Articulated Intelligence

What can Machine Learning do for (and to) the Social Sciences?

Étienne Ollion[1]

[1]CNRS (CREST) et École polytechnique

SASE Meetings, July 2020

1 Introduction: The Wave and the Currents

2 Preliminary Remarks about Machine Learning
- An Old Endeavor
- A Series of Booms and Bust
- What is Machine Learning?

3 Differences in Practice: Two Approaches to Data
- Supervised Setting
- Unsupervised Setting

4 Learning to Analyze, or Learning to Produce?
- Learning to Produce
- History of Political Journalism

5 Conclusion: AI, Why should we care?

Today's talk

Should we care about AI as (social) scientists?

- A Journey into a few Machine Learning Experiments
- Machine Learning for Quantitative Methods, and Beyond
- A Non-technical Approach

Thanks to talented collaborators:

- Julien Boelaert (Univ. of Lille)
- Salomé Do (Sciences Po-ENS)

Introduction : The Wave and the Currents

Introduction : The Wave and the Currents

P. Domingos, The Master Algorithm, 2015, p.13

Machine learning follows the same procedure [as classic statistics] of generating, testing, and discarding or refining hypotheses. But while a scientist may spend her whole life [doing so], machine learning can do it in a fraction of second

Machine learning is **"the scientific method on steroids"**. It is thus **"no surprise that it is revolutionizing science"**

Introduction : The Wave and the Currents

Introduction : The Wave and the Currents

- But many criticisms (since Breiman, 2001)
  - AI as a "black box" (Burell, 2016, Biau, 2012)
  - What to do with prediction in science?
  - Uncertain results

- External Criticisms
  - A restrictive definition of intelligence
  - The human in the machine (Casilli, 2019)

Introduction : The Wave and the Currents

- But many criticisms
  - AI as a "black box"
  - Uncertain optimality
  - What to do with prediction in science?
- External Criticisms
  - A restrictive definition of intelligence
  - The human in the machine

Figure: Doré's Deluge, 1865

Introduction : The Wave and the Currents

Today

- Presentation of relative strengths and weaknesses

- Argument: No great replacement
  - Relative incommensurability of these methods (question,outputs)
  - Main opportunity: Machine Learning to Produce (not Analyze) data

1 Introduction: The Wave and the Currents

2 Preliminary Remarks about Machine Learning
- An Old Endeavor
- A Series of Booms and Bust
- What is Machine Learning?

3 Differences in Practice: Two Approaches to Data
- Supervised Setting
- Unsupervised Setting

4 Learning to Analyze, or Learning to Produce?
- Learning to Produce
- History of Political Journalism

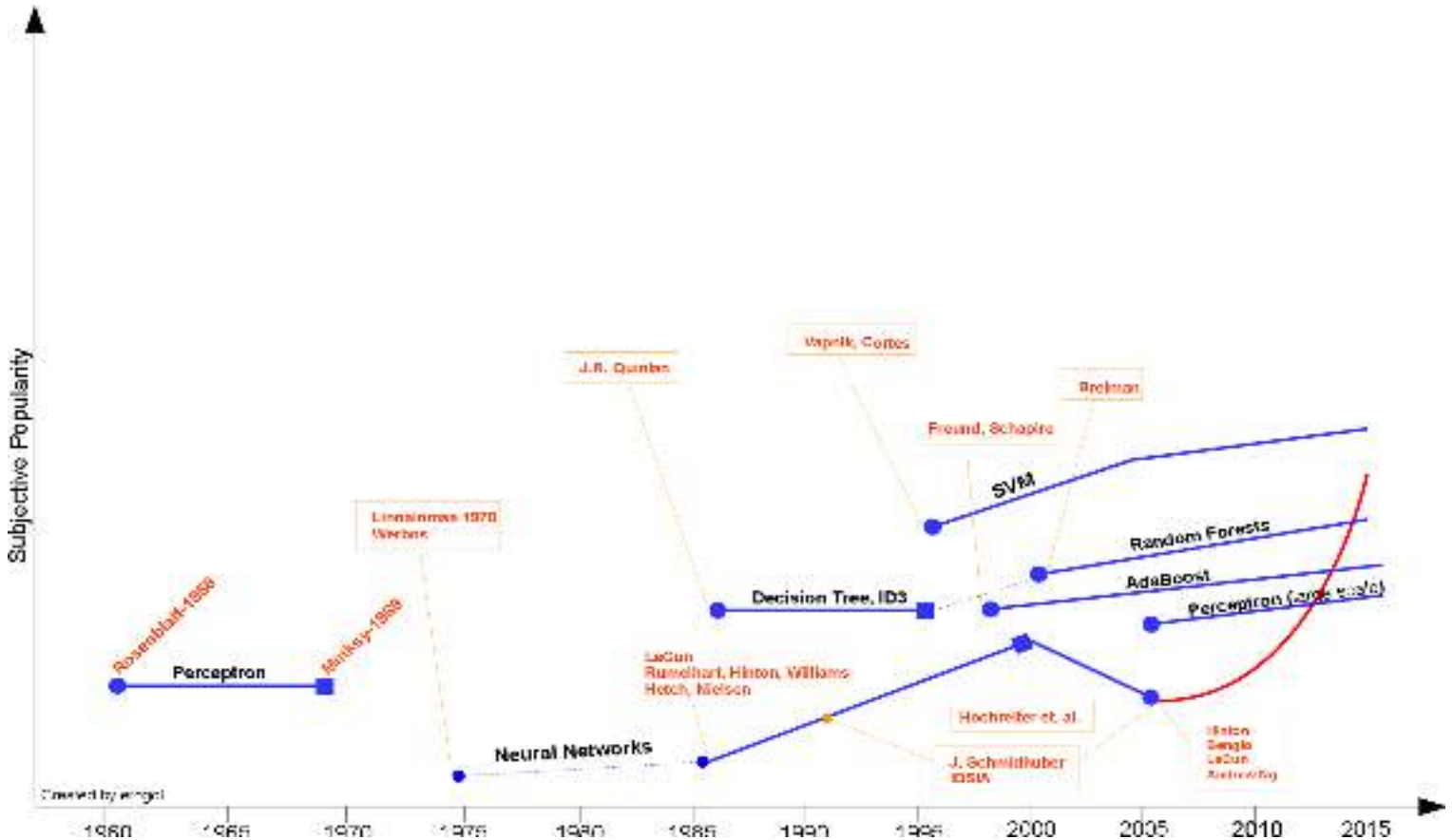
5 Conclusion: AI, Why should we care?

