Fall 2019

Big Data Processing and Analytics

Vassilis Christophides

christop@csd.uoc.gr

http://www.csd.uoc.gr/~hy562

University of Crete, Fall 2019

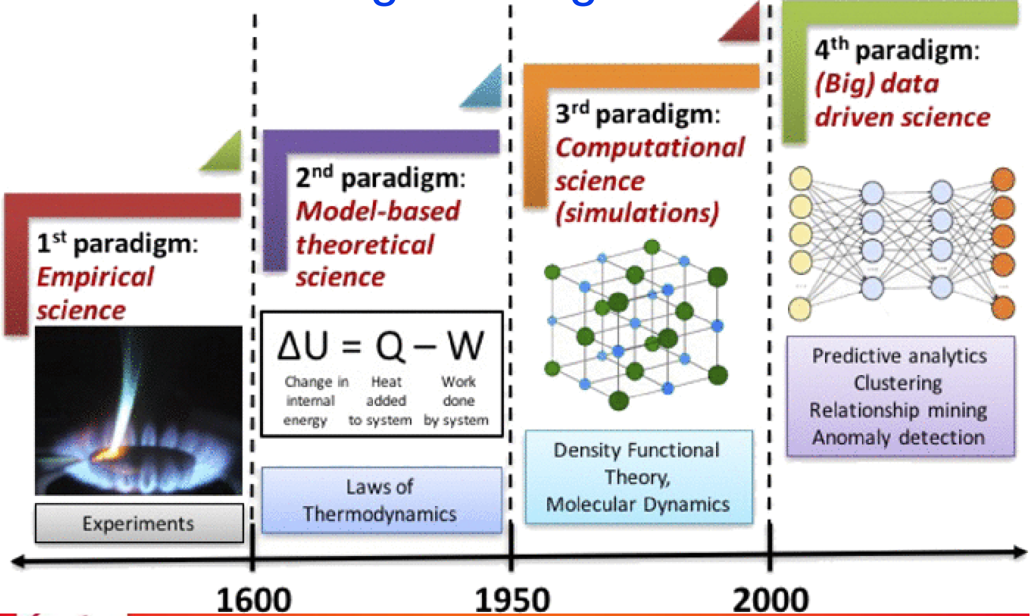
1

Fall 2019

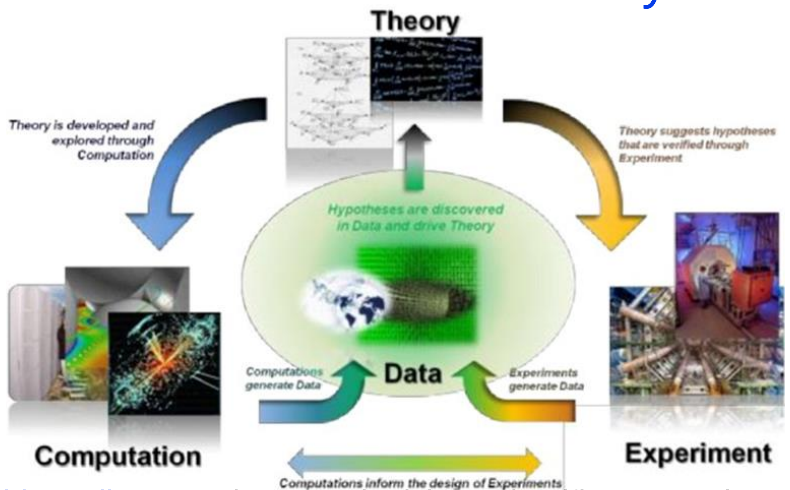
The Data Avalanche: From Science to Business

2

Shifting Paradigm in Sciences

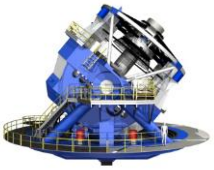


Data-driven Discovery



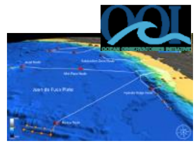
- Data-driven discovery is revolutionizing scientific exploration as well as engineering innovations
  - From hypothesis driven to hypothesis generating

