Python and Data Science (Live Coding)

Adam Richards and Elliot Cohen

09.13.2017

Why Data Science?

Well because of the data and because of the science

- Data science means different things to different people
  - Data engineering $\rightarrow$ data vis $\rightarrow$ predictive modeling
- Statistics, machine-learning, databases, web-development
- We are in an era of unprecedented data growth
  - And so few know how to effectively use data to gain insight
- Science is the proponent of truth through the communication of evidence

Why Data Science?

Well because of the data and because of the science

- Data science means different things to different people
  - Data engineering → data vis → predictive modeling
- Statistics, machine-learning, databases, web-development
- We are in an era of unprecedented data growth
  - And so few know how to effectively use data to gain insight
- Science is the proponent of truth through the communication of evidence

Companies need data scientists

The essential data science toolkit

Subject area mastery

- SQL and noSQL databases
- Associative statistics and hypothesis testing
- Unsupervised and supervised learning
- Data visualization
- Data products

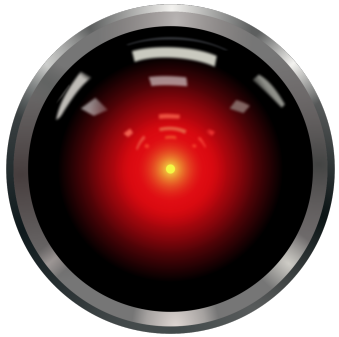
Programming mastery

Expert level proficiency in language that is useful for data science

Why Python?

- There are many languages of data science (Python, R,...)
- Ecosystem (NumPy, matplotlib, pandas)
- Readability, Flexibility
- Glue language
- Object-oriented and functional

What is NLP?

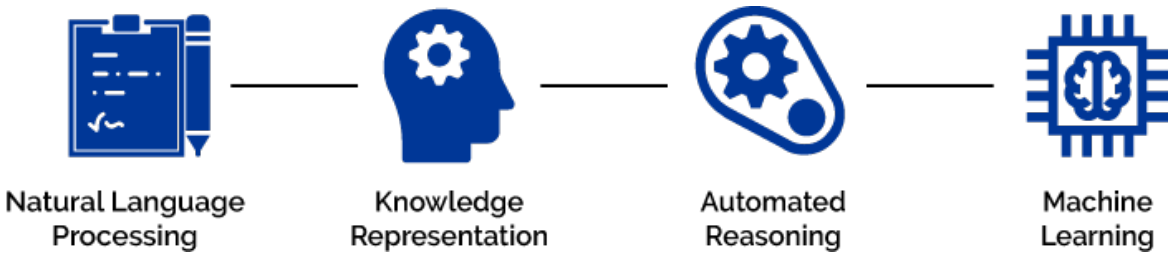


- Conversational Agents
  - Siri, Cortana, Google Now, Alexa
  - Talking to your car
  - Communicating with robots
- Machine Translation
  - Google Translate
  - Google's Neural Machine Translation
- Speech Recognition, Speech Synthesis
- Lexical Semantics, Sentiment Analysis
- Dialogue Systems, Question Answering

Conversation Agent Lecture - Dan Jurafsky

NLP and AI

The ultimate goal of NLP is to the fill the gap how the humans communicate (natural language) and what the computer understands (machine language).



Why Deep Learning Needed in NLP

- It uses a rule-based approach that represents Words as 'One-Hot' encoded vectors.
- Traditional method focuses on syntactic representation instead of semantic representation.
- Bag of words - classification model is unable to distinguish certain contexts.

http://www.datasciencecentral.com/profiles/blogs/overview-of-artificial-intelligence-and-role-of-natural-language

What is the meaning of this sentence?

I made her duck.

Challenges

Ambiguity

What does it mean when we say: 'I made her duck'

- I cooked waterfowl for her
- I cooked waterfowl belonging to her
- I created the (papier mache?) duck she owns
- I caused her to quickly lower her head or body
- I waved my magic wand and turned her into undifferentiated waterfowl

Other examples

- 'Court to try shooting defendant'
- 'Hospitals are sued by seven foot doctors'

This problem of determining which sense was meant by a specific word is formally known as word sense disambiguation

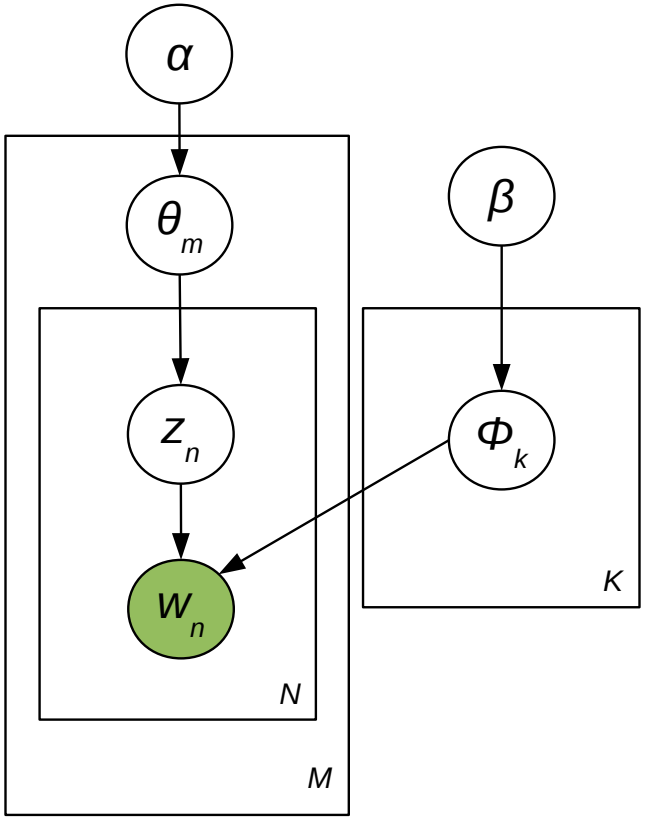
Other challenges include:

