New Algorithms for Complex Data

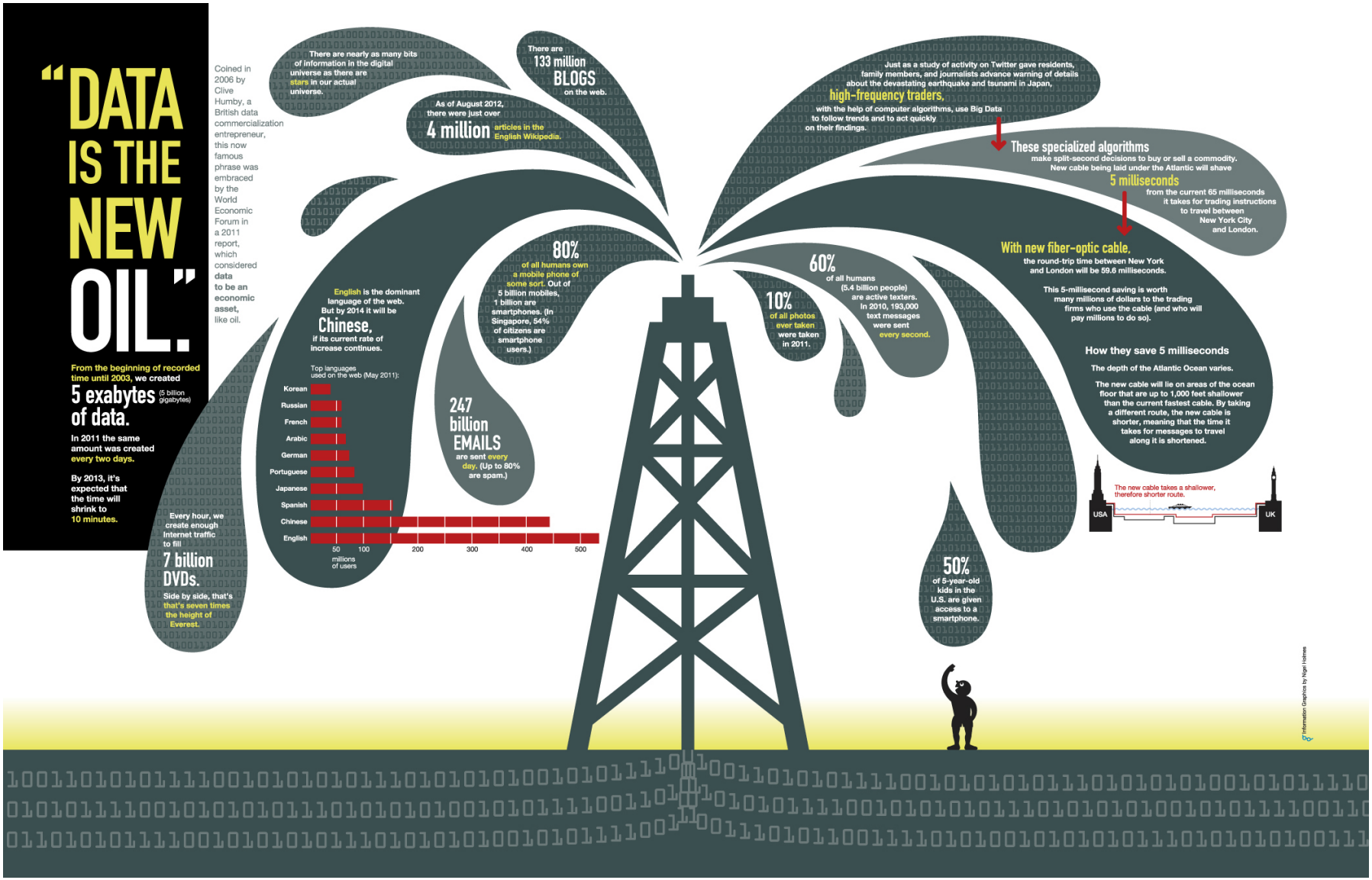
Eric Matzner-Løber, Nick Hengartner and Erwan Le Pennec
Rennes University (France), Los Alamos National Laboratory (USA),
Polytechnique (France)

