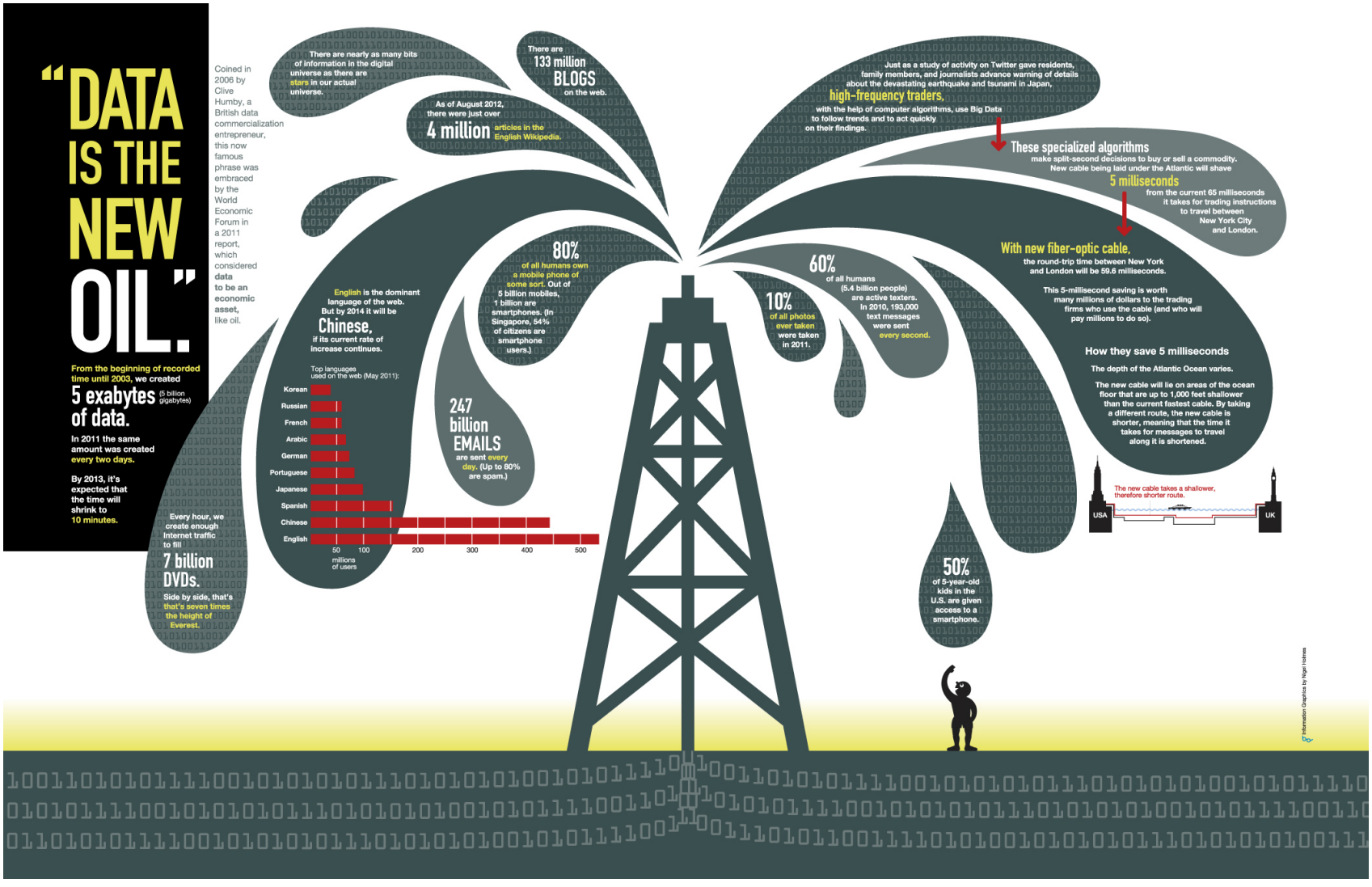
Data Science or Big Data

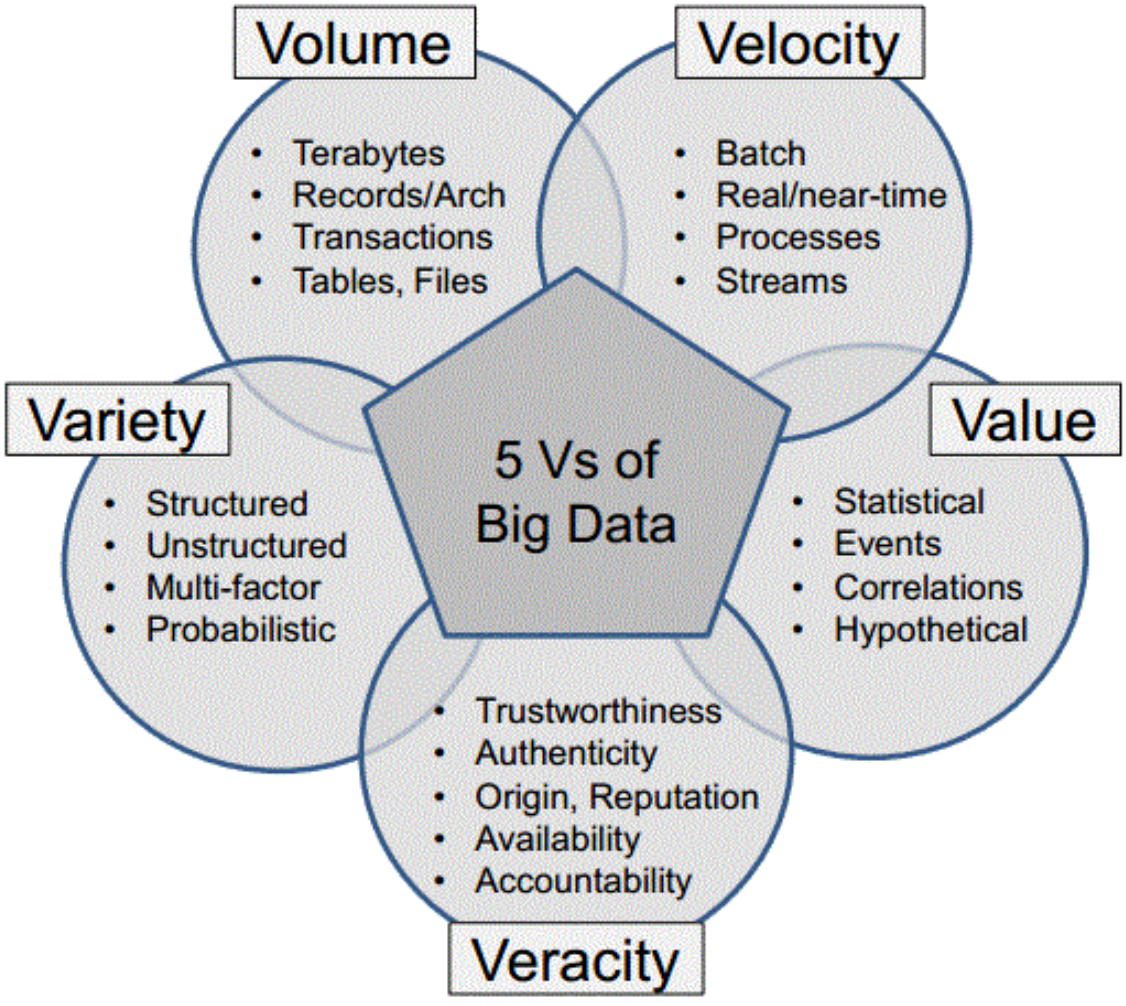
Erwan Le Pennec
Associate Professor
Applied Math Dpt

X Forum - 19/10/2016

Data is the new oil?



