



BROWN  
Computer Science

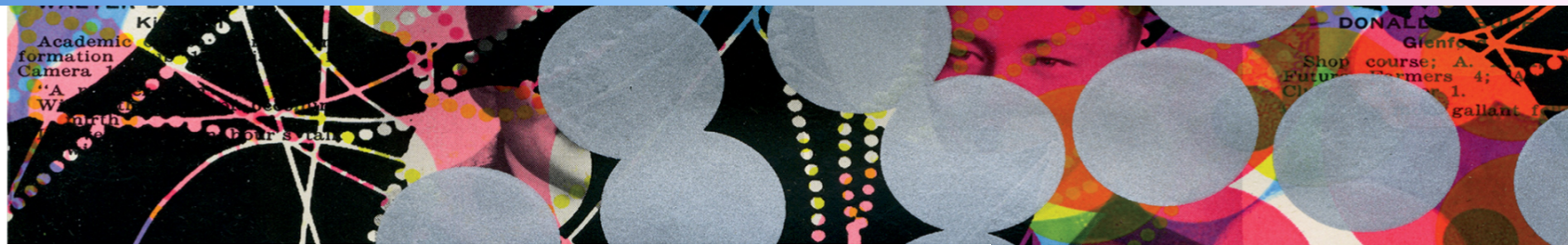
# CS1951A: Data Science

## Lecture 1: Introduction

Lorenzo De Stefani  
Spring 2022

# Overview

- Introduction: What is Data Science, and why should you care?
- Course organization
- Course contents

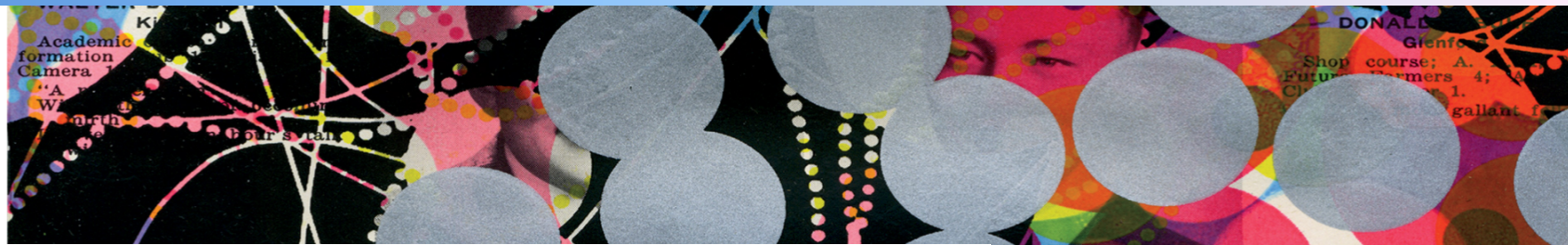


DATA

## Data Scientist: The Sexiest Job of the 21st Century

More than anything, what data scientists do is make discoveries while swimming in data. It's their preferred method of navigating the world around them. At ease in the digital realm, they are able to bring structure to large quantities of formless data and make analysis possible. They identify rich data sources, join them with other, potentially incomplete data sources, and clean the resulting set. In a competitive landscape where challenges keep changing and data never stop flowing, data scientists help decision makers shift from ad hoc analysis to an ongoing conversation with data.

Data scientists realize that they face technical limitations, but they don't allow that to bog down their search for novel solutions. As they make discoveries, they communicate what they've learned and suggest its implications for new business directions. Often they are creative in displaying information visually and making the patterns they find clear and compelling. They advise executives and product managers on the implications of the data for products, processes, and decisions.



DATA

## Data Scientist: The Sexiest Job of the 21st Century

More than anything, what data scientists do is make discoveries while swimming in data. It's their preferred method of navigating the world around them. At ease in the digital realm, they are able to bring structure to large quantities of formless data and make analysis possible. They identify rich data sources, join them with other, potentially incomplete data sources, and clean the resulting set. In a competitive landscape where challenges keep changing and data never stop flowing, data scientists help decision makers shift from ad hoc analysis to an ongoing conversation with data.


Data scientists realize that they face technical limitations, but they don't allow that to bog down their search for novel solutions. As they make discoveries, they communicate what they've learned and suggest its implications for new business directions. Often they are creative in displaying information visually and making the patterns they find clear and compelling. They advise executives and product managers on the implications of the data for products, processes, and decisions.



# What makes a “Data Scientist”?

## MODERN DATA SCIENTIST

Data Scientist, the sexiest job of 21st century requires a mixture of multidisciplinary skills ranging from an intersection of mathematics, statistics, computer science, communication and business. Finding a data scientist is hard. Finding people who understand who a data scientist is, is equally hard. So here is a little cheat sheet on who the modern data scientist really is.



- MATH & STATISTICS**
  - ☆ Machine learning
  - ☆ Statistical modeling
  - ☆ Experiment design
  - ☆ Bayesian inference
  - ☆ Supervised learning: decision trees, random forests, logistic regression
  - ☆ Unsupervised learning: clustering, dimensionality reduction
  - ☆ Optimization: gradient descent and variants
- DOMAIN KNOWLEDGE & SOFT SKILLS**
  - ☆ Passionate about the business
  - ☆ Curious about data
  - ☆ Influence without authority
  - ☆ Hacker mindset
  - ☆ Problem solver
  - ☆ Strategic, proactive, creative, innovative and collaborative
- PROGRAMMING & DATABASE**
  - ☆ Computer science fundamentals
  - ☆ Scripting language e.g. Python
  - ☆ Statistical computing package e.g. R
  - ☆ Databases SQL and NoSQL
  - ☆ Relational algebra
  - ☆ Parallel databases and parallel query processing
  - ☆ MapReduce concepts
  - ☆ Hadoop and Hive/Pig
  - ☆ Custom reducers
  - ☆ Experience with xaaS like AWS
- COMMUNICATION & VISUALIZATION**
  - ☆ Able to engage with senior management
  - ☆ Story telling skills
  - ☆ Translate data-driven insights into decisions and actions
  - ☆ Visual art design
  - ☆ R packages like ggplot or lattice
  - ☆ Knowledge of any of visualization tools e.g. Flare, D3.js, Tableau

