Sense and Sensibility in Data Science

Vicki Boykis | CapTech Consulting | @vboykis

About Me

{ "At": "CapTech. Others Talk, We Listen. "

"Data Science": ["Python", "R"],

"Data Engineering": ["AWS", "Spark"],

"Industries": ["Retail", "International Trade",

"Telecom", "Healthcare", "Finance"] }

Agenda

1. Hyped problems
2. Real problems
3. How to solve real problems

“Passion is inversely proportional to the amount of real information available.”

-Benford’s Law

Things we worry about but shouldn't

1. The end of human control over decisions
2. Outsourcing of human jobs
3. Robots killing us all

Hype: The end of human decision-making

- “Over half of all jobs will be automated within 10-20 years”
- " In the US alone, between 39 and 73 million jobs stand to be automated — making up around a third of the total workforce."

Source: The future of employment: How susceptible are jobs to computerization?
Frey and Osborne, 2013

Reality: The hard problems are still human problems

- Lunar Module Descent – Manual Mode
- GrubHub – Creating a data dictionary of foods
- Barnes Foundation
- The outsourcing study itself

Reality: The hard problems are still human problems

In this paper, we address the question: how susceptible are jobs to computerisation? Doing so, we build on the existing literature in two ways. First, drawing upon recent advances in Machine Learning (ML) and Mobile Robotics (MR), we develop a novel methodology to categorise occupations according to their susceptibility to computerisation.[1] Second, we implement this methodology to estimate the probability of computerisation for 702 detailed occupations, and examine expected impacts of future computerisation on US labour market outcomes.

Source: The future of employment: How susceptible are jobs to computerization? Frey and Osborne, 2013

Hype: Self-driving cars are going to radically change our society

- Elon Musk: We are two years away from sleeping in our cars
- "Autonomous cars are going to change everything"
  - Land use
  - Commuting marginal costs

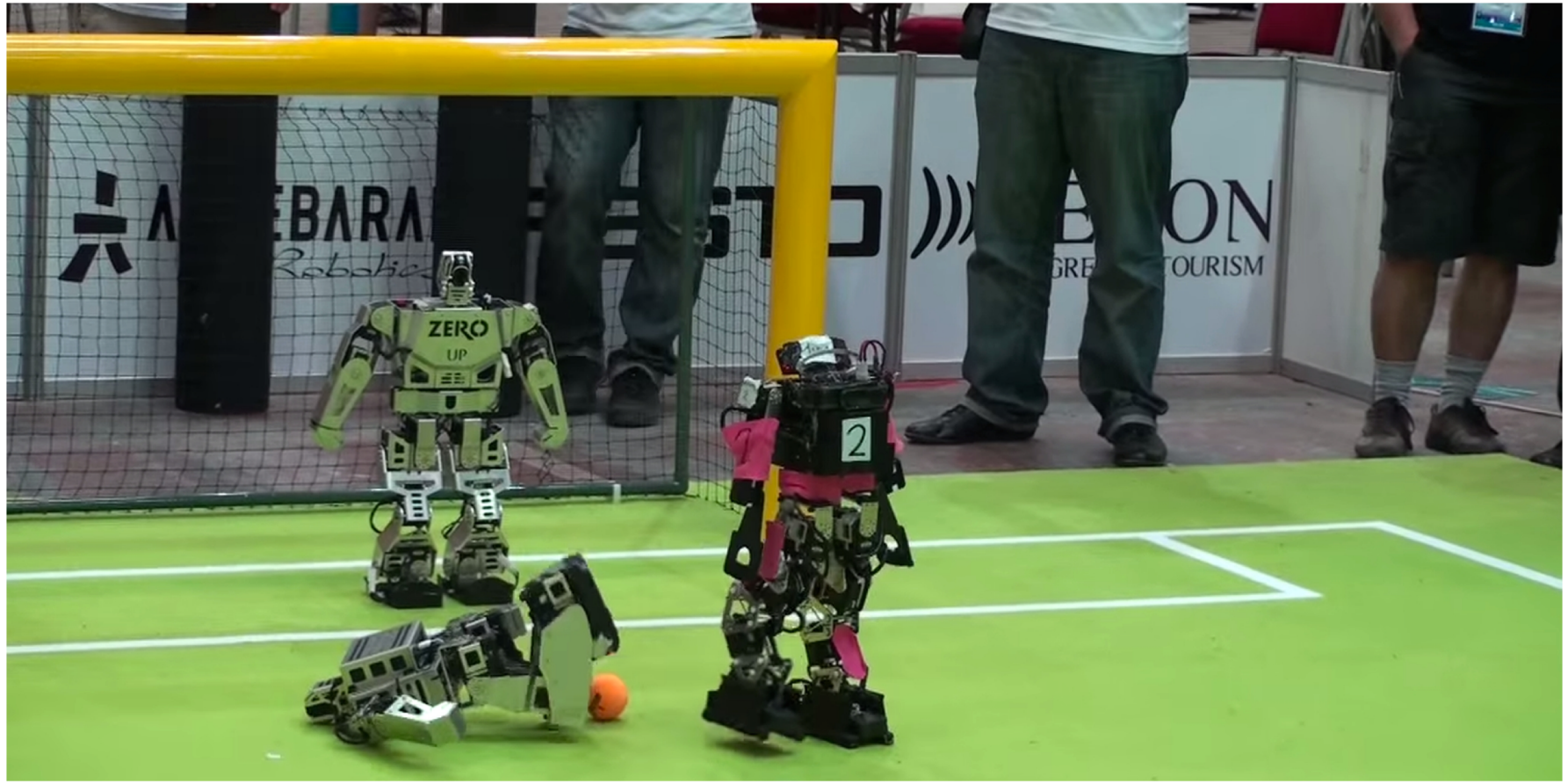
Reality: We still don't have enough data for self-driving cars to be accurate

