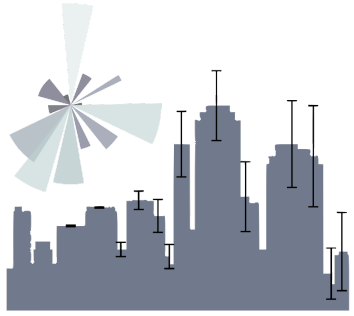
principles of Urban Science

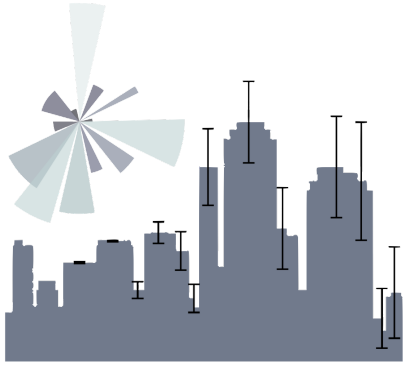


introduction

dr.federica bianco | *fbb.space* | *fedhere* | *fedhere*

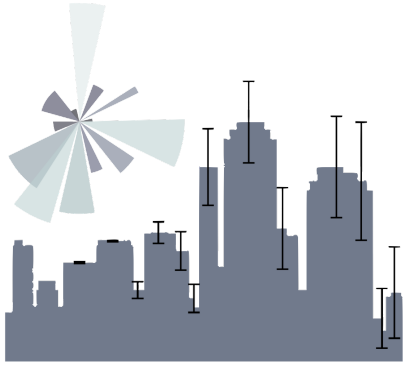
this slide deck: https://slides.com/federicabianco/pus2020_1

all slide decks will have this name pattern:
slides.com/federicabianco/pus2020_<week number>



let me introduce myself...

who am I?



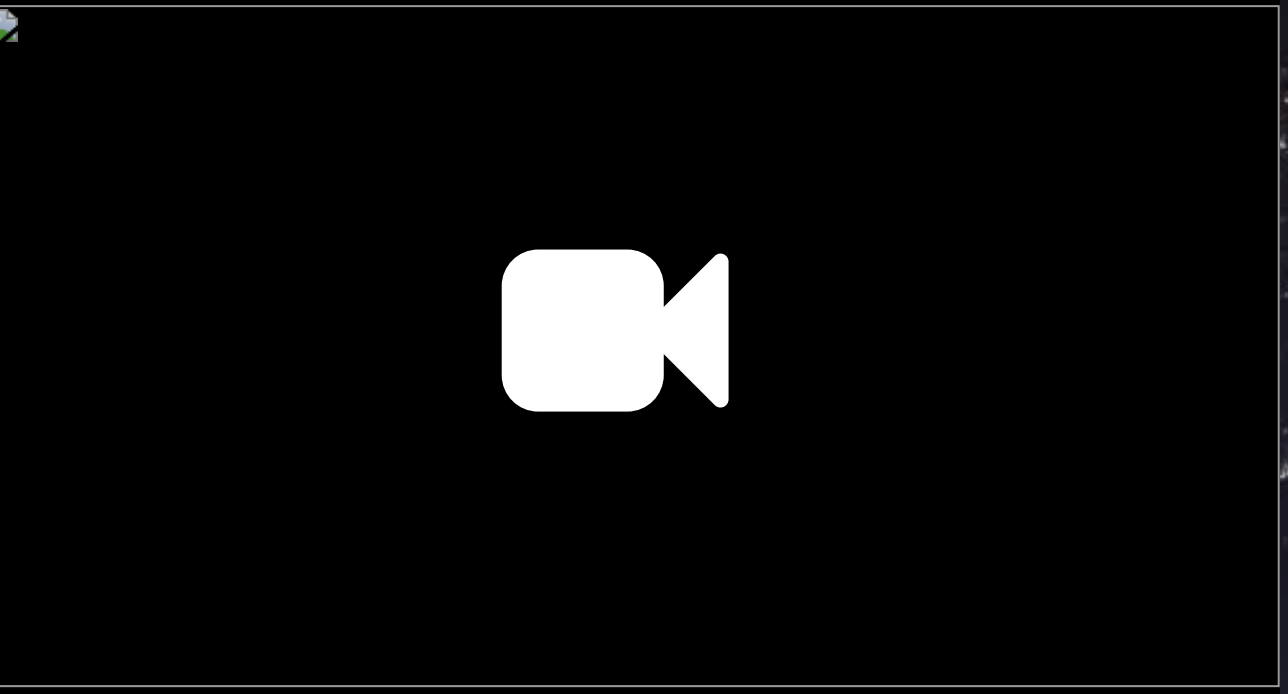
astrophysics -> data science

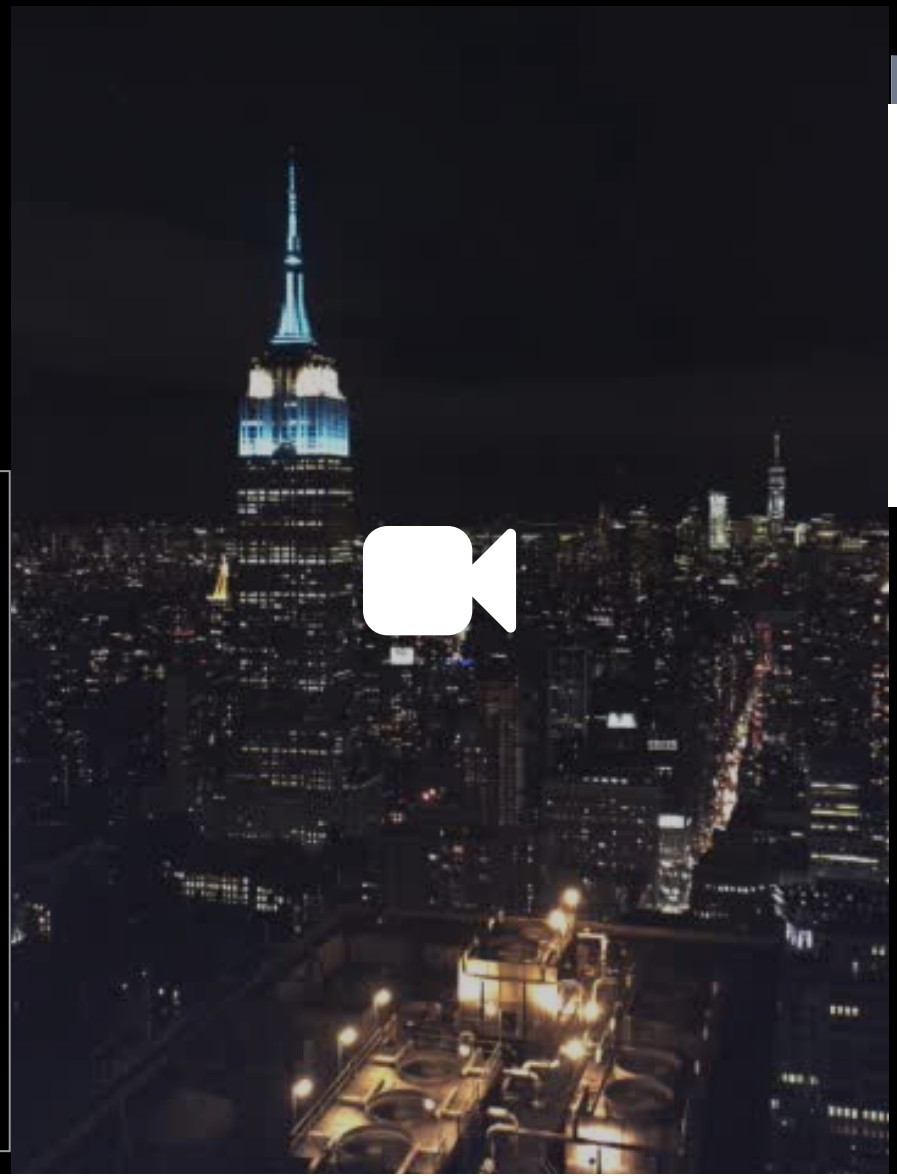
astrophysics stands our as an *observational*, rather than experimental science

to "observe" the natural status of a system provides advantages. To inquire the system about its status may provide a biased response

(e.g. surveys have many biases)

who am I?



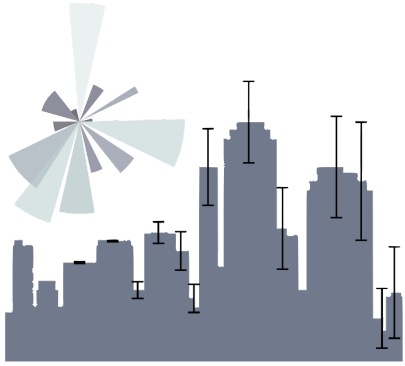


UO

CENTER FOR URBAN SCIENCE AND PROGRESS

cuspuo.org

who am I?



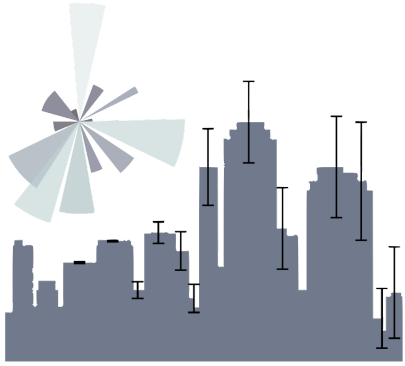
astrophysics -> data science

UD Data Science Institute - Inaugural event

what is data science? we have been using data in science the whole time, but with the volume, rate, and complexity of the current data we have to worry about things that we would neglect until now: what happens if our data has errors, what happens if we have missing data?