An Old Endeavor

- Starts after WWII
- A landmark seminar: McCarthy and the 1956 Dartmouth Summer Research Project on AI
  - "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".
- Many disconnected initiatives, a flurry of names

A Series of Booms and Bust

- "AI Summers" and "AI Winters"
  - An uneven development McCorduck, 2004; Cardon, 2018
  - Reasons for current success: computer power, masses of data...and lots of research money



What is Machine Learning?

attributed to Thomas, 1952

[Artificial intelligence] is the field of study that gives computers the ability to learn without being explicitly programmed

What is Machine Learning?

Past this point, many differences

- Connectionnist vs. Symbolic AI
- AI or Machine Learning?
- Countless applications, going in different directions
- Supervised vs. Unsupervised
  - Supervised $\sim$ Parametric Regression
  - Unsupervised $\sim$ Clustering, dimensionality reduction

1 Introduction: The Wave and the Currents

2 Preliminary Remarks about Machine Learning
- An Old Endeavor
- A Series of Booms and Bust
- What is Machine Learning?

3 Differences in Practice: Two Approaches to Data
- Supervised Setting
- Unsupervised Setting

4 Learning to Analyze, or Learning to Produce?
- Learning to Produce
- History of Political Journalism

5 Conclusion: AI, Why should we care?

Supervised Setting: Two approaches to Data

Say you want to capture the link between a variable (wage) and a series of others (socio-demographic)

- "Classic" setting (parametric regression)
  - Building on hypotheses, previous knowledge...
  - the research **specifies the functional form**

$$W = \alpha_1 Age + \alpha_2 Exp + \alpha_3 Educ + \alpha_4 Sex + \alpha_5 Child + \alpha_6 Region + \alpha_7 Family + \alpha_8 Citiz + \alpha_9 Occup + \alpha_1 0 Unemploy + \mathcal{E}$$

Supervised Setting: Two Approaches to Data

Say you want to capture the link between a variable (wage) and a series of others (socio-demographic)

- "Classic" setting (parametric regression)
- "Machine Learning" setting
  - Provide data to the algorithm...
  - ... which will look itself for the best fit
  - The algorithm **specifies itself the functional form**

Promise: "Universal Approximation"

Supervised Setting: Two Approaches to Data

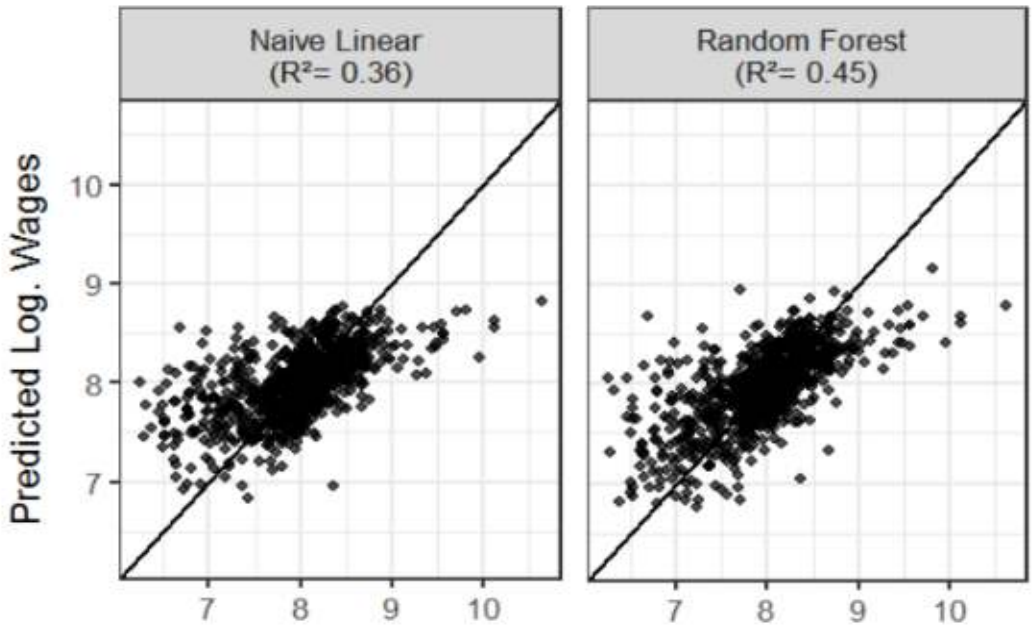
An empirical investigation in Sweden
Register-based country. 9M individuals, 300+ socio-demographic variables

