Big Data Processing and Analytics

Haridimos Kondylakis
http://www.csd.uoc.gr/~hy562
University of Crete
Fall 2024

What this Course is About

What You Will learn

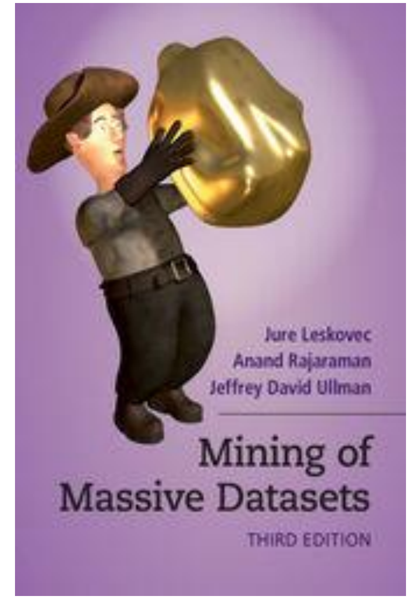
- Understand different models of computation:
  - MapReduce
  - Spark
- Mine different types of data:
  - Data is high dimensional
  - Data is infinite/never-ending
- Use different mathematical ‘tools’:
  - Hashing (LSH, Bloom filters)
  - Dynamic programming (frequent itemsets)
- Solve real-world problems:
  - Data Exchange
  - Schema Discovery
  - Data Summarization
  - Big Data in the Quantum Era

Prerequisites

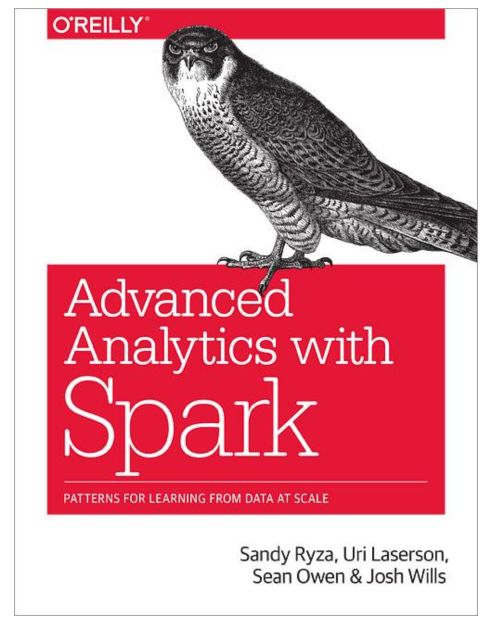
- Algorithms
  - Basic data structures, (dynamic programming)
- Basic probability
  - Typical distributions, maximum likelihood estimation (MLE), …
- Programming
  - We recommend Java, Python, or Scala
    - feel free to pick your own favorite

Course Textbooks

- *Jure Leskovec, Anand Rajaraman, Jeff Ullman. "Mining of Massive Datasets"* Cambridge University Press, 2020 https://www.cambridge.org/gr/academic/subjects/computer-science/pattern-recognition-and-machine-learning/mining-massive-datasets-3rd-edition
  - Free download *http://www.mmds.org*



- *Sandy Ryza, Uri Laserson, Sean Owen, Josh Wills. "Advanced Analytics With Spark: Patterns for Learning from Data at Scale" O'Reilly Media 2017* http://shop.oreilly.com/product/0636920035091.do



Tentative Course Schedule

- Week 1 (23/09-25/09) : Course Overview
- Week 2 (~~30/09~~-02/10) : Scalable Data Analytics (Assign. 1)
- Week 3 (07/10-09/10) : Finding Similar Items
- Week 4 (14/10-16/10) : Massive Data Processing (Assign. 1 due)
- Week 5 (21/10-23/10) : Extracting Association Rules (Assign. 2)
- Week 6 (~~28/10~~-30/10) : Streaming Analytics
- Week 7 (04/11-06/11) : Streaming Analytics
- Week 8 (~~11/11~~-13/11) : Semantic Summaries (Assign 2. due)
- Week 9 (18/11-20/11) : Schema Extraction
- Week 10 (25/11-27/11) : Data Exchange
- Week 11 (02/12-04/12) : Student paper presentations
- Week 12 (09/12-11/12) : Data Management in the Quantum Era
- Week 13 (16/12-18/12) : Student project presentations

- Lab 1 (04/10): MapReduce Programming
- Lab 2 (11/10): Programming in Spark
- Lab 3 (18/10): Assisting Lecture for Assign. 2
- Lab 4 (01/11): Intro to Data Frames and Spark SQL
- Lab 5 (08/11): Intro to Spark Streaming

Course Organization

- 2 Programming Exercises (30%): MapReduce & Spark
- 1 Research presentation (20%): Semantic Summarization
- Final Project (in Teams) (50%): Property Graphs Schema Extraction
  - Paper submission to ISWC/ESWC ☺
- TA: Zubaria Asma (csdp1232@csd.uoc.gr)

Words of Caution

- We can only cover a small part of the big data universe
  - Do not expect all possible architectures, programming models, theoretical results, or vendors to be covered
- This really is an algorithms course, not a basic programming course
  - But you will need to do a lot of non-trivial programming
- There are few certain answers, as people in research and leading tech companies are trying to understand how to deal with big data
- We are working with cutting-edge technology
  - Bugs, lack of documentation, new APIs
- In short: you will deal with inevitable frustrations and plan your work accordingly…
- …but if you can do that and are willing to invest the time, it will be a rewarding experience

Learning with examples!

- Understand different models of computation:
  - MapReduce
  - Spark
- Mine different types of data:
  - Data is high dimensional
  - Data is infinite/never-ending
- Use different mathematical ‘tools’:
  - Hashing (LSH, Bloom filters)
  - Dynamic programming (frequent itemsets)
- Solve real-world problems:
  - Data Ethics
  - Data Exchange
  - Schema Discovery
  - Data Summarization

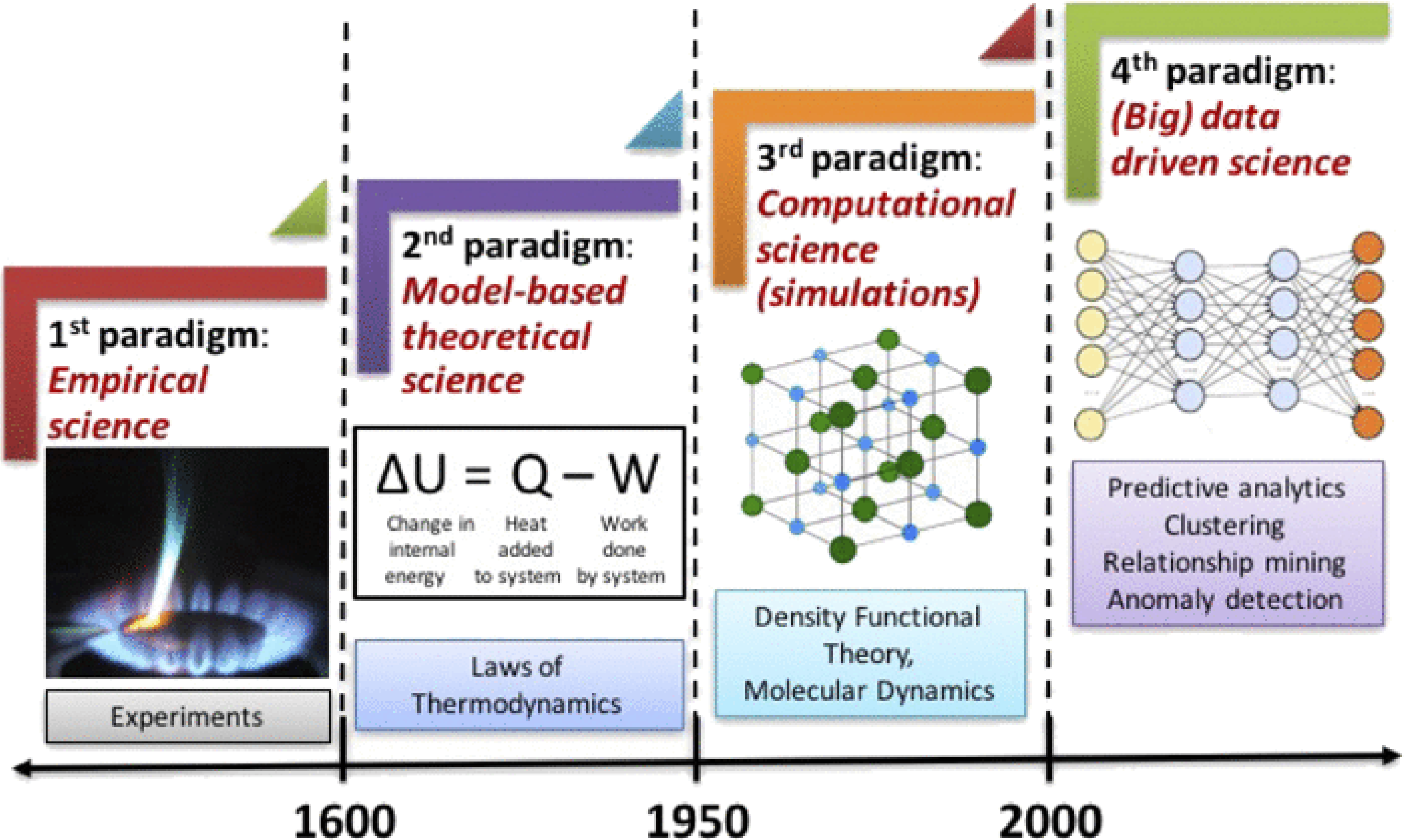
Hands-On “Game of Thrones”

- A network of character interactions from the novel "A Storm of Swords"

- Explore the dataset: https://bit.ly/3uatf5r
- We have an adjacency list of characters and their number of interactions throughout the text.
- Formulate teams of two-three persons
- Answer the following key questions
  - What key statistics can you provide?
  - How to identify key patterns in the data?
  - How to visualize data?
  - How to enable meaningful data exploration

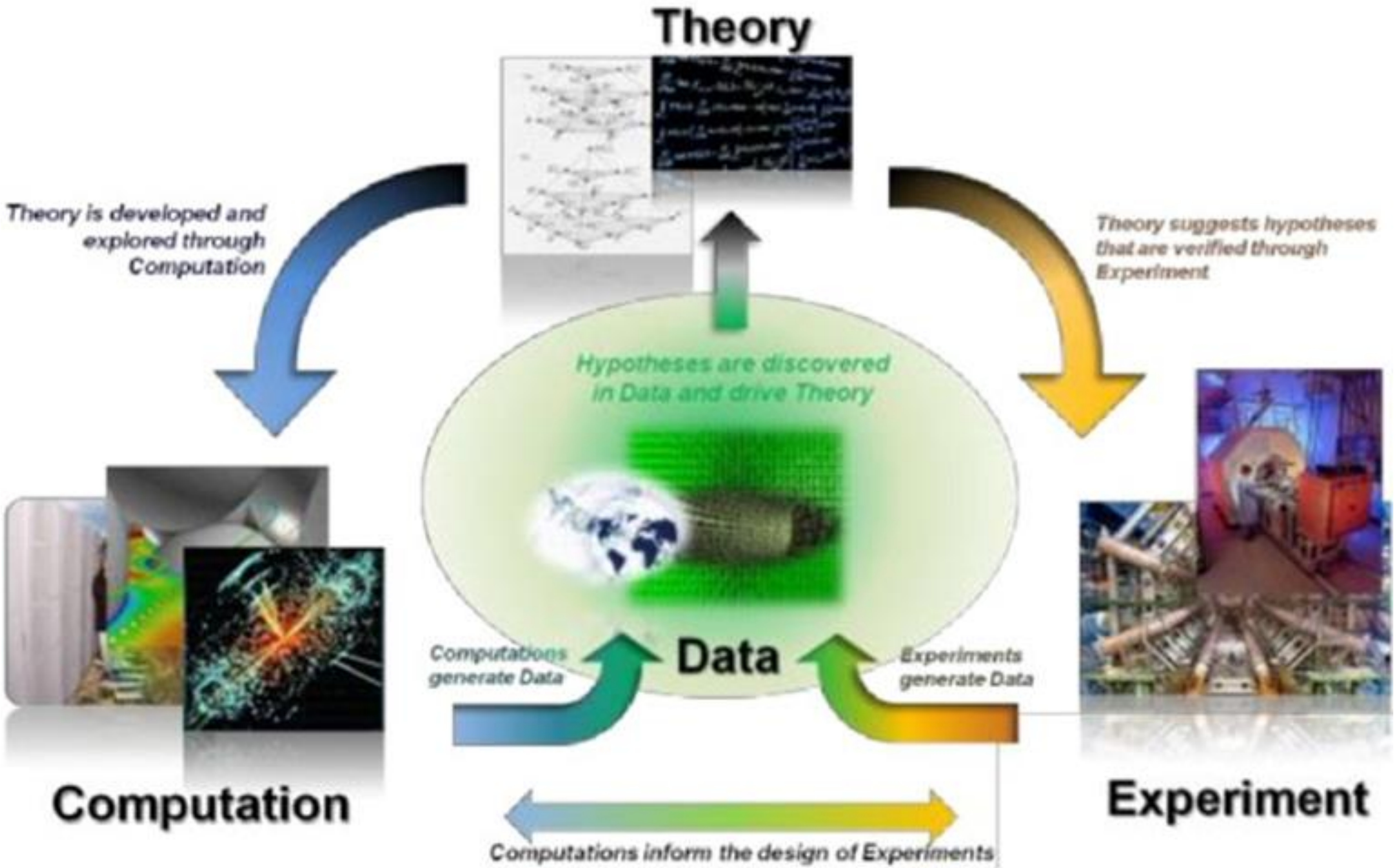
The Data Avalanche: From Science to Business

Shifting Paradigm in Sciences



https://aip.scitation.org/doi/10.1063/1.4946894

Data-driven Discovery



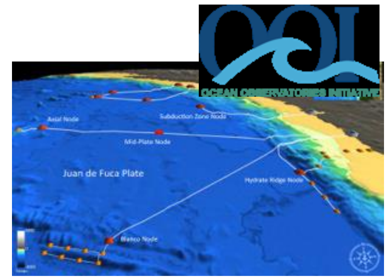
- Data-driven discovery is revolutionizing scientific exploration as well as engineering innovations
  - From hypothesis driven to hypothesis generating

R. Leland, R. Murphy, B. Hendrickson, K. Yelick, J. Johnson, J. Berry
Large-Scale Data Analytics & its Relationship to Simulation Jan. 2014

From “Data Poor” to “Data Rich” Scientific Research

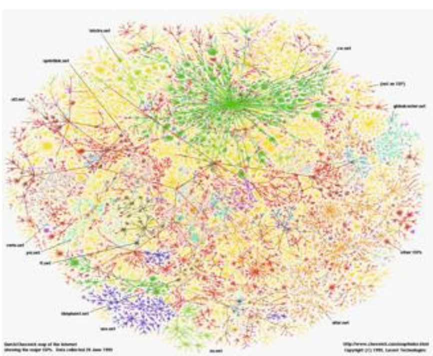
Astronomy: LSST

Physics: LHC

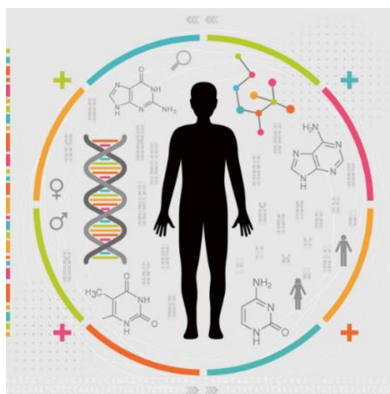


Oceanography

Biology: Sequencing



Sociology: The Web



Precision Medicine

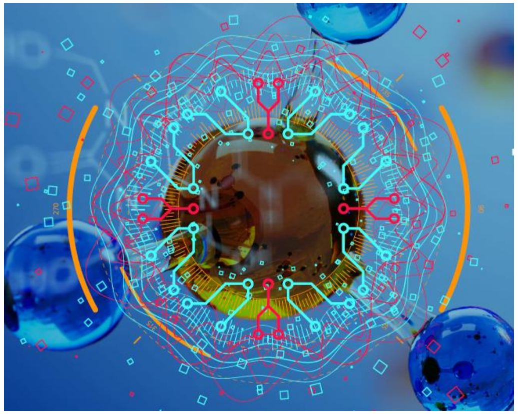
Neuroscience: EEG, fMRI

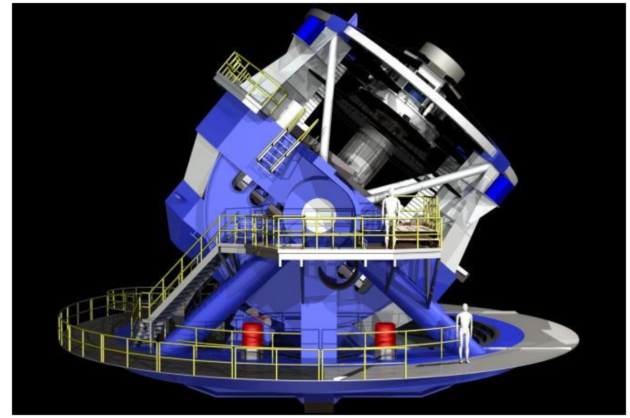
Sports

- Data deluge spans biology, climate, cosmology, materials, physics, …

M.Franklin Big Data Software: What’s Next? (and what do we have to say about it?) VLDB 2017

