

# Day in Life of Data Scientist

Women in Big Data

Becky Tucker  
July 25, 2018

NETFLIX

# Agenda

- Presentation (10-15 minutes)
- Hands-on tutorial (45 minutes)

# Data Science in One Slide

- Data manipulation/cleaning: SQL, sometimes Spark/PySpark/Scala
- Statistics and Probability: hypothesis testing, statistical significance, power studies, probability distributions, Bayesian methods
- Metric development: LTV, churn, CAC, NPS, DAU
- Analytics: Tableau, Mode, Periscope
- Supervised Machine Learning: linear/logistic regression, tree-based methods, kernel-based methods, neural networks, genetic algorithms
- Unsupervised Machine Learning: clustering, PCA/ICA/SVD, LDA and other latent variable models, collaborative filtering, anomaly detection
- Reinforcement Learning: Q-learning, transfer learning, multi-armed bandits
- Modeling techniques: Monte Carlo simulations, Markov chains, survival models, growth models
- Specializations: Natural Language Processing, Causal Inference, A/B testing, robotics/automation
- Programming skills: Python/R, sometimes C++/C, procedural and/or object oriented, including version control (git)
- Productionalization of models: writing APIs, model hosting, working with software/algorithm engineers, scheduling data pipelines and automated model retraining
- Domain expertise: bias in data, social/business consequences of bias in predictions
- Problem definition
- Communication with business partners (technical and nontechnical): memos, visualizations, slides, dashboards, presentations
- Project management

# Data Science

Can be any or all of those things, depending on:

- Your company
- Your team
- Your interests



# A month in *my* life as a data scientist

- A typical project takes weeks/months.
- Project arc:
  - Start with the business questions
  - Make a technical plan
  - Create a proof of concept
  - Iterate/validate until you succeed or decide it can't be done
  - Communicate results back to stakeholders
  - Productionalize the model

# A Typical Data Science Project (for me)

- **Start with the business questions:**
  - What is the business need?
  - What are we trying to do/measure/predict?
  - Who will be using this prediction or model?
  - What is a good proof of concept?
  - What is our metric of success?
  - Who is responsible for validation?
  - What is the time frame?
  - What other teams are involved?

# A Typical Data Science Project (for me)

- **Make your technical plan:**
  - What data do we have available? How clean is it?
  - What techniques have others used to solve problems like this?
  - What tools should we use to do the data engineering, modeling, and productionalization?
  - What error metric are we optimizing?
  - If there are several possible techniques, which are the most promising?

# A Typical Data Science Project (for me)

- **Create a proof of concept/iterate/validate:**
  - **SQL** to get/manipulate data (provided in clean, usable tables by data engineers)
  - Prototyping typically done in Jupyter notebooks (**Python**) with standard **ML** packages
    - Scikit-learn
    - XGBoost
    - Keras
    - Pandas
  - Early results often passed back and forth with stakeholders via spreadsheets and documents
  - Debugging/modeling suggestions from discussions and presentations to other data scientists
  - Early visualization is not fancy: Plots in Python or a basic **Tableau** dashboard
    - Matplotlib
    - Seaborn



# A Typical Data Science Project (for me)

- **Productionalize the model:**

- Create an **API** (microservice) for other teams to call your model and get predictions
  - May involve software/data engineering teams, may not.
- Decide on a **Service Level Agreement**
  - How quickly does this need to be fixed if it breaks?
  - Who needs to be informed if the model/service changes?
  - What is my responsibility for maintenance or improvements?
- Document with good comments and clean code in an internally available code repo
  - **Code reviews!**
- Document results with documents/presentations
- Presentations/emails to socialize the results or a new model

# Hands on Exercise: Proof of Concept

- Go to <https://tinyurl.com/NFLXWIBD2018>
- Click on
- Download Zip
- Extract downloaded zip file
- Start the terminal and CD to the location where you extracted the downloaded zip file contents
- Type the following in your terminal to launch notebooks  
jupyter notebook

# THANK YOU!

NETFLIX