POL40300: COMPUTATIONAL METHODS

Lecture 3 by Nikolai Gad
München, 2. November 2021

TODAY'S LECTURE

- Practicals
- What is Computational Methods?
- New subjects to study.
- New data (for PolSci)
- What can we use computational methods for?
- Next week

PRACTICALS

ONLINE TUTORIALS FOR TWO WEEKS

One student tested positive for Corona virus after attending class last week.

- All tutorials will be online until 10th November.
- So this week and next week.
- Find link to Zoom on Moodle
- Please only attend the tutorial class you are registered for!

HOW DID YOU FIND EXERCISE 2?

- Weekly polls on Moodle
- Poll about last week's tutorial online since yesterday
- Any other comments on tutorial?
- Always feel free to contact me with feedback on tutorials!

WHAT IS COMPUTATIONAL METHODS? SOME TERMINOLOGY

Not one right answer!

Today we will look at one take of this: *Brady, H. E. (2019). The Challenge of Big Data and Data Science. Annual Review of Political Science, 22(1), 297–323.*

BIG DATA

- A question of size?
- National Institute of Standards and Technology: Parallel computing needed instead of horizontal scaling.
- Brady: "A change in our cognitive environment" - The nature of big data is different:
  - Digitalised data: Structured and replicable.
  - Connectedness: (almost) Instant access and feedback.
  - Networked: Point-to-point and many-to-many, individualised/targeted mass communnication.
  - Automatically recorded in real-time.

DATA SCIENCE

Brady presents three views:

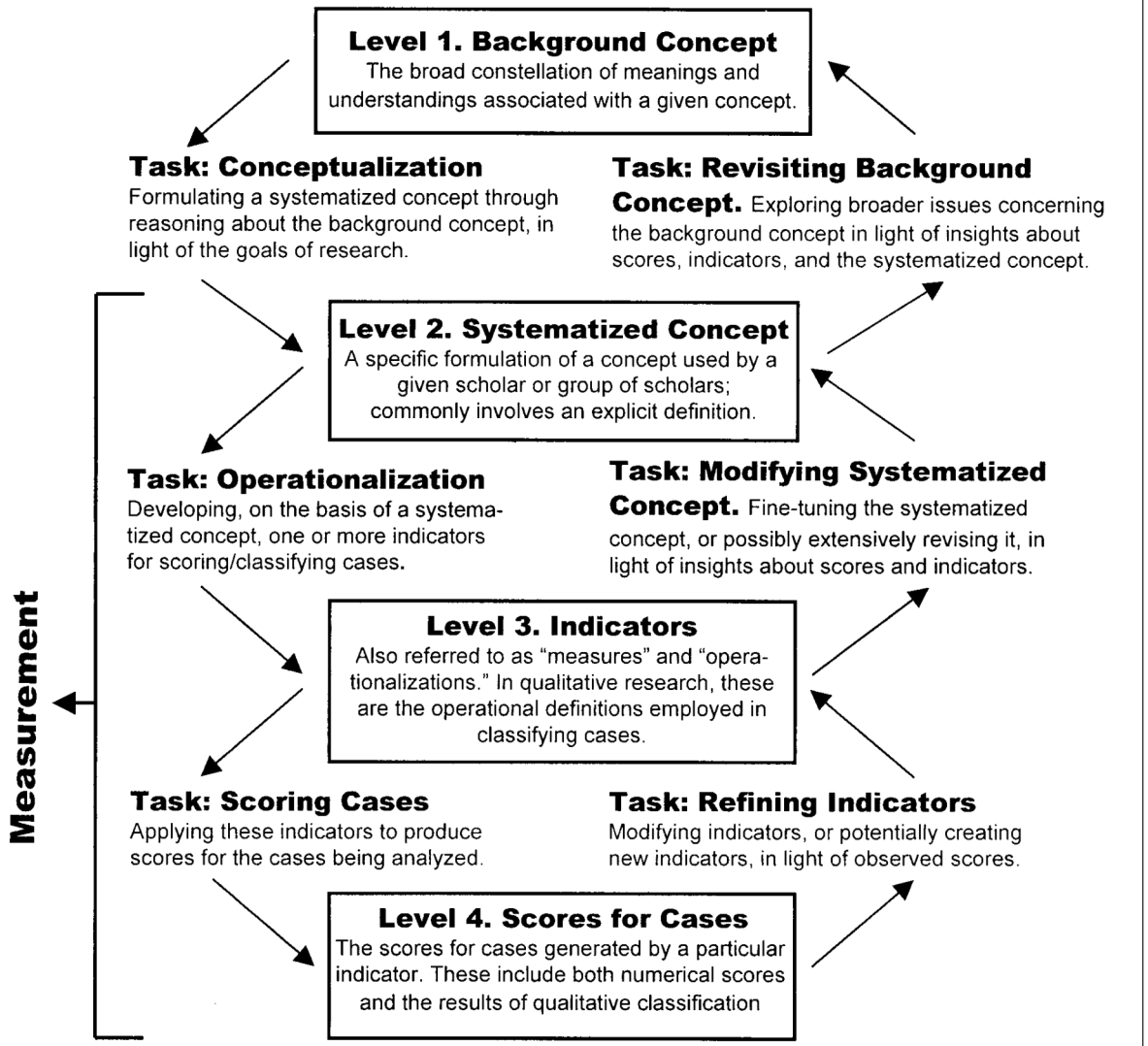
- Data Science as an umbrella containing skills from many disciplines (see table based on Donoho).
- Data Science as a set of skills or a seperate discipline.
- A new paradigm of data-driven science “that uses large collections of data to make scientific discoveries.” (based on Jim Gray)

