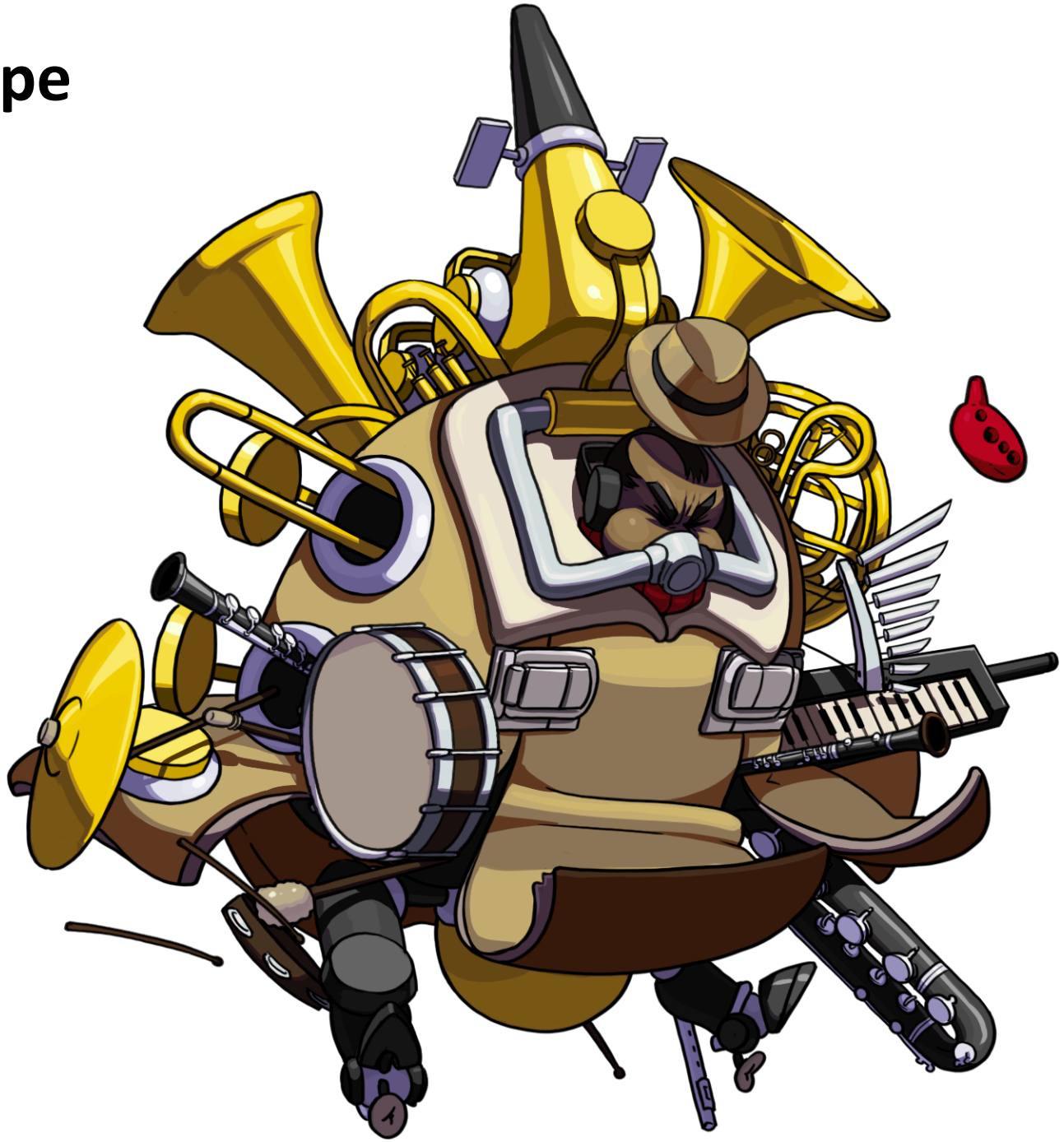
INTRODUCTION to BIG ~~BAND~~ ~~BANG~~ DATA

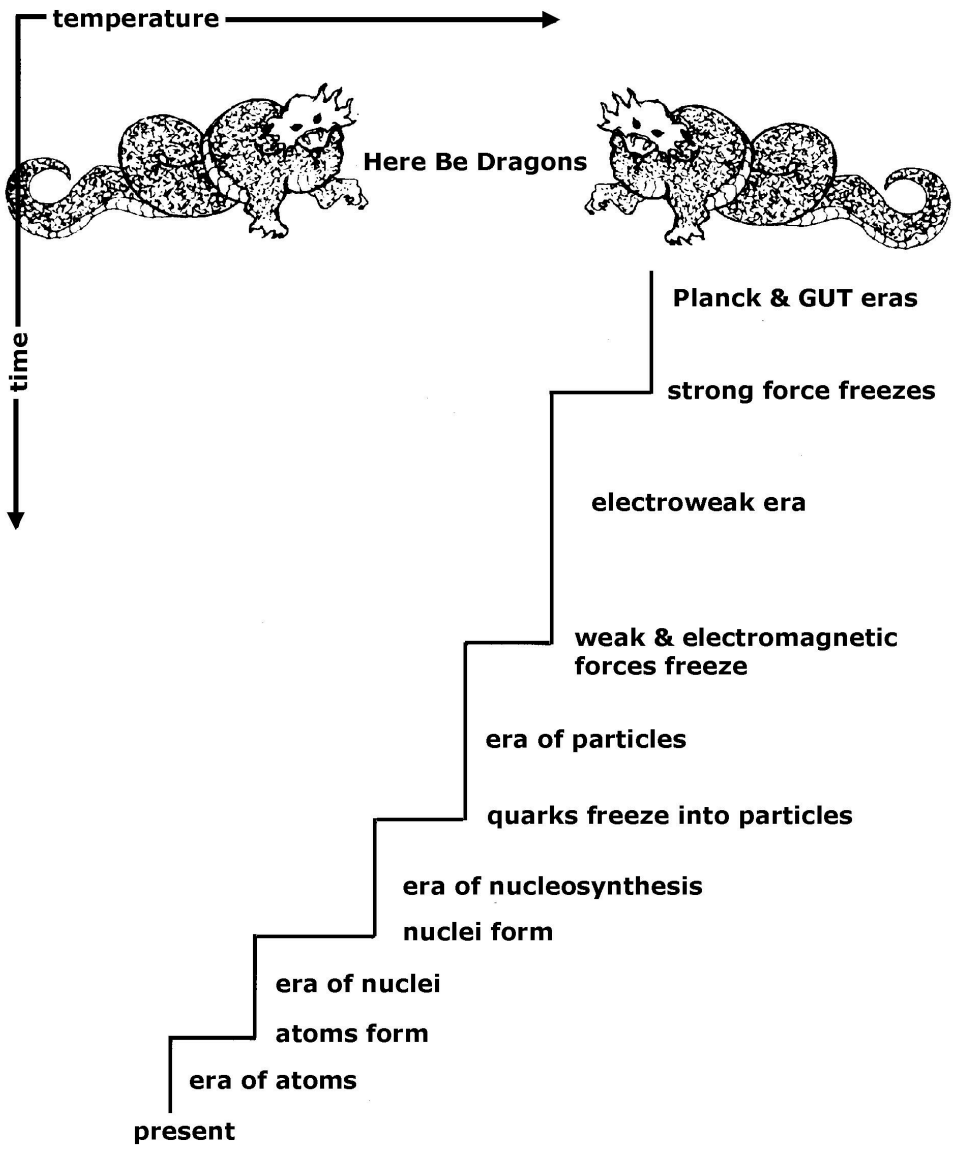
Yabebal Fantaye
Bruce Bassett
Pierre-Yves Lablanche

BIG BAND: type



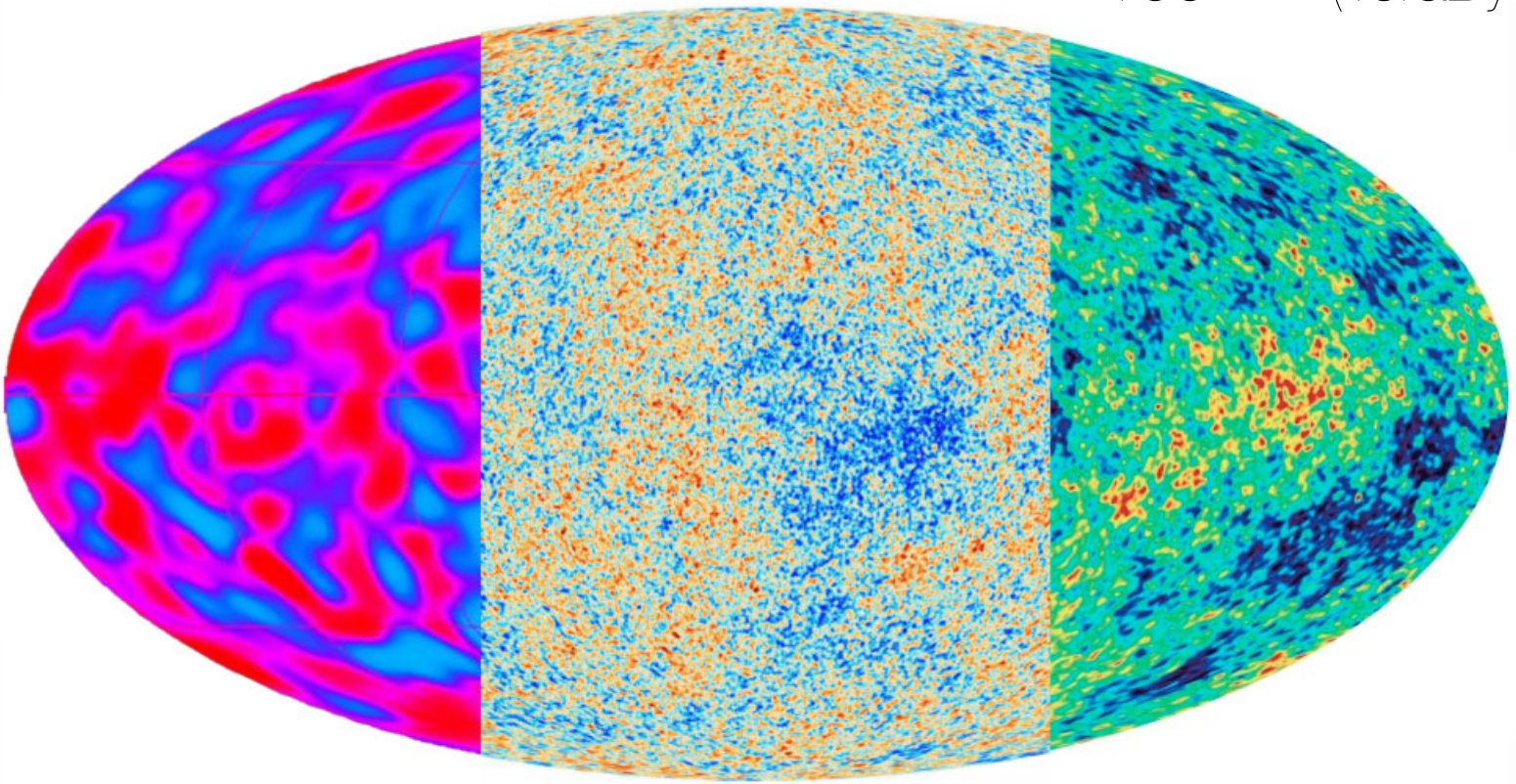
associated with jazz and the Swing Era typically consisting of percussion, brass, and woodwind instruments totalling **approximately 12 to 25 musicians**

BIG BANG: model



http://ircamera.as.arizona.edu/NatSci102/NatSci102/text/whybigbang.htm

1 1000 100 (TeraBytes)



