CSCI 3022: Intro to Data Science

Lecture 1: Introduction

Rachel Cox

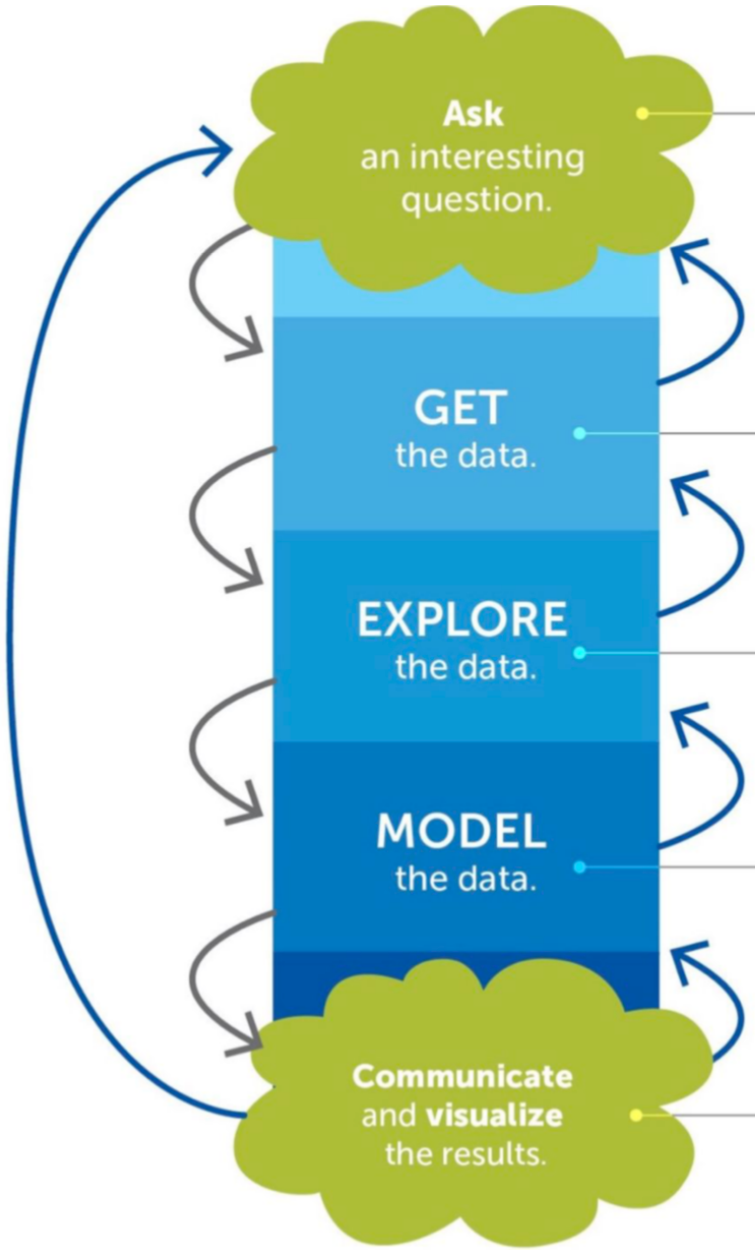
Department of Computer Science

Alan Dabney & Grady Klein

What is Data Science?

- Recovering insights & trends hiding within the data
- Using data to answer interesting questions
- Using data to understand the world around us
- Probability & Statistics (Math!)

What is Data Science?



What is the scientific **goal?**
What would you do if you had all the **data?**
What do you want to **predict** or **estimate?**

How were the data **sampled?**
Which data are **relevant?**
Are there **privacy** issues?

Plot the data.
Are there **anomalies?**
Are there **patterns?**

Build a model.
Fit the model.
Validate the model.

