Quantitative Text Analysis and Natural Language Processing using Python

Day 1

Joshua Cova and Luuk Schmitz

2026-01-22

Introduction

A quick round of introduction

- Do you have any experience with quantitative text analysis, natural language processing and/or programming?
- What brings you here?
- Do you already have text data that you can work with?

Workshop overview

Today (22 Jan)

- What is text-as-data? Conceptual foundations behind NLP
- How to use Python

Tomorrow (23 Jan)

- Introductory quantitative text analysis techniques and their implementation in Python
  - Frequency analysis
  - Sentiment analysis
  - Bag-of-words models
  - Text classification (Binary)
- Throughout this workshop we will combine theory (validity, reliability, research design), with methods and practice (Google Colab/Python)

Why text as data matters for social scientists?

- Text is everywhere:
  - Political speeches
  - Newspaper articles & Social media communication
  - Policy documents, court rulings
  - Interview transcripts and survey responses

- Text as an indicator of "latent" concepts → What is populism? How can we identify instances of populist communication?
- Great to analyze qualitatively, but hard to scale!
- NLP does not replace qualitative analysis, but *complements* it.

- Close vs. Distant Reading (Moretti, 2013)
  - Patterns, trends and frequencies across million of documents

Why is NLP important

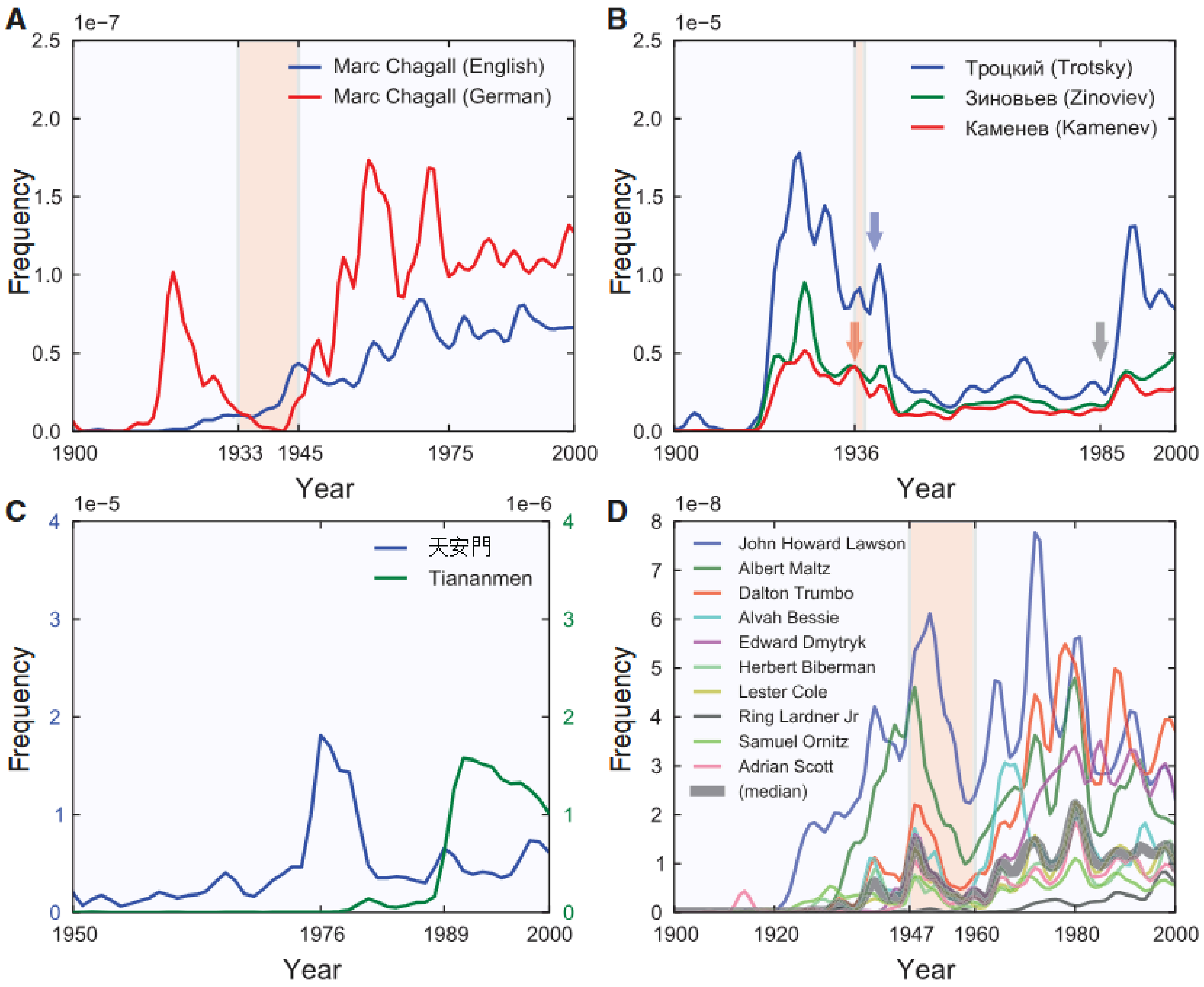
> **What is NLP?**
>
> "NLP enables computers […] to recognize, understand, generate text and speech by combining computational linguistics, the rule-based modeling of human language together with statistical modeling, machine learning and deep learning" (IBM)

