Introduction to Data Science

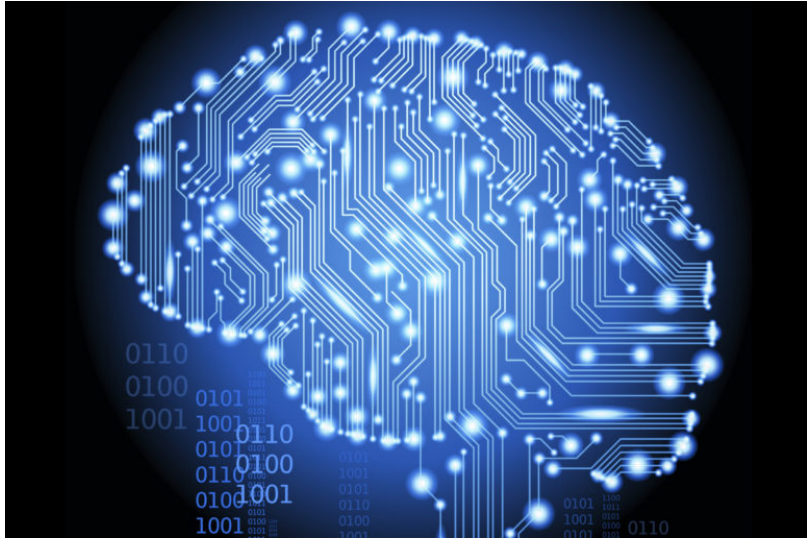
E. Le Pennec - A. Dieuleveut

CA - 08/09/2020

Outline

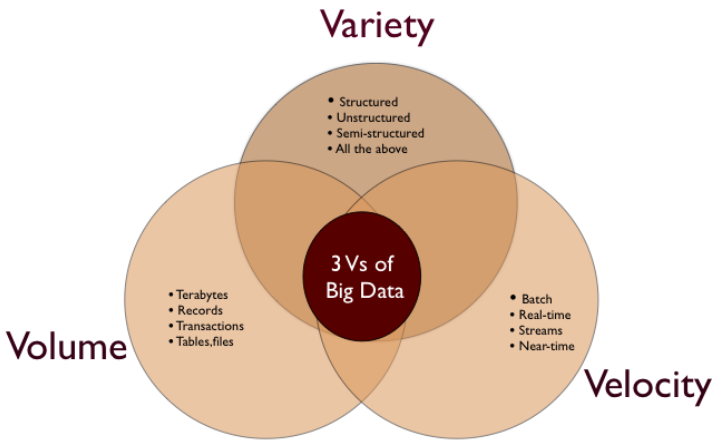
Outline

Artificial Intelligence?



Source: Usine Digitale

Big Data?

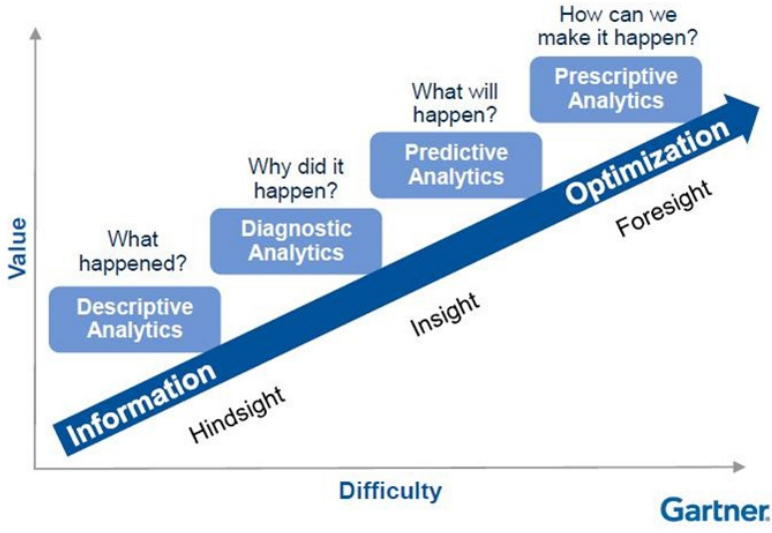


Source: Kavya Muthanna

Business Intelligence?

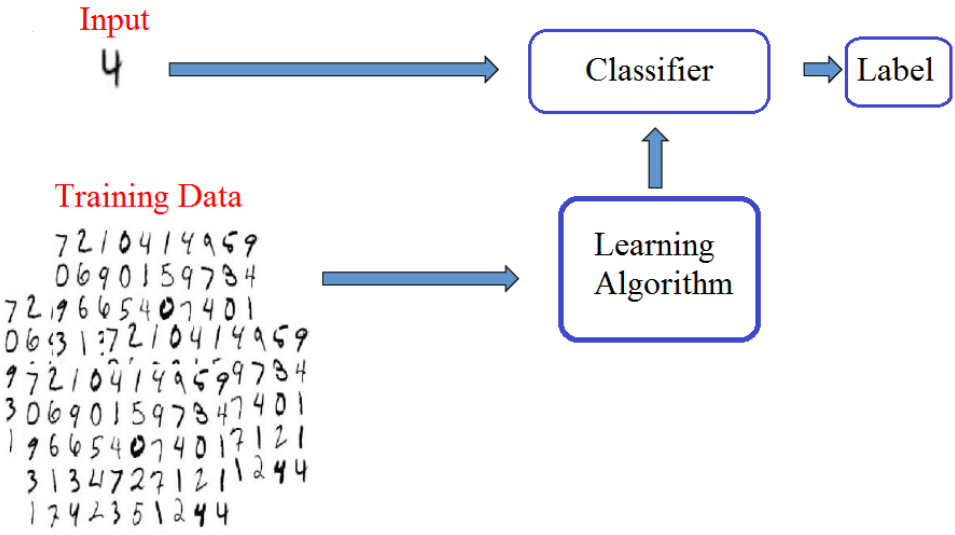
Source: inovaport

Statistics?



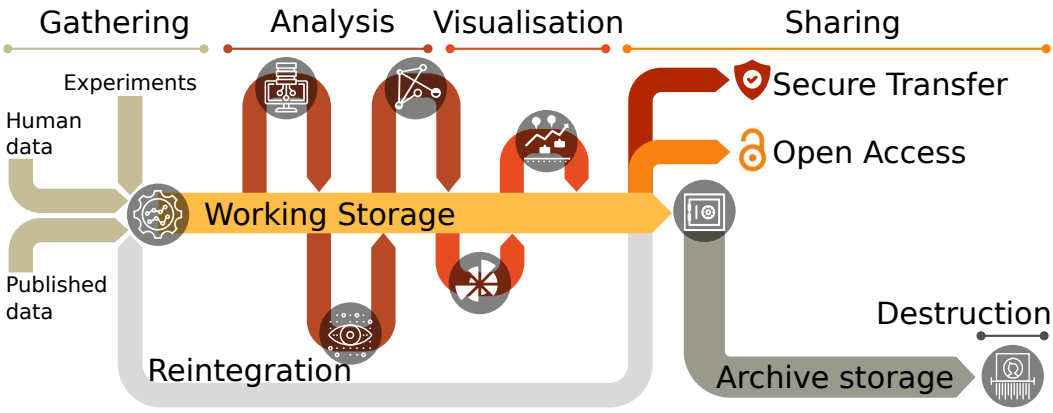
Source: Gartner

Machine Learning?



Source: Unknown

Data Management?



Source: Wikipedia

Data Science?

Big data

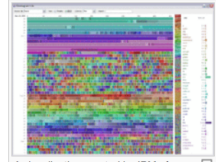
From Wikipedia, the free encyclopedia

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.

Source: Wikipedia

Wikipedia

Big data

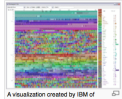
From Wikipedia, the free encyclopedia

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

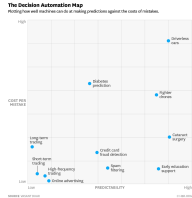


A visualization created by IBM of Wikipedia edits. At multiple terabytes in size, the text and images of Wikipedia are

- **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.
- **Data Management** comprises all disciplines related to managing data as a valuable resource.
- **Data science** is the study of the generalizable extraction of knowledge from data, yet the key word is science.

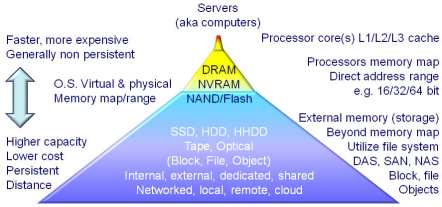
Source: Wikipedia

Wikipedia



- **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.
- **Data Management** comprises all disciplines related to managing data as a valuable resource.
- **Data science** is the study of the generalizable extraction of knowledge from data, yet the key word is science.

Wikipedia



- **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.
- **Data Management** comprises all disciplines related to managing data as a valuable resource.
- **Data science** is the study of the generalizable extraction of knowledge from data, yet the key word is science.

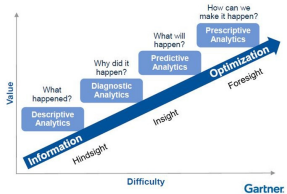
Source: storagenewsletter.com

Wikipedia

- **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.
- **Data Management** comprises all disciplines related to managing data as a valuable resource.
- **Data science** is the study of the generalizable extraction of knowledge from data, yet the key word is science.

Source: Investopedia

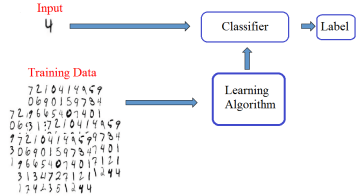
Wikipedia



- **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.
- **Data Management** comprises all disciplines related to managing data as a valuable resource.
- **Data science** is the study of the generalizable extraction of knowledge from data, yet the key word is science.

Source: Gartner

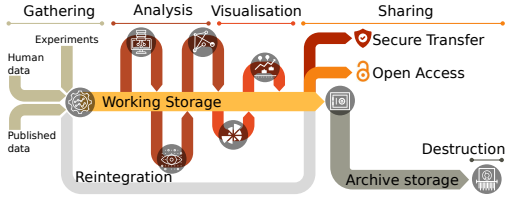
Wikipedia



- **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.
- **Data Management** comprises all disciplines related to managing data as a valuable resource.
- **Data science** is the study of the generalizable extraction of knowledge from data, yet the key word is science.