# Learning the language of data



# Learning the language of data

- We will see many useful tools!
- Data science is not just a composition of techniques
- It **transcends** its tools
- More like a “thinking approach”

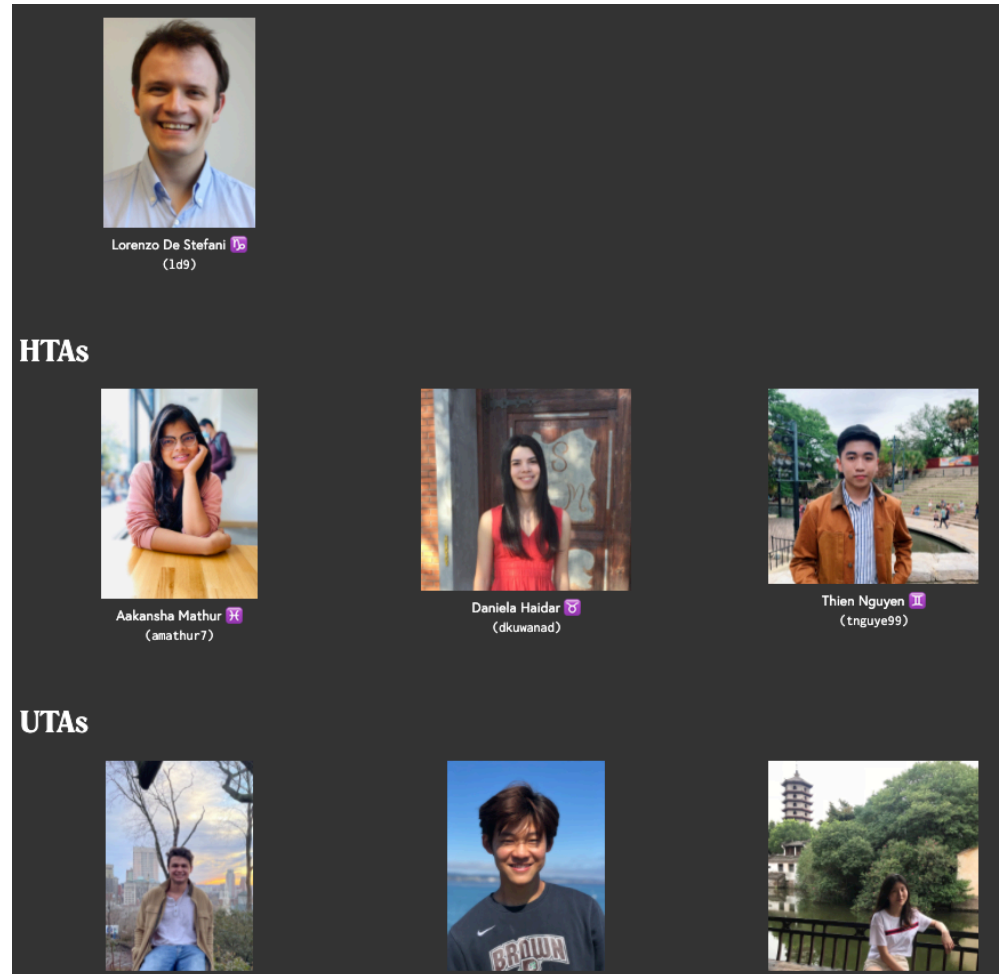


# Overview

- What is data science
- **Course organization**
- Cool facts about data science

# Couse Staff

- Instructor: Lorenzo De Stefani
- HTAs:
  - Aakansha Mathur
  - Daniela Haidar
  - Thien Nguyen
- UTAs:
  - Austin De Stefanis
  - Benjamin Shih
  - Ella Liang
  - Emily Ye
  - James (Yicheng) Shi
  - Kenya Kimata
  - Livia Gimenes
  - Dharam Madnani
  - Micah Buring
  - Nadav Druker
  - Nange Li
  - Robert Scheidegger
  - Shirley Loayza Sanchez
  - Sunny Li
  - Joanna Tasmin (also STA)
- STAs
  - Aanchal Sheth



# Course website

- <https://cs1951a-s21-brown.github.io/index.html>
- <https://piazza.com/class/kjqpfnm6w6b4f7>

cs1951a [home](#) [resources](#) [lectures](#) [assignments](#) [labs](#) [staff](#) [hours](#) [final projects](#) [archive](#)

## data science

Data is the core of all domains from material science to healthcare. Mastering big data requires a set of skills spanning a variety of disciplines, from distributed systems to statistics to machine learning. This course will provide an overview of the wide area of data science, with a particular focus on the tools required to store, clean, manipulate, visualize, model, and ultimately extract information from large amounts of data.

Our theme for this semester is Pandas (the animal) to honor the Python Data Analysis Library (aka pandas)!

Topics include:

- Database Design and SQL
- Web Scraping & Data Cleaning
- Hypothesis Testing
- Machine Learning
- Mapreduce
- Differential Privacy
- Correlation vs Causation

## final project

Throughout the entire course you will be working on a data science project which seeks to answer an interesting and important real-world question. You will be collecting your own data, cleaning it, modeling it, visualizing it, and finally presenting your results in a poster session at the end of the course. You will work in groups of four, and will be assigned a mentor TA to help you through the process.

Additionally, your project can be used as a capstone with just a few extra requirements, fully integrating what you will have learned in the course, and building a fully-functional data science application.

Check out the Final Project tab to learn more!