What did we **learn?**
Do the results make **sense?**
Can we tell a **story?**

Hypothesis

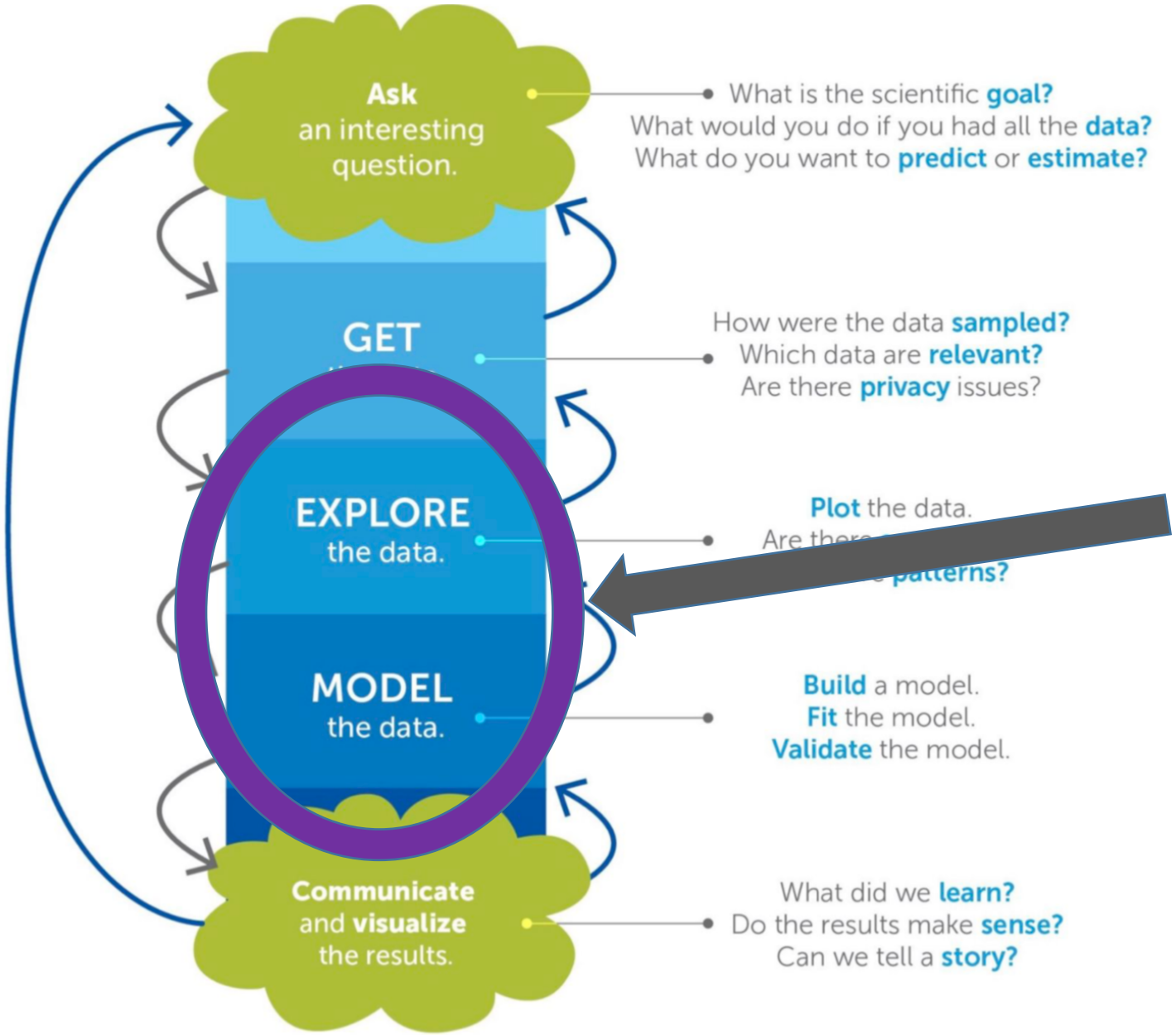
Observations

Analysis

Conclusions

Refinement (Do it all over again?)

What is Data Science?



We will focus largely on this part

- Exploration
  - Data Mining (**discover**)
- Modeling
  - Statistical analysis (**understand**)
  - Rudimentary machine learning (**predict**)

Course Topics

- **Exploratory data analysis**
  - Dig into your data, compute simple statistics, draw pictures and look for insight.

- **Cleaning, munging, and wrangling data**
  - Some data scientists report they spend up to 80% of their time cleaning data sets.

- **Probability theory and simulation**
  - Different probability distributions model different types of events.

- **Hypothesis testing and inferential statistics**
  - What is the data really telling us and how confident should we be in our conclusions?

- **Modeling and Classification**
  - Linear regression, multiple linear regression, logistic regression

What is Data Science?

