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.”

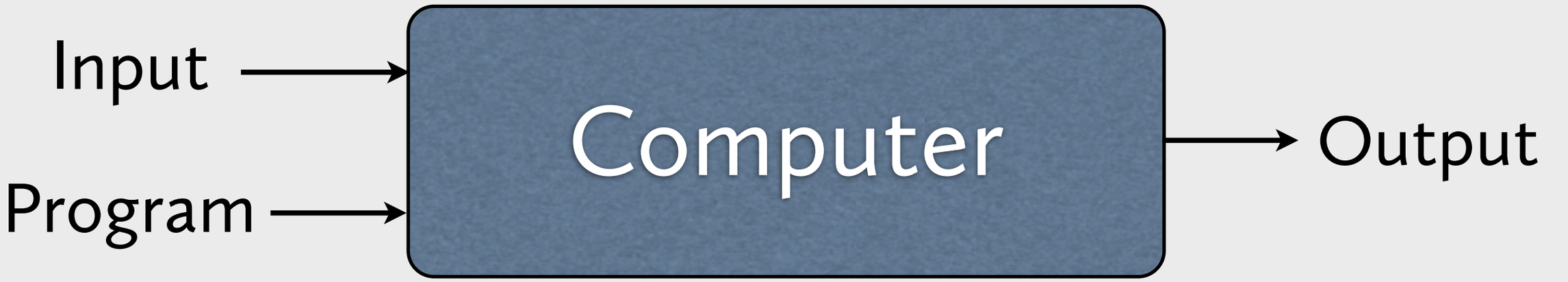
— *Arthur Samuel (1959)*

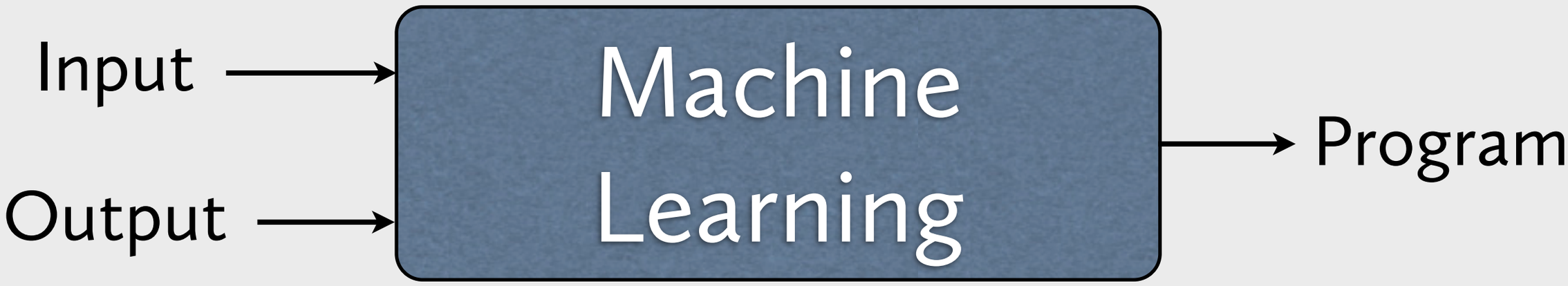
What's learning?

- A computer learns some task if its performance on this task improves with experience. (~Tom Mitchell, 1998)
- Finding a model that describes a given system only by observing it.
- A model = any relationship between the variables used to describe the system.

Two goals: make predictions and understand systems.

