DS PORTUGAL

Data Science NOW!

(and other not so deep thoughts about AI…)

Luis Moreira-Matias

Luis.Moreira.Matias@Gmail.com

Porto, Portugal .::. March, 2018

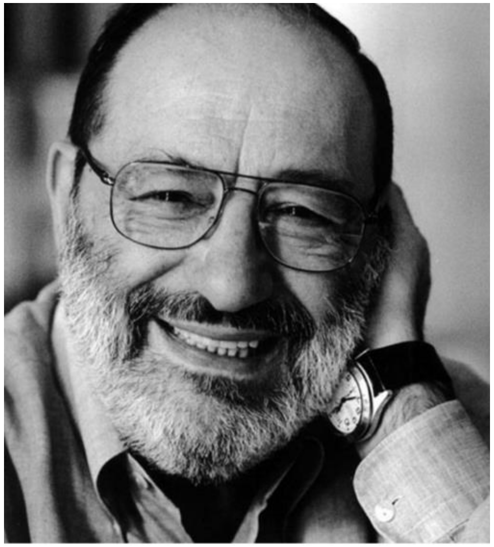
Outline

- DS – What is it and what serves for?
- DS??? How about Big Deep Data AI Learning??
- An example of a DS pipeline
    - Data Wrangling
    - Exploratory Data Analysis
    - Performance Evaluation and Interpretability
    - Deployment and ROI
- Turning the page towards Automated Analytics
- Case Studies
- Final Remarks

Computers do not solve everything...

“The computer is not an intelligent machine that helps stupid people.

It is a stupid machine that only works in the hands of smart people.”



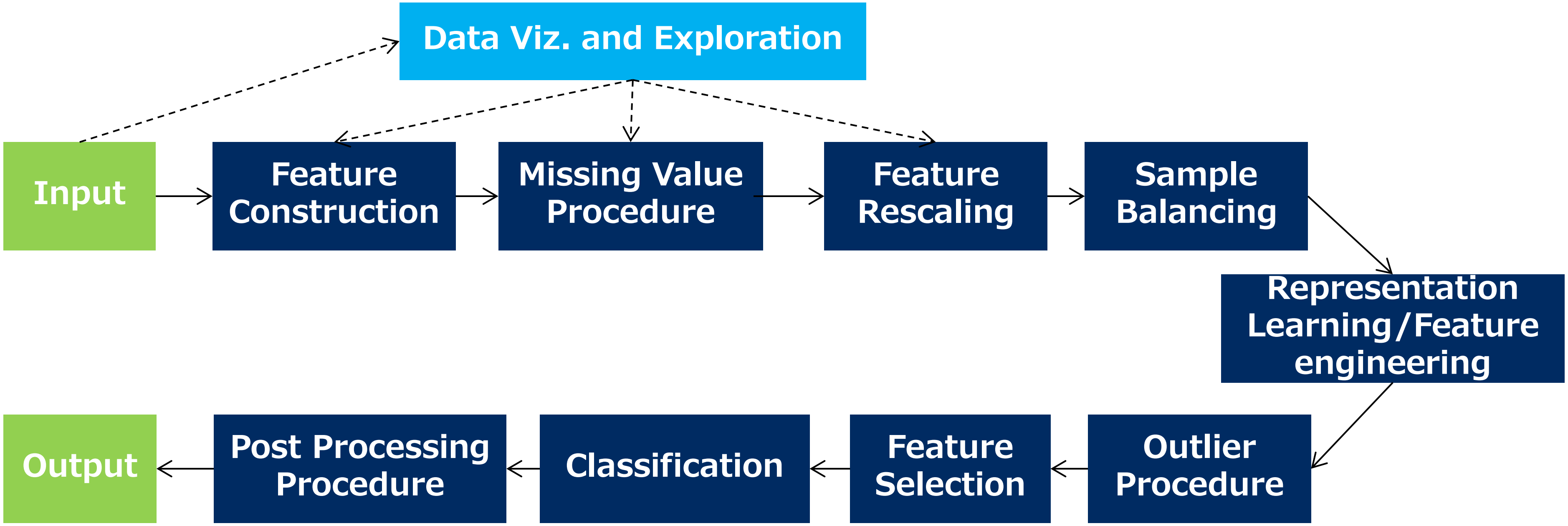
Umberto Eco

(1932-2016)
Italian Philosopher

What is DS about?

- **Data Science denotes the process of transforming raw data into knowledge in a (almost…) automated fashion.**
- **Related Disciplines:**
  - **Statistical Learning (incl. Bayesian Approaches)**
  - **Data Mining and Signal Processing**
  - **Mathematical Optimization**
  - **…among other Computer Science disciplines**

Typical Data Analytics Pipeline



Data Science NOW!

Where Does Knowledge Comes From?

Evolution

Experience

Culture

Computers

Tribute to Prof. Pedro Domingos (U. Washington)

Computers are everywhere...so is DS!

What is NOT DS about?

- **"I work with data…so I am a *Data Scientist*!"**

✖

- **Business Planning leveraging on Digital Data = Data Strategy**

- **Aggregate and Manually analyze (small) Data for Reporting and Decision Support purposes = Business Intelligence**