COBE
1993

PLANCK
2015

WMAP
2004

ESA & NASA

COSMOLOGY | THE EVOLUTION

Trying to describe the size of the Big Bang

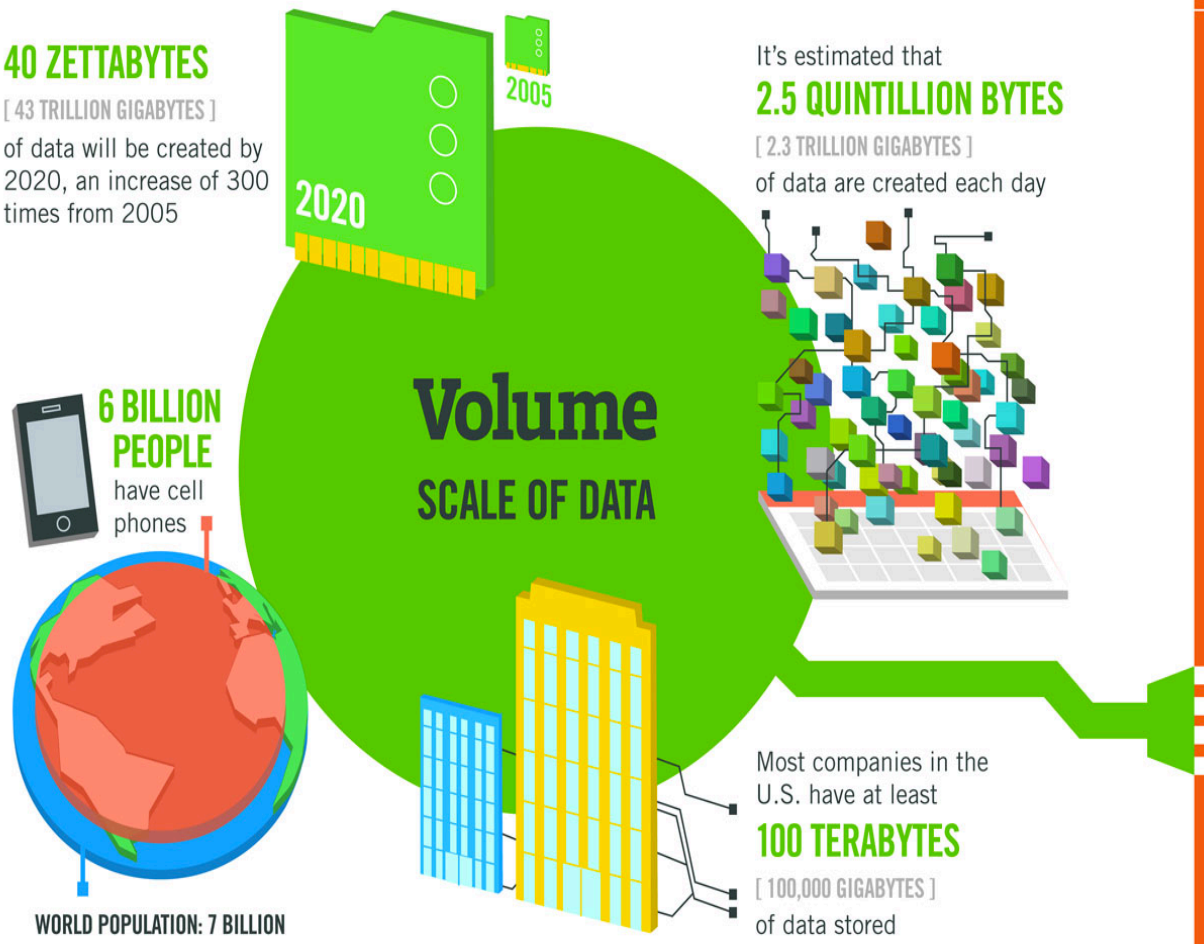
BIG DATA?


Veracity
Capture
Curation
Storage
Search
Velocity
BIG DATA
Sharing
Variety
Analysis
Volume
Visualization
Transfer

Outline

- Characteristics
- Use cases
- The secrete of Big Data
- Requirements

BIG DATA CHARACTERISTICS



The FOUR V's of Big Data

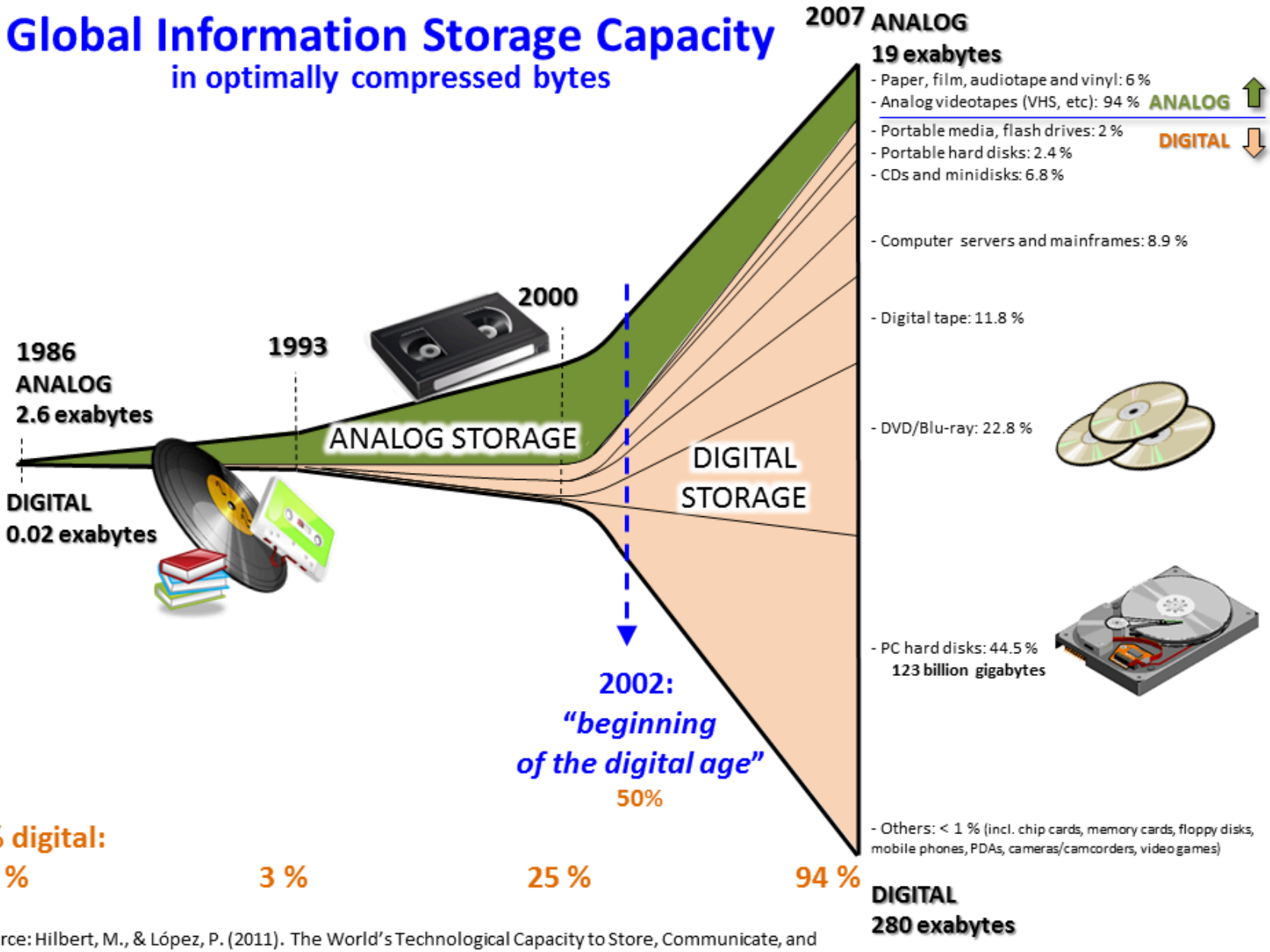
From traffic patterns and music downloads to web history and medical records, data is recorded, stored, and analyzed to enable the technology and services that the world relies on every day. But what exactly is big data, and how can these massive amounts of data be used?

As a leader in the sector, IBM data scientists break big data into four dimensions: **Volume, Velocity, Variety and Veracity**

IBM

Global Information Storage Capacity

in optimally compressed bytes



Source: Hilbert, M., & López, P. (2011). The World's Technological Capacity to Store, Communicate, and Compute Information. *Science*, 332(6025), 60 –65. http://www.martinhilbert.net/WorldInfoCapacity.html

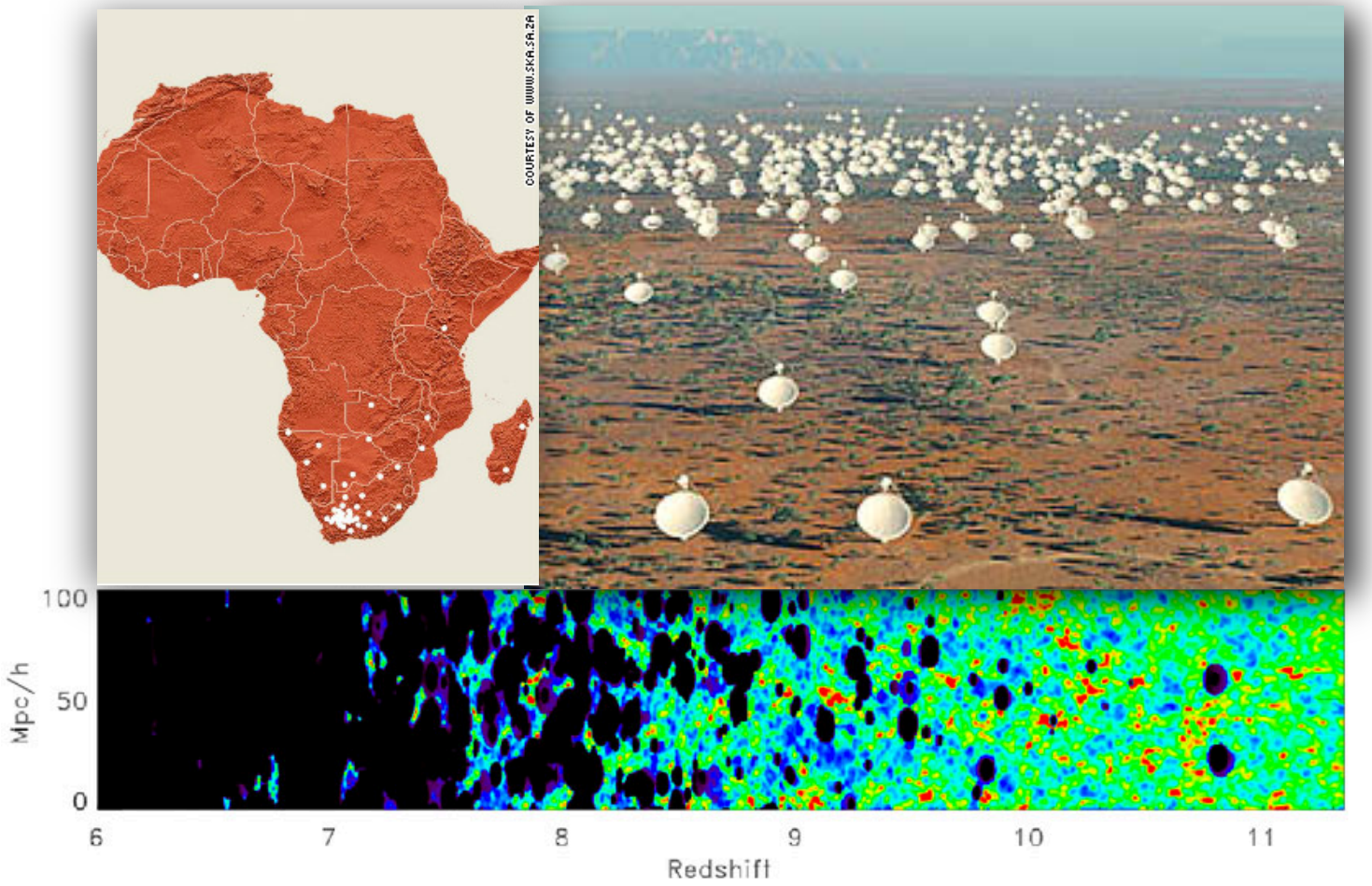
Information from the Internet of Things:

We have gone beyond the decimal system

Today data scientist uses Yottabytes to describe how much government data the NSA or FBI have on people altogether.