It's simple, really



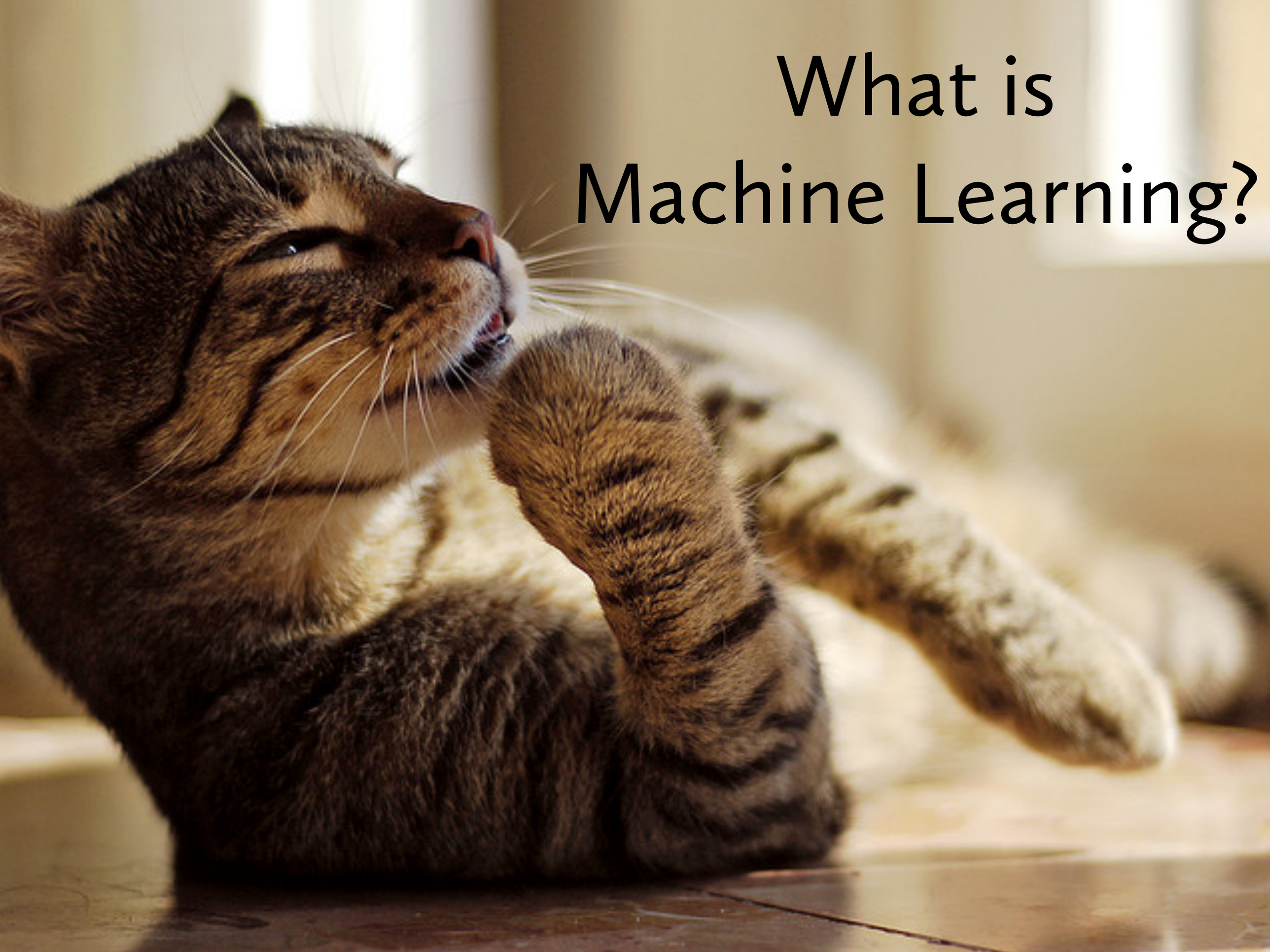


Did you mean...

- 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?

Machine Learning

tries to find patterns and correlations in data

That's, like, the definition.

Machine Learning

is not used in industry yet

There are mines operating without human intervention.

Machine Learning

goal is to simulate

the brain

No, check out AI Winter on wikipedia!

Machine Learning

is a black art more than a science

We have no idea what we're doing.

