Digital Strategies for the Social Sciences - 6

Étienne Ollion – CNRS

LOGISTICS

- 5 sessions, usual routine (morning lecture, afternoon lab)
- Before / after class
  - ✓ Sometimes a video to watch
  - ✓ Apply your skills on the data set of your choice
- As per usual, your feedback is key
- Watch out for changing rooms (indicated on the website)

Dealing with mass data: An introduction to quantitative description

Data analysis in the 21st Century

- A strong corpus of methods available, with demonstrated results
- But: data avalanche
  - ✓ Digitalization of everyday life and new information

DIGITAL STRATEGIES

Data analysis in the 21st Century

- A strong corpus of methods available, with demonstrated results
- But: data avalanche
  - ✓ The digitalization of everyday life leads to
    - More data
    - More variables (“organic data”, Groves 2011)
      * some of which are of uncertain relevance
      * The correlations between them may be hard to see
    - Tons of missing values

DIGITAL STRATEGIES

Data analysis in the 21st Century

- A strong corpus of methods available, with demonstrated results
- But: data avalanche
- Where to start? **Quantitative Description**

WHAT IS QUANTITATIVE DESCRIPTION ?

A rare term for a common practice

- Description of an object using numbers (**the source is indifferent**)
- Various techniques to uncover structures, patterns & processes

WHAT IS QUANTITATIVE DESCRIPTION ?

A rare term for a common practice

- Description of an object using numbers (**the source is indifferent**)
- Various techniques to uncover structures, patterns & processes

Specters of quantitative description

- Not Qualitative: A description with numbers
- Not Modeling: No *a priori* definition of a functional form (as in regressions)

WHAT IS QUANTITATIVE DESCRIPTION ?

QD is not limited to descriptive statistics

- Descriptive statistics
- Sequence analysis
- Some network analysis
- Dimensionality reduction
- Clustering
- Machine learning?

THE STANDING OF QUANTITATIVE DESCRIPTION

Very dependent on the discipline

- Virtually absent in Economics, quite present in History (if quantification happens at all).

In US Sociology, a recurrent but somewhat invisible practice

- Long regarded as “unsophisticated statistics” and used as a first step towards more “refined” analyses.
- A classic distinction from the 70s: Exploratory vs. Confirmatory Data Analysis (EDA/CDA)

THE EDA / CDA DEBATE

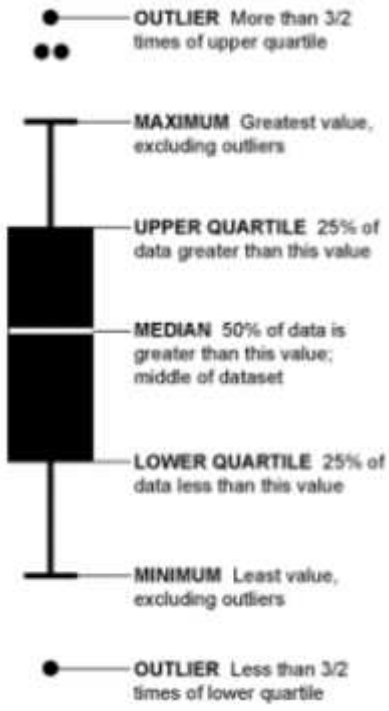
Started out on in the 1970s

- Growing number of available data, and dissemination of computers
- The rapid success of regression techniques in the social sciences
- A question: what to do with descriptive statistics

THE EDA / CDA DEBATE

Tukey's Take: Explore and Visualize

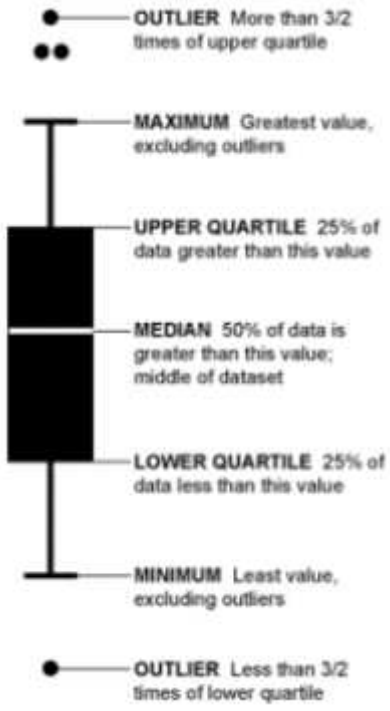
- John Tukey (1915-2000)
  - ✓ US Statistician, worked at Bell Labs & Princeton
  - ✓ Invented the Tukey plot
  - ✓ *Exploratory Data Analysis*, 1977