- Self-driving cars need a lot of "training data" to understand what they're seeing.
- Training a neural network that encompasses a ton of features takes a lot of time:

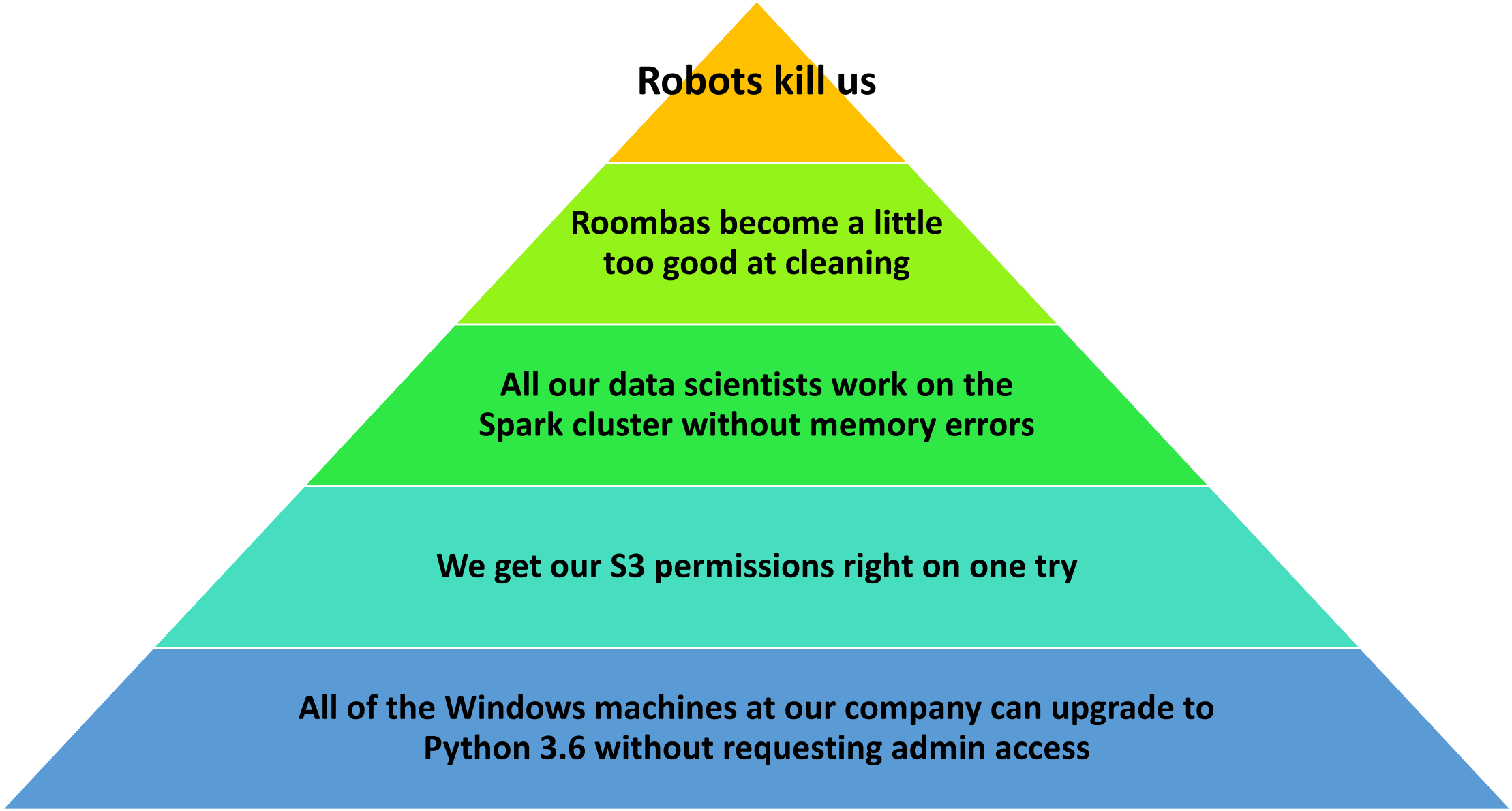
```
50 epochs * 104 TB of data / 19 MB/s = 9.1 years to
train an Inception-v3-like network once.
6.2 years for a ResNet-50-like network.
1.2 years for an AlexNet-like network.
```