Machine Learning

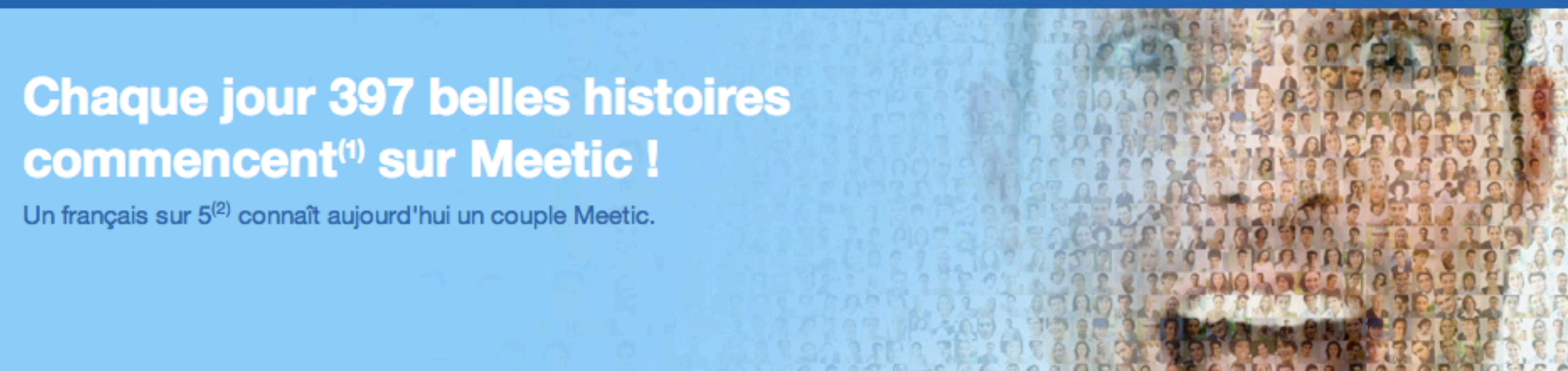
has no ethical ramifications yet

LOL

Machine Learning

can help you find

your life partner



Machine Learning

saves 5 million

lives a year

Machine Learning

makes 40 billion

USD a year

Google

Classification

(supervised learning)

Input	Output (labels)
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

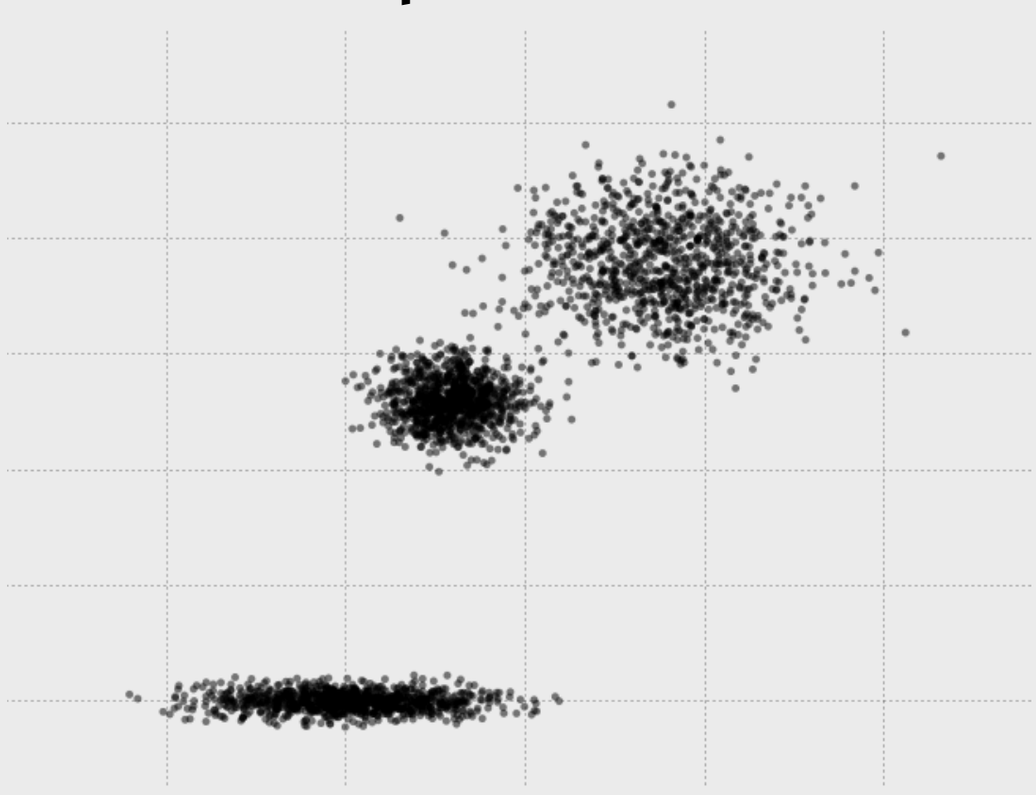
Machine learning: saving boobs without even touching them.