In the near future, **Brontobyte** will be the measurement to describe the type of sensor data that will be generated from the IoT (Internet of Things)

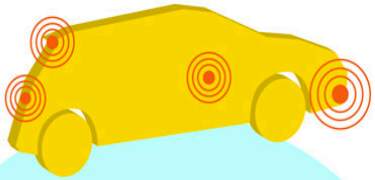
1 EXABYTE PER DAY
1,000,000,000,000,000,000



The New York Stock Exchange captures

1 TB OF TRADE INFORMATION

during each trading session



Modern cars have close to

100 SENSORS

that monitor items such as fuel level and tire pressure

Velocity

ANALYSIS OF STREAMING DATA

By 2016, it is projected there will be

18.9 BILLION NETWORK CONNECTIONS

– almost 2.5 connections per person on earth

Depending on the industry and organization, big data encompasses information from multiple internal and external sources such as transactions, social media, enterprise content, sensors and mobile devices. Companies can leverage data to adapt their products and services to better meet customer needs, optimize operations and infrastructure, and find new sources of revenue.

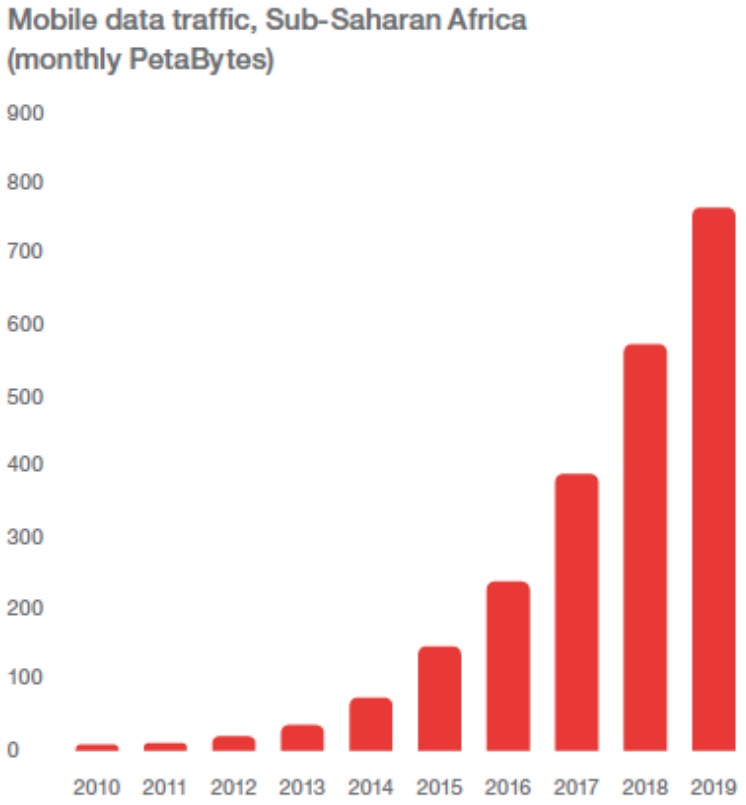
By 2015

4.4 MILLION IT JOBS

will be created globally to support big data, with 1.9 million in the United States

Sources: McKinsey Global Institute, Twitter, Cisco, Gartner, EMC, SAS, IBM, MEPTEC, QAS

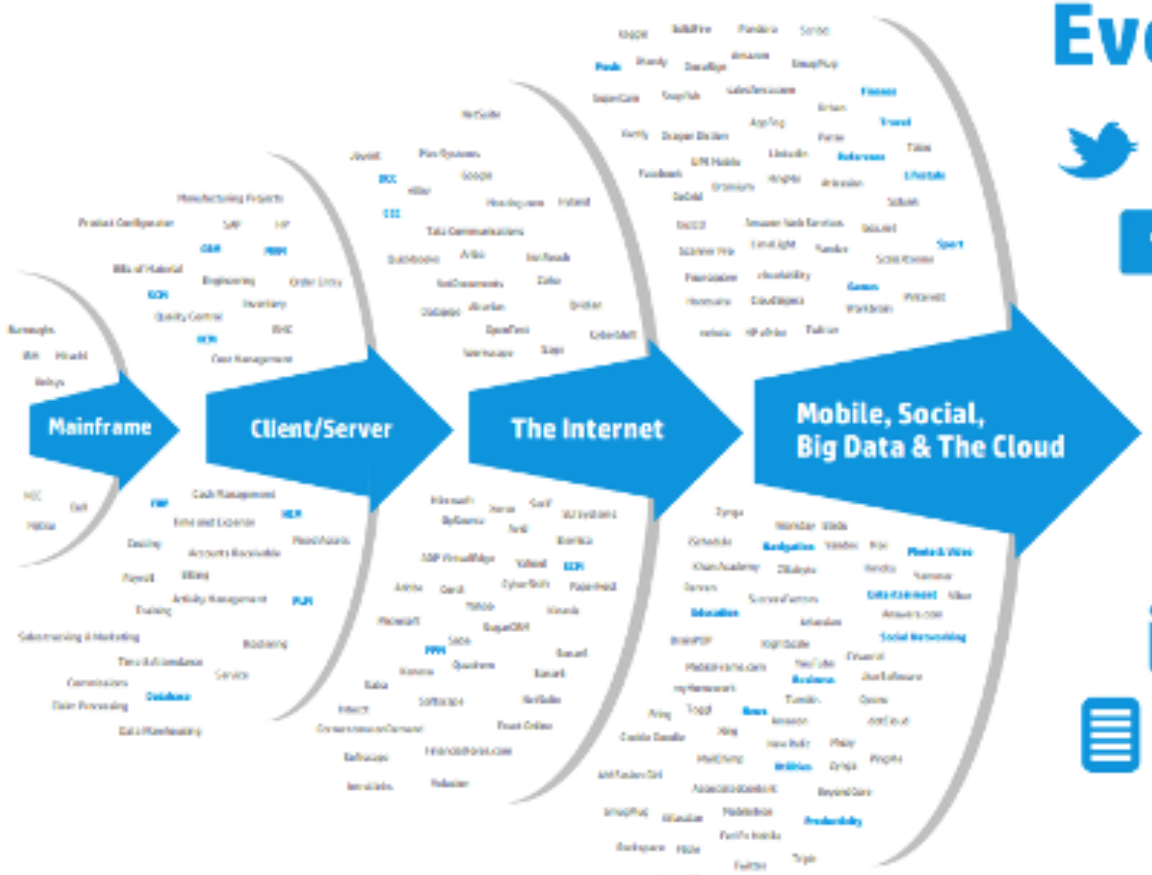
IBM



20X

growth in **mobile data traffic**
between 2013 and 2019

Ericsson ConsumerLab (2013)



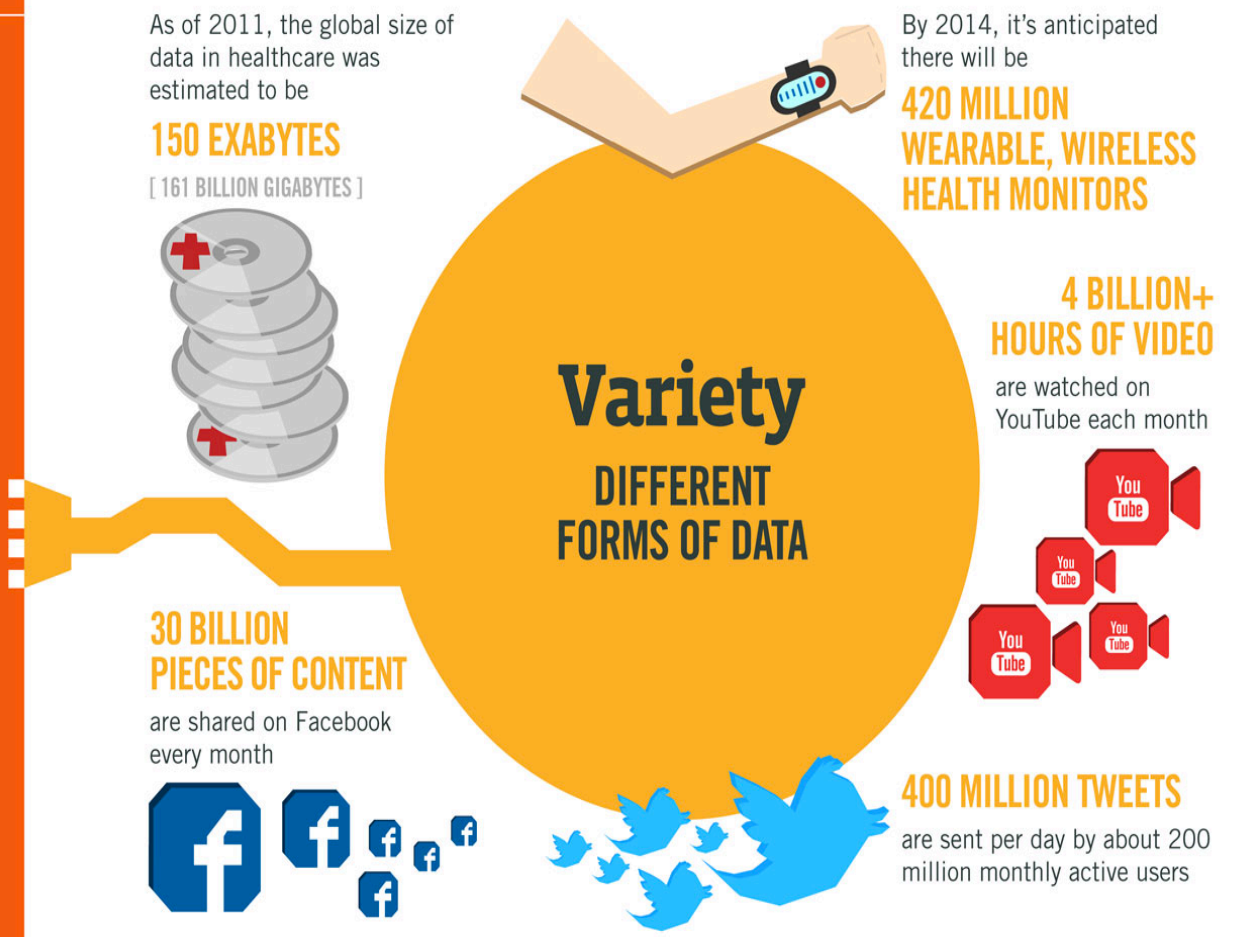
Every 60 seconds

- **98,000+** tweets
- **695,000** status updates
- **11million** instant messages
- **698,445** Google searches
- **168 million+** emails sent
- **1,820TB** of data created
- **217** new mobile web users

The FOUR V's of Big Data

From traffic patterns and music downloads to web history and medical records, data is recorded, stored, and analyzed to enable the technology and services that the world relies on every day. But what exactly is big data, and how can these massive amounts of data be used?

