AI for Health

Erwan Scornet (Maître de conférences, Ecole Polytechnique)
1 Introduction to AI
- No data project without data
- Data science, Machine Learning, Artificial Intelligence... Which term should we use?
- Different applications of AI

2 How does Machine Learning work?

3 Application of AI in health
- Radiology
  - Chest X-ray
  - Liver lesion segmentation
- Genomics
  - Gene network
  - Toxicogenetics
- Medical

4 Perspective and issues
- Limitations of data projects
- Data Project Organization
- Unveiling the mystery of Deep Learning
Where are the data?

Everywhere!

Good point: impossible to do a data project without data
Wooclap: What is the average quantity of data created per person per day in 2020?

Some scale (on average):
- An office document (Word, PowerPoint...): 321 KB
- 1 picture with a smartphone: 10 MB
- Recording a 3 hour zoom meeting: 1 GB
But how much data exactly?

Wooclap: What is the average quantity of data created per person per day in 2020?

Some scale (on average):
- An office document (Word, PowerPoint...): 321 KB
- 1 picture with a smartphone : 10 MB
- Recording a 3 hour zoom meeting: 1 GB

The quantity of data produced each day per person in 2020 is 2GB which is equivalent to

6h of zoom meeting recording or 200 pictures or 6,000 Office documents
EVERY DAY WE CREATE
2,500,000,000,000,000
(2.5 QUINTILLION) BYTES OF DATA

90% of the world's data today has been created in the last 2 years alone.

This would fill 10 million Blu-ray discs; the height of which stacked, would measure the height of 4 Eiffel Towers on top of one another.

BIG DATA:
Data stored grows
4X FASTER THAN WORLD ECONOMY

Substantial shift in
ECONOMIC POWER AND SOURCE OF ECONOMIC VALUE

Increasing quantity of data allows for
MORE QUALITATIVE APPROACH

Figure: OECD, 2019
Global Data Creation is About to Explode
Actual and forecast amount of data created worldwide 2010-2035 (in zettabytes)

1 zettabyte is equal to 1 billion terabytes.

Source: Statista Digital Economy Compass 2019
"Data is the new oil."

From the beginning of recorded time until 2003, we created 5 exabytes of data.

10% of the world's data was created in the past 2 years.

By 2020, it is expected that this time will shrink to 15 years.

Data is power.
Introduction to AI
- No data project without data
- Data science, Machine Learning, Artificial Intelligence... Which term should we use?
- Different applications of AI

How does Machine Learning work?

Application of AI in health
- Radiology
  - Chest X-ray
  - Liver lesion segmentation
- Genomics
  - Gene network
  - Toxicogenetics
- Medical

Perspective and issues
- Limitations of data projects
- Data Project Organization
- Unveiling the mystery of Deep Learning
When you're fundraising, it's AI / When you're hiring, it's ML / When you're implementing, it's linear regression / When you're debugging, it's printf()

Baron Schwartz, Twitter, Nov 2017
**Wooclap**: Assign to each term its corresponding definition.

- **Data Management**
  - Is the study of the collection, analysis, interpretation, presentation and organization of data.

- **Business Intelligence**
  - Comprises the strategies and technologies used by enterprises for the data analysis of business information.

- **Statistics**
  - Is part of a broader family of machine learning methods based on learning data representations, as opposed to task-specific algorithms.

- **Data science**
  - Is the study of the generalizable extraction of knowledge from data.

- **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.

- **Machine learning**
  - Comprises all disciplines related to handling data as a valuable resource.

- **Artificial Intelligence**
  - Is the subfield of computer science that gives computers the ability to learn without being explicitly programmed.

- **Deep Learning**
  - Aims at designing and studying *devices* that perceive its environment and take actions that maximize its chance of success at some goal.
Data terminology

- **Data Management** comprises all disciplines related to handling data as a valuable resource.

- **Business Intelligence** comprises the strategies and technologies used by enterprises for the data analysis of business information.

- **Statistics** is the study of the collection, analysis, interpretation, presentation and organization of data.

- **Data science** is the study of the generalizable extraction of knowledge from data.

- **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.

- **Machine learning** is the subfield of computer science that gives computers the ability to learn without being explicitly programmed.

- **Artificial Intelligence** aims at designing and studying *devices* that perceive its environment and take actions that maximize its chance of success at some goal.

- **Deep Learning** is part of a broader family of machine learning methods based on learning data representations, as opposed to task-specific algorithms.
On September 1955, a project was proposed by McCarthy, Marvin Minsky, Nathaniel Rochester and Claude Shannon introducing formally for the first time the term "Artificial Intelligence".

The study is to proceed on the basis of the conjecture that every aspect of learning or any other feature of intelligence can in principle be so precisely described that a machine can be made to simulate it. An attempt will be made to find how to make machines use language, form abstractions and concepts, solve kinds of problems now reserved for humans, and improve themselves.

Proposal for Dartmouth conference on AI (1956)
Misconception of AI

AI is about electronic device able to **mimic** human thinking:

- **Artificial intelligence**
- One famous class of AI algorithms are called **neural networks**.
- **Android** are close to humans in shape so they must think like humans.

Most AI algorithms do **not** aim at reproducing human reasoning.

*Artificial intelligence is the science of making machines do things that would require intelligence if done by men*

Marvin Minsky (1968)

*2001: A Space Odyssey*
What often happens is that an engineer has an idea of how the brain works (in his opinion) and then designs a machine that behaves that way. This new machine may in fact work very well. But, I must warn you that that does not tell us anything about how the brain actually works, nor is it necessary to ever really know that, in order to make a computer very capable. It is not necessary to understand the way birds flap their wings and how the feathers are designed in order to make a flying machine [...] It is therefore not necessary to imitate the behavior of Nature in detail in order to engineer a device which can in many respects surpass Nature’s abilities.