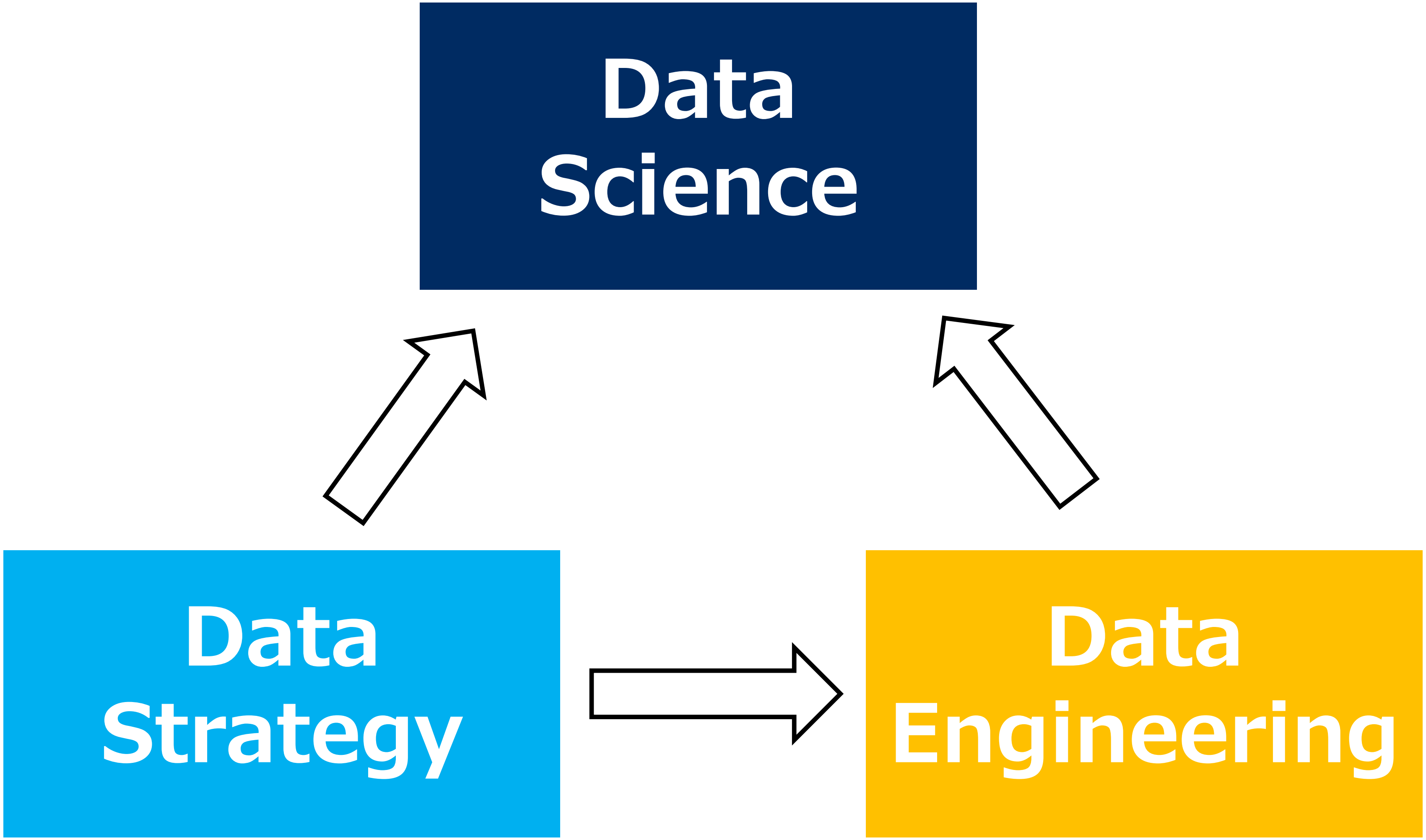
- **Processes of extracting, querying and displaying (possible large amounts of) data = Data Engineering**

- **Science of Data vs. Science with Data**

"Data Science: (not) the preferred nomenclature" –

Peter Flach (U. Bristol),
Editor-in-Chief of Machine Learning Journal
August, 2017

The III Pilars of a Data-Driven Company in 2018



Data Science

Data Strategy

Data Engineering

Why cant we just mix everything?

THIS IS BOB, OUR NEW CITIZEN DATA SCIENTIST. HE PREVIOUSLY WORKED AS A CITIZEN DENTIST AND A CITIZEN PILOT.
KDNUGGETS.COM

Why cant we just mix everything?

Job Specs of a **Lead Data Scientist** (based on a real world case)…

REQUIREMENTS

- *Capable of leadership (influence management) and pragmatism;*
- *5+ years' experience* ***prototyping*** *classification and regression models using machine software such as* ***scikit-learn, R, MATLAB, Octave, Weka, Mahout****, etc;*
- *Experience with* ***large-scale log processing*** *or big data including but not limited to* ***Elastic MapReduce, Hadoop Streaming, Pig, Hive, Spark****, etc;*
- *Fluency in a range of scripting languages, operating systems, and software platforms including but not limited to* ***Linux, Redis, Java, MySQL, Python, Pandas, AWS****;*
- *Experience in* ***Data Architecture, Data modeling, and Database design****;*
- *Experience working with relational and non-relational databases to retrieve structured, unstructured, and semi-structured data sets;*
- *Knowledge of business domains (Travel and Hospitality, Finance, Retail, etc.);*
- *Track record (4+ projects) of working directly with the customer (i.e. regularly facing customer directly) both remotely and on-site;*
- *Presale activities with exposure to customers on billable accounts;*
- ***Demonstrated experience with full life cycle of software development processes like RUP or "Waterfall", Agile (SCRUM, XP, etc.). Experience advocating and establishing an appropriate implementation methodology to successfully develop and deploy the solution;***
- *Demonstrated experience in solution cost estimation (including tools, tasks, complexity, labor and time) at coarse grain and fine grain levels, with supporting material evidence;*
- *Demonstrated experience in validating the overall solution from the perspective of performance, scalability, security and capacity.*
- ***Ability to FLY***☺

So...why are companies doing it?

Scarcity of Resources

- Adoption still low: only for **12% of use cases** defined by 3000 companies [1]
- More than **40%** of organizations using DS say "**the lack of adequate skills**" is a challenge [2]
- Expected market growth 2017-2022 @ 40% CAGR to $8.8B for ML alone [3]
- "Citizen data scientists" market will grow **5x faster** than "Professional data scientists" [4]

If everyone does it...why cant we too? *Just add up ".deep" or ".AI"...*

- **ML-as-a-service**: AWS's machine learning, IBM's Blue Mix, BigML, Theano, ...
- **Sales points**: Deep Learning sounds sexy, looks sexy and requires tons of resources...justifying sales volumes that would be impossible otherwise;

Reality...

"We have lots of data but we are **not able to use ML** or hire experts to do it so."

(Rome Transit Agency, Italy, April 2017)

[1] AI The Next Digital Frontier?, McKinsey Global Institute, July 2017
[2] How to do Machine Learning Without Hiring Data Scientists, Gartner, February 2017
[3] Machine Learning Market by Vertical, M&M, September 2017
[4] Predicts 2017: Analytics Strategy and Technology, Gartner, November 2016
[5] Data Scientists Automated and Unemployed by 2025, Data Science central, July 2017

DS Toolset

julia

Scala

python™

amazon web services™

TensorFlow

R

tableau

Apache mahout

hadoop

Data Science

Java

The C++ Programming Language

PostgreSQL

WEKA
The University
of Waikato

MySQL®

Azure

MATLAB
The Language of Technical Computing

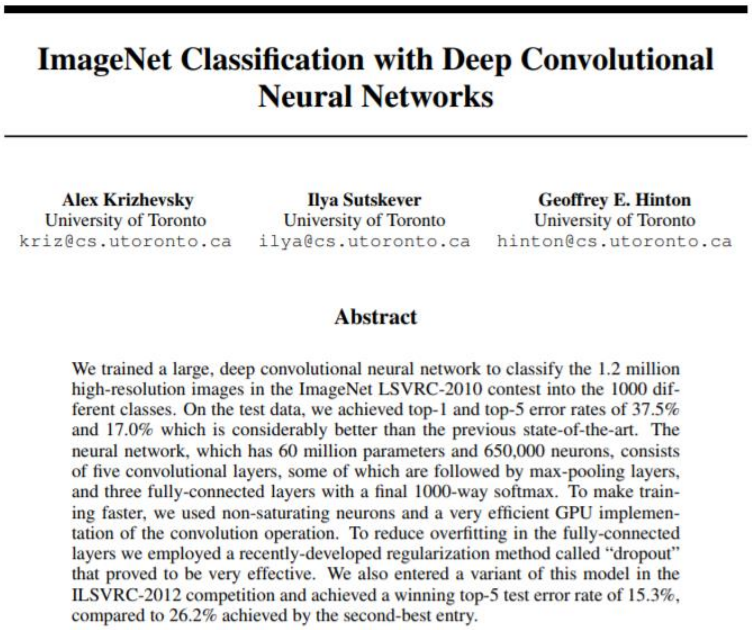
Flink

Wait a sec...was not Deep Learning the ultimate solution?

