R, Data Science and Teaching

Erwan Le Pennec

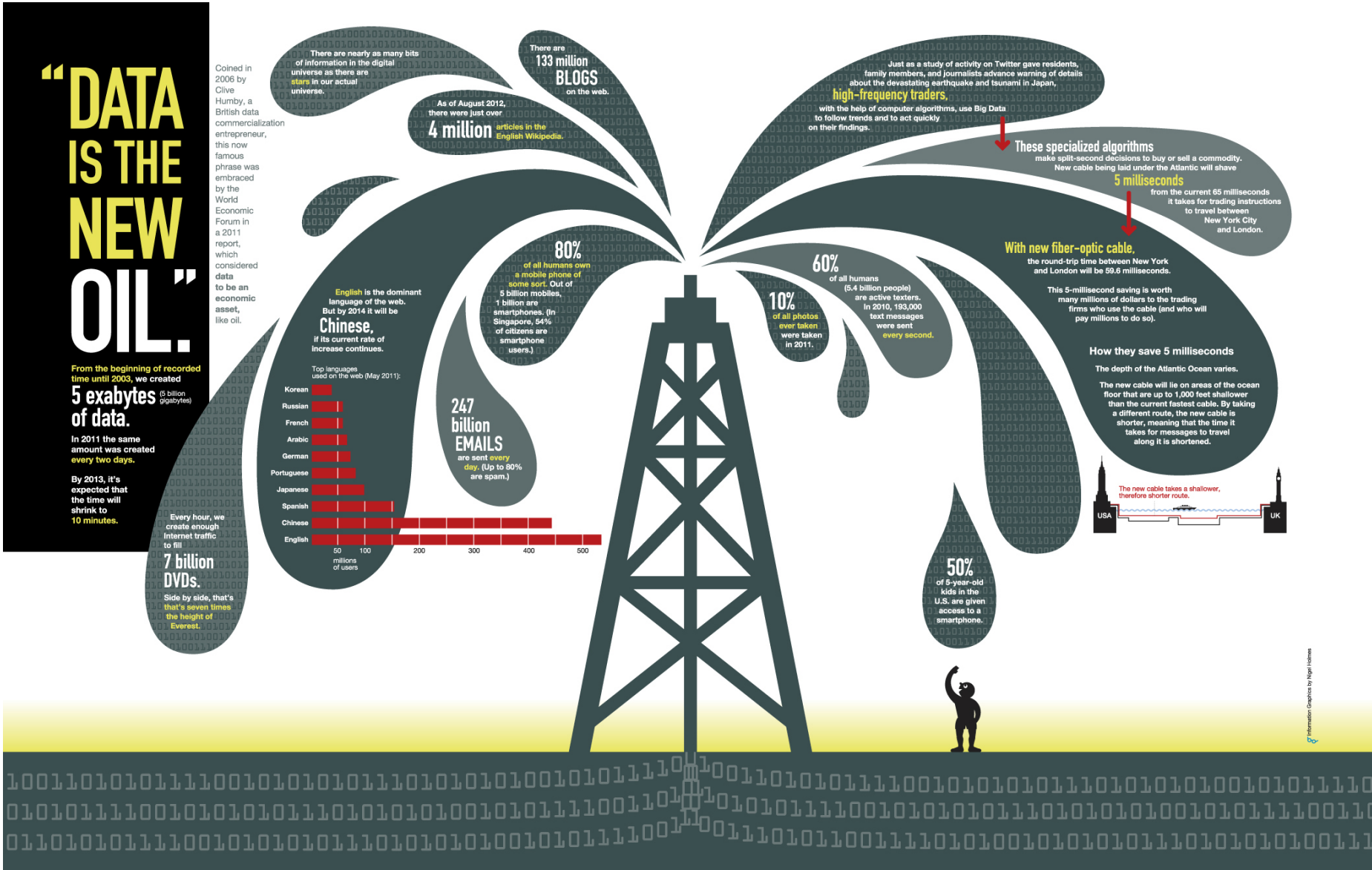
CMAP Ecole polytechnique

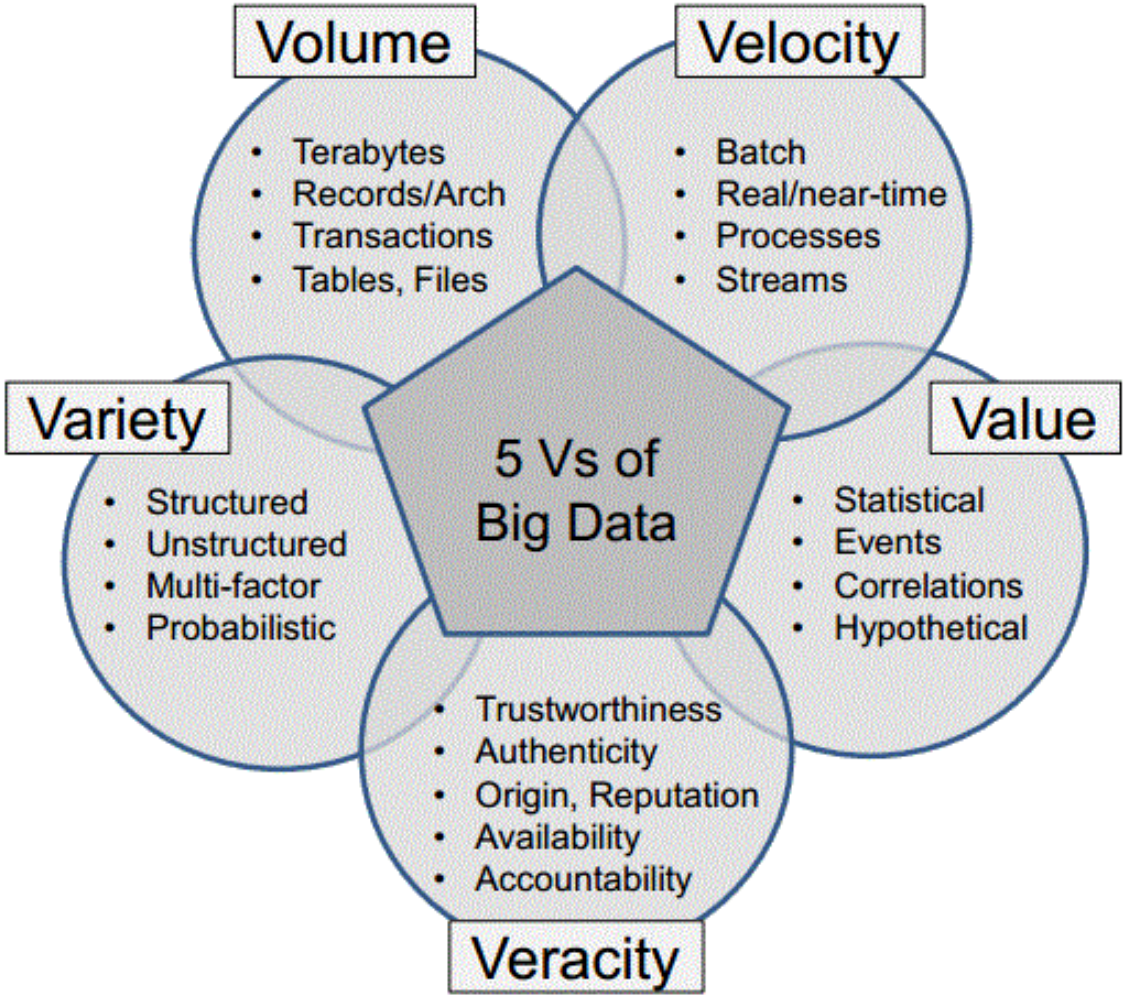
Rencontres R - 25/06/2015

Outline

Data Science and Big Data

Data is the new Oil!



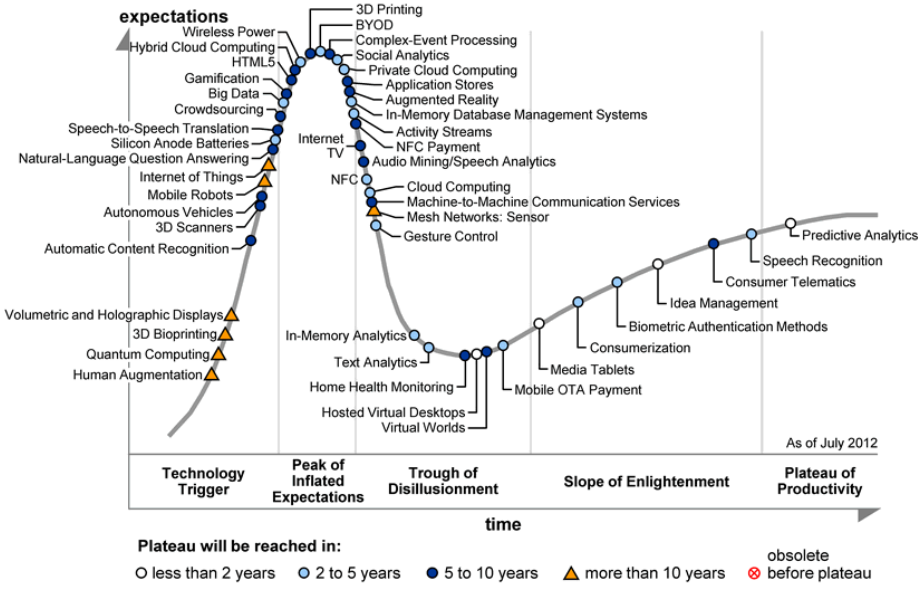

Volume
Terabytes
Records/Arch
Transactions
Tables, Files
Velocity
Batch
Real/near-time
Processes
Streams
Variety
Structured
Unstructured
Multi-factor
Probabilistic
5 Vs of Big Data
Value
Statistical
Events
Correlations
Hypothetical
Trustworthiness
Authenticity
Origin, Reputation
Availability
Accountability
Veracity

Data Science and Big Data

Lots of Words!

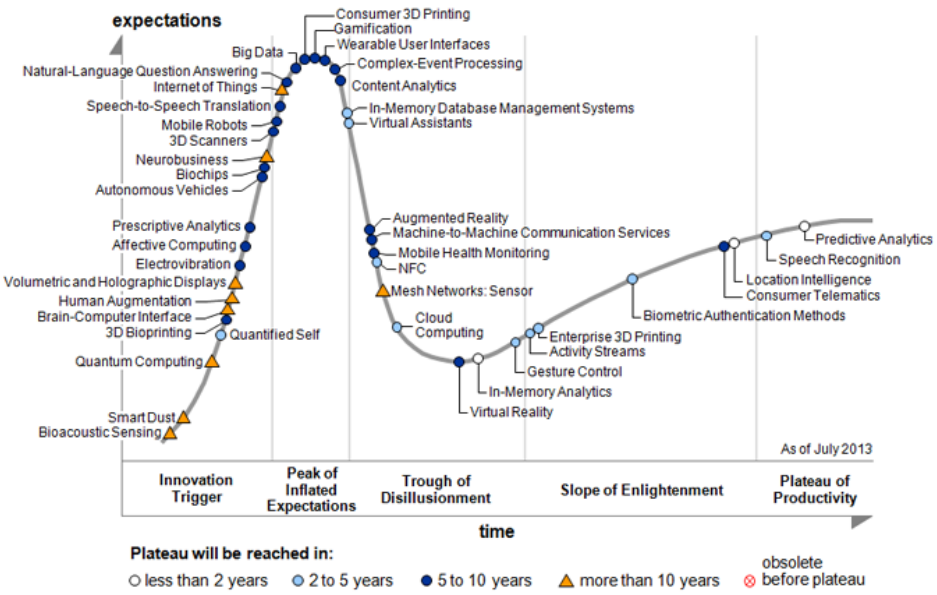
Data Science and Big Data

Don't believe the hype?



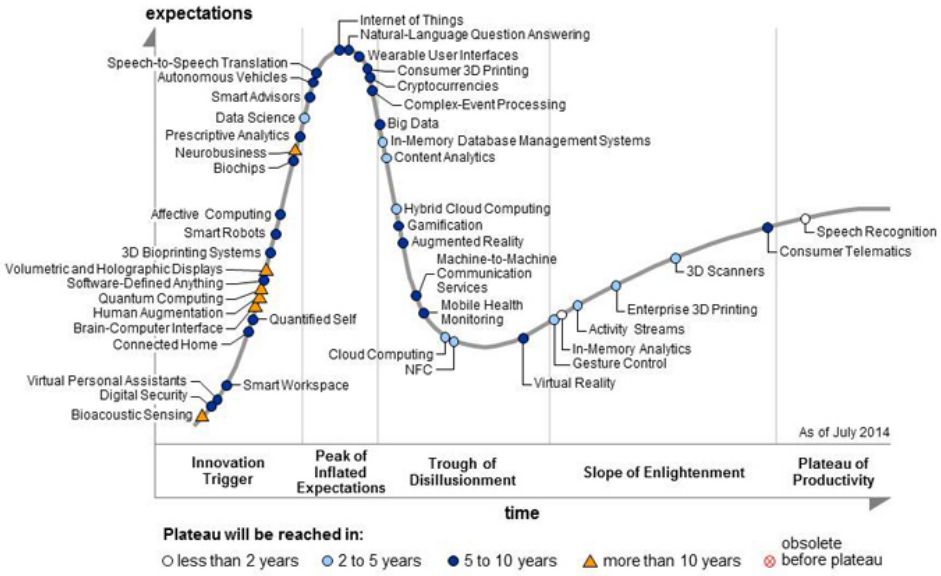
Data Science and Big Data

Don't believe the hype?



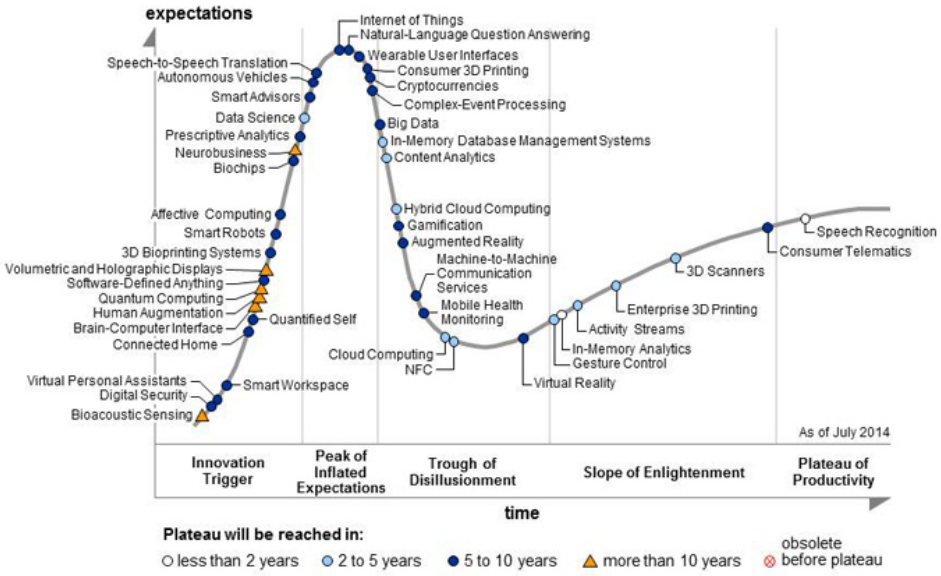
Data Science and Big Data

Don't believe the hype?



Data Science and Big Data

Don't believe the hype?



- Data Science and Big Data: Much more than a hype!

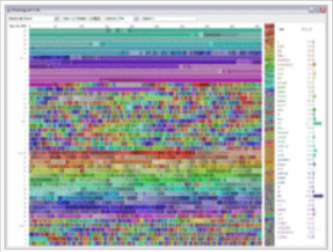
Big data

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]



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 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]

- **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.
- **Data science** is the study of the generalizable extraction of knowledge from data, yet the key word is science.
- **Statistics** is the study of the collection, analysis, interpretation, presentation and organization of data.

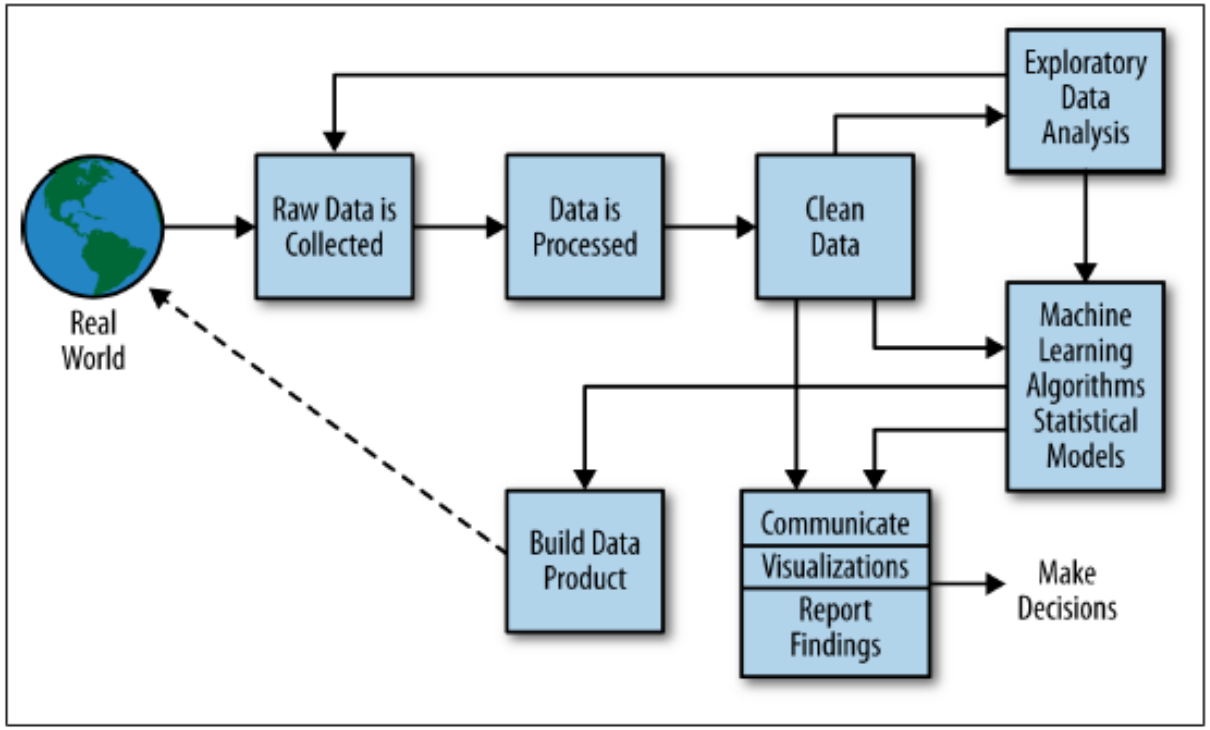


Figure 2-2. The data science process

Doing Data Science: Straight talk from the frontline

- Rachel Schutt, Cathy O'Neil - O'Reilly
- Art of data driven decision / evaluation.

Data everywhere

- Huge volume,
- Huge variety...

Affordable computation units

- Cloud computing
- Graphical Processor Units (GPU)...

- Growing academic and industrial interest!

Example of *off the shelves* solution

```
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()
}
```

Example of *off the shelves* solution

```
export AWS_ACCESS_KEY_ID=<your-access-keyid>
export AWS_SECRET_ACCESS_KEY=<your-access-key-secret>
cellule/spark/ec2/sparl-ec2 -i cellule.pem -k cellule -s <number of machines> launch <cluster-name>
ssh -i cellule.pem root@<your-cluster-master-dns>
spark-ec2/copy-dir ephemeral-hdfs/conf
ephemeral-hdfs/bin/hadoop distcp s3n://celluledecalcul/dataset/raw/train.csv /data/train.csv
scp -i cellule.pem cellule/challenge/target/scala-2.10/target/scala-2.10/challenges_2.10-0.0.jar

cellule/spark/bin/spark-submit \
        --class fr.cc.challenge.Preprocess \
        challenges_2.10-0.0.jar \
        /data/train.csv \
        /data/train2.csv

cellule/spark/bin/spark-submit \
       --class fr.cc.sparktest.LogisticRegression \
       challenges_2.10-0.0.jar \
       /data/train2.csv
```

$\Rightarrow$ Logistic regression for arbitrary large dataset!

Google

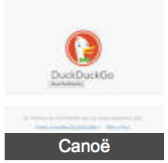
moteur de recherche

Web | Actualités | Images | Vidéos | Maps | Plus | Outils de recherche

Environ 10 100 000 résultats (0,24 secondes)

Moteur de recherche - Mozbot France - La recherche facile ...
www.mozbot.fr/
Moteur de recherche Mozbot en partenariat avec Brioude-Internet, Abondance et Google : résultats, synonymes, expressions connexes, statistiques mots clés, ...

Actualités correspondant à **moteur de recherche**



Le **moteur de recherche** DuckDuckGo bloqué en Chine
Le Monde - il y a 3 heures
Selon le site spécialisé TechInAsia, le **moteur de recherche** serait bloqué depuis le 4 septembre dans le pays. DuckDuckGo, qui se présente ...

L'Allemagne souhaite que Google dévoile les algorithmes ...
Clubic.com - il y a 5 jours

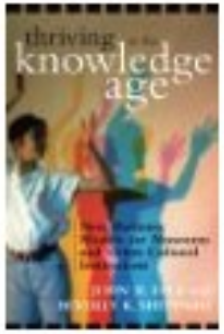
Plus d'actualités pour "**moteur de recherche**"

Moteur de recherche — Wikipédia
fr.wikipedia.org/wiki/**Moteur_de_recherche**
Un **moteur de recherche** est une application web permettant de retrouver des ressources (pages web, articles de forums Usenet, images, vidéo, fichiers, etc.) ...

Moteur de Recherche SEEK.fr ™
www.seek.fr/
Moteur de recherche alternatif français respectant la vie privée via un métamoteur utilisant les principaux **moteurs de recherche** ainsi qu'un annuaire ...
Metamoteur Web SEEK.fr - A Propos de Seek - Horoscope - Seek annuaire

More Ideas Based on Your Browsing History

You looked at



Thriving in the Knowledge Age: New... Paperback by John H. Falk
$29.95

You might also consider

Museum Administration: An Introduction Paperback by Hugh H. Genoways
~~$31.95~~ **$28.75**

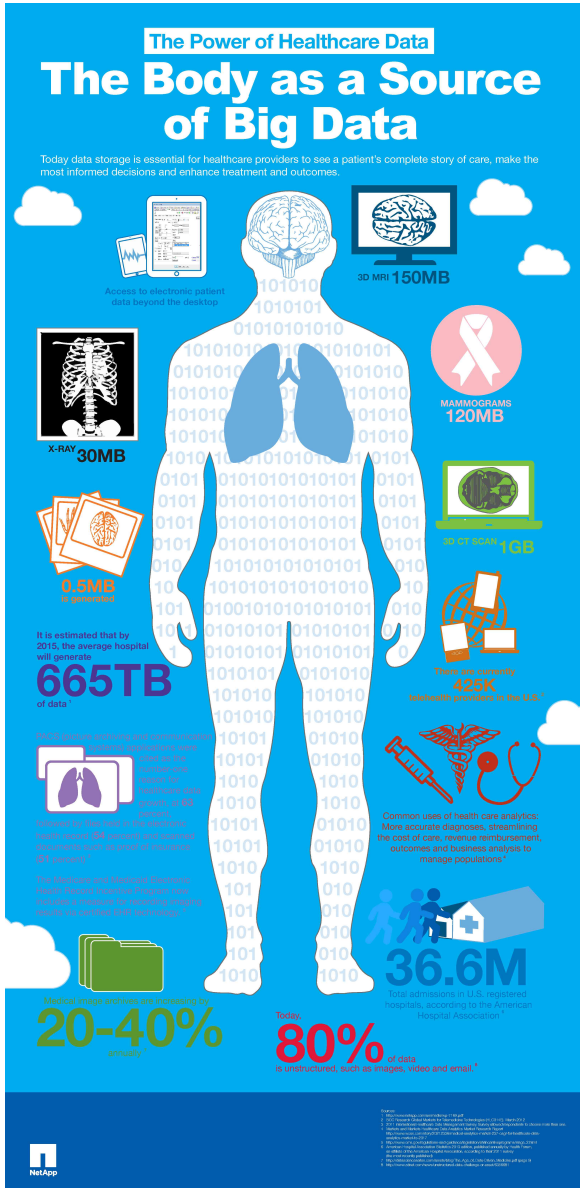
Exhibit Labels: An Interpretive Approach Paperback by Beverly Serrell
~~$34.95~~ **$27.85**

