



ML SYSTEMS DESIGN MEETUP GROUP

HETAV PANDYA

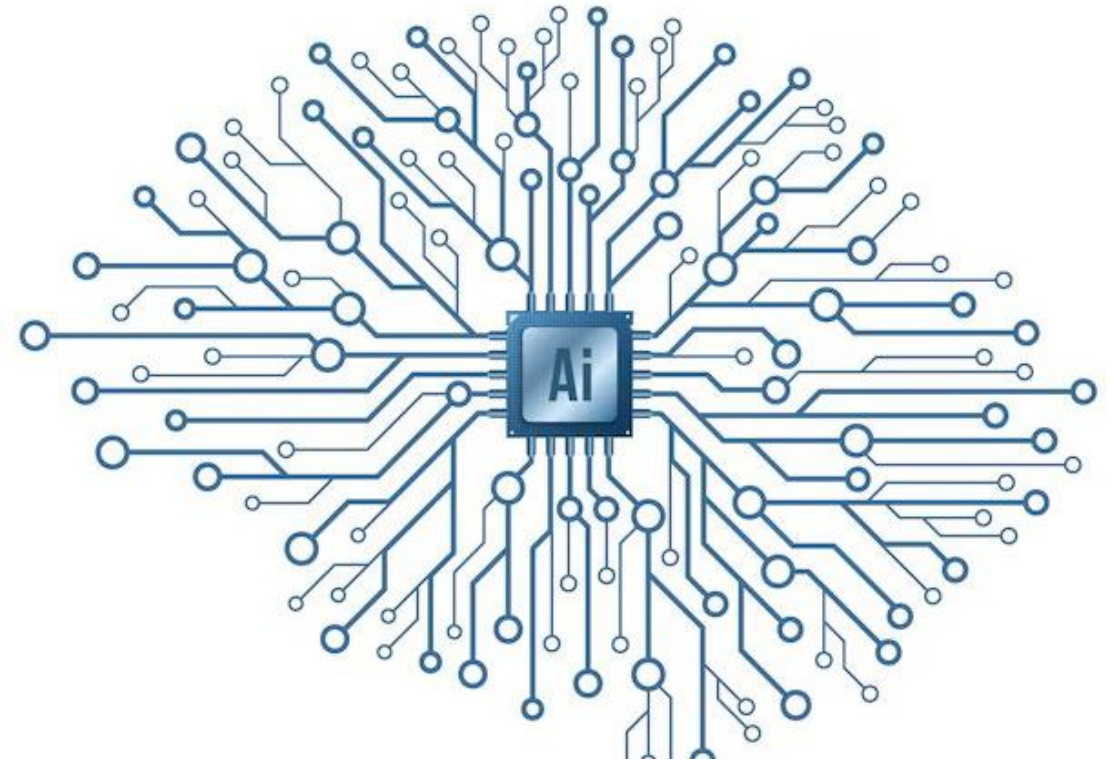
AGENDA

INTRODUCTION

OVERVIEW OF MACHINE LEARNING SYSTEMS

INTRODUCTION TO MACHINE LEARNING SYSTEMS DESIGN

OPEN Q&A





INTRODUCTION

- Welcome to ML Systems Design Meetup Group
- Designing Machine Learning Systems – Chip Huyen
- Two chapters every meeting
- Free Access – City Library
- Frequency – Biweekly - Monthly
- Questions: Meetup Event Chat

