Executive Briefing

Why managing machines is harder than you think

Peter Skomoroch - @peteskomoroch

Strata Data Conference, London - May 1, 2019

Background: Machine Learning & Data Products

Peter Skomoroch
@peteskomoroch

- Co-Founder and CEO of SkipFlag, Enterprise AI startup acquired in 2018 by Workday
- 18+ years building machine learning products
- Principal Data Scientist, ran Data Products team at LinkedIn. ML & Search at MIT, AOL, ProfitLogic
- Co-Host of O'Reilly AI Bots Podcast, Startup Advisor

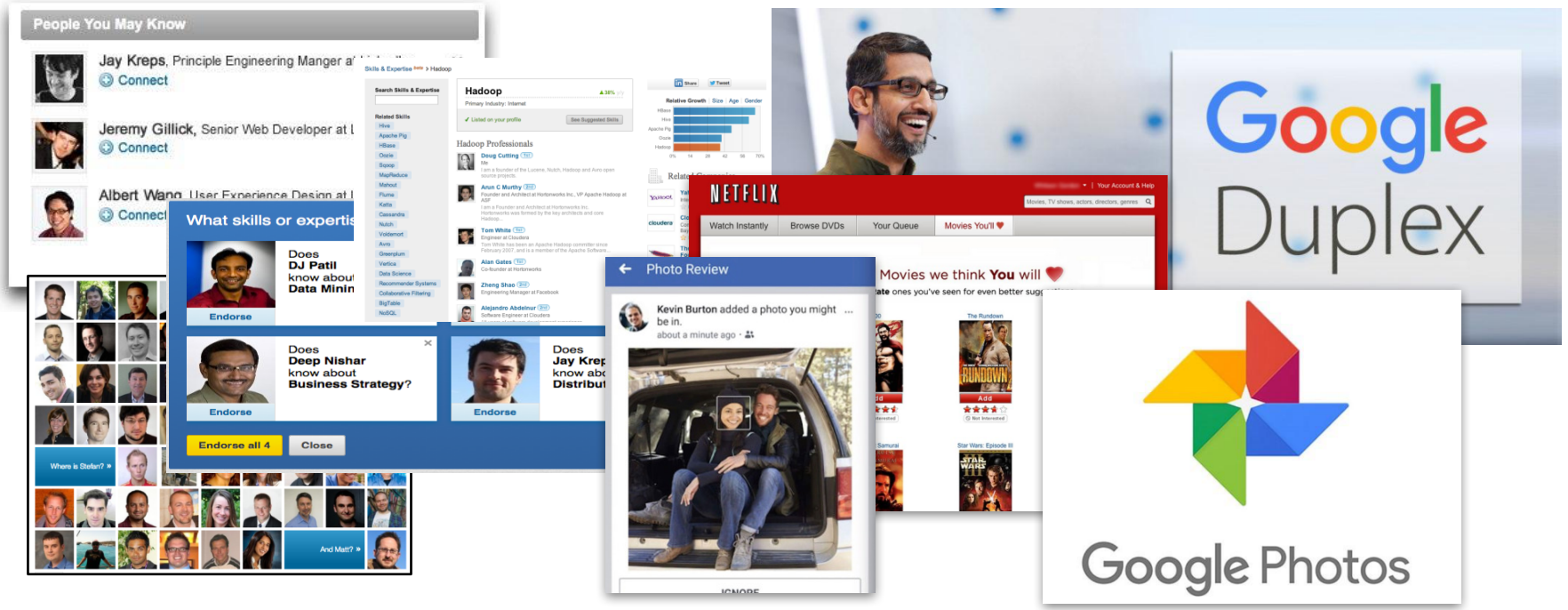
workday. SKIPFLAG LinkedIn AOL MIT

Better, Faster Decisions at Scale

- Machine learning drove massive growth at consumer internet companies over the last decade
- A wave of AI startups and vertical machine learning applications have emerged across other industries
- For many problems, machine learning makes better, faster, and more repeatable decisions at scale
- Amazon, Google, and Microsoft are now re-organizing themselves around AI