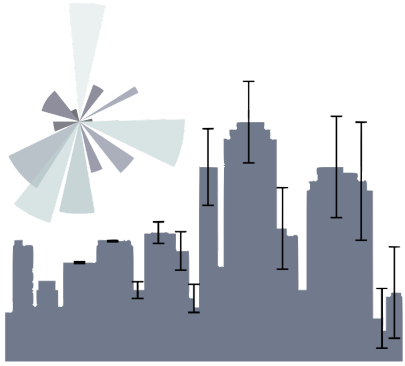
Lou Rossi, Mathematical Sciences
Chairperson, UD

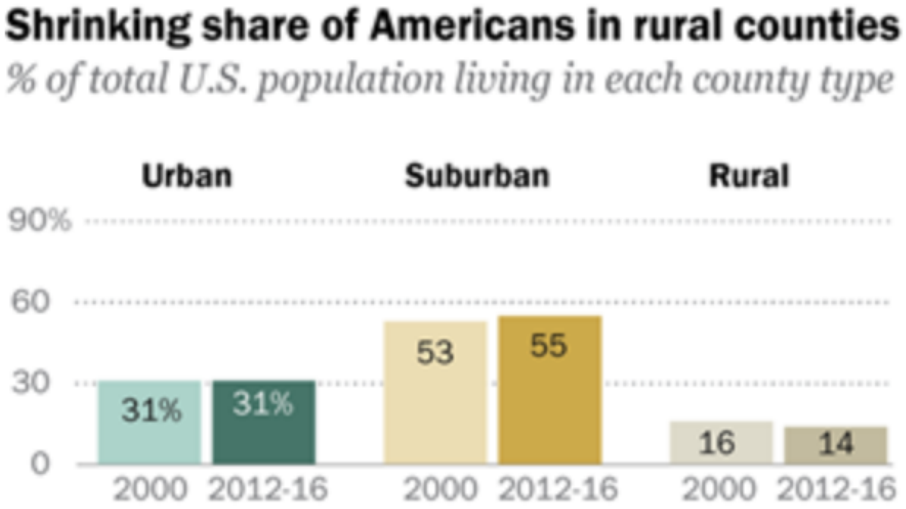
(astrophysicists have always worried about that)

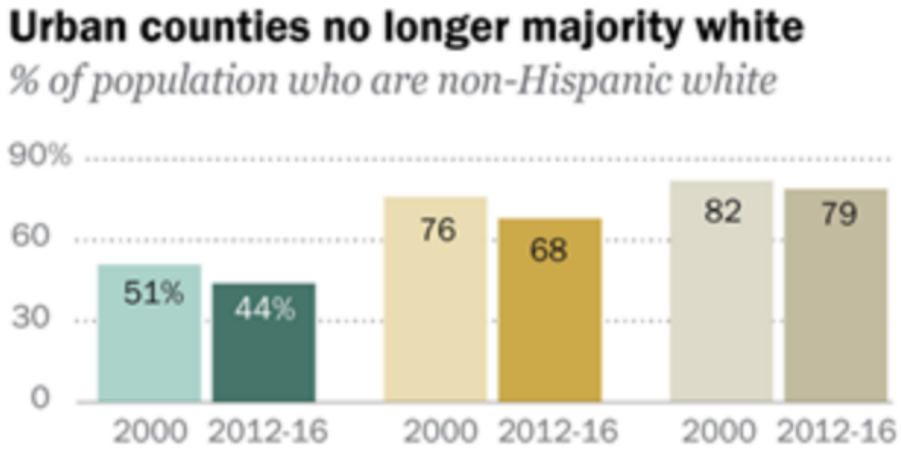


1 Urban Science

what can data science do for Cities?

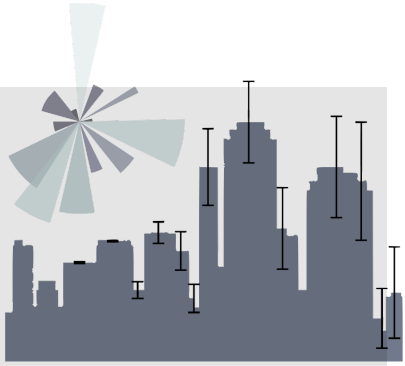




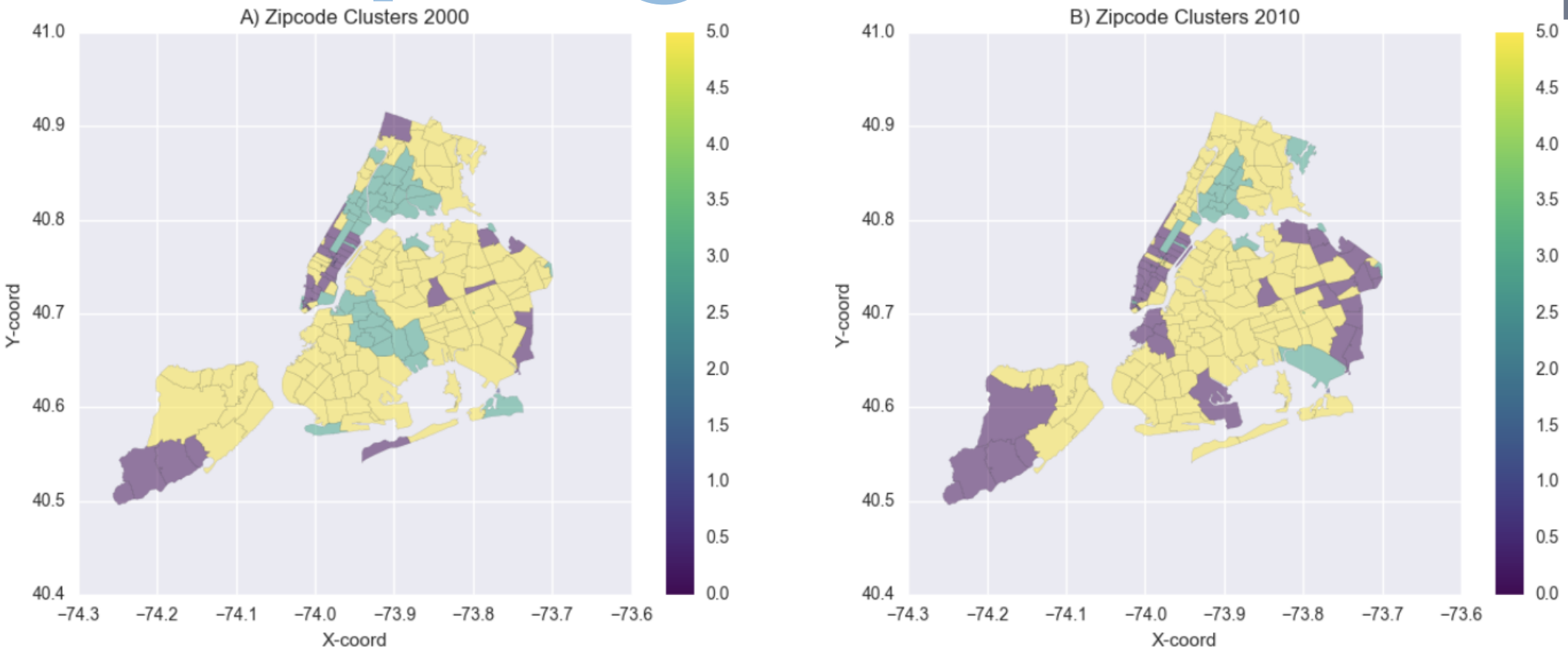


what can data science do for Cities?

https://opendata.cityofnewyork.us/data/

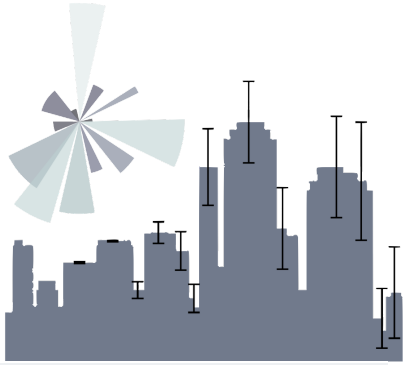


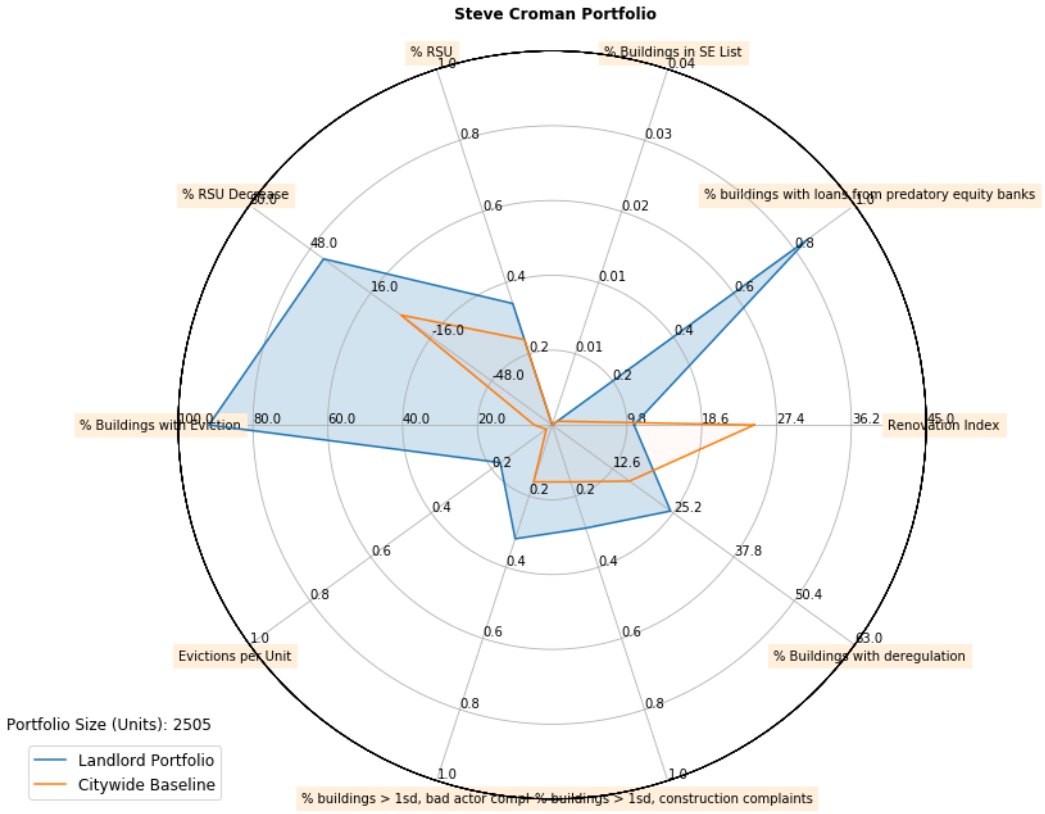
example: gentrification



change in business density, income, age, population, racial diversity measures gentrificatoin