Details in Boelaert and Ollion, The Great Regression, 2018

Supervised Setting: Two Approaches to Data

- On the classic question of wage (determinants well-known, see Mincer, 1974)
- 9 hand-selected variables

Supervised Setting: Two Approaches to Data



The ML algorithm significantly outperforms the parametric model

- Finds a better functional form
- Does it in a matter of seconds
- Without any knowledge of the data set

Supervised Setting: Two Approaches to Data

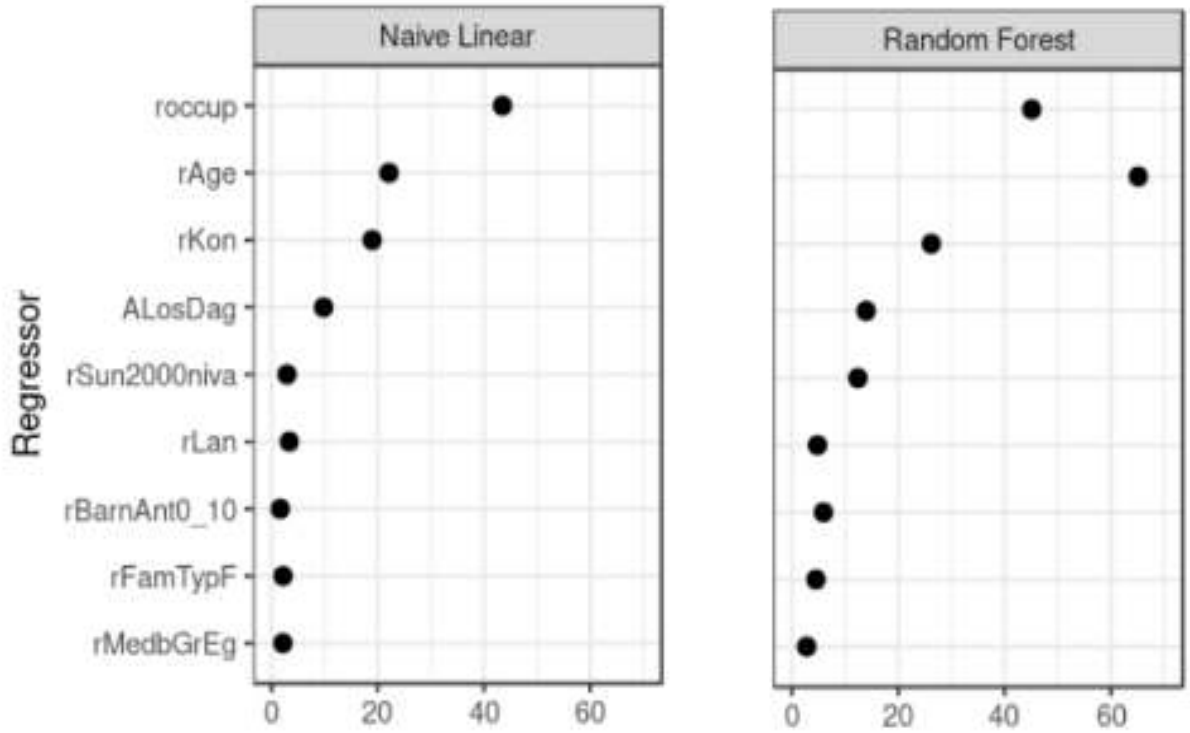
The Predictive Superiority of ML

- A quite classic result (see Salganik, 2020)

Supervised Setting: Two Approaches to Data

The Predictive Superiority of ML

- A quite classic result (see Salganik, 2020)
- More perks!
  - Going beyond prediction (variable importance)



