

DSSP14

Introduction to Data Science

E. Le Pennec



Fall 2019

1 Data Science

- AI? Big Data? BI? Statistics? ML? DS?
- Data Science Ecosystem
- Data Cycle
- A Focus on Machine Learning

2 Data Science Toolbox

- Computing and Distribution
- Database
- Data Science Languages
- DevOps

3 Data Scientists and Challenges

- Data Scientists
- Data Science Challenges

4 References

1 Data Science

- AI? Big Data? BI? Statistics? ML? DS?
- Data Science Ecosystem
- Data Cycle
- A Focus on Machine Learning

2 Data Science Toolbox

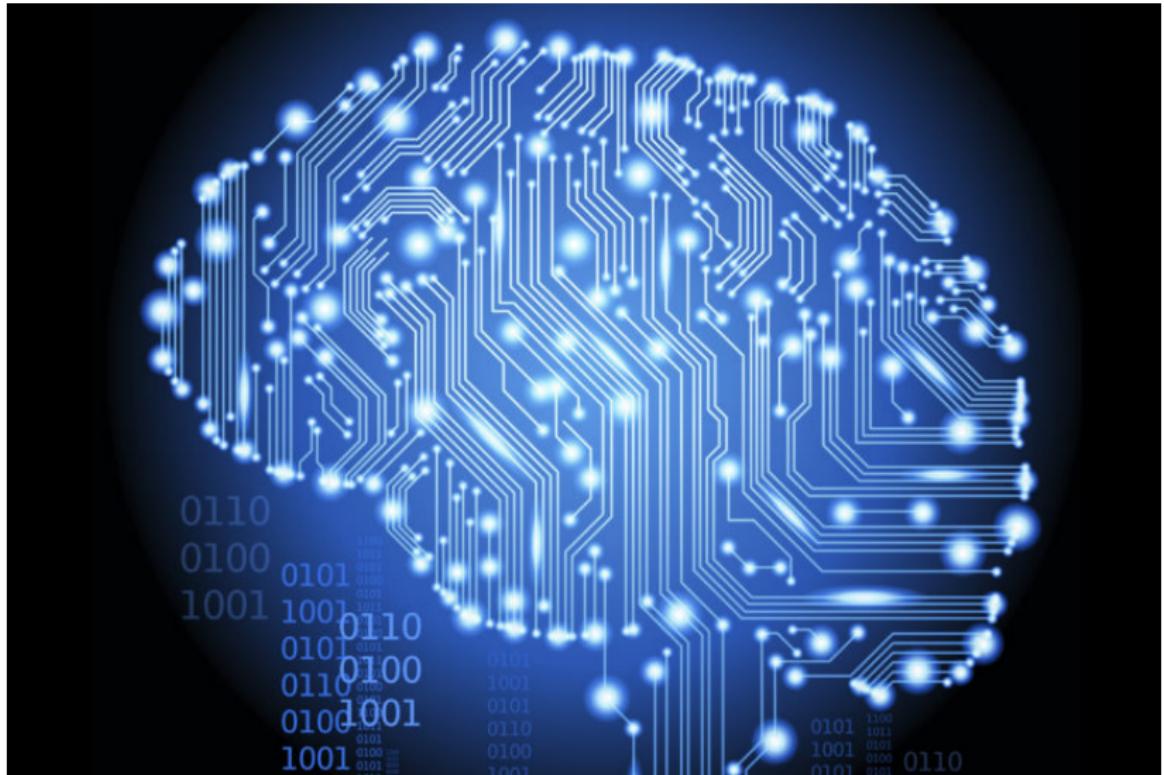
- Computing and Distribution
- Database
- Data Science Languages
- DevOps

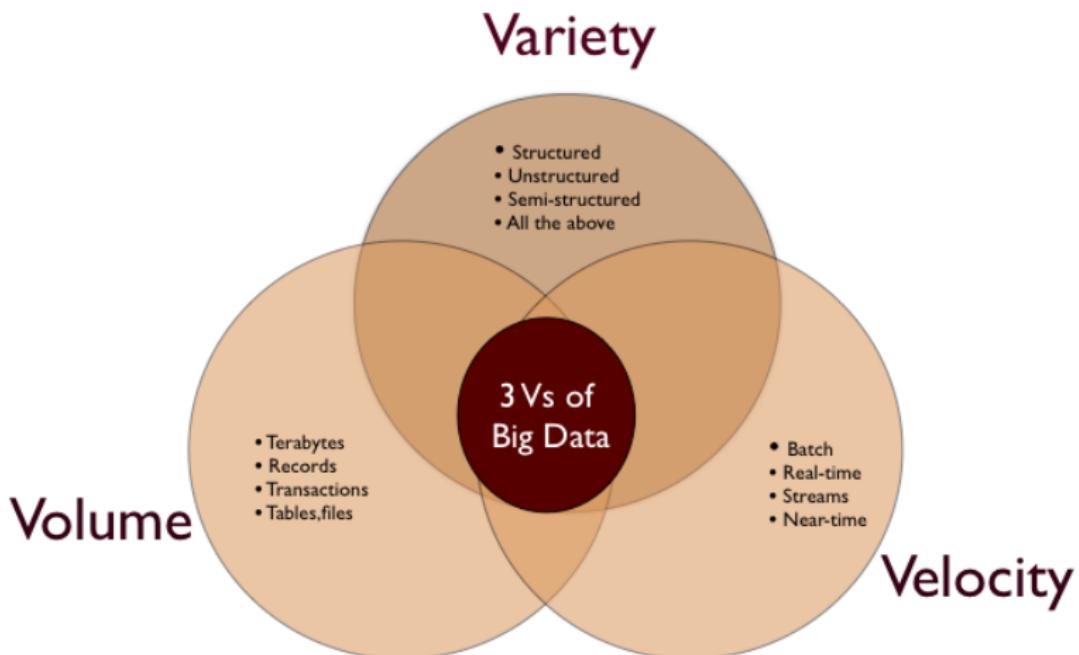
3 Data Scientists and Challenges

- Data Scientists
- Data Science Challenges

4 References

Artificial Intelligence?





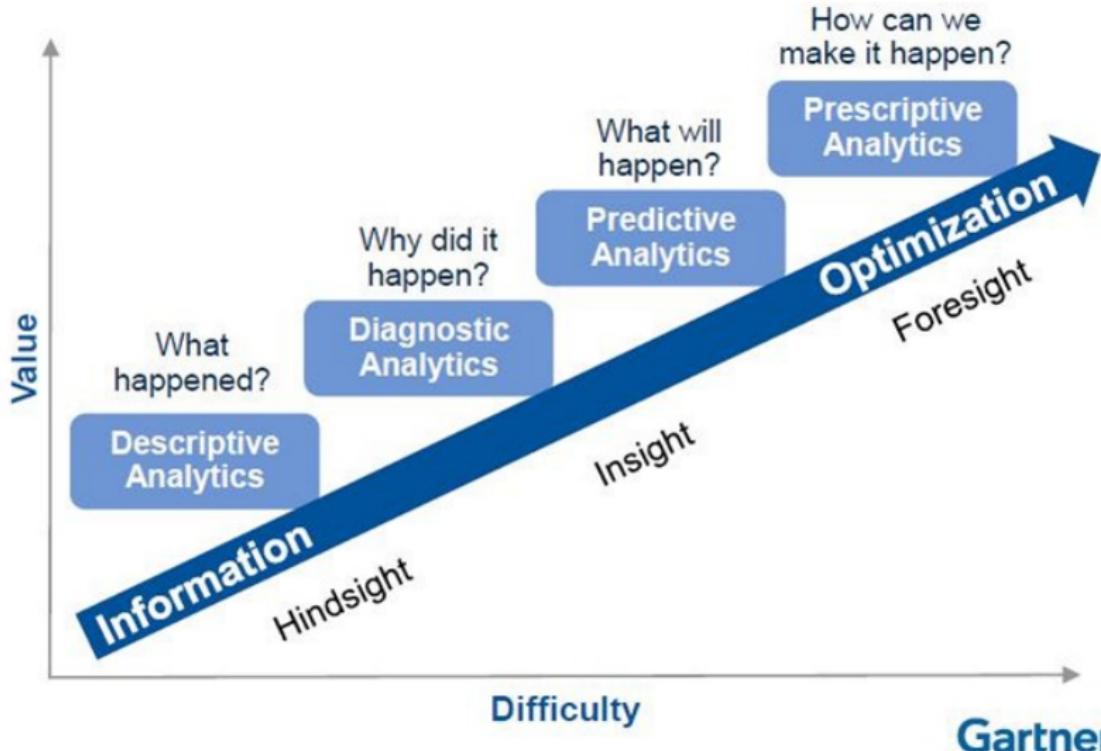
Business Intelligence?

Data Science



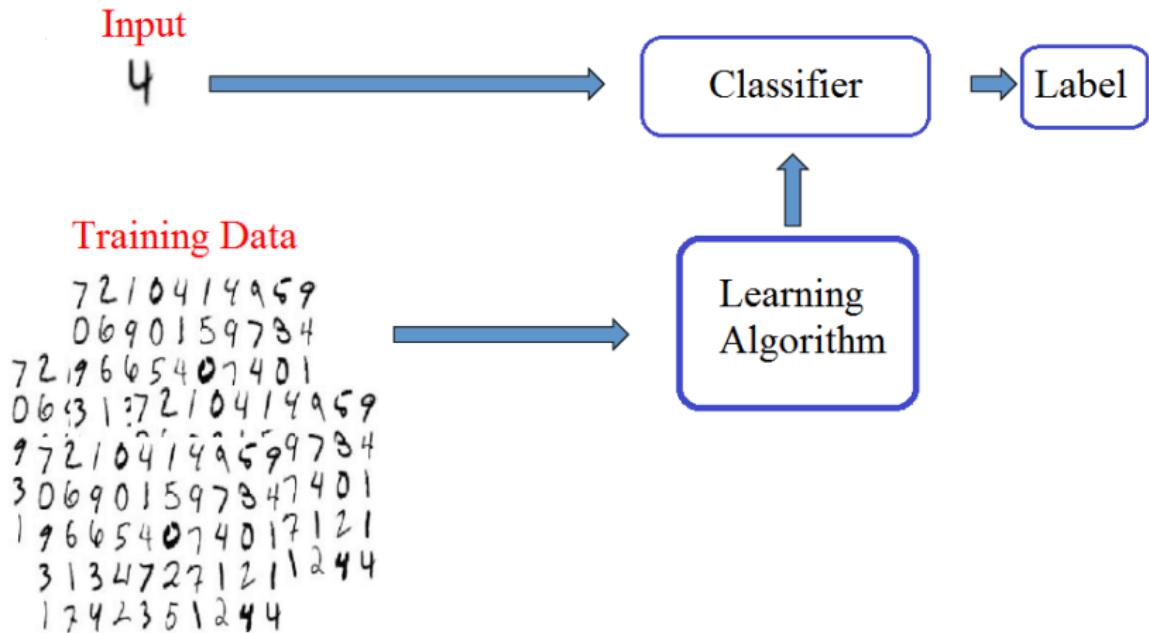
Statistics?

Data Science



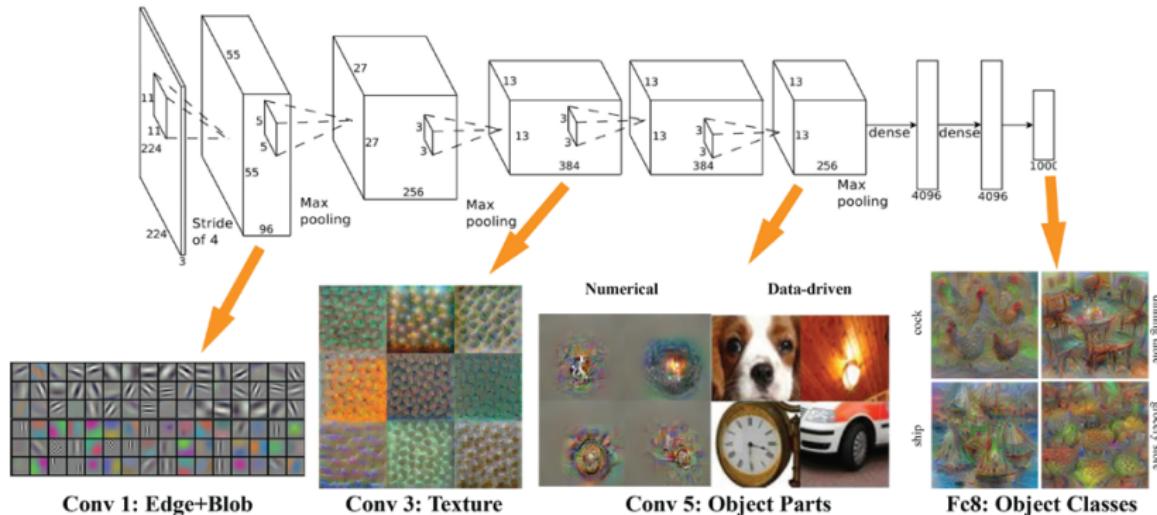
Machine Learning?

Data Science



Deep Learning?

Data Science



Data Science?

Data Science



Big data

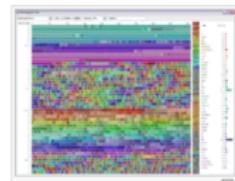
From Wikipedia, the free encyclopedia

This article is about large collections of data. For the band, see [Big Data \(band\)](#).

Big data^{[1][2]} is the term for a collection of [data sets](#) so large and complex that it becomes difficult to process using on-hand database management tools or traditional data processing applications. The challenges include capture, curation, storage,^[3] search, sharing, transfer, analysis^[4] and visualization. The trend to larger data sets is due to the additional information derivable from analysis of a single large set of related data, as compared to separate smaller sets with the same total amount of data, allowing correlations to be found to "spot business trends, determine quality of research, prevent diseases, link legal citations, combat crime, and determine real-time roadway traffic conditions."^{[5][6][7]}

As of 2012, limits on the size of data sets that are feasible to process in a reasonable amount of time were on the order of [exabytes](#) of data.^[8] Scientists regularly encounter limitations due to large data sets in many areas, including [meteorology](#), [genomics](#),^[9] [connectomics](#), complex physics simulations,^[10] and biological and environmental research.^[11] The limitations also affect [Internet search](#), [finance](#) and [business informatics](#). Data sets grow in size in part because they are increasingly being gathered by ubiquitous information-sensing mobile devices, aerial sensory technologies ([remote sensing](#)), software logs, cameras, microphones, [radio-frequency identification](#) readers, and [wireless sensor networks](#).^{[12][13]} The world's technological per-capita capacity to store information has roughly doubled every 40 months since the 1980s,^[14] as of 2012, every day 2.5 [exabytes](#) (2.5×10^{18}) of data were created.^[15] The challenge for large enterprises is determining who should own big data initiatives that straddle the entire organization.^[16]

Big data is difficult to work with using most relational database management systems and desktop statistics and visualization packages, requiring instead "massively parallel software running on tens, hundreds, or even thousands of servers".^[17] What is considered "big data" varies depending on the capabilities of the organization managing the set, and on the capabilities of the applications that are traditionally used to process and analyze the data set in its domain. "For some organizations, facing hundreds of gigabytes of data for the first time may trigger a need to reconsider data management options. For others, it may take tens or hundreds of terabytes before data size becomes a significant consideration."^[18]



A visualization created by IBM of Wikipedia edits. At multiple terabytes in size, the text and images of Wikipedia are a classic example of big data.

Big data

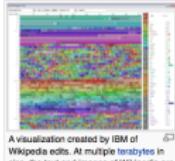
From Wikipedia, the free encyclopedia.

This article is about large collections of data. For the band, see [Big Data \(band\)](#).