Hype: Robots will kill us

Reality: AGI is still a long time away



What are people on the ground worried about?



The real problems in data work

1. Moving and keeping data in the cloud
2. Security and data privacy
3. Model intent

Problem: You are a large, traditional book retailer. You've been doing well in brick and mortar sales but now you need to compete against

Solution: Deep Learning

Deep Neural Networks for YouTube Recommendations

Paul Covington, Jay Adams, Emre Sargin
Google
Mountain View, CA
{pcovington, jka, msargin}@google.com

ABSTRACT

YouTube represents one of the largest scale and most sophisticated industrial recommendation systems in existence. In this paper, we describe the system at a high level and focus on the dramatic performance improvements brought by deep learning. The paper is split according to the classic two-stage information retrieval dichotomy: first, we detail a deep candidate generation model and then describe a separate deep ranking model. We also provide practical lessons and insights derived from designing, iterating and maintaining a massive recommendation system with enormous user-facing impact.

Keywords

recommender system; deep learning; scalability

1. INTRODUCTION

YouTube is the world's largest platform for creating, sharing and discovering video content. YouTube recommendations are responsible for helping more than a billion users discover personalized content from an ever-growing corpus of videos. In this paper we will focus on the immense impact deep learning has recently had on the YouTube video recommendations system. Figure 1 illustrates the recommendations on the YouTube mobile app home.

Recommending YouTube videos is extremely challenging from three major perspectives:

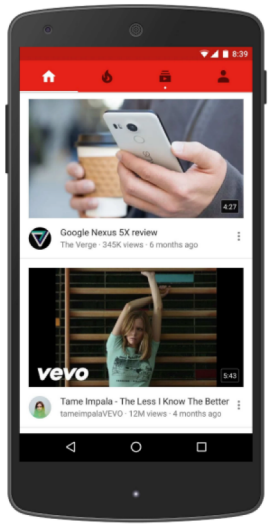


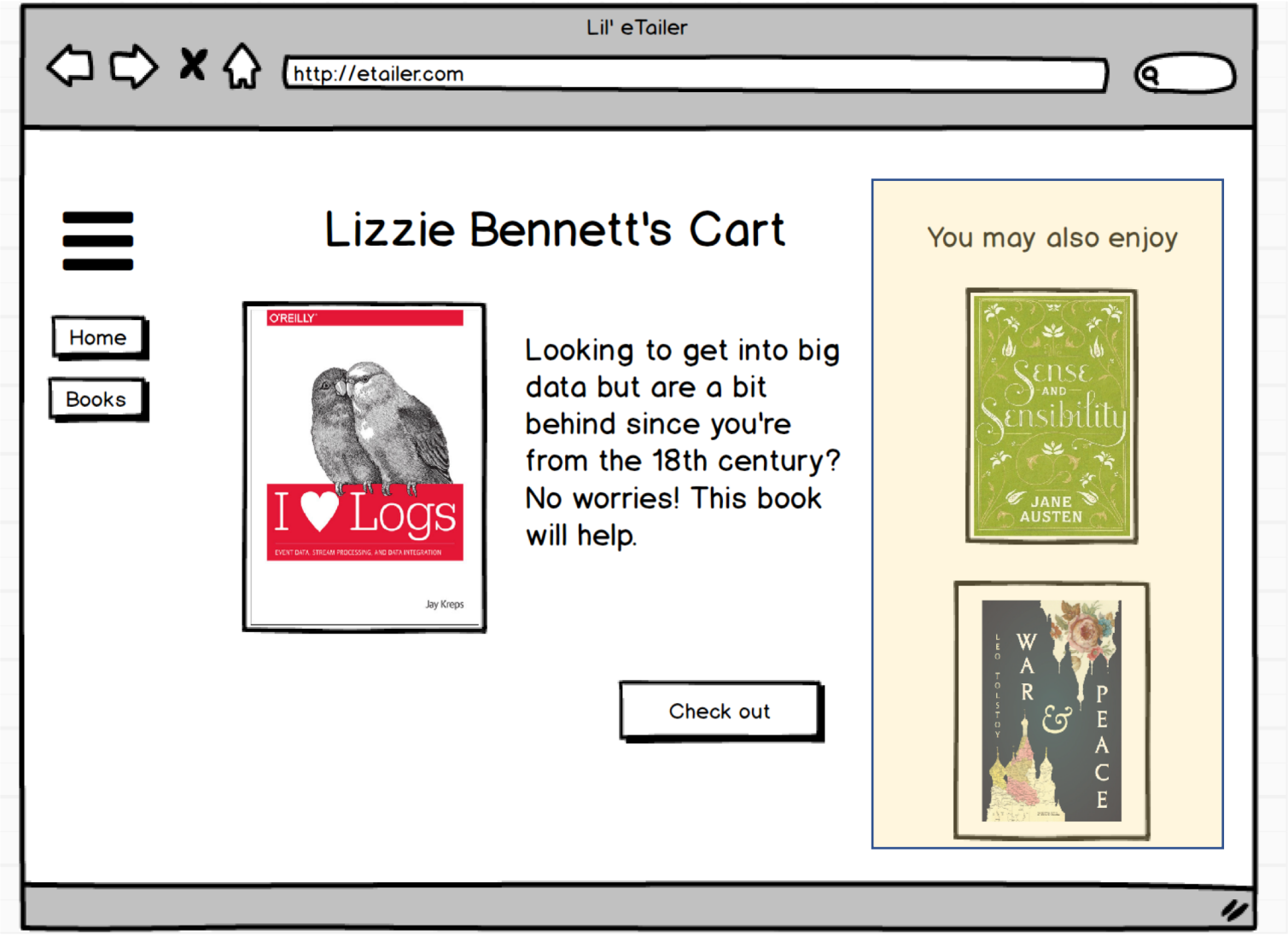
Figure 1: Recommendations displayed on YouTube mobile app home.

with well-established videos can be understood from an exploration/exploitation perspective.