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



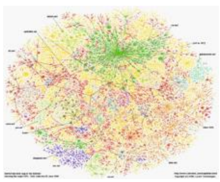
Astronomy: LSST

Physics: LHC



Oceanography

Biology: Sequencing



Sociology: The Web

Precision Medicine

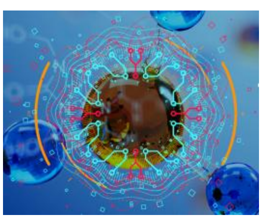
Neuroscience: EEG, fMRI

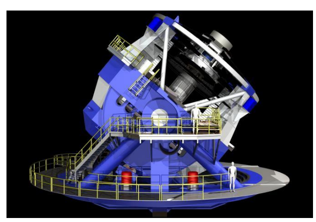
Sports

- 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

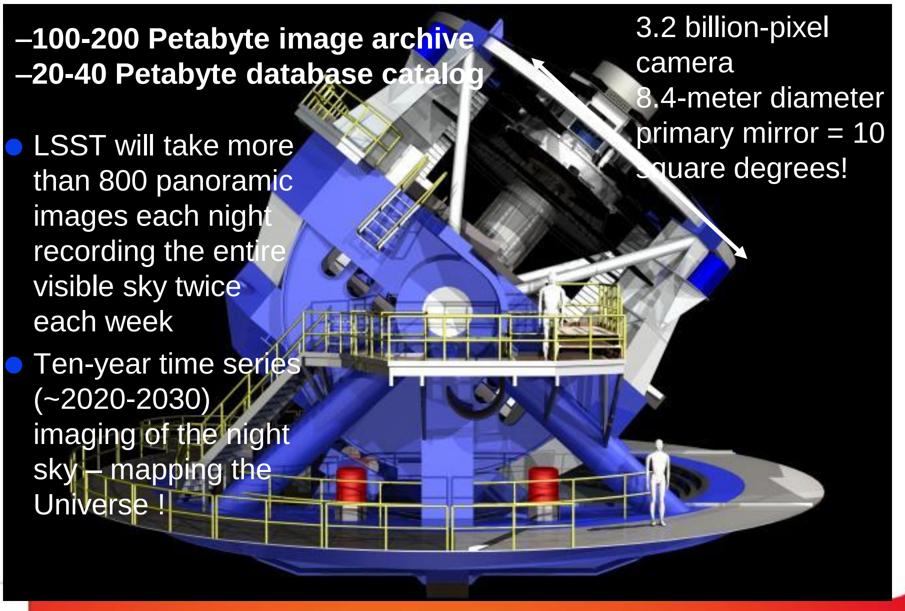




- Simulation Data: Increasing level of simulation detail and duration, as well as, model size by orders of magnitude!
- Experimental Data: Light sources, genome sequencing, next generation ARM radars, sky surveys, neuro-sensing and stimulation, …

- 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)



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

- 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 !

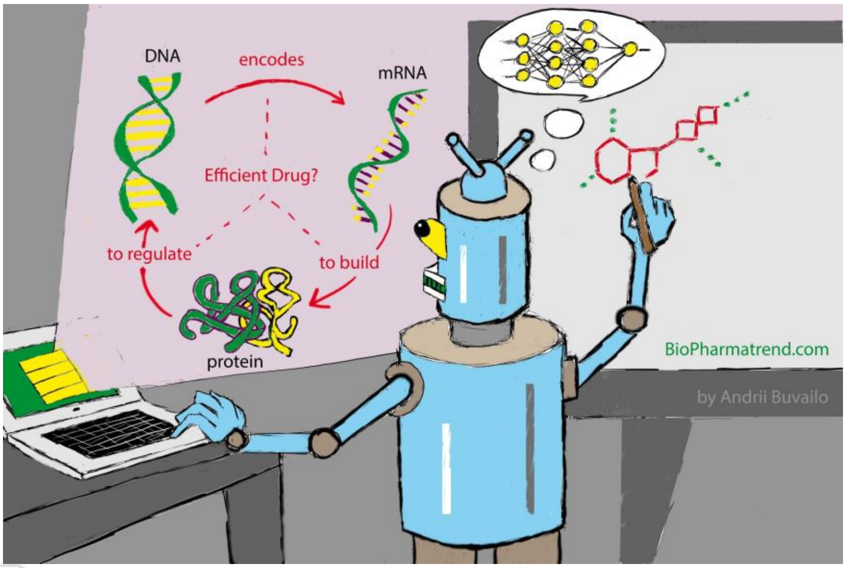
3.2 billion-pixel camera
8.4-meter diameter primary mirror = 10 square degrees!

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. Bauman posing with 5 petabytes of data necessary to image a black hole

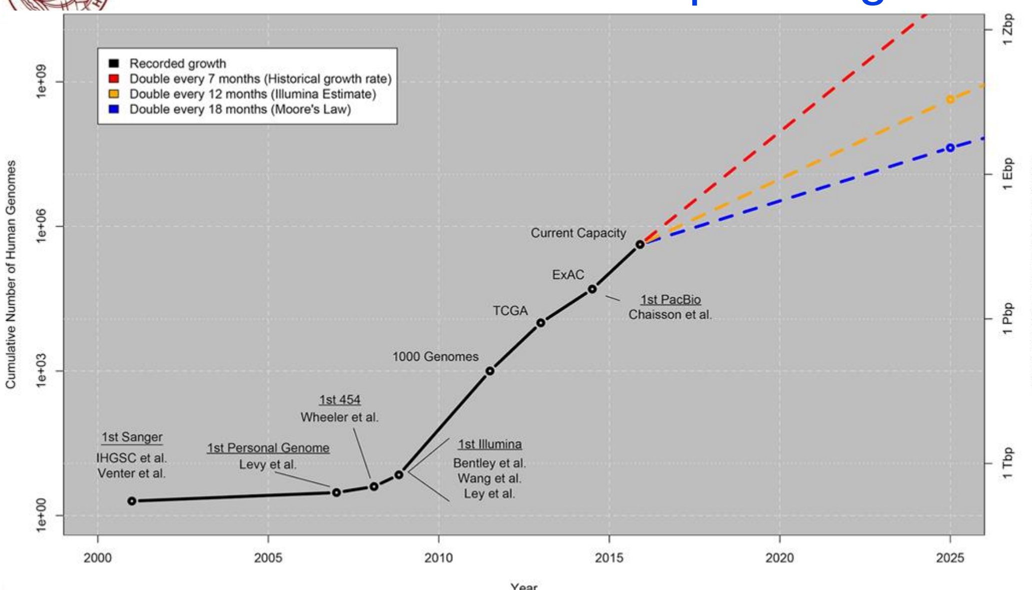
https://www.facebook.com/BusinessInsiderScience/videos/378897386038615

AI is Changing Drug Discovery!