The 5 Vs of Big Data



Lots of Words!

Doing Data Science

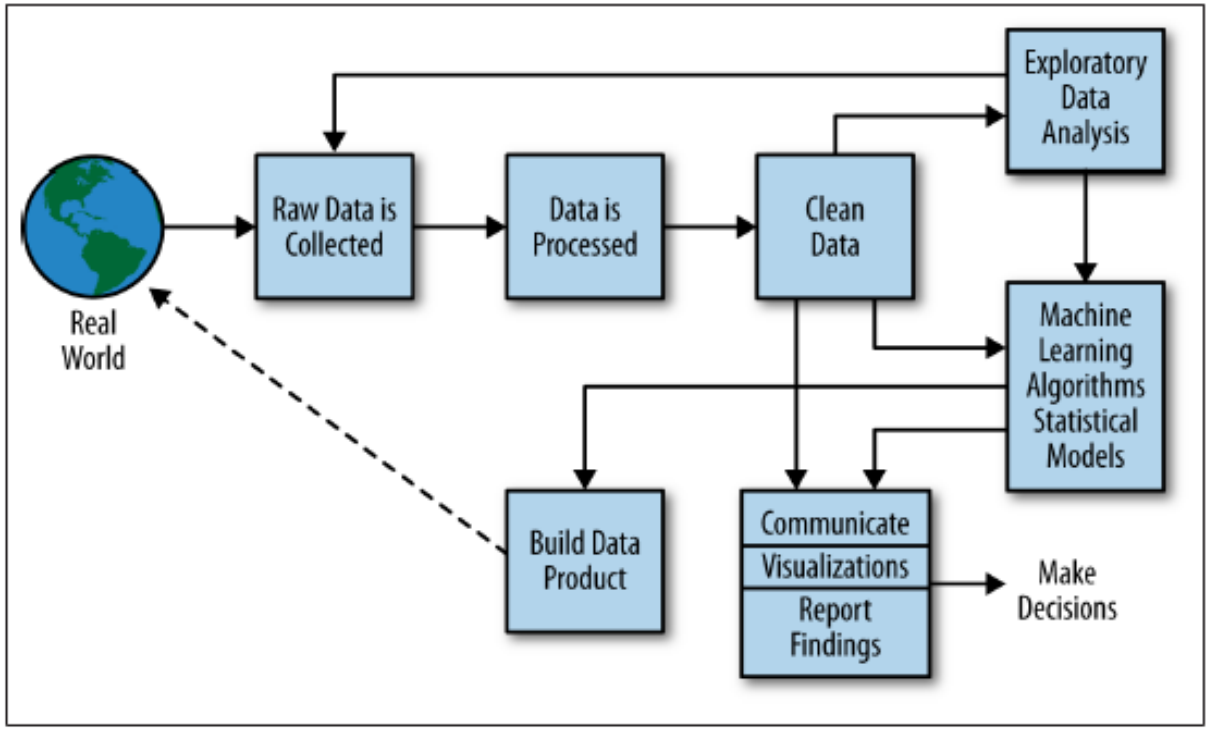


Figure 2-2. The data science process

- Doing Data Science: Straight talk from the frontline.
  - Rachel Schutt, Cathy O'Neil
  - O'Reilly

Data Science

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 Science

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...

Big Data is (quite) Easy

Example of *off the shelves* solution


```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()
}
```

