Data Science with Python

...

C'ville Data Science / PyCHO
December 2015

A few introductions…

Name: Patrick Harrison
Role: Data Scientist
Company: S&P Capital IQ and SNL

Current research interests... applications of data science to

- product recommendations
- personalized search
- document classification
- and more!

Academics background in Economics and Systems Engineering @ UVA

Wait, what was that company again?

SNL + S&P Capital IQ = S&P Capital IQ and SNL

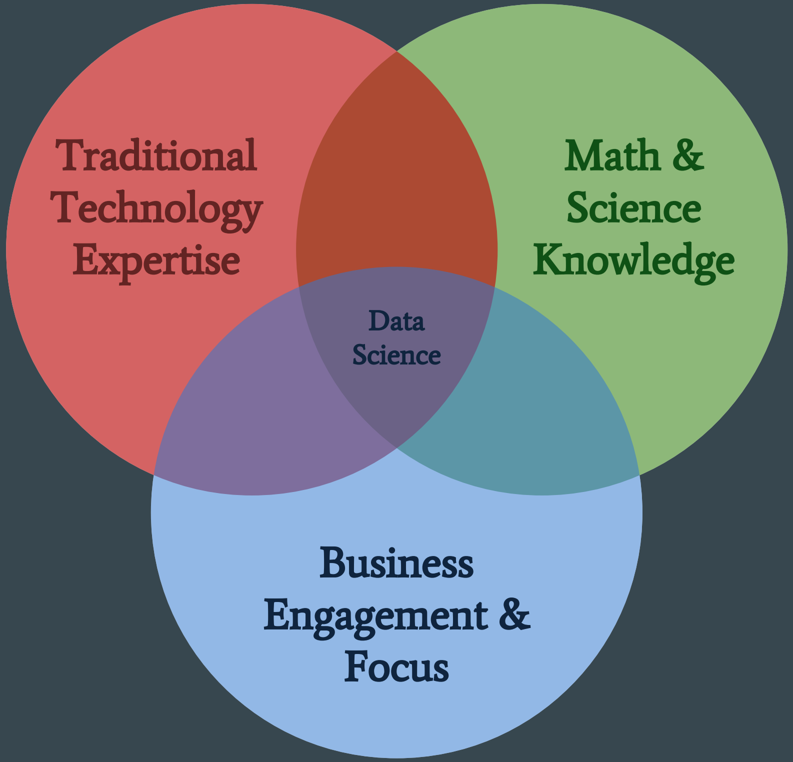
About S&P Capital IQ and SNL

- We are a global financial and business intelligence firm
- Data is our product — businesses and individuals subscribe to our data services
- We have some of the broadest, deepest, and most accurate datasets in the world about finance and business
- It’s a great place to be working on data science!

Data Science @ SPCIQ+SNL

python

DOMINO



Software Engineer · Data Engineer · **Data Scientist** · Applied Scientist · Research Scientist

What is data science anyway?

Data Science is the process of combining enriched source data with rigorous mathematical and scientific techniques to generate new, previously inaccessible insights that create value.

What is data science anyway?

Data Science is the process of combining **enriched source data** with rigorous mathematical and scientific techniques to generate new, previously inaccessible insights that create value.

What is data science anyway?

Data Science is the process of combining enriched source data with rigorous **mathematical and scientific techniques** to generate new, previously inaccessible insights that create value.

