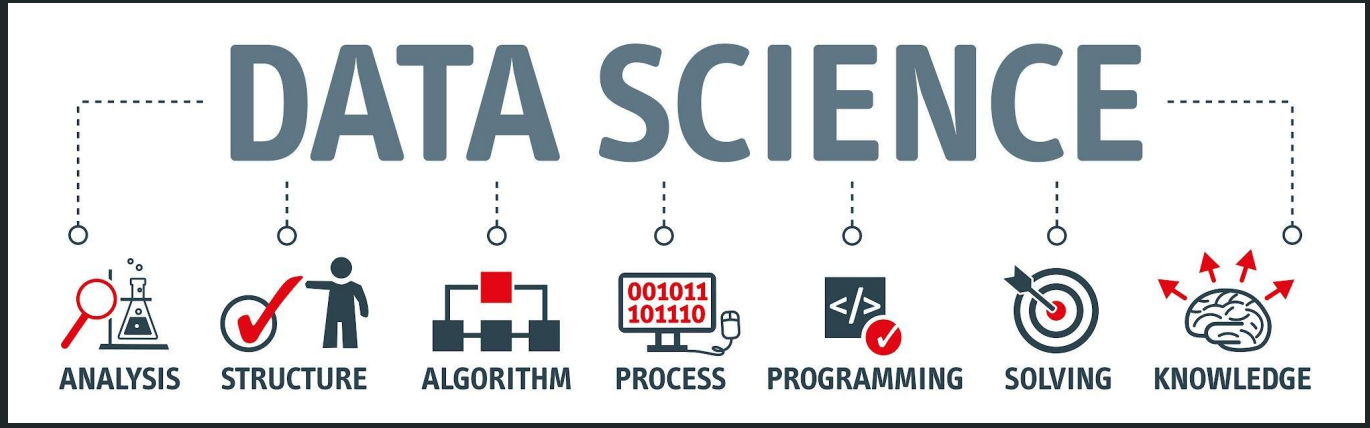


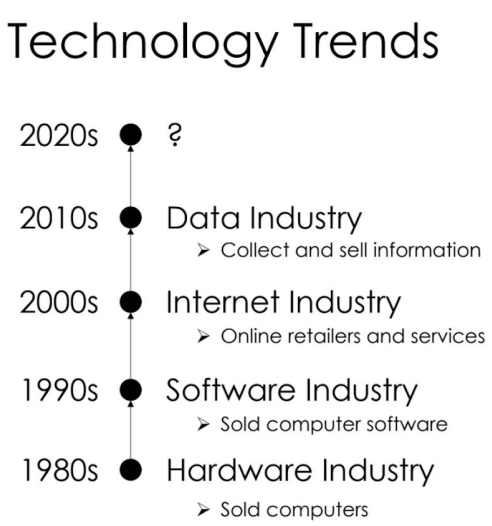
data science applications:
pandas and more

what is data science?

jack.wrenn.fyi/blog/brown-location-surveillance/

- at its heart, it is applied math and statistics!
- the amount of data we as a society have amassed is immense
  - what data have you generated today?
- having the computational skills enables you to work in any sector

amazing potential for good: healthcare sector
- single doctors / hospitals cannot sample enough data to see big trends
- researchers don't have access to the types of data hospitals have
- together, create network of data to find trends not visible to single doctors to create life saving policies



