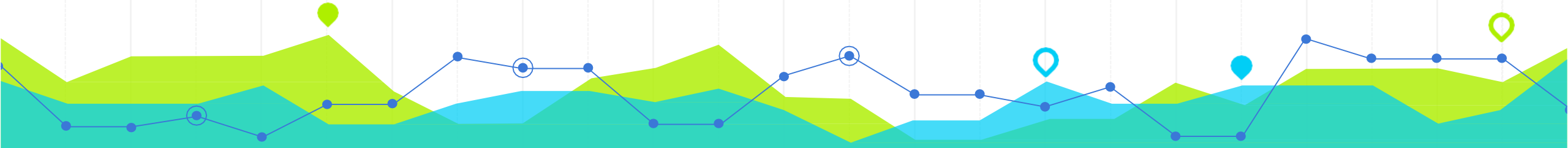
Introduction to Urban Data Science: Data, Interpretation, and Presentation

CRP/DESIGN 4680/5680

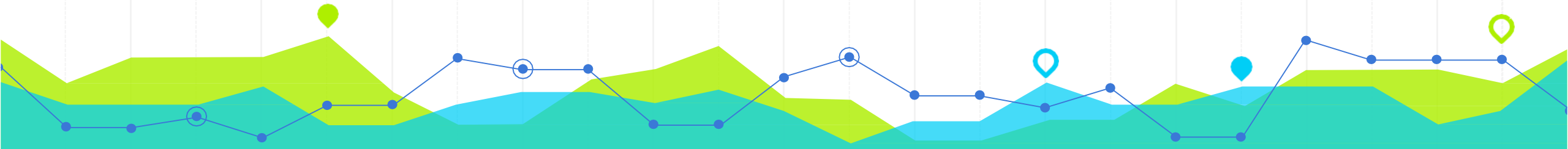


Lecture 1 Introduction

Wenzheng Li

Week 1: OUTLINE

- About the instructors
- First-class survey
- About the course
- Programming Platform Set-up



About the Lecturer 1

About Me

Wenzheng Li, Ph.D.
Visiting Lecturer, Cornell University

Research Interests

Regional Planning and Governance

- Sustainable urban forms in Sub-Saharan African cities
- Regional cooperation and coordination in Chinese regions

Methods:

- Urban Data Analytics,
- GIS/Remote Sensing,
- Econometrics

Working Experience

2018-2019—**Transportation planner**, Dept. Social Services, Tompkins County, NY

2015-2016—**GIS/Remote Sensing Analyst**, ToolGeo Company, Wuhan

Education Experience

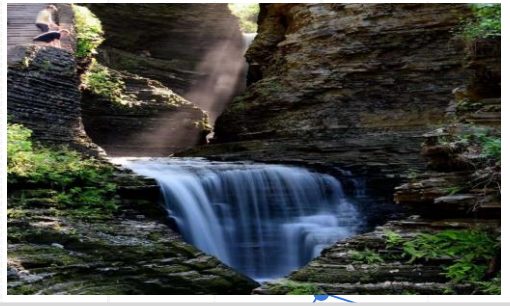
2019-2024—**Ph.D. in City and Regional Planning**, Cornell University

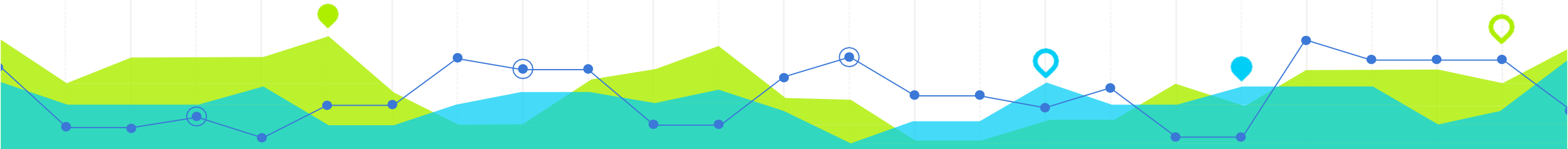
2016-2018—**Master in Regional Planning**, Cornell University

2012-2016—**B.S in Remote Sensing**, China University of Geoscience, China

Other roles...

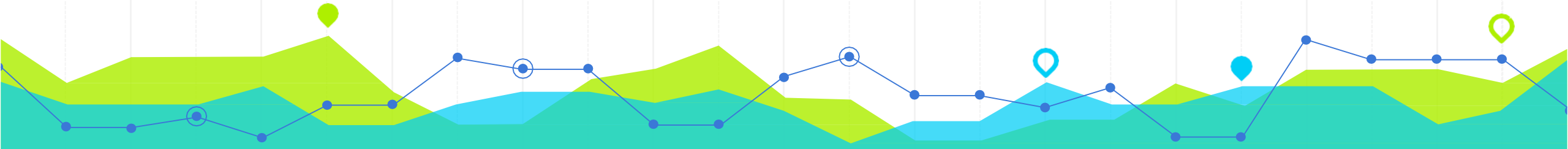
- Undergraduate Advisor
- Journal Reviewer
- Landscape Photographer & LEGO Fan





First-class survey 2

- Name, Major, Year of Study
- Why are you interested in the course?
- Previous experience with Python or any programming language, GIS, and Stats?



About the course

3

Urban Data Science

Urban Data Science

- Data gathering, preparation, and exploration
- Data representation and transformation
- Computing with data
- Data modeling
- Data visualization and presentation
- Science about data science

--- "*50 Years of Data Science*" Donoho (2017)

Article

B Urban Analytics and City Science

EPB: Urban Analytics and City Science
2019, Vol. 46(9) 1756–1768

DOI: 10.1177/2399808319882826
journals.sagepub.com/home/epb
SAGE

A roundtable discussion: Defining urban data science

Organizers

Wei Kang
University of California, Riverside, USA

Taylor Oshan
University of Maryland, USA

Levi J Wolf
University of Bristol, UK

Discussants

Geoff Boeing
University of Southern California, USA

Vanessa Frias-Martinez
University of Maryland Institute for Advanced Computer Studies, USA

Song Gao
University of Wisconsin, Madison, USA

Ate Poorthuis
Singapore University of Technology and Design, Singapore

Wenfei Xu
Columbia University, USA

…There is a distinction between “just” data science and “urban” data science. The distinction is the spatial component….

…Urban data science is data science applied to cities…

…Urban data science research must be embedded within the broader conversation in urban studies…

…Urban Data Science is an interdisciplinary study of applying computer science and statistical tools to understand urban issues and to inform urban decision-making…

……

Urban data science

1) the set of data analysis tools and methods used to understand a wide array of big data and big spatial data sources

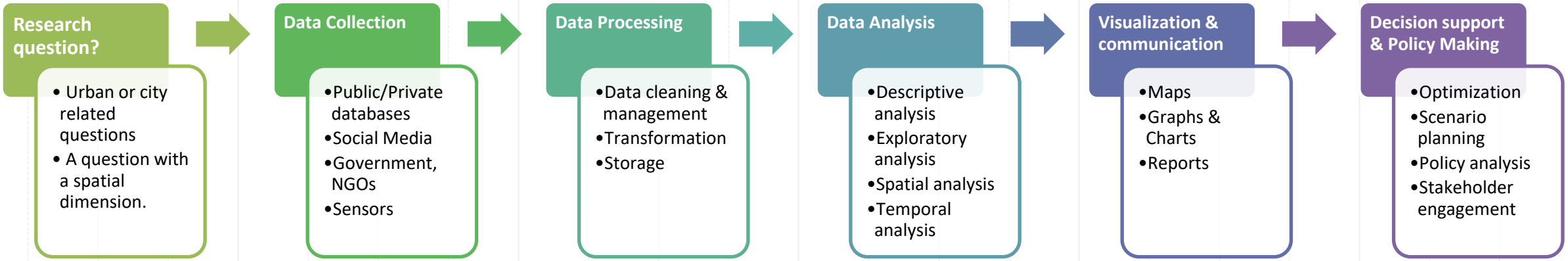
2) questions of urban development, structure, complexity, theory, policy, dynamics, and outcomes

Course objective:

(1) provide a toolkit to speak through data, code, statistics, and visualization.

(2) Using open-source data, Python and Jupyter Notebook, we will learn how to design testable research questions, collect and prepare data, apply relevant analytical techniques, present our process and results, and identify the limitations of quantitative analysis.

A personal laptop will be required.



