University of Colorado **Boulder**

CSCI 3022, Intro to Data Science
Spring 2019
Tony Wong

Lecture 1: Introduction

Introduction

What is *Data Science*?

Introduction

What is *Data Science*?

→ Making the invisible visible

→ Recovering insights/trends hiding within the data

→ Using data to answer interesting questions

→ **Catch-all: using data to understand the world around us**

Introduction

What is *Data Science*?

→ Making the invisible visible

→ Recovering insights/trends hiding within the data

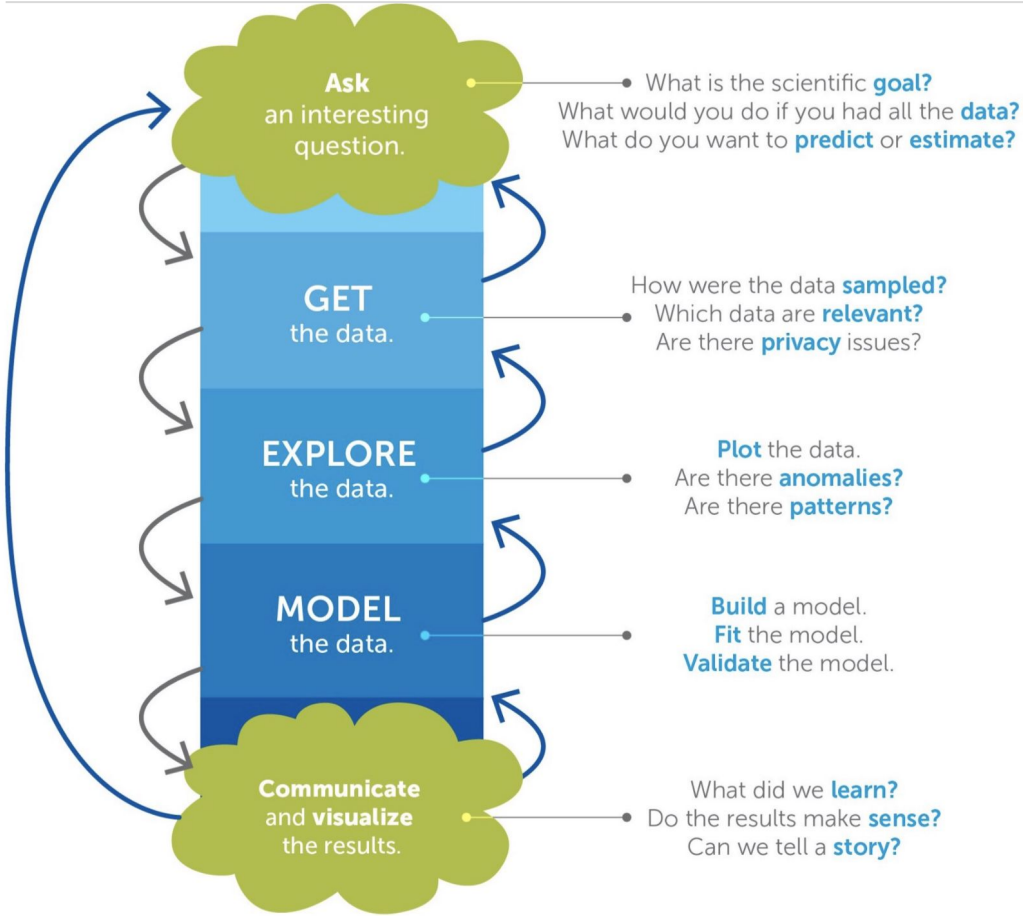
→ Using data to answer interesting questions

→ **Catch-all: using data to understand the world around us**

WARNING LABEL: we will do a lot of the "science" side of "data science"

→ Probability! Statistics! Math!