Source: Unknown

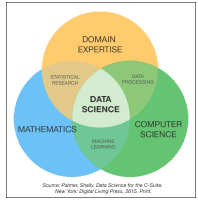
Wikipedia



- **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.
- **Data Management** comprises all disciplines related to managing data as a valuable resource.
- **Data science** is the study of the generalizable extraction of knowledge from data, yet the key word is science.

Source: Wikipedia

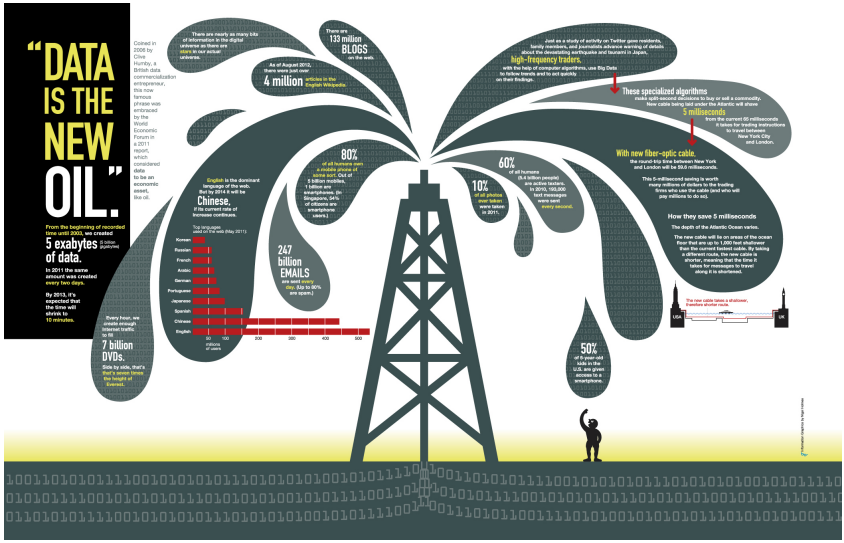
Wikipedia



- **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.
- **Data Management** comprises all disciplines related to managing data as a valuable resource.
- **Data science** is the study of the generalizable extraction of knowledge from data, yet the key word is science.

Source: Dr. Palmer

Data Is The New Oil?



Source: N. Holmes

Outline

Data Driven Approach

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

Data Driven Approach

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...
- Data Management...

Source: D. Donoho, 50 Years of Data Science

Large Scale ML Is (Quite) Easy


```scala
def run(params: Params) {
  val conf = new SparkConf()
    .setAppName(s"BinaryClassification with $params")
  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(s"Training: $numTraining, test: $numTest.")
  examples.unpersist(blocking = false)

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

  val algorithm = new LogisticRegressionWithSGD()
    algorithm.optimizer
      .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(s"Empirical CrossEntropy = ${myMetrics.crossEntropy()}.")
  println(s"Test areaUnderPR = ${metrics.areaUnderPR()}.")
  println(s"Test areaUnderROC = ${metrics.areaUnderROC()}.")

  sc.stop()
}
```

Spark — Lightning-fast cluster computing

stackoverflow

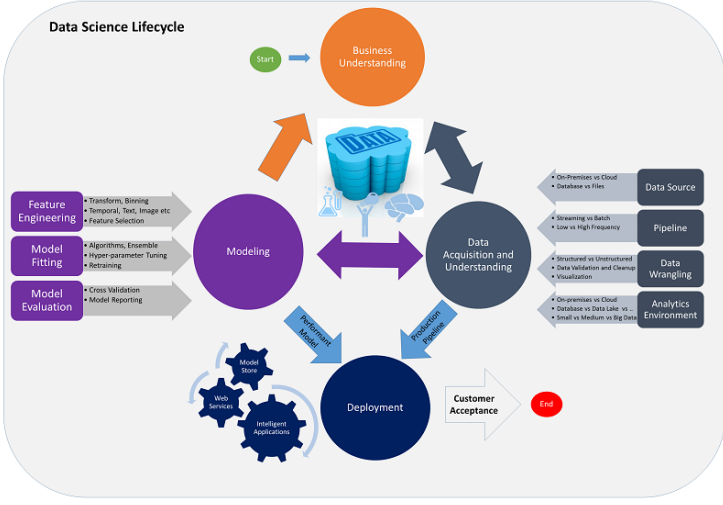
amazon web services

Example of off the shelves solution

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

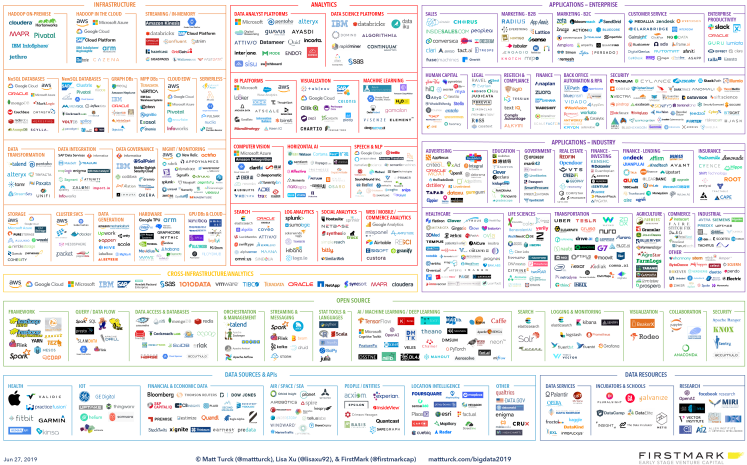
Sources: Unknown / 123rf.com / Unknown / Apache project / Stack Overflow / Amazon

Data Delivery Is (Quite) Complex!



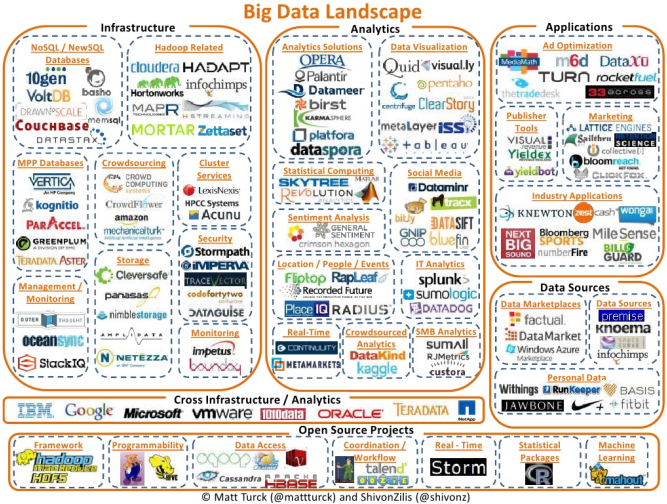
Source: Microsoft

Data Ecosystem Is (Quite) Complex!



Source: M. Turck

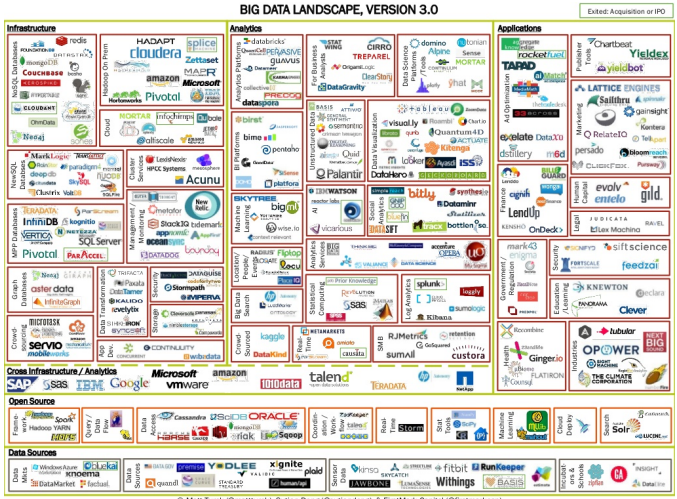
Data Ecosystem Is (Quite) Complex!



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

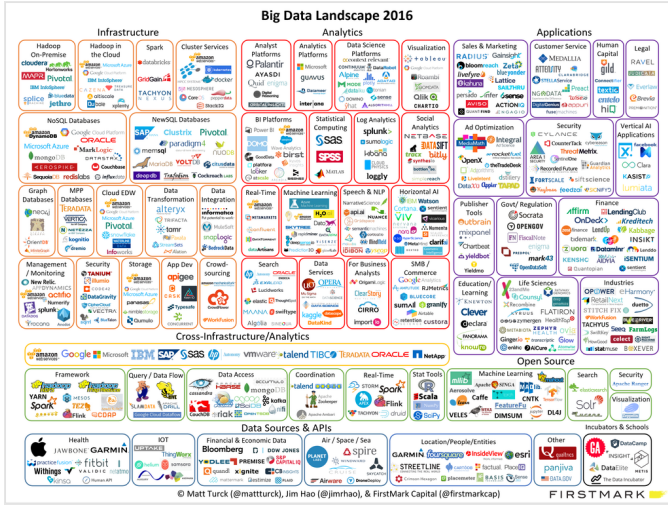
Source: M. Turck

Data Ecosystem Is (Quite) Complex!



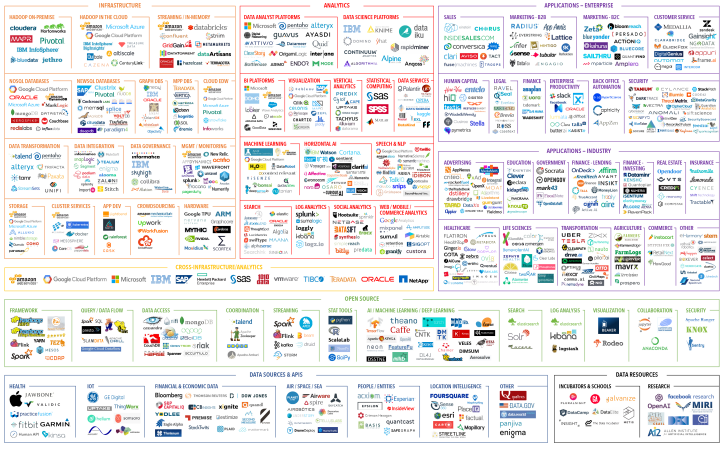
Source: M. Turck

Data Ecosystem Is (Quite) Complex!