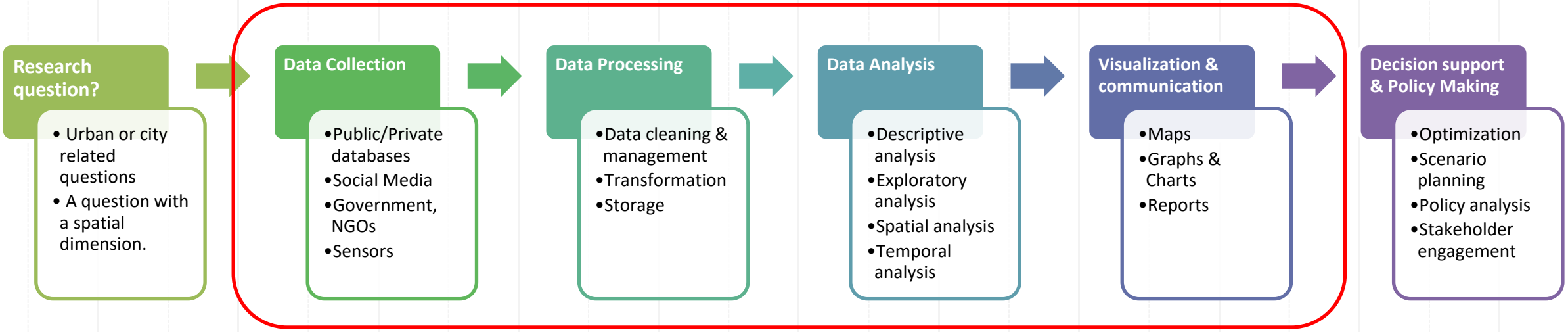
Urban Data Science Workflow

Course objective:

(1) provide a toolkit to speak through data, code, statistics, and visualization.

(2) Using open-source data, Python and Jupyter Notebook, we will learn how to design testable research questions, collect and prepare data, apply relevant analytical techniques, present our process and results, and identify the limitations of quantitative analysis.

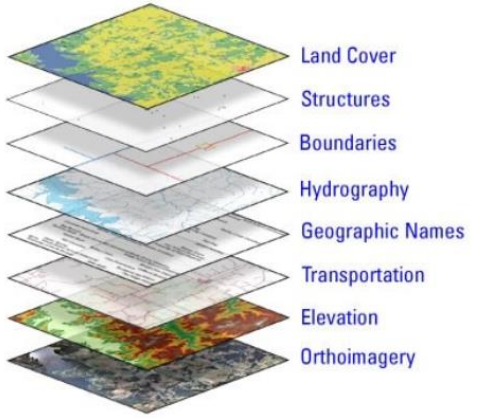
A personal laptop will be required.



Urban Data Science Workflow

New features in Urban Data Science?

- Geospatial Analysis
  - integrates spatial data and GIS tools to analyze location-based patterns and relationships.
- Big Data Integration
  - leverages massive, real-time datasets from diverse sources like social media and satellite imagery.
- Machine Learning, AI, and Predictive Analytics
  - Advanced algorithms enable pattern recognition, predictions, and modeling of complex urban systems.




BIG DATA
INTERNET
SOCIAL
BUSINESS
VISION
BRAND
WEB MARKETING
ORGANIZATION
SOLUTIONS
BRANDS
SERVICE
PROJECTS
CONTENT
BIG DATA
SOCIAL MEDIA
ADVERTISING
PROMOTION
TARGET
PLAN
E-MARKETING
COMMUNICATION
COMPUTER
BRANDING
CONSUMER
DEMAND
MARKETS
SERVICES
PROJECTS
MULTIMEDIA
NETWORK
WWW
INFORMATION TECHNOLOGY
CONSUMER
ORGANIZATION
PLANNING
EVENTS
PROGRAMMING
SOFTWARE
CODING
STATISTICS
APPS
SOCAIL NETWORKS
SOLUTIONS
RESEARCH
COM
ENGINEERING
WEB DEV
STRATEGY
WORLDWIDE
WEB SERVICES
BIG DATA
BUZZ
BEHAVIOUR
PRO
PRICING
B2B TARGET
PLANNING
MEDIA
MOBILE
INFORMATION
DIGITAL
SEGMENTATION
SOCIAL NETWORKS
CODING

Volume

– from Giga to Peta

BIG Data

Volume

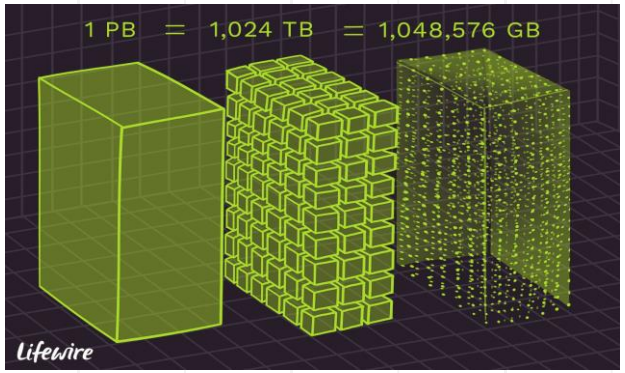
– from Giga to Peta

120GB

BIG Data

Volume

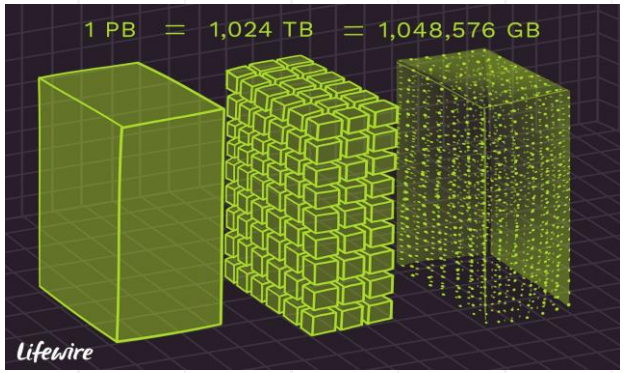
– from Giga to Peta



BIG Data

Volume

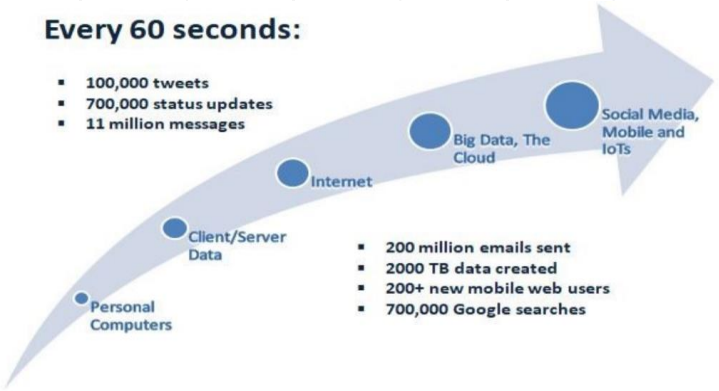
– from Giga to Peta



BIG Data

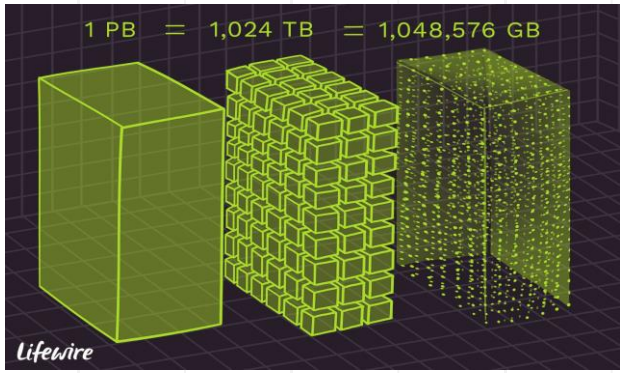
Velocity

–speed of data streaming in near real-time



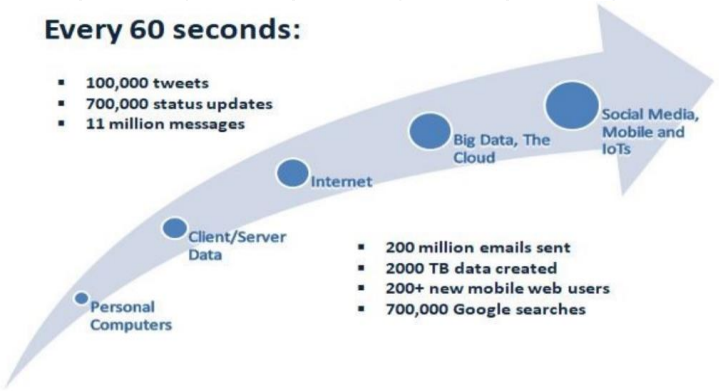
Volume

– from Giga to Peta



Velocity

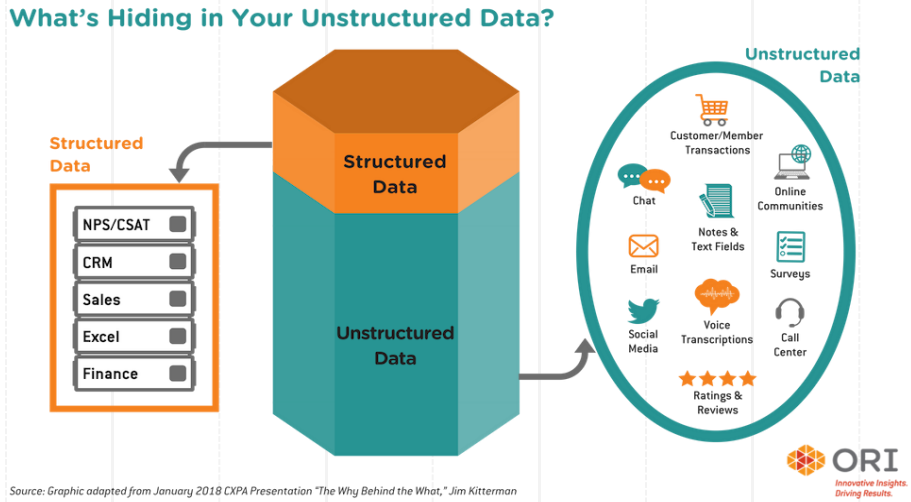
–speed of data streaming in near real-time



BIG Data

Variety

– Heterogeneous data and databases in different formats, structured and unstructured.