The AI/Deep Learning hype – Why and How?

- Early 201x - AI Challenges such as **Perception** and **Knowledge Representation** are still in *dark age:*
  - Approaches to Computer Vision are highly domain specific and requiring high human expertise;
  - Similar phenomenon happen with fields such as NLP, Ontologies and Speech;

- Then, at NIPS 2012...



ImageNet Classification with Deep Convolutional Neural Networks

Alex Krizhevsky University of Toronto kriz@cs.utoronto.ca	Ilya Sutskever University of Toronto ilya@cs.utoronto.ca	Geoffrey E. Hinton University of Toronto hinton@cs.utoronto.ca

Abstract

We trained a large, deep convolutional neural network to classify the 1.2 million high-resolution images in the ImageNet LSVRC-2010 contest into the 1000 different classes. On the test data, we achieved top-1 and top-5 error rates of 37.5% and 17.0% which is considerably better than the previous state-of-the-art. The neural network, which has 60 million parameters and 650,000 neurons, consists of five convolutional layers, some of which are followed by max-pooling layers, and three fully-connected layers with a final 1000-way softmax. To make training faster, we used non-saturating neurons and a very efficient GPU implementation of the convolution operation. To reduce overfitting in the fully-connected layers we employed a recently-developed regularization method called "dropout" that proved to be very effective. We also entered a variant of this model in the ILSVRC-2012 competition and achieved a winning top-5 test error rate of 15.3%, compared to 26.2% achieved by the second-best entry.

- Large amounts of data available;
- "New" Connectionist architecture feasible to train in reasonable time with GPU-enabled hardware – great sales point!!!
- It does *not* overfit ☺...

NVIDIA

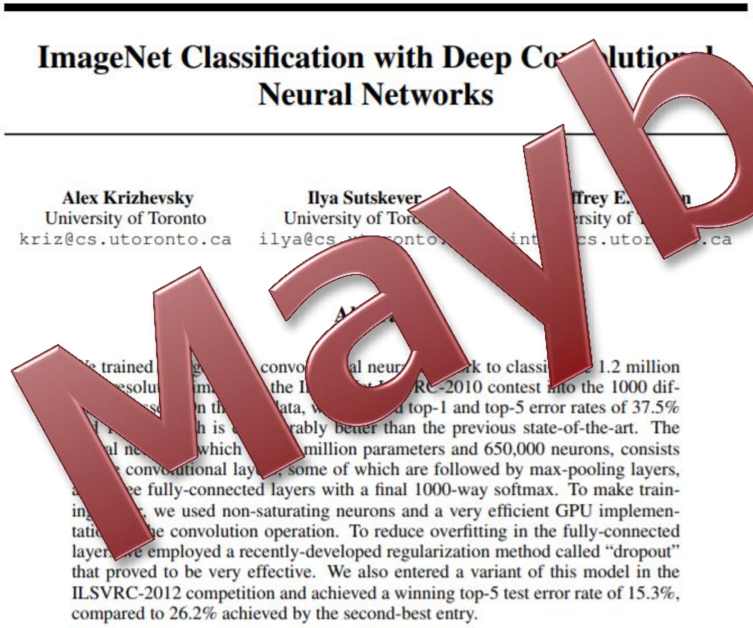
- The end of bias-variance tradeoff...just go ***deep**er* or ...**add more data**!

The AI/Deep Learning hype – Why and How?

Early 201x - AI Challenges such as **Perception** and **Knowledge Representation** are still in *dark age:*

- Approaches to Computer Vision are highly domain specific and requiring high human expertise;
- Similar phenomenon happen with fields such as NLP, Ontologies and Speech;

Then, at NIPS 2012…



ImageNet Classification with Deep Convolutional Neural Networks

Alex Krizhevsky
University of Toronto
kriz@cs.utoronto.ca

Ilya Sutskever
University of Toronto
ilya@cs.utoronto.ca

Geoffrey E. Hinton
University of Toronto
hinton@cs.utoronto.ca

Abstract

We trained a large, deep convolutional neural network to classify the 1.2 million high-resolution images in the ImageNet LSVRC-2010 contest into the 1000 different classes. On the test data, we achieved top-1 and top-5 error rates of 37.5% and 17.0% which is considerably better than the previous state-of-the-art. The neural network, which has 60 million parameters and 650,000 neurons, consists of five convolutional layers, some of which are followed by max-pooling layers, and three fully-connected layers with a final 1000-way softmax. To make training faster, we used non-saturating neurons and a very efficient GPU implementation of the convolution operation. To reduce overfitting in the fully-connected layers we employed a recently-developed regularization method called "dropout" that proved to be very effective. We also entered a variant of this model in the ILSVRC-2012 competition and achieved a winning top-5 test error rate of 15.3%, compared to 26.2% achieved by the second-best entry.

Large amounts of data available;

"New" Connectionist architecture feasible to train in reasonable time with GPU-enabled hardware – great sales point!!!

It does *not* overfit ☺…

The end of bias-variance tradeoff…just go ***deep**er* or …**add more data**!

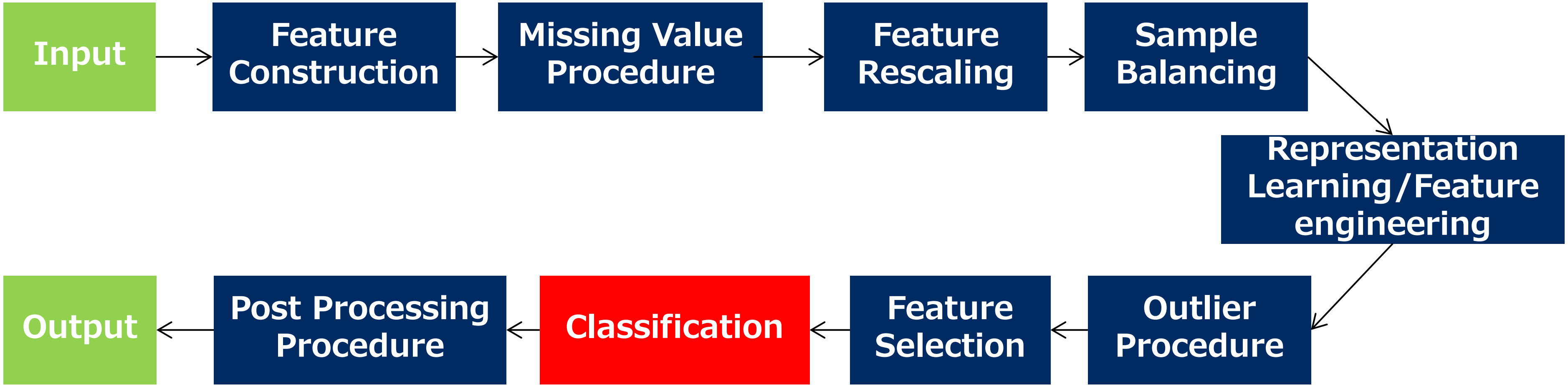
Maybe NOT!