Clustering

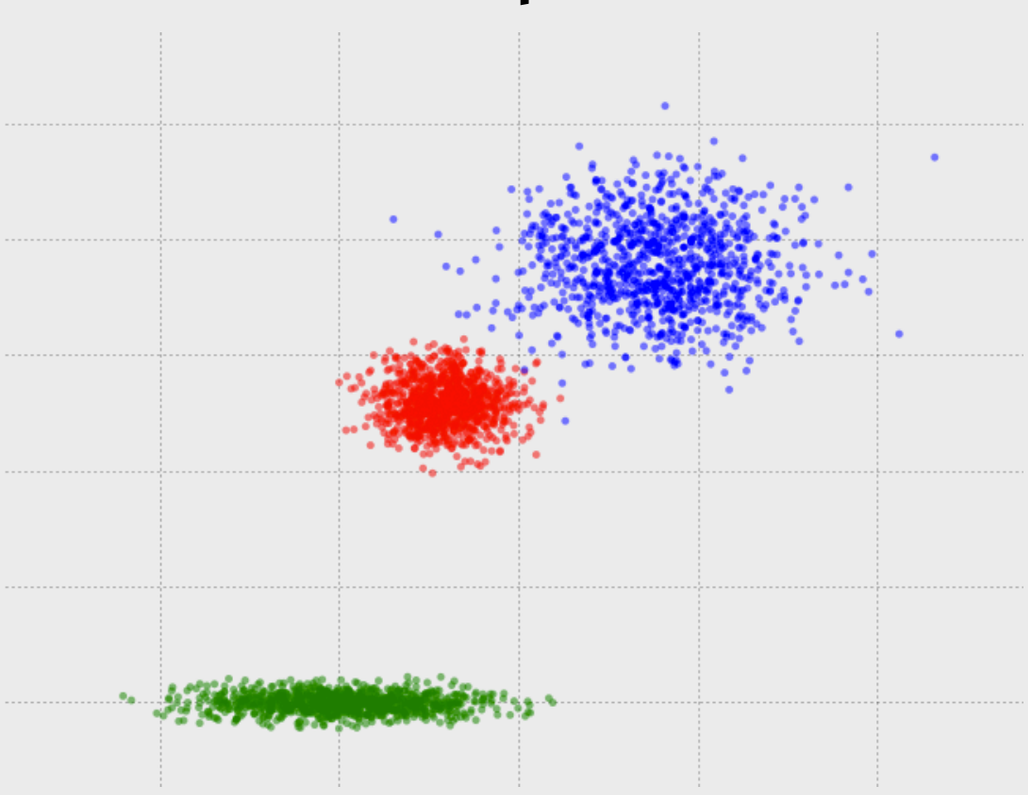
(unsupervised learning)

Like classification, but the labels are unknown.

Input



Output



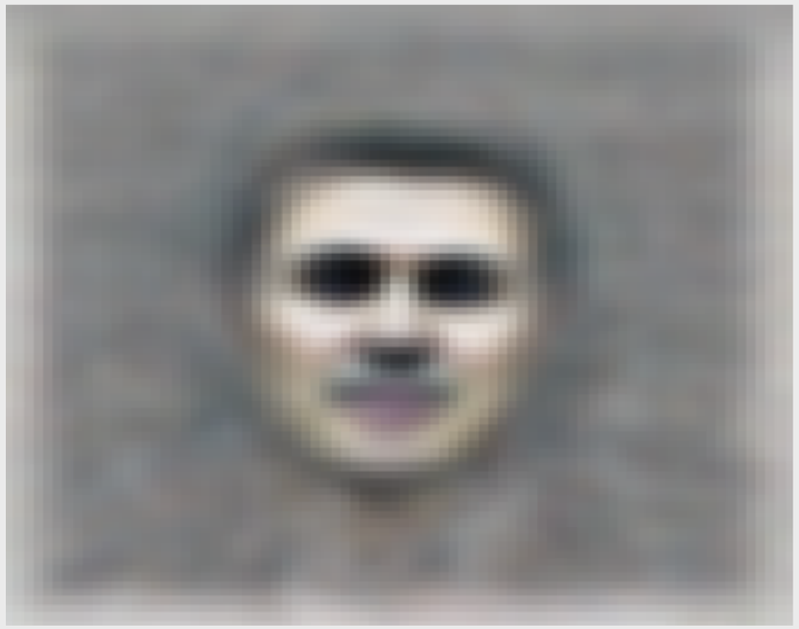
Clustering

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

You know it already: 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

When to use ML?