THE EDA / CDA DEBATE

Tukey's Take: Explore and Visualize

- EDA, An Ambivalent Enterprise
  - ✓ Advocated for descriptive statistics, developed tools and methods
  - ✓ But granted them a lower position in the research process
    - ◆ Assess the quality of the data
    - ◆ Generate hypotheses
    - ◆ First hand description
  - ✓ "A necessary, but insufficient, analysis"
  - ✓ cf. Kramer and Thiemann: "indiction" vs. "conviction"



THE STANDING OF QUANTITATIVE DESCRIPTION

The current return of quantitative description

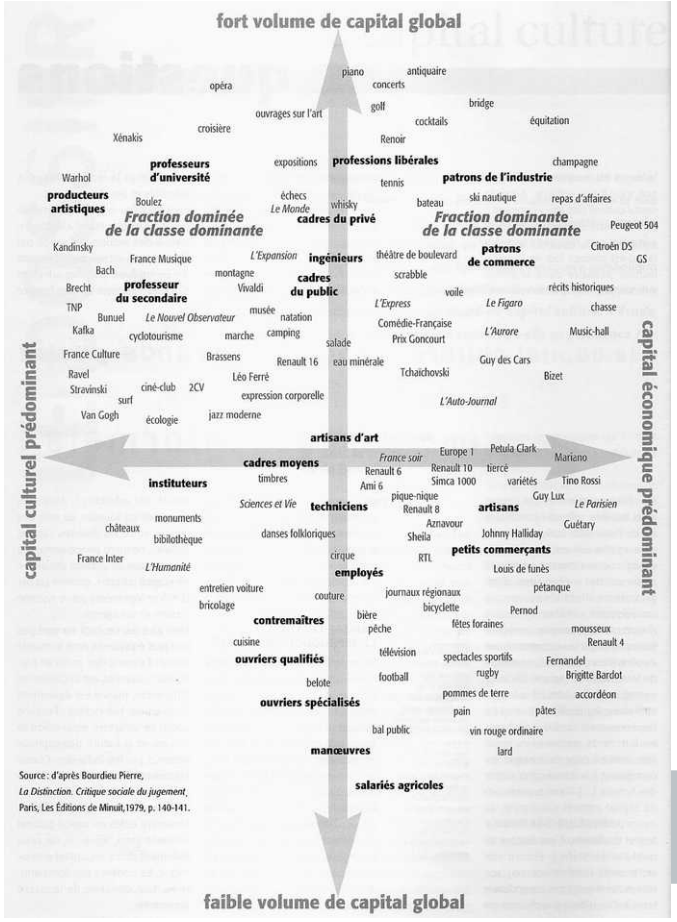
- The data avalanche
  - ✓ Digitalization of everyday life and new information
- The regression crisis
  - ✓ Classic criticisms (unrealism: Simiand 1903; oversimplification, Abbott 1988)
  - ✓ Issues with results, with replication (Ioannidis 2005)
  - ✓ Regressions don’t do well with “big data”
    - ◆ Masses & inference test: everything is significant
    - ◆ Masses of “organic data” (Grove, 2011)
    - ◆ Data on new phenomena: lack of established (parametric) knowledge

METHODS OF QUANTITATIVE DESCRIPTION?

Which methods can we use?

Geometric data analysis

- Dimensionality reduction, i.e. transforming a cloud of points into a 2D embedding.



An old endeavor, coming back

Intuition: Finding proximities and oppositions between individuals based on the frequency of their practices/ position (measured by variables)

Bourdieu, *Distinction*, 1979

Which methods can we use?

