Sense and Sensibility in Data Science

Vicki Boykis | CapTech Consulting | @vboykis

About Me

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

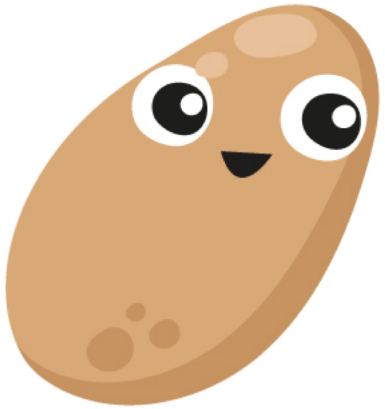
"Data Science": ["Python", "R"],

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

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

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

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



Douglas Adams on the

Silastic Armorfiends of Striterax

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

Things we worry about but shouldn’t

- Outsourcing of human jobs
- Self-driving cars
- Robots

The end of human-led decision making

Hype: Outsourcing of jobs to technology

- "Over half of all jobs will be automated within 10-20 years"

- "It's quite obvious that we should stop training radiologists" – Geoff Hinton

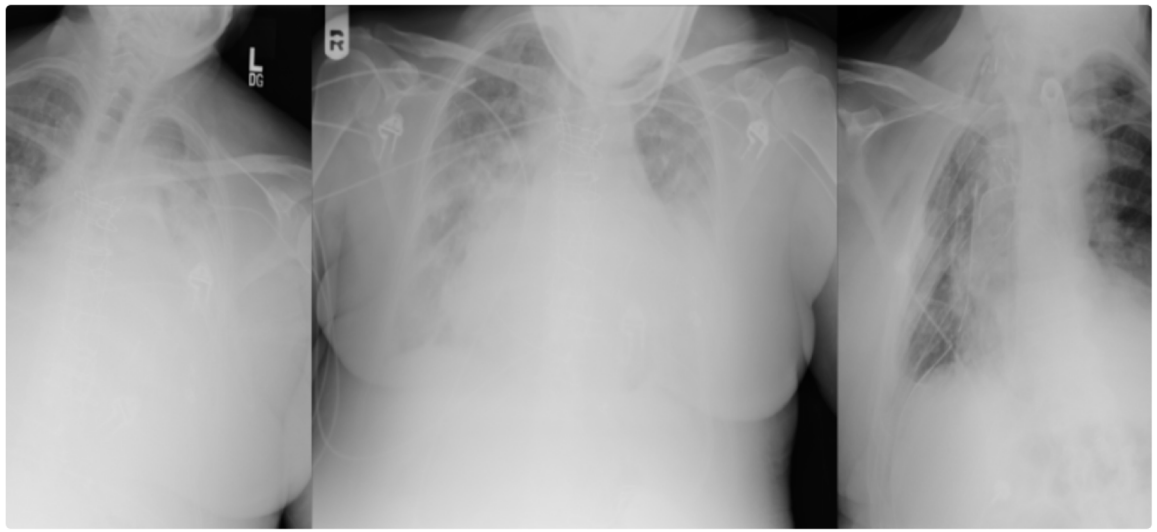
Source: The future of employment: How susceptible are jobs to computerization?
Frey and Osborne, 2013

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

Exploring the ChestXray14 dataset: problems



DECEMBER 18, 2017 ~ LUKEOAKDENRAYNER

A couple of weeks ago, I mentioned I had some concerns about the ChestXray14

Source

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
  - Free time increase

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:

```
1.13 years for Imagenet V3 or similar;
274 days for ResNer 50 or similar;
51 days for AlexNet or similar.
```

Reality: Humans still need to be in the loop

This means that a fleet of 100 cars instrumented with 5 cameras each will generate in excess of one million hours of video recording in a year. This data needs to be captured, transported from the car to the data center, stored, processed and used for training.