New Research Methods





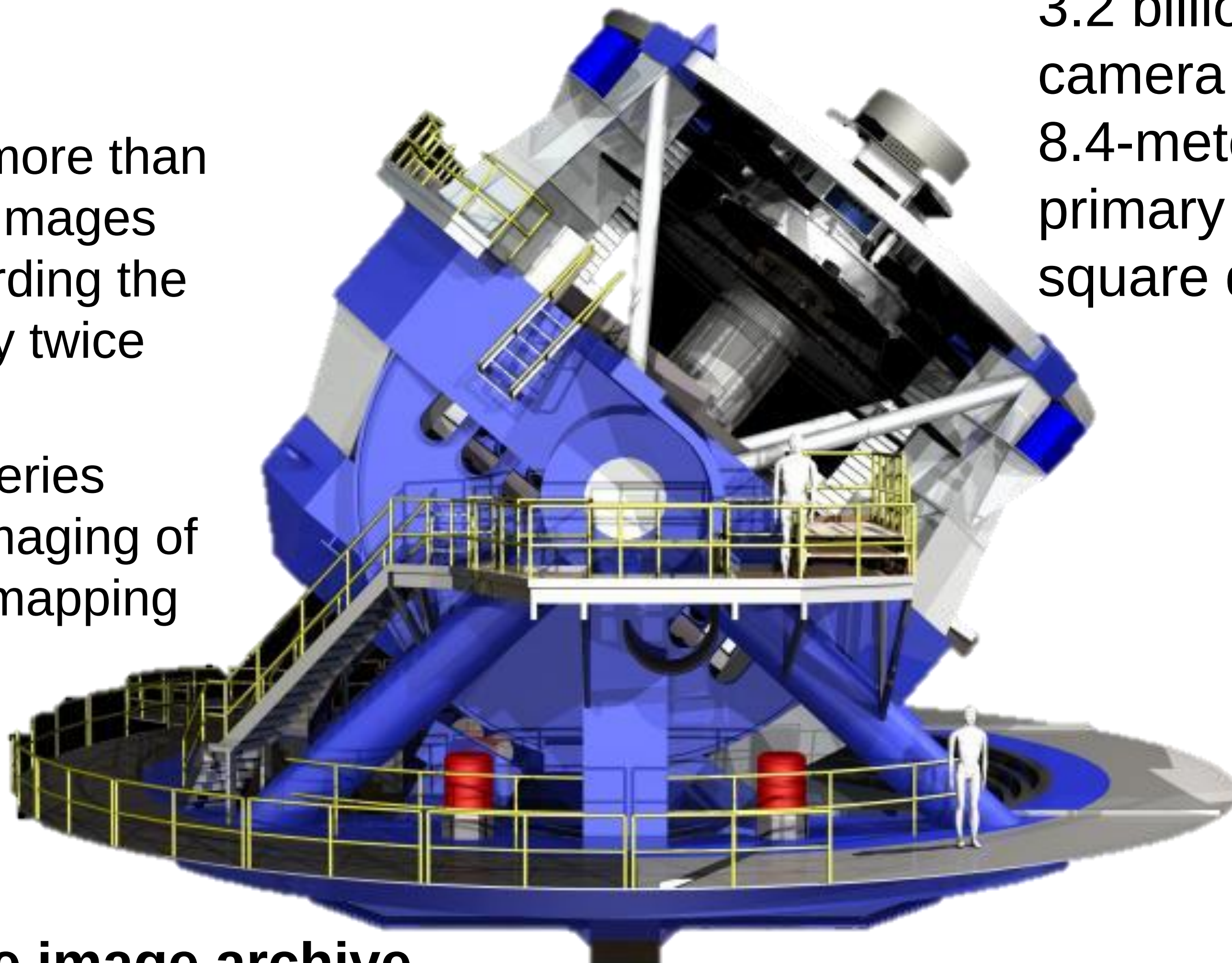
- Simulation Data: Increasing level of simulation detail and duration, as well as, model size by orders of magnitude!
- Experimental Data: Light sources, genome sequencing, next generation ARM radars, sky surveys, neuro-sensing and stimulation, …

- New research methods depend on coupling computation and experiment as well as on integrating data across sources and/or types

Large Synoptic Survey Telescope (LSST)

3.2 billion-pixel camera
8.4-meter diameter primary mirror = 10 square degrees!

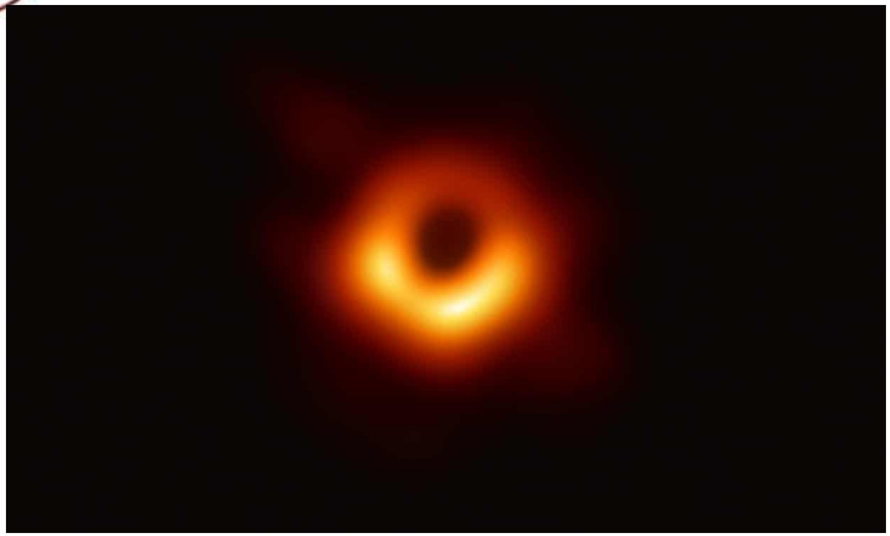
- LSST will take more than 800 panoramic images each night recording the entire visible sky twice each week
- Ten-year time series (~2020-2030) imaging of the night sky – mapping the Universe !



100-200 Petabyte image archive
20-40 Petabyte database catalog

www.lsst.org

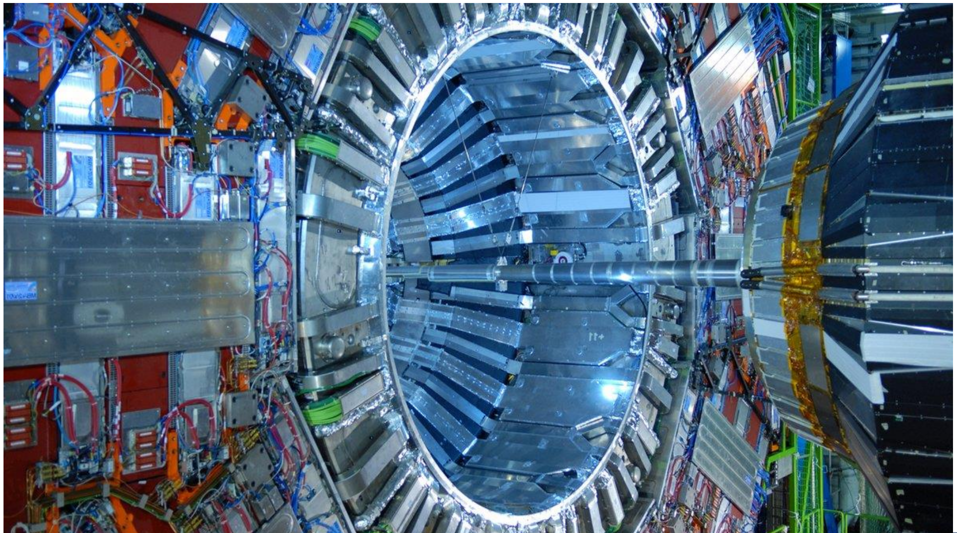
First Image of a Black Hole



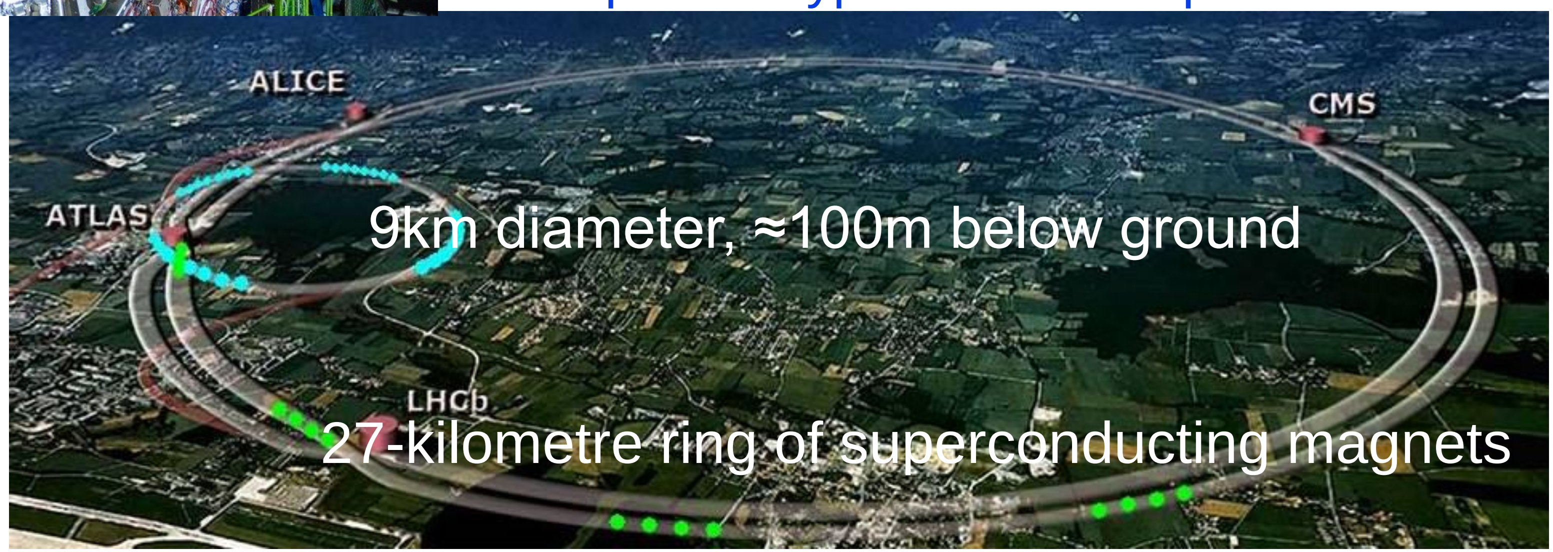
- Captured by the Event Horizon telescope (EHT), an NSF funded network of eight radio telescopes spanning locations from Antarctica to Spain and Chile, in an effort involving more than 200 scientists
  - achieved resolutions of 22.5 microarcseconds, enabling the array to resolve the event horizon of the black hole at the center of M87
  - a single-dish telescope would have to be 12000 km in diameter to achieve this same sharpness
- K. Bouman posing with 5 petabytes of data necessary to image a black hole

https://www.facebook.com/BusinessInsiderScience/videos/378897386038615

Large Hadron Collider (LHC)

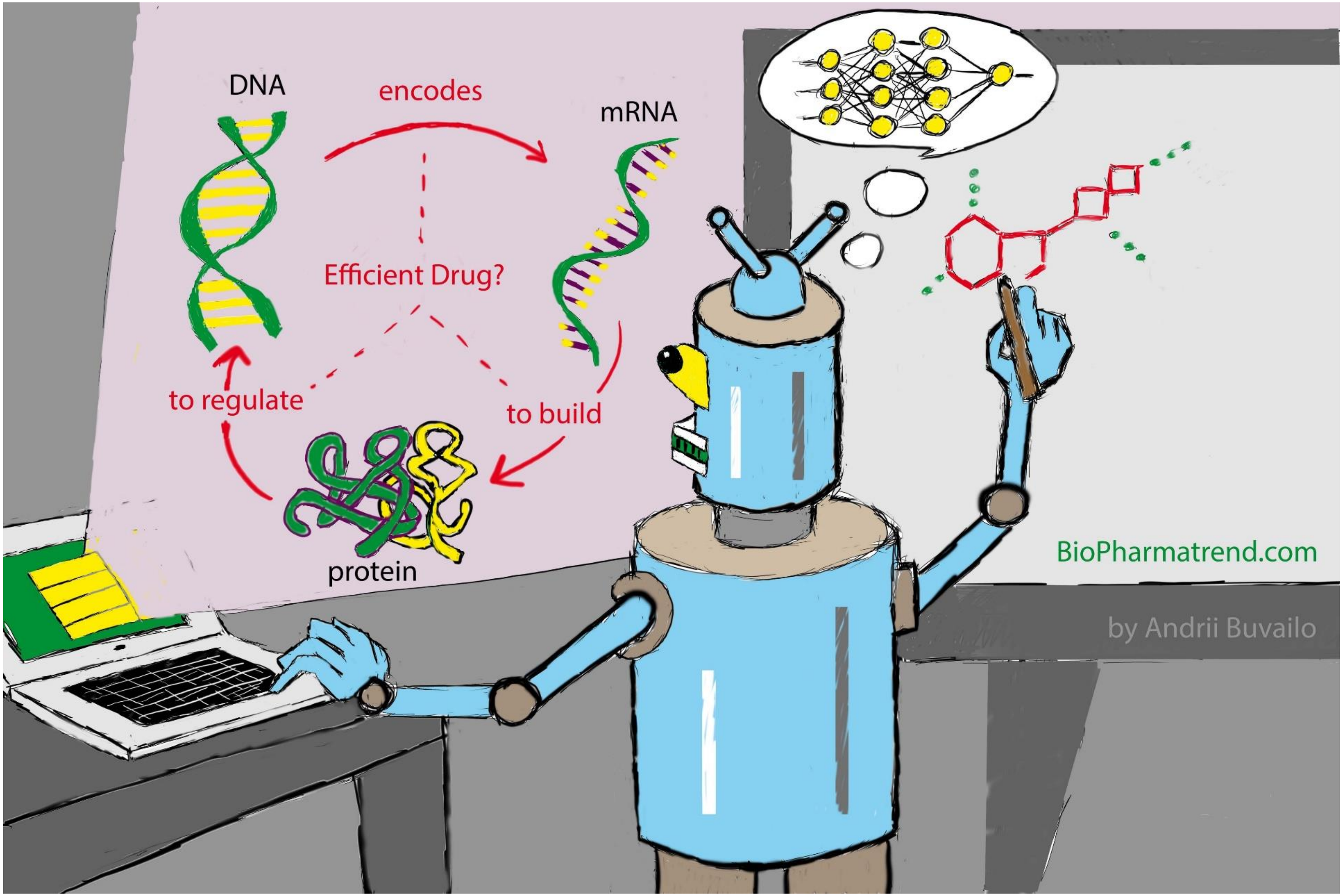


- Protons collide some 1 billion times per second where each collision produces about a megabyte of data
- Even after filtering out about 99% of it, scientists are left with around 30 petabytes each year to analyze for a wide range of physics experiments, including studies on the Higgs boson
  - reconstructing particle trajectories, the particle types and their speeds



