Introduction

Data Mining: Lecture 1

Course Instructor: Dr. Thomas Price

AI Academy

NC STATE UNIVERSITY

Welcome to AI Academy – Data Mining!

In this course you will learn to:

- Turn data into knowledge!
- Preprocess data to make it more useful.
- Train machine learning models to make predictions.
- Evaluate and compare models' performance.

Core Course Components

- **Program Specifics:** https://ai-academy.ncsu.edu/program-specifics/
- **Seminars**: Online video lessons at your own pace.
- **Workshops**: Live (Mon-Thu), collaborative practice sessions.
- **Assessments**: Individual weekly assignments. Not graded but will provide feedback.

Seminars

- We will have **two seminars** each week.
- Seminars consist of 3-5 videos, with short self-check questions after each.
- You should complete the seminars **before each workshop** (on your own time).
  - E.g. Complete Seminar 2 *before* Wednesday's workshop.

Note: This is Seminar 1, so there are no videos.

Workshops

- Workshop Schedule:
- Session 1
  - Monday: 6:00-8:00pm
  - Tuesday: 8:15-10:15pm
- Session 2
  - Wednesday: 6:00-8:00pm
  - Thursday: 8:15-10:15pm
- You can attend either session for a workshop.
- Mr. Martin will lead you through example exercises.
- You will complete practice written and programming exercises **working in small groups**.
- All exercises and assessments will be self-evaluated.

Assessment Problems

- We recommend that individual practice exercises are completed on your own.
- Discuss with mentors for questions.
- Turn in questions to be answered in the first seminar.

Each Week at a Glance

- **Monday**: Complete Seminar 1, attend Workshop 1.
- **Wednesday**: Complete Seminar 2, attend Workshop 2.
- **Friday**: Finish up any remaining workshop practice.
- **Sunday**: Finish individual assessment problems.

Collaboration & Support

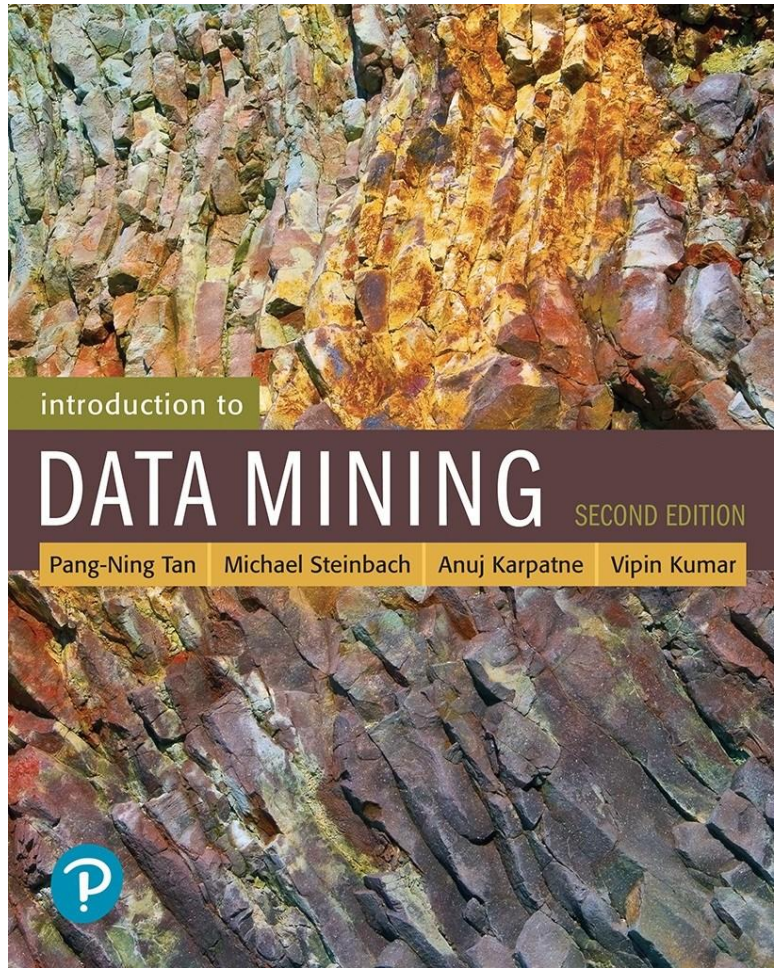
We are here to support you – please reach out for help!

- Post *public/conceptual* questions on Moodle forums.
- Email ced-aia-datamin-fall-2022-support@wolfware.ncsu.edu
  - Do not email instructors directly unless it is a private matter.
  - Please communicate private matters with Mr. Martin
- Post questions at the end of Seminars.
- Visit Office Hours.

Recommended Reading

Introduction to Data Mining
(Tan, et al.), 2nd Edition

- The weekly schedule lists required readings.
- You should complete these readings *before* the corresponding workshop.



Tour of Moodle

Questions?