Academics

Master's Degree

Certificate in Data Fluency

Data Science Fellows

Certificate in Data Fluency

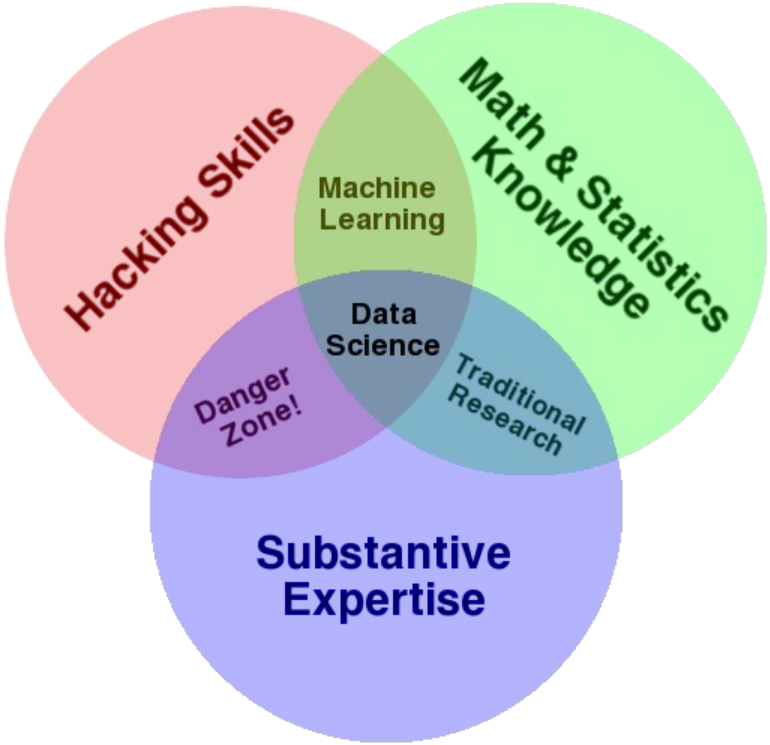
Description

The Certificate in Data Fluency is for undergraduate students who wish to gain fluency and facility with the tools of data analysis and its conceptual framework, but who are not pursuing a concentration in a data-intensive discipline. The program is designed to provide fundamental conceptual knowledge and technical skills to students with a range of intellectual backgrounds and concentrations, while emphasizing a critical liberal learning perspective. For more information about eligibility and requirements:

DATA FLUENCY CERTIFICATE INFO ▸

Purpose

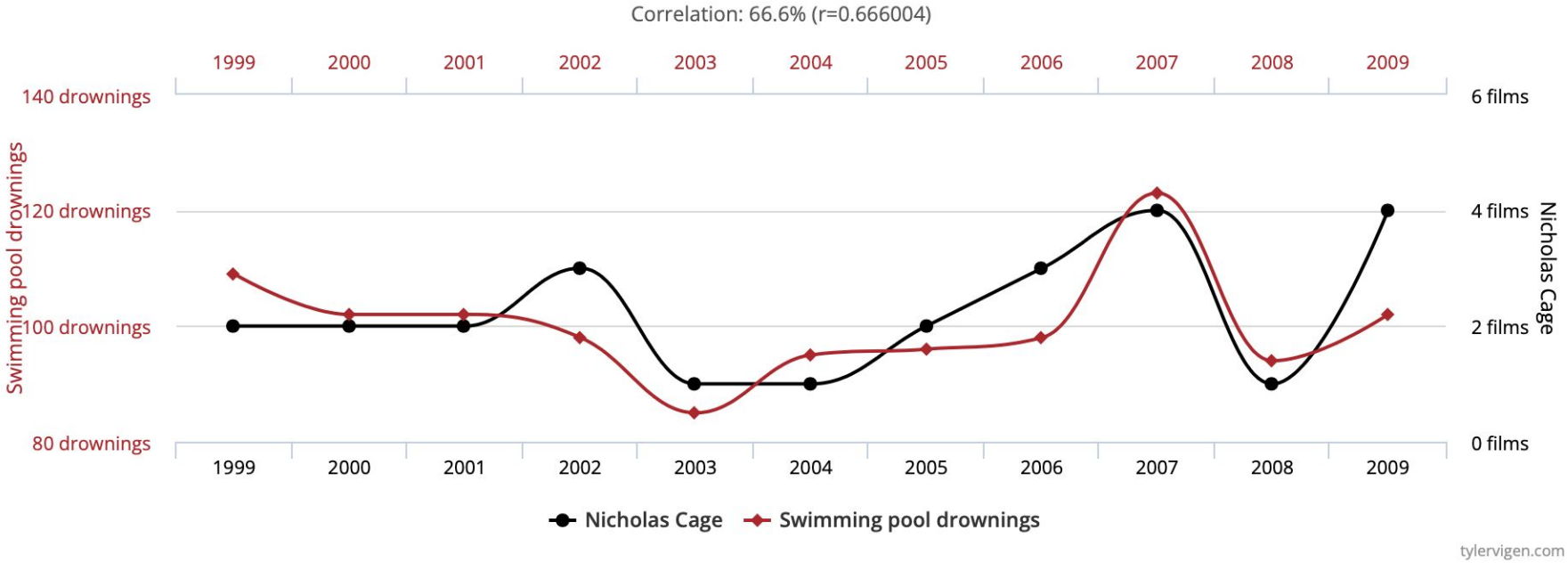
Data fluency implies a familiarity with data science and a basic competency working with data. Many disciplines now require an understanding of how data are collected, stored, analyzed, and visualized. The purpose of this certificate is to prepare