Goal: Fluency in the theoretical and computational aspects of data analysis

At the end of this course, you'll be able to:

1. Clean, munge, and wrangle data in Python and perform Exploratory Data Analysis.
2. Draw insight from data by computing and interpreting classic summary statistics.
3. Know the ins-and-outs of probability and how to use it to solve real-world problems.
4. Construct and analyze simple models to make predictions and inferences about data.
5. Perform statistical tests to determine if your conclusions are real or due to chance.
6. Tell compelling stories about data using modern visualization and presentation tools.

Course Logistics – Grading

Workload:

- **(35%) Homework Assignments (~every 2 weeks, lowest 1 dropped)**
- **(20%) Midterm exam**
- **(20%) Final exam (cumulative)**
- **(10%) Practicum 1 (~around 1st midterm)**
- **(10%) Practicum 2 (~around final)**
- **(5%) Quizlets on Canvas**

- Let me know of any special needs accommodations in a timely manner.

- Read the syllabus! Many more important details can be found there.

Course Logistics – Late Submissions

Late days

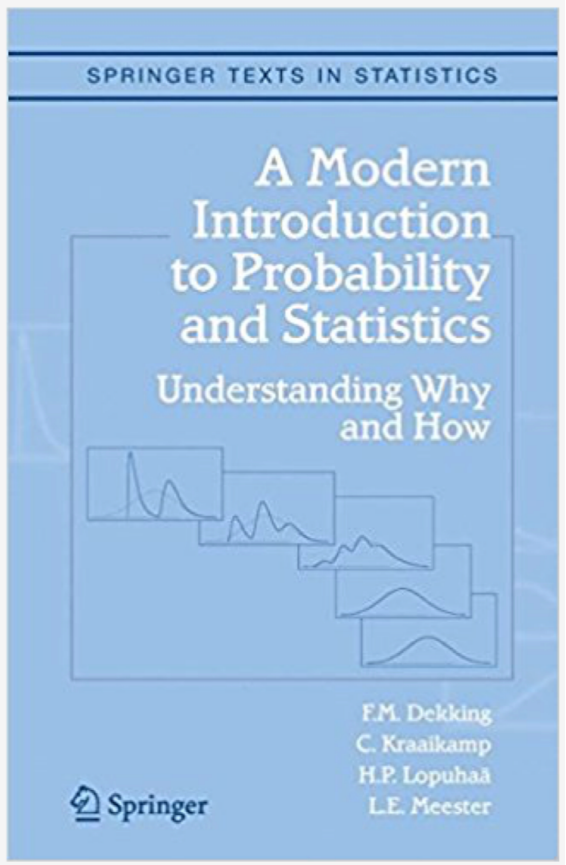
- (from the syllabus)

  You are allotted three late days that may be used for homework over the entire semester. Submitting an assignment between 1 second and exactly 24 hours late constitutes 1 late day; 24 hours and 1 second late is 2 late days. After you have expended your allotted late days late homework will not be accepted or graded. Your lowest homework score will be dropped.

- **TL/DR:**
  - 3 late days
  - Use em whenever you want for the HW
  - Can't use them for the Practicum or Quizlets
  - Canvas automatically flags late HW submissions. After a due date, we will record any late day usage.

Course Logistics – Book

<u>Textbook</u> – A Modern Introduction to Probability and Statistics (MIPS) by Dekking (et al.)



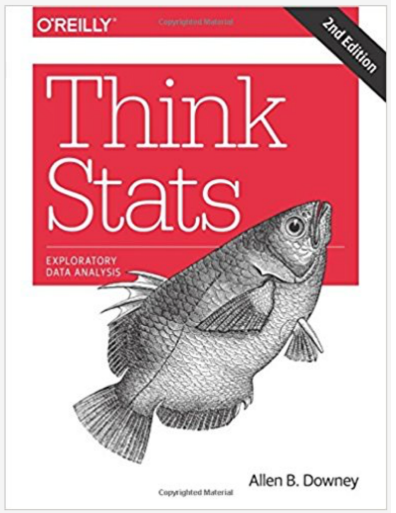
- International, old and PDF editions are ok, you may need to match sections.