Richard Feynman (1999)
AI technology - Autonomous cars

- Originates from 1920 (NY)
- First use of neural networks to control autonomous cars (1989)
- Four US states allow self-driving cars (2013)
- First known fatal accident (May 2016)
- Singapore launched the first self-driving taxi service (Aug. 2016)
- A Arizona pedestrian was killed by an Uber self-driving car (March 2018).
- System capable of analyzing entire word sequences (1980).
- Siri was the first modern digital virtual assistant installed on a smartphone (2011).
- Watson won the TV show Jeopardy! (2011)
Different uses of AI
The study found that Google Home performed the best, recognizing 98 per cent of topics accurately and providing advice that matched with Red Cross first aid guidelines 56 per cent of the time.

Alexa recognized 92 per cent of topics, and gave appropriate advice 19 per cent of the time.

The responses from Siri and Cortana were so low that researchers determined that they couldn't analyze them.
Different uses of AI

Erwan Scornet

AI for Health
Different uses of AI

Erwan Scornet

AI for Health
Different uses of AI

AI artwork sells for $432,500 — nearly 45 times its high estimate — as Christie’s becomes the first auction house to offer a work of art created by an algorithm.
Different uses of AI

Erwan Scornet

A Look Inside Augmented Analytics And Its Business Value In 2020

AI for Health
Different uses of AI

Erwan Scornet

AI for Health
Introduction to AI
- No data project without data
- Data science, Machine Learning, Artificial Intelligence... Which term should we use?
- Different applications of AI

How does Machine Learning work?

Application of AI in health
- Radiology
  - Chest X-ray
  - Liver lesion segmentation
- Genomics
  - Gene network
  - Toxicogenetics
- Medical

Perspective and issues
- Limitations of data projects
- Data Project Organization
- Unveiling the mystery of Deep Learning
Supervised learning

Training data

\[
\begin{array}{cccccccccc}
0 & 1 & 2 & 3 & 4 & 5 & 6 & 7 & 8 & 9 \\
1 & 2 & 3 & 4 & 5 & 6 & 7 & 8 & 9 & 0 \\
2 & 3 & 4 & 5 & 6 & 7 & 8 & 9 & 0 & 1 \\
0 & 1 & 2 & 3 & 4 & 5 & 6 & 7 & 8 & 9 \\
\end{array}
\]

Learning algorithm

- Logistic regression
- Random forests
- Neural networks
- ...

New input → Classifier → Output

(0,0,0,0,1,0,0,0,0,0)

A definition by Tom Mitchell (http://www.cs.cmu.edu/~tom/)

A computer program is said to learn from experience E with respect to some class of tasks T and performance measure P, if its performance at tasks in T, as measured by P, improves with experience E.
Three Kinds of Learning

Unsupervised Learning
- **Task:** Clustering/DR
- **Performance:** Quality
- **Experience:** Raw dataset (No Ground Truth)

Supervised Learning
- **Task:** Prediction
- **Performance:** Average error
- **Experience:** Predictions (Ground Truth)

Reinforcement Learning
- **Task:** Action
- **Performance:** Total reward
- **Experience:** Reward from env. (Interact. with env.)

**Timing:** Offline/Batch (learning from past data) vs Online (continuous learning)

Figure Source: BCG
Difficulties related to (Big) data

- The prediction must be **accurate**: difficult for some tasks like image classification, video captioning...
- Predictions must be **fast**: online recommendation should not take minutes.
- Data must be **stored** and **easily accessible**.
- It may be difficult to **access all data simultaneously**. Data may come sequentially.
- Data must be **clean**.
- Data should be **relevant**.

**Wooclap**: How would you evaluate the performance of an ML algorithm aiming at diagnosing a patient?
Various applications of AI in health

Wooclap: Can you think of one application of AI in health/medicine?
Wooclap: Can you think of one application of AI in health/medicine?

- **Diagnosis**
  - From 1970s, an expert system to identify blood infections.
  - Based on clinical data, omics data, medical imaging...

Various applications of AI in health
Various applications of AI in health

Wooclap: Can you think of one application of AI in health/medicine?

- **Diagnosis**
  - From 1970s, an expert system to identify blood infections.
  - Based on clinical data, omics data, medical imaging...

- **Predicting / Forecasting**
  - Patient journey (predict missing appointments, ER waiting time)
  - Forecasting glycemic concentration (D2P project, DiabeLoop)
  - Classifying signals: Brainwaves / ECG (Cardiologs)
Wooclap: *Can you think of one application of AI in health/medicine?*

- **Diagnosis**
  - From 1970s, an *expert system* to identify blood infections.
  - Based on clinical data, omics data, medical imaging...

- **Predicting / Forecasting**
  - Patient journey (predict missing appointments, ER waiting time)
  - *Forecasting glycemic concentration* (D2P project, DiabeLoop)
  - Classifying signals: Brainwaves / ECG (Cardiologs)

- **Drugs**
  - Determining *protein 3D shapes* (AlphaFold, 2020)
  - *Finding new drugs* (hit identification and de novo molecular design)
Various applications of AI in health

Wooclap: Can you think of one application of AI in health/medicine?

- Diagnosis
  - From 1970s, an expert system to identify blood infections.
  - Based on clinical data, omics data, medical imaging...

- Predicting / Forecasting
  - Patient journey (predict missing appointments, ER waiting time)
  - Forecasting glycemic concentration (D2P project, DiabeLoop)
  - Classifying signals: Brainwaves / ECG (Cardiologs)

- Drugs
  - Determining protein 3D shapes (AlphaFold, 2020)
  - Finding new drugs (hit identification and de novo molecular design)

- Robots
  - Surgical robots (robotic laparoscopes, sutures, cuts...)
  - Robots for rehabilitation, physical therapy
  - Prosthetic arm control
Various applications of AI in health

**Wooclap:** Can you think of one application of AI in health/medicine?

- **Diagnosis**
  - From 1970s, an *expert system* to identify blood infections.
  - Based on clinical data, omics data, medical imaging...