Big data^{[1][2]} is the term for a collection of data sets so large and complex that it becomes difficult to process using on-hand database management tools or traditional data processing applications. The challenges include capture, curation, storage,^[3] search, sharing, transfer, analysis,^[4] and visualization. The trend to larger data sets is due to the additional information derivable from analysis of a single large set of related data, as compared to separate smaller sets with the same total amount of data, allowing correlations to be found to "spot business trends, determine quality of research, prevent diseases, link legal citations, combat crime, and determine real-time roadway traffic conditions."^{[5][6][7]}

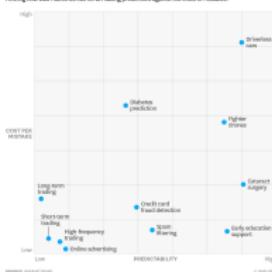
As of 2012, limits on the size of data sets that are feasible to process in a reasonable amount of time were on the order of exabytes of data.^[8] Scientists regularly encounter limitations due to large data sets in many areas, including meteorology, genomics,^[9] connectomics, complex physics simulations,^[10] and biological and environmental research.^[11] The limitations also affect Internet search, finance and business information. Data sets grow in size in part because they are increasingly being gathered by ubiquitous information-sensing mobile devices, aerial sensory technologies (remote sensing), software logs, cameras, microphones, radio-frequency identification readers, and wireless sensor networks.^{[12][13]} The world's technological per-capita capacity to store information has roughly doubled every 40 months since the 1980s,^[14] as of 2012, every day 2.5 exabytes (2.5×10^{18}) of data were created.^[15] The challenge for large enterprises is determining who should own big data initiatives that straddle the entire organization.^[16]

Big data is difficult to work with using most relational database management systems and desktop statistics and visualization packages, requiring instead "massively parallel software running on tens, hundreds, or even thousands of servers".^[17] What is considered "big data" varies depending on the capabilities of the organization managing the set, and on the capabilities of the applications that are traditionally used to process and analyze the data set in its domain. "For some organizations, facing hundreds of gigabytes of data for the first time may trigger a need to reconsider data management options. For others, it may take

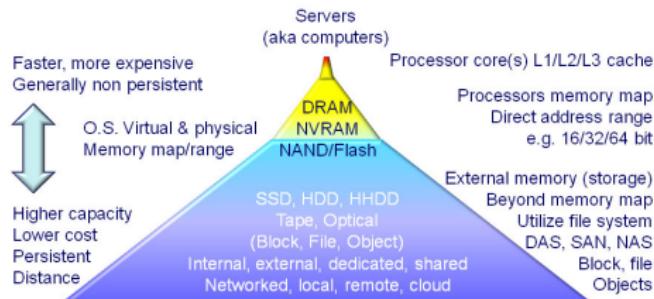


- **Artificial Intelligence** research is defined as the study of *intelligent agents*: any device that perceives its environment and takes actions that maximize its chance of success at some goal.
- **Big data** is an all-encompassing term for any collection of data sets so large and complex that it becomes difficult to process using traditional data processing applications.
- **Business Intelligence** comprises the strategies and technologies used by enterprises for the data analysis of business information.
- **Statistics** is the study of the collection, analysis, interpretation, presentation and organization of data.
- **Machine learning** is the subfield of computer science that gives computers the ability to learn without being explicitly programmed.
- **Deep Learning** is part of a broader family of machine learning methods based on learning data representations, as opposed to task-specific algorithms.
- **Data science** is the study of the generalizable extraction of knowledge from data, yet the key word is science.

The Decision Automation Map
Existing firms and machines can do at making predictions against the costs of inhibition.



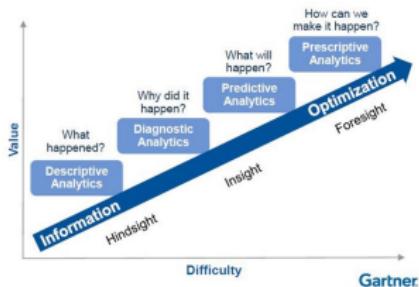
- **Artificial Intelligence** research is defined as the study of *intelligent agents*: any device that perceives its environment and takes actions that maximize its chance of success at some goal.
- **Big data** is an all-encompassing term for any collection of data sets so large and complex that it becomes difficult to process using traditional data processing applications.
- **Business Intelligence** comprises the strategies and technologies used by enterprises for the data analysis of business information.
- **Statistics** is the study of the collection, analysis, interpretation, presentation and organization of data.
- **Machine learning** is the subfield of computer science that gives computers the ability to learn without being explicitly programmed.
- **Deep Learning** is part of a broader family of machine learning methods based on learning data representations, as opposed to task-specific algorithms.
- **Data science** is the study of the generalizable extraction of knowledge from data, yet the key word is science.



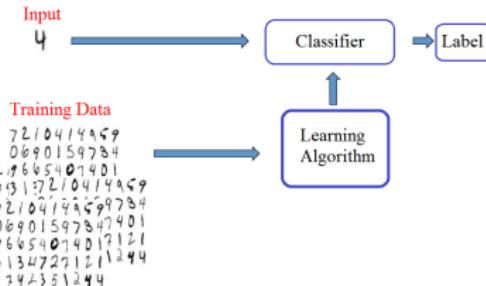
- **Artificial Intelligence** research is defined as the study of *intelligent agents*: any device that perceives its environment and takes actions that maximize its chance of success at some goal.
- **Big data** is an all-encompassing term for any collection of data sets so large and complex that it becomes difficult to process using traditional data processing applications.
- **Business Intelligence** comprises the strategies and technologies used by enterprises for the data analysis of business information.
- **Statistics** is the study of the collection, analysis, interpretation, presentation and organization of data.
- **Machine learning** is the subfield of computer science that gives computers the ability to learn without being explicitly programmed.
- **Deep Learning** is part of a broader family of machine learning methods based on learning data representations, as opposed to task-specific algorithms.
- **Data science** is the study of the generalizable extraction of knowledge from data, yet the key word is science.



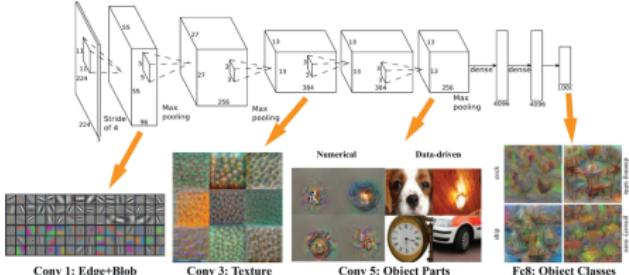
- **Artificial Intelligence** research is defined as the study of *intelligent agents*: any device that perceives its environment and takes actions that maximize its chance of success at some goal.
- **Big data** is an all-encompassing term for any collection of data sets so large and complex that it becomes difficult to process using traditional data processing applications.
- **Business Intelligence** comprises the strategies and technologies used by enterprises for the data analysis of business information.
- **Statistics** is the study of the collection, analysis, interpretation, presentation and organization of data.
- **Machine learning** is the subfield of computer science that gives computers the ability to learn without being explicitly programmed.
- **Deep Learning** is part of a broader family of machine learning methods based on learning data representations, as opposed to task-specific algorithms.
- **Data science** is the study of the generalizable extraction of knowledge from data, yet the key word is science.



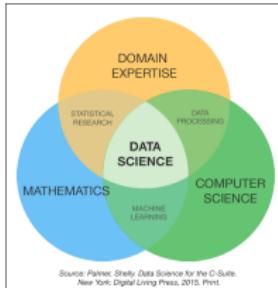
- **Artificial Intelligence** research is defined as the study of *intelligent agents*: any device that perceives its environment and takes actions that maximize its chance of success at some goal.
- **Big data** is an all-encompassing term for any collection of data sets so large and complex that it becomes difficult to process using traditional data processing applications.
- **Business Intelligence** comprises the strategies and technologies used by enterprises for the data analysis of business information.
- **Statistics** is the study of the collection, analysis, interpretation, presentation and organization of data.
- **Machine learning** is the subfield of computer science that gives computers the ability to learn without being explicitly programmed.
- **Deep Learning** is part of a broader family of machine learning methods based on learning data representations, as opposed to task-specific algorithms.
- **Data science** is the study of the generalizable extraction of knowledge from data, yet the key word is science.



- **Artificial Intelligence** research is defined as the study of *intelligent agents*: any device that perceives its environment and takes actions that maximize its chance of success at some goal.
- **Big data** is an all-encompassing term for any collection of data sets so large and complex that it becomes difficult to process using traditional data processing applications.
- **Business Intelligence** comprises the strategies and technologies used by enterprises for the data analysis of business information.
- **Statistics** is the study of the collection, analysis, interpretation, presentation and organization of data.
- **Machine learning** is the subfield of computer science that gives computers the ability to learn without being explicitly programmed.
- **Deep Learning** is part of a broader family of machine learning methods based on learning data representations, as opposed to task-specific algorithms.
- **Data science** is the study of the generalizable extraction of knowledge from data, yet the key word is science.

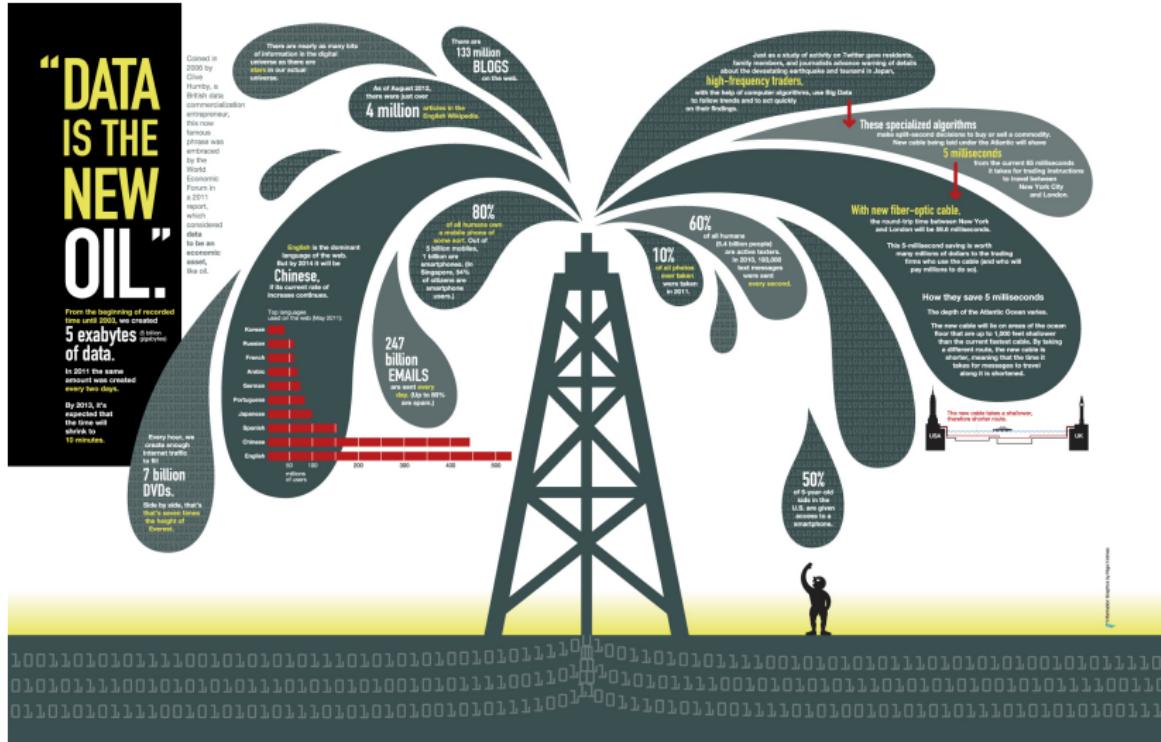


- **Artificial Intelligence** research is defined as the study of *intelligent agents*: any device that perceives its environment and takes actions that maximize its chance of success at some goal.
- **Big data** is an all-encompassing term for any collection of data sets so large and complex that it becomes difficult to process using traditional data processing applications.
- **Business Intelligence** comprises the strategies and technologies used by enterprises for the data analysis of business information.
- **Statistics** is the study of the collection, analysis, interpretation, presentation and organization of data.
- **Machine learning** is the subfield of computer science that gives computers the ability to learn without being explicitly programmed.
- **Deep Learning** is part of a broader family of machine learning methods based on learning data representations, as opposed to task-specific algorithms.
- **Data science** is the study of the generalizable extraction of knowledge from data, yet the key word is science.



- **Artificial Intelligence** research is defined as the study of *intelligent agents*: any device that perceives its environment and takes actions that maximize its chance of success at some goal.
- **Big data** is an all-encompassing term for any collection of data sets so large and complex that it becomes difficult to process using traditional data processing applications.
- **Business Intelligence** comprises the strategies and technologies used by enterprises for the data analysis of business information.
- **Statistics** is the study of the collection, analysis, interpretation, presentation and organization of data.
- **Machine learning** is the subfield of computer science that gives computers the ability to learn without being explicitly programmed.
- **Deep Learning** is part of a broader family of machine learning methods based on learning data representations, as opposed to task-specific algorithms.
- **Data science** is the study of the generalizable extraction of knowledge from data, yet the key word is science.

Data Is The New Oil?



1 Data Science

- AI? Big Data? BI? Statistics? ML? DS?
- **Data Science Ecosystem**
- Data Cycle
- A Focus on Machine Learning

2 Data Science Toolbox

- Computing and Distribution
- Database
- Data Science Languages
- DevOps

3 Data Scientists and Challenges

- Data Scientists
- Data Science Challenges

4 References

Major Influences

Four major influences act today:

- The formal theories of statistics
- Accelerating developments in computers and display devices
- The challenge, in many fields, of more and ever larger bodies of data
- The emphasis on quantification in an ever wider variety of disciplines

Major Influences - Tukey (1962)

Four major influences act today:

- The formal theories of statistics
 - Accelerating developments in computers and display devices
 - The challenge, in many fields, of more and ever larger bodies of data
 - The emphasis on quantification in an ever wider variety of disciplines
-
- He was talking of Data Analysis.
 - Data mining, Machine learning, Big Data, DS, AI...

Large Scale ML Is (Quite) Easy

Data Science



```
def run(params: Params) {
    val conf = new SparkConf()
        .setAppName("BinaryClassification with spark")
    val sc = new SparkContext(conf)

    Logger.getRootLogger.setLevel(Level.WARN)

    val examples = MLUtils.loadLibSVMFile(sc, params.input).cache()

    val splits = examples.randomSplit(Array(0.8, 0.2))
    val training = splits(0).cache()
    val test = splits(1).cache()
    val numTraining = training.count()
    val numTest = test.count()
    println("Training: " + numTraining, "test: " + numTest)
    examples.unpersist(blocking = false)

    val updater = params.optimizer match {
        case L1 => new L1Updater()
        case L2 => new SquaredL2Updater()
    }

    val algorithm = new LogisticRegressionWithSGD()
        .setNumIterations(params.numIterations)
        .setStepSize(params.stepSize)
        .setUpdater(updater)
        .setRegParam(params.regParam)
    val model = algorithm.run(training).clearThreshold()

    val prediction = model.predict(test.map(_.features))
    val predictionAndLabel = prediction.zip(test.map(_.label))

    val metrics = new BinaryClassificationMetrics(predictionAndLabel)
    val myMetrics = new MyBinaryClassificationMetrics(predictionAndLabel)

    println("Empirical CrossEntropy: " + myMetrics.crossEntropy())
    println("Test areaUnderPR: " + metrics.areaUnderPR())
    println("Test areaUnderROC: " + metrics.areaUnderROC())
}

sc.stop()
```



Example of off the shelves solution

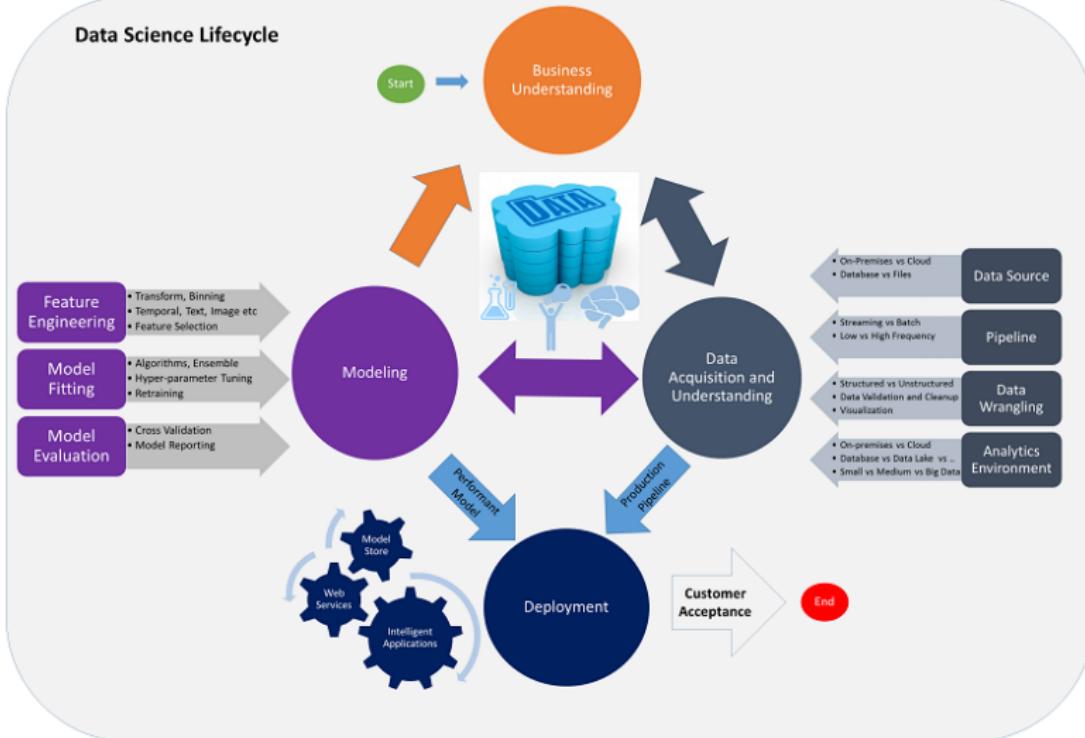
- Algorithm implementation + copy/paste + cloud computing.
- Machine learning on an arbitrary large dataset!

