

New Algorithms for Complex Data

Eric Matzner-Løber, Nick Hengartner and Erwan Le Pennec
Rennes University (France), Los Alamos National Laboratory (USA),
Polytechnique (France)

19/20 March 2015

- Third edition: 2011, 2013 and **2015!**
- New place (and one new *organizer*) but same spirit.
- Gather domain experts and give them the opportunity to talk to each other.
- Complex Data?

Data is the new Oil!

"DATA IS THE NEW OIL."

From the beginning of recorded time until 2000, we created **5 exabytes** of data. (5 billion gigabytes)

In 2011 the same amount was created every two days.

By 2013, it's expected that the time will shrink to **10 minutes**.

Every hour, we create enough internet traffic to fill **7 billion DVDs**.

Each day, AOL, eBay's, and Amazon's servers store the height of Mount Everest.

Created in 2008 by Clay Shirky, a leading data communication anthropologist, this new format primer was endorsed by the World Economic Forum in a 2011 report, which considered data to be an economic asset, like oil.

There are nearly as many bits of information in the digital universe as there are stars in our actual universe.

There are **135 million BLOGS** on the web.

As of August 2012, there were just over **4 million** articles in the English Wikipedia.

80% of all humans own a mobile phone. Out of 5 billion mobile phones, 1 billion are smartphones. In Singapore, 94% of citizens are smartphone users.

English is the dominant language of the web. But by 2014 it will be **Chinese**. If its current rate of increase continues.

This information came from the web, May 2011



247 billion EMAILS are sent every day (84 to 80% are spam).

60% of all business e-mail (pending) are deleted. In 2010, 100,000 lost messages were sent every second.

Just as a study of activity on Twitter gave insiders, family members, and journalists advance warning of attacks about the assassinating earthquake and tsunami in Japan, **high-frequency traders**, with the help of computer algorithms, use Big Data to follow trends and to act quickly on their findings.

These specialized algorithms make split-second decisions to buy or sell a commodity. These calls being sent under the Atlantic will shave **5 milliseconds** from the current 65 milliseconds it takes for trading instructions to travel between New York City and London.

With new fiber-optic cable, the record 39.5 time between New York and London will be 39 milliseconds. This 5-millisecond saving is worth many millions of dollars to the trading firms who use the cable (and who will pay millions to do so).

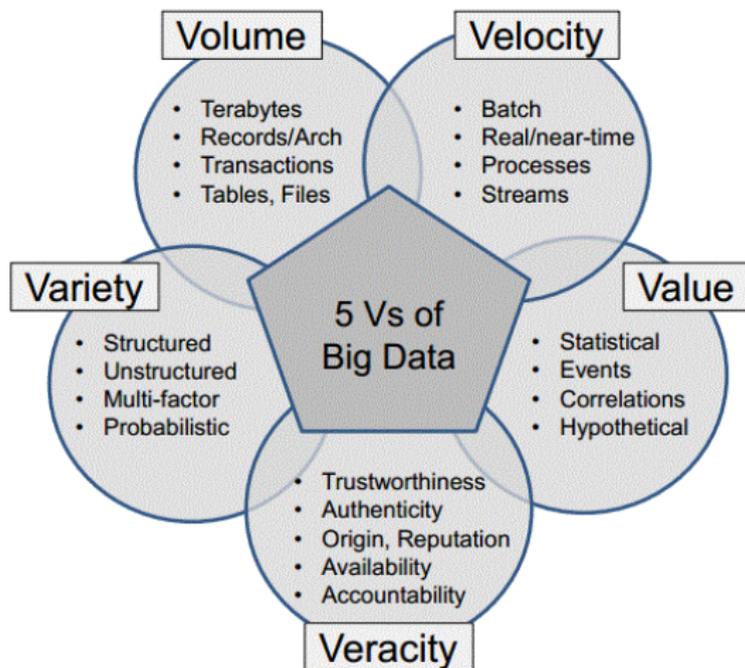
How they save 5 milliseconds

The depth of the Atlantic Ocean varies. The new cable will be on areas of the ocean floor that are up to 1,000 feet shallower than the current buried cable. By taking a different route, the new cable is shorter, meaning that the time it takes for messages to travel along it is shortened.