19/20 March 2015

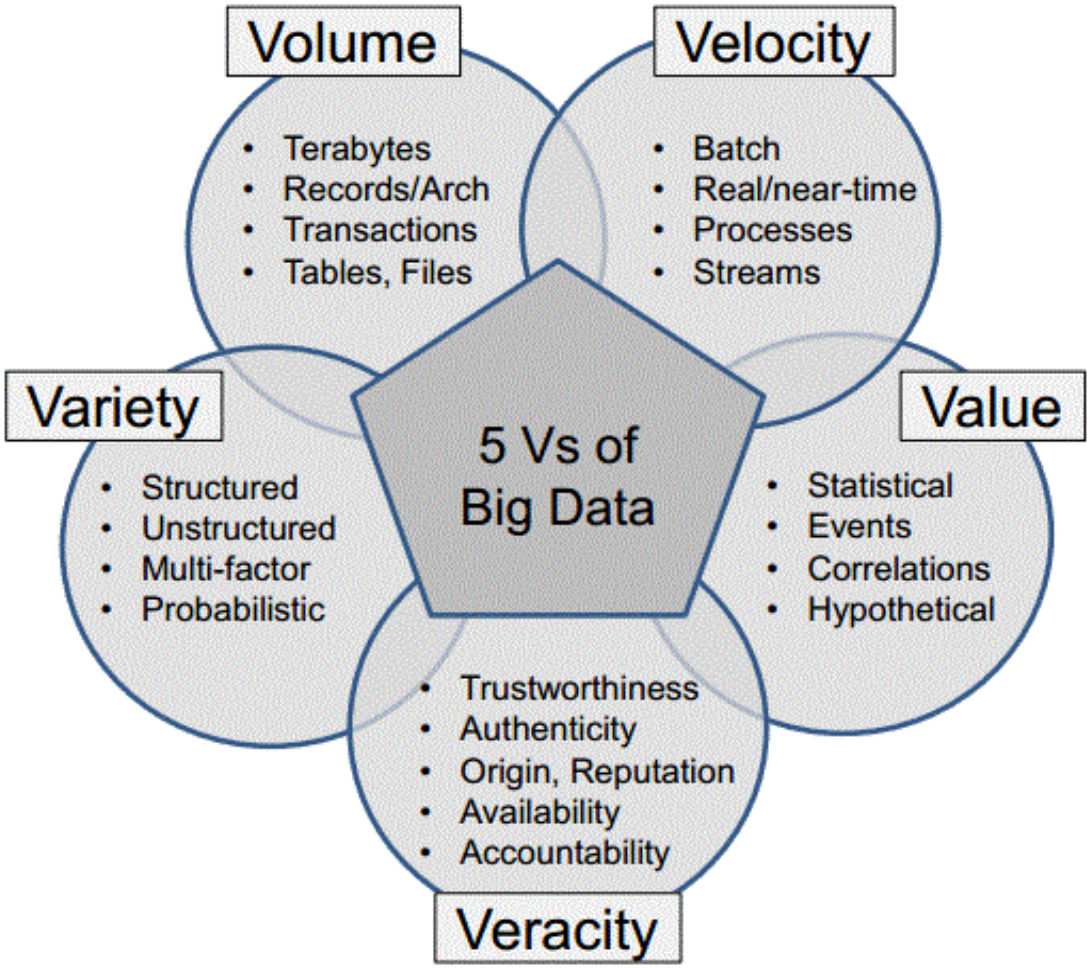
New Algorithms for Complex Data

- 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!