As a leader in the sector, IBM data scientists break big data into four dimensions: **Volume, Velocity, Variety and Veracity**



IBM

Veracity

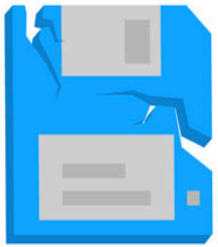
UNCERTAINTY OF DATA

1 IN 3 BUSINESS LEADERS

don't trust the information they use to make decisions

Poor data quality costs the US economy around

$3.1 TRILLION A YEAR



in one survey were unsure of how much of their data was inaccurate

IBM

7Bn people using technology means Big Data

VOLUME

In 2005, humankind created 150 exabytes of information.
In 2011, **1,200 exabytes were created.**

The Economist

VELOCITY

Worldwide digital content will **double in 18 months, and every 18 months thereafter.**

IDC

VARIETY

80% of enterprise data will be unstructured, spanning traditional and non traditional sources.

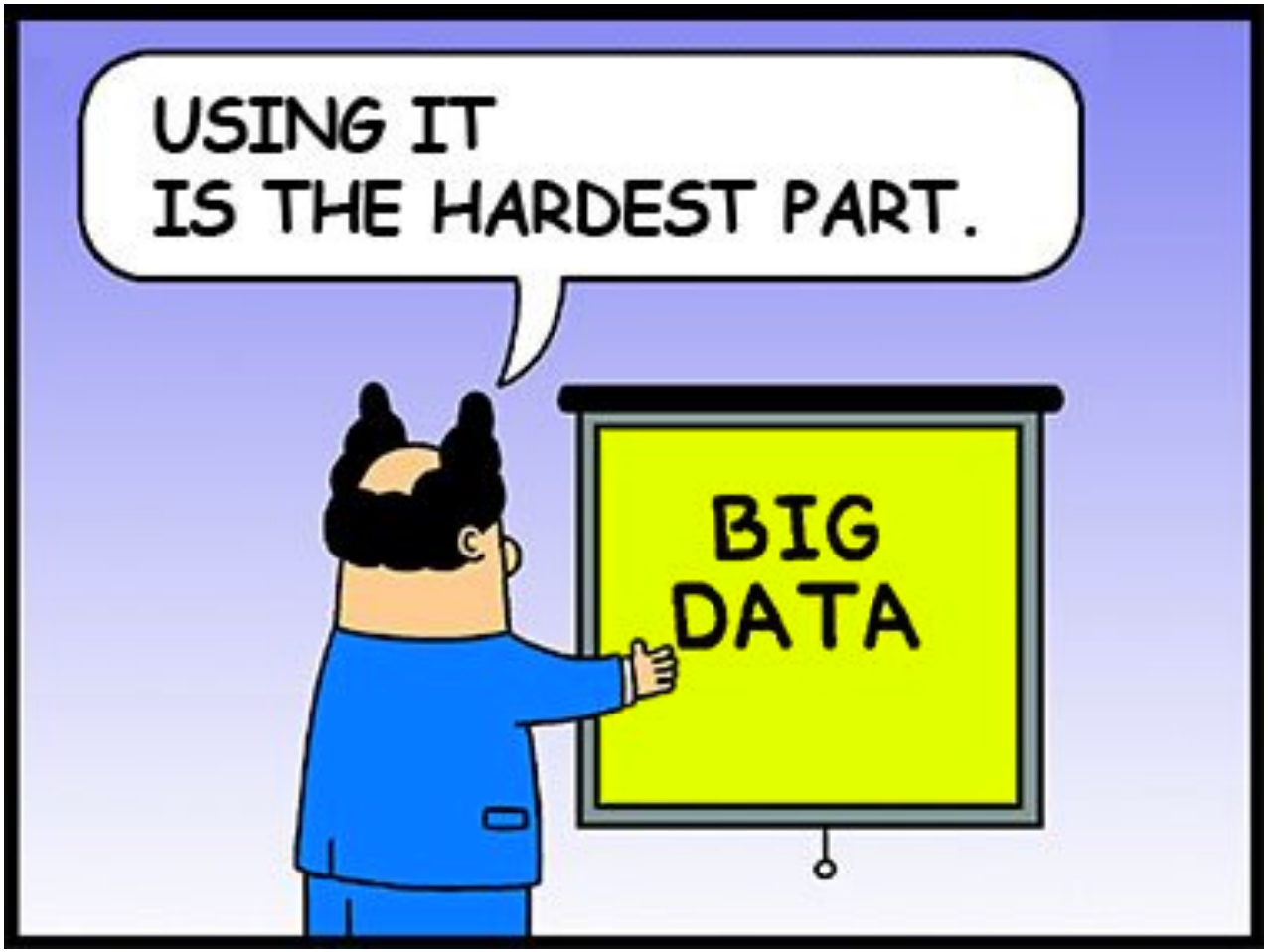
Gartner

TATA CONSULTANCY SERVICES
Experience certainty.

Discussion

What are the sources of Big Data in Africa?

BIG DATA
USE CASES


USING IT
IS THE HARDEST PART.
BIG
DATA

Big Data use cases

Big Data Exploration

Find, visualize, understand all big data to improve decision making. Big data exploration addresses the challenge that every large organization faces: information is stored in many different systems and silos and people need access to that data to do their day-to-day work and make important decisions.

Enhanced 360° View of the Customer

Extend existing customer views by incorporating additional internal and external information sources. Gain a full understanding of customers—what makes them tick, why they buy, how they prefer to shop, why they switch, what they'll buy next, and what factors lead them to recommend a company to others.

Security Intelligence Extension

Lower risk, detect fraud and monitor cyber security in real time. Augment and enhance cyber security and intelligence analysis platforms with big data technologies to process and analyze new types (e.g. social media, emails, sensors, Telco) and sources of under-leveraged data to significantly improve intelligence, security and law enforcement insight.

Operations Analysis

Analyze a variety of machine and operational data for improved business results. The abundance and growth of machine data, which can include anything from IT machines to sensors and meters and GPS devices requires complex analysis and correlation across different types of data sets. By using big data for operations analysis, organizations can gain real-time visibility into operations, customer experience, transactions and behavior.

Data Warehouse Modernization

Integrate big data and data warehouse capabilities to increase operational efficiency. Optimize your data warehouse to enable new types of analysis. Use big data technologies to set up a staging area or landing zone for your new data before determining what data should be moved to the data warehouse. Offload infrequently accessed or aged data from warehouse and application databases using information integration software and tools.

IBM

Big Data Use Cases II

Telecommunications Use Cases

Revenue assurance and price optimization
Customer churn prevention
Campaign management and customer loyalty
Call Detail Record (CDR) analysis
Network performance and optimization
Mobile User Location analysis

Government Use Cases

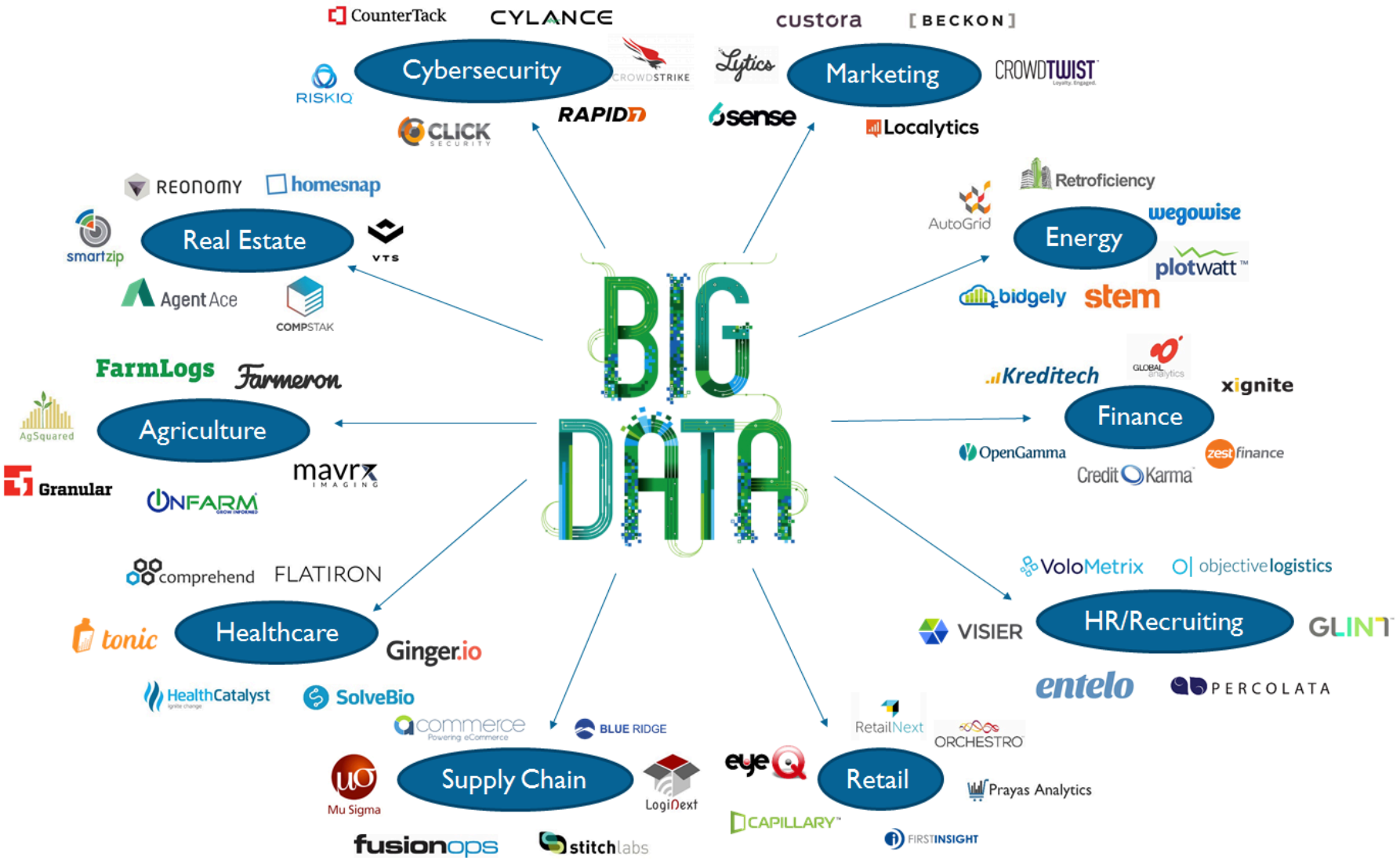
Fraud detection
Threat detection
Cybersecurity
Compliance and regulatory analysis

E-Commerce Use-Cases

Cross-Channel Analytics
Event Analytics
Recommendation Engines using Predictive Analytics
Right Offer at the Right time
Next Best Offer or Next Best Action (Ebay, Netflix, Amazon and Others)