50% of 8-year-old kids in the U.S. are given access to a smartphone.

The 5 Vs of Big Data



SO MUCH DATA, SO LITTLE TIME

There are over 1000 known archaeological sites in the park, but as you can see from the gaps on this map, most of the park has never been surveyed!



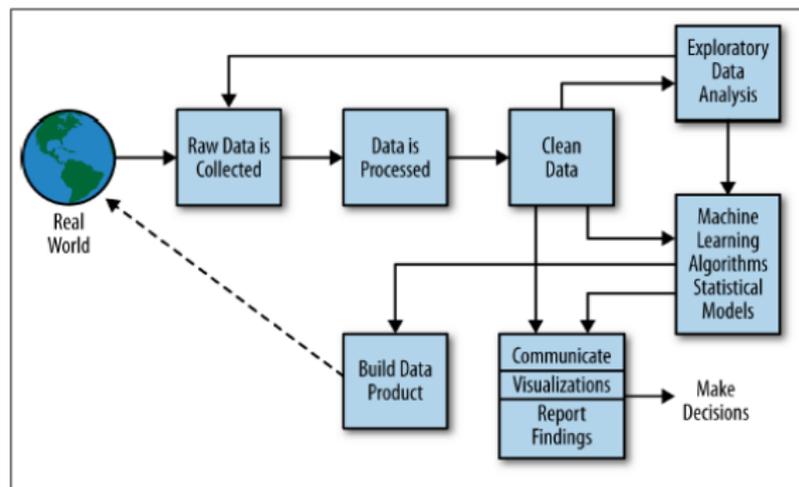


Figure 2-2. The data science process

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!