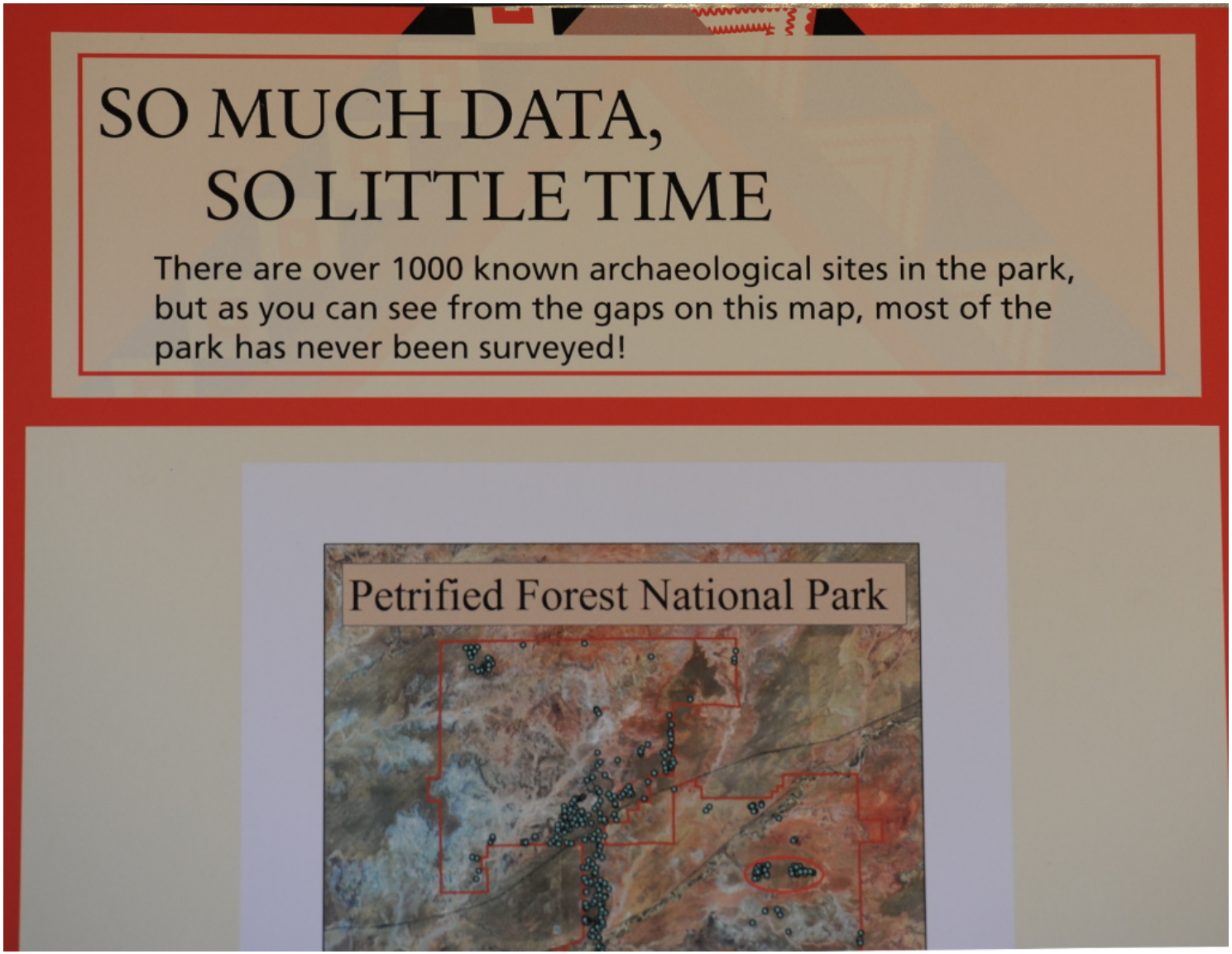


The 5 Vs of Big Data



Lots of Words!

Petrified Forest!



Doing Data Science

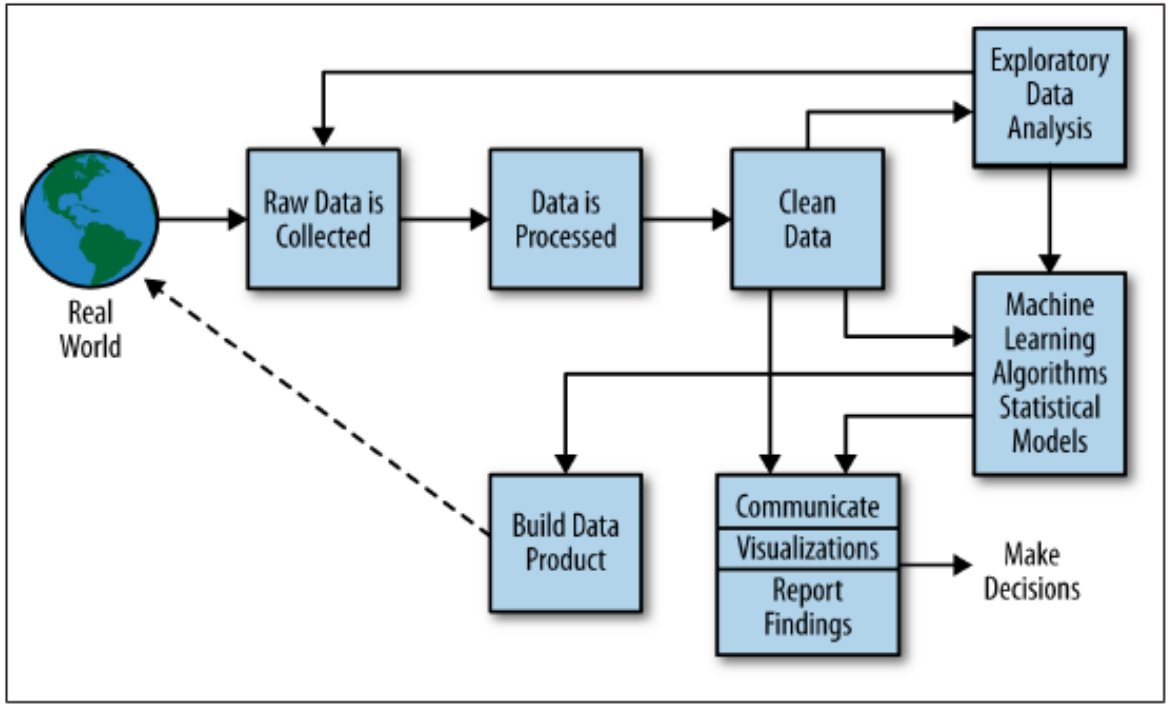


Figure 2-2. The data science process

Doing Data Science: Straight talk from the frontline

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

A new Context

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


```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()
}
```