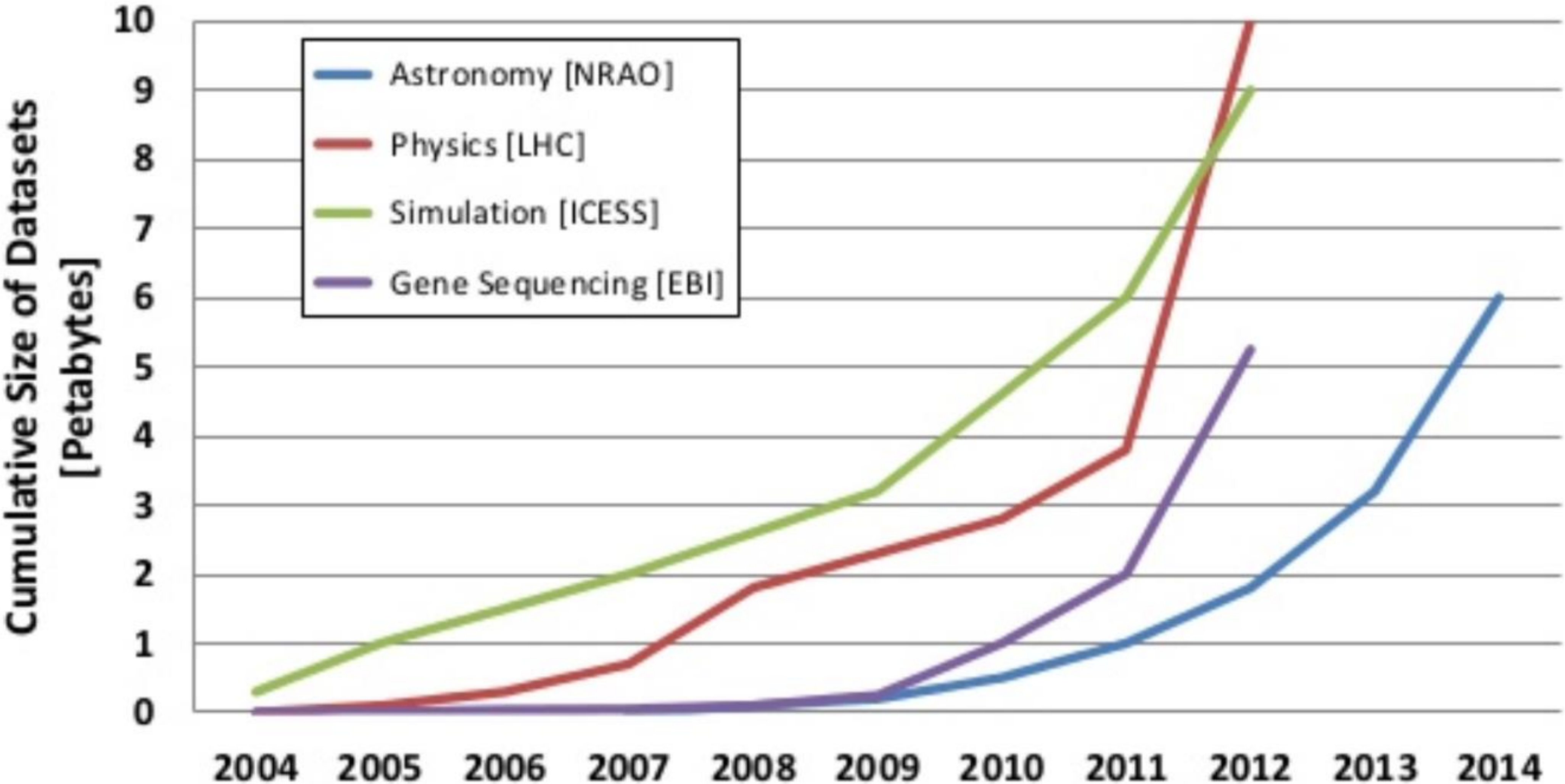
http://home.web.cern.ch/topics/large-hadron-collider

AI is Changing Drug Discovery!



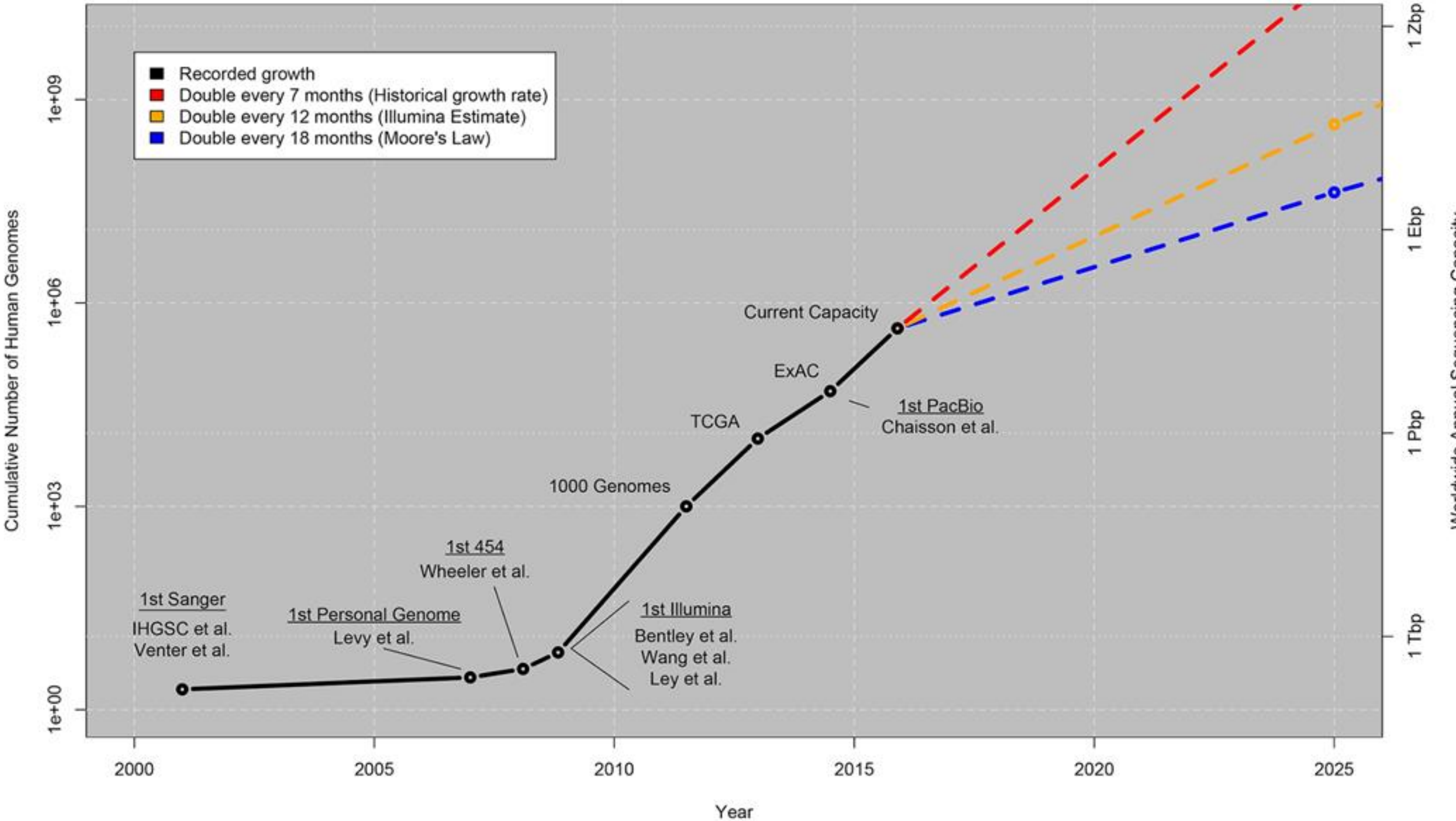
https://medium.com/@ABuvailo/artificial-intelligence-in-drug-discovery-2018-year-in-review-e17b99c99078

Scientific Data Grows Exponentially



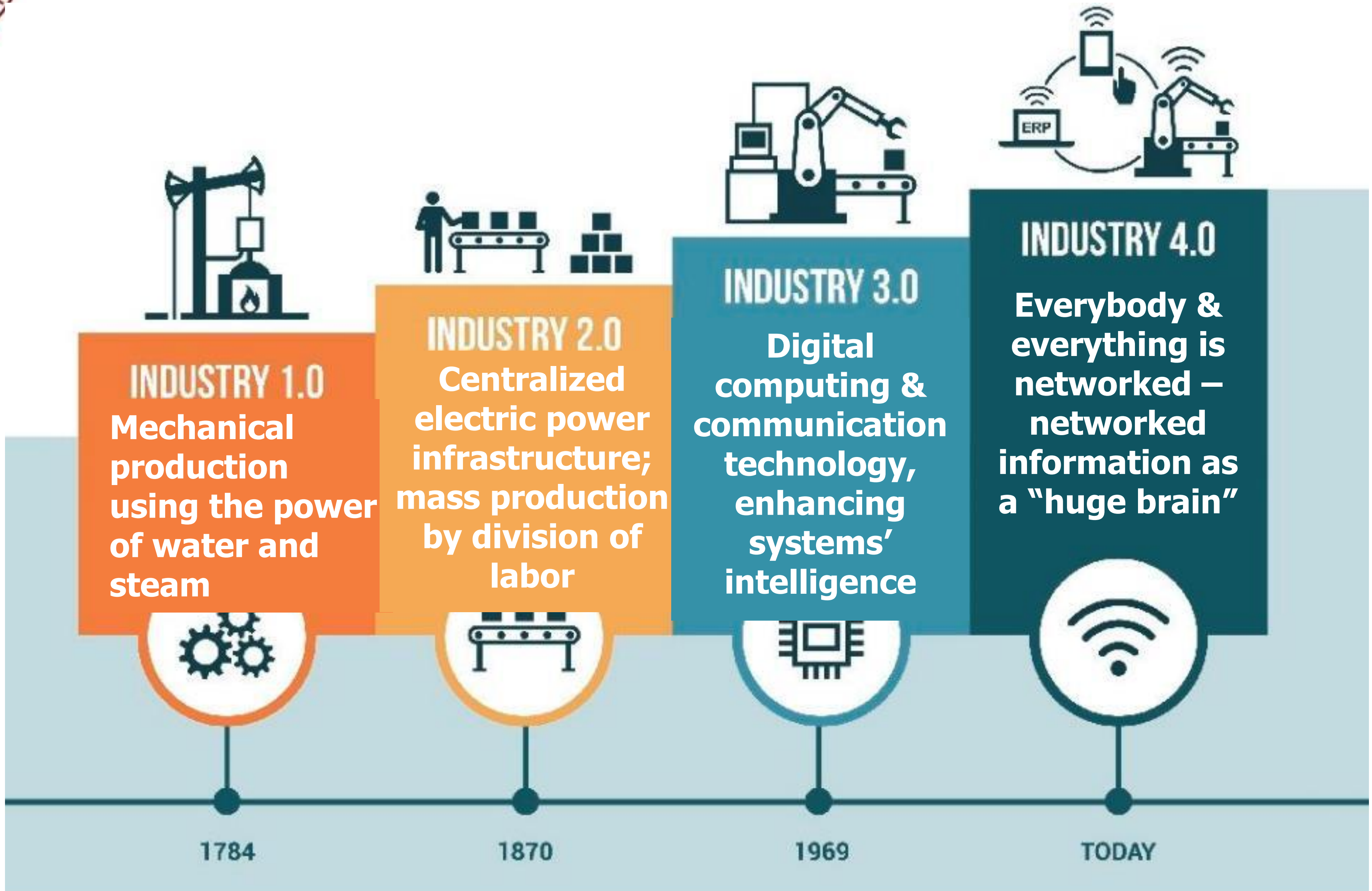
https://www.slideshare.net/eXascaleInfolab/braintalk-cuso-nm

Growth of DNA Sequencing



blogs.springeropen.com/springeropen/wp-content/uploads/sites/16/2018/01/bioData.png

The Four Industrial Revolutions



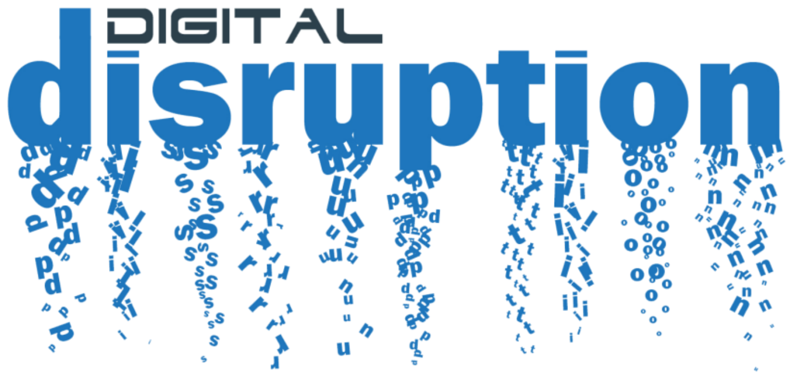
Henning Kagermann et.al., Recommendations for implementing the strategic initiative Industrie 4.0 Acatech, 2013

Digital Transformation of the Physical World

Industry	Past: Selling a Product	Future: a Service
Energy & utilities	Power networks/grids	On demand energy production/ consumption
Automotive	Cars	Transportation (assisted, autonomous driving)
Agriculture	Seeds	Crop Yields
Healthcare	Diabetes pumps	Diabetes cares
Food	Packaged goods	Nutrition
Cities	Physical Urban infrastructure / Facilities	Smart city e-services (street lighting, urban noise/pollution/ traffic monitoring, parking/waste management etc.)
….	….	….
IT Industry	Computers	Computation

- McKinsey, GE, IBM, Cisco et al. estimate hundreds of billion dollar savings/efficiency improvements in the next 10 years

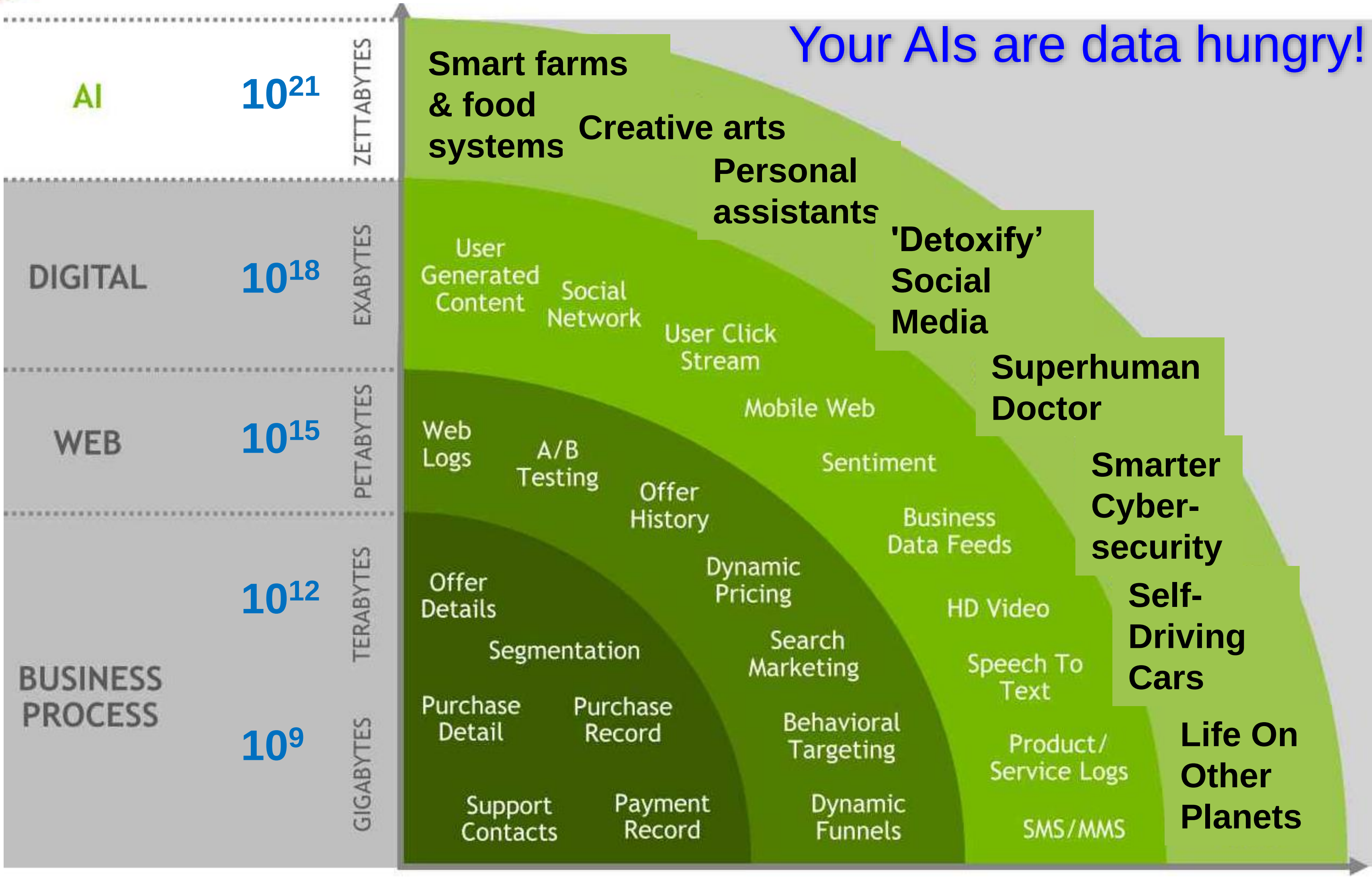
Digital Disruption Already Happening !



