Big data: an introduction

Riccardo Torlone

Università Roma Tre

ROMA TRE
UNIVERSITÀ DEGLI STUDI

Contents

- What?
- Where?
- Why?
- How?

A short video

(http://www.intel.it/content/www/it/it/big-data/big-data-101-animation.html)

"Big Data"??

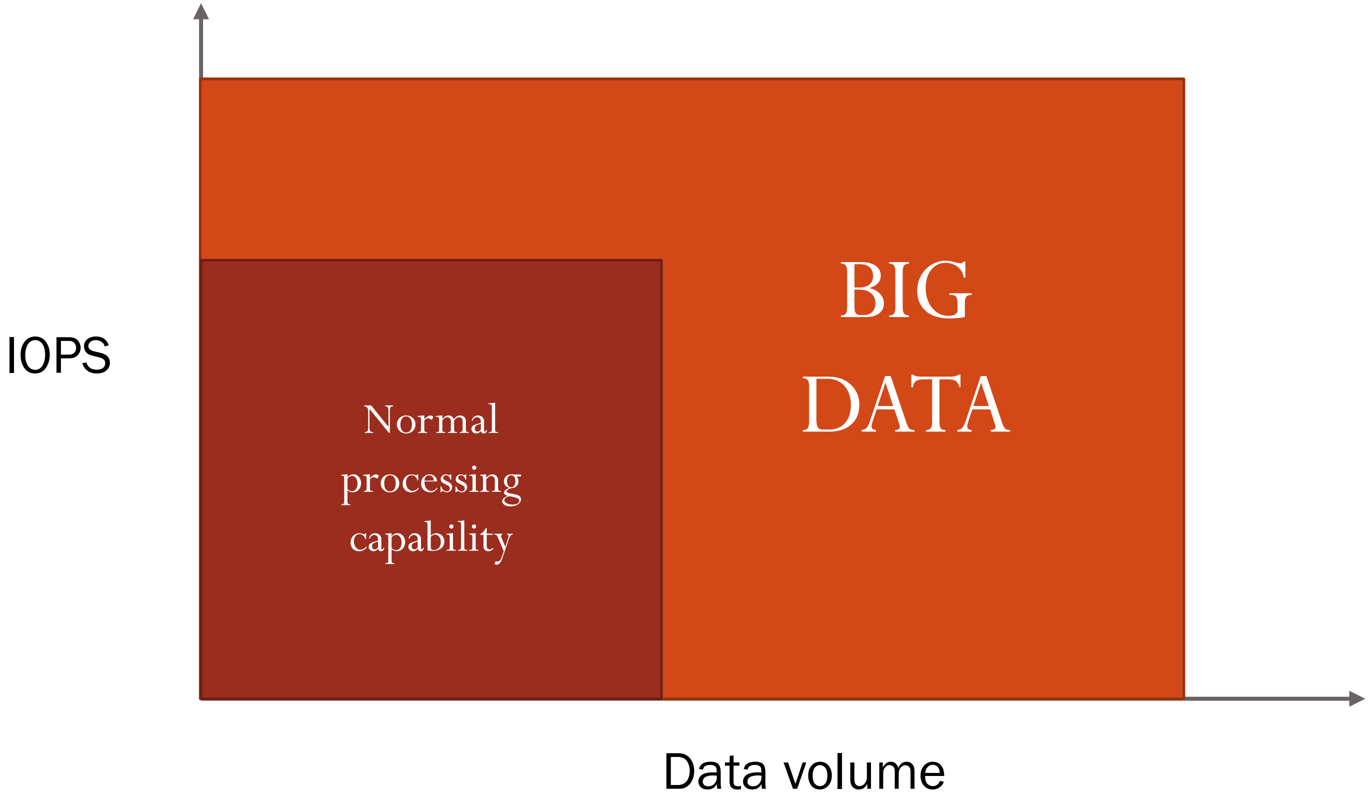
- Different definitions

“Big data exceeds the reach of commonly used hardware environments and software tools to capture, manage, and process it with in a tolerable elapsed time for its user population.” - Teradata Magazine article, 2011

“Big data refers to data sets whose size is beyond the ability of typical database software tools to capture, store, manage and analyze.” - The McKinsey Global Institute, 2012

“Big data is a field that treats of ways to analyze, systematically extract information from, or otherwise deal with data sets that are too large or complex to be dealt with by traditional data-processing application software.” - Wikipedia, 2019

When data become “Big”?



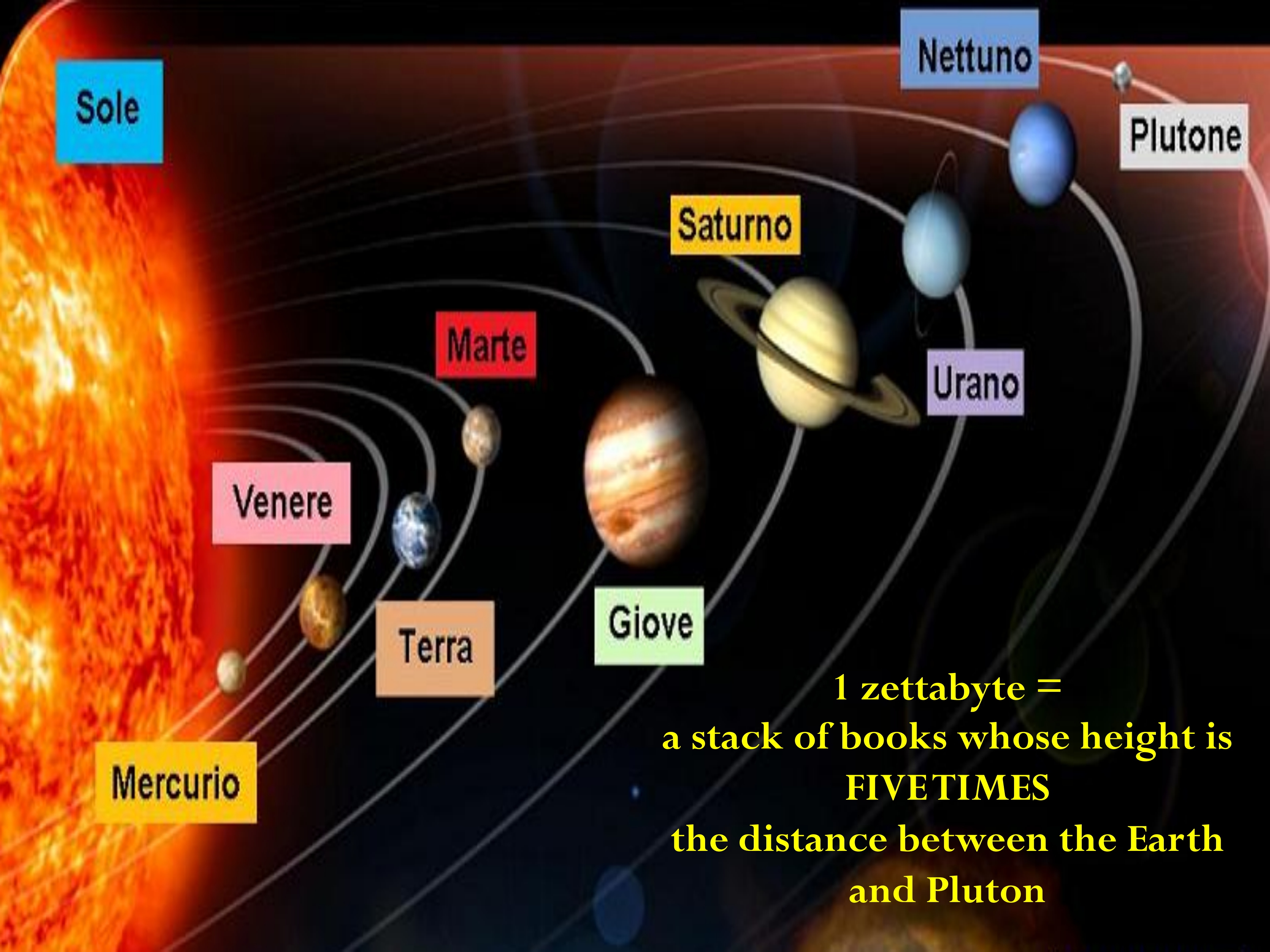
IOPS: Input/Output Operations Per Second

Some numbers

- How many data in the world?
  - 800 Terabytes, 2000
  - 160 Exabytes, 2006 (1EB = 10^{18}B)
  - 4.5 Zettabytes, 2013 (1ZB = 10^{21}B)
  - 44 Zettabytes by 2020
  - 163 Zettabytes by 2025
- How much is a zettabyte?
  - 1,000,000,000,000,000,000,000 bytes
- How many data in a day?
  - 2.5 Exabytes
  - 8 TB, Twitter
  - 50 TB, Facebook
- 90% of world's data:
  - generated over last two years!

How big is big data?

Let us try to make a stack of books containing a zettabyte of data

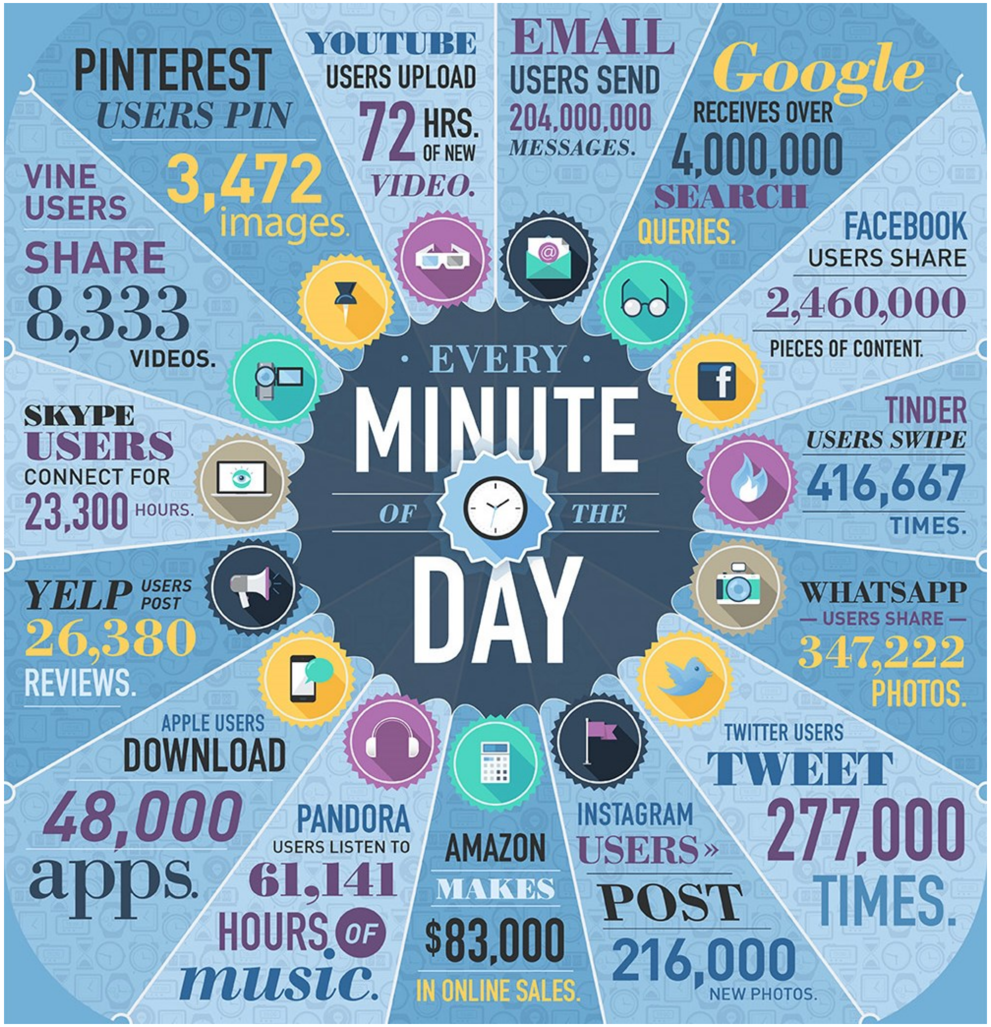


1 zettabyte =
a stack of books whose height is
FIVETIMES
the distance between the Earth
and Pluton

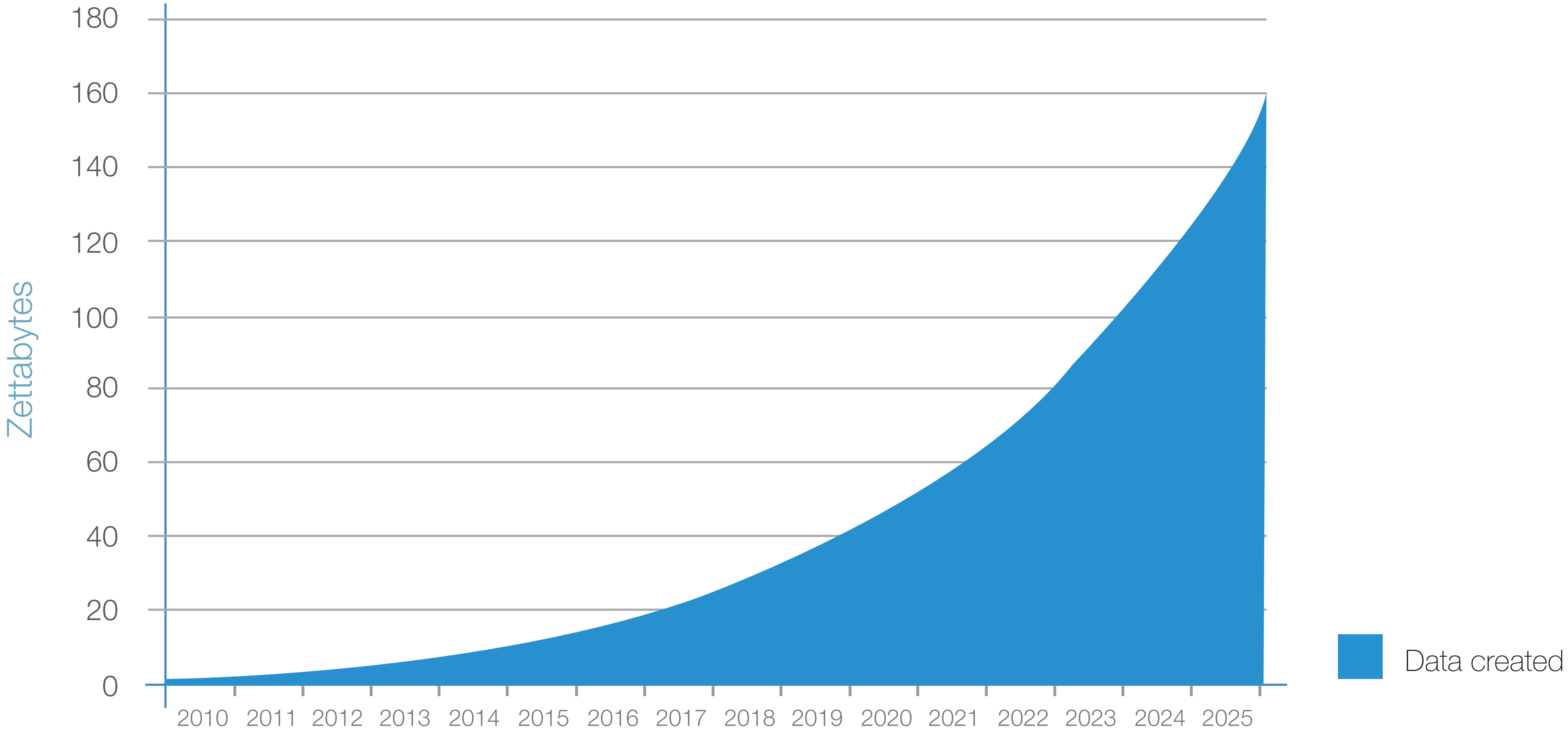
8 million of years of a video in UHD 8K format

Data grows fast!

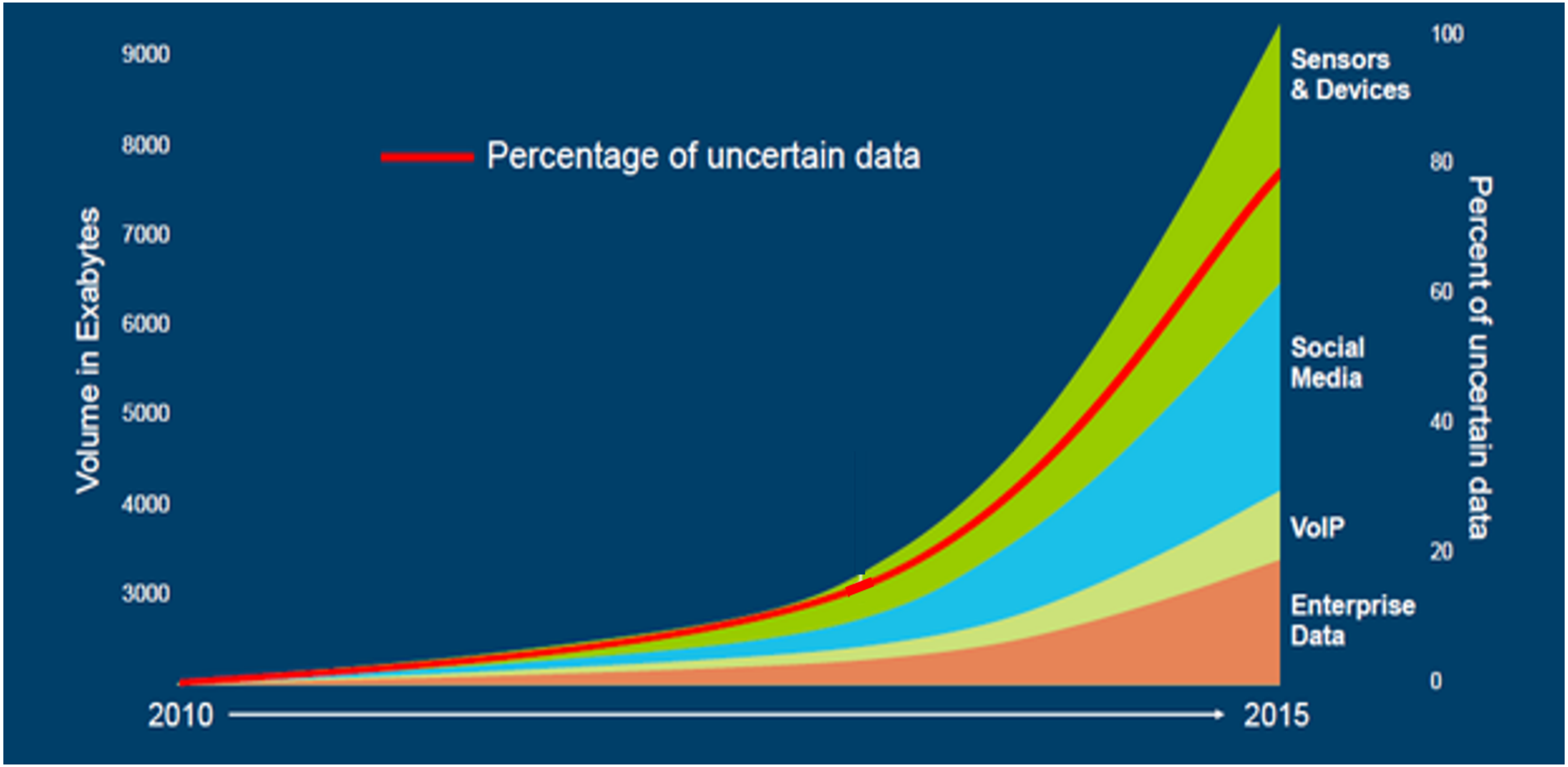
MORE IPHONES
ARE SOLD **THAN BABIES BORN**


EVERY MINUTE OF THE DAY
PINTEREST USERS PIN 3,472 images.
YOUTUBE USERS UPLOAD 72 HRS. OF NEW VIDEO.
EMAIL USERS SEND 204,000,000 MESSAGES.
Google RECEIVES OVER 4,000,000 SEARCH QUERIES.
FACEBOOK USERS SHARE 2,460,000 PIECES OF CONTENT.
TINDER USERS SWIPE 416,667 TIMES.
WHATSAPP USERS SHARE 347,222 PHOTOS.
TWITTER USERS TWEET 277,000 TIMES.
INSTAGRAM USERS» POST 216,000 NEW PHOTOS.
AMAZON MAKES $83,000 IN ONLINE SALES.
PANDORA USERS LISTEN TO 61,141 HOURS OF music.
APPLE USERS DOWNLOAD 48,000 apps.
YELP USERS POST 26,380 REVIEWS.
SKYPE USERS CONNECT FOR 23,300 HOURS.
VINE USERS SHARE 8,333 VIDEOS.

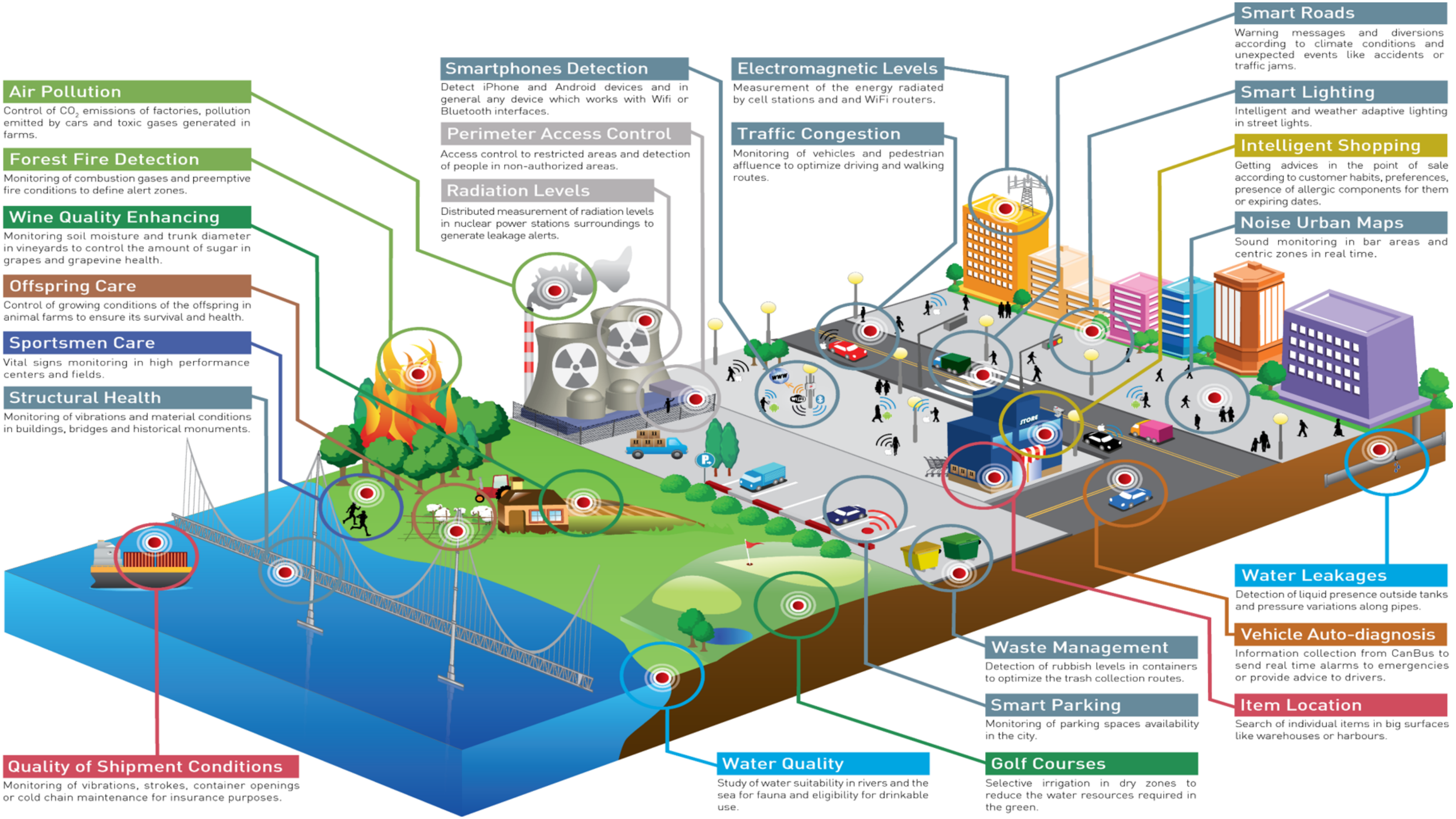
Growth


180
160
140
120
100
80
60
40
20
0
Zettabytes
2010
2011
2012
2013
2014
2015
2016
2017
2018
2019
2020
2021
2022
2023
2024
2025
Data created

Data sources



Proliferation of data sources



Air Pollution
Control of CO_2 emissions of factories, pollution emitted by cars and toxic gases generated in farms.