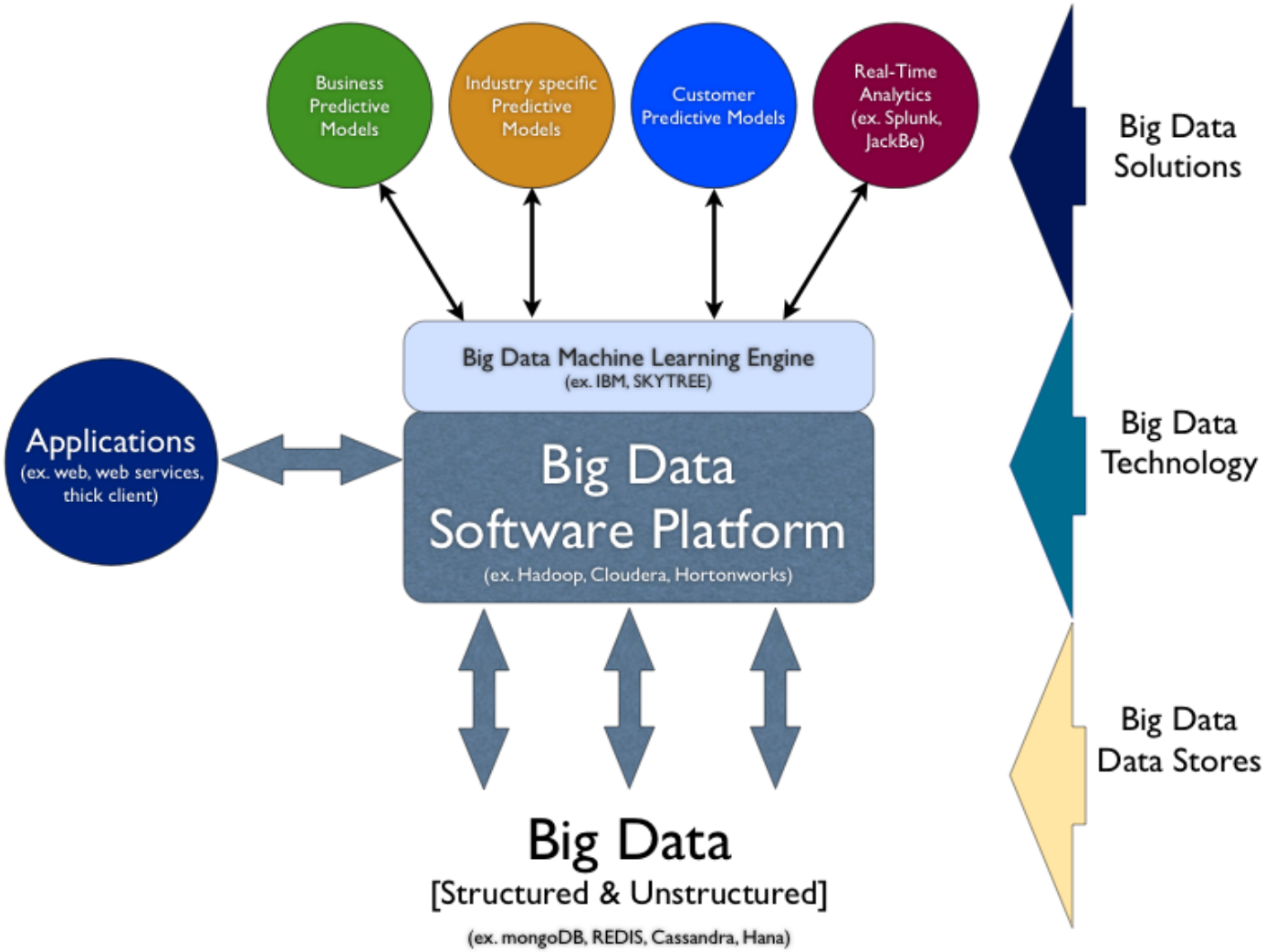
Big Data is (quite) Easy

Example of *off the shelves* solution

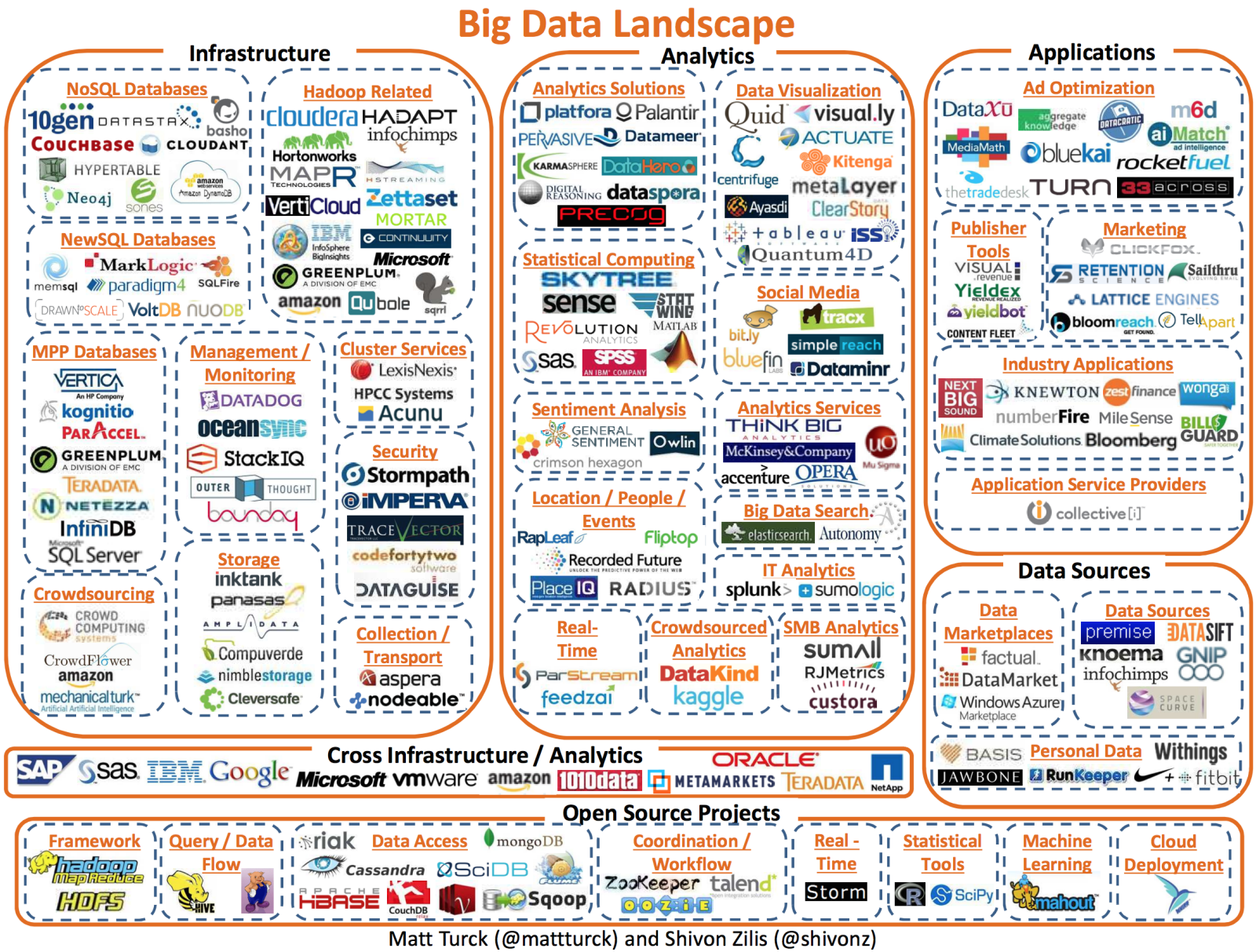

```
export AWS_ACCESS_KEY_ID=<your-access-keyid>
export AWS_SECRET_ACCESS_KEY=<your-access-key-secret>
cellule/spark/ec2/sparl-ec2 -i cellule.pem -k cellule -s <number of machines> launch <cluster-name>
ssh -i cellule.pem root@<your-cluster-master-dns>