Geometric data analysis

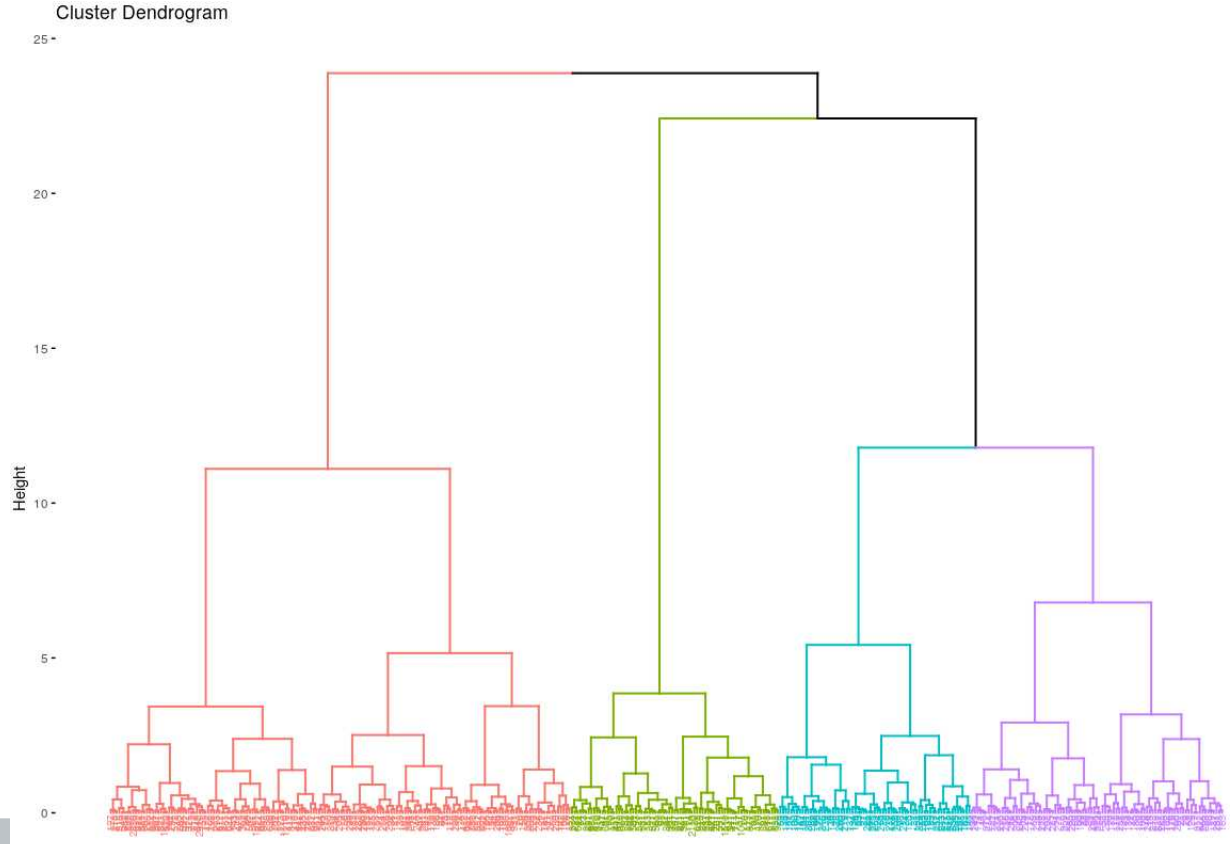
	TV Public	TV Private	CSpan	Radio Public	Radio Private	TV Cable	Radio Cable	Internet
MP1	1	2	5	5	3	1	1	0
MP2	0	0	0	0	0	0	0	0
MP3	0	0	1	0	0	0	0	0
...								
MP576	0	0	1	1	0	0	0	0
MP577	15	12	12	16	19	9	12	10

Media invitations of Mps in France

Which methods can we use?

Clustering

Grouping individuals together, based on a set of properties



Which methods can we use?