Importantly, since supervised learning algorithms are used, the data also needs to be annotated by humans. Marking every pedestrian, car, lane, and other details can become a significant bottleneck if the data annotation team is not adequately sized.

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

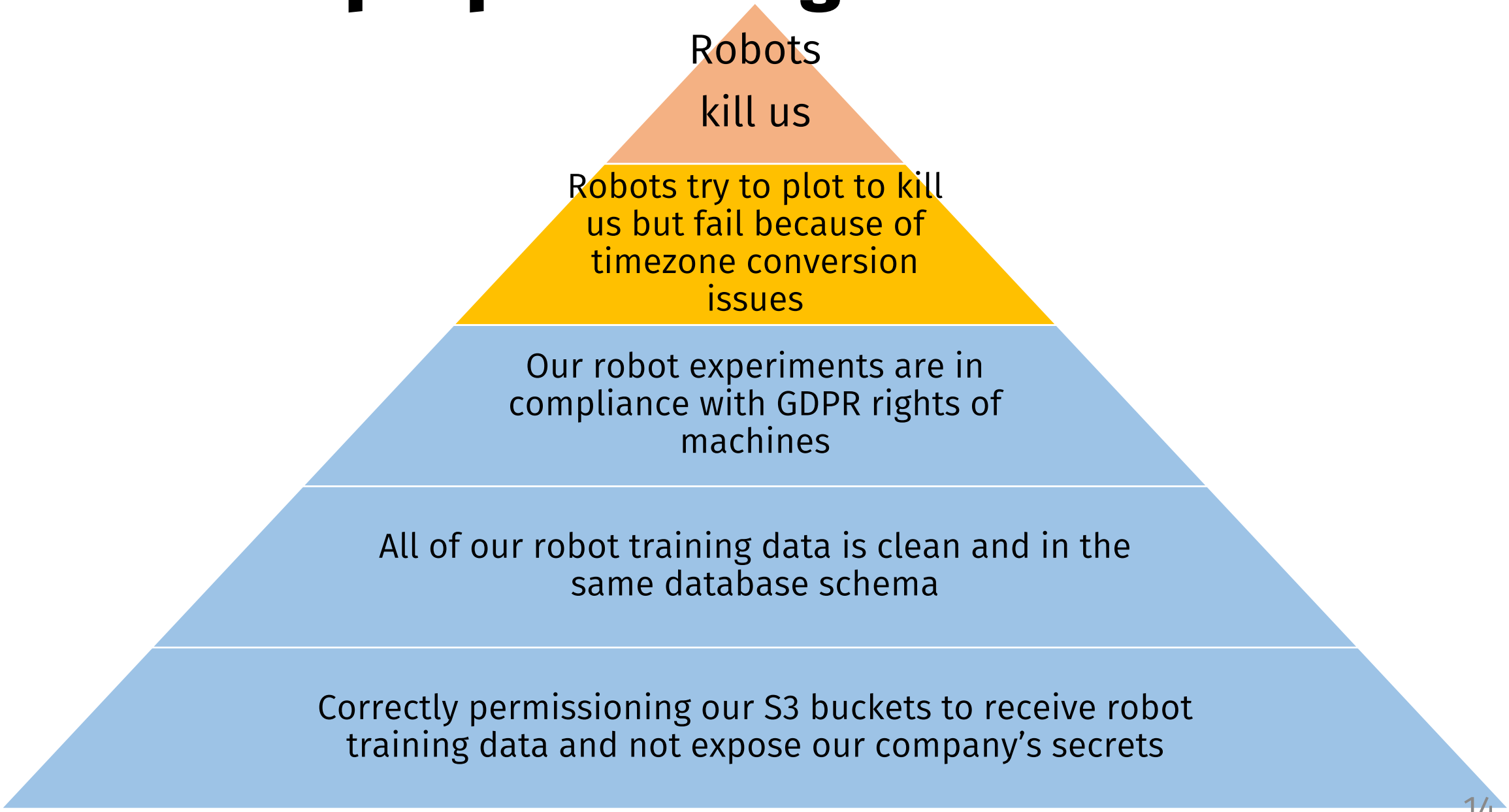
Hype: Robots will kill us

Reality: AGI is still a long time away

The real problems in data work

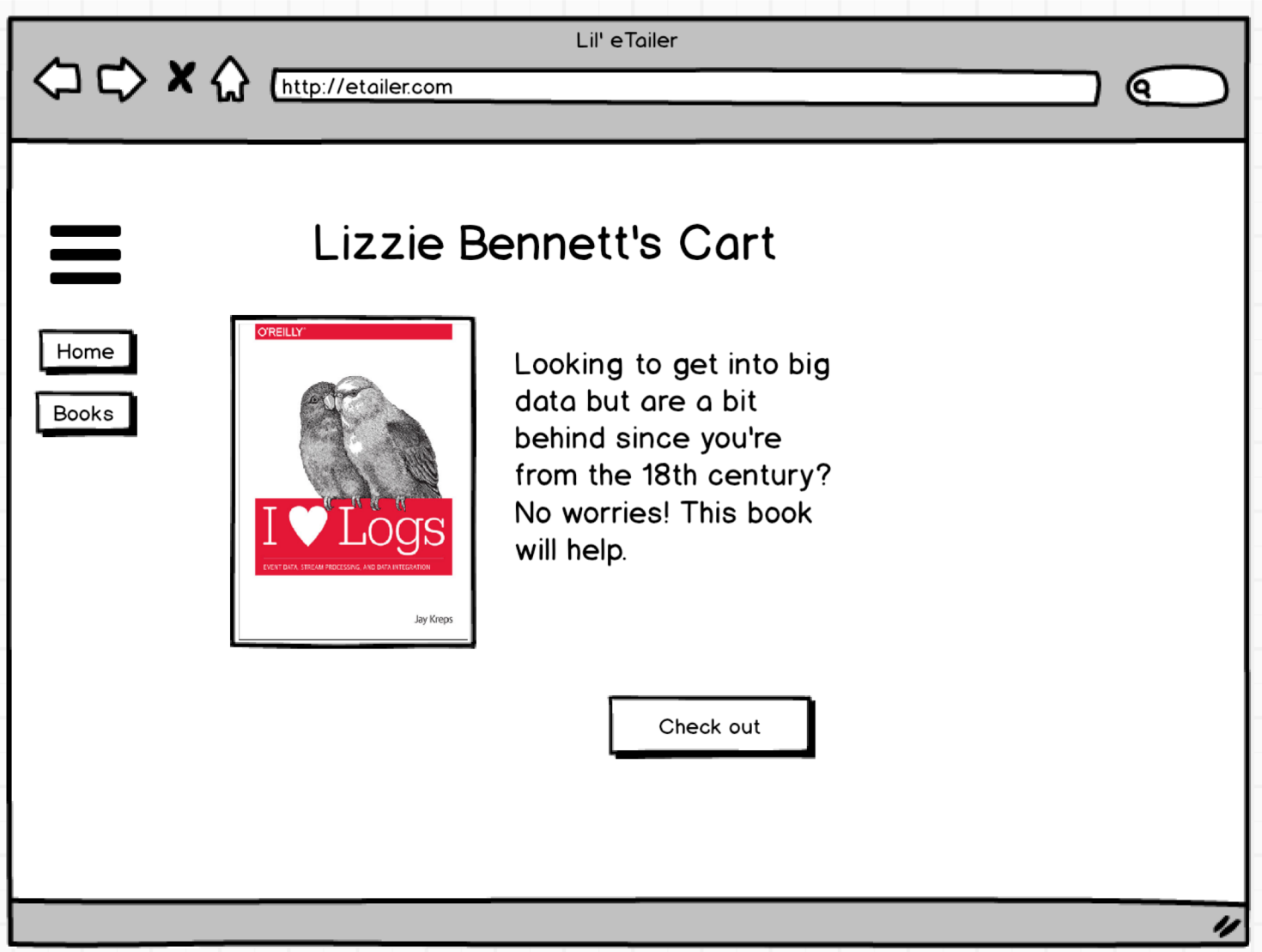
1. Moving and keeping data in the cloud
2. Security and data privacy
3. Model intent

What are people on the ground worried about?

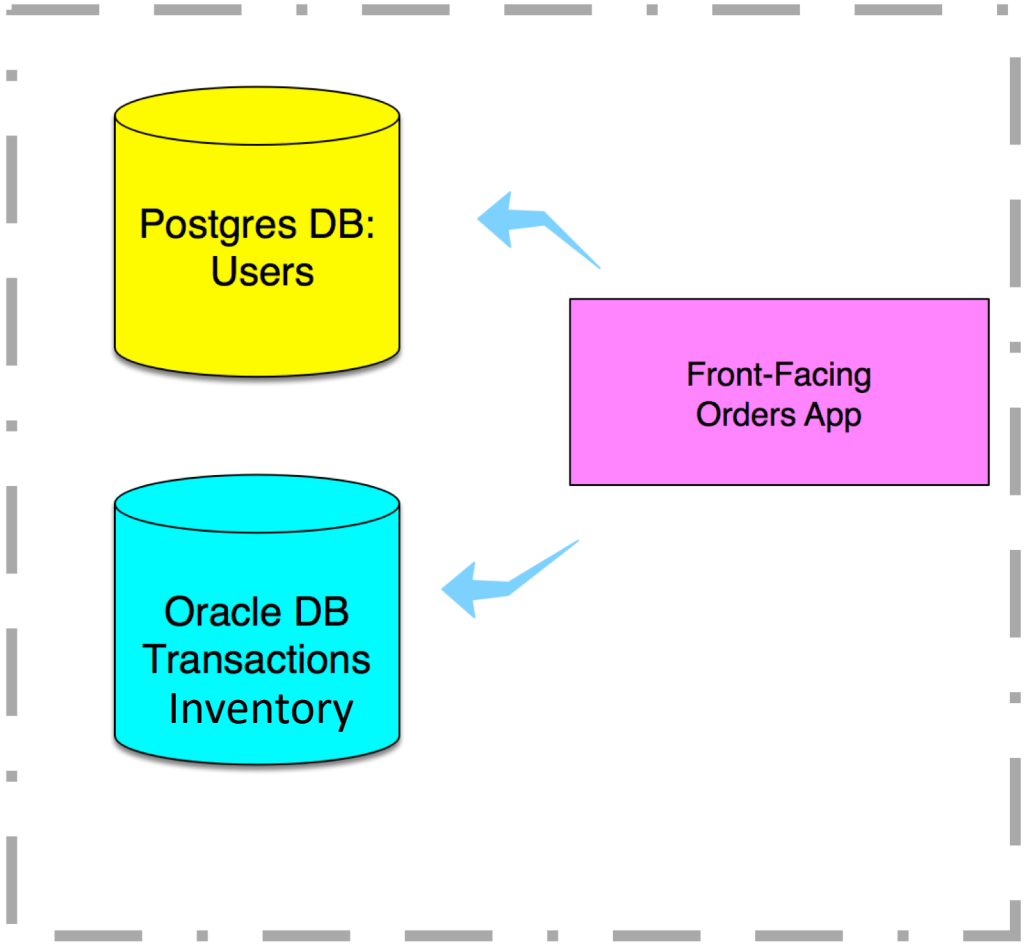


Problem: You are a large, traditional book retailer. You've been doing well in brick and mortar sales and ok your website but now you need to compete against

Your website



Tech Problem: You have a web app built on-prem and no recommendations.