spark-ec2/copy-dir ephemeral-hdfs/conf
ephemeral-hdfs/bin/hadoop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

A Complex Ecosystem!



A Complex Ecosystem!



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

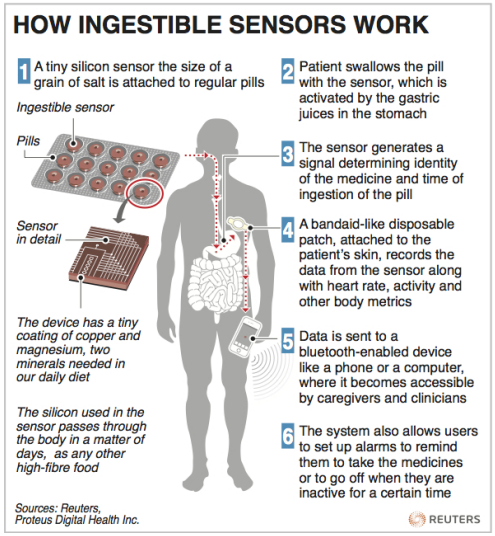
New Interdisciplinary Challenges

- 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

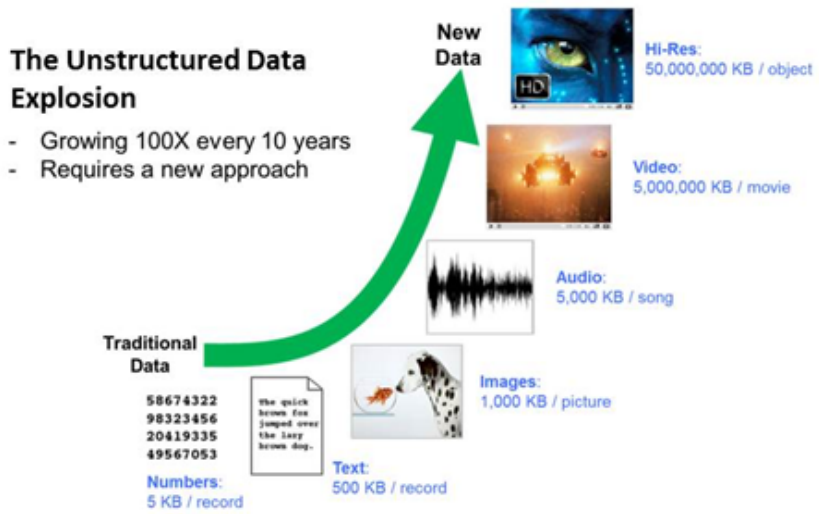