Data Science Is (Quite) Complex!

Data Science



Data Science Lifecycle



Data Ecosystem Is (Quite) Complex!

Data Science



DATA & AI LANDSCAPE 2019

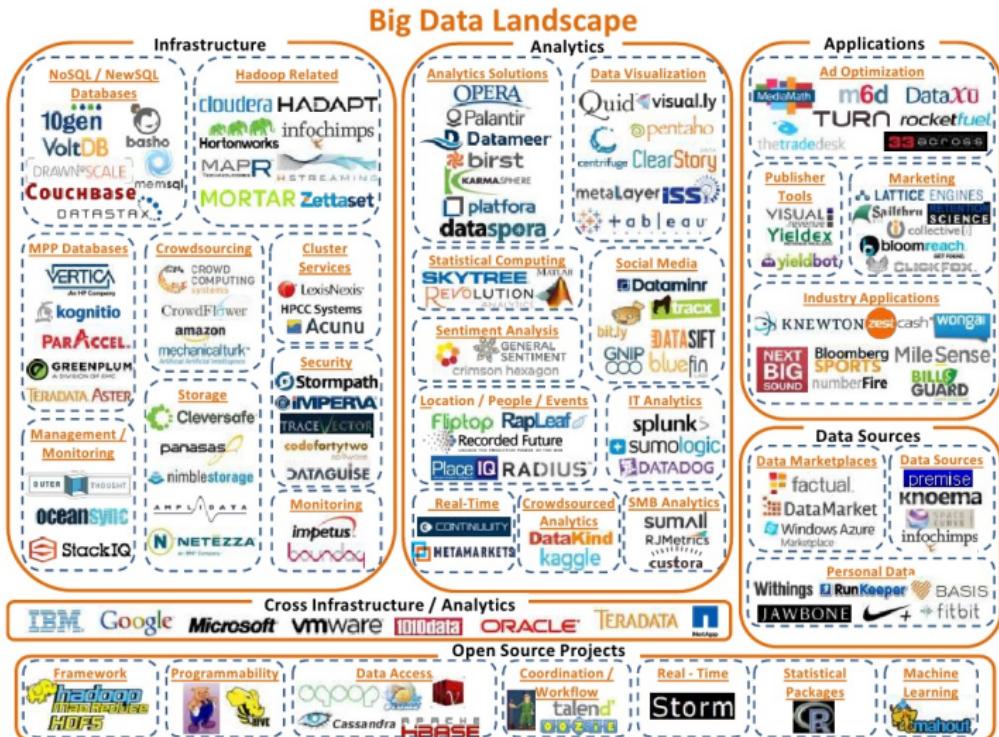


Jan 27, 2019

© Matt Turck (@mattturck), Lice Xu (@licaxu92), & FirstMark (@firstmarkcao) mattturck.com/bigdata2019

FIRSTMARK
FAMILY STAGED VENTURE CAPITAL

Data Ecosystem Is (Quite) Complex!

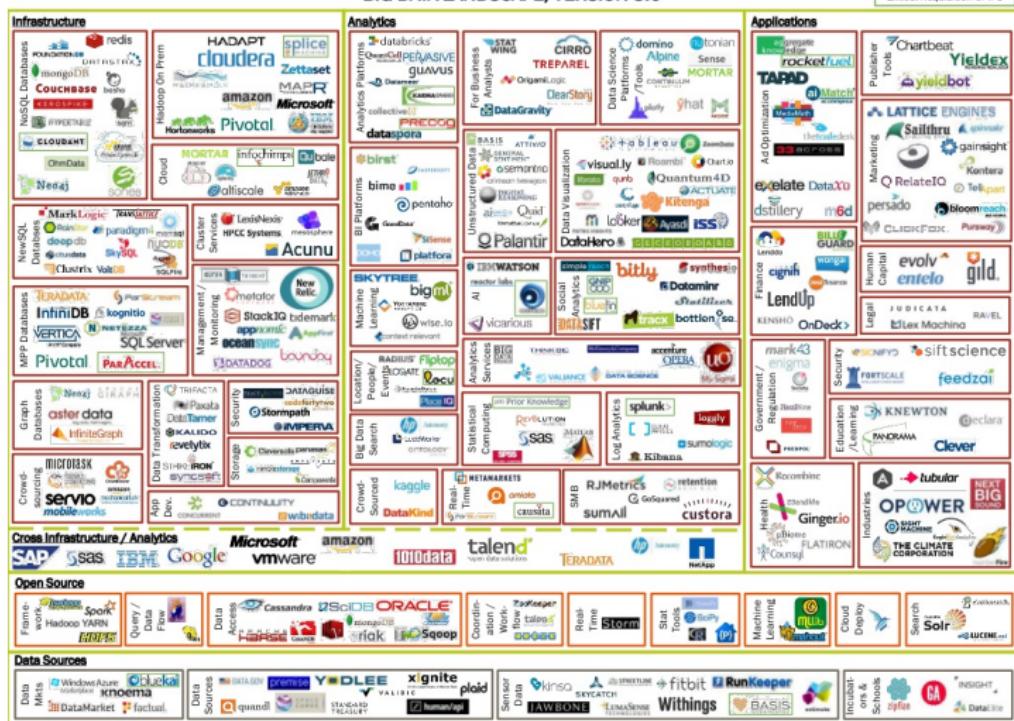


© Matt Turck (@mattturck) and ShivenZilis (@shivonz)

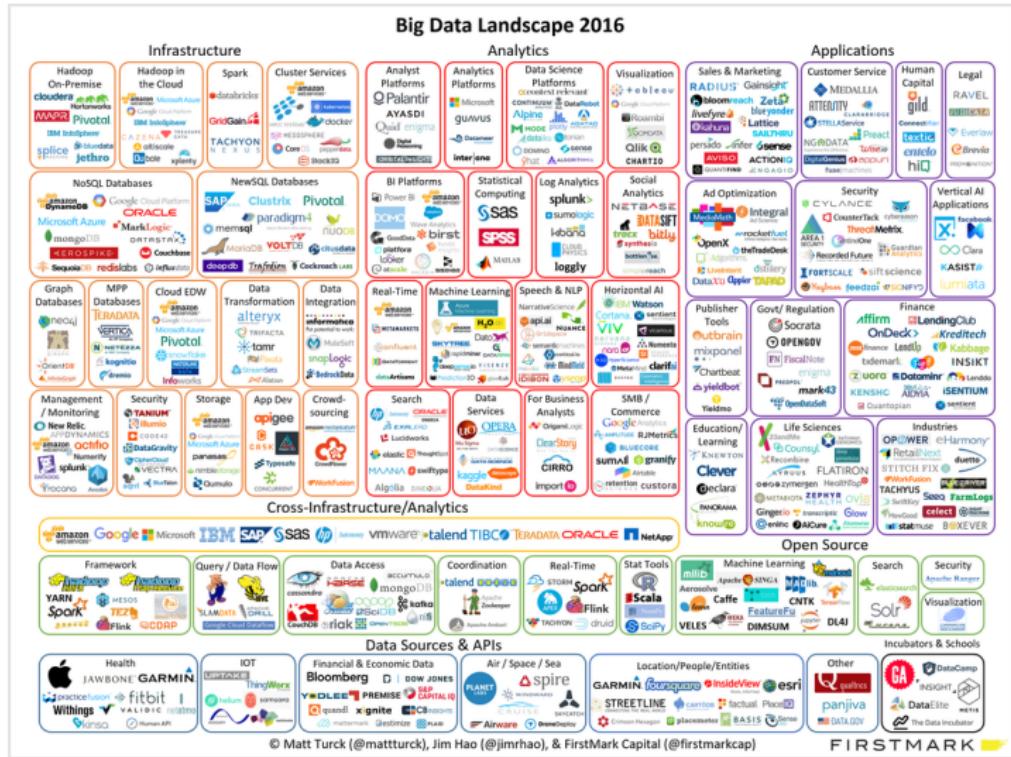
Data Ecosystem Is (Quite) Complex!

BIG DATA LANDSCAPE, VERSION 3.0

Exited: Acquisition or IPO



Data Ecosystem Is (Quite) Complex!



Data Ecosystem Is (Quite) Complex!



BIG DATA LANDSCAPE 2017



V2 - Last updated 5/3/2017

© Matt Turck (@mattturck), Jim Hao (@jimrhao), & FirstMark (@firstmarkcap) mattturck.com/biodata2017

FIRSTMARK

EARLY STAGE VENTURE CAPITAL

Source: M. Turck

Data Ecosystem Is (Quite) Complex!

Data Science



Final 2018 version, updated 03/15/2018

© Matt Turck (@mattturck), Remi Obawumi (@remi_ obawumi), & FirstMark (@firstmarkcap)

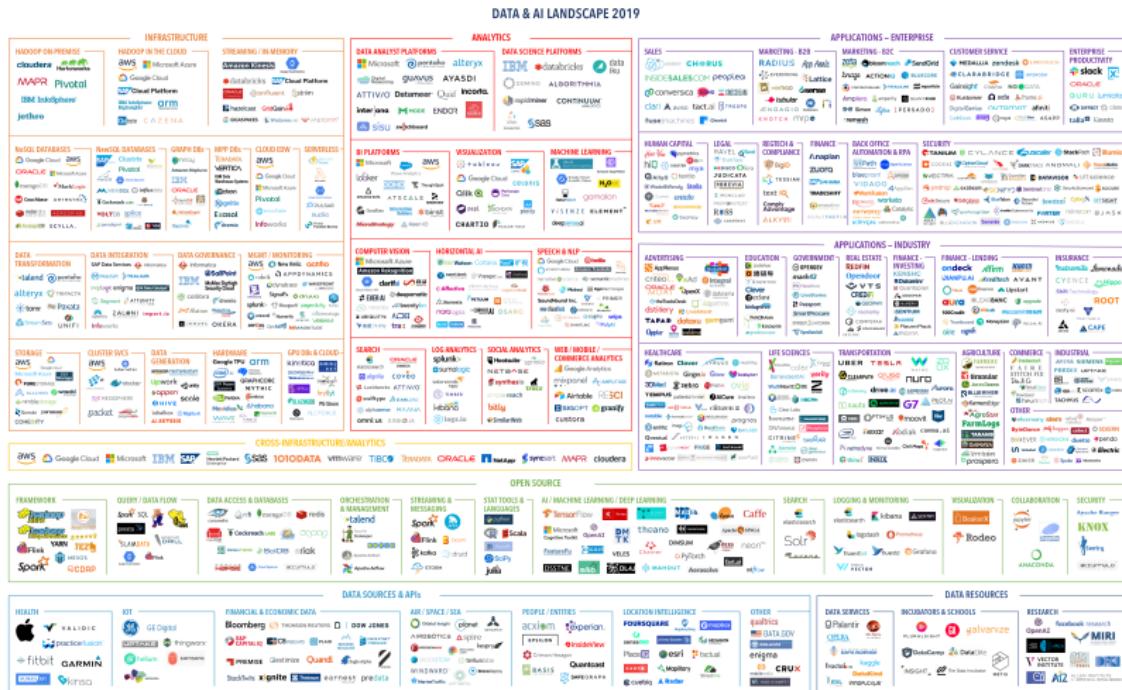
mattturck.com/bigdata2018

FIRSTMARK

Source: M. Turck

Data Ecosystem Is (Quite) Complex!

Data Science



Mo 27.2018

© Matt Turck (@mattturck), Lisa Xu (@lisaxy92), & FirstMark (@firstmarkcap)

mattturck.com/biodata2019

FIRSTMARK

Source: M. Turck

Data Ecosystem Is (Quite) Complex!

Data Science



 The Periodic Table of Data Science

An overview of key companies, resources and tools in data science (as of 4/12/2017)

Source: DataCamp

1 Data Science

- AI? Big Data? BI? Statistics? ML? DS?
- Data Science Ecosystem
- Data Cycle
- A Focus on Machine Learning

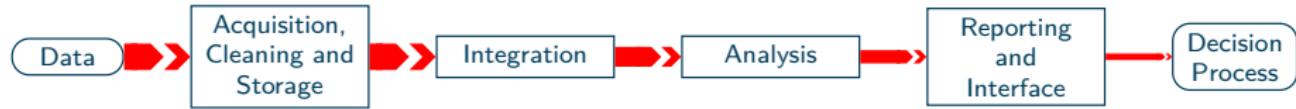
2 Data Science Toolbox

- Computing and Distribution
- Database
- Data Science Languages
- DevOps

3 Data Scientists and Challenges

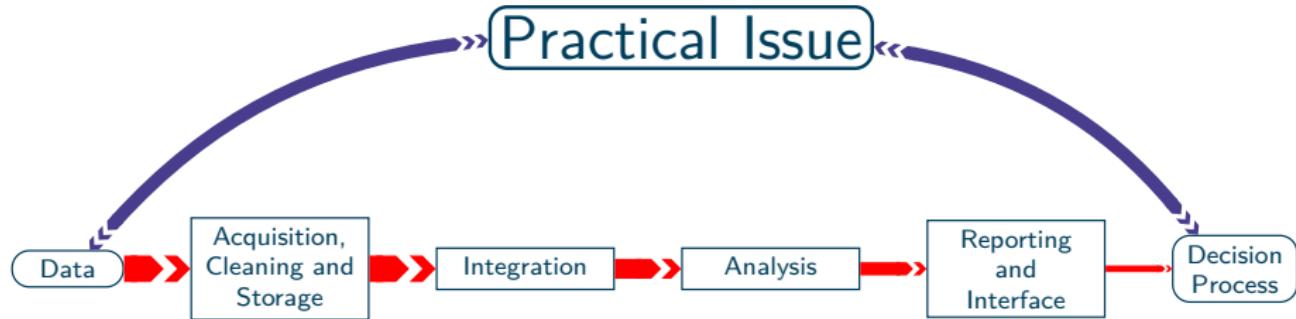
- Data Scientists
- Data Science Challenges

4 References



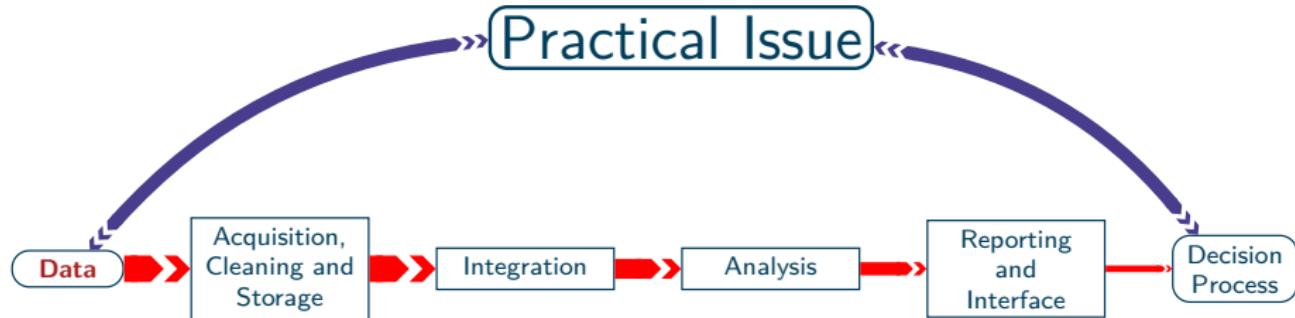
Data/Information Flow

- *Plumber vision*



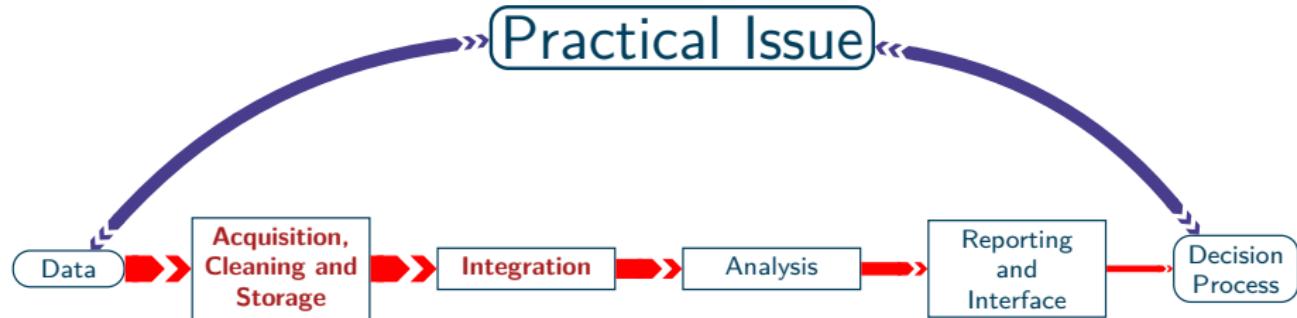
Data/Information Flow

- *Plumber vision*
- Driven by a goal



Data

- Raw material:
 - Proprietary / Market / Open
 - Structured and unstructured data
 - Various (file/source) format!
- Issues:
 - Availability / Legality / Acceptability (Ethics)
 - Quantity / Processing speed
 - Quality