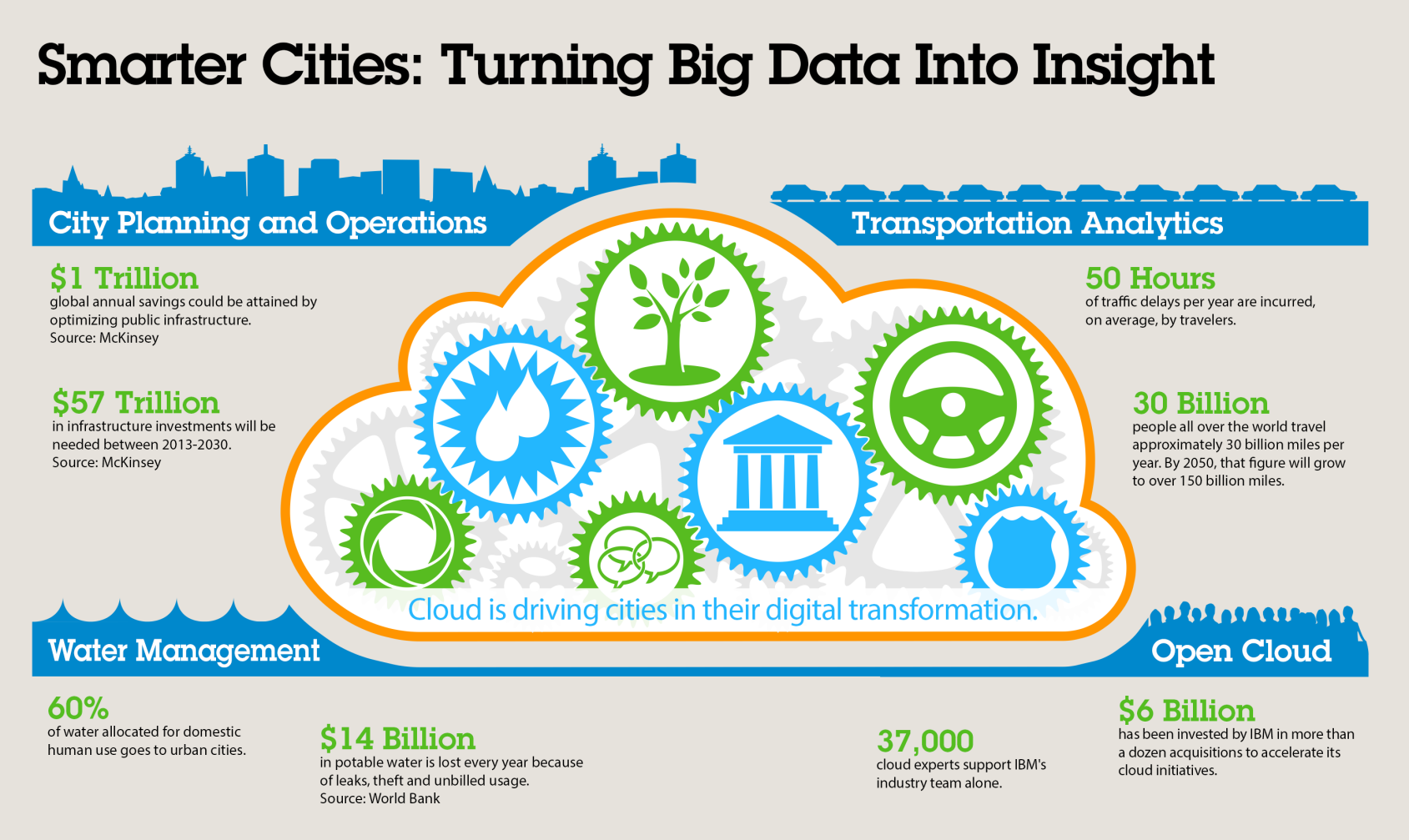
› Find similar items

Recommendations don't have to be about showing you more of the same...