Hacking Skills
Math & Statistics Knowledge
Substantive Expertise
Machine Learning
Data Science
Danger Zone!
Traditional Research

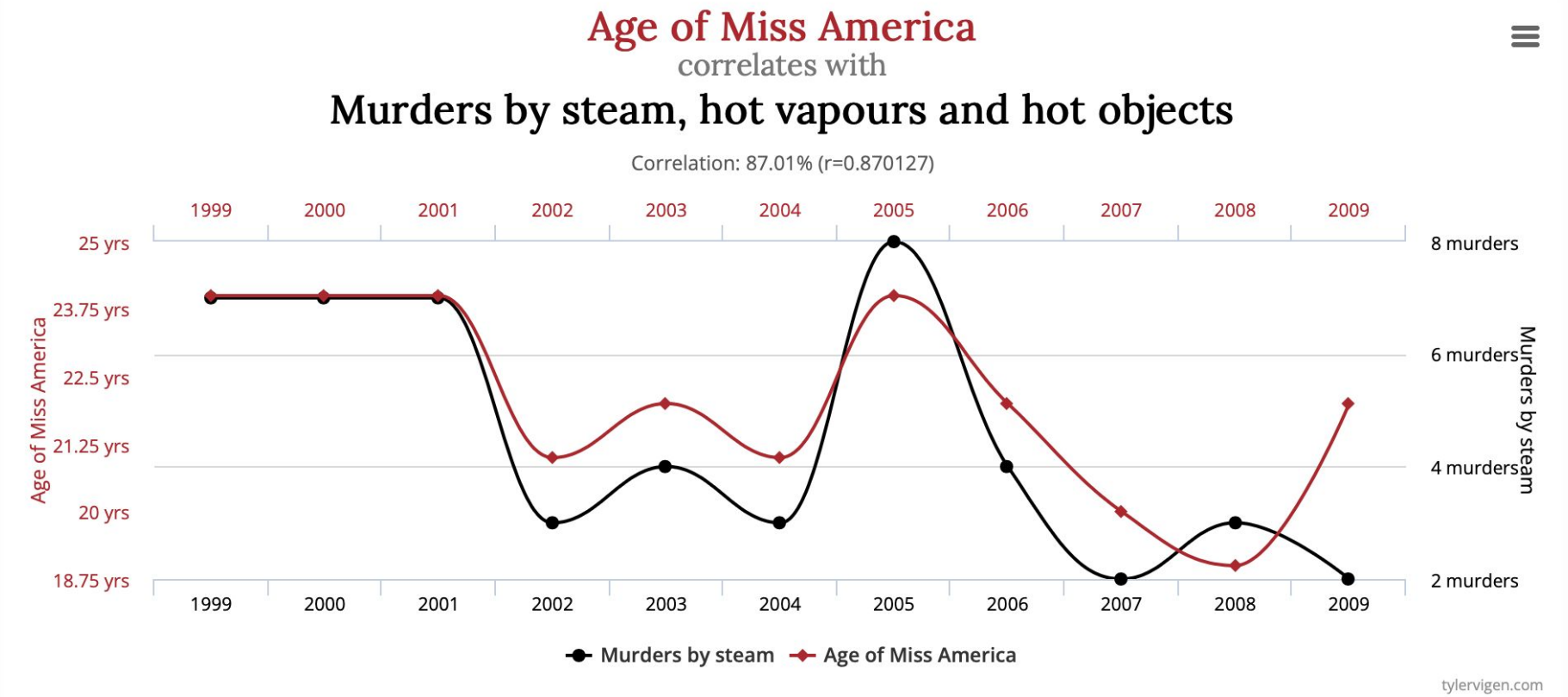
Number of people who drowned by falling into a pool