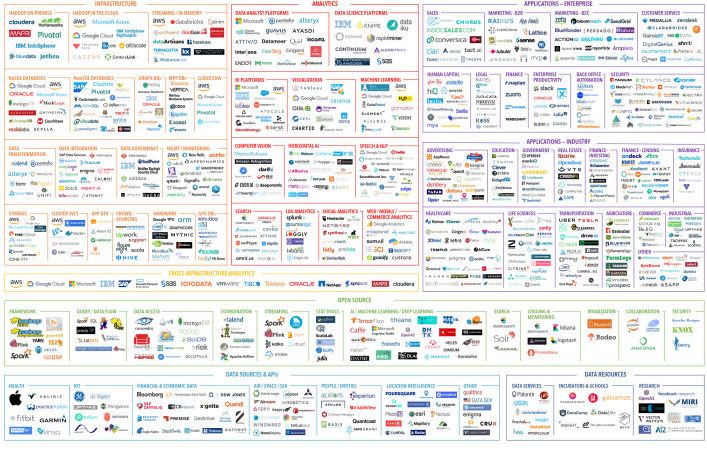
Source: M. Turck

Data Ecosystem Is (Quite) Complex!



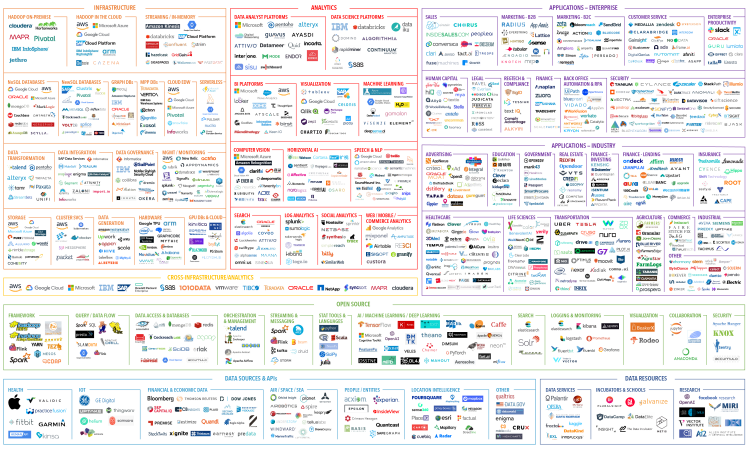
Source: M. Turck

Data Ecosystem Is (Quite) Complex!



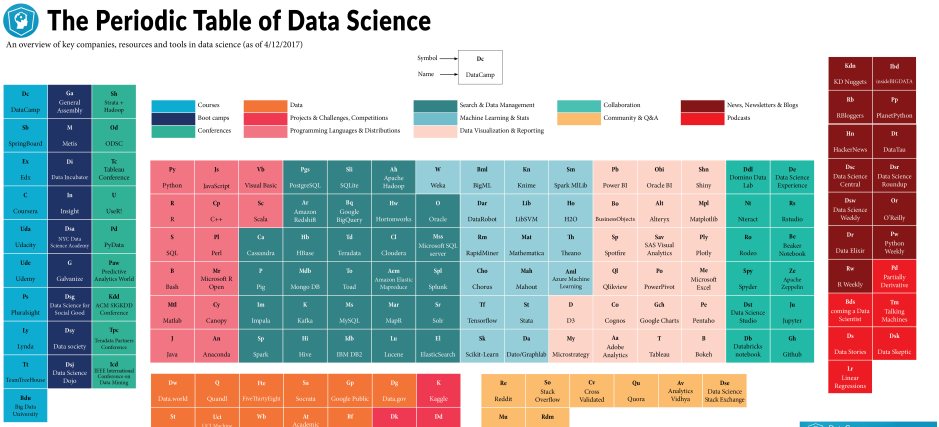
Source: M. Turck

Data Ecosystem Is (Quite) Complex!



Source: M. Turck

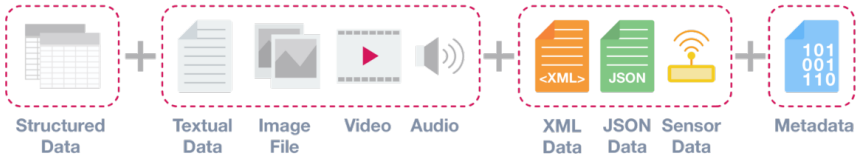
Data Ecosystem Is (Quite) Complex!



Source: DataCamp

Outline

Data Products - Data



Structured

- **Most convenient**
- Table

Unstructured

- **Most common**
- Text, Sound, Image, Video
- Graph

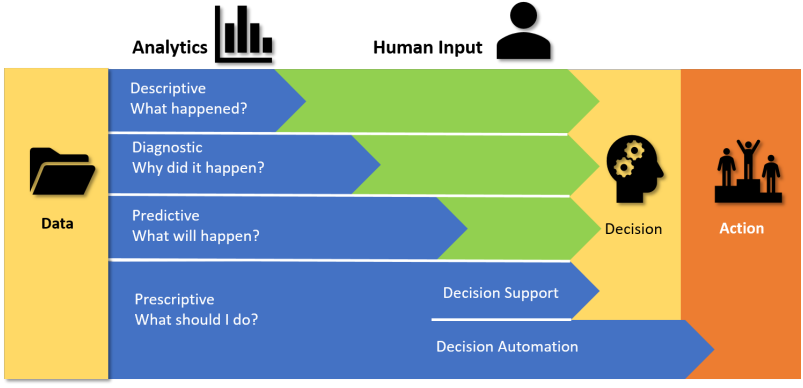
Semi-structured

- **Most volume?**
- Document (XML/JSON)
- Sensors/Logs

- Eventually everything has to be structured!
- Metadata are very important.

Source: Big Data Framework

Data Products - Goal



Goal

- Insight vs Action.
- Need to solve an issue.

Source: Gartner

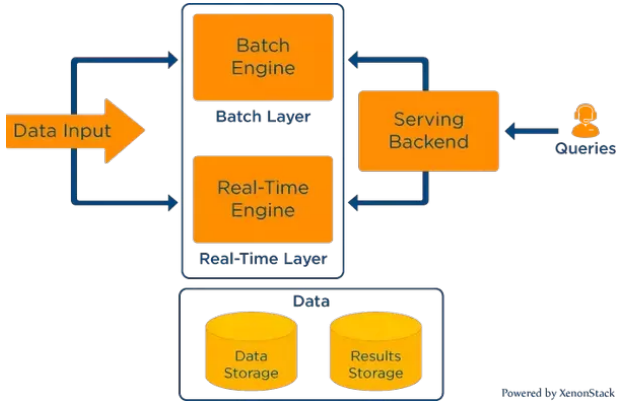
Data Products - Users

Users

- Human in the loop.
- Location (same office, company, clients...) and Expertise.

Source: Adobe

Data Products - Timeframe

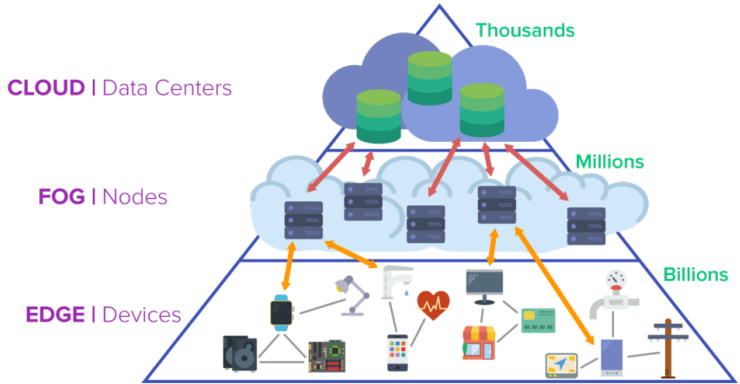


Timeframe

- Dual to size!
- Faster response and fresher data require much more work.

Source: Xenonstack

Data Products - Location

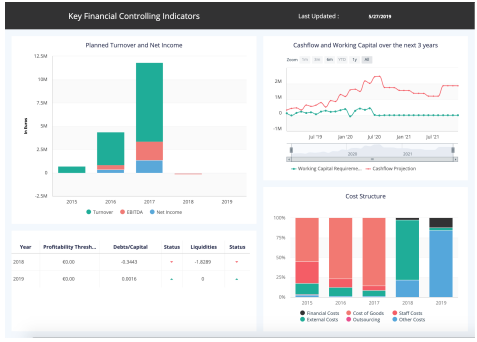


Location

- Where are the data and the computation units?

Source: PubNub

Monthly KPI Dashboard

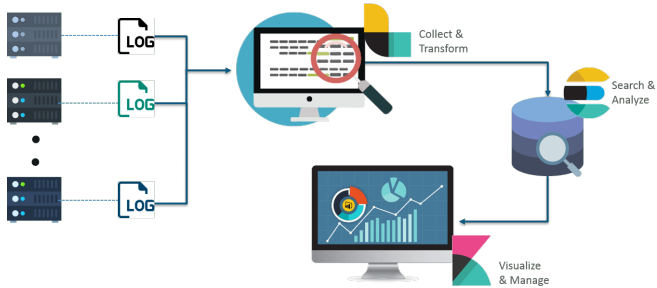


Monthly KPI Dashboard

- Using financial data to display important KPI for top managers every month in a slide.
- Automation to guaranty the quality of the data and the results.
- **Keywords:** ETL, Analytics, Dataviz

Source: decisyon

Realtime Log Dashboard



Realtime Log Dashboard

- Use log data to show the state of a system to IT in real time using on-premise tools.
- Automation to handle the huge volumetry.
- **Keywords:** Log, Stream

Source: edureka!