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:

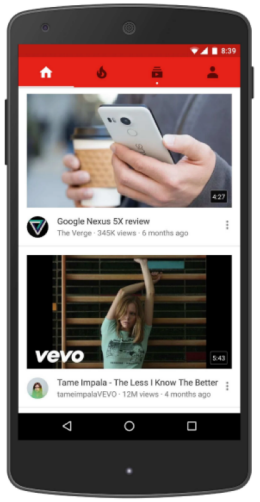


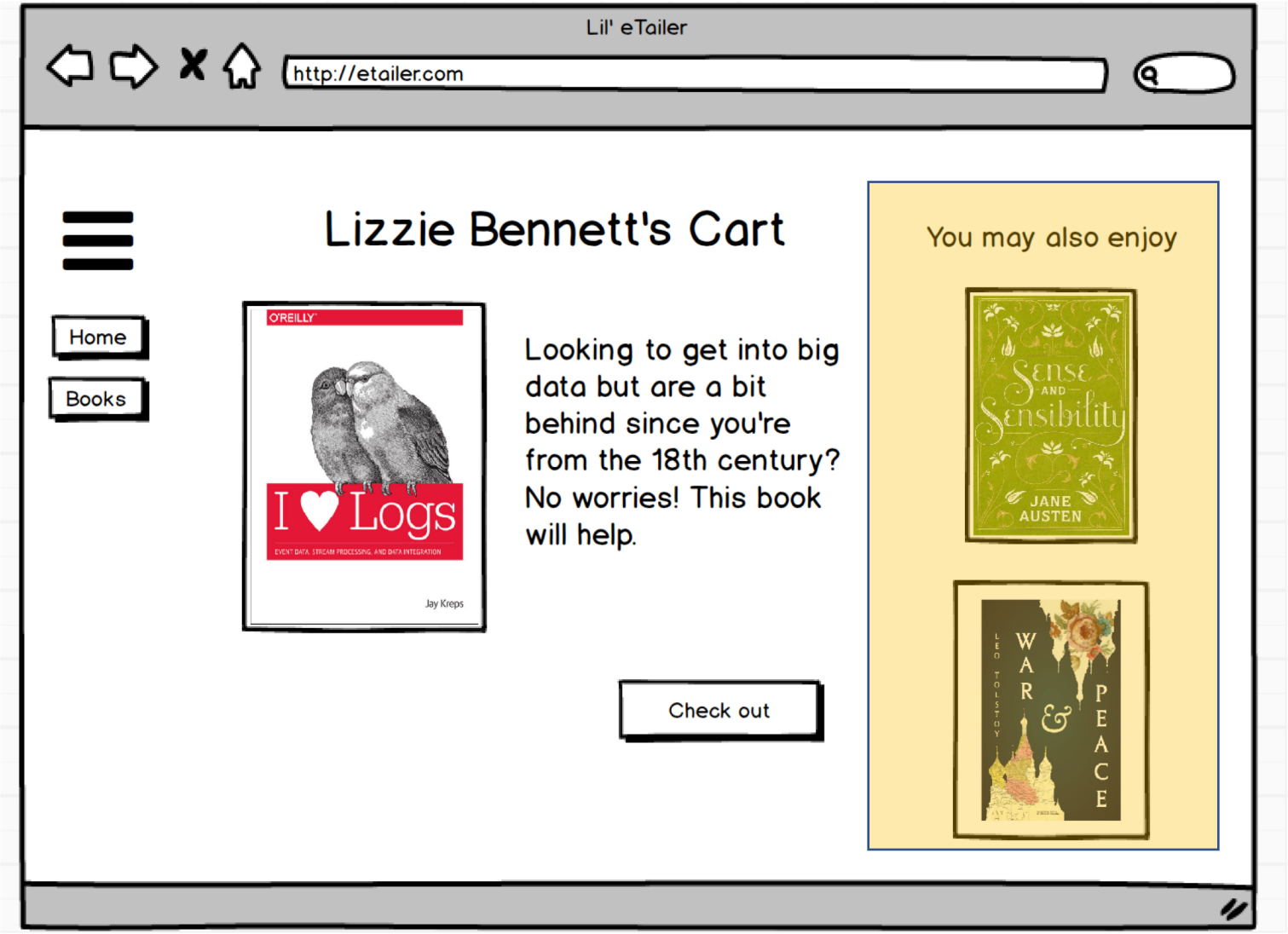
Figure 1: Recommendations displayed on YouTube mobile app home.