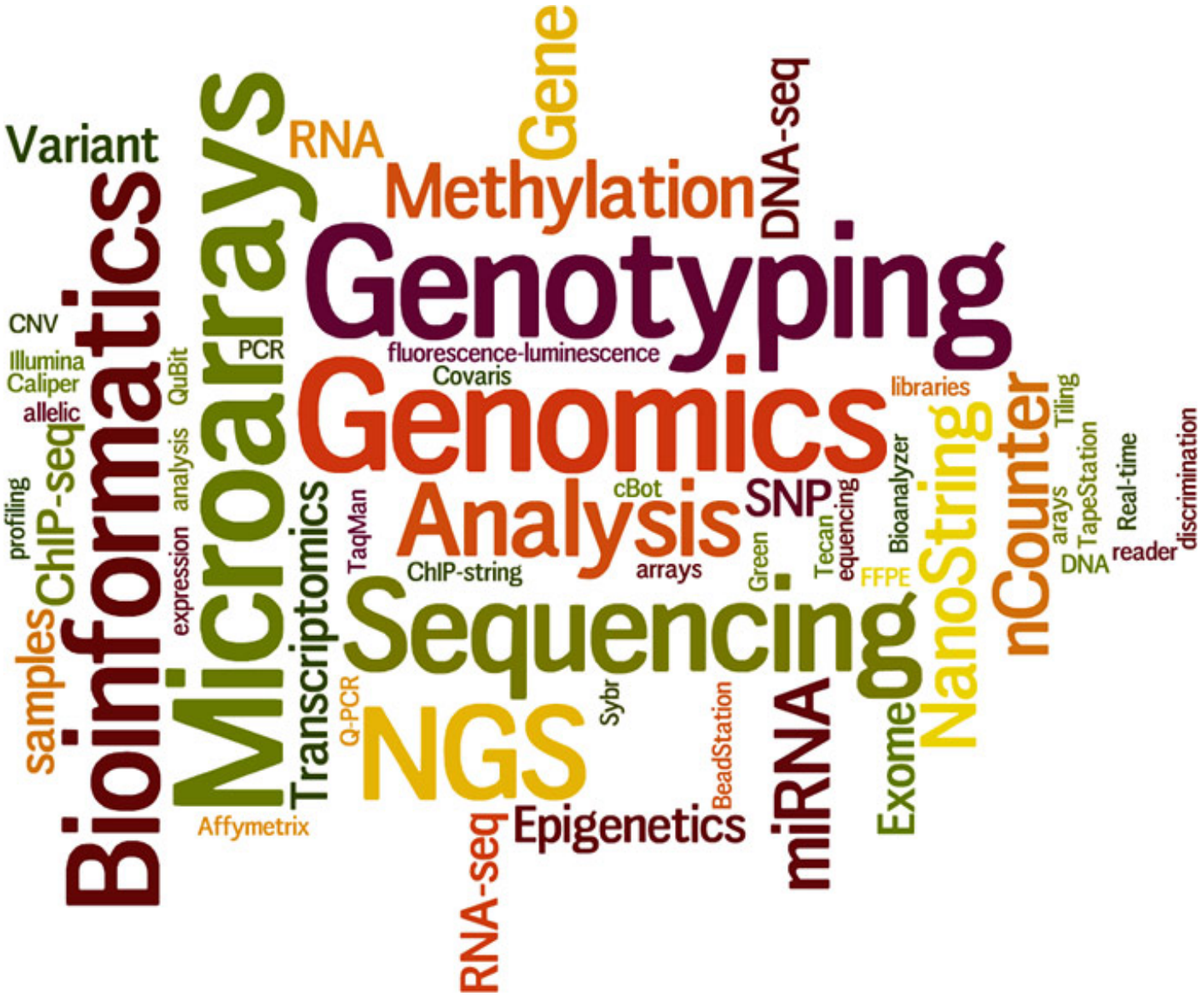
Data Science and Big Data

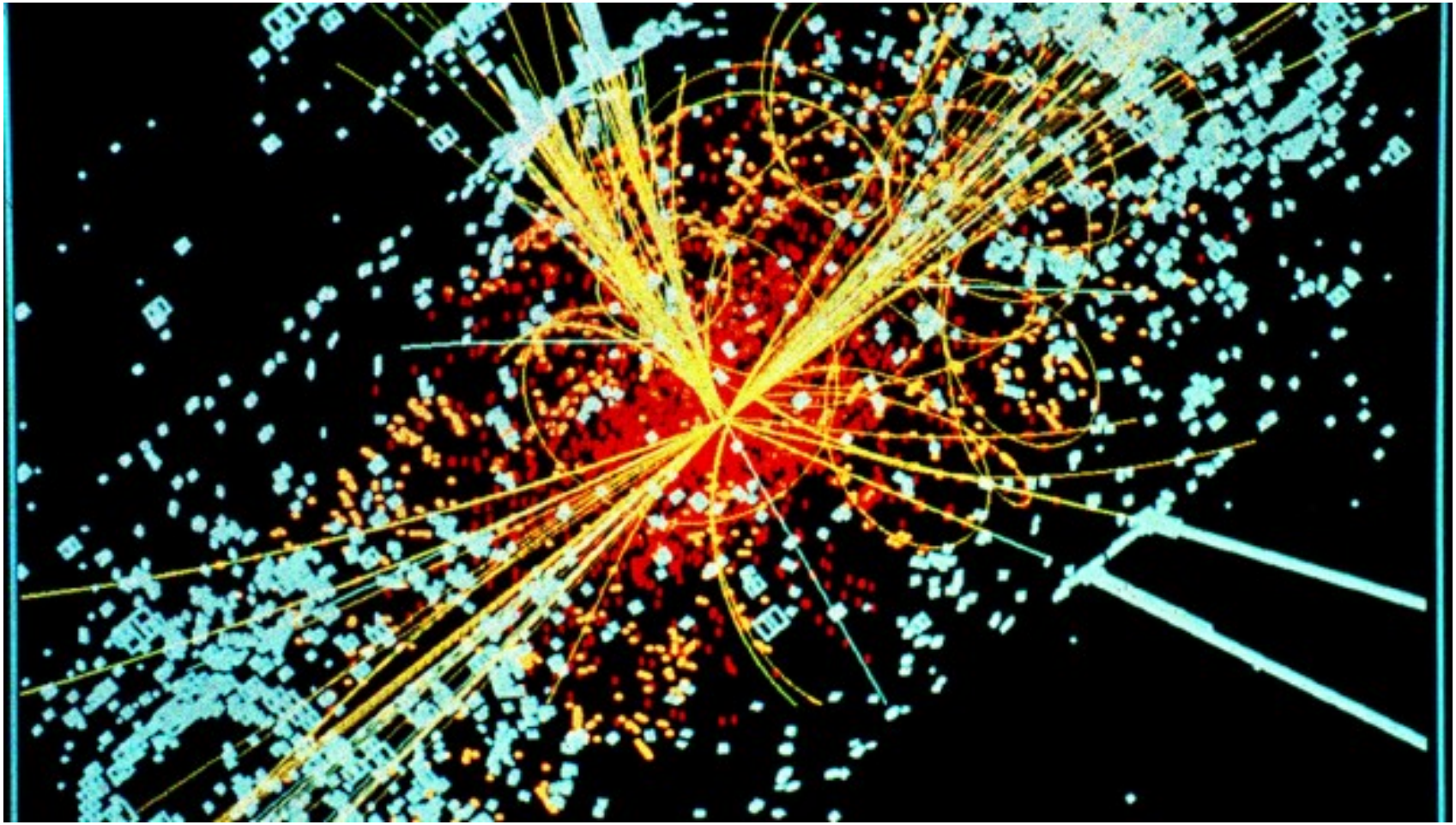
Health





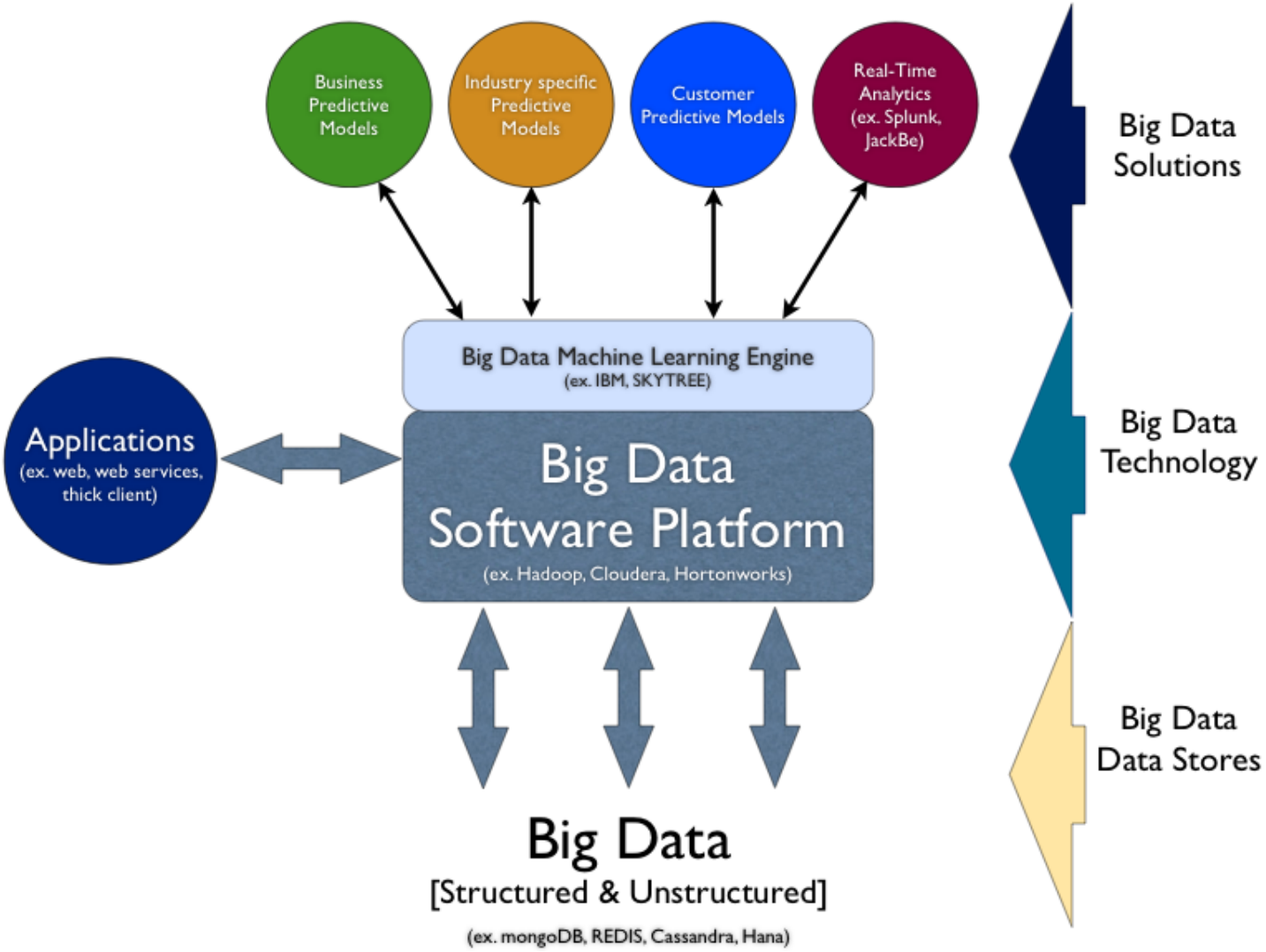
IBM Intelligent Operations software is designed with cities, for cities, to provide the tools to monitor, visualize and analyze vital city services such as water and wastewater systems, transportation, infrastructure planning, permit management and emergency response.


Variant
Bioinformatics
Microarrays
RNA
Gene
Methylation
DNA-seq
Genotyping
CNV
Illumina
Caliper
allelic
QuBit
PCR
fluorescence-luminescence
Covaris
Genomics
libraries
Tiling
nCounter
NanoString
Bioanalyzer
TapeStation
Real-time
discrimination
reader
DNA
arays
profiling
ChIP-seq
analysis
expression
samples
Transcriptomics
TaqMan
cBot
Analysis
SNP
equencing
Tecan
Green
FFPE
ChIP-string
arrays
Sequencing
Q-PCR
Sybr
NGS
BeadStation
miRNA
Exome
Affymetrix
Epigenetics
RNA-seq



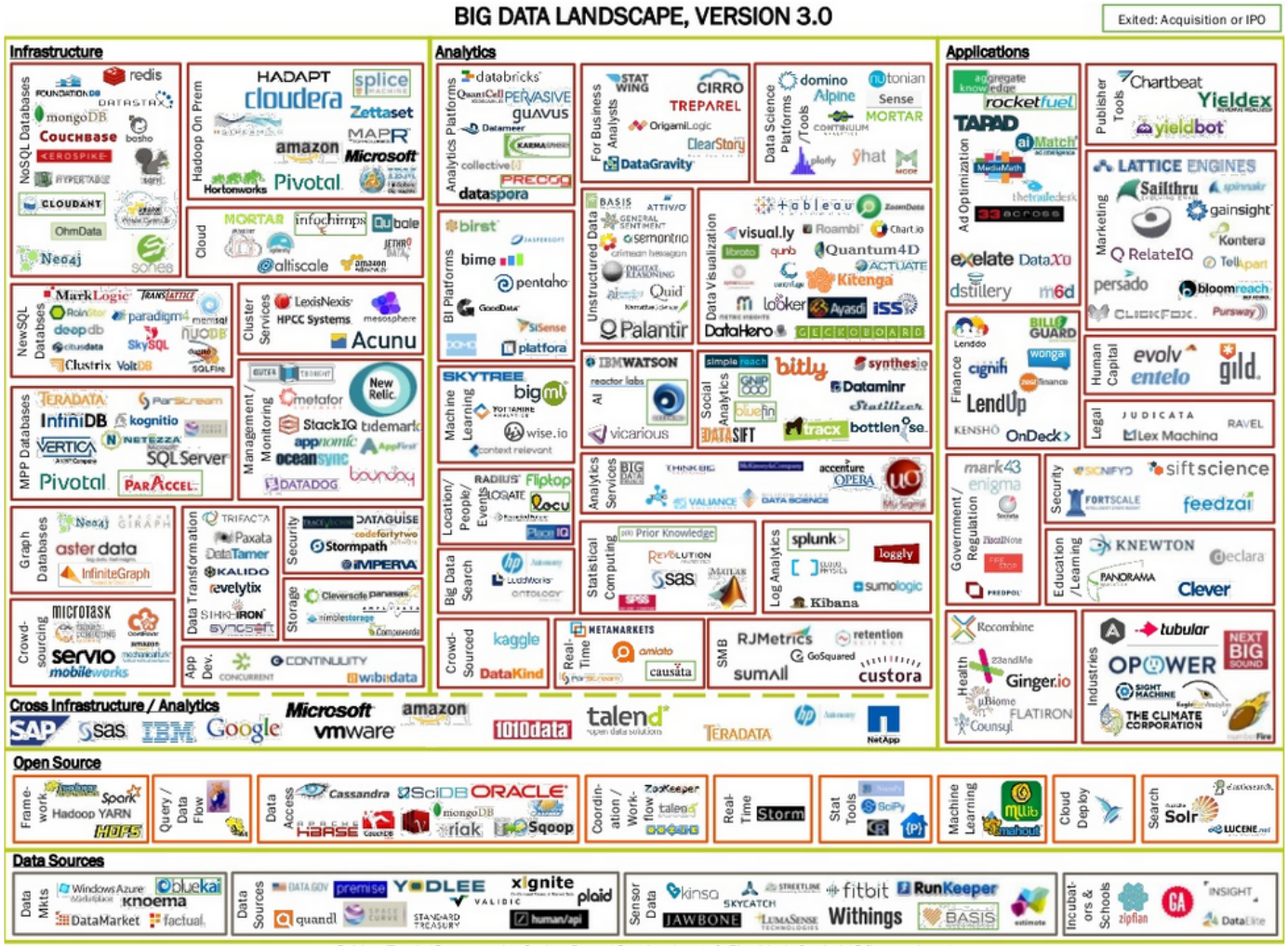
Data Science and Big Data

A Complex Ecosystem!



Data Science and Big Data

A Complex Ecosystem!