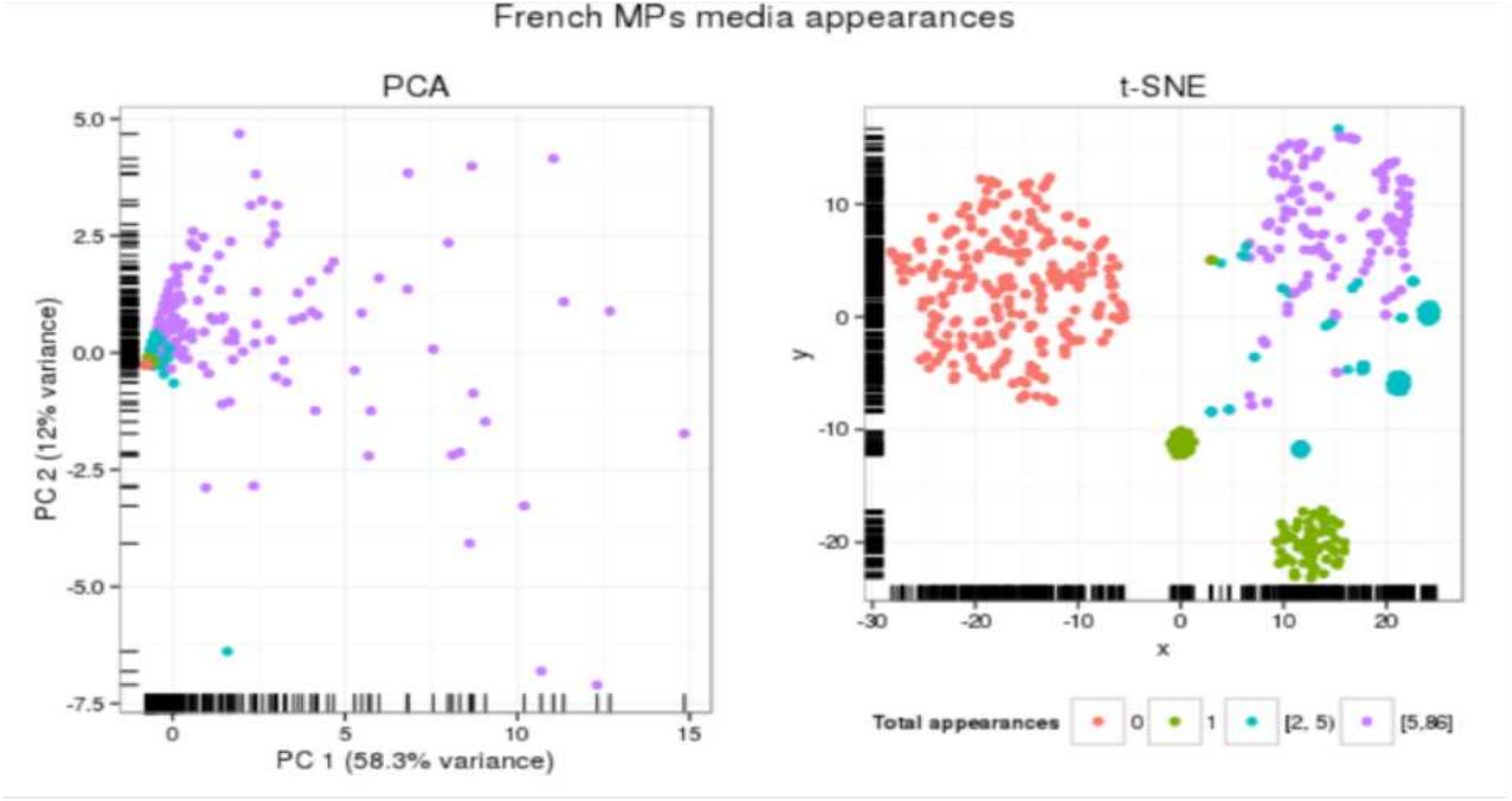
Clustering

Grouping individuals together, based on a set of properties

Methods of clustering?

Which methods can we use?

Non supervised Machine learning



Which methods can we use?