Big Data is (quite) Easy

Example of *off the shelves* solution



```
def run(params: Params) {
  val conf = new SparkConf()
    .setAppName(s"BinaryClassification with Sparsms")
  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()
    .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



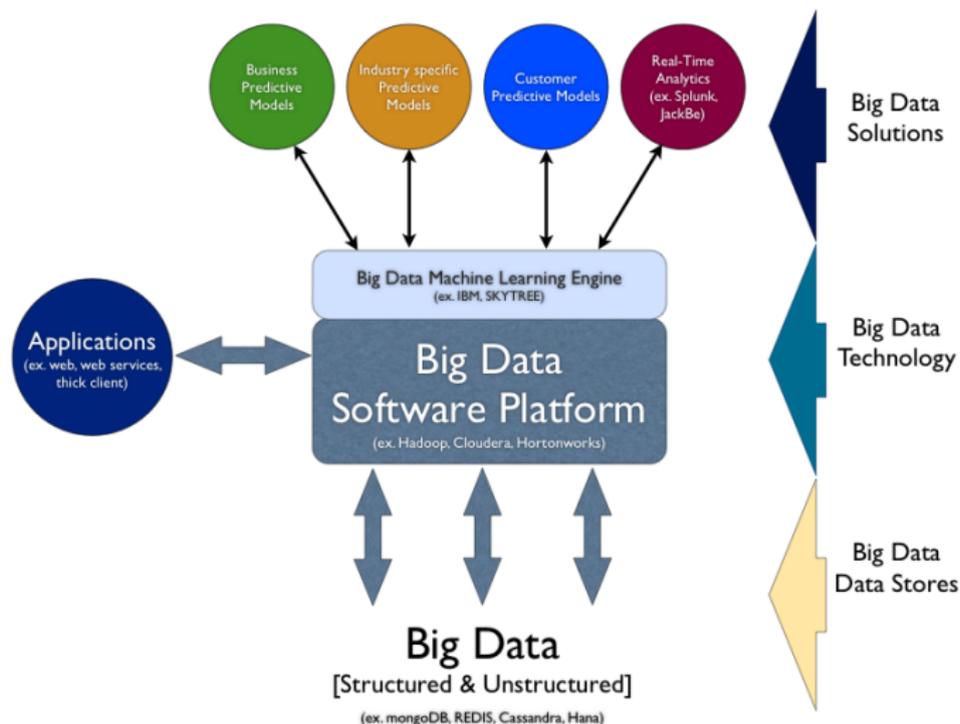
```
export AWS_ACCESS_KEY_ID=<your-access-keyid>
export AWS_SECRET_ACCESS_KEY=<your-access-key-secret>
cellule/spark/ec2/spar1-ec2 -i cellule.pem -k cellule -s <number of machines> launch <cluster-name>
ssh -i cellule.pem root@<your-cluster-master-dns>
spark-ec2/copy-dir ephemeral-hdfs/conf
ephemeral-hdfs/bin/hadoop 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
```

⇒ Logistic regression for arbitrary large dataset!

A Complex Ecosystem!



A Complex Ecosystem!

Big Data Landscape



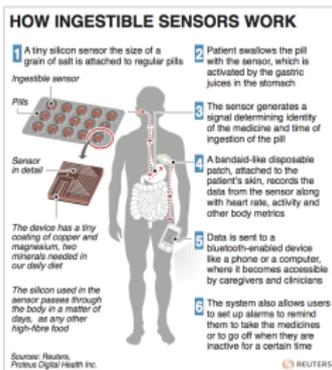
Matt Turck (@mattturck) and Shivon Zillis (@shivonz)

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



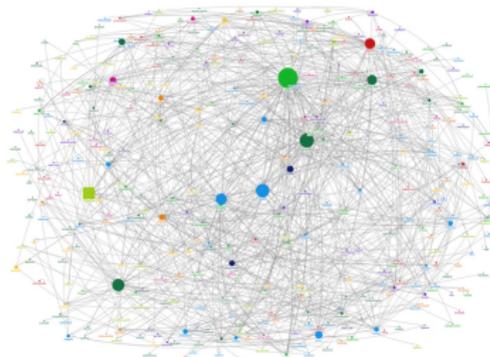
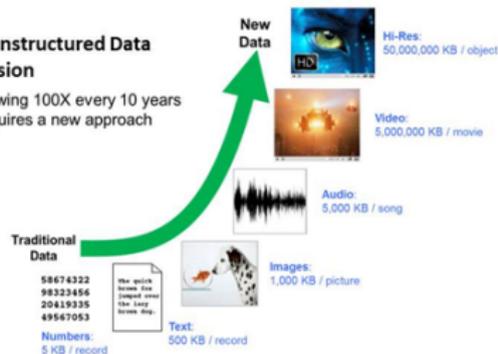
Some challenges

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

Unstructured Data

The Unstructured Data Explosion

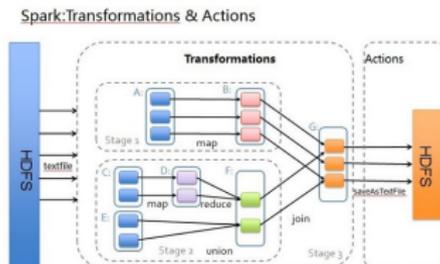
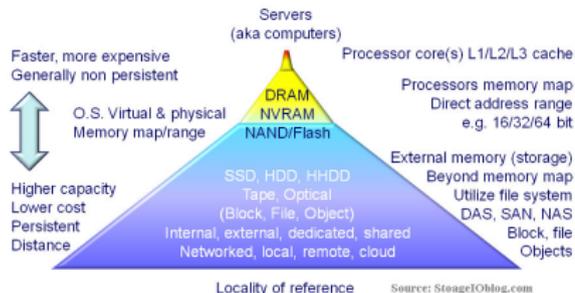
- Growing 100X every 10 years
- Requires a new approach



Some challenges

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

Huge Dataset



Some challenges

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

High Dimensional Data

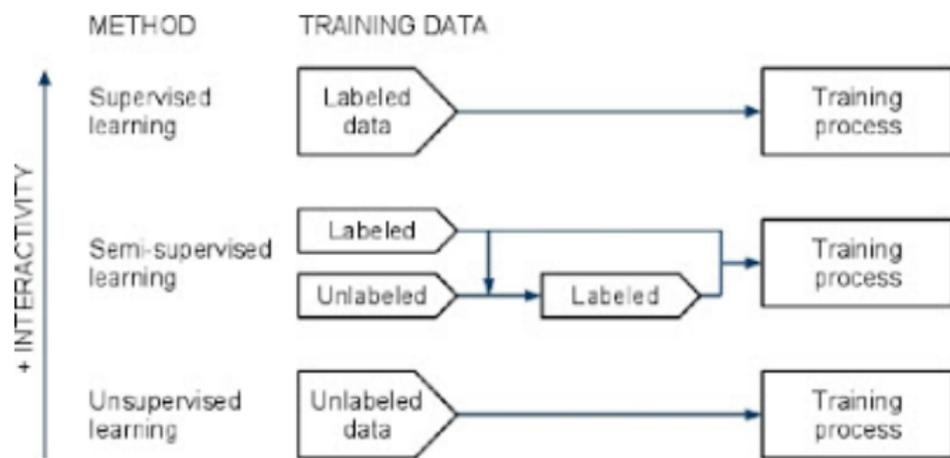
Main Paradigmatic Changes in Big Data Analytics Environment			
	Statistical Data Analysis <1985 (Pure Statistical Inference)	Business Intelligence 1985-2008 (Constrained Data Mining)	Big Analytics >2008 - up to now (Unconstrained Data Mining)
Data types	Homogeneous Structured Data (proprietary)	Homogeneous Structured & Homogeneous Unstructured Data, separately	Mix of Heterogeneous Unstructured & Structured Data (proprietary + open data)