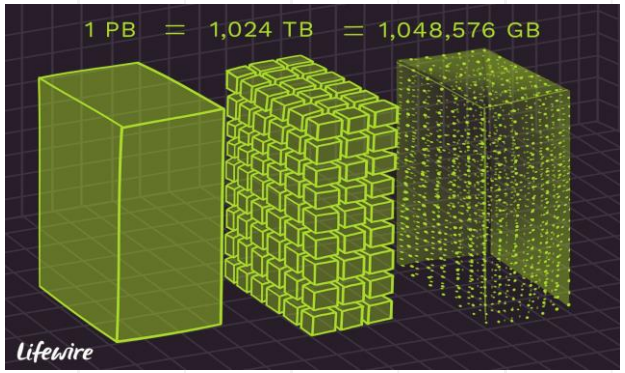


Source: Graphic adapted from January 2018 CXPA Presentation "The Why Behind the What," Jim Kitterman

BIG Data

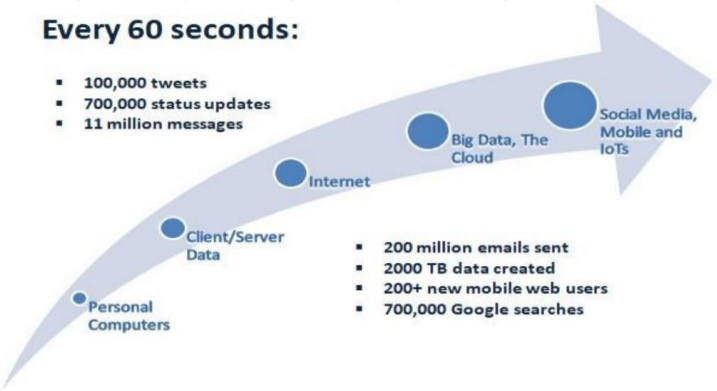
Volume

– from Giga to Peta



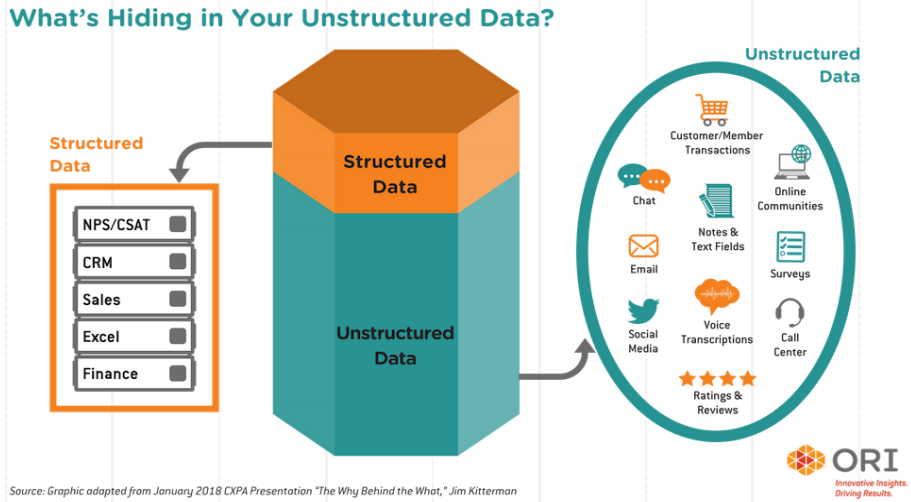
Velocity

–speed of data streaming in near real-time



Variety

– Heterogeneous data and databases in different formats, structured and unstructured.



Veracity?

–The degree of the data to be trusted? Inconsistency, incompleteness, ambiguity, latency, etc.

What is Machine Learning?

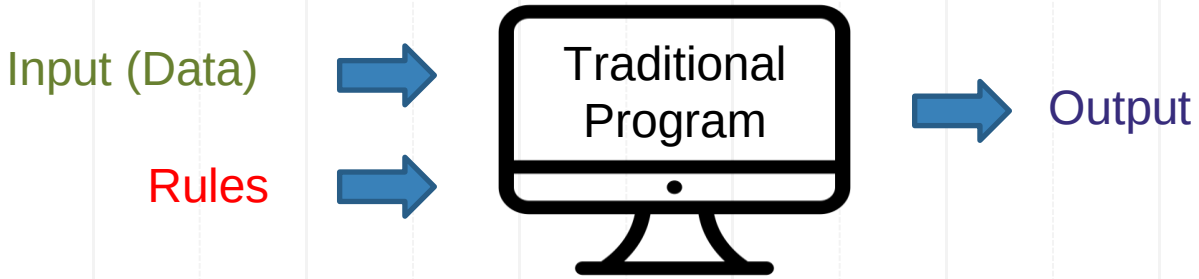
[Machine Learning is the] field of study that gives computers the ability to learn without being explicitly programmed.

--Arthur Samuel, 1959

*Samuel, Arthur (1959). "Some Studies in Machine Learning Using the Game of Checkers". IBM Journal of Research and Development. **3** (3): 210–229.*

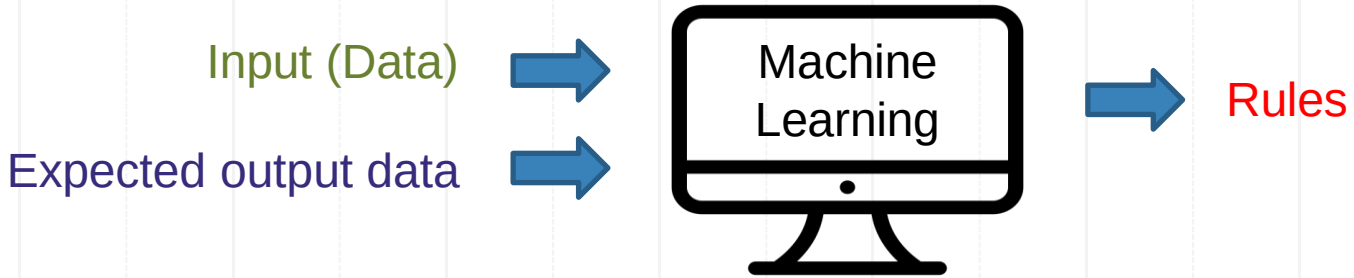
Traditional program:

developers give computers explicit instructions to follow.



Machine Learning:

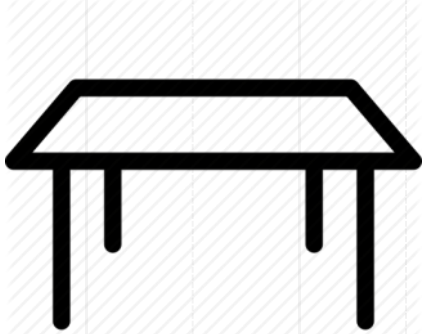
Machine learning uses algorithms to learn patterns from data and make predictions.

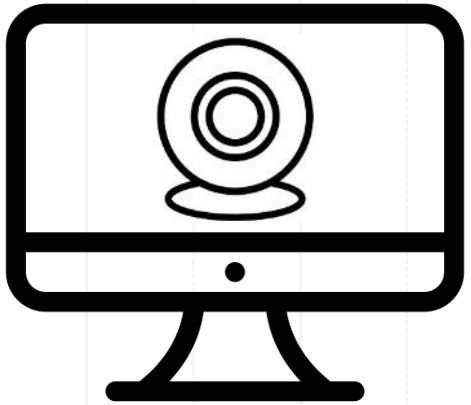


What is Machine Learning?