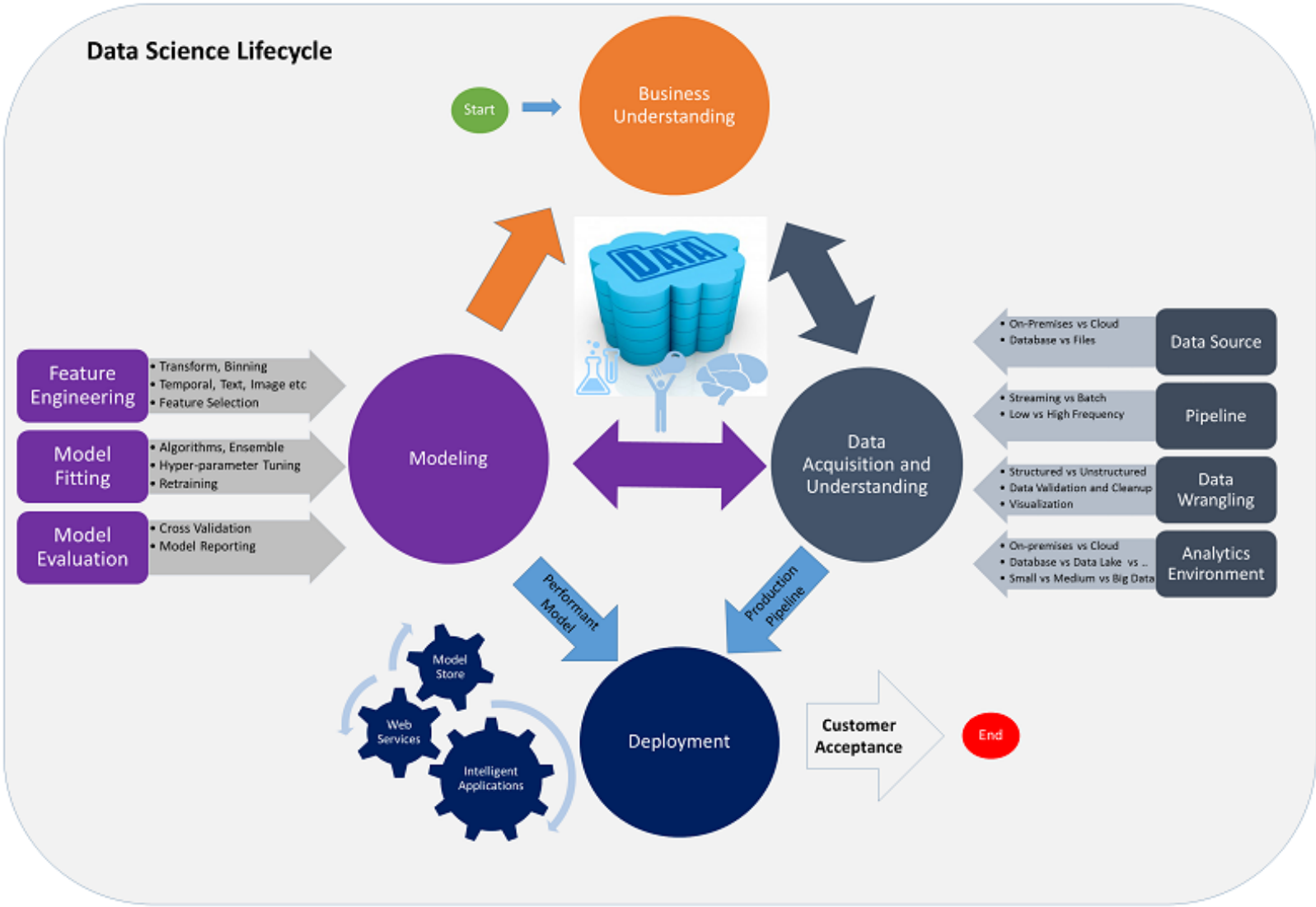
Big Data is (quite) Easy

Example of *off the shelves* solution

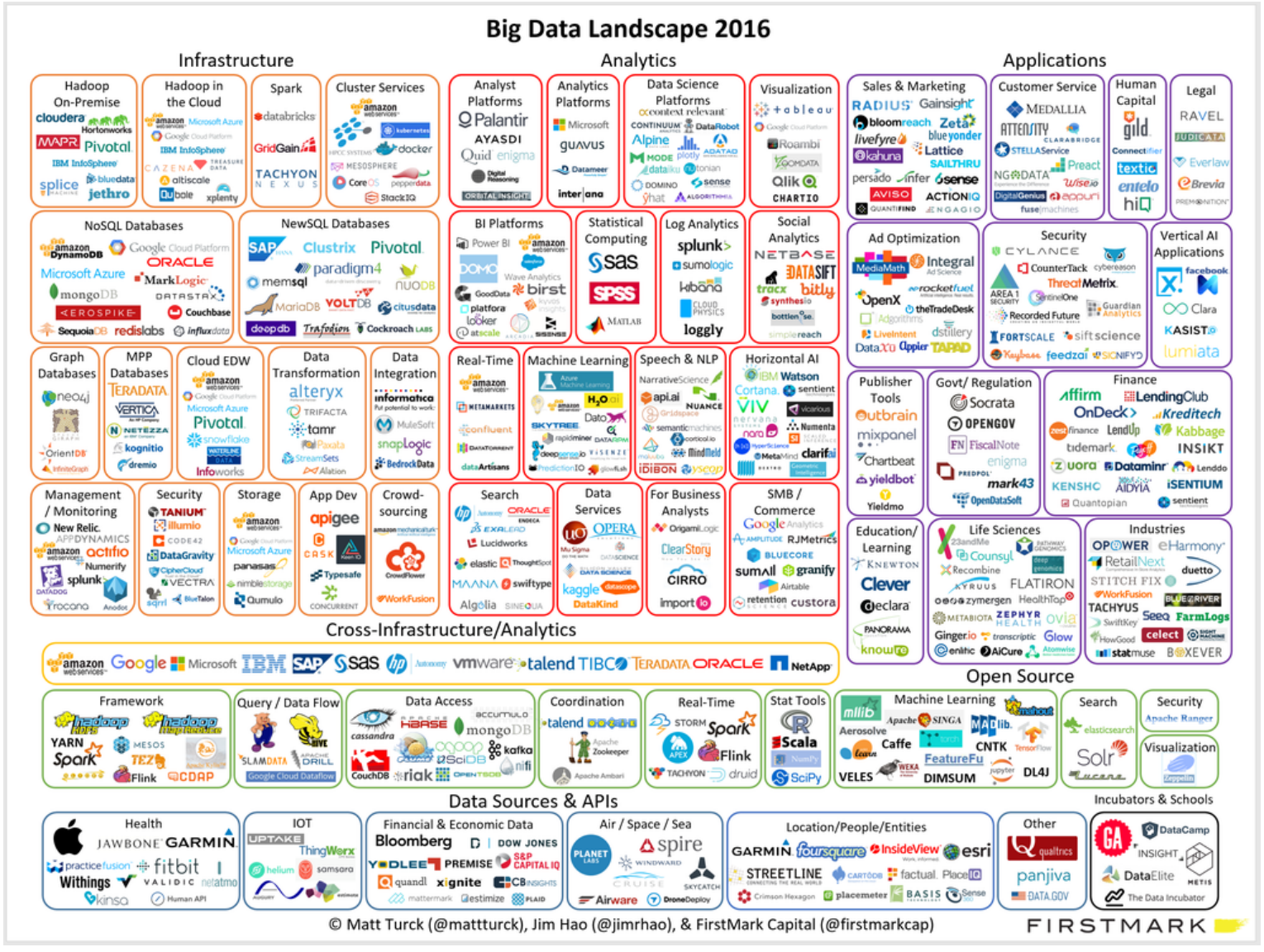

```
export AWS_ACCESS_KEY_ID=<your-access-keyid>
export AWS_SECRET_ACCESS_KEY=<your-access-key-secret>
cellule/spark/ec2/sparl-ec2 -i cellule.pem -k cellule -s <number of machines> launch <cluster-name>
ssh -i cellule.pem root@<your-cluster-master-dns>
spark-ec2/copy-dir ephemeral-hdfs/conf
ephemeral-hdfs/bin/hadoop 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!

Data Science is (quite) Complex!



Data Science is (quite) Complex!



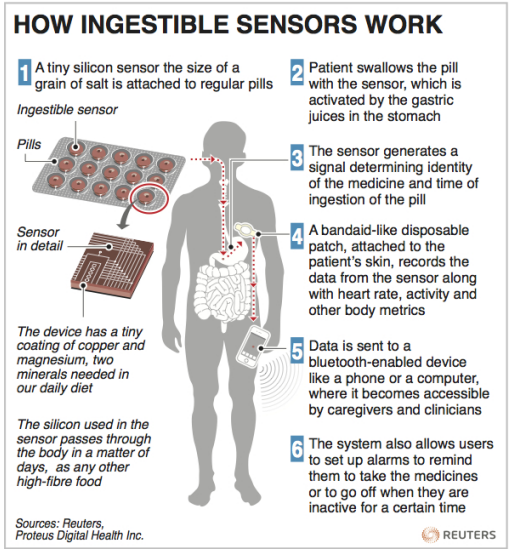
New Interdisciplinary Challenges

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

Some joint math/computer science challenges

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

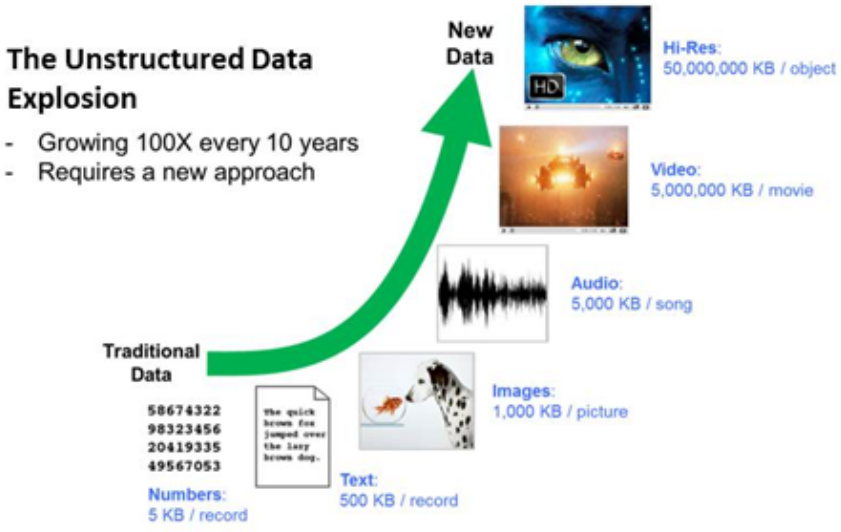
Data Acquisition

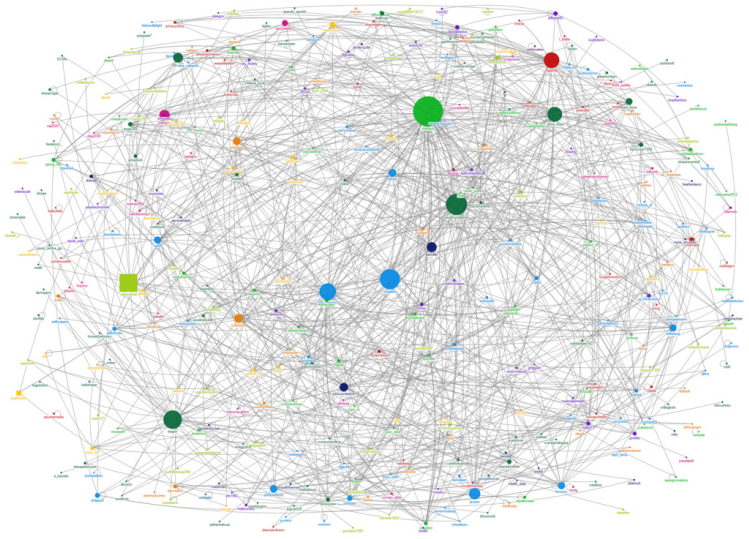


Some challenges

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