Data storing	Line & column dimensions fixed Flat Files, Hierarchical DBs, & first Relational DBs	Column dimensions fixed SQL DBs, MySQL, DB2, ORACLE & OLAP Cubes	No dimensions fixed NoSQL DBs-Column oriented DBs, object oriented DBs, etc.
Volume Cost/volume	Exponential cost decrease		Exponential volume increase
Basic Analytical Principles	Hypotheses driven mode: Power use of sampling Techniques	Mix Hypotheses driven & Data driven: Dimensions Reduction & Populations Segmentations	Full Data driven mode: Power use of learning techniques, mainly unsupervised
Main Algorithmic approaches	Regression Analysis, Factorial Analysis, Statistical Inference in sampling, Linear general Models, Decision Trees, Etc.	Clustering (K-means, K Neighbours), Classification & Support Vector Machines, Multi layers Neural Nets, Scoring Techniques, Sequential Patterns, etc.	Deep adaptive learning techniques, Auto encoded neural Nets, Huge Graph Modularization, & Visual Analytics, Full unsupervised linear Clustering, etc.
New types of Business deliverables	Score Cards, Decisional Models based on sampling	Populations Profiling: CRM, Churn & Attrition Analysis, Loyalty & Propensity Programs, Cross selling	Near real time analysis for: individuality, adaptive online marketing & sales, machine learning for various purposes, automated maintenance programs

THALES

Some challenges

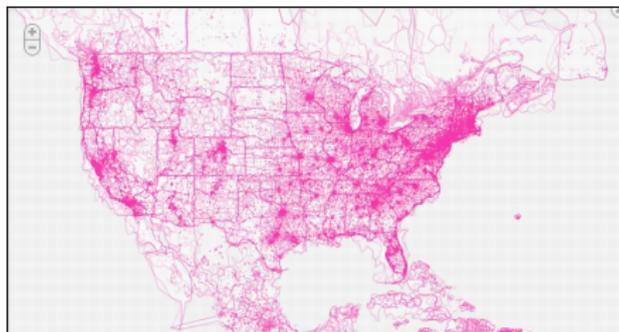
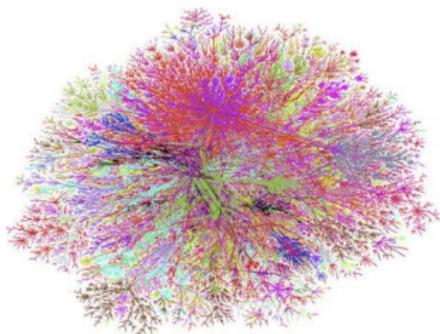
- How to describe (model) the data?
- How to reduce the data dimensionality?
- How to select/mix models?

Learning and Supervision



Some challenges

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



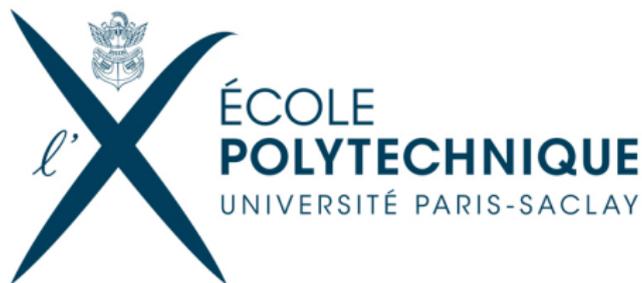
Some challenges

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

- **08h30-09h30. Breakfast** (rooms Puma B-C)
- **09h45-12h00. Big Data ?** (room Eagle B)
 - E. Le Pennec, Polytechnique, *Introduction*
 - M. Warren, LANL, *Big Data, or Astronomical Data*