The Age of Big Data

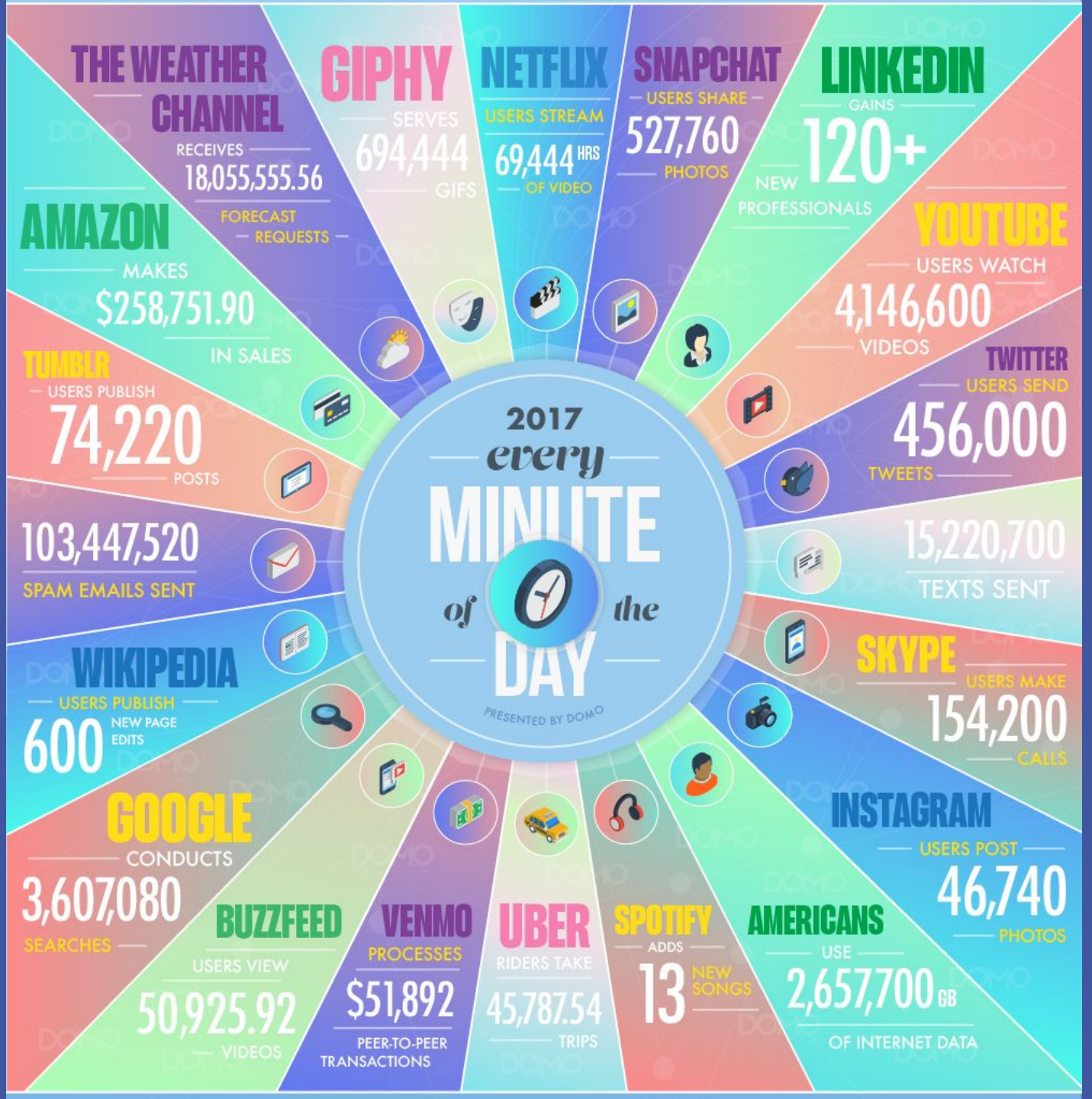
Every day, people create **2.5 quintillion (billion trillion) bytes** of data.