What is Deep Learning really about?

- Deep Learning is a first step to perform End-to-End Learning...wait a sec, what is **End-to-End Learning**?

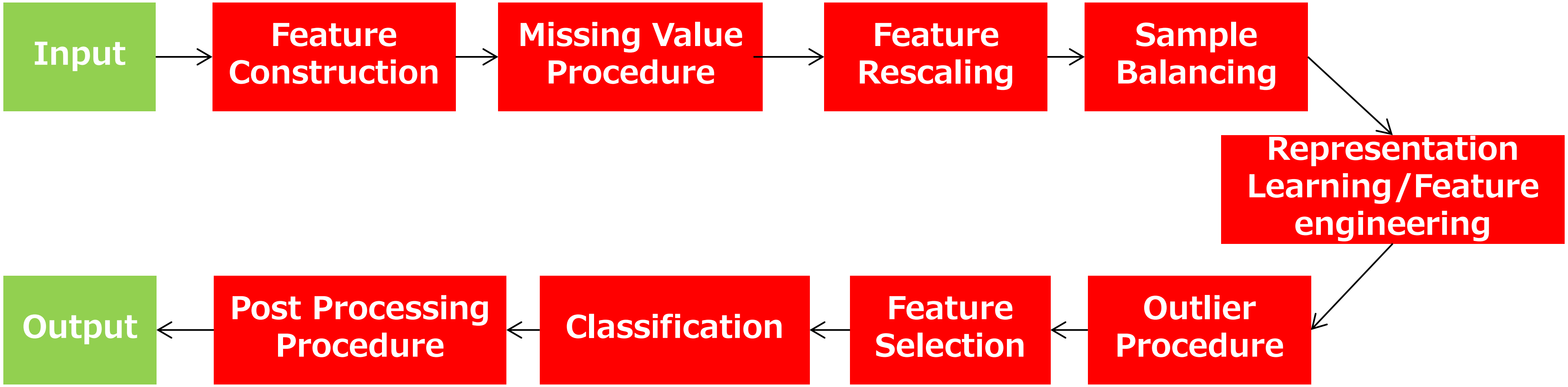
What is Deep Learning really about?

- Deep Learning is a first step towards End-to-End Learning...wait a sec, what is **End-to-End Learning**?



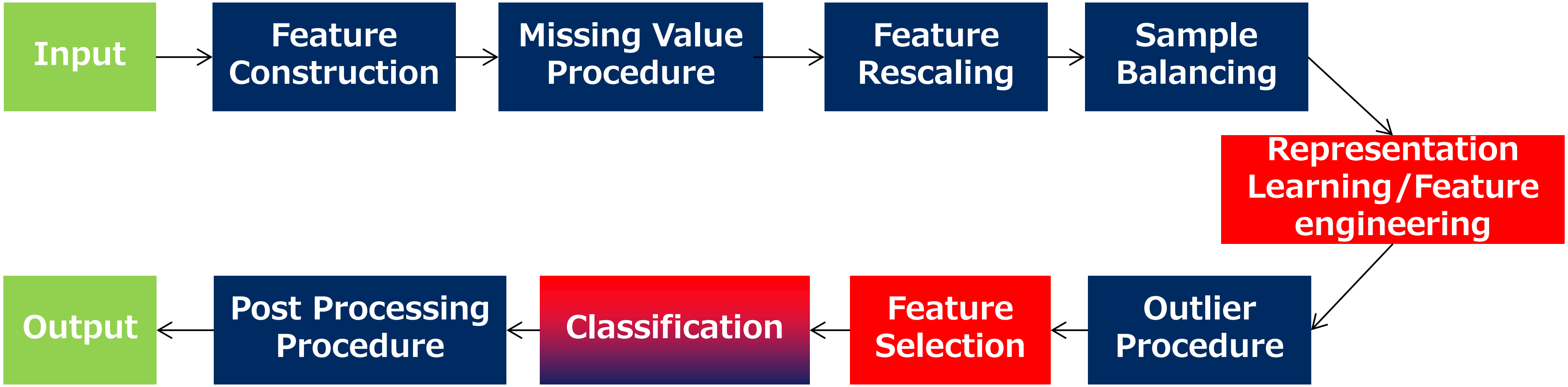
What is Deep Learning really about?

- Deep Learning is a first step towards End-to-End Learning...wait a sec, what is **End-to-End Learning**?