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

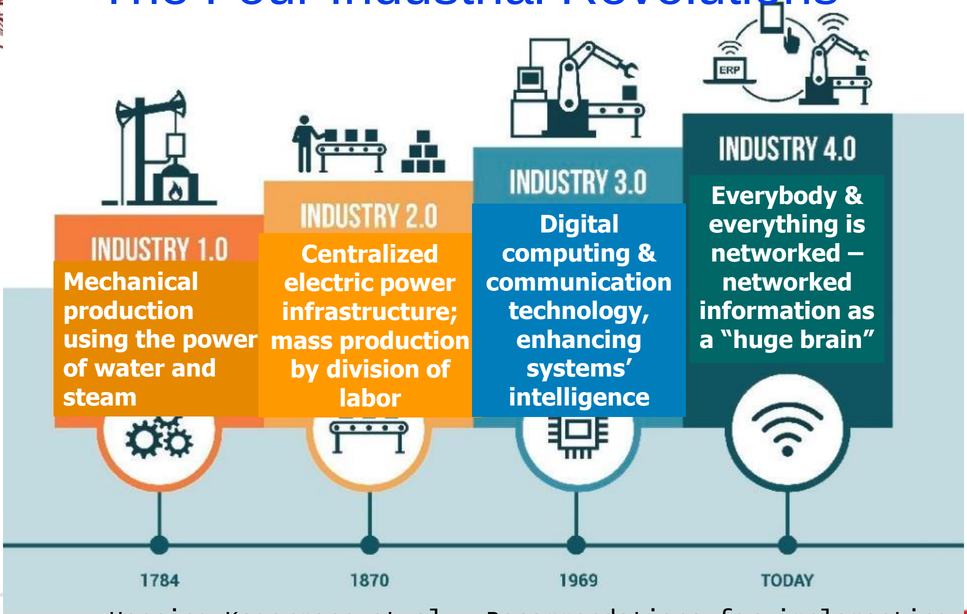
Growth of DNA Sequencing



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

Fall 2019

The Four Industrial Revolutions



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

15

Fall 2019

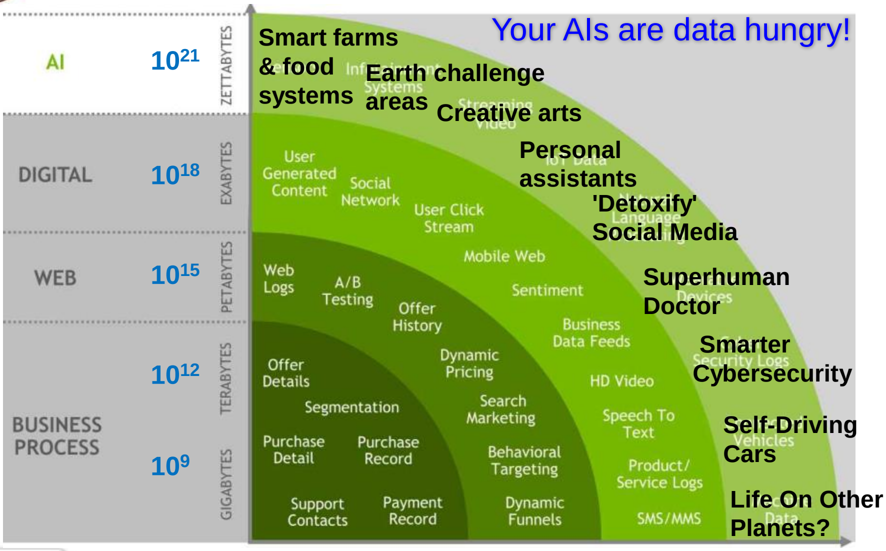
Digital Disruption Already Happening !

- Largest telco companies owns no telco infrastructure (Skype)
- World's largest movie houses 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)
- Faster growing banks have actually no money (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

17

The Data Tsunami: Transactions + Interactions + Observations



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

Driving Innovation with Big Data

Progress and Innovation is 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*

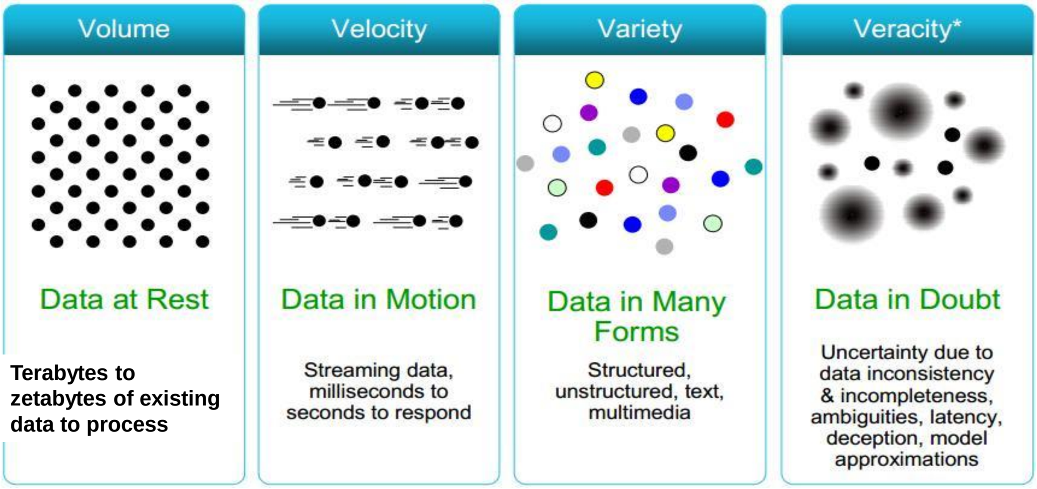
What Makes Data, “Big” Data?