- Sensors, mobile devices, online transactions, social networks

69K hours of video/min

46K photos/min

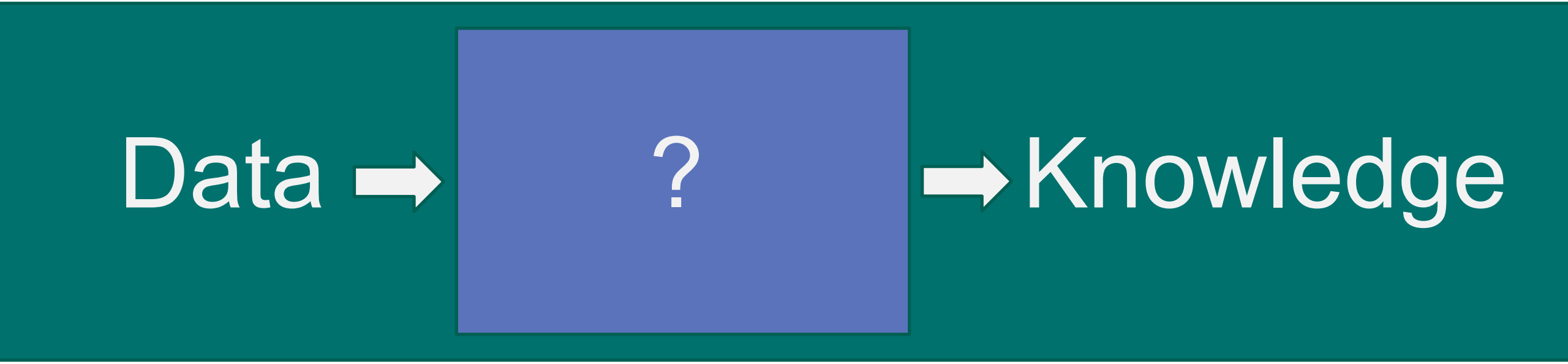
456K tweets/min



Data Never Sleeps 5.0, 2017

Data is not knowledge!

What is Data Mining?



Knowledge is actionable – it helps us (or computers) make decisions in the world!

What is Data Mining?

Knowledge is actionable – it helps us (or computers) make decisions in the world!

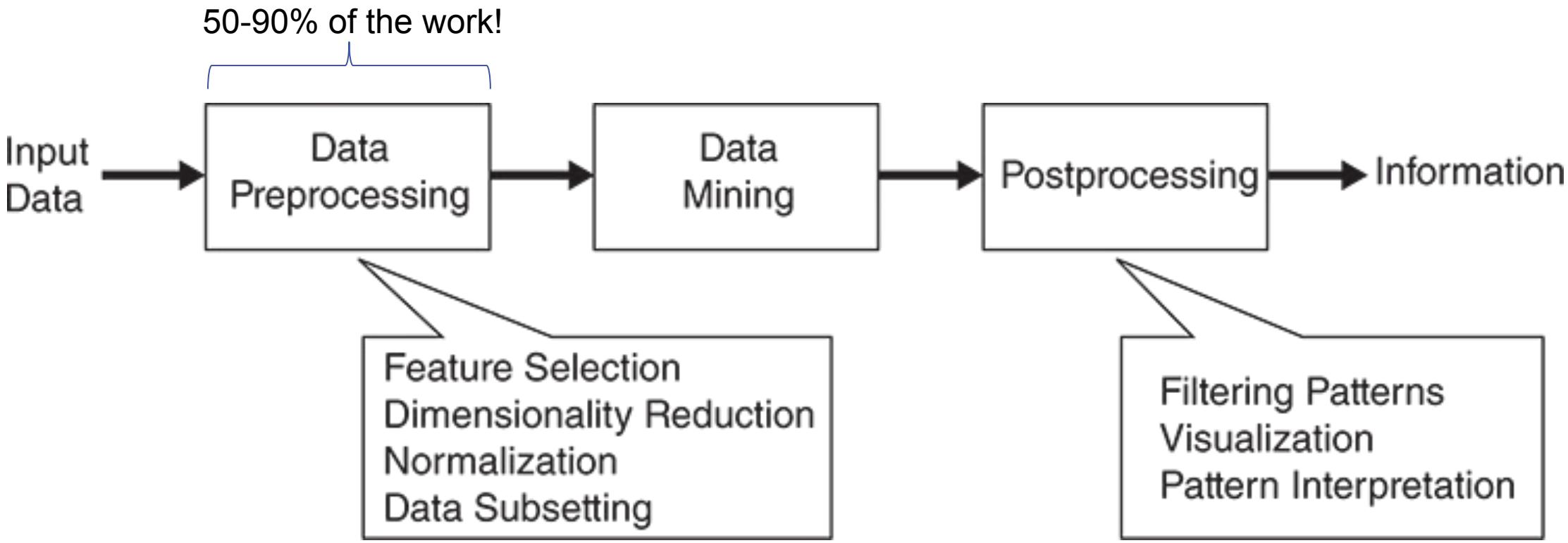
What is Data Mining?

- Design and Analysis of learning algorithms that:
  - **Automatically** or **semi-automatically**…
  - Discover useful **patterns or knowledge** (i.e. actionable, previously unknown)…
  - Using **large data repositories**.

What Data Mining Tasks do You Want to Do?