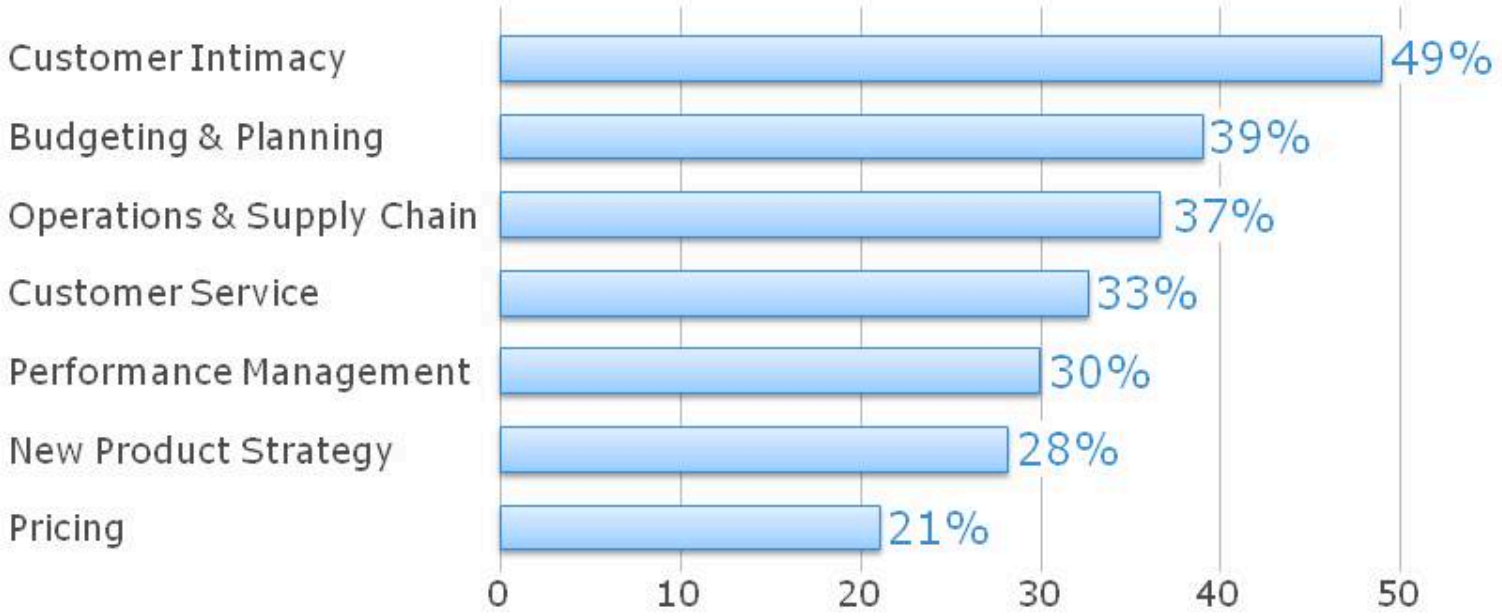
https://pramanicks.wordpress.com/tag/use-cases/

Startups Using Big Data



How Companies Are Using Big Data

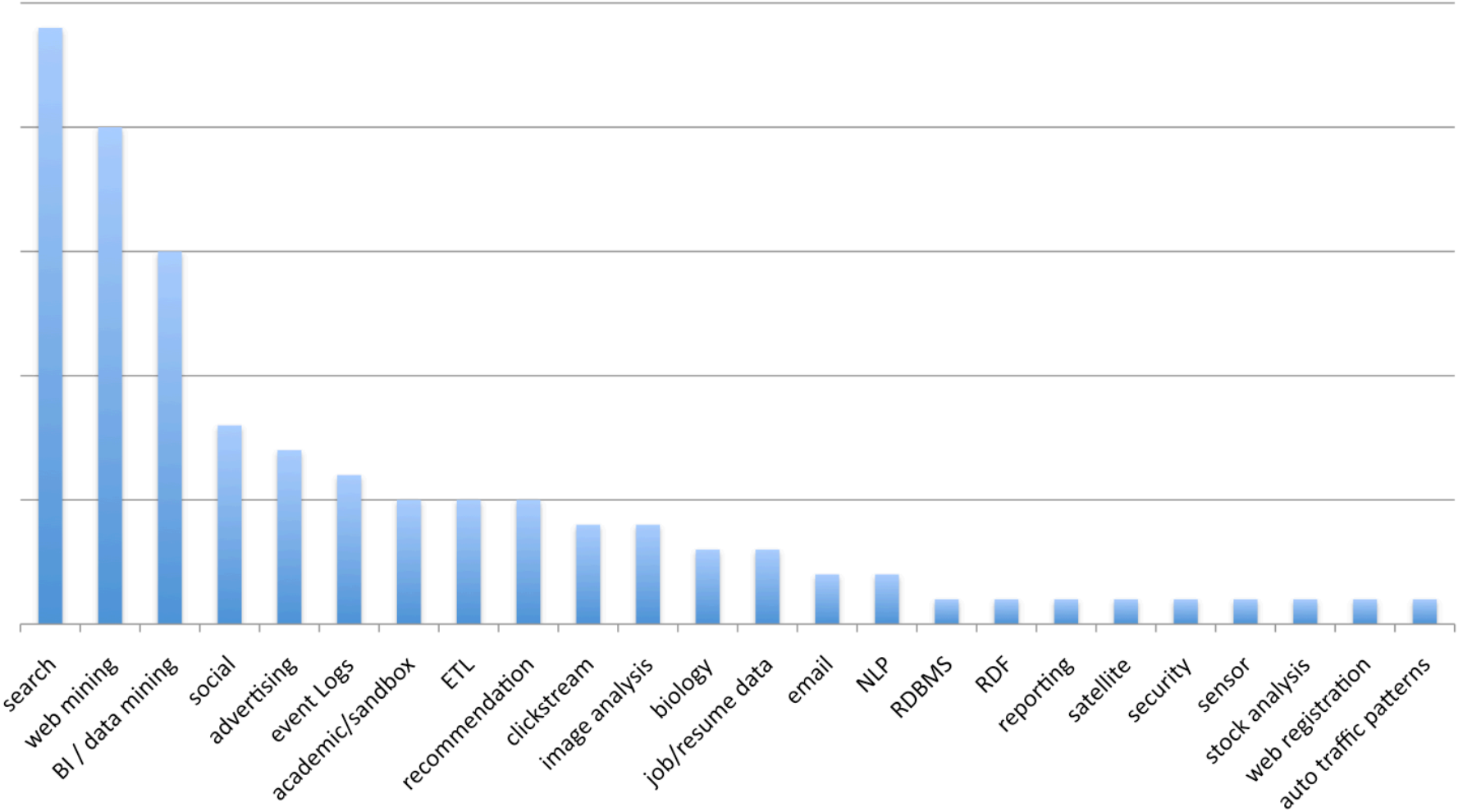
Functional Areas Where Companies Are Using Big Data



McKinsey Global Survey of 1,469 C-level executive respondents at a range of industries and company sizes, "Minding Your Digital Business," 2012.

EMC²

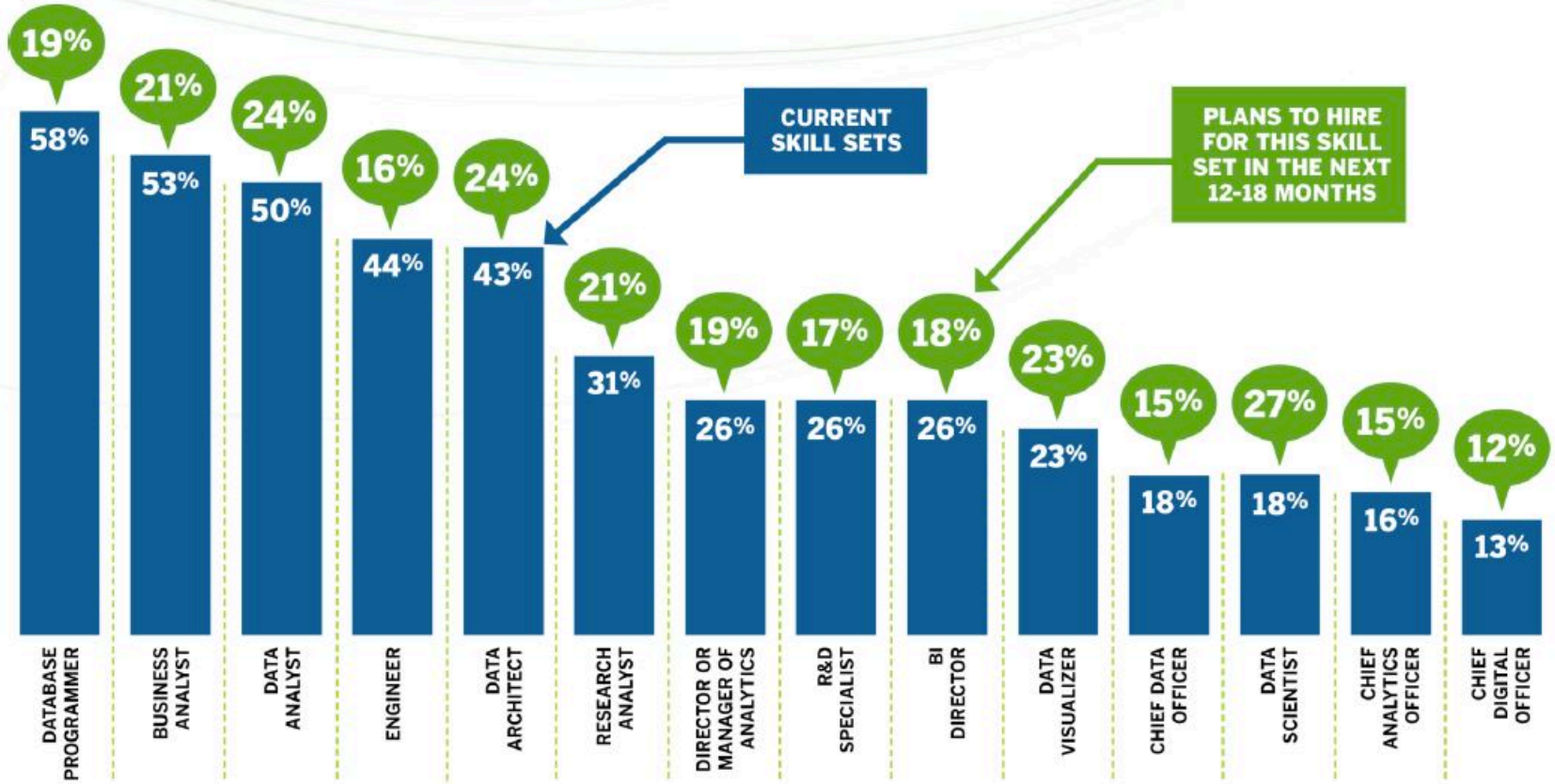
BIG DATA use-cases / applications from a sample of over 100 companies



http://jameskaskade.com/?p=2088

Staffing Up for Big Data Initiatives

IDG Enterprise
An IDG Communications Company



Q. With regard to big data initiatives, what skill sets does your organization currently have? AND **Q.** Which skill sets is your organization planning to hire within the next 12-18 months? BASE: Plans to deploy/implement big data projects.