Unstructured Data

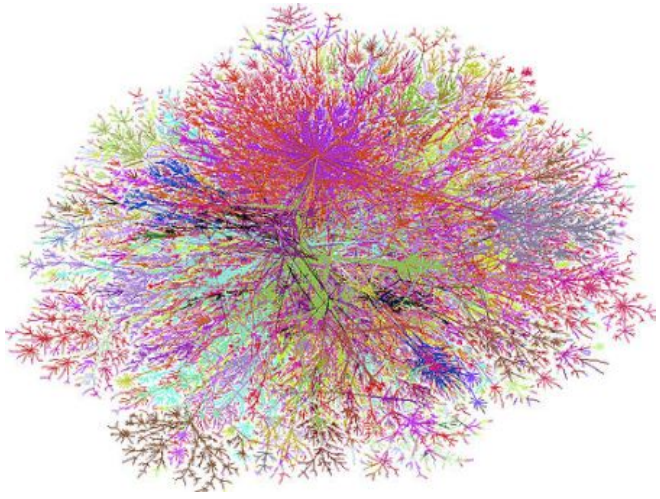


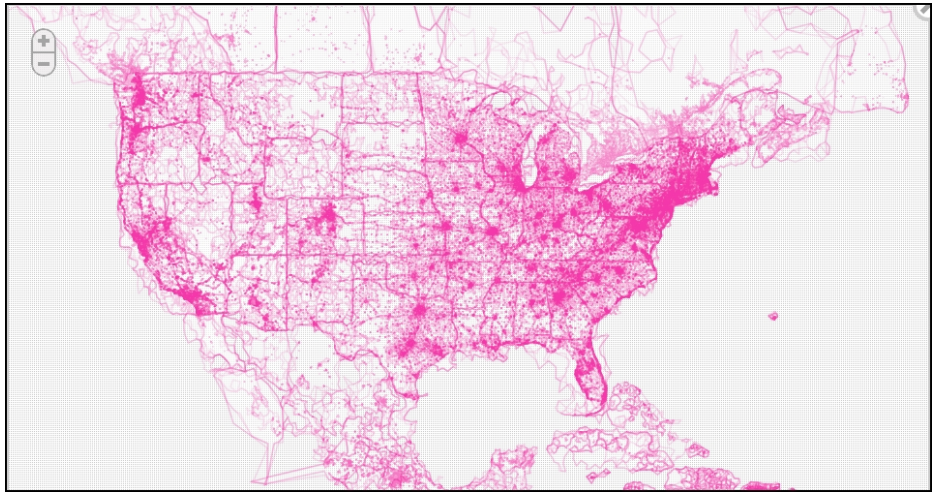


Some challenges

- 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?

Visualization

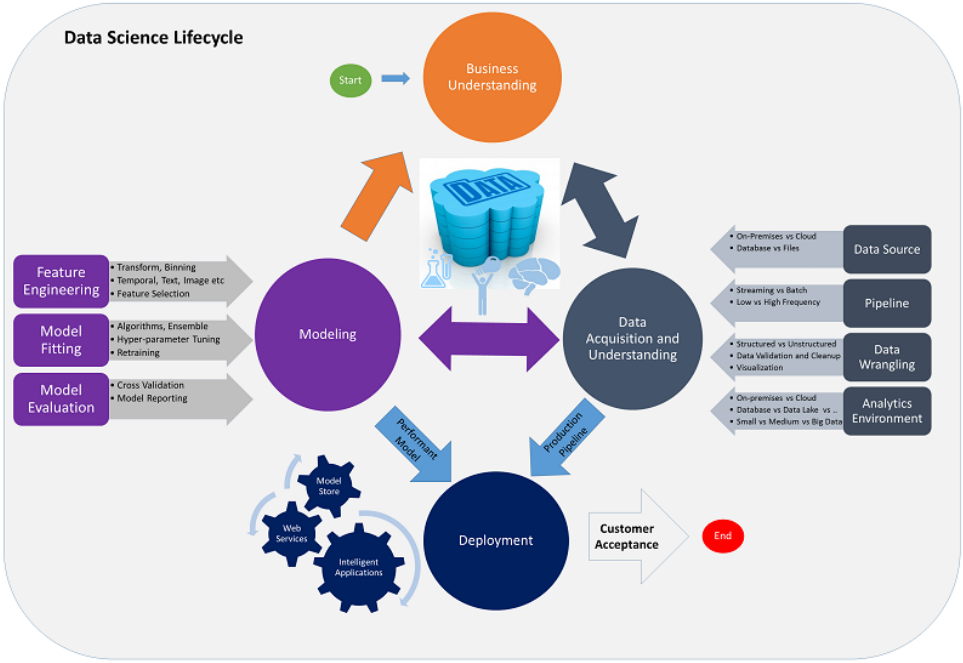




Some challenges

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

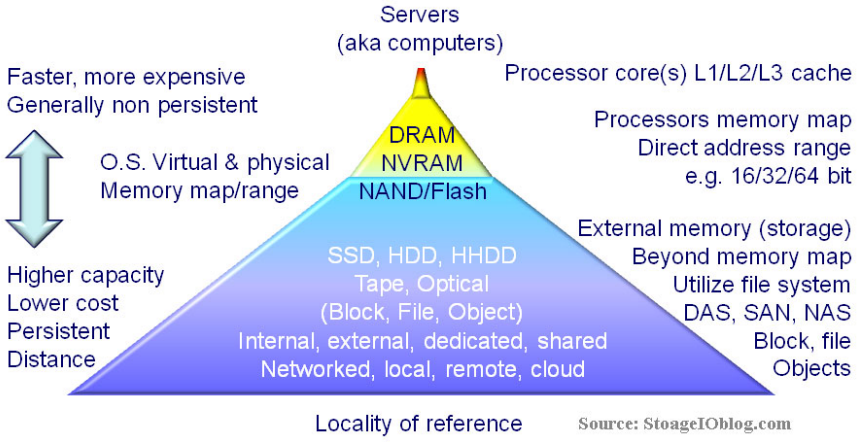
Software(s)

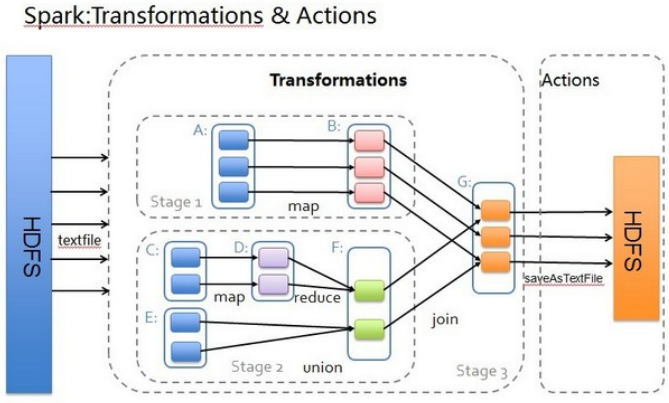


Some challenges

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

Huge Dataset (Big Data)





Some challenges

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

Domain Specific Knowledge

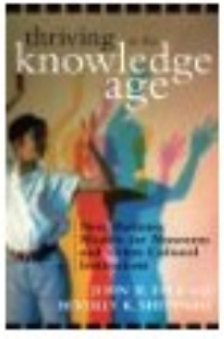
Some challenges

- How to optimize a given process?
- How to find new processes / new outcomes?
- How to completly change a chain of value?
- How to measure the performance?

Recommendation System

More Ideas Based on Your Browsing History

You looked at



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

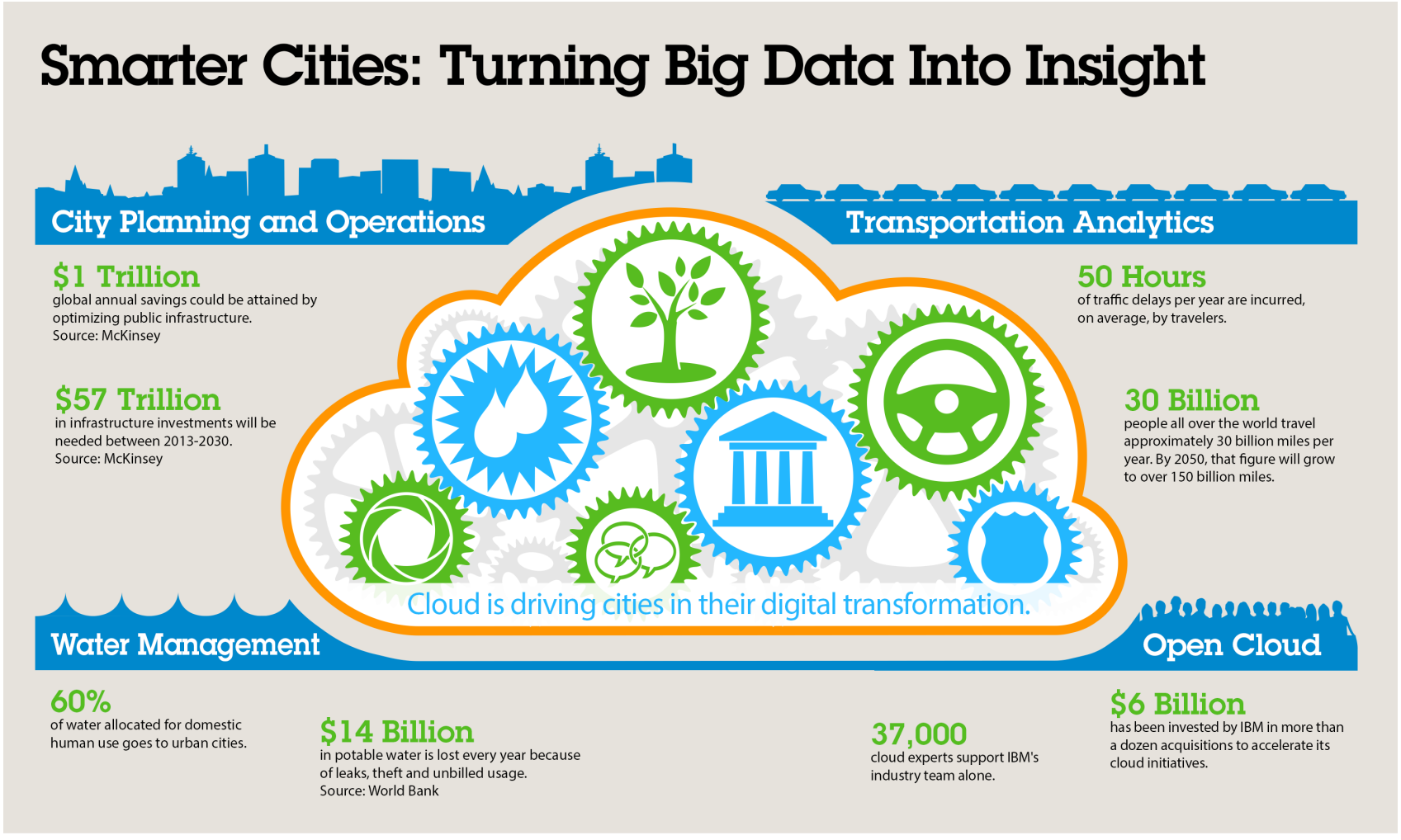
You might also consider

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**

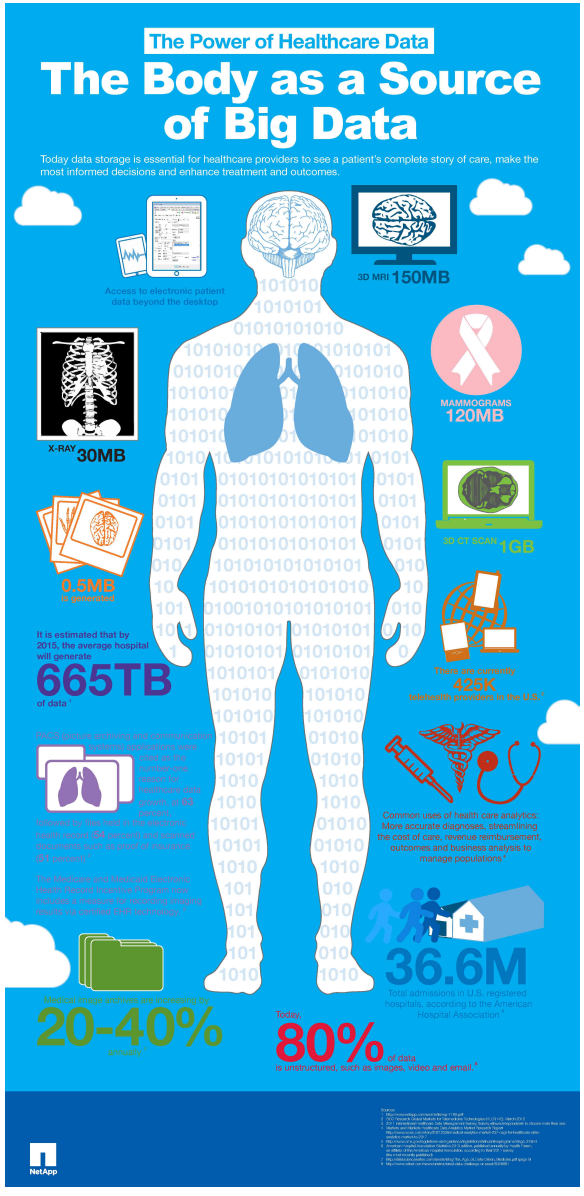
› Find similar items

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

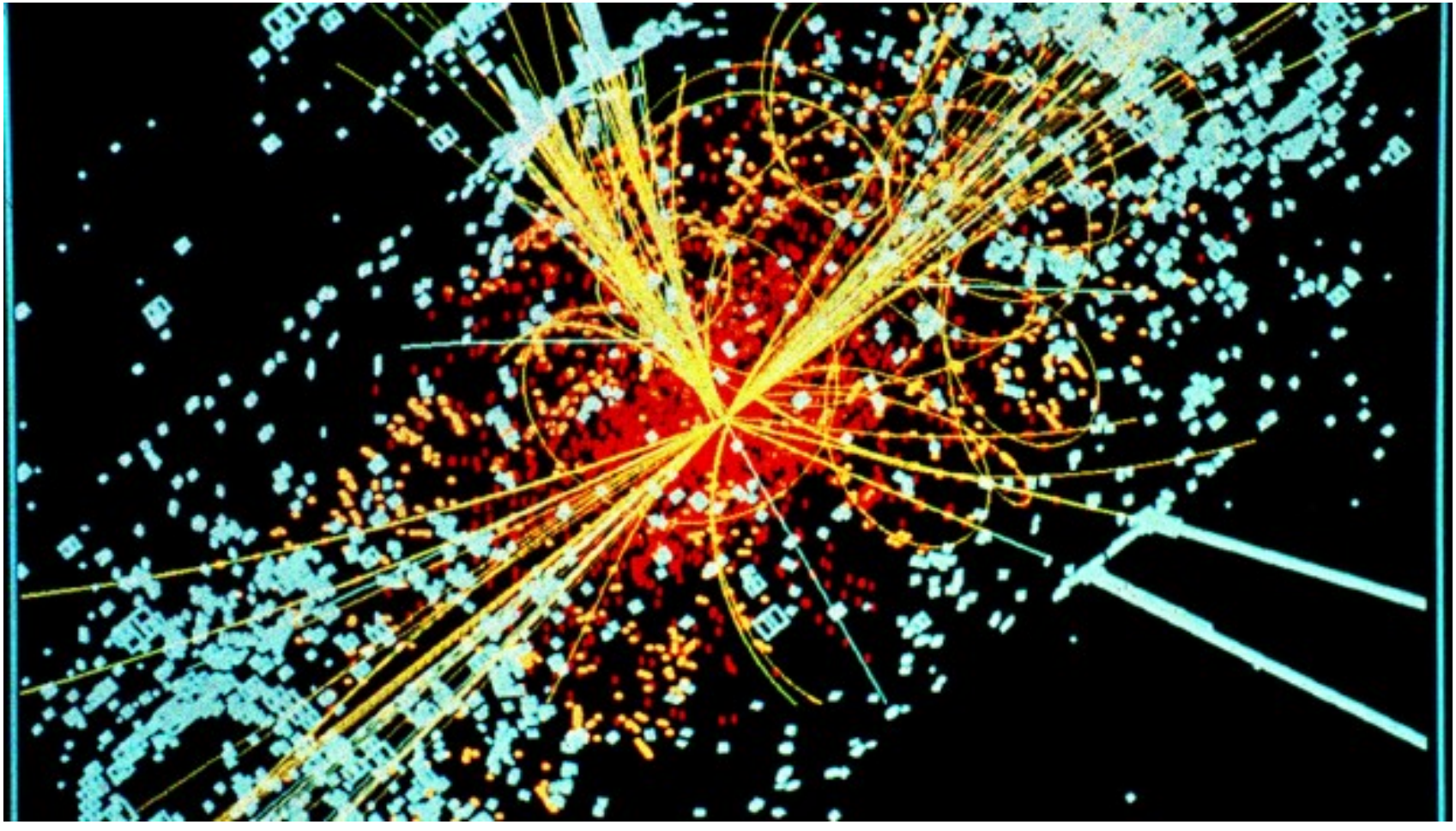


IBM Intelligent Operations software is designed with cities, for cities, to provide the tools to monitor, visualize and analyze vital city services such as water and wastewater systems, transportation, infrastructure planning, permit management and emergency response.

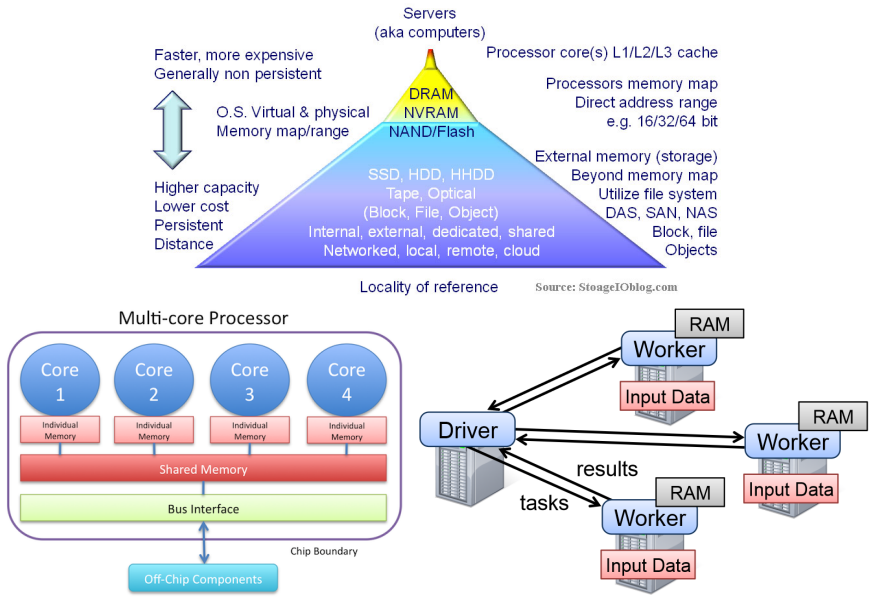
Health


The Power of Healthcare Data
The Body as a Source of Big Data
Today data storage is essential for healthcare providers to see a patient's complete story of care, make the most informed decisions and enhance treatment and outcomes.
Access to electronic patient data beyond the desktop
3D MRI 150MB
X-RAY 30MB
MAMMOGRAMS 120MB
3D CT SCAN 1GB
0.5MB is generated