Data Products

Automated systems that collect and learn from data to make user facing decisions with machine learning



Machine Learning Projects are Hard

- The transition to machine learning will be about 100x harder than the transition to mobile
- Companies that adopt an experimental culture can still succeed
- Some of the biggest challenges are organizational, not technical
- Data driven companies like Google and Facebook have a strategic advantage building ML products based on their data & compute assets, large user population, tracking & instrumentation, and AI talent

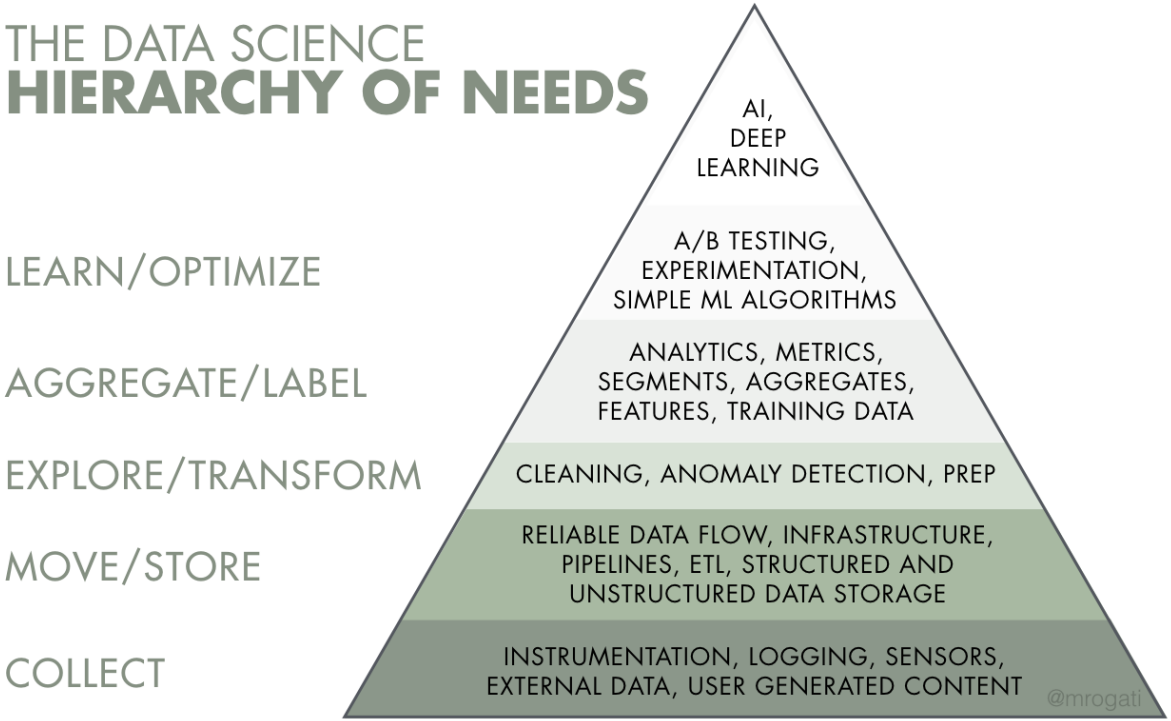
Experimental Culture

- Machine Learning shifts engineering from a deterministic process to a probabilistic one
- Take intelligent risks
- Most successful ML products are experiments at massive scale
- Companies driven by analytics and experimental insights are more likely to succeed

If you only do things where you know the answer in advance, your company goes away.