Supervised Setting: Two Approaches to Data

- More perks!
  - Empirical investigation of the effect of the variables

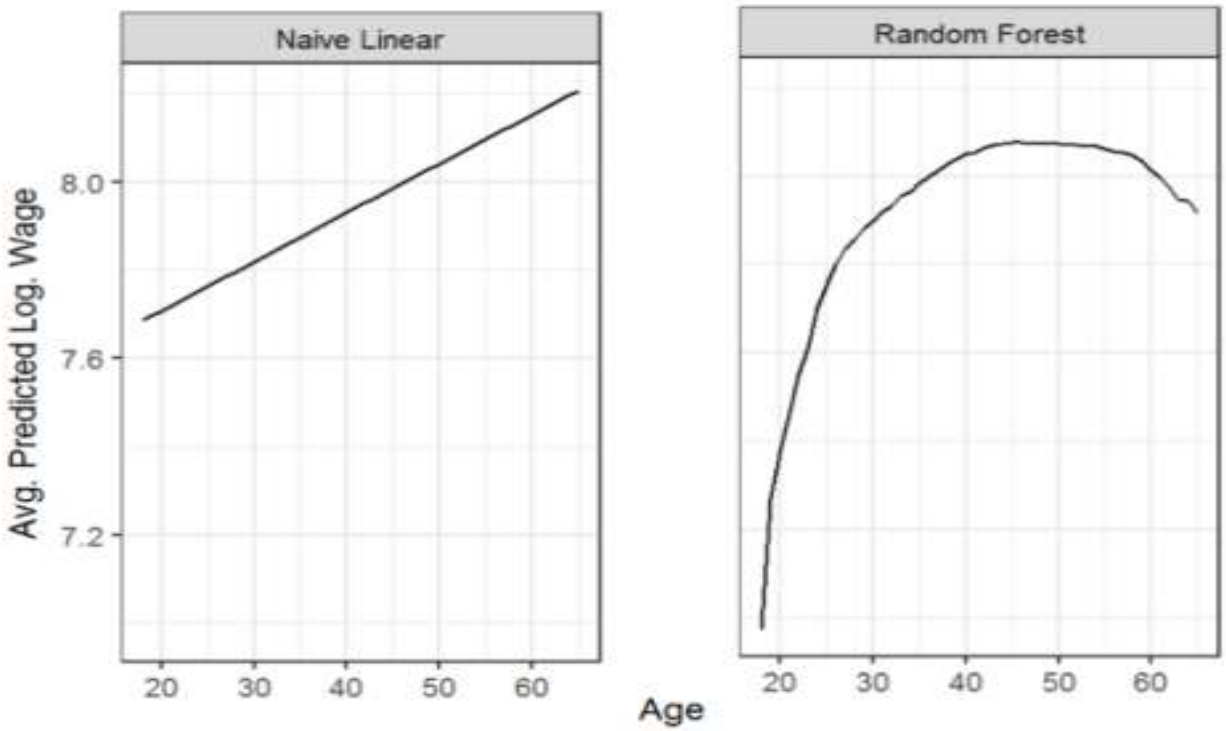


Figure: Partial dependence on Age

Supervised Setting: Two Approaches to Data

But one that does not go without issues

- Relative uncertainty about the results (no test, no demonstration)
- What about causality, the interpretation of coefficients?
- More complexity, less interpretability
- The problem of size

For more details, see (Boelaert, 2018)

Supervised Setting: Two Approaches to Data

Two ways of looking at data

- Demonstrative vs. Inductive
- Parcimonious vs. Complex
- Demonstrated vs. Empirically efficient

–> Two methods doing somewhat different tasks

Unsupervised Setting: Complexity vs. Interpretability

Unsupervised Machine Learning: Capturing Patterns in Data
Dimensionality Reduction

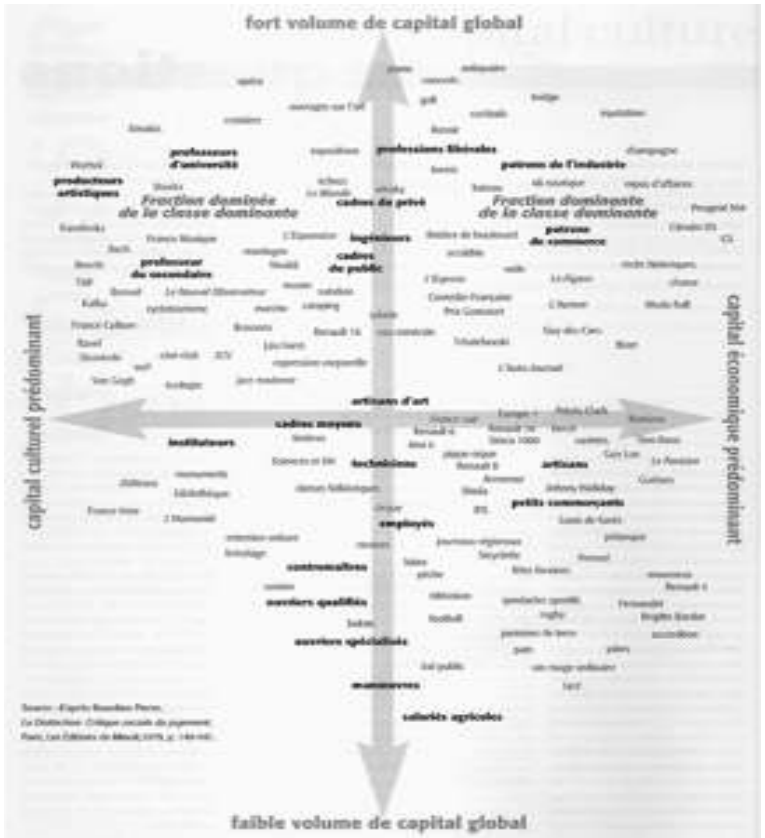


Figure: Bourdieu's GDA in *Distinction*, 1979

Unsupervised Setting: Complexity vs. Interpretability

A dataset about French MPs invitation in the media
623MPs (lines), 23 variables (one per channel, strong asymmetry)