Source: IDG Enterprise Big Data Study, 2014

Significant Big Data Investments Planned

IDG Enterprise
An IDG Communications Company

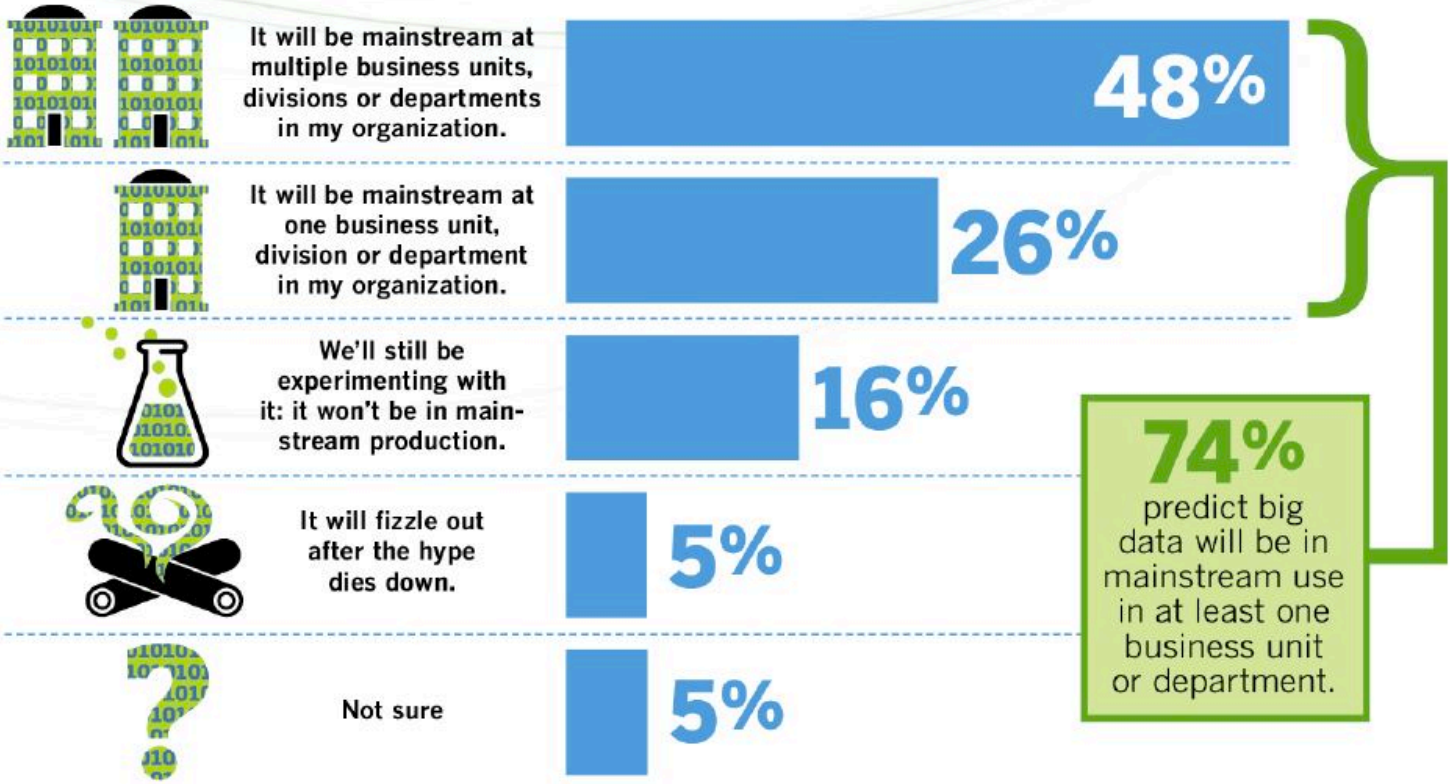
Company Spending			
	<1,000	1,000+	Total
$100 million or more	0%	5%	2%
$50 million - $99.9 million	1%	3%	2%
$10 million - $49.9 million	2%	9%	5%
$5 million - $9.9 million	3%	12%	7%
$1 million - $4.9 million	10%	17%	13%
$100,000 - $999,999	31%	24%	28%
Less than $100,000	31%	7%	19%
Not sure	22%	23%	24%

Q. Approximately how much will your organization spend on big data-related initiatives in the next 12 months? BASE: Plans to deploy/implement big data projects.

Source: IDG Enterprise Big Data Study, 2014

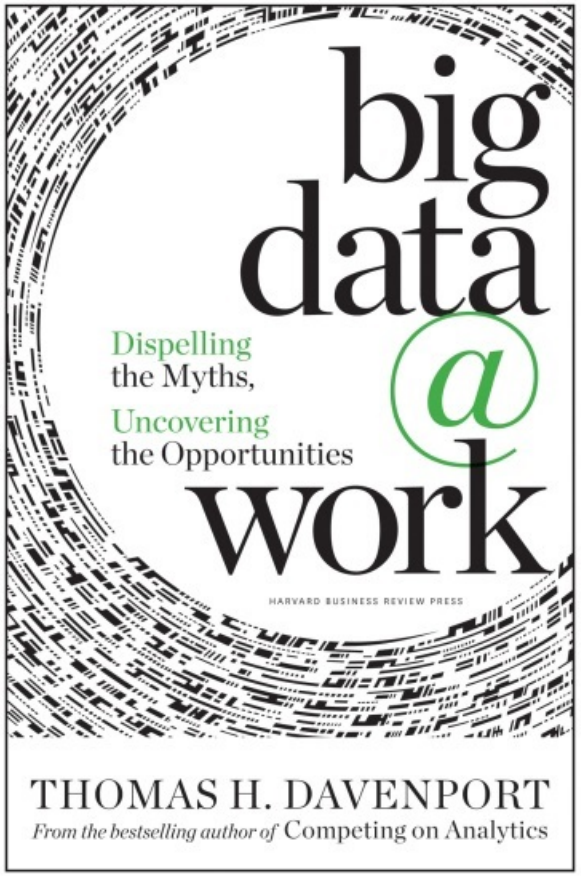
Big Data Is Here to Stay

IDG Enterprise
An IDG Communications Company



Q. Which of these statements best describes your prediction about big data at your organization three years from now? BASE: Plans to deploy/implement big data projects.

Source: IDG Enterprise Big Data Study, 2014



What separates Big Data from past marketing trends like online analytic processing (OLAP) or business intelligence (BI)?

"the *unstructured nature of Big Data* does put it in a different category. Basically, what it means is that you have to do a lot of things before you can do the kinds of analysis you describe. You have to get it into rows and columns.

If it's video you have to do facial recognition on it. If it's text you have to count word frequency, that sort of thing, before you can do any real analysis and make any decision on the basis of it.

The other difference is that because analytics have gotten a bit more advanced compared to OLAP and BI, it's a little bit more likely for people to do *statistical analysis* on it. There's a very high interest in *visual representation of Big Data*. That's not terribly different from BI, but there's a little stronger emphasis I would say."

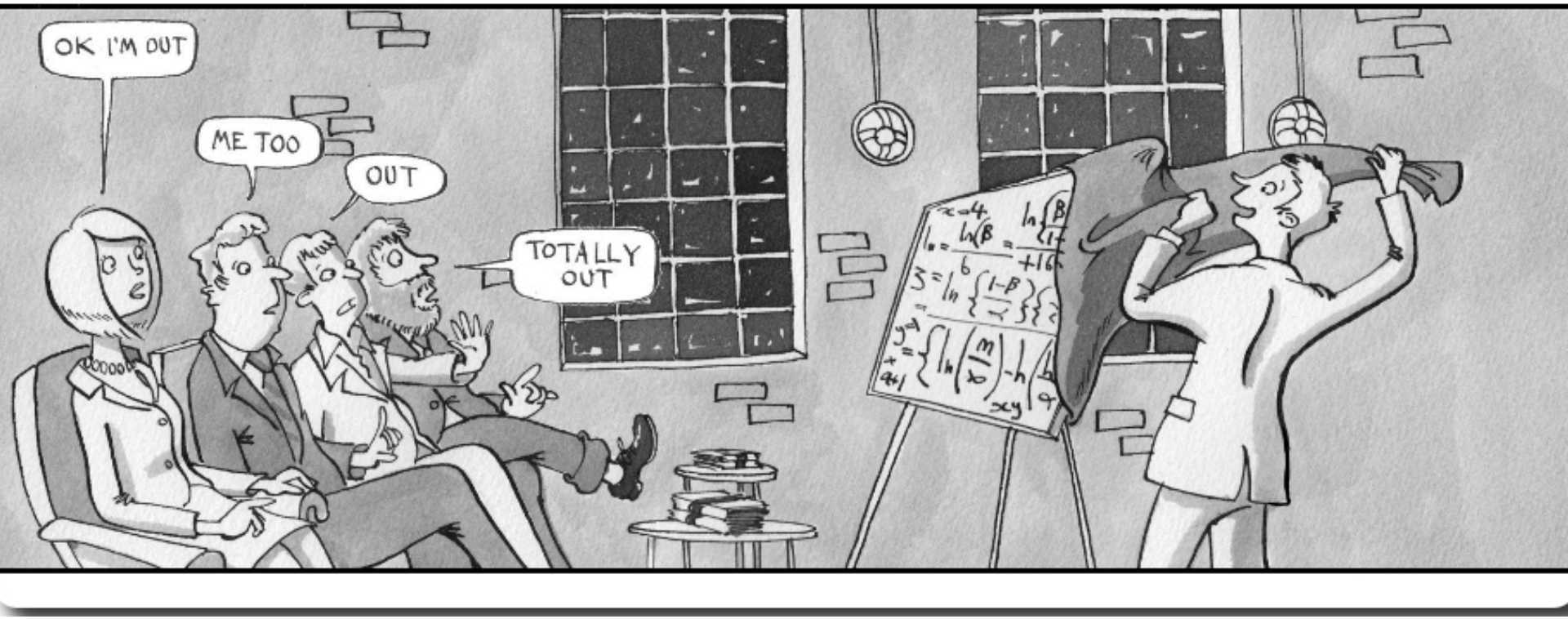
http://www.dmnews.com/direct-line-blog/qa-why-big-data-works/article/329975/

Discussion

What are the opportunities for Africa in Big Data analysis?

Break

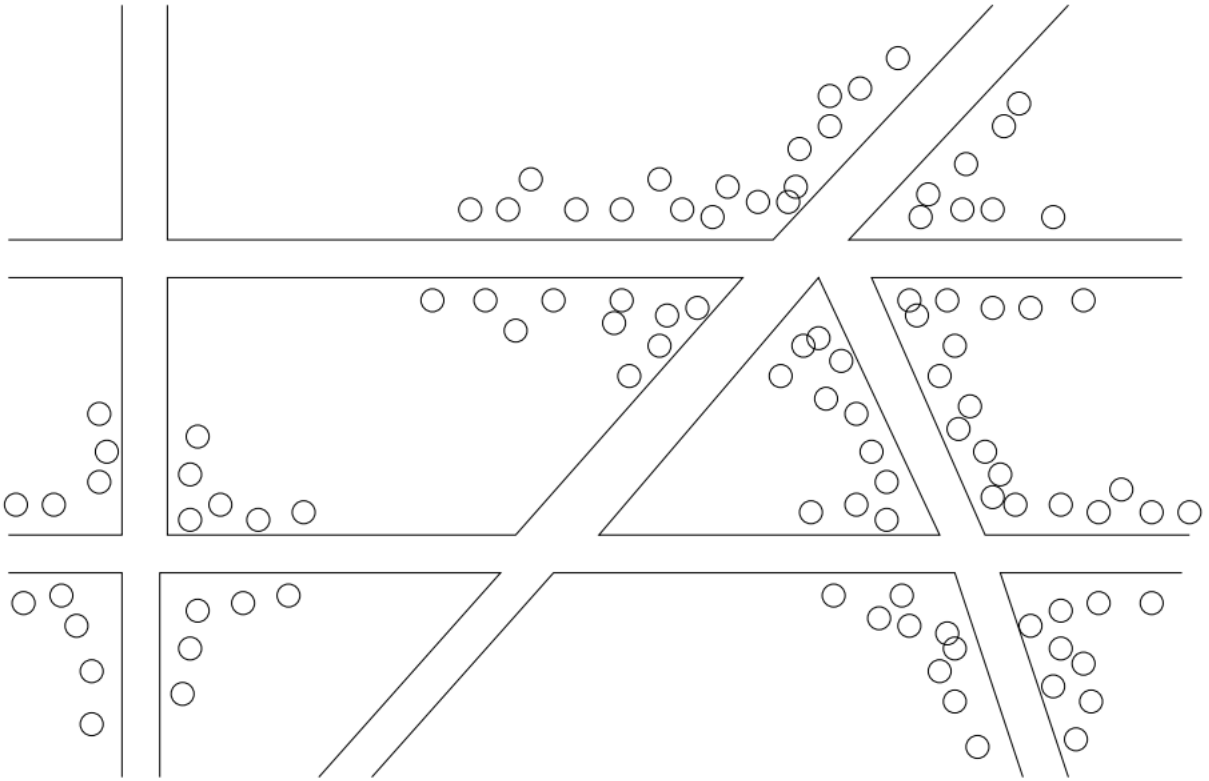
Big Data Revealed



http://www.thebigdatainsightgroup.com/site/article/big-data-talk-003-getting-board-big-data

The first “big data” example?

The Broadstreet cholera outbreak identified by John Snow



Clustering: the intersection was a well pump. Close to a cesspit.