- **Predicting / Forecasting**
  - Patient journey (predict missing appointments, ER waiting time)
  - *Forecasting glycemic concentration* (D2P project, DiabeLoop)
  - Classifying signals: Brainwaves / ECG (Cardiologs)

- **Drugs**
  - Determining *protein 3D shapes* (AlphaFold, 2020)
  - *Finding new drugs* (hit identification and de novo molecular design)

- **Robots**
  - *Surgical* robots (robotic laparoscopes, sutures, cuts...)
  - Robots for *rehabilitation, physical therapy*
  - *Prosthetic arm control*

- **Automation**
Various applications of AI in health

**Wooclap:** Can you think of one application of AI in health/medicine?

- **Diagnosis**
  - From 1970s, an expert system to identify blood infections.
  - Based on clinical data, omics data, medical imaging...

- **Predicting / Forecasting**
  - Patient journey (predict missing appointments, ER waiting time)
  - Forecasting glycemic concentration (D2P project, DiabeLoop)
  - Classifying signals: Brainwaves / ECG (Cardiologs)

- **Drugs**
  - Determining protein 3D shapes (AlphaFold, 2020)
  - Finding new drugs (hit identification and de novo molecular design)

- **Robots**
  - Surgical robots (robotic laparoscopes, sutures, cuts...)
  - Robots for rehabilitation, physical therapy
  - Prosthetic arm control

- **Automation**

- **Treatment efficiency** (causality)
1 Introduction to AI
- No data project without data
- Data science, Machine Learning, Artificial Intelligence... Which term should we use?
- Different applications of AI

2 How does Machine Learning work?

3 Application of AI in health
- Radiology
  - Chest X-ray
  - Liver lesion segmentation
- Genomics
  - Gene network
  - Toxicogenetics
- Medical

4 Perspective and issues
- Limitations of data projects
- Data Project Organization
- Unveiling the mystery of Deep Learning
Can Deep Learning be useful in radiology?

Deep Learning requires large annotated data sets (typically several millions of images, as in ImageNet)
637 X-ray images (6 chest pathologies):

- Right Pleural Effusion (73 images)
- Left Pleural Effusion (74 images)
- Right Consolidation (58 images)
- Left Consolidation (45 images)
- Cardiomegaly (154 images)
- Abnormal Mediastinum (145 images)
**Figure 2.** A schematic illustration of Donahue et al. (2013) CNN architecture and training process.

**Figure 3.** Feature extraction from Donahue et al. (2013) pre-trained CNN.
Convolutional Neural Network

Use SVM to predict a class based on the following features:

- Deep Learning features of different layers,
- Bag-of-Visual-Words (BoVW), particularly useful to categorize X-rays on organ level (ImageClef competitions, http://www.imageclef.org)
- GIST descriptors [4]
Figure 4. Features ranking, evaluated on the testing set, applied on the combined feature vector (BoVW, Deep features (Decaf5, Decaf6 and Decaf7), Gist). Ranking is performed on all examined identification cases. We use the following abbreviation: Ln for Decaf n.
### Results

**Table 1. AUC accuracy metric classification performance.**

<table>
<thead>
<tr>
<th>Descriptor</th>
<th>Right pleural effusion</th>
<th>Left pleural effusion</th>
<th>Right consolidation</th>
<th>Left consolidation</th>
</tr>
</thead>
<tbody>
<tr>
<td>GIST</td>
<td>0.85</td>
<td>0.79</td>
<td>0.77</td>
<td>0.41</td>
</tr>
<tr>
<td>BoVW</td>
<td>0.89</td>
<td>0.87</td>
<td>0.78</td>
<td>0.65</td>
</tr>
<tr>
<td>L5</td>
<td>0.91</td>
<td>0.81</td>
<td>0.80</td>
<td>0.75</td>
</tr>
<tr>
<td>L6</td>
<td>0.91</td>
<td>0.82</td>
<td>0.85</td>
<td>0.76</td>
</tr>
<tr>
<td>L7</td>
<td>0.90</td>
<td>0.79</td>
<td>0.75</td>
<td>0.79</td>
</tr>
<tr>
<td>L5 + L6 + GIST</td>
<td>0.92</td>
<td>0.82</td>
<td>0.83</td>
<td>0.68</td>
</tr>
<tr>
<td>L5 + L6 + L7</td>
<td>0.92</td>
<td>0.82</td>
<td>0.83</td>
<td>0.78</td>
</tr>
<tr>
<td>FS (5000)</td>
<td>0.93</td>
<td>0.82</td>
<td>0.84</td>
<td>0.78</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Cardiomegaly</th>
<th>Abnormal mediastinum</th>
<th>Healthy vs. pathology</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.96</td>
<td>0.73</td>
<td>0.88</td>
</tr>
<tr>
<td>0.94</td>
<td>0.74</td>
<td>0.85</td>
</tr>
<tr>
<td>0.95</td>
<td>0.79</td>
<td>0.90</td>
</tr>
<tr>
<td>0.94</td>
<td>0.80</td>
<td>0.90</td>
</tr>
<tr>
<td>0.93</td>
<td>0.79</td>
<td>0.89</td>
</tr>
<tr>
<td>0.96</td>
<td>0.79</td>
<td>0.91</td>
</tr>
<tr>
<td>0.94</td>
<td>0.80</td>
<td>0.91</td>
</tr>
<tr>
<td>0.95</td>
<td>0.80</td>
<td>0.92</td>
</tr>
</tbody>
</table>
Figure: Healthy vs. Pathology ROC;
1 Introduction to AI
- No data project without data
- Data science, Machine Learning, Artificial Intelligence... Which term should we use?
- Different applications of AI

2 How does Machine Learning work?

3 Application of AI in health
- Radiology
  - Chest X-ray
  - Liver lesion segmentation
- Genomics
  - Gene network
  - Toxicogenetics
- Medical

4 Perspective and issues
- Limitations of data projects
- Data Project Organization
- Unveiling the mystery of Deep Learning
Data:

- 131 abdominal CT scans
- The CT scans come with reference annotations of the liver and tumors done by trained radiologists (LiTS reference)
- Another set of annotations given by a medical-technical radiology assistant (MTRA)
Fully Convolutional Neural Networks

Based on U-net [6] designed specifically for Biomedical Image Segmentation.