## Prerequisites

## overview

**Instructor**  
Lorenzo De Stefani

**Instructor Office Hours**  
Fridays 4:30-6:30 pm  
[Schedule Zoom Meeting](#)

**HTA Mailing List**  
[cs1951aheadtas@lists.brown.edu](mailto:cs1951aheadtas@lists.brown.edu)

**Location & Time**  
Tuesdays & Thursdays 9:00 - 10:20am  
[Lecture Zoom Link](#)

**Course Materials**  
[Syllabus](#)  
[Collaboration Policy](#)  
[Piazza](#)  
[Anonymous Feedback Form](#)  
[Capstone Declaration Form](#)  
[Blocklist Form](#)  
[Extension Requests](#)

## grading

Below is the grading scheme for the course:

- Labs (10%)
- Assignments (50%)
- Final Project (40%)

# Grading

Three main components:

- 50%: Assignments
- 10%: Labs
- 40%: Project

# Lectures and recordings

- All lectures streamed live on Panopto
  - Live attendance is strongly encouraged
  - Join class meeting a couple of minutes in advance
  - Ask questions using the chat
- Recording of the class lectures will be available within 24 hrs of live classes on Panopto
  - We will have a pinned post on Ed
- Lecture slides will be uploaded before class
- Jupyter notebooks for code-along available before class

# Late days policy

- Assignments are due at 11:59 pm on the listed due date
- 3 late days total you can use at most 2 for any assignment
- Additional late days request must be accompanied by a Deans Note and can only be approved by Lorenzo
- SEAS accommodation: reach out to Lorenzo ASAP
- Assignments submitted past the deadline after late days will not be graded
- No late days for Final Project deliverables
- Timely communication is key!

# Collaboration policy

- Collaboration on assignments is allowed and encouraged 😊
- But! Your submission must be your own!
- Write your own solutions and your own code
  - Copy paste = BAD 😞
  - Do not search the web for solutions

# Please review sign and submit the syllabus



# Waitlist

- The class is currently full
- We have some little wiggle room with the cap
- Send requests via email and on CAB so we can keep track of them
- We will give priority fairly
  - Seniors > Juniors > Sophomores
  - Priority to students matching all prerequisites

# Capstone

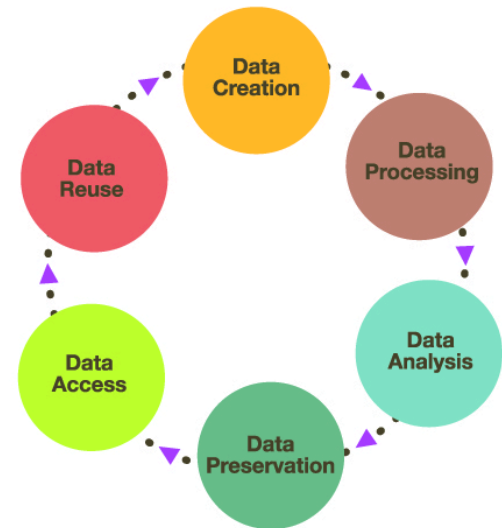
- Senior students can take CS1951A as a capstone
  - Please fill in the capstone form request on the website
- You will extend your project to include an interactive component or an extended analysis
  - If one person in a group plans to use the project for a capstone, the entire group will be held to the capstone standard
- All graduate students must complete the project at the capstone level

# Overview

- Intro: what is data science?
- Course organization
- **Course contents**

# What we are going to talk about

- Database design and SQL
- Data Cleaning
- MapReduce
- Hypothesis testing
- Machine Learning
- Data Visualization
- Crowdsourcing
- Causality/vs correlation
- Ethics of Data Analysis



# Handling and understanding data



## BIG DATA

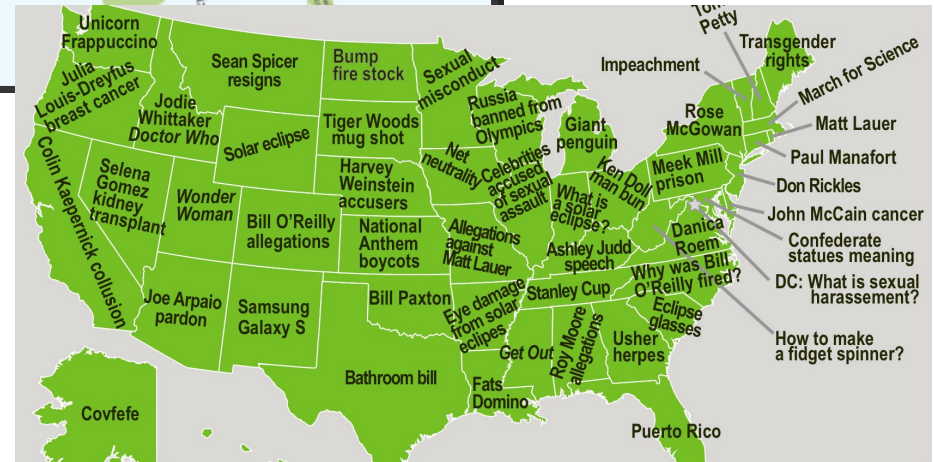
- Easy to acquire
- Hard to manage
- Consistency/cleanliness issues
- We need automatic ways to extract information
- Time sensitive
- Efficiency challenges
- Can provide lots of reliable insight!

## SMALL DATA

- Hard to acquire
  - “Precious data” (e.g. clinical trials, interviews, etc.)
- Easy to manage and review
- Long life cycle
- How can we make the most of it?
- Can we extract insights which are guaranteed to generalize to the larger world?



