

CAUSAL INFERENCE IN NATURAL LANGUAGE PROCESSING FOR ECONOMICS

(MSc in Economics, WiSo Universität Hamburg)

Summer Semester 2022

Instructor : Dr. Huyen Nguyen

Website: <http://huyenttnguyen.com>

E-mail: huyen.nguyen-1@uni-hamburg.de

COURSE SUMMARY

This course introduces causal inference techniques for social scientists through the lens of applied microeconometrics. The ability to determine causal pathways is instrumental for social science research and holds the key to understand the effects of policy. The ever-increasing availability of large-scale text corpora and computing power, along with advances in methods theory has boosted researchers' ability to estimate causal relationships, not just between policy variables and demographic information, but also hidden novel insights about relevant institutional and human patterns in texts.

By the end of the course, you are able to apply experimental and quasi-experimental econometric methods for the causal analysis of public policies, especially in the context of text as data. You can interpret complex econometric models and their results. Furthermore, you can develop an empirical strategy for causal identification and to select and apply the appropriate econometric method for a specific research question. By critically reading and evaluating state-of-the-art research articles in text as data application within a causal framework, you become familiar with developing and testing your own research questions and performing your own empirical work.

The first 4 lectures will go through the key fundamentals of contemporary causal inference techniques. Topics include, for instance: experiments, matching and weighing, instrument variables, regression discontinuity designs, panel data (differences-in-differences and synthetic control methods).

The subsequent lectures will explore various applications of Text as Outcome, as Treatment, and as Mediator, in the form of flipped classrooms. Examples include, but are not limited to: how texts on job ads impact different types of applicants; how the market reacts to FOMC releases; how deliberation affects opinion changes on a controversial topic. These sessions consist of a short lecture, followed by either: (1) a presentation by groups of 2-3 students on the required readings of that week and (2) critical discussions from remaining class members; or (2) an interactive coding exercise session in groups.

COURSE OBJECTIVES

- Comprehensive overview of contemporary causal inference methods in social science policy questions, especially in the context of text as data.
- Familiarity with statistical and practical issues around text data in policy impact evaluation.
- Ability to critically present relevant scientific articles/applications to peers and discuss presentations of other participants.
- Ability to fit, interpret and apply causal inference analysis techniques in chosen policy contexts to independent research projects.

PRE-REQUISITES

This course requires a basic understanding of Python, Stata, statistics, probability theories and applied econometric techniques used in social sciences. As a pre-requisite, participants are required to be familiar with:

- (1) the basics of text as data in Python (text data acquisition, cleaning, feature reduction, and basics of relevant supervised & unsupervised ML methods)
- (2) Stata (at the level obtained after BSc Econometrics courses)

You are encouraged to self-study/check your Python/text mining understanding from the previous course: <https://github.com/htn21uhh/Text-Analysis-for-Social-Sciences-in-Python>

ALL students are required to check their math and stats background. you don't have a solid background in calculus, linear algebra, and probability, read [part 1 from this online book](#).

COURSE LOGISTICS & HOUSE RULES

- 1) **Lectures** are held in Room WiWi 2019/2201 in VMP5 OR, unless otherwise noted, on Zoom **every Tuesdays from 08.15 to 10.45AM**, from 05.04.2022 to 12.07.2022.

There are two *exceptions*:

- (i) Tuesday 24.05: Pfingsten break.
- (ii) Tuesday 31.05: There will be individual group consultation appointments to discuss your mid-term project proposal assignment (more details below).

Lecturers are roughly structured as follows: 60' of lecture → 15' break → either:

- (i) 60' exercise. (Week 2 to 4)
- (ii) 30' of interactive presentation (from own groups/peers) → 30' of evaluation & discussion. (Week 5 onwards)

Students are expected to read the required readings and any exercises before class, as well as actively participating in all classes. Attendance on at least 11/14 sessions is expected.