- Largest telco company owns no telco infrastructure (Skype)
- World’s largest movie house owns no cinemas (Netflix)
- World’s most valuable retailer has no inventory (Alibaba)
- Most popular media owner creates no content (Facebook)
- World’s largest taxi company owns no vehicles (Uber)
- Largest accommodation provider owns no real estate (Airbnb)
- Fastest growing bank has no actual cash (Bitcoin)

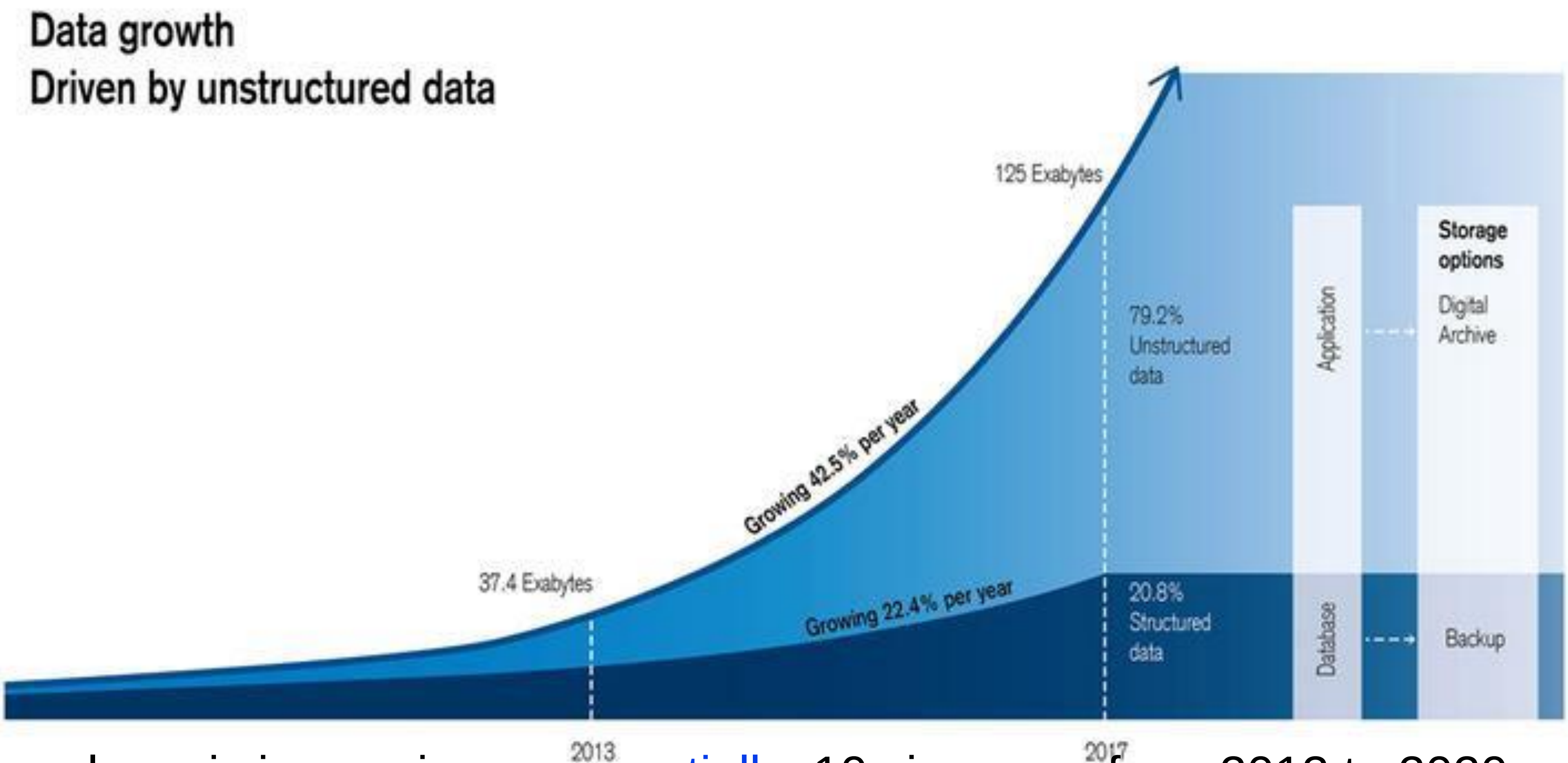
http://www.independent.co.uk/news/business/comment/hamish-mcrae/facebook-airbnb-uber-and-the-unstoppable-rise-of-the-content-non-generators-10227207.html

The Data Tsunami: Transactions + Interactions + Observations



https://www.slideshare.net/KeithKraus/gpuaccelerating-udfs-in-pyspark-with-numba-and-pygdf

Data Growth Over the Years



- Data volume is increasing exponentially: 10x increase from 2013 to 2020
- By 2025, about 25% of all data will be real time in nature out of which 95% of it will be generated by IoT!

https://www.emc.com/leadership/digital-universe/2014iview/executive-summary.htm

Driving Innovation with Big Data

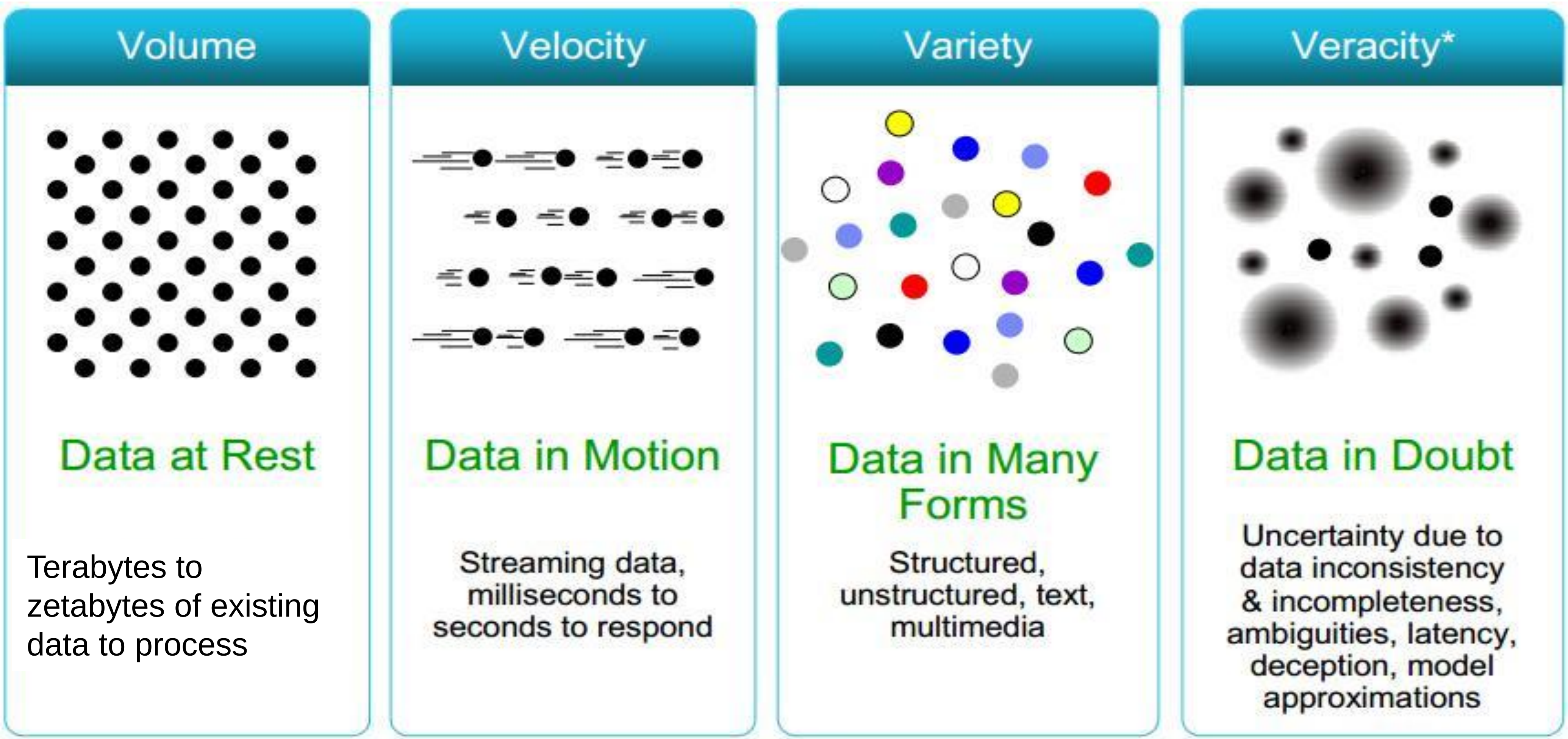
Progress and Innovation no longer hindered by the ability to collect data, but by the ability to *manage*, *analyze*, *summarize*, *visualize*, and *discover* knowledge from the collected data
in a *timely manner* and in a *scalable fashion*

What Makes Data, “Big” Data?

Definitions

- No single standard definition…
  - "Big Data" is data whose scale, diversity, and complexity require new architecture, techniques, algorithms, and analytics to manage it and extract value and hidden knowledge from it… (McKinsey Global Inst.)
  - "Big Data" is high-volume, high-velocity and high-variety information assets that demand cost-effective, innovative forms of information processing for enhanced insight and decision making (Gartner)

The Four V's of Big Data



www-05.ibm.com/fr/events/netezzaDM_2012/Solutions_Big_Data.pdf

Characteristics of Big Data: 1-Scale (Volume)

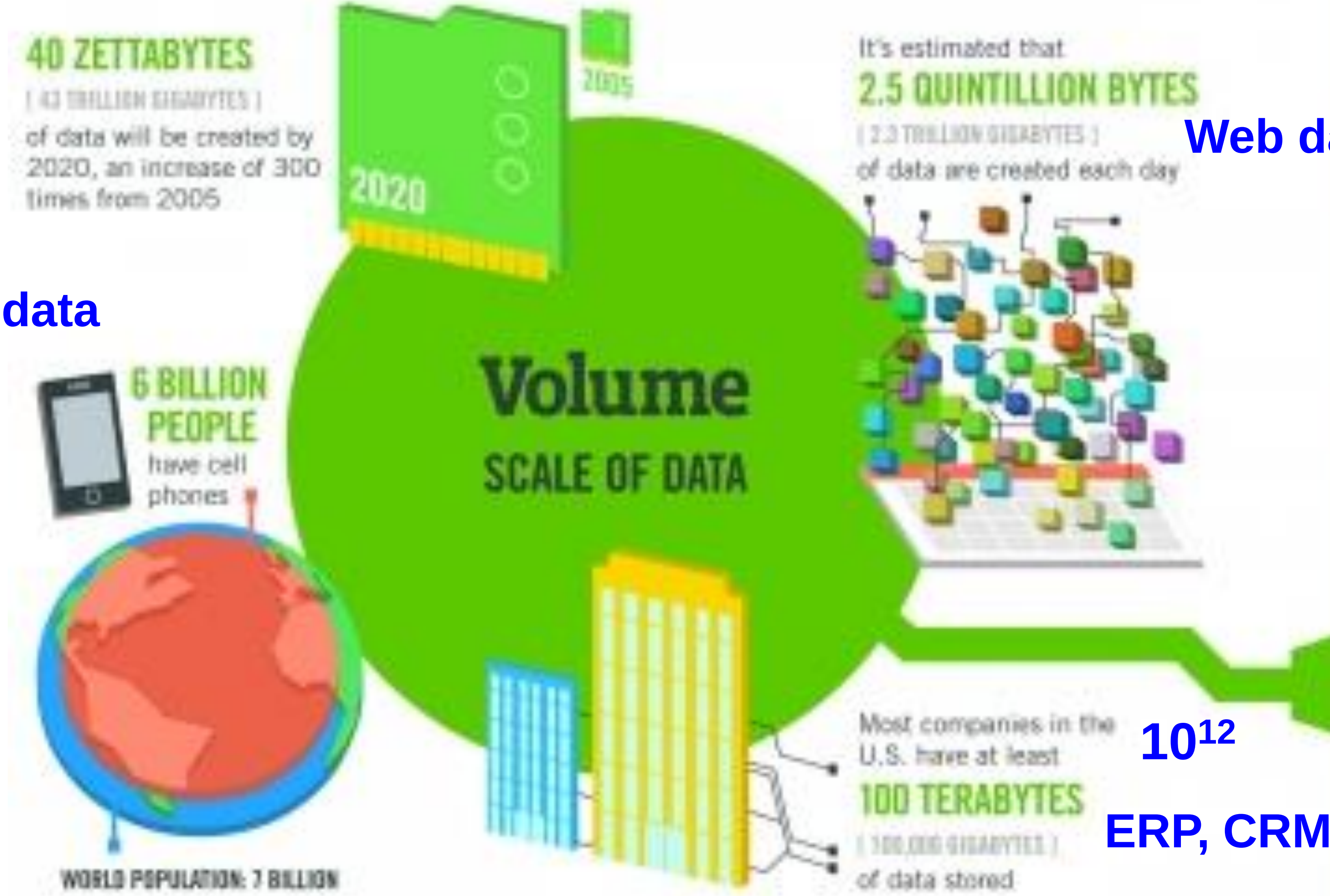
10^{21}

Mobile data

Web data

10^{12}

ERP, CRM data

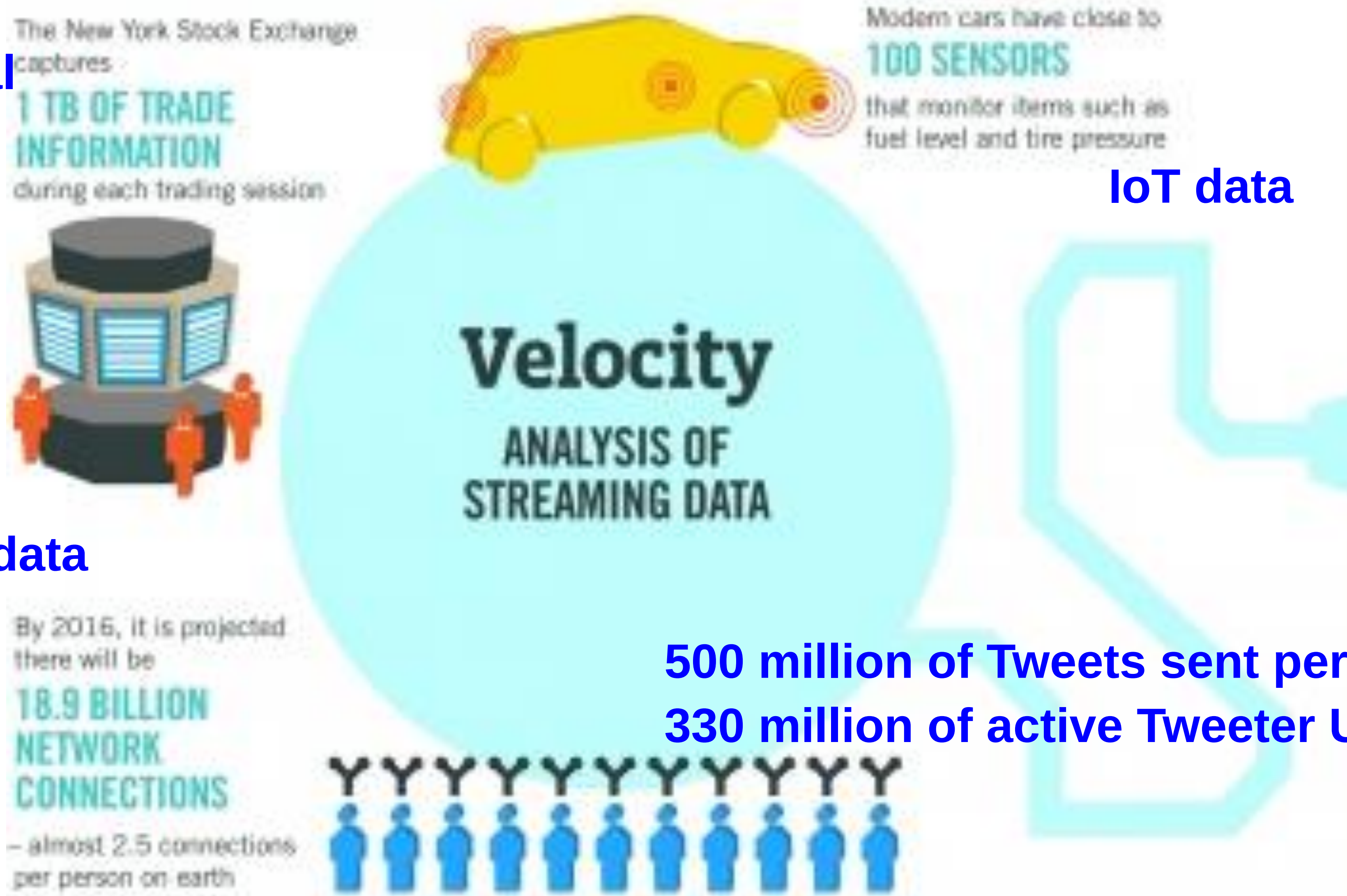


Too big: petabyte-scale collections or lots of (not necessarily big) data sets

Characteristics of Big Data: 2-Speed (Velocity)