Free PDF edition through CU (CU network, or VPN)

<u>And other reading:</u>

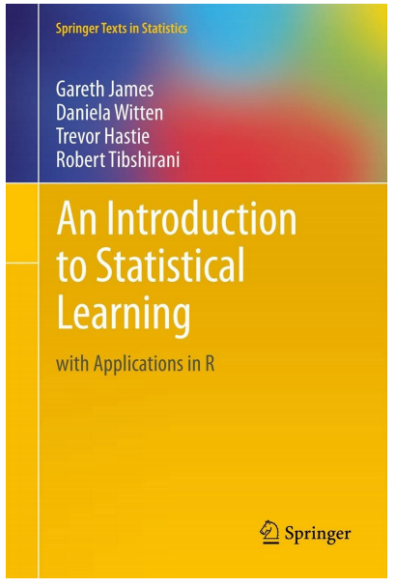
1) Think Stats (TS) by Downey

Free PDF here



2) Introduction to Statistical Learning (ISL)

PDF here



Course Logistics – Computing

- We will use **Python 3** and in particular **Numpy** and **Pandas**

- Lots of great data science libraries and decent plotting
- **We'll work exclusively in Jupyter Notebooks**
- Easiest way to get Jupyter is **Anaconda Python 3.6**; we strongly recommend you install local copy
- If not, you can use **Microsoft Azure** or **Google Colab** notebooks
- We will often work on problems in groups during class
- Bring a laptop, or have a friend with a laptop

Course Logistics – Computing

- Homework assignments will be done through Jupyter Notebooks

  Install Jupyter Notebook on your computer

  - Jupyter Notebook
  - Anaconda Python (includes Jupyter)

- Back your work up!
  - Github, Google Drive, SOMEthing
  - Make the repo **private** (collaboration policy)

Course Logistics – Academic Integrity

See the CU Academic Integrity Policy for more details. Here are some highlights.

- "Examples of cheating include: copying the work of another student during an examination or other academic exercise (includes computer programming)"

Bad

- "Examples of plagiarism include: [...] copying information from computer-based sources"

- For an assignment, Maciej and Felix work together to figure out how to implement the codes, but each works on their own computer and develops their own software.

Good

- For an assignment, Rhonda has a plan for how to implement an algorithm, but isn't sure how to manipulate a Python list in a particular way that she needs to. She searches the internet, finds a fix, and implements it in her code without copying it.

Course Logistics – Laptops in Class

"Results showed that students who used laptops in class spent considerable time multitasking and that the **laptop use posed a significant distraction to both users and fellow students**. Most importantly, the level of laptop use was negatively related to several measures of students learning, including self-reported understanding of course material and overall course performance."

https://www.sciencedirect.com/science/article/pii/S0360131506001436

Also: https://journals.sagepub.com/doi/pdf/10.1177/095679761667314

Also: https://www.sciencedirect.com/science/article/abs/pii/S0272775716303454

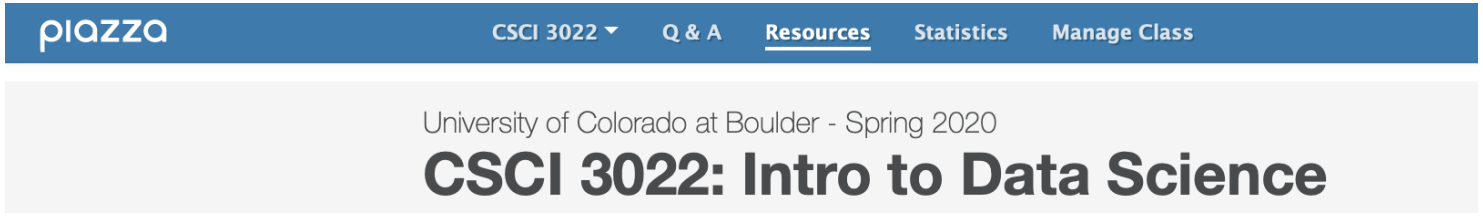
If you are going to use a laptop (aside from the Jupyter notebook times) …

1) **Sit in the back**
2) **Try to stay focused…**

Course Logistics – Platforms

1) Canvas – Assignments, solutions, Online Quizlets, Grades, Other Important Resources

2) Piazza– Class discussion forum Discuss work, but do not post solutions/vital code



3) Gradescope – Exam grading will be done here. Your paper exam will be returned to you as a digital PDF

Course Logistics – Before Next Class

- ❑ Read the Syllabus. You are responsible for knowing the information contained in this document.
- ❑ Make sure you can access the Canvas page.
- ❑ Check out the Piazza page.
- ❑ Install Anaconda (or other reliable Jupyter notebook method).
- ❑ Review and complete Numpy/Pandas tutorial
- ❑ Explore nb00 and nb01