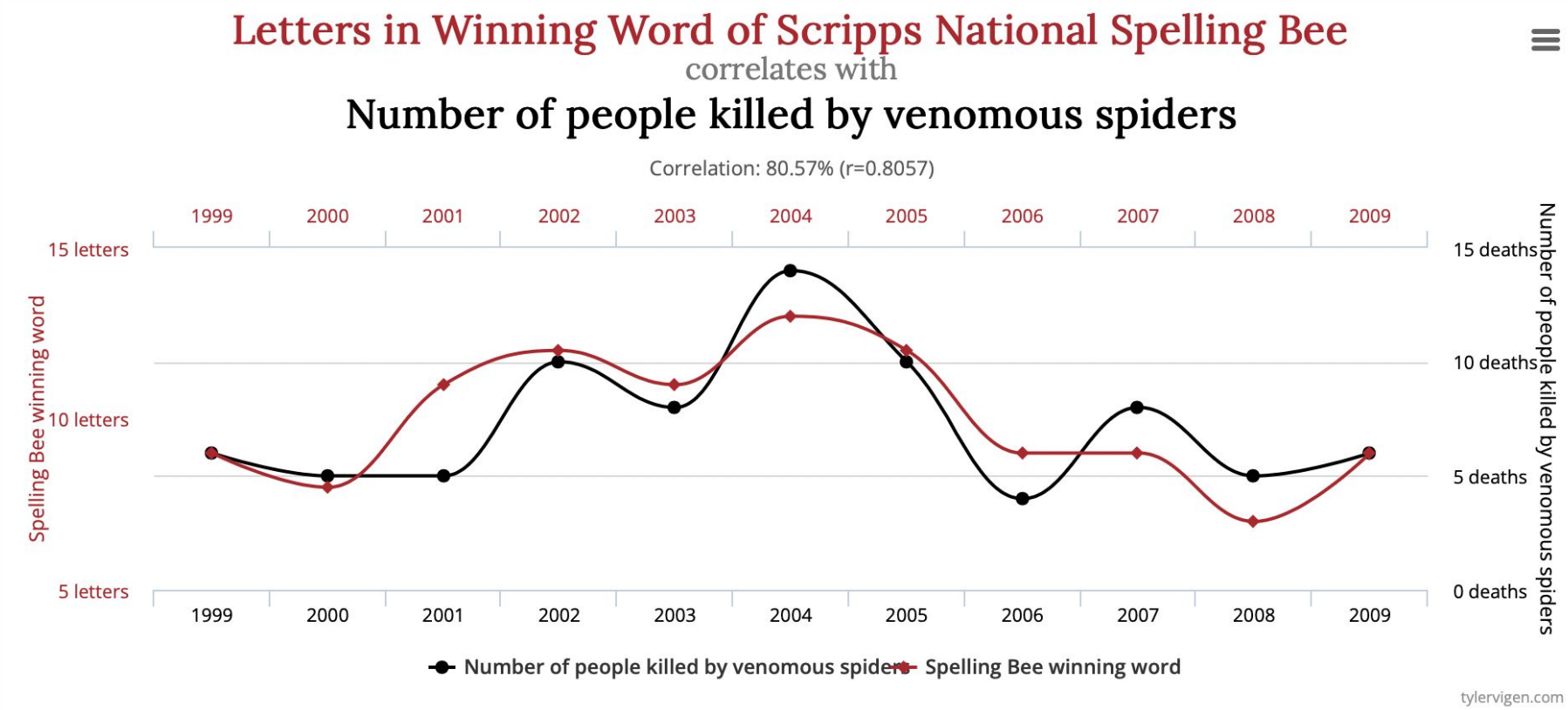
correlates with

Films Nicolas Cage appeared in



Data sources: Centers for Disease Control & Prevention and Internet Movie Database