It is estimated that by 2015, the average hospital will generate
665TB
of data
455K
Common uses of health care analytics: More accurate diagnoses, streamlining the cost of care, revenue reimbursement, outcomes and business analysis to manage populations
36.6M
Total admissions in U.S. registered hospitals, according to the American Hospital Association
Medical image archives are increasing by
20-40%
annually
Today,
80%
of data is unstructured, such as images, video and email.
NetApp

Physics



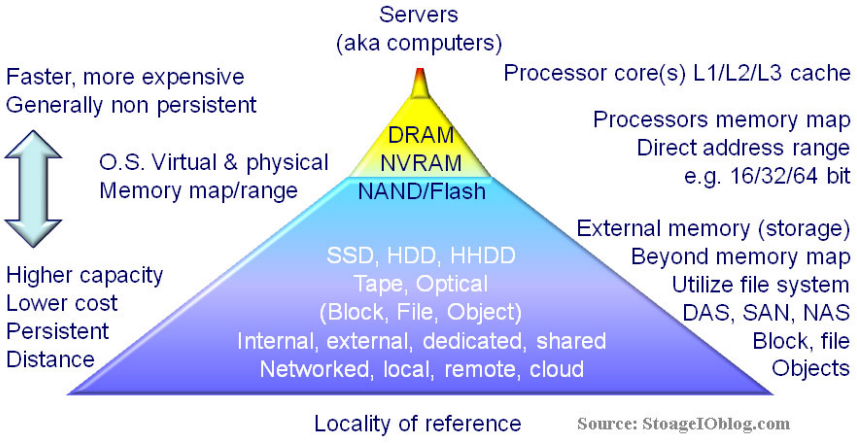
Big Data?

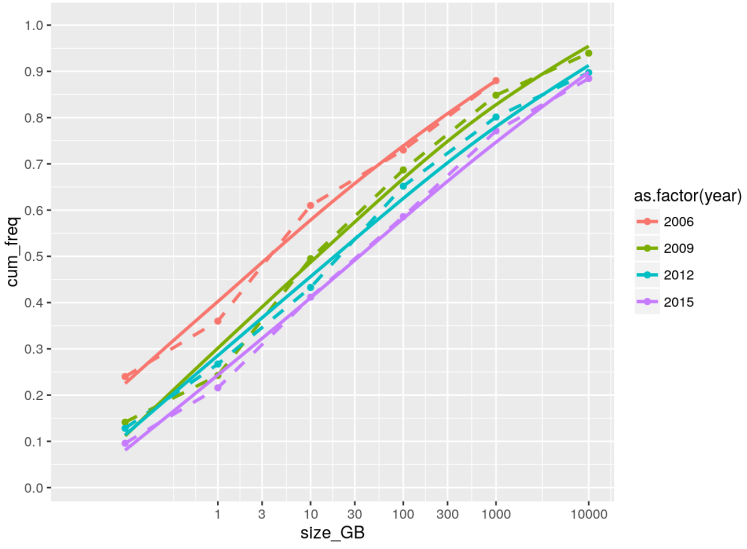


Hardware Constraints

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

Locality of Reference

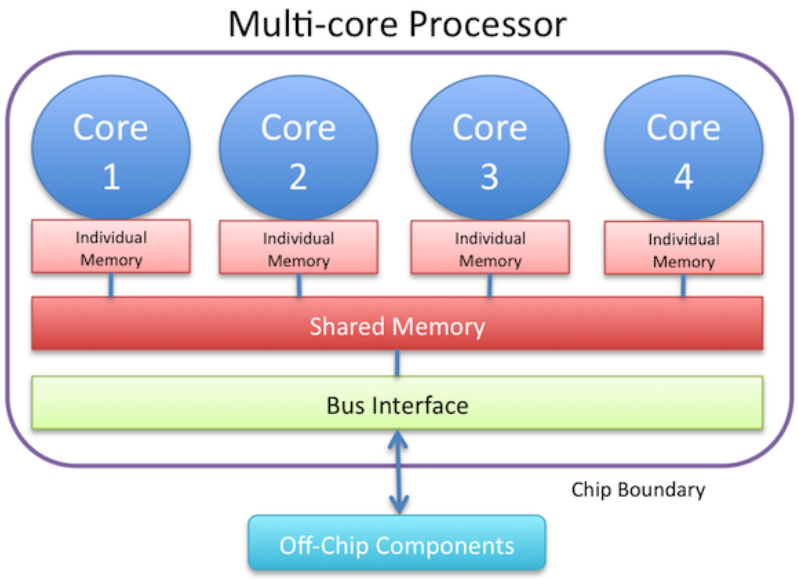




Memory Issue

- 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 growth **faster** than data growth (Death of big data?)
- Memory required may be **larger** than dataset (interactions...)

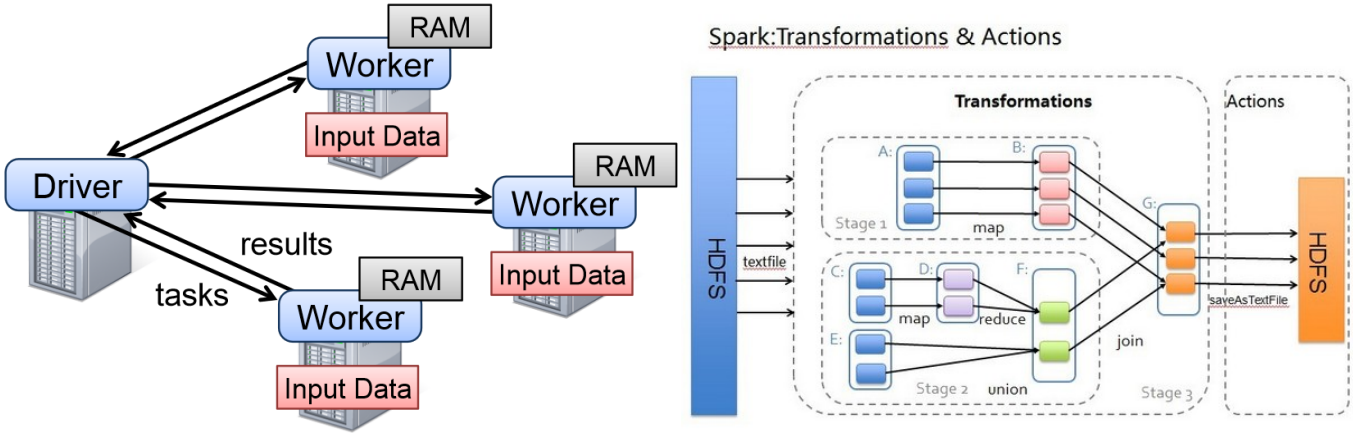