But at the end of the day...

> *“We are left with real phenomena but inadequate language.” (p. 304)*

Brady, H. E. (2019). The Challenge of Big Data and Data Science. Annual Review of Political Science, 22(1), 297–323. https://doi.org/10.1146/annurev-polisci-090216-023229

FIGURE 1. Conceptualization and Measurement: Levels and Tasks



Adcock, R., & Collier, D. (2001). Measurement Validity: A Shared Standard for Qualitative and Quantitative Research. American Political Science Review, 95(3), 529–546.

TRADITIONAL POLSCI METHODS:

- Research Question
- Hypothesis
- Conceptualisation
- Operationalisation
- Measurement / Data

DATA-DRIVEN METHODS:

- Data first!
- What does it represent? (Operationalisation)
- What does it mean? (conceptualisation)
- What can it tell us? (hypothesis)
- What can we learn from this? (RQ)

“We are left with real phenomena but inadequate language.” (p. 304)

Brady, H. E. (2019). The Challenge of Big Data and Data Science. Annual Review of Political Science, 22(1), 297–323. https://doi.org/10.1146/annurev-polisci-090216-023229

"Traditional approach loosely based on Adcock, R., & Collier, D. (2001). Measurement Validity: A Shared Standard for Qualitative and Quantitative Research. American Political Science Review, 95(3), 529–546."

NEW SUBJECTS TO STUDY

Big Data and Data Science (and the developments in computation that has enabled these phenomona) also brings with them new *political* issues to study. Some examples Brady mentions:

- Cyber War
- Smart Cities
- Precision Medicine
- (Online) media
- Robots and AI

NEW DATA AVAILABLE (FOR POLITICAL SCIENCE)

ADMINISTRATIVE DATA (REGISTER DATA)

Long history in PolSci, but new possibilities with computational methods:

- Larger quantities and more complex analysis
- Combining datasets (not always trivial to do though)
- Be aware of reliability of this kind of data!
- Absence of data.
- Also be aware of ethics and privacy regulation.

INTERNET DATA

(Almost) everything we do online leaves a digital trace.

- One-click actions.
- Fine grained time series data.
- Online experiments (be aware of ethics here though!)
- Real-time and unfiltered data collection.

INTERNET DATA

Pros:

- Data about whole networks
- Real-time data
- Hidden patterns

Cons/Challenges:

- Unrepresentative samples:
  - Internet users generally not representative.
  - User of specific platforms even less so.
  - Nudging by platforms affects online behaviour.
- Absence of data.
- Access to data.