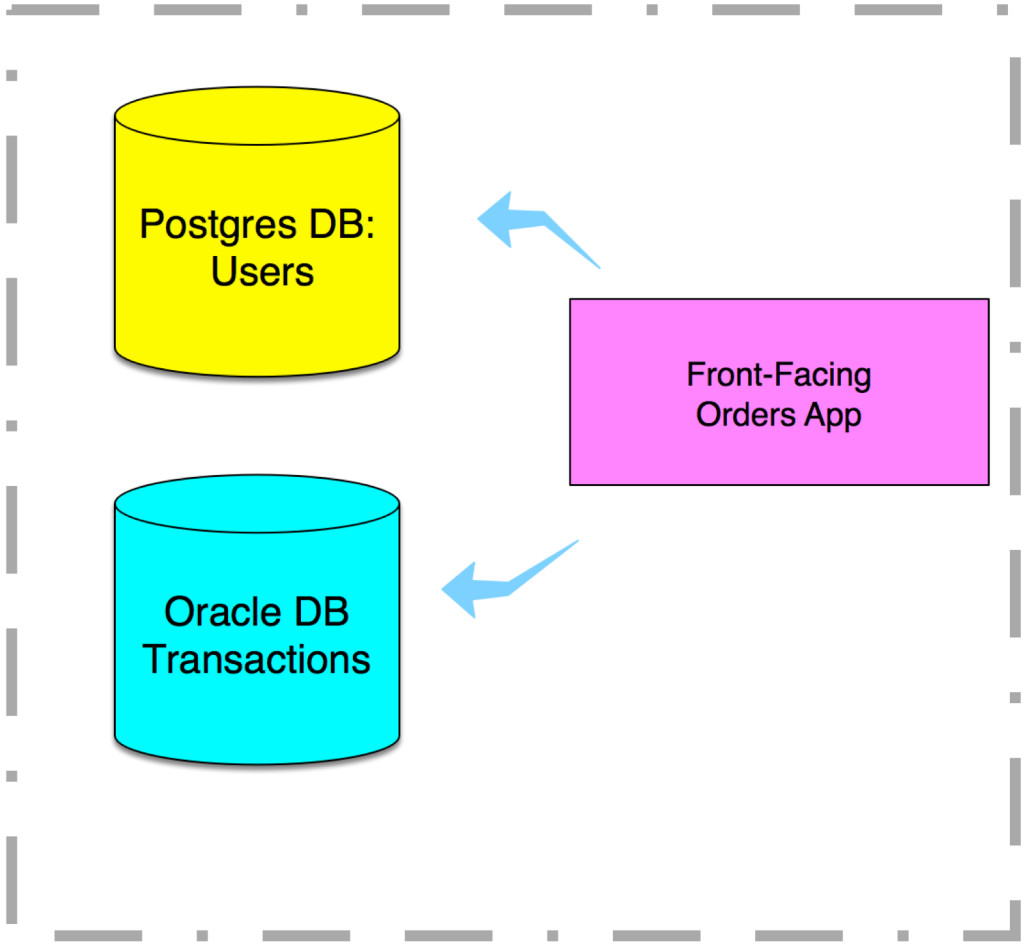
- *Noise:* Historical user behavior on YouTube is inher-