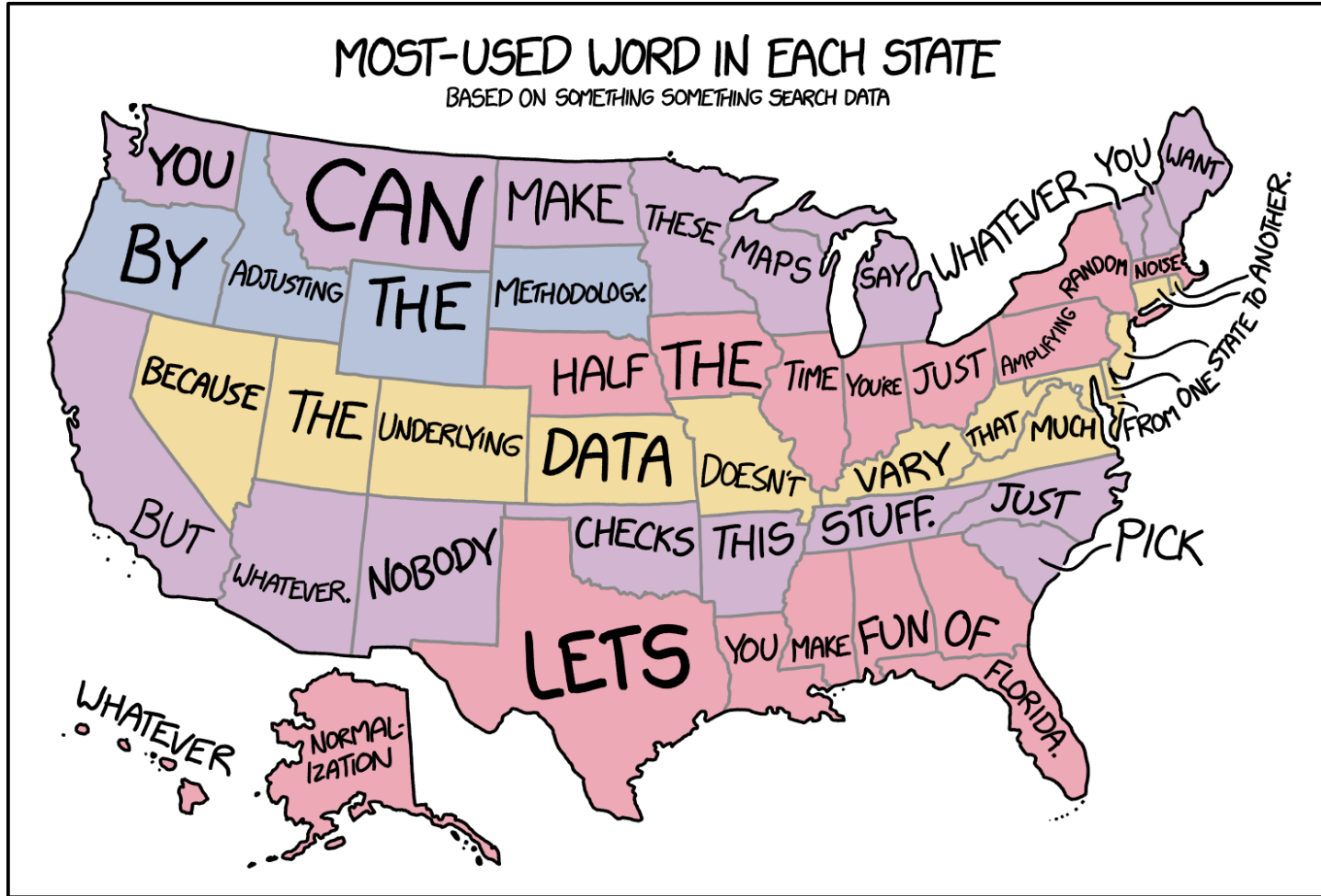
# Data "Science"?



