Machine Learning in the Trenches

Will Landecker
@thewill
Who am I?

• Reed College (math)

• PhD Computer Science under Melanie Mitchell

• Data scientist @ twitter

• Data scientist @ lyft
data science?
data science?

1. what is data science?

2. what do I do?

3. how to prepare?
circle + triangles
circle + triangles
euros, backward square root, 7
steam logo?
euros, backward square root, 7
circle + triangles
steam logo?
euros, backward square root, 7
circle + triangles
this is a river.
takeaway: people are excited about data science, but many of them have no idea what it is.
Technical things that data scientists are sometimes asked to do:

- Machine learning implementation & research
- Statistics
- Experiment analysis
- Analytics (counting things)
- Data visualization
- Building ETL pipelines; data engineering
- Software engineering
Technical things that data scientists are sometimes asked to do:

- Machine learning implementation & research
- Statistics
- Experiment analysis
- Analytics (counting things)
- Data visualization
- Building ETL pipelines; data engineering
- Software engineering
Questions to ask yourself when looking for a “more science-y” job

Is there data?
If not, nothing to analyze.

Is there a data infrastructure team?
If not, it will probably be me.

Are there other data scientists?
If not, less support and more road bumps.

Are there interesting opportunities for personal development & growth in the “less science-y” jobs?

Technical things that data scientists are sometimes asked to do:

- Machine learning implementation & research
- Statistics
- Experiment analysis
- Analytics (counting things)
- Data visualization
- Building ETL pipelines; data engineering
- Software engineering
Dirty secret: the majority of work most data scientists do isn't heavy duty ML, sophisticated algorithms, or cutting edge deep learning.

On good days, I'm happy to figure out what a good denominator is.

@oceankidbilly the "I count things" joke is played out, but seriously there's a lot of counting things.

@TomAugspurger don't forget installing things.
What I do

• Machine learning (implementation & research)
  - Choosing / modifying / creating learning algorithms & models
  - Feature engineering

• Finding sources of useful data (including labels!)

• Scientific communication

• Translating business goals into objectives functions that can be optimized

• Implementation (sometimes)
What I do

I want to eliminate spam.
What I do

I want to eliminate spam.

\[ \mathcal{L}(x, \lambda) = \frac{1}{2} \| x - x_0 \|_2^2 + \lambda (\Phi x - y). \]
• How can machine learning / statistics solve this problem?

• What are the features? What are the labels?

• What imbalances & biases exist in that data?

• What will make for a successful product? What objective function can I use as a proxy?

• Given everything above, what’s the best optimization algorithm for the job?

• What is the right operating point for this classifier?

• What are the shortcomings of this solution?

• How will it be implemented in production?
Awesome new cross-stitched addition to the office, courtesy of @ansate! So happy about this!!
How old is each user?
How old is each user?
Industry ML vs. Academic ML
(for me)

- Finding features & labels
- Very unbalanced data
- Semi supervised
- Communication
- Tools
- Data Privacy
Industry ML vs. Academic ML (for me)

• Finding features & labels
• Very unbalanced data
• Semi supervised
• Communication
• Tools
• Data Privacy

How old is each user?
Industry ML vs. Academic ML
(for me)

- Finding features & labels
- Very unbalanced data
- Semi supervised
- Communication
- Tools
- Data Privacy

- AUC, F1, Precision @ X% Recall, etc
- How does the metric of success translate to the loss function & the optimization algorithm?
- Sampling, weighting, calibrating
Industry ML vs. Academic ML
(for me)

- Finding features & labels
- Very unbalanced data
- Semi supervised
- Communication
- Tools
- Data Privacy

- Unsupervised feature engineering
Industry ML vs. Academic ML (for me)

- Finding features & labels
- Very unbalanced data
- Semi supervised
- Communication
- Tools
- Data Privacy

- Expertise != attacking others (teach others without being a jerk)
- Advocate for the data
Industry ML vs. Academic ML
(for me)

• Finding features & labels
• Very unbalanced data
• Semi supervised
• Communication

• Tools

• Data Privacy

- Python: scikit-learn, PuLP, tensorflow,…
- Scalding / Spark / Hadoop
- Databases
Industry ML vs. Academic ML (for me)

- Finding features & labels
- Very unbalanced data
- Semi supervised
- Communication
- Tools
- Data Privacy
Technical subjects for data science interviews

Machine learning (classifiers & learning algorithms; semi / un-supervised; loss functions; metrics)

• Probability theory, quantifying uncertainty (expected values; variance)

• Statistical inference (causal inference; experiment design & analysis; confidence intervals; estimation)

• Optimization (objective functions; numeric methods; convexity; (in)stability; discrete optimization)

• Algorithms, coding

• **Bonus:** the linear / logistic regression interview
Non-technical subjects for data science interviews

- Communication, empathy, compassion
- Passion for what the company does
- Ethics
- Broadly: *am I excited about working with this person?*
Resources:
Cosma Shalizi’s textbook
http://stat.cmu.edu/~cshalizi/ADAfEPoV
Erin Shellman’s blog
http://www.erinshellman.com/
Trey Causey’s blog
treycausey.com
Podcast:
Partially Derivative

Thanks for listening. Questions?