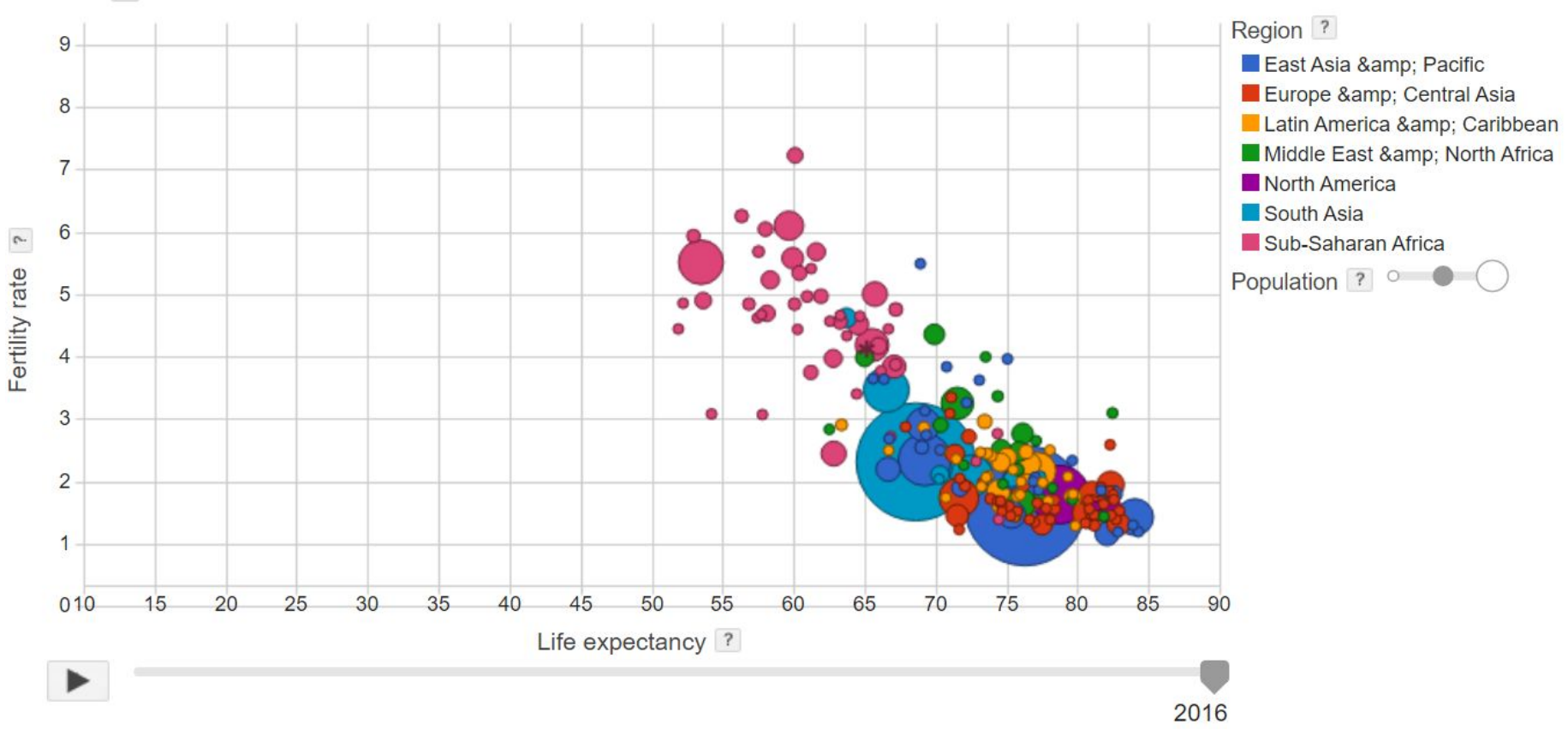
Think of 2-3 types of data mining tasks you are interested in conducting.

The Data Mining Pipeline



Dataset Example

Google Public Data Visualizations:
https://www.google.com/publicdata/directory



Small Group Exercise

- Explore the data visualization on your own and identify an interesting trend.
- Discuss with others. Did they see something you missed?
- Link to Data: go.ncsu.edu/google-data

Data Means People

- Data often reflects the experiences and impacts of real people.
- Data does not exist in a vacuum.
- Similarly, **what we do with that data can have real impact on people.**

Examples of Data Mining

AI Academy

Document Classification

Spam Filters

Welcome to New Media Installation: Art that Learns

Hi everyone,

Welcome to New Media Installation:Art that Learns

The class will start tomorrow.
Make sure you attend the first class, even if you are on the Wait List.
The classes are held in Doherty Hall C316, and will be Tue, Thu 01:30-4:20 PM.

By now, you should be subscribed to our course mailing list: 10615-announce@cs.cmu.edu.

Natural _LoseWeight SuperFood Endorsed by Oprah Winfrey, Free Trial 1 bottle, pay only $5.95 for shipping mfw rlk Spam | X

=== Natural WeightL0SS Solution ===

Vital Acai is a natural WeightL0SS product that Enables people to lose wieght and cleansing their bodies faster than most other products on the market.

Here are some of the benefits of Vital Acai that You might not be aware of. These benefits have helped people who have been using Vital Acai daily to Achieve goals and reach new heights in there dieting that they never thought they could.

* Rapid WeightL0SS

Spam or
Not Spam?

Stock Market Prediction

Many Many More…

- Text Mining
- Medical Data Analysis
- Social Networks
- Education
- …

What is Data Mining?

