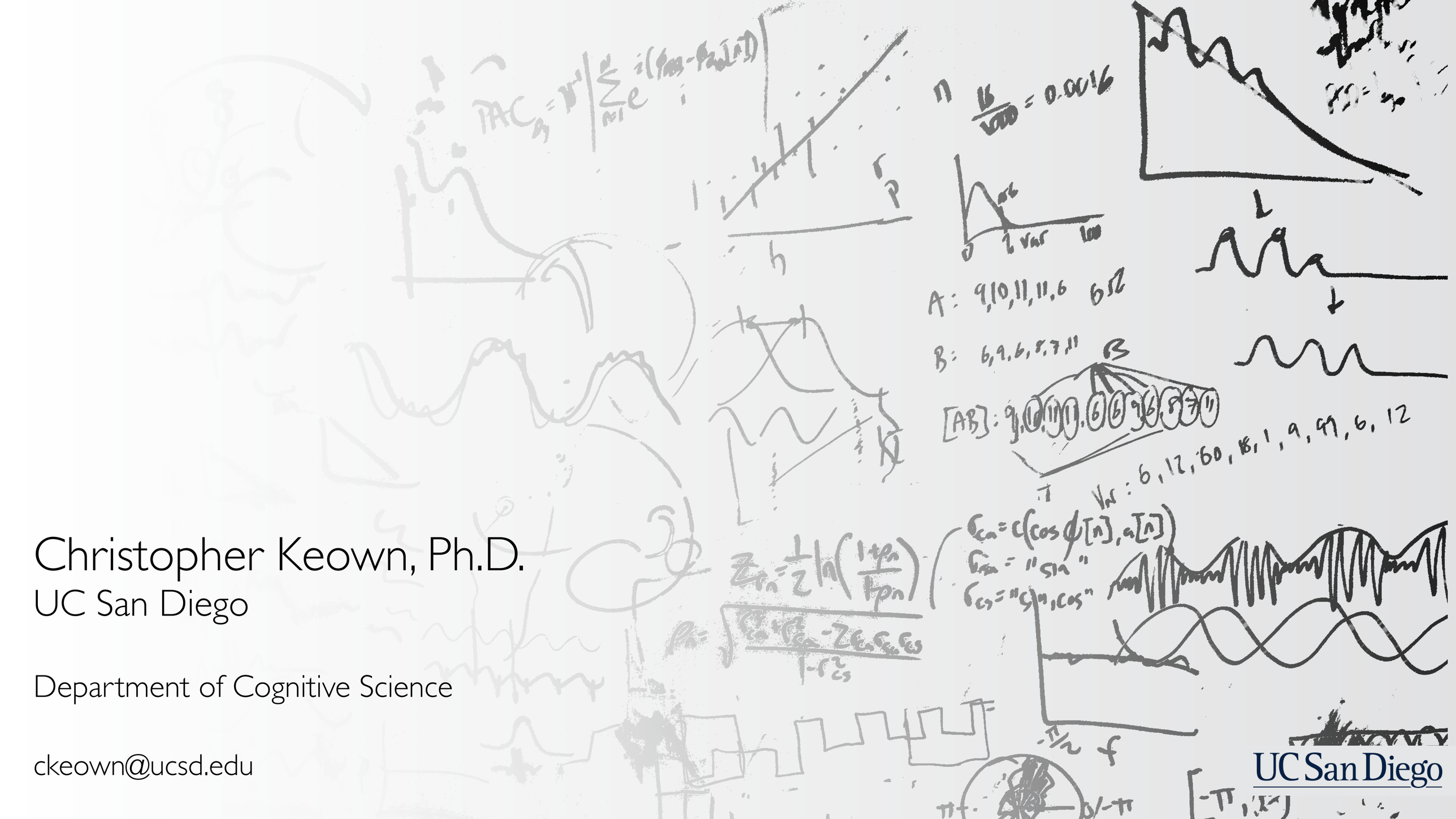


Christopher Keown, Ph.D.

UC San Diego

Department of Cognitive Science

ckeown@ucsd.edu

UC San Diego

Administrative stuff

- Waitlist
- Everyone should be on Piazza now

Administrative stuff

All office hours are now posted on Piazza!