Financial data

IoT data



Social data

500 million of Tweets sent per Day

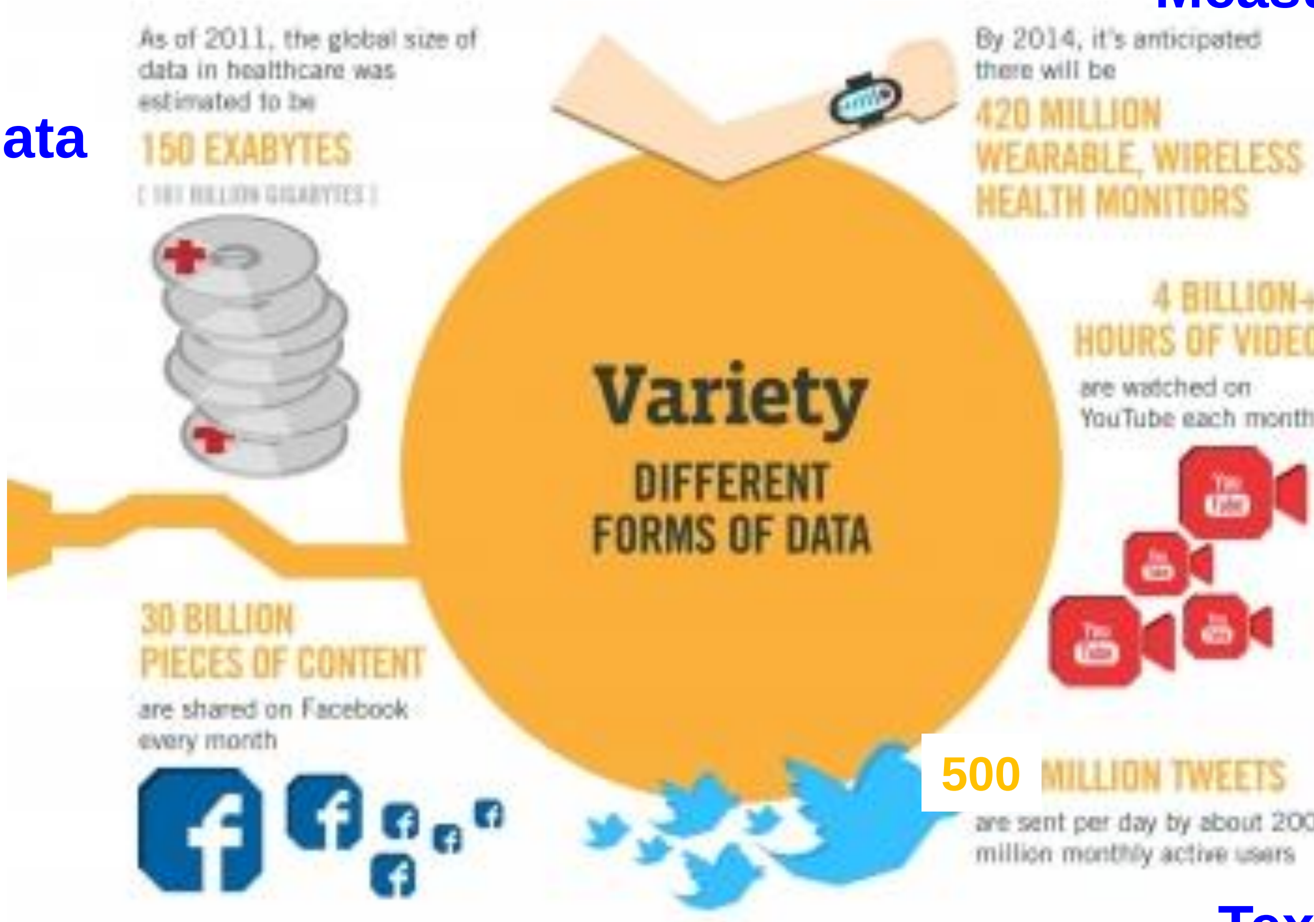
330 million of active Tweeter Users

Too *fast*: needs to be processed quickly and react promptly

Characteristics of Big Data: 3-Complexity (Variety)

Measurement data

Medical Imaging data



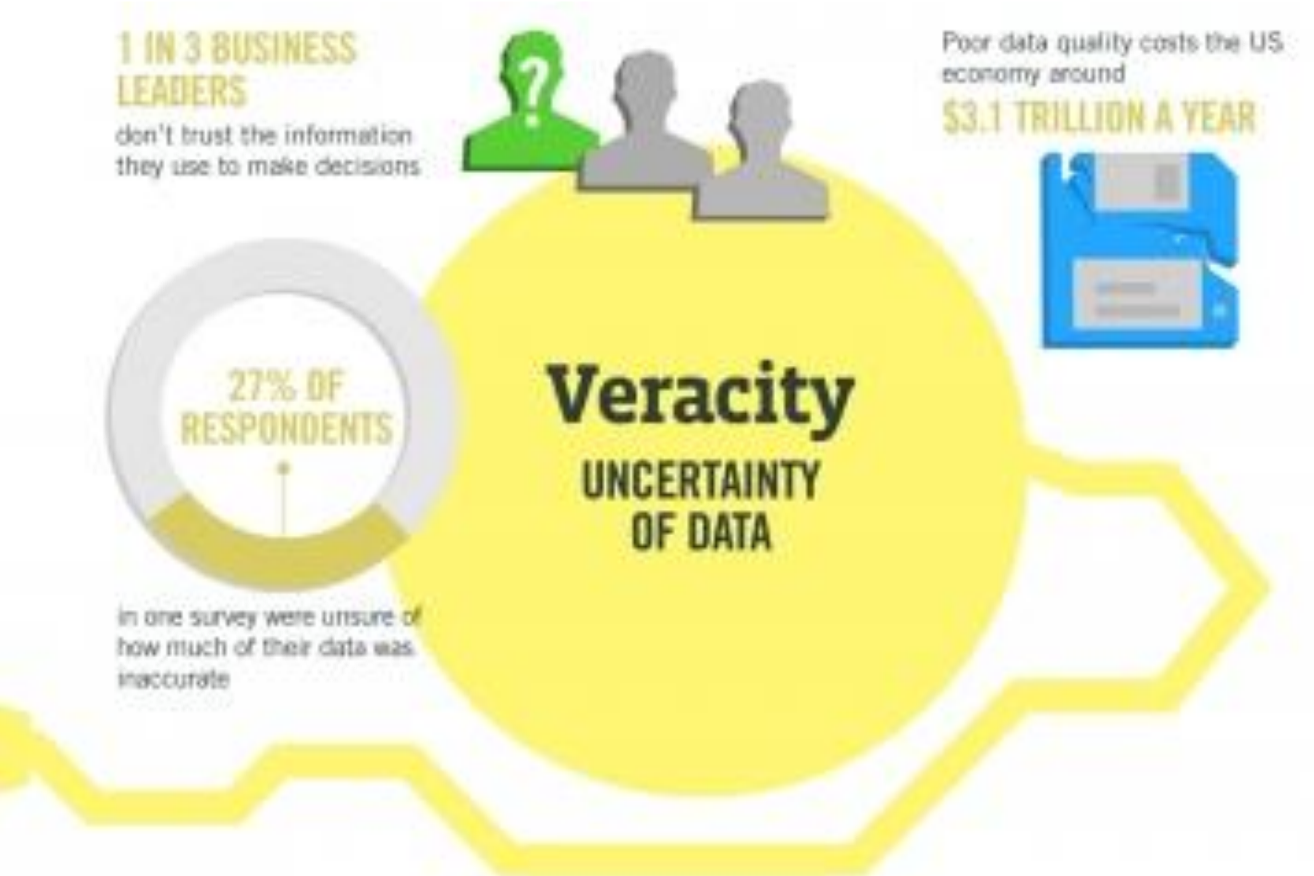
Video data

Textual data

Textual data

Too diverse: does not fit neatly in an existing tool

Characteristics of Big Data: 4-Quality (Veracity)

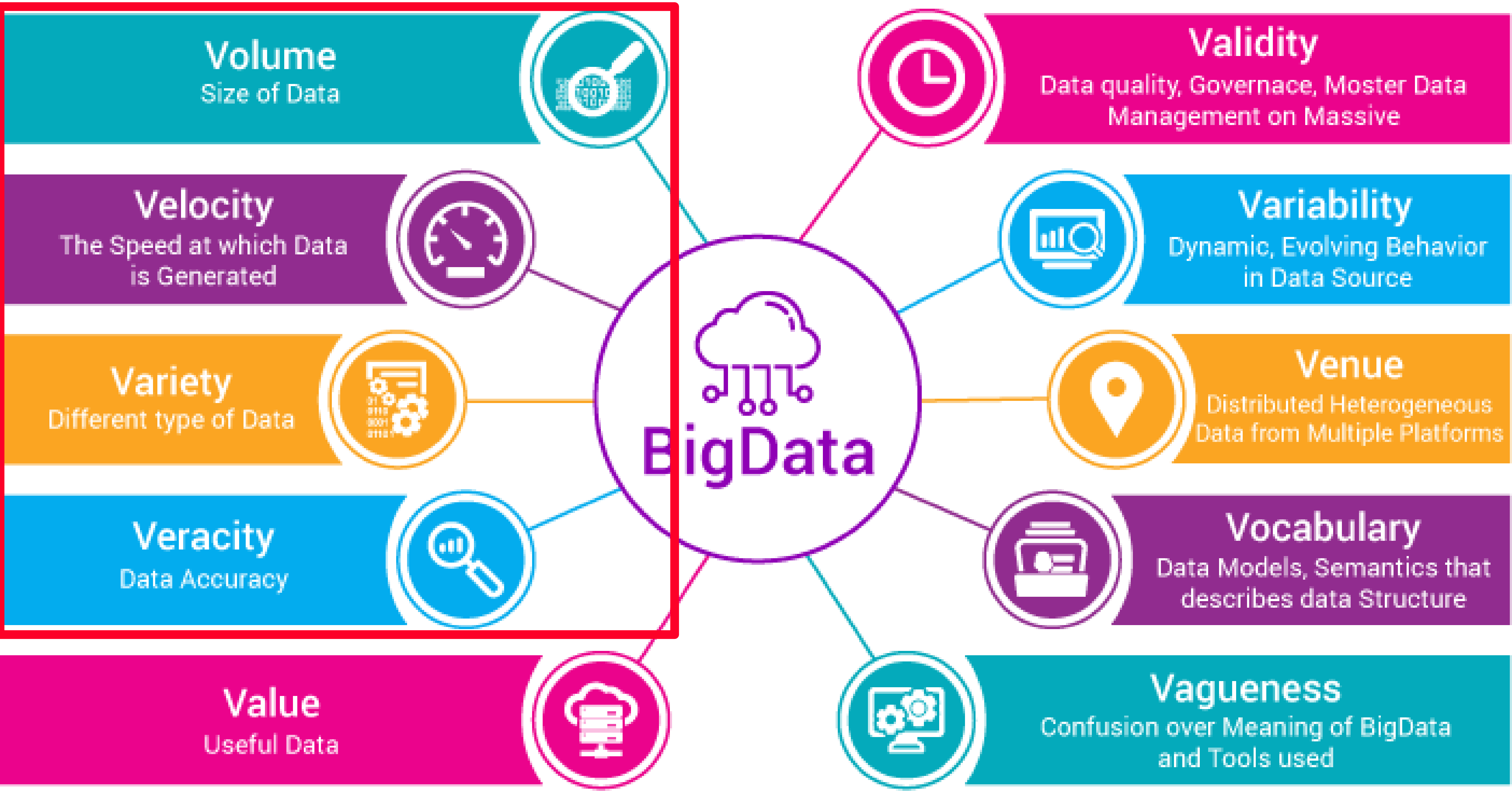


Many sources of online information:
are all these sources equally
- **accurate**
- **up-to-date**
- and **trustworthy**?