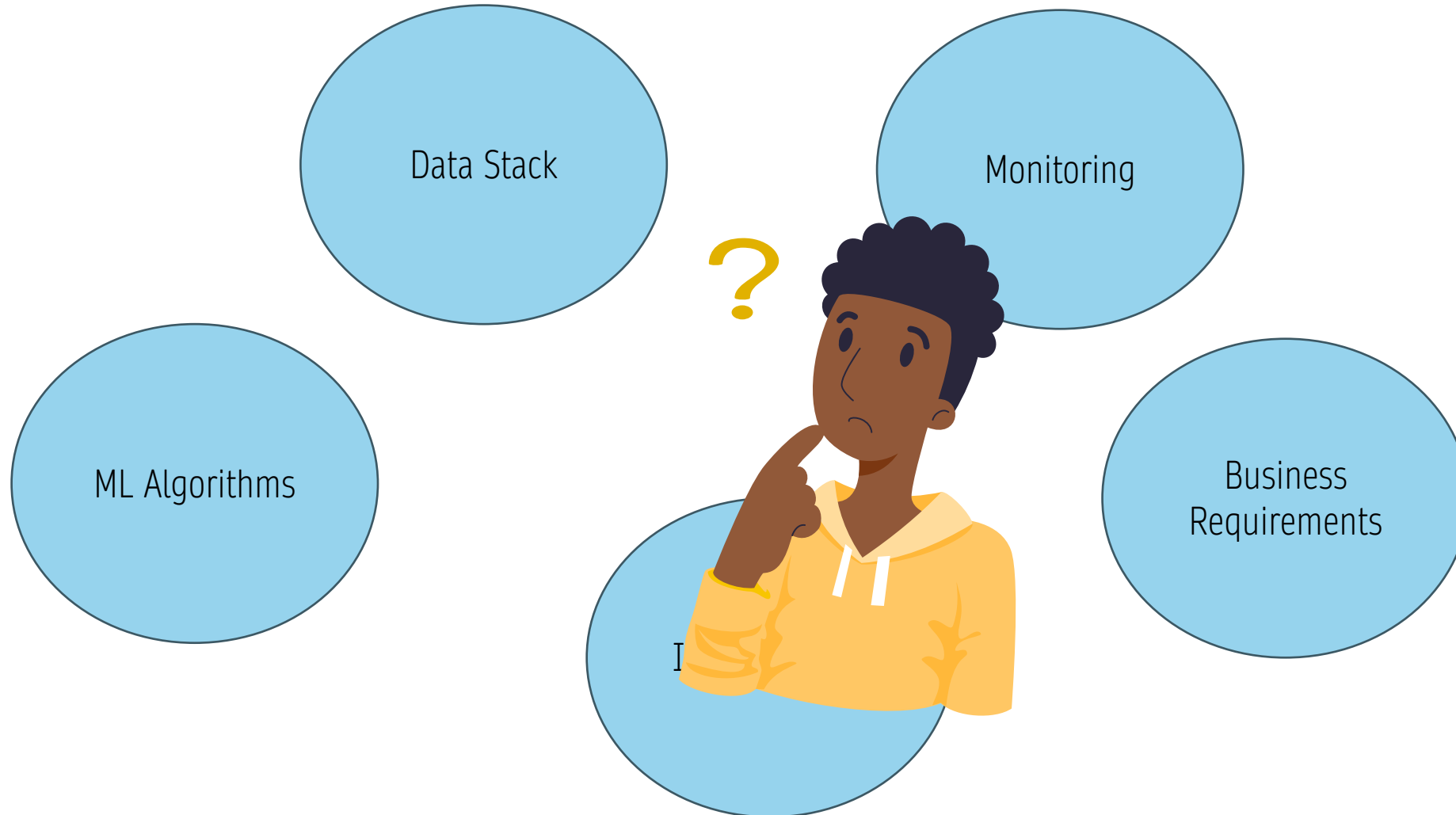


FREE ACCESS

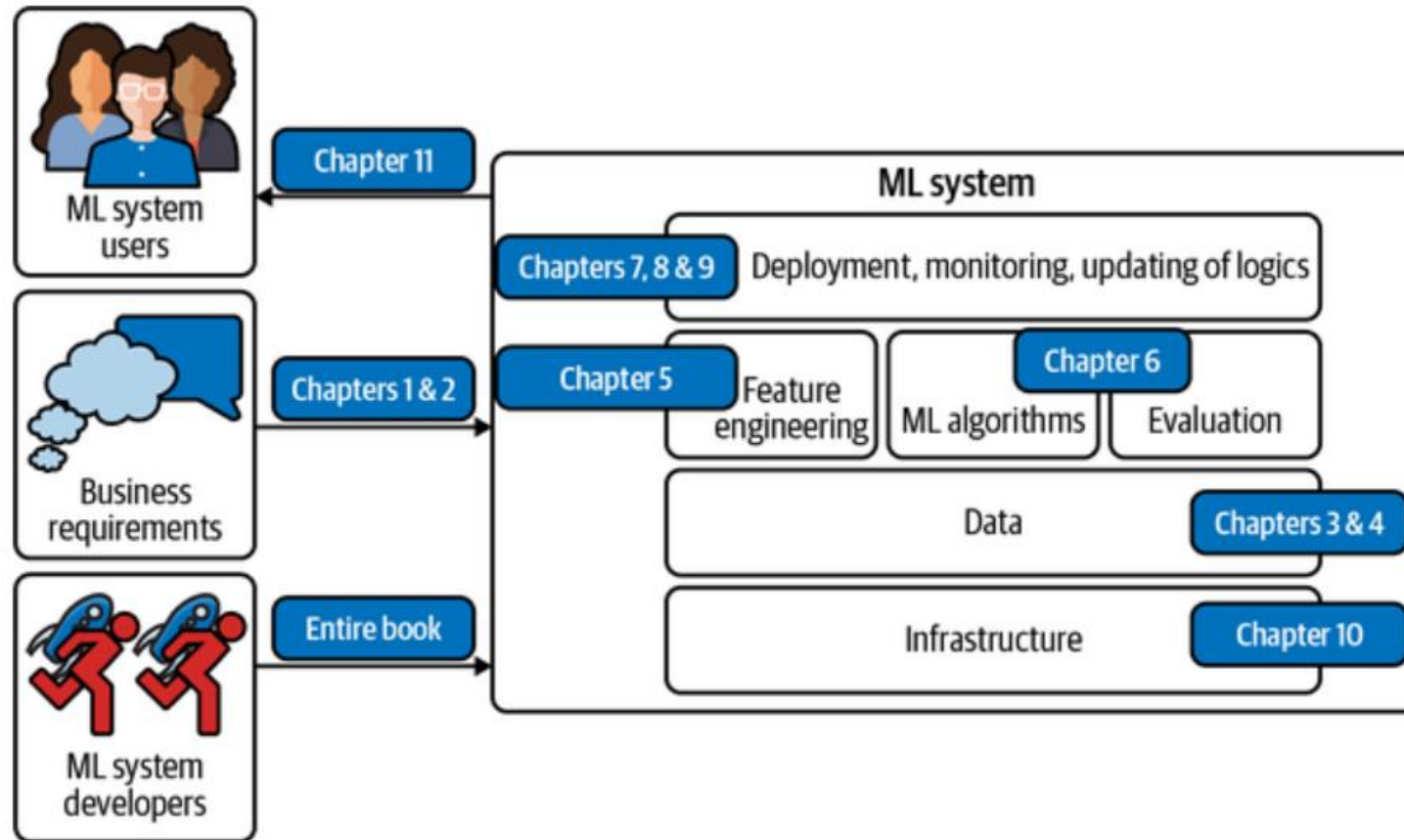


Via Burnaby Public Library

OVERVIEW OF MACHINE LEARNING SYSTEMS



THE BOOK IN A NUTSHELL



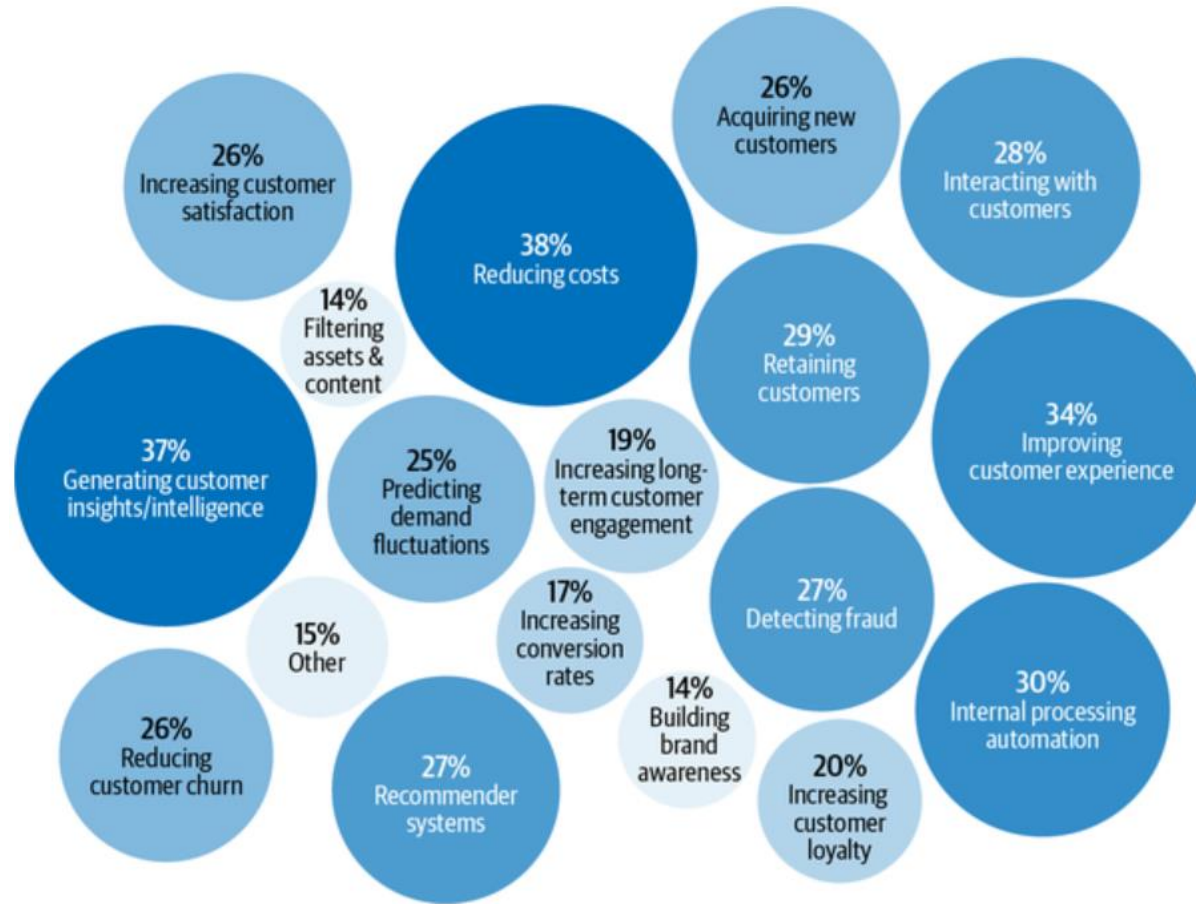
WHERE CAN ML HELP?

Machine learning is an approach to (1) *learn* (2) *complex patterns* from (3) *existing data* and use these patterns to make (4) ~~*predictions*~~ on (5) *unseen data*.

WHERE CAN ML HELP?

- It's repetitive
- The cost of wrong predictions is low
- It is required at scale
- Patterns are constantly changing

WHERE CAN IT HELP YOU?



WHERE CAN IT HELP YOU?

As of 2019, the average cost for an app to acquire a user who'll make an in-app purchase is \$86.61. The acquisition cost for Lyft is estimated at \$158/rider. This cost is so much higher for enterprise customers. Customer acquisition cost is hailed by investors as a startup killer. Reducing customer acquisition costs by a small amount can result in a large increase in profit