Figure 1. Overview of the neural network architecture. The numbers denote the feature map count.
Figure 5. Neural network (black) compared with the LiTS (white) annotations. (a) Case with 0.85 dice/case (b,c) Cases with 19 and 16 FPs (d) Case where a small tumor was not detected (e,f) Case where tumor segmentation strongly differed on consecutive slices.
Comparison between Human and Computer recognition performances

<table>
<thead>
<tr>
<th></th>
<th>Recall</th>
<th>Recall ≥ 10 mm</th>
<th>FP per case</th>
<th>Dice per case</th>
<th>Dice per correspondence</th>
<th>Merge error</th>
<th>Split error</th>
</tr>
</thead>
<tbody>
<tr>
<td>Human vs. Human</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>MTRA (LiTS)</td>
<td>0.92</td>
<td>0.94</td>
<td>2.6</td>
<td>0.70 ± 0.27</td>
<td>0.72 ± 0.11</td>
<td>11</td>
<td>5</td>
</tr>
<tr>
<td>LiTS (MTRA)</td>
<td>0.62</td>
<td>0.85</td>
<td>0.3</td>
<td>0.70 ± 0.27</td>
<td>0.72 ± 0.11</td>
<td>5</td>
<td>12</td>
</tr>
<tr>
<td>Computer vs. Human</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>FCN (MTRA)</td>
<td>0.47</td>
<td>0.75</td>
<td>4.7</td>
<td>0.53 ± 0.37</td>
<td>0.72 ± 0.11</td>
<td>7</td>
<td>13</td>
</tr>
<tr>
<td>FCN (LiTS)</td>
<td>0.72</td>
<td>0.86</td>
<td>4.6</td>
<td>0.51 ± 0.37</td>
<td>0.65 ± 0.16</td>
<td>12</td>
<td>14</td>
</tr>
<tr>
<td>FCN + RF (LiTS)</td>
<td>0.63</td>
<td>0.77</td>
<td>0.7</td>
<td>0.58 ± 0.36</td>
<td>0.69 ± 0.18</td>
<td>11</td>
<td>10</td>
</tr>
</tbody>
</table>

**Table 1.** Mean metric values for human vs. human and computer vs. human comparisons. The parentheses denote the dataset used as a reference for the computation of evaluation metrics.
**Introduction to AI**
- No data project without data
- Data science, Machine Learning, Artificial Intelligence... Which term should we use?
- Different applications of AI

**How does Machine Learning work?**

**Application of AI in health**
- Radiology
  - Chest X-ray
  - Liver lesion segmentation
- Genomics
  - Gene network
  - Toxicogenetics
- Medical

**Perspective and issues**
- Limitations of data projects
- Data Project Organization
- Unveiling the mystery of Deep Learning
Different fields in Omics [3]

- **Transcriptomics**
  - gene expression analysis
  - non-coding RNA analysis
  - gene-fusion detection
  - mRNA splice-analysis
  - RNA editing

- **miRNomics**
  - expression analysis
  - miRNA identification
  - miRNA editing

- **Epigenomics**
  - targeted bisulfite sequencing
  - ChIP-Seq
  - MeDIP-Seq, methylCap-seq
  - whole genome bisulfite sequencing

- **Genomics**
  - targeted sequencing
  - whole-exome sequencing
  - whole-genome sequencing

Information content, experimental costs, current clinical utility.
[5] Effect of low irradiation on Saccharomyces cerevisiae strains: 6 irradiated groups vs 11 non-irradiated groups.

Aim: Detect the irradiation looking only at the transcriptional changes.
1 Introduction to AI
   • No data project without data
   • Data science, Machine Learning, Artificial Intelligence… Which term should we use?
   • Different applications of AI

2 How does Machine Learning work?

3 Application of AI in health
   • Radiology
     • Chest X-ray
     • Liver lesion segmentation
   • Genomics
     • Gene network
     • Toxicogenetics
   • Medical

4 Perspective and issues
   • Limitations of data projects
   • Data Project Organization
   • Unveiling the mystery of Deep Learning
What is toxicogenetics?

- Different responses to drugs or environmental chemicals according to genotypes
- Aim: provide a personalized treatment for patients
DREAM is a Dialogue for Reverse Engineering Assessments and Methods. The main objective is to catalyze the interaction between experiment and theory in the area of cellular network inference and quantitative model building in systems biology.

Dialogue for Reverse Engineering Assessments and Methods

DREAM8 Challenges Are Open

We are pleased to announce that the DREAM8 Challenges are now open for participation. During the “Challenge season” spanning from June 10 to September 15, 2013, Sage Bionetworks and DREAM will run the following three Challenges:

1. HPN-DREAM Breast Cancer Network Inference Challenge – Infer the signaling networks in breast cancer cell lines
2. NIEHS-NCATS-UNC DREAM Toxigenetics Challenge – Predict individual response to environmental and pharmaceutical chemicals
3. The Whole-Cell Parameter Estimation DREAM Challenge – Infer the kinetic parameters underlying biological processes in whole cell models

To sign up for a Challenge, and access the data sets and descriptions of the DREAM8 Challenges, please go to https://www.synapse.org/

Erwan Scornet AI for Health
DREAM8 Toxicogenetics challenge

Toxicogenetics Challenge Data

- Chemical descriptors
  - 10K attributes

- Genotypes
  - Not available
  - 1.3M SNPs

- RNASeq
  - 46K transcripts

- LCLs
  - 337 LCLs

Cytotoxicity data (EC_{10})