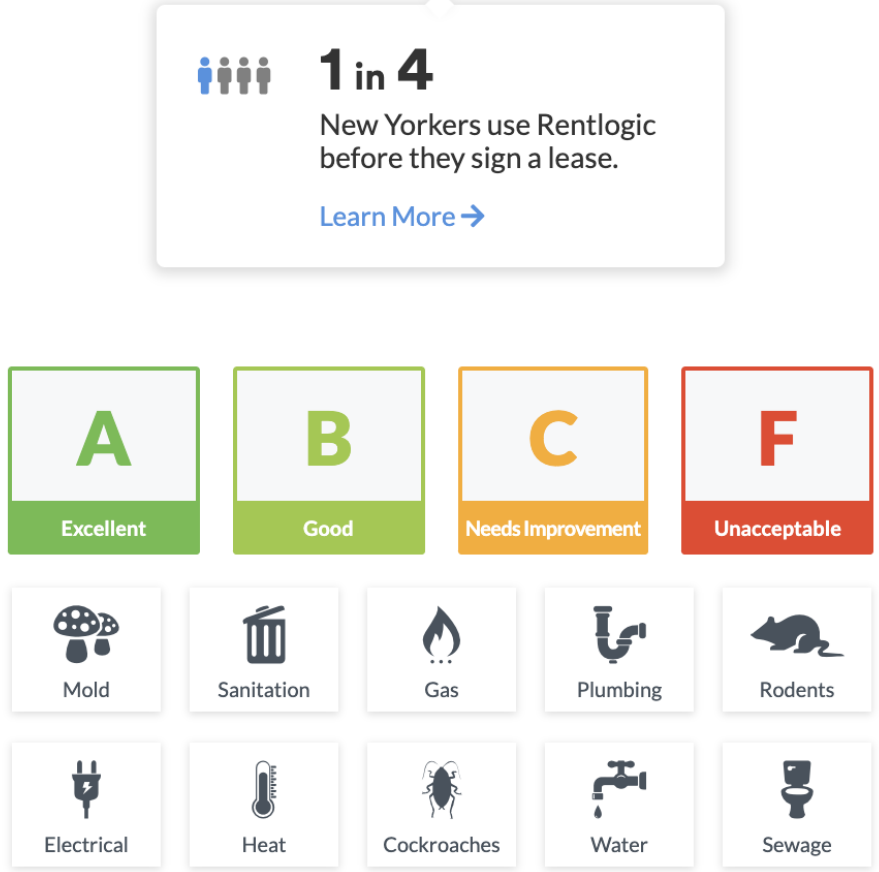
example: real estate





identifying predatorial landlords

(CUSP capstone)

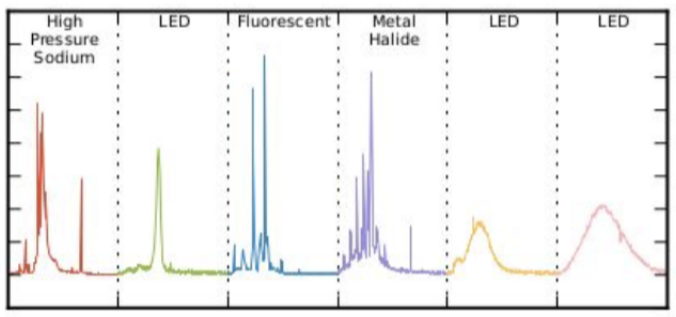


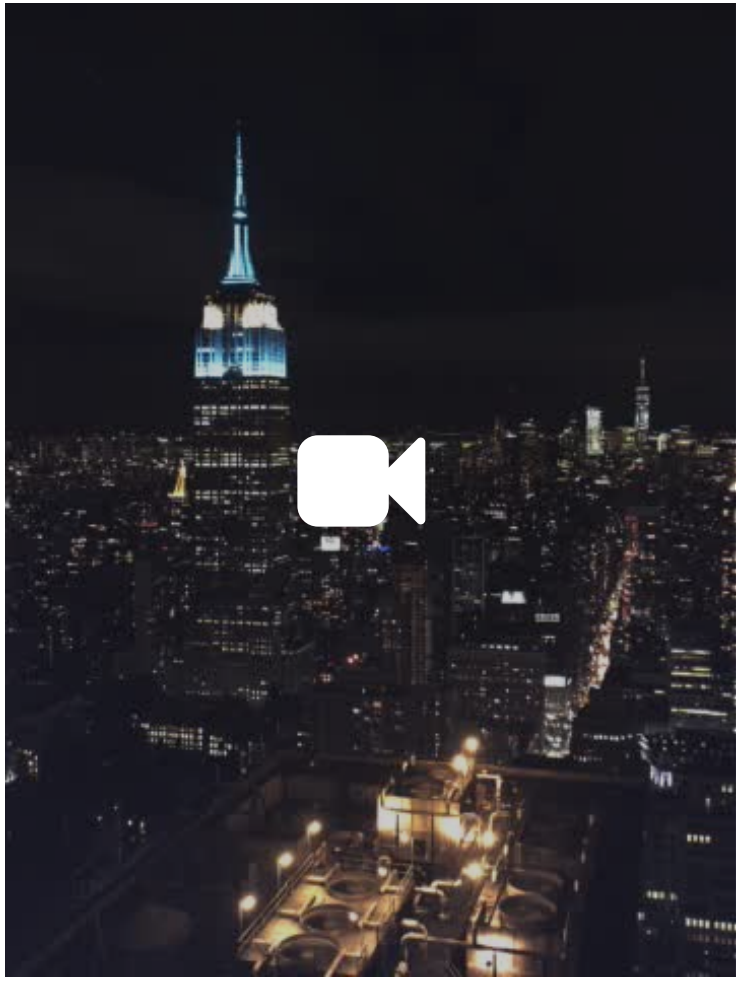
apps to evaluate rental deals

https://rentlogic.com/

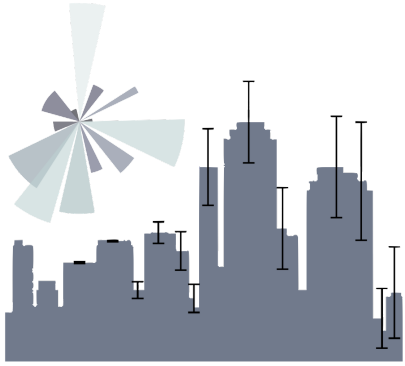
example: energy

New York City Lighting Technologies



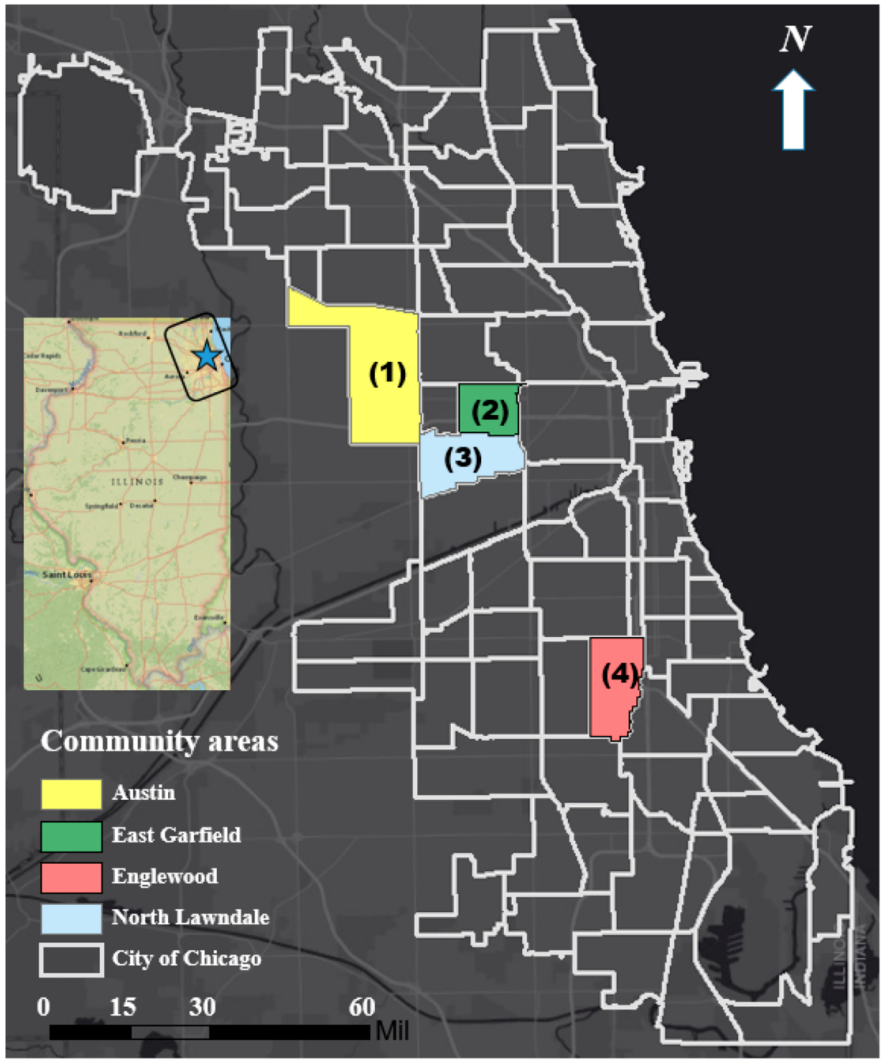


example: pollution



https://fbb.space/PUS2020/data/original,434253,10-20.gif

example: crime



SOCIAL INNOVATION NARRATIVES

Journal of
COMMUNITY SAFETY & WELL-BEING

The ethical dangers and merits of predictive policing

Moish Kutnowski

ABSTRACT

Predictive policing is an emerging law enforcement technique that uses data and statistical analysis to aid in the identification of criminal activity. Its intention is to proactively reduce crime by providing police forces with likely areas of high risk variables. While this is a noble pursuit, every new tool must be accompanied by the ethical considerations of its potential consequences. Predictive policing is still in its infancy, borne from crime analysis and big data; however, the Western criminal justice system in the traditional sense is a reactive institution with a diverse history. The use of predictive policing presents a new challenge for law enforcement in that it allows for a divergence from the distinct reality of modern policing. Using the United States as an example of the dangers and flaws of predictive policing as a discretionary tool used to justify questionable processes and biases, this paper will analyze the potential opportunity that predictive policing and new holistic forms of law enforcement and community safety initiatives can use in partnerships with communities and policy makers.

Key Words Predictive policing; big data; community safety; technology; policy planning; ethical policing; collaboration.

Journal of CSWB. 2017 Mar;2(1):13-17 www.journalcswb.ca

INTRODUCTION

Predictive policing is a relatively new phenomenon borne from Big Data analytics and trend forecasting. Big Data is the collection of large and complex datasets that are more accessible due to modern advanced technologies. In application, predictive policing uses analytical techniques on Big Data to detect probable criteria for police via statistical prediction (Perry, 2013). The larger conversation revolving around Big Data from a consumer and marketing perspective is highly scrutinized for privacy and accuracy concerns. This scrutiny becomes quickly critical with regards to predictive policing; law enforcement groups generally consider it a boon, while civil liberty groups generally consider it a bane. This paper will outline some of the ethical considerations vis-à-vis the dangers and merits of predictive policing, given the limited information available in this evolving field.