WHERE CAN HELP YOU?

The cost of acquiring a new user is approximated to be 5 to 25 times more expensive than retaining an existing one. *Churn prediction* is predicting when a specific customer is about to stop using your products or services so that you can take appropriate actions to win them back.

CS101: INTRO TO ML

	Research	Production
Requirements	State-of-the-art model performance on benchmark datasets	Different stakeholders have different requirements
Computational priority	Fast training, high throughput	Fast inference, low latency
Data	Static ^a	Constantly shifting
Fairness	Often not a focus	Must be considered
Interpretability	Often not a focus	Must be considered

INFERENCE VS THROUGHPUT

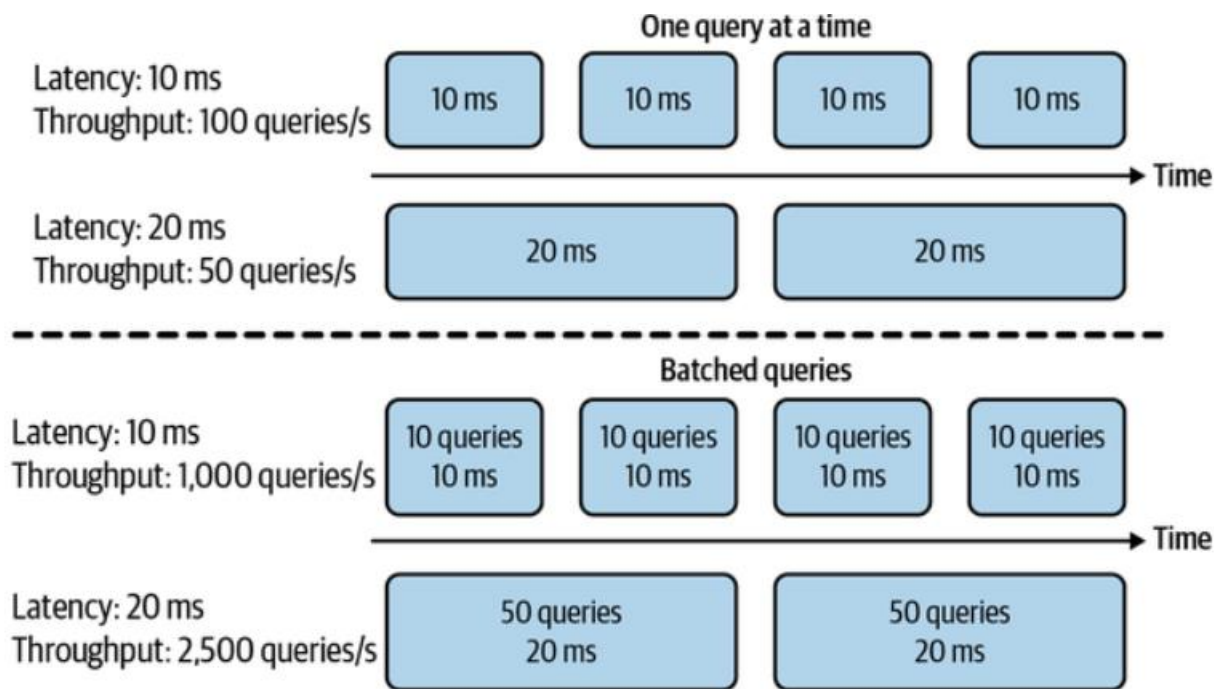


Figure 1-4. When processing queries one at a time, higher latency means lower throughput. When processing queries in batches, however, higher latency might also mean higher throughput.

INFERENCE VS THROUGHPUT

In 2019, Booking.com found that an increase of about 30% in latency cost about 0.5% in conversion rates—“a relevant cost for our business.” In 2016, Google found that more than half of mobile users will leave a page if it takes more than three seconds to load. Users today are even less patient.

INTERPRETABILITY

A light blue cloud shape with a black outline, containing the text "SHAP".

SHAP

“Suppose you have cancer and you have to choose between a black box AI surgeon that cannot explain how it works but has a 90% cure rate and a human surgeon with an 80% cure rate. Do you want the AI surgeon to be illegal?”

A light blue cloud shape with a black outline, containing the text "CAPTUM".

CAPTUM

A light blue cloud shape with a black outline, containing the text "LIME".

LIME

ML VS SOFTWARE*

I need data, more data, good data, consistent data!

VS

I need precise, concise and maintainable logic

ML VS SOFTWARE*



ML VS SOFTWARE*

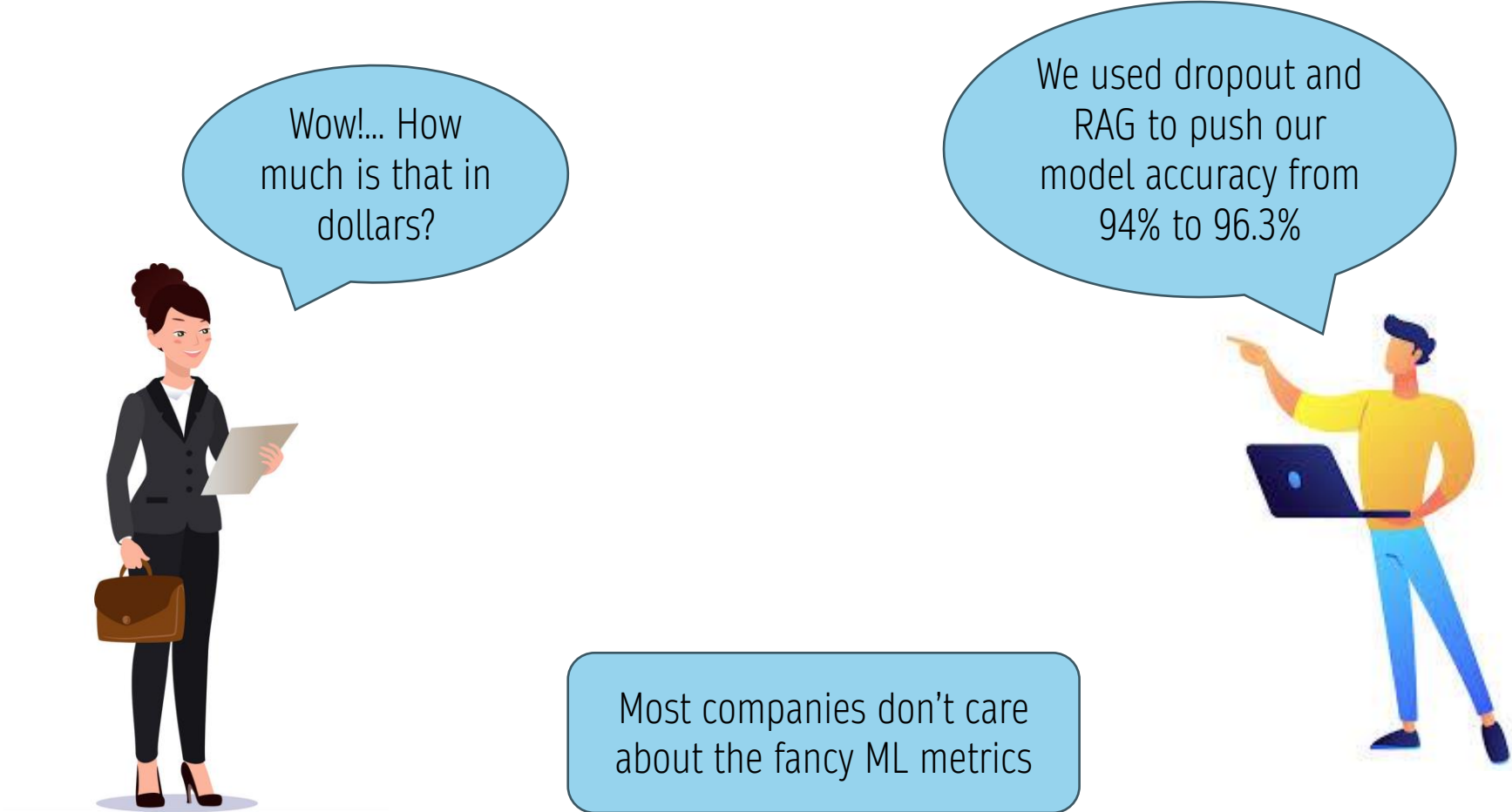
Need more resources, RAM,
CPU, GPU, TPU!

VS

Why?

I am usually content with
4GB RAM... unless I am
chrome

THE HARSH TRUTH



Wow!... How much is that in dollars?

We used dropout and RAG to push our model accuracy from 94% to 96.3%

Most companies don't care about the fancy ML metrics

WHERE IS MY MONEY

The ultimate goal of any project within a business is, therefore, to increase profits, either directly or indirectly: directly such as increasing sales (conversion rates) and cutting costs.

Netflix measures the performance of their recommender system using *take-rate*: the number of quality plays divided by the number of recommendations a user sees. The higher the take-rate, the better the recommender system.

ML + ? = SUCCESS

- Experimentation
- Beta Releases
- A/B Testing
- Consumer Surveys
- Understanding revenue and cost structure
 - Automation is efficient
 - Redundancy is painful
 - Repetition is an opportunity

RETURNS OVER TIME

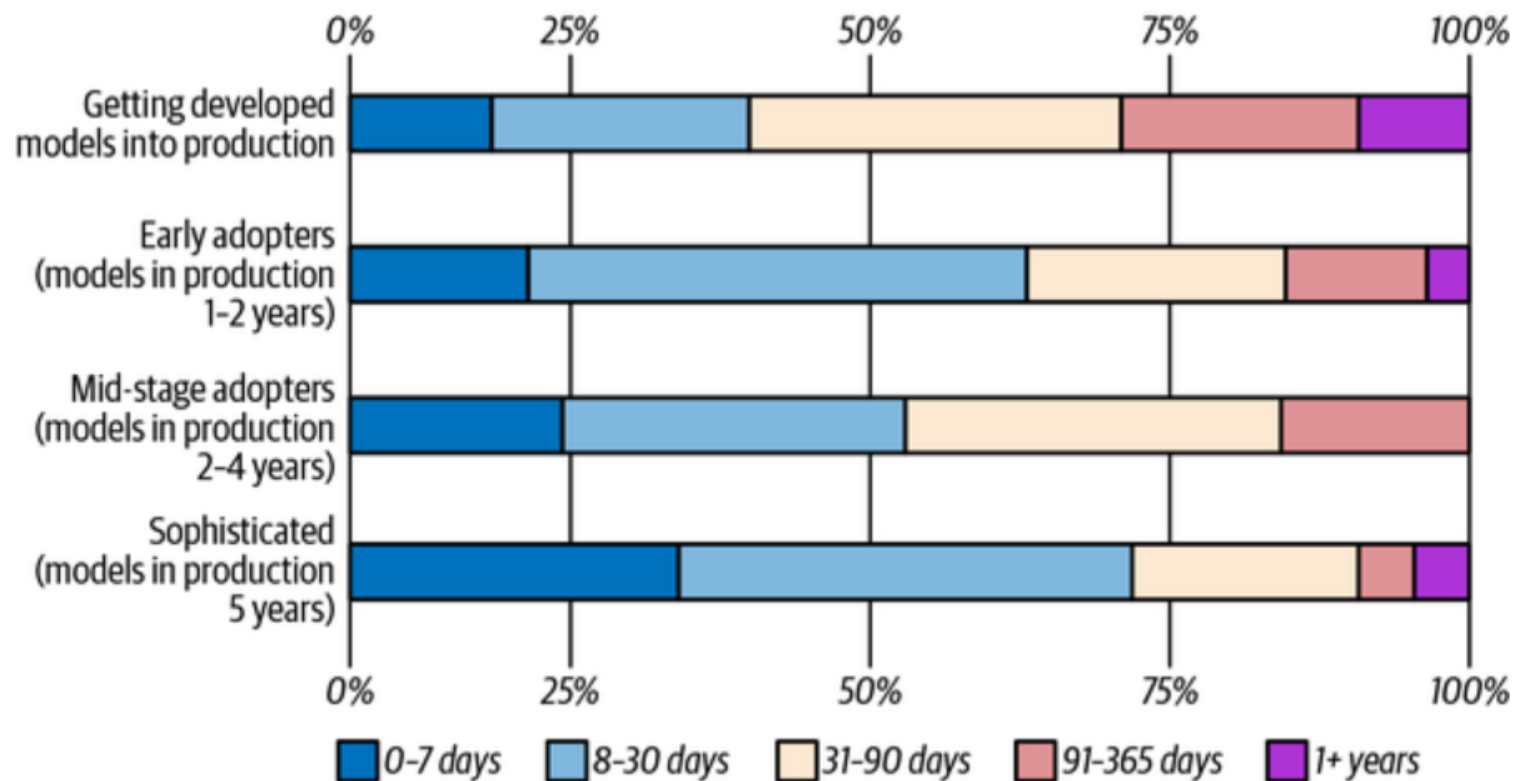
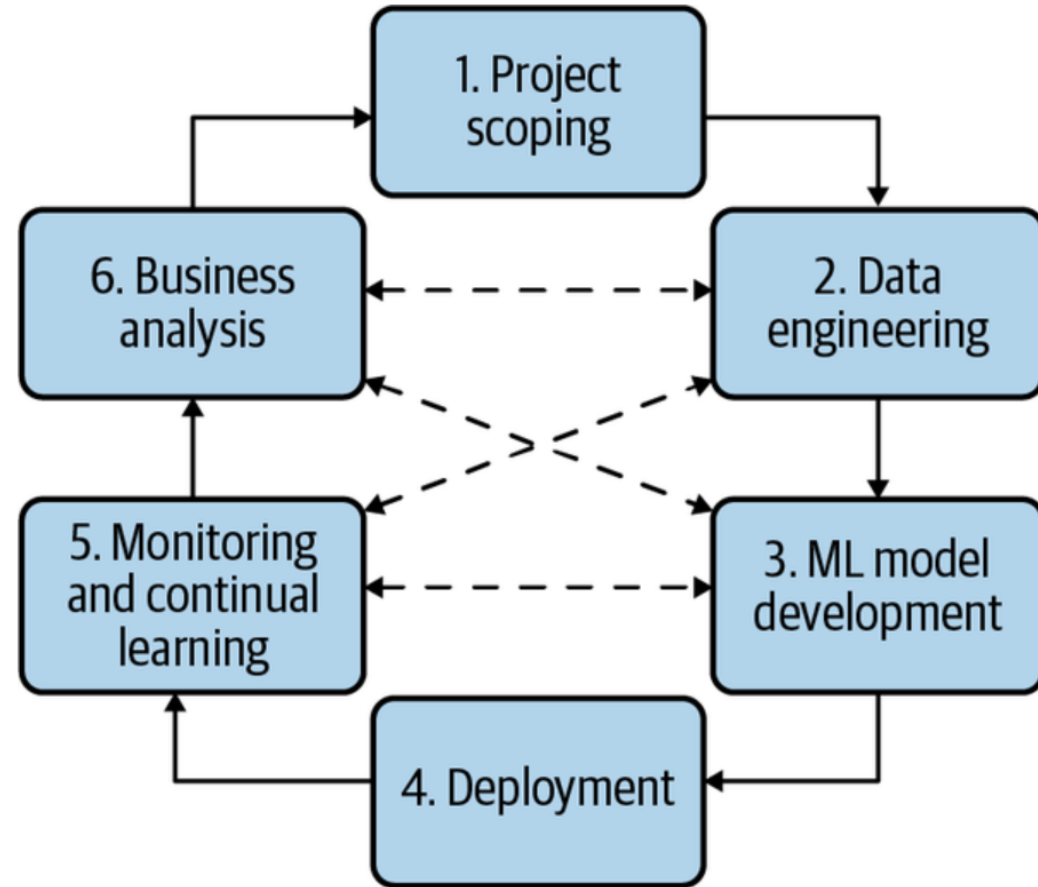


Figure 2-1. How long it takes for a company to bring a model to production is proportional to how long it has used ML. Source: Adapted from an image by Algorithmia

REQUIREMENTS OF ML SYSTEMS

- Reliability
- Scalability
- Maintainability
- Adaptability



HOW RELIABLE IS YOUR SYSTEM?

With traditional software systems, you often get a warning, such as a system crash or runtime error or 404. However, ML systems can fail silently. End users don't even know that the system has failed and might have kept on using it as if it were working.

No errors, it must be working fine!



SYSTEM SCALING

At peak, your system might require 100 GPUs (graphics processing units). However, most of the time, it needs only 10 GPUs. Keeping 100 GPUs up all the time can be costly, so your system should be able to scale down to 10 GPUs.

AWS!

Even Amazon fell victim to this when their autoscaling feature failed on Prime Day, causing their system to crash. An hour of downtime was estimated to cost Amazon between \$72 million and \$99 million

FIND IT TO FRAME IT

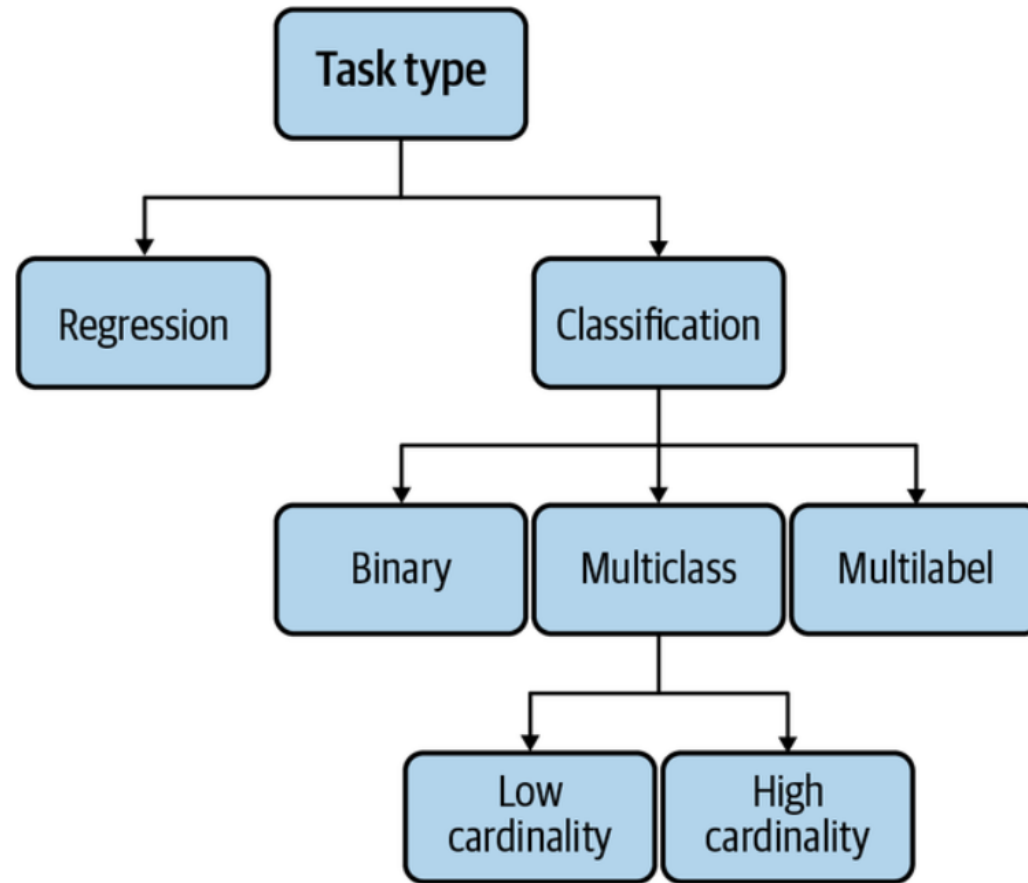


Figure 2-3. Common task types in ML

WHAT IS YOUR OBJECTIVE?

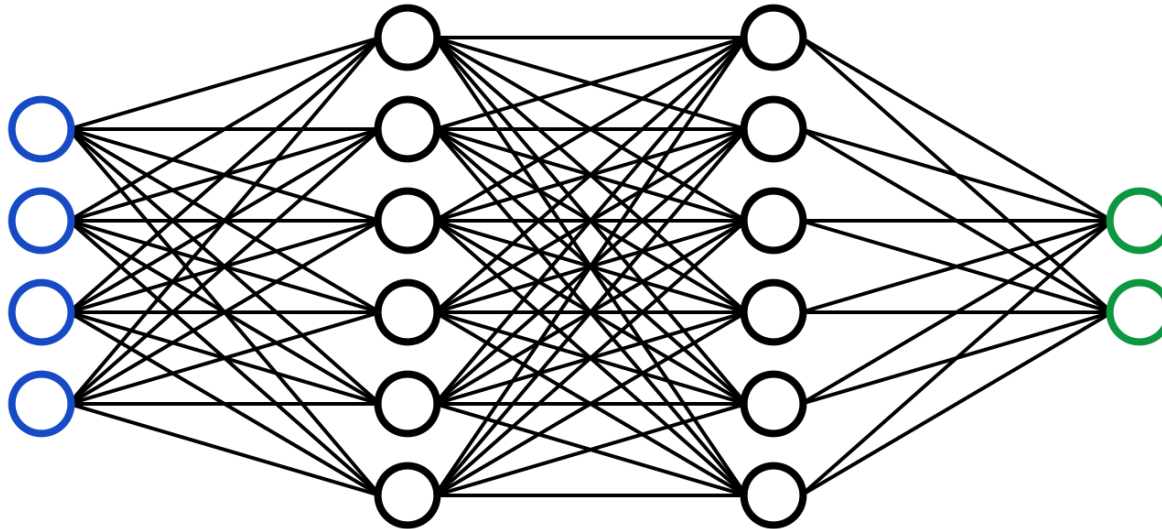


WHAT IS YOUR OBJECTIVE?

$$\alpha * health + \beta * money + \gamma * Happiness$$

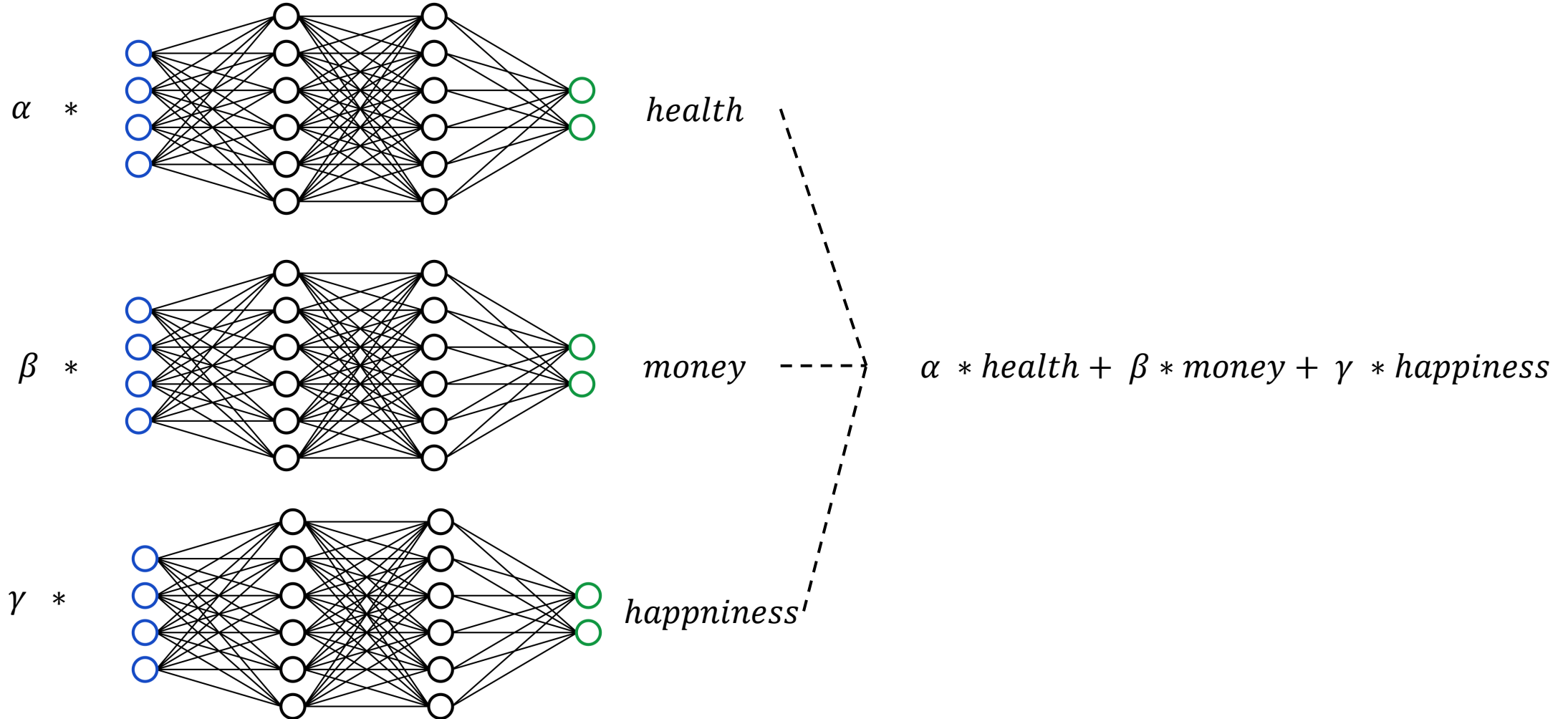


WHAT IS YOUR OBJECTIVE?



$$\alpha * health + \beta * money + \gamma * Happiness$$

WHAT IS YOUR OBJECTIVE?