We put the “science” in “data science”



Hypothesis

⇩

Observations

⇩

Analysis

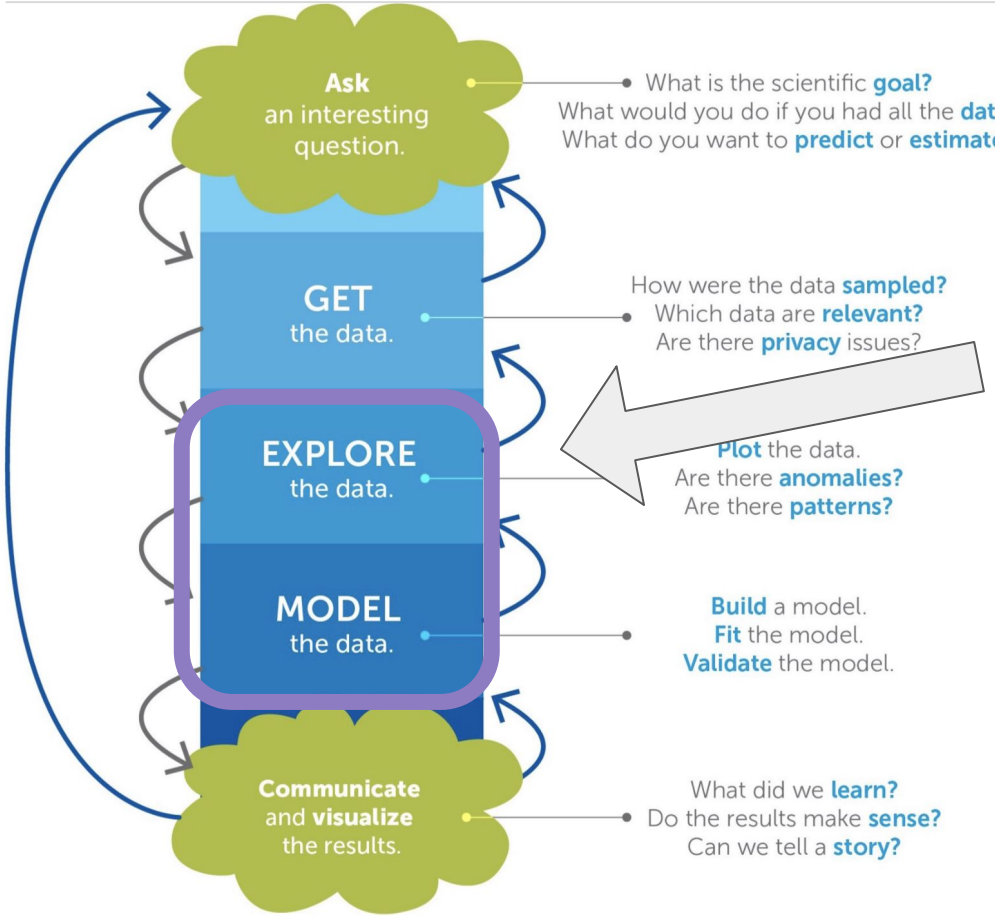
⇩

Conclusions

⇩

Refinement (Do it all over again?)

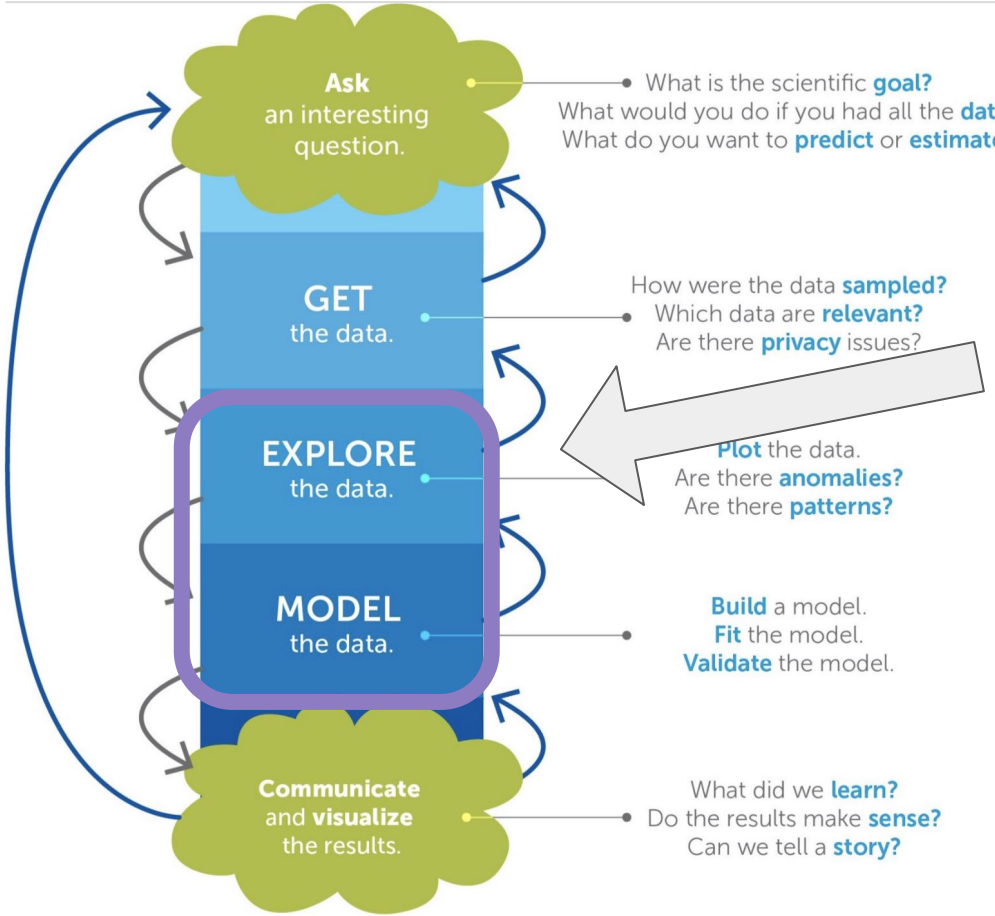
We put the “science” in “data science”