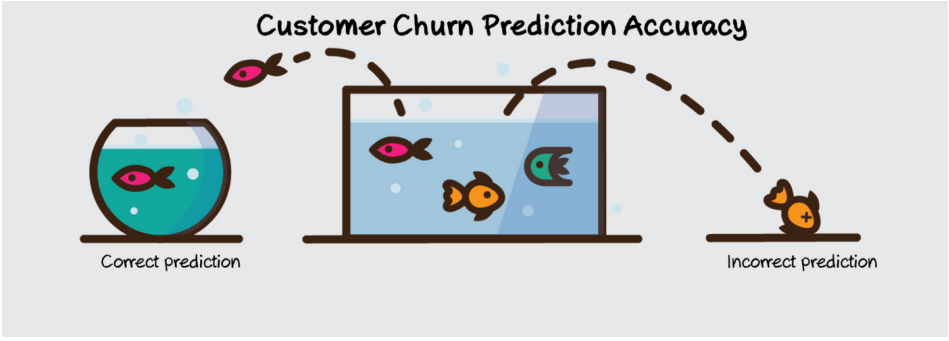
On-demand Legal Document Generation

On-demand Legal Document Generation

- Use raw data to legal document template for a lawyer on-demand using a local database.
- First draft to be edited by the lawyer.
- **Keywords:** RPA, NLP

Source: Amicus Attorney

Weekly Churn Prediction

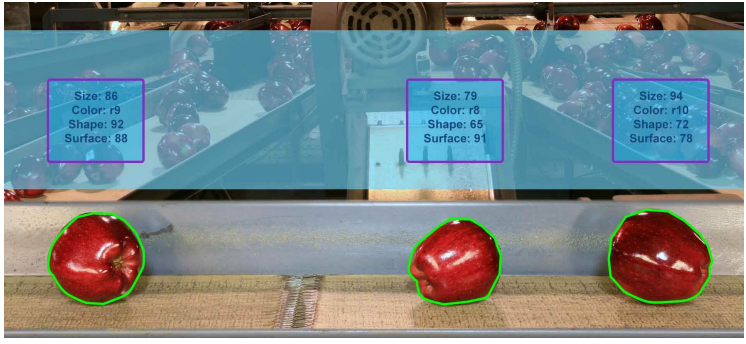


Weekly Churn Prediction

- Using consumer characteristics and history to give a churn score to the marketing every week using the cloud.
- Automation to scale to the volumetry but no strategy recommendation.
- **Keywords:** Machine Learning, Cloud

Source: Parthasarathy

Realtime Automatic Fruit Sorting

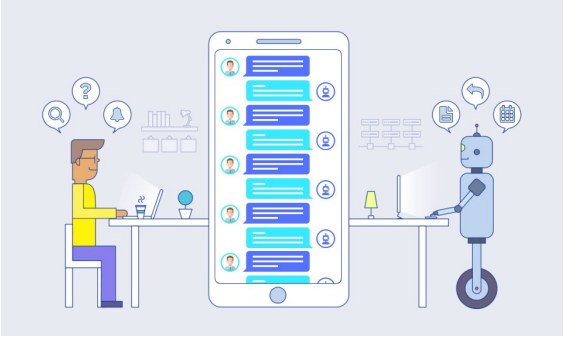


Realtime Automatic Fruit Sorting

- Using camera to sort fruits in a plant in realtime using local computers with GPU.
- Automation to reduce cost.
- **Keywords:** Computer Vision, Deep Learning, GPU

Source: BitRefine

Realtime Chatbot

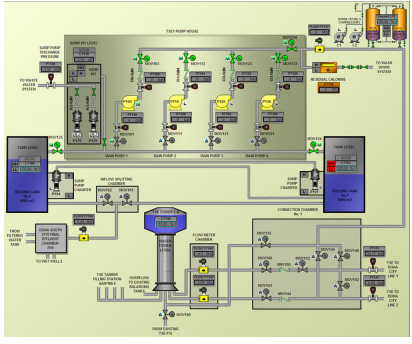


Realtime Chatbot

- Use previous interactions to predict answer to a consumer question in real time using the cloud.
- Reduce human interaction cost.
- **Keywords:** NLU, Cloud

Source: M'Bufung

Realtime Anomaly Detection



Realtime Anomaly Detection

- Use production data to detect anomalies in a plant in real time on a Scada system.
- Reduce failure cost.
- **Keywords:** Unsupervised, Scada

Source: Wikipedia

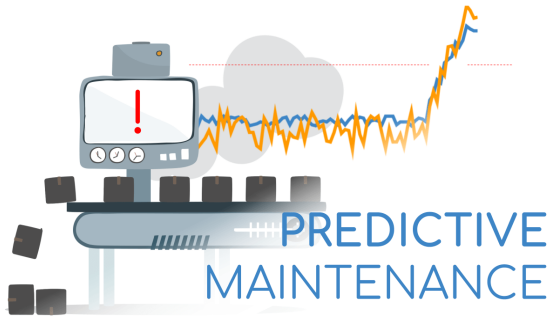
On-demand Fraud Detection

On-demand Fraud Detection

- Use claim and client data to detect fraud for an insurer on-demand using on-premise resources
- First automated pass on the claims.
- **Keywords:** Unsupervised, NLP

Source: MoneyKama

Prescriptive Maintenance



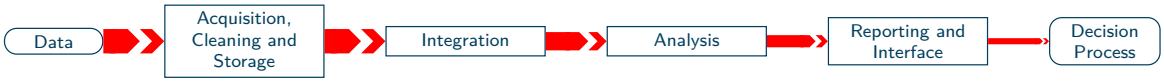
Prescriptive Maintenance (Not yet available. . .)

- Use data to devise and apply the best maintenance plan in a plant using IOT.
- Reduce maintenance cost.
- **Keywords:** Reinforcement Learning, IOT

Source: CSS Electronics

Outline

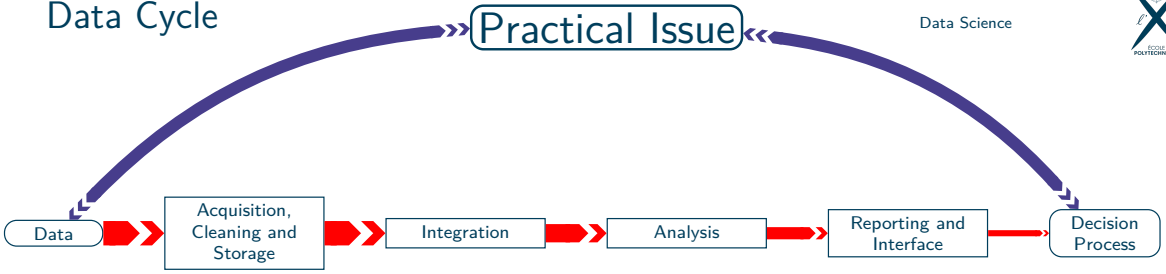
Data Cycle



Data/Information Flow

- **Plumber** vision

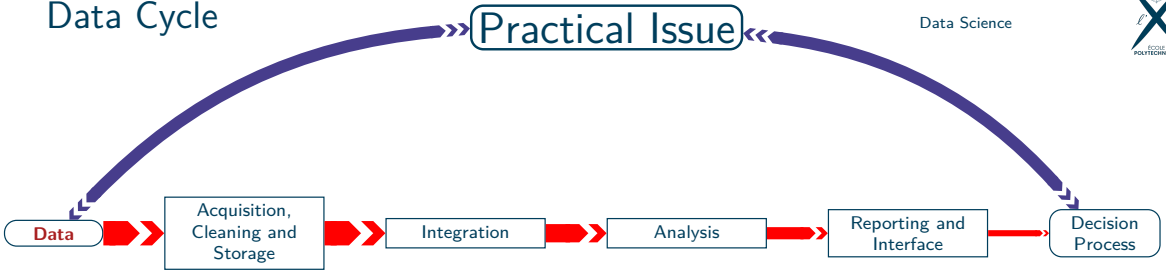
Data Cycle



Data/Information Flow

- **Plumber** vision
- Driven by a goal

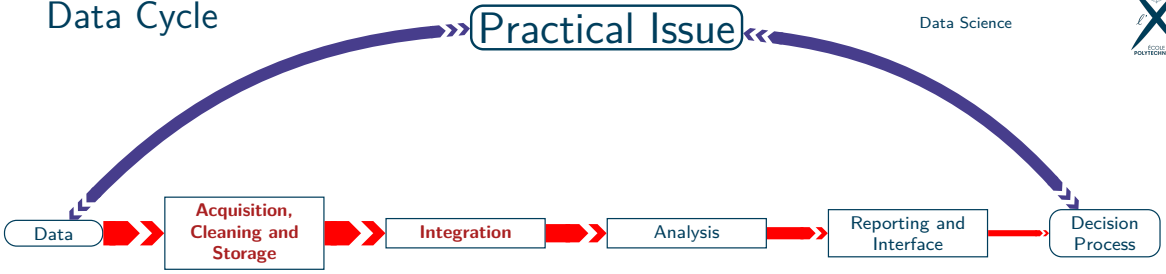
Data Cycle



Data

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

Data Cycle



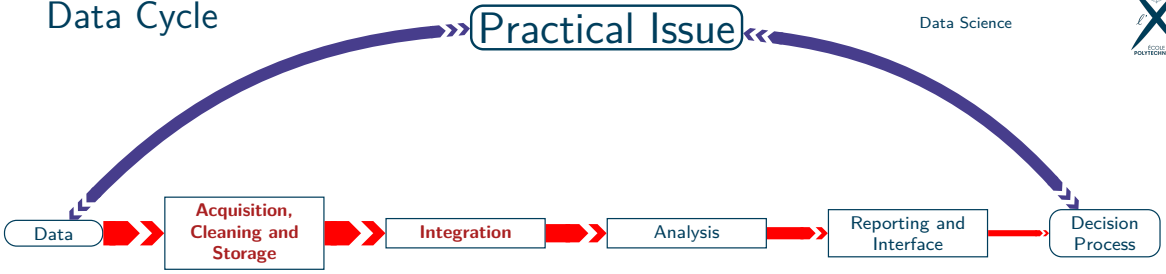
Acquisition, Cleaning, Storage (ETL)

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

Integration

- Analysis tool specific data preparation.