HarooBotter

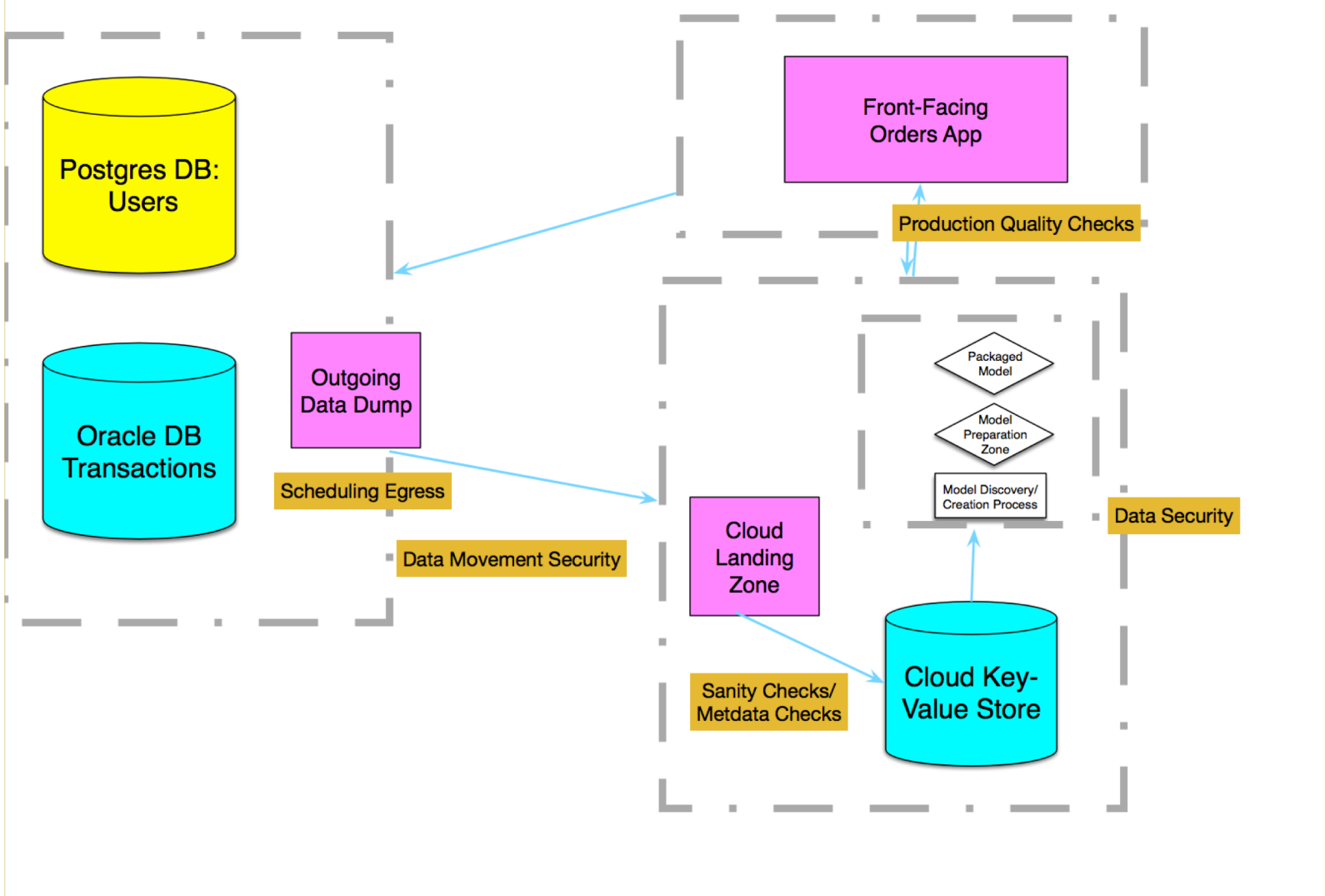
A sorting hat for your book preferences



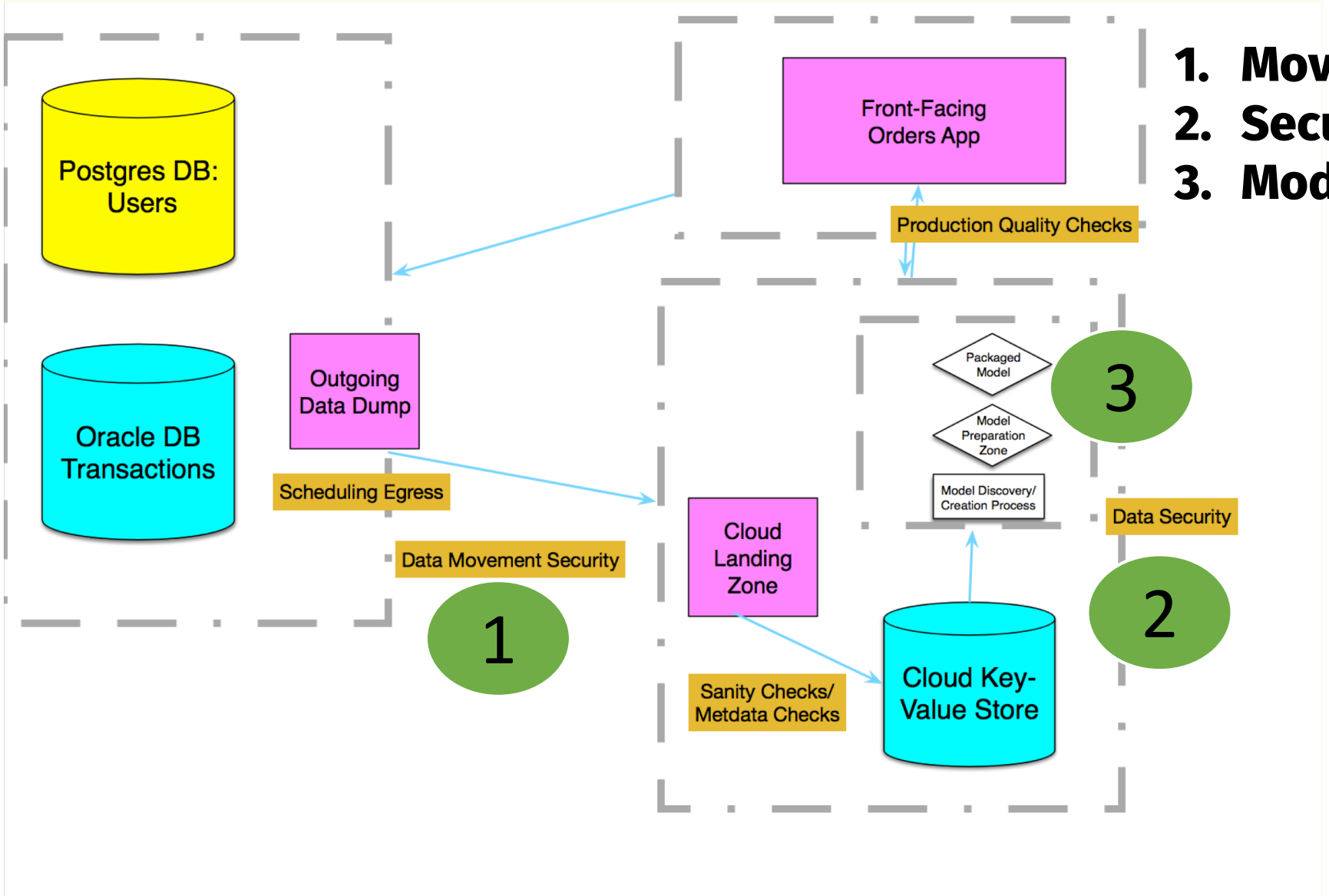
Tech Problem: You have an in-house web app and no recommendations.



Solution: A scalable cloud-based application that interfaces with your existing web app to offer personalized recommendations using deep neural nets.



You now have three new problems.



1. **Moving data in the cloud**
2. **Security and data privacy**
3. **Model interpretability**

Problem: Moving data to the cloud

- It takes time: the average time before respondents thought their IT budgets would be 80% cloud-based was 15 months
- Considerations
  - Schema migration
  - Metadata migration
  - Estimating costs
  - Analysts now become developers

Moving data to the cloud

Problem

- Which tool should you use?
- Which cloud provider should you use?
- Move data from Postgres and Oracle to AWS
  - Throughput?
  - Schemas?
- Schema and metadata management
- How much data movement needs to happen

Solution

- You need people who have been through the process before or do a lot of trial and error

Problem: Data security and privacy

- In the cloud – responsible for your own security
  - List of S3 hacks
    - personal information of 57 million Uber users and driver's license numbers
- Data breaches
  - Equifax data breach
  - GDPR

Data security and privacy

Problem

- We need to keep data secure in transit
- We can perform recommendations, but only on a subset of users
  - Export issues

Solution

- Differential privacy
- Data anonymization and hashing
- GANs

Problem: Neural Network intent

- The Barnes Foundation model
- The lack of data problem
- Spotify Playlist Recommender

Model Intent

Problem

- The model recommends books that you've already read
- Puts you into a filter bubble
- Not enough data available
- The model recommends
- "I like this, but I don't know why"

Solution

- Get people involved
- Instrumentation
- "Amazon does not know about books I buy or obtain in other capacities, and it assumes that I'm voracious for more from authors I've already purchased. "
- a simple model could do better

“It is a mistake to think you can solve any major problems just with potatoes.”
– Douglas Adams

Thanks!

Vicki Boykis | CapTech Consulting | @vboykis

In

tion. Doing so, we build on the existing literature in two ways. First, drawing upon recent advances in Machine Learning (ML) and Mobile Robotics (MR), we develop a novel methodology to categorise occupations according to their susceptibility to computerisation.[1] Second, we implement this methodology to estimate the probability of computerisation for 702 detailed occupations, and examine expected impacts of future computerisation on US labour market outcomes.