ECE 29595
Introduction to Data Science
Instructor: Milind Kulkarni
Fridays, 1:30–2:20
what is data?
lots of different definitions
humans have used data forever

- Ever since Thag Simmons first thought, “Last time, we only sent two people to hunt the smilodon. Maybe this time we should send three?”
why do we use data?

- Analyzing data helps us make decisions and take actions
what has changed?

• There’s a lot more data, and we’re trying to do more with it!
a parable of purdue professors

Prof. Milind Kulkarni (ECE) builds systems to make data analyses run faster

Prof. Bryan Pijanowski (Forestry) collects sound recordings from forests to study ecological change

Prof. Seungyoon Lee (Comm) analyzes social media behavior to understand how social networks help people process information

Prof. Jennifer Neville (CS) builds new machine learning tools to study graphs and networks

Prof. Stanley Chan (ECE and Stats) develops new algorithms for extracting data and signals from noisy images

Are they doing data science?
what is data science?

• Collecting data from a wide variety of sources and putting them into a consistent format?
• Making observations about patterns in data?
• Visualizing trends in data?
• Making predictions about what will happen in the future?
• Identifying similarity between data points?
• Developing new machine learning and data mining algorithms?
• Accelerating analysis algorithms?
data science is a lot of things

- using analyses to make predictions
- identifying patterns in data
- visualizing data
- building systems for data analysis
- collecting/organizing data
- interpreting data
- ethics
- analyzing data
- privacy concerns
- writing data analyses
data science is a lot of things

- using analyses to make predictions
- identifying patterns in data
- visualizing data
- building systems for data analysis
- collecting/organizing data
- interpreting data
- ethics
- analyzing data
- privacy concerns
- writing data analyses
This is one of three one-credit classes that cover data science topics:

- **PHIL 293** – Ethics for Data Sciences
- **ILS 295** – Introduction to Data management
syllabus break!
data analysis in “practice”

• Lets say we have a data set of applicants to Purdue

<table>
<thead>
<tr>
<th>Name</th>
<th>High school GPA</th>
<th>SAT Math</th>
<th>SAT R/W</th>
<th>Residence</th>
</tr>
</thead>
<tbody>
<tr>
<td>Jane Doe</td>
<td>4.7</td>
<td>760</td>
<td>700</td>
<td>Indiana</td>
</tr>
<tr>
<td>Purdue Pete</td>
<td>3.5</td>
<td>680</td>
<td>620</td>
<td>Indiana</td>
</tr>
<tr>
<td>B. O. Iler</td>
<td>3.0</td>
<td>800</td>
<td>650</td>
<td>Michigan</td>
</tr>
<tr>
<td>Engy Neer</td>
<td>4.2</td>
<td>750</td>
<td>590</td>
<td>N.C.</td>
</tr>
<tr>
<td>…</td>
<td>…</td>
<td>…</td>
<td>…</td>
<td>…</td>
</tr>
</tbody>
</table>

• What might we want to learn about them?
descriptive statistics

• Which students come from which states?
• What is the distribution of GPAs? SAT scores?
• Can build histograms — but how do we know how big to make the buckets?
• How do Purdue applicants compare to the national average?
  • Mean GPA of applicants: 3.6
• Is this high or low?
  • Can *sample* GPA of all high school students (randomly collect 1000 GPAs)
  • Mean GPA is 3.4
• Does this mean Purdue students have a higher GPA on average?
• Need more information!
  • Need to know about *variance* of the data (what is the spread of GPAs)
  • Need to know the *confidence interval* (what is the likely range of the true mean GPA?)
making predictions

• Can we predict how successful a particular applicant might be at Purdue?

• Idea: look at the application statistics of the current seniors and see if there is a relationship between their statistics and their Purdue GPA

• One way to find a relationship is using linear regression

• Might tell you something like: “a Purdue student’s GPA is predicted mostly by their high school GPA, and not very much by their SAT score”
classifying students

- What if I want to make admissions decisions more quickly
- Predict whether a student should be accepted or not
- Idea: compare each applicant to past applicants that were admitted and those that were rejected
  - See whether this applicant is more similar to other admitted applicants, or to rejected applicants
- This is a \textit{k-nearest neighbor} classifier
grouping students

• What if I just want to know if there are different groups of students

• Idea: see if students are clustered together in some way
  • Some students look more like “nearby” students than students that are “far away”
  • Questions: what features of students should you consider (e.g., maybe don’t consider something like hair color!)

• This is k-means clustering