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)

8

The Four V's of Big Data



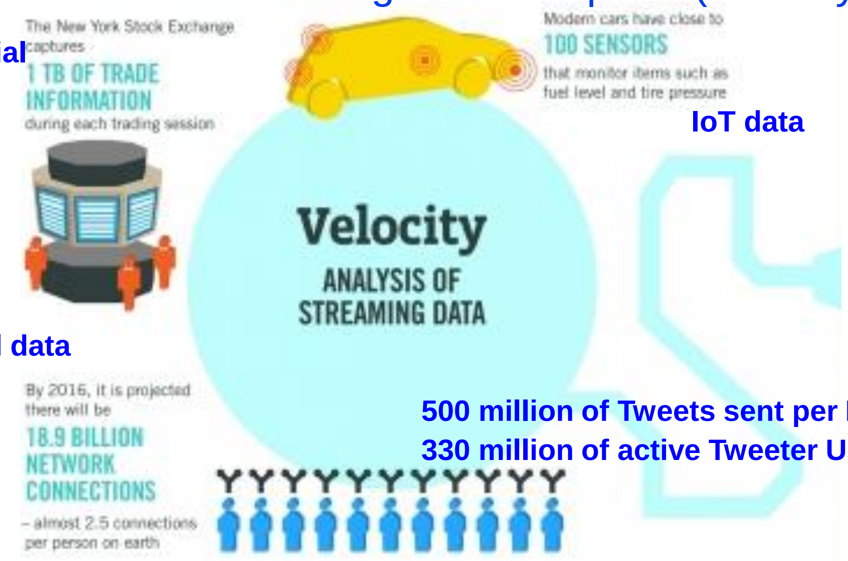
www-05.ibm.com/fr/events/netezzaDM_2012/Solutions_Big_Data.pdf

Characteristics of Big Data: 1-Scale (Volume)

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

Fall 2019

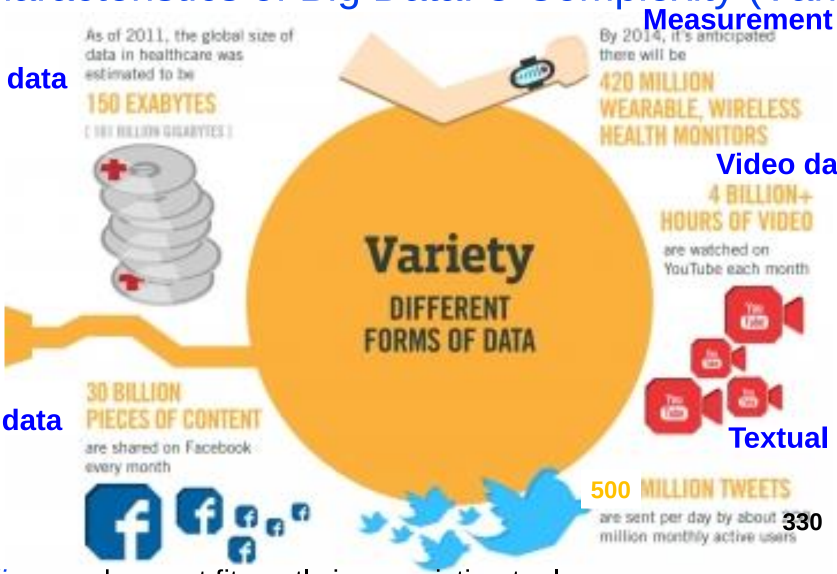
Characteristics of Big Data: 2-Speed (Velocity)



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

Fall 2019

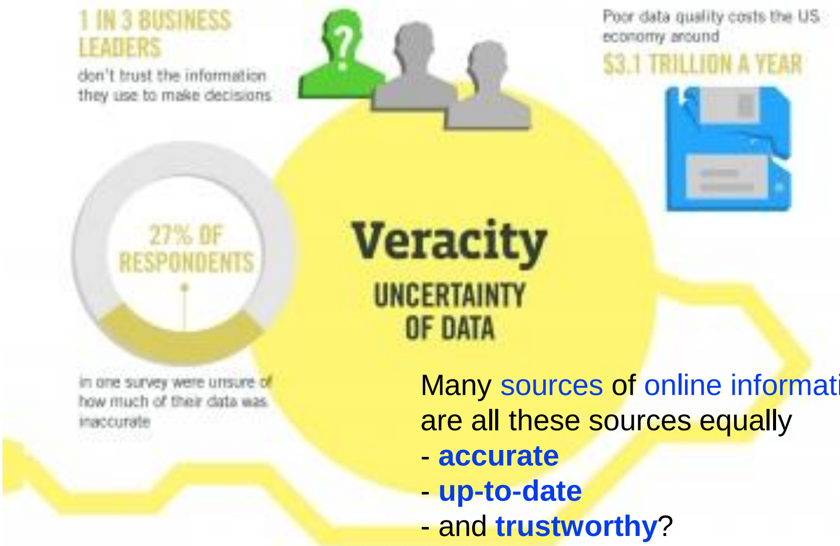
Characteristics of Big Data: 3-Complexity (Variety)



- *Too diverse*: does not fit neatly in an existing tool

Fall 2019

Characteristics of Big Data: 4-Quality (Veracity)



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

- *Too crappie*: needs to assess their quality

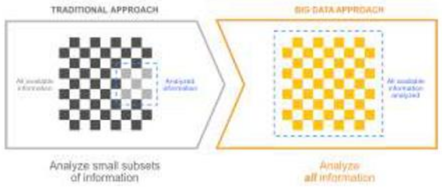
Fall 2019

Big Data Characteristics & Challenges

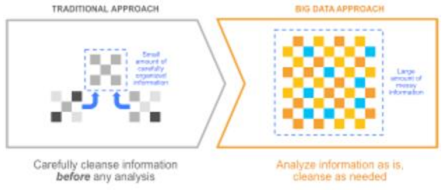
Characteristic	Description	Challenges	Root Cause
Volume	**Over increasing amount** of data that must be *ingested*, *processed* & *analyzed.* A single machine can not manage large volumes of data efficiently.	Constantly scale hardware and software infrastructure to accommodate very large storage spaces	High number of data sources High resolution sensors
Velocity	**Fast data** is being **ingested** and need to be **transformed into insight at a high speed.**	Support streaming/ online pprocessing & Real-time Analytics	High-rate data acquisition, low cost of hardware
Variety	Degree of **diversity** (in terms of **content** and **structuring**) of data from sources both inside and outside an organization	Cope with Multi-Modality, Complex interactions and Implicit Semantics	Social media Scientific data Video M2M / IoT
Veracity	**Quality** and **trustworthiness** of data	Curate data for missing, duplicate, erroneous values, enchase data traceability	Crowd data production, Human & Machine Sensing