UCSD Data Science Society

- *First meeting of the quarter this Thursday*
- *Refresher on Python program*
- *More info on the club's Facebook page*

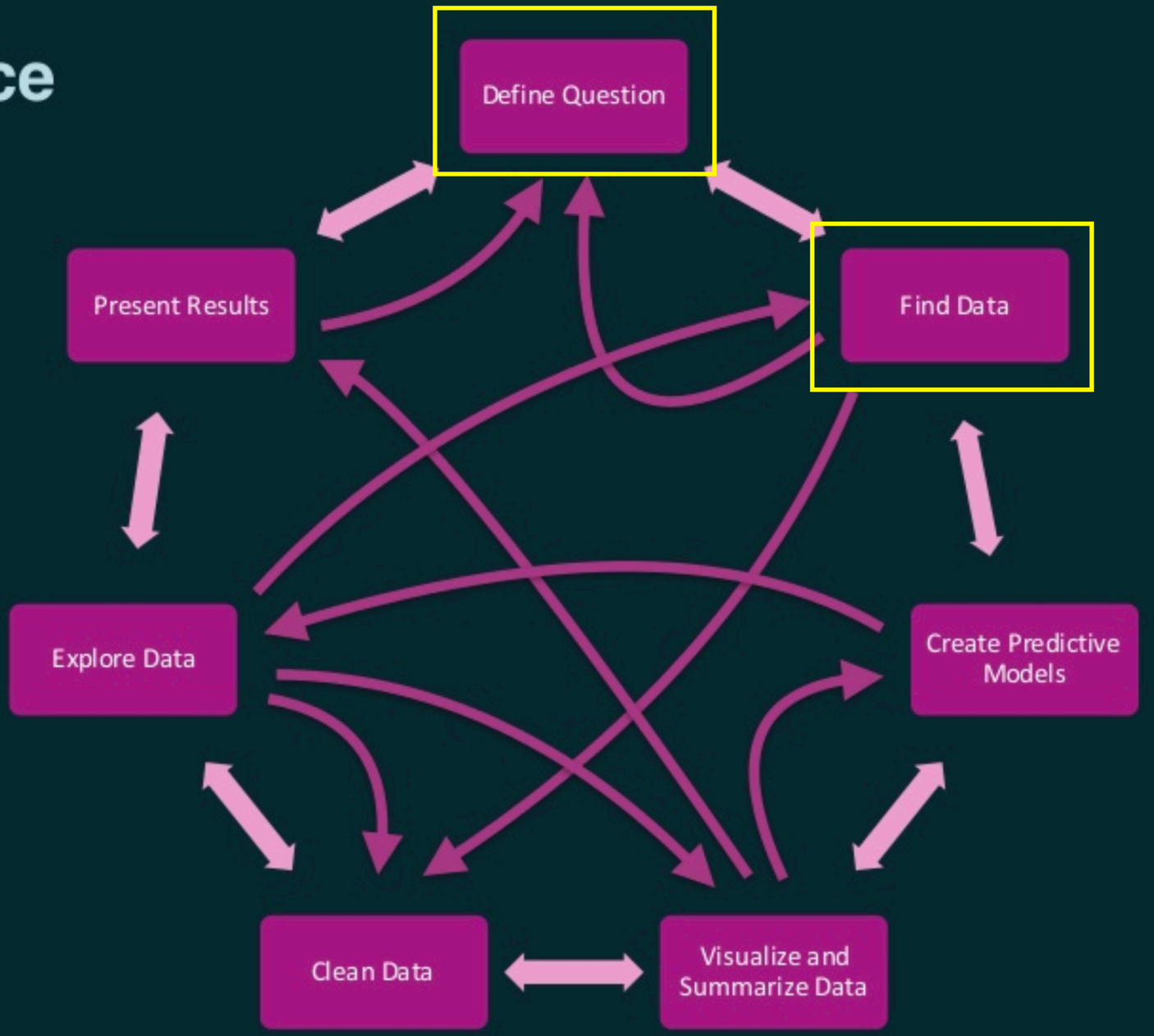
UC San Diego

Assignment 1

- *It's now live.*
- *It's due Sunday, 1/20/2019 at 11:59:59 PM*
- *The final notebook will be uploaded to Triton Ed.*

Assignment 1

Data Science Workflow



@MargrietGr

Cogs108 Capstone Final Project

The 108 Capstone Project will give you the chance to explore a topic of your choice and to expand your analytical skills. By working with real data of your choosing you can examine questions of particular interest to you.