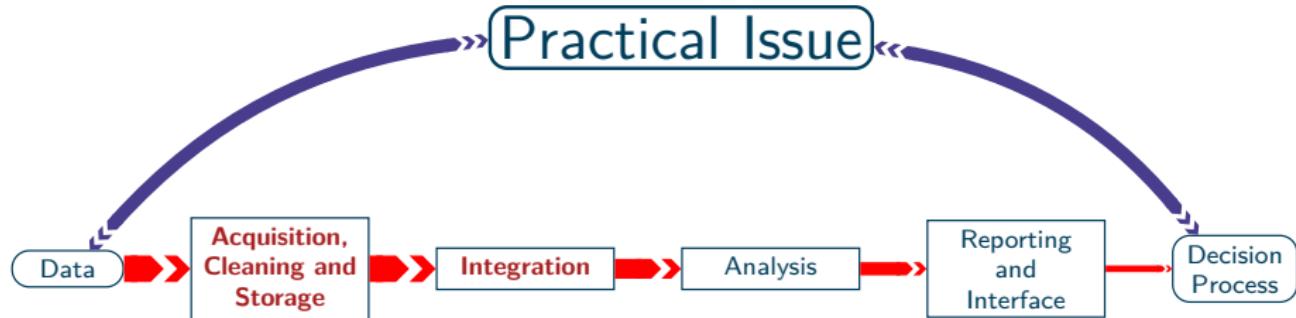


Acquisition, Cleaning, Storage (ETL)

- Getting data from the sources.
- Storage issue and processing.
- Cleaning and formatting.

Integration

- Analysis tool specific data preparation.

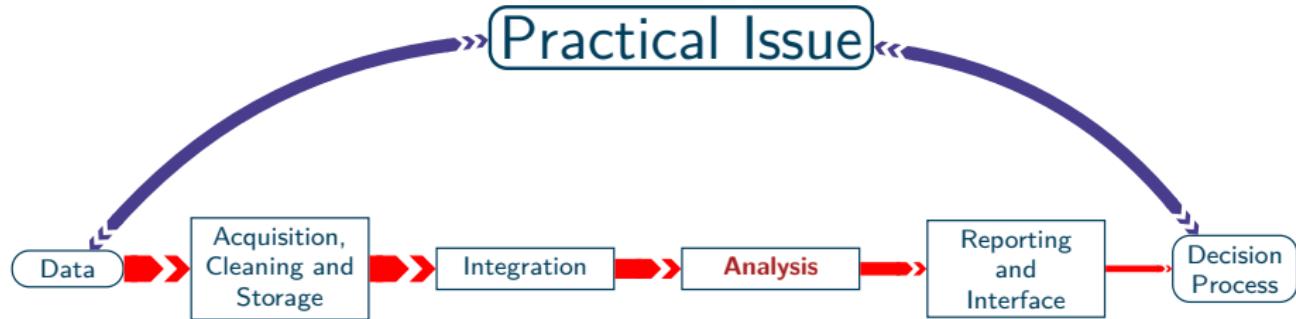


Acquisition, Cleaning, Storage (ETL) (Data management)

- Getting data from the sources.
- Storage issue and processing.
- Cleaning and formatting.

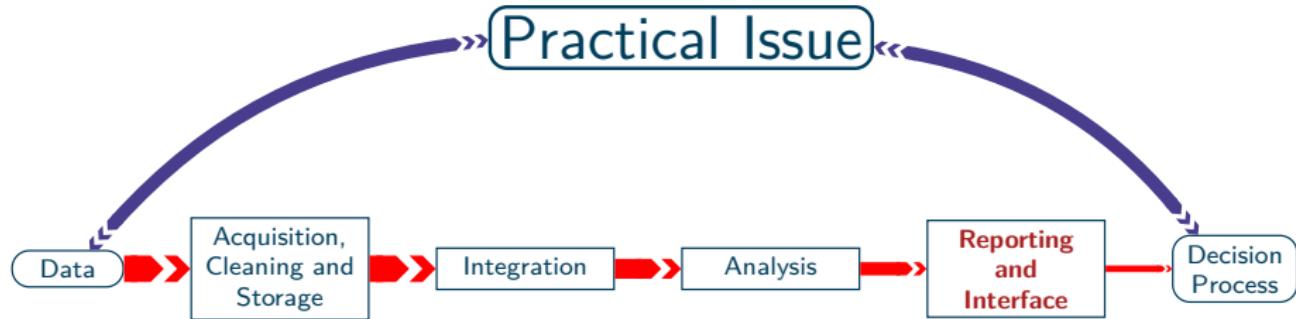
Integration (Plumbing issue)

- Analysis tool specific data preparation.
- Time consuming!
- **Big Data:** hardware is the limit (time/volume)



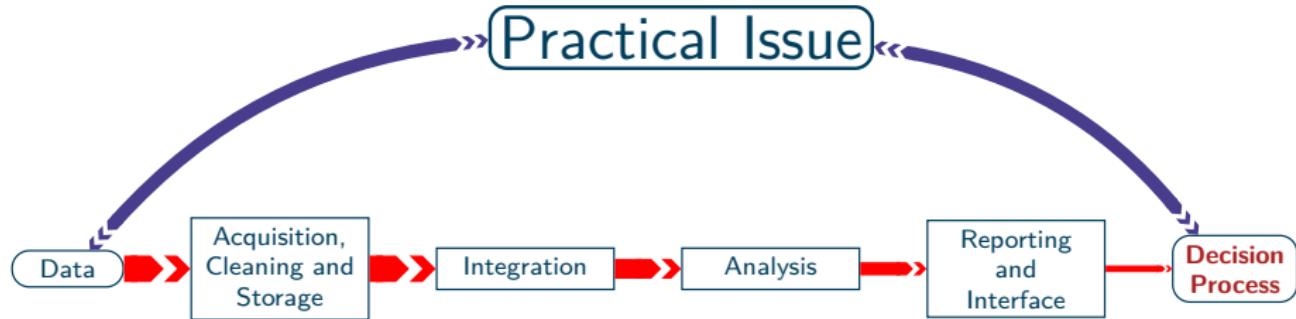
Analysis

- Knowledge extraction from the data
- Statistics, Machine Learning, AI...
- **Big Data:** hardware is the limit (time/volume)



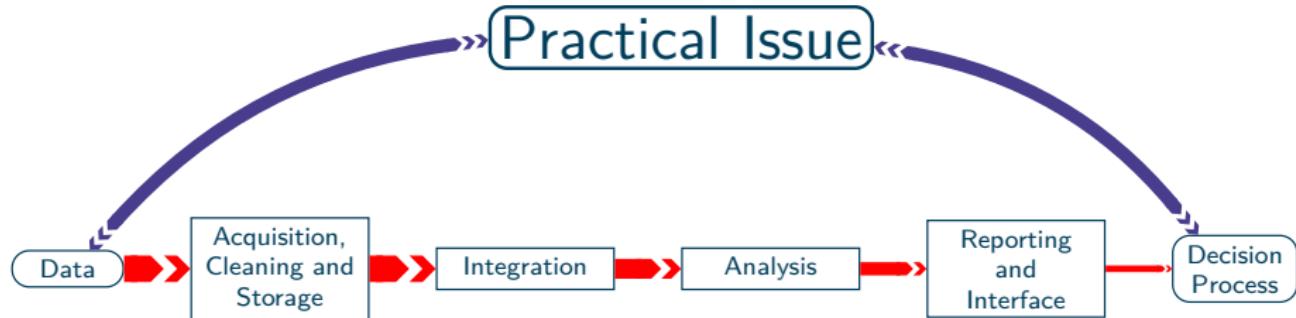
Reporting and Interface

- (Human) Reporting: visualization, report, app...
- (Computer) Interface: dedicated tool, Application Programming Interface (API)
- Also useful for data exploration
- Some key aspects:
 - Sharing (and Collaboration)
 - Deployment



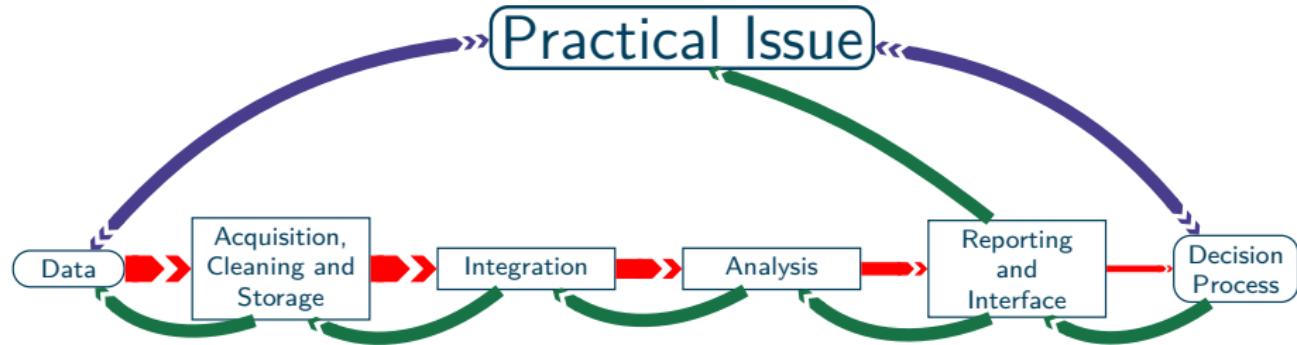
Decision and Goal Oriented Analysis

- Better decisions: **Value**
- Need to have a problem/question!
- No good answer without a well specified question!
- Need to define a success criterion!
- Feedback everywhere...



Data/Information Flow

- *Plumber vision*
- Driven by a goal
- **Iterative and interactive process**



Data/Information Flow

- *Plumber vision*
- Driven by a goal
- **Iterative and interactive process**

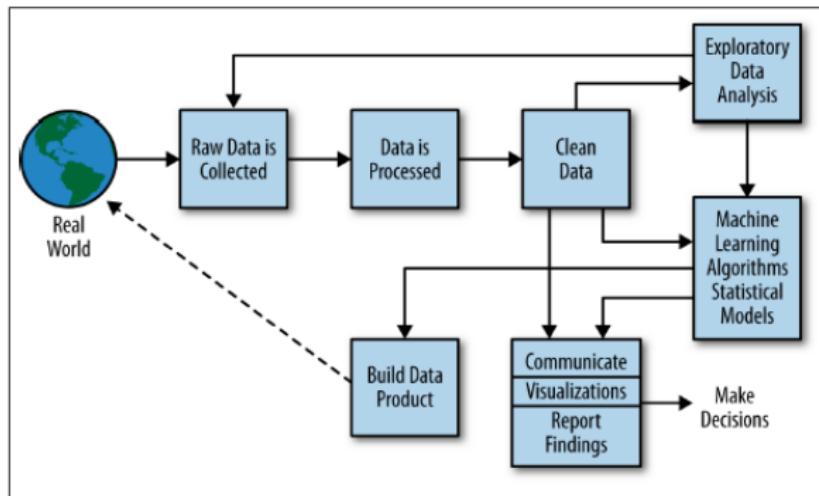


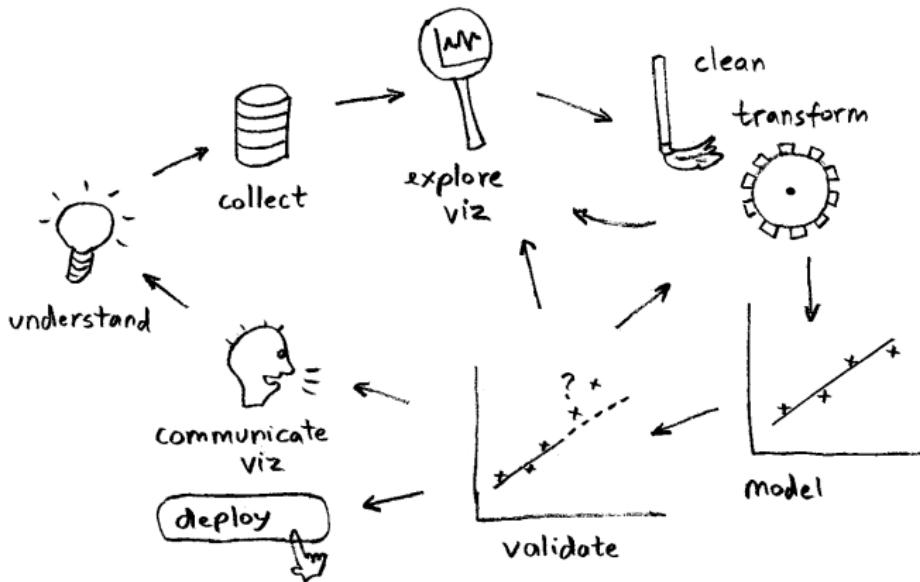
Figure 2-2. The data science process

Doing Data Science: Straight talk from the frontline

- Rachel Schutt, Cathy O'Neil
- Columbia (2012), O'Reilly (2013)

Data Science in 1 Slide

Data Science



CRISP-DM

- Cross-industry standard process for data mining (1999)
- Adapted by Szilard Pafka.

1 Data Science

- AI? Big Data? BI? Statistics? ML? DS?
- Data Science Ecosystem
- Data Cycle
- A Focus on Machine Learning

2 Data Science Toolbox

- Computing and Distribution
- Database
- Data Science Languages
- DevOps

3 Data Scientists and Challenges

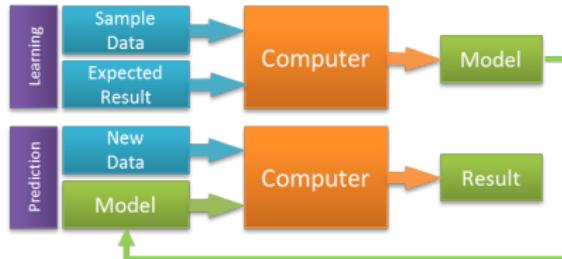
- Data Scientists
- Data Science Challenges

4 References

Traditional modeling:



Machine Learning:



A definition by Tom Mitchell
(<http://www.cs.cmu.edu/~tom/>)

A computer program is said to learn from **experience E** with respect to some **class of tasks T** and **performance measure P**, if its performance at tasks in T , as measured by P, improves with experience E.

News Clustering

Data Science



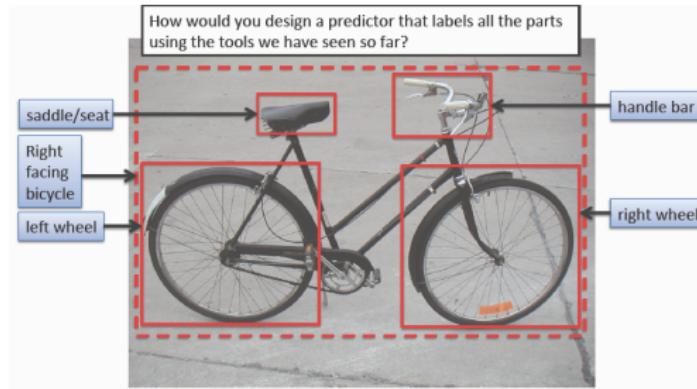
The screenshot shows the Google News interface. On the left, a sidebar lists 'SECTIONS' including Top Stories, World, U.S., Business, Technology, Entertainment, Sports, Science, and Health, with 'Top Stories' currently selected. The main area displays 'Top Stories' with two main articles:

- Sarah Huckabee Sanders rips CNN, media at heated briefing** (Fox News - 1h ago)
RELATED COVERAGE:
Three journalists leaving CNN after retracted article
Highly Cited - CNINMoney - Jan 26, 2017
Reporter accuses White House of 'flaming' media tensions in heated exchange
The Hill (blog) - 1h ago
- A Time Magazine with Trump on the cover hangs in his golf clubs. It's fake.** (Washington Post - 4h ago)
RELATED COVERAGE:
Donald Trump Time Magazine Covers: See Them All | Time.com
Most Referenced - Time Magazine - 1h ago

On the right, there are sidebar sections for 'Google News' (with a link to 'View more'), 'Read our blk' (with a link to 'In the News'), and 'In the News' (listing Executive I, Serena Williams, Syria, Google, John McElroy, and Bashar al-Assad).