- Training Set
  - 487 Cell Lines
  - 106 chemicals

- Test Set
  - Subchallenge 1
    - 156 chemicals
  - Subchallenge 2
    - 884 Cell Lines
The two G projects

The 1000 Genomes Project
- Decreasing of sequencing cost allow to sequence a large number of people
- Find most genetic variants that have frequencies of 1%.
- Data are publicly available at http://www.1000genomes.org/data

The Geuvadis Project
- European project in high-throughput sequencing
- RNA sequencing of 465 cell lines belonging to the 1000 Genomes Project
- So RNA data for non-european people are missing in the challenge
- Data publicly available at http://www.geuvadis.org
Our strategy

- Need to integrate large-scale, heterogeneous data with missing information → kernels
- Share information across chemicals → multitask learning

So we developed a joint model on both patients and drugs.
Kernel Trick

cell line descriptors

drug descriptors
Kernel Trick

cell line descriptors

Kcell

Kdrug

drug descriptors

Kernel Trick
Kernel Trick

cell line descriptors

Kcell

kernelized

Kdrug

kernel bilinear regression

\hat{f}

drug descriptors
Kernel Trick

Kernel choice?
- descriptors
- data integration
- missing data

Kcell

Kdrug

kernelized

kernel bilinear regression

\( \hat{f} \)
Introduction to AI
- No data project without data
- Data science, Machine Learning, Artificial Intelligence... Which term should we use?
- Different applications of AI

How does Machine Learning work?

Application of AI in health
- Radiology
  - Chest X-ray
  - Liver lesion segmentation
- Genomics
  - Gene network
  - Toxicogenetics
- Medical

Perspective and issues
- Limitations of data projects
- Data Project Organization
- Unveiling the mystery of Deep Learning
15000 patients / 250 variables / 11 hospitals, from 2011

- 4000 new patients / year

<table>
<thead>
<tr>
<th>Center</th>
<th>Accident</th>
<th>Age</th>
<th>Sex</th>
<th>Weight</th>
<th>Height</th>
<th>BMI</th>
<th>BP</th>
<th>SBP</th>
</tr>
</thead>
<tbody>
<tr>
<td>1 Beaujon</td>
<td>Fall</td>
<td>54</td>
<td>m</td>
<td>85</td>
<td>NR</td>
<td>NR</td>
<td>180</td>
<td>110</td>
</tr>
<tr>
<td>2 Lille</td>
<td>Other</td>
<td>33</td>
<td>m</td>
<td>80</td>
<td>1.8</td>
<td>24.69</td>
<td>130</td>
<td>62</td>
</tr>
<tr>
<td>3 Pitie Salpetriere</td>
<td>Gun</td>
<td>26</td>
<td>m</td>
<td>NR</td>
<td>NR</td>
<td>NR</td>
<td>131</td>
<td>62</td>
</tr>
<tr>
<td>4 Beaujon</td>
<td>AVP moto</td>
<td>63</td>
<td>m</td>
<td>80</td>
<td>1.8</td>
<td>24.69</td>
<td>145</td>
<td>89</td>
</tr>
<tr>
<td>6 Pitie Salpetriere</td>
<td>AVP bicycle</td>
<td>33</td>
<td>m</td>
<td>75</td>
<td>NR</td>
<td>NR</td>
<td>104</td>
<td>86</td>
</tr>
<tr>
<td>7 Pitie Salpetriere</td>
<td>AVP pedestrian</td>
<td>30</td>
<td>w</td>
<td>NR</td>
<td>NR</td>
<td>NR</td>
<td>107</td>
<td>66</td>
</tr>
<tr>
<td>9 HEGP</td>
<td>White weapon</td>
<td>16</td>
<td>m</td>
<td>98</td>
<td>1.92</td>
<td>26.58</td>
<td>118</td>
<td>54</td>
</tr>
<tr>
<td>10 Toulon</td>
<td>White weapon</td>
<td>20</td>
<td>m</td>
<td>NR</td>
<td>NR</td>
<td>NR</td>
<td>124</td>
<td>73</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>SpO2</th>
<th>Temperature</th>
<th>Lactates</th>
<th>Hb</th>
<th>Glasgow</th>
<th>Transfusion</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>97</td>
<td>35.6</td>
<td>&lt;NA&gt;</td>
<td>12.7</td>
<td>12</td>
</tr>
<tr>
<td>2</td>
<td>100</td>
<td>36.5</td>
<td>4.8</td>
<td>11.1</td>
<td>15</td>
</tr>
<tr>
<td>3</td>
<td>100</td>
<td>36</td>
<td>3.9</td>
<td>11.4</td>
<td>3</td>
</tr>
<tr>
<td>4</td>
<td>100</td>
<td>36.7</td>
<td>1.66</td>
<td>13</td>
<td>15</td>
</tr>
<tr>
<td>6</td>
<td>100</td>
<td>36</td>
<td>NM</td>
<td>14.4</td>
<td>15</td>
</tr>
<tr>
<td>7</td>
<td>100</td>
<td>36.6</td>
<td>NM</td>
<td>14.3</td>
<td>15</td>
</tr>
<tr>
<td>9</td>
<td>100</td>
<td>37.5</td>
<td>13</td>
<td>15.9</td>
<td>15</td>
</tr>
<tr>
<td>10</td>
<td>100</td>
<td>36.9</td>
<td>NM</td>
<td>13.7</td>
<td>15</td>
</tr>
</tbody>
</table>

http://www.traumabase.eu/fr_FR
**Major trauma**: any injury that endangers the life or the functional integrity of a person. Road traffic accidents, interpersonal violence, self-harm, falls, etc
→ hemorrhage and traumatic brain injury.

Patient prognosis can be improved: **standardized and reproducible procedures** but **personalized** for the patient and the trauma system.