UC San Diego

Final Project - Objectives

- Identify the problems and goals of a ***real*** situation and dataset.
- Choose an appropriate approach for formalizing and testing the problems and goals, and be able to articulate the reasoning for that selection.
- Implement your analysis choices on the dataset.
- Interpret the results of the analyses.
- Contextualize those results within a greater scientific and social context, acknowledging and addressing any potential issues related to privacy and ethics.
- Work effectively to manage a project as part of a team.

Final Project - Grading

- 5% of your class grade will be from the project proposal
- 30% of your class grade will be from the final project

Final project grade breakdown

Category	Percentage of Project Grade
Introduction and Background	**10%**
Data Description	**10%**
Data Cleaning/Pre-processing	**10%**
Data Visualization	**15%**
Data Analysis and Results	**25%**
Privacy/Ethics Considerations	**15%**
Conclusions and Discussion	**15%**

Deadlines and action items

- *Proposal due Monday, February 18th @ 11:59 pm (Week 6)*
- *Due Thursday, March 21st @ 11:59 pm (Finals Week)*

Final Project - Teams

- To accomplish this you will work in teams of 3 to 6 students to conceive of and carry out an analysis project.
- Everyone must be part of a group (even if you really, really, really, really don't want to).

Final Project - The Proposal

Project Proposal - Detailed Description

For the Project Proposal you need to write a report, in the style outlined below, about how you might approach your question of interest. Specifically, every Report must contain seven sections, briefly outlined here, with more specific direction provided in the proposal template notebook:

1) Research Question: What's your question?
2) Hypothesis: What's your prediction?
3) Dataset(s): What data will you use to answer your question? Describe your dataset(s).
4) Background: Why is this question of interest, what background information led you to your hypothesis, and why is this important?
5) Proposed Methods: What methods will you use to analyze your data?
6) Ethics: Acknowledge and address any potential ethics and privacy issues related to your project.
7) Discussion: Discuss the potential impact of your project, as well as trying to anticipate any problems you may encounter.

Source: https://github.com/COGS108/Projects/blob/master/CapstoneProject.pdf

Objective for today

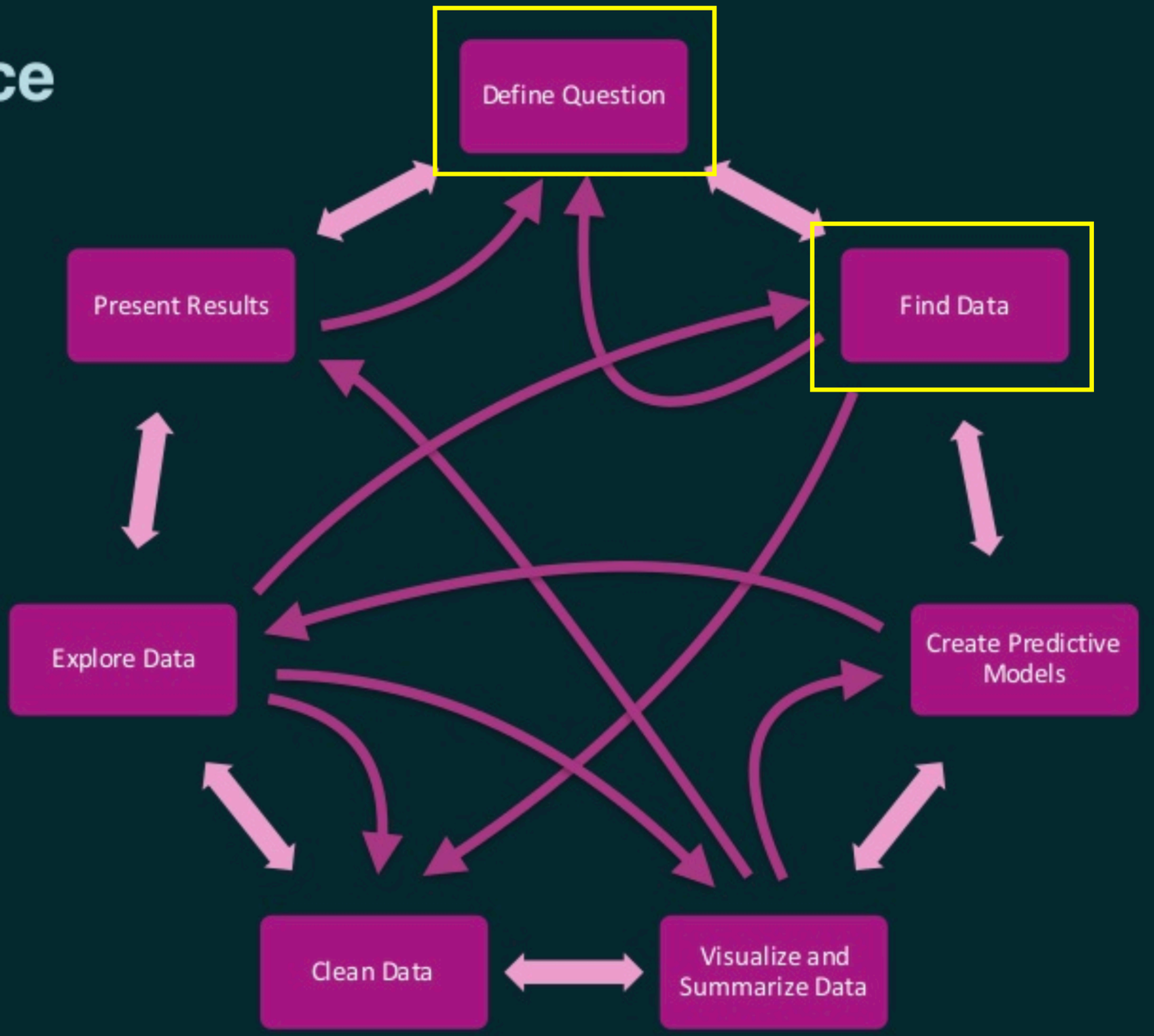
Data science is a way of thinking!
Let's learn how data scientists think.