Machine learning is useful when:

- Humans don't know how to do (navigating on Mars)
- Humans don't know how they do (speech recognition)
- Humans are too slow (routing on a network)
- Humans can't cope with system size (weather forecasts)
- Humans are too expensive (drones, Foxconn)

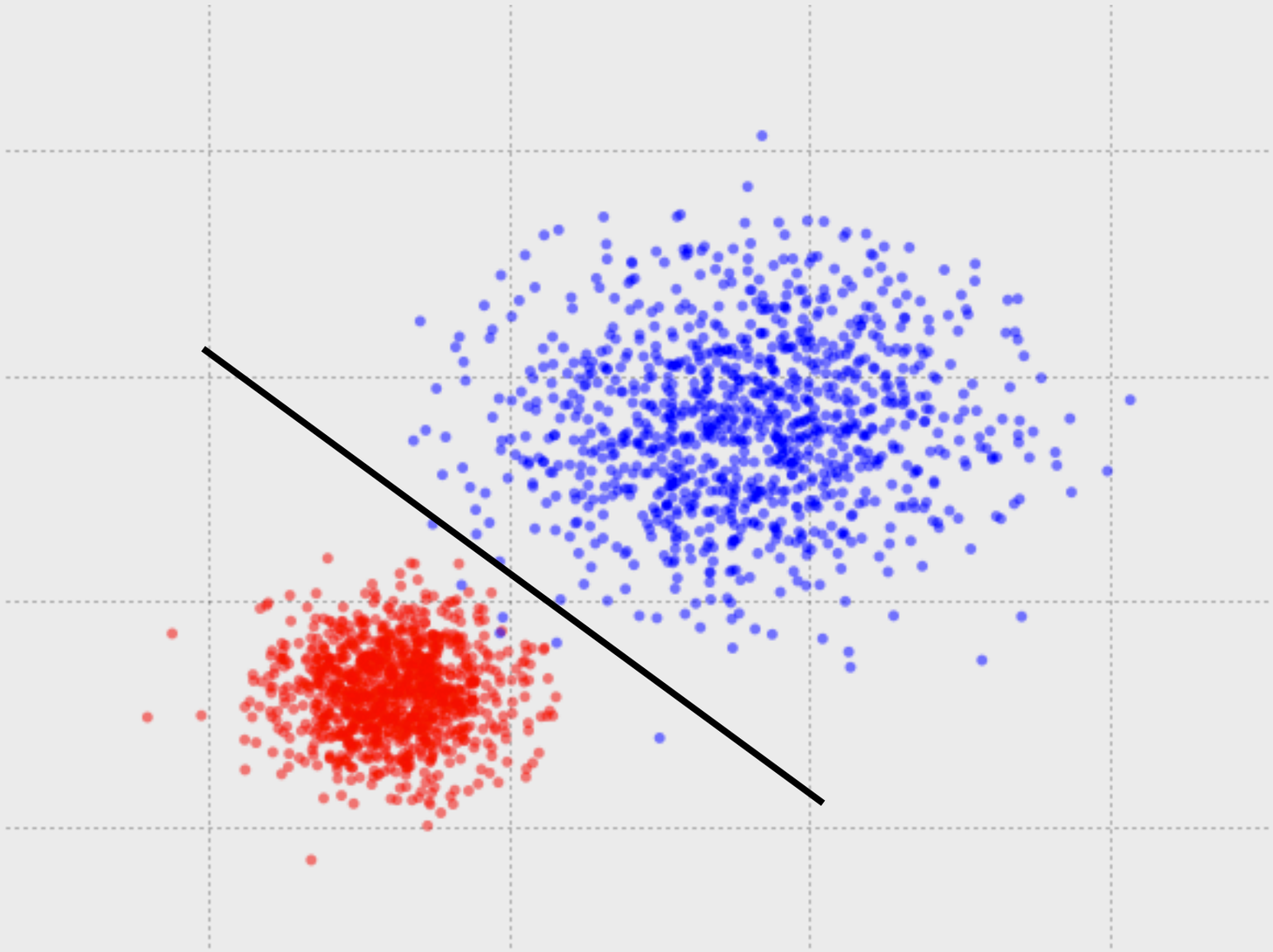
Ridge

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?