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

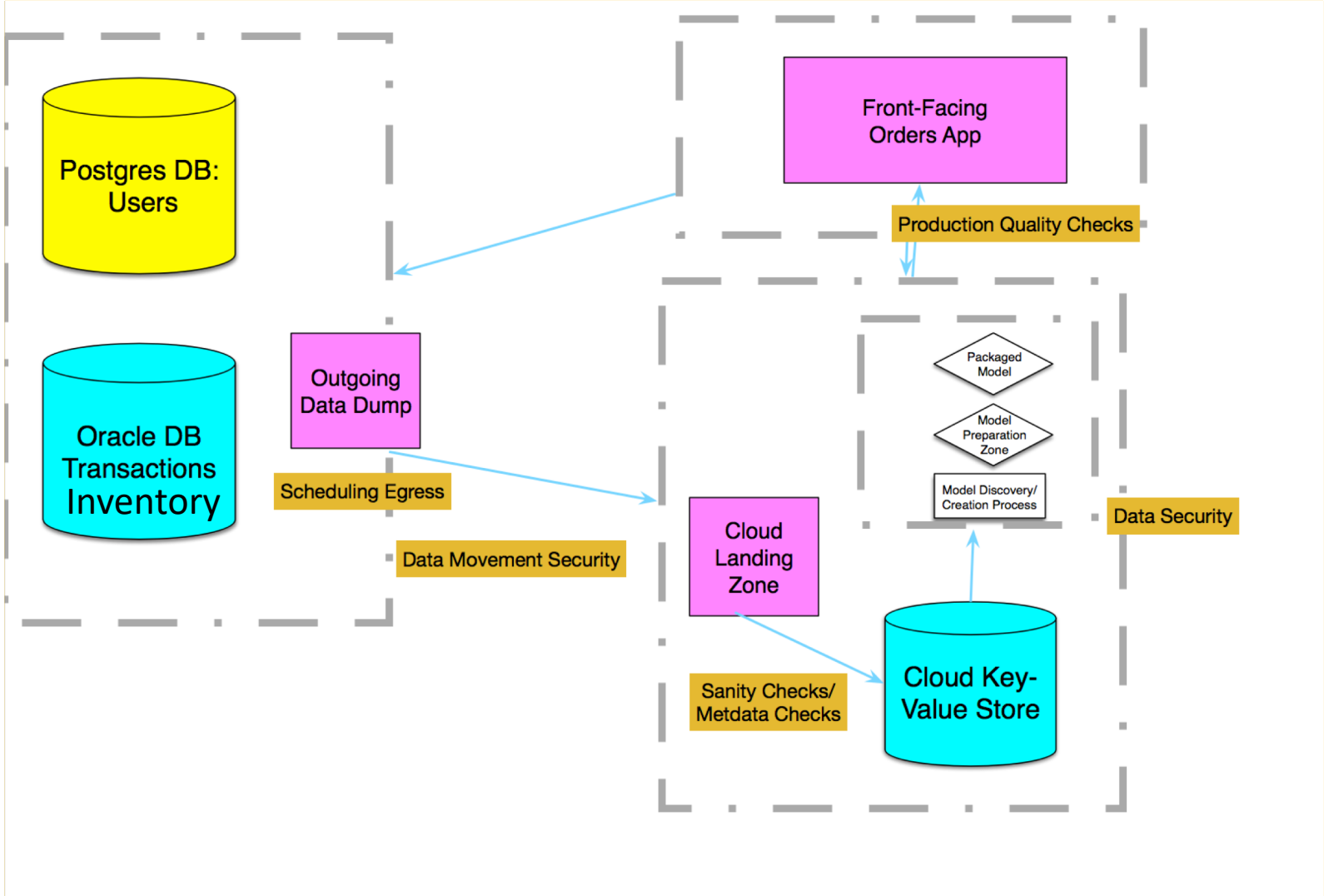
- *Noise:* Historical user behavior on YouTube is inher-

HariBotter

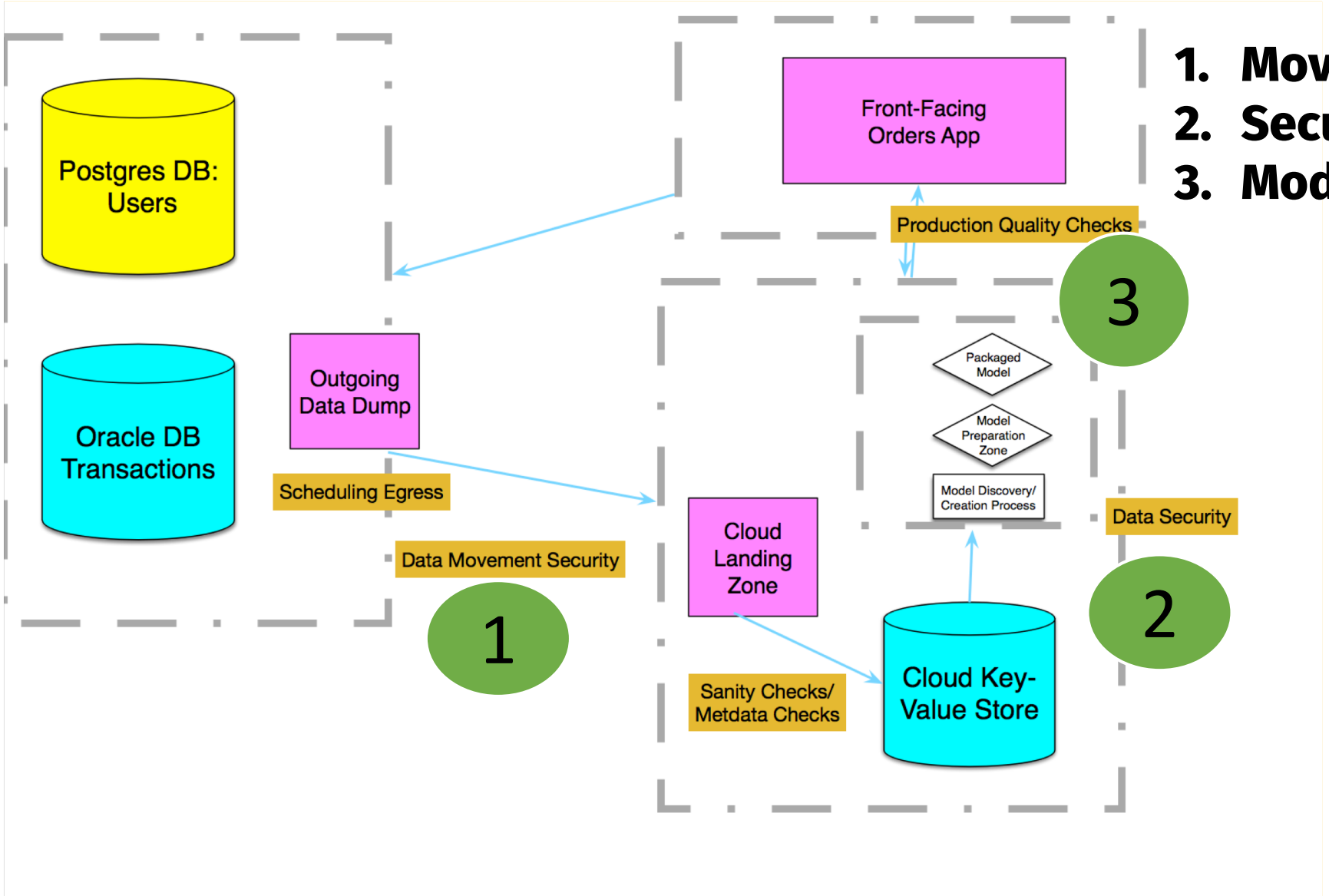
A.I.-based sorting hat for your book preferences