Enables:

- Systematic analysis of large-scale textual data
- Replicable and transparent coding procedures

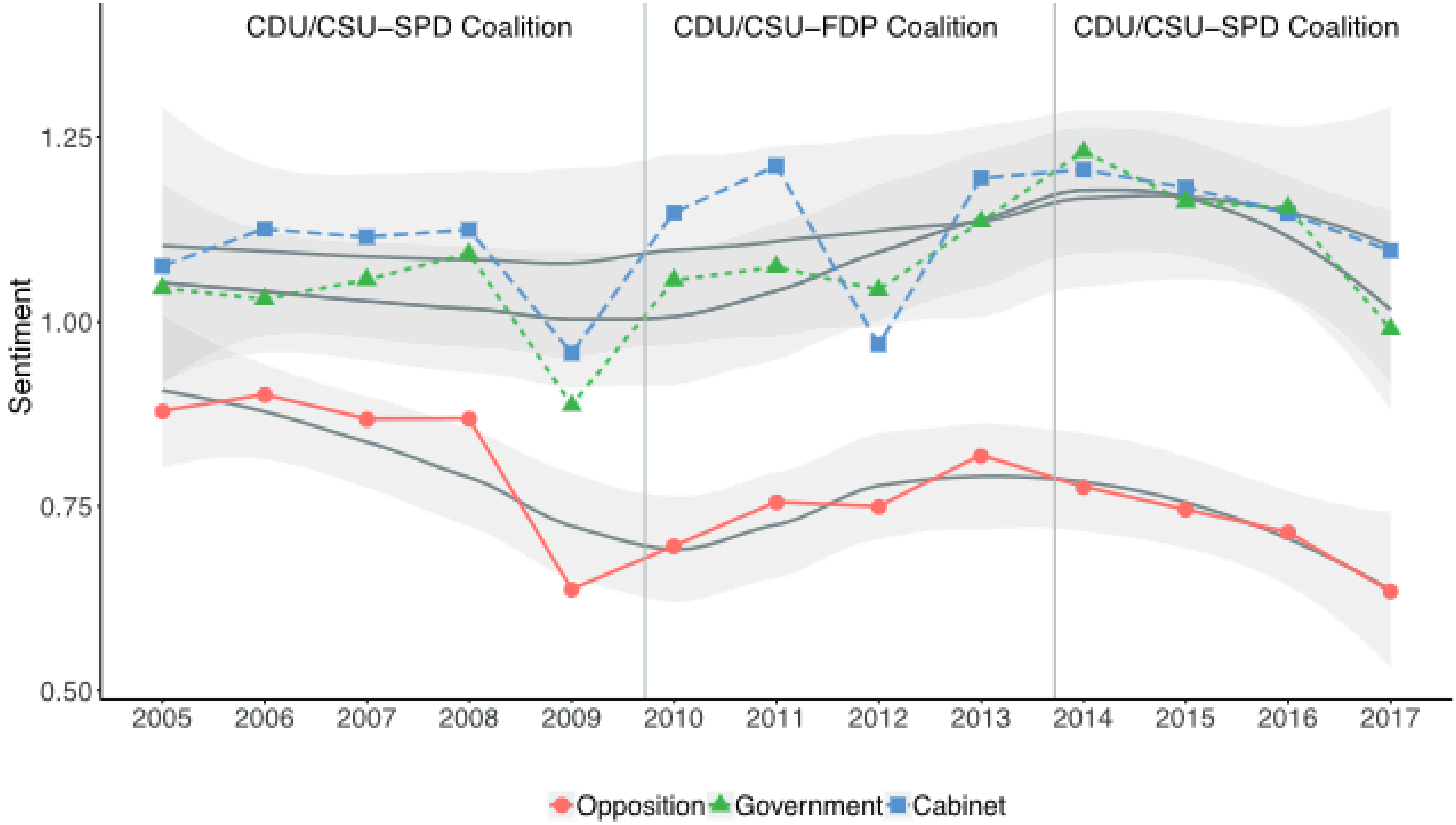