A news clustering algorithm:

- **Task:** group article corresponding to the same news
- **Performance:** quality of the clusters
- **Experience:** set of articles



A detection/recognition algorithm:

- **Task**: say if an object is present or not in the image
- **Performance**: number of errors
- **Experience**: set of previously seen labeled images

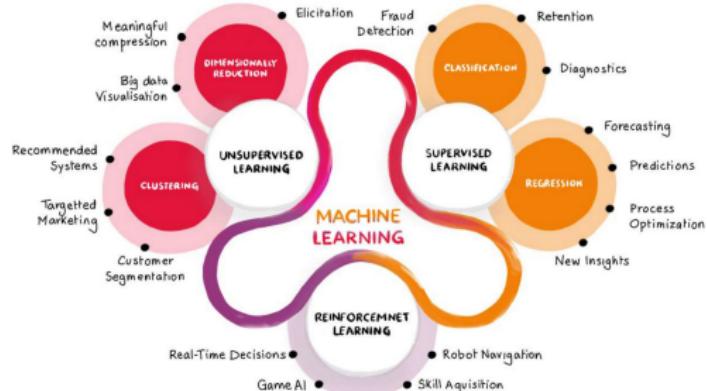


A robot endowed with a set of sensors and an online learning algorithm:

- **Task:** play football
- **Performance:** score evolution
- **Experience:**
 - current environment and outcome,
 - past games

Three Kinds of Learning

Data Science



Unsupervised Learning

- Task:** Clustering/DR
- Performance:** Quality
- Experience:** Raw dataset
(No Ground Truth)

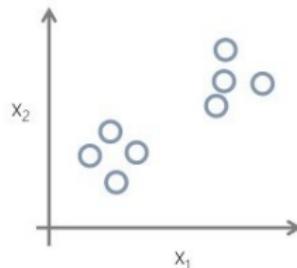
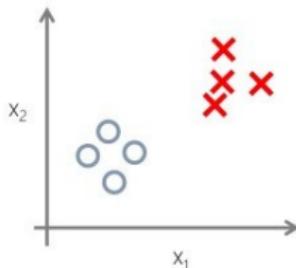
Supervised Learning

- Task:** Prediction
- Performance:** Average error
- Experience:** Predictions
(Ground Truth)

Reinforcement Learning

- Task:** Action
- Performance:** Total reward
- Experience:** Reward from env.
(Interact. with env.)

- Timing:** Offline/Batch (learning from past data) vs Online (continuous learning)
- Implicit stationarity assumption:** Tomorrow is the same as yesterday!



Supervised Learning (Imitation)

- **Goal:** Learn a function f predicting a variable Y from an individual \underline{X} .
- **Data:** Learning set with labeled examples (\underline{X}_i, Y_i)
- **Assumption:** Future data behaves as past data!
- **Predicting is not explaining!**

Unsupervised Learning (Structure Discovery)

- **Goal:** Discover a structure within a set of individuals (\underline{X}_i) .
- **Data:** Learning set with unlabeled examples (\underline{X}_i)
- Unsupervised learning is not a well-posed setting....

Machine Can and Cannot



Machine Can

- Forecast (Prediction based on the past)
 - Detect some changes
 - Memorize/Reproduce
 - Take a decision very quickly
 - Learn from huge dataset
 - Optimize a single task
 - Replace/Help some humans

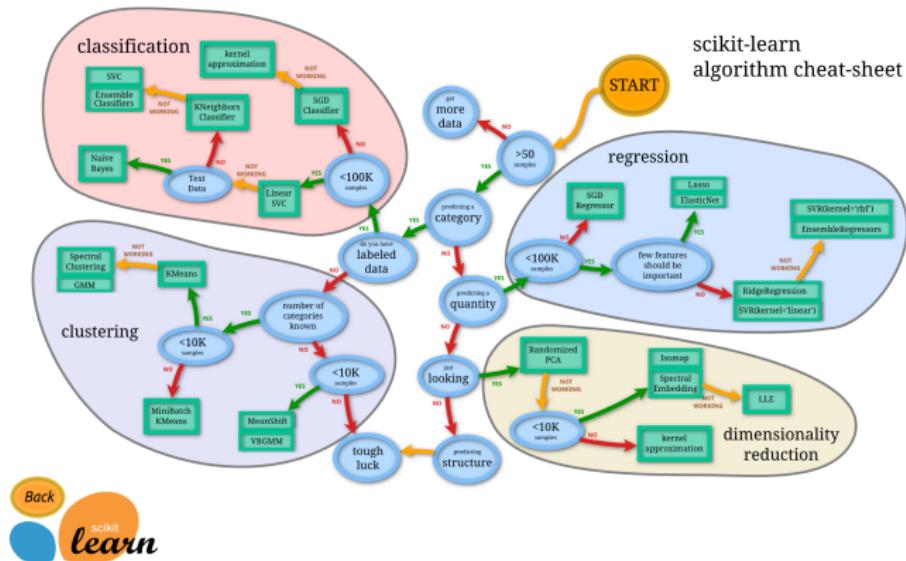
Machine Cannot

- Predict something never seen before
 - Detect any new behaviour
 - Create something brand new
 - Understand the world
 - Get smart really fast
 - Go beyond their task
 - Kill all humans

- Some progresses but still very far from the *singularity*...

Machine Learning

Data Science

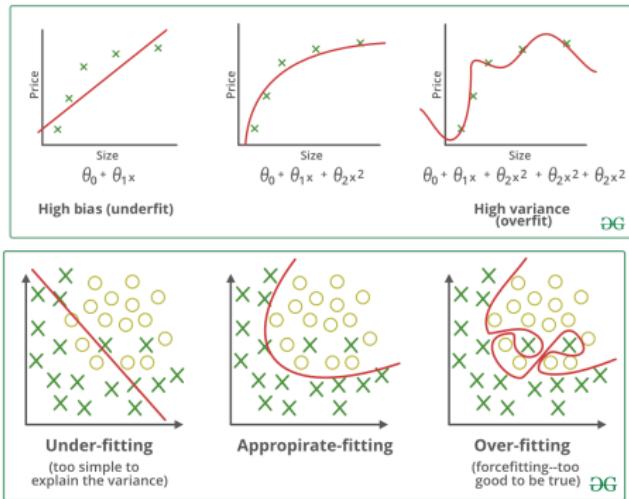


ML Methods

- Huge catalog of methods,
 - Need to define the performance,
 - Feature design...

Under and Over Fitting

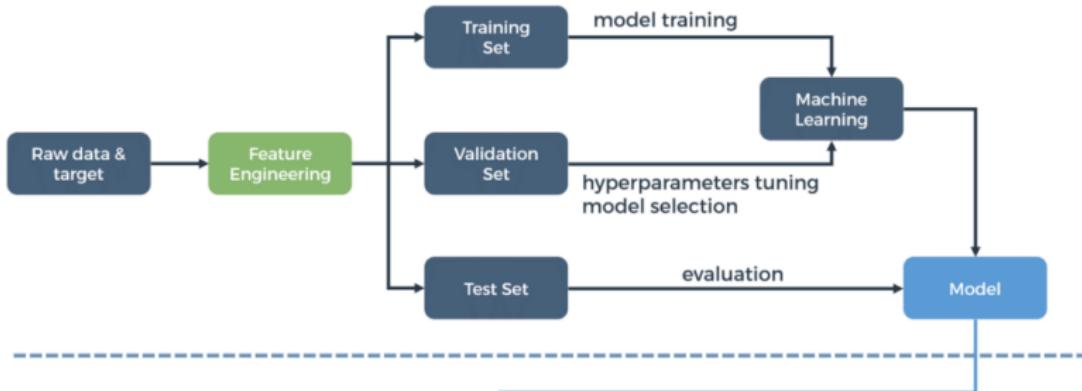
Data Science



Finding the Right Complexity

- What is best?
 - A simple model that is stable but false? (*oversimplification*)
 - A very complex model that could be correct but is unstable? (*conspiracy theory*)
- Neither of them: tradeoff that depends on the dataset.

TRAINING

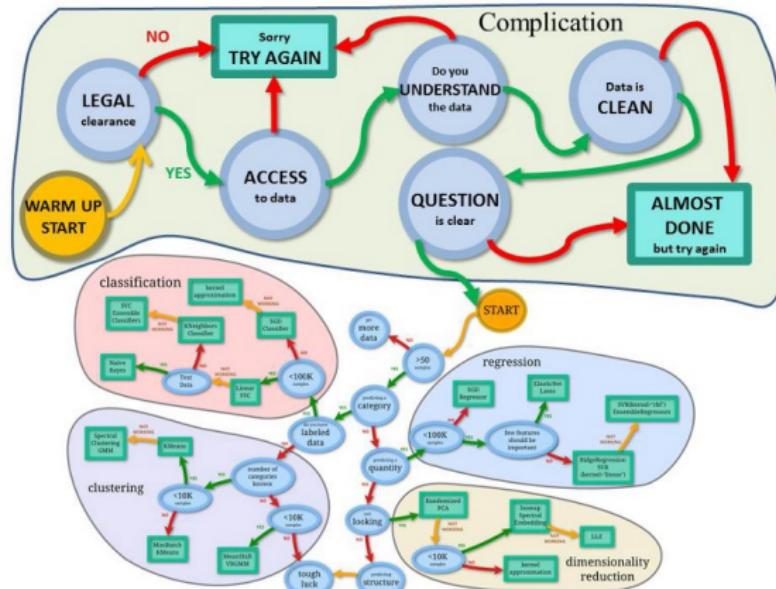


PREDICTING



Learning pipeline

- Test and compare models.
- Deployment pipeline is different!



Main DS difficulties

- Figuring out the problem,
- Accessing the data,
- Not the ML part!

1 Data Science

- AI? Big Data? BI? Statistics? ML? DS?
- Data Science Ecosystem
- Data Cycle
- A Focus on Machine Learning

2 Data Science Toolbox

- Computing and Distribution
- Database
- Data Science Languages
- DevOps

3 Data Scientists and Challenges

- Data Scientists
- Data Science Challenges

4 References

1 Data Science

- AI? Big Data? BI? Statistics? ML? DS?
- Data Science Ecosystem
- Data Cycle
- A Focus on Machine Learning

2 Data Science Toolbox

- Computing and Distribution
- Database
- Data Science Languages
- DevOps

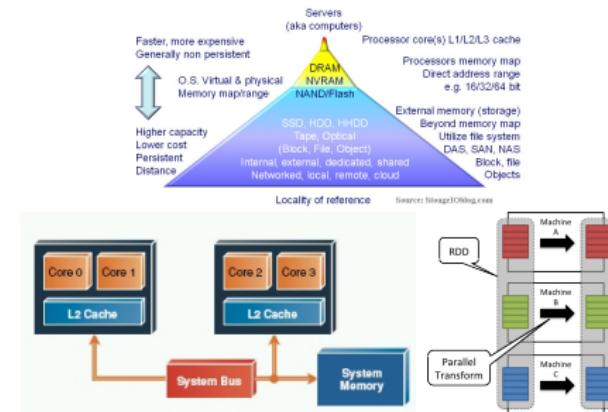
3 Data Scientists and Challenges

- Data Scientists
- Data Science Challenges

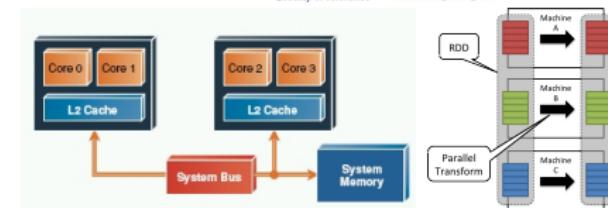
4 References

Big Data?

Data Science Toolbox



Source: StorageIOblog.com



Hardware Constraints

- All the computations are done in a core using data stored somewhere nearby.
- Constraints:
 - Data access / storage (Locality of Reference).
 - Multiple core architecture (Parallelization).
 - Cluster (Distribution)

Possible Issue

- Coding issue?
- IO issue?
- Processing issue?
- Data storage issue?

Enhancement?

- Better algorithm/language/library? (code optimization)
- Better memories usage? (locality of reference)
- Better CPU usage? (parallelization)
- More computers? (distribution)
- Better computing infrastructure? (hardware)

What could be limiting?

Data Science Toolbox



Possible Issue

- **Coding issue?**
- IO issue?
- Processing issue?
- Data storage issue?

Enhancement?

- **Better algorithm/language/library? (code optimization)**
- Better memories usage? (locality of reference)
- Better CPU usage? (parallelization)
- More computers? (distribution)
- Better computing infrastructure? (hardware)

Possible Issue

- Coding issue?
- IO issue?
- Processing issue?
- Data storage issue?

Enhancement?

- Better algorithm/language/library? (code optimization)
- Better memories usage? (locality of reference)
- Better CPU usage? (parallelization)
- More computers? (distribution)
- Better computing infrastructure? (hardware)

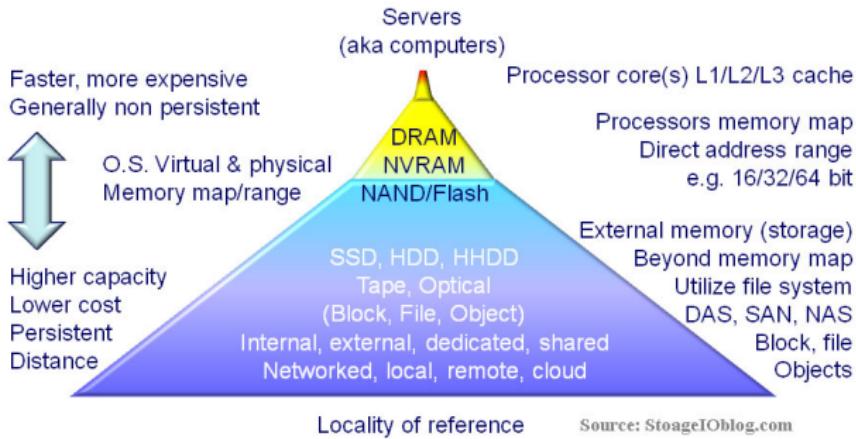
What could be limiting?

Possible Issue

- Coding issue?
- **IO issue?**
- Processing issue?
- Data storage issue?

Enhancement?

- Better algorithm/language/library? (code optimization)
- **Better memories usage? (locality of reference)**
- Better CPU usage? (parallelization)
- More computers? (distribution)
- Better computing infrastructure? (hardware)



Size hierarchy

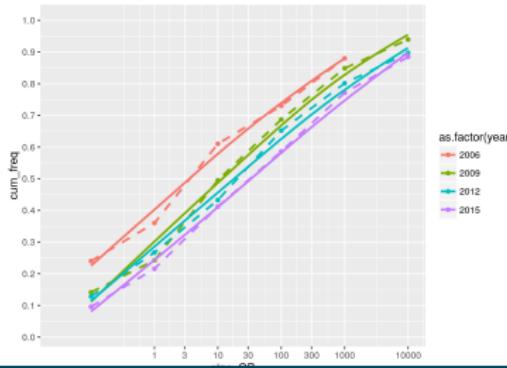
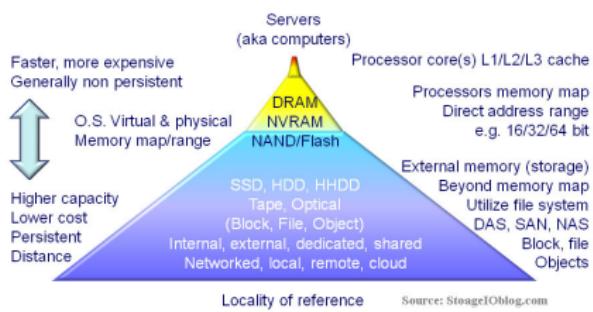
CPU register	64 b × 16
Level 1 cache access	8-128 kb
Level 2 cache access	32-1024 kb
Level 3 cache access	1-8 MB
Main memory access	4-64 GB
Solid-state disk I/O	250 GB - 4 TB
Rotational disk I/O	500 GB - 8 TB

Speed hierarchy

1 CPU cycle	0.3 ns	1 s
Level 1 cache access	0.9 ns	3 s
Level 2 cache access	2.8 ns	9 s
Level 3 cache access	12.9 ns	43 s
Main memory access	120 ns	6 min
Solid-state disk I/O	50-150 μ s	2-6 days
Rotational disk I/O	1-10 ms	1-12 months
Internet: SF to NYC	40 ms	4 years
Internet: SF to UK	81 ms	8 years
Internet: SF to Australia	183 ms	19 years
OS virtualization reboot	4 s	423 years
SCSI command time-out	30 s	3000 years
Hardware virtualization reboot	40 s	4000 years
Physical system reboot	5 m	32 millenia

Locality Of Reference

Data Science Toolbox



Memory Issue

- Data should be as **close** as possible from the core.
- **Ideal case:** dataset in the **memory of a single computer**.
- **Useless** if data used only once... (bottleneck = disk)
- **Split and Apply:** split the data in piece and work independently on each piece.
- Memory required may be **larger** than dataset (interactions...)
- Memory growth **faster** than data growth (Death of big data?)