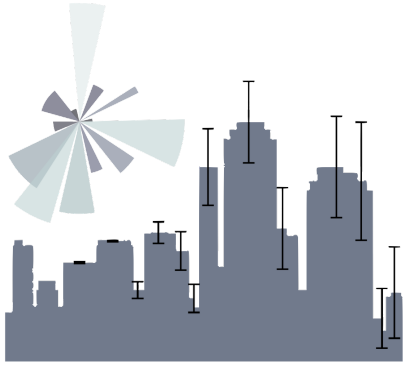
There are some basic assumptions about how predictive policing is perceived that must be considered initially from a common, but not exhaustive, standpoint. There is a distinct bias against using data analytics to enforce a rule of law, whereas data analytics for the purposes of health, marketing, and economic development are generally supported. The assumption is that policing, and/or the activities therein, are fundamentally different from other activities. The normative paradigm would then be the divide between encouragement in community initiatives (like health and economics) and commandment (like policing). However, that could extend further within a realist context; the ability to enforce the rule of law is fundamental to social cohesion and rests solely in the state. While this may form the basis of a realist's perspective, an alternative perspective derived from a humanist's consideration would indicate that policing by its very nature rests outside of the social boundaries; it is unique only to policing. It would be antithetical for policing to conform to a universal maxim as it is responsive to state of nature behaviour within the confines of civil society.

The actual statistical model used in predictive policing depends on the Gaussian function as an integral kernel (Rosser and Cheng, 2016). This function acts as a probability estimate with limited and random variables, such as regional information, in order infer patterns, such as crime. While the use of statistical probabilities is ubiquitous in industry, it provides a questionable benefit for law enforcement. The difficulty with statistics for prediction, and Gaussian probability specifically, is its vulnerability to spurious variables which undermine the fundamental premise of predictability (Taleb, 2007).

The promise of predictive policing is alluring because it promises preventative measures, but in reality the justice system (and law enforcement) is reactive, which does not mean that community safety or policing necessarily is as well.

Correspondence to: Moish Kutnowski, Ottawa ON, Canada.
E-mail: Moish.kutnowski@gmail.com

https://www.researchgate.net/publication/333498701_The_ethical_dangers_and_merits_of_predictive_policing

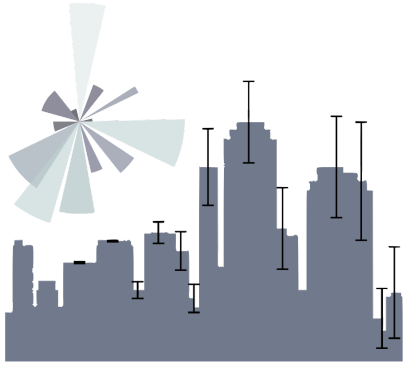


PSU: principle of urban science

GENERATING AND INTERPRETING EVIDENCE FOR EVIDENCE-BASED DECISIONS IN URBAN ENVIRONMENTS

Code of Conduct

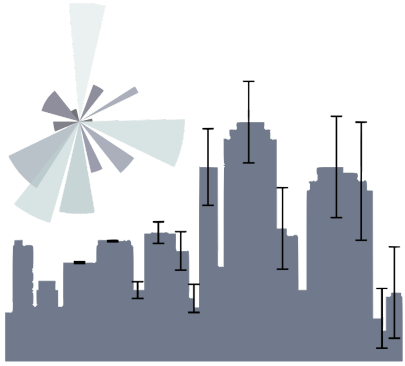
https://github.com/fedhere/PUS2020_FBianco/blob/master/CodeofConduct.md



Diversity is considered a resource that enriches us culturally and intellectually in this class.
I expect to see a supportive, collaborative attitude from all of you, to assure we maintain and foster a learning environment that leads to rigor, excellence, and happiness. No instances of harassment or attempts to marginalize students will be tolerated in my class. No instances of bullyong, or cyber-bullying. If you have concerns, if you do not feel safe or are made to feel unwelcome, please come talk to me - keeping in mind that I am a mandatory TitleIX reporter.

https://forms.gle/p7ckqNhaje8CpFeZ8

Code of Conduct



https://github.com/fedhere/PUS2020_FBianco/blob/master/CodeofConduct.md

Microaggressions

We all come from different places and have different level and kind of privilege, and oppression.

Because of this it is impossible to completely understand another person's vulnerabilities.

This leads to *microaggressions:* ways in which we perpetrate oppressions in subtile, often unintentional ways.

An example of a microaggression, one that was exposed in the UD students's letter to the administration, may be to ask an asian student to adopt a western name with an easier pronunciation. The person that ask likely does not realize tht this imposes the concept that there is a "norm" that is europocentric and white.

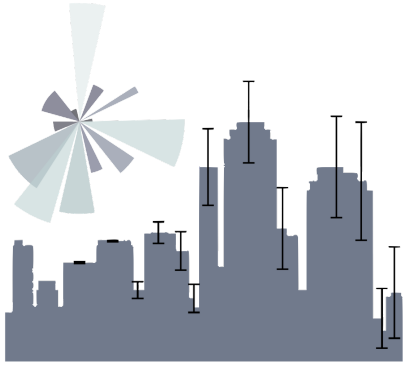
I am a white middle-class-raised, middle-class woman. In that I have a lot of privilege. If with my privilege I mess up and "committ" a microaggression, ***please call me out on it***! You can turn on your microphone and say "ouch" if you are comfortable to do it in front of the class. Or you can talk to me personally after class. If class members do it let them know in the same way. Ask me if you want my intervention if you prefer not doing it directly.

If someone points out that you something you "said" or "did" is a microaggression don't be defensive! Listen and commit to understand. There should be no assumption that you had malignant intensions. This way we can move forward and grow together.

https://yth.org/new-campaign-draws-attention-micro-aggressions/

Objectives

https://github.com/fedhere/PUS2020_FBianco/blob/master/CodeofConduct.md



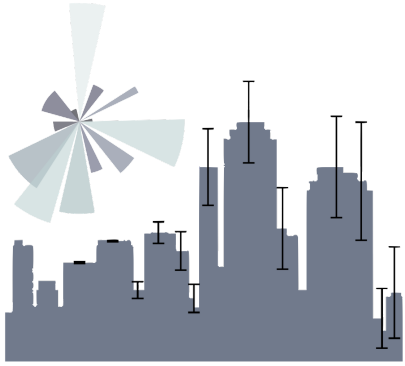
Learn data science methodology including statistics and machine learning, focusing on time series and geospatial analysis.

Each method will be approached as it applies to existing data and problems, and problems will be explored in multiple urban contexts, generating comparative studies.

- exploit open data collections
- gather data and prepare data
- extract statistical information
- model mechanisms

- elements of machine learning
- visualize data and present evidence,
- interpret evidence presented by their peers

working in groups



Whe are not objective.

Even when we attempt to have a scientific approach and leave our background behind, the questions that we ask, the way in which we try to answer them is influenced by who we are and our experience.

Working in groups we remove some of this bias.

Also, even tho I am the teacher, peer learning is a more effective way to learn!

Be respectful, be kind, be considerate, be openminded.

working in groups

Because we are working in groups, please if at all possible turn your camera on!

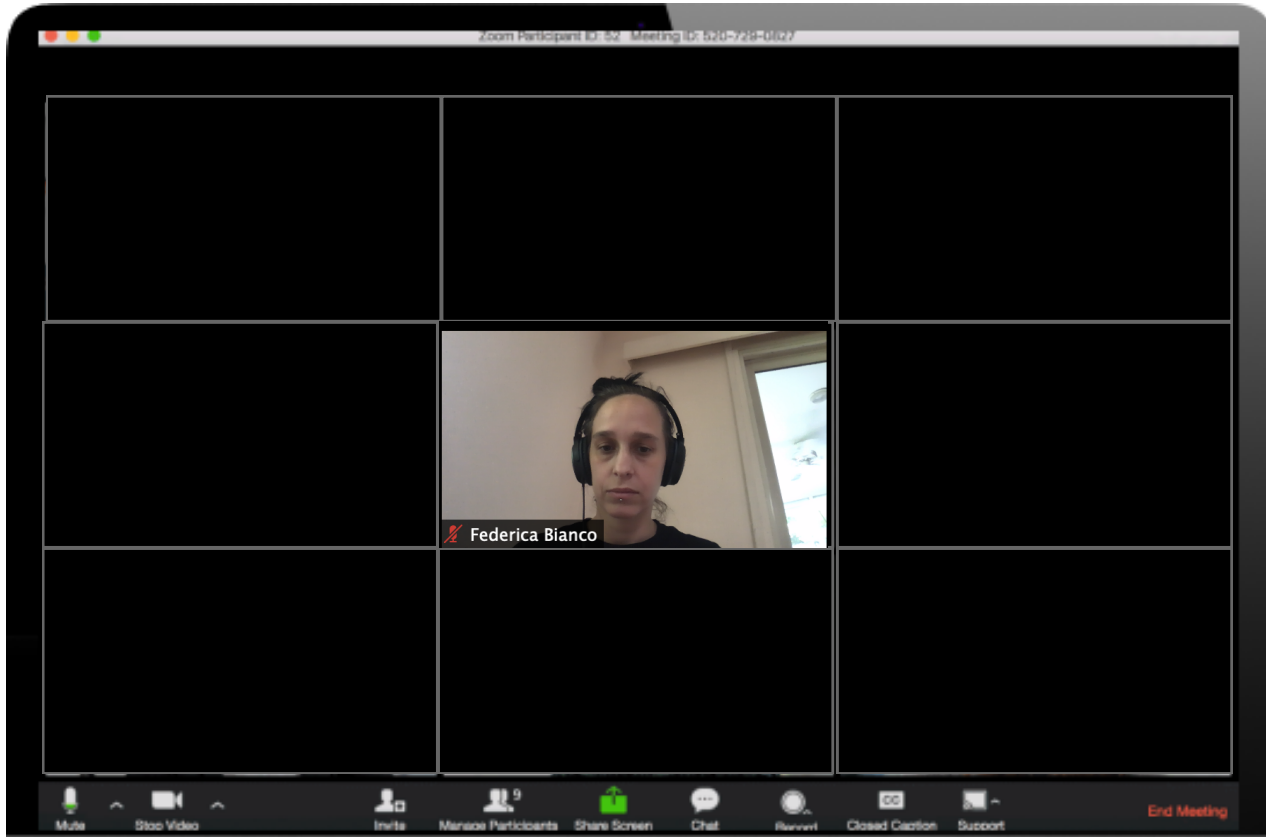
This is not a book-based class where I teach conventional lectures - it is hard for me to know if I am being clear without seeing you

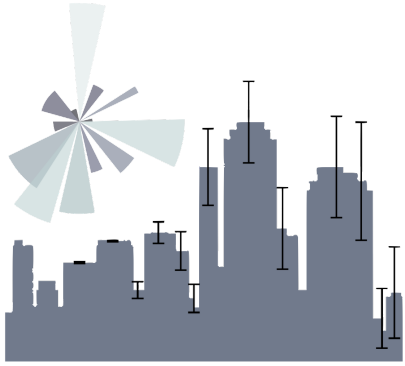
It is hard to work with other people that you have never seen: we form mental pictures of people around us that help us understand them.

