What is Machine Learning?

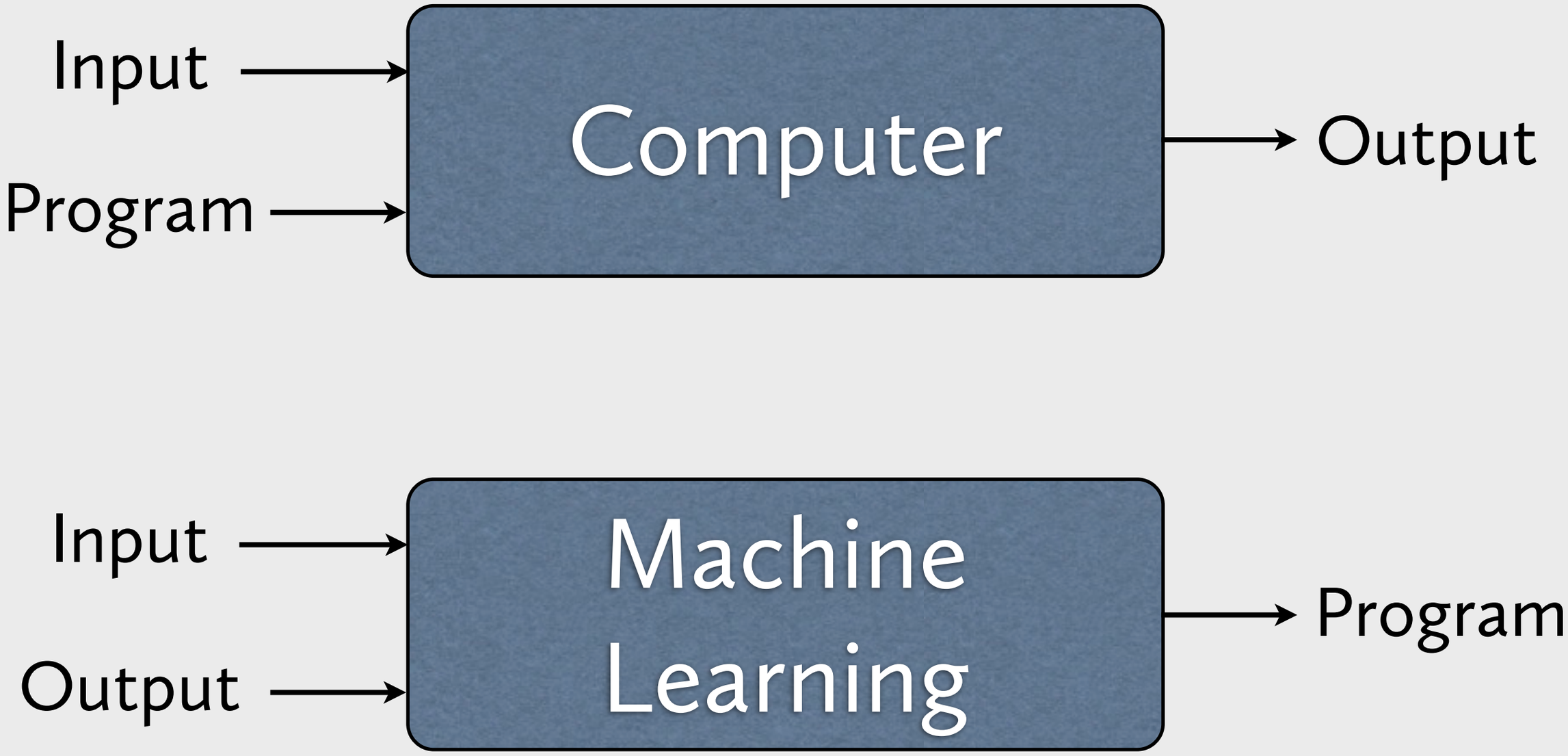
(and should I care?)

Charles-Pierre Astolfi, 4A0, cpa@crans.org
wiki.crans.org/CharlesPierre

« Field of study that gives the computer the ability to learn without being explicitly programmed. »

— Arthur Samuel (1959)

It's simple, really



Did you mean...

- Machine learning (ML)
- Data science
- Data mining
- Big data
- Data analytics
- Statistics
- Artificial Intelligence

3 simple questions

- What's ML?
- What do people do with ML?
- Is the law something for boring assholes who want to impede innovation?

What is Machine Learning?

Pop quiz!

Pop quiz!

Machine Learning...

Pop quiz!

Machine Learning...

- aim is to model the brain.