Data Cycle



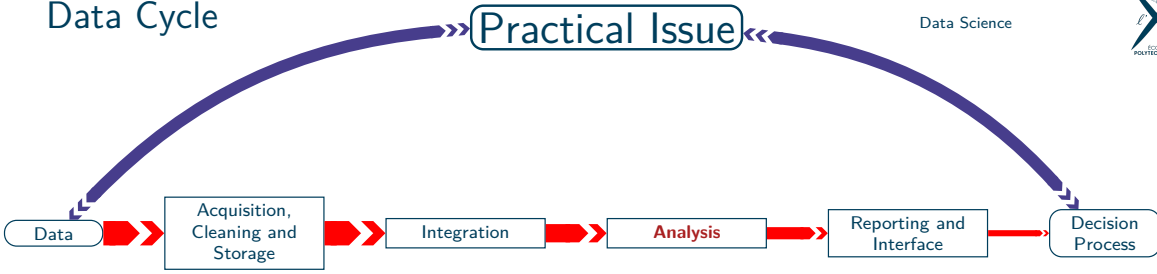
Acquisition, Cleaning, Storage (ETL) (Data Management)

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

Integration (Plumbing issue)

- Analysis tool specific data preparation.
- Time consuming!

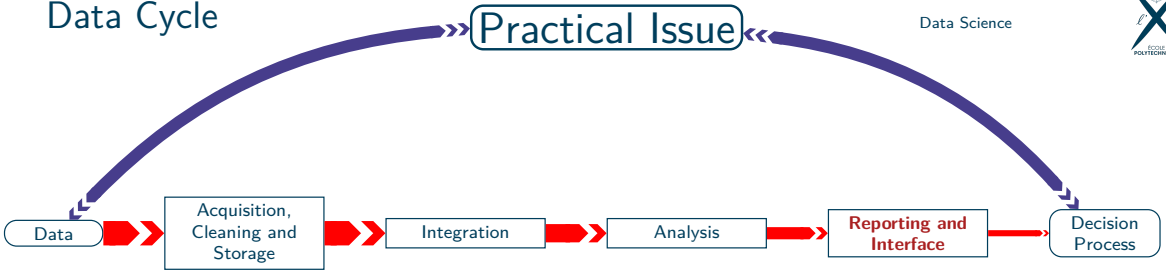
Data Cycle



Analysis

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

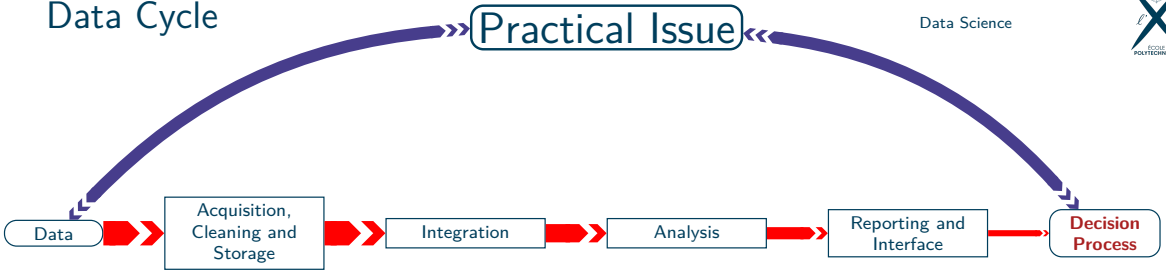
Data Cycle



Reporting and Interface

- (Human) Reporting: visualization, report, app...
- Important also during Exploratory Data Analysis
- (Computer) Interface: dedicated tool, Application Programming Interface (API)
- Some key aspects:
  - Sharing (and Collaboration)
  - Deployment / Continuous Integration
  - Maintenance / Continuous Amelioration

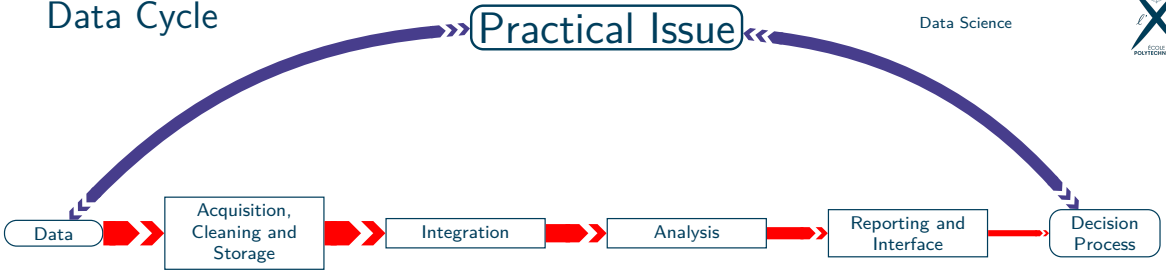
Data Cycle



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!

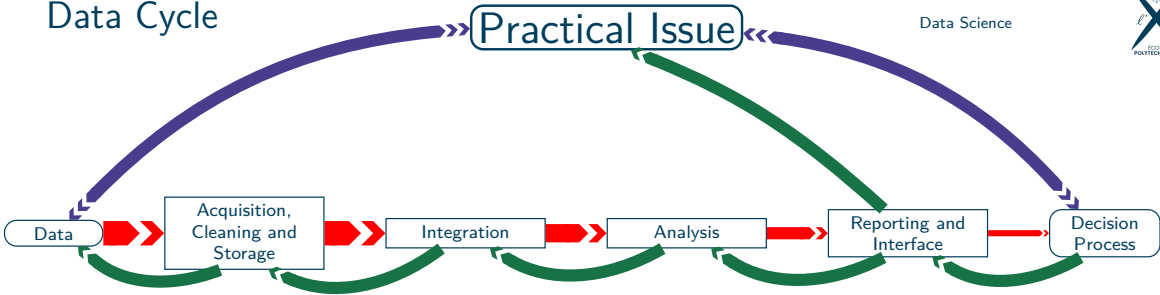
Data Cycle



Data/Information Flow

- **Plumber** vision
- Driven by a goal

Data Cycle



Data/Information Flow

- **Plumber** vision
- Driven by a goal

- **Iterative and interactive process**

Data Cycle (Data Science)

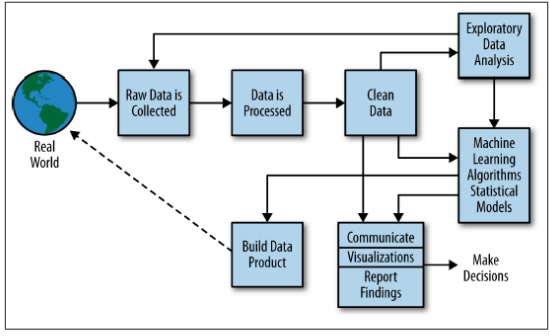


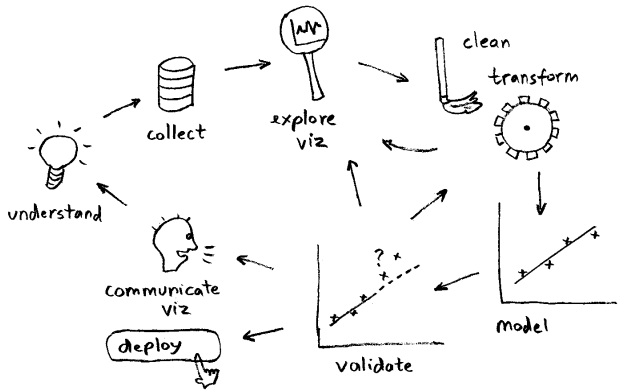
Figure 2-2. The data science process

Data Science

- Final product point of view

Source: S. Schutt & C. O'Neil

Data Cyle (Data Mining)

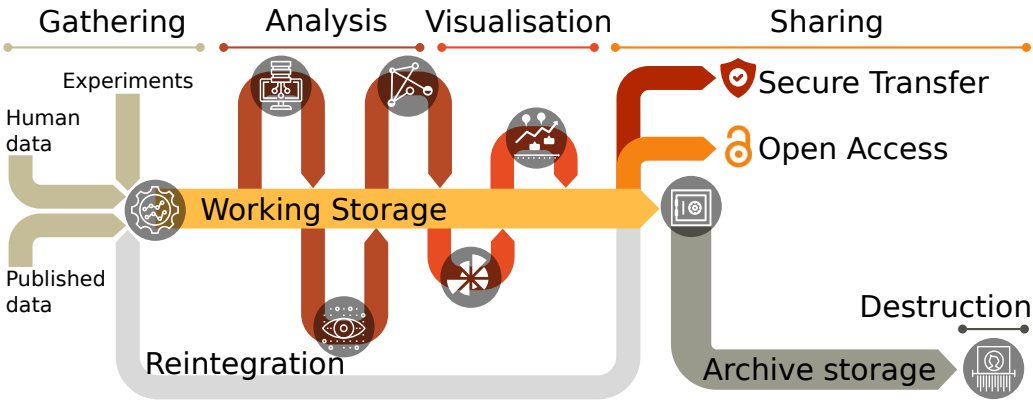


Data Mining

- CRISP-DM: Cross-industry standard process for data mining (1999)
- Data exploitation point of view

Source: Sz. Pafka

Data Cyle (Data Management)



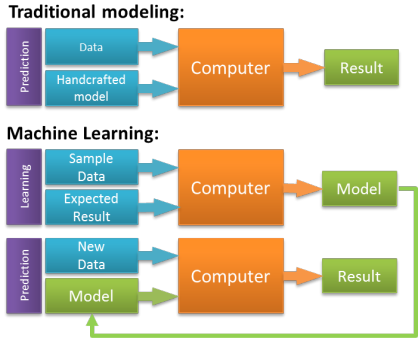
Data Management

- Data path point of view

Source: Wikipedia

Outline

Machine Learning

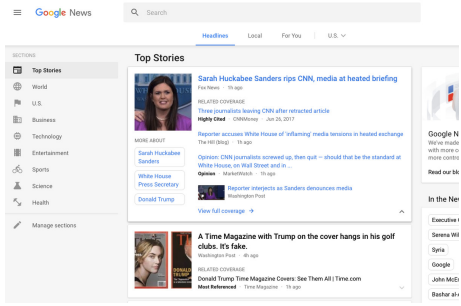


A definition by Tom Mitchell (http://www.cs.cmu.edu/~tom/)

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.