I realize you may have good reasons for not wanting to let people into your home (privacy) or your bandwidth may make it impossible. Not so good reasons tho:

- want to do something else but still get credits for attendance

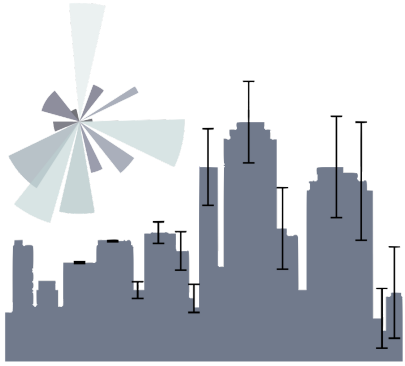
- did not feel like showering : preparing for class helps you learn





3 grading

grades

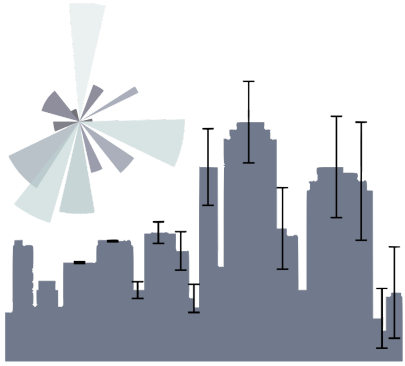


https://github.com/fedhere/PUS2020_FBianco/blob/master/README.md

- 15% pre-class questions
- 15% labs performance and participation
- 20% homework
- 20% midterm
- 30% final

grades: quiz

https://github.com/fedhere/PUS2020_FBianco/blob/master/README.md



PUI2020 example class quiz

* Required

Email address *

Your email

Is clustering a supervised or unsupervised learning technique *

- Supervised learning
- Unsupervised learning

what is kNearestNeighbors? *

- a distance metric or loss function
- a machine learning model
- a feature

Name at least two type sof data-science analysis/data-science question according to J. Leek & R. Peng'2 2015 paper. *

Your answer

- **15% pre-class questions**
- 15% labs performance and participation
- 20% homework
- 20% midterm
- 30% final

- there is no book

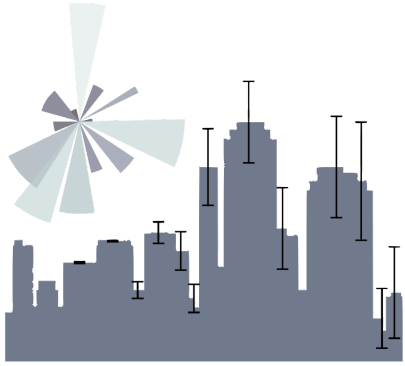
- slides are the basis for the study material

- reviewing material as we go is necessary because each topic builds on the previous one

quizzes have questions about the material covered in class and assigned reading

<- example of class quiz https://forms.gle/Z3tcC3zP2rSrLkot7

grades: lab performance



Each class we will discuss a topic and then I will split you into groups to practice on it.

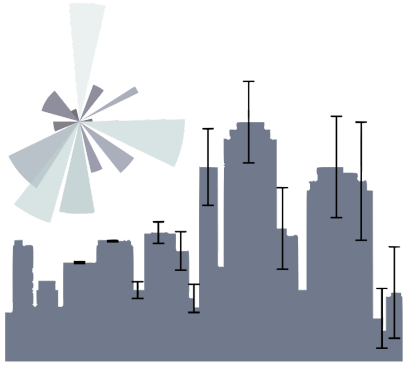
Groups will be randomly assembled. I will pop in each breakout room to help you as you need.

In the breakout rooms work collaboratively! be generous with your time and with your skills! If you know something you will solidify your knowledge by helping your classmates.

If you do not something there is nothing to be embarrassed: let your classmates help you!

- 15% pre-class questions
- **15% labs performance and participation**
- 20% homework
- 20% midterm
- 30% final

grades: homework

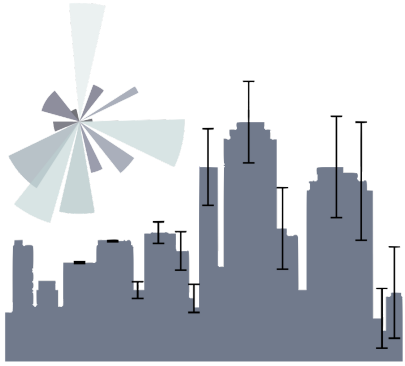


All assignments will be based on real data and real problems. Each assignment will intail

- writing code to collect data
- writing code to explore the statistical properties of the data
- writing code to answer a specific question about the data as assigned by me
- formulating another question of your choosing
- writing code to answer this question
- writing code to visualize the results and communicate them

- 15% pre-class questions
- 15% labs performance and participation
- **20% homework**
- 20% midterm
- 30% final

grades: homework



All assignments will be based on real data and real problems. Each assignment will intail

work in groups

deliver individual notebooks

include in the "github repo" a statement of what you did and what you were and were not responsible for

CitiBike HW - v1

Question

Are CitiBike's easing commuter journey's across the East River?

Hypothesis

- H0: The probability of a citibike subscriber crossing the East River in a given month is **independent** of whether the trip is taken during rush hour
- H1: The probability of a citibike subscriber crossing the East River in a given month is **not independent** of whether the trip is taken during rush hour

Project work balance

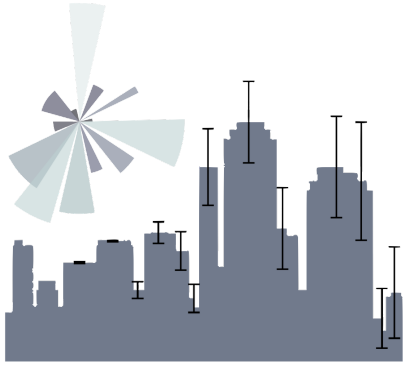
hypothesis generation

Max, Arno, Clayton discussed and equally shared hypothesis generation. Max had the original idea of looking at bridges as he is an avid CitiBike user

Tasks

1. Clayton is tagging trips as cross east river or not
2. Max is defining historic hours as "on peak" or "not on peak"
3. Arno completes a chi-square test of our hypothesis

grades: homework



All assignments will be based on real data and real problems. Each assignment will intail

- writing code to collect data
- writing code to explore the statistical properties of the data
- writing code to answer a specific question about the data as assigned by me
- formulating another question of your choosing
- writing code to answer this question
- writing code to visualize the results and communicate them

```
fig = pl.figure(figsize=(15,5))
ax = fig.add_subplot(131)
bkmerged.plot(column="unemployedF", ax=ax, label=True)
ax.set_title("unemployment rate")
ax.axis('off')

ax = fig.add_subplot(132)
bkmerged.plot(column="condition", ax=ax, label=True)
ax.set_title("parks condition")
ax.axis('off')

ax = fig.add_subplot(133)
bkmerged.plot(column="parkcount", ax=ax, label=True)
ax.set_title("park number")
ax.axis('off');
```

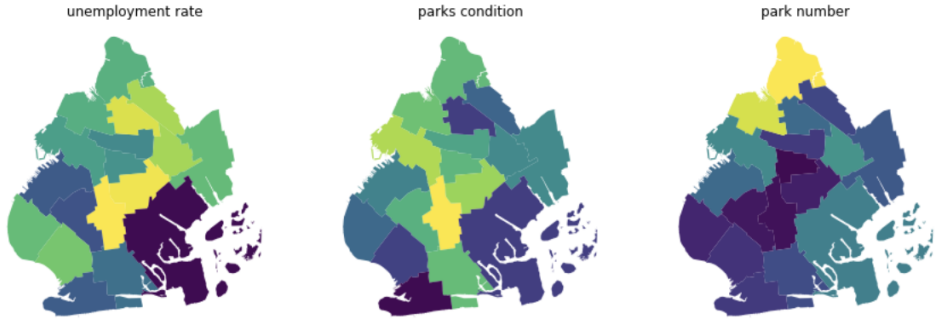
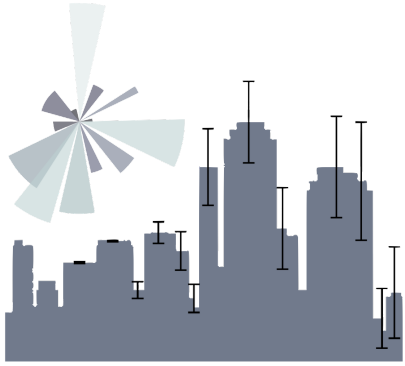


Figure 4 distribution of unemployment rate, average condition of the parks, and number of parks by PUMA (see figure 3). Yellow corresponds to the highest values of each variablem, purple to the lowest. Ranges are reported below. Unemployment rate is maximal in central Brooklyn, and minimal in the area surrounding JFK, Conversely, the average park condition seems to be maximal in central Brooklyn (which is in tension with our thesis) but the number of parks per PUMA (where parks are assigned to a PUMA based on the location of the park center) shows a minimum in central Brooklyn and maximum in the notheast PUMAs. Note, however, that loosing Prospect park during the merge significantly affect our inference. However, if prospect park appeared as a single unit in the park inspection data, likely our inference would not be affected.

grades: homework



All assignments will be based on real data and real problems.

For each step of the assignment you should create one or more cells of code, and you should produce one or more outputs.