<https://www.dailydot.com/unclick/state-googled-2017>

<http://nerdgeeks.co/us-state-words-map>

# Data "Science"?

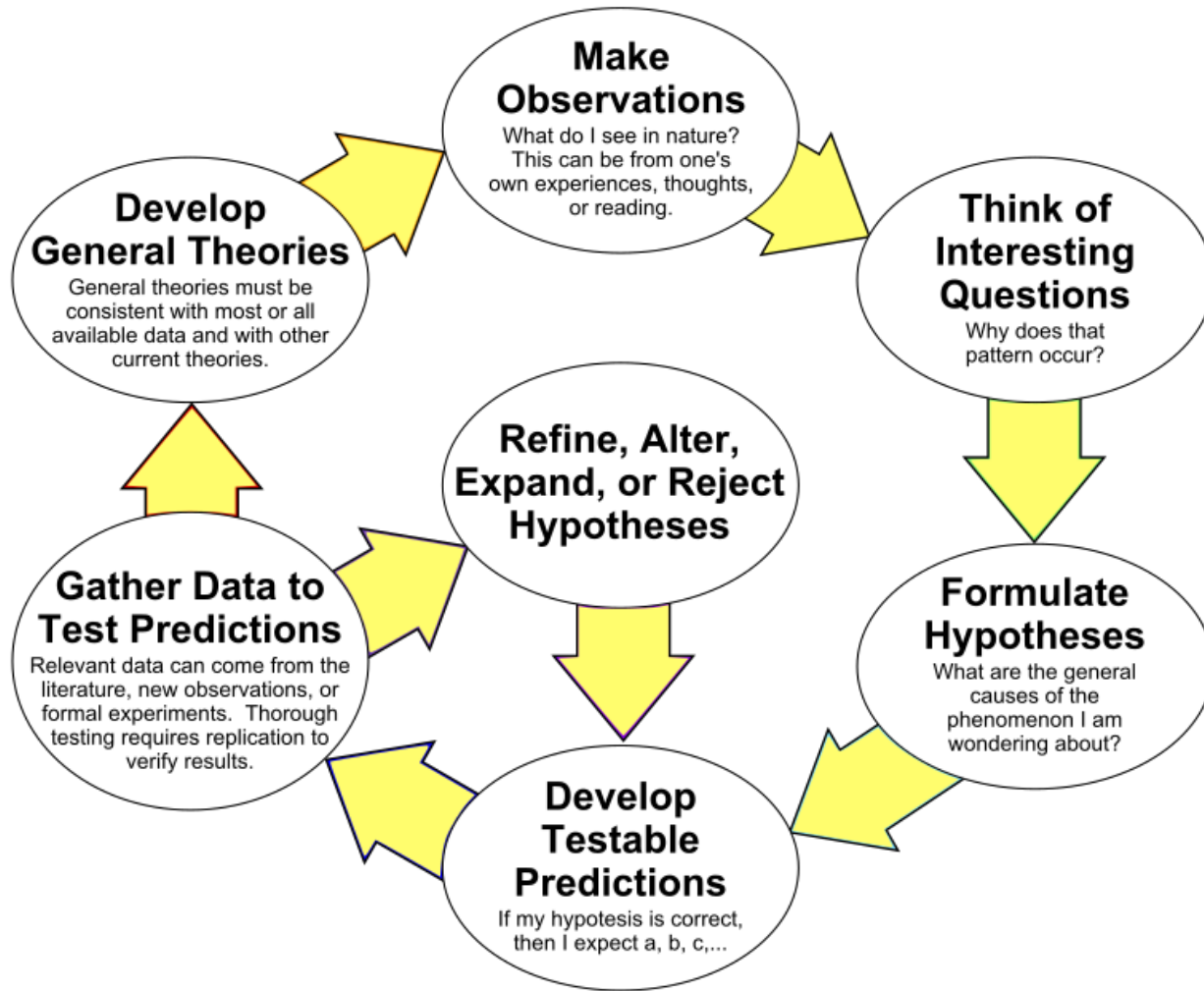


# Science and Art

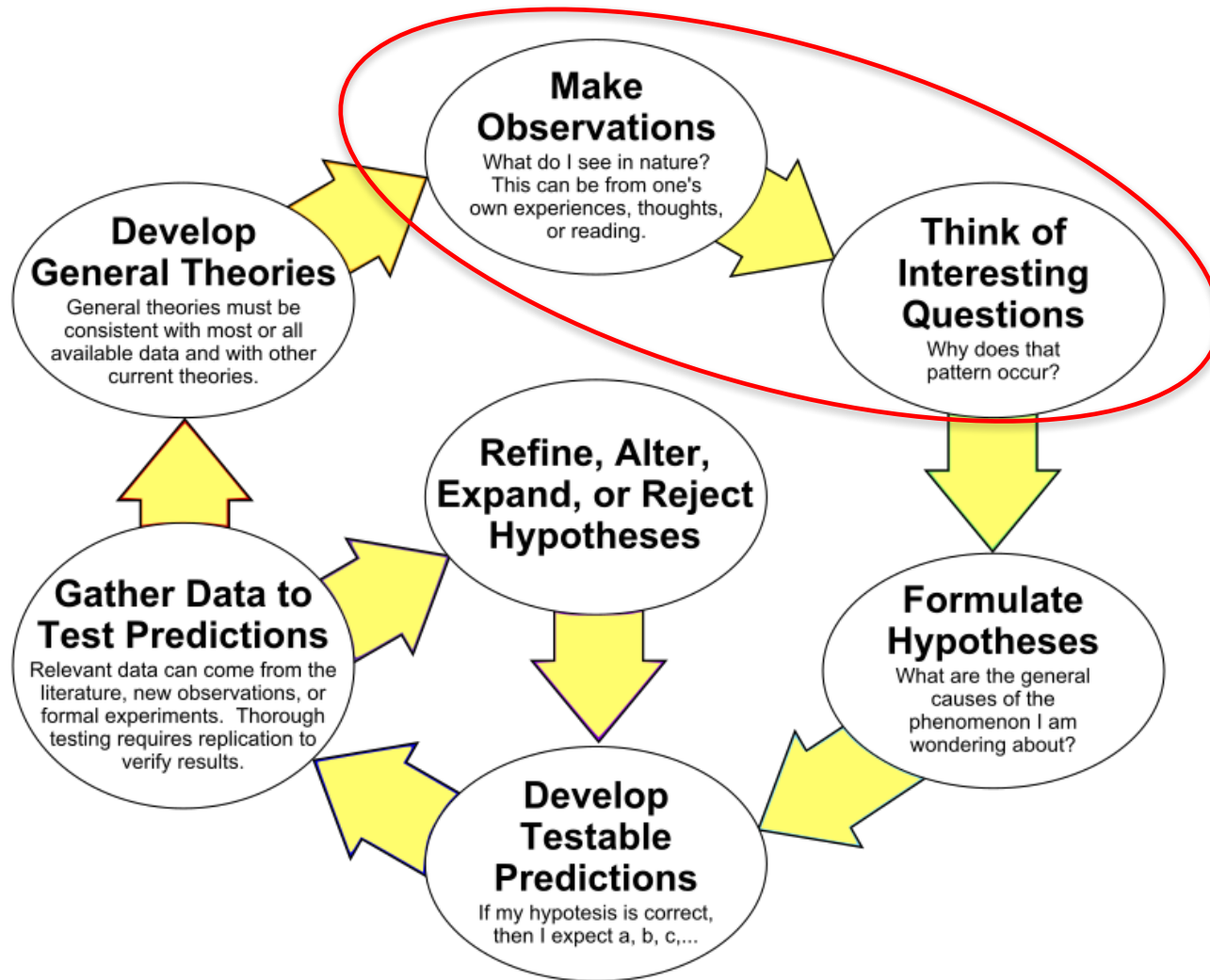
- Intuition and empirical techniques play a huge role in Data Science
  - What is the best approach to analyze the given data?
  - What are interesting insights to evaluate?
  - How do I define the “interest” of a discovery
  - How can I effectively present the results of my analysis?