We've Moved into a New Era of Data Analytics

Look At All The Data

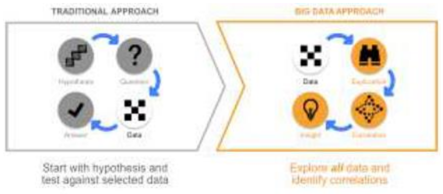


Look Even At Dirty & Noisy Data



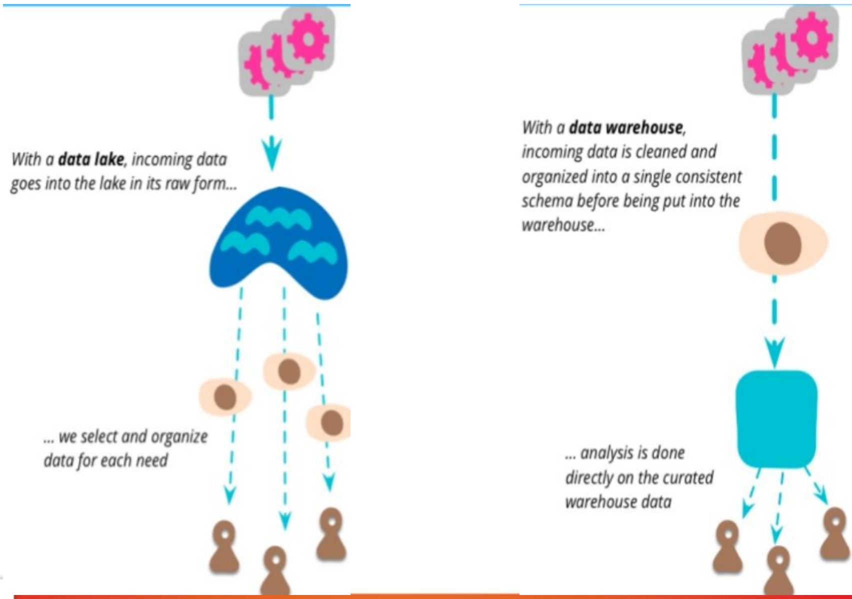
Leverage Data as it is Captured

Let Data Lead the Way



@ 2014 IBM Corporation

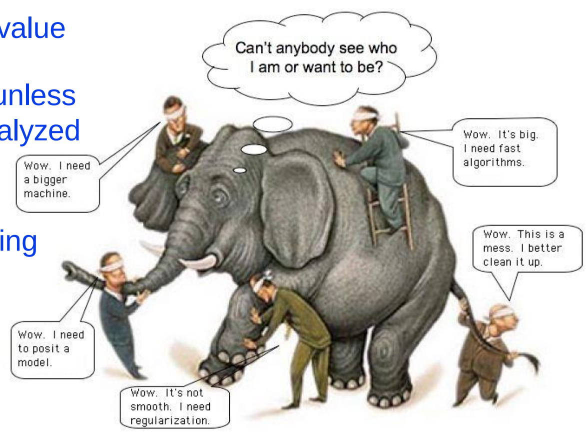
Data Lakes vs Data Warehouse



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

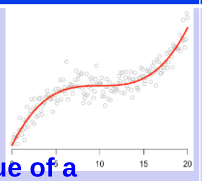
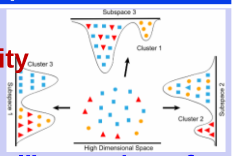
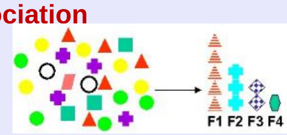


- Grand challenge for data science and engineering:
  - ◆Empower a wide range of users to explore and obtain trustworthy, actionable insights from big data

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

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

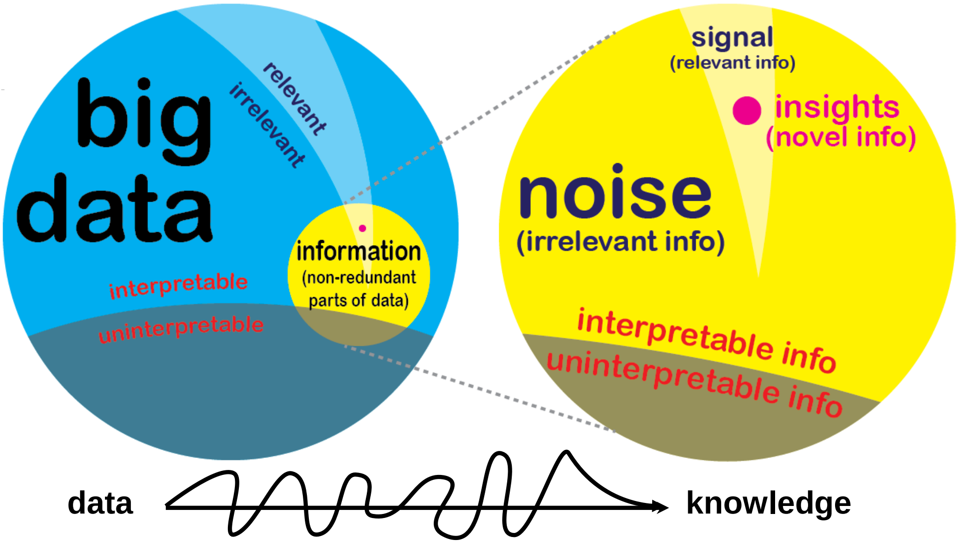
The WRONG Picture!