Data acquisition

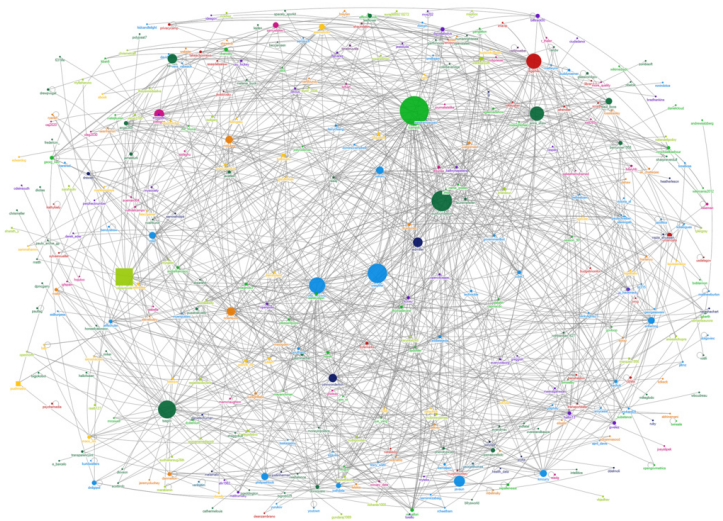


Some challenges

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

Unstructured Data

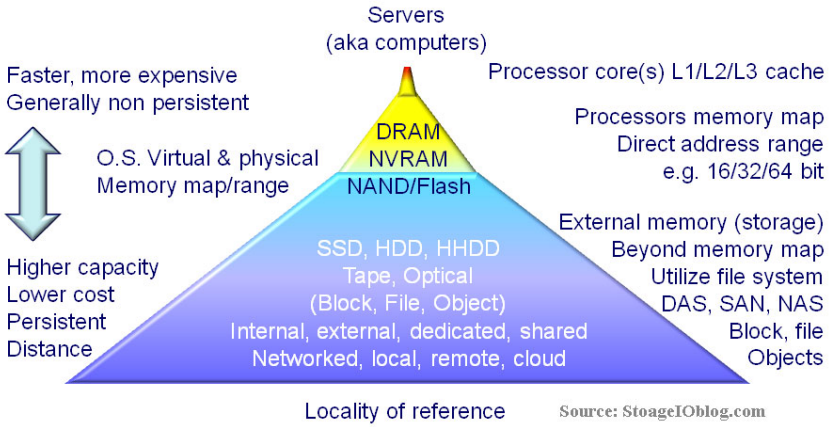


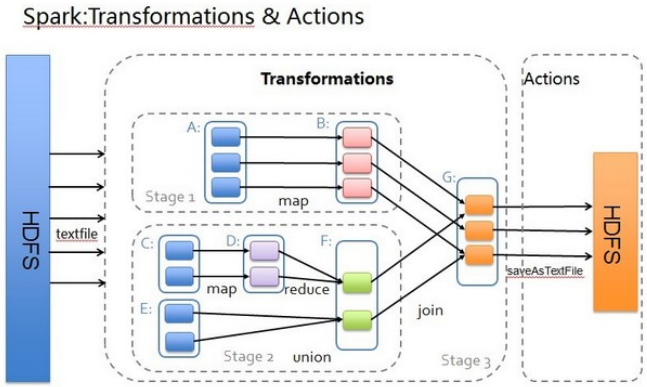


Some challenges

- 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?

Huge Dataset

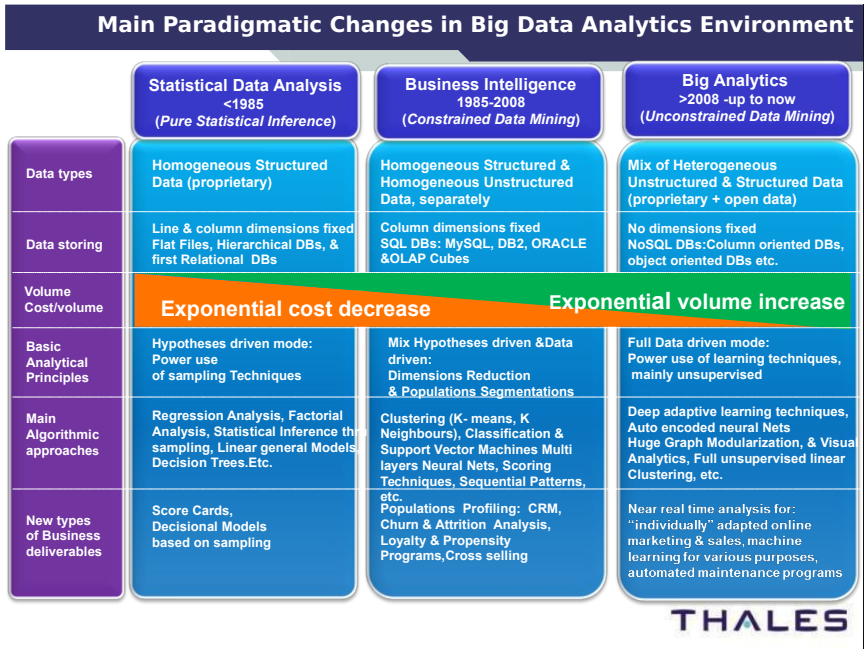