# The scientific method



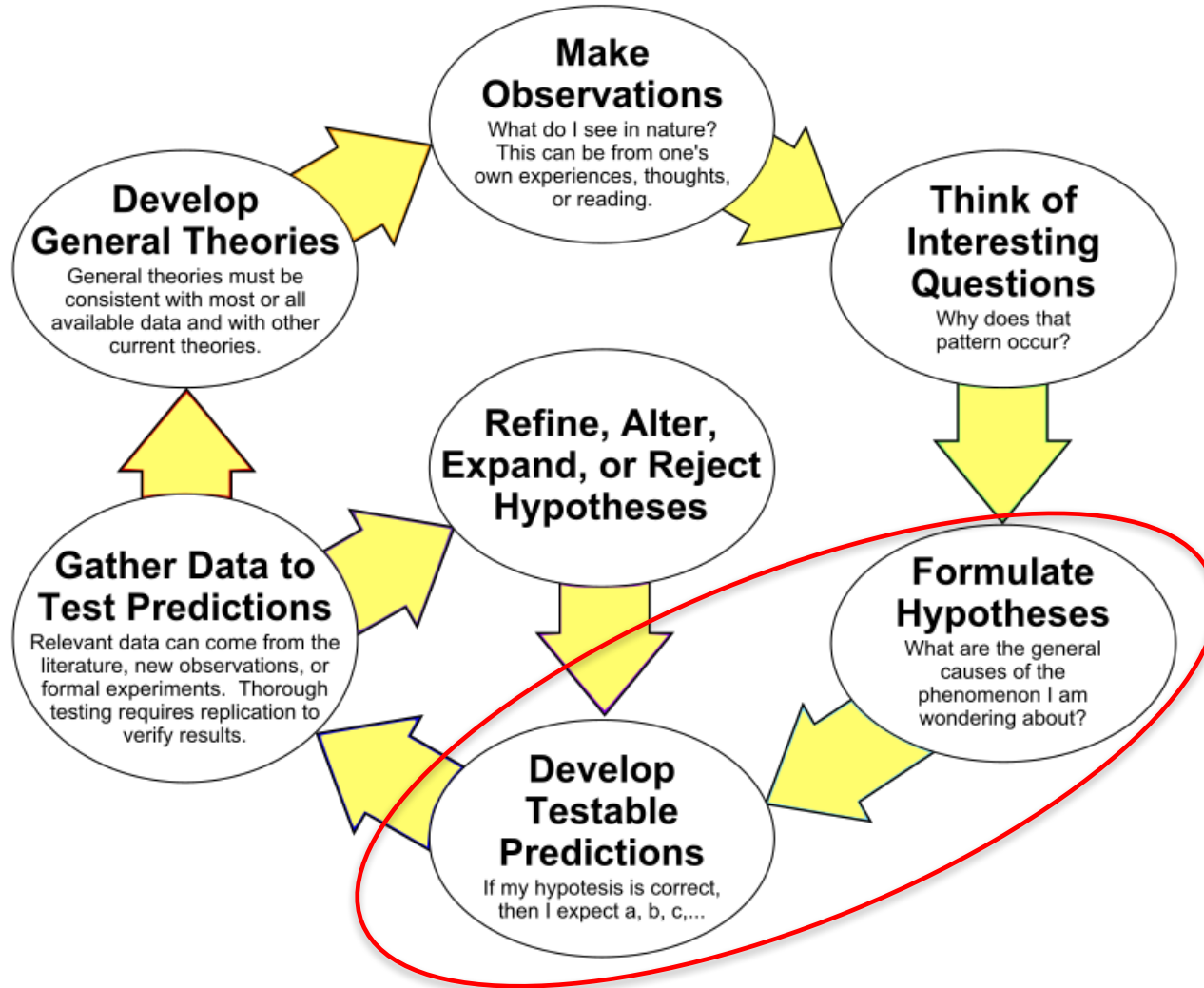
# The scientific method



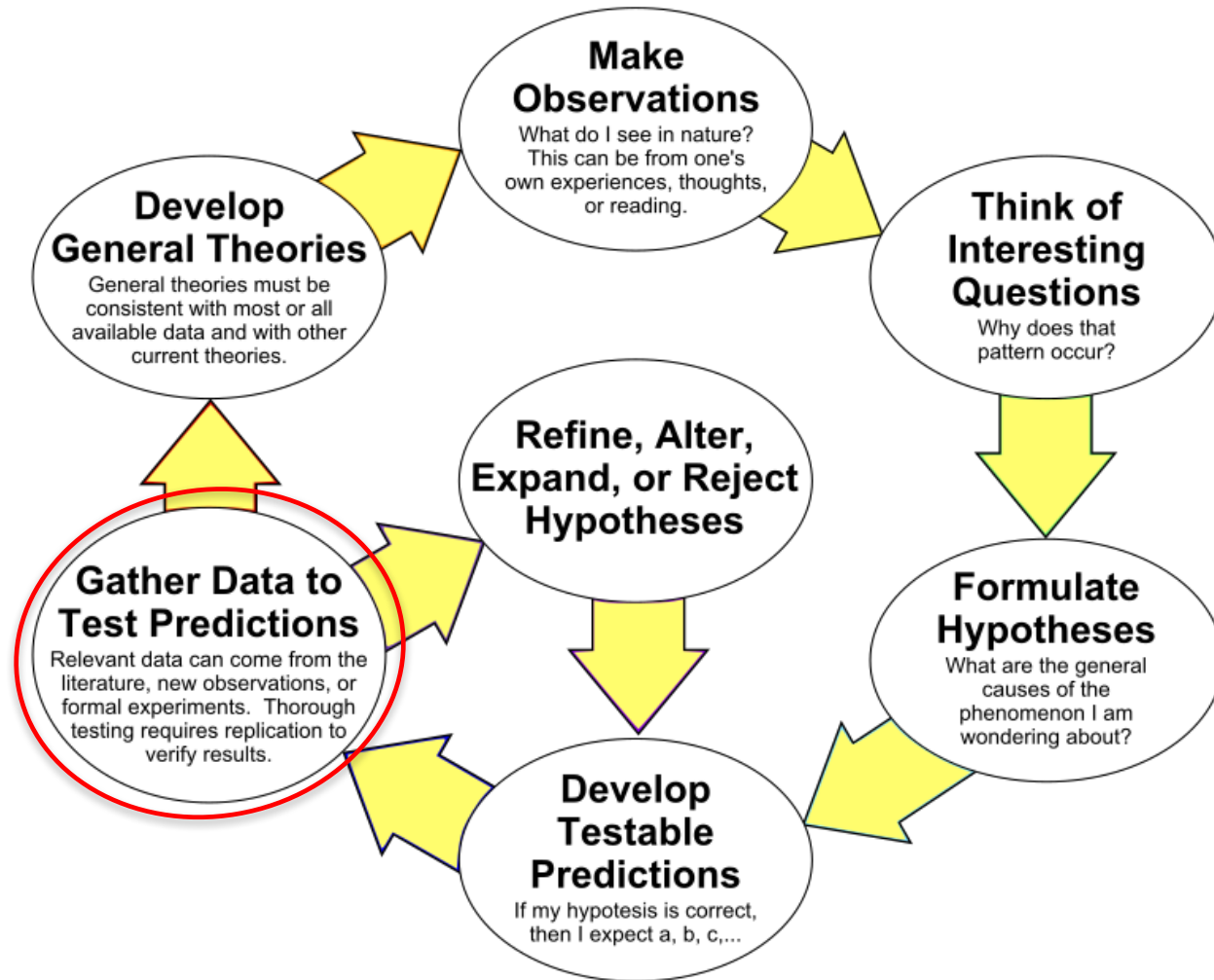
# The Scientific method of data science

- Theoretical models and analysis help us understand the validity of our insight
  - Exploratory analysis (even when it involves the biggest of data) is meant to *\*form\** a hypothesis, not test one
  - Good experimental design and **rigorous statistics** are essential if we want to make claims about how the world works
  - “All models are **wrong**, but some are **useful**.”