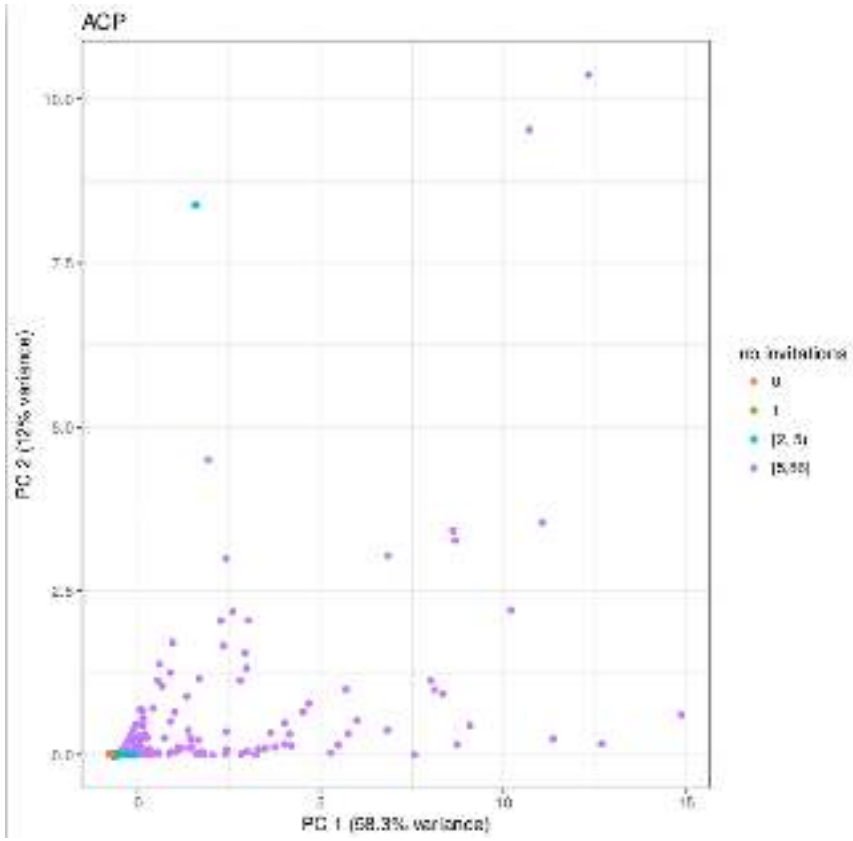


Figure: PCA on Media Invitations of French MPs

Unsupervised Setting: Complexity vs. Interpretability

A dataset about French MPs invitation in the media (2012-1015)
623MPs (lines), 23 variables (one per channel, strong asymmetry)

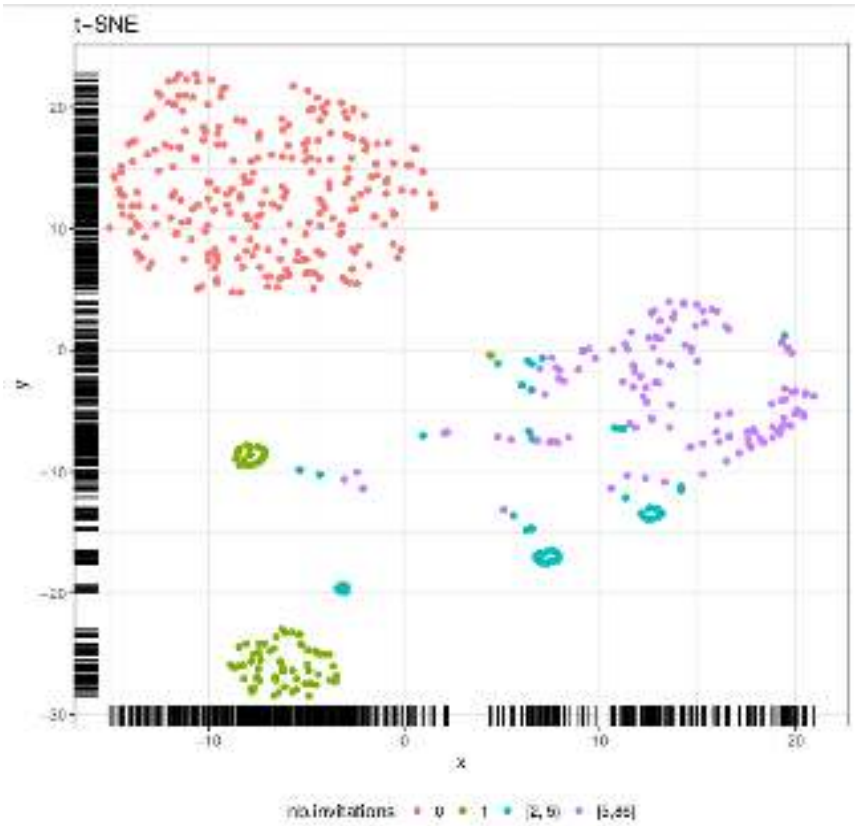
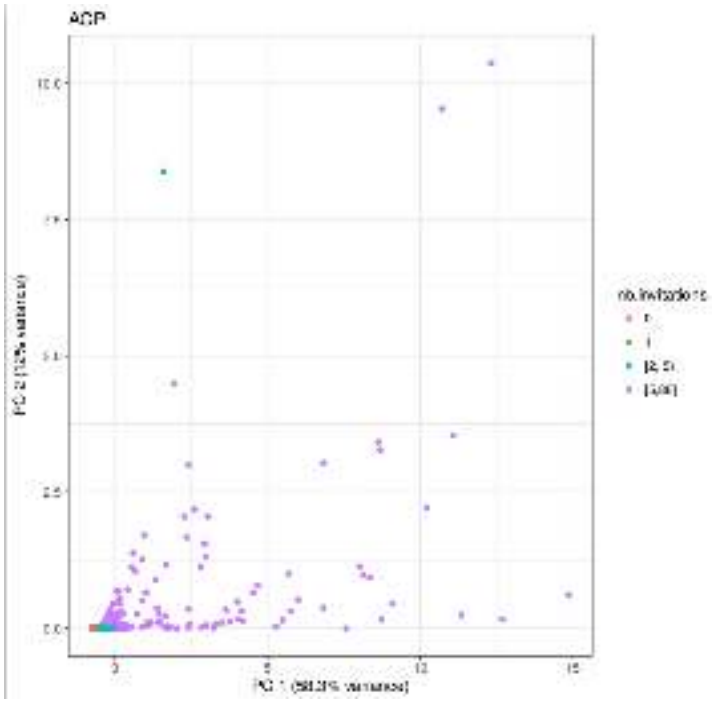
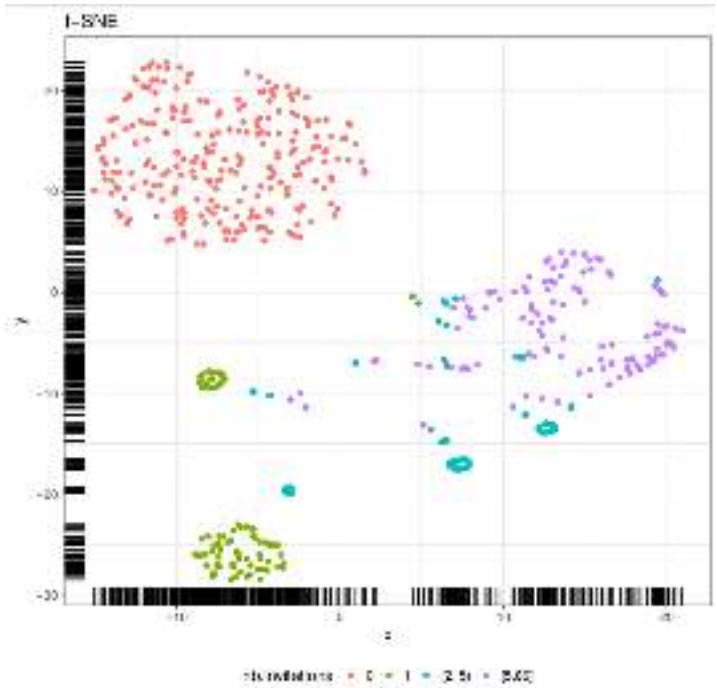