Pop quiz!

Machine Learning...

- aim is to model the brain.
- tries to find patterns and correlations.

Pop quiz!

Machine Learning...

- aim is to model the brain.
- tries to find patterns and correlations.
- is not used in industry yet.

Pop quiz!

Machine Learning...

- aim is to model the brain.
- tries to find patterns and correlations.
- is not used in industry yet.
- is a black art more than a science.

Pop quiz!

Machine Learning...

- aim is to model the brain.
- tries to find patterns and correlations.
- is not used in industry yet.
- is a black art more than a science.
- has no ethical ramifications (yet).

Pop quiz!

Machine Learning...

- aim is to model the brain.
- tries to find patterns and correlations.
- is not used in industry yet.
- is a black art more than a science.
- has no ethical ramifications (yet).
- can help you find your life partner.

Pop quiz!

Machine Learning...

- aim is to model the brain.
- tries to find patterns and correlations.
- is not used in industry yet.
- is a black art more than a science.
- has no ethical ramifications (yet).
- can help you find your life partner.
- saves 5 millions lives a year.

Pop quiz!

Machine Learning...

- aim is to model the brain.
- tries to find patterns and correlations.
- is not used in industry yet.
- is a black art more than a science.
- has no ethical ramifications (yet).
- can help you find your life partner.
- saves 5 millions lives a year.
- makes 40 billions dollars a year.

What is Machine Learning?

~~Science~~ Black art with the goal:

- Classify data.
- Capture characteristics from empirical data.
- Generate data “in the style of” what has been seen.
- Learn to take decisions based on the past course of actions.

Classification (and ranking)

Clustering

Regression

Reinforcement learning

Classification

(supervised learning)

Input	Output
Age + Year of operation + Number of axillary nodes detected	0 if the patient died within 5 years 1 if the patient survived 5 years or longer

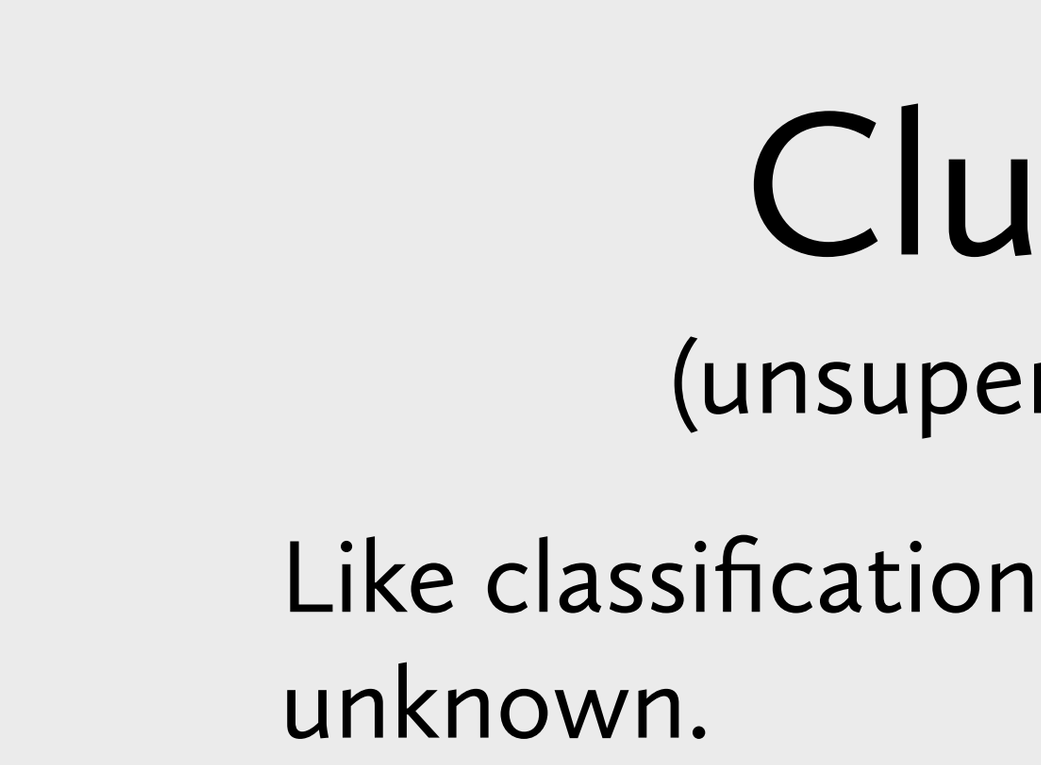
Machine learning: saving boobs without even touching them.