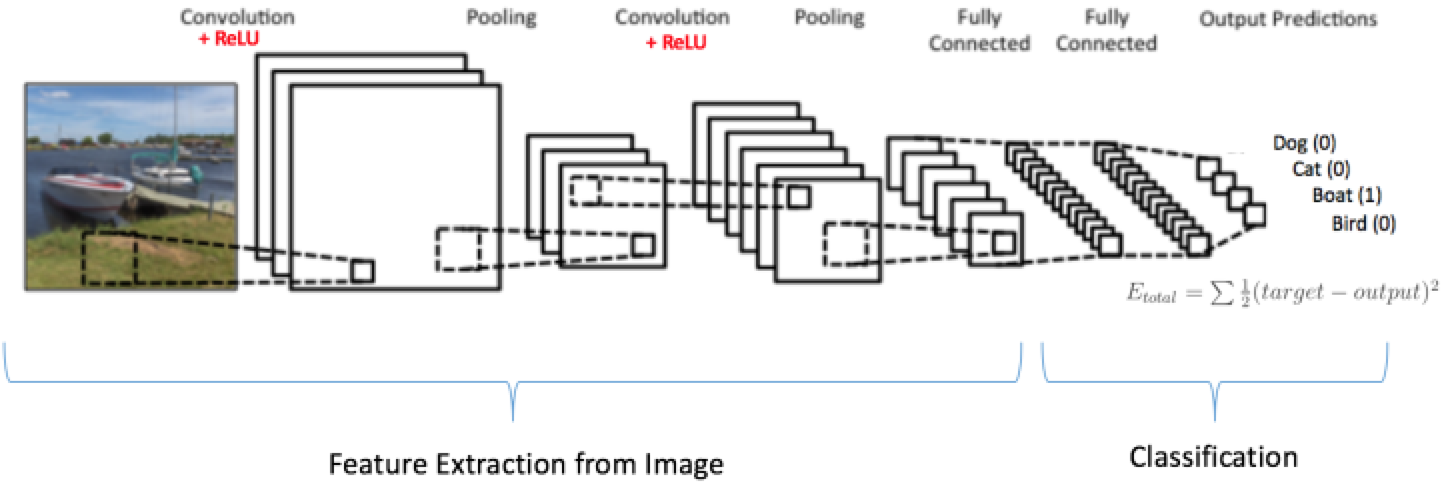


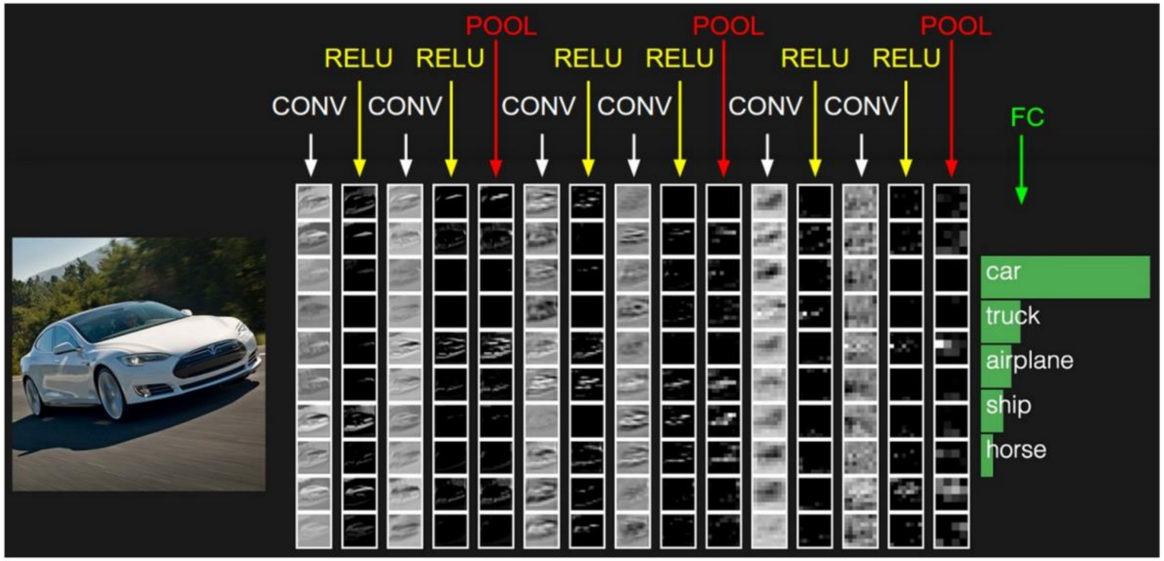
What is Deep Learning really about?

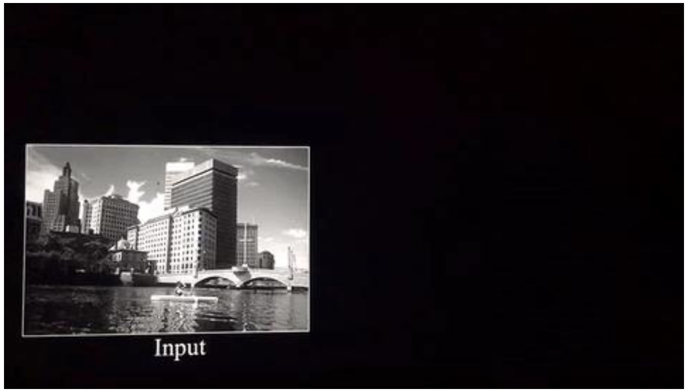
- Deep Learning is a first step towards End-to-End Learning;
- It is about to jointly learn an euclidean representation (e.g. vectorial) of feature space and a dependency function of the target with respect to this new representation – **in a single optimization problem**!
- Most prominent connectionist approaches are CNNs (SoA performance in Vision) and LSTM (arguably SoA performance in NLP).



What is Deep Learning really about? – Example with CNN







REFERENCES:

- CS231n Convolutional Neural Networks for Visual Recognition, Stanford
- Clarifai / Technology
- Deep Learning Methods for Vision, CVPR 2012 Tutorial

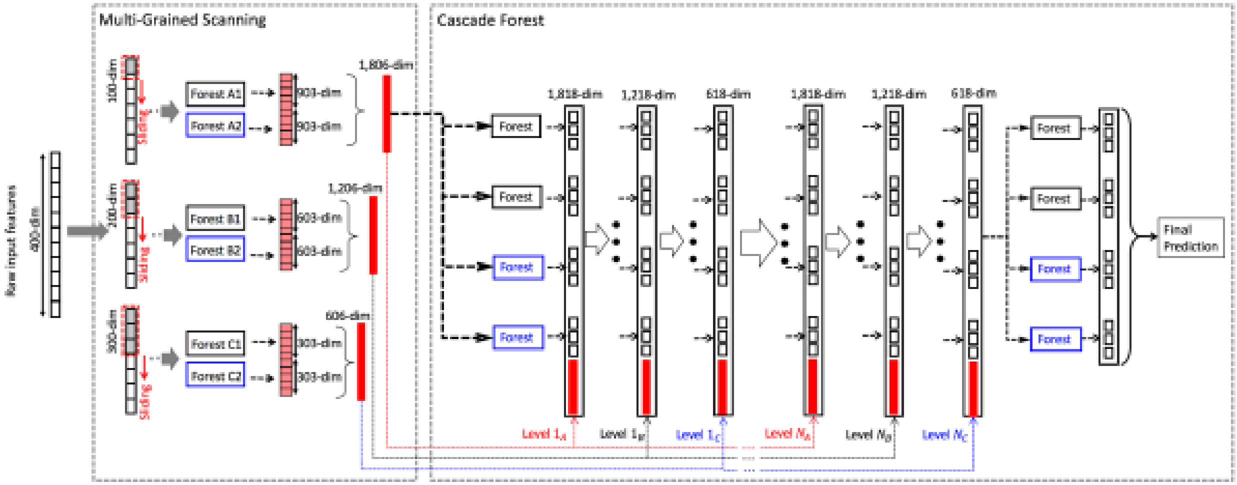
Pitfalls of Connectionist Approaches to Deep Learning