Data Mining Tasks

- **Supervised Learning**: Prediction
  - Learning patterns from **known pairs** (X-Y) to predict unknown or future value of variables (Y').
  - Classification, Regression
- **Unsupervised Learning**: Description
  - Find human-interpretable patterns that describe the data or patterns in it
  - Clustering, Dimensionality Reduction

Supervised Learning: Prediction

Feature Space $\mathcal{X}$

Label Space $\mathcal{Y}$

"Sports"
"News"
"Science"
...

Discrete Labels
Classification

Share Price
"$ 24.50"

Continuous Labels
Regression

Task: Given $X \in \mathcal{X}$, predict $Y \in \mathcal{Y}$.

Unsupervised Learning: Prediction

AKA: "Learning without a teacher"

Feature Space $\mathcal{X}$

Word distribution
(Probability of a word)

Task: Given $X \in \mathcal{X}$, learn $f(X)$.

Unsupervised Learning

Clustering: Group similar things (e.g. images)

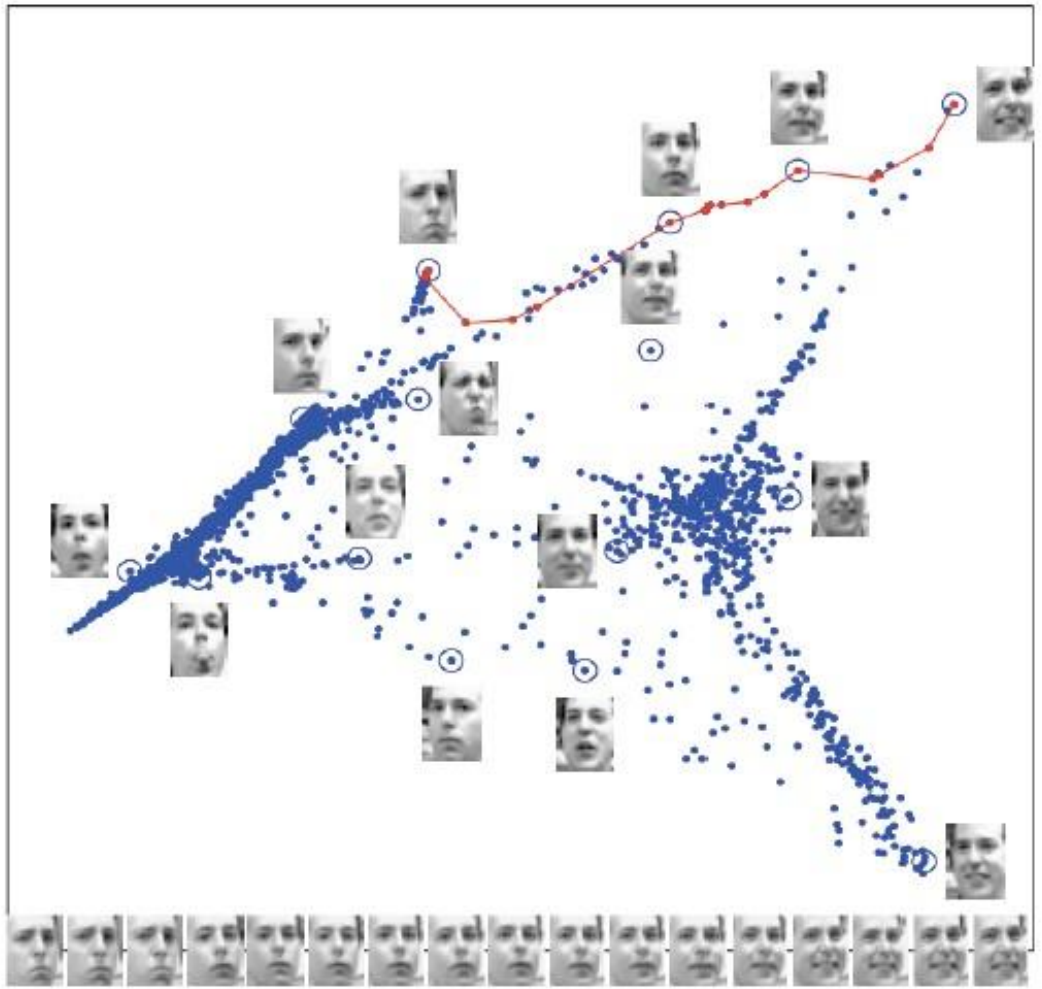
[Goldberger et. al]

Unsupervised Learning

Dimension Reduction [Saul & Roweis, 2003]

Images have thousands or millions of pixels.

Can we give each image a coordinate, such that similar images are near each other?