Forest Fire Detection
Monitoring of combustion gases and preemptive fire conditions to define alert zones.

Wine Quality Enhancing
Monitoring soil moisture and trunk diameter in vineyards to control the amount of sugar in grapes and grapevine health.

Offspring Care
Control of growing conditions of the offspring in animal farms to ensure its survival and health.

Sportsmen Care
Vital signs monitoring in high performance centers and fields.

Structural Health
Monitoring of vibrations and material conditions in buildings, bridges and historical monuments.

Smartphones Detection
Detect iPhone and Android devices and in general any device which works with Wifi or Bluetooth interfaces.

Perimeter Access Control
Access control to restricted areas and detection of people in non-authorized areas.

Radiation Levels
Distributed measurement of radiation levels in nuclear power stations surroundings to generate leakage alerts.

Electromagnetic Levels
Measurement of the energy radiated by cell stations and and WiFi routers.

Traffic Congestion
Monitoring of vehicles and pedestrian affluence to optimize driving and walking routes.

Smart Roads
Warning messages and diversions according to climate conditions and unexpected events like accidents or traffic jams.

Smart Lighting
Intelligent and weather adaptive lighting in street lights.

Intelligent Shopping
Getting advices in the point of sale according to customer habits, preferences, presence of allergic components for them or expiring dates.

Noise Urban Maps
Sound monitoring in bar areas and centric zones in real time.

Water Leakages
Detection of liquid presence outside tanks and pressure variations along pipes.

Vehicle Auto-diagnosis
Information collection from CanBus to send real time alarms to emergencies or provide advice to drivers.

Item Location
Search of individual items in big surfaces like warehouses or harbours.

Quality of Shipment Conditions
Monitoring of vibrations, strokes, container openings or cold chain maintenance for insurance purposes.

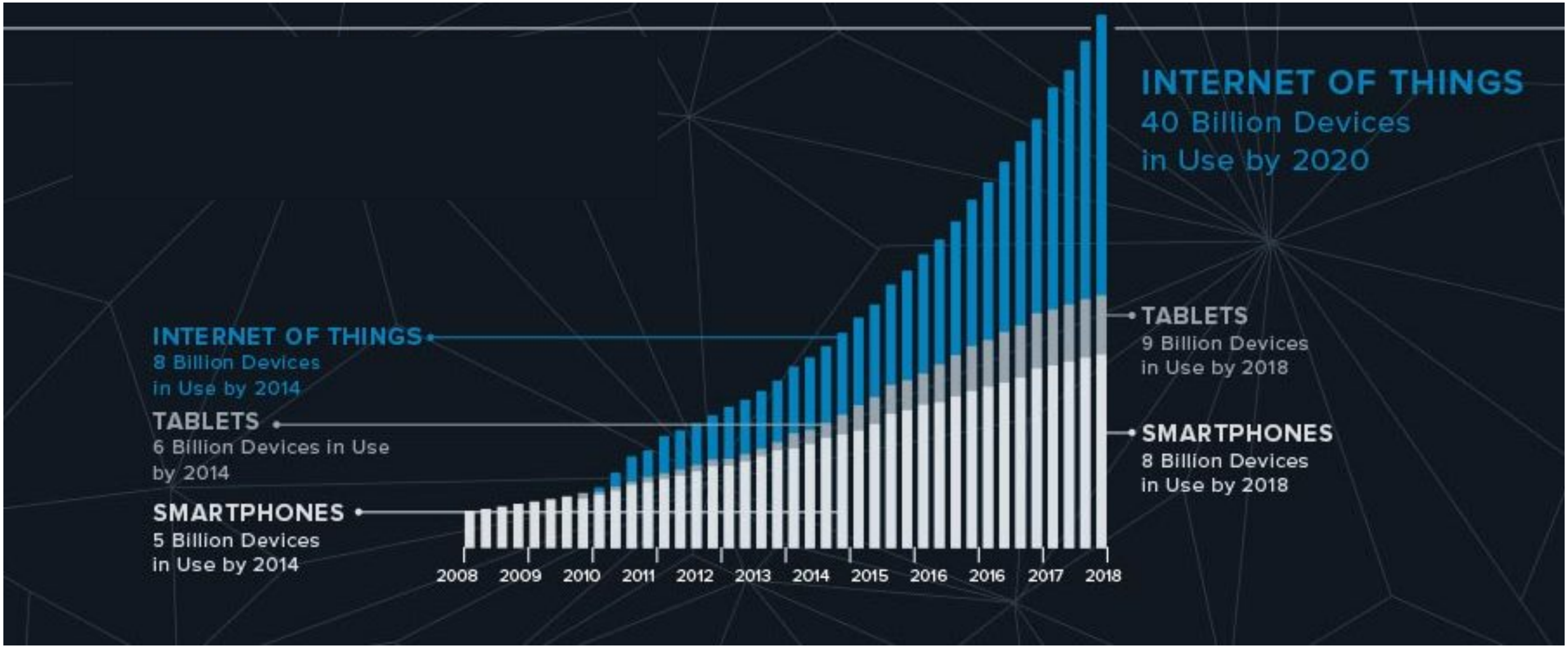
Water Quality
Study of water suitability in rivers and the sea for fauna and eligibility for drinkable use.

Waste Management
Detection of rubbish levels in containers to optimize the trash collection routes.

Smart Parking
Monitoring of parking spaces availability in the city.

Golf Courses
Selective irrigation in dry zones to reduce the water resources required in the green.

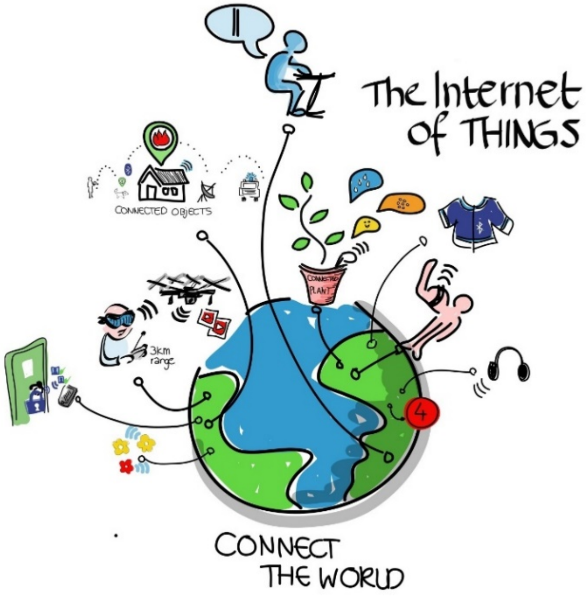
Global Internet device forecast

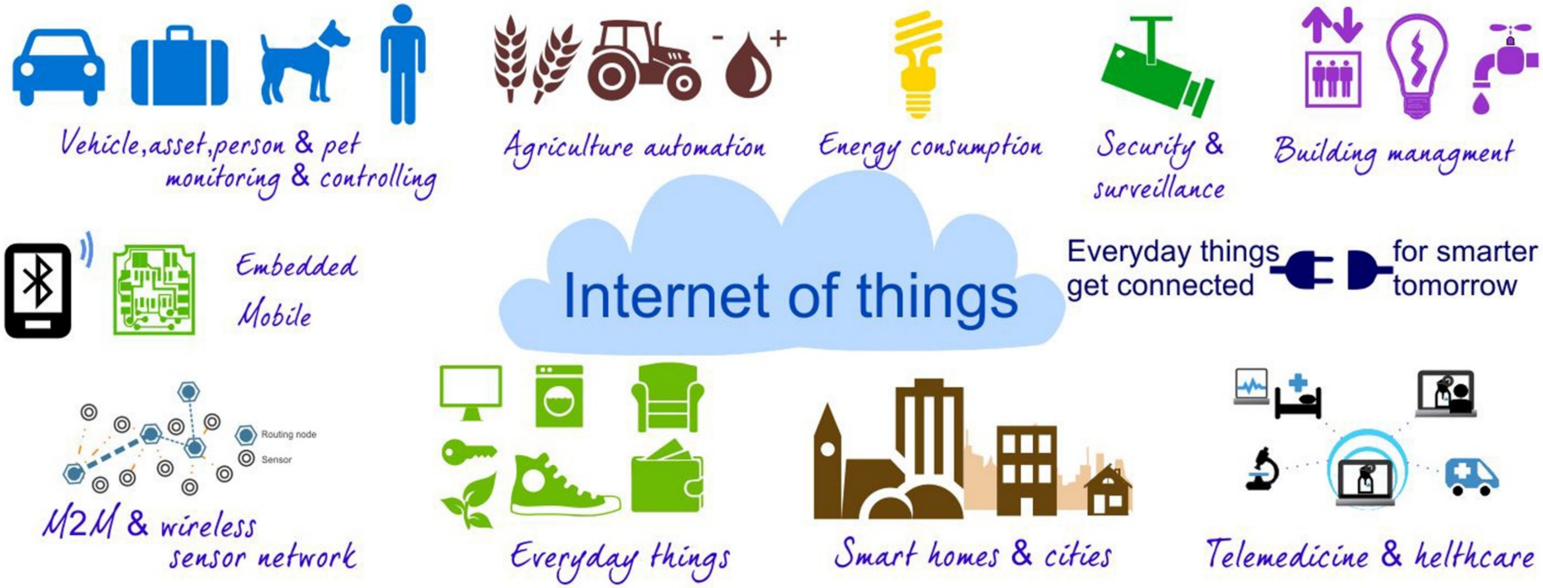


Internet of Things

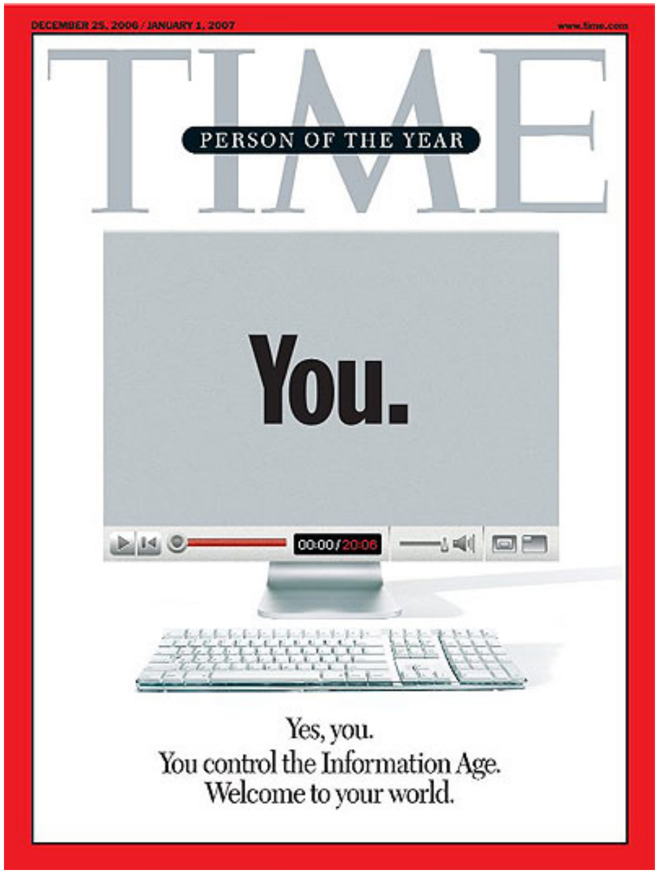
'There will be as many as **40 TO 80 BILLION** connected objects by 2020.

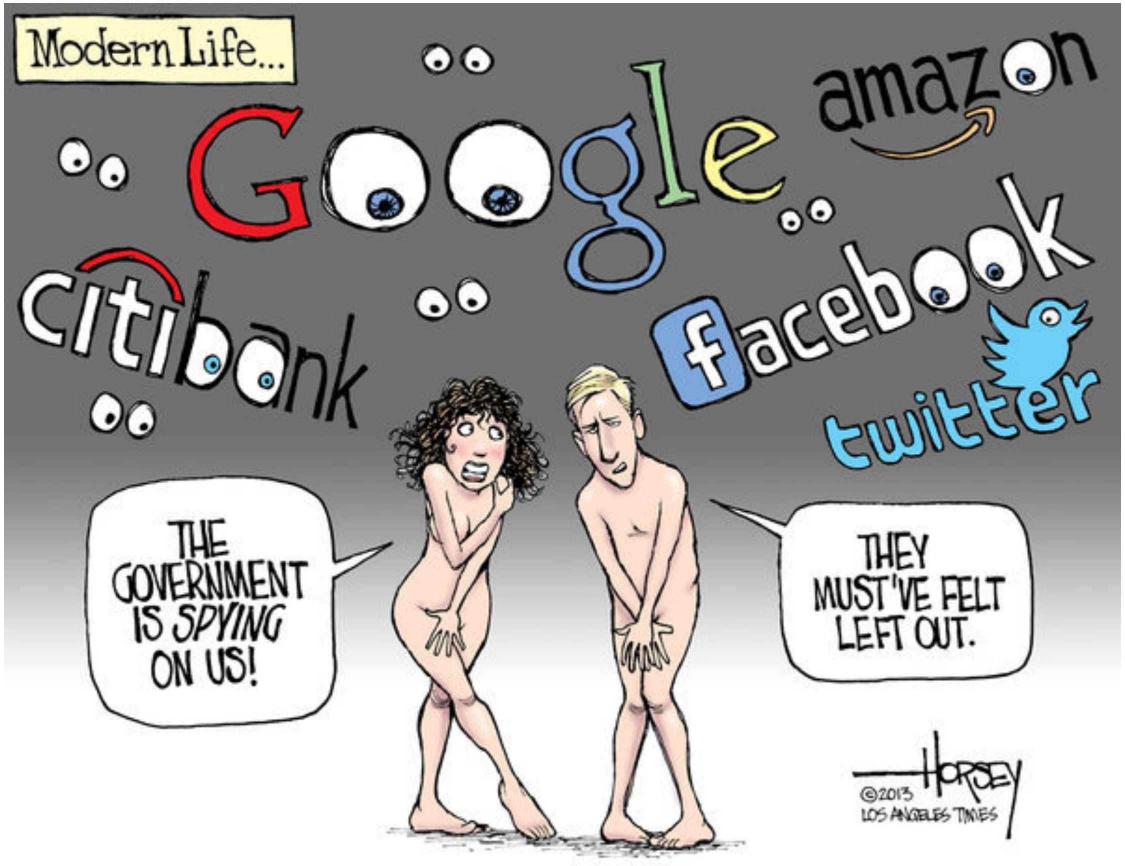
There will be **10** connected objects for every man, woman, and child on the **PLANET.**





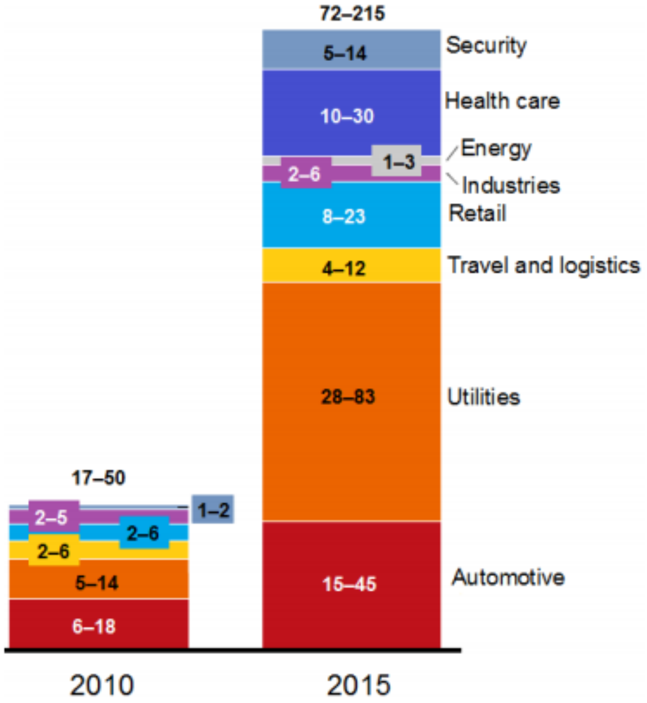
User-generated content

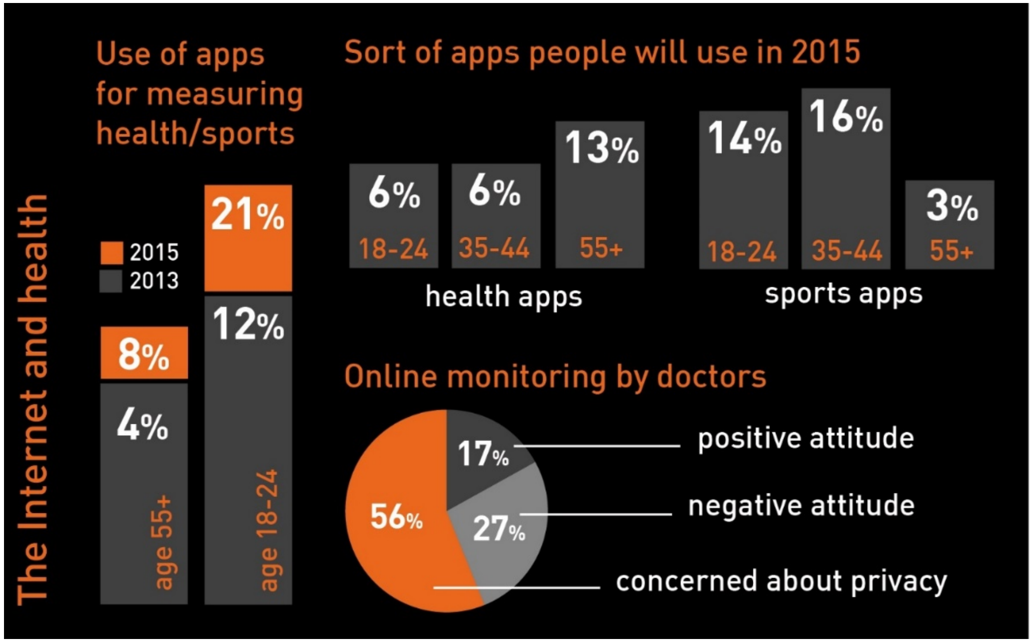




Horizontal spread

Data is central to all of our existences, whether we're a giant enterprise or an individual person





Data types

	Video	Image	Audio	Text/ numbers
Banking	▓	▓	▓	█
Insurance	░	░	░	█
Securities and investment services	░	░	░	█
Discrete manufacturing	▓	▓	░	█
Process manufacturing	▓	▓	░	█
Retail	▓	░	░	█
Wholesale	░	░	░	█
Professional services	▓	▓	▓	█
Consumer and recreational services	▓	░	▓	▓
Health care	░	█	░	█
Transportation	░▓	░▓	░	█
Communications and media[2]	█	▓	█	█
Utilities	░▓	░▓	░	█
Construction	░	▓█	░	▓
Resource industries	▓	▓	░	█
Government	█	▓	█	█
Education	█	▓	█	▓

Penetration

█ High

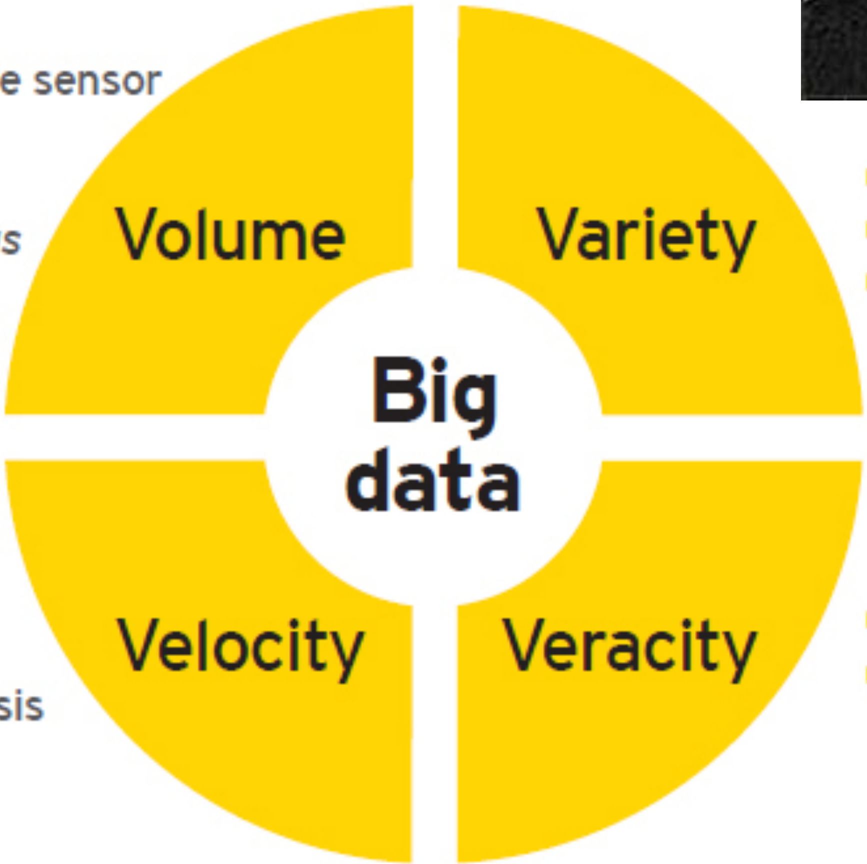
▓ Medium

░ Low

The four "V's" of Big Data

- Not just a matter of volume..

- Click stream
- Active/passive sensor
- Log
- Event
- Printed *corpus*
- Speech
- Social media
- Traditional



- Unstructured
- Semi-structured
- Structured

- Speed of generation
- Rate of analysis

- Untrusted
- Uncleansed

What is more important?

- The “Big”
- The “Data”
- Both
- Neither

What is more important?

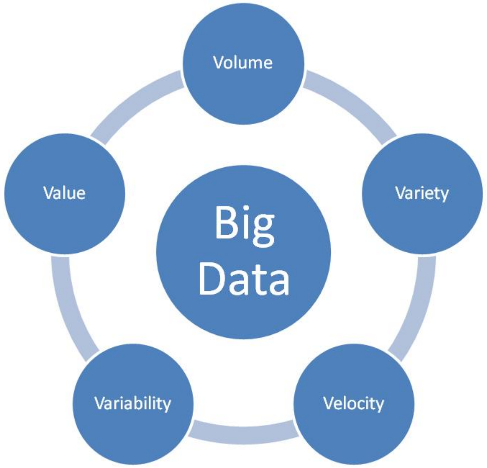
- The “Big”
- The “Data”
- Both
- Neither

What organizations do with big data

"Data is not information, information is not knowledge, knowledge is not understanding, understanding is not wisdom"
Cliff Stoll

Big Data: V^4+VALUE

- Volume： Terabyte(10^{12}), Petabyte(10^{15}), Exabyte(10^{18}), Zettabyte (10^{21})
- Variety: Structured, semi-structured, unstructured; Text, image, audio, video, record
- Velocity: Periodic, Near Real Time, Real Time
- Veracity: Quality of the data can vary greatly
- Value: Big data can generate huge competitive advantages



What’s new?

The wide availability of data allows us to apply more sophisticated models and you get much more accurate results than in the past!