Age of Miss America
correlates with
Murders by steam, hot vapours and hot objects
Correlation: 87.01% (r=0.870127)
1999
2000
2001
2002
2003
2004
2005
2006
2007
2008
2009
25 yrs
23.75 yrs
22.5 yrs
21.25 yrs
20 yrs
18.75 yrs
Age of Miss America
8 murders
6 murders
4 murders
2 murders
Murders by steam
Murders by steam
Age of Miss America
tylervigen.com
Data sources: Wikipedia and Centers for Disease Control & Prevention

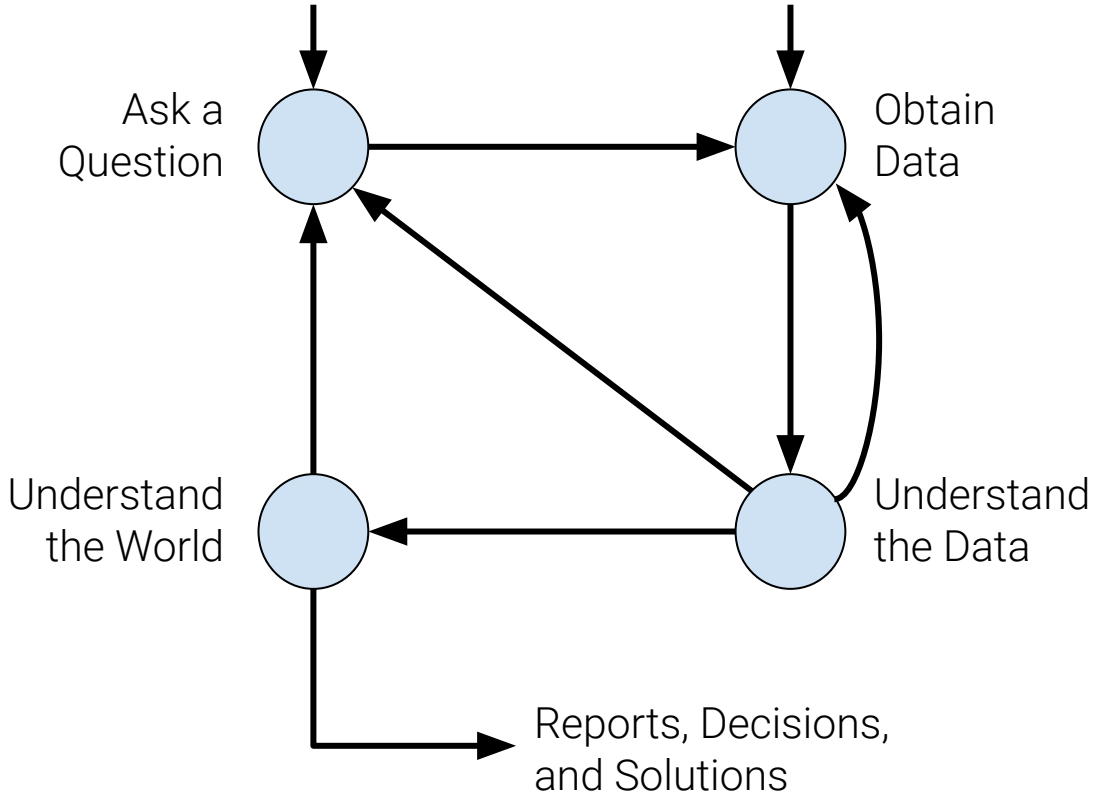


tylervigen.com

Data science lifecycle

The data science lifecycle is a **high-level description** of the data science workflow.