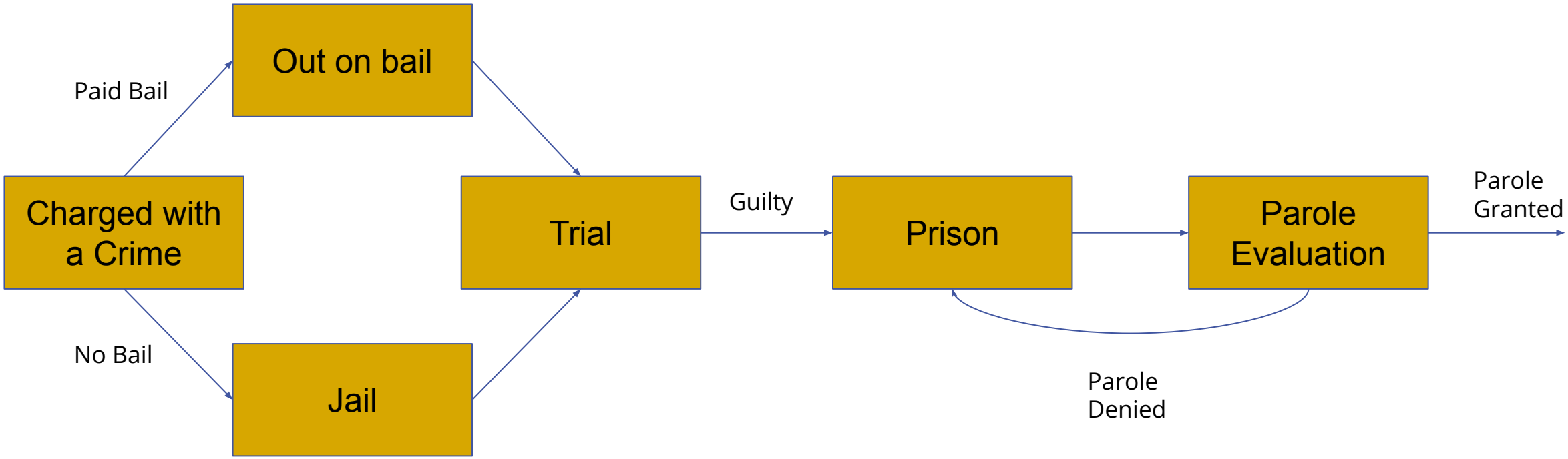


Think Back to your Data Mining Tasks

*Label them as **supervised, unsupervised,** or **other***

Data Ethics Case Study

Context: U.S. Criminal Justice System



How much bail?

Grant parole?

AI Academy

NC STATE

Bail Risk Prediction

Goal: Given a person's history, predict whether they are safe to release on bail.

The Data

ID	Sex	Location of Crime	Type of Crime	Bail?
1	Male	92nd St.	Robbery	Yes
2	Male	61st St.	Arson	No
3	Female	36th St.	Larceny	Yes
...	...	...	...	...
96	Female	55th St.	Assault	No
97	**Male**	**63rd St.**	**Larceny**	**????**

Group Activity - 2 Minutes

Would you want an algorithm helping to make this decision? Take 2 minutes to talk with your group about possible pros and cons.

https://youtu.be/eWDpOpnONL4?t=49

Machines outperform Judges

Human Decisions and Machine Predictions

Jon Kleinberg, Himabindu Lakkaraju, Jure Leskovec,
Jens Ludwig & Sendhil Mullainathan

Kleinberg is a Computer Scientist who trained an algorithm to make bail decisions.

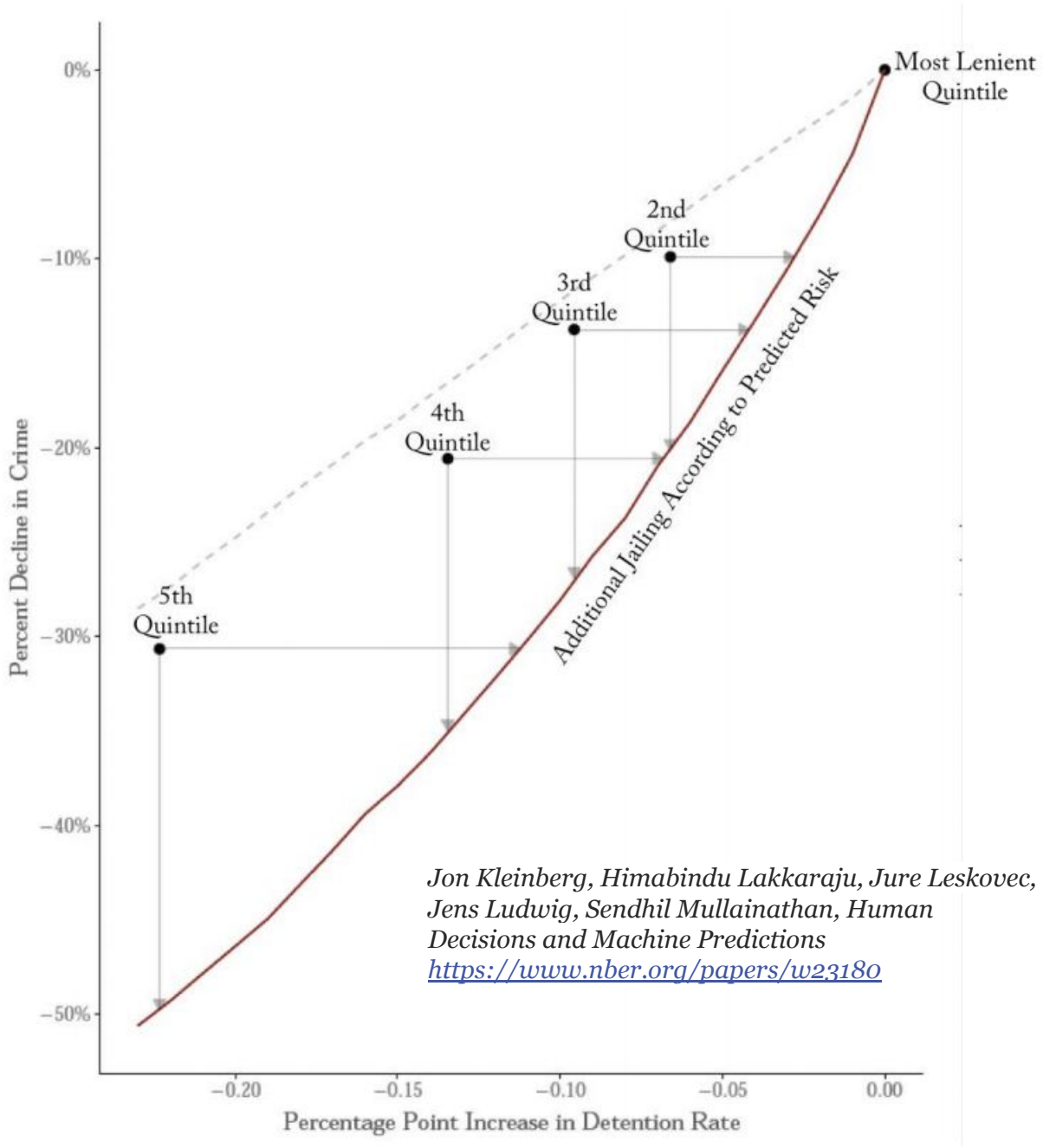
- He compared the algorithm to real judges' decisions in New York courts