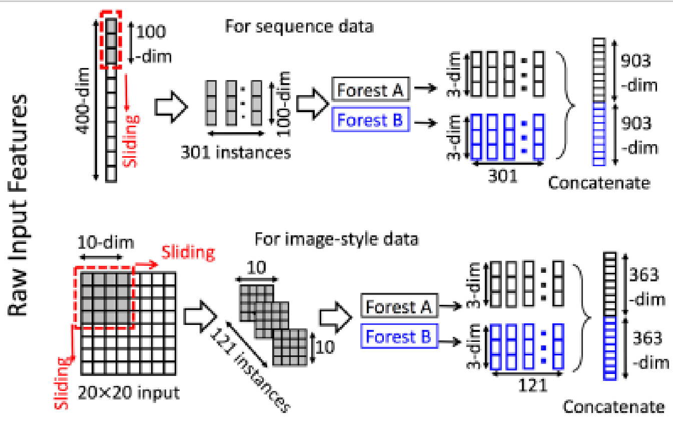
- They rely on **extremely large** datasets of **labeled** data that for many problems may never be physically or economically available.
- They take **long training times** on *expensive* GPUs
- The modelling-to-production phase is exacerbated because there are also a **very large number of hyperparameters (including network topology)** which are not altogether well understood and require multiple attempts to get right (yes, well...we still need to handle bias-variance tradeoff after all);
- It is still true that some of these configurations **fail to train** at all losing all the value of the time and money invested => like all the neural nets trained with backpropagation, the underlying optimization problem lacks a **theoretical guarantee of convergence**;

Don’t they know that? Yes, but ... ☺

Ref: David Perkins, The Economist “The world’s most valuable resource is no longer oil...” Accessed March 2018

A note on alternative Deep Learning architectures

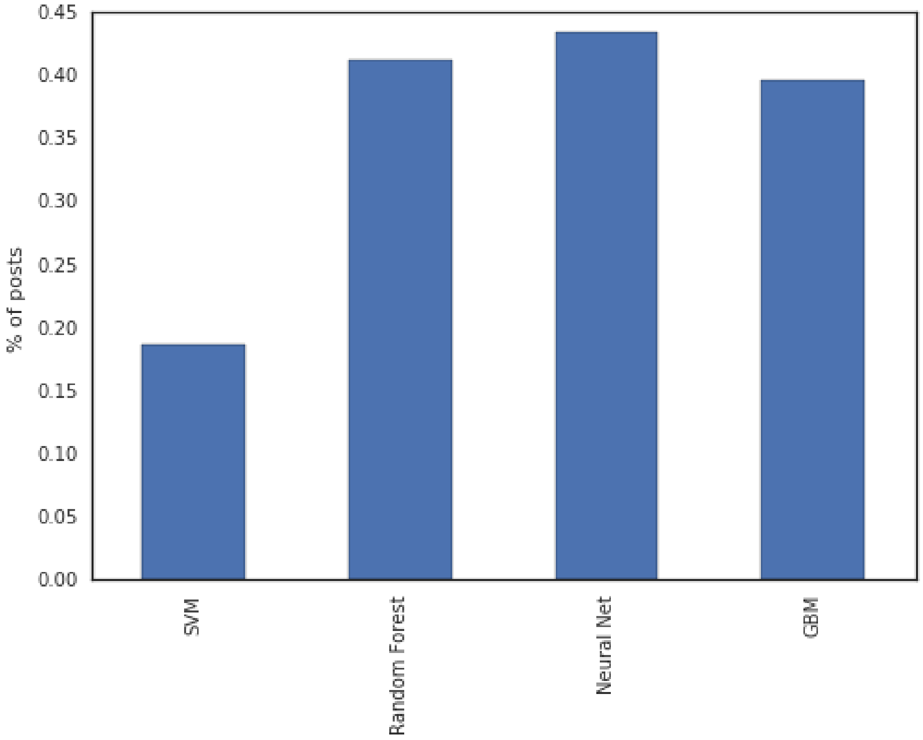


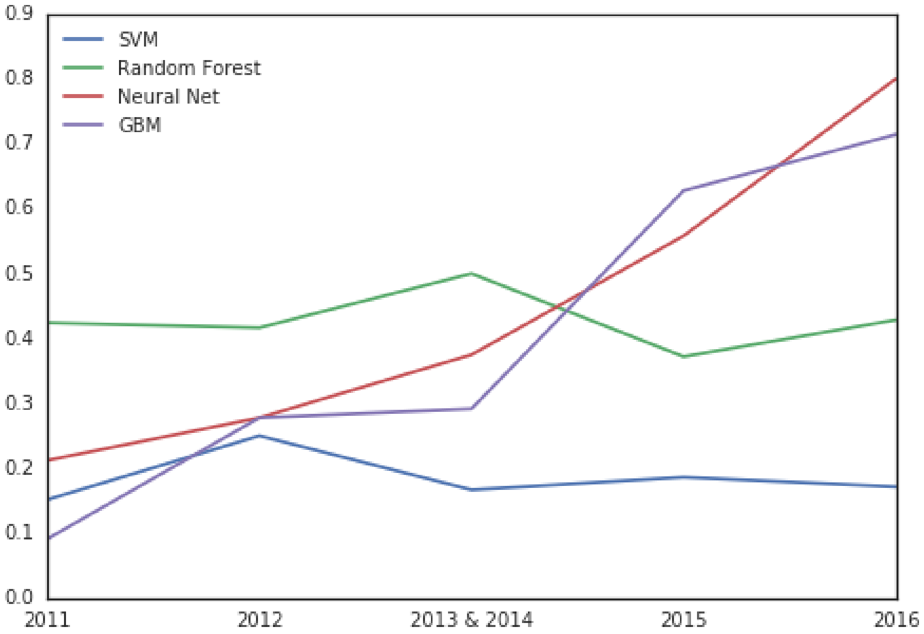


- Zhi-Hua Zhou recently proposed **gcForests**, a deep learning approach based on decision trees;
- It operates by combining random forests with fully randomized trees;
- It requires less data, less hardware and can be easily fully parallelized;
- It has lower sensitiveness to hyperparameter settings;
- It does not overcome CNNs/LSTMs…but it presents a top/competitive performance in some image/text datasets for classification;

Ref: Zhi-Hua Zhou *et al* - Deep Forest: Towards An Alternative to Deep Neural Networks (2017, arxiv)

Fiction vs. Reality





- The right ML algorithm **still depends** on the task at hand;
- More than 50% of winning solutions of Kaggle Competitions in the last 5 years **did not included any neural network**;
- There is no free lunch – deep or not ☺

Ref: [What algorithms are most successful in Kaggle? (Accessed 03/2018)]