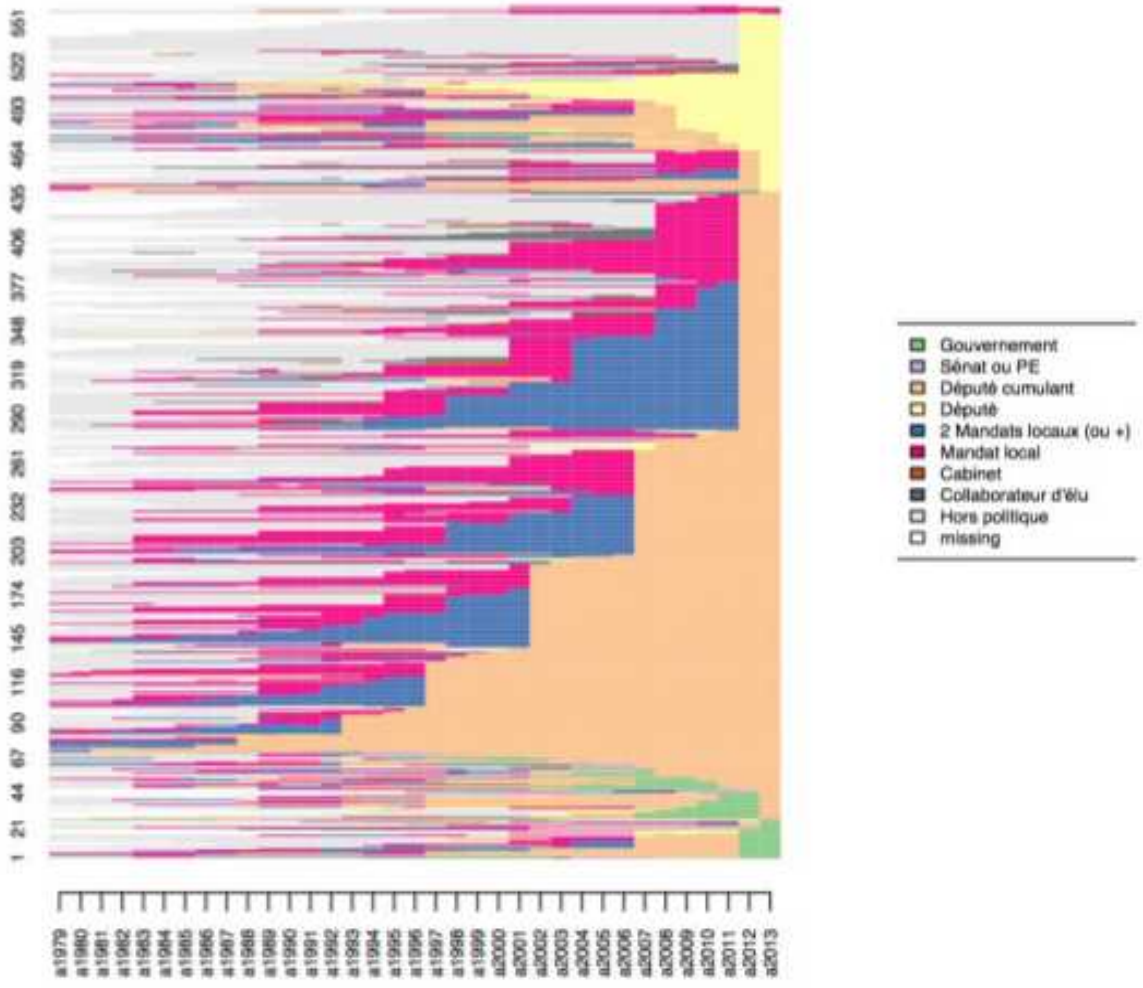
Sequence Analysis

Capturing life-course sequences

nid	1978	1979	1980	1981	1982	...	Election year	Party	Gender
1	mp	mp	mp	mp	mp		1978	Right	M
2	manager	manager	manager	mp	mp		1981	Right	M
3	teacher	staffer	staffer	mp	mp		1978	Left	M
4	mp	mp	mp	retired	retired		2012	Left	F
5									

Which methods can we use?