It is a capital mistake to theorize before one has data

Anthony Goldbloom

The bigger the data set you have, the more accurate the predictions about the future will be

In God we trust; all others must bring data

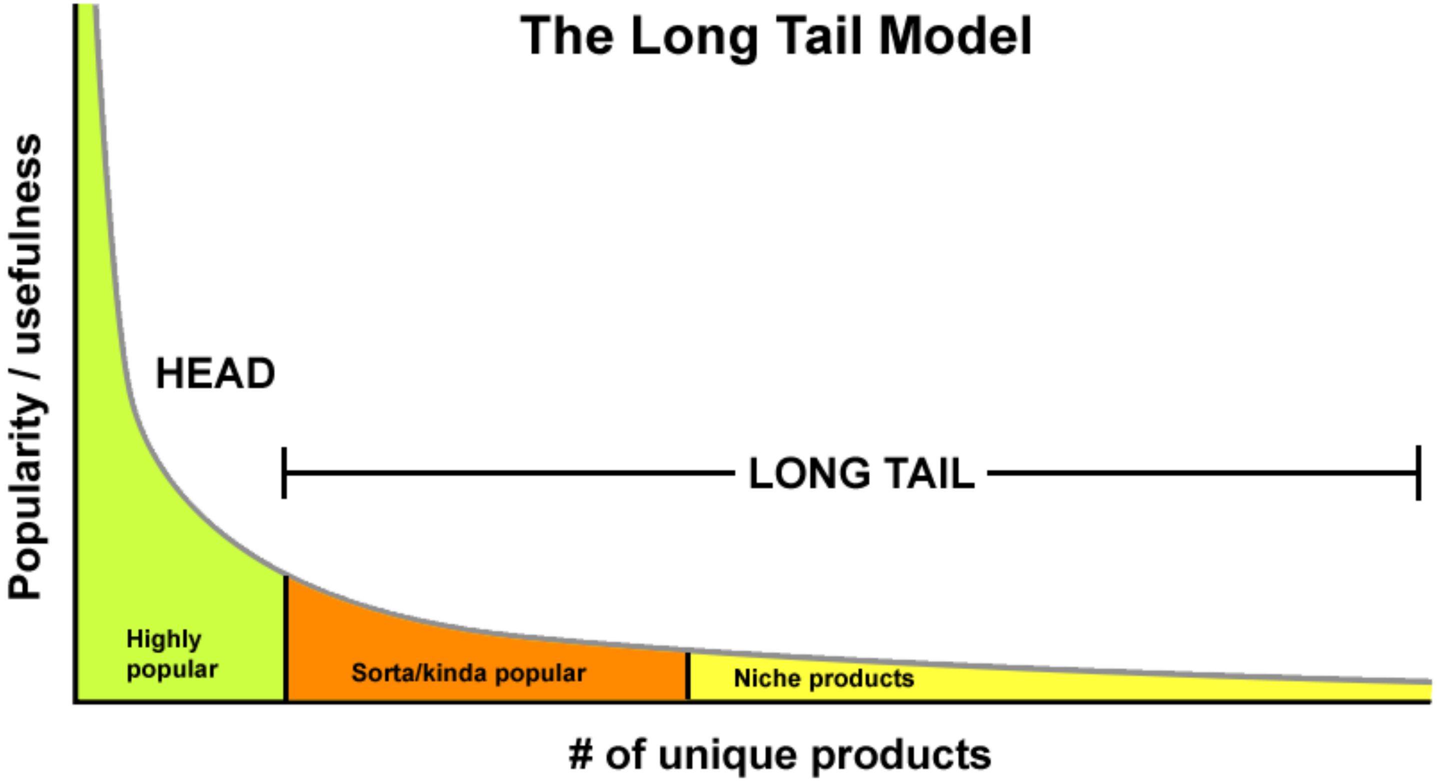
William Deming

Bigger = Smarter!

Bigger = Smarter?

- YES!
  - algorithms work much better
  - tolerate errors
  - discover the "long tail" and "corner cases"
- BUT:
  - more heterogeneity
  - data grows faster than energy on chip
  - still need humans to ask right questions

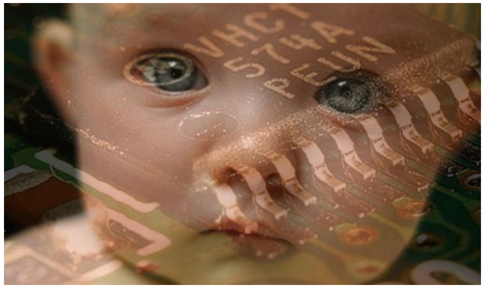
Big tail



“We sold more items today that didn't sell at all yesterday than we sold today of all the items that did sell yesterday” – Amazon employee.

Why now?

- Because we have data
  - Data born already in digital form
  - 40% of data growth per year



- Because we can

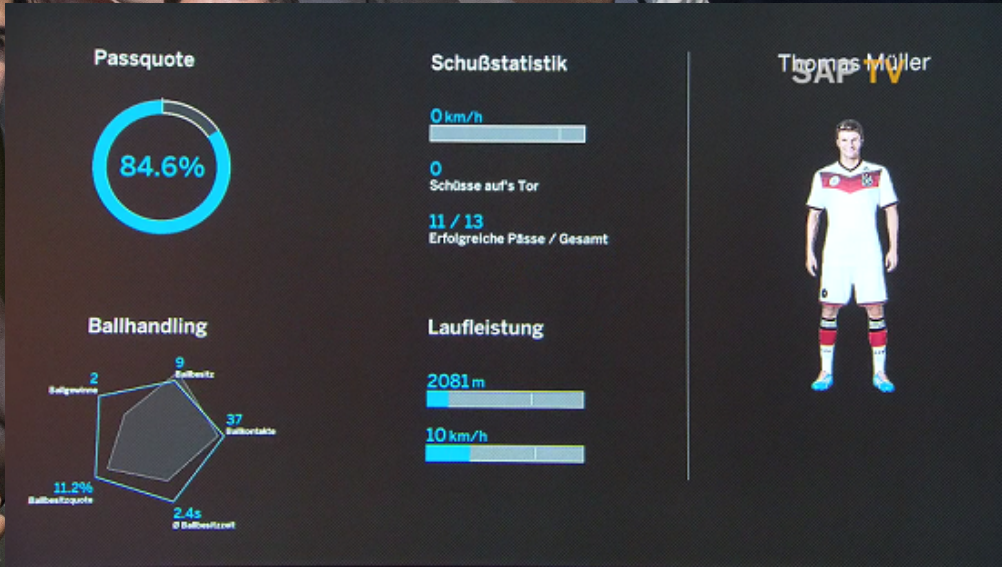
  - 300$ for a drive in which to store all the music of the world
  - >40 years of Moore's Law → large computational resources
  - 68% of companies have invested in big data in 2018
  - 57 billions $ invested in big data in 2018

- “Because we reached dead end with logic”

A simple example of bigger=smarter

- Google Translate
  - you collect snippets of translations
  - you match sentences to snippets
  - you continuously debug your system
- Why does it work?
  - there are tons of snippets on the Web
  - the accuracy improves as the training set grows

A success story



Passquote

84.6%

Schußstatistik

0 km/h

0
Schüsse auf's Tor

11 / 13
Erfolgreiche Pässe / Gesamt

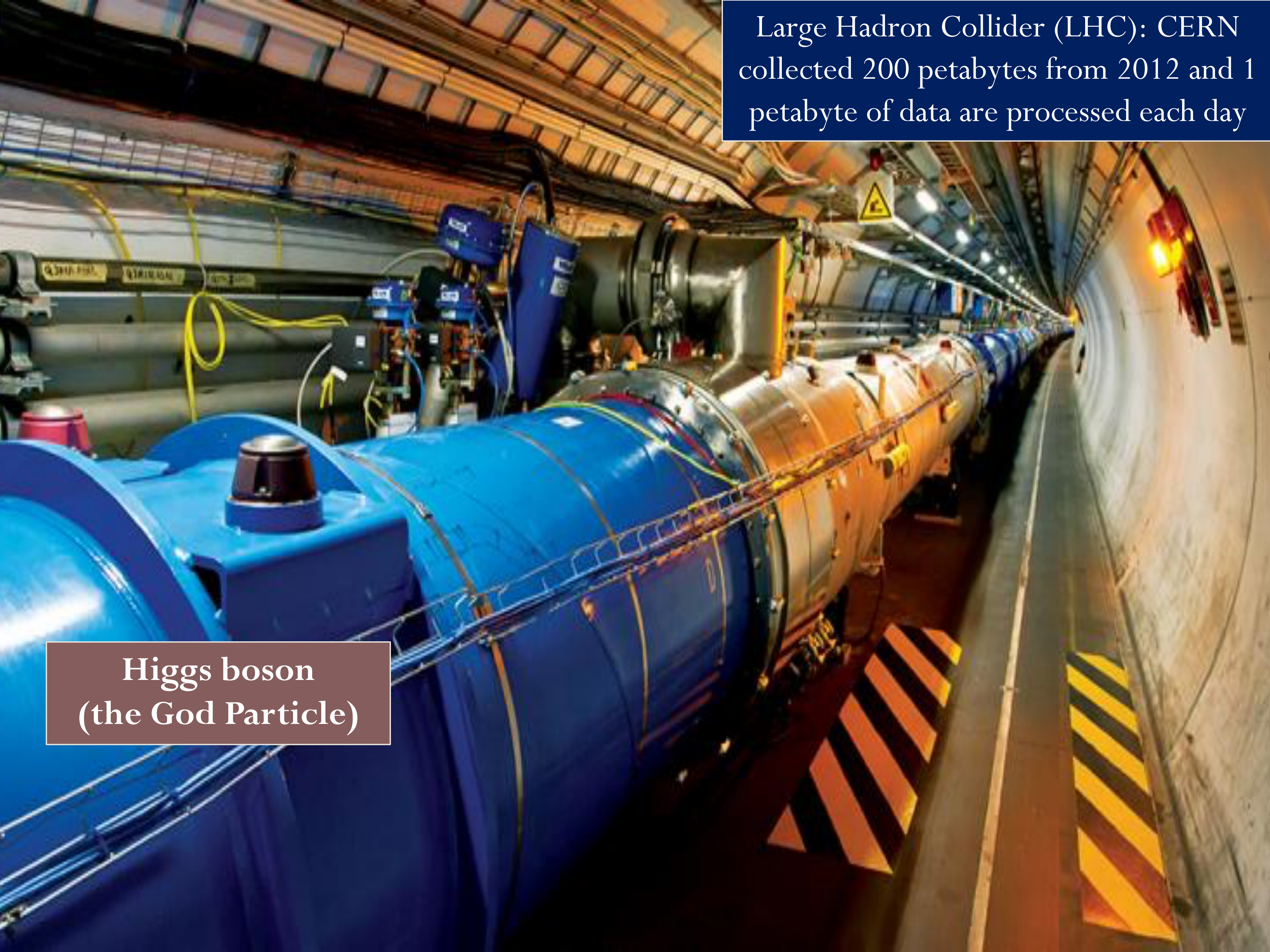
Thomas Müller

Ballhandling

Laufleistung

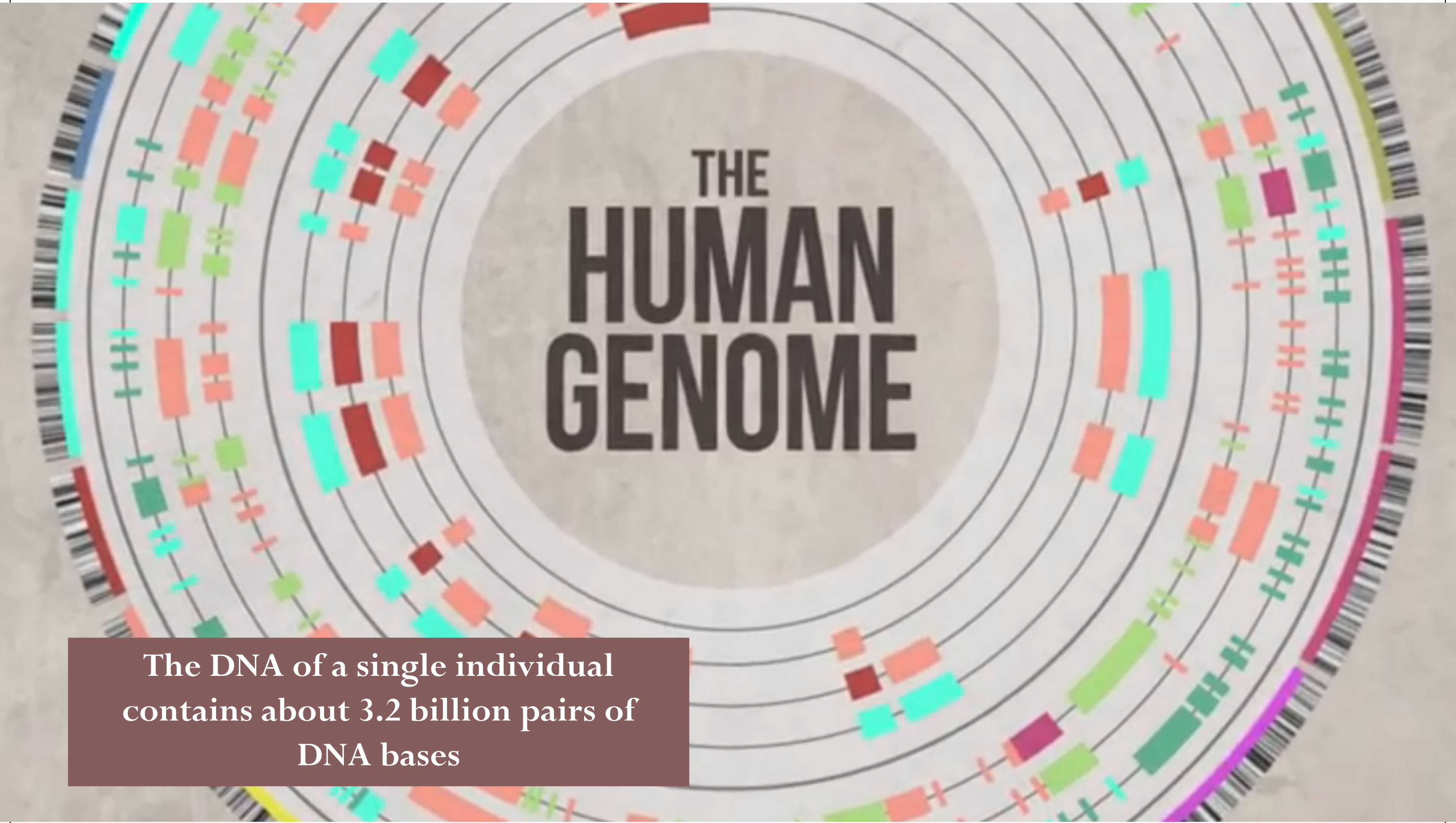
2081 m

10 km/h



Large Hadron Collider (LHC): CERN collected 200 petabytes from 2012 and 1 petabyte of data are processed each day

Higgs boson (the God Particle)



The DNA of a single individual contains about 3.2 billion pairs of DNA bases

..many other stories in very diverse sectors

- Crime Prevention in Los Angeles
- Diagnosis and treatment of genetic diseases
- Investments in the financial sector
- Generation of personalized advertising
- Astronomical discoveries
- …

- but….

Use cases

Today’s Challenge	New Data	What’s Possible
Healthcare Expensive office visits	Remote patient monitoring	Preventive care, reduced hospitalization
Manufacturing In-person support	Product sensors	Automated diagnosis, support
Location-Based Services Based on position	Real time location data	Geo-advertising, traffic, local search
Public Sector Standardized services	Citizen surveys	Tailored services, cost reductions
Retail One size fits all marketing	Social media	Sentiment analysis segmentation

Potential value

US health care

- $300 billion value per year
- ~0.7 percent annual productivity growth

Europe public sector administration

- €250 billion value per year
- ~0.5 percent annual productivity growth

Global personal location data

- $100 billion+ revenue for service providers
- Up to $700 billion value to end users

US retail

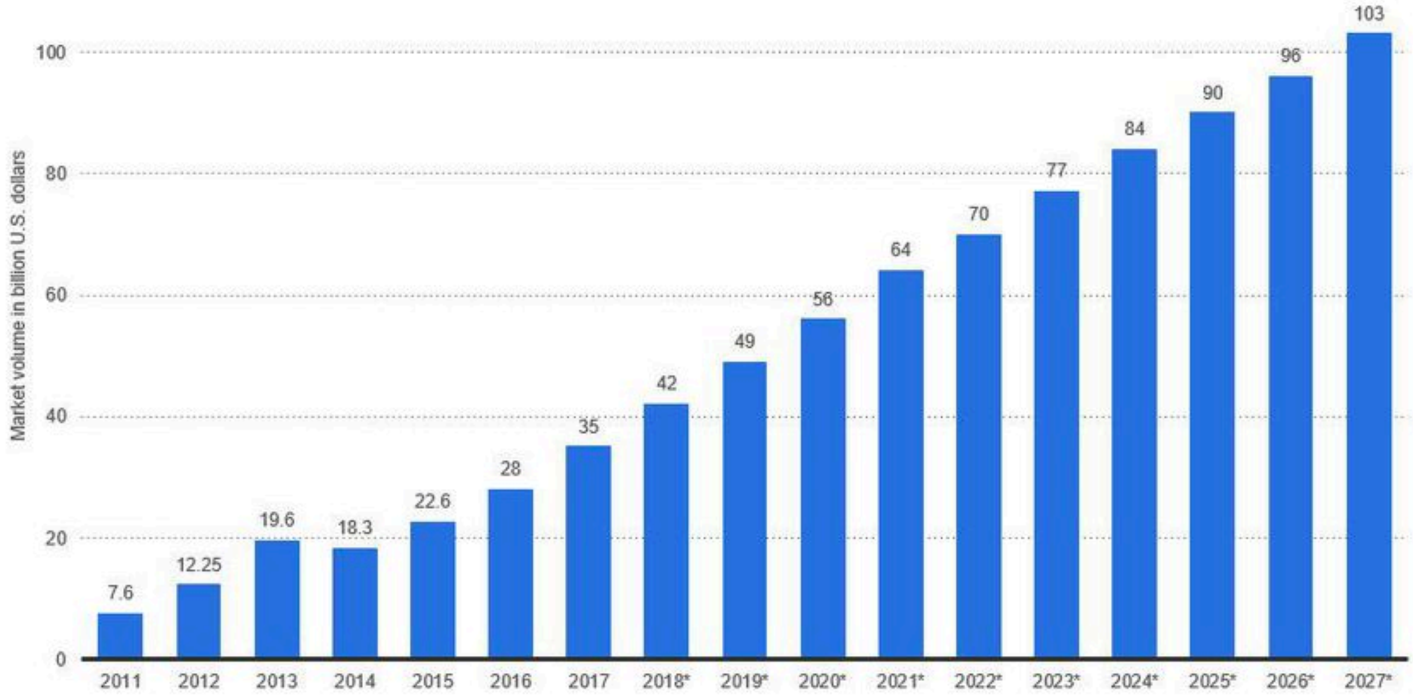
- 60+% increase in net margin possible
- 0.5–1.0 percent annual productivity growth

Manufacturing

- Up to 50 percent decrease in product development, assembly costs
- Up to 7 percent reduction in working capital

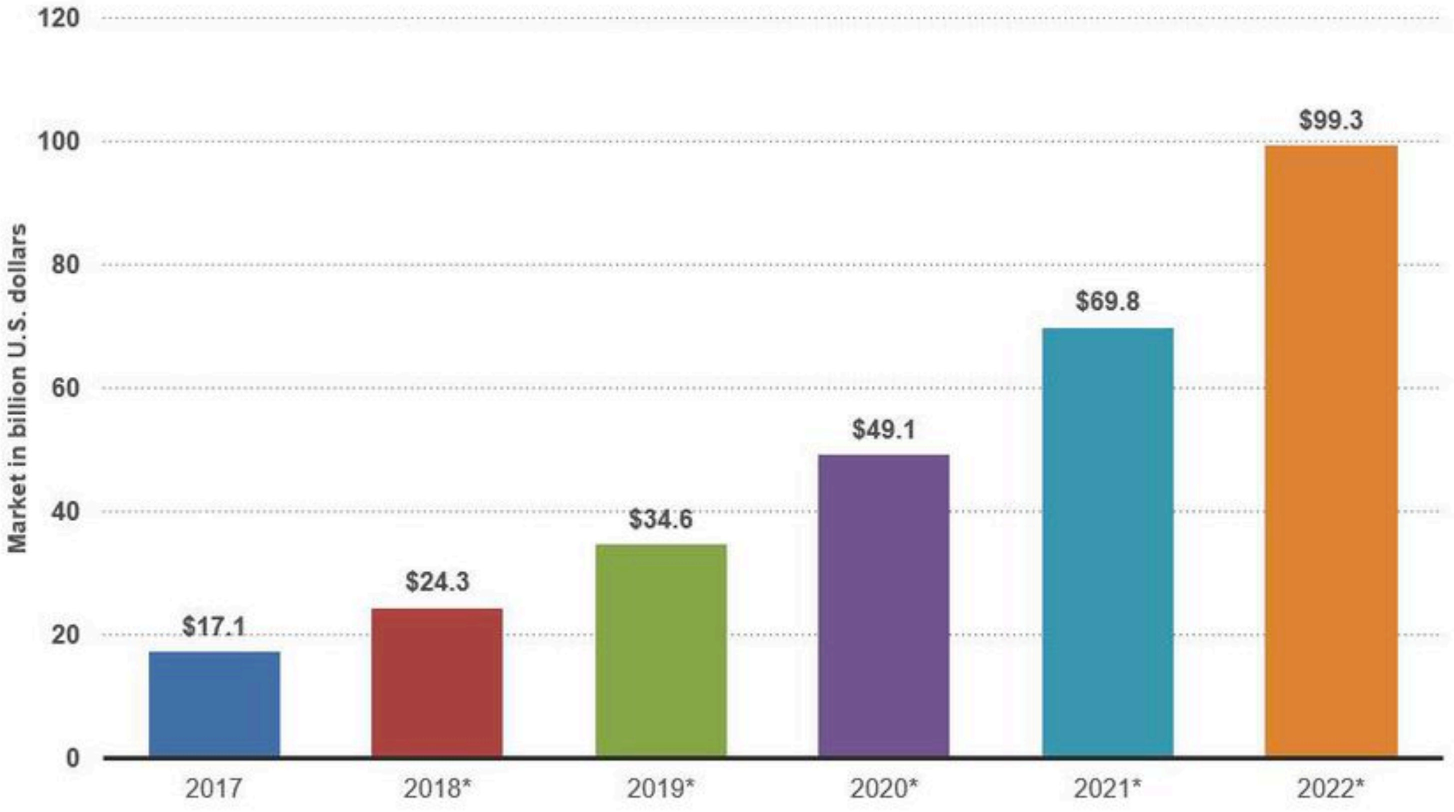
Forecast Revenue Big Data Market Worldwide 2011-2027

Big Data Market Size Revenue Forecast Worldwide From 2011 To 2027 (in billion U.S. dollars)