Big Data

Deep Insights

- Incorrect conclusions can lead to bad decisions

Big Data vs Deep Insights



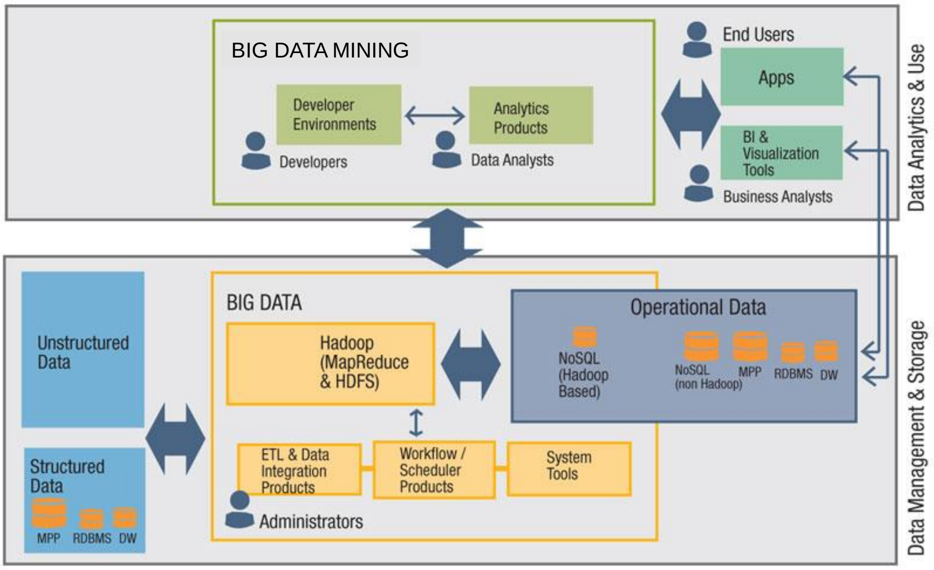
Data exploration is hard regardless of whether data are big or small !

The TRUE Picture!

The time for developing an analysis (with small data)

The time for developing an analysis (with big data)

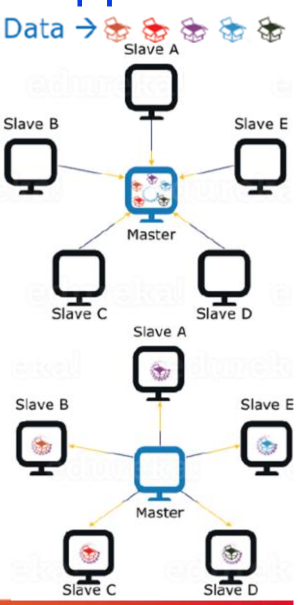
Big Data Processing & Analytics



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 Mbit/s, which is about 12.5MB/s
  - ◆Traditional hard disks provide a lot of storage and transfer speeds around 40-60MB/s

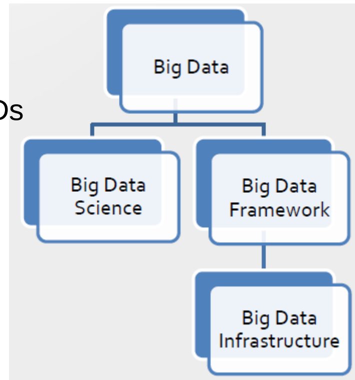


https://www.edureka.co/blog/mapreduce-tutorial

What we Need to Make Sense of Big Data?

New Computing Frameworks:

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



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

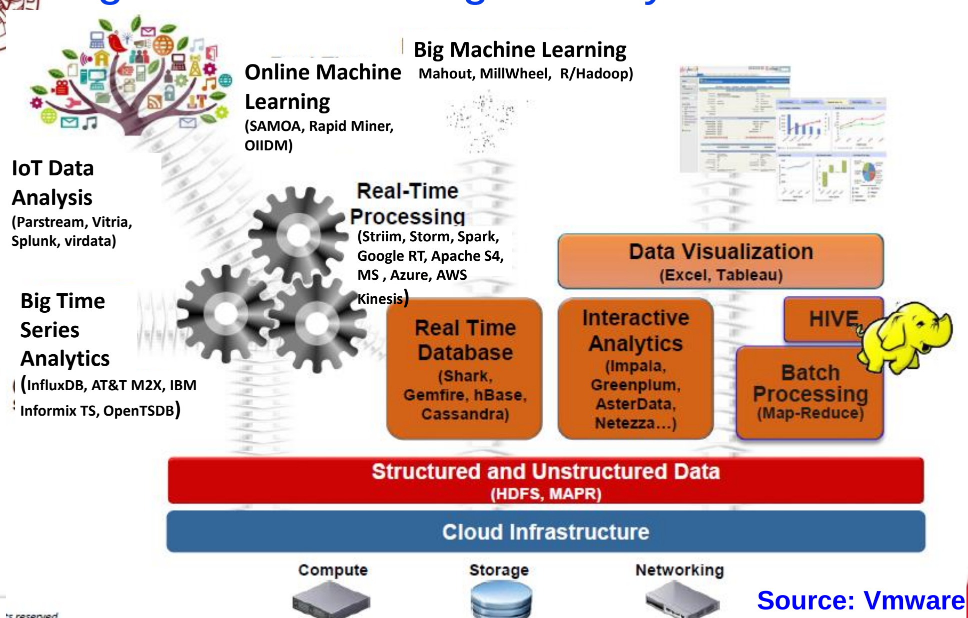
But also *new* Approaches/Algorithms!