Clustering

(unsupervised learning)

Like classification, but the labels are unknown.

Input



Output

Clustering

State of the art:

- Andrew Ng & al. trained an unsupervised large-scale (16,000 cores) neural network

- This is a neuron that detects faces

- Precision: 19% on 22000 classes.

Regression

- Like classification, but one has to predict a value rather than a label.
- E.g.: given some statistics about crime in a neighborhood, predict the number of crimes next year.
- E.g.: Predict the temperature tomorrow

Reinforcement learning

- Predictions are decisions!
- Demo: Pendulum swing up learning
- There's this guy, Pavlov...
- Kids!

Let's recap

If I'm given...	My predictions are...	Then I'm doing...
Vectors	(Known) finite set of labels	Classification
	(Unknown) finite set of labels	Clustering
	Real value	Regression
Past events	Actions	Reinforcement learning

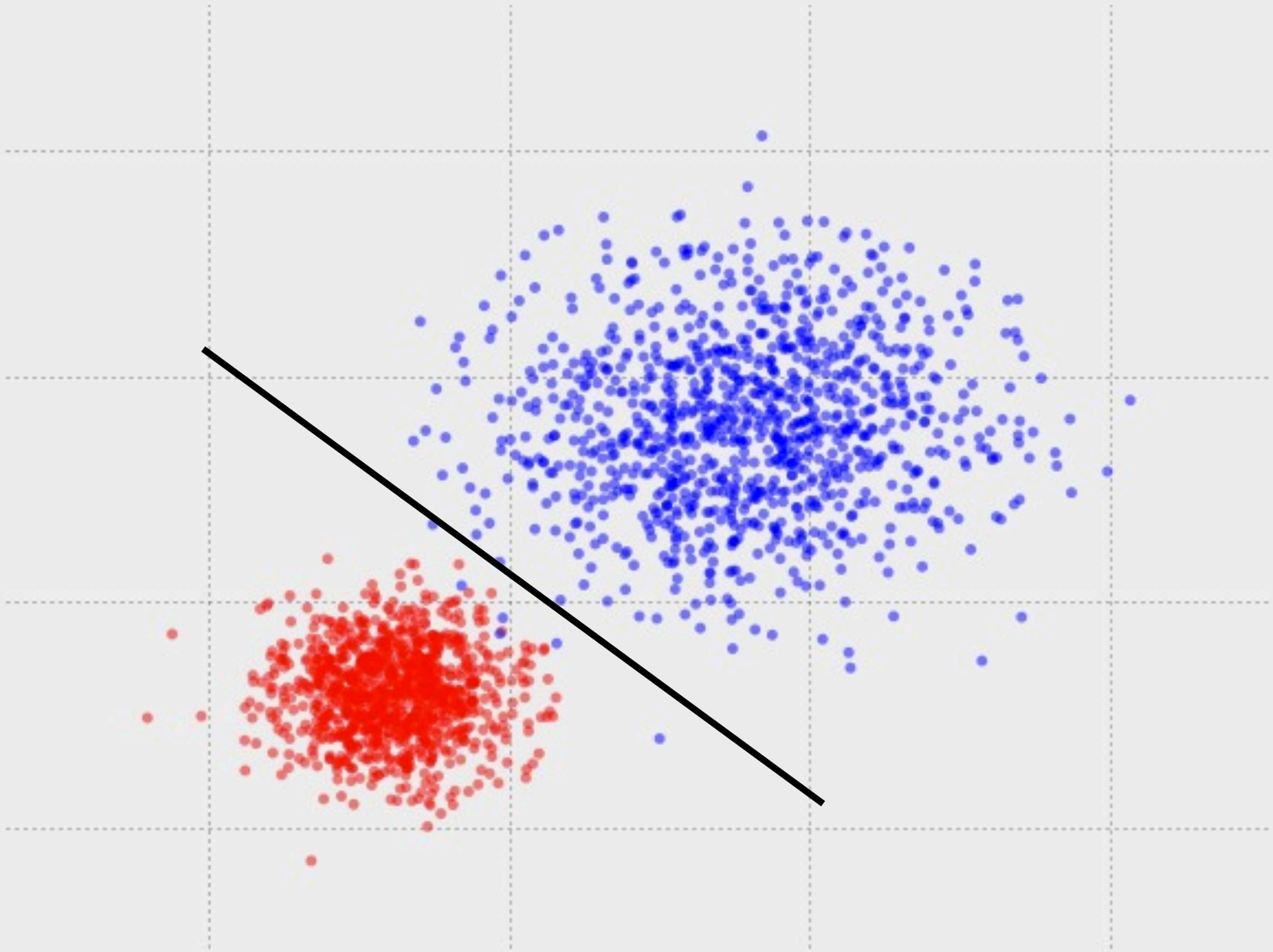
Ridge

Given $\mathbf{X} \in \mathbb{R}^n \times \mathbb{R}^m$ (training data)
and $Y \in \mathbb{R}^n$ (outcomes),