Big Data and Hadoop Market Size Forecast Worldwide 2017-2022

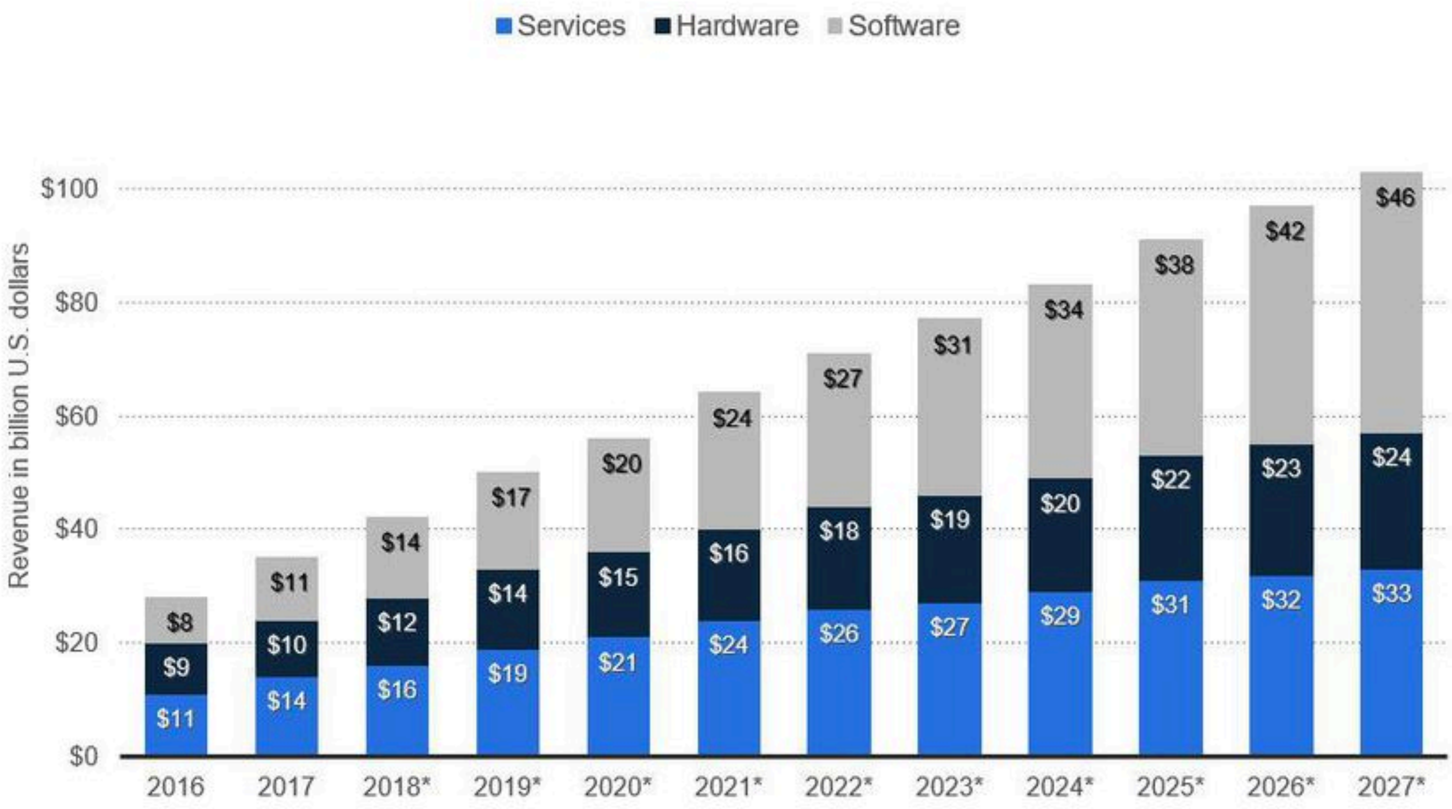
Size of Hadoop and Big Data Market Worldwide From 2017 To 2022 (in billion U.S. dollars)



statista

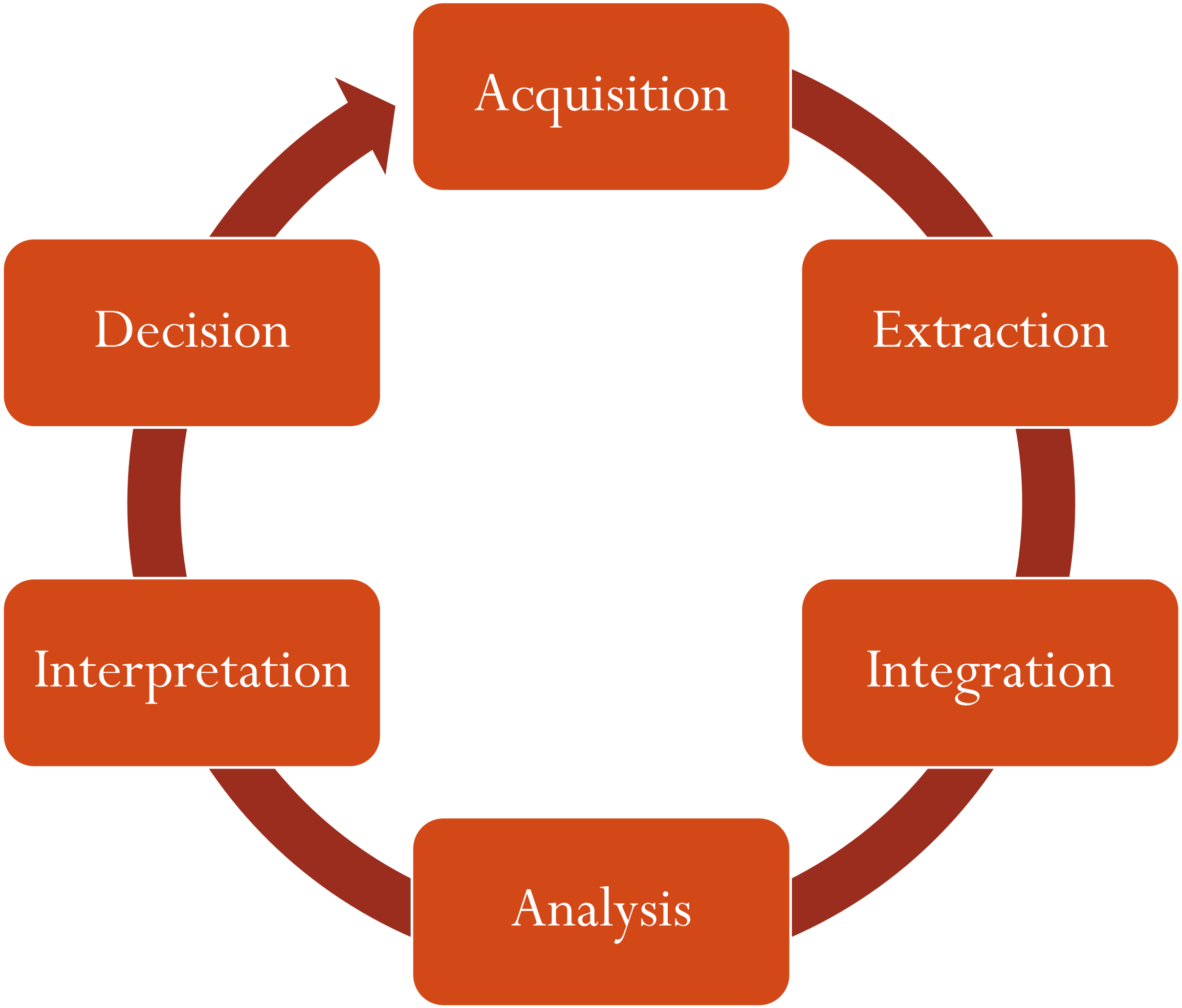
Global Big Data Revenue 2016-2027, by type

Big Data Revenue Worldwide from 2016 to 2027, by major segment (in billion U.S. dollars)



statista

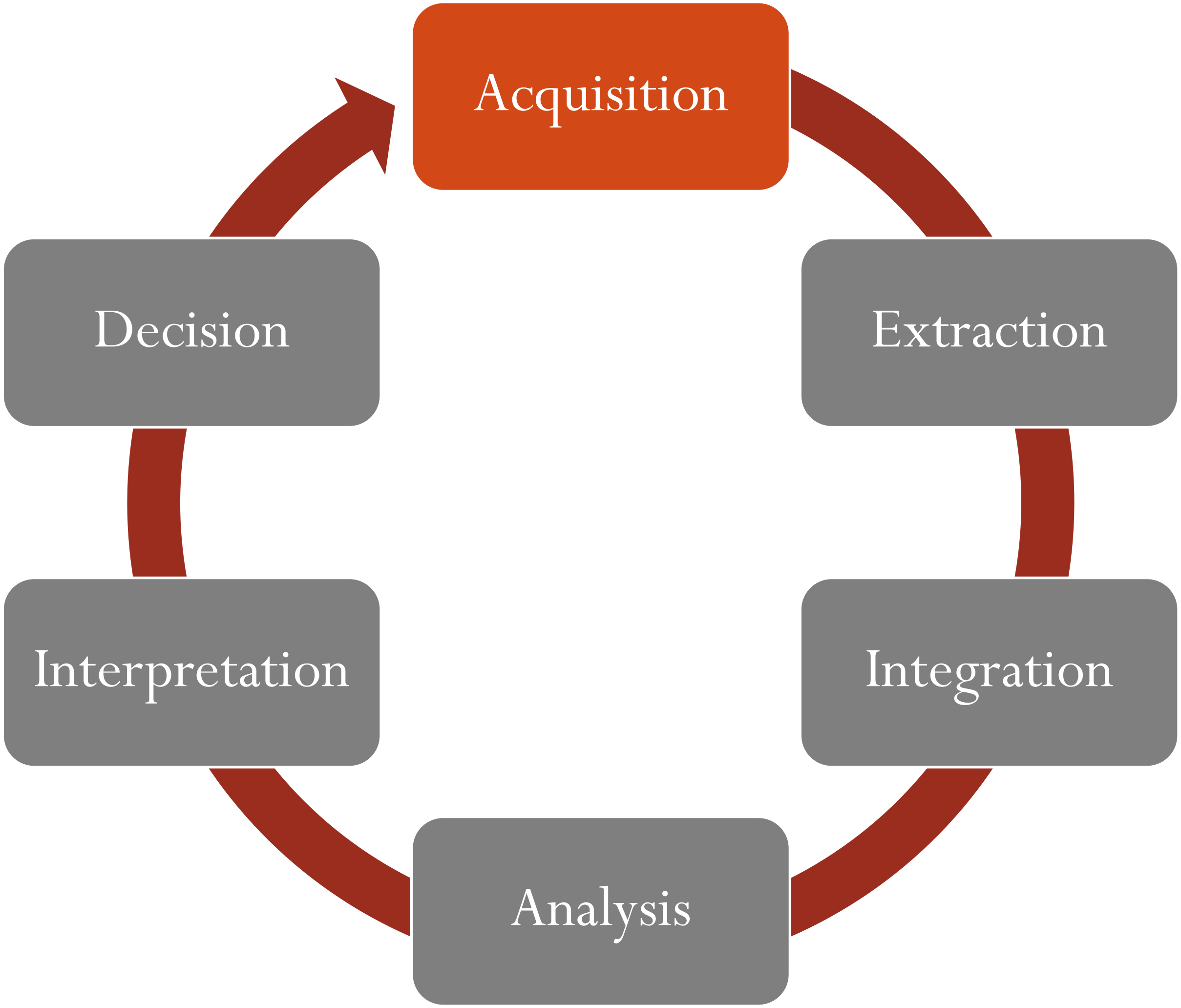
The big data process



Goal:

to make effective strategic decisions exploiting the availability of big data

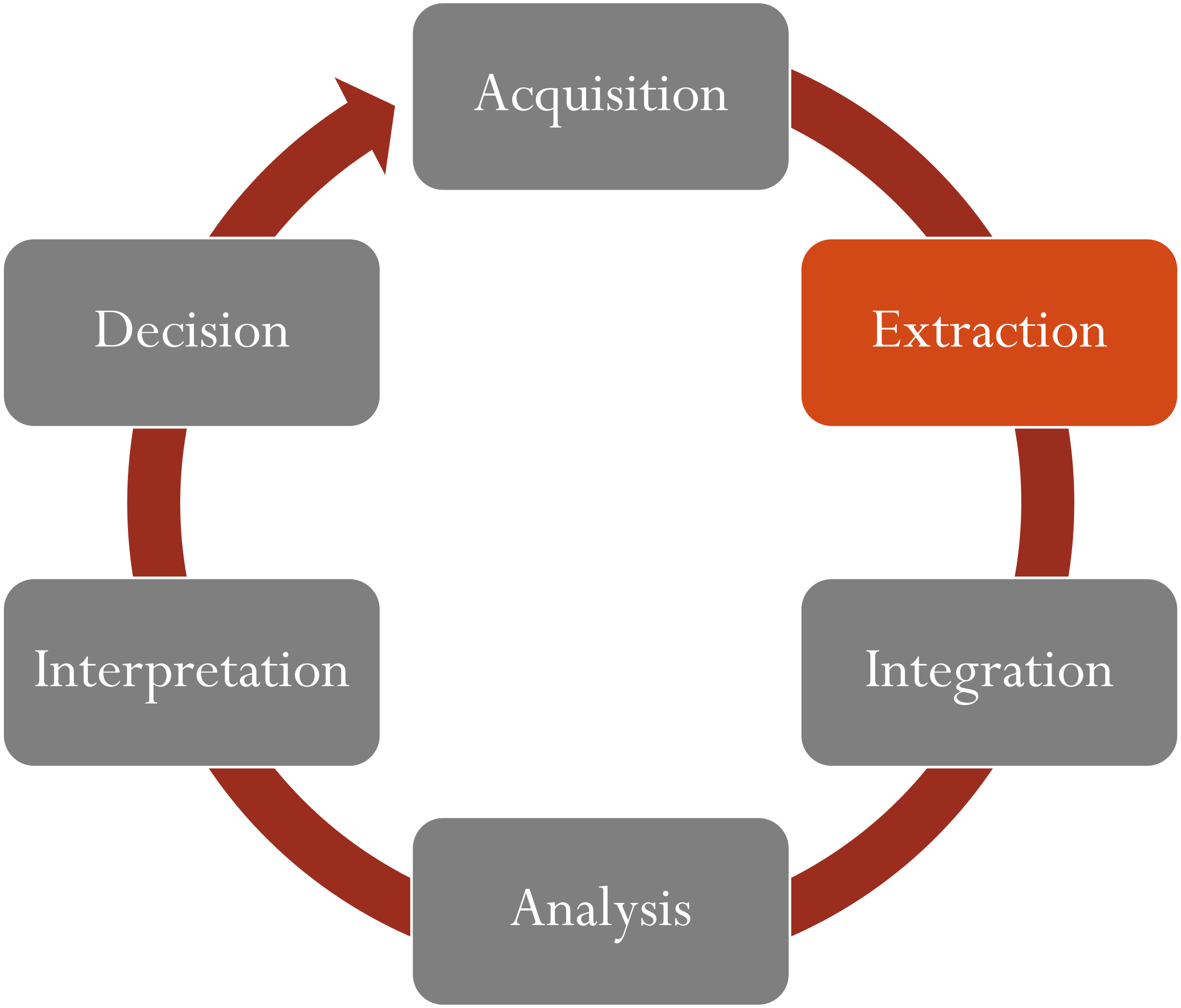
Big Data in action



Requires:

- selection
- filtering
- Metadata generation
- managing provenance

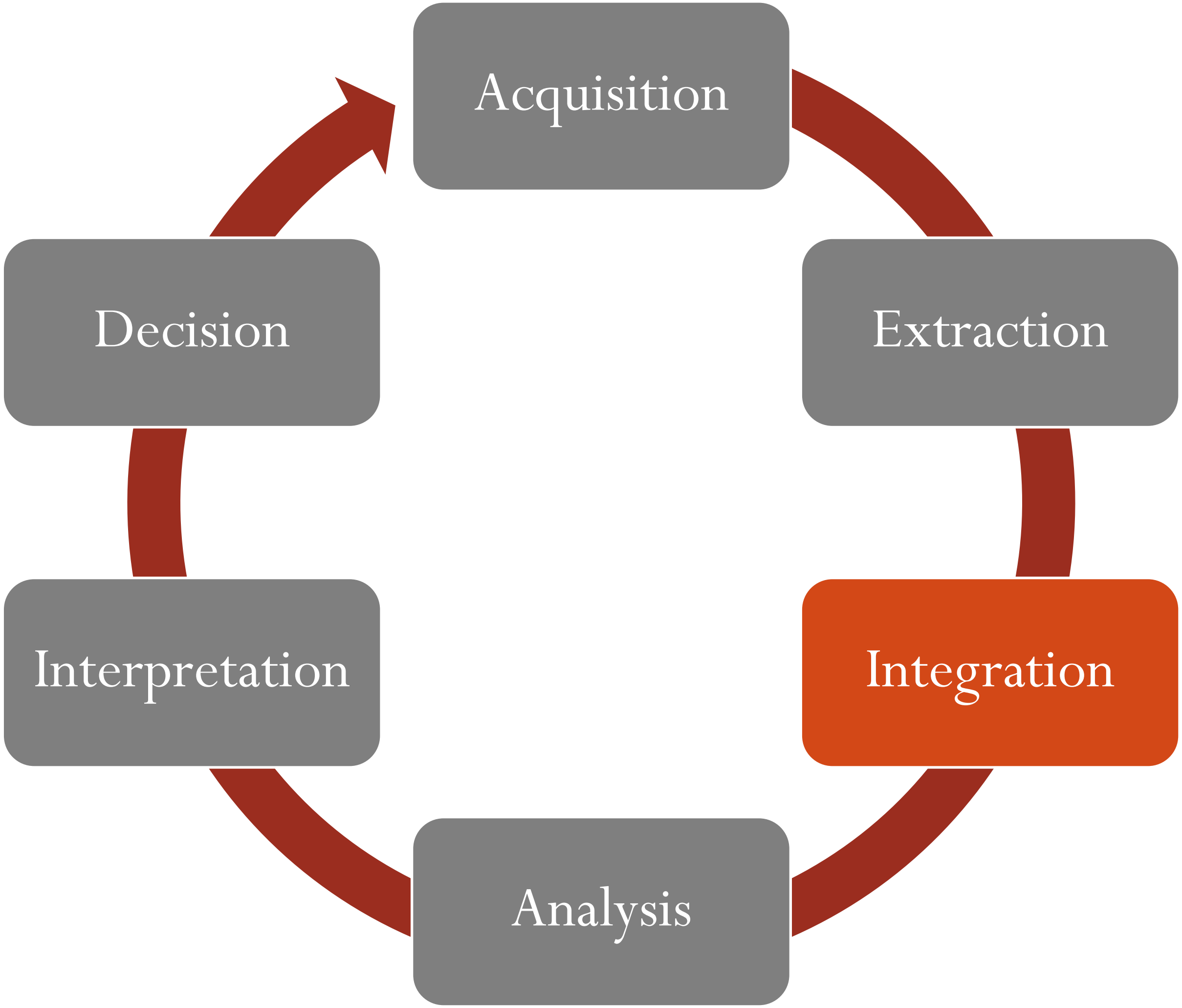
Big Data in action



Requires:

- transformation
- normalization
- cleaning
- aggregation
- error handling

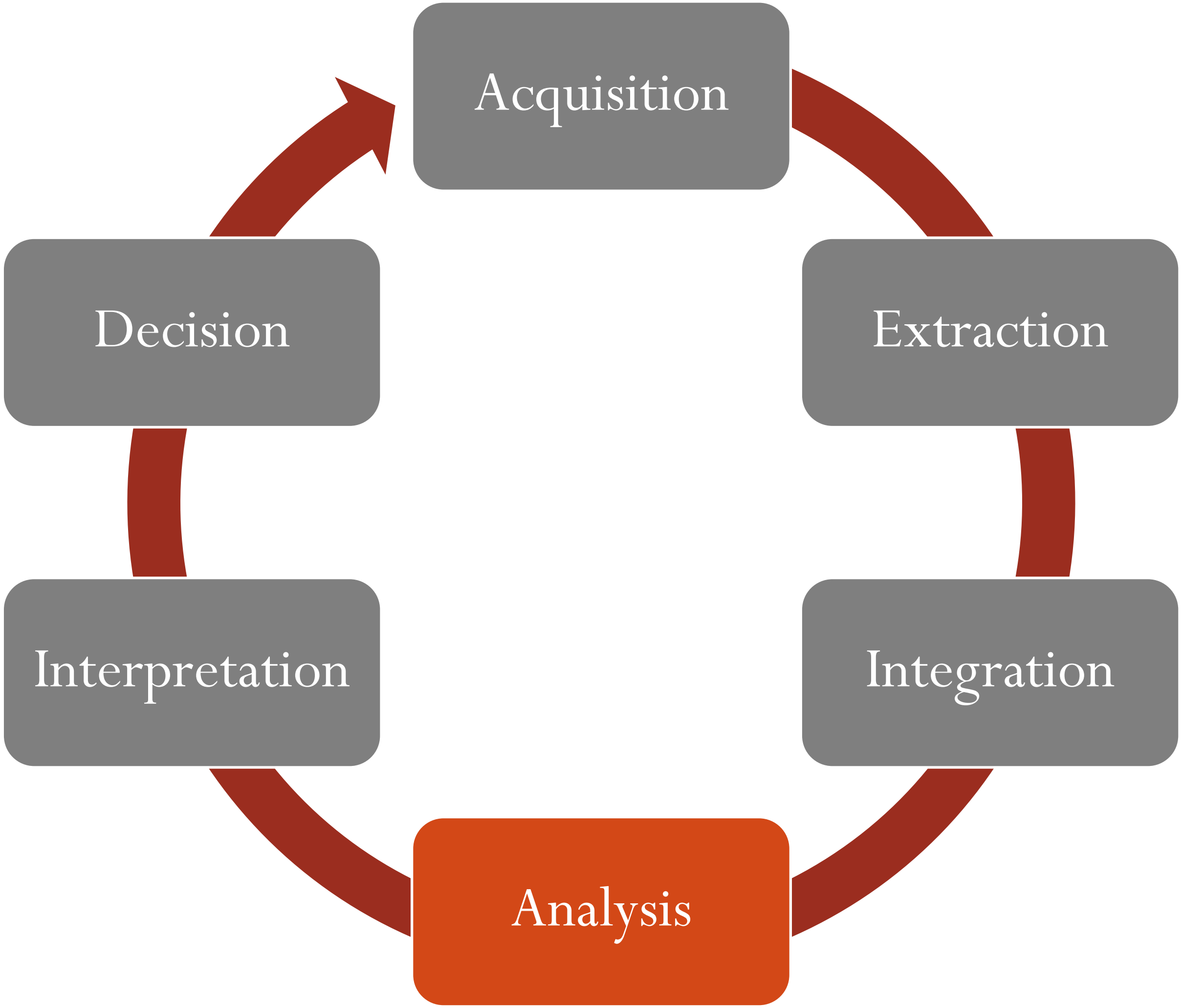
Big Data in action



Requires:

- standardization
- conflict management
- reconciliation
- mapping definition

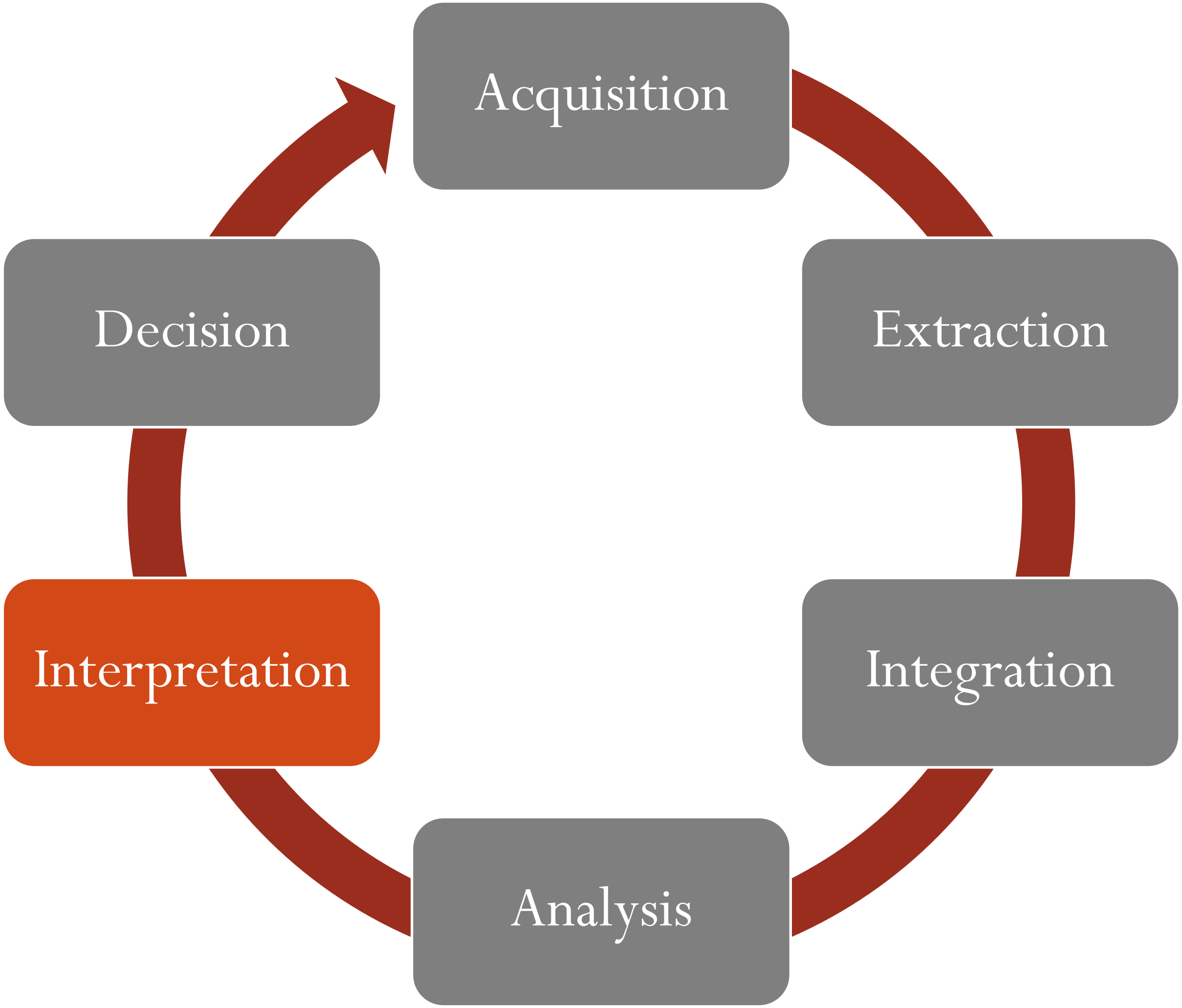
Big Data in action



Requires:

- exploration
- data mining
- machine learning
- visualization

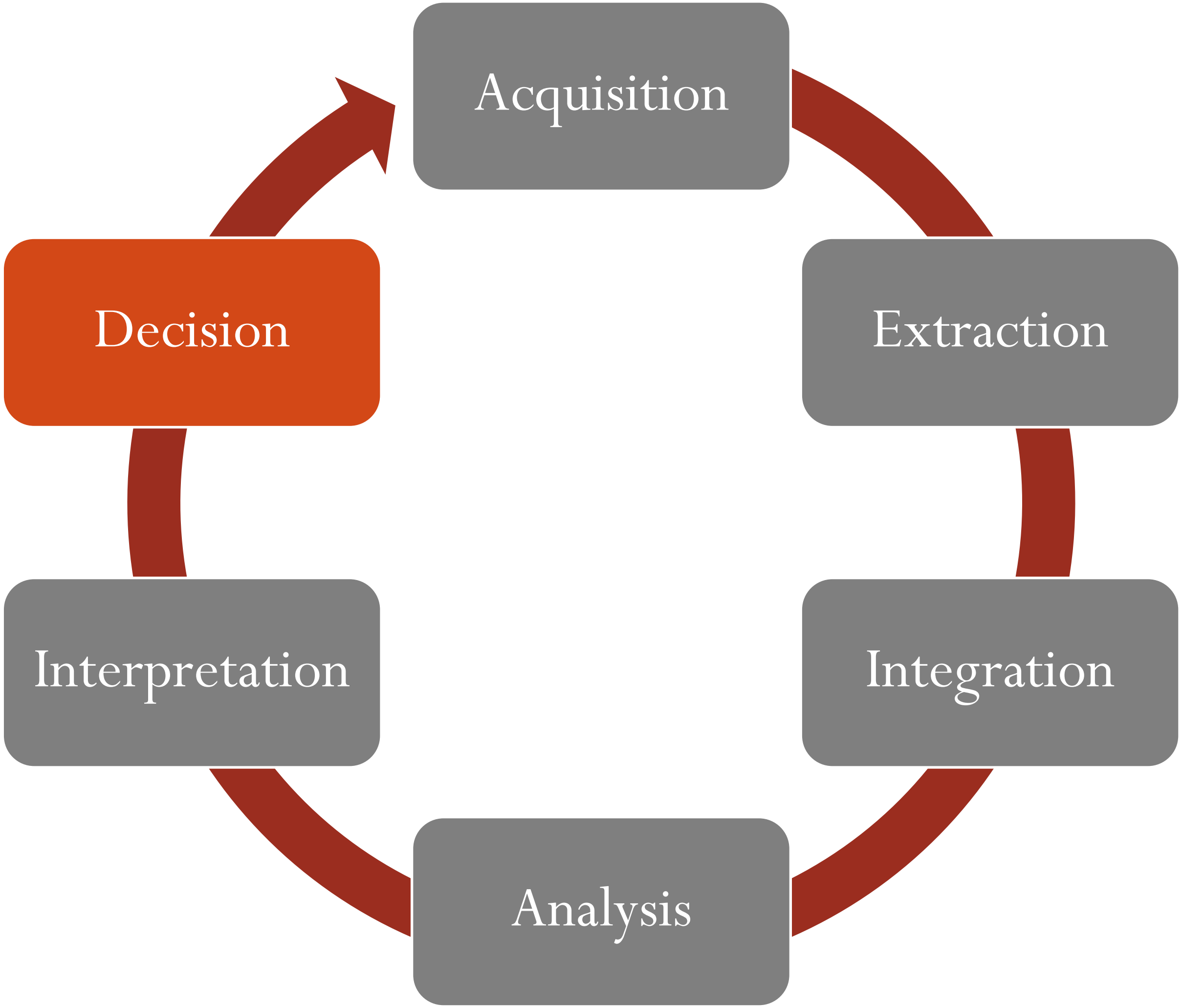
Big Data in action



Requires:

- Knowledge of the domain
- Knowledge of the provenance
- Identification of patterns of interest
- Flexibility of the process

Big Data in action



Requires:

- managerial skills
- continuous improvement of the process

A simple example of a big data process

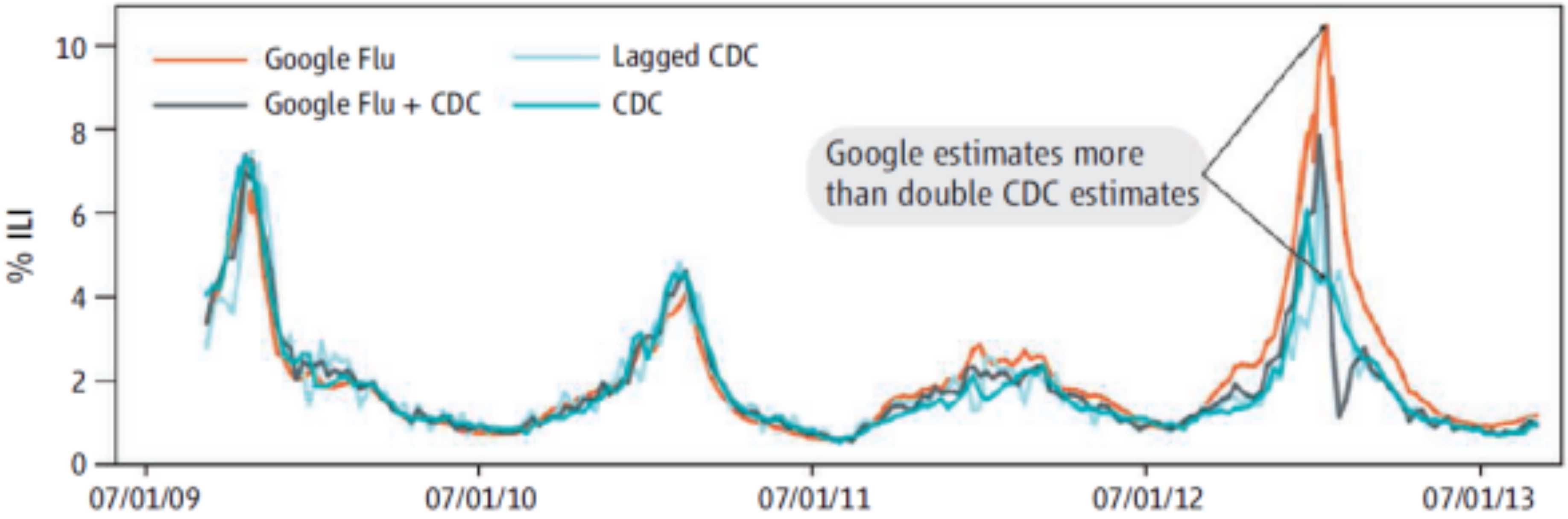
- Problem: The sale of lollipops is going down!
- Acquisition:
  - Sales by customer, region and time
  - Surveys of users
  - Social networks
- Extraction:
  - Data loading from receipts
  - Automatic reading of questionnaires
  - Data extraction from twitter
- Integration:
  - On the basis of user types
- Analysis:
  - lollipops bought by people older than 25
  - lollipops preferred by people younger than 10
- Interpretation:
  - Moms believe: lollipops = bad teeth
  - Boys and girls believe that lollipops are for babies
- Decision:
  - We make lollipops without sugar
  - We ask dentists to advertise our lollipops
  - We make commercials targeted to boys and girls

Risks and Challenges of Big Data

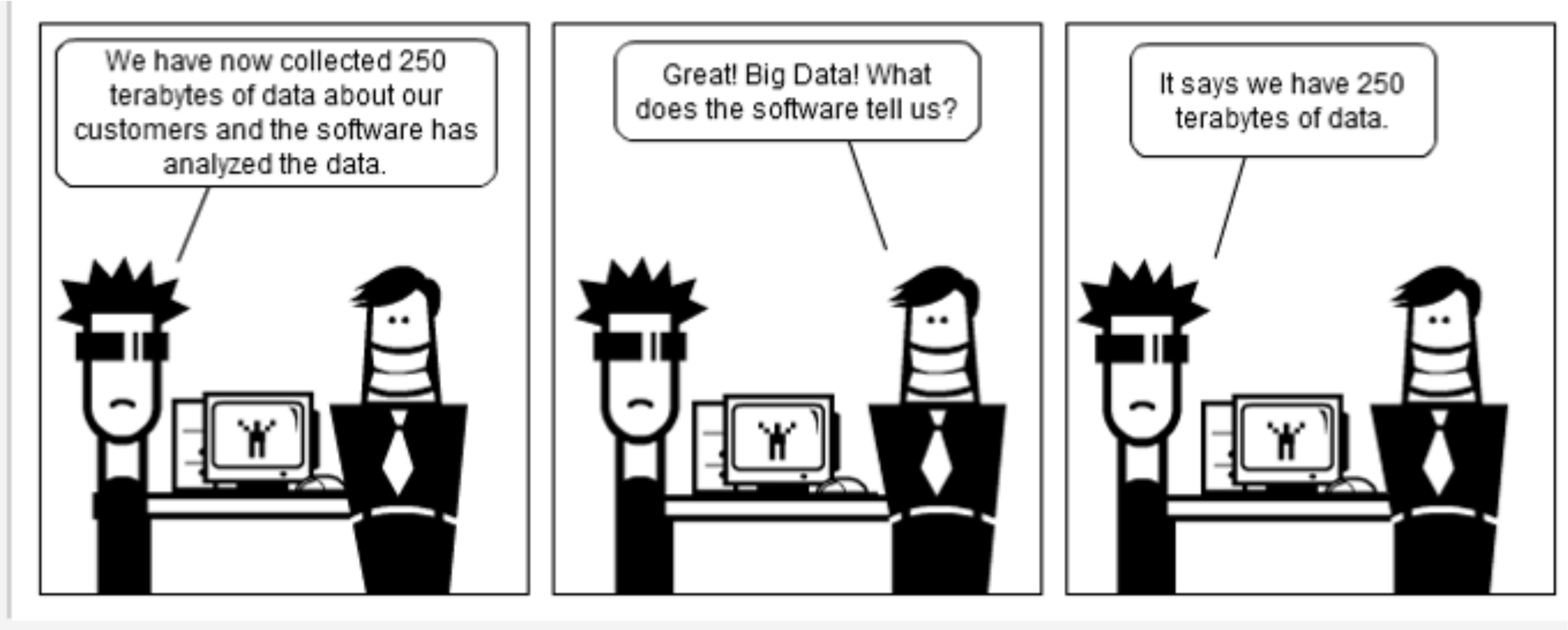
- Performance, performance, performance!
  - Data grows faster than energy on chip
  - Efficiency
  - Scalability
- Effectiveness
- Heterogeneity
- Privacy
- Costs

Effectiveness: a failure story

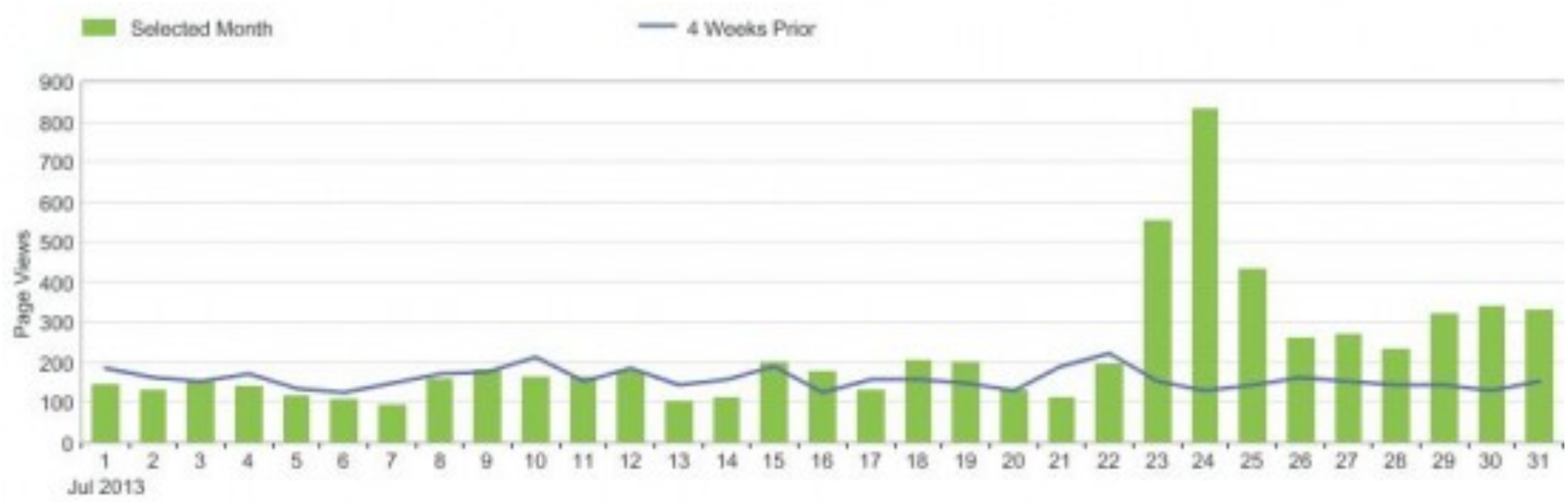
- Google Flu Trends
  - over-estimated the prevalence of flu for 100 out of 108 weeks



Big problem: understanding the output


We have now collected 250 terabytes of data about our customers and the software has analyzed the data.
Great! Big Data! What does the software tell us?
It says we have 250 terabytes of data.

Unexpected results

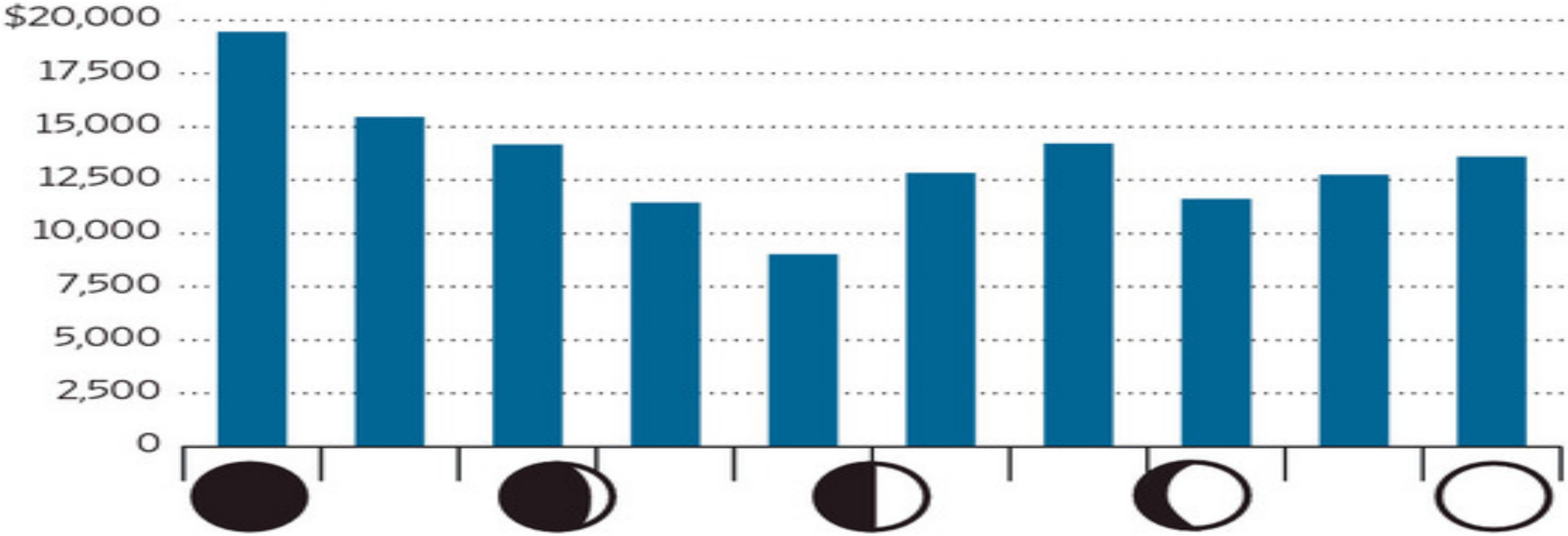


Page Views Report | All Visits (No Segment) | July 2013 | Graph generated by Adobe Analytics at 5:23 PM EDT, 13 Aug 2013

Surprising behaviors

Moon Metrics

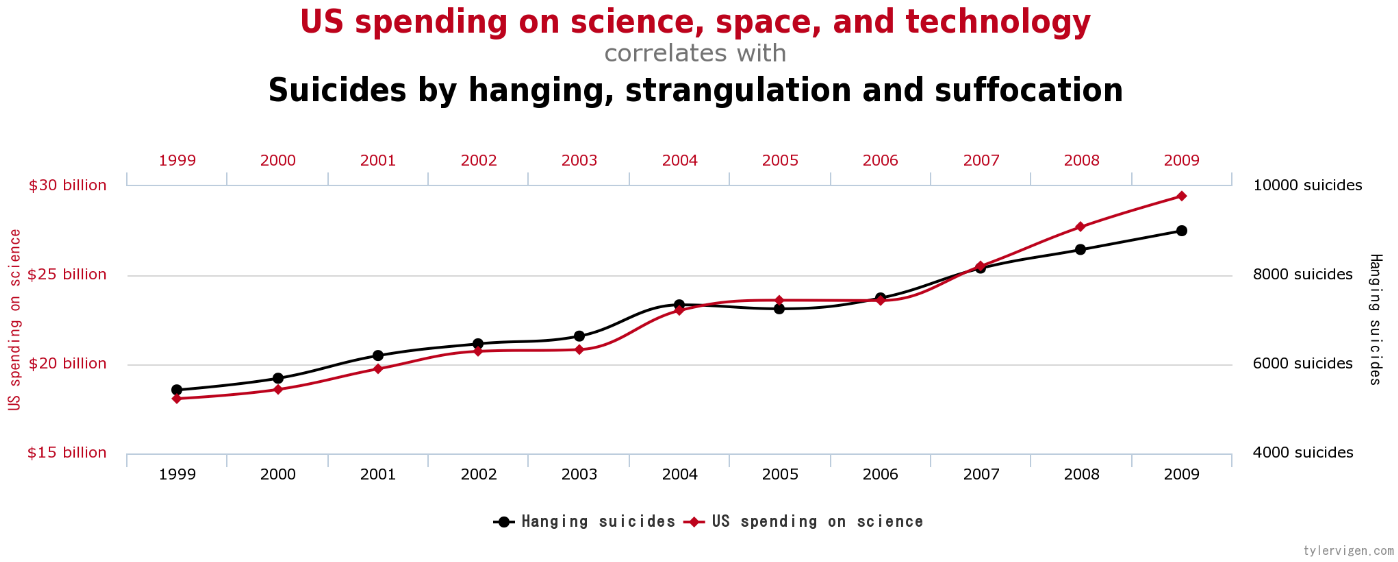
The average value of deals closed by salespeople over nine years in one study peaked during a new moon at more than twice the value during a half moon and 43% higher than the value during a full moon.

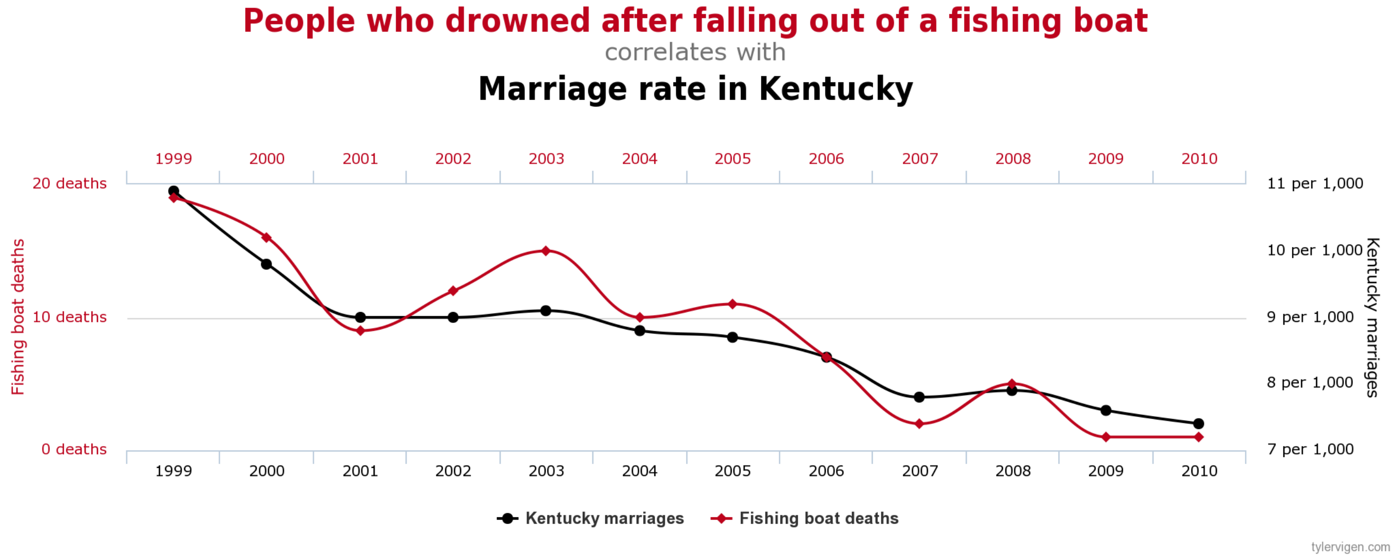


Source: InsideSales.com study of 1,675 deals in various industries, weighted toward business services, technology and financial services.

