Data Literacy 2022*
Syllabus

Simon Döbele†
Konstantin Lehmann‡
Sebastian Zezulka§

03/05/2022

1 Seminar Overview

Scientific, policy, as well as corporate decisions should be based on good evidence and solid reasoning. Often, this requires decision makers to consider large amounts of data and to use quantitative methods to analyse them. This course equips students with solid foundations to apply, communicate and reflect on quantitative methods. It enables students to become data literate along three dimensions. First, an introduction to standard machine learning methods. Second, practical skills in structuring a data science project and communicating the results. Third, understanding the philosophical reflections on the applied methods. A special emphasis of the course is to give students a very first hands-on experience with Python. On Sunday, we will offer three parallel workshops on philosophy of statistics, practical data visualisation, and a data science project.

2 Dates, Expectations, and Requirements

Participation in the online seminar *Introduction to Python* on Friday, 06.May 2022, 09.00-12.00h and the weekend seminar from 10.-12.June 2022 in Bayreuth is obligatory. A non-mandatory online Q&A session will be offered on Wednesday, 25.May 2022, 17.00-18.30h to help with the data science practicals.
To be successful in this course, students must come prepared to class. A student is adequately prepared if she has carefully completed all of the *required* readings or practical assignments and has compiled some comments or questions. **All data science practicals and short answers are due by Friday, 10.06.2022 at 10am sharp.**

*Seminar 50033, Philosophy & Economics, Uni Bayreuth
†simon.doebele@student.kit.edu
‡konstantin.lehmann@tu-dortmund.de
§sebastian.zezulka@student.uni-tuebingen.de

2.1 Preparation for Introduction to Python - May 6th

Please join the discord-server, which serves as a forum for you to ask each other and the instructors questions and discuss solutions to the coding tasks among you.
Please ensure that you have a google account (if you do not have one yet). This is needed so that you can open a Colab notebook. Test your success by following this link and clicking: $File \rightarrow NewNotebook$.

2.2 Requirements

- 2 CP: You need to (1) do the *required* readings for the philosophy and machine learning part, (2) answer the respective questions regarding those readings and hand them in by June 10th, 10am (submit all answers as a single pdf), (3) submit the data science practicals (Colab and Datacamp) by June 10th, 10am, and (4) participate actively in the seminar.
- 5 CP: Same as 2 CP + (5) write either a philosophical essay or a report on a data science project you did *after* the seminar.

2.3 Colab Coding Practicals

These practicals will be described in the respective Colab notebooks. They consist of the following two Colab notebooks:

1. Python Tutorial (with the tasks given to you during the Tutorial, to be submitted after the tutorial and before the seminar)
2. General Python Programming (will be sent to you in a separate e-mail)

2.4 Datacamp Coding Practicals

For parts of the seminar we use datacamp. If you follow this link, you can create a *free* datacamp-account using your Uni Bayreuth mail account. There you find assignments you need to complete before the 10.June, 10am sharp. The order of the assignments is:

1. NumPy
2. Dictionaries & Pandas
3. Loading Data in pandas
4. Plotting Data with Matplotlib
5. Python Data Science Toolbox (Part 1)
6. Exploratory Data Analysis
7. Introduction to Data Visualization with Seaborn

8. Regression
9. Classification

Of course you can do other courses on datacamp as well.

2.5 Presentations

Students can earn a bonus of 0.3 on their grade. To do so, they have to give a short presentation in class on one of the following topics.

- **For the philosophy of statistics part:** present one of the required readings from Sessions 3.3, 3.5, or 3.9. Please end the presentation with 2/3 questions or critical remarks that serve as an opener for the debate in class.
- **For the data visualisation part:** present one of two topics on data visualization for Session 3.7:
    1. Correct representation of data. Sources are Franconeri et al. (2021). "Avoid common illusions and misperceptions" (p.118-119), and the two articles of week 6 of the course Calling Bullshit: "Misleading axes" and "Proportional Ink".
    2. Properly visualising uncertainty. Source here is again Franconeri et al. (2021). "Common Visualisation of uncertainty are often misinterpreted" (p.140-147).
- **For the machine learning part:** you may choose one of the following four topics (1a, 1b, 2a or 2b). Please end the presentation with 2/3 questions that serve as an opener for the debate in class.
    1. If you prefer to talk about Logistic Regression, given the required readings in sections 3.6:
        (a) Explain the different types of gradient descent, their respective advantages and disadvantages.
        (b) Give an overview of regularization methods and why we use them.
    2. If you prefer to talk about Clustering, given the required readings in sections 3.8:
        (a) Read also subchapters 7.3.4 and 7.3.5 in the respective required reading (4 pages). Give an overview of the BFR algorithm and compare it to k-means.
        (b) Read also subchapter 7.4 in the respective required reading (4 pages).Give an overview of the CURE algorithm and compare it to k-means.

The presentation should be brief (7-10 min) and cover the main argument of the assigned text. No slides are required. Registration for the individual presentations takes place *after* the *Introduction to Python* seminar on Friday, 06.05.2022 via discord.

3 Reading List

3.1 Introduction to Python Tutorial

Friday, 06.05., 9am s.t. to 12 noon

Required readings:

- None.

Further readings and Course recommendations:

- The introduction is heavily based on two courses: Harvard University's CS50 and UC Berkeley's CS61A. You can find many more programming exercises there.
- Composing Programs - An Introduction to Python
- H. Abelson, G. Sussman, and with Julie Sussman. MIT Press/McGraw-Hill, Cambridge, 2nd Editon edition, (1996)
- O'Reilly: Learning Python. Lutz, Mark (2013)
- Deeper Understanding of numpy arrays

3.2 Q&A session: data science practicals

Wednesday, 25.05., 17h s.t.

Required readings:

- None. Please prepare questions and if possible share them on discord before the session.

3.3 Inductive Risk and Value Judgments

Friday, 10.06., 12h c.t.

Required readings:

- Schurz, Gerhard (2019). Hume's problem solved: The optimality of meta-induction. MIT Press. **Chapter 1.**
- Steele, Katie (2012). The scientist qua policy advisor makes value judgments. Philosophy of Science, 79(5), 893-904.
  - Sketch Hume's second argument against the justification of inductive inferences from observed regularities as presented by Schurz (2019).
  - If policy-makers were also scientific experts, would the problem of inductive risk be solved?

Further readings (non obligatory):

- Rudner, Richard (1953). The scientist qua scientist makes value judgments. Philosophy of science, 20(1), 1-6.
- Biddle, J. B., and Kukla, R. (2017). The geography of epistemic risk. Exploring inductive risk: Case studies of values in science, 215-237.
- Parker, W. (2014). Values and uncertainties in climate prediction, revisited. Studies in History and Philosophy of Science Part A, 46, 24-30.

3.4 Probability Theory and basic Statistics

Friday, 10.06., 14h c.t.

3.5 The value-free ideal

Friday, 10.06., 16h c.t.

Required readings:

- Bright, Liam Kofi (2018). Du Bois' democratic defence of the value free ideal. Synthese, 195(5), 2227-2245.
  - Name and briefly explain the three lines of argument against the *value-free ideal* presented in Bright (2018).
  - Why did Du Bois defended the *value-free ideal*? Give one argument/reason.

Further readings (non obligatory):

- Anderson, E. (2004). Uses of value judgments in science: A general argument, with lessons from a case study of feminist research on divorce. Hypatia, 19(1), 1-24.

- Douglas, H. (2017). Science, values, and citizens. In Eppur si muove: Doing history and philosophy of science with Peter Machamer (pp. 83-96). Springer, Cham.
- Holst, C., and Molander, A. (2017). Public deliberation and the fact of expertise: making experts accountable. Social Epistemology, 31(3), 235-250.

3.6 Machine Learning I: Supervised Learning

Saturday, 11.06., 10h c.t.

Required readings:

- Chapters 5 - 5.9 and 5.11 of: Speech and Language Processing (3rd ed. draft). Dan Jurafsky and James H. Martin (2021)
  - Explain (in two sentences) when you use the sigmoid function in logistic regression and when the softmax function.
  - Think of how a sentiment classification algorithm could be used for a start-up in the area of financial engineering, a.k.a. predictive finance. You should describe what you would input to the algorithm, what the outputs are and why that is of value to the start-up. (open-ended).

Further readings (non obligatory):

- Chapters 4 - 4.3 of: Gareth James, Daniela Witten, Trevor Hastie, Robert Tibshirani. An Introduction to Statistical Learning: with Applications in R. New York: Springer, 2013.
- (Math-heavy:) Chapters 4.4 - 4.4.4 of Hastie, T., Tibshirani, R.,, Friedman, J. (2001). The Elements of Statistical Learning. New York, NY, USA: Springer New York Inc.
- if you need some recap on linear algebra (especially for the lab): CS229 Machine Learning
- (Advanced:) Deisenroth, M. P., Faisal, A. A.,, Ong, C. S. (2020). Mathematics for Machine Learning. Cambridge University Press.
- (Advanced:) Bishop, Christopher M. (2006). Pattern recognition and machine learning. New York: Springer, 2006.

3.7 Data Visualisation

Saturday, 11.06., 12h c.t.

Required Readings:

- None.

Further readings (non obligatory):

- Franconeri et.al (2021). The Science of Visual Data Communication: What Works. In Psychological Science in the Public Interest, 22(3), 110-161, https://journals.sagepub.com/doi/10.1177/15291006211051956

3.8 Machine Learning II: Unsupervised Learning: Clustering

Saturday, 11.06., 14h c.t.

Required readings:

- Chapters 7 - 7.2.3 and 7.3 - 7.3.3 of: Mining Massive Datasets A. Rajaraman, J. Leskovec, and J. Ullman. (2014)
  - Describe in two sentences the two options of how to initialize the clustroids in k-means.
  - Think about the following thought experiment: You are in a d-dimensional space (where $d \geq 2$). If you had to choose between a measurement device that tells you in which direction to go vs. a measurement device that tells you how far you are from your goal, which of the two would you choose? Provide a drawing that explains your reasoning.

Further readings (non obligatory):

- Chapter 12.4 of: Gareth James, Daniela Witten, Trevor Hastie, Robert Tibshirani. An Introduction to Statistical Learning: with Applications in R. New York: Springer, 2013.

3.9 Fairness in Machine Learning

Saturday, 11.06., 16h c.t.

Required readings:

- Barocas, Solon; Moritz Hardt, and Arvind Narayanan (2019). Fairness and Machine Learning. **Chapter 2: Classification**.
  - Why does it not suffice for fair outcomes to remove all sensitive features (such as race or gender) from the data sets?

– Give a short and intuitive description of the three fairness criteria *Independence*, *Separation*, and *Sufficiency*. One/Two sentences per criterion are sufficient.

Further readings (non obligatory):

- Binns, Reuben (2018). Fairness in Machine Learning: Lessons from Political Philosophy. In *Proceedings of Machine Learning Research 81:149-159*. `https://proceedings.mlr.press/v81/binns18a.html`.
- Williamson, R. (2020). Process and Purpose, Not Thing and Technique: How to Pose Data Science Research Challenges. Harvard Data Science Review. `https://doi.org/10.1162/99608f92.6e525663`.
- Green, B., & Hu, L. (2018). The myth in the methodology: Towards a recontextualization of fairness in machine learning. In *Proceedings of the machine learning: the debates workshop*.

3.10 Workshop: Data Visualisation

Sunday, 12.06., 10-14h

- Please create an account on kaggle.com.

3.11 Workshop: Data Science Project

Sunday, 12.06., 10-14h

Required readings:

- None. It is recommended to do some more programming exercises, as described in the further readings under 3.1.

Further readings (non obligatory):

- If you want to work more like a professional data scientist, you may download Visual Studio Code or even PyCharm using their Free Educational Licences. If you want, you can try to open your .ipynb notebooks there (google how to).
- Then, you may try to convert your .ipynb files to python files (simply by copy pasting python code into a new file with .py ending).
- Next, read about Defining Main Functions in Python and on that same link, execute.
- You may even explore and create an account on Github: google how to use git add, git commit and git push from the command line, and publish your Colab Notebooks (or even .py files) so you can showcase your work and collaborate with others.

- Note: You do *not* have to hand-in anything for this. It is for your own practice and might make the workshop more engaging.

3.12 Workshop: Philosophy of Statistical Inference I

Sunday, 12.06., 10h c.t.

Required readings:

- Nosek, B. A., et al. (2022). Replicability, robustness, and reproducibility in psychological science. Annual review of psychology, 73, 719-748.

Further readings (non obligatory):

- Wasserstein, R. L., Schirm, A. L., & Lazar, N. A. (2019). Moving to a world beyond "$p < 0.05$". The American Statistician, 73(sup1), 1-19.
- Ioannidis, J. P. (2005). Why most published research findings are false. PLoS medicine, 2(8), e124.
- Romero, Felipe, and Sprenger, Jan (2021). Scientific self-correction: the Bayesian way. Synthese, 198(23), 5803-5823.

Introductions to Philosophy of Statistics (more general):

- Sober, E. (2008). Evidence and evolution: The logic behind the science. Cambridge University Press. **Chapter 1**.
- Romeijn, Jan-Willem, "Philosophy of Statistics", The Stanford Encyclopedia of Philosophy (Spring 2022 Edition), Edward N. Zalta (ed.).
- Mayo, D. G., & Cox, D. R. (2006). Frequentist statistics as a theory of inductive inference. In *Optimality* (pp. 77-97). Institute of Mathematical Statistics.
- Liu, K., & Meng, X. L. (2016). There is individualized treatment. Why not individualized inference?. Annual Review of Statistics and Its Application, 3, 79-111.

3.13 Workshop: Philosophy of Statistical Inference II

Sunday, 12.06., 12h c.t.

Required readings:

- Mayo, Deborah (2020). P-values on trial: Selective reporting of (best practice guides against) selective reporting.

Further readings (non obligatory):

- Sprenger, J. (2016). Bayesianism and Frequentism in Statistical Inference.
- Imbens, G. W. (2021). Statistical significance, p-values, and the reporting of uncertainty. Journal of Economic Perspectives, 35(3), 157-74.
- Betensky, R. A. (2019). The p-value requires context, not a threshold. The American Statistician, 73(sup1), 115-117.

3.14 Seminar Closing

Sunday, 12.06., 14h c.t.