Some motivation

Culturnomics: Using QTA to trace human history



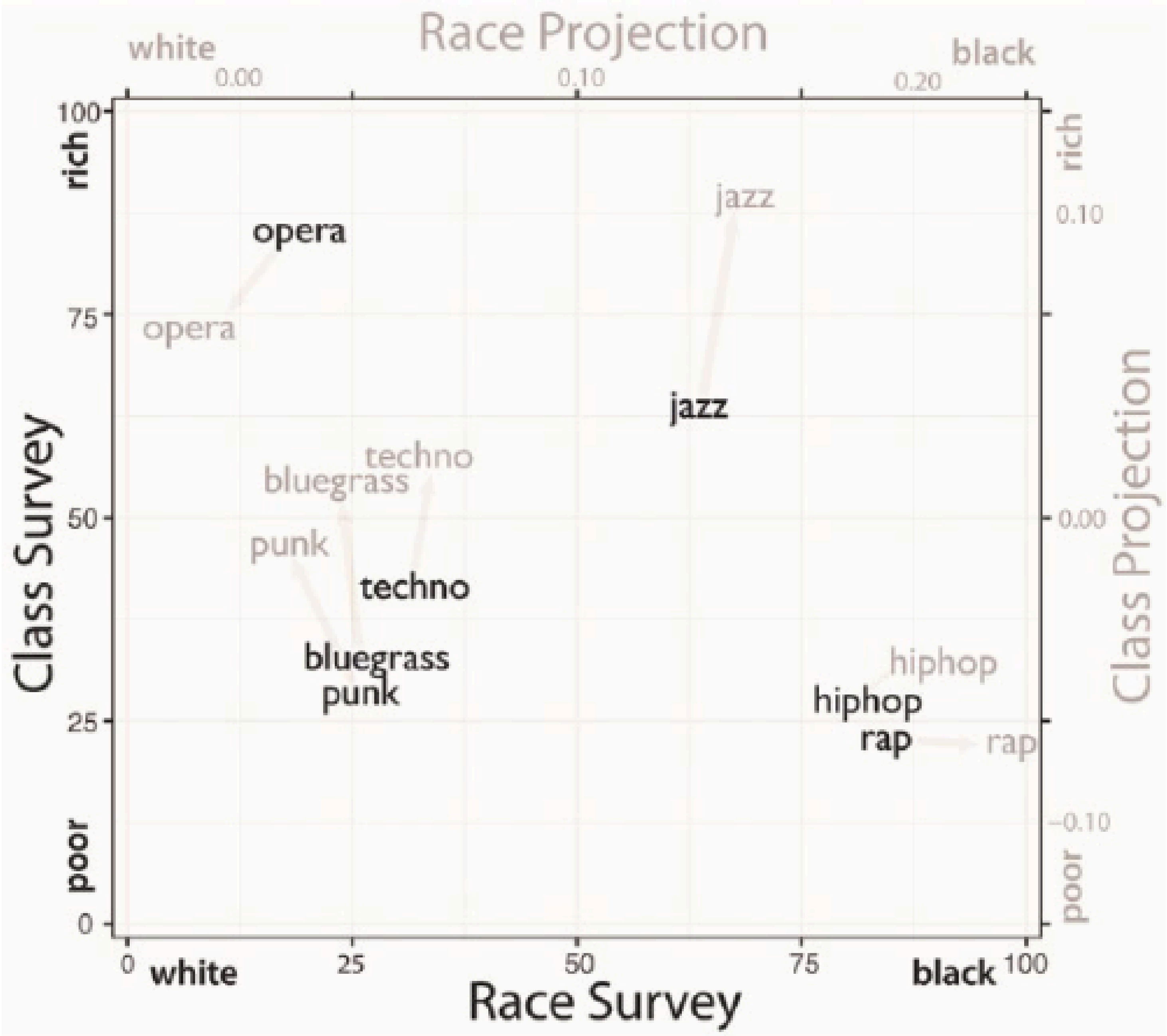
Michel et al. (2011)