Possible Issue

- Coding issue?
- IO issue?
- Processing issue?
- Data storage issue?

Enhancement?

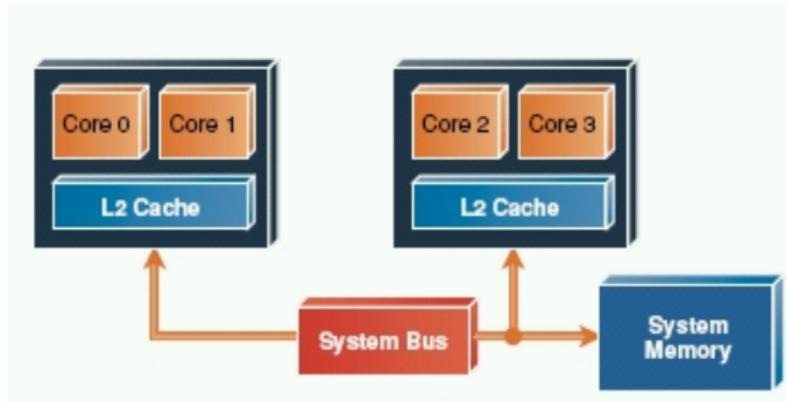
- Better algorithm/language/library? (code optimization)
- Better memories usage? (locality of reference)
- Better CPU usage? (parallelization)
- More computers? (distribution)
- Better computing infrastructure? (hardware)

Possible Issue

- Coding issue?
- IO issue?
- **Processing issue?**
- Data storage issue?

Enhancement?

- Better algorithm/language/library? (code optimization)
- Better memories usage? (locality of reference)
- **Better CPU usage? (parallelization)**
- More computers? (distribution)
- Better computing infrastructure? (hardware)



Speed Issue

- **Parallelization:** Modern computer have **several cores**.
- **HPC / DS setting:** CPU bound tasks / IO bound tasks.
- **Data science:** Often **embarrassingly parallel** setting (no interaction between tasks).
- Not always acceleration due to **IO limitation!**

Possible Issue

- Coding issue?
- IO issue?
- Processing issue?
- Data storage issue?

Enhancement?

- Better algorithm/language/library? (code optimization)
- Better memories usage? (locality of reference)
- Better CPU usage? (parallelization)
- More computers? (distribution)
- Better computing infrastructure? (hardware)

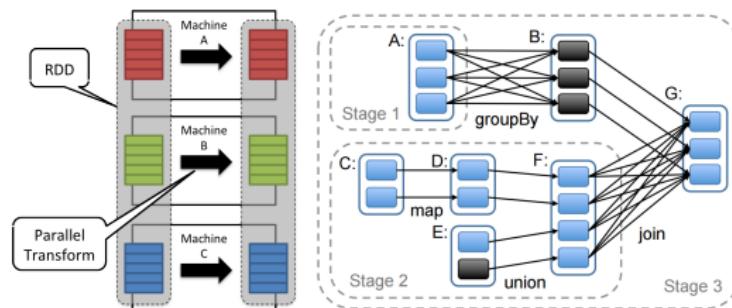
What could be limiting?

Possible Issue

- Coding issue?
- **IO issue?**
- **Processing issue?**
- **Data storage issue?**

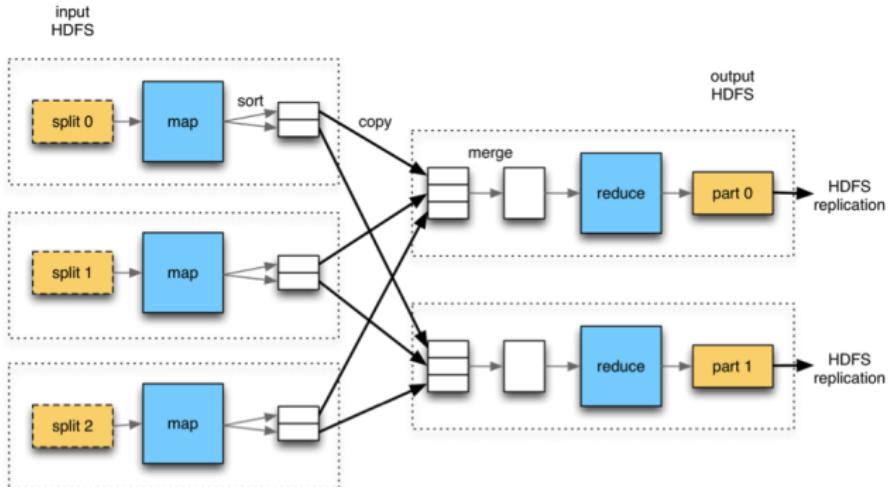
Enhancement?

- Better algorithm/language/library? (code optimization)
- Better memories usage? (locality of reference)
- Better CPU usage? (parallelization)
- **More computers? (distribution)**
- Better computing infrastructure? (hardware)



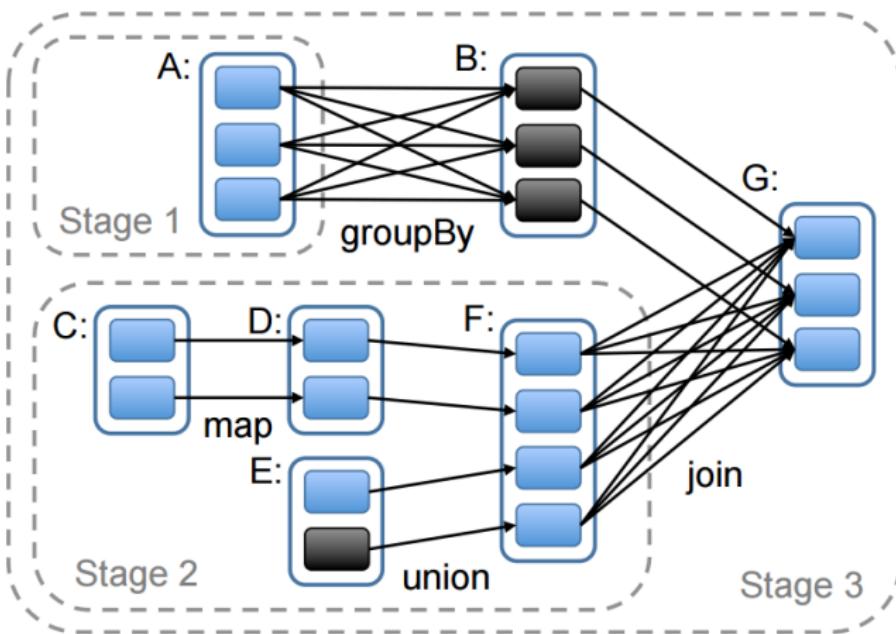
True Big Data Setting

- Computation in a **cluster**:
 - Distribution of the **data**(DS),
 - or/and distribution of the **computation** (HPC)
- **Hadoop/Spark** realm.
- Locally **parallel in memory** computation are faster... if data used more than once.
- Real **challenge** when **not embarrassingly parallel** (interaction...)



Hadoop

- Implementation of (classical) Map/Reduce algorithm.
- Data transfer through disk and networked file system!
- Main contribution: Node failure handling and ecosystem.



Spark

- More flexible algorithm structure (DAG).
- In Memory: cache some objects in memory...

Possible Issue

- Coding issue?
- IO issue?
- Processing issue?
- Data storage issue?

Enhancement?

- Better algorithm/language/library? (code optimization)
- Better memories usage? (locality of reference)
- Better CPU usage? (parallelization)
- More computers? (distribution)
- Better computing infrastructure? (hardware)

Possible Issue

- Coding issue?
- **IO issue?**
- **Processing issue?**
- **Data storage issue?**

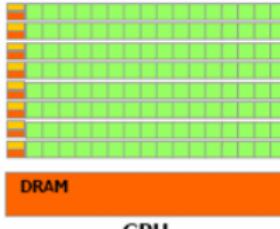
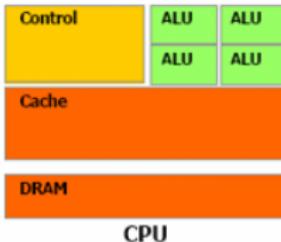
Enhancement?

- Better algorithm/language/library? (code optimization)
- Better memories usage? (locality of reference)
- Better CPU usage? (parallelization)
- More computers? (distribution)
- **Better computing infrastructure? (hardware)**



RAM and SSD

- The larger and the faster the better...
- Quite cheap nowadays.

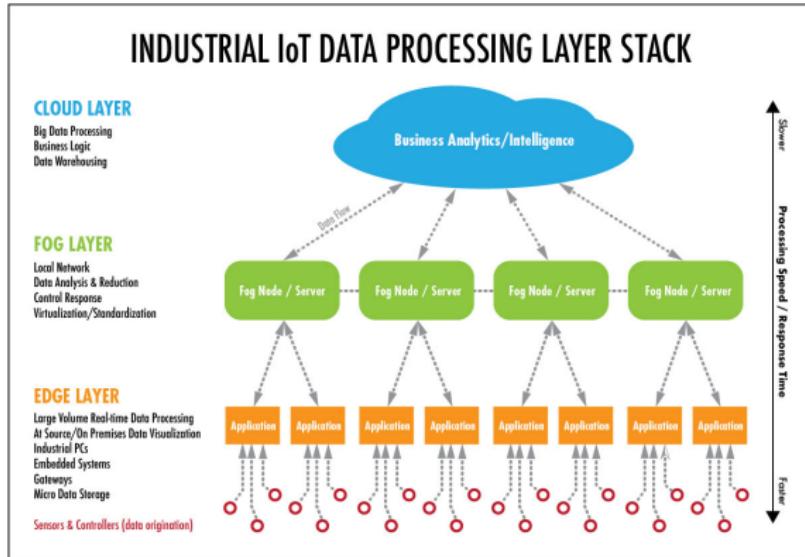


PU: CPU, GPU, FPGA, ASICS

- More than one processor architecture.
- Flexibility vs performance.
- Parallelism: $\text{CPU} < \text{GPU} < \text{FPGA} < \text{ASICS}$.

Cluster

- More computers...
- IO is important!



IOT: Computing and Bandwidth constraints

- Hierarchical distributed computing scheme.
- Do as much as possible locally to reduce communication issues (bandwidth/latency).

1 Data Science

- AI? Big Data? BI? Statistics? ML? DS?
- Data Science Ecosystem
- Data Cycle
- A Focus on Machine Learning

2 Data Science Toolbox

- Computing and Distribution
- Database
- Data Science Languages
- DevOps

3 Data Scientists and Challenges

- Data Scientists
- Data Science Challenges

4 References

Possible Issue

- Coding issue?
- IO issue?
- Processing issue?
- Data storage issue?

Enhancement?

- Better algorithm/language/library? (code optimization)
- Better memories usage? (locality of reference)
- Better CPU usage? (parallelization)
- More computers? (distribution)
- Better computing infrastructure? (hardware)

Possible Issue

- Coding issue?
- IO issue?
- Processing issue?
- **Data storage issue?**

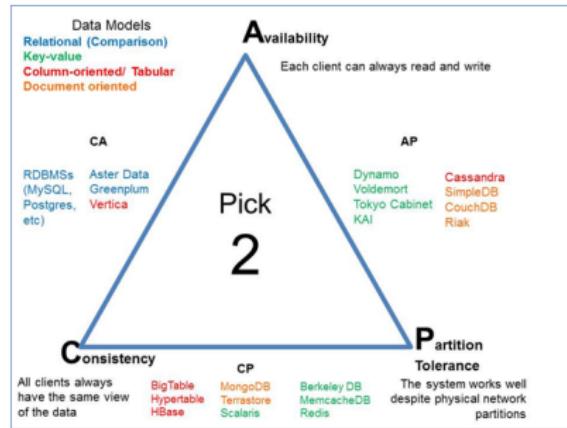
Enhancement?

- Better algorithm/language/library? (code optimization)
- Better memories usage? (locality of reference)
- Better CPU usage? (parallelization)
- More computers? (distribution)
- Better computing infrastructure? (hardware)
- **Better data storage? (database)**



(SQL?) Databases

- Most convenient tool to store/access data.
- Abstraction of the implementation that eases the use.
- Lot of knowledge inside.



SQL

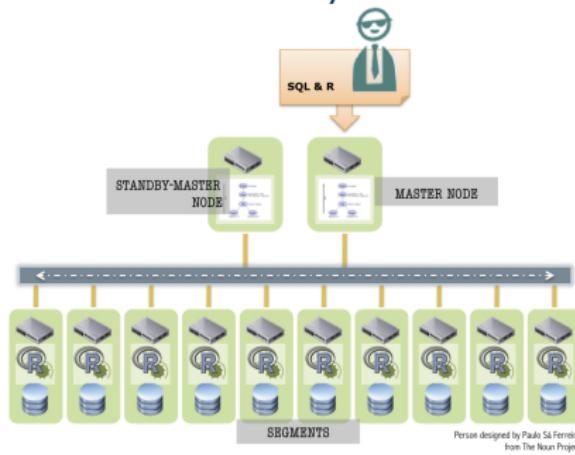
- Most classical design,
- Limitation due to the CAP theorem.
- Hard to distribute without asking less...

NoSQL (Not only SQL!)

- Relaxation on the consistency to ease distribution.
- Simplification/modification of the stored data to ease the use.

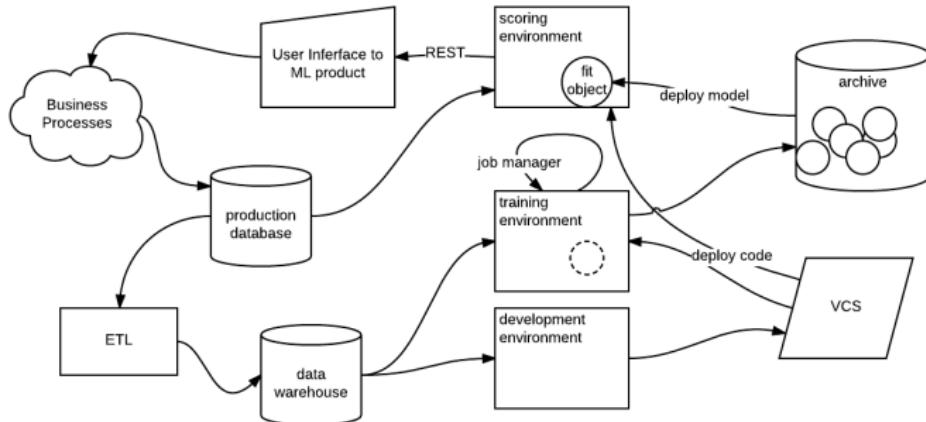
UDF: DB as a Distributed System

Data Science Toolbox



Database and User Defined Function

- Allow to define complicated function that can be run in the server of the DB.
- Idea: minimize the data transport by moving only the answer.
- PostgreSQL, SqlServer, Oracle, Teradata, HAWQ, SAP Hana...
- Require some privileges...