Parallelization



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**!

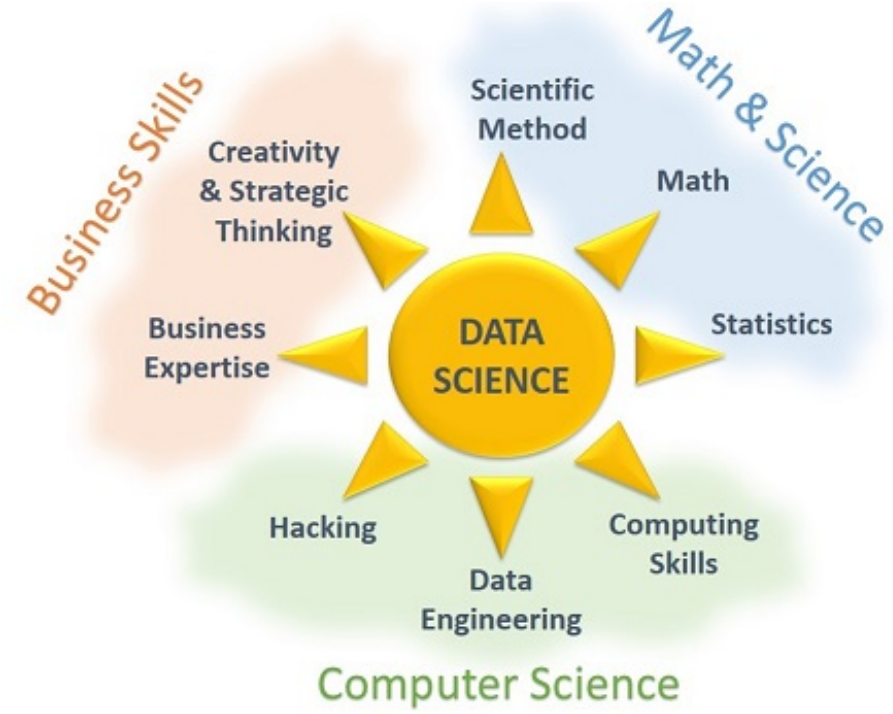
Distribution



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...)

Data Scientists!

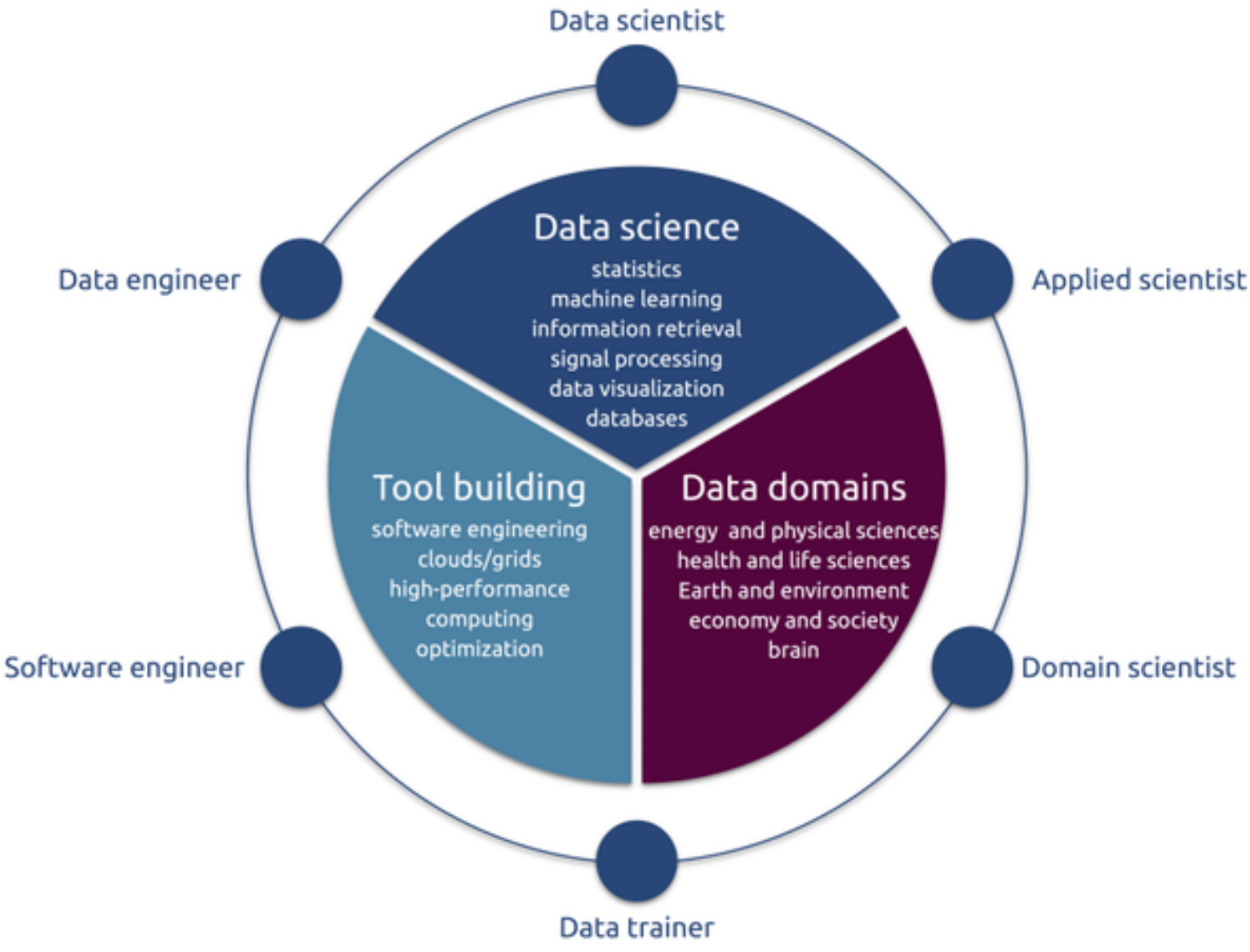


Data Scientist

- **Mix** of Math/Statistics/ML, Computer Science, Business (and Communication) skills.
- **Hard to be an expert of everything!**

- Importance of balanced **teams**.

More than one Type of Data Scientists?



Data Science Teaching at X

- 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 Science at X

X/HEC Msc Big Data for Business

- 2 year program with both the technical skills and the strategic mindset for a business career with a Big Data expertise.

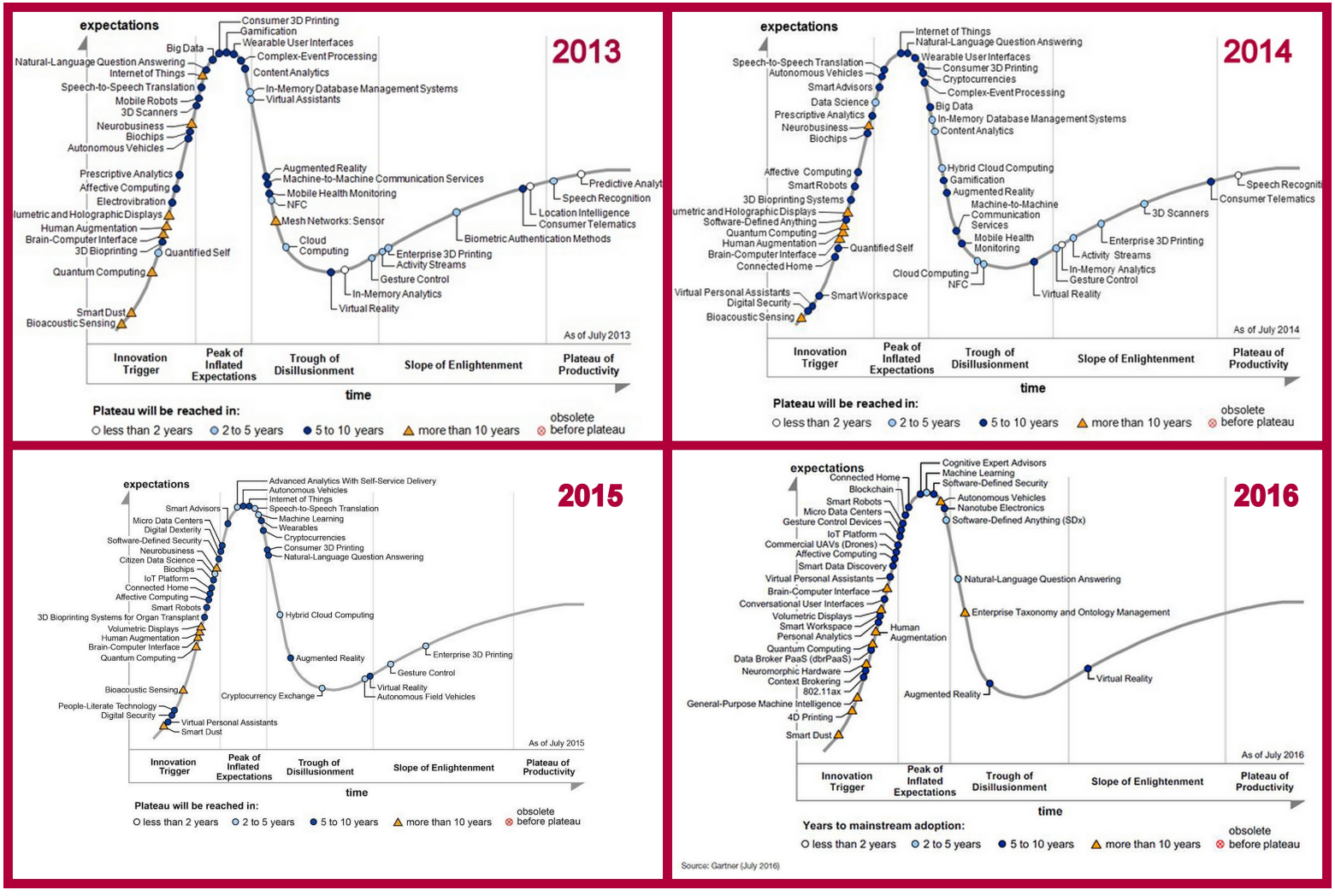
Data Sciences Starter Program

- 18 days continuous training program.
- RNCP certification.

Data Science Initiative at X

- More than **12 permanent researchers**, **20 Phd students/Post doc**, **10 engineers**!
- **Teaching and Research** (projects, collaborations, chairs...)
- Examples:
  - Collaboration with CNAM.
  - **Data Scientist** chair (X, Keyrus, Orange, Thales)
  - **Data Science for Insurance Sector** chaire (Axa)
  - ...

Don't Believe the Hype?



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

Take Away Message

Data Science is here to stay...

- **How to enhance decisions based on data!**
- **Business/Communication** as important as **ML/Stat/CS**.
- If you are interested, **learn and try!**:
  - **Initial training:** Masters, Lectures...,
  - **Online training:** Books, Mooc...,
  - **Continuous training:** DSSP, Telecom...
  - **Practice:** R/Python, Hadoop/Spark, Kaggle/Challenge Data, Hackathon/RAMP...
- Lot of **opportunities** for a variety of **profiles**:
  - **Data Scientist** (strong in Stat/ML and Algorithm)
  - **Domain Scientist** (strong in Business/Sciences)
  - **Data Engineer** (strong in Development/Hardware)