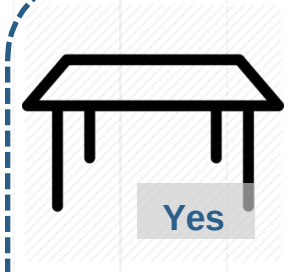
- **Example: Image Recognition**

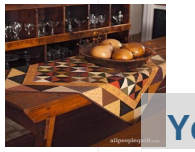
  Task: whether there is a table in the image

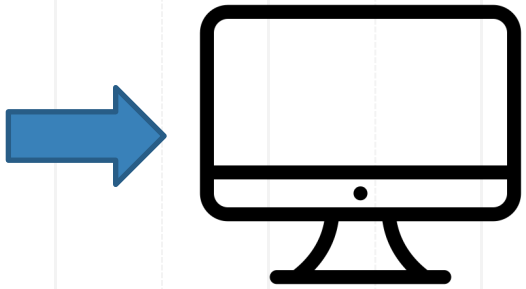




A table detector based on machine learning

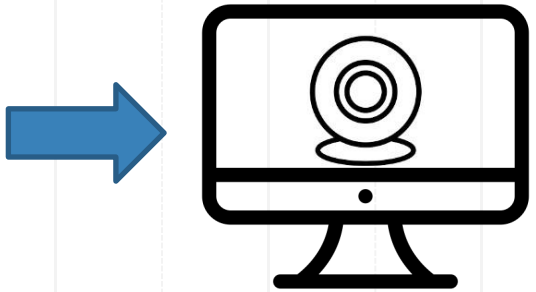






Common features:
Flat Surface?
Legs or Support?
Height Consistency?
Edges and Corners?
Shadows and Reflections?

Training



Yes

features

No

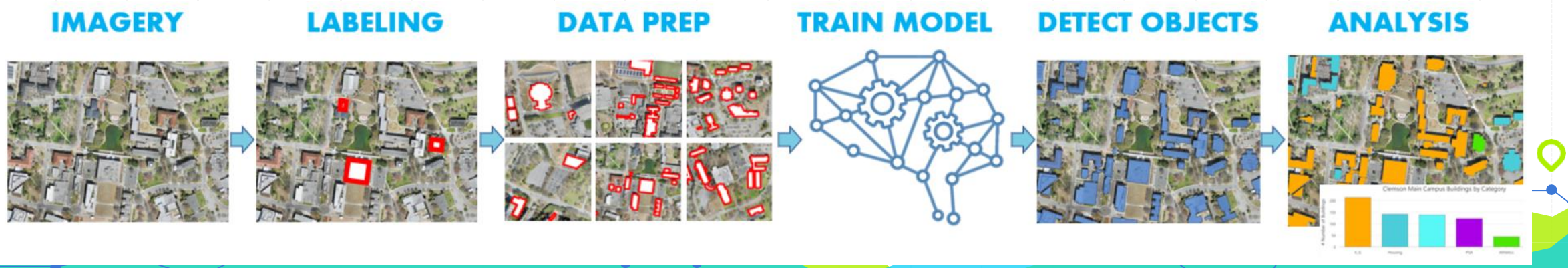
features

Prediction

Applications-computer vision, object detection, and deep learning

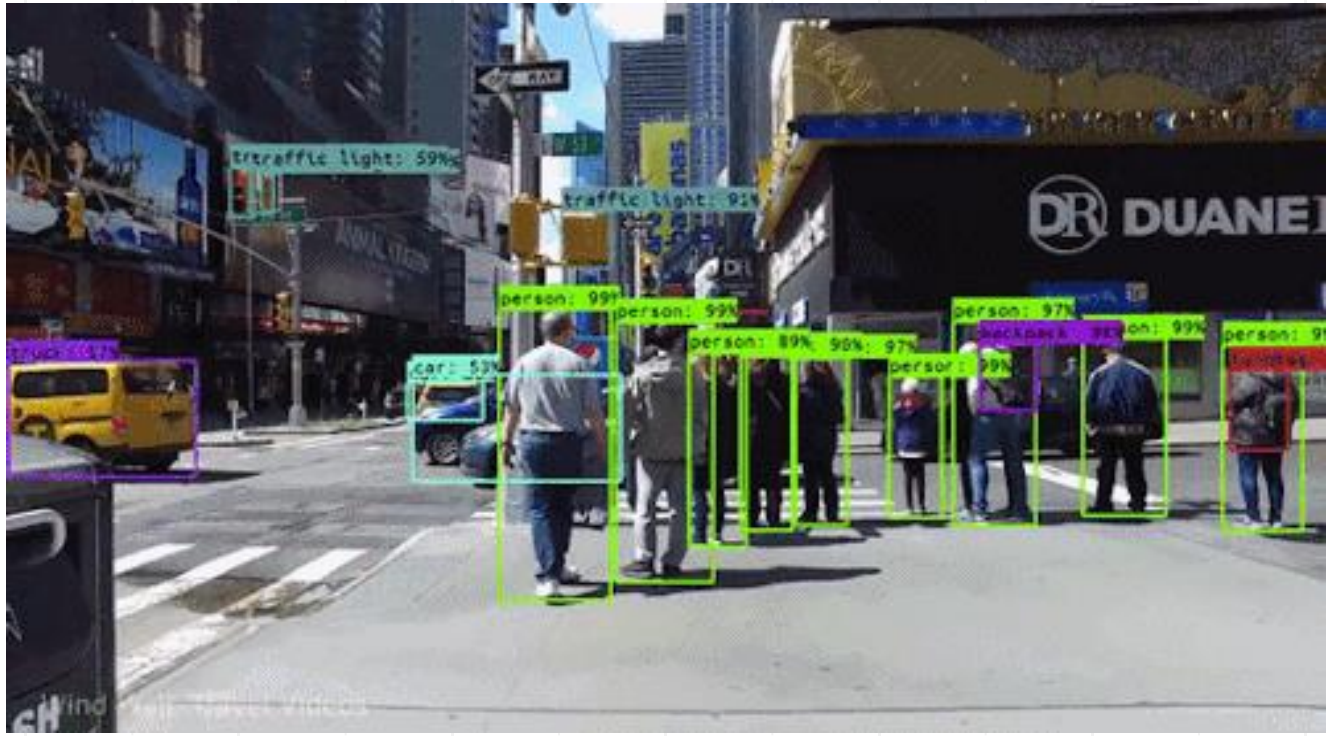
Model the real world for prediction
Aerial imagery is used to extract imagery of buildings and roads in Grenada to identify the population and infrastructure at risk for landslides.

Deep learning workflow: building detection



Applications-computer vision, object detection, and deep learning

- For ex: image segmentation
- Guess what we use for autonomous vehicles to detect objects in space?



Machine Learning?

- **Applied science**: hypothesis testing, model assumptions, explanation, and interpretation

- In machine learning, different than in applied statistics, we are less interested in what these parameters are, and more in how well they can
    - Make **predictions**
    - Describe **underlying structures or characteristics in the data**

Scope of this course

Section 1: Introduction to Python and Data Techniques

- Basics of Python
- Data management
- Non-spatial data visualization

Week	Tuesday	Thursday
	Section 1: Introduction to Python and Data Techniques	
1	**Jan 21** **Introductions and Course Overview** • Read over the syllabus together • Introducing ourselves • Open science and the modern urban data science software stack	**Jan 23** **Coding environment setup** • Setting up your Python and Anaconda coding environment **Lab Session:** • Setting up the coding environment. • Organize notebooks through markdown
2	**Jan 28** **Basics of Python:** • Basic syntax; • Variables and flows; • List, tuple, dictionary, set; • If-statement and for-loop.	**Jan 30** **Data management using *Pandas* 1** • Python Packages • Basics of Pandas: DataFrame, import and export datasets, built-in functions
3	**Feb 4** **Data management using *Pandas* 2** • filtering a DataFrame: indexing and slicing • Data cleaning	**Feb 6** **Data management using *Pandas* 3** • Linking datasets, • overlaying and aggregating data, • re-classifying data with pandas
4	**Feb 11** **Data visualization:** • Basic plots using *Pandas, Matplotlib*, and *Seaborn* • Customizing your plots • Interactive visualization using *Folium* and *Bokeh*	**Feb 13**

Python Basics

1.1 Variables and data type

1.1.1 Creating a variable
1.1.2 Data types
1.1.3 Data type conversion

1.2 Operators
¶

1.2.1 Arithmetic operators
1.2.2 Comparison operators
1.2.3 Logical operators

1.3 List

1.3.1 Defining a list
1.3.2 List concatenation
1.3.3 Subscript indices and slices (IMPORTANT)

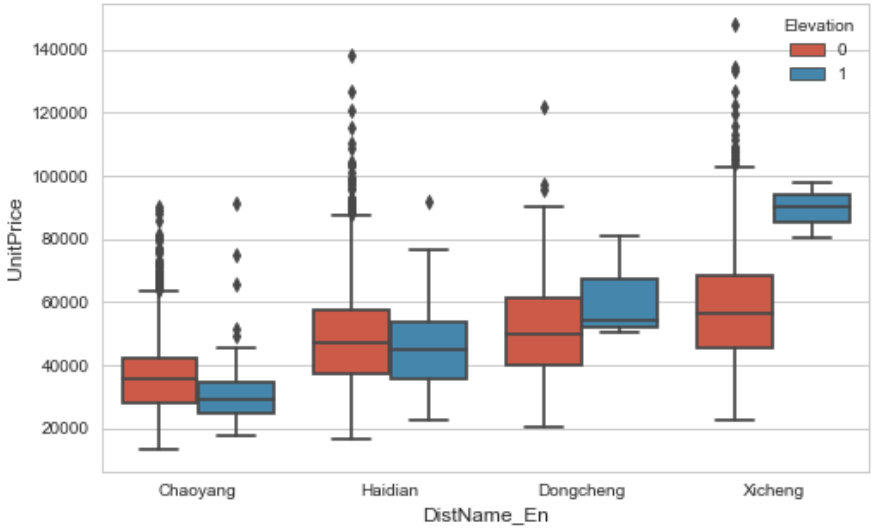
1.4 String

1.5 Dictionary

1.6 if statement

1.7 for-loop

Visualization



Data Management

	HouseID	CommunityID	TotalPrice	TransYear	Bedroom
0	BJFT84326414	1544	1400010.56	2012	2
1	BJCP84958845	2606	1800066.00	2012	3
2	BJDX84905788	2264	1350038.34	2012	2
3	BJFT00386624	3621	1800006.91	2012	2
4	BJCY84713854	1127	1970019.58	2012	1