Data Science Architecture

- Usage dependent architecture!
- **Finding a good architecture is complex**

1 Data Science

- AI? Big Data? BI? Statistics? ML? DS?
- Data Science Ecosystem
- Data Cycle
- A Focus on Machine Learning

2 Data Science Toolbox

- Computing and Distribution
- Database
- **Data Science Languages**
- DevOps

3 Data Scientists and Challenges

- Data Scientists
- Data Science Challenges

4 References

DS Toolkit

- Several tools

- Script/Code
- R/Python
- Glue to libraries
- ETL, ML, Viz,
Report..
- Limited size

Language/
Environment

DS Toolkit

- Several tools

Database

- Database...
- SQL/NoSQL
- Data Storage.
- ETL
- Very limited algorithmic

DS Toolkit

- Several tools

Distributed System

- Distributed Computing
- Hadoop/Spark
- Big Data!
- ETL, ML, Cleaning
- Limited distributed algorithmic

DS Toolkit

- Several tools

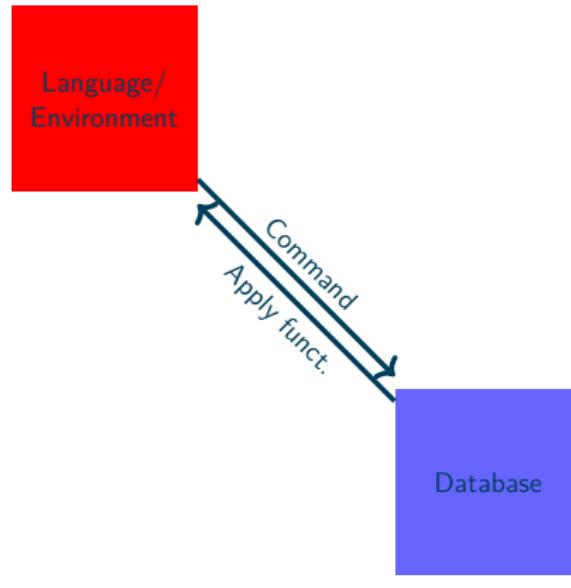
- Graphical Interface / API
- Javascript, D3.js / REST, SOAP
- User Interface / Prog. Interface
- Viz, Report / Analytics
- Limited flexibility, reproducibility / Specific use

GUI / API

DS Toolkit

- Several tools

- Script/Code
- R/Python
- Glue to libraries
- ETL, ML, Viz,
Report..
- Limited size



- Database...
- SQL/NoSQL
- Data Storage.
- ETL
- Very limited
algorithmic

DS Toolkit

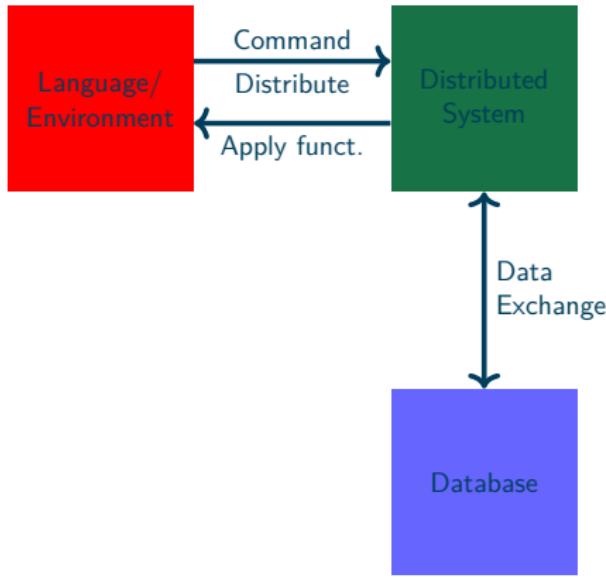
- Several tools working together

Data Science Tools

Data Science Toolbox



- Script/Code
- R/Python
- Glue to libraries
- ETL, ML, Viz, Report..
- Limited size



- Distributed Computing
 - Hadoop/Spark
 - Big Data!
 - ETL, ML, Cleaning
 - Limited distributed algorithmic
-
- Database...
 - SQL/NoSQL
 - Data Storage.
 - ETL
 - Very limited algorithmic

DS Toolkit

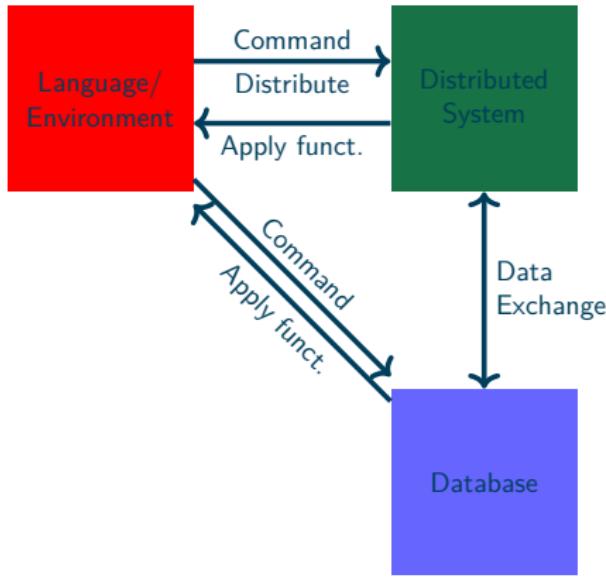
- Several tools working together

Data Science Tools

Data Science Toolbox



- Script/Code
- R/Python
- Glue to libraries
- ETL, ML, Viz, Report..
- Limited size



- Distributed Computing
 - Hadoop/Spark
 - Big Data!
 - ETL, ML, Cleaning
 - Limited distributed algorithmic
-
- Database...
 - SQL/NoSQL
 - Data Storage.
 - ETL
 - Very limited algorithmic

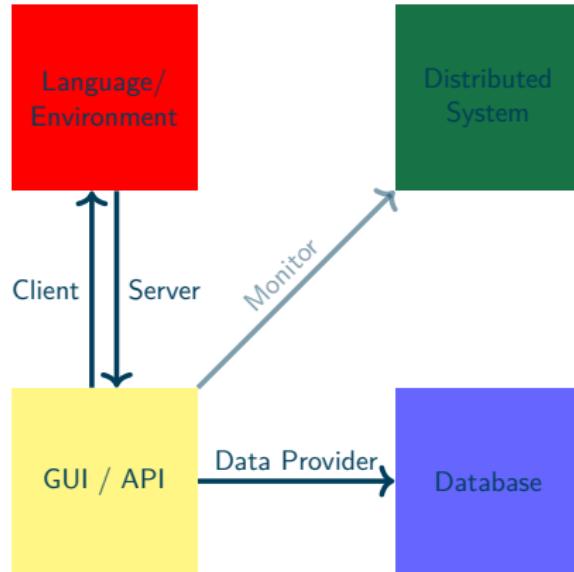
DS Toolkit

- Several tools working together

Data Science Tools

- Script/Code
- R/Python
- Glue to libraries
- ETL, ML, Viz, Report..
- Limited size

- Graphical Interface / API
- Javascript, D3.js / REST, SOAP
- User Interface / Prog. Interface
- Viz, Report / Analytics
- Limited flexibility, reproducibility / Specific use



- Distributed Computing
- Hadoop/Spark
- Big Data!
- ETL, ML, Cleaning
- Limited distributed algorithmic

- Database...
- SQL/NoSQL
- Data Storage.
- ETL
- Very limited algorithmic

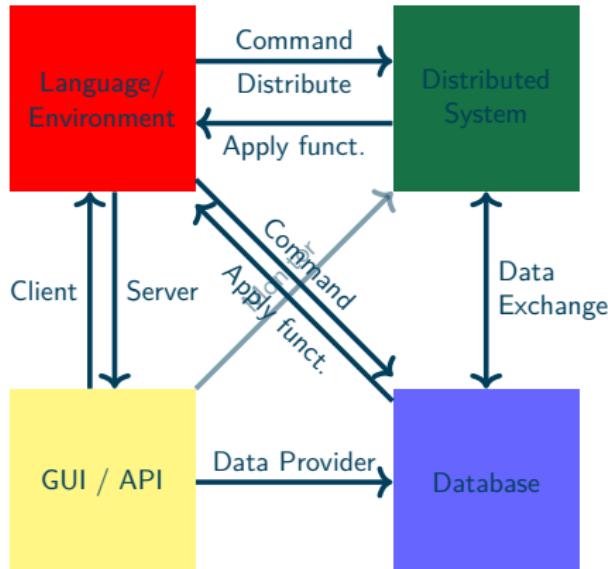
DS Toolkit

- Several tools working together

Data Science Tools

- Script/Code
- R/Python
- Glue to libraries
- ETL, ML, Viz, Report..
- Limited size

- Graphical Interface / API
- Javascript, D3.js / REST, SOAP
- User Interface / Prog. Interface
- Viz, Report / Analytics
- Limited flexibility, reproducibility / Specific use



- Distributed Computing
- Hadoop/Spark
- Big Data!
- ETL, ML, Cleaning
- Limited distributed algorithmic

- Database...
- SQL/NoSQL
- Data Storage.
- ETL
- Very limited algorithmic

DS Toolkit

- Several tools working together



Polyglot

- Datascience languages:
 - Python,
 - R,
 - Julia???
- What differences?



Python

- Project initiated by G. Von Rossum at the beginning of the 90's.
- Interpreted language written in C.
- Widely used in various area.
- Clear and very simple syntax.
- Can be interfaced with a lot of languages.
- Allow to write concise program with a high level of abstraction.
- Available on Unix, Windows, OS X...



R

- Project initiated by R. Ihaka and R. Gentleman at the beginning of the 90's.
- Interpreted language written in C and Fortran.
- Widely used in various data science area.
- Can be interfaced with a lot of languages.
- Allow to write concise program with a high level of abstraction.
- Available on Unix, Windows, OS X...



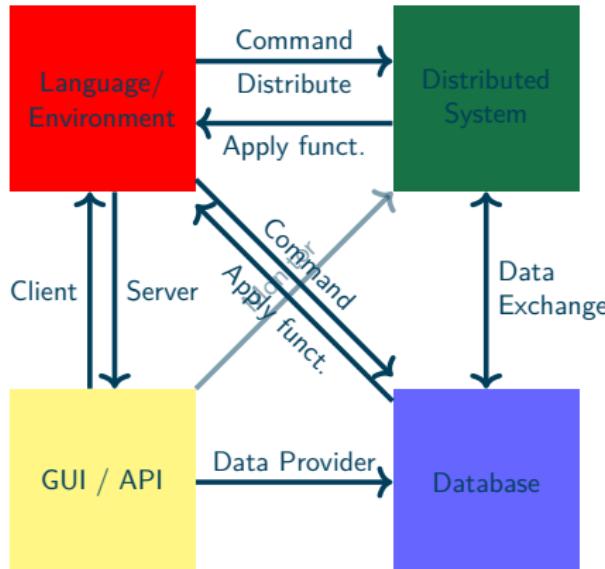
Julia

- Project initiated at the MIT in the 10's.
- Interpreted language written in C and C++.
- Clear and simple syntax (with type declaration).
- Can be interfaced with most classical language.
- Allow to write concise program with a high level of abstraction.
- Available on Unix, Windows, OS X....
- JIT compiler
- Easy HPC computing?!

Data Science Tools

- Script/Code
- R/Python
- Glue to libraries
- ETL, ML, Viz, Report..
- Limited size

- Graphical Interface / API
- Javascript, D3.js / REST, SOAP
- User Interface / Prog. Interface
- Viz, Report / Analytics
- Limited flexibility, reproducibility / Specific use



- Distributed Computing
- Hadoop/Spark
- Big Data!
- ETL, ML, Cleaning
- Limited distributed algorithmic

- Database...
- SQL/NoSQL
- Data Storage.
- ETL
- Very limited algorithmic

DS Toolkit

- Several tools working together
- Language: R, Python or Julia?

Folklore Comparison

- R is the language of choice for exploration/prototyping/POC
- Python can scale up to a software (gui...)
- Julia not very common

Data Analysis Comparison

- **Data type:**
 - R and Python can handle quantitative and qualitative variables.
 - R handles natively missing data.
- **ML Library:**
 - Python has a homogeneous ML library (`scikit-learn`)
 - R has much more libraries.
- **Almost equivalent libraries:**
 - for visualization, data manipulation...
 - distributed computation (Hadoop and Spark or HPC).
- **Choice mostly based on the local ecosystem (software and users)**

1 Data Science

- AI? Big Data? BI? Statistics? ML? DS?
- Data Science Ecosystem
- Data Cycle
- A Focus on Machine Learning

2 Data Science Toolbox

- Computing and Distribution
- Database
- Data Science Languages
- DevOps

3 Data Scientists and Challenges

- Data Scientists
- Data Science Challenges

4 References

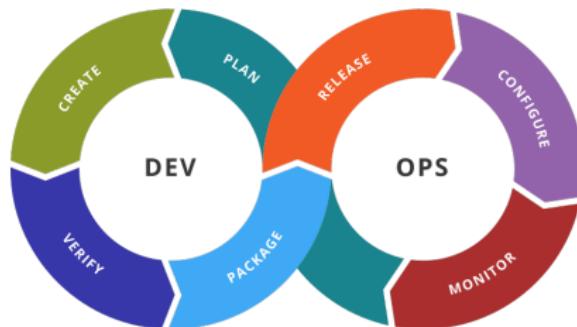
DevOps and ML?

Data Science Toolbox



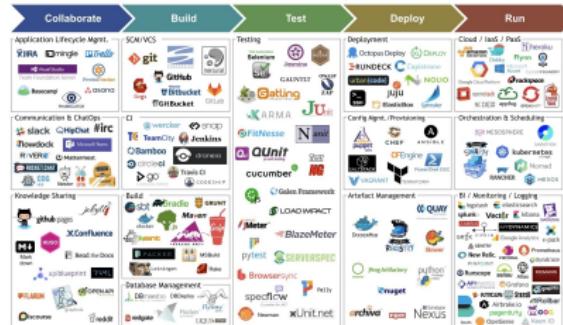
DevOps

- Combination of Software Development and IT Operations.
 - *a set of practices intended to reduce the time between committing a change to a system and the change being placed into normal production, while ensuring high quality*
 - Combine tools and mindset!
 - Directly applicable to most ML projects.



Much more than technical tools!

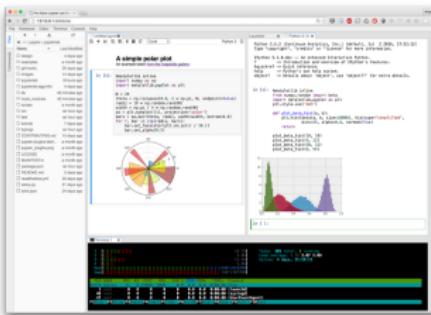
- **Culture:** Cooperation / Learning / Blamelessness / Empowerment
- **Automation:** Tools / Tests / Package / Configuration
- **Monitoring:** Dashboard / Post Mortem
- **Sharing:** Goals / Practice / Learning



Lots of tools for each step!

- **Collaborate:** Lifecycle mgmt, Communication, Knowledge sharing
- **Build:** SCM/VCS, CI, Build, DB mgmt
- **Test:** Testing
- **Deploy:** Deployment, Config mgmt, Artefact mgmt
- **Run:** Cloud/*aas, Orchestration, Monitoring

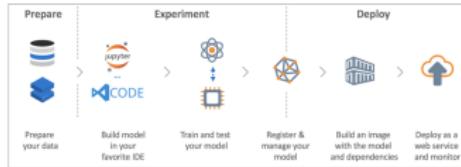
- Tool choice depends on the context.
- Good usage is more important than the tool itself.



- Code are meant to be used/shared/reused.

Good practice

- Versioning,
- Documentation,
- Testing,
- Packaging,
- Training,
- Continuous integration?



- Models are meant to be used/shared/reused.

Good practice

- Versioning (dataset/models),
- Artifact mgmt,
- Documentation,
- Testing/Monitoring,
- Training,
- Continuous integration.
- Transfer learning...

1 Data Science

- AI? Big Data? BI? Statistics? ML? DS?
- Data Science Ecosystem
- Data Cycle
- A Focus on Machine Learning

2 Data Science Toolbox

- Computing and Distribution
- Database
- Data Science Languages
- DevOps

3 Data Scientists and Challenges

- Data Scientists
- Data Science Challenges

4 References

1 Data Science

- AI? Big Data? BI? Statistics? ML? DS?
- Data Science Ecosystem
- Data Cycle
- A Focus on Machine Learning

2 Data Science Toolbox

- Computing and Distribution
- Database
- Data Science Languages
- DevOps

3 Data Scientists and Challenges

- Data Scientists
- Data Science Challenges

4 References

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

DOMAIN KNOWLEDGE & SOFT SKILLS

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



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 withaaS like AWS

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 visualization tools e.g. Flare, D3.js, Tableau

MarketingDistillery.com is a group of practitioners in the area of e-commerce marketing. Our fields of expertise include marketing strategy and optimization, customer tracking and analytics, predictive analytics and econometrics, data warehousing and big data systems, marketing channel insights in Paid Search, SEO, Social, CRM and brand.