What is driving big data - II

- Make a list of every intimate aspect of your life.

- These *probably* can be inferred (with enough good data and good analysis) from apparently completely independent actions and you will not be aware of the link.

For example…

Gender

Childhood

Sexuality

Financial status

Proclivities

Hobbies

Beliefs

Addictions…

How you walk…

How you write…

How your eyes move…

How you breathe…

The tone of your voice…

Your friends…

Why is this?

We use the same parts of the brain for many “trivial” activities as we do for “important” activities.

By observing “trivial” activities (perhaps in great detail) one can decode and infer the important ones…

Gender from pronouns

Feature	Definition	Genre	Female μ ± stderr	Male μ ± stderr	t-test	Female median	Male median	Mann-Whitney U test
1p	I, me, my, mine, myself, we, us, our, ours, ourselves	Nonfic	149 ± 14	86 ± 8	p<0.0002	66.7	50.2	p<0.1
1p-sing	I, me, my, mine, myself	Nonfic	98.8 ± 11	45.0 ± 6.3	p<0.0001	31.0	18.8	p<0.005
1p-plu	we, us, our, ours, ourselves	Nonfic	49.7 ± 4.5	40.9 ± 3.4	**n/s**	27.8	23.7	**n/s**
2p	you, your, yours, yourself	Nonfic	63.9 ± 8.0	30.0 ± 5.2	p<0.0005	16.7	3.9	p<0.0001

Argamon et al - "Gender, Genre, and Writing Style in Formal Written Texts"

Machine **Human**

Data Management

Human Cognition

Data Mining

Perception

Machine Learning

Visualization

Story Telling

Business Intelligence

Decision Making

Theory

Statistics

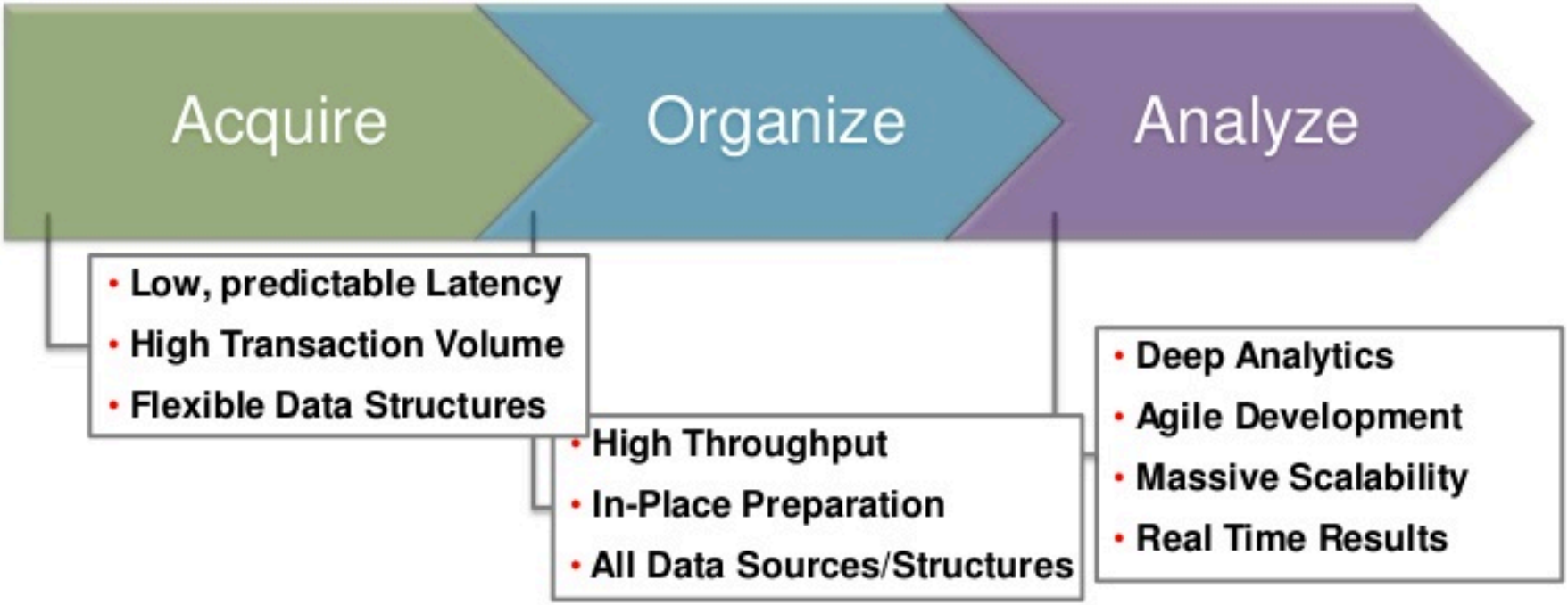
Data Science

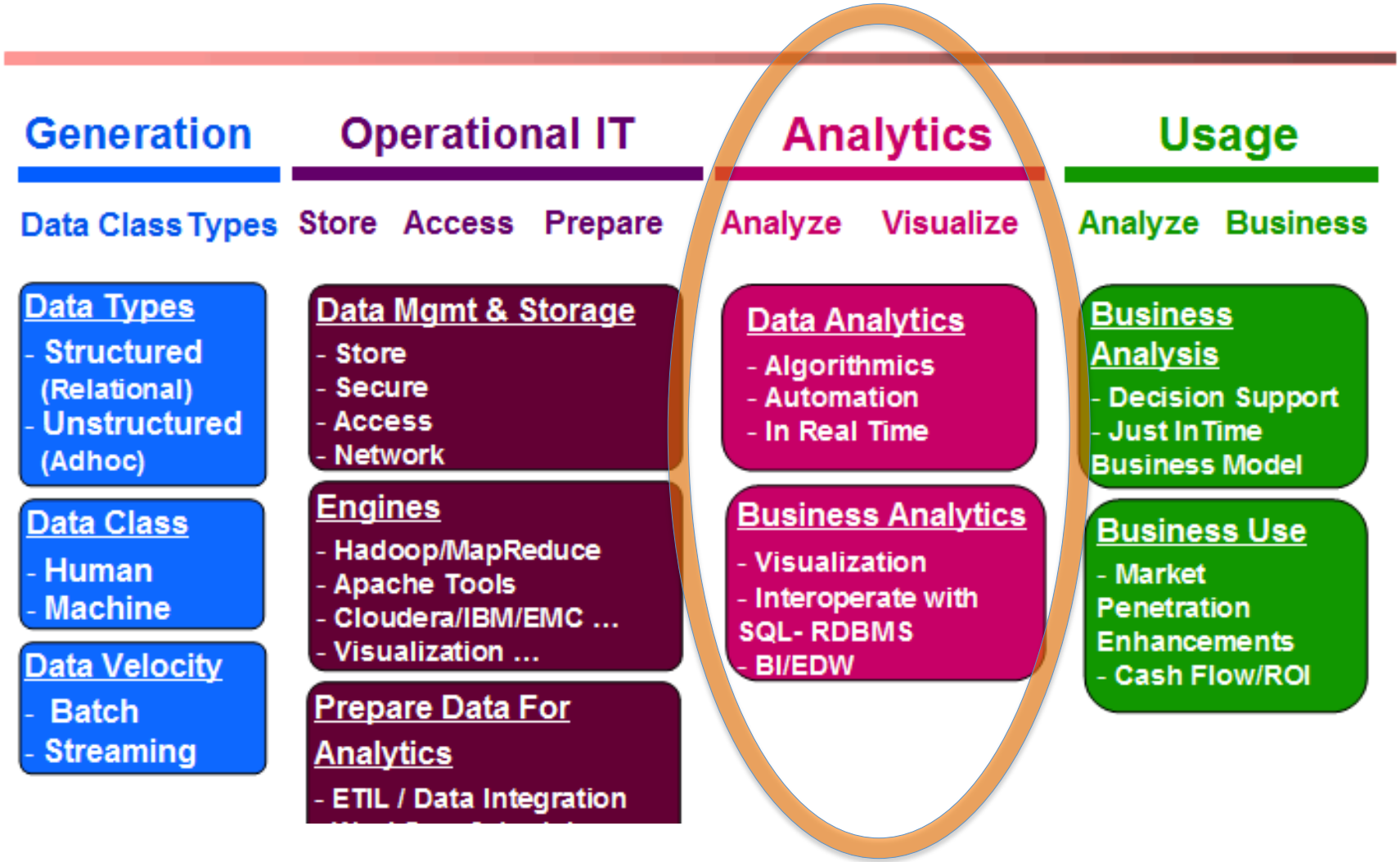
Hanspeter Pfister
& Joe Blitzstein

Any question?