1-on-1 with a DS project

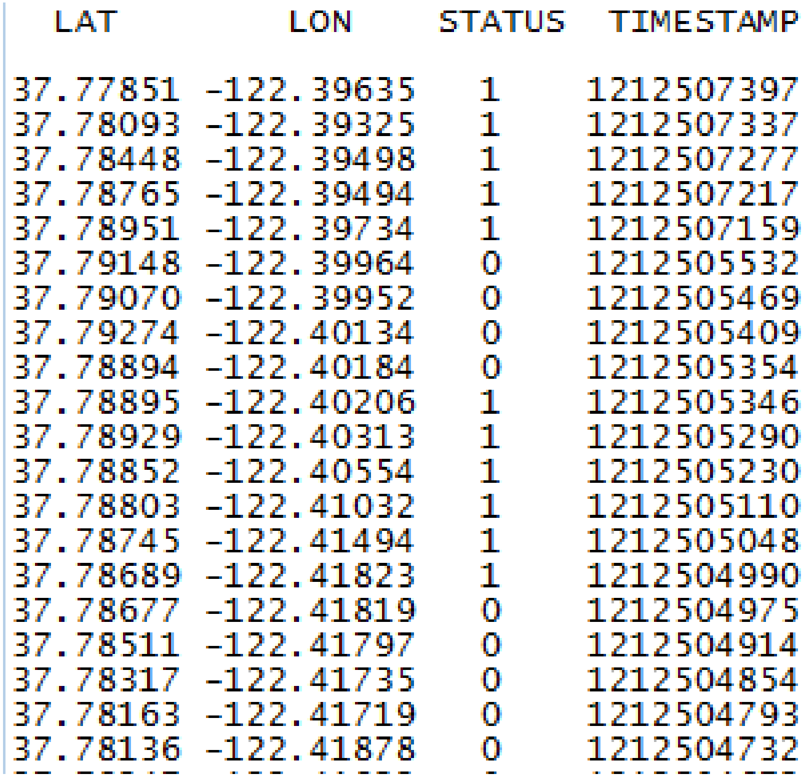
Case Study: Telematics Data from a Taxi Fleet

Business Question (from client): *How can we improve our business leveraging on this operational data?*

DS Project Stages:

1. Data type recognition;
2. Visualization and Data Wrangling;
3. Query and Problem Formulation;
4. Feature Engineering/Selection;
5. Algorithm Selection and Hyperparameter Tuning;
6. Post-Processing;

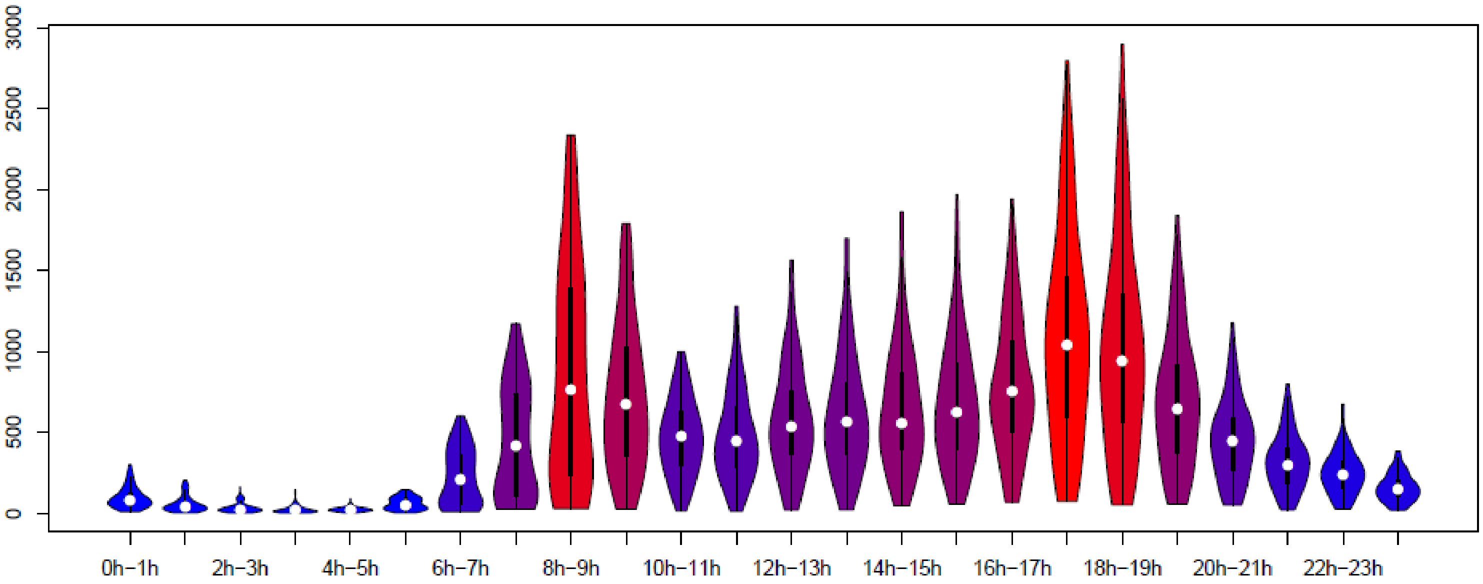
1-on-1: Data type recognition



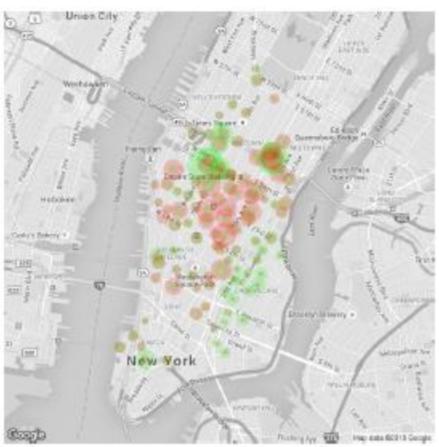
LAT	LON	STATUS	TIMESTAMP
37.77851	-122.39635	1	1212507397
37.78093	-122.39325	1	1212507337
37.78448	-122.39498	1	1212507277
37.78765	-122.39494	1	1212507217
37.78951	-122.39734	1	1212507159
37.79148	-122.39964	0	1212505532
37.79070	-122.39952	0	1212505469
37.79274	-122.40134	0	1212505409
37.78894	-122.40184	0	1212505354
37.78895	-122.40206	1	1212505346
37.78929	-122.40313	1	1212505290
37.78852	-122.40554	1	1212505230
37.78803	-122.41032	1	1212505110
37.78745	-122.41494	1	1212505048
37.78689	-122.41823	1	1212504990
37.78677	-122.41819	0	1212504975
37.78511	-122.41797	0	1212504914
37.78317	-122.41735	0	1212504854
37.78163	-122.41719	0	1212504793
37.78136	-122.41878	0	1212504732

- Each Vehicle broadcasts one message (a line) like this each 15 seconds;
- Naturally, we are facing a sequential/time series data source;
- Even if in this case this recognition is trivial, there are projects where this choice is crucial. Good example? treat trajectories as images to explore directionality in CNN;

1-on-1: Visualization, Data Wrangling and Query



- Spatial and Temporal Seasonal patterns;



Union City
New York

1-on-1: Visualization, Data Wrangling and Query

Possible Query: How many passengers will demand a taxi in each city area?