Too crappy: needs to assess their quality

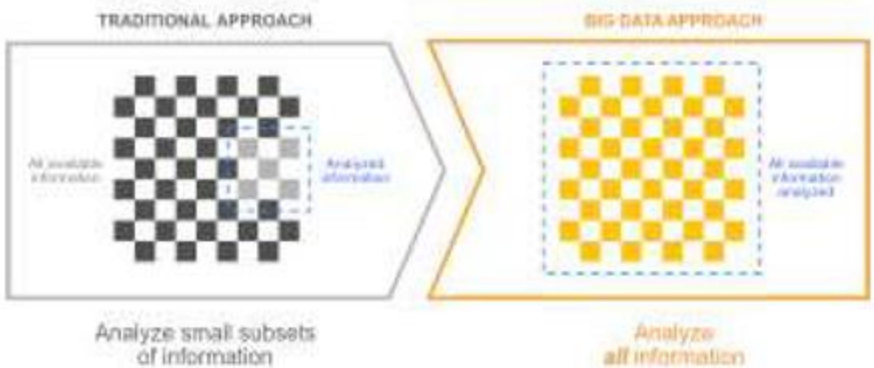
Other Big Data Qualities

most popular ones



A New Era of Data Analytics

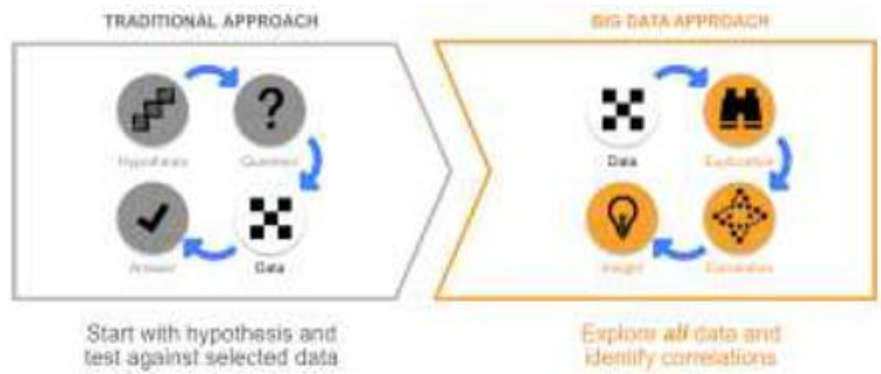
Look At All The Data



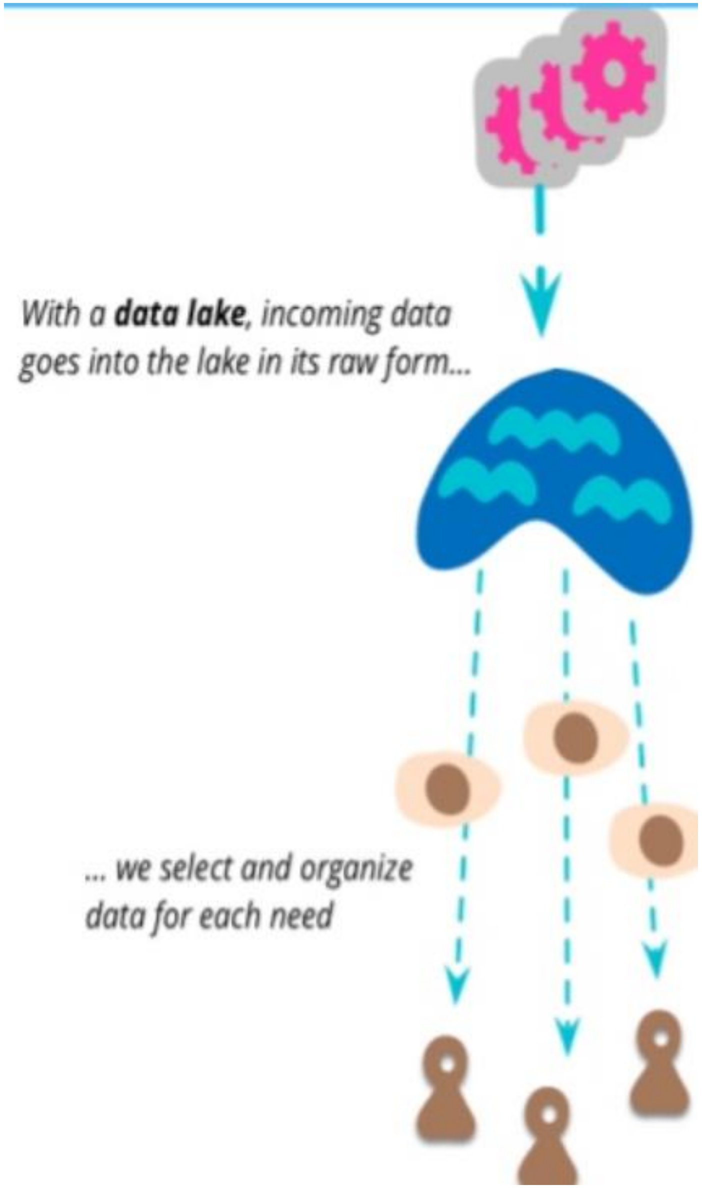
Look Even At Dirty & Noisy Data

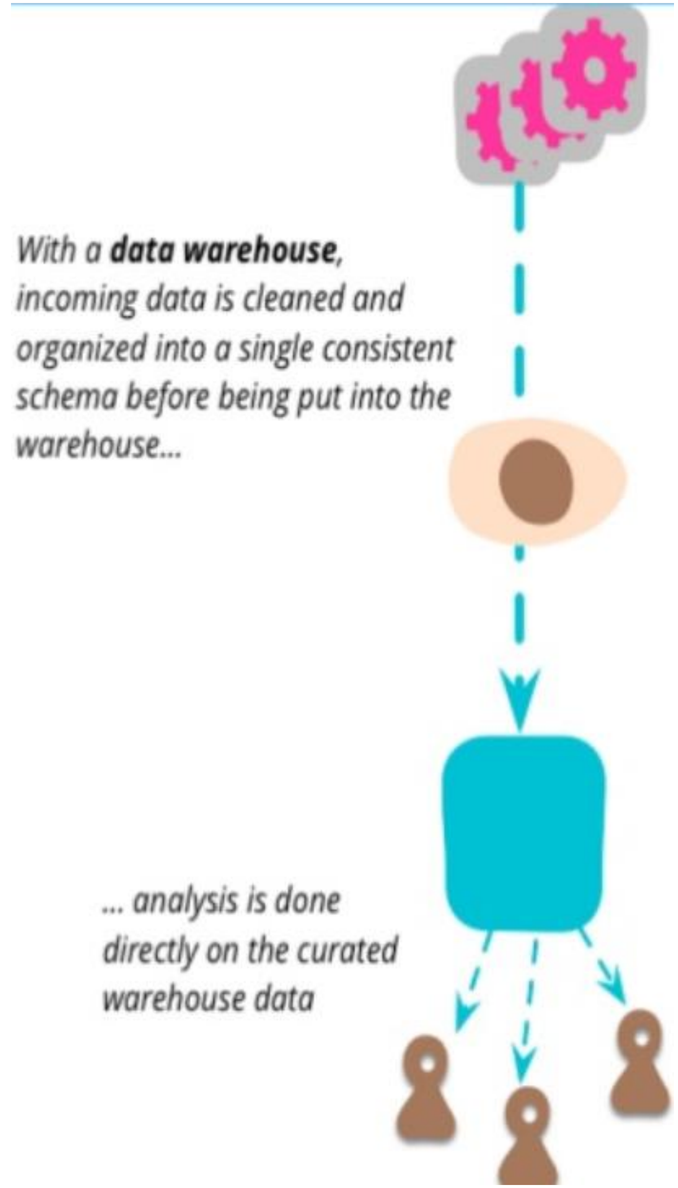
Leverage Data as it is Captured

Let Data Lead the Way



Data Lakes vs Data Warehouses

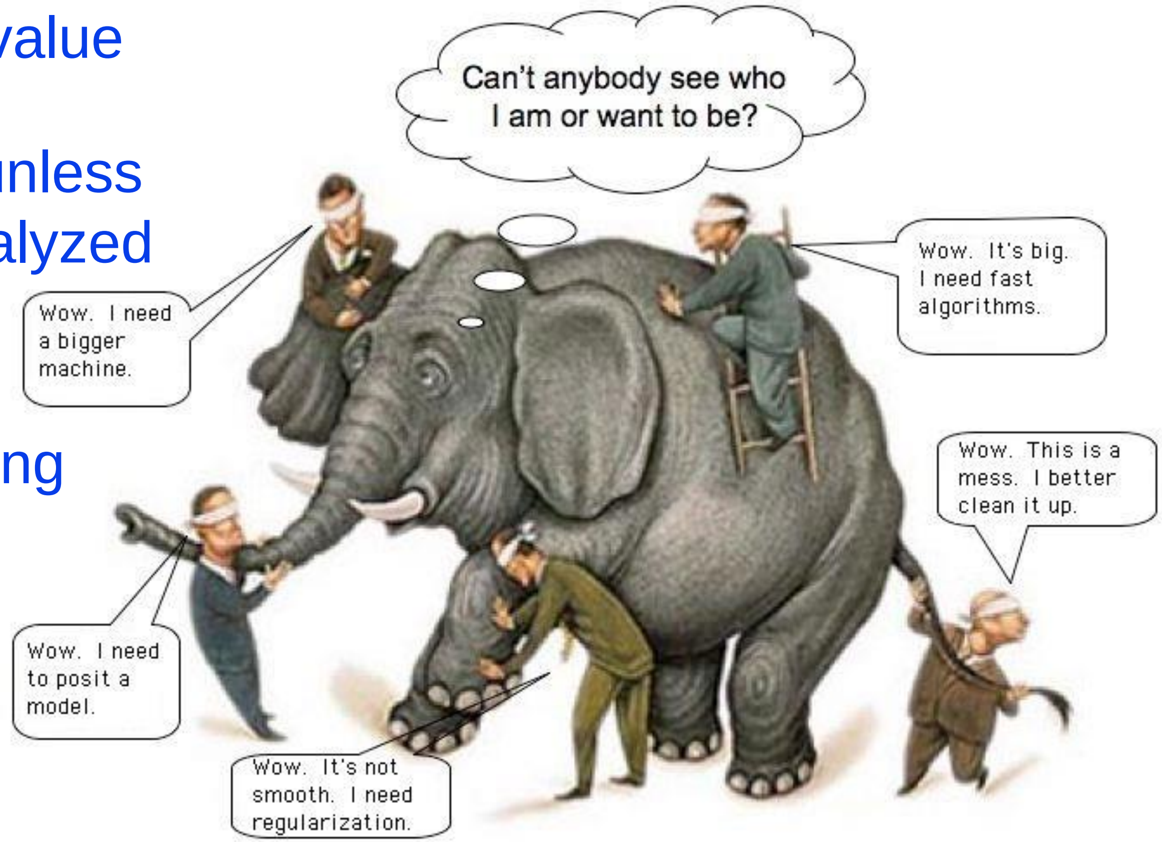




Big Data Mining

What to Do with Big Data?

- Data contains knowledge and value
- Nobody knows what's in data unless it has been processed and analyzed
- Data value for:
  - Faster, better decision making
  - Cost savings
  - New products and services

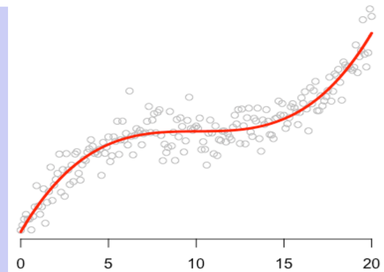
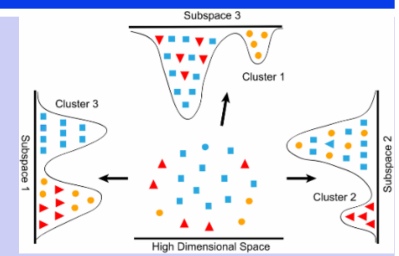
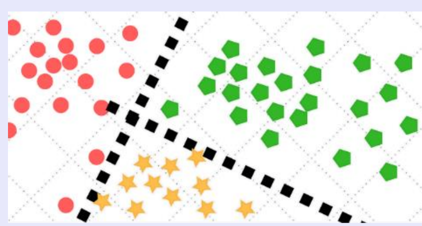
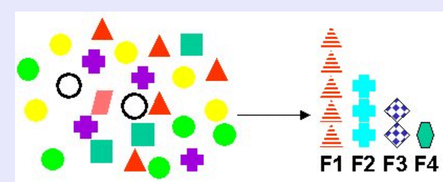


- Grand challenge for data science and engineering:
  - Empower a wide range of users to explore and obtain trustworthy, actionable insights from big data

Data Mining Methods

Predictive: Use some variables to predict *unknown* or *future* values of other variables

Descriptive: Find human-interpretable *patterns* that describe the data

	Supervised	Unsupervised
Continuous	**Regression**  **Predict the value of a continuous variable**	**Clustering & Dimensionality Reduction**  **Finds "natural" grouping of instances given unlabeled data**
Categorical	**Classification**  **Predict the label of an instance from pre-label (classified) instances**	**Frequent Patterns & Association Rules**  **Discover interesting co-occurrence relations between variables**

Data Analysis: ERP & CRM Examples

Who are our lowest/highest margin customers ?

What is the most effective distribution channel?

Who are my customers and what products are they buying?

What product prom- -otions have the biggest impact on revenue?

Which customers are most likely to go to the competition ?

What impact will new products/services have on revenue and margins?

Large-Scale, Real-World Analytics

Question	Method
How do I segment my customers?	K-means Clustering
How is product ownership distributed across customer segments?	SQL, Cumulative Distribution Functions
Does this product appeal to some segments more than others?	Log-likelihood
What new products should I offer my customers?	Cosine similarity, k-Nearest Neighbors, Matrix factorization
Which campaign is working better?	Mann-Whitney U Test
How do I target my marketing efforts towards customers most likely to churn?	Logistic Regression
What are my customers saying about the new product launch?	NLP, sparse vectors
How can I identify fraudulent activity?	Classification, Logistic Regression

Tools and Technologies for Big Data Steven Hillion V.P. Analytics EMC Data Computing Division 2011

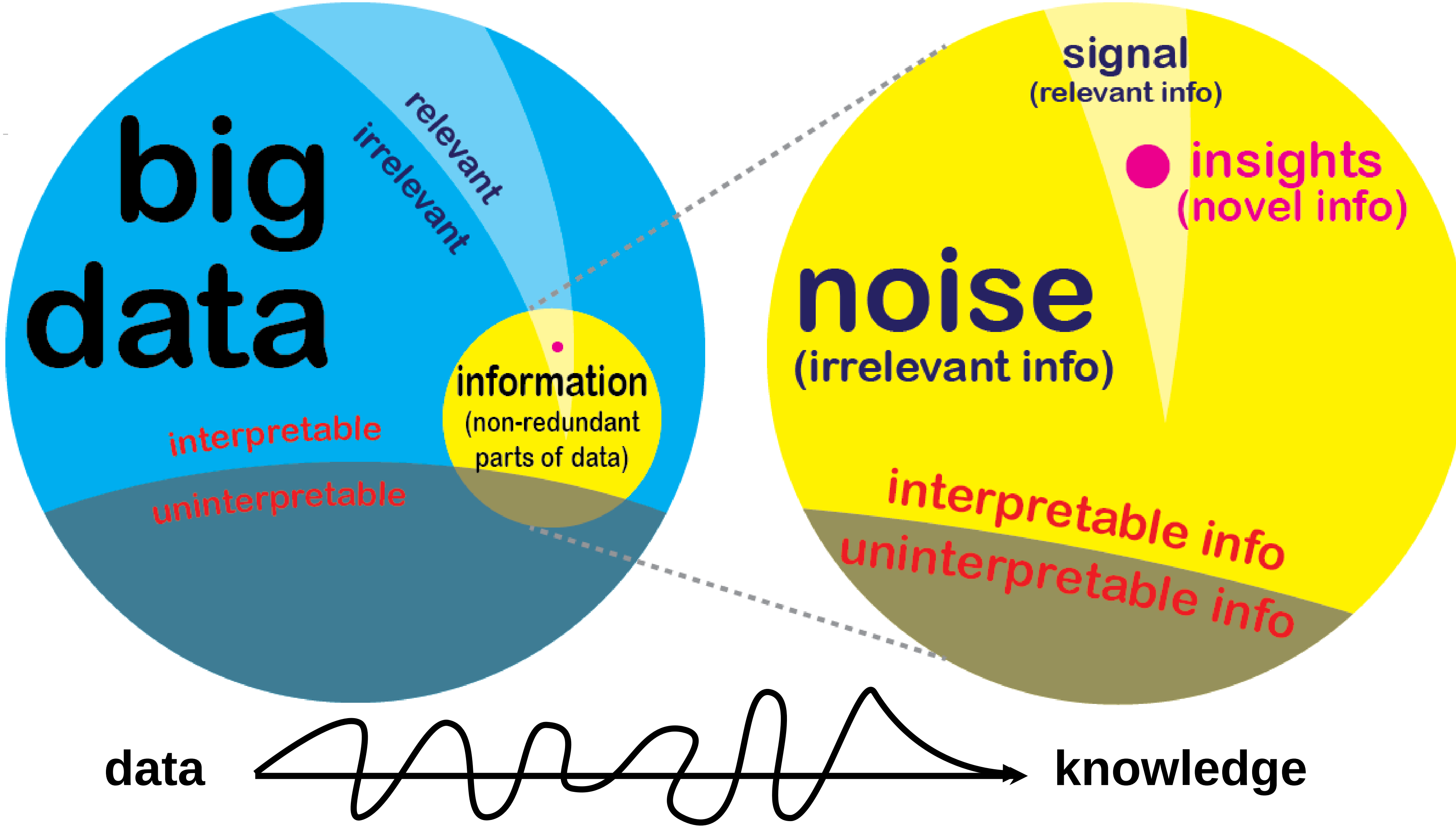
The WRONG Picture!

Big Data

Deep Insights

- Incorrect conclusions can lead to bad decisions

Big Data vs Deep Insights

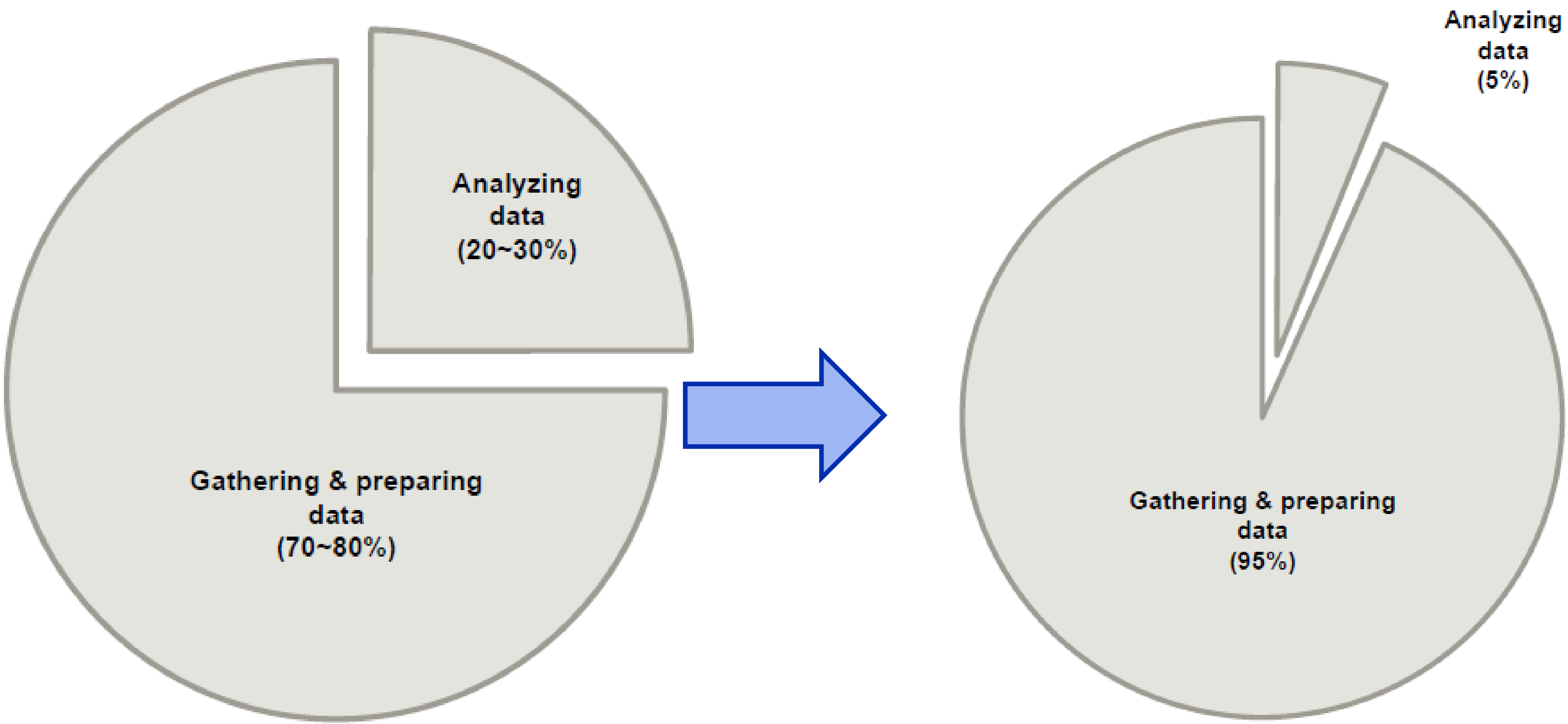


Data exploration is hard regardless of whether data are big or small !