Data scientists are curious!

Data scientists are curious and always ask questions:

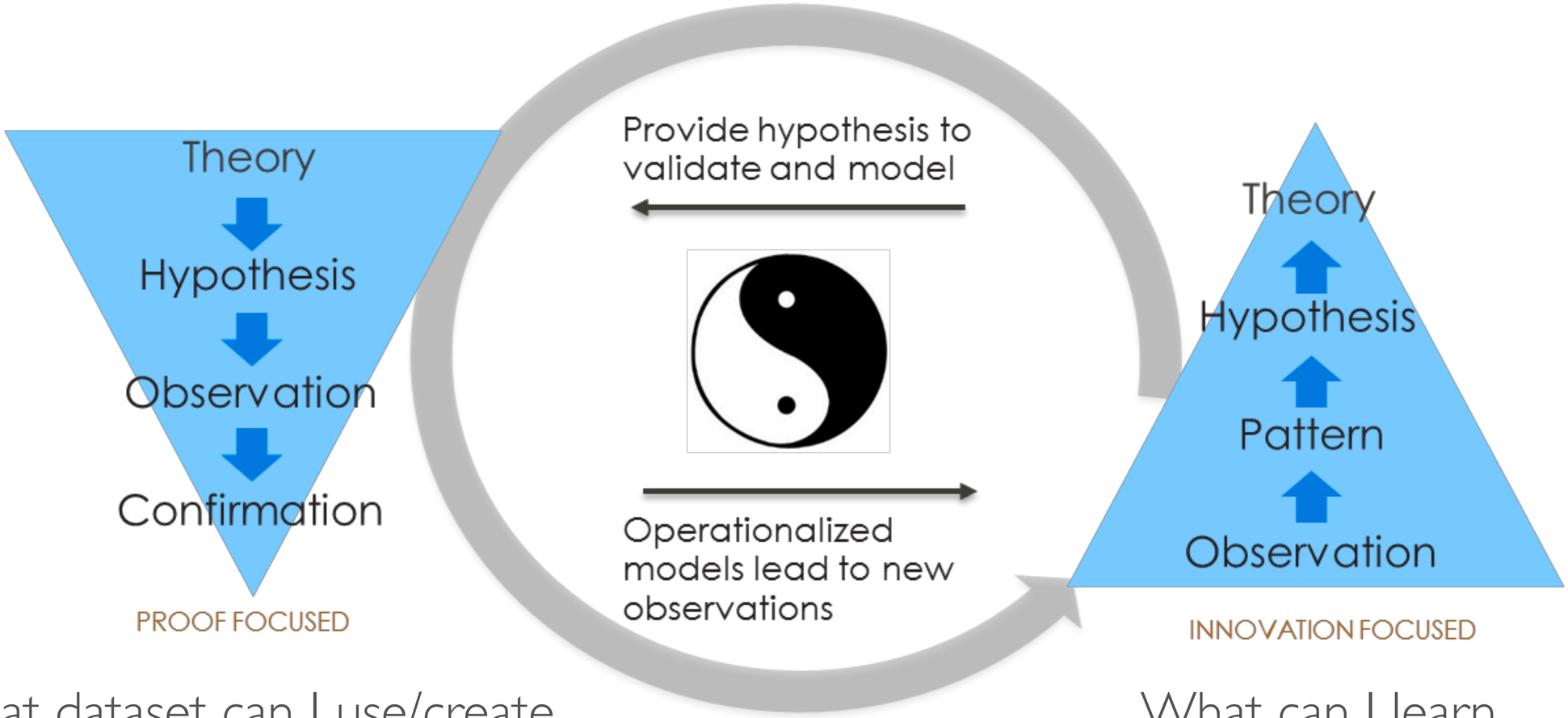
- *What things might you be able to learn from a given data set?*
- *What do you really want to know about the world?*
- *What will it mean to you once you find out?*