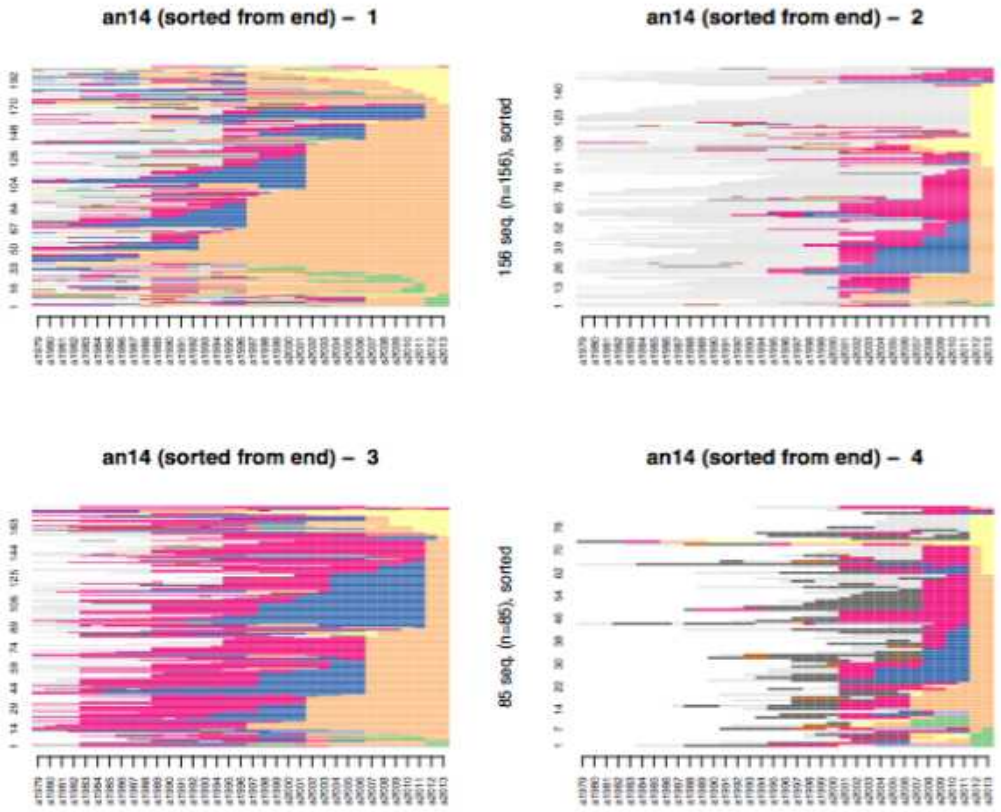
Sequence Analysis



MPs, 2012 election

Which methods can we use?

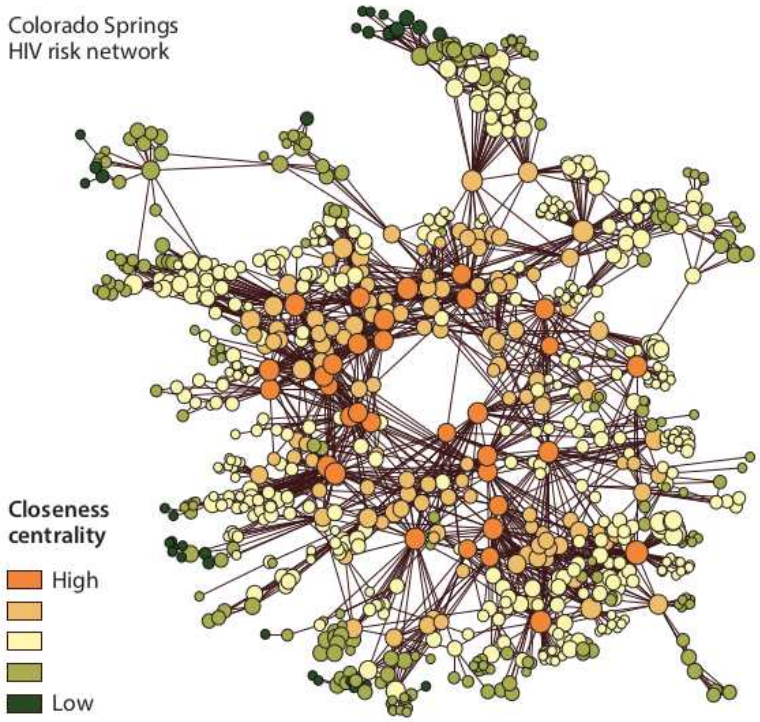
Sequence Analysis



Careers of French MPs, in Ollion *et al.* 2017

Which methods can we use?

Network analysis



Bearman *et al.*, “Chains of Affection,” *AJS*, 1996 (as per Healy & Moody, 2014)

Quantitative descriptions, what for?

- Exploratory analysis, debiasing of data set
- As a means to use another approach
- As an end

This period

* Geometric data analysis (PCA, MCA)
* Clustering (Partitioning, Hierarchical)
* How to revisit this with machine learning?

6 – Intro (+ Textual Analysis with ML)

7 – Intro to PCA

8 – PCA, MCA advanced

9 – Clustering

10 – Non supervised Machine learning

Validation

Validation – Week 2

A project exploring a dataset

Write a report using one of the 4 datasets (not all 4!)

Validation – Week 2

* Describe the data you will use
* Correct it if need be (NA removal, imputation, aggregation, else ?)
* Find a question to answer
* Offers elements of response using methods seen in class (dimensionality reduction, clustering).
    - Describe the technique briefly
    - Explain why you think it is relevant on this data
    - Apply it, making sure you provide your reader with enough information to 1. understand & 2. trust you (but not too much → Place in appendix).
    - Evoke possible limitations
    - Suggest conclusions.

Validation – Week 2

A project exploring a dataset

<u>To get an B</u>

* Use one method
  - Justify. Why can you do this. Why is it interesting?
  - Did you learn something? (what?)

* Offer an interpretation
  - Assess the role of certain variables over others
  - Offer plausible explanations

* Assess the limits of the methods

To get an A

* Combine two (or more) methods.