 - S. Skillman, KIPAC / Stanford / SLAC, *Big Open Data: Hardware vs Software*
- **12h00-13h30. Lunch**
- **13h30-15h00. Computational Statistics Challenges**
 - J. Bruer, CALTEC, *Designing Statistical Estimators that Balance Sample Size, Risk, and Computational Cost*
 - F. Pourkamali-Anaraki, Univ. Colorado, *Efficient Algorithms for Analyzing large high-dimensional datasets via randomized sketching*
- **15h00-15h30. Coffee Break** (in front of Eagle B)
- **15h30-17h00. Health Applications**
 - M. Cuggia, Univ. Rennes, *Health BigData : context, use cases and challenges*
 - C. Lambert, Univ. New Mexico, *The local control method for treatment comparisons in large scale EHR/claims data*
- **18h30-... Welcome Drinks and Dinner**

- **07h30-08h30. Breakfast**
- **08h30-10h00. Implementation**
 - M. Turk, Univ. of Illinois, *Data Services: A Disrupted Industry*
 - N. Halko, SpotRight, *Near Real Time Analysis of Web Scale Social Data*
- **10h00-10h30. Coffee Break**
- **10h30-12h30. Sparse signal**
 - J. Bruna, Berkeley, *Signal recovery from scattering representations*
 - S. Sardy, Genève Univ. Switzerland, *Quantile universal threshold for efficient high-dimensional model selection*
 - D. Moody, LANL, *Adaptative sparse signal for discrimination of atellite-based radiofrequency recordings of lightning events*
- **12h30-14h00. Lunch**
- **14h00-15h30. Algorithms**
 - M. Challacombe, LANL, *Opportunities for generalized N-Body solvers in the materials genomics problem*
 - B. Kegl, X and CNRS Paris, *Learning to discover: signal/background separation and the Higgs boson challenge*
- **15h30-16h30. Coffee break and discussions**

Looking for interns for next spring?



- Polytechnique: french *grande école* (selective engineer school / university)
- Data Science initiative:
 - Research: joint applied math and computer science team (already 10 permanent researchers)
 - Teaching: master program, continuous training
- Lots of very good students with a strong mathematical background looking for internships!