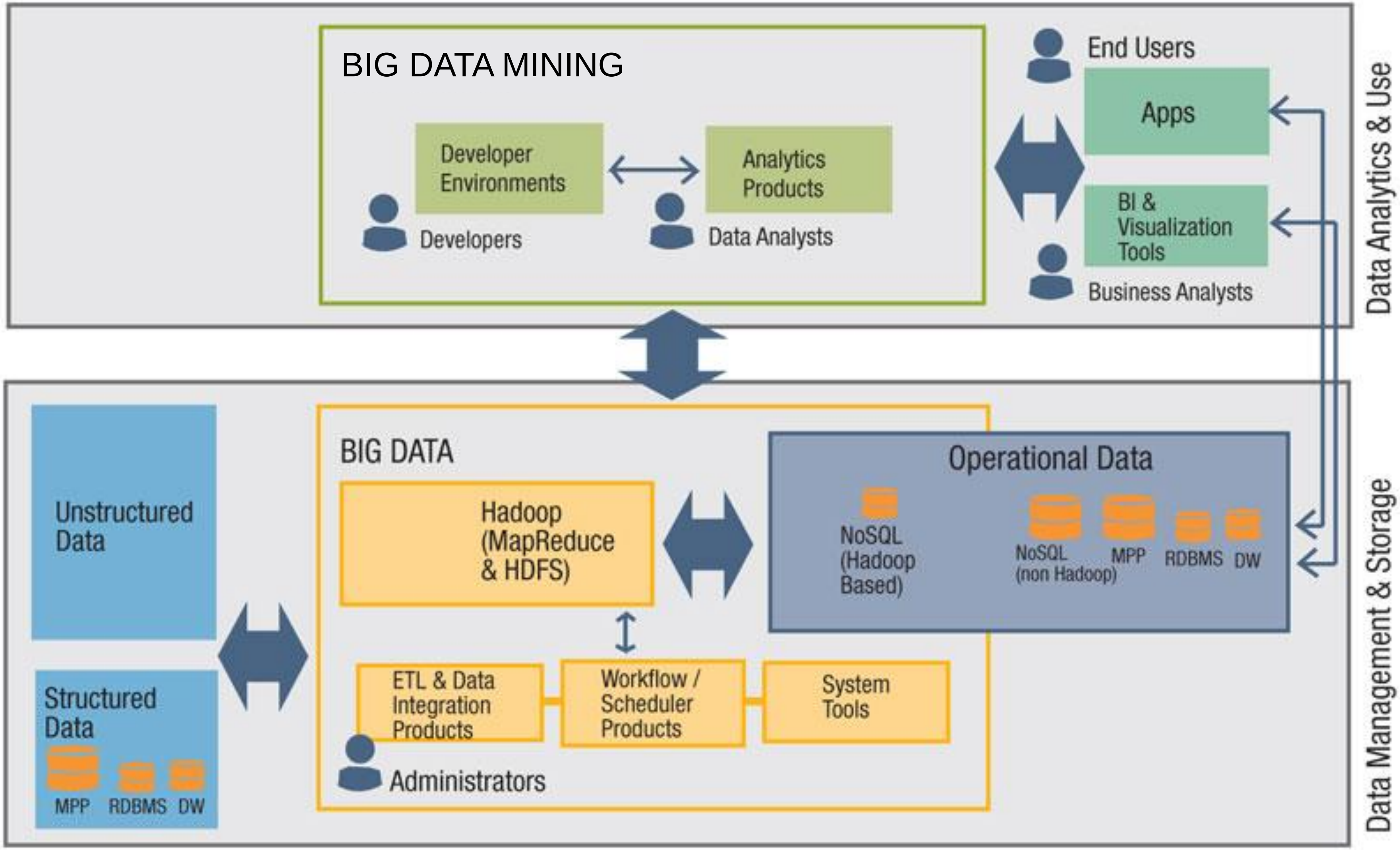
The TRUE Picture!

The time for developing an analysis (with small data)

The time for developing an analysis (with big data)

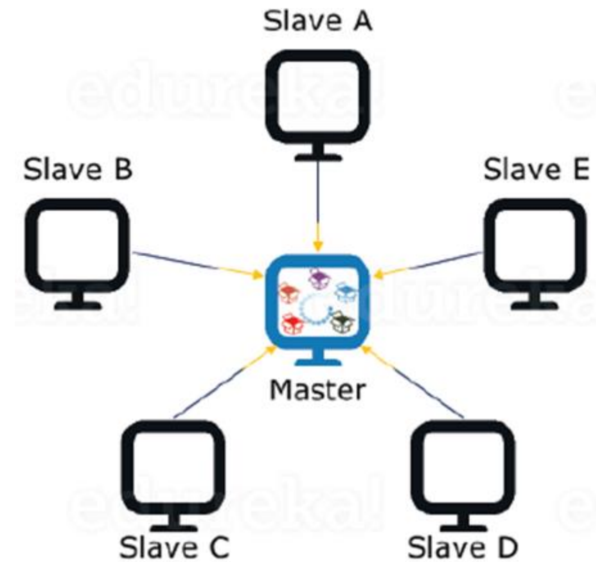


Big Data Processing & Analytics

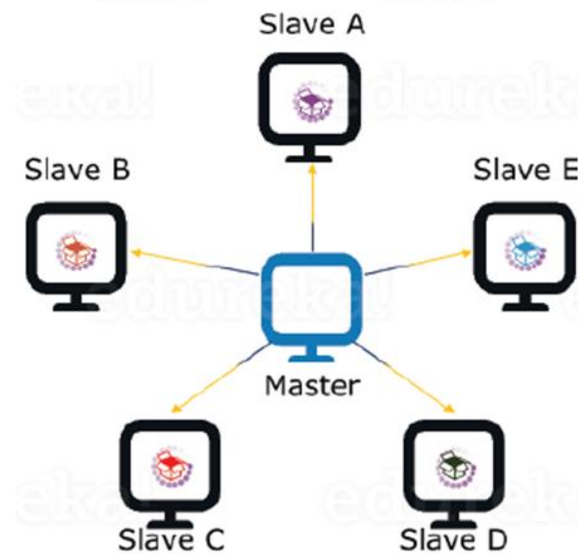


Traditional vs. Map/Reduce Approach

- Don’t move data to workers… Move workers to the data!
  - Store data on the local disks for nodes in the cluster
  - Start up the workers on the node that has the data local!

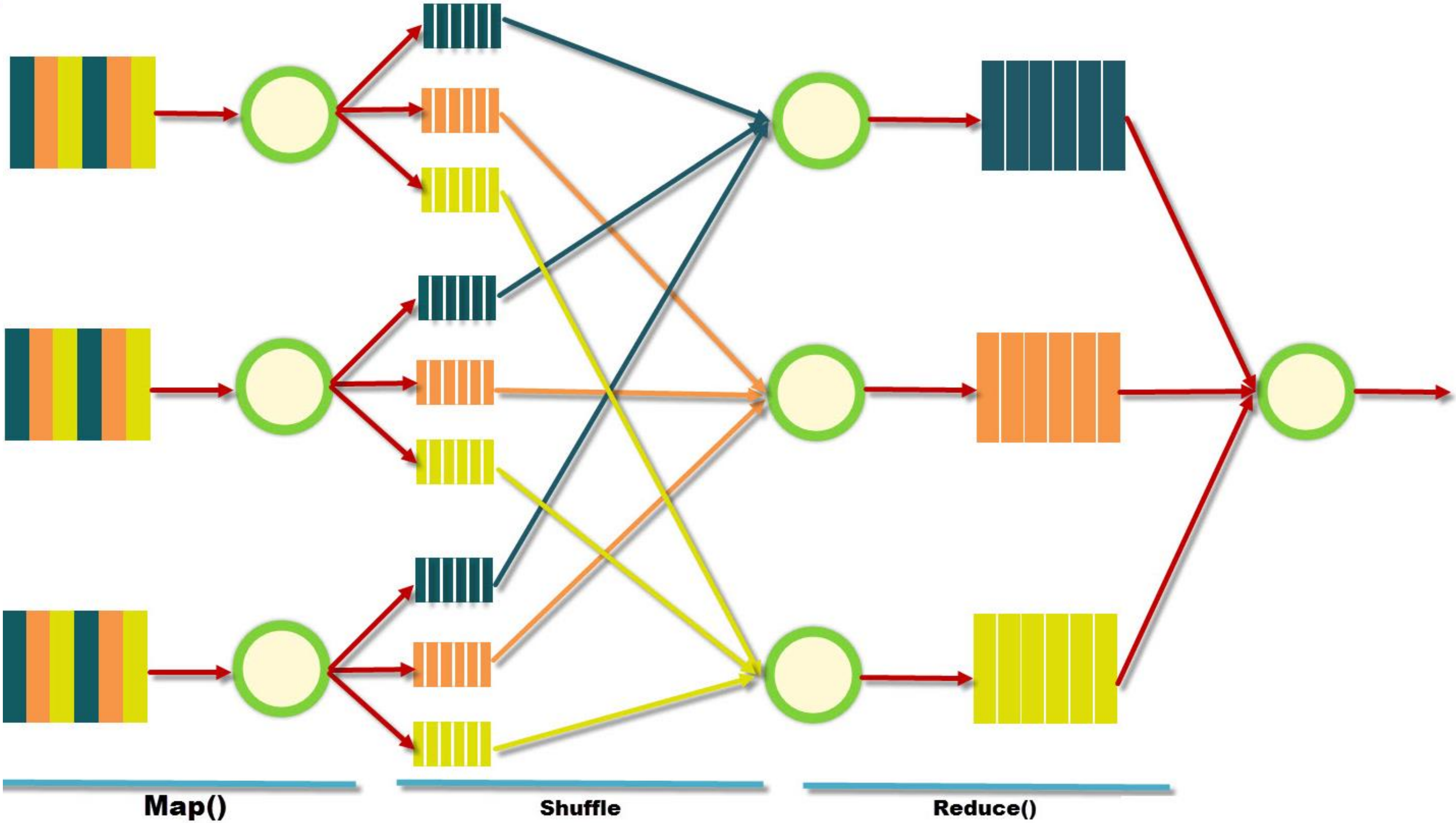


- Why?
  - Not enough RAM to hold all the data in memory
  - Common local-area network (LAN) speeds go up to 100 Mbps, which is about 12.5MB/s
  - Traditional hard disks provide a lot of storage and transfer speeds around 40-60MB/s

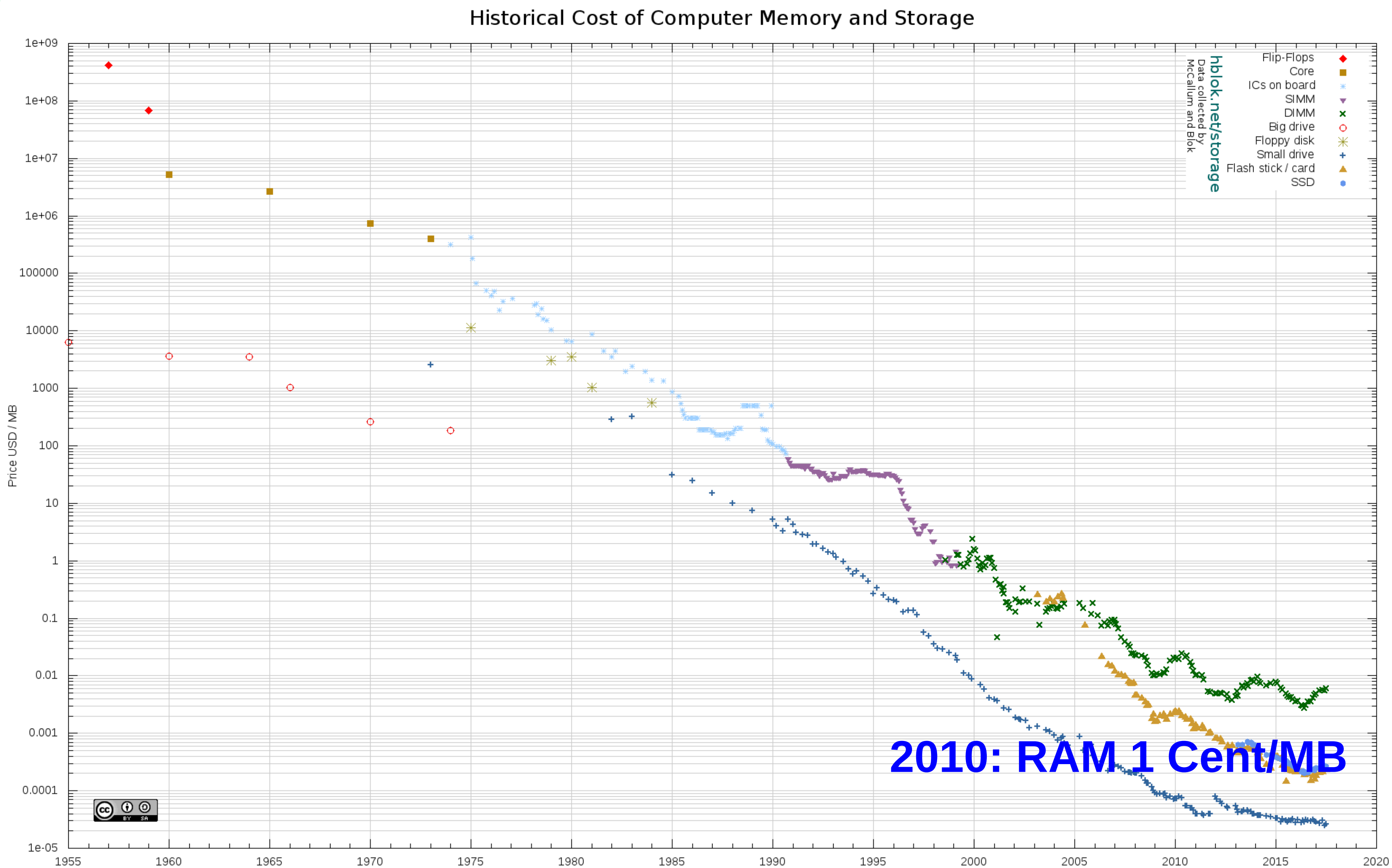


https://www.edureka.co/blog/mapreduce-tutorial

Analyzing Big Data using Map/Reduce



https://blog.sqlauthority.com



https://hblok.net/blog/posts/2017/12/17/historical-cost-of-computer-memory-and-storage-4/

What we Need to Make Sense of Big Data?