Data Science Workflow



@MargrietGr

Hypothesis driven vs. data driven



What dataset can I use/create to answer this problem?

Be creative and pragmatic!

What can I learn from this dataset?

If I had an hour to solve a problem and my life depended on it, I would use the first 55 minutes determining the proper question to ask, for once I know the proper question, I could solve the problem in less than five minutes. —Einstein

What makes a question a good question?

The data science oracle

When choosing your question, imagine that you are approaching an oracle that can tell you anything in the universe, as long as the answer is a number or a name. It's a mischievous oracle, and its answer will be as vague and confusing as it can get away with. You want to pin it down with a question so airtight that the oracle can't help but tell you what you want to know.

Be your own devil's advocate

The data science oracle

Examples of poor questions that leave wiggle room for useless answers:

- What can my data tell me about my business?
- What should I do?
- How can I increase my profits?

Examples of good questions where the answer is impossible to avoid:

- How many Model 3s will Tesla sell in San Diego during the third quarter?
- How many students will apply for admission to UCSD in 2019?
- How many students should UCSD admit in 2019 for a target class size of 5000?

IMDB

The Godfather (1972)

9.2/10
1,399,281
Rate This

R | 2h 55min | Crime, Drama | 24 March 1972 (USA)

Watch Now
From $2.99 (SD) on Prime Video

ON DISC

The aging patriarch of an organized crime dynasty transfers control of his clandestine empire to his reluctant son.

Director: Francis Ford Coppola
Writers: Mario Puzo (screenplay by), Francis Ford Coppola (screenplay by) | 1 more credit »
Stars: Marlon Brando, Al Pacino, James Caan | See full cast & crew »

Al Pacino (I)

Top 500

Actor | Producer | Soundtrack

One of the greatest actors in all of film history, Al Pacino established himself during one of cinema's most vibrant decades, the 1970s, and has become an enduring and iconic figure in the world of American movies. Alfredo James Pacino was born on April 25, 1940 in Manhattan, New York City, to an Italian-American family. His parents, Rose (Gerardi)...
See full bio »

Born: April 25, 1940 in Manhattan, New York City, New York, USA

More at IMDbPro »
Contact Info: View agent, publicist, legal on IMDbPro

1741 photos | 99 videos »

IMDB

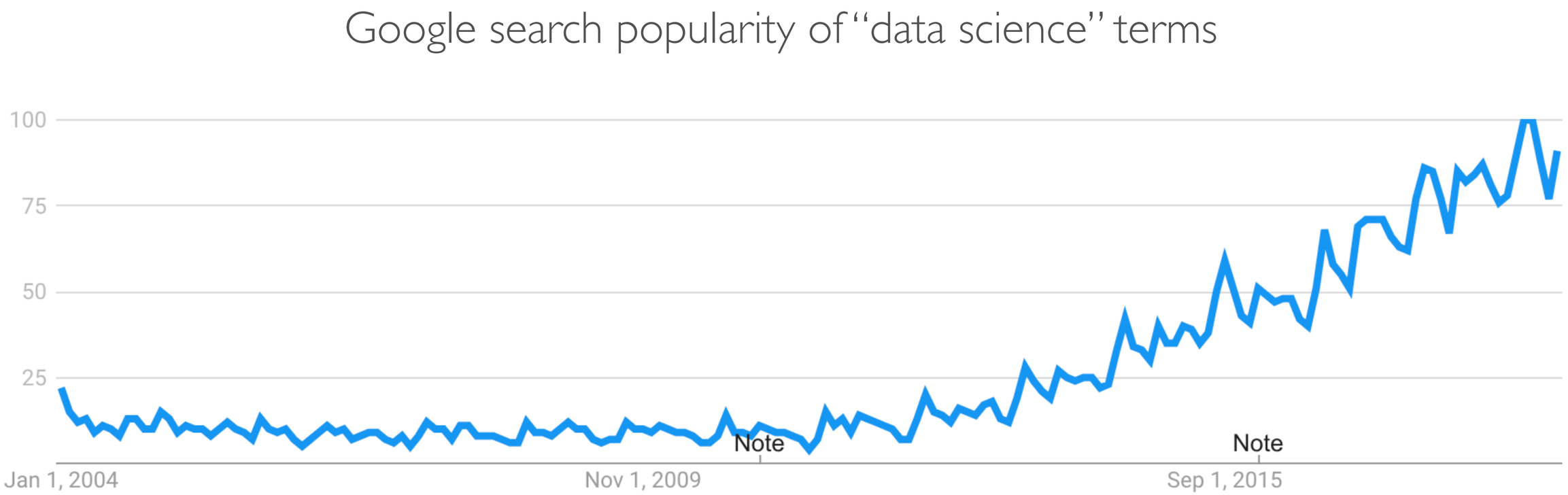
Narrow questions:

- *Which actor played in the most movies each year for each genre?*
- *Which movie genre has the highest grossing film each year?*

More general questions:

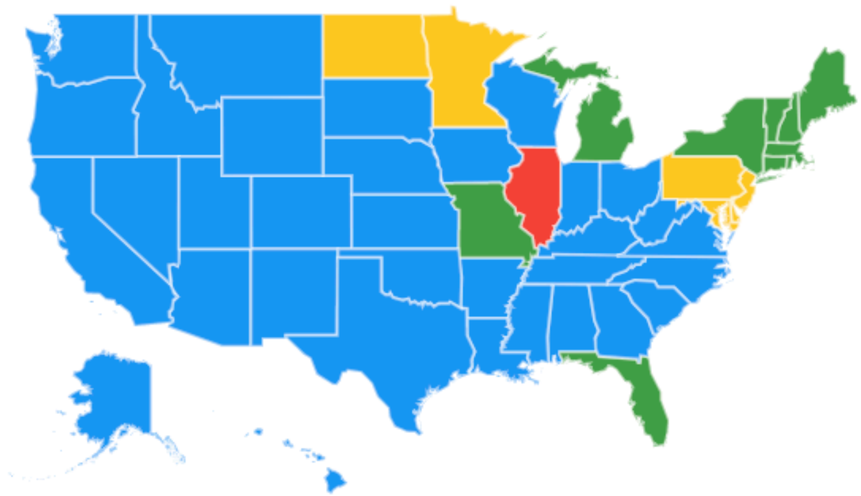
- *How often does having an all-star cast produce a block-buster film?*
- *How do Hollywood movies compare to Bollywood movies, in terms of ratings, budget, and gross?*
- *Do movie stars live longer or shorter lives than bit players, or compared to the general public?*

Google Trends



If you had access to all of the google search data (key words, location and date/ time of the search), what questions would you ask?

Search interest in the most searched teams, past week



- Dallas Cowboys
- Chicago Bears
- Philadelphia Eagles
- New England Patriots
- Baltimore Ravens

Finding Miss Universe

The annual Miss Universe competition aspires to identify the most beautiful woman in the world. Can computational models predict who will win a beauty contest?

Nature of a data scientist

Data scientists:

- *are data driven.*
- *care about answers. They analyze data to discover something about how the world works.*
- *care about whether the results make sense, because they care about what the answers mean.*
- *are comfortable with the idea that data has errors.*
- *know nothing is ever completely true or false in science, while everything is either true or false in computer science or mathematics.*

Data Science vs. Computer Science

- *Data scientist's job is to turn numbers into insight*
- *Software engineers are focused on building and optimizing systems:*
  - *Setting up an API to access data through*
  - *Make data processing and access faster*
- *They collaborate!*

Final project - action items

- *Start identifying questions and datasets*
- *Look for team members — post on Piazza*
- *No changes to teams after week 6, once the proposal is submitted*