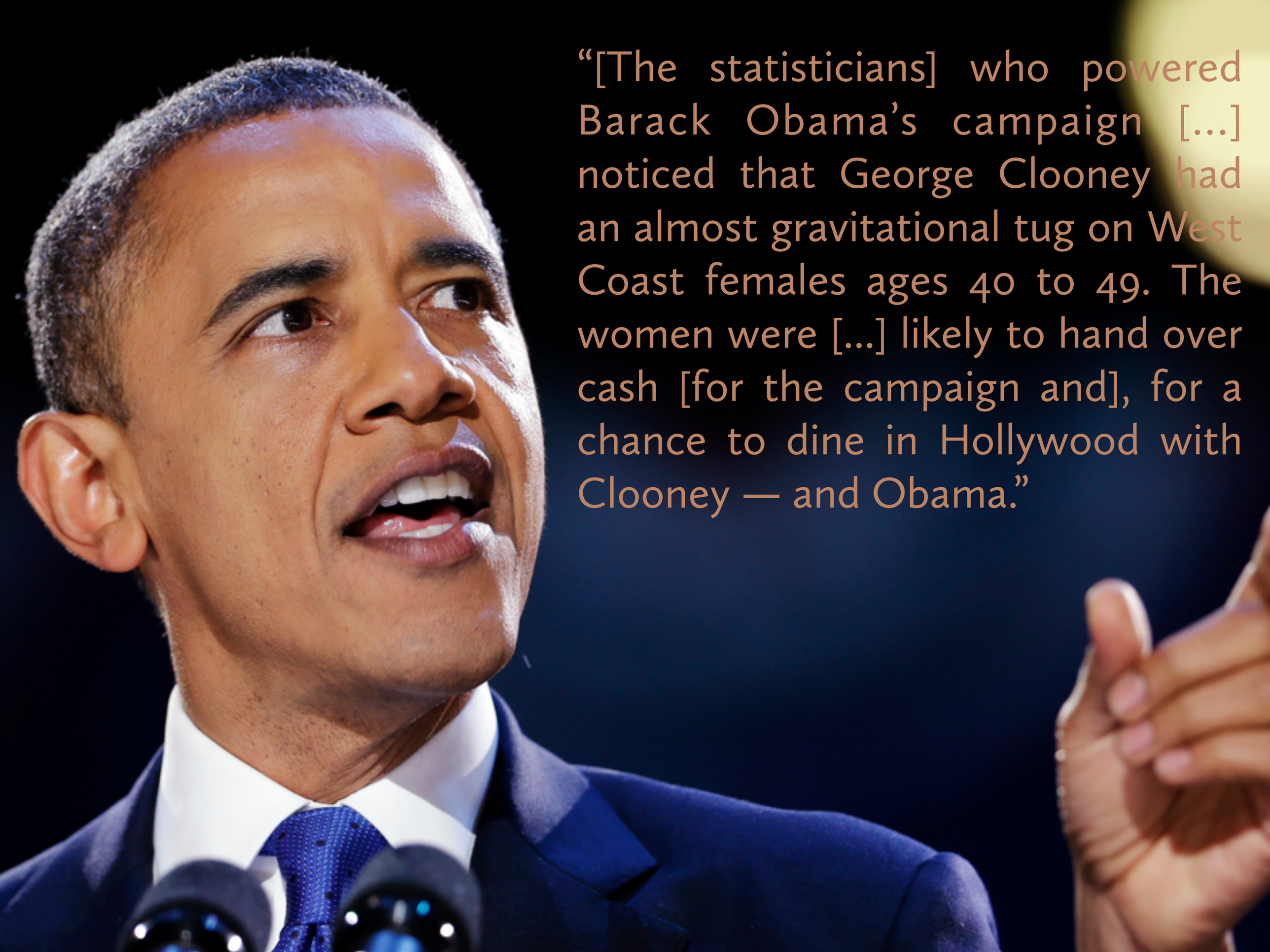

HETEMEEL.COM
I WANT YOU
TO REVEAL YOUR PERSONAL
LIFE AND CLICK ON ADS!