Some challenges

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

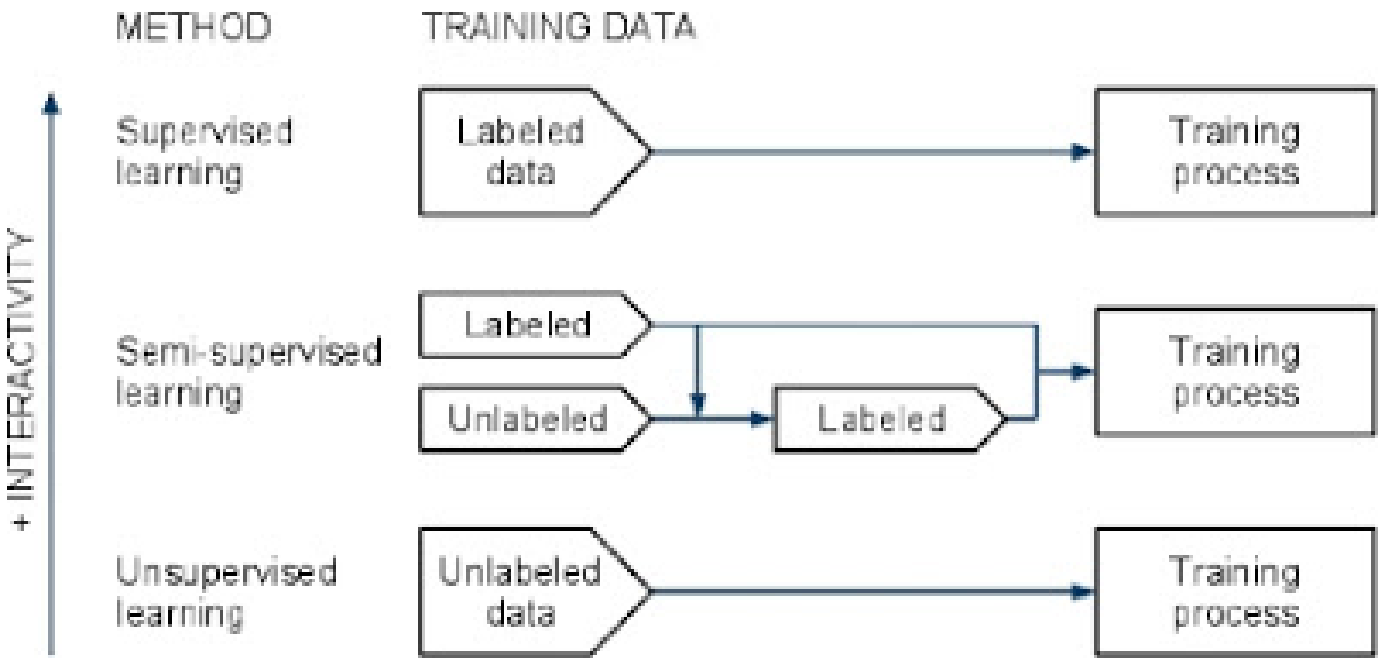
High Dimensional Data



Some challenges

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

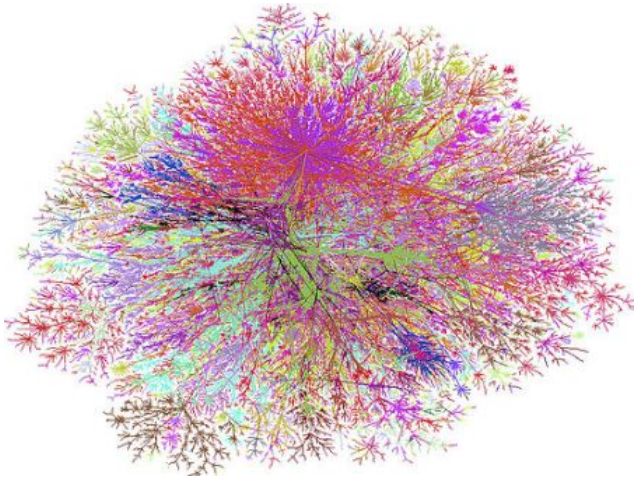
Learning and Supervision

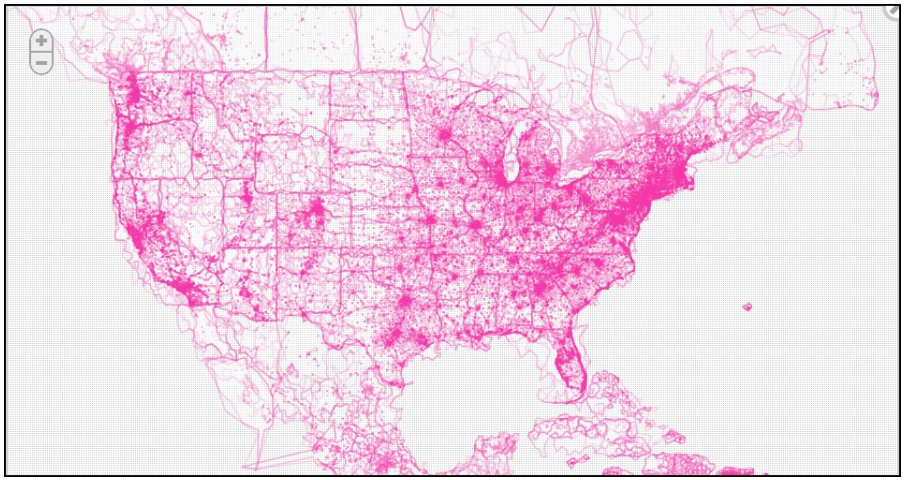


Some challenges

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