2) **Materials** (slides, exercise links, data sources, tips) are uploaded **on GitHub**

https://github.com/htn21uhh/Causal_Inference_NLP_Economists

- **Lecture slide** will be added to GitHub → Lectures folder every Wednesday morning, unless otherwise noticed. The list of required readings and exercise links are in the detailed syllabus overview below.
- **Exercises** (on Jupyter notebook/Google Colab and Stata) that are used in class (and accessible after class for your own practice) will be updated in Github → Exercises folder.
- **Papers** (required readings for presentations & peer critic presentation each week, from Week 5 onwards):
 - Most of the readings can be found in this comprehensive list <https://github.com/causaltext/causal-text-papers>.
 - Exact/Any other readings per week can be found in the Group Matching Google Sheet OR the Syllabus below.
- **Data sources and tips** are updated regularly in the detailed Syllabus below as the course proceeds.

3) **Forum** (introduction, teammate mix-and-match, exchange of ideas, code bugs) **on Slack channel:**

https://join.slack.com/t/uni-hamburg-world/shared_invite/zt-16xt0sqhs-Og~ctCHgWoZiYujAgnHCXw

How to join in this Slack channel? Use your @uhh account to create an account, register and log into the channel (Desktop apps are also downloadable) using the above link.

There are 4 channels, as follows:

- **#introduce_yourself**: [COMPULSORY, **during 1st session**] Introduce yourself, your course expectations and levels of understanding for text as data and causal inference framework
- **#group_mix_and_match** [COMPULSORY, **deadline in Week 3**]: For the paper presentation, peer project deliverables

- *#qna_code_bugs*: The Q&A space for code issues, solutions, Stackoverflow/Stata materials and any useful exercises you found.
- *#research_ideas_plans*: The space to discuss/pitch research project ideas, interesting papers, planning, and recording resources.

NOTE: Announcement and notices will be done mainly via STINE messages.

4) Course setup

- From Week 2 to Week 4, we will go through the step-by-step exercises together in class, and you need to go through any remaining parts on your own before the next lecture.
- From Week 5 onwards, assigned groups X needs to prepare a 30' mini lecture/presentation on the required readings (research paper(s) applying text in a causal context in real life); whereas assigned group Y will need to prepare a critique discussion of group X presentation (30'), along with relevant questions and discussions of the readings (30'). The group pair X-Y is as follows.

| WEEK | Presenting Group | Response Group |
|-----------------|-------------------------------------|----------------|
| Week 5 (03.05) | G1 | G8 |
| Week 6 (10.05) | G2 | G7 |
| Week 7 (17.05) | G3 | G6 |
| Week 8 (31.05) | Group Research Project Consultation | |
| Week 9 (07.06) | G4 | G1 |
| Week 10 (14.06) | G5 | G2 |
| Week 11 (21.06) | G6 | G3 |
| Week 12 (28.06) | G7 | G4 |
| Week 13 (05.07) | G8 | G5 |

- There can be 8 groups in total, with max 3 students/group. You can find your group members with similar research interests on the Slack channel, and then sign up in the following Google Sheet
<https://docs.google.com/spreadsheets/d/1EEUJp1SrZ05BwEUH2u3CbYwMykmgEUhZdn2SqrmeaRs/edit?usp=sharing>
(!)Kindly **DO NOT CHANGE** the responses of other classmates
- Based on your indicated group interest, I will assign suitable papers for your group to read and prepare a mini-lecture (30'/group) to your class on the assigned date. Another group will prepare a response/critic presentation and a discussion plan of this group(30')
- Your tasks as a group include:
 - Midterm pitch recording (5') of your NLP causal inference research design
 - Midterm 1-page research design proposal
 - Final 25' research project presentation

- Final proof-of-concept analysis plan
- Research Paper presentation (30')
- Peer Response presentation & discussion (30')

COURSE GRADING POLICY & WORKLOAD

All course deliverables are done in **groups of three students**. In general, your course commitment is on average between 7.5 – 10 hours/week, depending on your starting level.