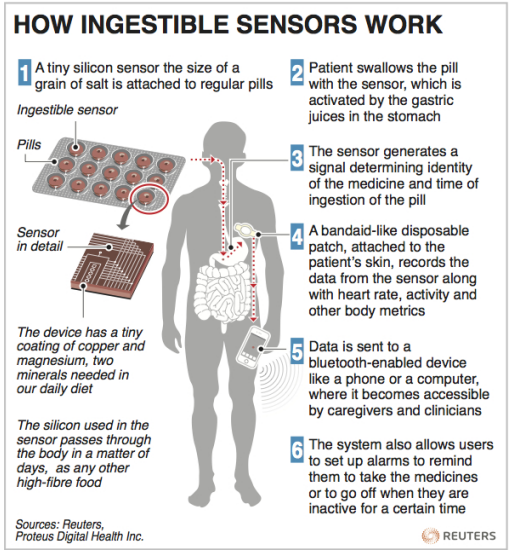


© Matt Turck (@mattturck), Sutian Dong (@sutiandong) & FirstMark Capital (@firstmarkcap)

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

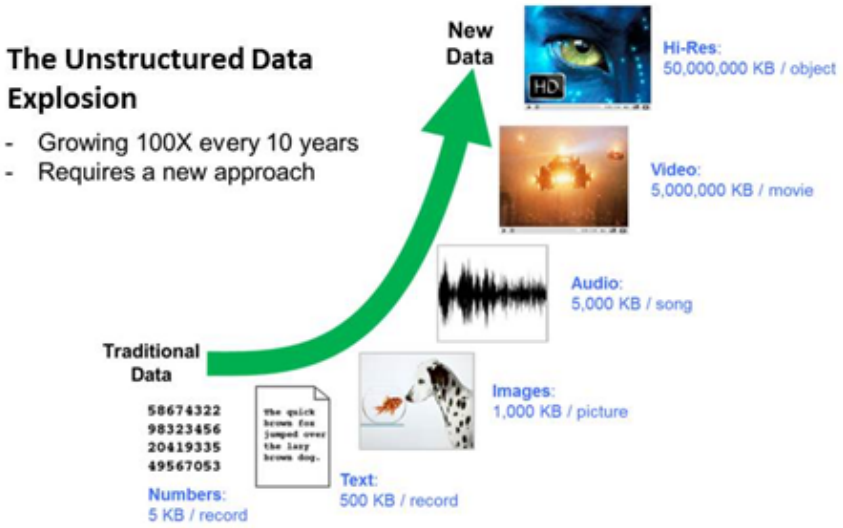
Some joint math/computer science challenges

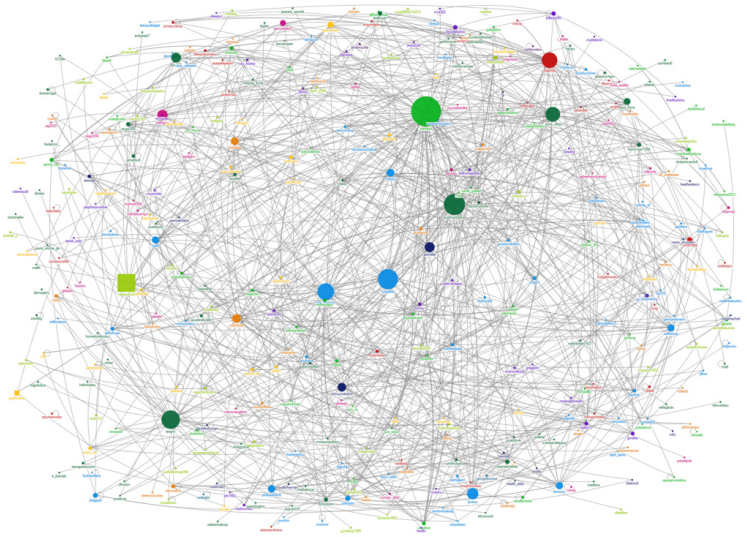
- Data acquisition
- Unstructured data and their representation
- Huge dataset and computation
- High dimensional data and model selection
- Learning with less supervision
- Visualization
- Software(s)...



Some challenges

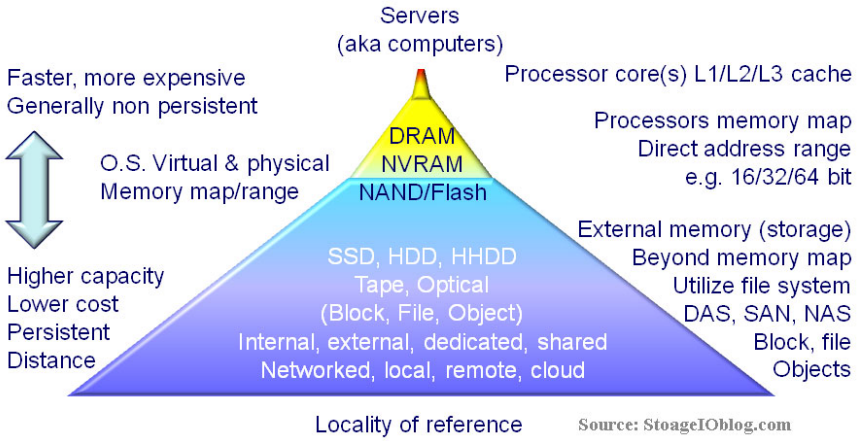
- How to measure new things?
- How to choose what to measure?
- How to deal with distributed sensors?
- How to look for new sources of informations?

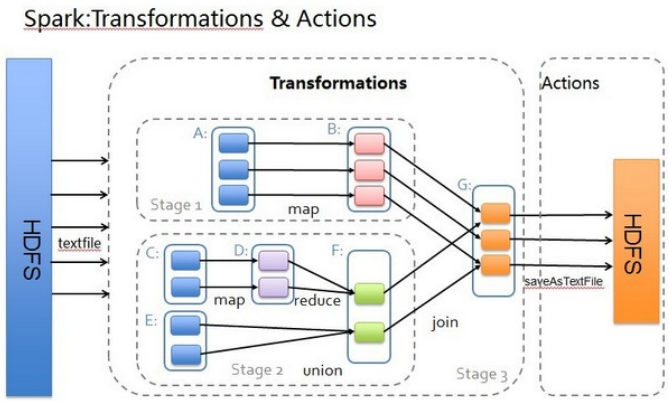




Some challenges

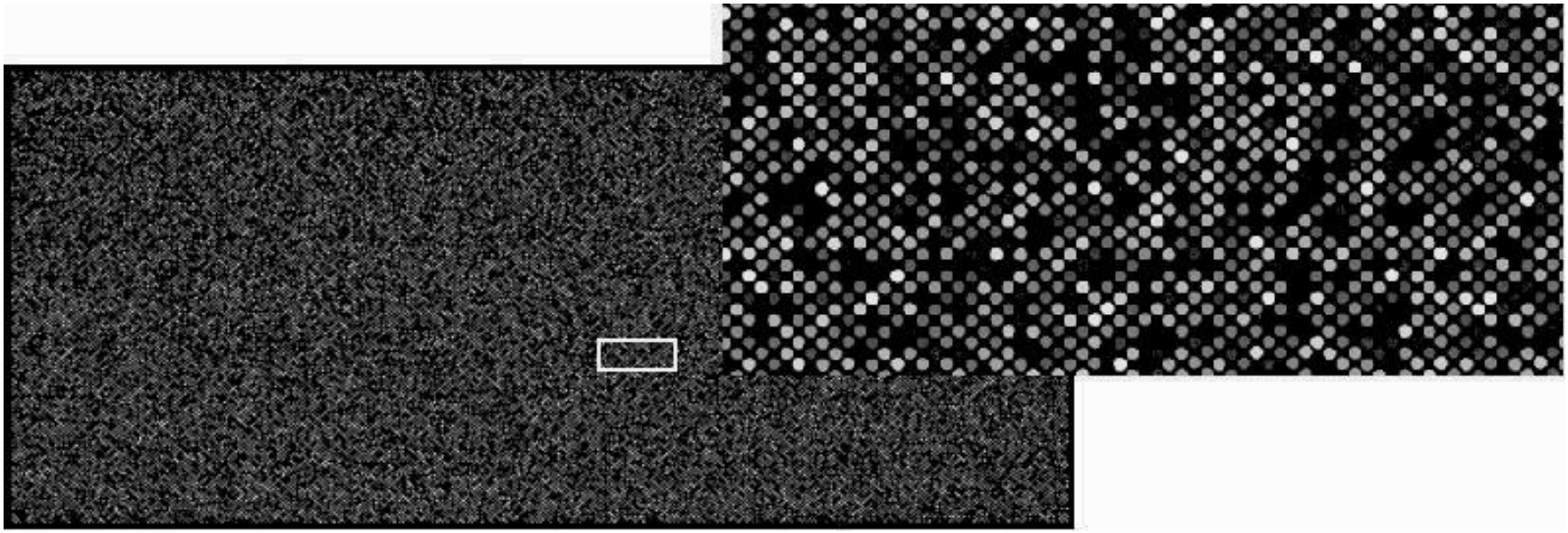
- How to store efficiently the data?
- How to describe (model) them to be able to process them?
- How to combine data of different nature?
- How to learn dynamics?