We will focus largely on **this part**

- Exploration
  - Data mining
- Modeling
  - Statistical analysis
  - Rudimentary machine learning

We put the “science” in “data science”



We will focus largely on **this part**

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

Foundations

realms	topics
probability	EDA, null models/hypotheses, Markov models
statistical inference	averages, regression models, max. likelihood estimates
optimization and calculus	model fitting, math shortcuts
linear algebra	any time we have a matrix (... or can make one!)
computer science	data structures, rapid estimation, simulation

Game plan

Week	Date	nb	txt	Topic	Slides	Hmwk
1	8.27			Course & Computing Introduction		
	8.29		16.1-3	EDA and Summary Statistics		
	1.26		2	Introduction to Probability		
2	9.03			**LABOR DAY - NO CLASS**		
	9.05		16.4-16.7	EDA and Data Visualization		hw1 posted
	9.07			**Data Wrangling**		
3	9.10		2,3	**How to Python**		
	9.12		6	Axioms and Theorems of Probability		
	9.14		3	**Stochastic Simulation**		hw1 due
4	9.17		4	Bayes' Rule and Intro to PDFs		hw2 posted
	9.19		4,5	Discrete RVs, PMFs, CMFs		
	9.21			Discrete RVs Strike Back		
5	9.24		5	**Return of the Discrete RVs**		
	9.26			Continuous RVs Awaken, PDFs, CDFs		
	9.28		5	**The Last Continuous RVs**		hw2 due
6	10.01		7	Expectation		hw3 posted
	10.03			Variance		
	10.05		5.5	**More Expectation & Variance**		
7	10.08			The Normal Distribution		
	10.10		14	**MIDTERM EXAM REVIEW**		
	10.10			The Central Limit Theorems		
	10.12			**MIDTERM EXAM (PM)**		hw3 due
8	10.15		23,24	**The Central Limit Theorem and You**		hw4 posted
	10.17		23,24	Inference and CI Intro		
	10.19			Two-Sample CIs		
9	10.22		25,26	**CI's in the Wild**		
	10.24		25,26	Hypothesis Testing Intro		
	10.26			p-Values		hw4 due
10	10.29		27	**Practical HT & p**		hw5 posted
	10.31			Small-sample HT		
	11.02			TBD		
11	11.05		18,23.3	Bootstrap Intro		
	11.07			**Bootstrap and Small n HT**		
	11.09		27	OLS/SLR Regression		hw5 due
12	11.12			Inference in SLR		hw6 posted
	11.14			*Hands on inference in SLR*		
	11.16			MLR		
13	11.19			**FALL BREAK - NO CLASS**		
	11.21			**FALL BREAK - NO CLASS**		
	11.23			**FALL BREAK - NO CLASS**		
14	11.26		ISL Ch3	Inference in MLR		practicum posted
	11.28		ISL Ch3	More MLR and ANOVA I		
	11.30		ISL Ch3	ANOVA II		hw6 due
15	12.03			**ANOVA + Inference in MLR**		
	12.05			Logistic Regr. & Classification		
	12.07			**Logistic Regr. & Classification**		
16	12.10			Solution Techniques and SGD		
	12.12			**FINAL EXAM REVIEW**		practicum due
X	12.XX			**FINAL EXAM **		