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)



"[The statisticians] who powered Barack Obama's campaign [...] noticed that George Clooney had an almost gravitational tug on West Coast females ages 40 to 49. The women were [...] likely to hand over cash [for the campaign and], for a chance to dine in Hollywood with Clooney — and Obama."

Funfact

(14 dec. 2012)

There are 800 000 books available on Amazon...

...that will only be written and printed after you have purchased it.

Subjects includes financial reports, crosswords, rare diseases...

They are generated by an algorithm that processes data available on the internet and rewrites it, as to avoid plagiarism.

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

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)
- Make US army copters fly
- 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)*

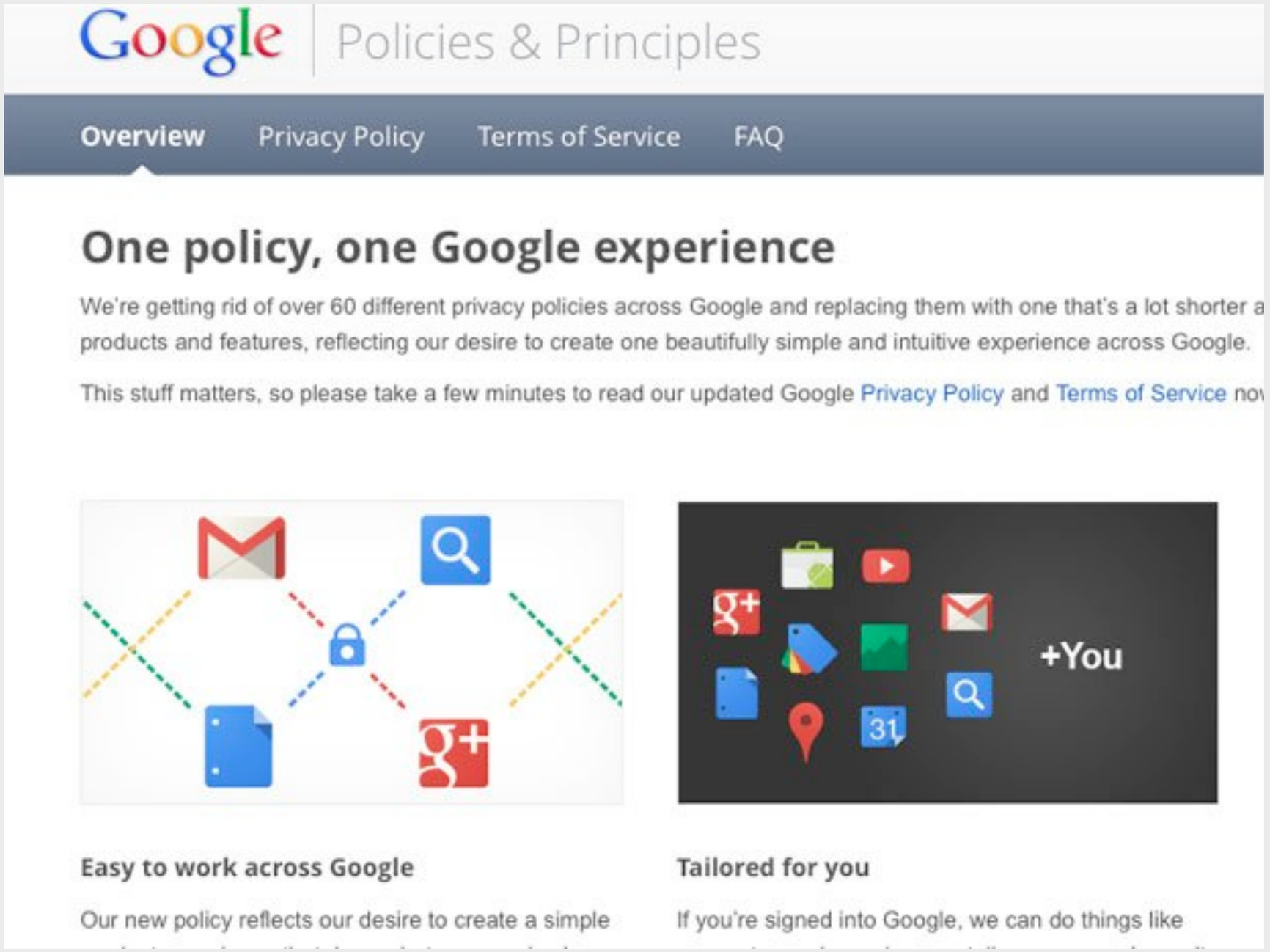
It’s just a technology

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

Some legal issues

- Eugenism!
  (My ML algorithm says it's very likely for me to have a ginger)
- 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!

