Data challenge 2 (JBG050)

Handbook

Bennett Kleinberg, Laura Genga

Last update: 23 April, 2024

Welcome to this year's Data Challenge 2 course (JBG050)!

This handbook provides the details necessary to start this course, the structure and timeline of the course, and the grading criteria with which we will assess your work.

The course builds on your previous data challenges and the courses you took as part of your BSc in Data Science. A key difference to previous data challenge courses is that you are now working on a real problem provided by stakeholders, which implies that you must first translate the stakeholder's problem into a *data-science problem*.

This year's problem was designed in collaboration with the Metropolitan Police Service (MPS) in London and the London Mayor's Office for Policing and Crime (MOPAC). You will have the chance to ask the key decision-makers within these organisations questions.

1 Problem description

The topic for your project occupies law enforcement agencies and - at a broader level - national policy-making in many countries, including the United Kingdom. While there has been a sole focus on solving and reducing crimes for a long, a shift is underway so that law enforcement organisations increasingly include trust and confidence in their priorities. MOPAC lists trust and confidence as one of their key priorities as follows:

> Making sure that all Londoners feel that the police treat you fairly, provide a professional service and are dealing with the local crime and antisocial behaviour issues that matter to you.

Understanding what drives trust and potentially predicting trust in the police could help key decision-makers focus time and resources on areas and problems that need attention (e.g., where trust is corrosive or generally low). Similarly, on the operational level, a police officer who is informed about the (lack of) trust in their neighbourhood can focus on those constituents of trust that need attention. Put simply, a nuanced understanding of trust is essential for a modern police force and a livable global city like London.

Your task as a group is to contribute to this issue using data science techniques. At the same time, we want you to consider the ethical implications that your decisions could have when applied in real life.

We will provide you with data from the Public Attitude Survey (PAS), a detailed survey of 19,200 London residents and point you to crime data. Your task is to contribute to a better understanding of trust and confidence in the police in London. The PAS and crime data (see below) are the starting point for your analysis.

The specific overarching question for your project is: **How can data science contribute to a better understanding of the determinants and predictors of trust and confidence in the Metropolitan Police Service London?**

After reading through the current document, you will have the chance to ask questions to the teaching team and to the stakeholder (The Metropolitan Police London) in the introductory plenary session.

1.1 Expectations

To answer this question most usefully, your projects should focus on specific sub-questions of that overarching question. Rather than answering many sub-questions, it may be wise to choose just one or two and provide a deep dive into them. Your project should also result in a clear proposal of how the MPS can address a low or altogether lack of trust and confidence.

For the most viable solutions, your group should have an eye on what is known about trust and confidence (see the required background reading).

1.2 Questions to consider

The questions below function as inspiration. You can use them, adapt them, or simply take them as initial ideas. You do not have to make them a central part of your project, but you are free to do so if you want to.

1. How can non-traditional variables predict trust and confidence? Is there any signal in open-source data beyond the already known predictors, such as demographic variables?
2. To what extent can we predict "soon to be much lower in trust" areas? Similar work has been done on violent crimes (e.g., Sutherland et al., 2020)
3. How do changes in crime predict trust? How does this differ between different regions in London?

You may also want to think about the following for your analysis and project outcomes:

- incorporating (all or a selection of types of) crime (counts or volume) in your analysis
- using an adequate and justifiable spatial and temporal granularity in your data
- ensuring your data aggregation is meaningful - both methodologically and from a practical angle

1.3 The data

Your initial source of data for this project is the Public Attitude Survey and a dataset (obtainable from the open data archive of https://data.police.uk/) that spans all police-reported crimes since December 2010 in the whole of England, Wales and Northern Ireland (in total more than 70 million cases). In the crime dataset, each reported crime is defined as belonging to a specific crime type (e.g., violent crime, burglary) and provided with spatiotemporal details. Importantly, to preserve the privacy of affected individuals, the data are aggregated to a monthly level (e.g., 3rd of Feb, 2015 is aggregated to Feb, 2015) and - on the spatial dimension - to a *lower-super output area* (LSOA)[1]. The LSOAs are units provided by the Ordnance Survey and used for census-related purposes. You can find details about that anonymisation procedure below in the suggested reading. In England, there are currently 32,844 LSOAs.

The PAS data are provided at the borough level.

For your project, you are free to use other official and publicly available datasets that could help you provide better insights for the stakeholder. For example, the UK Office of National Statistics makes available so-called *deprivation indices* which measure societal well-being for each LSOA, and you can obtain various other datasets on LSOAs.

1.4 Accessing the data

- The Public Attitude Survey data are available via https://data.london.gov.uk/dataset/mopac-surveys
- You can access the data for download from the police.uk data repository via this link: https://data.police.uk/data/

[1] Before that spatial aggregation is happening, each case's geospatial coordinates are slightly randomised.

- In addition to crime (and anti-social behaviour) incident data, the police.uk data archive also provides you with the case outcomes (e.g. whether a crime incident went to court) and stop-and-search data. You can - but do not have to - incorporate and/or use that data for your project.
- The London Datastore has plenty of relevant data sources that can be accessed and might be suitable for your specific sub-questions.
- The Office for National Statistics also offers a range of high-quality datasets that could be useful depending on your sub-question(s).
- Data on spatial boundaries of police forces and neighbourhoods: https://data.police.uk/data/boundaries/
- Additional data that could be useful is the UK's societal well-being data measured through the "index of multiple deprivations" (IMD). You can find these data here: https://opendatacommunities.org/def/concept/folders/themes/societal-wellbeing. More information on the IMD can be found at https://www.gov.uk/guidance/english-indices-of-deprivation-2019-mapping-resources#indices-of-deprivation-2019-explorer-postcode-mapper

You can use any other official, publicly available data for your project.

1.5 Background reading

- The anonymisation procedure of the police.uk data: https://data.police.uk/about/#anonymisation
- Tompson et al. (2014). UK open source crime data: accuracy and possibilities for research. Cartography and Geographic Information Systems. https://www.tandfonline.com/doi/full/10.1080/15230406.2014.972456
- Hough, M. (2012). Researching trust in the police and trust in justice: A UK perspective. Policing and Society, 22(3), 332–345. https://doi.org/10.1080/10439463.2012.671826
- Jackson, J., & Bradford, B. (2010). What is Trust and Confidence in the Police? Policing, 4(3), 241–248. https://doi.org/10.1093/police/paq020

Reference guide on forecasting:

- Hyndman & Athanasopoulos (2018). Forecasting: Principles & Practice. Freely available at https://otexts.com/fpp2/[2]

1.6 Software/resources for this project

You are free to use any software that you have access to. As a group, you will have to share your documented code with us via GitHub at the end of this course.

2 London trip

In the past years for this course, the two best groups travelled to London and presented the findings to an audience of decision-makers from a policy, operational and strategic level of the Metropolitan Police at their headquarters in New Scotland Yard.

This trip is again planned for this year. The universities (TiU and TU/e) will pay for for the trip (travel, accommodation and subsistence) and the two winning groups will be able to fine-tune their presentation skills in a workshop before the London trip. Details about this trip will follow.

[2]This book is R-based, but the principles and underlying statistical foundations are useful even if you work in python

3 Structure of this course

3.1 Differences to previous "Data Challenges"

This course is set up with uncertainty by design.

The fundamental objective of this course is to take you a step further to a real-world Data Science project with all the issues and uncertainties it brings.

In prior Data Science courses, including the other Data Challenges, you have been learning Data Science in an academic environment as follows: I am given problem X to solve, so I have to find the solution for X.

Data Science practice is nothing like that at all.

In practice, stakeholders state X as the final objective to solve, but along the way, you figure out that before solving X, you first have to solve Y and Z, and then time runs out. This makes real-life data science projects messy and introduces uncertainty. The outcome is often not reaching the objective but moves you closer to a solution. And that is all everyone expects.

We designed this course, "Data Challenge 2", to take you a step further to real-world Data Science in practice in a safe environment by giving you a real-life case but providing you with supervision and guidance along the way.

The objectives for this course are that you learn to:

- translate a stakeholder's problem into a Data Science problem
- specify a sub-problem from a broader problem
- apply Data Science techniques to address that sub-problem
- gradually uncover the issues that have to be addressed before "solving" the problem
- refine your project to make a meaningful step towards X
- handle and resolve uncertainty (experiencing that uncertainty is a necessary part of it)

It is important to re-iterate the following: not all problems have a clear solution (no one expects this from you in messy real-world data science).

Thus, while the assignment makes you experience uncertainty, the course is exactly about you going through it. We specifically designed the course to give you time to make mistakes and to help you learn from them.

The tutors and lecturers for this course are there to help you along the way.

3.2 Lecturers

- Dr Bennett Kleinberg, Associate Professor in Behavioural Data Science, Department of Methodology and Statistics, Tilburg University (responsible lecturer and coordinator)
- Dr Laura Genga, Assistant Professor in Information Systems, Department of Industrial Engineering and Innovation Sciences, TU/e

3.3 Tutors

Below you can find a list of the tutors.

- Andreea Murariu: a.e.murariu@student.tue.nl
- Cameron Dougherty: c.c.dougherty@student.tue.nl
- Maxwell Litsios: m.l.h.litsios@student.tue.nl
- Robert Druga-Tache: r.druga.tache@student.tue.nl
- Fedra de Haan: f.a.c.d.haan@student.tue.nl
- Stijn van de Ven: s.v.d.ven1@student.tue.nl

- Agustin Bejar Kurtin: a.bejar.kurtin@student.tue.nl
- Igor Dmochowski: i.f.dmochowski@student.tue.nl

All of the tutors have experience with Data Challenge 2 and are an asset of this course. The tutors will guide each group in weekly meetings.

3.4 Group formation

The groups have been formed by us (using random group member assignment), and you can find your group on Canvas.

3.5 Working with your group

All activities for this course run on campus (all in Eindhoven).

Your tutor will schedule weekly meetings with your group and will let you know in which room these take place. The time for group meetings is in line with the official timetable:

- Wednesdays: 13:30 - 17:30 (Eindhoven campus)
- Fridays: 8:45 - 12:45 (Eindhoven campus)

You do not have to fill the whole time with the group meetings, but all meetings with your tutor should happen in these time blocks in your rooms. You are free (and will have) to hold additional group meetings.

3.6 The meeting log

It is important that you carefully prepare your group meetings and log what has been discussed in them. To facilitate this, each group is expected to keep a meeting log (of each meeting, including the meetings with tutors, lectures and you group alone).

You can see below that part of your grade is the meeting log (i.e., you have to submit your group's meeting log, and every group member has to have played the role of log keeper at least once).

The log serves three purposes:

1. it sets an agenda of points you want to address in the meeting (keep in mind that time is of the essence for such projects)
2. it provides clarity for all group members of what has been discussed
3. it ensures that tasks are distributed, and every group member knows who is preparing what

The log for each meeting must include the following:

- General information
  - Group members present
  - Group members absent (incl. the reason)[3]
  - Date and time of the meeting
  - Location of meeting
  - Name of log keeper[4]
- Meeting content
  - Agenda of the meeting
  - Summary of points discussed (this can be in bullet points)
  - Tasks for the next meeting (who will do what?)

[3]Note that it is expected that you attend every group meeting. If there are valid reasons why you cannot attend a meeting, you have to discuss this with your group tutor in advance and obtain their approval.

[4]Note: every group member is expected to prepare the log for at least one meeting

You need to submit the meeting logs at the end of the course. *It is important that you log each meeting directly after it happened.* Make sure to also share the log with your group and tutor.

Note: the meeting log keeper role is not the same as the SCRUM master role.

4 Timelines

4.1 General schedule

We will have four plenary sessions. In your group, you will hold two group meetings every week, one of which is with your group's tutor.

Week	Day	Date	Activity
1	Wed	24/04/2024	Plenary session 1
1	Fri	26/04/2024	Group work
2	Wed	01/05/2024	Group work
2	Fri	03/05/2024	Group work
3	Wed	08/05/2024	Group work
3	Fri	10/05/2024	UNI CLOSED
4	Wed	15/05/2024	Plenary session 2
4	Fri	17/05/2024	Group work
5	Wed	22/05/2024	Group work
5	Fri	24/05/2024	Group work
6	Wed	29/05/2024	Plenary session 3
6	Fri	31/05/2024	Group work
7	Wed	05/06/2024	Group work
7	Fri	07/06/2024	Walk-in feedback
8	Wed	12/06/2024	Group work
8	Fri	14/06/2024	Group work
9	Wed	19/06/2024	Presentations
9	Fri	21/06/2024	Presentations

- The plenary sessions will be held from 13:30 - 15:30 on the TU/e campus.
- Details on the walk-in feedback session will be shared in due course.
- The presentations will be held on the TU/e campus on the 19th and 21st of June.

4.2 Milestones

The timeline below is a suggestion of milestones. These are not harsh deadlines but give you a way to better navigate the problem space as a group.

Week	Milestones for this week
1	read in the core data / understand the core data / clean the data / pitch first thoughts / read key materials
2	determine the specific sub-questions you want to answer / decide on the unit of analysis / add other datasets where applicable and relevant / think about the specific approach you want to take
3	decide on final dataset / gather additional data if needed / prepare data for your analysis / pitch analysis ideas
4	run the analysis / use the data to answer your sub-questions
5	refine the the analysis

Week	Milestones for this week
6	draw conclusions from the data / answer your sub-questions
7	refine analysis / revise and adjust conclusions and suggestions / prepare presentation
8	finalise the presentation / work on presentation and technical report and ethics reflection
9	Presentations

5 Grading

5.1 Deliverables and deadlines

Four deliverables count towards your final grade. You can find details on how each component is weighted in the grading criteria section.

1. The group presentation (deadline: 18 June 2024, 23:59h) - 50% of the final grade
2. The final report (deadline: 21 June 2024, 23:59h) - 20% of the final grade
3. The discussion of ethical consideration (deadline: 21 June 2024, 23:59h) - 15% of the final grade
4. General progress as a group (to be evaluated by the group tutors) - 5% of the final grade
5. The meeting log (updated continuously, submitted at the end of the course) - 5% of the final grade
6. Code documentation on GitHub - 5% of the final grade

You will receive a grade as a group. If problems arise within a group (e.g., if a group member disengages), we may opt for individualised grade adjustments.

5.1.1 The group presentation

The final presentation is targeted at a stakeholder audience and should include:

- a recap of the problem and your specific sub-problem
- the aims of your project
- the decisions made with the data
- the findings of your analysis
- 3 core conclusions and 3 recommendations (derived from the conclusions)

The presentation must be no longer than 15 minutes. The speaking time should be distributed equally among the group members. The slides used for your presentation should be uploaded to Canvas by the 18th of June, 23:59h.

The presentation will be presented on campus on the 19th and 21st of June.

Tournament-style presentations:

The presentations will be used to determine which groups will travel to London to present their final work to the police at New Scotland Yard (HQ of the Metropolitan Police). That tournament may take the following form:

- Round 1: each group is part of a pool of other groups and holds a presentation (15 mins) in front of the teaching team and the pool of students
- Round 2: from each pool, the best groups are selected to present in a final in front of the whole teaching team and all students. Groups in the final will have three additional minutes to present their work (18 minutes). This second round takes two days after the first round.
- Decision: the two best groups from the second round will travel to London.

5.1.2 The final report

The final report (written as a group) should cover:

- a brief introduction to the problem
- details on your data science approaches
- a detailed, statistical evaluation of your approach(es)
- a link to the GitHub page where your code and data are located and documented[5]
- an interpretation of your finding(s)
- at least two limitations of your approach and suggestions (for each limitation) how these could be fixed in the future

Using the provided template, the report should not be longer than **six** pages (including references and the reference list). The report must be submitted before the deadline as a pdf file named "group_X_technicalreport.pdf" (where X is replaced with your group number).

Template for the technical report

You should use the following Overleaf template: https://www.overleaf.com/latex/templates/acl-rolling-review-template/jxbhdzhmcpdm - taken from the proceedings of the Association of Computational Linguistics conference (ACL). When submitting your final report, make sure to use the non-anonymised version and list all group members as authors. The permitted page length includes the in-text references + the reference list. Your submission should be as a pdf via Canvas

Guide to GitHub:

In case you are unfamiliar with GitHub, have a look at these tutorials/guides:

- GitHub Quickstart Hello World
- GitHub Tutorial - Beginner's Training Guide
- GitHub desktop GUI

You can (but do not have to) use GitHub for your whole data and code flow. All we require is that you share a link to a public GitHub repository with the code to run your analyses and the data used.

5.1.3 Discussion of ethical considerations

A big challenge for real-world data science projects is the balance between stakeholder (in the widest sense) interests and potential ethical problems that may arise either directly or indirectly as a consequence of the project outcomes. In this course, you are working an a particularly challenging topic around police data, which has long been a controversy, but has also resulted in several safe-guarding frameworks to ensure proper applications of automated decision-making frameworks in the area of law-enforcement.

A whole body of work - both from legal scholars as well as computational crime scientists - has critically examined the role of predictive approaches in policing. Some core papers (incl. rare empirical evidence), are listed hereafter:

- Albert Meijer & Martijn Wessels (2019) Predictive Policing: Review of Benefits and Drawbacks, International Journal of Public Administration, 42:12, 1031-1039, DOI: 10.1080/01900692.2019.1575664
- Lyria Bennett Moses & Janet Chan (2018) Algorithmic prediction in policing: assumptions, evaluation, and accountability, Policing and Society, 28:7, 806-822, DOI: 10.1080/10439463.2016.1253695
- P. Jeffrey Brantingham, Matthew Valasik & George O. Mohler (2018) Does Predictive Policing Lead to Biased Arrests? Results From a Randomized Controlled Trial, Statistics and Public Policy, 5:1, 1-6, DOI: 10.1080/2330443X.2018.1438940
- Susser, Daniel, Predictive Policing and the Ethics of Preemption (June 29, 2021). Ben Jones and Eduardo Mendieta (eds.), The Ethics of Policing: New Perspectives on Law Enforcement (NYU Press), 2021, Available at SSRN: https://ssrn.com/abstract=3875917

[5]The code should enable someone else to fully replicate your findings

One of the most influential frameworks to think about ethics in data applications in UK policing is the ALGO-CARE framework (Oswald et al., 2018) that is now common practice and required for all police data science projects in the UK.[6]

Your task for the ethical considerations report is as follows:

- use the ALGO-CARE framework[7] on your group's project to *identify*, *discuss*, and *make suggestions for improvements* of potential ethical problems
- choose 3 different ALGO-CARE criteria and for each of them:
  - describe how each comes back in your project work (= identifying the problem)
  - discuss the implications of each ethical reservation/problem (e.g., who is affected, how are they affected?)
  - provide a suggestion for mitigating each ethical concern (i.e. what could be done to remove your concerns?) and mention the impact that the decision has had (or might have had) on the project outcomes
- discuss ethical considerations of technical design choices of your project (e.g., what are the implications of a specific analytical procedure?)

Further relevant literature that could help you:

- Brayne, Sarah, Alex Rosenblat, and Danah Boyd. "Predictive Policing". Data & Civil Rights: a new era of policing and justice. October 27, 2015. https://datacivilrights.org/pubs/2015-1027/Predictive_Policing.pdf
- B. Green, "Data Science as Political Action: Grounding Data Science in a Politics of Justice," in Journal of Social Computing, vol. 2, no. 3, pp. 249-265, September 2021, doi: 10.23919/JSC.2021.0029. https://ieeexplore.ieee.org/abstract/document/9684742

Page limit: 3 pages. A template is provided below. The report must be submitted before the deadline as a pdf file named "group_X_ethics.pdf" (where X is replaced with your group number).

Template ethics report

You should use the following Overleaf template: https://www.overleaf.com/latex/templates/acl-rolling-review-template/jxbhdzhmcpdm - taken from the proceedings of the Association of Computational Linguistics conference (ACL). When submitting your final report, make sure to use the non-anonymised version and list all group members as authors. The permitted page length includes the in-text references + the reference list. Your submission should be as a pdf via Canvas

5.1.4 The meeting log

See the details above –> The meeting log. The meeting log will need to be submitted via Canvas.

5.1.5 General progress as a group

In close coordination with your weekly supervisors, each group will receive a lightweight grade for general progress made throughout the course.

5.2 Grading criteria

The grading criteria listed below per assignment will be used to determine the final weighted grade per assignment.

[6] If you are interested in a more extensive review of "data analytics and algorithms in policing", you can have a look at this article from RUSI or the briefing version of this report here.

[7] Marion Oswald, Jamie Grace, Sheena Urwin & Geoffrey C. Barnes (2018) Algorithmic risk assessment policing models: lessons from the Durham HART model and 'Experimental' proportionality, Information & Communications Technology Law, 27:2, 223-250, DOI: 10.1080/13600834.2018.1458455

Once all assignments have been graded, the full final grade will be the weighted average of the assignments.

Each rubric is scored on a scale from 1 to 5 as follows:

- 5 = very good/excellent
- 4 = good
- 3 = as expected
- 2 = poor
- 1 = very poor/absent

5.2.1 Late submissions and formal requirements

The rules below apply to all assignments.

Late submission policy: If you submit an assignment late, your grade will be deducted by 10% for each day (starting directly after the deadline and counting for each begun period of 24 hours) that is between the original submission deadline and your submission. For example, if the deadline is the 8 Feb., 11:59 pm and you submit on 9 Feb., 0:01 am, your original grade (say: 7.5) will be deducted by 10% (here: 0.75) to 6.75. If you submitted on 11 Feb., 3 pm (= 3 days late), the grade would be reduced by 30% (here: 2.25) to 5.25.

Formal requirements policy: Each assignment has some formal requirements re. page length, presentation duration, speaking time per person, etc. These are stipulated for each assignment. If these are not met (e.g., exceeding page length or presentation time), grade deductions will apply. The grade deductions will be made in consultation with the entire teaching team.

5.2.2 The group presentation (50%)

Criterion	Explanation	Weight
Problem definition	The degree to which a specific sub-problem was identified and the justification provided for that decision	10%
Stakeholder clarity	The degree to which the presentation identifies and speaks the stakeholder's (i.e. non-technical) language	10%
Data decisions	The degree to which adequate decisions re. the dataset(s) have been made with the data in line with the sub-problem	20%
Data reasoning	The degree to which there is a logical flow from problem to data to analysis to findings	20%
Conclusions and recommendations	The quality of the conclusions and subsequent recommendations and the degree to which each conclusion is directly backed up by the findings	30%
Quality of the presentation	The quality of the presentation, slides, language on the slides, visuals and overall layout of the presentation	10%

5.2.3 The final report (20%)

Criterion	Explanation	Weight
Data science techniques	The adequacy and correctness of the statistical and technical details for the data science techniques applied	30%
Statistical evaluation	The quality, thoroughness and correctness of the statistical evaluation and interpretation of the models	30%
Limitations and future work	The quality of the limitations identified and the suggestions for future work	20%
Quality of the report	The degree to which the report is written clearly, without mistakes, formatted properly, aided by visuals, and presented in academic language	20%

5.2.4 The ethical discussion (15%)

Criterion	Explanation	Weight
Description of ethical concerns	The adequate identification of ethical concerns of the project along three self-chosen ALGO CARE elements	20%
Discussion of the implications of ethical concerns	Overall quality of the critical reflection on and discussion of the implications of the three chosen ALGO CARE elements	30%
Suggestions for the mitigation of the ethical concerns	Overall quality of the mitigation ideas for the three chosen ALGO CARE elements	20%
Discussion of technical design choices	The degree to which technical/analytical design choices are discussed regarding their implication for ethics in this project	15%
Overall quality of the report	The degree to which the report is written clearly, without mistakes, formatted properly, and presented in academic style and language	15%

5.2.5 Group progress (5%)

The degree to which the group has made progress in tackling the problem and working as a data science group (i.e., not a collection of individuals).

The grade will be evaluated on the rubric scale by your group tutor:

- 5 = very good/excellent
- 4 = good
- 3 = as expected
- 2 = poor
- 1 = very poor/absent

5.2.6 The meeting log (5%)

Submission of the meeting log that details every meeting held.

This grade will be be binary (sufficient/insufficient quality) and accounts for 5% of the final group grade. If no meeting log is submitted, the final group grade will be deducted by up to 1.0.

5.2.7 Code documentation on GitHub (5%)

The quality of the GitHub documentation of your code. The repository should be sufficiently detailed for someone else to replicate your project.

This grade will be be binary and assess the quality of the GitHub repository (sufficient/insufficient quality) and accounts for 5% of the final group grade.