Exploratory data analysis

Probability theory & simulation

Hypothesis testing, inference

Modeling, classification, prediction

Game plan

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 [illegible] are real or due to chance
6) Tell compelling stories ab[illegible]tion and presentation tools

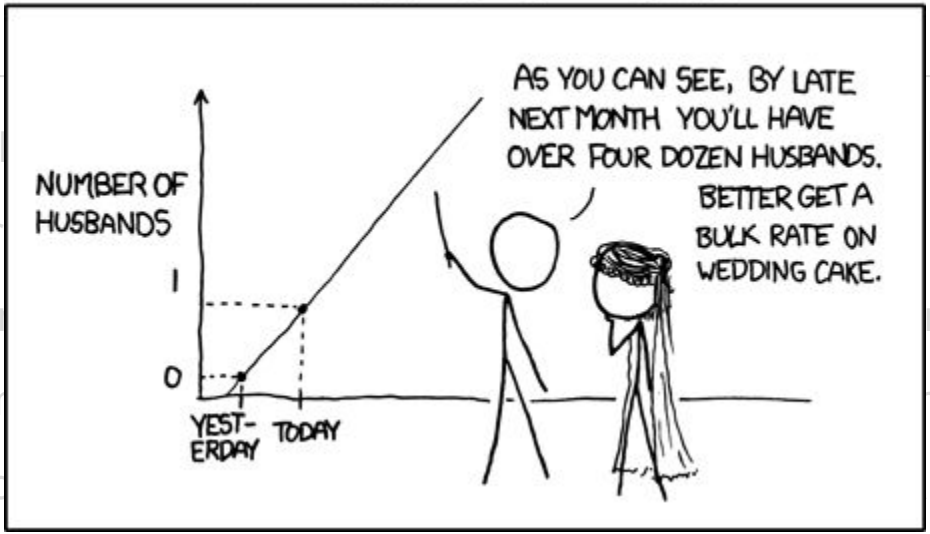
Game plan

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 t re real or due to chance
6) Tell compelling stories al ion and presentation tools

Game plan



Goal: Fluency in the t lysis

At the end of this cour

1) Clean, munge, a ry Data Analysis
2) Draw insight from mary statistics
3) Know the ins-and l-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

Computing

- We will use **Python 3** and in particular **Numpy** and **Pandas**
- Lot's of great data science libraries and decent plotting
- We'll exclusively work in Jupyter Notebooks
- Jupyter is ubiquitous DS collaboration and communication tool. Easiest way to get both is **Anaconda Python 3.6**
- We strongly recommend you install local copy
- If not, you can use **Microsoft Azure** or **Google Colab** notebooks
- Often work on problems in groups in class
- Bring a laptop or have a buddy with a laptop

Computing

- Homework assignments will be done through **Jupyter Notebooks** and submitted through **Canvas**

  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)

Laptops

"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 student learning, including self-reported understanding of course material and overall course performance."*

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

Also: http://journals.sagepub.com/doi/pdf/10.1177/0956797616677314

And: http://www.sciencedirect.com/science/article/pii/S0272775716303454 (... and others…)

Laptops

"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 student learning, including self-reported understanding of course material and overall course performance."*

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

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

Some logistics

Workload:

- (35%) Homework assignments (~every 2 weeks, lowest dropped, late days)
- (20%) Midterm exam
- (20%) Final exam (cumulative)
- (10%) Practicum 1 (~midterm)
- (10%) Practicum 2 (~final)
- (5%) Quizlets (Canvas)

≥ 55% exam average required to earn a C- or higher in the class

Let me know about any special needs in a timely manner

Read the syllabus! More details can be found there regarding course policies

Some logistics

Workload:

- (35%) Homework assignments (~ever
- (20%) Midterm exam
- (20%) Final exam (cumulative)
- (10%) Practicum 1 (~midterm)
- (10%) Practicum 2 (~final)
- (5%) Quizlets (Canvas)