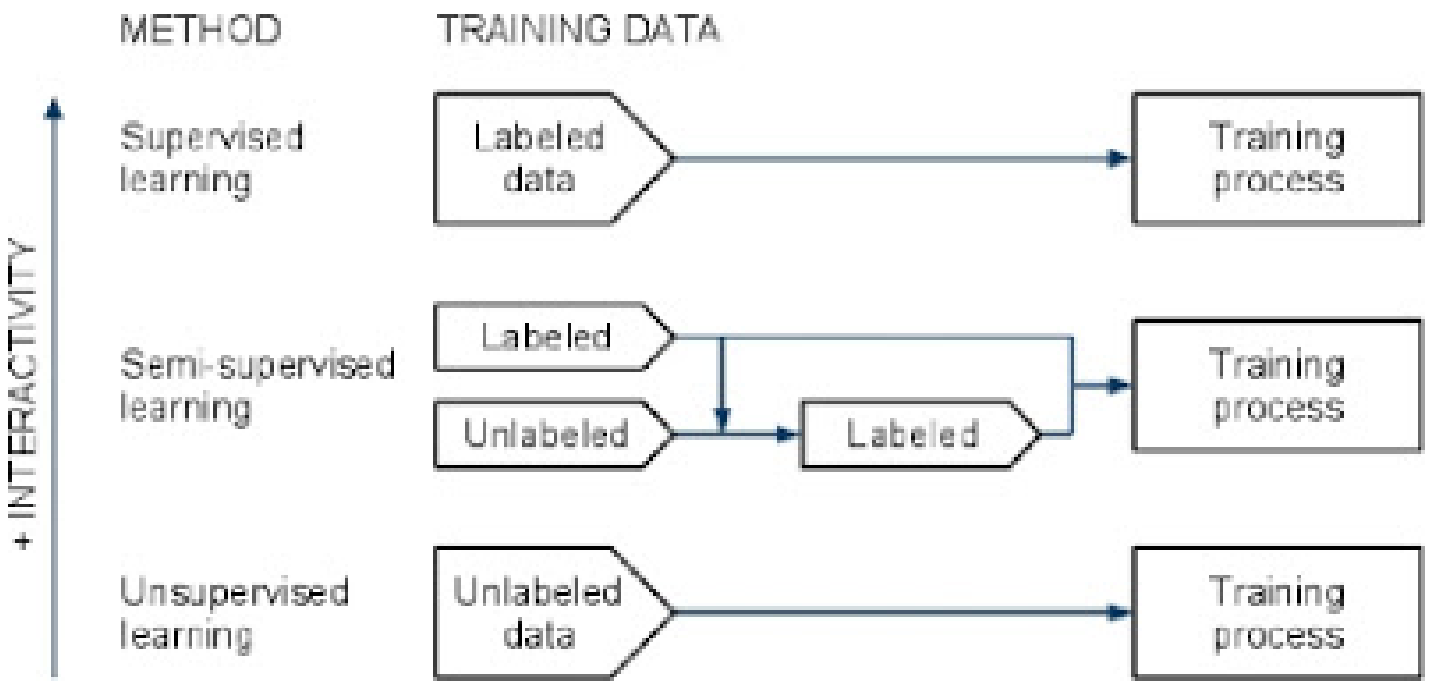
Some challenges

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



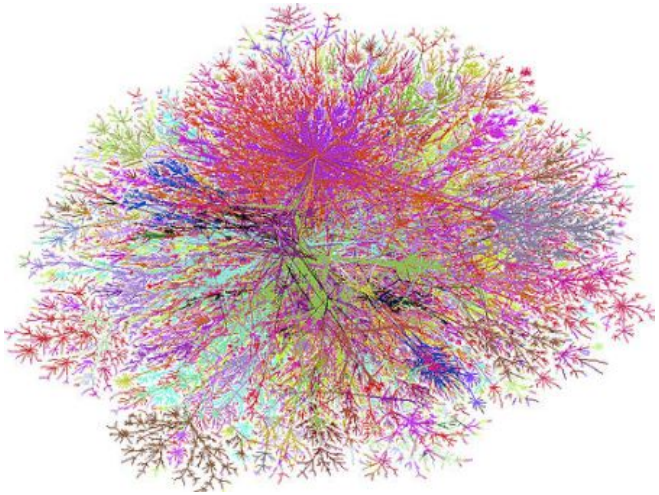
Some challenges

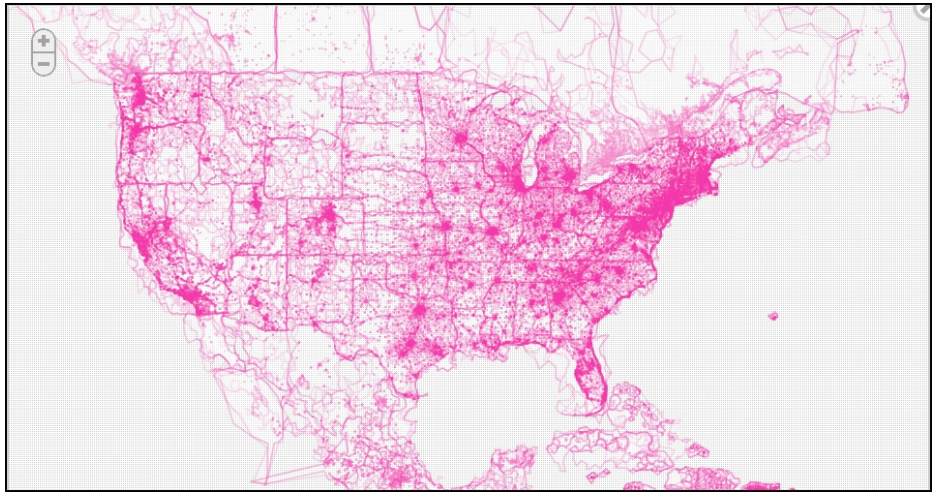
- How to describe (model) the data?
- How to reduce the data dimensionality?
- How to select/mix models?



Some challenges

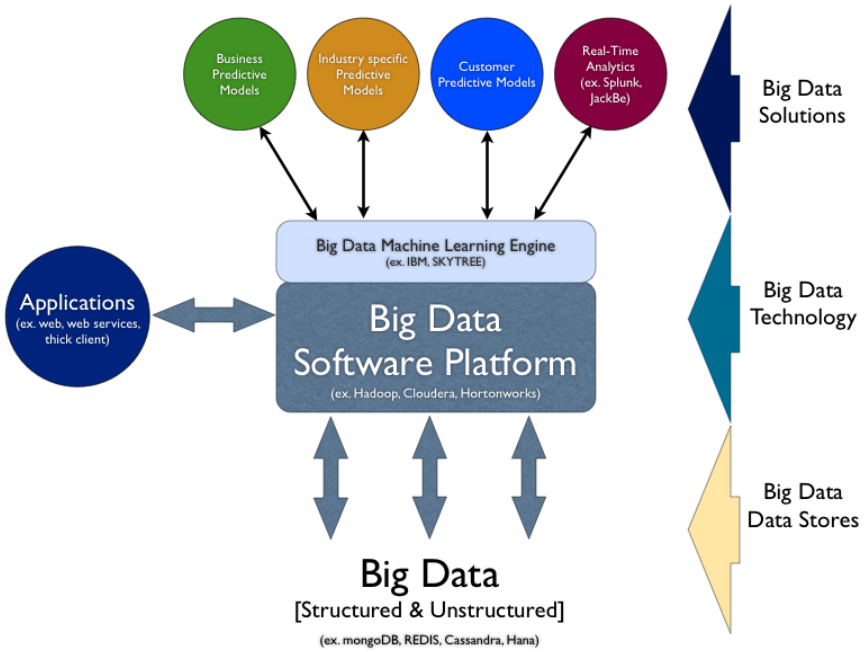
- How to learn with the less possible interactions?
- How to learn simultaneously several related tasks?