Sentiment analysis in action



Proksch et al. (2018)

Enter word embeddings: The changing meaning of class



Kozlowski et al. (2019)

What NLP can and cannot do

What NLP can do

✓ Detect patterns, trends, and differences in language use

✓ Systematically code text at scale

Strong link between theory, data and method

It is crucial to understand the building blocks of NLP to understand how LLMs work (upcoming workshop in the spring)

The Python programming language

- Widely used in data science and NLP
- Large ecosystem of libraries:
  - pandas, numpy (data handling)
  - nltk, spaCy (NLP)
  - scikit-learn (machine learning)
- Open-source and reproducible
- Strong community support

Afternoon overview

- Why theory matters before code
- Validity in text analysis
- Reliability and the human baseline
- Deductive vs. inductive approaches
- Preview of Day 2

Why theory matters before code

The same data, different findings

Consider two studies of European Parliament speeches on economic policy:

- Study A finds **increasing polarization** since 2010
- Study B finds **convergence toward centrist positions**

Both use the same corpus. Both use “state-of-the-art” NLP methods.

What went wrong?

The measurement problem, amplified

> "[Quantitative text analysis methods] are best thought of as *amplifying* and *augmenting* careful reading and thoughtful analysis."
>
> — Grimmer & Stewart (2013)

Core social science concerns remain:

- What are we measuring?
- Does our operationalization capture the concept?
- Would another researcher reach the same conclusion?

The danger of “just running the model”

`corpus → preprocessing → algorithm → results → paper`

At every arrow, you make choices:

- Which texts to include?
- How to tokenize, lemmatize, filter?
- Which algorithm, which parameters?
- How to interpret output?

Each choice can flip your findings.

Validity in text analysis

What are we actually measuring?

When we count words or classify documents, we're making claims about **meaning**.

But meaning is:

- Context-dependent
- Culturally situated
- Often ambiguous
- Not directly observable

Two types of validity

Semantic validity

Does the method capture the meaning we intend?

- Does “positive” in a sentiment dictionary mean what we think?
- Do our “populist” keywords actually indicate populism?

Construct validity

Does the operationalization map onto the theoretical concept?

- Is “word frequency” a valid proxy for “salience”?
- Does “sentiment score” capture “economic confidence”?

Case study: Measuring "populism"

What makes a speech "populist"?