≥ 55% exam average required to earn a (

Let me know about any special needs in a

Read the syllabus! More details can be found there regarding course policies

Quizlets:

- 5-15 minutes
- Build off of examples, concepts from class
- I'll announce them in class or on Canvas some days…
- … and they'll be due by noon on the next class meeting day
- I'll also add to the course calendar once assigned (<u>your responsibility to check it</u>)
- Like a pop quiz that you can do in your pajamas over the course of two days

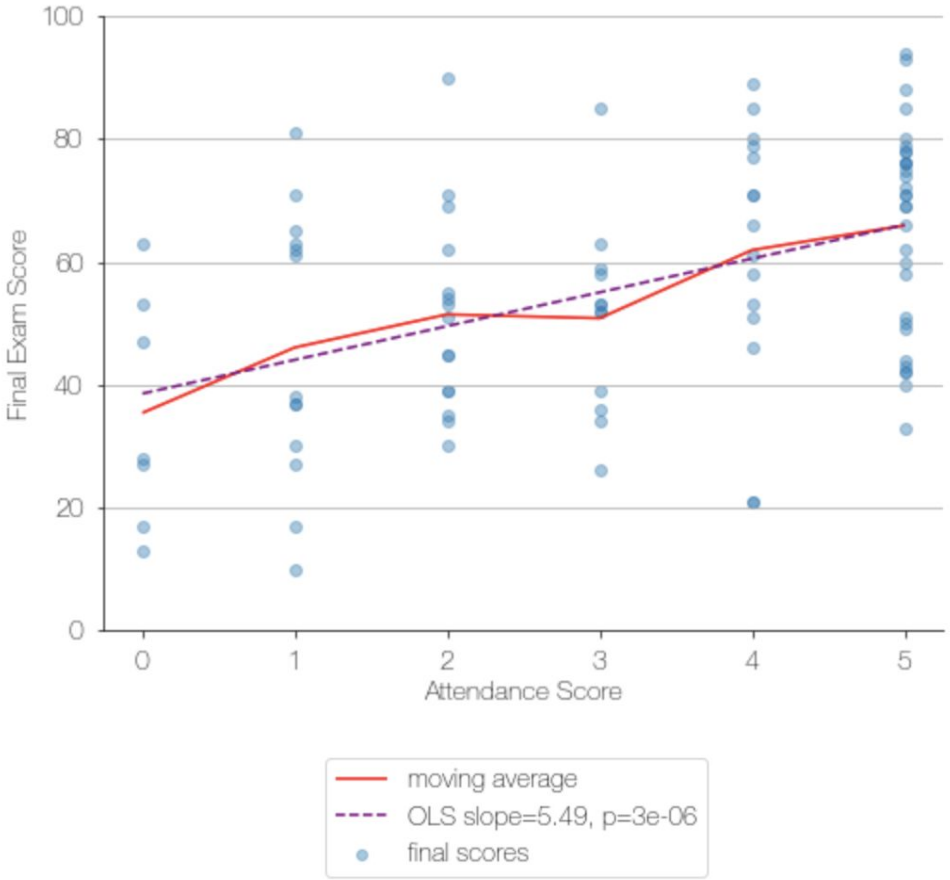
Late days

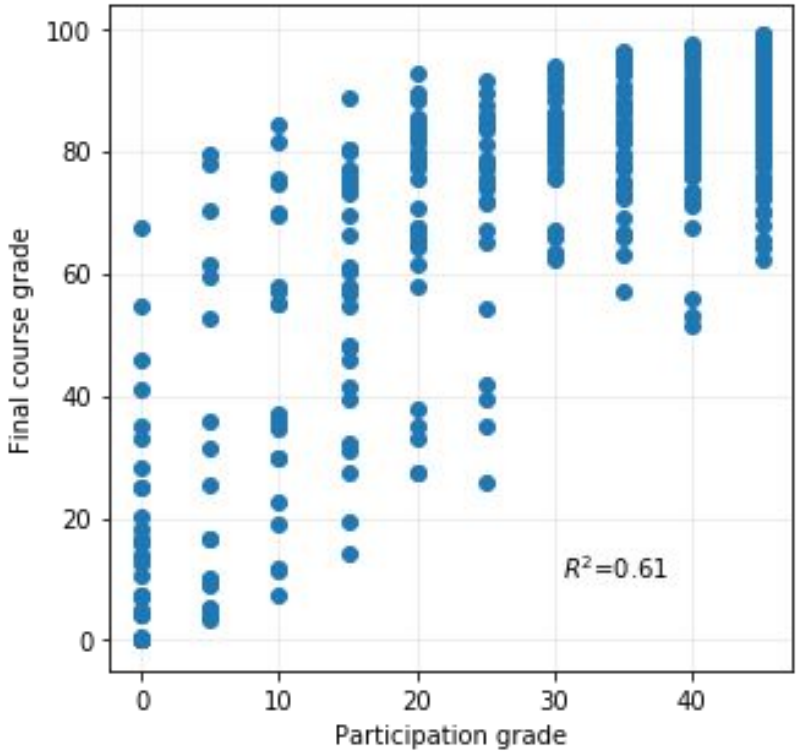
- (from the syllabus)