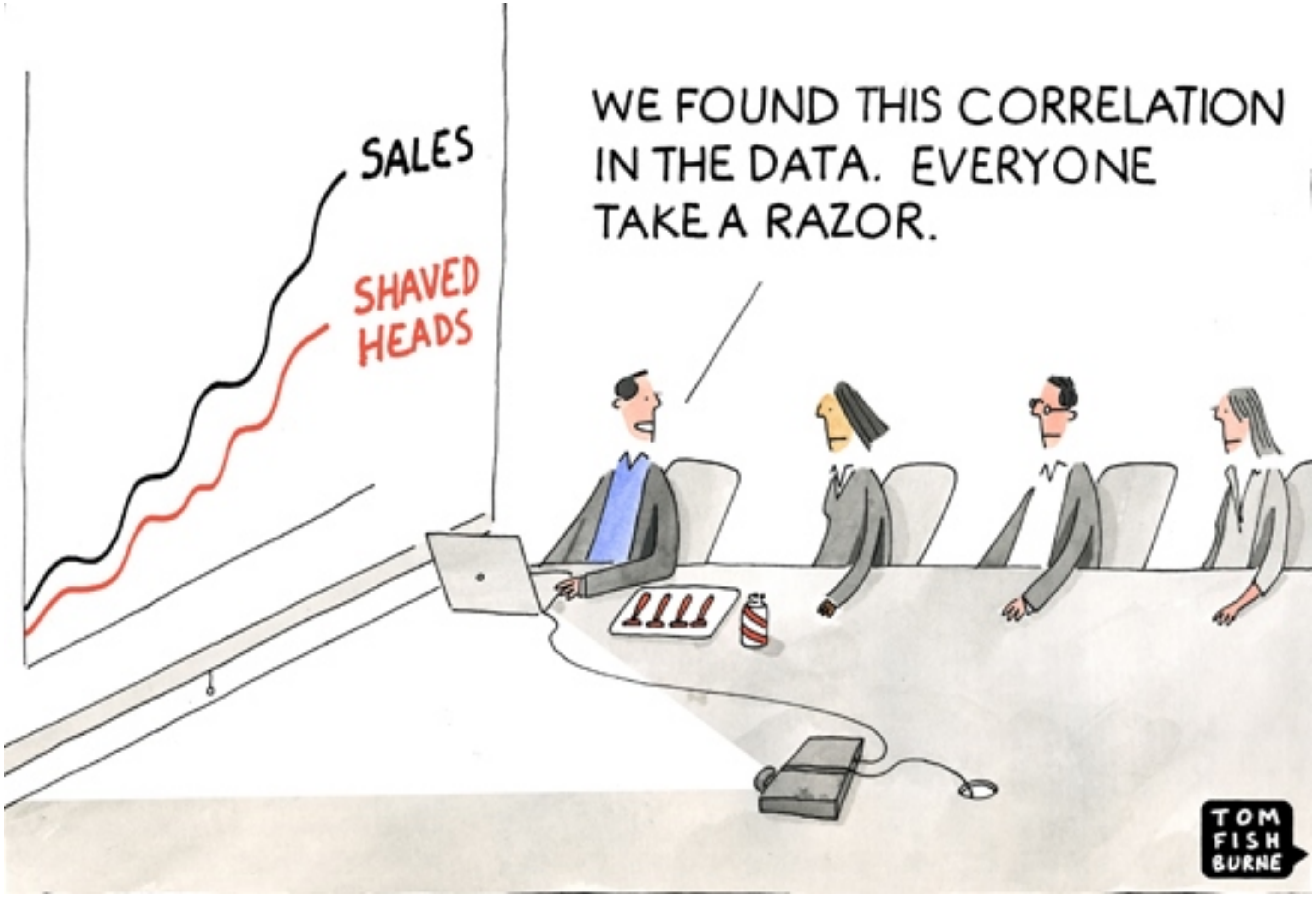
The Wall Street Journal

Strange correlations





Risks of bad interpretation


SALES
SHAVED HEADS
WE FOUND THIS CORRELATION IN THE DATA. EVERYONE TAKE A RAZOR.
TOM FISH BURNE

Video

Privacy: Unpleasant drawbacks

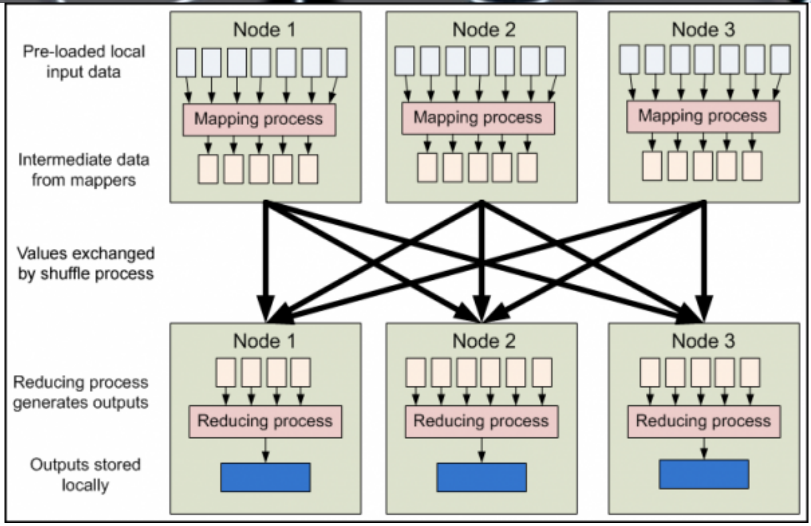
- AOL search data leak (NYT, 8/9/2006)
- Anonymous Netflix vs IMDb database (Wired, 12/13/2007)
- Why Johnny Can't Browse The Internet In Peace (Forbes, 8/1/2012)
- How Companies Learn Your Secrets (NYT, 16/2/2012)
- Facebook–Cambridge Analytica scandal (The Guardian, 2017)

Performance: taming Big Data..

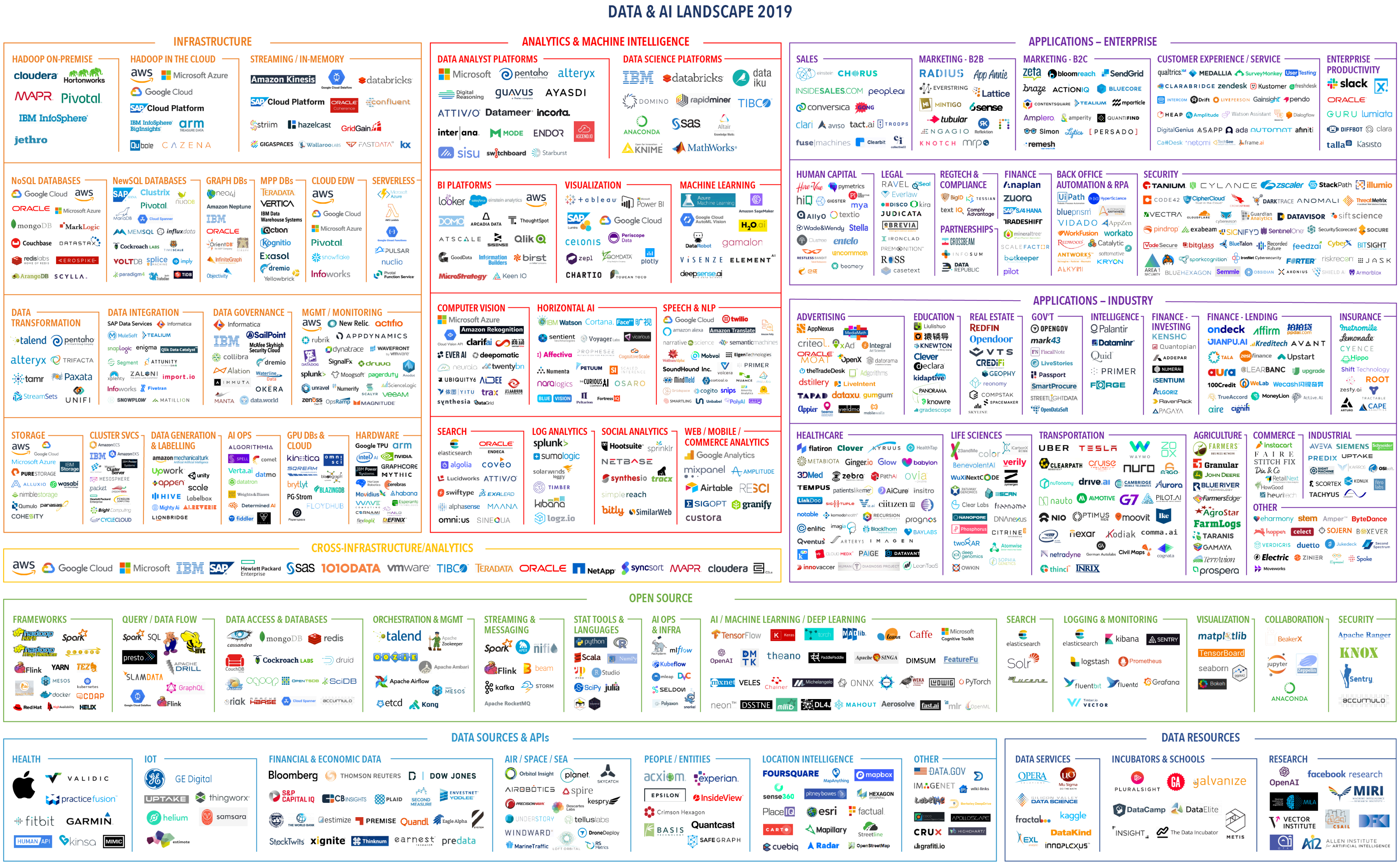
BIG
DATA

Distribution of resources and services

- Distributed Architecture
  - Clusters of computers that work together to a common goal
  - Scale out not up!
- Fault- tolerance
  - Resource replication
  - Eventual consistency
- Distributed processing
  - Shared-nothing model
  - New programming paradigms



The Big Data Landscape



DATA & AI LANDSCAPE 2019

INFRASTRUCTURE

HADOOP ON-PREMISE · HADOOP IN THE CLOUD · STREAMING / IN-MEMORY

NoSQL DATABASES · NewSQL DATABASES · GRAPH DBs · MPP DBs · CLOUD EDW · SERVERLESS

DATA TRANSFORMATION · DATA INTEGRATION · DATA GOVERNANCE · MGMT / MONITORING

STORAGE · CLUSTER SVCS · DATA GENERATION & LABELLING · AI OPS · GPU DBs & CLOUD · HARDWARE

ANALYTICS & MACHINE INTELLIGENCE

DATA ANALYST PLATFORMS · DATA SCIENCE PLATFORMS

BI PLATFORMS · VISUALIZATION · MACHINE LEARNING

COMPUTER VISION · HORIZONTAL AI · SPEECH & NLP

SEARCH · LOG ANALYTICS · SOCIAL ANALYTICS · WEB / MOBILE / COMMERCE ANALYTICS

APPLICATIONS – ENTERPRISE

SALES · MARKETING - B2B · MARKETING - B2C · CUSTOMER EXPERIENCE / SERVICE · ENTERPRISE PRODUCTIVITY

HUMAN CAPITAL · LEGAL · REGTECH & COMPLIANCE · PARTNERSHIPS · FINANCE · BACK OFFICE AUTOMATION & RPA · SECURITY

APPLICATIONS – INDUSTRY

ADVERTISING · EDUCATION · REAL ESTATE · GOV'T · INTELLIGENCE · FINANCE - INVESTING · FINANCE - LENDING · INSURANCE

HEALTHCARE · LIFE SCIENCES · TRANSPORTATION · AGRICULTURE · COMMERCE · INDUSTRIAL · OTHER

CROSS-INFRASTRUCTURE/ANALYTICS

OPEN SOURCE

FRAMEWORKS · QUERY / DATA FLOW · DATA ACCESS & DATABASES · ORCHESTRATION & MGMT · STREAMING & MESSAGING · STAT TOOLS & LANGUAGES · AI OPS & INFRA · AI / MACHINE LEARNING / DEEP LEARNING · SEARCH · LOGGING & MONITORING · VISUALIZATION · COLLABORATION · SECURITY

DATA SOURCES & APIs

HEALTH · IOT · FINANCIAL & ECONOMIC DATA · AIR / SPACE / SEA · PEOPLE / ENTITIES · LOCATION INTELLIGENCE · OTHER

DATA RESOURCES

DATA SERVICES · INCUBATORS & SCHOOLS · RESEARCH

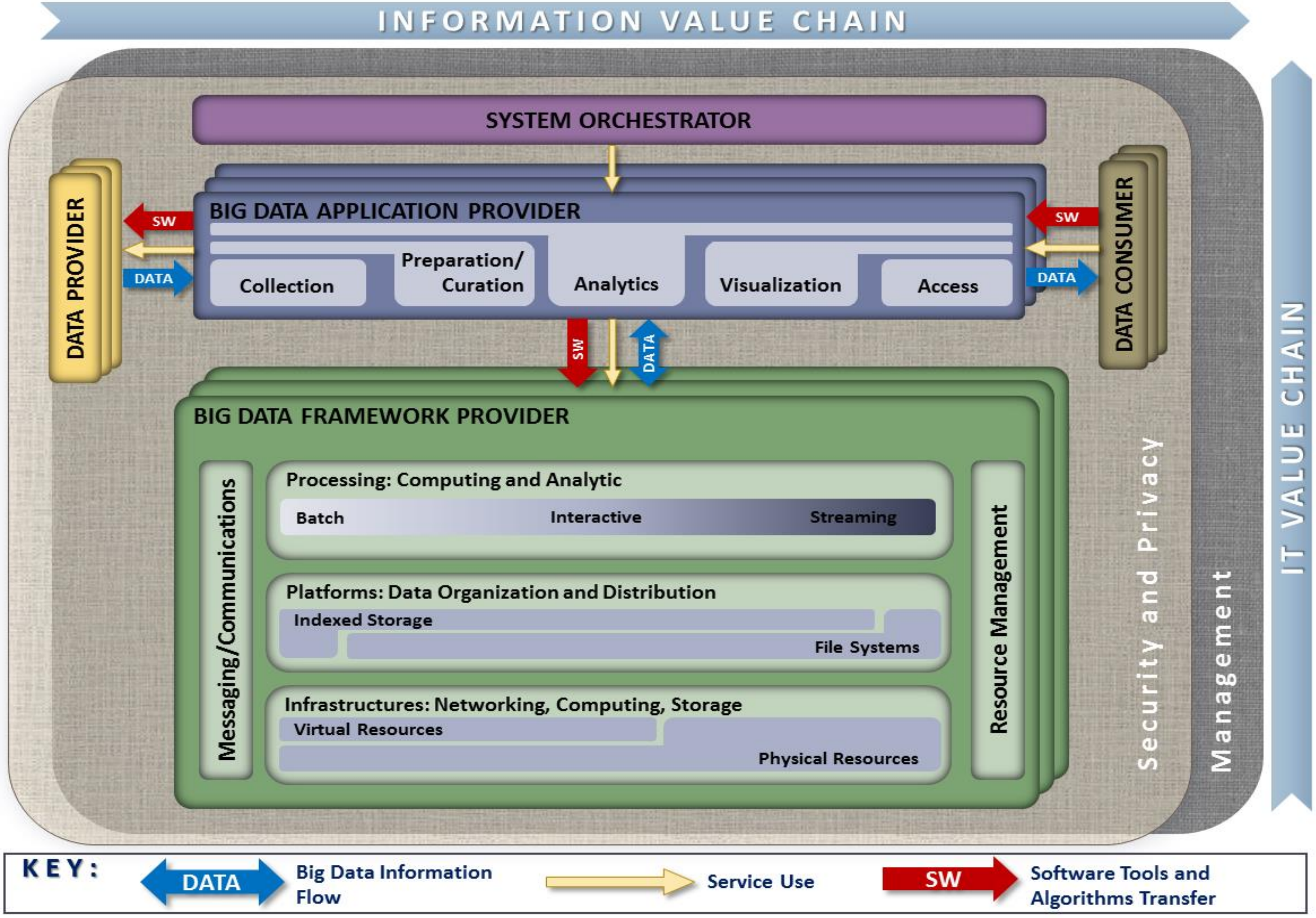
July 16, 2019 - FINAL 2019 VERSION

© Matt Turck (@mattturck), Lisa Xu (@lisaxu92), & FirstMark (@firstmarkcap)

mattturck.com/data2019

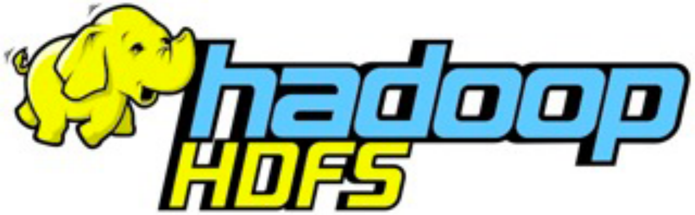
FIRSTMARK
EARLY STAGE VENTURE CAPITAL

NIST Big Data Reference Architecture



The New Software Stack

- New programming environments designed to get their parallelism not from a supercomputer but from computing clusters
- Bottom of the stack: distributed file system (DFS)
  - We have a winner!

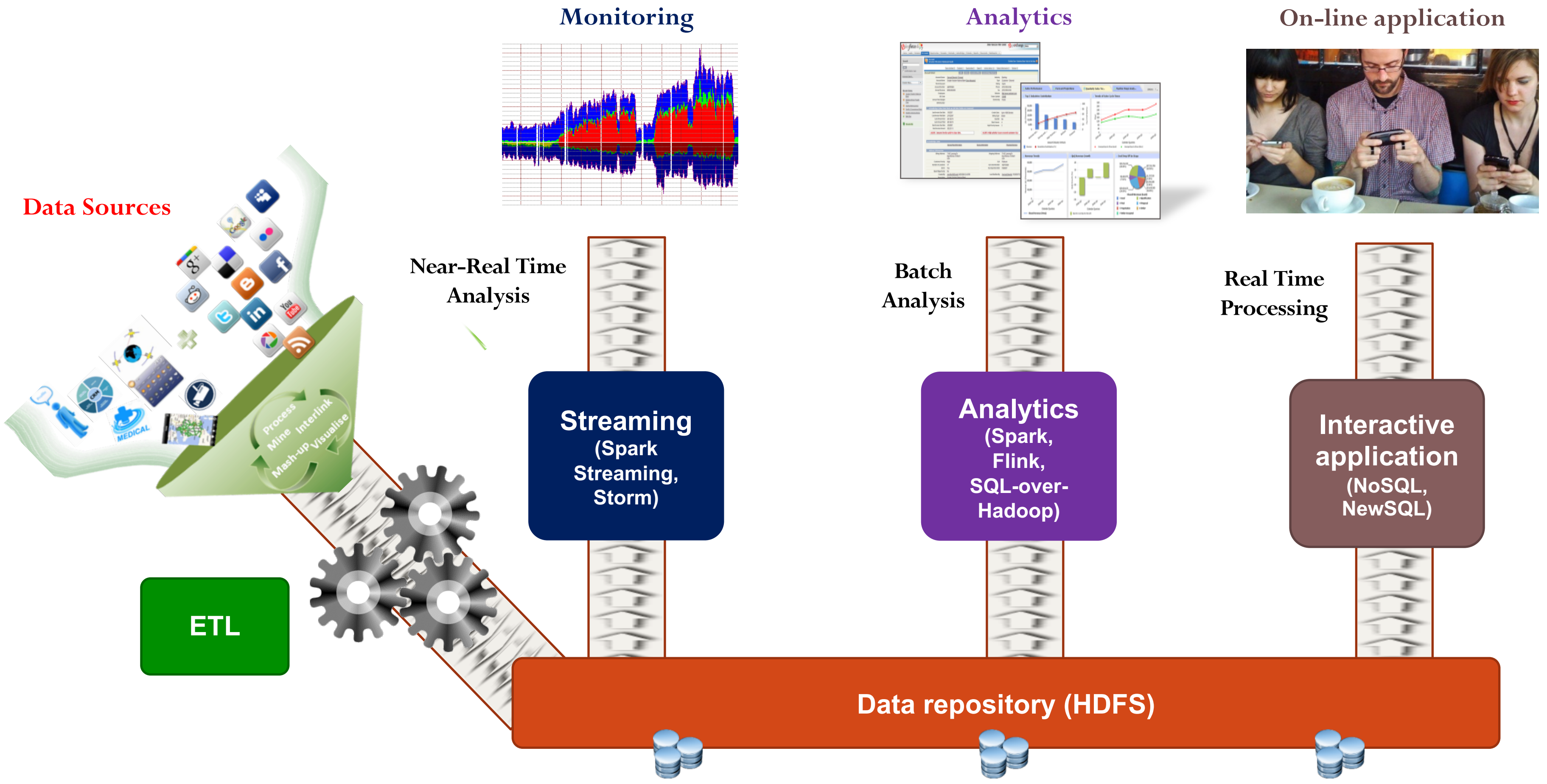


On the top of HDFS

- Hundreds of different (high-level) programming solutions

- Three main scenarios:
  - Analytics (batch)
    - collecting, transforming, and modeling data with the goal of discovering useful information and supporting decision-making
    - Append-only I/O, not necessarily persistent data, no ACID transactions
  - Streaming (near-real-time)
    - processing stream data (sequence of data elements made available over time) with the goal of monitoring and analyzing data on the fly via time windows
    - Stream I/O, possibly persistent data for analytics, no ACID transactions
  - Interactive (real-time)
    - processing data and returning the results quickly to affect the environment at that time (e-commerce, search engines, booking, …)
    - Read/write I/O, persistent data, (soft)ACID transactions

The Big Data flow



Monitoring

Analytics

On-line application

Data Sources

Near-Real Time Analysis

Batch Analysis

Real Time Processing

Streaming (Spark Streaming, Storm)

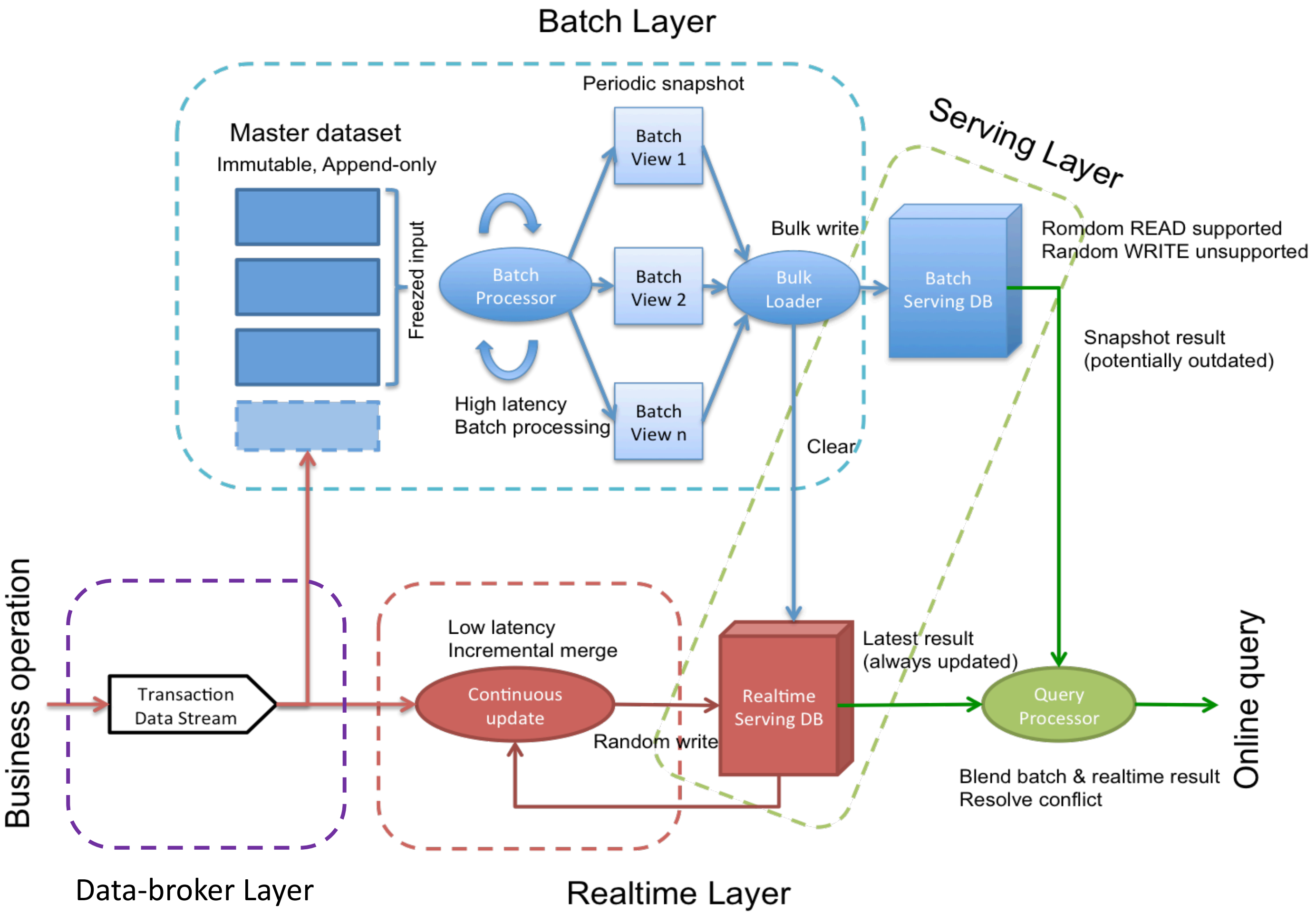
Analytics (Spark, Flink, SQL-over-Hadoop)

Interactive application (NoSQL, NewSQL)

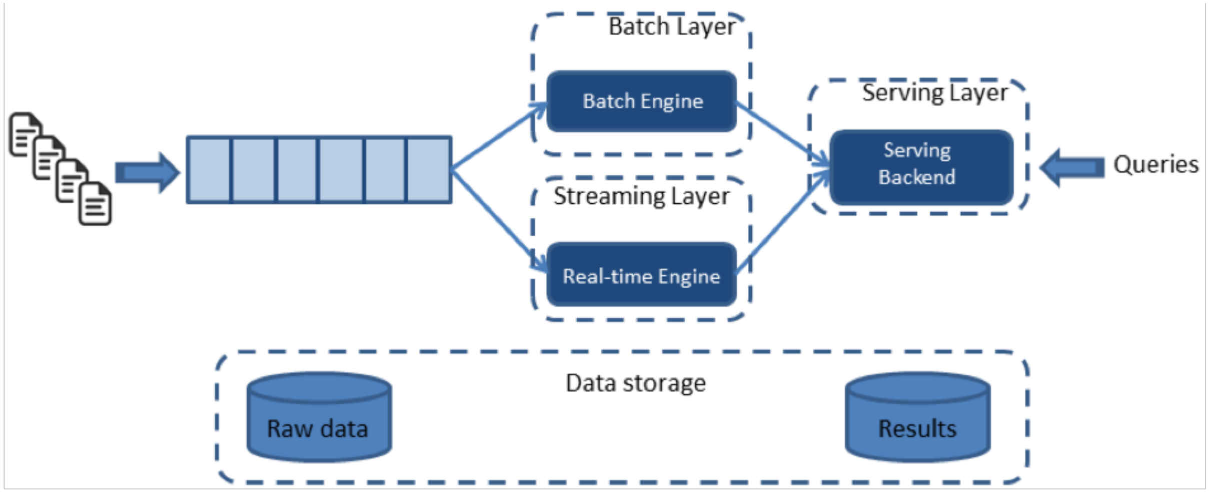
ETL

Data repository (HDFS)