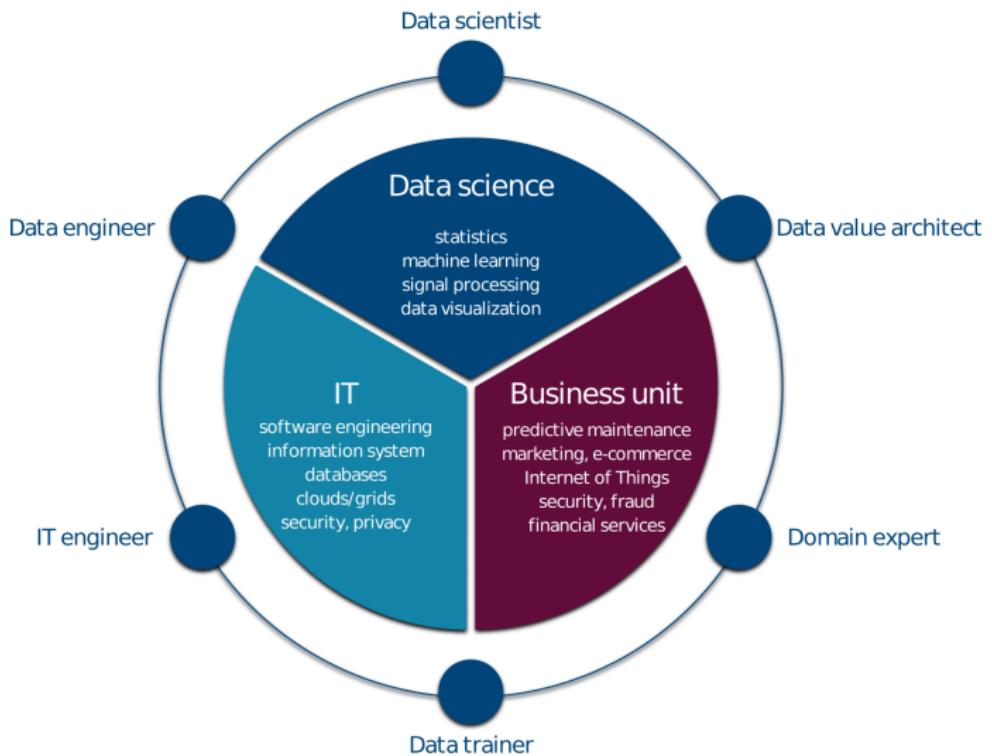
Marketing
DISTILLERY
40 experts | 20 countries

Data Scientist

- **Mix** of various skills.
- **Hard to be an expert of everything!**

More Than One Type Of Data Jobs!

DS / Challenges



Much More Variety?

DS / Challenges



DATA SCIENTIST
AS RARE AS UNICORNS

Languages
R, SAS, Python, Matlab, SQL, Hive, Pig, Spark

Skills & Talents

- Distributed computing
- Predictive modeling
- Story-telling and visualizing
- Math, Stats, Machine Learning

Role
Cleans, manages and organizes big data

Mindset
Curious data wizard

Hired By
Google, Microsoft, IBM

STATISTICIAN
HISTORIC LEADERS OF DATA

Languages
R, SAS, SPSS, Matlab, Scala, Python, Perl, Hive, Pig, Spark, SQL

Skills & Talents

- Statistical theory & methodology
- Data mining & machine learning
- Distributed Computing (Hadoop)
- Database systems (SQL and NO SQL based)
- Cloud tools

Role
Collects, analyzes and interprets qualitative as well as quantitative data with statistical theories and methods

Mindset
Logical and entrepreneurial stats genius

Hired By
LinkedIn, Johnson & Johnson, PEPSICO

DATA ANALYST
DATA DETECTIVE

Languages
R, Python, HTML, JavaScript, C/C++, SQL

Skills & Talents

- Spreadsheet tools (e.g. Excel)
- Database systems (SQL and NO SQL based)
- Communication & visualization
- Math, Stats, Machine Learning

Role
Collects, processes and performs statistical data analyses

Mindset
Intuitive data junkie with high "figure-it-out" quotient

Hired By
IBM, DHL

DATA ENGINEER
SOFTWARE ENGINEER BY TRADE

Languages
SQL, Hive, Pig, R, Matlab, SAS, SPSS, Python, Java, Ruby, C++, Perl

Skills & Talents

- Develops, contracts, tests and maintains data infrastructure (such as databases and large-scale processing systems)
- Database systems (SQL & NO SQL based)
- Data modeling & ETL tools
- Data APIs
- Data warehousing solutions

Role
Develops, contracts, tests and maintains data infrastructure (such as databases and large-scale processing systems)

Mindset
All-purpose everyman

Hired By
Spotify, Facebook, a

DATA ARCHITECT
THE CONTEMPORARY DATA MODELLER

Languages
SQL, XML, Hive, Pig, Spark

Skills & Talents

- Data warehousing solutions
- In-depth knowledge of database architecture
- Extraction Transformation and Load (ETL), spreadsheet and BI tools
- Data modeling
- Systems developer

Role
Creates blueprints for data management systems to integrate, centralize, protect and maintain data sources

Mindset
Inspiring architect with a love for data architecture design patterns

Hired By
VISA, Coca-Cola, logitech

BUSINESS ANALYST
CHANGE AGENT

Languages
SQL

Skills & Talents

- Basic tools (e.g. MS Office)
- Date visualization tools (e.g. Tableau)
- Contextual listening and storytelling
- Business intelligence understanding
- Data modeling

Role
Improves business processes as intermediary between business and IT

Mindset
Resilient project juggler

Hired By
USER, ORACLE

DATA AND ANALYTICS MANAGER
DATA SCIENCE TEAM LEADER

Languages
SQL, R, SAS, Python, Matlab, Java

Skills & Talents

- Database systems (SQL and NO SQL based)
- Leadership & project management
- Interpersonal communication
- Data mining & predictive modeling

Role
Manages a team of analysts and data scientists

Mindset
Data Wizards' Cheerleader

Hired By
GOLDSFIRE, slack, MOTOROLA SOLUTIONS

Several Profiles

- Several type of problem / several type of tools
- Much more variety than this...
- Importance of balanced **teams**.

1 Data Science

- AI? Big Data? BI? Statistics? ML? DS?
- Data Science Ecosystem
- Data Cycle
- A Focus on Machine Learning

2 Data Science Toolbox

- Computing and Distribution
- Database
- Data Science Languages
- DevOps

3 Data Scientists and Challenges

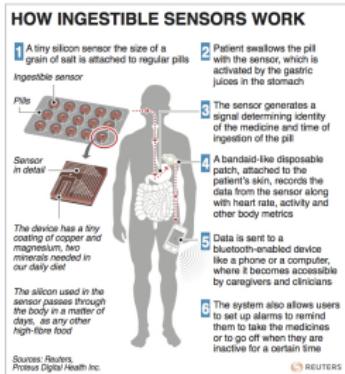
- Data Scientists
- Data Science Challenges

4 References

- Applied math **AND** Computer science
- Huge importance of domain specific knowledge: physics, signal processing, biology, health, marketing...

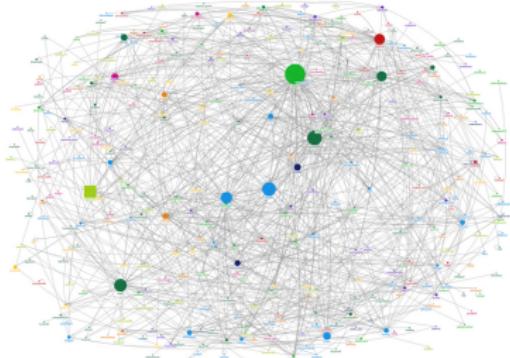
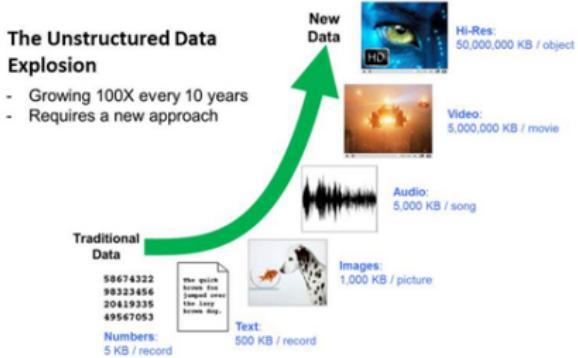
Some joint math/computer science challenges

- Data acquisition
- Unstructured data and their representation
- Huge dataset and computation
- High dimensional data and model selection
- Learning with less supervision
- Visualization
- Software(s)
- Domain specific issue!
- Regulation and ethics



Some challenges

- How to measure new things?
- How to choose what to measure?
- How to deal with distributed sensors?
- How to look for new sources of information?
- How to plan some experiments?

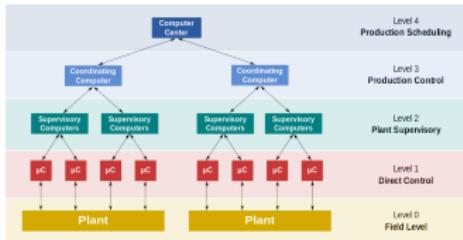
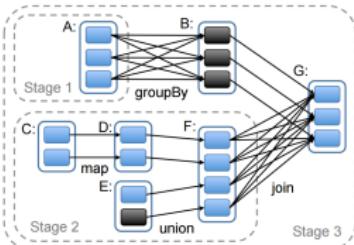
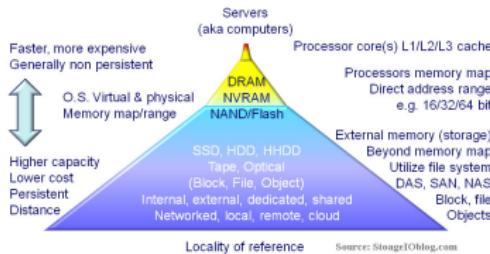


Some challenges

- How to store efficiently the data?
- How to describe (model) them to be able to process them?
- How to combine data of different nature?
- How to learn dynamics?

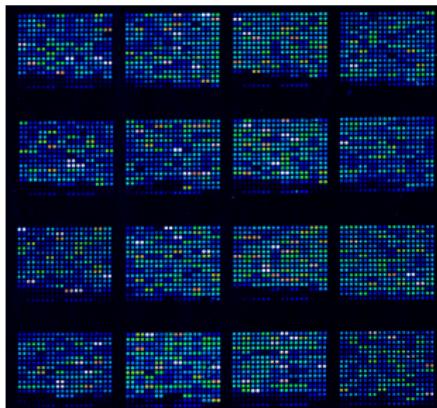
Huge Dataset and Computation

DS / Challenges



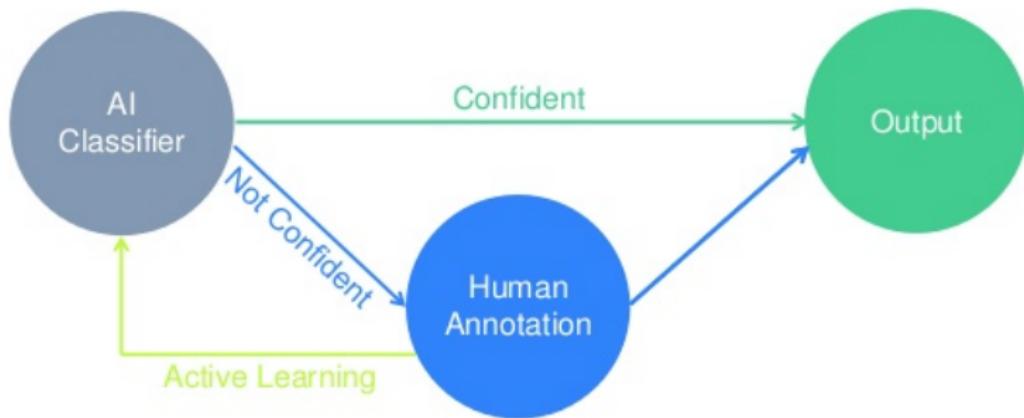
Some challenges

- How to take into account the locality of the data?
- How to take into account hardware constraints?
- How to construct distributed architectures?
- How to design adapted algorithms?



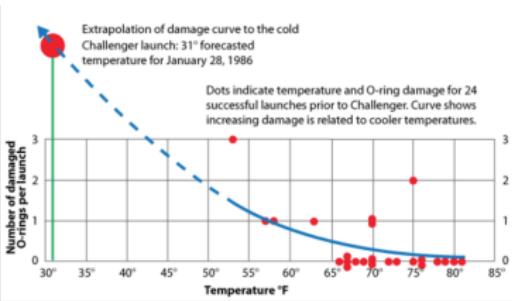
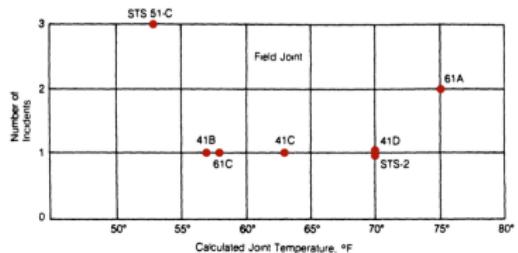
Some challenges

- How to model (describe) the data (and the world)?
- How to incorporate expert knowledge?
- How to avoid the curse of dimensionality?
- How to prescribe the best acquisition strategy?



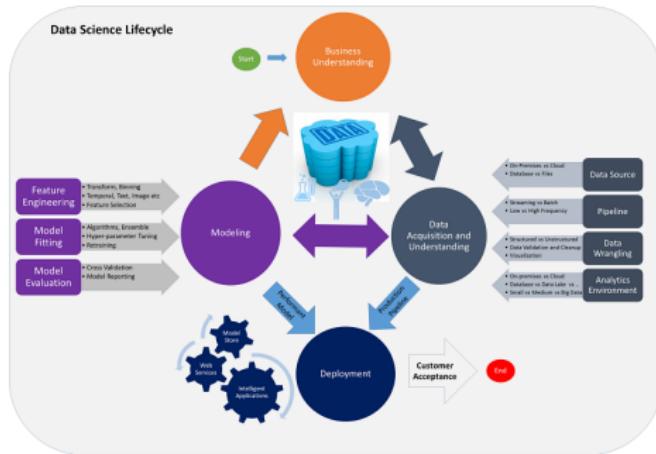
Some challenges

- How to learn with the less possible interactions?
- How to learn simultaneously several related tasks?



Some challenges

- How to look at the data?
- How to present results?
- How to help taking better informed decision?



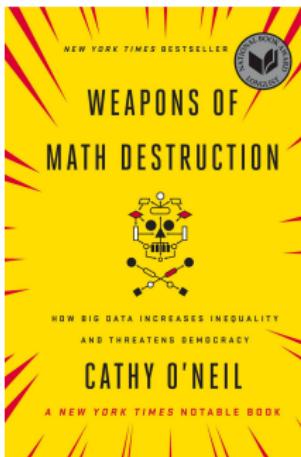
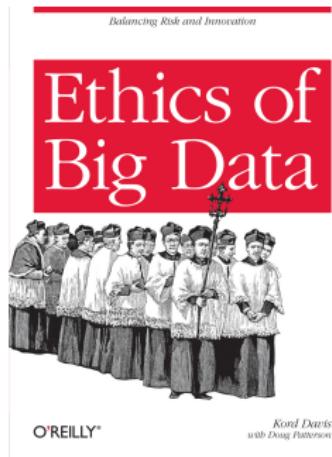
Some challenges

- How to construct a consistent ecosystem?
- How to construct interoperable systems?
- How to build a team of users?



Some challenges

- How to find the real problem at hand?
- How to incorporate human expertise?
- How to measure the performance?



Some challenges

- How to comply with regulation?
- How to answer public concerns?
- How to avoid ethical issues?

1 Data Science

- AI? Big Data? BI? Statistics? ML? DS?
- Data Science Ecosystem
- Data Cycle
- A Focus on Machine Learning

2 Data Science Toolbox

- Computing and Distribution
- Database
- Data Science Languages
- DevOps

3 Data Scientists and Challenges

- Data Scientists
- Data Science Challenges

4 References

References

References



R. Schutt and C. O'Neil.
Doing Data Science: Straight talk from the frontline.
O'Reilly, 2014



T. Hastie, R. Tibshirani, and J. Friedman.
The Elements of Statistical Learning.
Springer Series in Statistics, 2009



V. Kale.
Big Data Computing: a Guide for Business and Technology Manager.
CRC Press, 2016



S. Sakr.
Big Data 2.0 Processing Systems: A Survey.
Springer, 2016



G. Harrison.
Next Generation Databases: NoSQL, NewSQL, and Big Data.
Apress, 2015



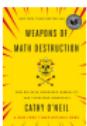
J. Davis and K. Daniels.
Effective DevOps.
O'Reilly, 2016



H. Wickham and G. Grolemund.
R for Data Science.
O'Reilly, 2017



A. Géron.
Hands-On Machine Learning with Scikit-Learn and TensorFlow.
O'Reilly, 2017



C. O'Neil.
Weapons of Math Destruction.
Crown, 2016



K. Davis.
Ethics of Big Data.
O'Reilly, 2012