The course will be graded on the usual grading scale with passing grades from 1.0 (very good) to 4.0 (sufficient), and with a failing grade 5.0 (insufficient). The grades are determined as follows:

- 30%: Midterm pitch recording (5') + 1-page NLP causal inference research design proposal

(Deadline: 23.59 Tuesday 18.05)

- 40%: Group research project oral presentation (25') + proof-of-concept analysis plan
(Deadline: 23.59 Thursday 30.06)
- 15%: Research paper presentation (30')
(own group chosen date, from Week 5 onwards)
- 15%: Peer evaluation oral presentation (30')
(own group chosen date, from Week 5 onwards)

Submission rule:

Paper presentation: For presentation groups, please discuss with your assigned peer evaluation group in that same week to arrange such that you hand over your presentation slides to them promptly. This allows them to properly prepare a response presentation and discussion plan, right after your presentation.

Midterm & final submissions: For every hour of late submission of the midterm pitch & proposal, everyone in the team will receive a 5% reduction of the score. Should you have verifiable extenuating circumstances (e.g. illness, personal loss, hardship, or caring duties), an extension can be granted. In such cases, please contact the course instructor as soon as possible before the deadline.

You can find more detailed instructions on the pitch, proposals, the consultation sessions in a separate file in Github/Open Olat.

GETTING STARTED

1) Python

The examples in the course will mainly use Python. Additionally, when we move to regression exercises, Stata might be used in economic examples. You are strongly recommended to install and set up Python before the course, using the instructions in the following links:

[Python Setup Instructions](#)
[Codecademy Online Python Course](#)

2) Jupyter Notebook

Read the following step-by-step blog to get familiar with Jupyter Notebook.

<https://medium.com/codingthesmartway-com-blog/getting-started-with-jupyter-notebook-for-python-4e7082bd5d46>

For installation and trying out Jupyter Notebook, read this:

<https://jupyter.readthedocs.io/en/latest/install/notebook-classic.html>

<https://satyaborg.com/posts/accessing-virtualenv-in-jupyter-notebooks>

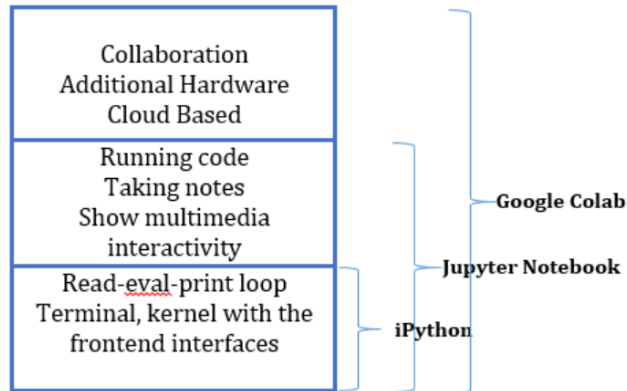
3) Google Colab

If Jupyter Notebook on the web does not work for you during class exercise sessions, [use Google Colab](#).

Google Colab is a specialized version of Jupyter Notebook, which runs on the cloud and offers free computing resources. It does not require a setup, plus the notebooks that you will create can be simultaneously edited by your team members – in a similar manner you edit documents in Google Docs. The greatest advantage is that Colab supports most popular [machine learning libraries](#) which can be easily loaded in your notebook. In order to use Google Colab, you will need a Google Account. Your files are by default linked to your Google Drive. For more information and instructions, please check the following link:

https://colab.research.google.com/?utm_source=scs-index

The following diagram summarizes the relationship between iPython, Google Colab and Jupyter Notebook:



4) **StackOverflow** <https://stackoverflow.com/>

Your go-to website (in addition to Google) for ANY programming bug, installation, package issues. Good programming and data wrangling skills = good Googling/StackOverflow skills.

NOTE! I will IGNORE any questions and e-mails related to programming/installation/package issues. The answers are highly likely available on StackOverflow or from your classmates (use Slack forum to ask, or ask them directly in class).

SYLLABUS & MATERIALS

This detailed syllabus is subject to changes as the course proceeds. I will update it regularly on the Github repository, please check it on a weekly basis.

| Week | Date & location | Topic & Content | Readings | Practice |
|------|-----------------|--|---|---|
| 1 | Tue 05.04 | <p>Introduction and course logistics</p> <p>Causal inference ft. text as data overview</p> | <p>[Lecture video on Causal Inference, MIT CourseWave] https://www.youtube.com/watch?v=gRkUhg9Wb-I</p> <p>The State of Applied Microeconometrics: Causality and Policy Evaluation (Athey & Imbens, JEP 2017)</p> | <p>[Python] A beginner's guide to Python https://wiki.python.org/moin/BeginnersGuide</p> <p>[Python] Python for Data Science Cheat Sheets</p> <p>[Text analysis]</p> |

| | | | | |
|---|---|--------------------------------------|--|---|
| | | | | <p>Check Chapter 4 to 6 of the NLTK book</p> <p>Text Analysis Starter Guide</p> <p>[Stata] Treatment effects/causal inference video library https://www.stata.com/features/treatment-effects/</p> <p>[OPT] Stata to Python equivalent http://www.danieldsullivan.com/pages/tutorial_stata_to_python.html</p> |
| 2 | Tue 12.04 | Causal inference – Text as Treatment | <p>Text and Causal Inference: A Review of Using Text to Remove Confounding from Causal Estimates Katherine A. Keith, David Jensen, and Brendan O'Connor</p> <p>How to Make Causal Inferences Using Texts Naoki Egami, Christian J. Fong, Justin Grimmer, Margaret E. Roberts, and Brandon M. Stewart</p> | Tutorial: Causal inference and Machine Learning in Practice with Industry cases (Microsoft, Tripadvisor and Uber) |
| 3 | Tue 19.04 DEADLINE @ 23.59: group matching + paper choice | Causal inference – Text as Outcomes | | |
| 4 | Tue 26.04 | Causal inference – | | |

| | | | | |
|----|---|---|--|--|
| | | Text as Mediators | | |
| 5 | Tue 03.05 (presentationG1 + response G8) | Causal inference – Application Text as Outcomes | TBD (based on interests) | |
| 6 | Tue 10.05 (presentationG2 + response G7) | Application Text as Treatment | | |
| 7 | Tue 17.05 (presentationG3 + response G6) <u>DL: midterm pitch & research design</u> | Application Text as Treatment | TBD (based on interests) | |
| | Tue 24.05 | Pfingsten Break | | |
| 8 | Tue 31.05 | Consultation on project midterm deliverables | | |
| 9 | Tue 07.06 (presentation G4 + response G1) | Application Text as Treatment | | |
| 10 | Tue 14.06 (presentationG5 + response G2) | Application Text as Outcome | | |
| 11 | Tue 21.06 (presentationG6 + response G3) | Application Text as Outcomes | | |
| 12 | Tue 28.06 (presentation G7 + response G4) | Application Text as Mediators | | |
| 13 | Tue 05.07 (presentation G8 + response G5) | Application Text as Mediators | | |
| 14 | Friday 15.07 (or TBD date) (FINAL PRESENTATION) | | Check your <u>assigned group time slot & Zoom link</u> [GOOGLE DOC] | |

SUPPLEMENTARY MATERIALS FOR DATA ANALYSIS

The course will be mainly based on the weekly lecture slides and research articles, but the following books and code exercises can be used as reference along with the slide content.

Natural Language Processing in Python, Third Edition, available at nltk.org/book.

- Aurelien Geron, Hands-On Machine Learning with Scikit-Learn & TensorFlow, O'Reilly 2017 ([link](#))
- [Jupyter notebooks Github for Geron's book](#).
- [Google Developers Text Classification Guide](#) (This guide contains some practical tips and code examples for using text data)
- For Python syntax programming, this book Fluent Python (O'Reilly 2015) serves as a good guideline
<https://www.oreilly.com/library/view/fluent-python/9781491946237/>