  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 attendance





Some logistics

We'll use **Canvas** to manage course content

https://canvas.colorado.edu/courses/24706

- Assignments, solutions, and other resources will be posted here

Piazza: Ask questions in Q & A forum (and answer other students' questions!)

- Discuss work, but **do not post solutions/vital code**
- Send **private messages** to faculty instead of email (keeps things organized)

https://piazza.com/colorado/spring2019/csci3022

Canvas landing page:

1) Link to lecture schedule, including **slides, in-class notebooks,** and **assignments**
2) Link to course syllabus
3) Lots of cheatsheets and tutorials
4) Office hours spreadsheet

Some logistics

We'll use **Canvas** to manage course-related communication and content

https://canvas.colorado.edu/courses/24706

- Assignments, solutions, and other resources will be posted here

Piazza: Ask questions in Q & A forum (and answer other students' questions!)

- Discuss work, but **do not post solutions/vital code**
- Send **private messages** to faculty instead of email (keeps things organized)

https://piazza.com/colorado/spring2019/csci3022

Canvas landing page:

1) Link to lecture schedule, including **slides**, **in-class notebooks**, and **assignments**
2) Link to course syllabus
3) Lots of cheatsheets and tutorials
4) Office hours spreadsheet

Remember when I said “You must be willing to struggle a little” ?

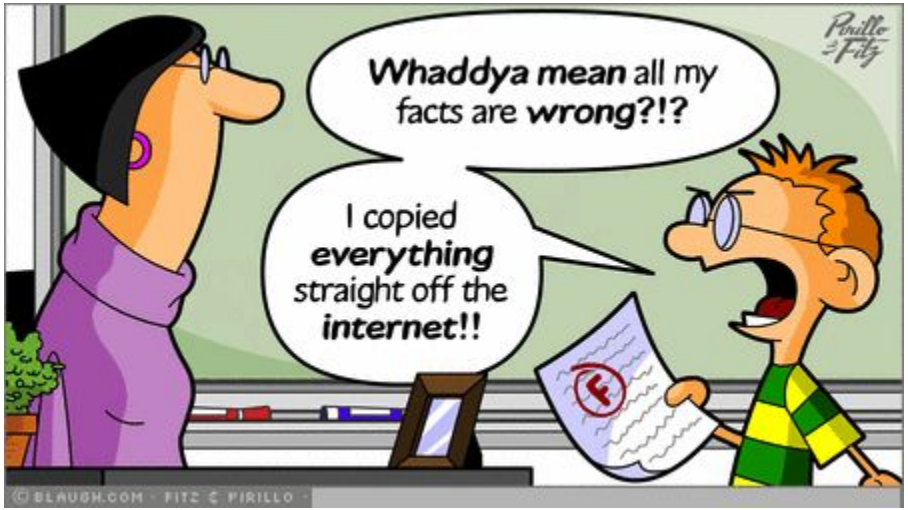
When you’re asking for help, be sure to explain…

- what you’re trying to do
- what you ***think*** should happen
- what you get instead (copy/pastes or screenshots work well)
- what all you have tried
    - if you haven’t tried anything, **try *some*thing** first

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)”
- “Examples of plagiarism include: […] copying information from computer-based sources”