"Our results suggest potentially large welfare gains:

- "a policy simulation shows **crime can be reduced by up to 24.8%** with **no change in jailing** rates,
- or **jail populations can be reduced by 42.0%** with **no increase in crime** rates."
- "Such gains can be had while also significantly reducing the **percentage of African-Americans and Hispanics in jail**"

Judges vs Algorithms

- Judges vary in how strict they are
- If we go from the most lenient group to 2nd most lenient, we see some reduction in crime, and also more people in jail
- But it's not much better than randomly jailing people (gray)
- An algorithm (red line) does much better



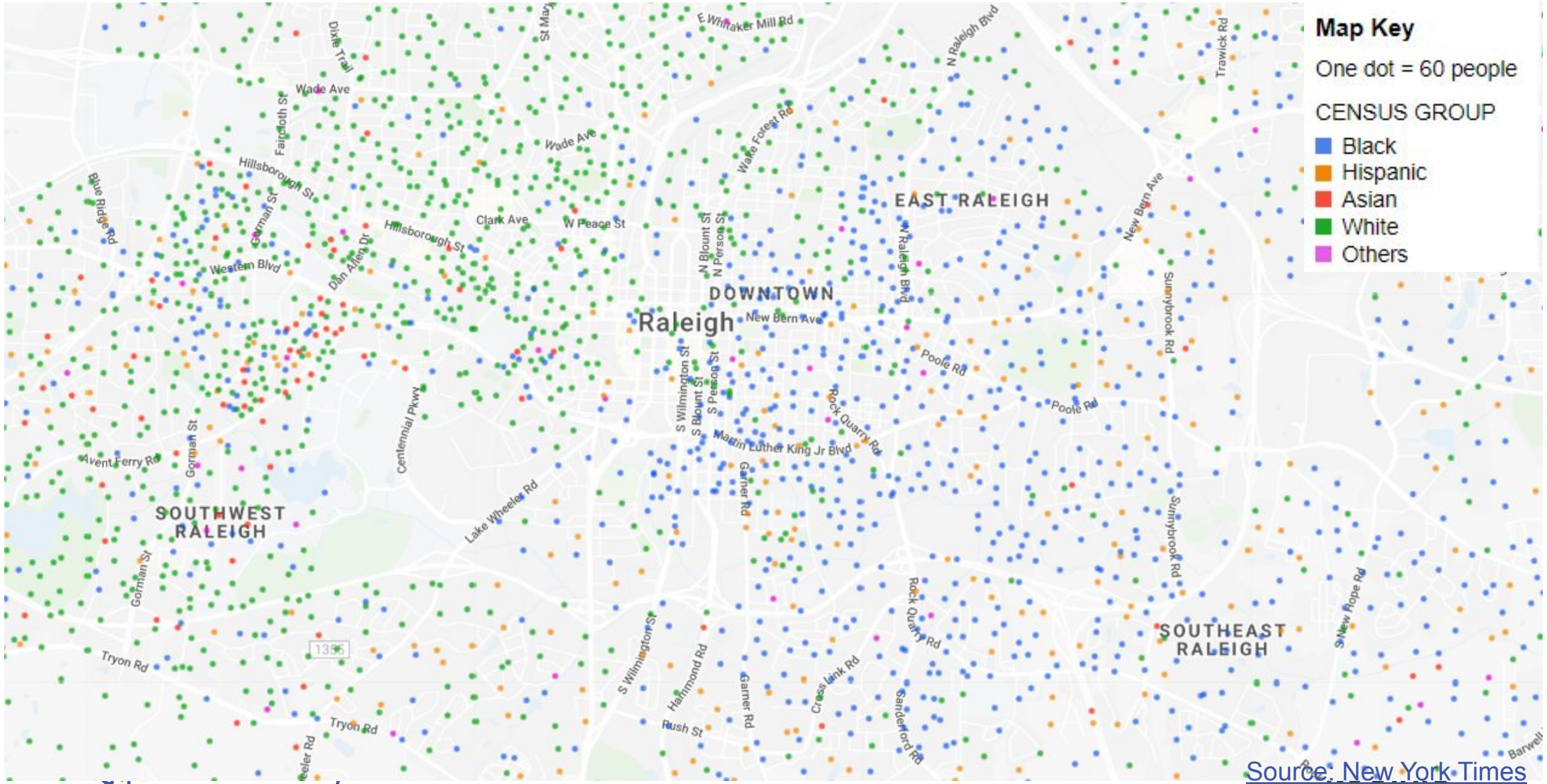
Jon Kleinberg, Himabindu Lakkaraju, Jure Leskovec, Jens Ludwig, Sendhil Mullainathan, Human Decisions and Machine Predictions
https://www.nber.org/papers/w23180

