info about material presented in this lecture:

IDC (2012) Digital Universe in 2020 - http://www.emc.com/collateral/analyst-reports/idc-the-digital-universe-in-2020.pdf
Diya Soubra (2012) "The 3Vs that define Big Data" - http://www.datasciencecentral.com/forum/topics/the-3vs-that-define-big-data
Self Quantification - http://www.topcoder.com/blog/big-data-mobile-sensors-visualization-gamification-quantified-self/
Materials Genome Project - http://www.theverge.com/2013/9/26/4766486/materials-genome-initiative-mit-and-harvard
Pay as You Go - http://www.businessweek.com/articles/2012-10-15/pay-as-you-drive-insurance-big-brother-needs-a-makeover
Montgomery County Traffic Cameras - http://www6.montgomerycountymd.gov/tmctmpl.asp
Trending Now: Using Social Media to Predict and Track Disease - Outbreaks http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3261963/
Health Map - http://healthmap.org/en/
Manhunt - Boston Bombers. Aired May 29, 2013 on PBS - Nova Program on Facial Recognition - http://www.pbs.org/wgbh/nova/tech/manhunt-boston-bombers.html
Netflix and Big Data - http://technologyadvice.com/wp-content/themes/newta/how-netflix-is-using-big-data-to-get-people-hooked-on-its-original-programming.html#.Ul1WlBAlgnZ
Data Never Sleeps - http://www.domo.com/learn/infographic-data-never-sleeps
Google Research Publication: The Google File System - http://research.google.com/archive/gfs.html
Google Research Publication: MapReduce - http://research.google.com/archive/mapreduce.html

Links to more info about material presented in this talk:

Sorting the World: Google Invents New Way to Manage Data - http://www.wired.com/science/discoveries/magazine/16-07/pb_sorting
Hadoop - http://hadoop.apache.org/
What lies at the core of Hadoop? http://blog.enablecloud.com/2012/06/what-lies-at-core-of-hadoop.html
Distributed Average Consensus with Least-Mean-Square Deviation - http://www.stanford.edu/~boyd/papers/pdf/lmsc_mtns06.pdf
HDFS in Cartoon -
https://docs.google.com/file/d/0B-zw6KHOtbT4MmRkZWJjYzEtYjI3Ni00NTFjLWE0OGItYTU5OGMxYjc0N2M1/edit?usp=drive_web&pli=1
Big-Data Computing: Creating revolutionary breakthroughs in commerce, science, and society - Computer Research Association CCC led paper - http://www.cra.org/ccc/resources/ccc-led-white-papers
How Quantum Computers and Machine Learning Will Revolutionize Big Data - Quanta Magazine - http://www.wired.com/wiredscience/2013/10/computers-big-data/
Predictive Apps http://www.information-management.com/blogs/application-developers-ignore-big-data-at-your-own-peril-10024904-1.html
Real World Use of Big Data – Ford Focus - http://www.stthomas.edu/gradsoftware/files/BigData_RealWorldUse.pdf
Eatery Massive Health Experiment - https://eatery.massivehealth.com/
Waze – www.waze.com
CS2013 Ironman v1.0 draft - http://ai.stanford.edu/users/sahami/CS2013/

Questions?