Introduction to data science

Patrick Shafto

Department of Math and Computer Science

Plan for today

- Questions
- HW: Questions that you wish to ask of your data set. Specifically write:
    - What is the question?
    - What data will be used to answer the question?
    - Describe a second way to answer the same question?
    - What are assumptions and limitations of the approach?
    - What is the statement that you will be able to make after doing the analysis:

What is data science?

Why use the term rather than “statistics” or “data mining” or “analytics”?

Asking and answering questions

Asking and answering questions

- Some suggest that there are X types of questions data science can answer.
    - “It might surprise you, but there are only five questions that data science answers:
        - Is this A or B?
        - Is this weird?
        - How much – or – How many?
        - How is this organized?
        - What should I do next?”
- These people are wrong. More importantly, they are answering the wrong question.

Asking and answering questions

- Is this A or B? Is this weird? How much – or – How many? How is this organized? What should I do next?"
- These people are wrong. More importantly, they are answering the wrong question.
- This assumes that the only thing a data scientist can do is apply algorithms. (A very impoverished set, as well.)
- In reality, good data science should assist organizations in collecting data that answer the questions that they need to answer.

Types of questions

- Causal questions:
    - Why did this happen?
    - What causes what?
- Correlational questions:
    - What happens together?
- Predictive questions:
    - These are typically correlational questions (with some rare exceptions for temporal data)

Questions <=> Methods

- Answers to causal questions require that you conduct an experiment.
    - Simply analyzing available data will (almost) never answer such a question
- Correlational / predictive questions can be answered with whatever data are available, assuming
    - The relationship of interest is observed in the data in a way that the variables of interest can be put into correspondence
    - Missing data don't compromise one's ability to obtain a valid answer

Asking and answering questions

- Causal or correlational:
    - Is this A or B?
    - Is this weird?
    - How much – or – How many?
    - How is this organized?
    - What should I do next?

Good questions

- Are precise:
    - They can be answered with yes/no, a single number, or similar
- Can be answered with available (or collectable) data
    - To check, look for the answer to your question in the data that you have

Hurricane Frances was on its way, barreling across the Caribbean, threatening Florida's Atlantic coast. Residents made for higher ground. Far away, in Bentonville, Ark., executives at Wal-Mart Stores decided that the situation offered a great opportunity for one of their newest data-driven weapons, predictive technology.

A week before landfall, Wal-Mart's chief information officer, pressed her staff to come up with forecasts based on what had happened when Hurricane Charley struck. Backed by the trillions of bytes' worth of shopper history that is stored in Wal-Mart's data warehouse, she felt that the company could "start predicting what's going to happen, instead of waiting for it to happen," as she put it.

Why might they want to generate such predictions?

What assumptions are they making?

What might they try to predict?

What kind of data would they need to draw upon to make such predictions?

- **HW: Pick one data set, write notebook that downloads and cleans the data (for general purpose analyzing)**

- Netflix data
    - https://www.kaggle.com/netflix-inc/netflix-prize-data/data
- Yahoo finance
    - https://pypi.python.org/pypi/yahoo-finance
- IMF data
    - https://briandew.wordpress.com/2016/05/01/machine-reading-imf-data-data-retrieval-with-python/
- NYC open data
    - https://opendata.cityofnewyork.us/data/#datasetscategory
    - Examples:
        - http://blog.nycdatascience.com/student-works/r-shiny/noise-coming-case-study-nycs-311-noise-complaints/
        - http://blog.nycdatascience.com/student-works/new-york-city/

- Netflix data
    - https://www.kaggle.com/netflix-inc/netflix-prize-data/data
    - README

- Yahoo finance
  - https://pypi.python.org/pypi/yahoo-finance

Example: Yahoo! Inc. (YHOO)

```
>>> from yahoo_finance import Share
>>> yahoo = Share('YHOO')
>>> print yahoo.get_open()
'36.60'
>>> print yahoo.get_price()
'36.84'
>>> print yahoo.get_trade_datetime()
'2014-02-05 20:50:00 UTC+0000'
```

Available methods

- `get_price()`
- `get_change()`
- `get_percent_change()`
- `get_volume()`
- `get_prev_close()`
- `get_open()`
- `get_avg_daily_volume()`
- `get_stock_exchange()`
- `get_market_cap()`
- `get_book_value()`
- `get_ebitda()`
- `get_dividend_share()`
- `get_dividend_yield()`
- `get_earnings_share()`
- `get_days_high()`
- `get_days_low()`
- `get_year_high()`
- `get_year_low()`
- `get_50day_moving_avg()`
- `get_200day_moving_avg()`
- `get_price_earnings_ratio()`
- `get_price_earnings_growth_ratio()`
- `get_price_sales()`
- `get_price_book()`
- `get_short_ratio()`
- `get_trade_datetime()`
- `get_historical(start_date, end_date)`
- `get_info()`
- `get_name()`
- `refresh()`

- IMF data
    - https://www.imf.org/en/Data

- NYC open data
    - https://opendata.cityofnewyork.us/data/#datasetscategory
    - Examples:
        - http://blog.nycdatascience.com/student-works/r-shiny/noise-coming-case-study-nycs-311-noise-complaints/
        - http://blog.nycdatascience.com/student-works/new-york-city/

- NYC open data
    - https://data.cityofnewyork.us/Social-Services/311-Service-Requests-from-2010-to-Present/erm2-nwe9

- **HW: Pick one data set, write notebook that downloads and cleans the data (for general purpose analyzing)**

- Netflix data
    - https://www.kaggle.com/netflix-inc/netflix-prize-data/data
- Yahoo finance
    - https://pypi.python.org/pypi/yahoo-finance
- IMF data
    - https://briandew.wordpress.com/2016/05/01/machine-reading-imf-data-data-retrieval-with-python/
- NYC open data
    - https://opendata.cityofnewyork.us/data/#datasetscategory
    - Examples:
        - http://blog.nycdatascience.com/student-works/r-shiny/noise-coming-case-study-nycs-311-noise-complaints/
        - http://blog.nycdatascience.com/student-works/new-york-city/

- Questions that you wish to ask of your data set. Specifically write:
    - What is the question?
    - What data will be used to answer the question?
    - Describe a second way to answer the same question?
    - What are assumptions and limitations of the approach?
    - What is the statement that you will be able to make after doing the analysis:

HW

- Questions that you wish to ask of your data set. Specifically write:
    - What is the question?
    - What data will be used to answer the question?
    - Describe a second way to answer the same question?
    - What are assumptions and limitations of the approach?
    - What is the statement that you will be able to make after doing the analysis: