

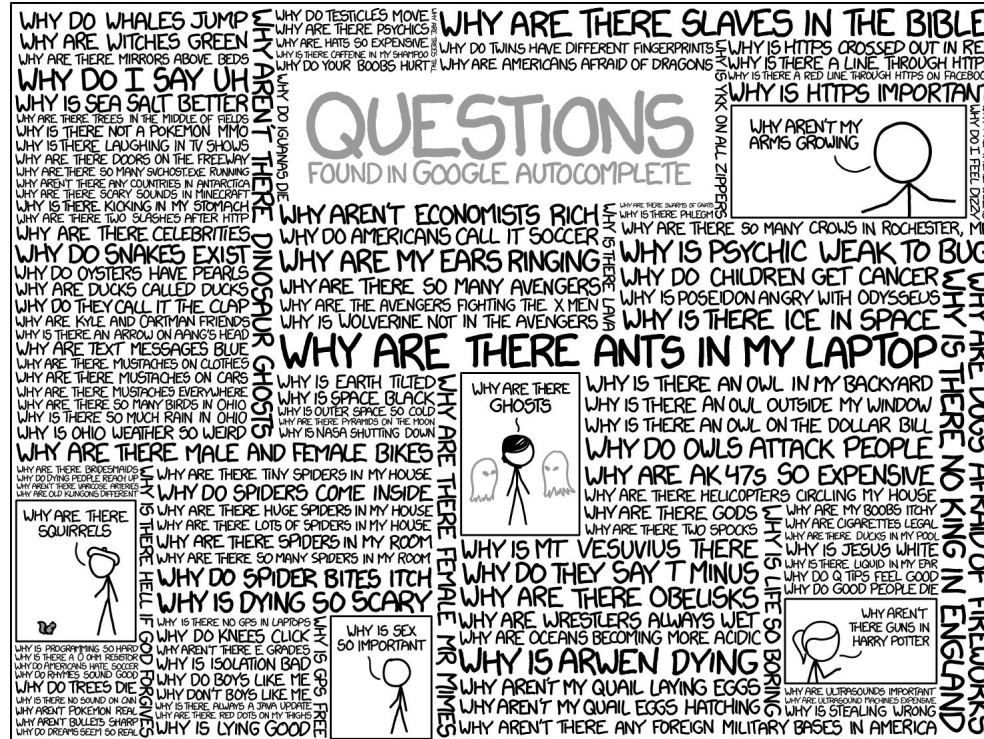
03_Data_Sci_Questions

Kyle Shannon

kshannon@ucsd.edu

Dept. of Cognitive Sciences

UC - San Diego



What You Will Learn

- ☐ Explain the data science process (high level)
- ☐ What are good and bad data science questions
- ☐ How to go from *general* to *specific* when asking questions
- ☐ Real world data science questions

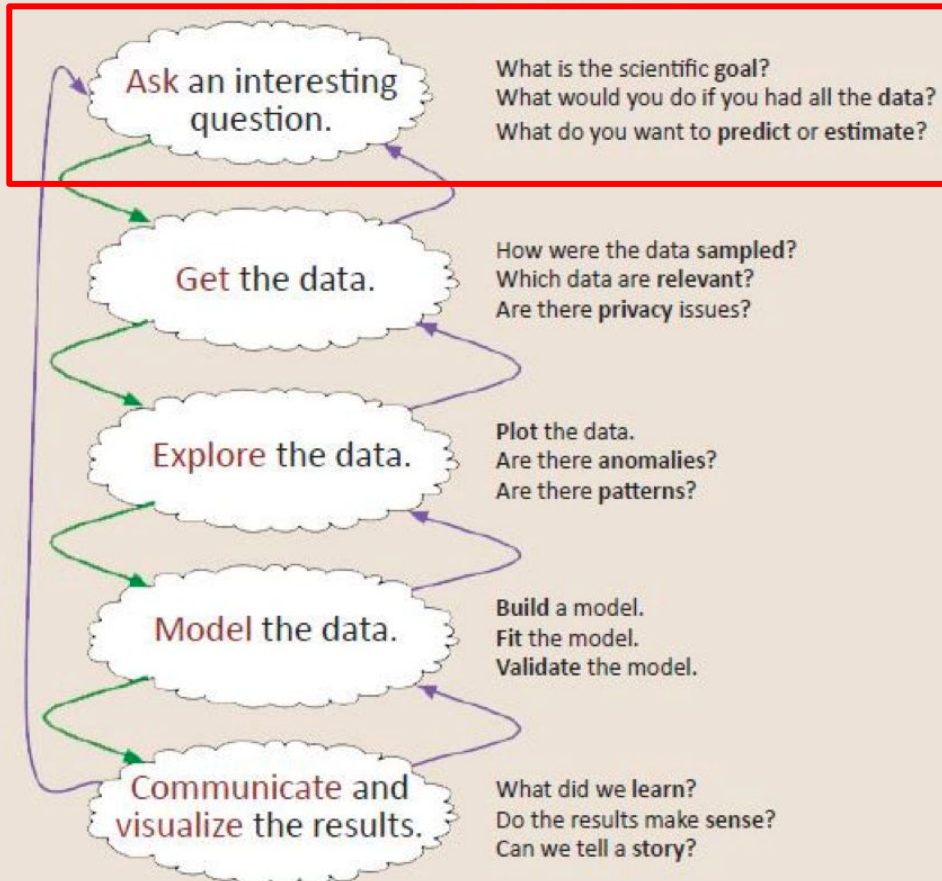
The Data Science Process

(high level)

Nature of a data scientist

- Data-driven decisions for some stakeholder, end user, or client
- Considers domain expertise + analytic expertise
- Desire to uncover what secrets the data is hiding
- Care about the downstream consequences of their results/models
- Understand that raw data is not perfect, it is often incomplete and misleading
- Knows analytic results are not binary (especially when involving humans), but rather shades of confidence.
- Communicates ideas effectively to a broad audience

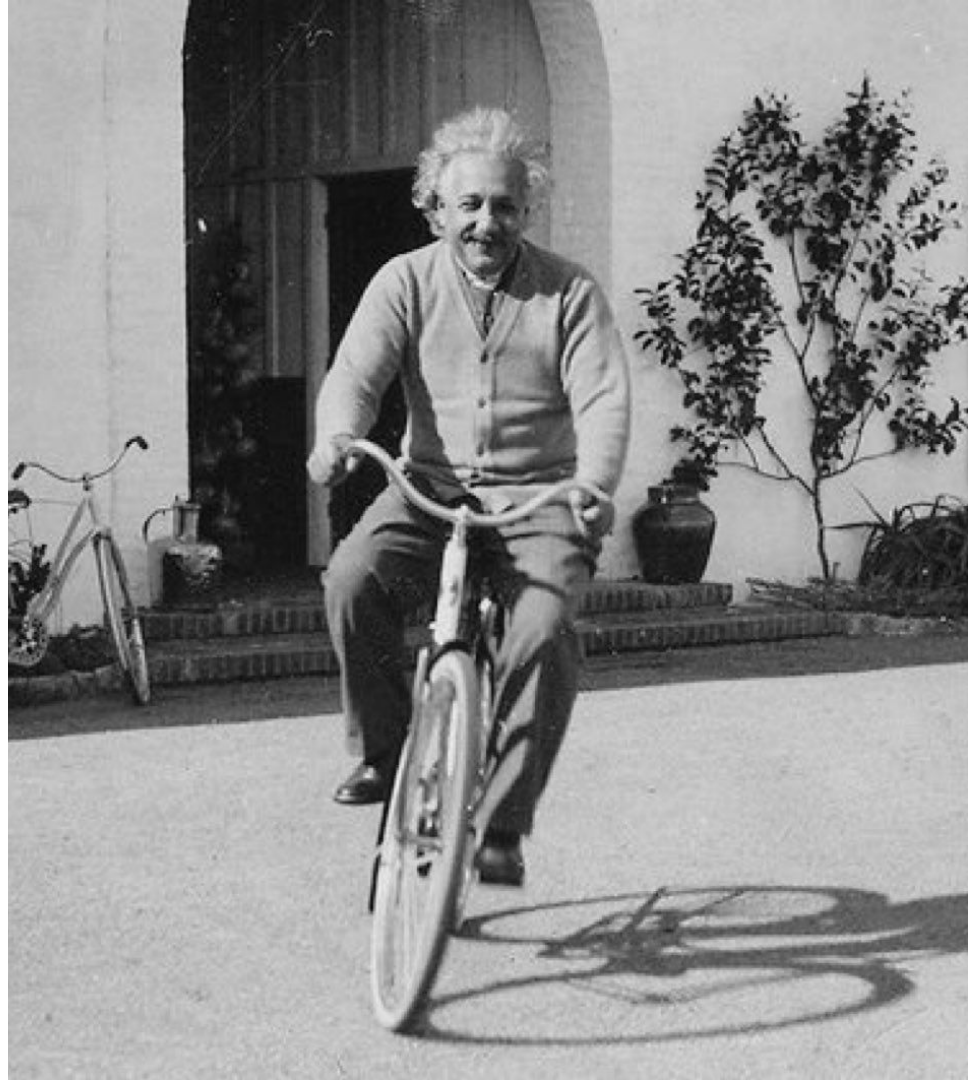
The Data Science Process



Joe Blitzstein and Hanspeter Pfister, created for the Harvard data science course <http://www.cs109.org/>

A General Pattern...
But it always starts with
a Question!

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 You Will Learn

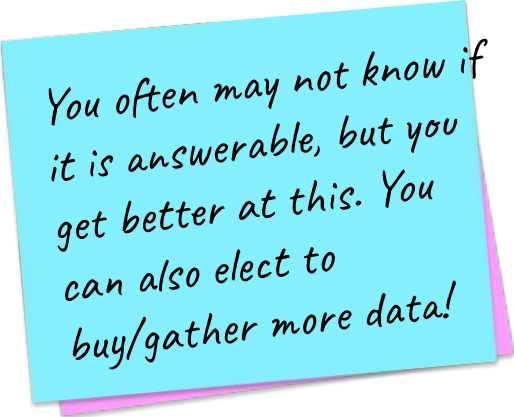
- ☒ Explain the data science process (high level)
- ☐ What are good and bad data science questions
- ☐ How to go from *general* to *specific* when asking questions
- ☐ Real world data science questions

So what makes a good
question?

(what about bad ones?)

Data Science questions should...

- **Be answerable with (available/attainable) data**
- Specify what's being measured
- Have relevance to someone or something
- Be specific



*You often may not know if
it is answerable, but you
get better at this. You
can also elect to
buy/gather more data!*

Own It

Internal data
from operations

METHOD

1



METHOD

2



Gather it

Manual collection or
primary research

METHOD

3



Borrow it

Partner data sources or
open data sources

METHOD

4



Pay for it

Third-party data that you
buy from data vendors and
aggregators

You need data...



Data Not Available

The requested data does not exist for
Feature Overview

Data Science questions should...

- Be answerable with (available/attainable) data
- **Specify what's being measured**
- Have relevance to someone or something
- Be specific

Did I mention?

BE SPECIFIC :)

Specifying what you're going to measure is important

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

- What can my data tell me about my business?
- What data should I continue to collect or cut?
- 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?

Key Performance Indicators

Definition and Examples

A quantifiable measure a company uses to determine how well it's meeting its operational and strategic goals.



A sales team might track **new revenue**



A customer support team might measure the **average on-hold time** for customers



A marketing group will look at the contribution of **marketing generated sales leads**



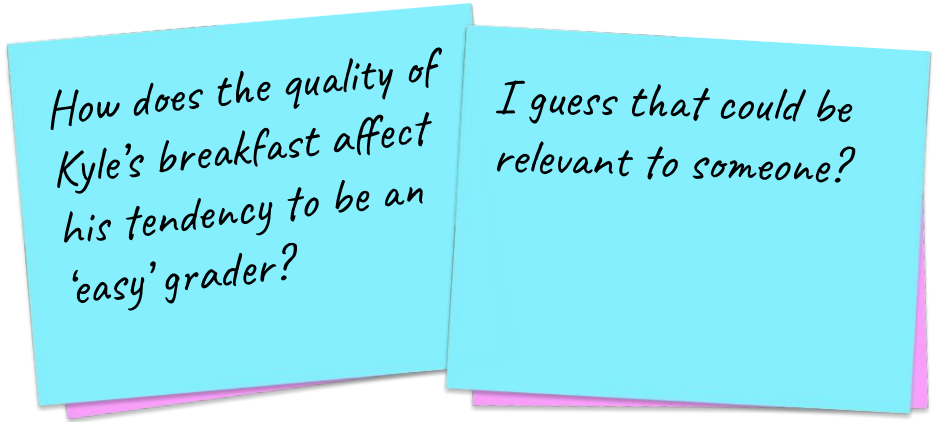
Human resources will look at **employee engagement**



Other areas of the business will look at the **efficiency of processes**

Data Science questions should...

- Be answerable with (available/attainable) data
- Specify what's being measured
- **Have relevance to someone or something**
- Be specific



*How does the quality of
Kyle's breakfast affect
his tendency to be an
'easy' grader?*

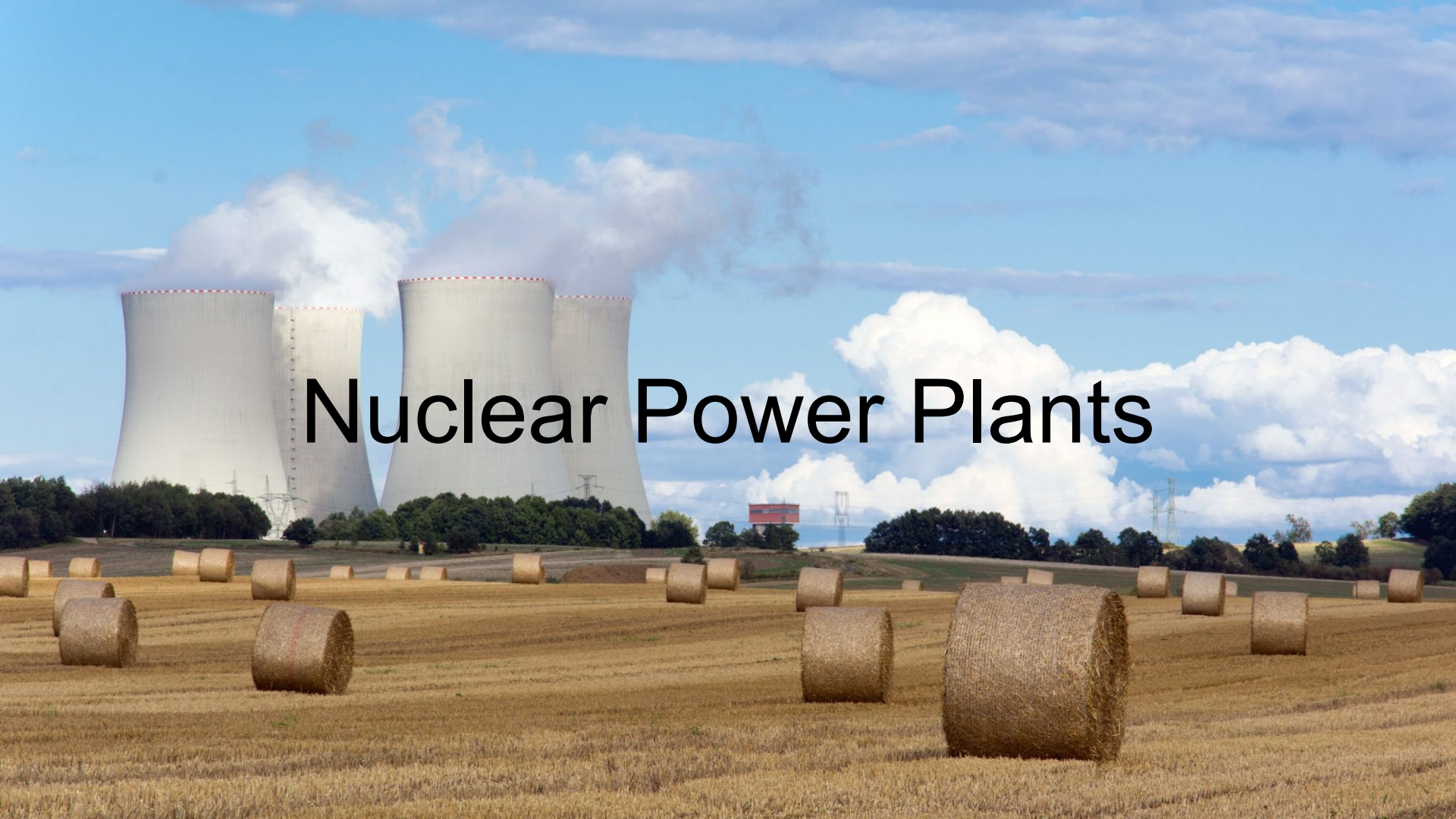
*I guess that could be
relevant to someone?*

So who or what are
stakeholders?

Less Grossman!



Nuclear Power Plants



Stakeholder 1:

.gov



Stakeholder 2: Executives





DAILY

EXPRESS

Wednesday April 30 1986 • 20p CDD in Eire

THE VOICE OF BRITAIN



Russia admits worst A-plant disaster ever

NUCLEAR NIGHTMARE IS HERE

- More than 3,000 reported dead
- Thousands more are doomed
- Help us plea goes to the West

By MICHAEL EVANS

AS MANY as 3,000 people could have died in the nightmare of the Russian nuclear disaster, Western diplomats in Moscow were reporting last night.

Ten thousand more could die from radiation and cancer.

One of the four reactors at the Chernobyl nuclear power complex 80 miles from Kiev is still burning.

Russia yesterday asked the West for help in dealing with the worst nuclear accident ever. Despite the fact that Tass, the official Soviet newsagency, is claiming that only two people died.

Soviet leaders, quoted in New York, put the figure at 2,100 dead. Another source in Kiev said up to 3,000 had died and 15,000 people were being evacuated from the area in hurriedly-constructed buses and lorries.

RADIATION

Many may already be under sentence of death. Medical experts say there could be 10,000 fatalities over the next 10 years as radiation-induced cancers take their toll.

Ireland appears to be safe. The danger of a high cloud of radiation which passed over Scandinavia is now being blown back across Russia—and over the North Pole towards America.

Water supplies are contaminated around Kiev.

Radiation-watchers at Loughborough in County Down, Harwell and Abingdon in Oxfordshire are all monitoring radioactivity levels.

Suicide squads on edge of hell: Pages 2 and 3



THE LAST FAREWELL

THE Queen, the Queen Mother and other Royals made their farewells to the Duchess of Windsor. Jean Bank reports on Page 6.



SEALED WITH A KISS

KISSERS from the stars of East-Enders as they celebrate winning the award of top BBC programme of the year. See Page 7.