5 rows × 30 columns

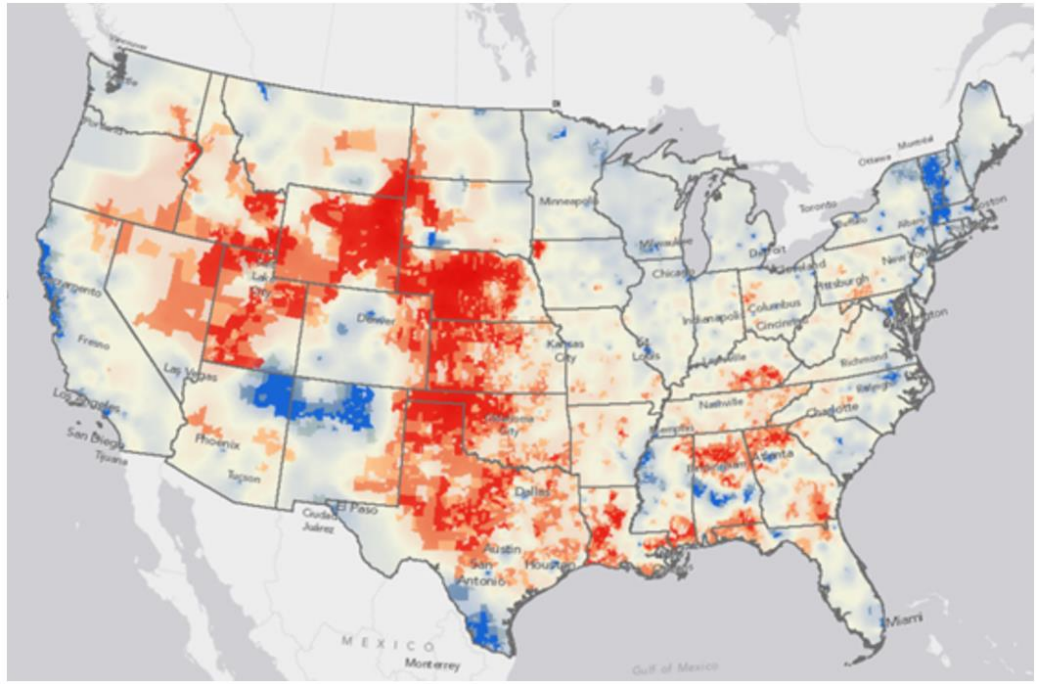
Scope of this course

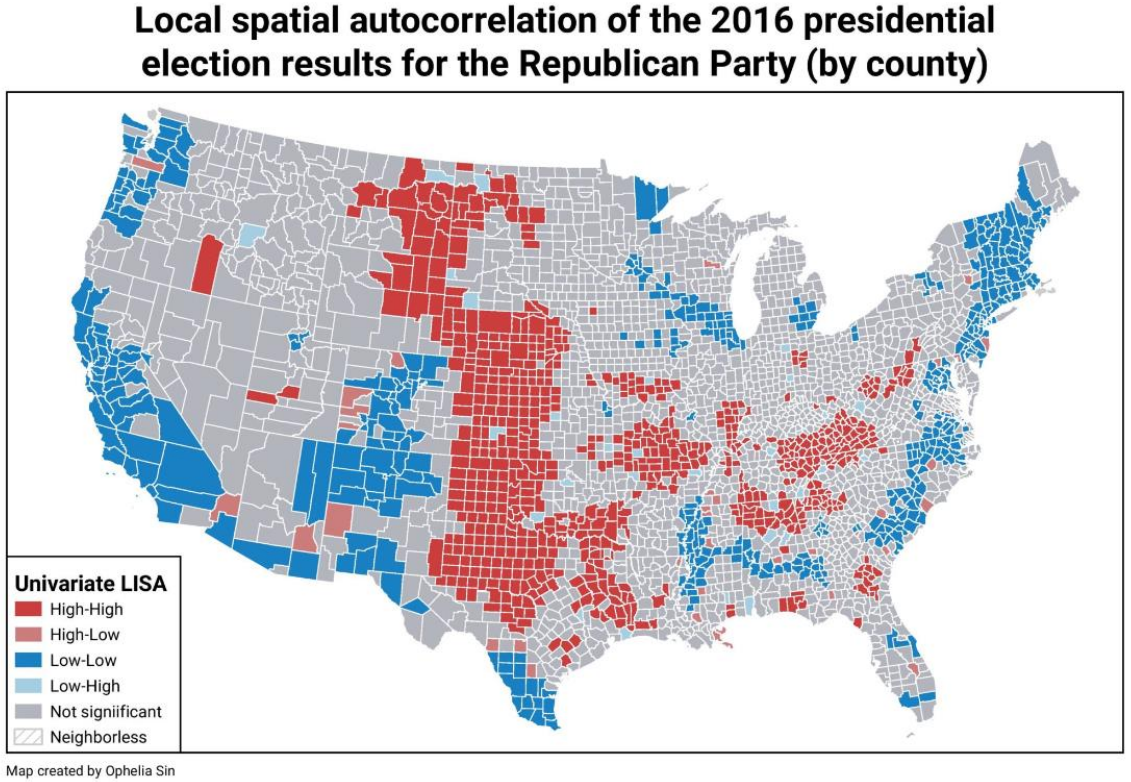
Section 2: Exploratory Spatial Data Analysis (ESDA) and Spatial Econometrics

- Geospatial data management and operation
- Spatial data visualization (mapping)
- Spatial autocorrelation
- Spatial regression (econometrics)
- Data collection

Week	Tuesday	Thursday
5	**Feb 18** **NO CLASS – FEBRUARY BREAK**	**Feb 20** **Geospatial operations 1:** • Basics of GeoPandas • Geometry and Projection • Spatial join
6	**Feb 25** **Geospatial operations 2:** • Spatial data visualization • Choropleth maps using *GeoPandas*	**Feb 27** **Spatial data analysis 1:** Spatial weights with *pysal*
7	**Mar 4** **Spatial data analysis 2:** Spatial autocorrelation with *pysal*	**Mar 6** **Spatial data analysis 3:** Point pattern analysis
8	**Mar 11** **Regression 1:** Linear regression with *statsmodels* and *scikit-learn*	**Mar 13** **Regression 2:** Spatial regression with *pysal*
9	**Mar 18** **Data Collection 1:** • Google Map APIs for geocoding and distance calculation	**Mar 20** **Data Collection 2:** • Web-scraping or using OSMnx package to obtain OpenStreetMap dataset

Moran’s I and spatial dependence analysis





2008 and 2016 Presidential Election Results with red areas reflect strong Republican party wins and blue areas reflect strong Democratic Party wins

Scope of this course

Section 3: Machine Learning

- Unsupervised learning: Dimensionality reduction and Clustering
- Supervised learning: decision trees, random forest, classification, cross-validation
- Guest speakers: (1) computer version and deep learning; and/or (2) natural language processing

Week	Tuesday	Thursday
10	**Mar 25** **Unsupervised learning 1**: Dimensionality reduction and K-means clustering with *scikit-learn*	**Mar 27** **Unsupervised learning 2:** Spatial clustering through DBSCAN
11	**Apr 1** **NO CLASS – SPRING BREAK**	**Apr 3** **NO CLASS – SPRING BREAK**
12	**Apr 8** **Supervised learning 1:** Ensemble learning with decision trees and random forest models with scikit-learn	**Apr 10** **Supervised learning 2:** Regression vs classification, model selection, bias-variance tradeoff, and cross-validation with scikit-learn
13	**Apr 15** **Special topics**: • Spatial networks or • Web-scraping or • Guest speaker: Interactive mapping • Guest speaker: Natural language processing	**Apr 17** **Guest Speaker:** Prof.Waishan Qiu from the University of Hong Kong

Class Structure

Weeks 1-13: Tuesdays and Thursdays (8:40-9:55am)

- lecture (concepts) + codebook (code explanation) + in-class exercise

Week 14-16: Tuesdays and Thursdays (8:40-9:55am)

- in-class work and one-on-one final project meeting

Lab session: Thursday (4:30-5:20pm); Can we change to 4:45-5:35pm

- practice and review the concepts,
- discuss weekly in-class exercises and homework assignments

Course Pre-requisites

This course is designed for masters students and upper class undergraduate students.