New Computing Frameworks:

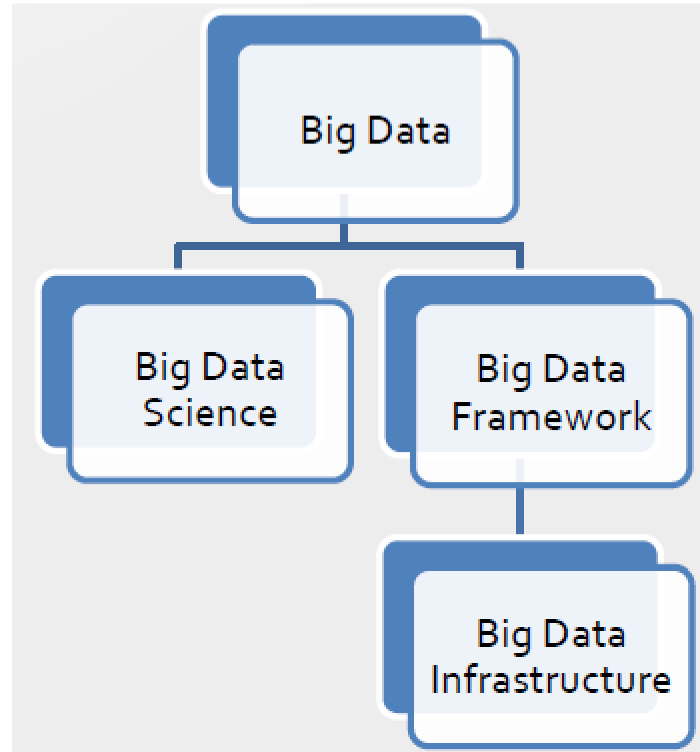
- Parallel/Distributed architectures: Cloud, HPC, MapReduce (Apache Hadoop, Spark), …
- Storage solutions: NoSQL, column stores, RDDs
- Processing Languages: Spark SQL, GraphX, Streaming, …

But also *new* Approaches/Algorithms!

- To *explore* and *process* big data
    - *integrate*, *curate*, *prepare*, …
- To *mine* data in Big Data frameworks

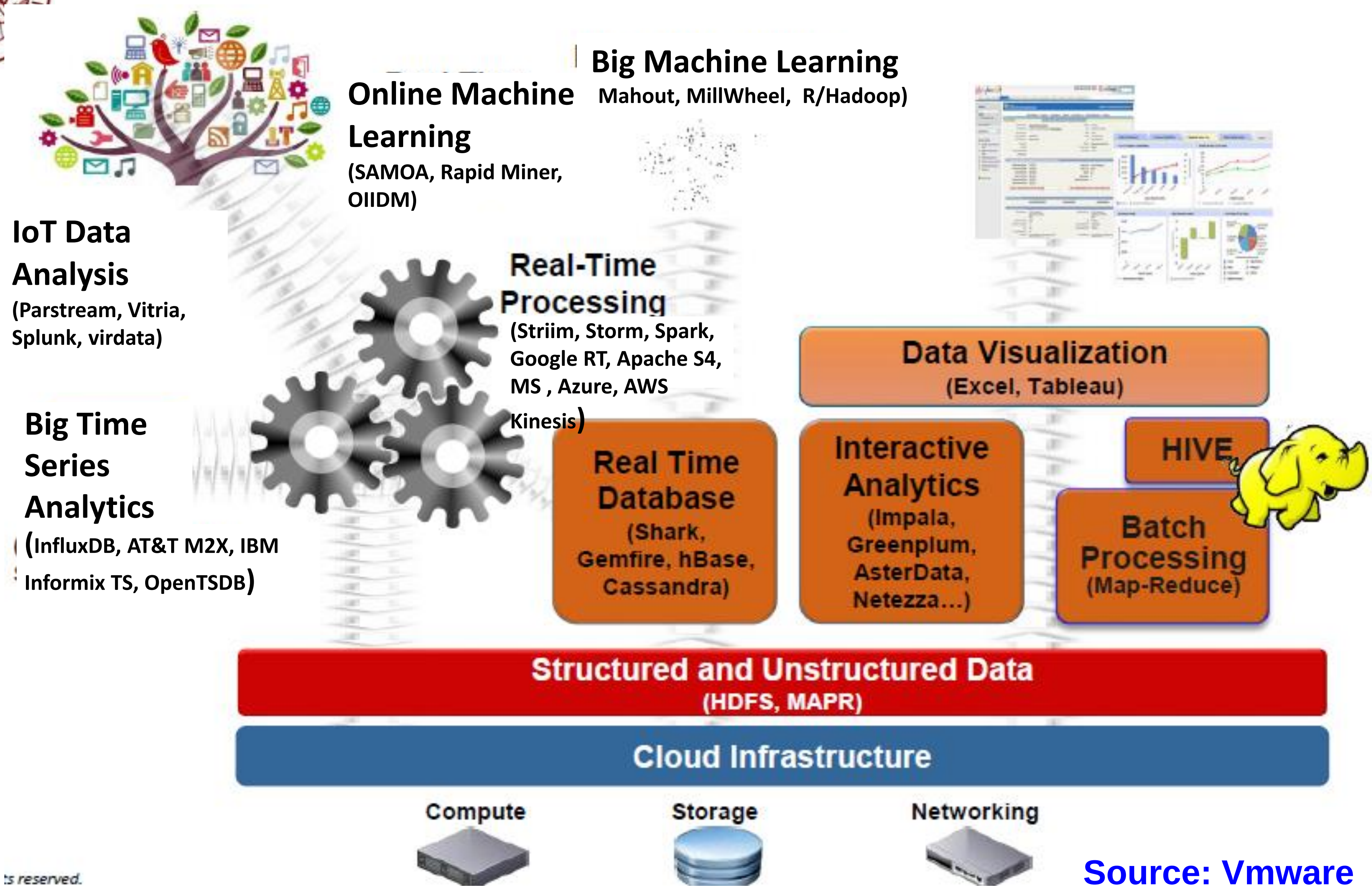
Several software libraries exist but there is *no one-size-fits-all solution*!

- often, you have to build your own…



M. Cooper & P. Mell Tackling Big Data NIST Information Technology Laboratory Computer Security Division

Big Data Processing & Analytics Platforms

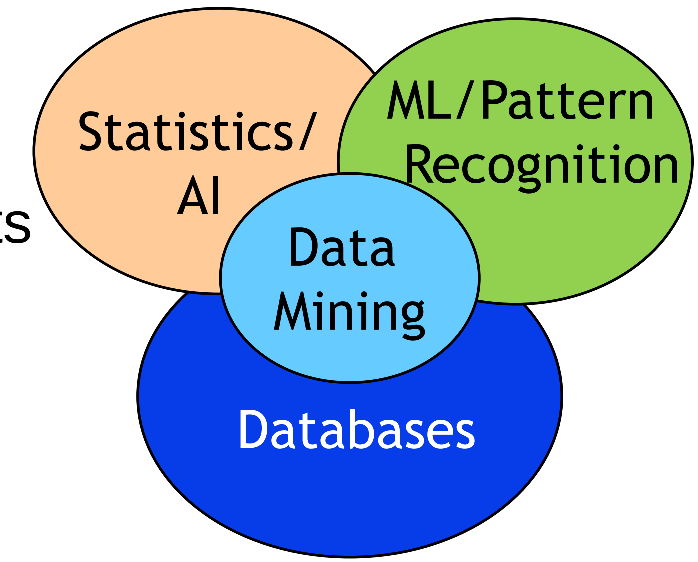


Source: Vmware

The Big Data Mining Mindset

- Data mining overlaps with:
  - Databases (DB): Large-scale data, simple queries
  - Machine Learning (ML): Small data, Complex models
  - Computer Science Theory: (Randomized) Algorithms

- Big Data urges for a cross-culture curriculum stressing on
  - Scalable Systems
  - Algorithmic Thinking
  - Computing Architectures
  - Automation for Handling Very Large Datasets



Big Data and its Relation to Statistics

- Statistical methods are the core of what Big Data is today
- A statistician will typically assume that datasets she/he deals with will fit into the main memory on a single machine
- Statistics extract most information from a very sparse and expensive to acquire typically small dataset
- However, now we move from a data poor regime to a data rich regime
- The goal is not anymore about new fancy mathematical method to squeeze more information from a small dataset
- The goal is now to about to build new engineering tools to process very large datasets
- Similarly like statisticians, visualization specialist are less concerned with massive datasets that span across hundreds/thousands of machines on the Internet

Big Data and its Relation to Business Intelligence (BI)

- BI aims at descriptive statistics with data with high information density to measure things, detect trends etc.
- Big Data targets inductive statistics with data with low information density whose huge volume allow to infer laws (regressions…)
- Software stack designed for BI is very specific and not very adaptable when requirements change
    - Data warehouse and specific dashboards and reports that consume data from the data warehouse in order to answer specific questions
- Software stack designed for BI is not applicable to Big Data problems where changing requirements is a norm
- BI engineers do not consumer their own products and make the decisions themselves, while Big Data analysts do

Big Data and its Relation to Data Engineering

- DB engineers and administrators posses a lot of skills to make them appropriate to Big Data tasks
- However, they are focused on a particular data model which is usually the relational one (columns and rows)
- Big data analysts deal with heterogeneous data sources that may include video, audio, text, graphs, images, structures and unstructured data, etc.
  - The relational data model may not be appropriate for some sources
- To a DB person, data mining is an extreme form of analytic processing – queries that examine large amounts of data
  - Result is the query answer
- However, to a ML person, data-mining is the inference of models – ML algorithms = “engine” to solve ML models
  - Result is the parameters of the model

Hadoop MR is not Suited to Iterative ML

- Typically we want to analyse a dataset by accessing data several times
  - Many trial-and-error steps, easy to get lost…
- Most existing data mining/ML methods were designed without considering data access and communication of intermediate results
  - They *iteratively* use data by assuming they are readily available

- Hadoop is not efficient at iterative programs
  - need *many map-reduce phases*
  - HDFS disk I/O becomes bottleneck!

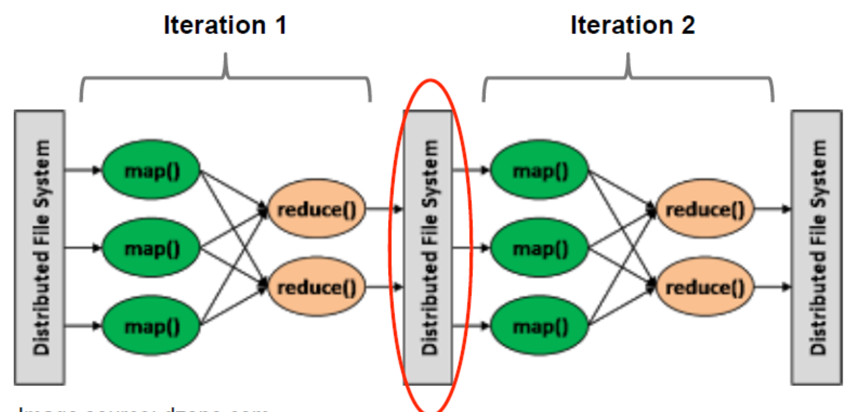


Image source: dzone.com

HDFS Bottleneck

MapReducable?

	One Iteration	Multiple Iterations	Not good for MapReduce
Clustering	Canopy	KMeans	
Classification	Naïve Bayes, kNN	Gaussian Mixture	SVM
Graphs		PageRank	
Information Retrieval	Inverted Index	Topic modeling (PLSI, LDA)	

- One-iteration algorithms are perfect fits
- Multi-iteration algorithms are OK, but…
  - a small amount of data has to be synchronized across iterations (typically via the file system)
- Some Algorithms are not Good for the MapReduce computing paradigm
  - when a large amount of data has to be synchronized across iterations

The Big ML Research

- Roughly there are two types of approaches
  - Parallelize existing (single-machine) algorithms (data, model, hybrid)
  - Design new algorithms particularly for massively parallel settings
  - of course there are things in between
- To have technical breakthroughs in big-data analytics, we should know both algorithms and systems well, and consider them together

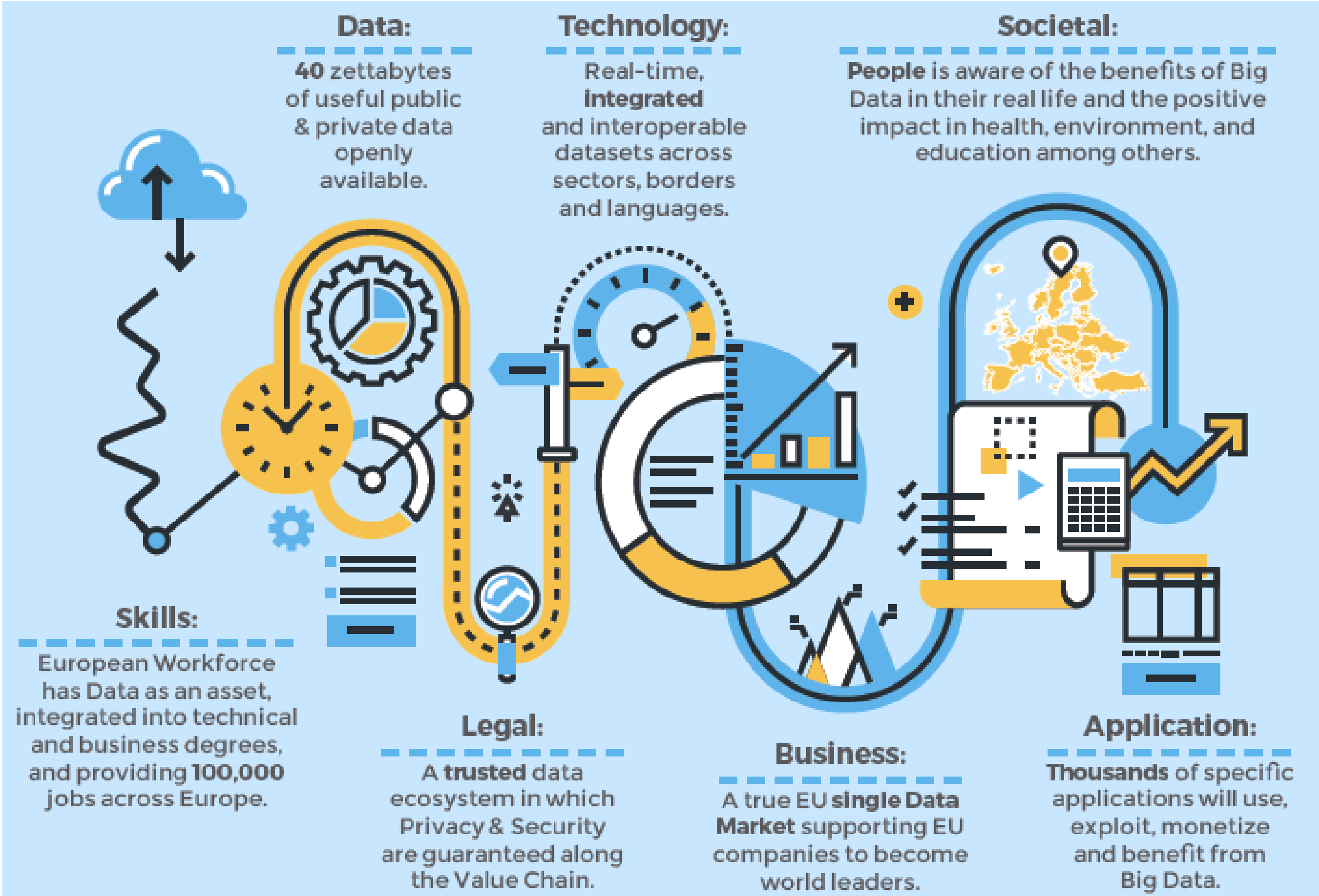
E. P. Xing, Q. HoA, Big ML Software for Modern ML Algorithms, IEEE Big Data 2014

References

- CS246: Mining Massive Datasets. Jure Leskovec, Stanford University 2020
- CS9223 – Massive Data Analysis. J. Freire & J. Simeon, New York University 2013
- CS6240: Parallel Data Processing in MapReduce. Mirek Riedewald, Northeastern University 2014
- Big Data Infrastructures: Exploiting the Power of Big Data. T. Sellis School of CS & IT Athens 2015
- CS525: Special Topics in DBs Large-Scale Data Management Advanced Analytics on Hadoop. Mohamed Eltabakh, Spring 2013
- Big-data Analytics: Challenges and Opportunities. Chih-Jen Lin, National Taiwan University 2014
- Knowledge Discovery and Data Mining. Evgueni Smirnov, Maastricht University 2013

Questions?

Big Data Value Vision for 2020



www.ijcai-18.org/wp-content/uploads/2018/07/3_BDVA_IJCAI_July-2018-LLB.pdf