Do you remember agreeing to this on March 2012?



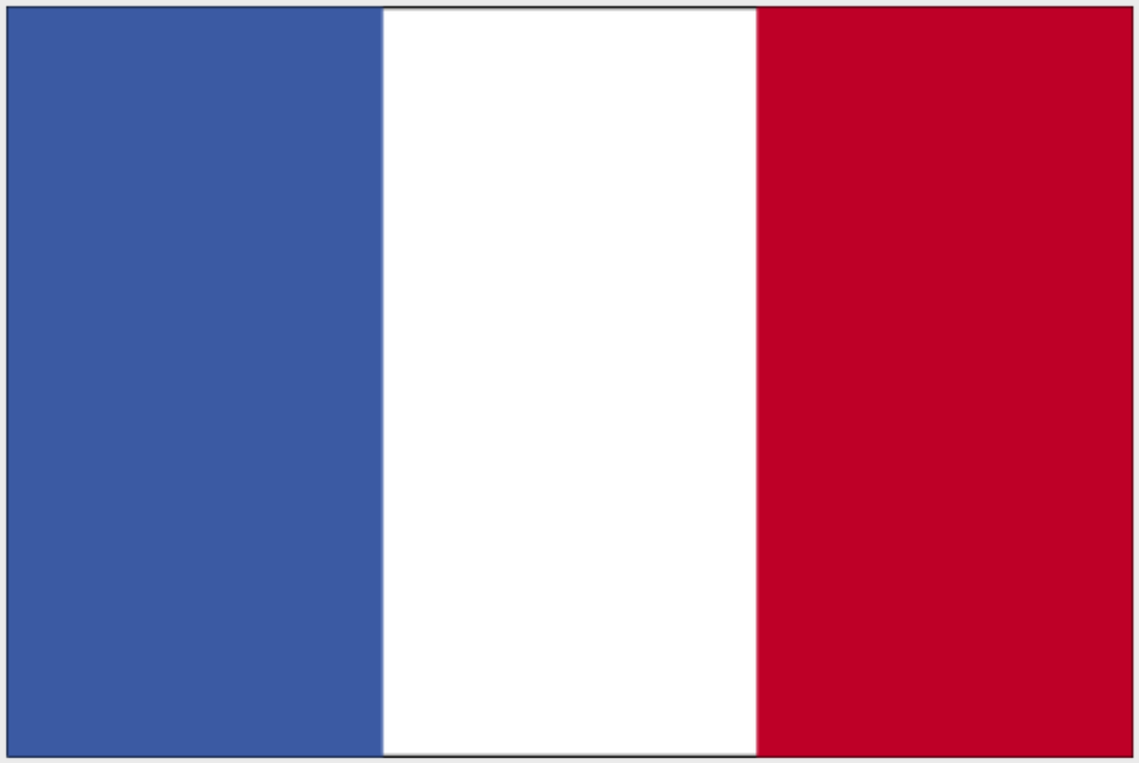
CNIL's (EU's) opinion

- Google's new privacy policy: incomplete information and uncontrolled combination of data across services.
- Google provides insufficient information to its users on its personal data processing operations.
- Google should therefore modify its practices when combining data across services for these purposes.
- Google does not provide retention periods.
- (a lot more actually)
- This has been anounced in october and nothing has changed.

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.
- My zip code
- The kind of music I listen to.
- The movies I saw.
- My browser's version.
- The pages I've liked.
- My IP address. (CNIL and CJUE says yes, Cour d'appel de Paris says no)

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

The EU opinion is that ML will turn all of this into PII.

"[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)

All in all . . .

It's the future. Deal with it.

It's just a technology, with a very broad scope.

It brings issues that we, as a society, will have to spot, understand and sort out.

I think we're done here.

Questions?

(and thank you!)

Cats by Maccio Capatonda on flickr, Dilbert comic by Scott Adams, Chickens by Doug Savage.
Couldn't find sources for other pictures.

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
- A must read
- CNIL's report « Vie privée à l'horizon 2020 »