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

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

- often, you have to build your own…

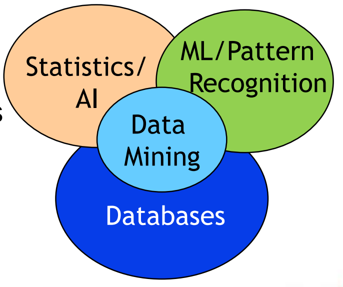
Big Data Processing & Analytics Platforms



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)

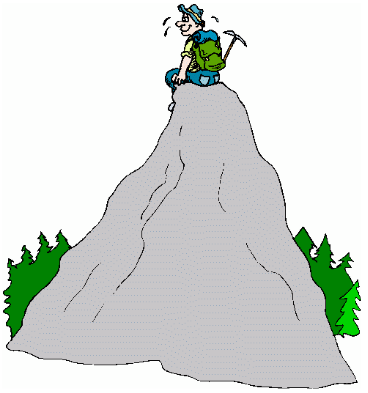
- BI aims at descriptive statistics with data with high information density to measure things, detect trends etc.
- Big Data targets inductive statistics with data with low information density whose huge volume allow to infer laws (regressions…)
- Software stack designed for BI is very specific and not 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

20

ML Computation vs. Traditional Programming

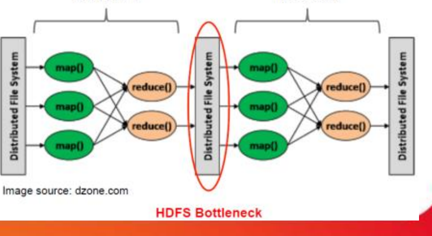


ML Program:
optimization-centric and iterative convergent

Traditional Program:
operation-centric and deterministic

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!



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 fits
  - but a small amount of data have to be synchronized across iterations (typically via the file system)
- Some Algorithms are not Good for the Mad/Reduce computing paradigm
  - when a large amount of data have to be synchronized across iterations

Why Need new Big ML Systems?

ML engineer's view

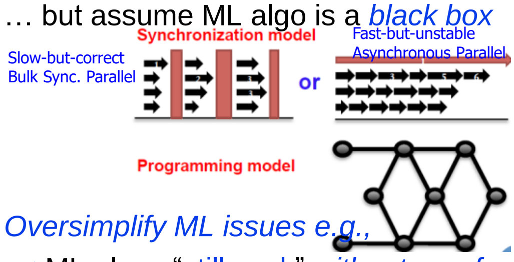
- Focus on
  - correctness,
  - fewer iterations to converge
- … but assume an *ideal system*

```
for (t = 1 to T) {
      doThings()z
      parallelUpdate(x,θ)
      doOtherThings()
}
```

- Oversimplify systems issues, e.g.,
  - need machines to perform consistently
  - need to sync parameters any time

Systems engineer's view

- Focus on
  - High iteration throughput (more iterations per sec)
  - strong fault-tolerant atomic ops
- … but assume ML algo is a *black box*



- *Oversimplify ML issues e.g.,*
  - ML algos "still work" *without proof* under different execution models
  - "easy to rewrite" in chosen abstraction (MapR, vertex, etc.)

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

22

An Alg/Sys INTERFACE for Big ML

Alone, neither side has full picture ...
New opportunities exist in the middle!

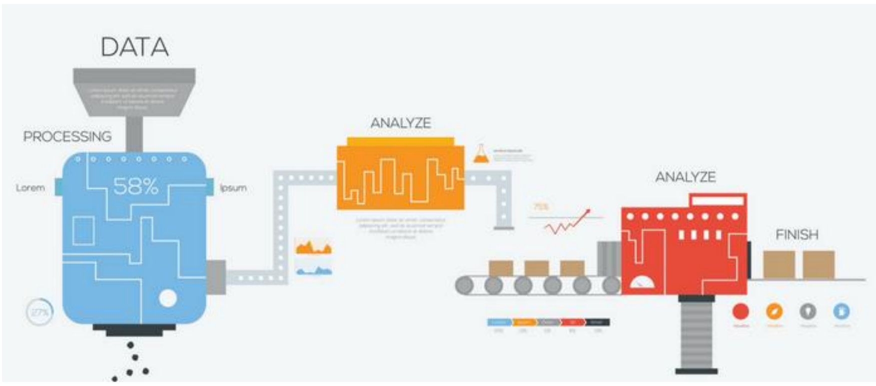
Resource Constraints:
- **Computational Efficiency (noise tolerant algorithms)**
- **Communication Cost**
- **Statistical Efficiency**
- **Human Labelling Effort**
- **Data Confidentiality , Security**

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

Fall 2019

What this Course is About?