Visualization





Some challenges

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

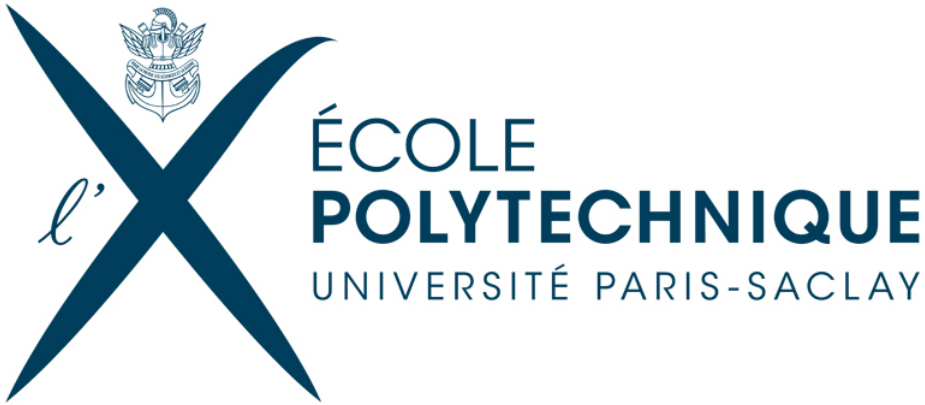
Thursday March 19th

- **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**

Friday March 20th

- **07h30-08h30. Breakfast**
- **08h30-10h00. Implementation**
  - M. Turk, Univ. of Illinoy, *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!