Jeff Bezos
Founder, Chairman & CEO of Amazon.com

Data Pipelines & Analytics Before AI



Credit: @mrogati

ML Algorithms Need Lots of Labelled Data

```
<p>I saw a sweet <a href="https://en.wikipedia.org/wiki/Ford_Mustang">Mustang</a>
on the highway yesterday.</p>
```

```
<p>The <a href="https://en.wikipedia.org/wiki/Mustang">mustang</a> is a free-
roaming horse of the American West.</p>
```

Use anchor links on a giant Web crawl for supporting evidence

Common Crawl

Common Crawl: ~4B pages monthly

Combined Pools of Data Give Better Results

https://www.flickr.com/photos/nakrnsm/3814916578

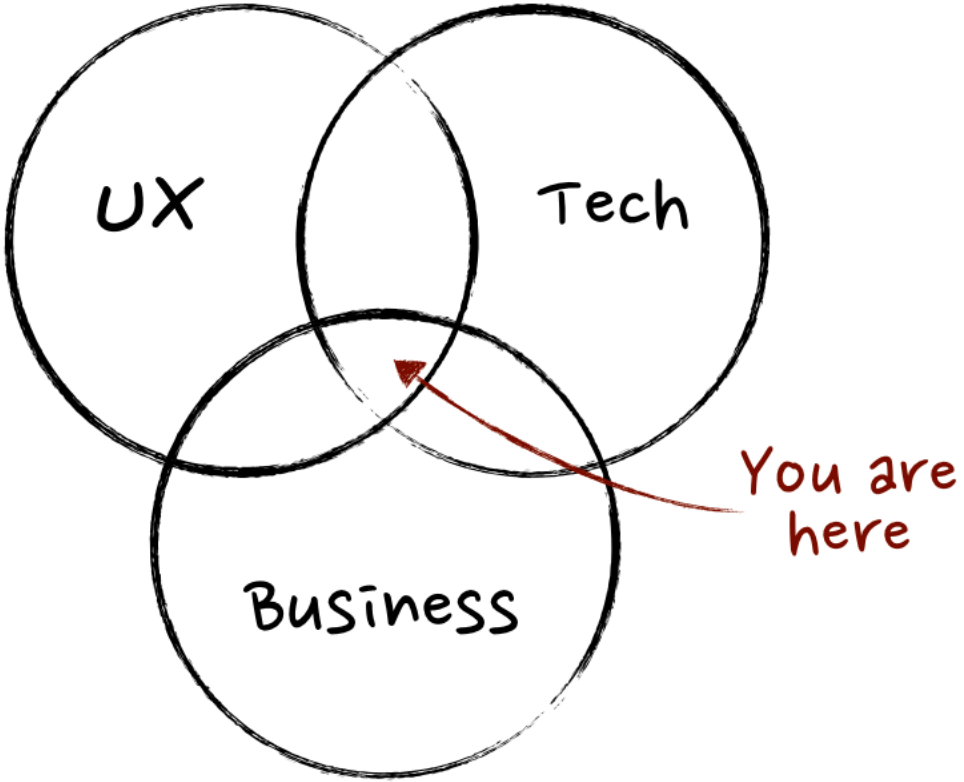
- Learning patterns across large numbers of customers is the power behind recommendations from companies like Amazon and Netflix
- The more precise or nuanced a prediction, the more data will need to be pooled
- You need **large amounts** of **labelled** training data
- Transfer learning may help push these limits further

Democratize Data Access

- Allow teams across your company to combine real data to improve their product areas, design with data, and discover new insights
- Share derived data and input features for ML models across teams
- At LinkedIn we had a rich repository of signals like connection strength, inferred skills, and other datasets that greatly accelerated new product development
- Empower small teams to build things quickly and compound returns on feature engineering & derived data

See https://www.confluent.io/ebook/i-heart-logs-event-data-stream-processing-and-data-integration/