The most common outputs are going to be the
1) result of statistical tests (e.g. p-values, classifications) or

2) plots

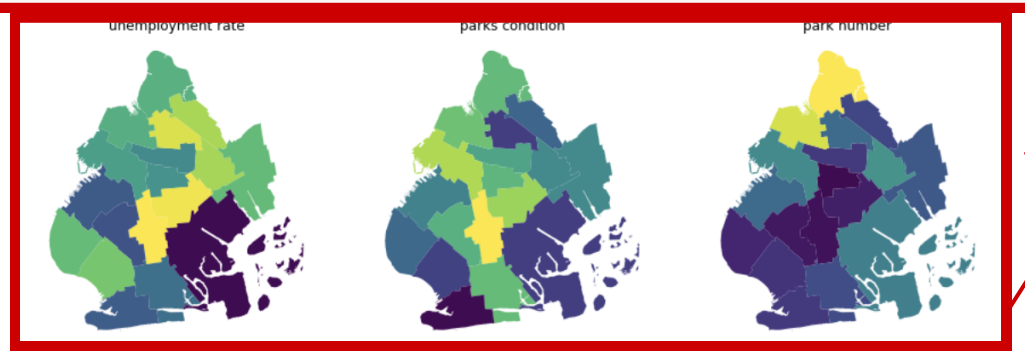
The grading will be done on the interpretation of the statistical results and of the figures as written in the captions

code

```
fig = pl.figure(figsize=(15,5))
ax = fig.add_subplot(131)
bkmerged.plot(column="unemployedF", ax=ax, label=True)
ax.set_title("unemployment rate")
ax.axis('off')

ax = fig.add_subplot(132)
bkmerged.plot(column="condition", ax=ax, label=True)
ax.set_title("parks condition")
ax.axis('off')

ax = fig.add_subplot(133)
bkmerged.plot(column="parkcount", ax=ax, label=True)
ax.set_title("park number")
ax.axis('off');
```

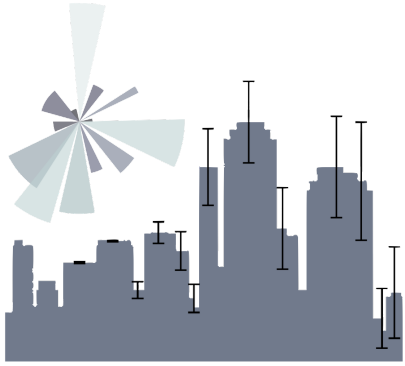


figure

Figure 4 distribution of unemployment rate, average condition of the parks, and number of parks by PUMA (see figure 3). Yellow corresponds to the highest values of each variablem, purple to the lowest. Ranges are reported below. Unemployment rate is maximal in central Brooklyn, and minimal in the area surrounding JFK, Conversely, the average park condition seems to be maximal in central Brooklyn (which is in tension with our thesis) but the number of parks per PUMA (where parks are assigned to a PUMA based on the location of the park center) shows a minimum in central Brooklyn and maximum in the notheast PUMAs. Note, however, that loosing Prospect park during the merge significantly affect our inference. However, if prospect park appeared as a single unit in the park inspection data, likely our inference would not be affected.

caption

grades: homework



All assignments will be based on real data and real problems.

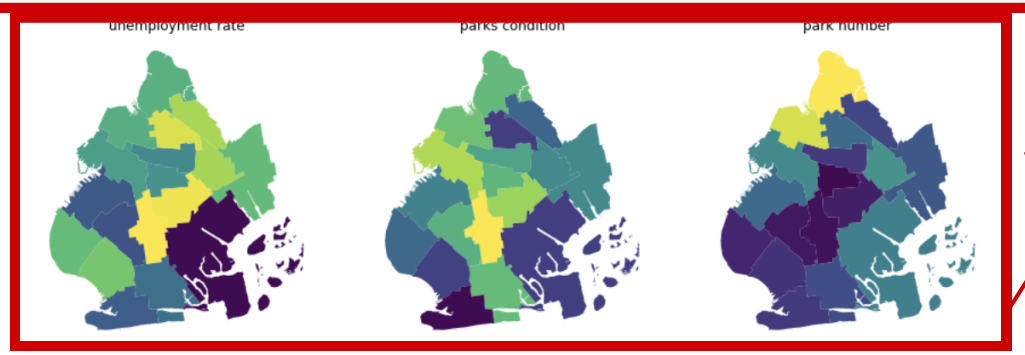
The question you choose to ask and answer with your group will be discussed in class on the next meeting.

code

```
fig = pl.figure(figsize=(15,5))
ax = fig.add_subplot(131)
bkmerged.plot(column="unemployedF", ax=ax, label=True)
ax.set_title("unemployment rate")
ax.axis('off')

ax = fig.add_subplot(132)
bkmerged.plot(column="condition", ax=ax, label=True)
ax.set_title("parks condition")
ax.axis('off')

ax = fig.add_subplot(133)
bkmerged.plot(column="parkcount", ax=ax, label=True)
ax.set_title("park number")
ax.axis('off');
```

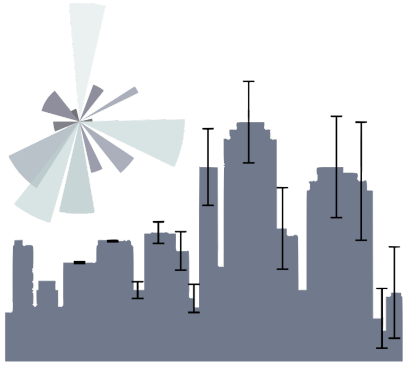


figure

Figure 4 distribution of unemployment rate, average condition of the parks, and number of parks by PUMA (see figure 3). Yellow corresponds to the highest values of each variablem, purple to the lowest. Ranges are reported below. Unemployment rate is maximal in central Brooklyn, and minimal in the area surrounding JFK, Conversely, the average park condition seems to be maximal in central Brooklyn (which is in tension with our thesis) but the number of parks per PUMA (where parks are assigned to a PUMA based on the location of the park center) shows a minimum in central Brooklyn and maximum in the notheast PUMAs. Note, however, that loosing Prospect park during the merge significantly affect our inference. However, if prospect park appeared as a single unit in the park inspection data, likely our inference would not be affected.

caption

grades: midterm & final

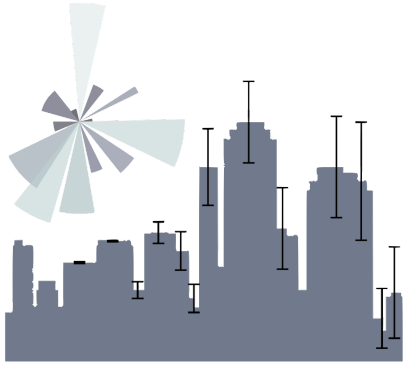


Midterm1: a project proposal for a project that you will perform in the second half of the semester

~~Midterm2~~: ~~a short in-class project.~~

Final: a project proposal and presentation

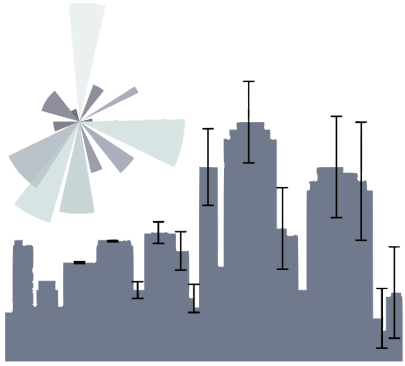
- 15% pre-class questions
- 15% labs performance and participation
- 20% homework
- **20% midterm**
- **30% final**



4 tech details

python - collab - github - slack - canvas

tech details



if you cannot code in python:

start here https://developers.google.com/edu/python/introduction

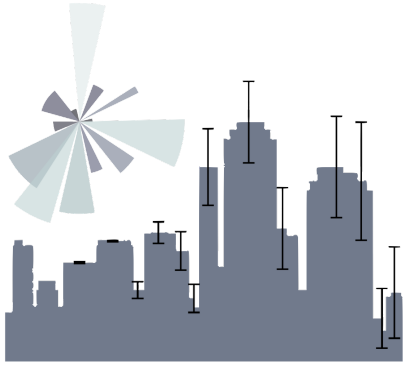
and here Beginning Python Visualization, 2009

Python bootcamps (tentative dates):

Friday 09/04 12-4PM

Friday 09/11 12-4PM

tech details



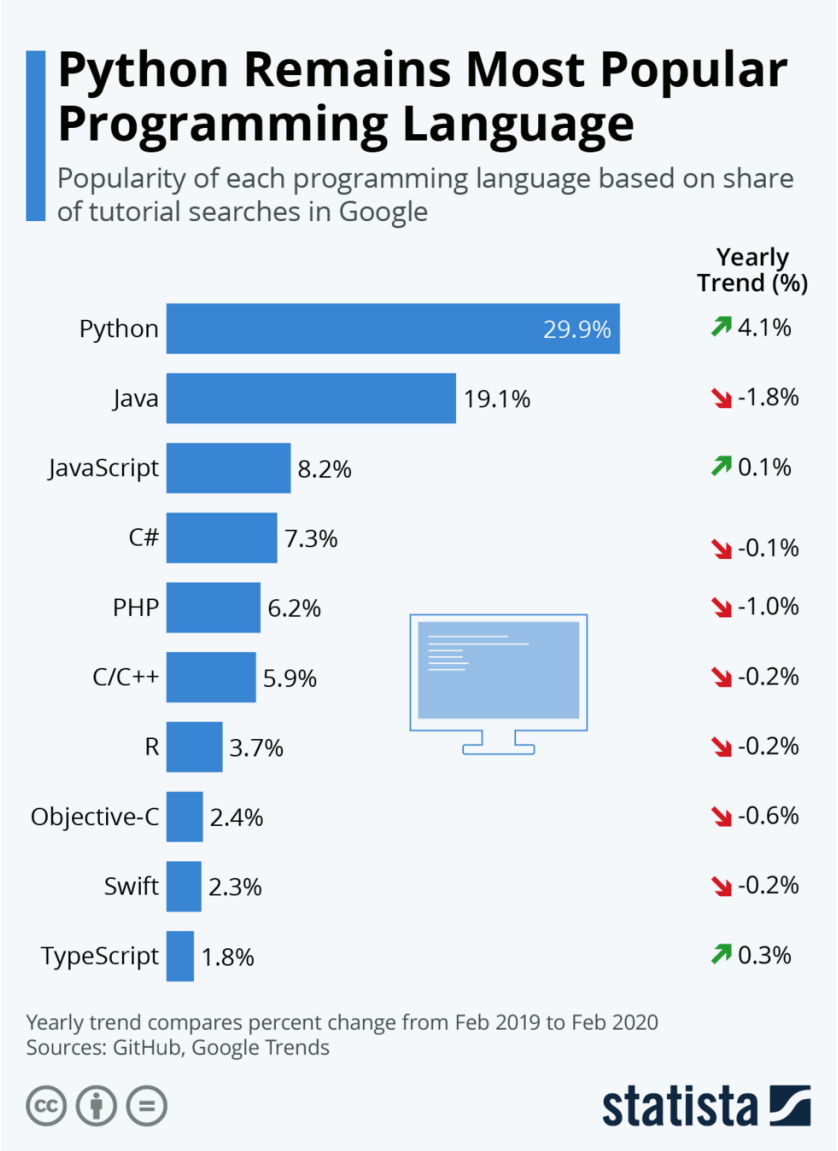
python™

why python??

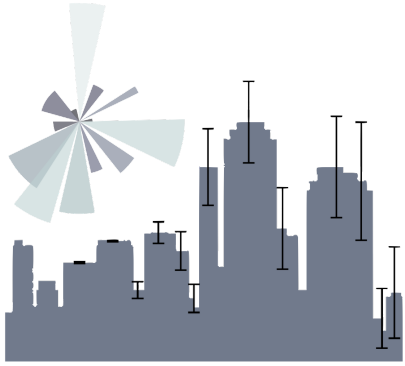
open source

huge community of developers

extremely popular (great skill on the job market!)



tech details



What is Colaboratory?

Colaboratory, or "Colab" for short, allows you to write and execute Python in your browser, with

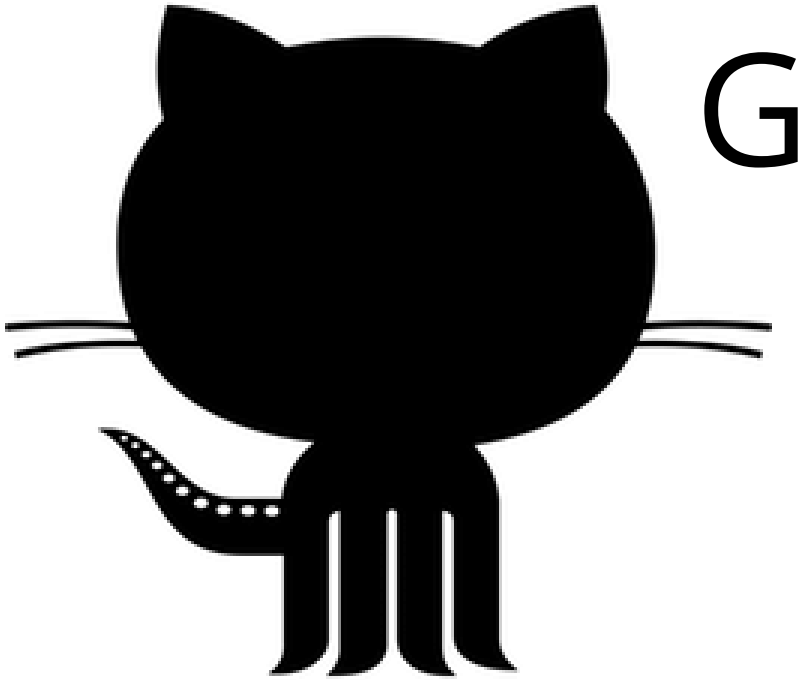
- Zero configuration required
- Free access to GPUs
- Easy sharing

Whether you're a **student**, a **data scientist** or an **AI researcher**, Colab can make your work easier. Watch Introduction to Colab to learn more, or just get started below!

https://colab.research.google.com/notebooks/intro.ipynb

DEMO TIME!

tech details



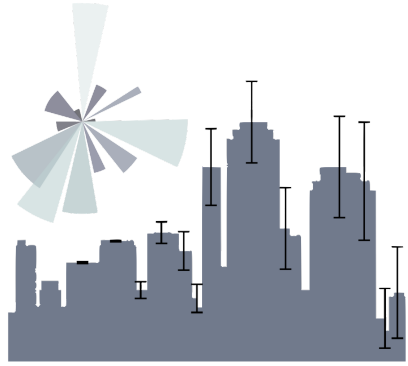
GitHub

distributed version control system: a version of the files on your local computer is made also available at a central server. The history of the files is saved remotely so that any version (that was checked in) is retrievable. Others can access and generate their versions of the files enabling collaborative work.

https://github.com/fedhere/PUS2020_FBianco

version control + collaborative coding

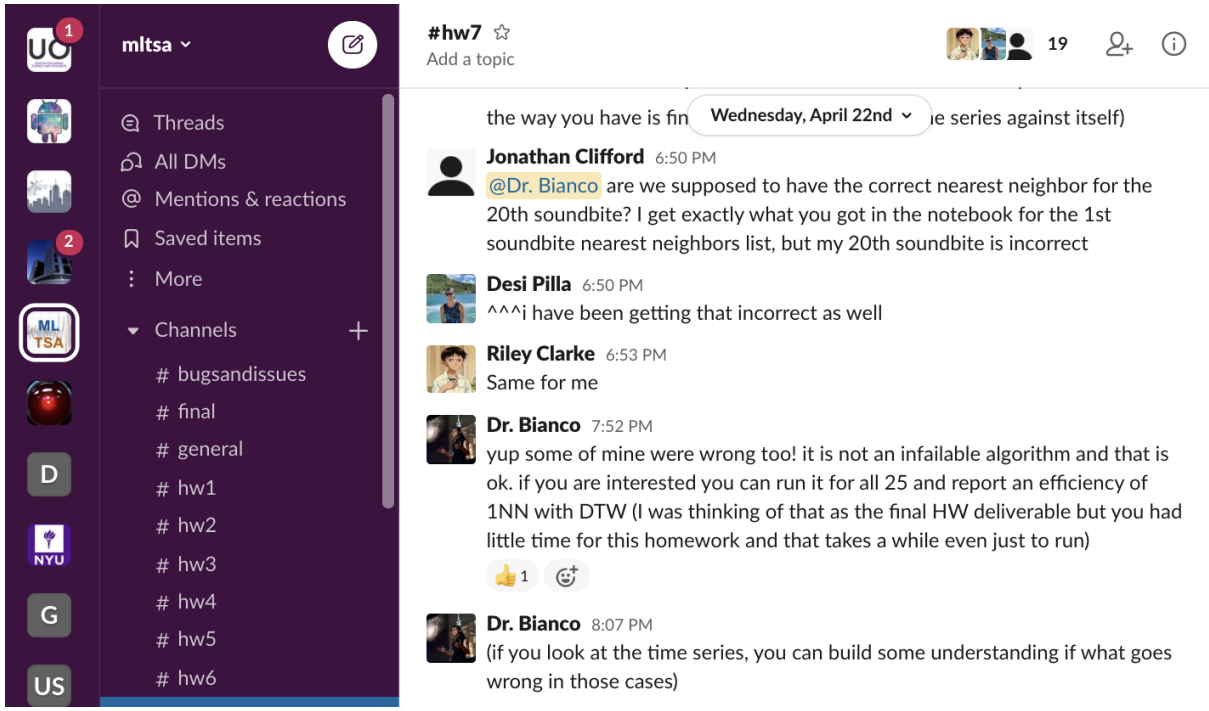
tech details



slack

https://join.slack.com/t/pus2020/shared_invite/zt-gzw71ste-9J7yGeLQAvmzZbLYuEri~g

practical and effective team communication tool



tech details

Canvas

used for grades and lecture videos

20F-SPPA667-011: SEMINAR

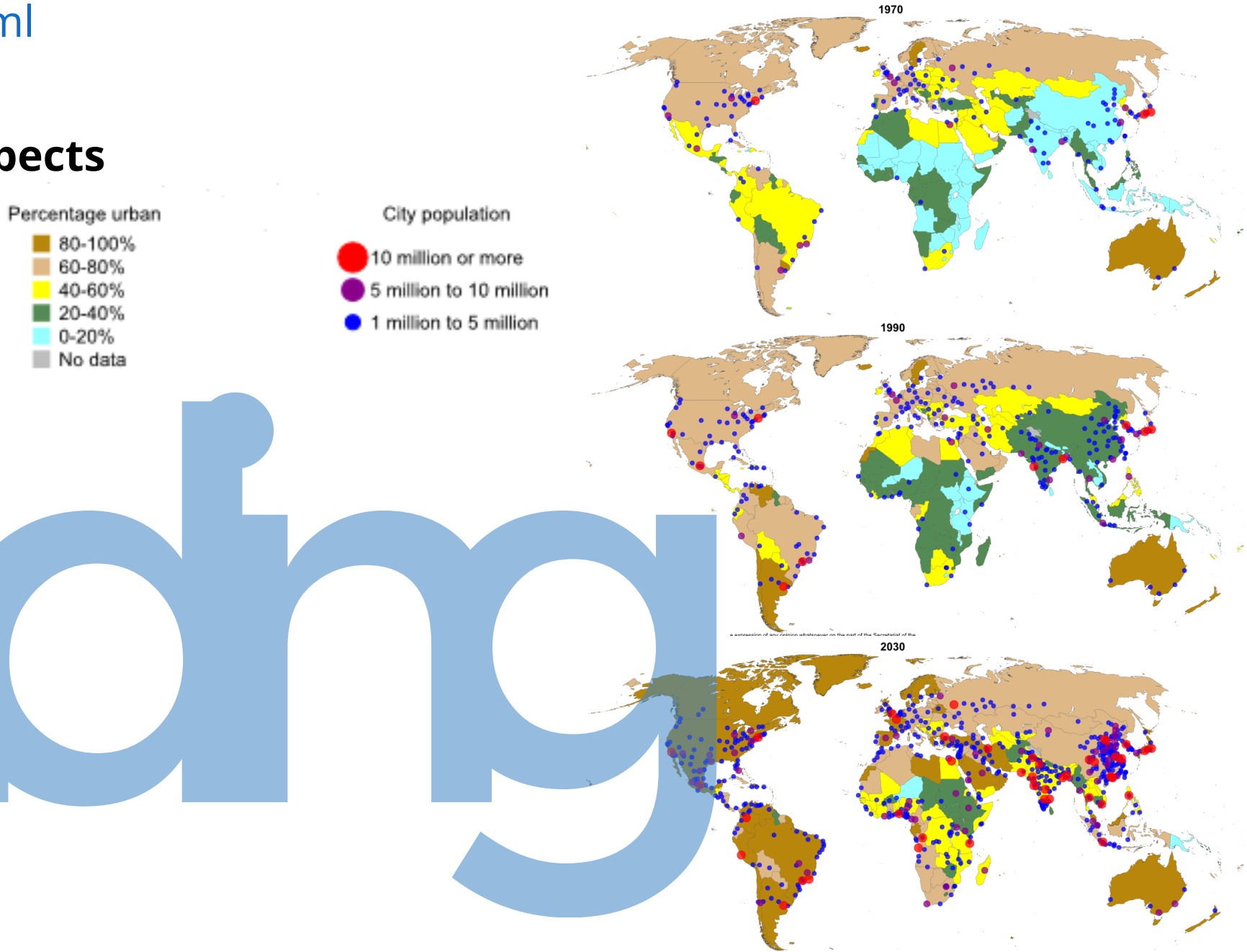
Jump to Today | Edit

GENERATING AND INTERPRETING EVIDENCE FOR EVIDENCE-BASED DECISION IN URBAN ENVIRONMENTS

A.K.A. PRINCIPLES OF URBAN SCIENCE - UDel 20

https://www.un.org/development/desa/en/news/population/2018-revision-of-world-urbanization-prospects.html

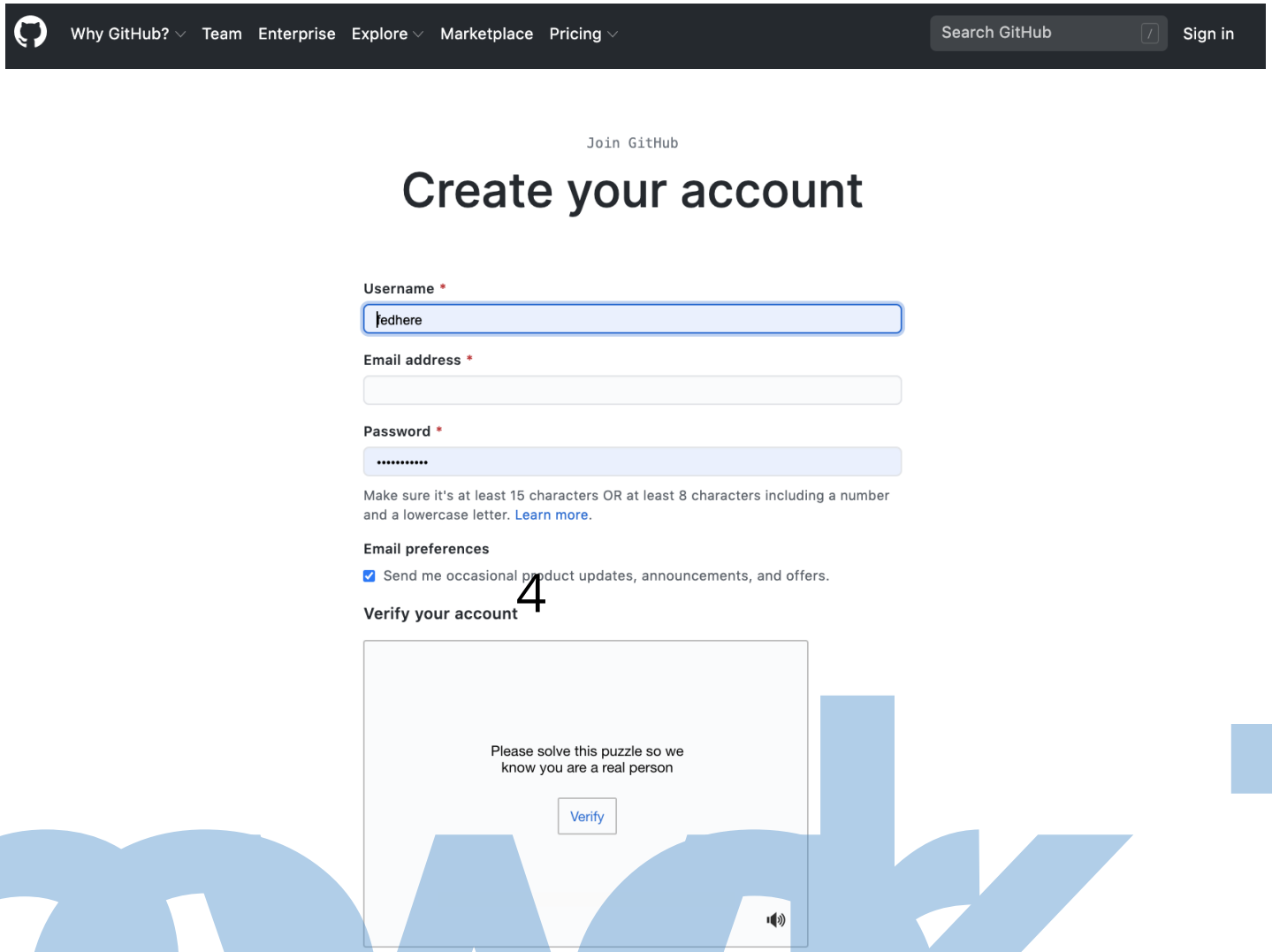
World Urbanization Prospects



reading

Create an account on github

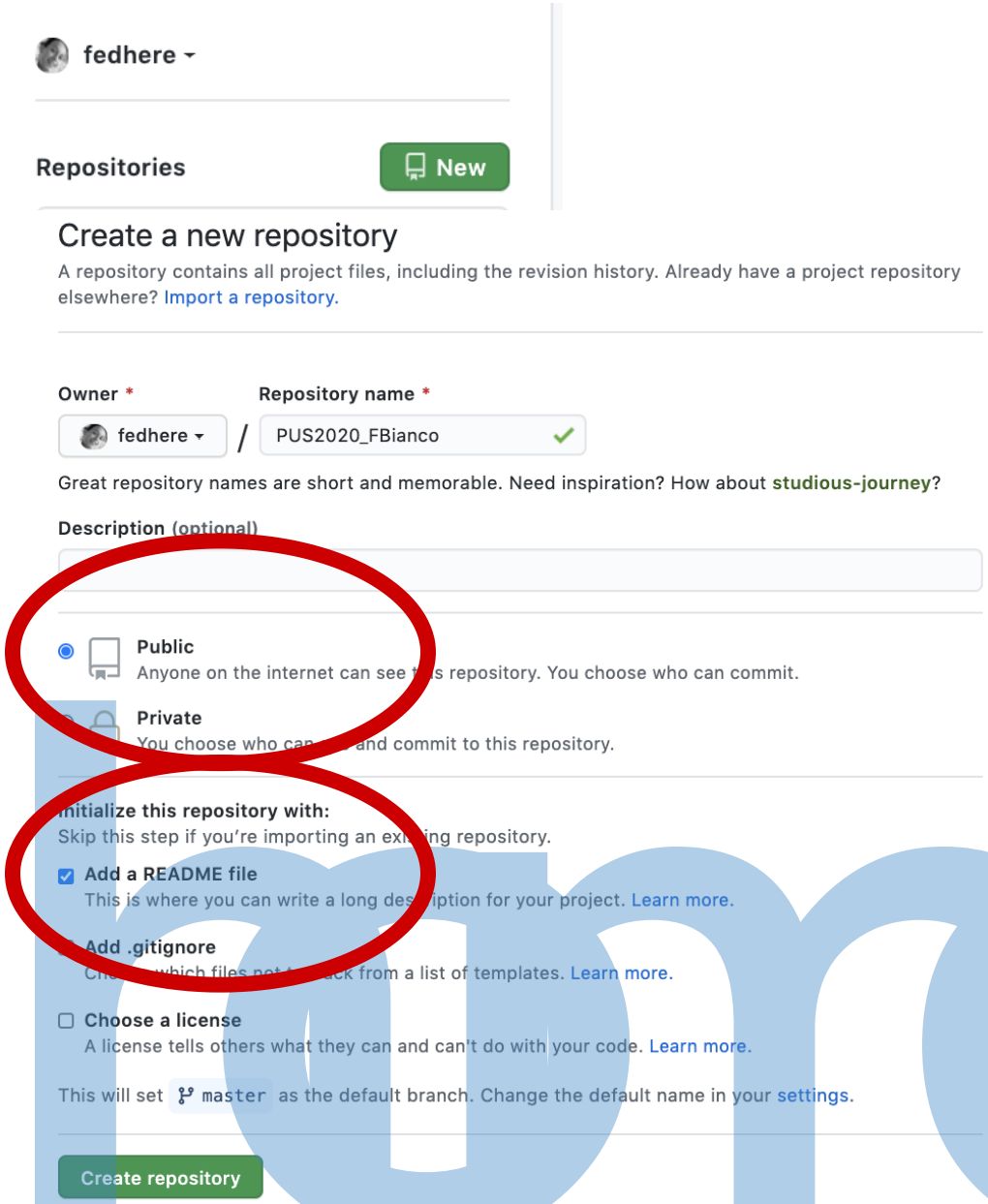
https://github.com/



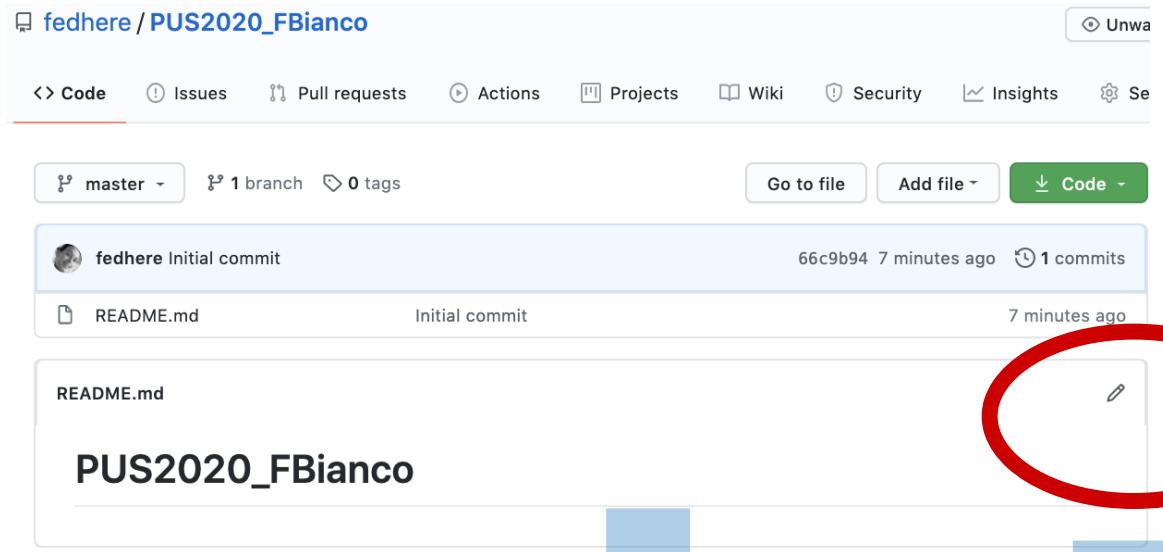
4

homework 1

Create a github repo PUS2020_<Firstname Initial><Last name>



Add a readme that describes why you are taking the class and what you expect to learn from it.



homework 1

Read the class Code of Conduct and answer questions about it in this form.
Include the URL of your PUS repo with the form response

https://forms.gle/W6QjYjLGWp7nF7w37