- CRP4080/5080 (**Intro to GIS**) or an equivalent course is a prerequisite for the course.
- Additionally, I assume you have some basic statistics knowledge, such as descriptive statistics, hypothesis testing, basic regression and some familiarity using spreadsheet software (Excel, Google spreadsheets).
- Prior or concurrent coursework in quantitative methods, visualization, and programming is recommended.

Assignments and Grading

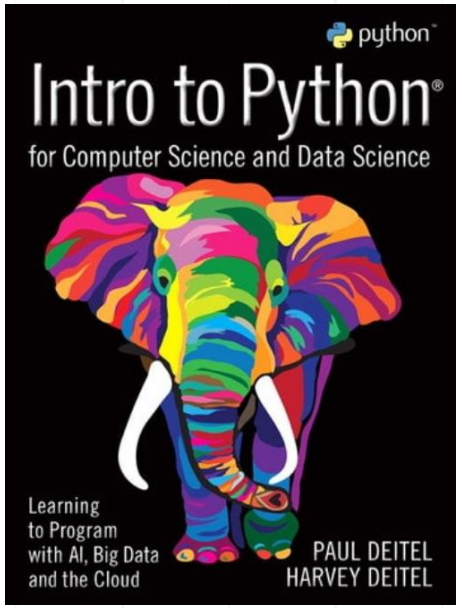
- (15%) **Weekly In-class exercises**
- (35%) **Homework**: There will be 4-5 HW assignments.
- (10%) **Class attendance and participation**
- (40%) **Final Project**: Students will develop a research project (individual/group of 2)
  - (5%) **Proposal:** due on **March 28 at 11:59pm**
  - (5%) **Presentation**: you will present your project at the end of the semester. The presentation should include the research question, data, descriptive analysis, methods, and **preliminary** results.
  - (30%) **Paper**: A written paper (about 15 pages) including final results (due **May 16 at 11:59pm**).

Textbook and help resources

No specific Python textbooks are required for this class. But some books could be helpful. *Rather than reading them from start to finish, use them as reference guide or dictionary.*

(1) Python Data Science Handbook by Jake VanderPlas (available via Canvas)

(2) Intro to Python by Paul Deitel & Harvey Deitel



Week 1: OUTLINE

- About the instructors
- First-class survey
- About the course
- Programming Platform Set-up

TA and instructor office hours

Instructor
Office Hours: Monday 2:30 – 4:30pm and Wednesday 11:00am-1:00pm in Sibley Hall 214. Book a time here

TA and GTRS
Yujin Hazel Lee (TA) yl3276@cornell.edu
Office hours: Thursday 5:30-6:30 pm in Sibley Hall 305
Xi Guan (GTRS) xg298@cornell.edu
Office hours: Tuesday 4:30-5:30pm in Sibley Hall 305

Generative AI Policy

Generative AI

Tools such as ChatGPT can often facilitate the generation of functions, processes, and frameworks in coding. They can be a useful aid in our analytical process. As such, we will learn how to work with ChatGPT in our homework assignments. *The use of generative artificial intelligence (AI) tools is permitted for coding with proper attribution*. Additionally, as I mentioned in the "Academic Integrity" section, be prepared to verbally explain what your code is doing. There are some very fundamental concepts in Python, machine learning, and regression that I would like you to actually learn.

Undergrad vs. Grad

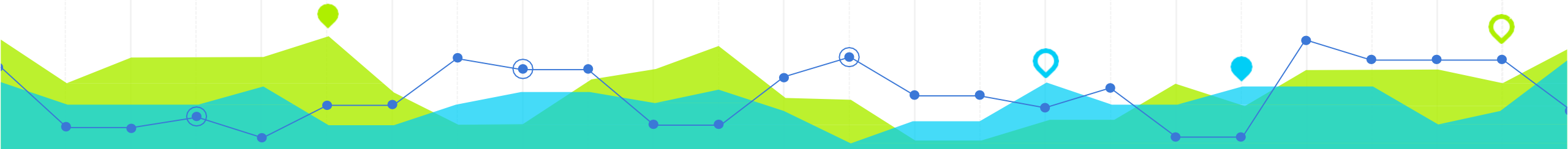
There is an undergrad and graduate version of this course. My expectations for graduate students enrolled in this course is for the final projects to be more research-oriented. I will expect you to complete a literature review for your proposal in order to justify the research question you investigate.

A note about learning to code

1) Experiment with small changes in existing code.
2) Read the error and decide if it is helpful or not. Trust the error message.
3) Look for **typos** in the code.
4) Search for the issue on Google - lead you to sites such as Stack Overflow or Medium, which provides code snippets and sometimes step-by-step instructions on how to resolve your question. Try to be specific in your search. Do not be afraid to sound silly. My search generally involves the following keywords:
    a. [language or tool] ex: “Python”, “Pandas”, “Matplotlib”
    b. [function or action] ex: “plt.subplots”, “plotting multiple plots in one figure”
    c. [error or issue] ex: “plots are tiny”, “not showing all plots”, etc.
5) If trying to implement a fairly standard process, look through our class notebooks or the readings. There are often code snippets for reference there.
6) Ask classmates.
7) If none of the above is fruitful,

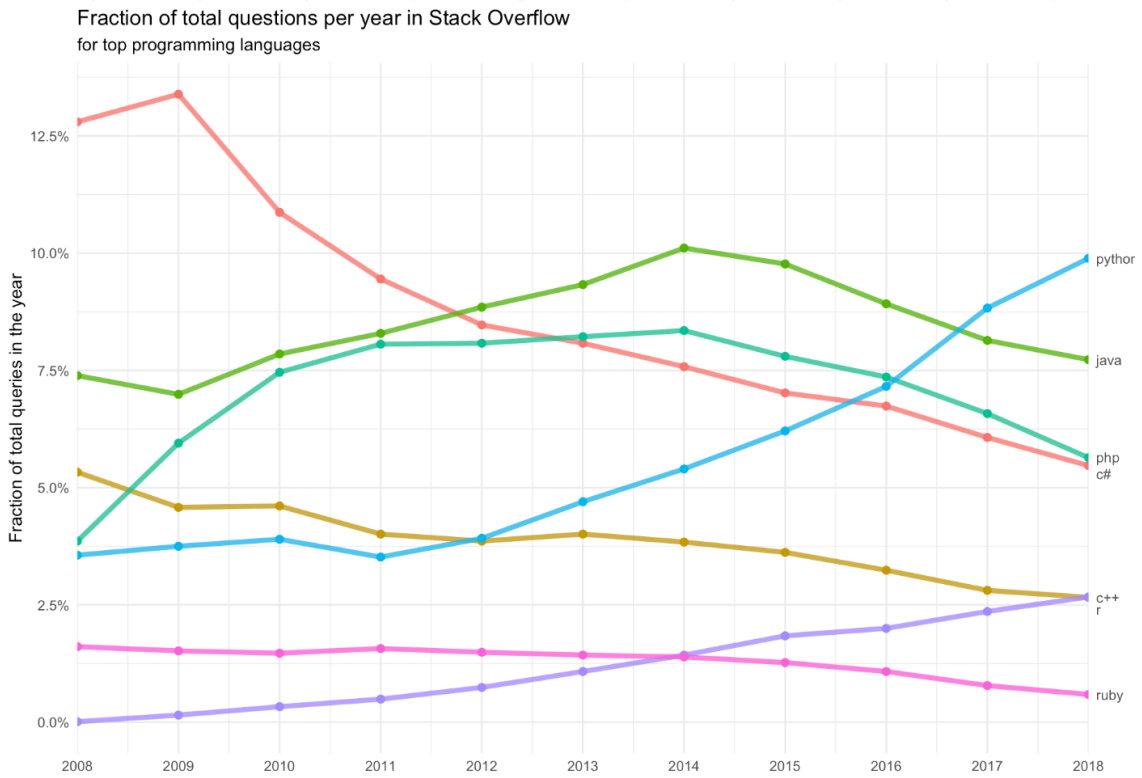
    a) come to our TAs and my office hours.