Different operationalizations:

1. **Dictionary approach**: Count populist keywords (Rooduijn & Pauwels 2011)
2. **Anti-elite rhetoric**: Classify attacks on elites
3. **People-centrism**: References to "the people" vs. institutions
4. **Full definition**: Anti-elite AND people-centric AND Manichean

The populism measurement problem

Approach	Captures	Misses
Keywords	Explicit populist language	Implicit populism, dog whistles
Anti-elite	Criticism of establishment	People-centrism dimension
People-centric	Appeals to “the people”	Elite criticism dimension
Full definition	Theoretical completeness	Complexity, reliability

Your turn: Concepts in your research

Think about a concept you want to measure in your own research:

- How would you know if a text "contains" this concept?
- What words or patterns would indicate it?
- What would be a false positive? False negative?
- How context-dependent is the concept?

Reliability and the human baseline

The agreement problem

Before we ask “does the machine code correctly?”, we need to ask:

Do humans agree on what “correct” means?

This is harder than it sounds.

An exercise in disagreement

Consider these sentences about the economy:

1. “Inflation remains elevated but shows signs of moderating”
2. “The labor market is resilient despite headwinds”
3. “Consumer confidence fell less than expected”

Is the sentiment **positive**, **negative**, or **neutral**?

Why simple agreement is misleading

If two coders agree 80% of the time, is that good?

It depends on the base rate.

If 90% of documents are “negative”:

- Always guessing “negative” yields 81% agreement
- 80% agreement might be worse than chance

Inter-coder reliability measures

Cohen's Kappa (two coders)

$$\kappa = \frac{p_o - p_e}{1 - p_e}$$

- p_o = observed agreement
- p_e = expected agreement by chance

Krippendorff's Alpha (multiple coders, various scales)

- Handles missing data
- Works for nominal, ordinal, interval scales
- Generally preferred in content analysis

The "gold standard" problem

Who decides what's correct?

Options:

1. **Expert coding**: Authoritative but expensive, potential bias
2. **Majority vote**: Democratic but can miss subtle cases
3. **Adjudication**: Resolve disagreements through discussion
4. **Probabilistic labels**: Model uncertainty explicitly

Each has tradeoffs for training and evaluating automated methods.

Deductive vs. inductive approaches

Two philosophies of text classification

Deductive (theory-driven)

Start with predefined categories from theory

→ “I know what I’m looking for”

Inductive (data-driven)

Let patterns emerge from the text

→ “Show me what’s there”

Deductive approaches

Logic: Theory defines categories → operationalize as patterns → apply to texts

Classic example: Dictionary methods

- Sentiment dictionaries (positive/negative word lists)
- Policy dictionaries (Manifesto Project)
- Domain-specific dictionaries (financial, medical, political)

Strengths and weaknesses of deductive methods

Strengths

- Transparent and replicable
- Theoretically grounded
- No training data required
- Easy to understand and critique

Weaknesses

- Miss context: “not good” ≠ negative?
- Polysemy: “bank” (financial vs. river)
- Domain dependence: “viral” (disease vs. marketing)
- Fixed vocabulary: can’t adapt to new language

Inductive approaches

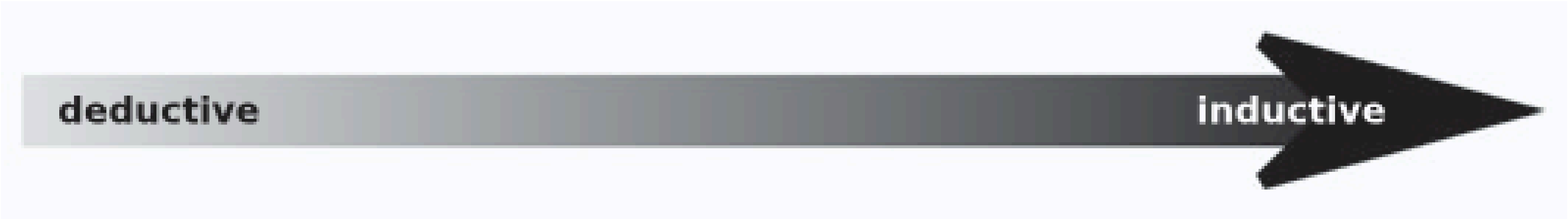
Logic: Analyze texts → discover patterns → interpret categories