Resulting problems:

- Temporal and Spatial Segmentation – how to split the city in regions and which time horizons to use?
- Missing data – how about the passengers that didn't got a vehicle because none were available in their area?
- **Supervised Learning - How to model the dependency between future and past demand quantities?** (one-time ahead forecasts)

1-on-1: Feature Engineering, Selection and Prob. Formulation

- Each ROI will have a discrete time series where each term correspond to the number of services per time slot;
- To predict the next term of each series is a special Regression problem known as **point forecast;**
- Model-based approach: **VAR.** Reasons? **Simplicity and Explanatory Power**
- Feature Selection done through **L1-penalty**;

Problem

$$\mathbb{Y} = \begin{pmatrix} y_1^1 & y_2^1 & \cdots & y_{t-1}^1 & y_t^1 \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ y_1^i & \cdots & \cdots & y_{t-1}^i & y_t^i \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ y_1^K & \cdots & \cdots & y_{t-1}^K & y_t^K \end{pmatrix} = \begin{pmatrix} \mathbf{Y}_1 \\ \vdots \\ \mathbf{Y}_{t-j} \\ \vdots \\ \mathbf{Y}_t \end{pmatrix}^T$$

Model

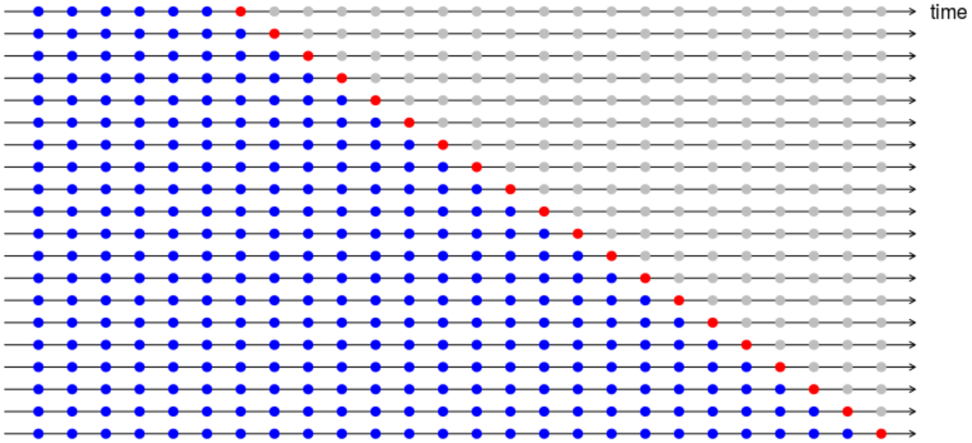
$$\begin{aligned} \hat{\mathbf{Y}}_{t+1} &= \mathbf{Y}_t{}^T \Phi_0 + \mathbf{Y}_{t-1}{}^T \Phi_1 + \cdots + \mathbf{Y}_{t-p+1}{}^T \Phi_{p-1} + \epsilon_t \\ &= \sum_{j=1}^{p} \mathbf{Y}_{t-j+1}{}^T \Phi_{j-1} + \epsilon_t \end{aligned}$$

Optimization

$$\min_{\mathbf{\Phi}} \left\{ l(\hat{\mathbf{Y}}_{t+1}, \mathbf{Y}_{t+1}) + \tfrac{1}{2} ||\mathbf{\Phi}||\beta^2 \right\}$$

1-on-1: Algorithm Selection and Hyperparameter Tuning

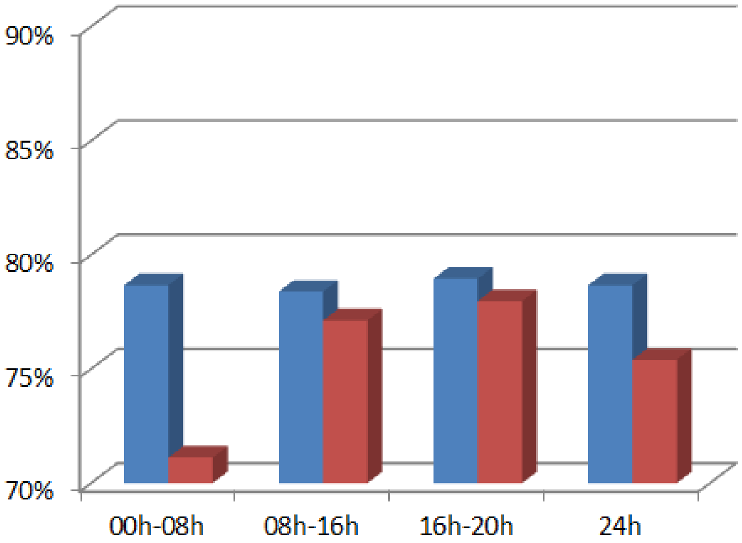
- Two classical solutions for solvers: **least squares** OR **stochastic gradient descent**;
- We would took the latter so we can be adaptive to drift (a story to another talk...)...but we end testing both;
- Hyperparameter Tuning is made classically for BETA using grid search. Generalization error estimation made using TSCV (time series cross validation).
- Post processing? Just round the output to integer domain ☺



1-on-1: Results and Lessons Learned

Results

- Results on test set: blue (LS) and red (SGD);
- Evaluation Metric: 1-sMAPE



1-on-1 with a Data Science Project: Lessons Learned

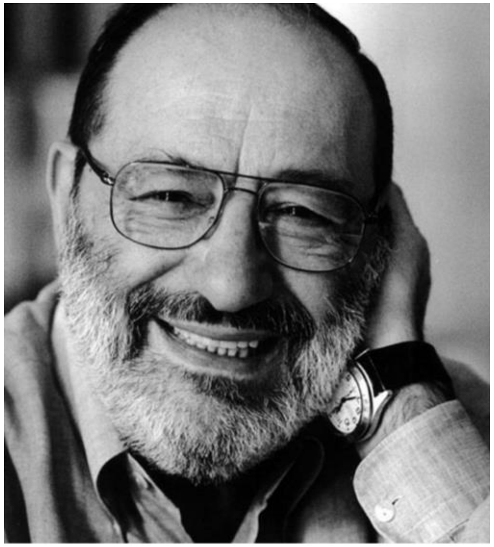
- A careful exploration of your dataset throughout simple experiments and visualization tools can help you towards taking good modelling decisions;
- Simple decisions such as regularization mechanism or optimization solver can have huge impacts on your KPIs;
- Today, DS Pipelines are still very human dependent;

Wait a sec...why can’t a computer do that?

“Tl pute t an
ir at
he ple.

It is a that
c ir ls
art pe

What if we can CHANGE that?



Umberto Eco

(1932-2016)
Italian Philosopher

AutoML Success...hum, notable Cases

DataRobot

scikit learn

MLBox

Cloud AutoML

The Automatic Statistician

An artificial intelligence for data science

7 Tomatoes or … Claps Time!

1. DS is about to transform raw data into knowledge automatically;

2..Data Science is one of the most thrilling industries worldwide.
Why? Computing Power and loads of digital data available!

3. The lack of experts pushes organizations to stretch its definition way beyond its real scope;

4. Deep Learning? Don’t fall by jumping in the hype.
Neural nets are super cool…but it is just yet another tool.

5. On a real-world project, a careful preparation and good modelling decisions make huge difference;

6. Be aware: Automation is the new battleground!!!

7. AI goes way beyond ML.
We will only make it real by addressing its different challenges together!