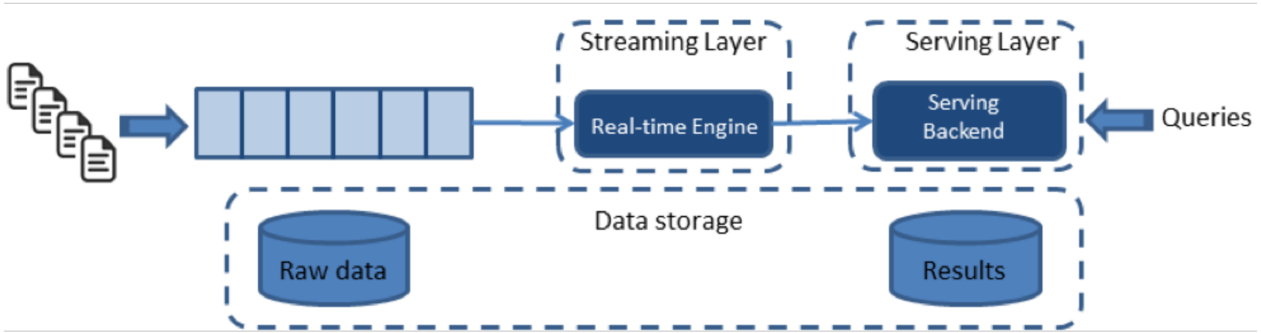
The lambda architecture for analytics



Lambda vs kappa architecture

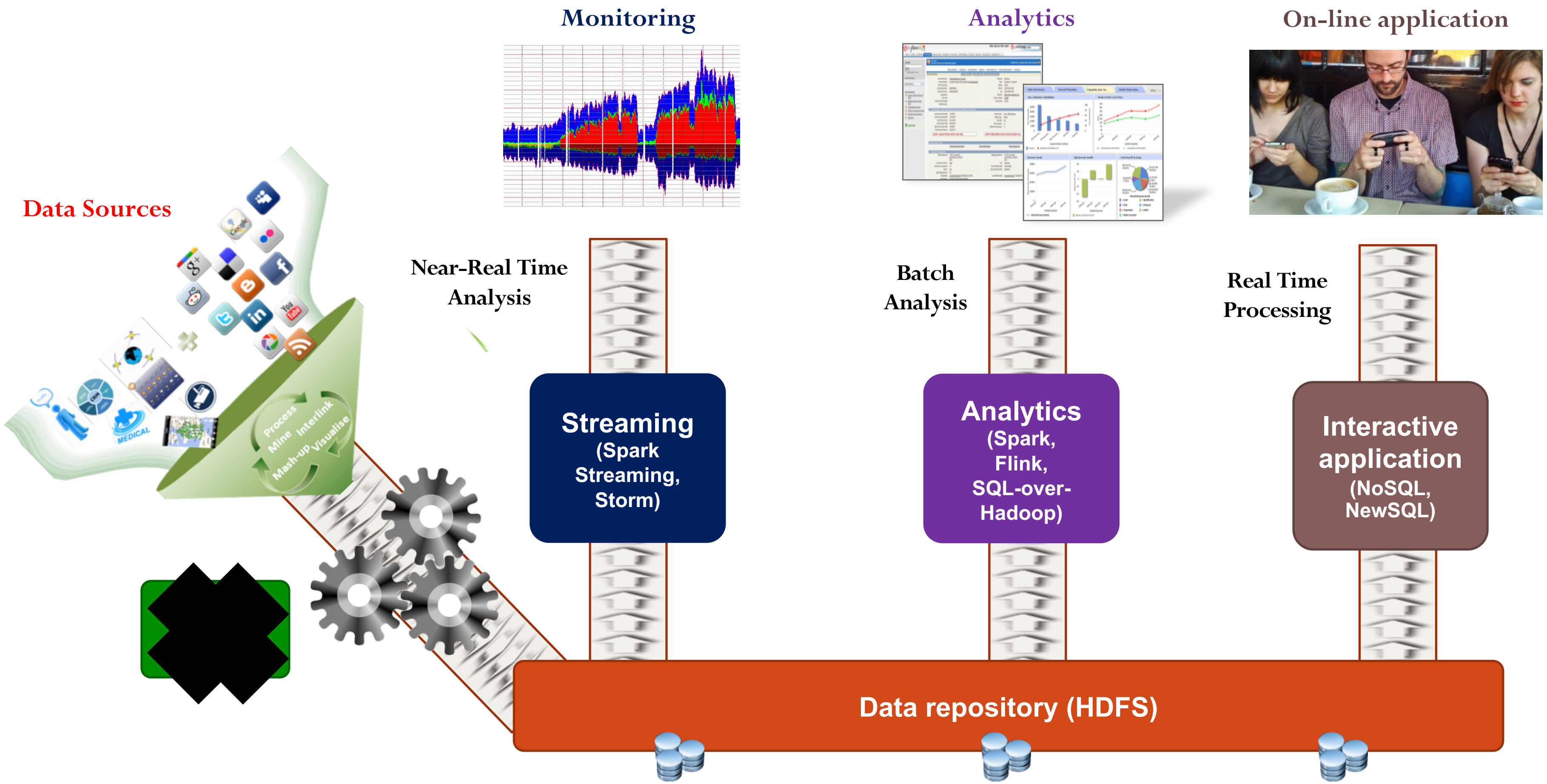


Architettura lambda



Architettura kappa

A recent trend for Data Collection



Monitoring

Analytics

On-line application

Data Sources

Near-Real Time Analysis

Batch Analysis

Real Time Processing

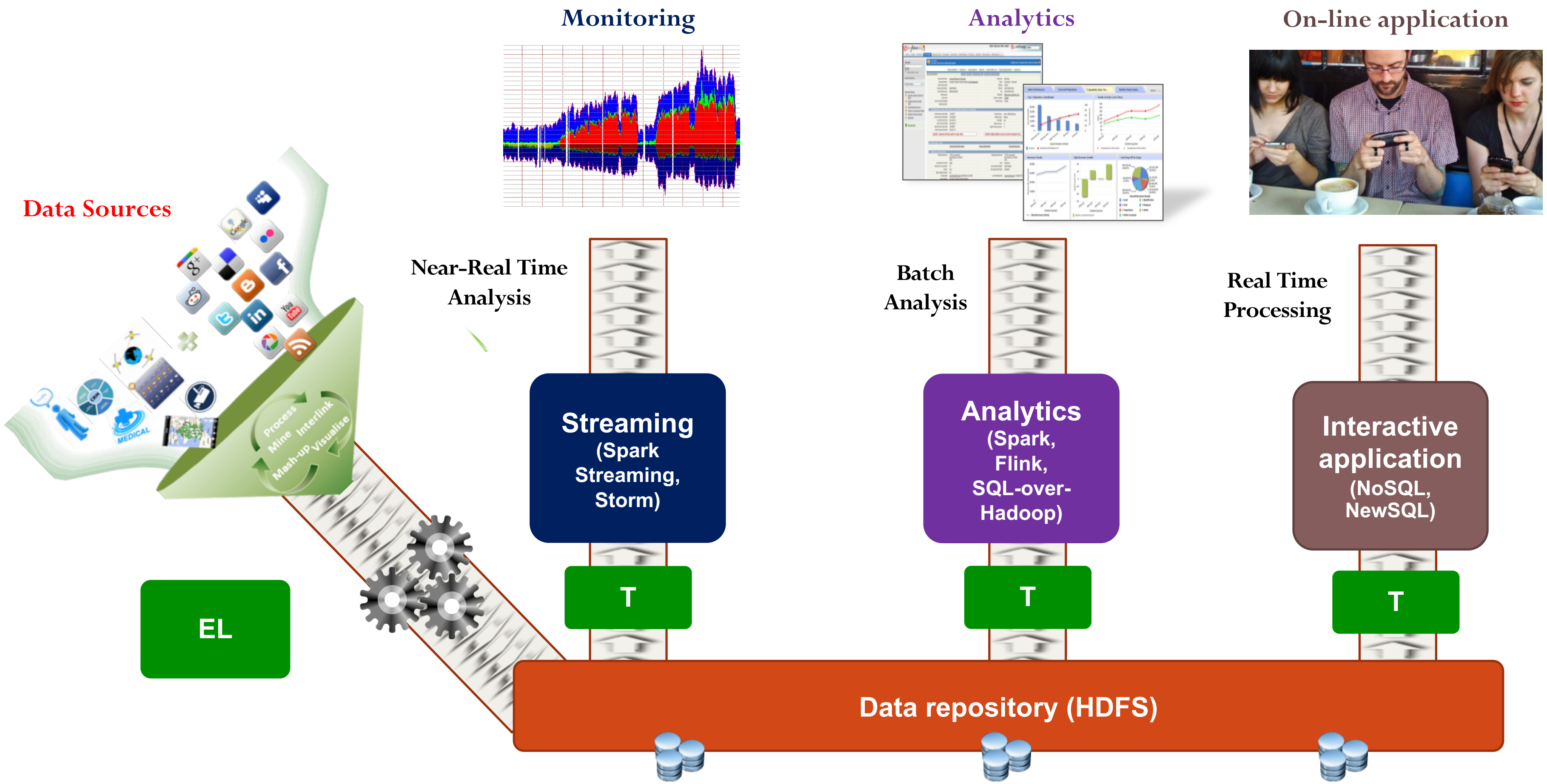
Streaming
(Spark Streaming, Storm)

Analytics
(Spark, Flink, SQL-over-Hadoop)

Interactive application
(NoSQL, NewSQL)

Data repository (HDFS)

Or better...



Monitoring

Analytics

On-line application

Data Sources

Near-Real Time Analysis

Batch Analysis

Real Time Processing

Streaming (Spark Streaming, Storm)

Analytics (Spark, Flink, SQL-over-Hadoop)

Interactive application (NoSQL, NewSQL)

EL

T

T

T

Data repository (HDFS)

Data Lake

"A data lake is a single store of all enterprise data including raw copies of source system data used for tasks such as reporting, visualization, and analytics. A data lake can include structured data (relational tables), semi-structured data (CSV, logs, XML, JSON), unstructured data (emails, documents, PDFs) and binary data (images, audio, video)" - Wikipedia, 2020

"A data lake is a centralized repository that allows you to store all your structured and unstructured data at any scale. You can store your data as-is, without having to first structure the data, and run different types of analytics (from dashboards and visualizations to big data processing, real-time analytics, and machine learning) to guide better decisions." - AWS, 2020

Data lake

1 The incoming flow represents multiple raw data archives ranging from emails, spreadsheets, social media content, etc.

STRUCTURED DATA

1. Information in rows and columns
2. Easily ordered and processed with data mining tools

UNSTRUCTURED DATA

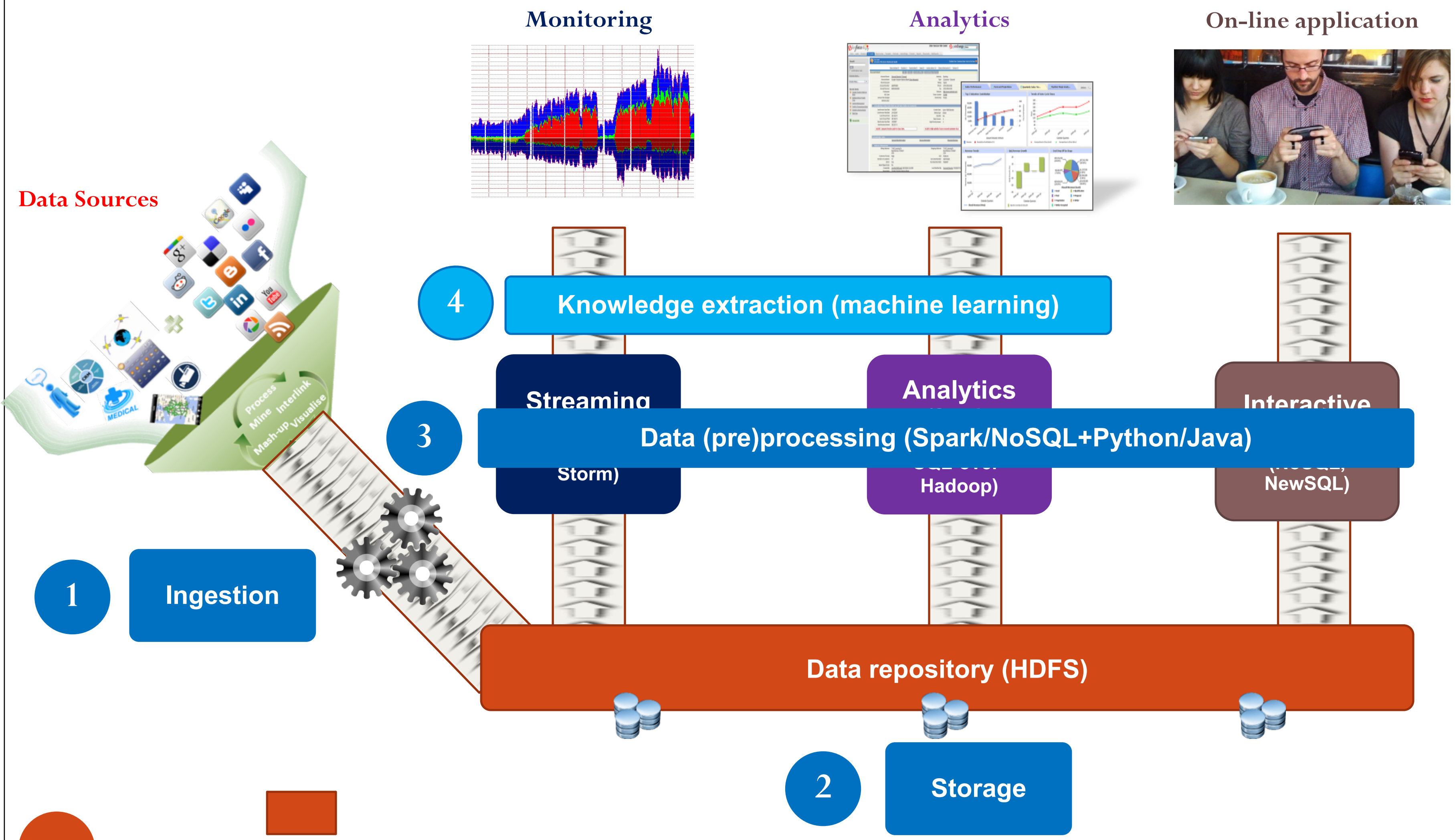
1. Raw, unorganized data
2. Emails
3. PDF files
4. Images, video and audio
5. Social media tools

2 The reservoir of water is a dataset, where you run analytics on all the data.

3 The outflow of water is the analyzed data.

4 Through this process, you are able to “sift” through all the data quickly to gain key business insights.

The software stack for data analytics



Monitoring

Analytics

On-line application

Data Sources

4 Knowledge extraction (machine learning)

Streaming

Storm)

Analytics

Hadoop)

Interactive

NewSQL)

3 Data (pre)processing (Spark/NoSQL+Python/Java)

1 Ingestion

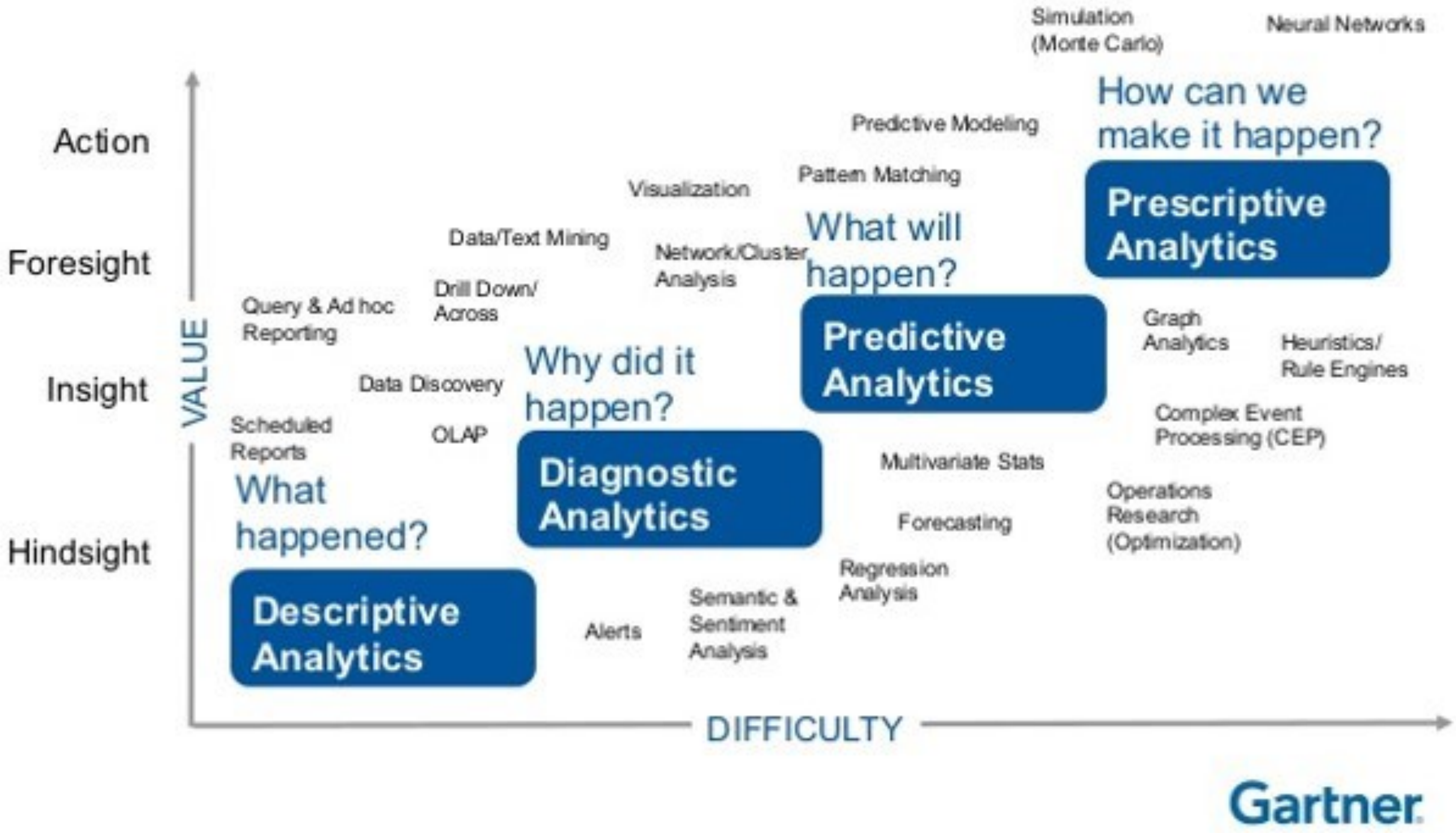
Data repository (HDFS)

2 Storage

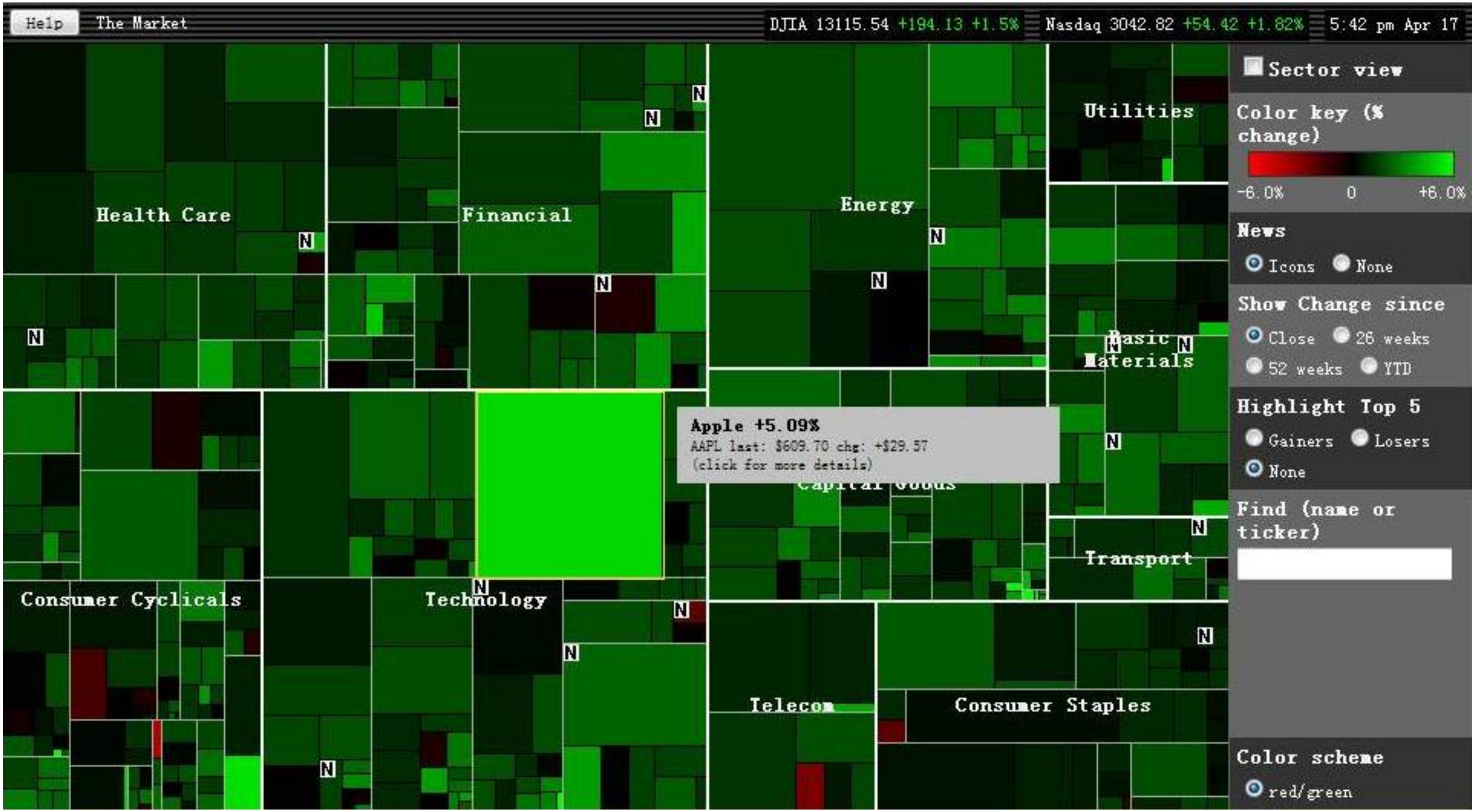
Techniques for big data analysis

- Extract, transform, and load (ETL)
- Data fusion and data integration
- **Data Management** (focus of the course)
- Analytics
  - Data mining
    - Association rule learning
    - Classification
    - Cluster analysis
    - Regression
  - Machine learning
    - Supervised learning
    - Unsupervised learning
- Crowdsourcing
- …

Goals of analytics



Visualization is fundamental



Help The Market

DJIA 13115.54 +194.13 +1.5% Nasdaq 3042.82 +54.42 +1.82% 5:42 pm Apr 17

Apple +5.09%
AAPL last: $609.70 chg: +$29.57
(click for more details)

□ Sector view

Color key (% change)