Source: azavea

Article Clustering



An article clustering algorithm:

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

Source: theverge.com

Object Detection

A detection algorithm:

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

Source: MyCarDoesWhat.org

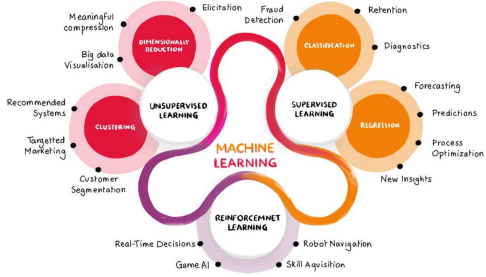
A Robot that Learns

A robot endowed with a set of sensors playing football:

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

Source: Rutgers Prep School

Three Kinds of Learning



Unsupervised Learning

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

Supervised Learning

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

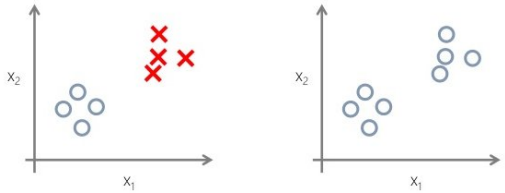
Reinforcement Learning

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

- **Timing:** Offline/Batch (learning from past data) vs Online (continuous learning)

Source: BCG

Supervised and Unsupervised



Supervised Learning (Imitation)

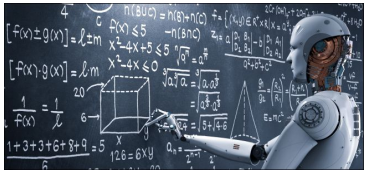
- **Goal:** Learn a function f predicting a variable Y from an individual $\underline{X}$.
- **Data:** Learning set with labeled examples $(\underline{X}_i, Y_i)$
- **Assumption:** Future data behaves as past data!
- **Predicting is not explaining!**

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...

Source: KDnuggets

Machine Can and Cannot



Machine Can

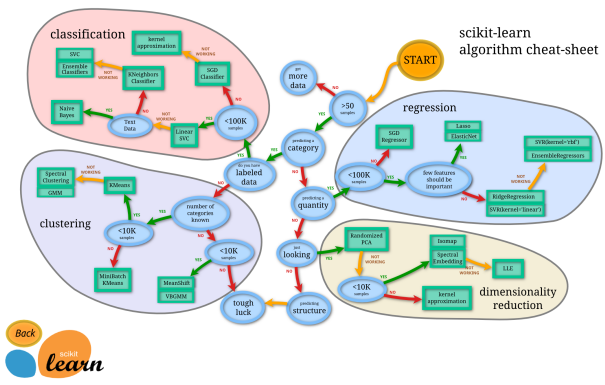
- Forecast (Prediction using the past)
- Detect some changes
- Memorize/Reproduce
- Take a decision very quickly
- Learn from huge dataset
- Optimize a single task
- Replace/Help some humans

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

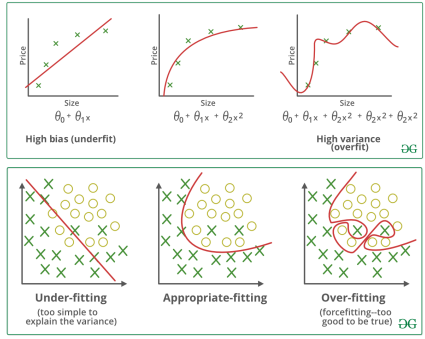


ML Methods

- Huge catalog of methods,
- Need to define the performance,
- Numerous tricks: feature design, hyperparamter selection...

Source: scikit-learn.org

Under and Over Fitting

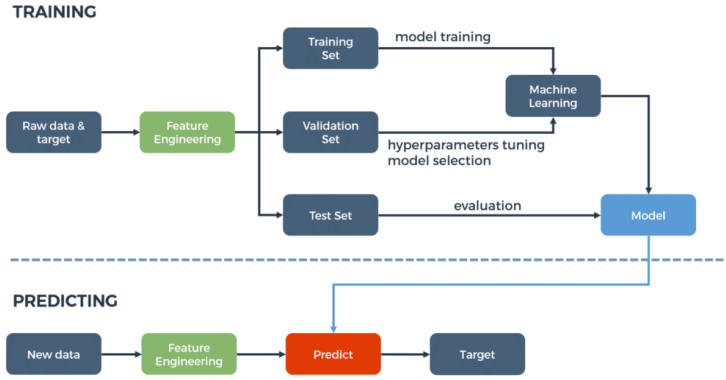


Finding the Right Complexity

- 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.

Source: geeksforgeeks.com

ML Pipeline



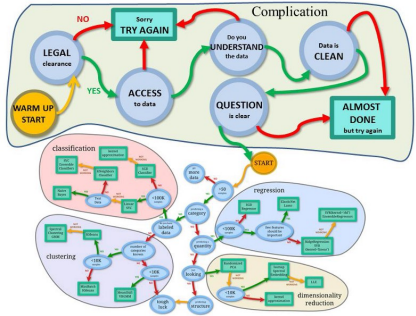
Learning pipeline

- Test and compare models.

- Deployment pipeline is different!

Source: CDiscount

Data driven approach $\neq$ ML



Main data driven approach difficulties

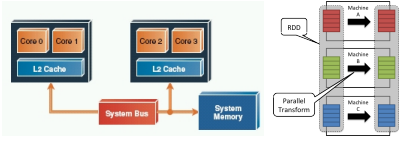
- Figuring out the problem,
- Formalizing it,
- Storing and accessing the data,
- Deploying the solution,
- Not (always) the ML part!

Source: Ch. Bourguignat

Outline

Outline

Big Data?



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)

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)

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)

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)

What could be limiting?

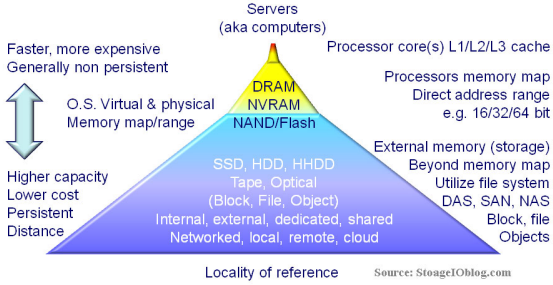
Possible Issue

- 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)

Memories



Size hierarchy

CPU register	64 b $\times$ 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

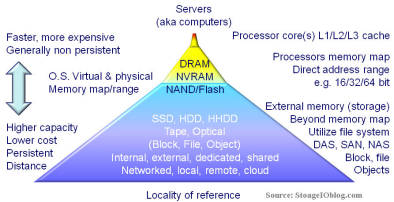
Source: storageioblog.com

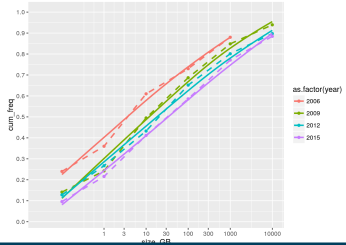
Memories

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	CPU bound
Level 3 cache access	12.9 ns	43 s	
Main memory access	120 ns	6 min	
Solid-state disk I/O	50-150 μs	2-6 days	
Rotational disk I/O	1-10 ms	1-12 months	
Internet: SF to NYC	40 ms	4 years	IO bound
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





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?)

Sources: storageioblog.com / Sz. Pafka

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)

What could be limiting?

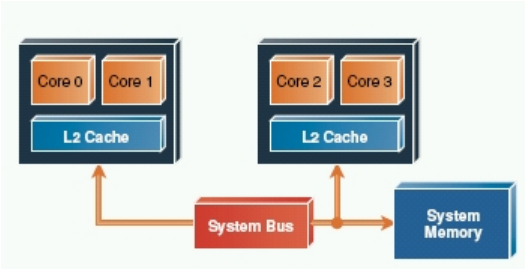
Possible Issue

- 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)

Parallelization



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**!

Source: Unknown

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)

What could be limiting?

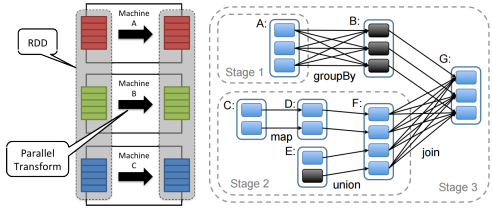
Possible Issue

- 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)

Distribution

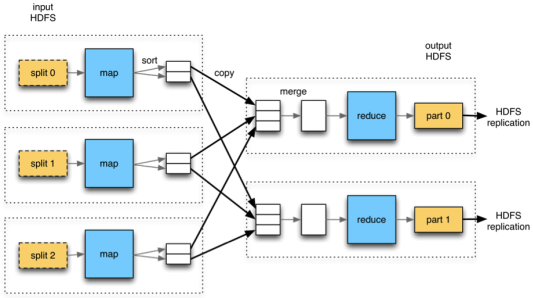


True Big Data Setting

- 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...)

Source: R. Ho / N. McBurnett

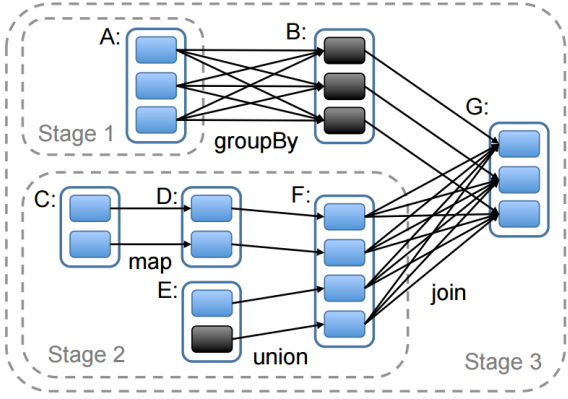
Hadoop And Map/Reduce



Hadoop

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

Source: Cloudera



Spark

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

Source: N. McBurnett

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)

What could be limiting?

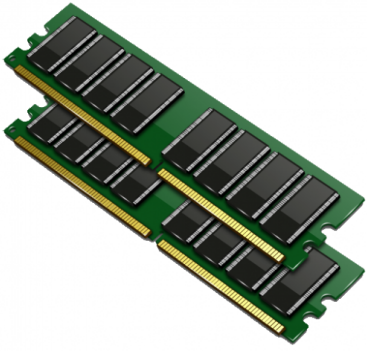
Possible Issue

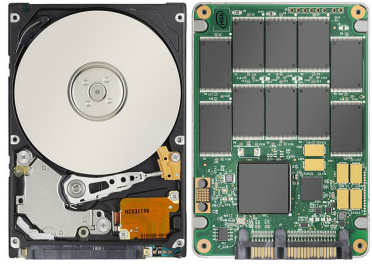
- Coding issue?
- **IO issue?**
- **Processing issue?**
- **Data storage issue?**

Enhancement?