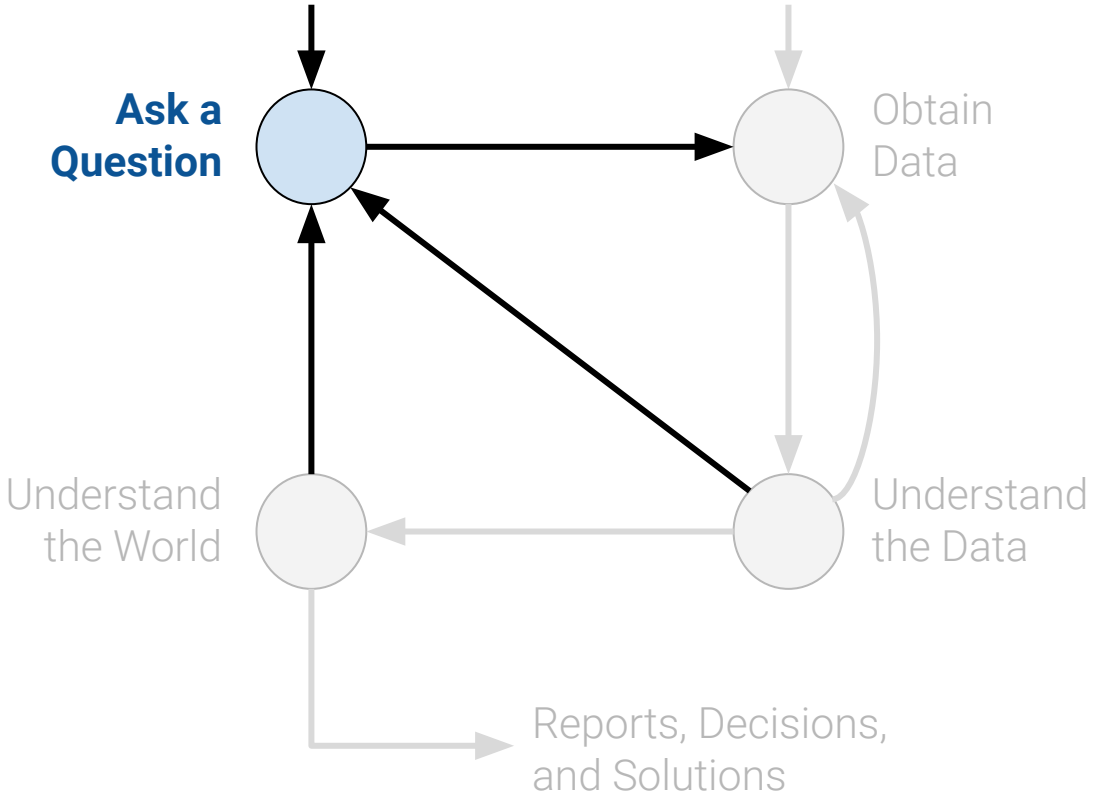
Note the two distinct entry points!



slides taken from Josh Hug @ uc berkeley

1. Question/Problem Formulation

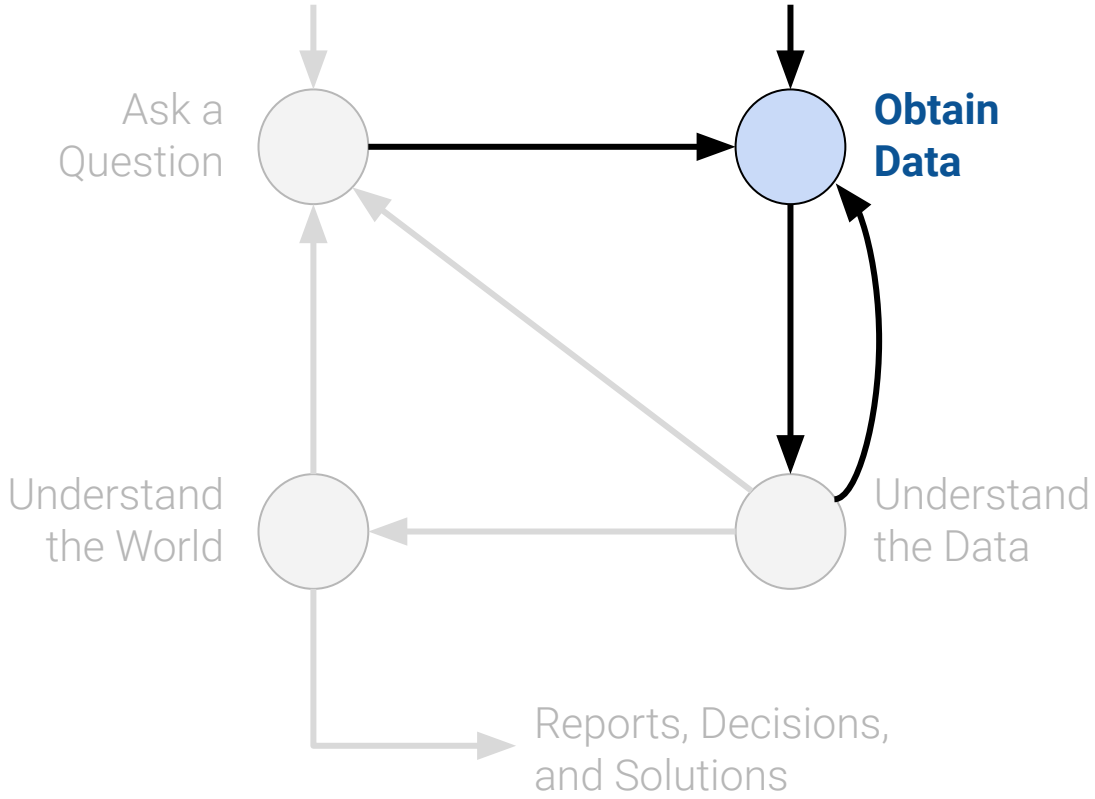
- What do we want to know?
- What problems are we trying to solve?
- What are the hypotheses we want to test?
- What are our metrics for success?