AI Academy

https://youtu.be/W6rC7EzBPdo

Can we remove demographic attributes?

ID	Sex	Location of Crime	Type of Crime	Bail?
1	Male	92nd St.	Robbery	Yes
2	Male	61st St.	Arson	No
3	Female	36th St.	Larceny	Yes
...	...	...	...	...
96	Female	55th St.	Assault	No



Source: New York Times

Is the algorithm based on race?

The *age* of secrecy and unfairness in recidivism prediction

Cynthia Rudin*, Caroline Wang† Beau Coker†

Later academic analysis disputes ProPublica's results

- "we show that COMPAS does not necessarily depend on race, contradicting ProPublica’s analysis"
- However, this does not mean the algorithm did not have disparate impact based on race.

Paired Activity - 2 Minutes

Discuss again: Would you want an algorithm helping to make this decision? Has your opinion changed?

Ethics of Data Mining

- Data represent real people, places and events.
- Machine learning is powerful - and therefore requires responsible use.
- Algorithms – and datasets – are capable of bias and often reflect the biases of their designers.

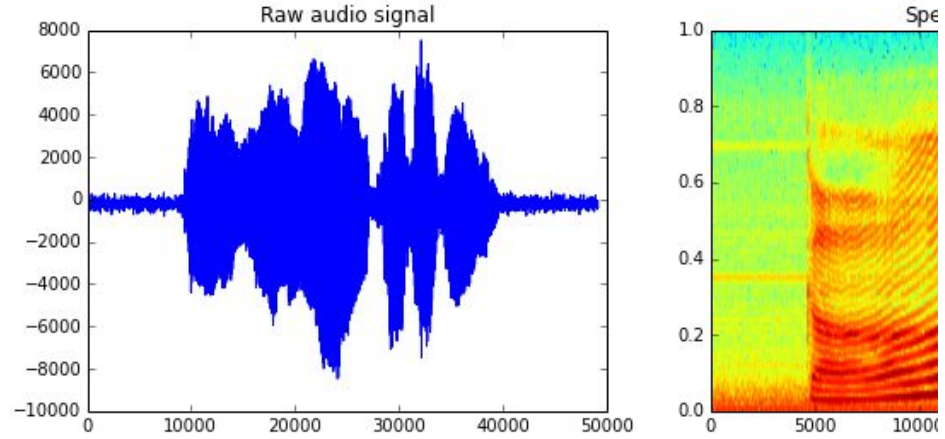
Workshops

Workshops at a Glance

You will work with **Jupyter Notebooks**

- An industry standard tool
- Visualizes output as your work
- Integrates code, results and documentation

```
In [2]: %matplotlib inline
from matplotlib import pyplot as plt
fig, (ax1, ax2) = plt.subplots(1, 2, figsize=(12, 4))
ax1.plot(x); ax1.set_title('Raw audio signal')
ax2.specgram(x); ax2.set_title('Spectrogram');
```



Workshops at a Glance

- **Example** Problems
  - Done together with Mr. Martin
- **Group** Exercises
  - PDF of written problems
  - Jupyter Notebook of programming problems
- **Assessment** Task

AI Academy

go.ncsu.edu/aiacademy

NC STATE UNIVERSITY

Stay Connected

Travis Martin

Assistant Instructor
tmmarti5@ncsu.edu