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

Memories





RAM and SSD

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

Source: Unknown / Unknown

Processing Units

PU: CPU, GPU, FPGA, ASICS

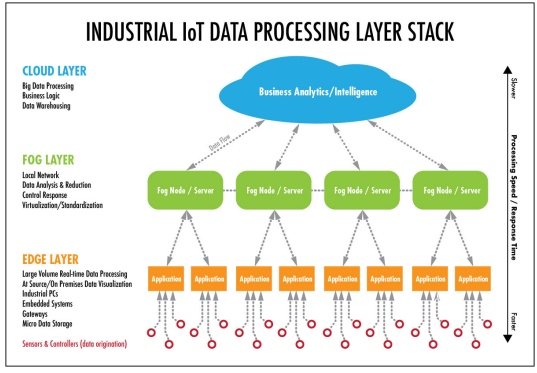
- More than one processor architecture.
- Flexibility vs performance.
- Parallelism: CPU $<$ GPU $<$ FPGA $<$ ASICS.

Cluster

- More computers...
- IO is important!

Source: Nvidia / advancedclustering.com

IOT Edge/Fog/Cloud



IOT: Computing and Bandwidth constrains

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

Source: WinSystems

Outline

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)

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)
- **Better data storage? (database)**

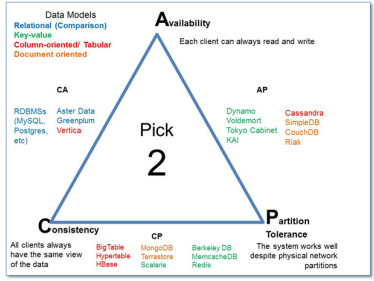
Databases

(SQL?) Databases

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

Source: CodingGeek

SQL/NoSQL



SQL

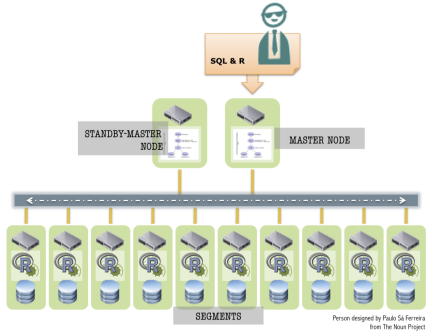
- Most classical design,
- Limitation due to the CAP theorem.
- Hard to distribute without asking less...

NoSQL (Not only SQL!)

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

Source: flux7.com

UDF: DB as a Distributed System

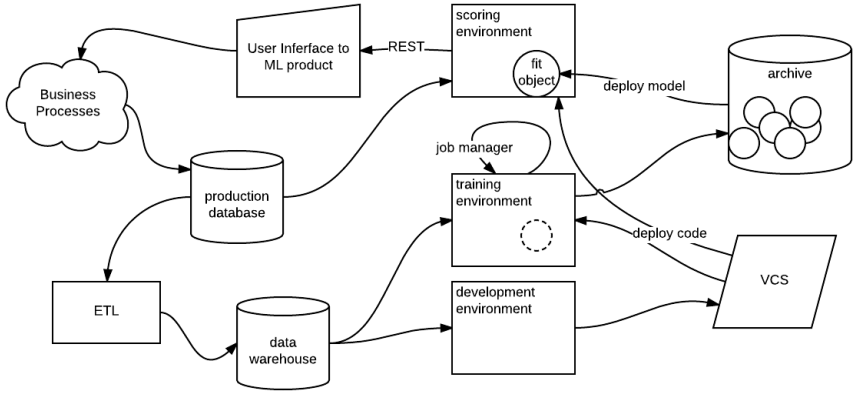


Database and User Defined Function

- Allow to defined 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 priviledges...

Source: Unknown

DS Architecture



Data Science Architecture

- Usage dependent architecture!

- **Finding a good architecture is complex**

Sources: Sz. Pafka

Outline

Data Science Tools

DS Toolkit

- Several tools

Data Science Tools

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

DS Toolkit

- Several tools

Data Science Tools

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

DS Toolkit

- Several tools

Data Science Tools

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

DS Toolkit

- Several tools

Data Science Tools

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

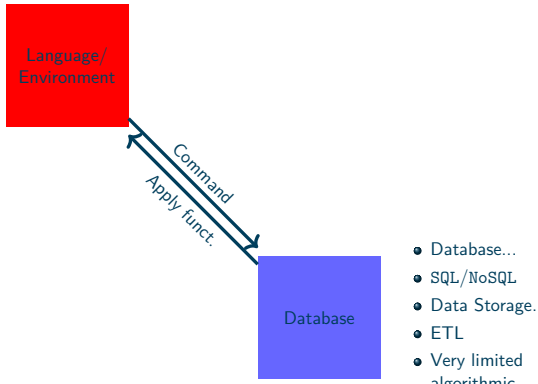
GUI / API

DS Toolkit

- Several tools

Data Science Tools

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



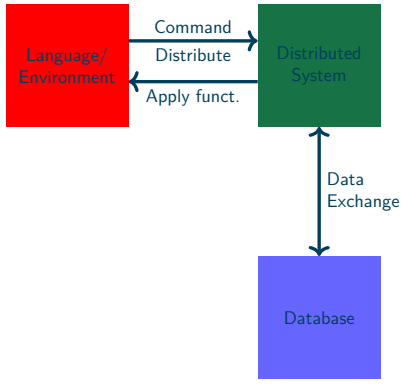
- Database...
- `SQL`/`NoSQL`
- Data Storage.
- ETL
- Very limited algorithmic

DS Toolkit

- Several tools working together

Data Science Tools

- 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

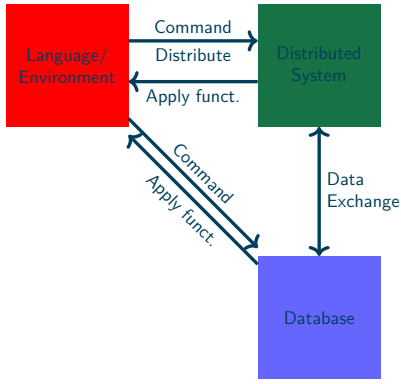
- Database...
- `SQL/NoSQL`
- Data Storage.
- ETL
- Very limited algorithmic

DS Toolkit

- Several tools working together

Data Science Tools

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



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

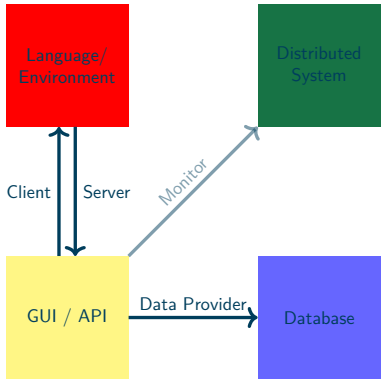
- Database...
- `SQL/NoSQL`
- Data Storage.
- ETL
- Very limited algorithmic

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, reproductibility / Specific use

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

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

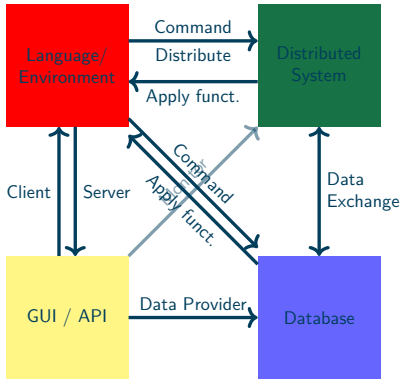
DS Toolkit

- 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, reproductibility / 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

Languages

Polyglot

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

Sources: python.org / r-project.org / julialang.org

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`...

Source: python.org

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`...

Source: r-project.org

Julia

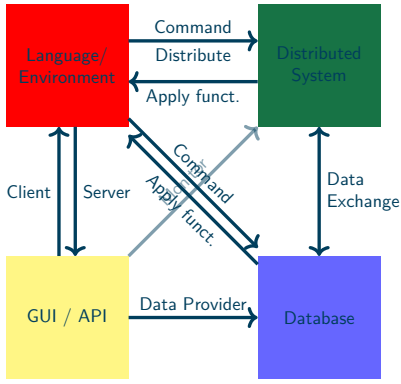
- Project initiated at the MIT in the 10's.
- Interpreted language written in C and C++.
- Clear and simple syntax (with type declaration).
- Can be interfaced with most classical language.
- Allow to write concise program with a high level of abstraction.
- Available on Unix, Windows, OS X...
- JIT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

Source: julialang.org

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, reproductibility / 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?

Comparison

Folklore Comparison

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

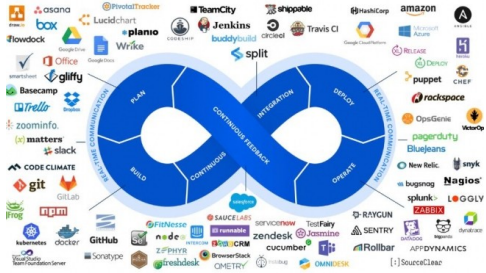
Data Analysis Comparison

- **Data type:**
  - `R` and `Python` can handle quantitative and qualitative varibles.
  - `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)**

Outline

DevOps and Data?

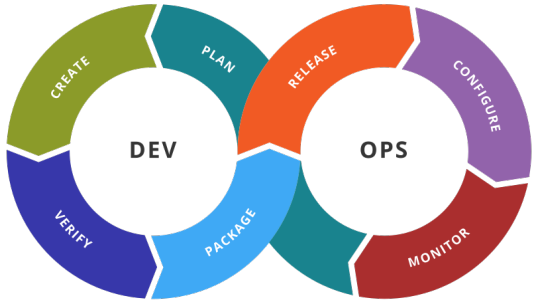


DevOps

- 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!

Source: Atlassian?

DevOps Mindset

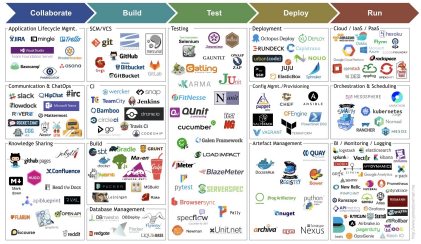


Much more than technical tools!