Fall 2019

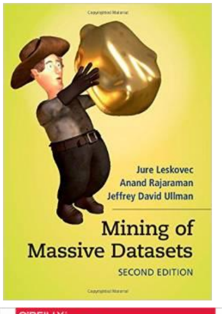
What You Will learn

- Understand different models of computation:
  - MapReduce
  - Streams and online algorithms
- 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:
  - Duplicate document detection
  - Market Basket Analysis

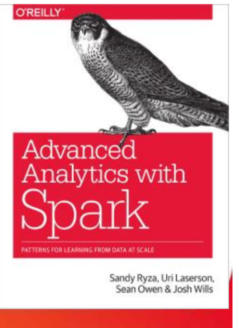
Prerequisites

- Algorithms
  - Dynamic programming, basic data structures
- Basic probability
  - Moments, typical distributions, maximum likelihood estimation (MLE), …
- Programming
  - You can do programming assignments in any language we support (Python, Java, C, C++, C# it is your choice)
    - We recommend Python, Java and C#

Course Text Books

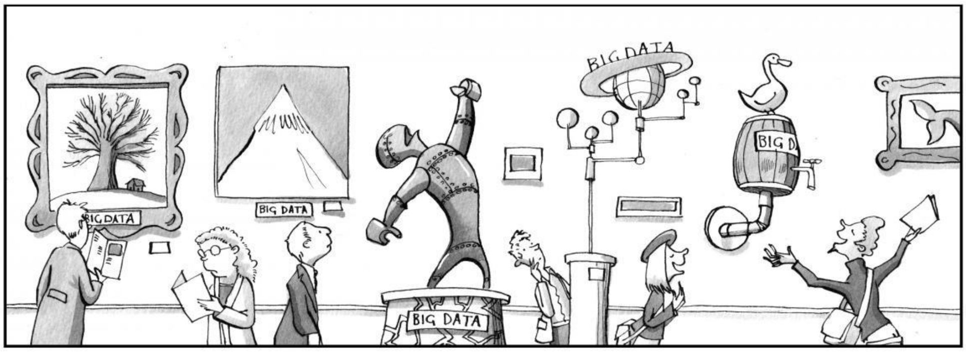


- *Jure Leskovec, Anand Rajaraman, Jeff Ullman. "Mining of Massive Datasets"* Cambridge University Press, 2014 http://www.cambridge.org/gr/academic/subjects/computer-science/knowledge-management-databases-and-data-mining/mining-massive-datasets-2nd-edition
  - Free download *http://www.mmds.org*



- *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

Course Organization



- 3 Programming Exercises (30%) : SPARK
- 1 research presentation (20%): Modern ML pipelines and trustful AI
- Final Examination (50%)
- TA: Michail Giannoulis (giannoulis@csd.uoc.gr)
  Serafim Mustakas (mustakas@csd.uoc.gr)

Tentative Course Schedule

- Lecture 1 (24-26/09): Course Overview
- Lecture 2 (01-03/10): Scalable Data Analytics using Spark
- Lecture 3 (08-10/10): Finding Similar Items
- Lecture 4 (15-17/10): Massive Data Processing
- Lecture 6 (22-24/10): Extracting Association Rules
- Lecture 7 (29-31/10): Analysing Data Streams
- Lecture 8 (05-07/11): Analysing Data Streams
- Lecture 9 (12-14/11): IoT Data Analytics
- Lecture 9 (19-21/11): Responsible Big Data Analytics
- Lab 1 (09/10) Introduction to Map-Reduce Programming
- Lab 2 (16/10) Programming in Spark Scala
- Lab 3 (23/10) Assisting Lecture for Ass 1
- Lab 4 (30/10) Intro to DataFrames and Spark SQL
- Lab 6 (6/11) Assisting Lecture for Ass 2
- Lab 7 (13/11) Streaming Programming in Spark
- Lab 7 (20/11) Assisting Lecture for Ass 3
- Students presentations (03, 05, 10, 12, 17, 19 /12)

© NY Times

Fall 2019

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 Hadoop API
- In short: you have to be able to 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

69

Fall 2019

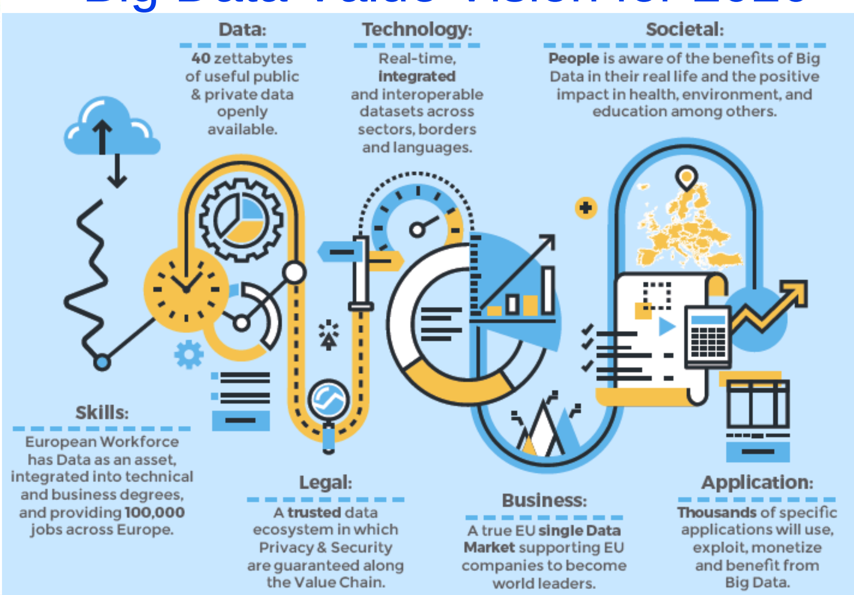
References

- CS246: Mining Massive Datasets Jure Leskovec, Stanford University, 1014
- CS9223 – Massive Data Analysis J. Freire & J. Simeon New York University Course 2013
- CS 6240: 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, 2015 Athens
- 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 Department of Computer Science National Taiwan University August 30, 2014
- Knowledge Discovery and Data Mining Evgueni Smirnov Maastricht School on Data Mining Department of Knowledge Engineering, Maastricht University, Maastricht, The Netherlands August 27 - August 30, 2013

70

Questions?

Big Data Value Vision for 2020



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