Classic example: Topic models (LDA)

- No predefined categories
- Algorithm finds word clusters
- Researcher interprets what topics “mean”

Deductive vs. Inductive summary page

	Methodological approach		
	Counting and Dictionary	*Supervised Machine Learning*	*Unsupervised Machine Learning*
Typical research interests and content features	visibility analysis sentiment analysis subjectivity analysis	frames topics gender bias	frames topics
Common statistical procedures	string comparisons counting	support vector machines naive Bayes	principal component analysis cluster analysis latent dirichlet allocation semantic network analysis



Boumans and Trilling (2018)

Strengths and weaknesses of inductive methods

Strengths

- Discover unexpected patterns
- No need for predefined categories
- Can capture complex, corpus-specific structure
- Useful for exploration

Weaknesses

- Interpretation is subjective
- Results can be unstable (different runs → different topics)
- Researcher degrees of freedom in choosing parameters
- Hard to validate—what makes a topic “correct”?

Choosing your approach

Question	Deductive	Inductive
Clear concept definition?	✓ Required	✗ Not needed
Labeled data available?	✗ Not needed	✗ Not needed
Need transparency?	✓ High	✗ Lower
Exploratory research?	✗ Not ideal	✓ Good fit
Confirmatory research?	✓ Good fit	✗ Risky

Disciplinary context

In political science and economic sociology:

Deductive methods align with hypothesis-testing traditions

- Manifesto Project coding scheme
- Policy Agendas Project
- Lexicoder Sentiment Dictionary

Inductive methods align with interpretive traditions

- Discovering frames in media coverage
- Identifying discourse coalitions
- Exploratory analysis of new corpora

Wrapping up

Key takeaways

1. **Computational methods scale the measurement problem**—they don’t solve it
2. **Validity** concerns whether we measure what we intend
3. **Reliability** requires human agreement as baseline
4. **Deductive vs. inductive** reflects deeper epistemological choices
5. **Every preprocessing and modeling choice matters**

Preview: Day 2 afternoon

Tomorrow afternoon we'll get hands-on with:

- **Classification metrics**: precision, recall, F1
- **Document-term matrix**: representing texts as numbers
- **TF-IDF**: weighting words by importance
- **Bag-of-words models**: the foundation of text classification
- **Bridge to embeddings**: where we're headed next

A question to take with you

Think about your own research:

What’s one concept you want to measure in text?

- Do you lean deductive or inductive?
- What would reliable human coding look like?
- What errors would be most costly for your inference?

References I

- Boumans, J.W. and Trilling, D., 2018. Taking stock of the toolkit: An overview of relevant automated content analysis approaches and techniques for digital journalism scholars. *Rethinking research methods in an age of digital journalism*, pp.8-23.
- Grimmer, J., & Stewart, B. M. (2013). Text as data: The promise and pitfalls of automatic content analysis methods for political texts. *Political Analysis*, 21(3), 267-297.
- Kozlowski, A.C., Taddy, M. and Evans, J.A., 2019. The geometry of culture: Analyzing the meanings of class through word embeddings. *American Sociological Review*, 84(5), pp.905-949.
- Krippendorff, K. (2018). *Content Analysis: An Introduction to Its Methodology*. Sage.
- Michel, J.B., Shen, Y.K., Aiden, A.P., Veres, A., Gray, M.K., Google Books Team, Pickett, J.P., Hoiberg, D., Clancy, D., Norvig, P. and Orwant, J., 2011. Quantitative analysis of culture using millions of digitized books. *Science*, 331(6014), pp.176-182.

References II

- Proksch, S.O., Lowe, W., Wäckerle, J. and Soroka, S., 2019. Multilingual sentiment analysis: A new approach to measuring conflict in legislative speeches. *Legislative Studies Quarterly*, 44(1), pp.97-131.
- Rooduijn, M., & Pauwels, T. (2011). Measuring populism: Comparing two methods of content analysis. *West European Politics*, 34(6), 1272-1283.