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

Source: Wikipedia

DevOps Tools



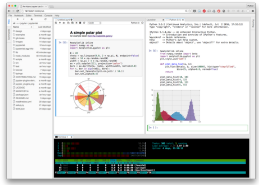
Lots of tools for each step!

- **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 that the tool itself.

Source: James Bowman

Code and DevOps



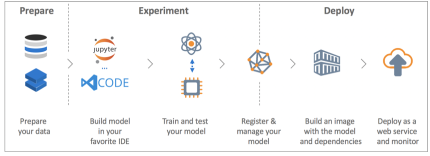
- Code are meant to be used/shared/reused.

Good practice

- Versioning (Code),
- Documentation,
- Testing,
- Packaging,
- Continuous Integration/Continuous Deployment,
- Human Training

Source: Chris Villala

Models and MLOps



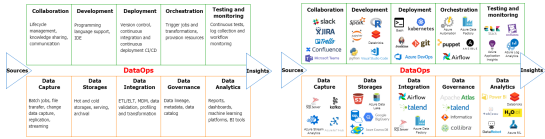
- Models are meant to be used/shared/reused.

Good practice

- Versioning (Dataset/Models/Code),
- Artifact mgmt,
- Documentation,
- Training/Testing/Monitoring,
- Human Training,
- Continuous Integration/Continuous Deployment

Source: Azure

Data and DataOps



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

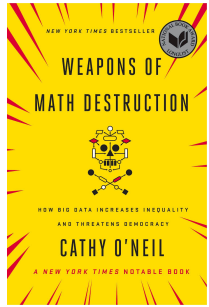
Good practice

- Versioning (Dataset/Models/Code),
- Documentation/Governance,
- Training/Testing/Monitoring,
- Packaging (Dataset/Tools),
- Human Training,
- Continuous Integration/Continuous Deployment.

Source: Valdas Maksimavicius

Outline

Sociology, Regulation and Ethics



Some challenges

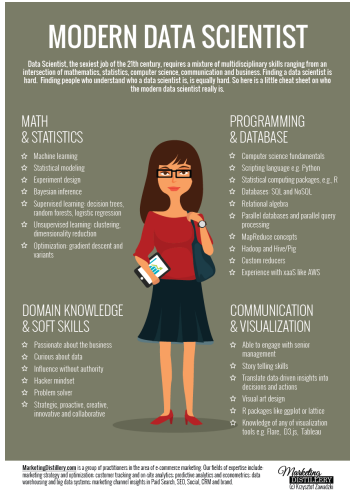
- How to embed humans in the process?
- How to share the benefits?
- How to be fair?
- How to comply with regulation? / How to answer public concerns?
- How to avoid ethical issues?

Sources: D.J. Patil / C. O'Neill

Outline

Outline

Skills

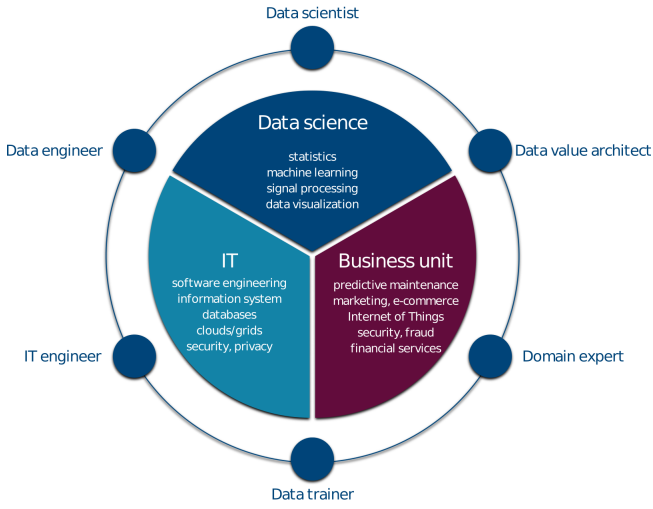


Data Scientist

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

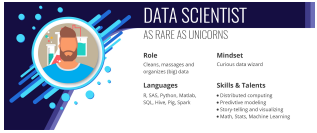
Source: MarketingDistillery.com

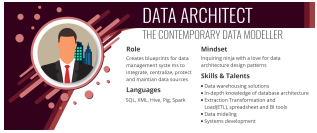
More Than One Type Of Data Jobs!



Source: B. Kegl - CDS Paris Saclay

Much More Variety?





Several Profiles

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

Source: imarticus

Outline

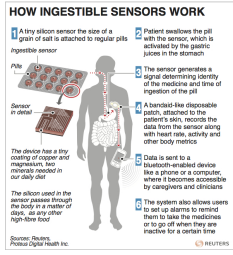
Data Challenges

- Applied math **AND** Computer science...
- **AND** Domain Specific Knowledge

Some joint Math/CS/Domain challenges

- Data acquisition
- Big Data: Huge dataset and computation issues
- Unstructured data
- Small/Fat Data: High dimension
- Learning, Training set and Supervision
- Visualization and UX
- Data Architecture, Software(s) and Deployment
- Domain Specific Knowledge!
- Sociology, Regulation and Ethics

Data Acquisition

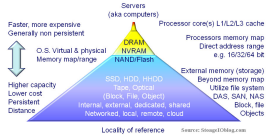


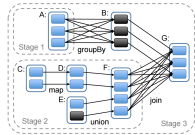
Some challenges

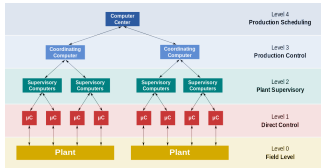
- How to measure new things?
- How to choose what to measure?
- How to deal with distributed sensors?
- How to look for new sources of information?
- How to plan some experiments?

Source: Reuters, Proteus Digital Health

Big Data: Huge Dataset and Computation Issues





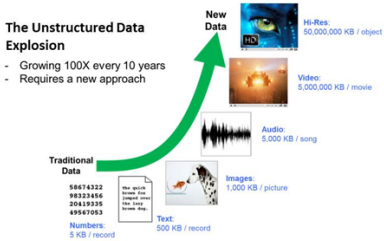


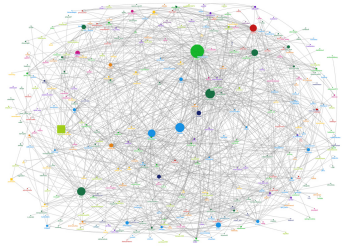
Some challenges

- 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?

Sources: storageioblog.com / N. McBurnett / Wikipedia

Unstructured Data



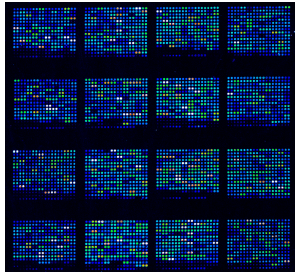


Some challenges

- 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?

Sources: L. Schubert & K. Jeffrey / Dataversity

Small/Fat Data: High Dimension

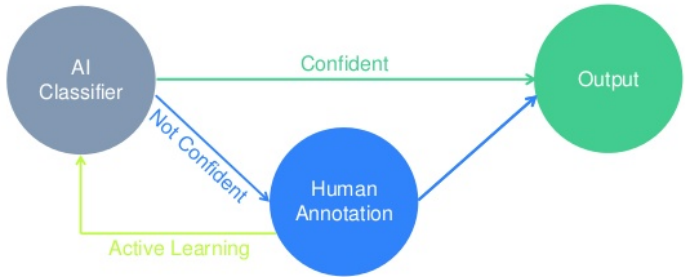


Some challenges

- 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?

Source: A. Nantel

Learning, Training Set and Supervision

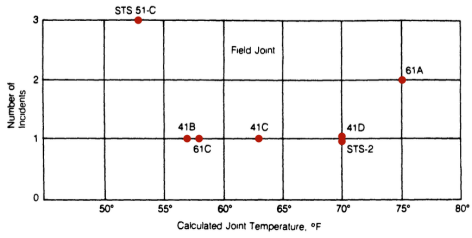


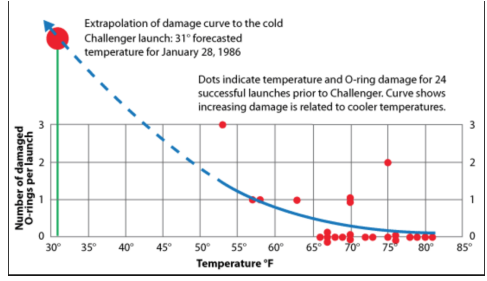
Some challenges

- How to learn with the less possible data/interactions?
- How to reuse a model trained on a first task?
- How to learn simultaneously several related tasks?

Source: crowdflowder.com

Visualization and UX



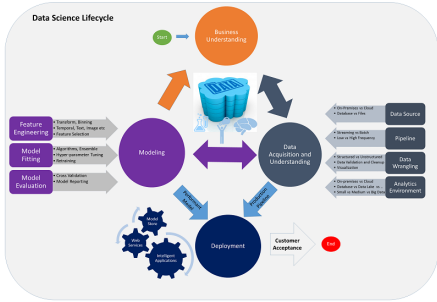


Some challenges

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

Sources: NASA / E. Tufte

Data Architecture, Software(s) and Deployment



Some challenges

- How to construct a consistent ecosystem?
- How to construct interoperable systems?
- How to guaranty the quality?
- How to build a data team?
- How to attract users?

Source: Microsoft

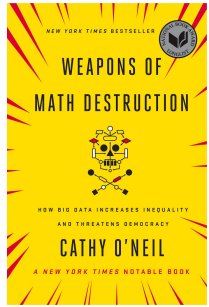
Domain Specific Knowledge

Some challenges

- How to find the real problem at hand?
- How to build product?
- How to measure the performance?
- How to measure the ROI?
- How to embed human expertise?

Source: Shutterstock

Sociology, Regulation and Ethics



Some challenges

- How to embed humans in the process?
- How to share the benefits?
- How to be fair?
- How to comply with regulation? / How to answer public concerns?
- How to avoid ethical issues?

Sources: D.J. Patil / C. O'Neill

Outline

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, Keras and TensorFlow (2nd ed.)
O'Reilly, 2019

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

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