# The scientific method

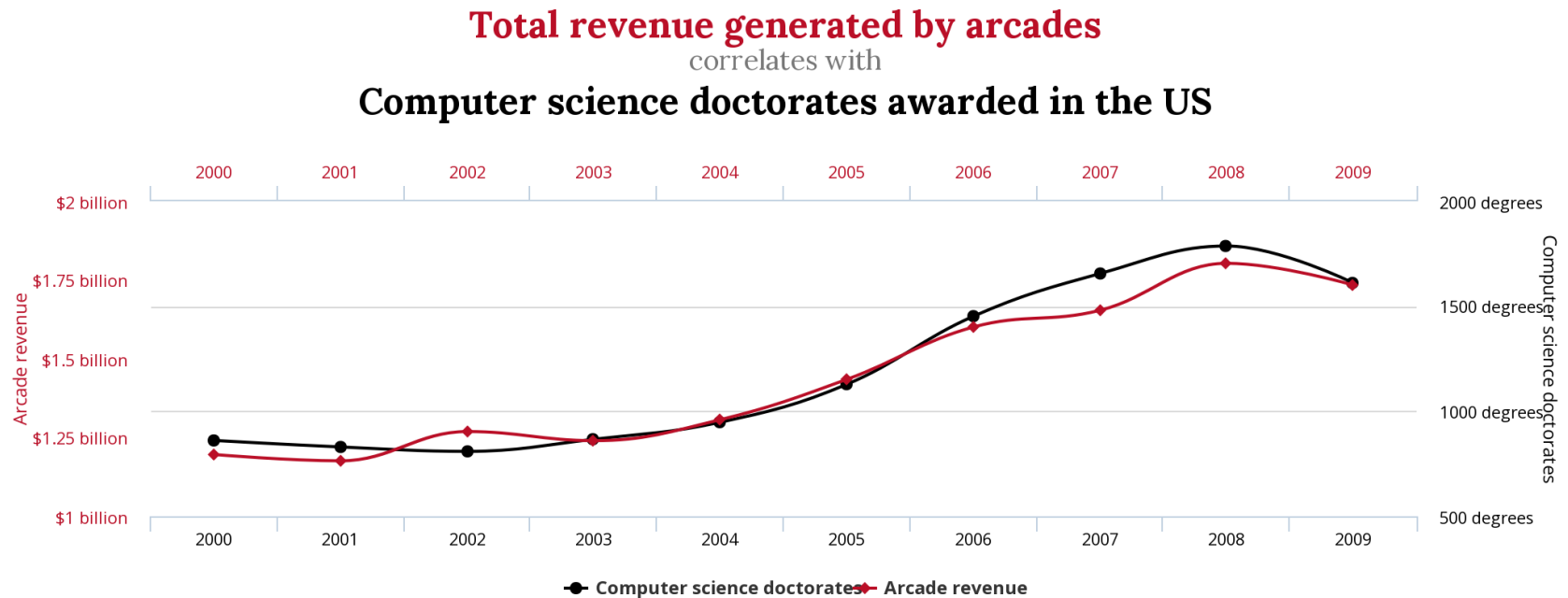


# The scientific method



# The Scientific method of data science

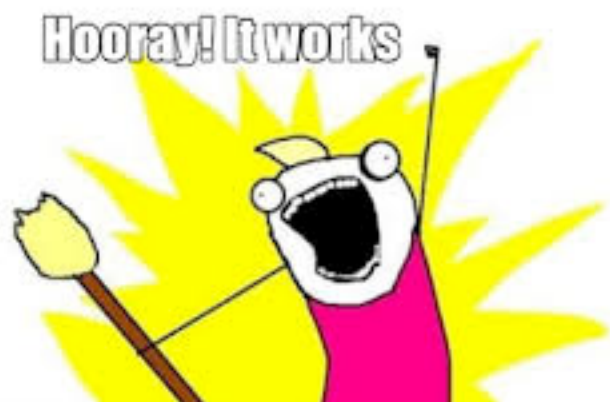
Intuition can be useful but also very dangerous!



... just because something looks right or relevant it does not mean that it is!

# The problem with heuristic and intuition

- We have plenty of great more or less complicated heuristics that work great!



That's great!  
How does it work?

# The problem with heuristic and intuition

- We have plenty of great more or less complicated heuristics that work great!

**I mean....it works! YAY!**

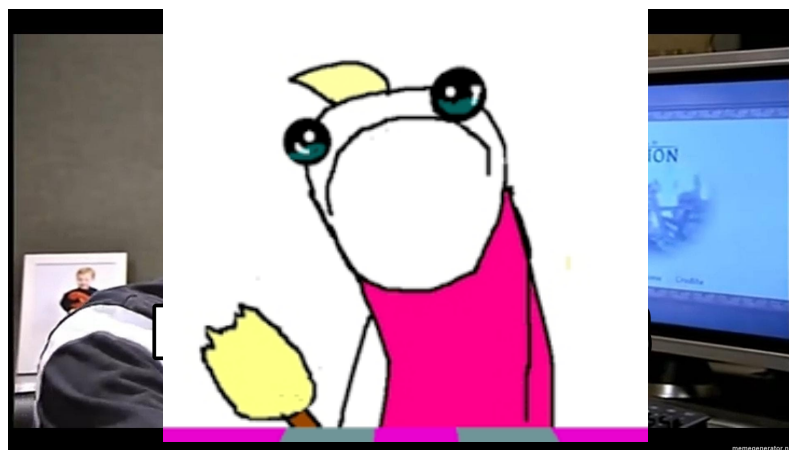


OK I get it!  
But how are you  
sure that it is  
working correctly?