Figure: tSNE on Media Invitations of French MPs

Unsupervised Setting: Complexity vs. Interpretability

- More Flexibility –> Better rendering of variations
- More Flexibility –> Less interpretability (no axis, distance is meaningless)





See (Boelaert and Ollion, 2020), email if you want a copy

1 Introduction: The Wave and the Currents

2 Preliminary Remarks about Machine Learning
- An Old Endeavor
- A Series of Booms and Bust
- What is Machine Learning?

3 Differences in Practice: Two Approaches to Data
- Supervised Setting
- Unsupervised Setting

4 Learning to Analyze, or Learning to Produce?
- Learning to Produce
- History of Political Journalism

5 Conclusion: AI, Why should we care?

Learning to Produce

"Produce data": Using ML to extract information from large corpora

- Leverage the predictive power of the algorithm

Example: History of Political Journalism

- History of Political Journalism, with a focus on narration
  - Question: How do journalists talk about politics?
  - Looking at the text of newspapers (not authors, not metadata)
- Problem: Thousands of articles
- Classic solution: Sampling/hand annotation/ interpolation (resp. hire an army of research assistants)
- ML Solution: Train an algorithm so that it does the job for you
  - Show a limited number of examples to the algorithm
  - Make sure it has learnt adequately
  - Have it predict on the whole corpus

Example: History of Political Journalism

- Task: finding instances of "Off the Record"
  - Quoting unnamed sources, to reveal backstage aspects of politics
- Problem: many ways to introduce "Off the record" speech
  - "According to an unnamed source", "An aide to the President revealed that", "Persons briefed on the matter", etc...
- Solution: Train an algorithm
  - Carefully select instances of "Off" in a set of articles
  - 2400 instances hand-labelled, French highbrow newspaper *Le Monde*

History of Political Journalism

In Practice: a few hours of tedious annotation (~ 12h in this case)

« *C'est une drôle d'ambiance* » qui règne en ce moment dans les couloirs du pouvoir, souffle un conseiller ministériel : « *On travaille comme si de rien n'était, on prévoit des choses pour septembre, alors qu'on ne sait pas ce qu'il va se passer.* » Emmanuel Macron a promis de dessiner un « *nouveau chemin* » dans son quinquennat, début juillet, et personne ou presque, au sein de l'exécutif, n'a d'assurances sur sa place dans le futur attelage. « *On ne pense qu'à ça et on ne parle que de ça* », confie un ministre. Gouvernement, majorité, parti… Les soutiens du chef de l'Etat anticipent « *un changement à tous les étages* ». Selon son entourage, ce dernier compte se laisser le temps de la réflexion jusqu'à « *fin juin, début juillet* ».