slides taken from Josh Hug @ uc berkeley

2. Data Acquisition and Cleaning

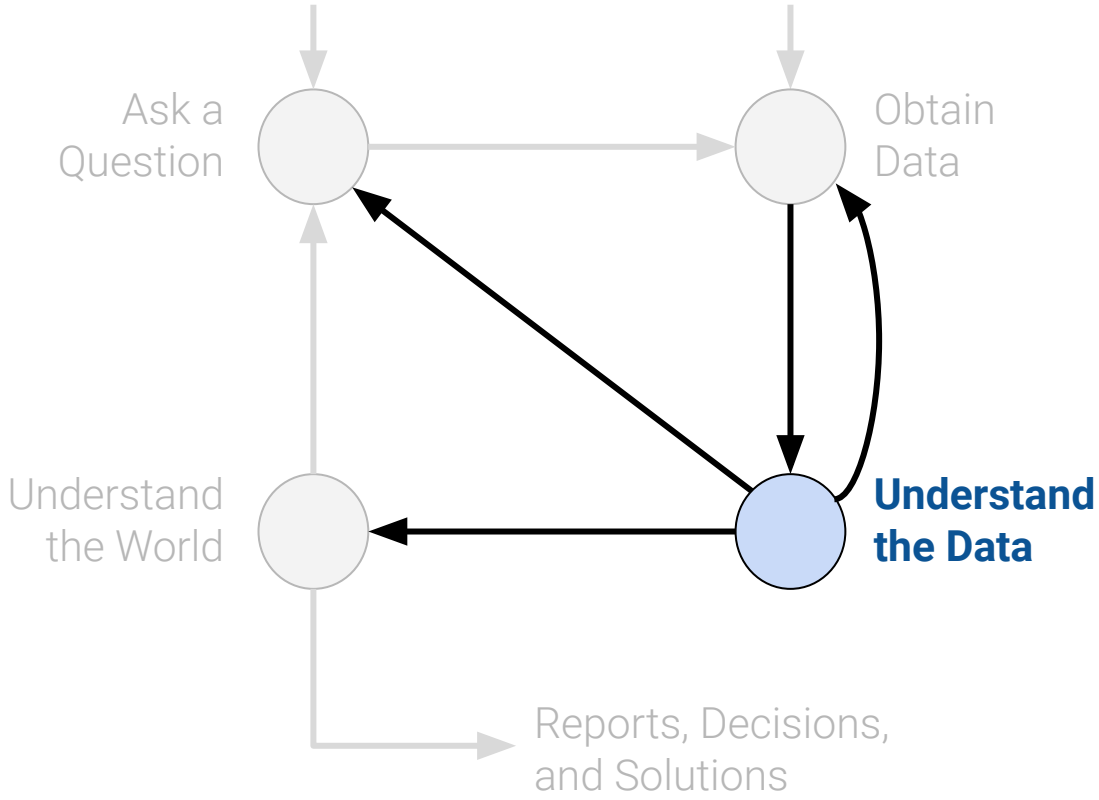
- What data do we have and what data do we need?
- How will we sample more data?
- Is our data representative of the population we want to study?



slides taken from Josh Hug @ uc berkeley

3. Exploratory Data Analysis & Visualization

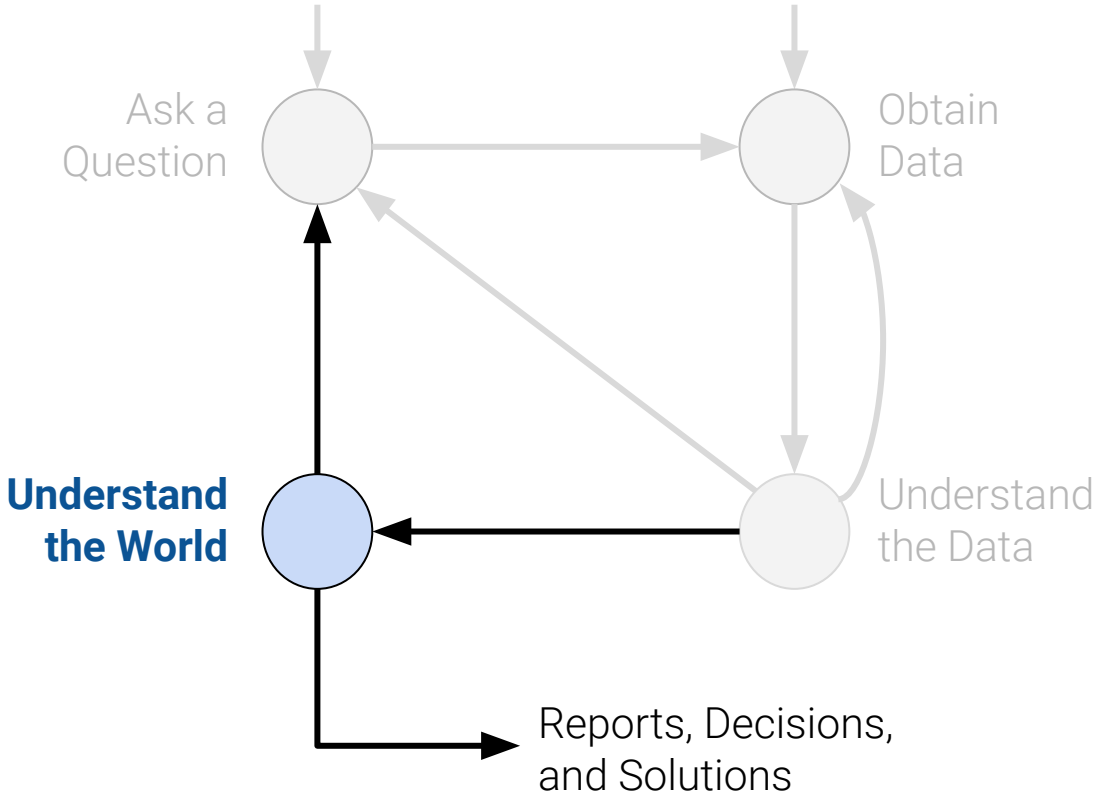
- How is our data organized and what does it contain?
- Do we already have relevant data?
- What are the biases, anomalies, or other issues with the data?
- How do we transform the data to enable effective analysis?



slides taken from Josh Hug @ uc berkeley

4. Prediction and Inference

- What does the data say about the world?
- Does it answer our questions or accurately solve the problem?
- How robust are our conclusions and can we trust the predictions?



slides taken from Josh Hug @ uc berkeley

Dataset Today

bit.ly/datasci_clps950

can we predict the per capita death rate from cancer in US counties using other population level data? what are some powerful data science tools built into python?

what now?

- how might we apply what we have found in this analysis?
  - policy proposals, “teeth” for legislation
  - targeted information campaigns
  - directed monetary support
  - more research into causes (especially if the data we have is not predictive!)
  - more complex models
  - collaborations with other groups