TEXT DATA

Long history of text analysis in PolSci (despite what Brady insinuates), but computation provides new methods to collect and analyse (large quantities of) text:

- Automatic content analysis
- Word frequencies and lexical analysis
- Classification (supervised machine learning)
- Clustering (unsupervised machine learning)
- Ideological scaling (supervised machine learning)
- (Dimension reduction (unsupervised machine learning))

New sources of text (Brady does not mention these, but nonetheless significant to computational methods):

- Immense amounts of new political text produced.
- Automated collection of text data.

TEXT DATA

Pros:

- Cost! Less time consuming and researcher time equals money.
- Reliability
- Replicability

Cons - “all quantitative models of language are wrong – but some are useful” quote from Grimmer & Stewart (2013):

- Validity can be tricky
- (Do they actually bring anything new to PolSci?)

SENSOR, AUDIO, VIDEO, AND OTHER DATA

- Sensor data
- Video and audio
- (Images and static visuals)

WHAT CAN WE USE COMPUTATIONAL METHODS FOR (IN POLSCI)?

DATA SCIENCE COMES FROM DIFFERENT FIELDS

Computer scientists:	Statisticians (PolSci methodologists):
Pattern recognition	Hypothesis testing of causal relationships
Predictions	Explanation

“Whatever the reason, deep learning methods seem to work remarkably well for pattern recognition problems, but their interoperation is often difficult given their arcane complexity. ***They are better at yielding predictions than explanatory insights.” (p. 315)***

Brady, H. E. (2019). The Challenge of Big Data and Data Science. Annual Review of Political Science, 22(1), 297–323. https://doi.org/10.1146/annurev-polisci-090216-023229

https://xkcd.com/1838/

FOUR BASIC PROBLEMS IN (QUANTITATIVE) EMPIRICAL RESEARCH

1. Forming concepts (and measuring them)
2. Reliable descriptive inference
3. Causal inferences from past experience
4. Predicting the future

FORMING CONCEPTS (AND MEASURING THEM)

Factor analysis, principal component analysis, and (simple) cluster analysis have a long history in PolSci.
Machine learning is great at pattern recognition.
COMPLETELY DATA-DRIVEN APPROACHES FOR CONCEPT FORMATION STILL VIEWED WITH SOME (JUSTIFIED) SCEPTICISM THOUGH!

Feeds into discussion about inductive vs. deductive research methods, and pragmatic vs. theoretic concept formation.

RELIABLE DESCRIPTIVE INFERENCE

Sometimes (too often) big data are presented as population data.
Unrepresentative data and irrelevant variables gives biased results.
A LOT OF DATA AND A LOT OF VARIABLES DOES NOT EQUAL REPRESENTATIVENESS!

CAUSAL INFERENCES FROM PAST EXPERIENCE

Probably where machine learning and other computational methods have least to offer.
Look out for statements like “[a] worldview built on the importance of causation is being challenged by a preponderance of correlations”
CORRELATION DOES NOT EQUAL CAUSATION!

=> Sampling and variable selection still needs to be done thoroughly to check for spuriousness (despite what some data scientists migth tell you).

PREDICTING THE FUTURE

Machine Learning and in particular deep learning methods have turned out to be incredible good at this (sometimes).
But often difficult to interpret what is going on in the resulting algorithms.
So why do we need to know what caused something if we (ie our algorithms) can predict the outcome with enough data anyway?

- We still want to learn and understand (purpose of any science/research).
- Biased (discriminatory) algorithms.
- Critical research: Difficult to question existing social structures if we don't understand them, but can only predict the outcome.

SO CAN COMPUTATIONAL METHODS ACTUALLY TEACH US ANYTHING NEW?

To some extent the jury is still out on this.
But for now we can note that Machine Learning alone are mainly good at:

- Automated pattern recognistion, and
- Prediction

...which again can be used as part of a research design to do more:

- Measure new variables
- Measure them more precisely (?)
- Classify/measure larger datasets

THIS WEEK'S TUTORIAL

Install package for use next week.
From next week onwards lectures will be more directly related to tutorials.

NEXT WEEK: TEXT AS DATA

Some parts of assigned text a bit technical, but additional text explain some things in more simple language.