Figure: Annotating text

History of Political Journalism

In Practice: Assessing the quality of the prediction

	Precision	Recall	F1-score	Support
Off the record	0.80	0.78	0.79	2041

Figure: Quality assessment (between 0 and 1)

Tedious, but cheaper, nicer, and better results than with an army of research assistants

History of Political Journalism

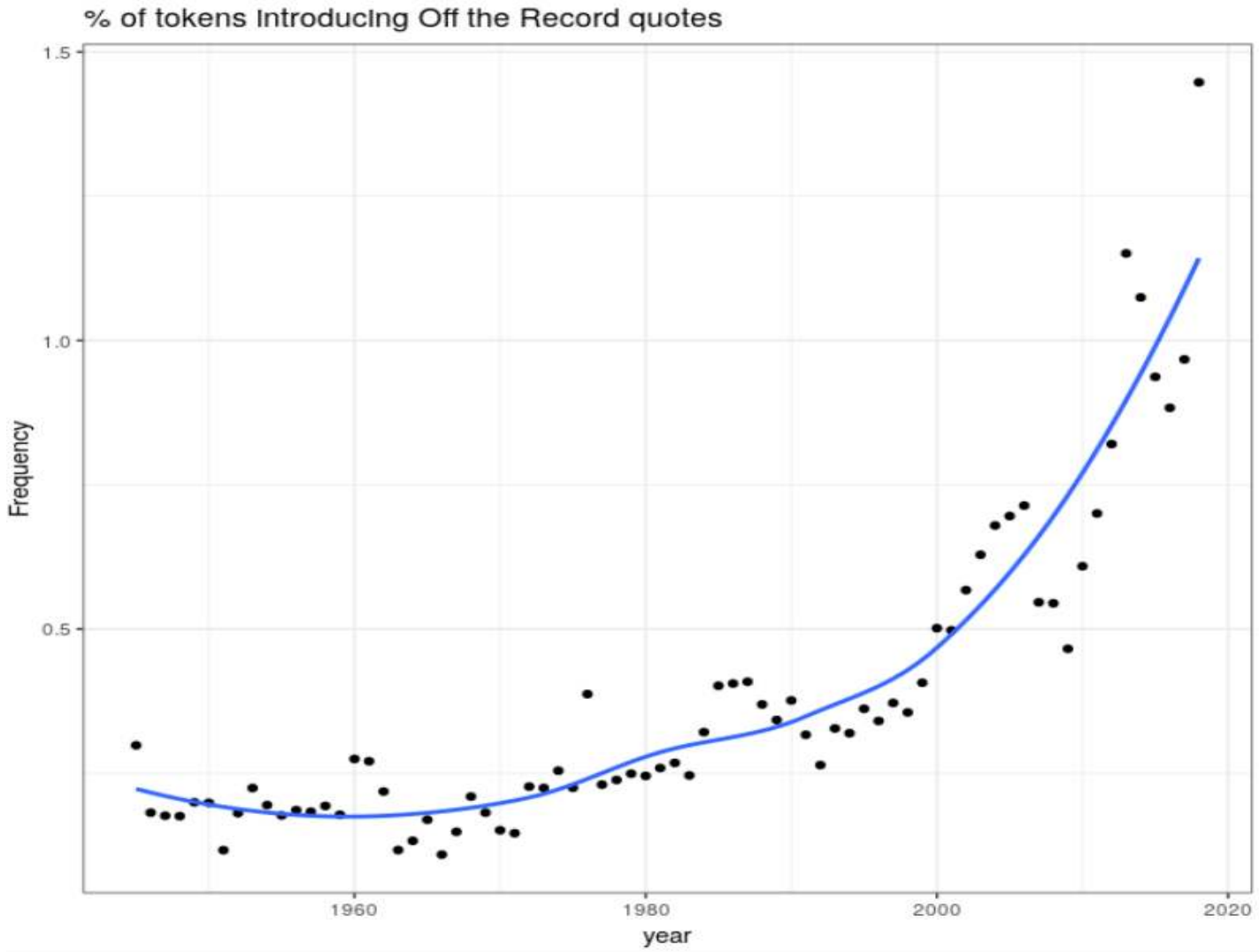
In Practice: Output

SELON nos [OFF] informations [OFF] obtenues [OFF] dans [OFF] l [OFF] ' [OFF] entourage [OFF] du [OFF] ministre [OFF] de [OFF] l [OFF] ' [OFF] économie [OFF] et [OFF] des [OFF] finances [OFF] , c'est le vendredi 3 septembre - et non le jeudi 2, comme cela l'avait été dit jusqu'à présent - que Nicolas Sarkozy annoncera sa décision sur son avenir au sein de l'Union pour un mouvement populaire (UMP). Il devrait confirmer sa candidature, que [OFF] l [OFF] ' [OFF] un [OFF] de [OFF] ses [OFF] proches [OFF] présente [OFF] comme [OFF] « [OFF] plus que probable », à la présidence de l'UMP, vacante depuis la démission d'Alain Juppé. Toujours selon [OFF] les [OFF] mêmes [OFF] sources [OFF] , cette annonce devrait prendre la forme d'un simple communiqué.

3 sequences of 'off the record'
32 tokens (out of 101)
2 misclassified tokens (1FP, 1FN)

History of Political Journalism

In Practice: Results



Producing Data

- Detect presence /absence (women in films, people in pictures)
- Transcribe writing style (German Kurrentschrift)
- Detect roles, moods
- Transcribe automatically audio, video
- Classify according to your own classification scheme

Becomes all the more relevant as we have growing masses of digital data

Producing data as a way out of the aforementioned conundrums

- Black box is not a problem anymore, so long as the prediction is good
- Validation is easy (just look)
- A method amenable to any type of (social) scientific research

1 Introduction: The Wave and the Currents

2 Preliminary Remarks about Machine Learning
- An Old Endeavor
- A Series of Booms and Bust
- What is Machine Learning?

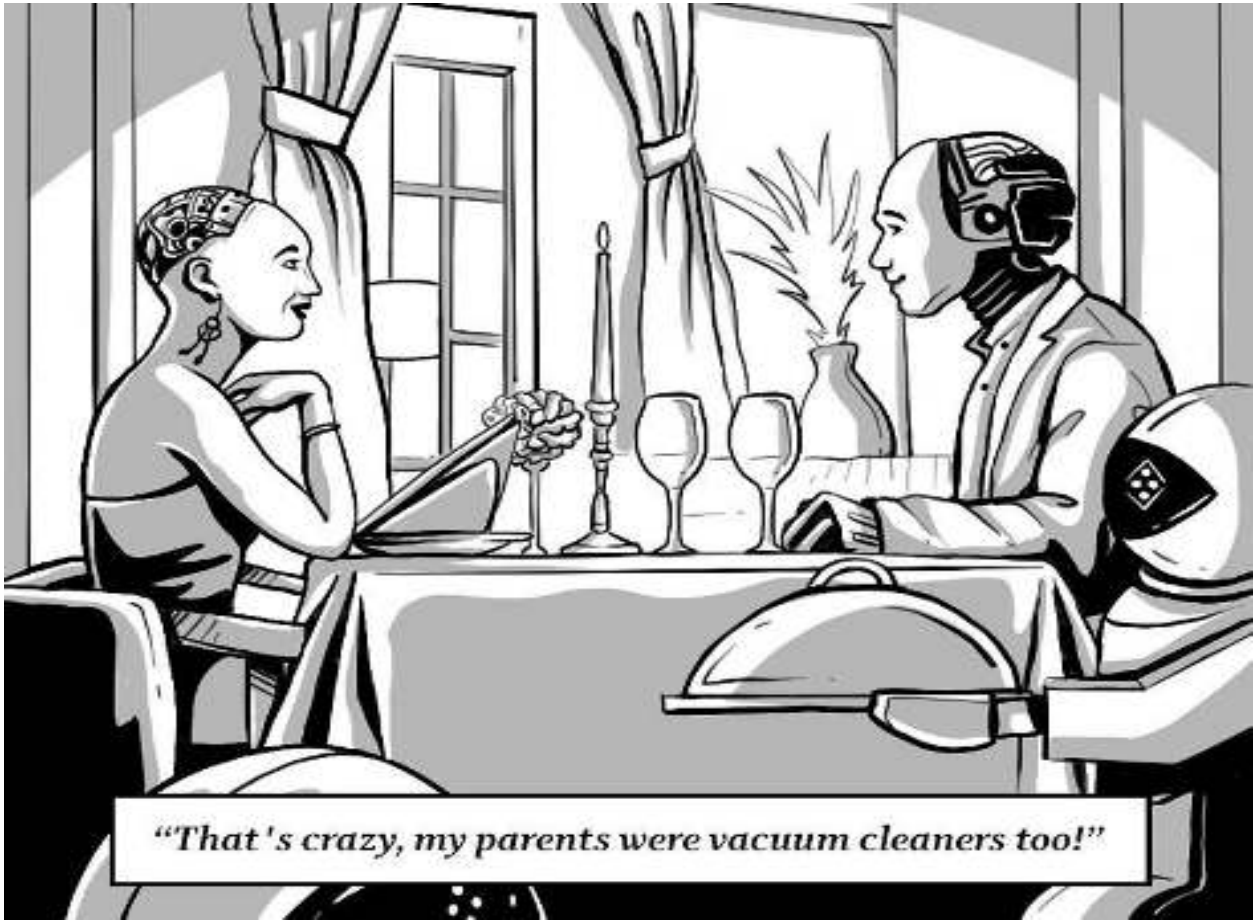
3 Differences in Practice: Two Approaches to Data
- Supervised Setting
- Unsupervised Setting

4 Learning to Analyze, or Learning to Produce?
- Learning to Produce
- History of Political Journalism

5 Conclusion: AI, Why should we care?

Conclusion AI, Why should we care?

- Reconsidering our statistical routines
- Bridging the gap between our theories and our methods
- Extracting knowledge from masses of data



Q&A

References I

Biau, Gérard (2012). "Analysis of a Random Forest Model". *The Journal of Machine Learning Research*.

Boelaert Julien; Ollion, Etienne (2018). "The Great Regression. Machine Learning, Econometrics, and the Future of Quantitative Social Sciences". *Revue Française de Sociologie*.

Boelaert Julien; Ollion, Etienne (2020). "How to Represent the Social Space. Methods and Concepts for an Enriched Topography of Fields". *working paper*.

Breiman, Leo (2001). "Statistical Modelling: The Two Cultures". *Statistical Science*.

Burell, Jenna (2016). "How the Machine Thinks. Understanding Opacity in Machine Learning Algorithms". *Big Data and Society*.

Cardon Dominique; Cointet Jean-Philippe; Mazière, Antoine (2018). "Neurons Spike Back". *Réseaux*.

Casilli, Antonio (2019). *A quoi rêvent les algorithmes*. Le Seuil.

McCorduck, Pamela (2004). *Machines Who Think. A Personal Inquiry into the History and Prospects of Artificial Intelligence*. Taylor and Francis.

References II

Salganik, Matthew et al. (2020). "Measuring the predictability of life outcomes with a scientific mass collaboration". *PNAS*.