Find w that satisfies:

$$\min_w \sum_{i=1}^{n} (Xw - Y)_i^2 + \alpha \sum_{i=1}^{m} w_i^2$$

Ridge results



ML drawbacks

- No silver bullet. (SVM? Ridge? Lasso? Random Forests? Deep learning?)
- NP-Hardness is often an issue.
- Even for heuristics, complexity is usually more than linear.
- It's hard to get clean data.
- It's hard to select the right features.
- It's often hard to understand your predictive model.
- It's next to impossible to ensure statistical significance.
- There's this thing we call the "Curse of dimensionality"...


What do
people do
with ML?


HETEMEEL.COM
I WANT YOU
TO REVEAL YOUR PERSONAL
LIFE AND CLICK ON ADS!

google.com/ads/preferences

Below you can edit the interests and inferred demographics that Google has associated with your cookie:

Category	
Arts & Entertainment - Events & Listings - Concerts & Music Festivals	Remove
Arts & Entertainment - Events & Listings - Ticket Sales	Remove
Arts & Entertainment - Movies - Science Fiction & Fantasy Films	Remove
Business & Industrial - Transportation & Logistics - Urban Transport	Remove
Computers & Electronics - Consumer Electronics - ... - Handheld Game Consoles	Remove
Hobbies & Leisure - Outdoors	Remove
Internet & Telecom - ... - Search Engine Optimization & Marketing	Remove
Law & Government - Government - Legislative Branch	Remove
News - Business News	Remove
Travel - Bus & Rail	Remove
Demographics - Gender - Male	Remove

(inferred from your behavior on the web)

ML Applications

- Finding conservation equations for the double pendulum (a chaotic dynamic system!)
- Web search
- Providing love and sex (meetic, eharmony and okcupid hire a lot of ML people!)
- Discriminate gender on Twitter
  Most common words for females:
  "!, love, :), haha, so"
  For males: "Goog, googl, google, http"
- Apple's Siri, Google Now
- iPhone's auto correct (I don't know for android)

ML Applications (cont'd)

- Automated mining: Rio Tinto and Nicta
- Web search: Google
- Ad selection: Google, Facebook
- Medical research
- Machine Vision: Driverless cars, animal census via drones, face detection
- Speech Recognition: Help desks, banking.
- Killer drones (in development)
- Intelligence agencies!
- Snail mail: address recognition
- Sentiment mining: who's thinking what?
- Recommender systems: Netflix (1M$ prize), Air France
- Automated translation
- Rare event detection (people fighting on CCTV)
- Stock prediction
- Logistics
- Energy consumption prediction
- Weather forecasting
- Signal analysis (RADARs)
- Behavior analysis
- Understand abstract art
- Job finding
- Obama's camaign (2012)
- Antivirus / firewall
- Infinite Gangnam style
- Hospital logistics + Flight logistics by GE : 500kUSD
- Drug design
- Detect penises


Is it all legal?

« [Your credit card limit has been lowered because] other customers who have used their card at establishments where you recently shopped have a poor repayment history with American Express. »

— American Express (to Kevin Johnson, 2008)

It’s just a technology

It's just a technology

- That works at an unprecedented scale.

It’s just a technology

- That works at an unprecedented scale.
- That the general audience doesn’t know much about.

It’s just a technology

- That works at an unprecedented scale.
- That the general audience doesn’t know much about.
- That works with a media on which proving that something has been done is virtually impossible.

It’s just a technology

- That works at an unprecedented scale.
- That the general audience doesn’t know much about.
- That works with a media on which proving that something has been done is virtually impossible.
- For which accountability is not clearly defined.

It’s just a technology

- That works at an unprecedented scale.
- That the general audience doesn’t know much about.
- That works with a media on which proving that something has been done is virtually impossible.
- For which accountability is not clearly defined.
- Which changes data analysis economics entirely.

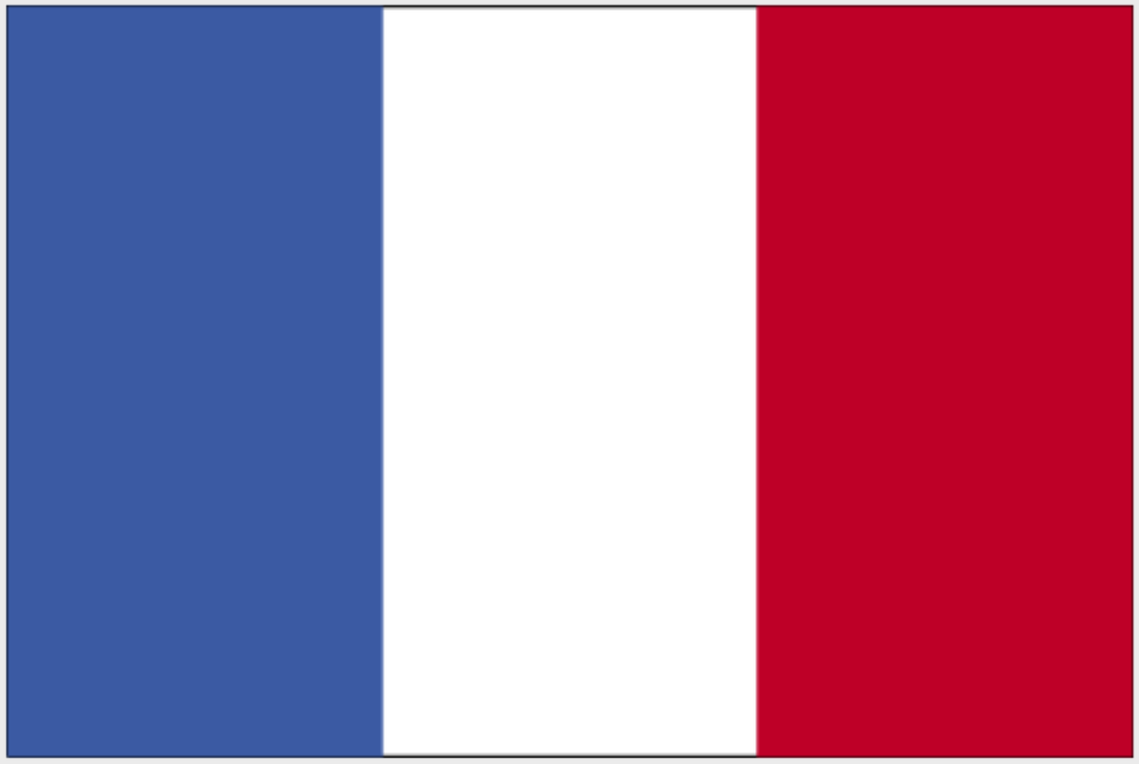
Some legal issues

- Eugenism!
  (My ML algorithm says it's very likely that my child will have such traits)
- Discrimination!
  (My ML algorithm says it's a bad idea to loan money to black people)
- Proof killer!
  (That's not me speaking on this record but a machine that learned to speak like me)
- Privacy on the internet!
  (Let's focus on that)

Legal

- In France, Loi Informatique et libertés (1978) roughly implies that:
  - No decision should rely upon an automatic system.
  - You can't do ML without users' consent if you hold Personally Identifiable Information (PII).
  - What can be collected is defined by the intended use.
  - Collection of PII is stricly supervised.
- In France, privacy is part of the law. (Art 9 du Code Civil : « Chacun a droit au respect de sa vie privée. »)
- More or less the same laws in all EU.

FUCK YEAH FRANCE!



You got my back!

NOPE.

What is PII?

This is PII.

- First and last name
- Address
- Email
- Phone number
- Date and place of birth
- SSN
- Credit card number
- Photo
- DNA
- Fingerprints
- License plate

Is this PII?

- How I walk.
- How I speak.
- How I write.
- Whom I'm friends with.
- What I like.
- My browser's cookies.
- The kind of music I listen to.
- The movies I see.
- My browser's version.
- The pages I've liked.
- My IP address. (CNIL says yes, Cour d'appel de Paris says no)

My *opinion* is that ML will turn all of this into PII.

And that's also the EU's opinion

"[The definitions] leave to interpretation whether [personal data] includes information that can be used to identify a person with high probability but not with certainty..."

—*EU report on the Right to be forgotten*

So...