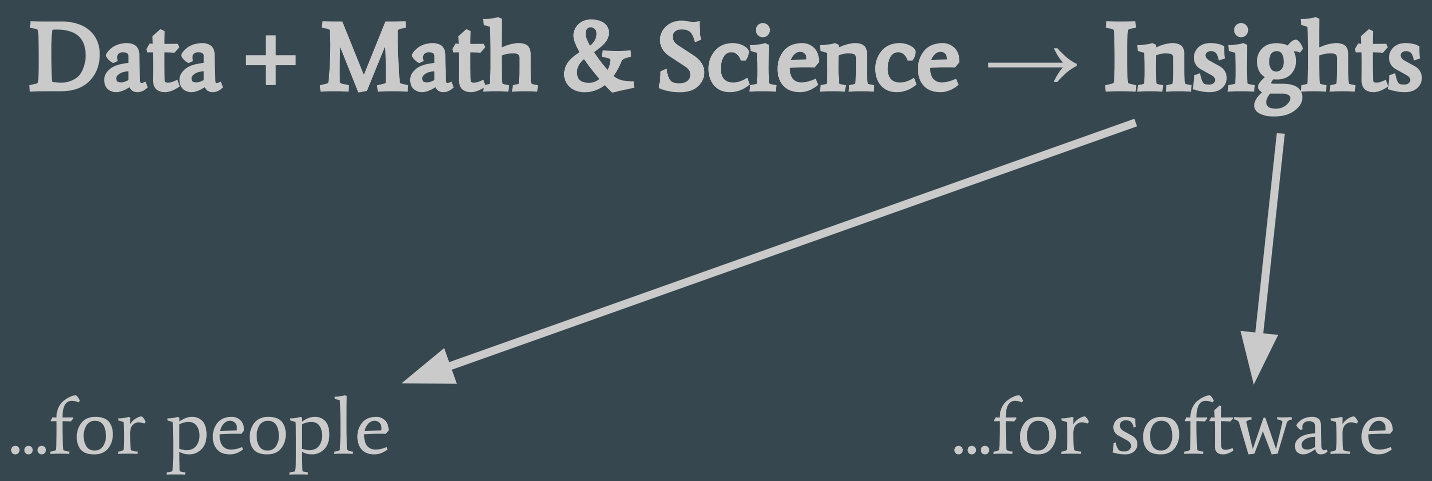
What is data science anyway?

Data Science is the process of combining enriched source data with rigorous mathematical and scientific techniques to generate new, previously inaccessible **insights that create value**.

What is data science anyway?

Data Science is the process of combining enriched source data with rigorous mathematical and scientific techniques to generate **new, previously inaccessible** insights that create value.

Or, put another way…



Data + Math & Science → Insights

…for people

…for software

Data + Math & Science → Insights

- Structured Data
- Time Series
- Text
- Social
- Images
- Audio
- Video
- Often helpful to enrich/link data across different datasets or domains
- …what if there's a lot?

Data + Math & Science → Insights

- Statistical Modeling
- Machine Learning
  - Classification
  - Regression
  - Clustering
- Natural Language Processing
- Graph Analytics
- Deep Learning
- Experimental Design
- Often projects will tap multiple techniques

Data + Math & Science → Insights

- Better personalize a product experience for individual users
- Detect the main themes of a document before reading it
- Predict which patients are most likely to be readmitted to the hospital after discharge
- Estimate how likely a customer is to leave your service, or respond to a marketing campaign, or upgrade a product, or...
- Write descriptive captions for images before seeing them
- ...

“there’s an app for that”

“there’s ~~an app~~ for that”

a Python package

A brief overview of the Python data ecosystem…

Doing Work

- Jupyter
- Domino
- yhat

Math & Analysis

- NumPy
- Pandas

Statistical Modeling

- SciPy
- StatsModels

Machine Learning

- scikit-learn
- Keras & Theano

NLP

- NLTK
- Gensim
- spaCy

Graph Analytics

- Networkx

Parallelism

- PySpark
- Joblib
- Dask

Visualization

- matplotlib
- Bokeh
- Seaborn

and much more…

DEMOS

Observations…

- If you need to do a common thing, it almost certainly already exists in Python
    - …or at least the high-level building blocks you need do
- Emerging data science development platforms such as Domino and yhat make some hard things easy for data scientists
- If you have gigabyte-scale data, you can get pretty far on modern machines
- Machine learning is pretty sweet