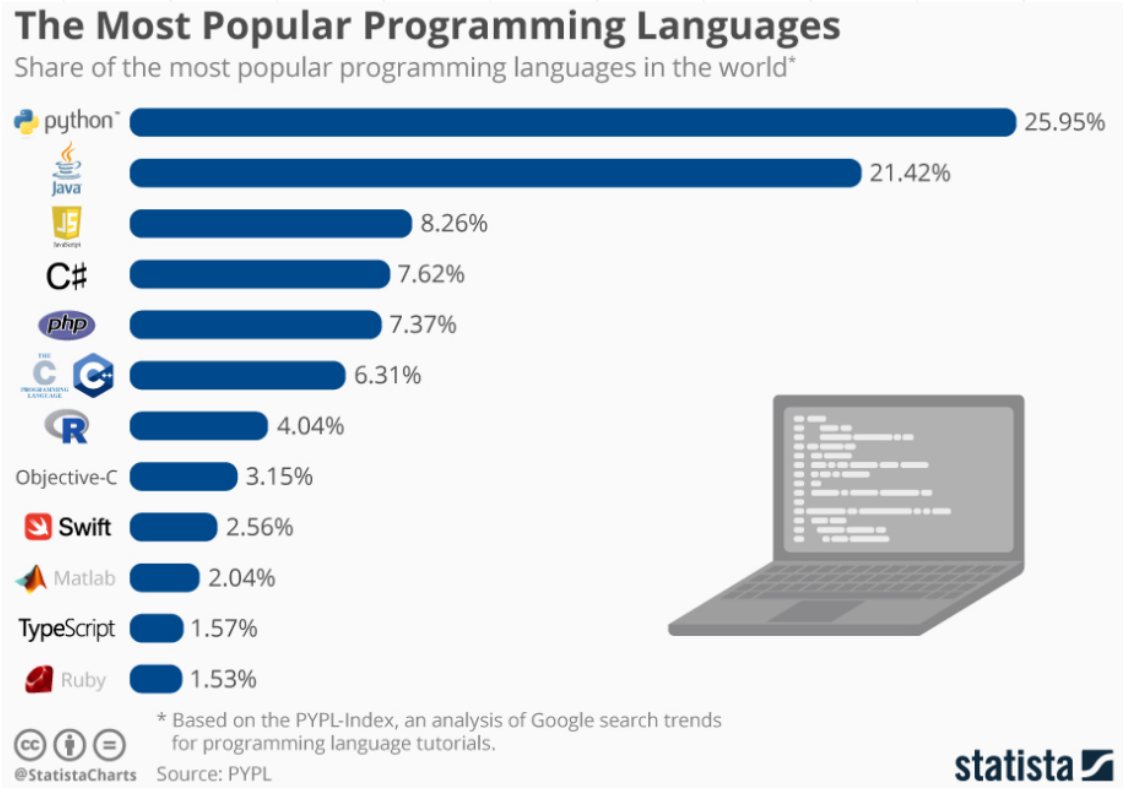
    b) you may want to message Yujin (Hazel), Xi, or me on the discussion session via Canvas with the specific task and the relevant code snippet either as a screenshot or a Github Gist. Do not send code in the body of an email as rich text editors often add hidden formatting that can introduce new code errors.



Python Platform Set-up 4

Why Python?





https://www.statista.com/chart/16567/popular-programming-languages/

https://stackoverflow.blog/2017/09/06/incredible-growth-python/

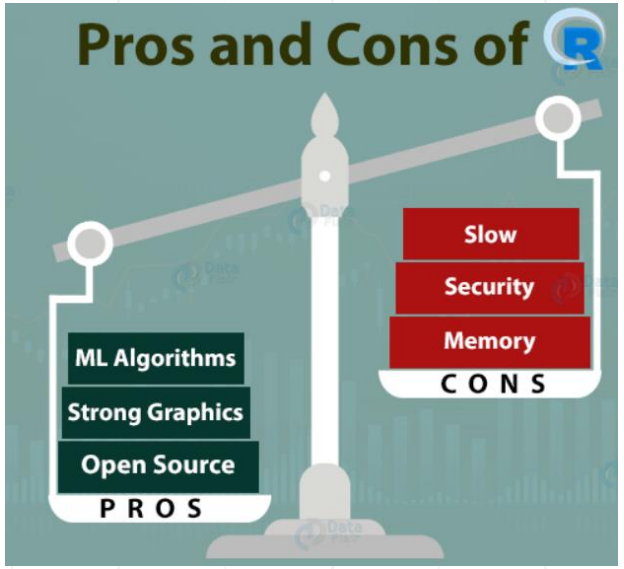
Python

- A general-purpose Programing Language
- Efficient for repetitive tasks
- Able to handle “big data”
- An integrated platform for workflows (e.g., collect, manage, and visualize data)
- A strong community developing powerful tools (Python libraries)

Limits

1,048,576 rows

16,384 columns



Python Platform Set-up

Building the Programming Platform:

ANACONDA.

Pandas
Data analysis and manipulation

Matplotlib
Data visualisations

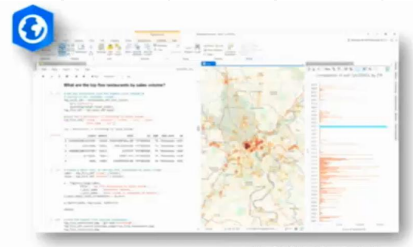
NumPy
Mathematical functions

SeaBorn
Data visualisations

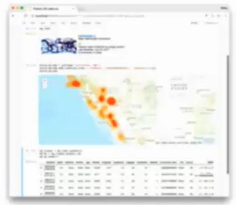
Python packages

Jupyter

- Interactive, typically browser-based environment
- Allows you to combine code and markdown cells
  - Code, images, videos, gifs, text, etc.
- IDE or not an IDE?



ArcGIS Notebooks in Pro



Jupyter Notebook

Jupyter Lab

ArcGIS Notebooks in Online and Enterprise

Python Packages (libraries)

libraries (packages)? ready-to-use functions (solutions) to common programming problems. They have not been installed. We need to *install* them before using them.

Python: Coding language you will use to create analyses

+

Pandas/Geopandas: open-source package that allows you to work with (geospatial) data and operations, much like you would a table or spreadsheet

/GEOS, PROJ, and GDAL in C/C++ allowing you to manipulate geometries, using map projections and calculate areas, and read/write vector and raster data.

+

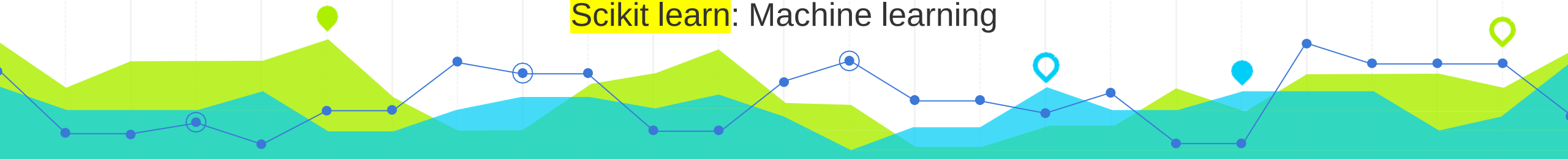
Matplotlib/seaborn/pydeck: Visualization

+

Pysal (Python Spatial Analysis Library): geospatial analysis

+

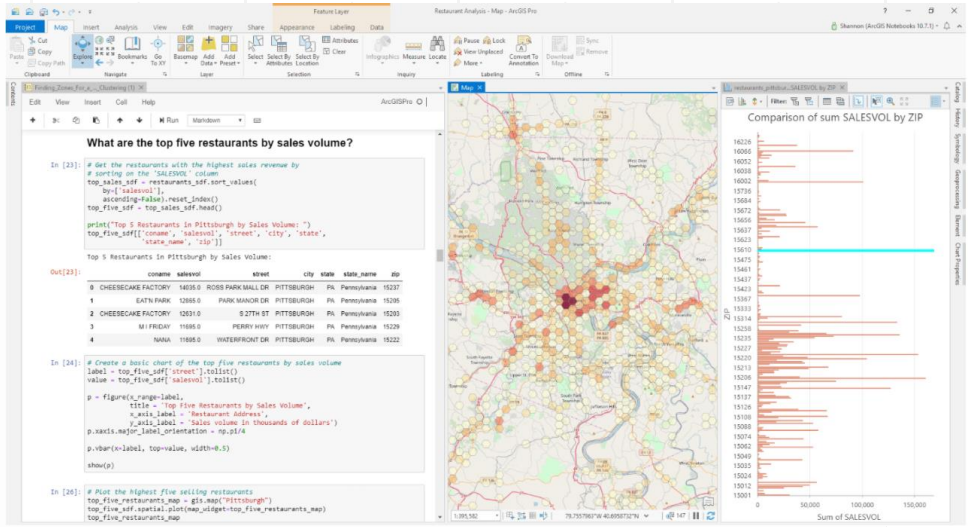
Scikit learn: Machine learning

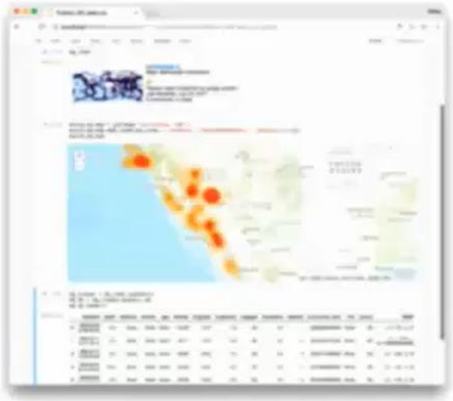


Jupyter

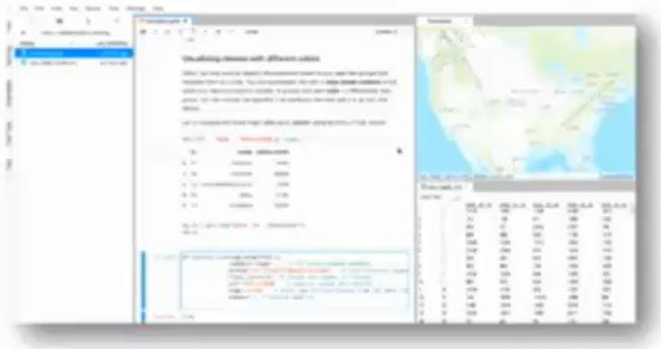
- Interactive, typically browser-based environment
- Allows you to combine code and markdown cells
  - Code, images, videos, gifs, text, etc.
- IDE or not an IDE?

  Integrated development environment (IDE)

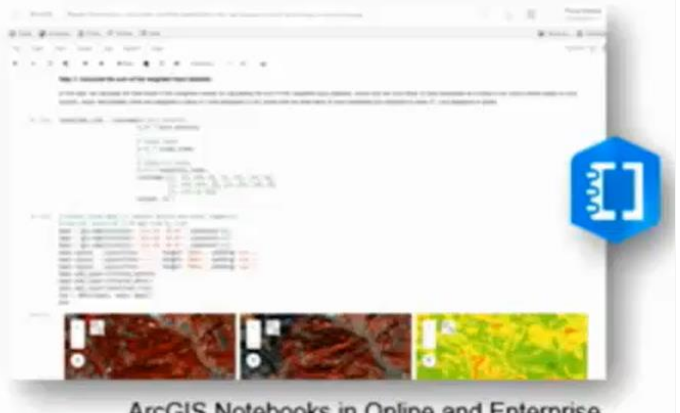




Jupyter Notebook



Jupyter Lab



ArcGIS Notebooks in Online and Enterprise

Example: https://jupyter.org/try-jupyter/notebooks/?path=notebooks/Intro.ipynb

Today's task (for both the lecture and the Lab): Programming Platform Set-up

- Install Anaconda + Python
- Install the Python virtual environment
- Exercise with Jupyter Notebook

Please follow the instruction (Week1_2_Setup).

Week1_inclass_exercise due: 11:59pm, Friday, Jan 24

Demo – jupyter notebook

For more functions regarding the Jupyter Notebook Markdown, please refer to:
https://www.youtube.com/watch?v=uVLzL5E-YBM

For basic Markdown syntax:

https://www.markdownguide.org/basic-syntax/
https://jupyter.org/try-jupyter/notebooks/?path=notebooks/Intro.ipynb

Questions?

Reference

- Batty, M. (2019). Urban analytics defined.
- Batty, M. (2017). The future journal.
- Donoho, D. (2017). 50 years of data science. *Journal of Computational and Graphical Statistics*, 26(4), 745-766.
- Organizers, Kang, W., Oshan, T., Wolf, L. J., Discussants, Boeing, G., ... & Xu, W. (2019). A roundtable discussion: Defining urban data science. *Environment and Planning B: Urban Analytics and City Science*, 46(9), 1756-1768.

Programming? How to?

- Experiment with small changes in existing code.
- Read and google errors closely and get familiar with online help forums.
- In this course, all the codebooks will be uploaded on Canvas before class. However, as a Python beginner, you should follow the instructor in the class demo and type in the code by yourself. Please DO NOT copy and paste the code directly.

- **Task**: filter spam emails

```
if conditions:
    it is a spam
else:
    it is not a spam
```

Spam emails

What are the ***conditions***?

- bad words
- ads (what are the ads)?
- suspicious senders?
- in title? In email body?

Not possible to write a rule to exhaust all cases!

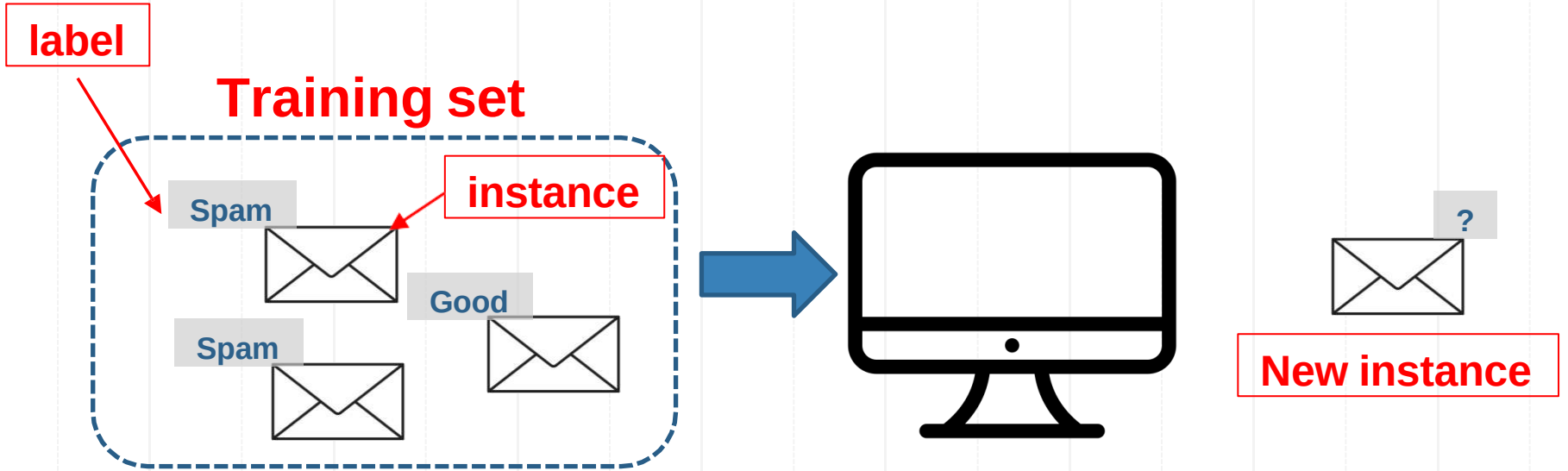
Nonspam emails

Machine Learning Types

1. Supervised Learning: machine is trained with human supervision with a “teacher”, (the training set is labeled)
2. Unsupervised Learning: machine is trained without human supervision without a “teacher”, (the training set is not labeled)
3. Semisupervised Learning
4. Reinforcement Learning

Supervised Learning

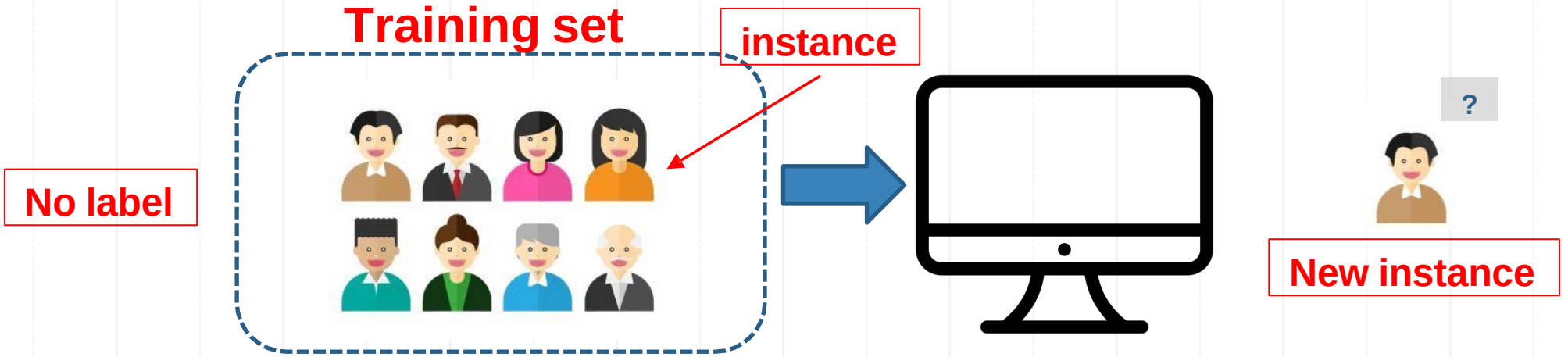
- Training



1. Supervised Learning: machine is trained with human supervision with a "teacher", (the training set is labeled)

Unsupervised Learning

- Training



2. Unsupervised Learning: machine is trained without human supervision without a “teacher”, (the training set is not labeled)