- Part of speech tagging
- Syntactic disambiguation (The I made her duck example)

Speech and Language Processing (Jurafsky and Martin)

Demo 1

Exploratory data analysis

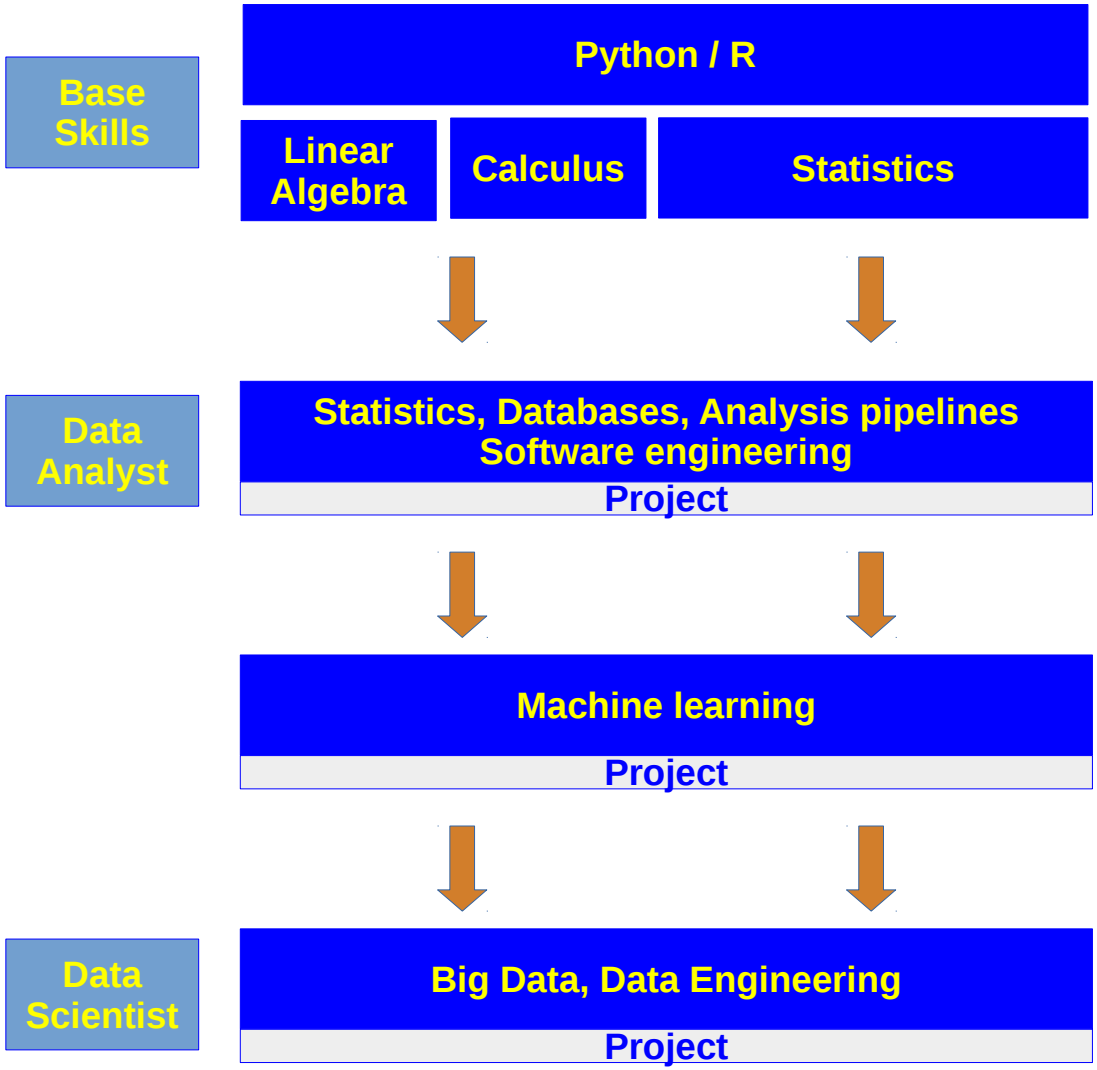

α
θ m
β
z n
Φ k
w n
N
K
M

Demo 2

Finding hidden topics

Overview

1. Version control based workflow
2. Jupyter notebook
3. Data types (lists, arrays, data frames)
4. EDA - working with data frames, plotting
5. Working with text (latent topics, clustering)
6. Data visualization


Base Skills
Python / R
Linear Algebra
Calculus
Statistics
Data Analyst
Statistics, Databases, Analysis pipelines
Software engineering
Project
Machine learning
Project
Data Scientist
Big Data, Data Engineering
Project