⇒ Can AI help?
<table>
<thead>
<tr>
<th>Center</th>
<th>Accident</th>
<th>Age</th>
<th>Sex</th>
<th>Weight</th>
<th>Height</th>
<th>BMI</th>
<th>BP</th>
<th>SBP</th>
</tr>
</thead>
<tbody>
<tr>
<td>Beaujon</td>
<td>Fall</td>
<td>54</td>
<td>m</td>
<td>85</td>
<td>NR</td>
<td>NR</td>
<td>180</td>
<td>110</td>
</tr>
<tr>
<td>Lille</td>
<td>Other</td>
<td>33</td>
<td>m</td>
<td>80</td>
<td>1.8</td>
<td>24.69</td>
<td>130</td>
<td>62</td>
</tr>
<tr>
<td>Pitie Salpetriere</td>
<td>Gun</td>
<td>26</td>
<td>m</td>
<td>NR</td>
<td>NR</td>
<td>NR</td>
<td>131</td>
<td>62</td>
</tr>
<tr>
<td>Beaujon</td>
<td>AVP moto</td>
<td>63</td>
<td>m</td>
<td>80</td>
<td>1.8</td>
<td>24.69</td>
<td>145</td>
<td>89</td>
</tr>
<tr>
<td>Pitie Salpetriere</td>
<td>AVP bicycle</td>
<td>33</td>
<td>m</td>
<td>75</td>
<td>NR</td>
<td>NR</td>
<td>104</td>
<td>86</td>
</tr>
<tr>
<td>Pitie Salpetriere</td>
<td>AVP pedestrian</td>
<td>30</td>
<td>w</td>
<td>NR</td>
<td>NR</td>
<td>NR</td>
<td>107</td>
<td>66</td>
</tr>
<tr>
<td>HEGP</td>
<td>White weapon</td>
<td>16</td>
<td>m</td>
<td>98</td>
<td>1.92</td>
<td>26.58</td>
<td>118</td>
<td>54</td>
</tr>
<tr>
<td>Toulon</td>
<td>White weapon</td>
<td>20</td>
<td>m</td>
<td>NR</td>
<td>NR</td>
<td>NR</td>
<td>124</td>
<td>73</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>SpO2</th>
<th>Temperature</th>
<th>Lactates</th>
<th>Hb</th>
<th>Glasgow</th>
<th>Transfusion</th>
</tr>
</thead>
<tbody>
<tr>
<td>97</td>
<td>35.6</td>
<td>&lt;NA&gt;</td>
<td>12.7</td>
<td>12</td>
<td>yes</td>
</tr>
<tr>
<td>100</td>
<td>36.5</td>
<td>4.8</td>
<td>11.1</td>
<td>15</td>
<td>no</td>
</tr>
<tr>
<td>100</td>
<td>36</td>
<td>3.9</td>
<td>11.4</td>
<td>3</td>
<td>no</td>
</tr>
<tr>
<td>100</td>
<td>36.7</td>
<td>1.66</td>
<td>13</td>
<td>15</td>
<td>yes</td>
</tr>
<tr>
<td>100</td>
<td>36</td>
<td>NM</td>
<td>14.4</td>
<td>15</td>
<td>no</td>
</tr>
<tr>
<td>100</td>
<td>36.6</td>
<td>NM</td>
<td>14.3</td>
<td>15</td>
<td>yes</td>
</tr>
<tr>
<td>100</td>
<td>37.5</td>
<td>13</td>
<td>15.9</td>
<td>15</td>
<td>yes</td>
</tr>
<tr>
<td>100</td>
<td>36.9</td>
<td>NM</td>
<td>13.7</td>
<td>15</td>
<td>no</td>
</tr>
</tbody>
</table>

⇒ **Predict** whether to start a blood transfusion, to administer fresh frozen plasma, etc...
<table>
<thead>
<tr>
<th>Center</th>
<th>Accident</th>
<th>Age</th>
<th>Sex</th>
<th>Weight</th>
<th>Height</th>
<th>BMI</th>
<th>BP</th>
<th>SBP</th>
</tr>
</thead>
<tbody>
<tr>
<td>Beaujon</td>
<td>Fall</td>
<td>54</td>
<td>m</td>
<td>85</td>
<td>NR</td>
<td>NR</td>
<td>180</td>
<td>110</td>
</tr>
<tr>
<td>Lille</td>
<td>Other</td>
<td>33</td>
<td>m</td>
<td>80</td>
<td>1.8</td>
<td>24.69</td>
<td>130</td>
<td>62</td>
</tr>
<tr>
<td>Pitie Salp.</td>
<td>Gun</td>
<td>26</td>
<td>m</td>
<td>NR</td>
<td>NR</td>
<td>NR</td>
<td>131</td>
<td>62</td>
</tr>
<tr>
<td>Beaujon</td>
<td>AVP moto</td>
<td>63</td>
<td>m</td>
<td>80</td>
<td>1.8</td>
<td>24.69</td>
<td>145</td>
<td>89</td>
</tr>
<tr>
<td>Pitie Salp.</td>
<td>AVP bike</td>
<td>33</td>
<td>m</td>
<td>75</td>
<td>NR</td>
<td>NR</td>
<td>104</td>
<td>86</td>
</tr>
<tr>
<td>Pitie Salp.</td>
<td>AVP ped.</td>
<td>30</td>
<td>w</td>
<td>NR</td>
<td>NR</td>
<td>NR</td>
<td>107</td>
<td>66</td>
</tr>
<tr>
<td>HEGP</td>
<td>White</td>
<td>16</td>
<td>m</td>
<td>98</td>
<td>1.92</td>
<td>26.58</td>
<td>118</td>
<td>54</td>
</tr>
<tr>
<td>Toulon</td>
<td>White</td>
<td>20</td>
<td>m</td>
<td>NR</td>
<td>NR</td>
<td>NR</td>
<td>124</td>
<td>73</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>SpO2</th>
<th>Temperature</th>
<th>Lactates</th>
<th>Hb</th>
<th>Glasgow</th>
<th>Transfusion</th>
</tr>
</thead>
<tbody>
<tr>
<td>97</td>
<td>35.6</td>
<td>&lt;NA&gt;</td>
<td>12.7</td>
<td>12</td>
<td>yes</td>
</tr>
<tr>
<td>100</td>
<td>36.5</td>
<td>4.8</td>
<td>11.1</td>
<td>15</td>
<td>no</td>
</tr>
<tr>
<td>100</td>
<td>36</td>
<td>3.9</td>
<td>11.4</td>
<td>3</td>
<td>no</td>
</tr>
<tr>
<td>100</td>
<td>36.7</td>
<td>1.66</td>
<td>13</td>
<td>15</td>
<td>yes</td>
</tr>
<tr>
<td>100</td>
<td>36</td>
<td>NM</td>
<td>14.4</td>
<td>15</td>
<td>no</td>
</tr>
<tr>
<td>100</td>
<td>36.6</td>
<td>NM</td>
<td>14.3</td>
<td>15</td>
<td>yes</td>
</tr>
<tr>
<td>100</td>
<td>37.5</td>
<td>13</td>
<td>15.9</td>
<td>15</td>
<td>yes</td>
</tr>
<tr>
<td>100</td>
<td>36.9</td>
<td>NM</td>
<td>13.7</td>
<td>15</td>
<td>no</td>
</tr>
</tbody>
</table>

⇒ **Presence of missing data:** Not Recorded, Made, Applicable, etc.
Percentage of missing values

Variable
- null.data
- na.data
- nr.data
- nf.data
- imp.data
How to solve this problem?

1. Delete all missing values → bad idea!

- Impute data with your favorite imputation method
  - Replace all missing values by the mean/median/mode of the corresponding variable.
  - Good point: the mean/median/mode is unchanged!
  - Bad point 1: the variance of the imputed data is lower than reality
  - Bad point 2: structure of dependence between variables is destroyed.

- Design methods that handle missing values.
How to solve this problem?

1. Delete all missing values → bad idea!

2. Impute data with your favorite imputation method
   - Replace all missing values by the mean/median/mode of the corresponding variable.
     - Good point: the mean/median/mode is unchanged!
     - Bad point 1: the variance of the imputed data is lower than reality
     - Bad point 2: structure of dependence between variable is destroyed.
   - Multiple imputation.
How to solve this problem?

1. Delete all missing values → bad idea!

2. Impute data with your favorite imputation method
   - Replace all missing values by the mean/median/mode of the corresponding variable.
     - Good point: the mean/median/mode is unchanged!
     - Bad point 1: the variance of the imputed data is lower than reality
     - Bad point 2: structure of dependence between variable is destroyed.
   - Multiple imputation.

3. Design methods that handle missing values.
Introduction to AI
- No data project without data
- Data science, Machine Learning, Artificial Intelligence... Which term should we use?
- Different applications of AI

How does Machine Learning work?

Application of AI in health
- Radiology
  - Chest X-ray
  - Liver lesion segmentation
- Genomics
  - Gene network
  - Toxicogenetics
- Medical

Perspective and issues
- Limitations of data projects
- Data Project Organization
- Unveiling the mystery of Deep Learning
Cost of storing data

Money: 300,000 US dollars in Google Cloud to store 1 petabyte during one year.

Data centers and environment
- 2% of the total electricity consumption in the US.
- 626 billion liters of water.
- 2% of total global greenhouse emissions.

Erwan Scornet
AI for Health
Money: 300.000 US dollars in Google Cloud to store 1 petabyte during one year.
Cost of storing data

- **Money**: 300,000 US dollars in Google Cloud to store 1 petabyte during one year.

- **Data centers and environment**
  - 2% of the total electricity consumption in the US.
  - 626 billion liters of water.
  - 2% of total global greenhouse emissions.
The Shift Project [https://theshiftproject.org/lean-ict/]
Female drivers and right front passengers are approximately 17 percent more likely to be killed in a car crash than a male occupant of the same age.

Any seatbelt-wearing female vehicle occupant has 73 percent greater odds of being seriously injured in a frontal car crash than the odds of a seatbelt-wearing male occupant being injured in the same kind and severity of crash.

Sources: NHTSA and the journal Traffic Injury Prevention

Analysis of crash and injury data compiled from the National Automotive Sampling System Crashworthiness Data System for the years 1998 to 2015.
What is COMPAS?

Correctional Offender Management Profiling for Alternative Sanction used in US justice courts to predict the reoffending probability.
What is COMPAS?
Correctional Offender Management Profiling for Alternative Sanction used in US justice courts to predict the reoffending probability.

Assessing the fairness of COMPAS

(A) Calibration
Given a score, the percentage of black people who reoffend is the same as the percentage of white people who reoffend.

(B) Parity - False Positive rate
The false positive rates (probability of being classified at risk while being not at risk) are the same for the group of black people and white people.

(C) Parity - False Negative rate
The false negative rates (probability of being classified not at risk while being at risk) are the same for the group of black people and white people.
(A) According to Northpoint, **COMPAS is calibrated**. Among defendants who scored a seven on the COMPAS scale, 60 percent of white defendants reoffended, which is nearly identical to the 61 percent of black defendants who reoffended.

(B) According to ProPublica, **COMPAS does not satisfy parity** for false Positive rate. Among defendants who ultimately did not reoffend, blacks were more than twice as likely as whites to be classified as medium or high risk (42 percent vs. 22 percent).

(C) Parity - False Negative rate

**Theorem:** assume that reoffending cannot be exactly predicted via the input features (life is always a bit random), then there is no algorithm that satisfies (A), (B), (C).

Washington Post: A computer program used for bail and sentencing decisions was labeled biased against blacks. It’s actually not that clear.
Ethical issues - Bias

- Car crash tests lead to unfair vehicles.
  → Debia**s the data**: collect more, better quality, better representativity.