* Offer an interpretation

* Write this as a research paper

i) find a relevant question
ii) evoke some literature
iii) advance knowledge
[(iv) publish it in *Science* and get a nobel Prize]
all of this before **June 2nd, 23:59**)

Report:
- circa 7 pages (1.5 spaced).
- Appendix possible.
- Clear code for replication

Turn in: June, 2nd 2024 23:59

The datasets

1. « MP practices » (Ollion *et al.*, 2020)
   - How do MPs do politics ?
   - A year of practices (2017-2018), in the new legislature
   - 23 variables, 569 MPs

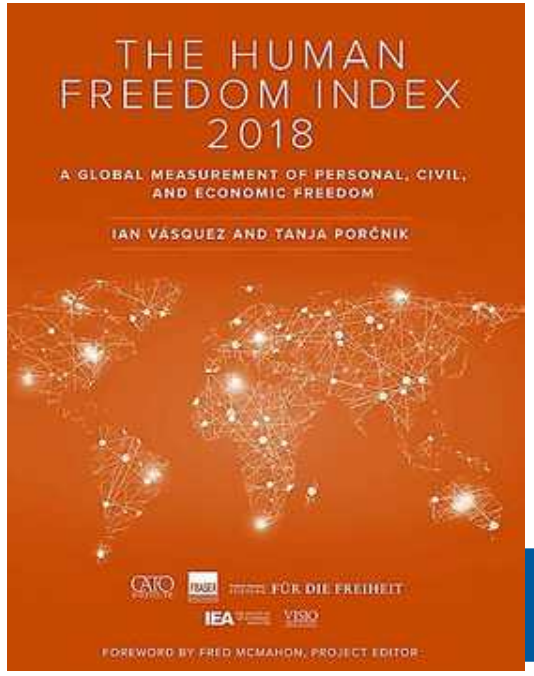
The datasets

2. « Measuring Socialism » (Joseph Cohen, 2018)

- No consensus about socialism (Venezuela? Sweden?)
- 41 countries, 2015-2017
- 243 variables about the role of the State in the economy (public expenditures, revenues, investment, workforce…)
- Difficulties : missing values

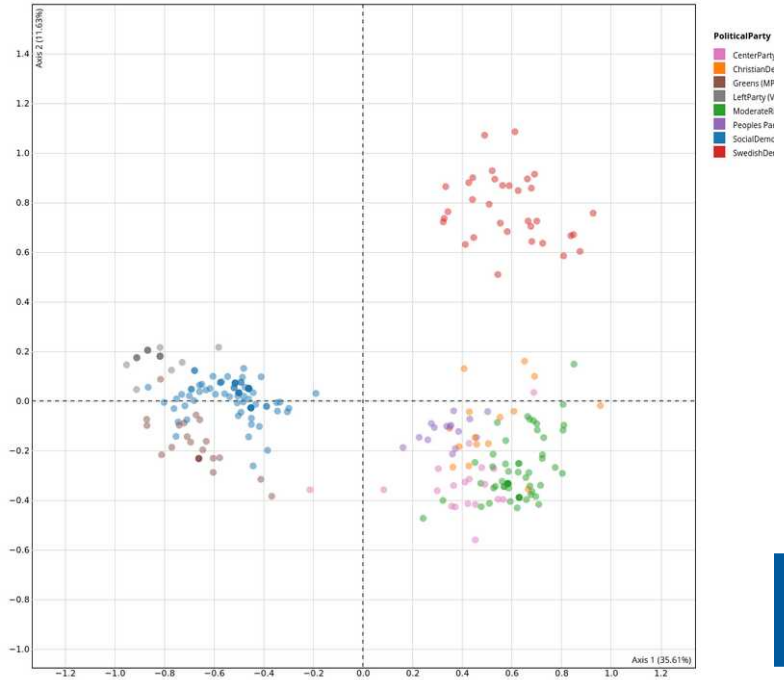
The datasets

3. Human Freedom Index (Cato institute, 2018)
   - Indicators about types of freedom
   - Data for 2008-2016, 128 countries
   - Difficulties : diachronic data, missing values



The datasets

4. Swedish MPs, 2014 & 2018
   - Political Positions of Parties in Sweden
   - Difficulties : diachronic data, missing values, different populations & questions
   - (need for a good research question)



For next session

1. Watch the video (see course website)