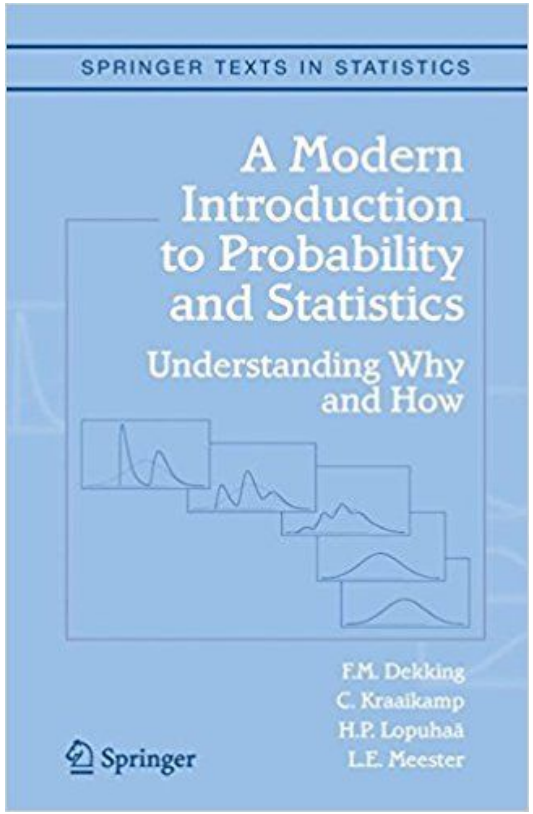
Academic Integrity

Example 1: For an assignment, Chris searches the internet for relevant codes and copy-pastes them into his Jupyter Notebook. He properly cites the source of the codes.

Example 2: 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.

Example 3: 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 Reading



A Modern Introduction to Probability and Statistics (MIPS) by Dekking (et al.)

- International, old and PDF editions are okay
- Just be sure to match sections

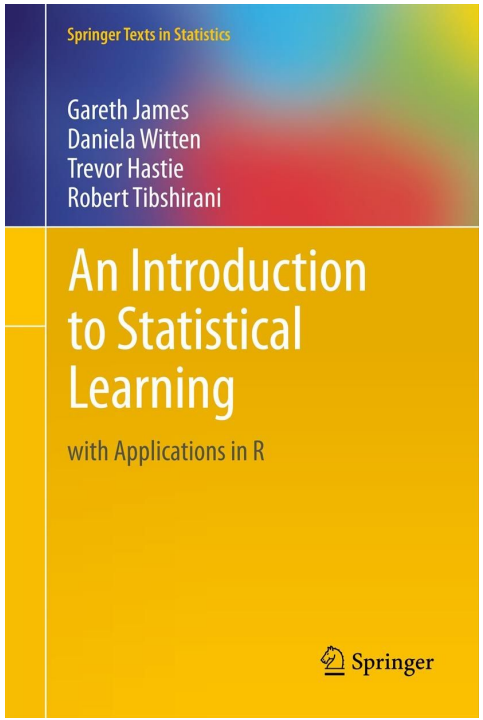
Free PDF edition through CU (CU network, or VPN):
→ https://www.springer.com/us/book/9781852338961

Additional reading will be linked to the course calendar as needed

Additional Reading

1) Think Stats by Downey (“TS”)

   PDF: here!



2) Introduction to Statistical Learning (“ISL”)

   PDF: here!

Hi! I'm Tony.

- Call me **Tony**. Or **Dr. Wong** if you're more comfortable with that. I don't care.
- Second year teaching in **CS**

- **Before this:** Postdoc at **Penn State**. And taught Earth Science
  Grad student in **Applied Math**. And taught Calc/Diff Eq
- **Research interests:**
  - Computational: Uncertainty quant., Markov chains, (Bayesian) model calibration
  - Applications: Storm surge/sea-level projections, coastal flood risk

- **Office:** ECOT 623
- **Office hours:** M/W/Th 11-12

Now…

Let’s get to work!

Get out your laptop, or -- better yet -- partner up with someone with a laptop and work together!

1) Numpy and Pandas tutorial
2) nb01 notebook

Before next class:

1) Make sure you can access the Canvas page and **read the syll**
2) Set up **some way** to back up your work
3) Install Anaconda (or other reliable Jupyter notebook method)
4) Review and complete Numpy/Pandas tutorial
5) Explore nb01