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



You have three new problems to think about.



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

Problem: Moving data to the cloud

- Planning a migration takes time. 15 months from initialization phase to 80% of a budget being cloud spending.
- Considerations
  - Picking tooling
  - Configuring infrastructure and permissions
  - Stakeholder buy-in
  - Schema migration
  - Metadata migration
  - Estimating costs
  - Finding cloud-native talent

Moving to the cloud

Problem

"I need to blend data from Postgres and Oracle and move into the cloud."

Solution

- Study cloud-native architectures to leverage the strengths:
  - Disposable resources
  - Multiple ways of imports
- Use available metadata management solutions like Glue
- Allow end users to be able to manage data inputs
- Increase cloud literacy across the organization

Problem: Data security and privacy

- Responsibility for security doesn't end where the cloud begins
  - Alteryx as a partner of Experian revealed financial info of 123 million American households
- Data breaches
  - Equifax data breach
- GDPR

Data security and privacy

Problem

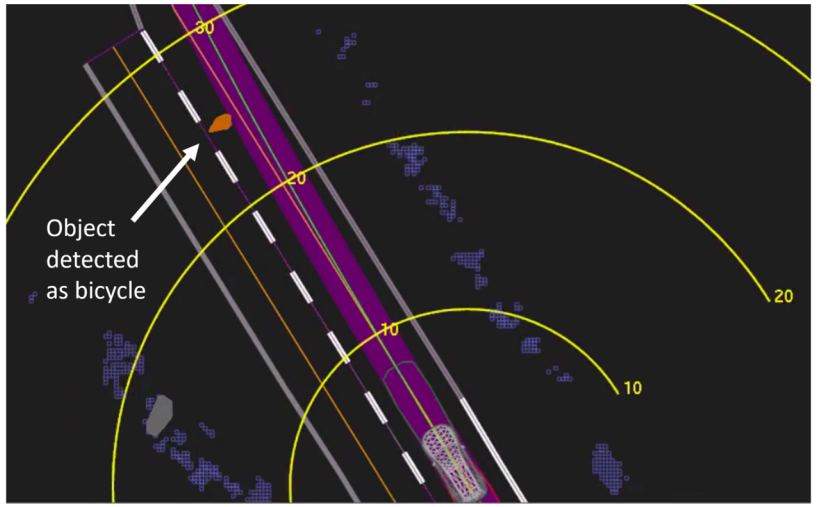
"We don't want to have our user data compromised. "

Solution

- The less data you store, the easier it can be
  - Keep windows of history, then send to cold storage
- Differential privacy
  - Adding noise to a sample of usage patterns
- Synthetic data creation
  - GANs

Problem: Having neural nets make decisions like humans

- The Barnes Foundation model
  The algorithm led to recommendations of paintings that were similar
- Self-driving cars
  The Uber NSTF report showed that the car tried several times to classify the object before, finally stopping. Humans work differently.



Model Intent

Problem

- "The model only recommends books in the same genre and we can't figure out how to change that."

Solution

- Get people involved earlier in the process
  - User testing
  - Data validation
- Collecting a lot of varied training data
- Use simpler models that are easier to interpret

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

Thanks!

Vicki Boykis | CapTech Consulting | @vboykis