BIG DATA
Requirements

Big Data: Infrastructure Requirements



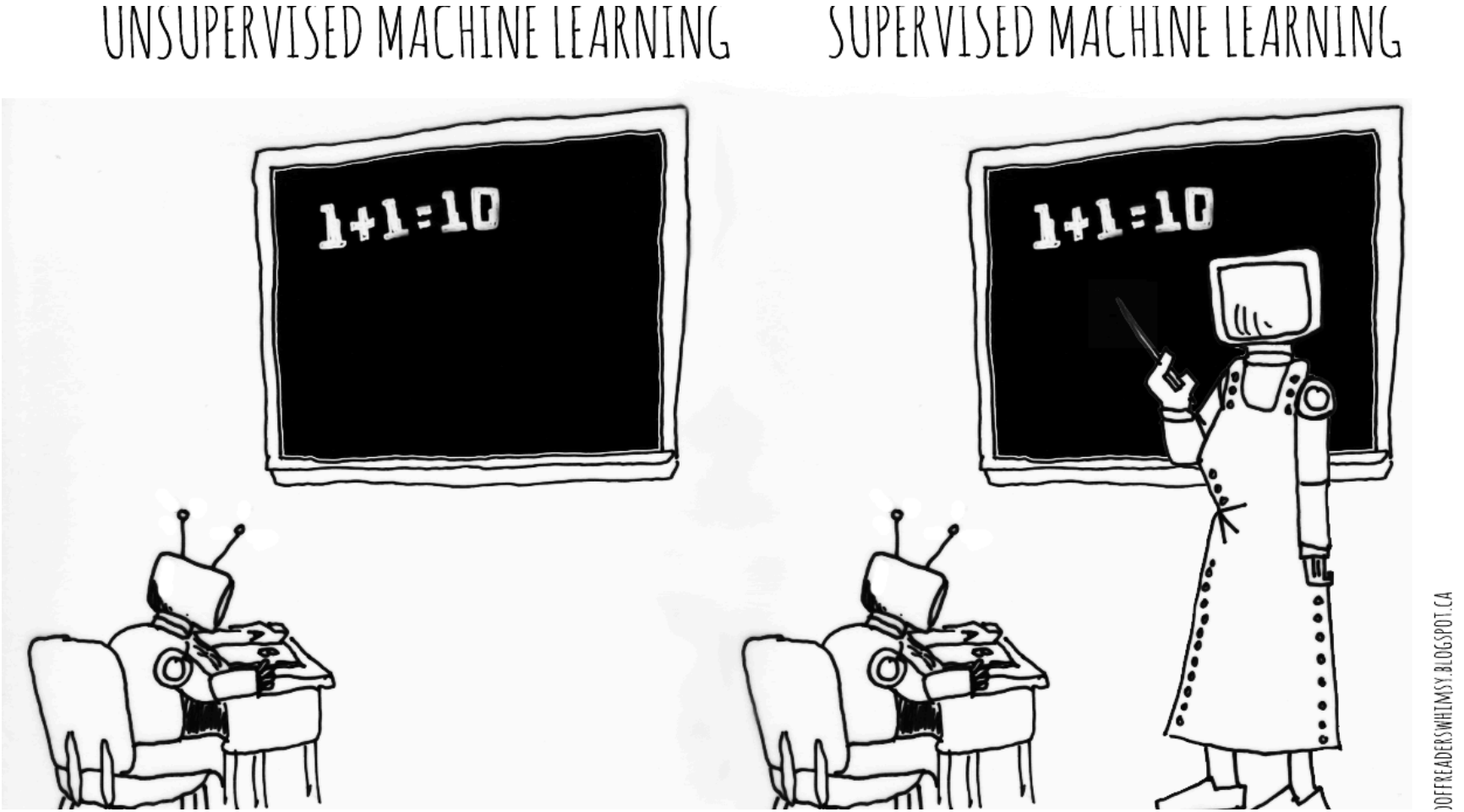


DATA SCIENCE

Machine Learning comes here

http://www.imexresearch.com/newsletters/BDEconomics.html

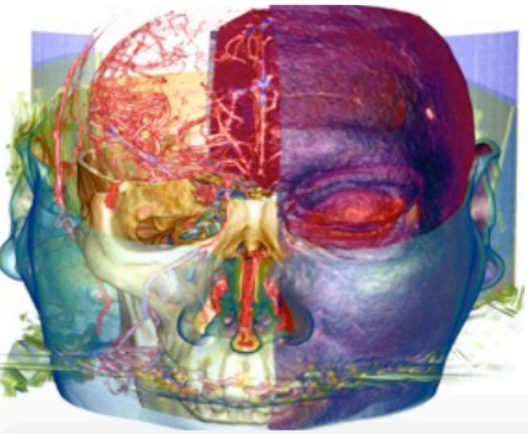
Machine Learning

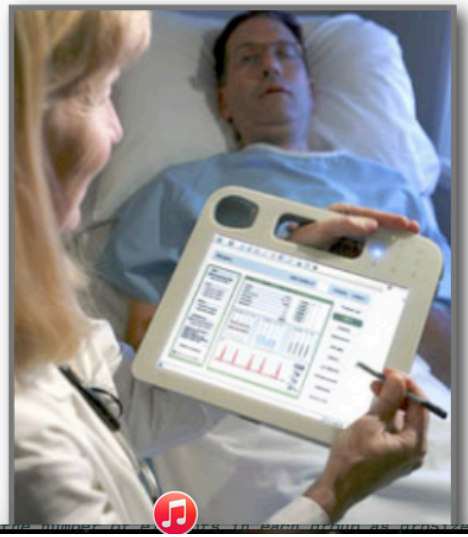


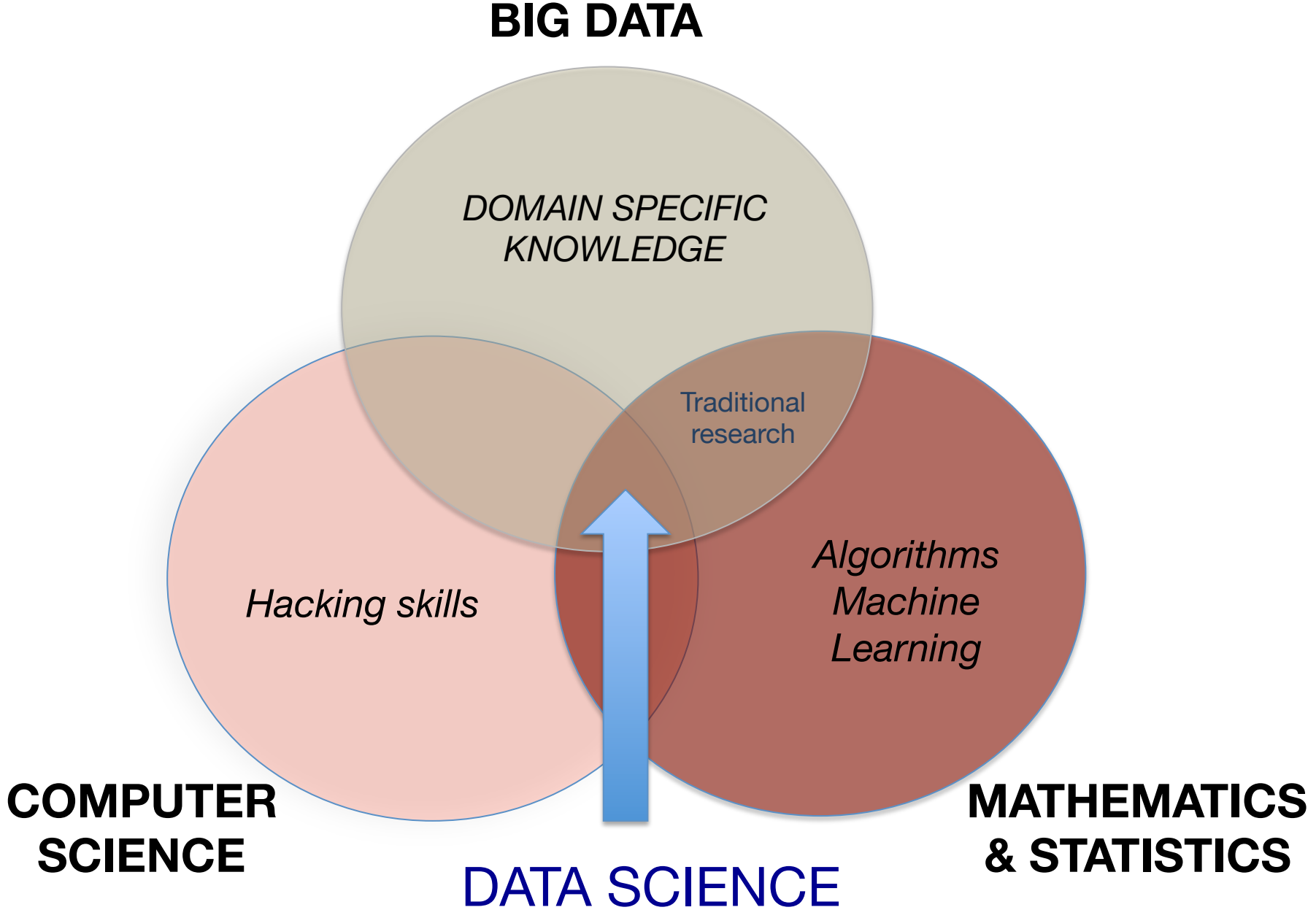
Data Science

Data Science

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





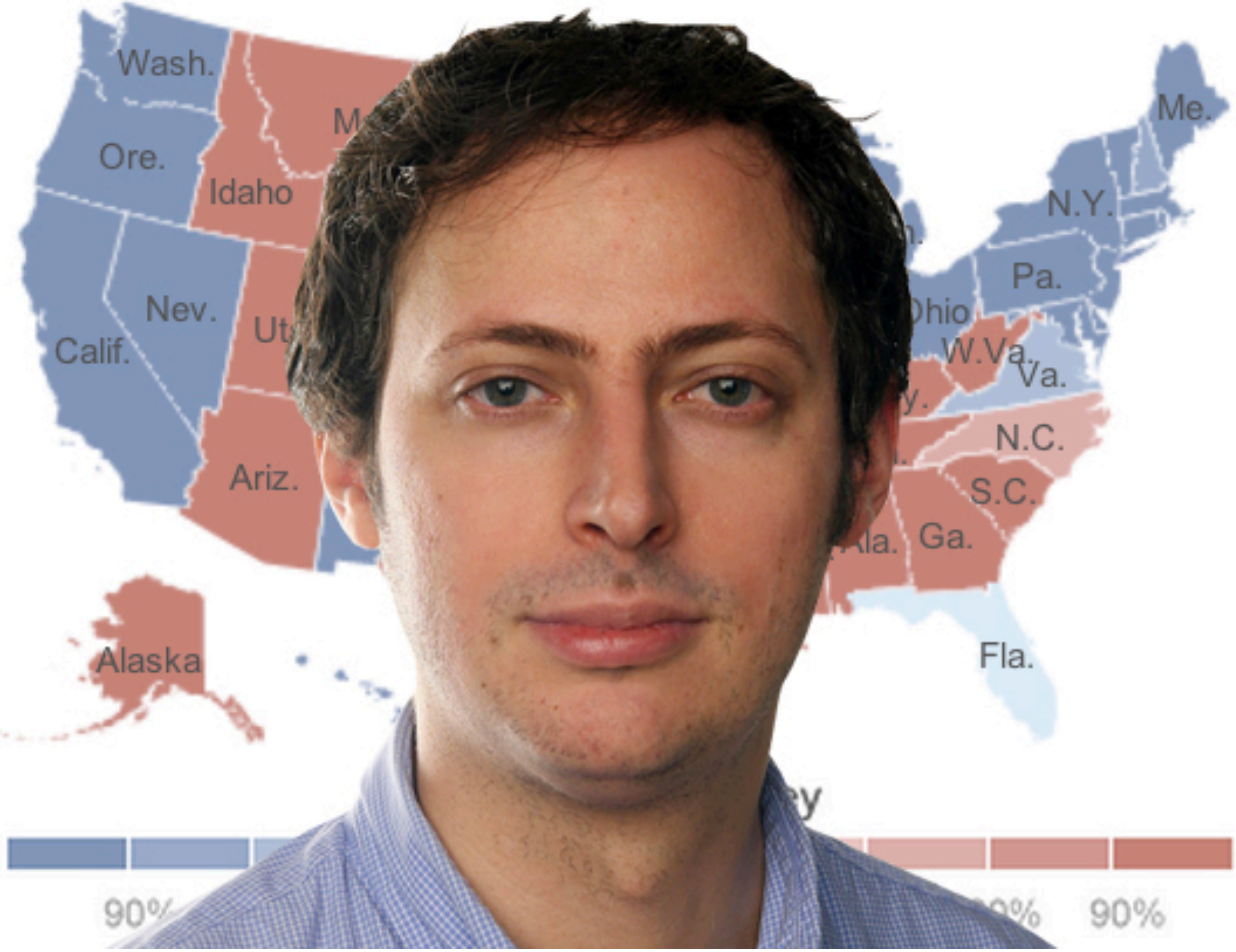

BIG DATA
DOMAIN SPECIFIC KNOWLEDGE
Traditional research
Hacking skills
Algorithms Machine Learning
COMPUTER SCIENCE
MATHEMATICS & STATISTICS
DATA SCIENCE

You are a data scientist!

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

- Josh Blumenstock

Nate Silver: data scientist



Named as one of The World's 100 Most Influential People by Time

Nate Silver: a data science witch

2012 U.S. elections

Nate Silver's Map

The Actual Map

Nate Silver prediction:
Barack Obama a 90.9% chance of winning a majority of the 538 electoral votes
predicted the winner of every one of the 50 states and the District of Columbia

Nate Silver Key Principles

- use many data sources
- 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
- have good communication skills (What does a 60% probability even mean? How can we visualize, validate, and understand the conclusions?)

To learn more about how to predict elections check http://www.padjo.org/2014-10-30/

Reading Assignment

Why Big Data is a Big Deal

http://harvardmagazine.com/2014/03/why-big-data-is-a-big-deal

http://b-i.forbesimg.com/marketshare/files/2013/05/5_big_data.jpg

The course

- Review of basic python with ipython notebook
- Introduction to Pandas
- Introduction to APIs
- Machine Learning using scikit-learn
- Kaggle Titanic challenge
- Project: Twitter data mining


OUR BIG DATA ANALYSIS TELLS US THAT ONLY THE TOP PERFORMERS LEAVE FOR HIGHER PAY.
SINCE YOU'RE STILL HERE, IT MEANS YOUR PERFORMANCE IS AVERAGE AT BEST.
THAT'S NOT FAIR!
THAT'S WHAT ALL THE AVERAGE PEOPLE SAY.
Dilbert.com DilbertCartoonist@gmail.com
1-27-14 ©2014 Scott Adams, Inc./Dist. by Universal Uclick