- Correctional Offender Management Profiling for Alternative Sanctions (Compas) used in the US.
  → Debias the algorithm: twist predictions to annihilate one bias.

- Social Credit System / DeepNude
  → Impact on society: do we want these algorithms in our life?
1 Introduction to AI
- No data project without data
- Data science, Machine Learning, Artificial Intelligence... Which term should we use?
- Different applications of AI

2 How does Machine Learning work?

3 Application of AI in health
- Radiology
  - Chest X-ray
  - Liver lesion segmentation
- Genomics
  - Gene network
  - Toxicogenetics
- Medical

4 Perspective and issues
- Limitations of data projects
- Data Project Organization
- Unveiling the mystery of Deep Learning
Wooclap: Order the following different steps of a data project.

- Data collection
- Model evaluation
- Predictive modeling
- Continuous Optimization
- Solution Deployment
- Data Wrangling (gathering data in a usable format)
- Business understanding
- Testing / validation

You can also mention how each step interacts with the others.
Data Project Framework

Figure: http://www.anovaanalytics.com/data-science-consulting/
Data Science in 1 Slide

Figure: Source: Sz. Pafka

CRISP-DM
- Adapted by Szilard Pafka.
Data Scientists and Challenges

Data Scientist

- Mix of various skills.
- Hard to be an expert of everything!
Wooclap: Assign the following job names to the job descriptions below.

Data engineer, business analyst, statistician, data and analytics manager, data scientist, data architect, data analyst.
Different occupations in a data project

Several Profiles

- Several kind of problem / several kind of tools
- Much more variety than this...
- Importance of balanced teams.
Different training and different occupations

Introduction to AI
- No data project without data
- Data science, Machine Learning, Artificial Intelligence... Which term should we use?
- Different applications of AI

How does Machine Learning work?

Application of AI in health
- Radiology
  - Chest X-ray
  - Liver lesion segmentation
- Genomics
  - Gene network
  - Toxicogenetics
- Medical

Perspective and issues
- Limitations of data projects
- Data Project Organization
- Unveiling the mystery of Deep Learning
**Number Recognition**

- **Data:** Annotated database of images (each image is represented by a vector of $28 \times 28 = 784$ pixel intensities)
- **Input:** Image
- **Output:** Corresponding number
Fundamental elements

Convolution
Fundamental elements

Convolution

Max-Pooling
Convolutional neural network

Image

28 x 28

Convolution
padding = 1, kernel = 3x3, stride = 1
+ ReLU

32 x 28 x 28

Max pooling
Kernal = 2x2, Stride = 2

32 x 14 x 14

Convolution
padding = 1, kernel = 3x3, stride = 1
+ ReLU

64 x 14 x 14

Max pooling
Kernal = 2x2, Stride = 2

64 x 7 x 7

Flatten

3136 x 128

128 x 10

0

1

9
The 82 patterns misclassified by LeNet5. Below each image is displayed the correct answer (left) and the prediction (right). These errors are mostly caused by genuinely ambiguous patterns, or by digits written in a style that are under represented in the training set.
Other generic applications of CNN


[Krizhevsky 2012]

[Image: More images of horse, dog, and people with bounding boxes indicating their locations.]

[Faster R-CNN - Ren 2015]

[Image: A scene with a vehicle and people with their bounding boxes.]

[NVIDIA dev blog]
Stephen Hawking BBC, Dec 2 2014

The development of full artificial intelligence could spell the end of the human race. We cannot quite know what will happen if a machine exceeds our own intelligence, so we can't know if we'll be infinitely helped by it, or ignored by it and sidelined, or conceivably destroyed by it.
Take-home messages

- No data projects without data
  - A lot of data are available in the world
  - Difficulty of gathering the relevant ones and cleaning them (70% of the data project)
  - Environmental/Technical point of view: collecting/creating data is expensive for the planet (and the company)

- Different terms used in a data project
  - Keep in mind that AI has nothing to do with intelligence.
  - AI does not mimic human reasoning
Take-home messages

- No data projects without data
  - A lot of data are available in the world
  - Difficulty of gathering the relevant ones and cleaning them (70% of the data project)
  - Environmental/Technical point of view: collecting/creating data is expensive for the planet (and the company)

- Different terms used in a data project
  - Keep in mind that AI has nothing to do with intelligence.
  - AI does not mimic human reasoning

- How does Machine learning works?
  - Machine learning requires data to detect and learn patterns in the data.
  - Different tasks can be solved depending on the data (supervised, unsupervised, images, texts...)
  - Different tasks cannot be solved with ML notably if relevant information are not inside the collected data
  - Specific questions require specific data
Limitations of ML

- **Data may be biased** because our world is, and data are nothing but a reflection of it.
- Detecting and removing these biases is tricky but very important if individuals are impacted by the ML solution.
- **ML may have trouble to adapt in temporal changes.** ML has a tendency to reproduce the past.
Take-home message II

- Limitations of ML
  - Data may be biased because our world is, and data are nothing but a reflection of it.
  - Detecting and removing these biases is tricky but very important if individuals are impacted by the ML solution.
  - ML may have trouble to adapt in temporal changes. ML has a tendency to reproduce the past.

- Data cycle and Data jobs
  - A data cycle is composed of iterations, nothing is ever over.
  - Business analysis is very important through the cycle
  - Many different actors are involved in a data project
  - Good communication is required!
Limitations of ML

- **Data may be biased** because our world is, and data are nothing but a reflection of it.
- Detecting and removing these biases is tricky but very important if individuals are impacted by the ML solution.
- **ML may have trouble to adapt in temporal changes.** ML has a tendency to reproduce the past.

Data cycle and Data jobs

- A data cycle is composed of iterations, **nothing is ever over.**
- Business analysis is very important through the cycle
- Many different actors are involved in a data project
- **Good communication is required!**

Data Science is evolving constantly.

- New opportunities appear
- New challenges are detected
- Need for adaptability
Thank you!