Product Management for Machine Learning



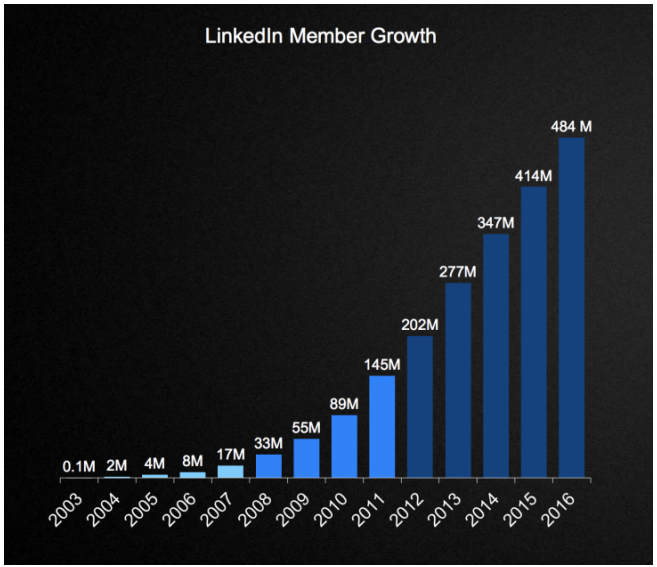
- A Data Product Manager (PM) has core product skills (strategy, roadmaps, prioritization, etc.) along with an intuitive grasp of ML
- They help identify and prioritize the highest value applications for machine learning and do what it takes to make them successful

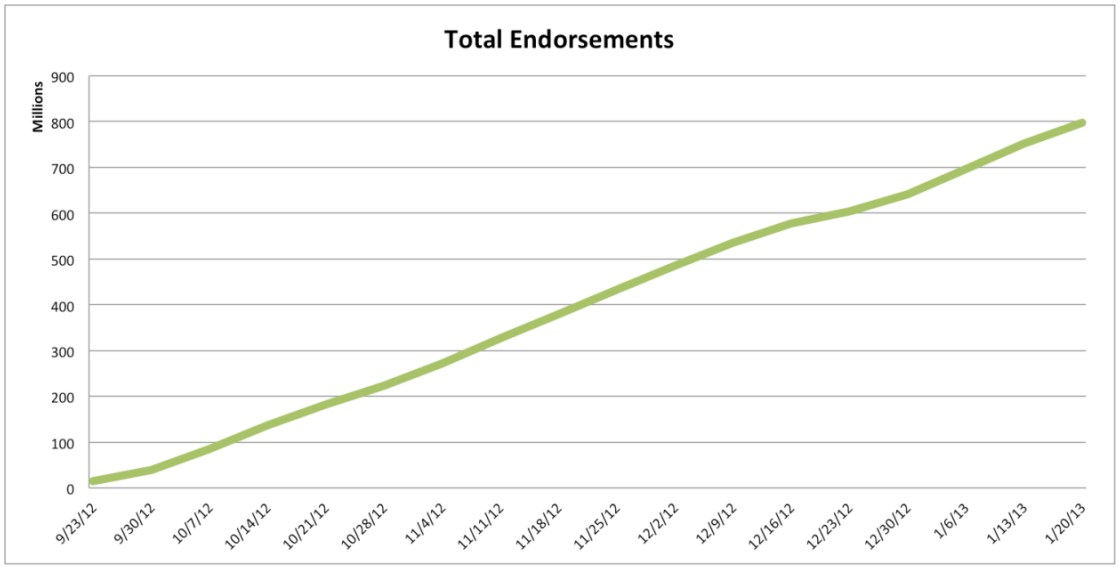
Image source: Martin Eriksson https://www.mindtheproduct.com/2011/10/what-exactly-is-a-product-manager/

Good ML Product Managers Have Data Expertise

- Know the difference between easy, hard, and impossible machine learning problems
- Even if something is feasible from a machine learning perspective, the level of effort may not justify building the feature
- Know your company's data inside and out including quality issues, limitations, biases, and gaps that need to be addressed
- Develop an intuitive understanding of your company's data and how it can be used to solve customer problems

Apply ML to a Metric the Business Cares About





Machine Learning Product Development

1. Verify you are solving the right **problem**
2. Theory + **model design** (in parallel with UI design)
3. **Data collection**, labelling, and cleaning
4. **Feature engineering**, model training, offline validation
5. **Model deployment**, monitoring & large scale training
    - Iterate: repeat process, refine live model & improve
    - 80% of effort and gains come from iterations after shipping v 1.0
    - Use derived data from the system to build new products

ML Adds Uncertainty to Product Roadmaps

- PMs are often uncomfortable with expensive ideas that have an uncertain probability of success
- Many organizations will struggle to justify the expense of projects that require significant research investment upfront
- Some ML products may need to be split into time boxed projects that get to market in a shorter time frame
- What can you productize now vs. much later on?
- Keep track of dependencies on other teams and have a “Plan B”

Data Quality & Standardization

- Guide user input when you can
- Use auto suggest fields
- Validate user inputs, emails
- Collect user tags, votes, ratings
- Track impressions, queries, clicks
- Sessionize logs
- Disambiguate and annotate entities (company names, locations, etc.)

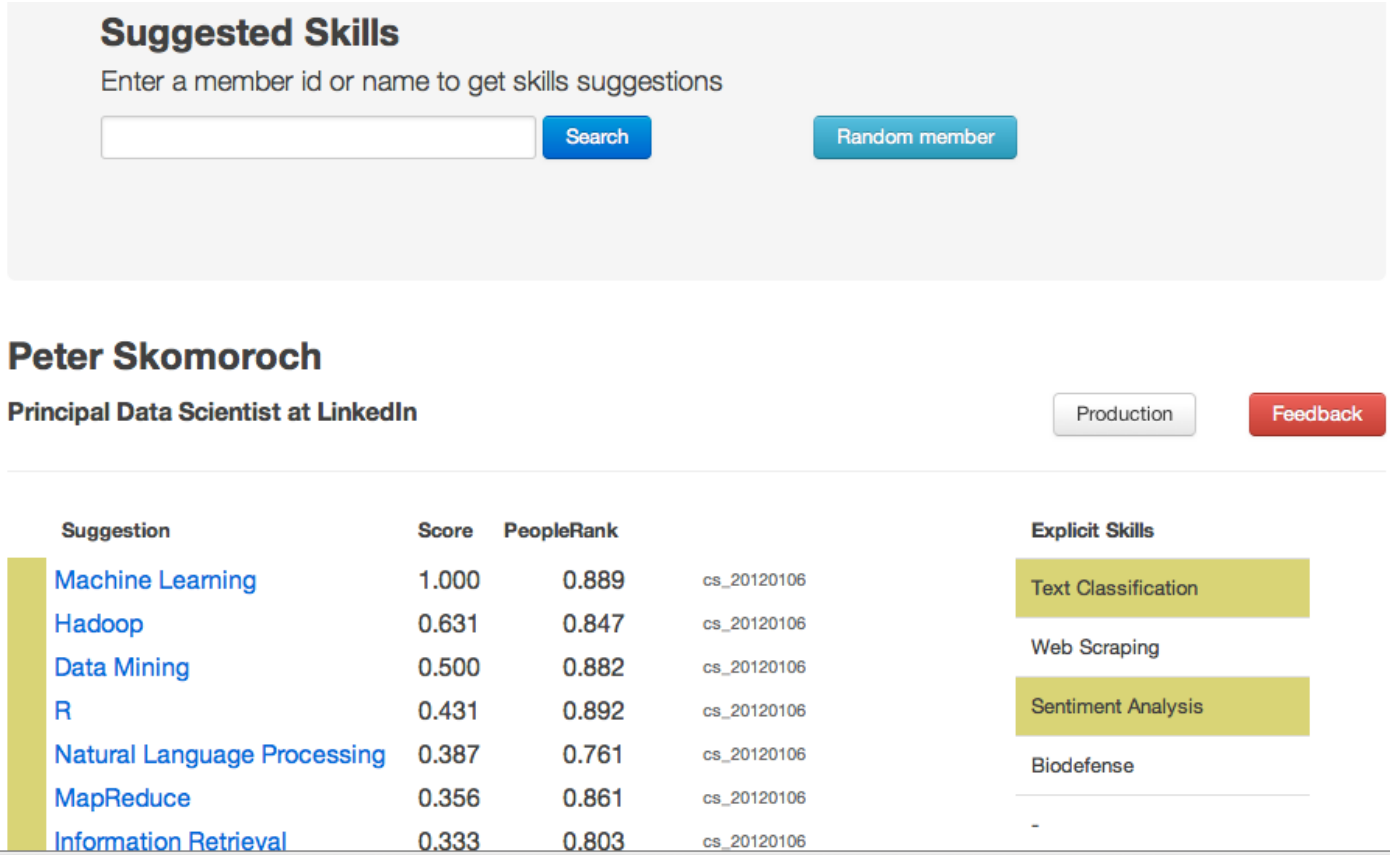
Every single company I've worked at and talked to has the same problem without a single exception so far — poor data quality, especially tracking data

Ruslan Belkin
VP of Engineering, Salesforce.com

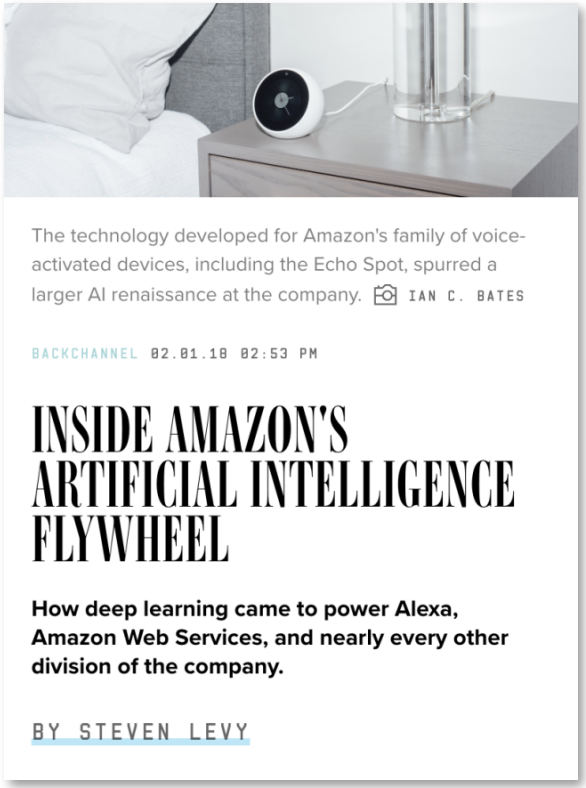
Testing Machine Learning Products

- Algorithm work that drags on without integration in the product where it can be seen and tested by real users is risky
- Ship a complete MVP in production ASAP, benchmark, and iterate
- Beware unintended consequences from seemingly small product changes
- Remember the prototype is not the product - see what happens when you use a more realistic data set or scale up your inputs
- Real world data changes over time, ensure your model tests and benchmarks keep up with changes in underlying data
- Machine learning systems tend to fail in unexpected ways

Look at Your Input Data & Prediction Errors



Flywheel Effects & Data Products



The technology developed for Amazon's family of voice-activated devices, including the Echo Spot, spurred a larger AI renaissance at the company. IAN C. BATES

BACKCHANNEL 02.01.18 02:53 PM

INSIDE AMAZON'S ARTIFICIAL INTELLIGENCE FLYWHEEL

How deep learning came to power Alexa, Amazon Web Services, and nearly every other division of the company.

BY STEVEN LEVY

- Users generate data as a side effect of using most software products
- That data in turn, can improve the product's algorithms and enable new types of recommendations, leading to more data
- These "Flywheels" get better the more customers use them leading to unique competitive moats
- This works well in platforms, networks or marketplaces where value compounds

* https://medium.freecodecamp.org/the-business-implications-of-machine-learning-11480b99184d

Final Thoughts

- Machine learning products are hard to build, but within reach of teams who invest in data infrastructure
- Some of the biggest challenges are organizational, not technical
- Good product leaders are a key factor in shipping successful ML products
- Find a machine learning application with a direct connection to a metric your organization values and ship it

Send me questions! **@peteskomoroch**

Q&A / Discussion