Stakeholder 3:

The Media

Stakeholder 4:

The Public



Stakeholder 5: The Scientists





Stakeholder 6:
Investors

Data Science questions should...

- Be answerable with (available/attainable) data
- Specify what's being measured
- Have relevance to someone or something
- **Be specific**

Did I mention?

BE SPECIFIC :)

I did!

Nailing down the right question: **politics**

Too-vague: What impacts American politics?

... Does pop culture have an impact on American politics?

... Do American TV shows have an **impact** on American politics?

... Does South Park **affect** American politics?

... Is there a **relationship between** words in South Park episodes and American politics?

... Is there a relationship between the sentiment of political words in South Park and American politics?

Better: Is there a relationship between the sentiment of political words in South Park and America's presidential approval rating?

Nailing down the right question: **cause of death**

Too-vague: What gets attention in the news?

... Do terrorist attacks get reported too much?

... Is there a relationship between the number of people who die relative to the amount of media attention a story gets?

... What causes of death are over reported in the news relative to CDC death data? Underreported?

... Is there a relationship over time between cause of death terms in the NYT, The Guardian, and Google trends data relative to data from the CDC?

Nailing down the right question: **policing**

Too-vague: Why isn't police response time always the same?

... How can we improve police response time?

... Do crime levels and time of day affect response time?

... Where should police cars be stationed, accounting for crime levels and time of day, to make police response times equitable?

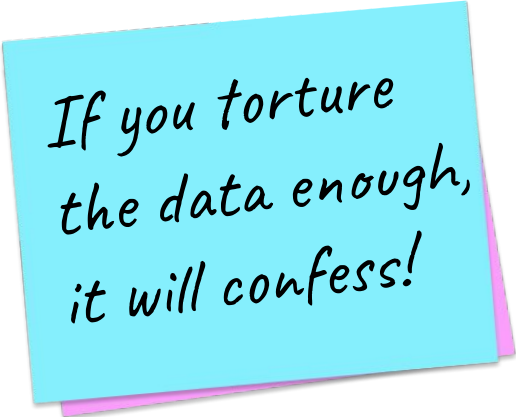
better: Where should police cars be stationed, accounting for crime levels and time of day, to make police response times equitable throughout San Diego?

Often Data availability leads
the questioning...

What You Will Learn

- ☒ Explain the data science process (high level)
- ☒ What are good and bad data science questions
- ☒ How to go from *general* to *specific* when asking questions
- ☐ Real world data science questions

Your question must fit the data.
You can't force the data to fit
your question.



*If you torture
the data enough,
it will confess!*



You're interested in learning more about age in US politics

Which of the following is the BEST data science question?



A

How old are
Congress members?

B

How many people
are in Congress
currently?

C

What is best about
US politics? What
is worst?

D

What should I learn
about US politics
age and where
should I learn that
information?

E

How has the
average age of
members in
Congress changed
over time?

Real World Questions

A Walmart VP asked, to increase target offering ad revenue, how can we best segment our customers across geographic areas within any given product?

Real World Questions

FiveThirtyEight asked is Uber drivers/trips increased Manhattan rush hour traffic volume?

Real World Questions

Google asked if we can forecast web traffic using time series analysis, to help manage server load/balance and resource allocation.

Real World Questions

A data scientist (johnny wales) asked if we can build a model and analysis to determine if print news articles are fake or legitimate.

Real World Questions

A book seller asked if there a way to determine the best ecommerce strategy for selling used books on amazon vs another site.

What You Will Learn

- ☑ Explain the data science process (high level)
- ☑ What are good and bad data science questions
- ☑ How to go from *general* to *specific* when asking questions
- ☑ Real world data science questions

Assignment 1

For your first assignment, you'll have to:

1. **Ask a DataSci Question**
2. Think about a Hypothesis
3. Background info/research
4. Where can data be found?
5. Ethics