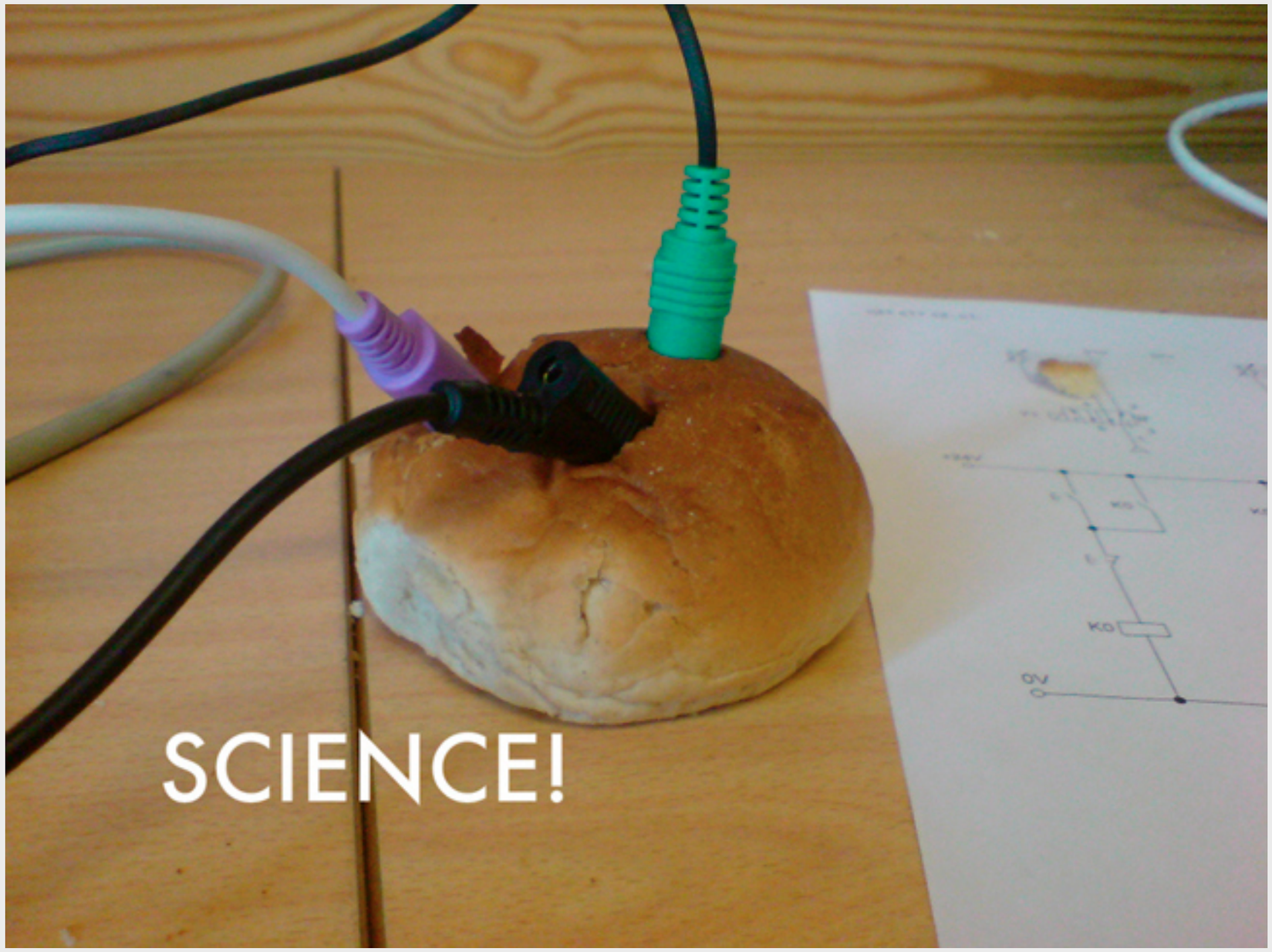
- Sensible regulation and laws about data storage, retrieval, (simple) analysis...
- But not ready for the firepower ML brings (see ENISA's reports)
- Economic incentive to collect and use data on a large scale (it's cheap!)

I think we're done here.

Questions?

(and thank you!)

Cats by Maccio Capatonda on flickr, Dilbert comic by Scott Adams



Science. It’s surprising.

Where do I start?

- Books
  - ML in action
  - Elements of statistical learning (theoretical!)
- Programming libraries
  - python with scikit learn (and its excellent tutorial)
  - R (and its libraries)
- Communities
  - reddit.com/r/machinelearning
  - quora.com
  - crossvalidated.com
  - kaggle.com