Some challenges

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

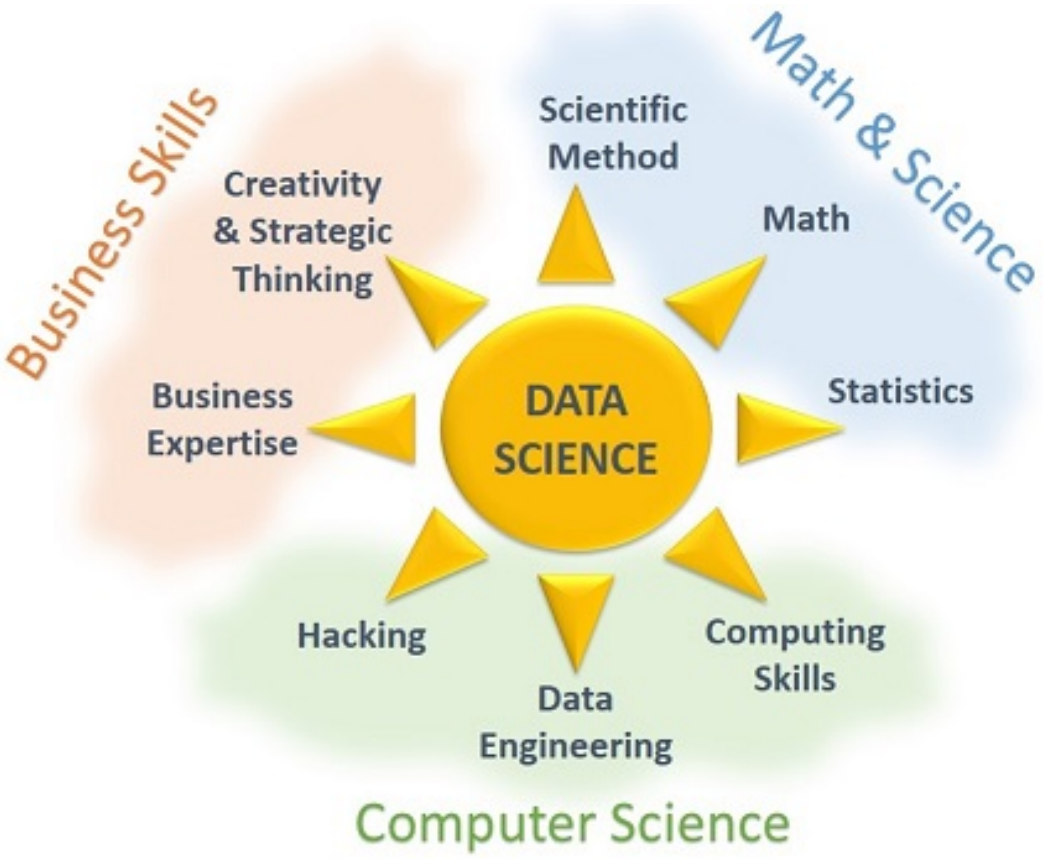


Some challenges

- How to construct a consistent ecosystem?
- How to construct interoperable systems?

Teaching and R

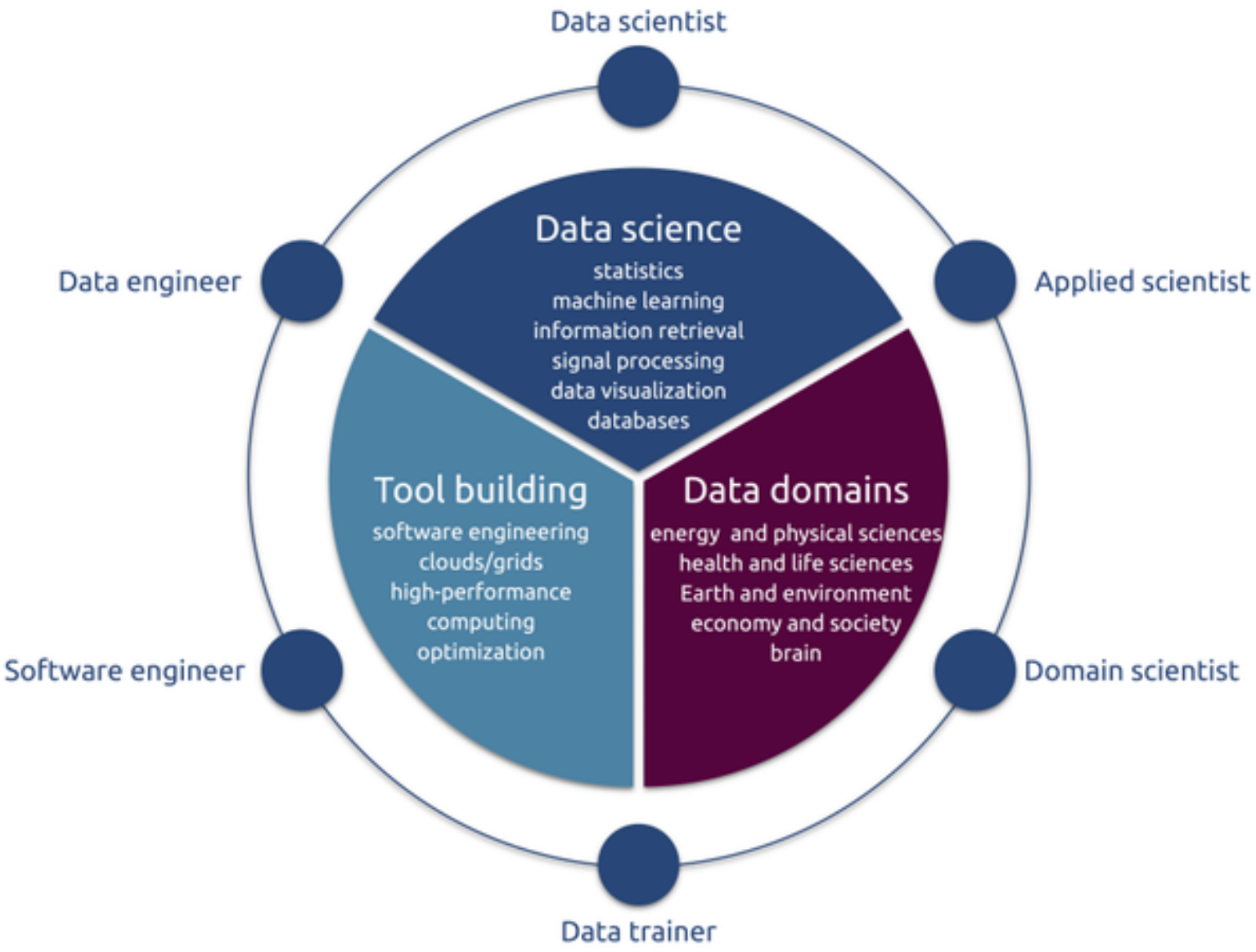
Data Scientists!



- No one masters all the skills!
- Importance of teams.

Teaching and R

More than one type of Data Scientists?



- Focus on the **Data Scientist** profile...

3 years program for **X student**

- 2A: Fundamentals (Math, Stat., Learn., CS) + 1 year project
- 3A: Dedicated track (Appl. Math/CS) + Projects + Internship
- 4A: **Data Sciences Master**

Data Sciences track of the **Master** Mathematics and Application of **Paris Saclay**

- Operated by Polytechnique in collaboration with Telecom, ENSAE, Paris Sud and ENS Cachan.
- Data Scientist training: statistical learning, machine learning, optimization, Big Data technologies...

Data Sciences Starter Program

- 20 days continuous training program.

Data Science Initiative at X

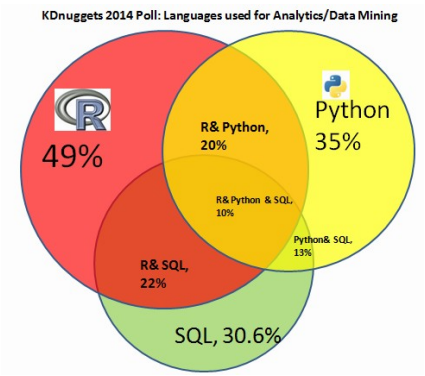
- Support for both teaching and research (projects, collaborations, chairs...)
- Teaching supported by the **Data Scientist** chair (X, Keyrus, Orange, Thales)

R

- Domain Specific Language
- Used in Data Science
- Package Ecosystem:
  - Dataframe (and datatable) + dplyr
  - Ggplot2
  - Huge statistical / machine learning package ecosytem
- Object oriented? + lazy evaluation
- Prototyping?

Python

- Generalist Language
- Used in Data Science
- Package ecosytem:
  - Dataframe + pandas
  - Matplotlib, Bokeh...
  - Scikit-Learn...
- Object oriented
- Production?



Teaching choice

- **R** and **Python**!
- Expose the students to the two most used DS frameworks!
- **Python** already taught!
- Focus on **R** on the Statistic and Machine Learning courses to make them learn a new language.
- Presentation of the **Scikit-learn** + **Python** framework.
- More balanced for the continuous training program.

Personal choice

- **R**:
  - Huge package collection,
  - Graphics and interaction,
  - Flexible language
- Rstudio and Hadleyverse...

Working environment

- RStudio:
  - Well designed IDE
  - Platform independent
  - Rmarkdown!
- Packages:
  - `dplyr` (and friends) for data frame management
  - `ggplot2` (and friends) for graphics
  - `caret` (and friends) for learning

- Not necessarily the best choice for everything or everyone....
- Choice based on the idea of a systematic and coherent syntax...
- R seen as a **glue tool** more than a programming language...
- as S was designed by J. Chambers at Bell Labs in 1976!
- Importance of **literate programming and reproducible science**!

Principles

- Provide a comprehensive script using Rmarkdown
- Let the students manipulate and modify it
- No formal **R** course!
- Introduction to literate programming and reproducible science!

- Future: Mix Rmarkdown / Jupyter ?

Examples

- Illustration of the classical classification methods (Master Data Sciences, Paris Saclay)
- Velib scraping (DSSP, Ecole Polytechnique)
- Credit scoring (Master Économie et Société, Paris Ouest) (A. Fermin)

- Data Science is here to stay...
- R **and** Python are here to stay...
- New ways of teaching Data Science are appearing.
- Interplay between theory and practice.
- Rmarkdown and Jupyter are a good start!