-6.0% 0 +6.0%

News
● Icons ○ None

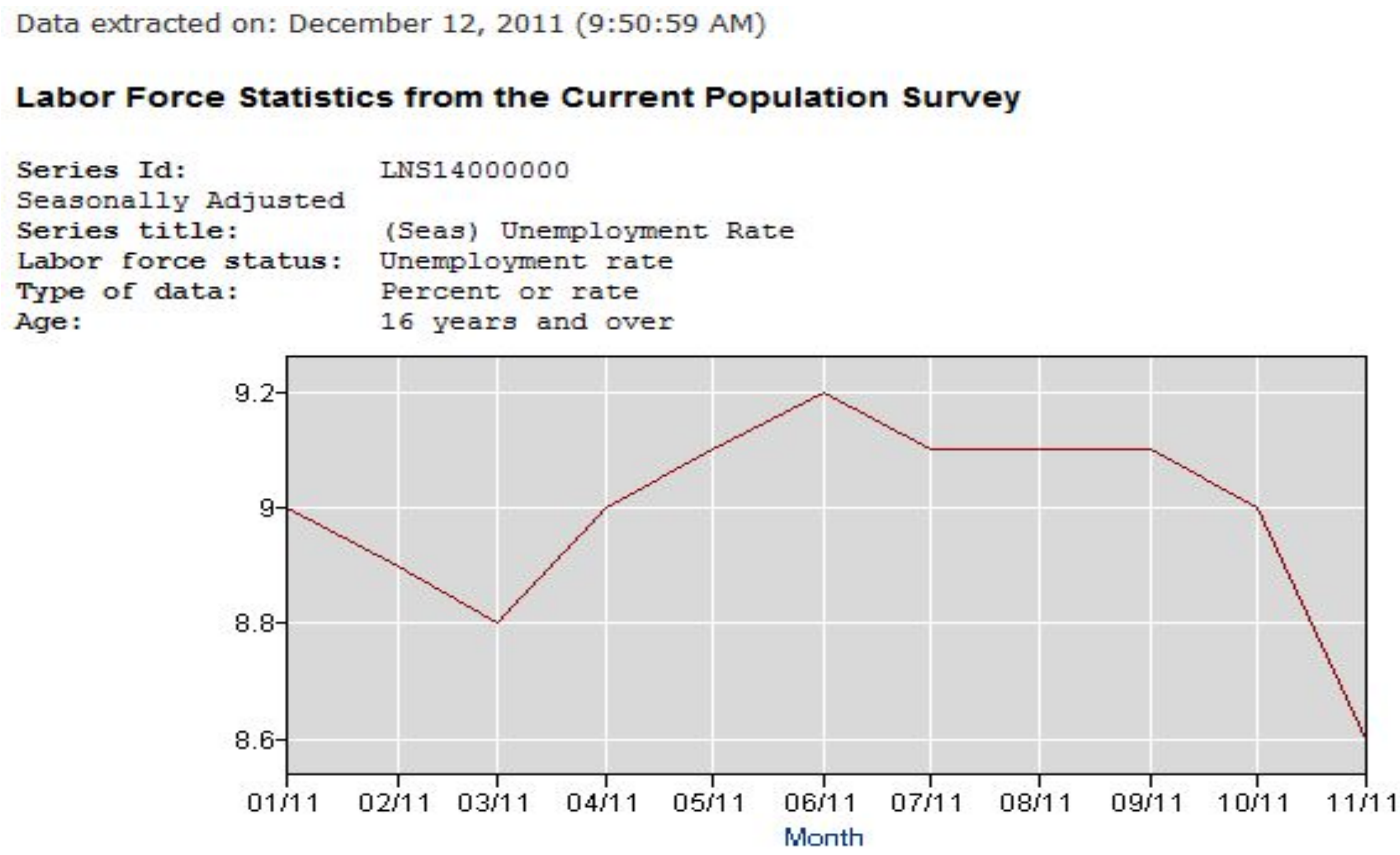
Show Change since
● Close ○ 26 weeks
○ 52 weeks ○ YTD

Highlight Top 5
○ Gainers ○ Losers
● None

Find (name or ticker)

Color scheme
● red/green

But be careful!!



Data extracted on: December 12, 2011 (9:50:59 AM)

Labor Force Statistics from the Current Population Survey

```
Series Id:              LNS14000000
Seasonally Adjusted
Series title:           (Seas) Unemployment Rate
Labor force status:     Unemployment rate
Type of data:           Percent or rate
Age:                    16 years and over
```

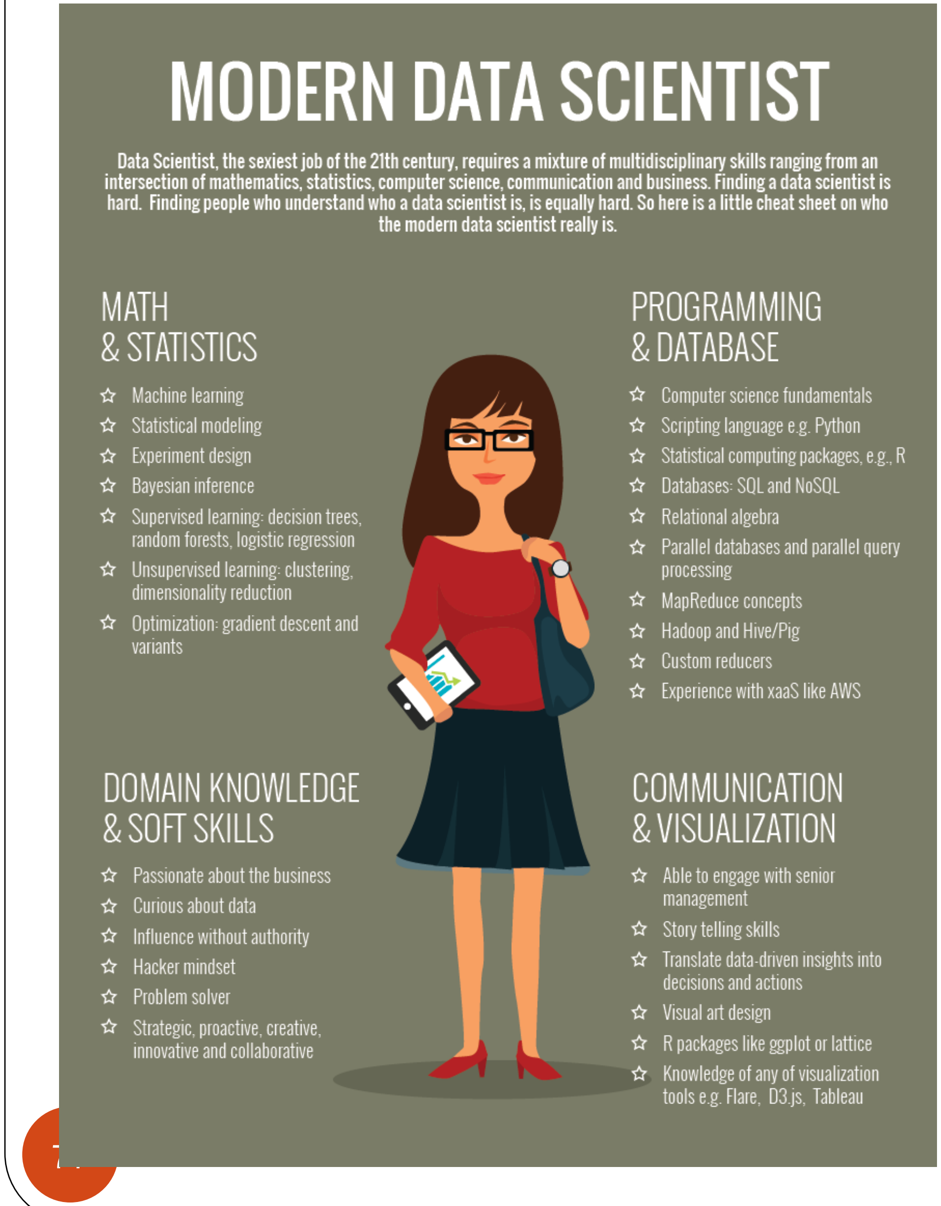
What's wrong with this chart??

Data scientist: a brand new profession

- Data Scientist: The Sexiest Job of the 21st Century [Harward Business Review 2013]
- Data scientist? A guide to 2015's hottest profession [Mashable 2015]
- “It’s official – data scientist is the best job in America” [Forbes, 2016]

Video

Skills of data scientists



MODERN DATA SCIENTIST

Data Scientist, the sexiest job of the 21th century, requires a mixture of multidisciplinary skills ranging from an intersection of mathematics, statistics, computer science, communication and business. Finding a data scientist is hard. Finding people who understand who a data scientist is, is equally hard. So here is a little cheat sheet on who the modern data scientist really is.

MATH & STATISTICS

- ☆ Machine learning
- ☆ Statistical modeling
- ☆ Experiment design
- ☆ Bayesian inference
- ☆ Supervised learning: decision trees, random forests, logistic regression
- ☆ Unsupervised learning: clustering, dimensionality reduction
- ☆ Optimization: gradient descent and variants

PROGRAMMING & DATABASE

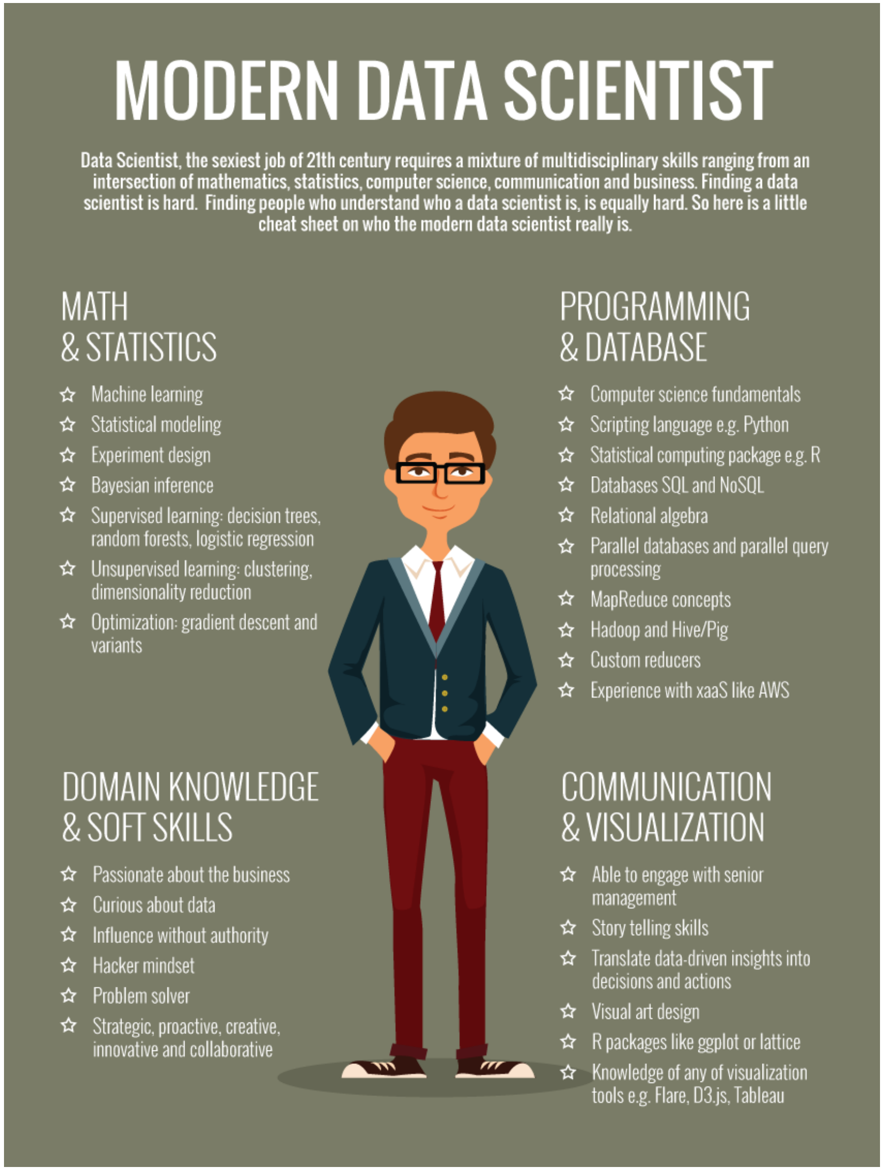
- ☆ Computer science fundamentals
- ☆ Scripting language e.g. Python
- ☆ Statistical computing packages, e.g., R
- ☆ Databases: SQL and NoSQL
- ☆ Relational algebra
- ☆ Parallel databases and parallel query processing
- ☆ MapReduce concepts
- ☆ Hadoop and Hive/Pig
- ☆ Custom reducers
- ☆ Experience with xaaS like AWS

DOMAIN KNOWLEDGE & SOFT SKILLS

- ☆ Passionate about the business
- ☆ Curious about data
- ☆ Influence without authority
- ☆ Hacker mindset
- ☆ Problem solver
- ☆ Strategic, proactive, creative, innovative and collaborative

COMMUNICATION & VISUALIZATION

- ☆ Able to engage with senior management
- ☆ Story telling skills
- ☆ Translate data-driven insights into decisions and actions
- ☆ Visual art design
- ☆ R packages like ggplot or lattice
- ☆ Knowledge of any of visualization tools e.g. Flare, D3.js, Tableau



MODERN DATA SCIENTIST

Data Scientist, the sexiest job of 21th century requires a mixture of multidisciplinary skills ranging from an intersection of mathematics, statistics, computer science, communication and business. Finding a data scientist is hard. Finding people who understand who a data scientist is, is equally hard. So here is a little cheat sheet on who the modern data scientist really is.

MATH & STATISTICS

- ☆ Machine learning
- ☆ Statistical modeling
- ☆ Experiment design
- ☆ Bayesian inference
- ☆ Supervised learning: decision trees, random forests, logistic regression
- ☆ Unsupervised learning: clustering, dimensionality reduction
- ☆ Optimization: gradient descent and variants

PROGRAMMING & DATABASE

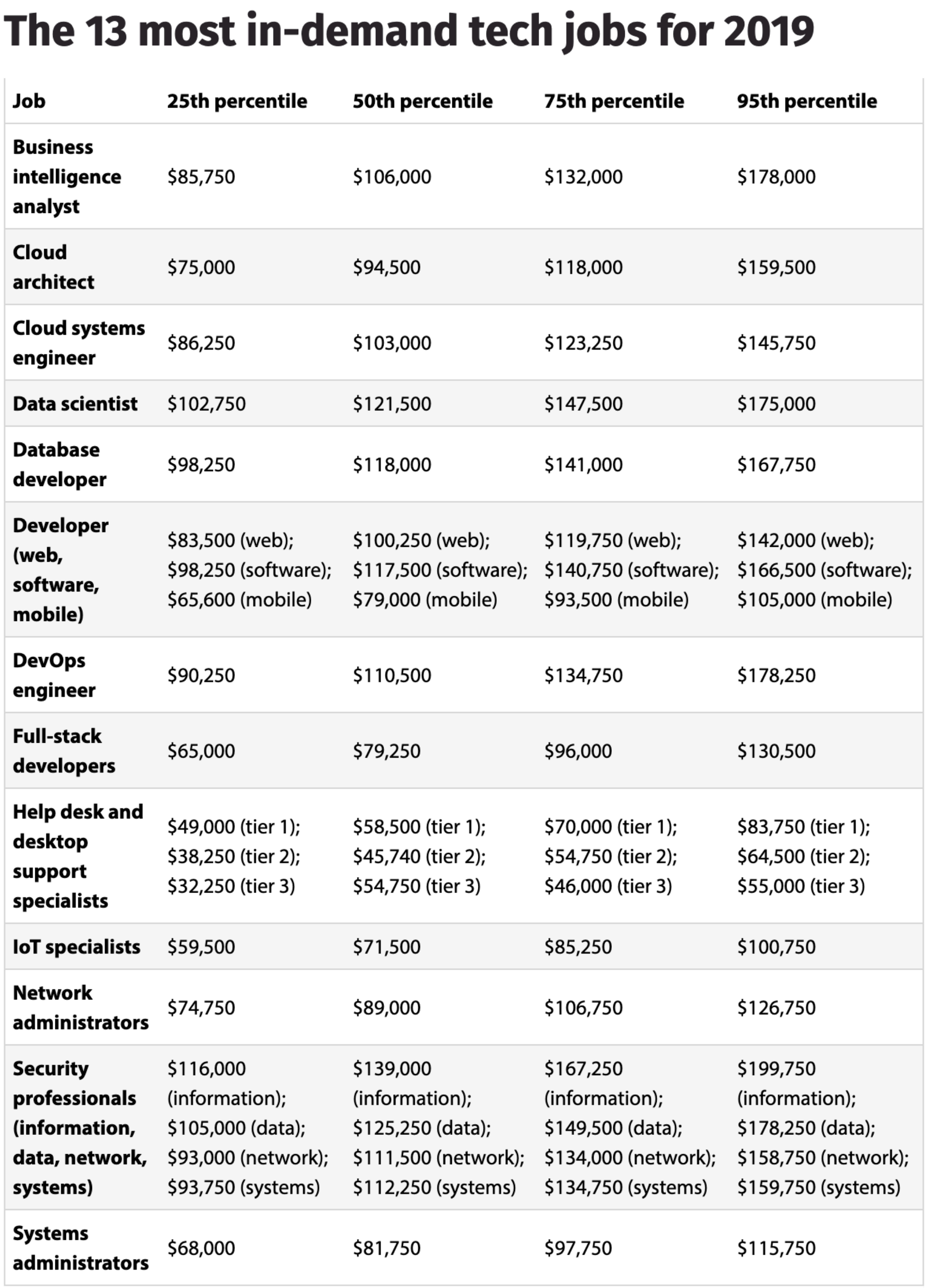
- ☆ Computer science fundamentals
- ☆ Scripting language e.g. Python
- ☆ Statistical computing package e.g. R
- ☆ Databases SQL and NoSQL
- ☆ Relational algebra
- ☆ Parallel databases and parallel query processing
- ☆ MapReduce concepts
- ☆ Hadoop and Hive/Pig
- ☆ Custom reducers
- ☆ Experience with xaaS like AWS

DOMAIN KNOWLEDGE & SOFT SKILLS

- ☆ Passionate about the business
- ☆ Curious about data
- ☆ Influence without authority
- ☆ Hacker mindset
- ☆ Problem solver
- ☆ Strategic, proactive, creative, innovative and collaborative

COMMUNICATION & VISUALIZATION

- ☆ Able to engage with senior management
- ☆ Story telling skills
- ☆ Translate data-driven insights into decisions and actions
- ☆ Visual art design
- ☆ R packages like ggplot or lattice
- ☆ Knowledge of any of visualization tools e.g. Flare, D3.js, Tableau



The 13 most in-demand tech jobs for 2019

Job	25th percentile	50th percentile	75th percentile	95th percentile
Business intelligence analyst	$85,750	$106,000	$132,000	$178,000
Cloud architect	$75,000	$94,500	$118,000	$159,500
Cloud systems engineer	$86,250	$103,000	$123,250	$145,750
Data scientist	$102,750	$121,500	$147,500	$175,000
Database developer	$98,250	$118,000	$141,000	$167,750
Developer (web, software, mobile)	$83,500 (web); $98,250 (software); $65,600 (mobile)	$100,250 (web); $117,500 (software); $79,000 (mobile)	$119,750 (web); $140,750 (software); $93,500 (mobile)	$142,000 (web); $166,500 (software); $105,000 (mobile)
DevOps engineer	$90,250	$110,500	$134,750	$178,250
Full-stack developers	$65,000	$79,250	$96,000	$130,500
Help desk and desktop support specialists	$49,000 (tier 1); $38,250 (tier 2); $32,250 (tier 3)	$58,500 (tier 1); $45,740 (tier 2); $54,750 (tier 3)	$70,000 (tier 1); $54,750 (tier 2); $46,000 (tier 3)	$83,750 (tier 1); $64,500 (tier 2); $55,000 (tier 3)
IoT specialists	$59,500	$71,500	$85,250	$100,750
Network administrators	$74,750	$89,000	$106,750	$126,750
Security professionals (information, data, network, systems)	$116,000 (information); $105,000 (data); $93,000 (network); $93,750 (systems)	$139,000 (information); $125,250 (data); $111,500 (network); $112,250 (systems)	$167,250 (information); $149,500 (data); $134,000 (network); $134,750 (systems)	$199,750 (information); $178,250 (data); $158,750 (network); $159,750 (systems)
Systems administrators	$68,000	$81,750	$97,750	$115,750

After this course

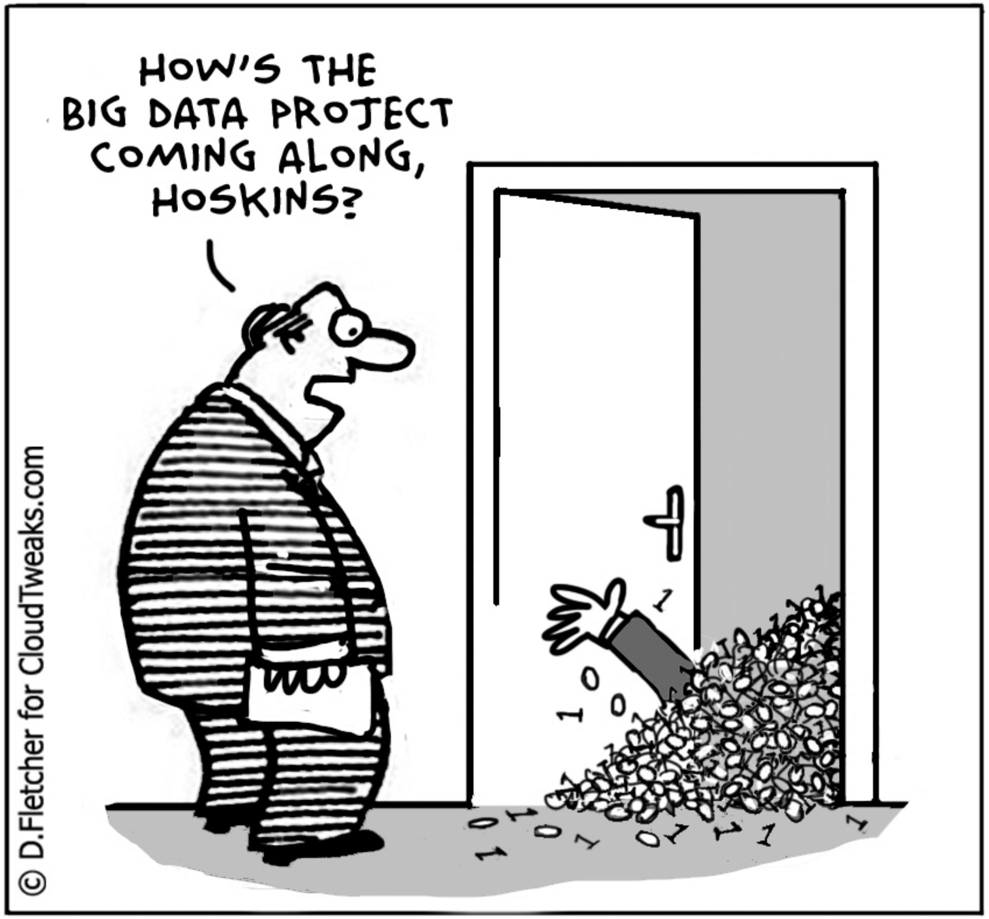
© 2014 Ted Goff www.tedgoff.com

"So you want to hire me as a Data Scientist for Intelligent Virtualized Deep Machine Learning Real-time Big Data in the Cloud for Social Networks? Ok, but if you also want Hadoop, increase my salary by 50%."

Conclusions

- We live in the era of Big Data
- Wide range of availability in different areas
- Big opportunities to solve big problems
- They can create value
- The challenge is how to manage and use them
- New technologies are needed
- Methodological aspects are important
- A rapidly evolving area
- Data scientists: the current hottest profession in IT

So, let us face big data projects..



© D.Fletcher for CloudTweaks.com

..with a Bruce Willis attitude!

BIG DATA IS GOING TO IMPACT US SOON!
CALL BRUCE WILLIS!!
DATA
www.nuxeo.com
Soulcié

References

- "Big Data: The next frontier for innovation, competition, and productivity". Rapporto McKinsey&Company, 2012.
- "Challenges and Opportunities with Big Data". A community white paper developed by leading researchers across the United States, 2012.
- "Taming The Big Data Tidal Wave: Finding Opportunities in Huge Data Streams with Advanced Analytics". Bill Franks, John Wiley & Sons, 2012.