# The problem with heuristic and intuition

- We have plenty of great more or less complicated heuristics that work great!



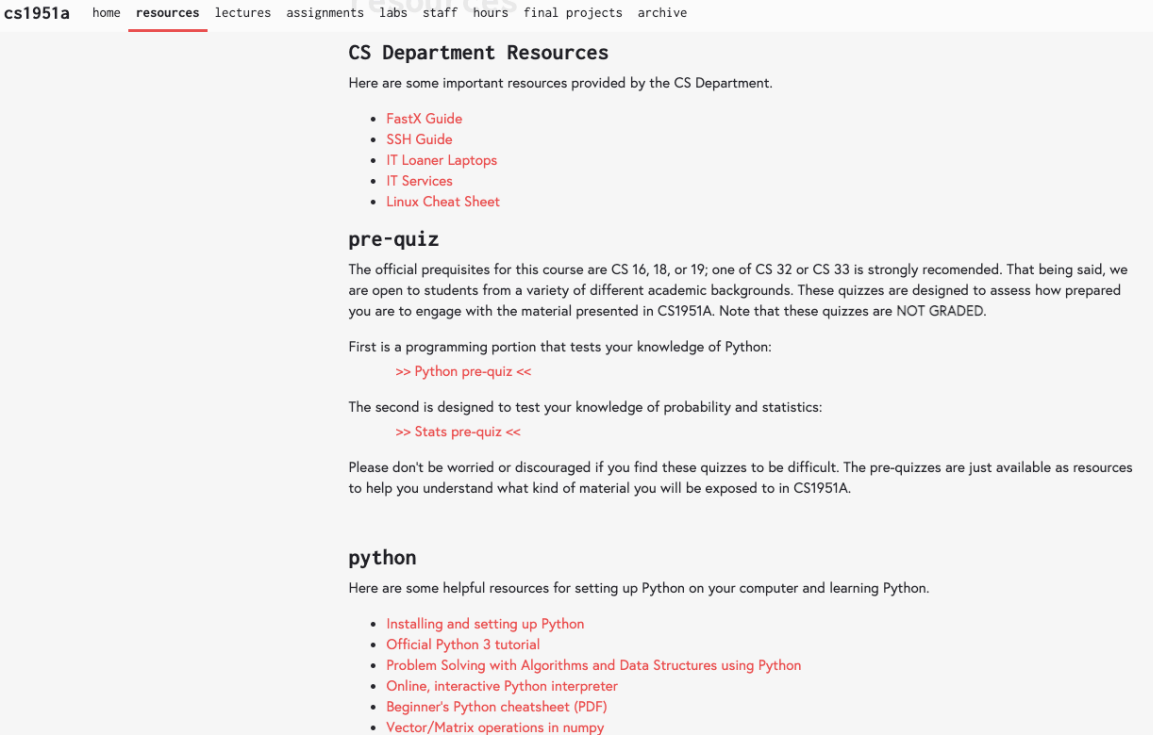
You really have no idea, have you?

# The benefit of rigorous analysis

- Helps us understand why our methods are working!
  - Guarantees on the **significance** and **correctness** of our insights
  - **Generalizability** of our discoveries
  - Distinguish “true” phenomena from **noise in the data**
  - Understanding of the properties of the population being considered
  - Crucial for **reproducibility**

# What do I need to know?

- Programming experience is very useful
  - Python 3.7
  - Use the resources on the website
- Background in statistics



The screenshot shows the 'resources' page of the CS1951A course website. The navigation bar includes links for 'home', 'resources', 'lectures', 'assignments', 'labs', 'staff', 'hours', 'final projects', and 'archive'. The main content is titled 'CS Department Resources' and lists several helpful links: 'FastX Guide', 'SSH Guide', 'IT Loaner Laptops', 'IT Services', and 'Linux Cheat Sheet'. Below this, there are two 'pre-quiz' sections. The first is for Python, and the second is for statistics. A note states that these quizzes are not graded but are intended to help students understand the material. At the bottom, there is a 'python' section with links to resources for setting up Python, including an official tutorial, a problem-solving guide, an interactive interpreter, a beginner's cheatsheet, and a guide to numpy operations.

cs1951a home resources lectures assignments labs staff hours final projects archive

### CS Department Resources

Here are some important resources provided by the CS Department.

- [FastX Guide](#)
- [SSH Guide](#)
- [IT Loaner Laptops](#)
- [IT Services](#)
- [Linux Cheat Sheet](#)

#### pre-quiz

The official prerequisites for this course are CS 16, 18, or 19; one of CS 32 or CS 33 is strongly recommended. That being said, we are open to students from a variety of different academic backgrounds. These quizzes are designed to assess how prepared you are to engage with the material presented in CS1951A. Note that these quizzes are NOT GRADED.

First is a programming portion that tests your knowledge of Python:

[>> Python pre-quiz <<](#)

The second is designed to test your knowledge of probability and statistics:

[>> Stats pre-quiz <<](#)

Please don't be worried or discouraged if you find these quizzes to be difficult. The pre-quizzes are just available as resources to help you understand what kind of material you will be exposed to in CS1951A.

#### python

Here are some helpful resources for setting up Python on your computer and learning Python.

- [Installing and setting up Python](#)
- [Official Python 3 tutorial](#)
- [Problem Solving with Algorithms and Data Structures using Python](#)
- [Online, interactive Python interpreter](#)
- [Beginner's Python cheatsheet \(PDF\)](#)
- [Vector/Matrix operations in numpy](#)

# To do now

- Make sure you are registered on Ed
- Submit signed syllabus
- Consult resources on the website
- Apply for a lab slot
- Start thinking on the group project
  - Find collaborators
  - Share ideas