MIND VS DATA

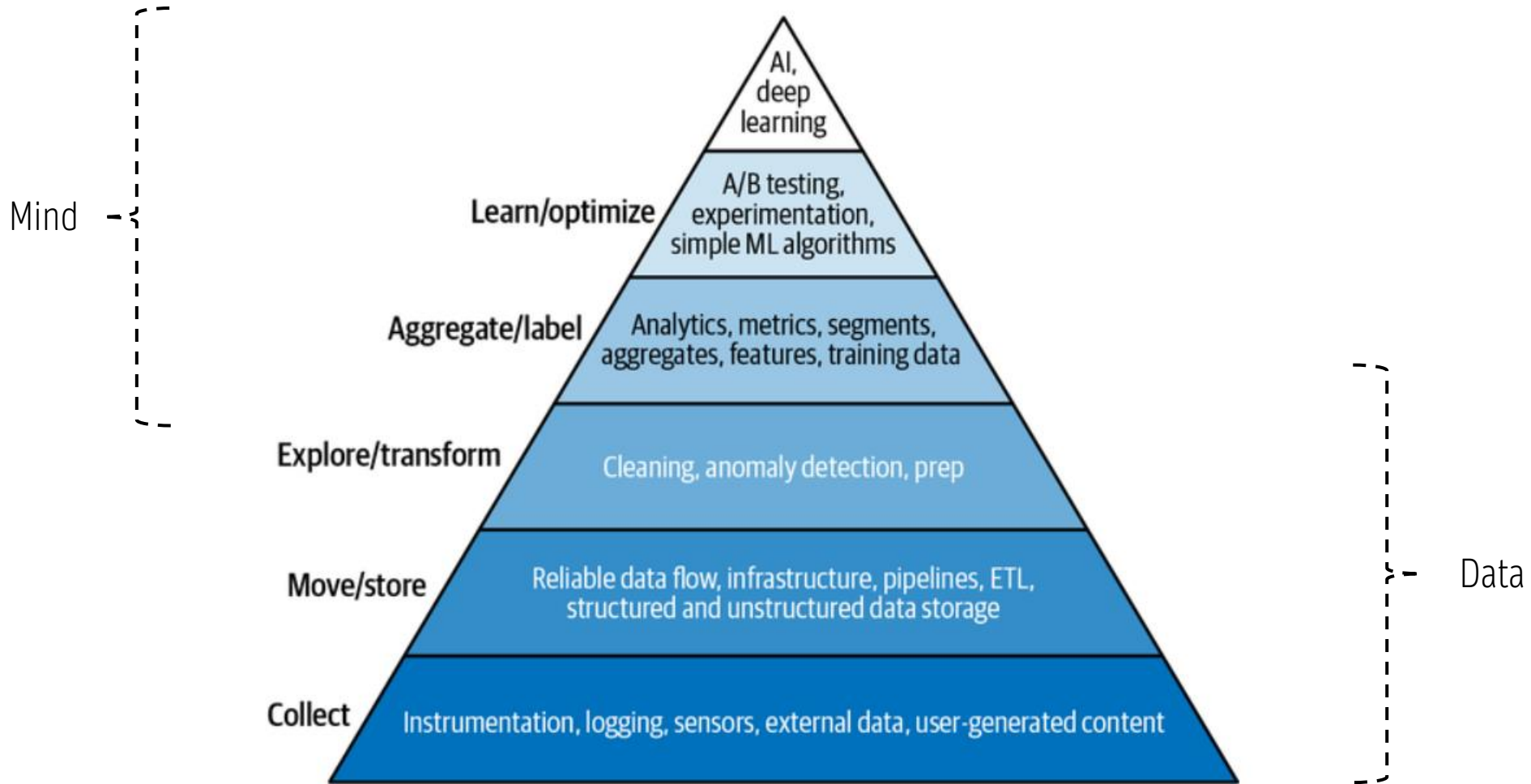


Figure 2-7. The data science hierarchy of needs. Source: Adapted from an image by Monica Rogati²²

YOUR VOICE MATTERS!

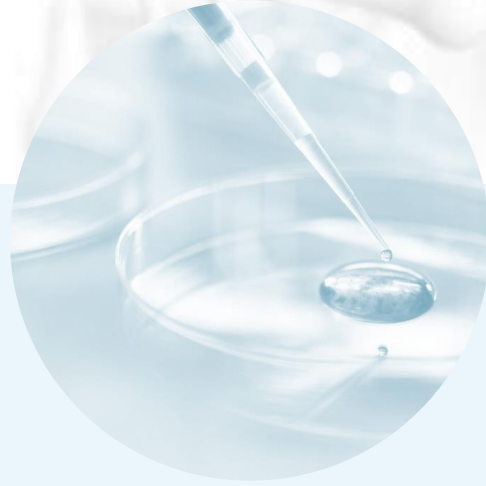
Please take some time to
fill up our very very short
feedback form 😊



If you would like to
connect with me, feel free
to scan this!

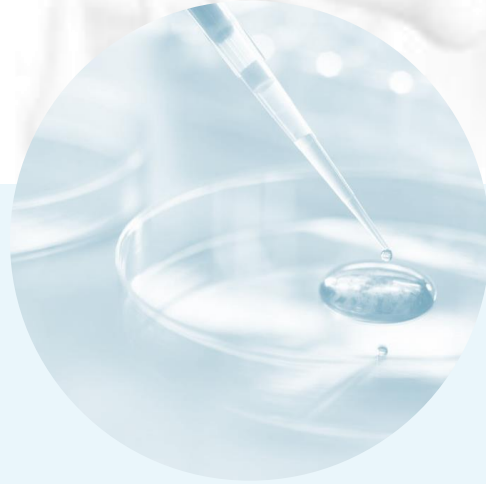


THANK YOU!



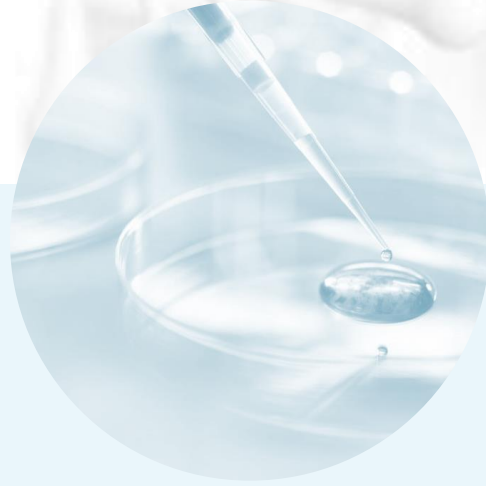
Q&A TIME





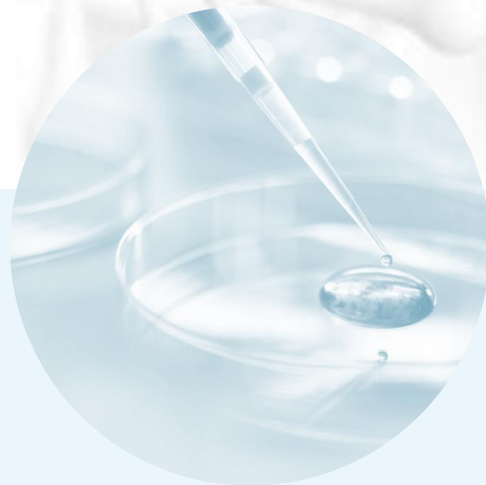
HOW CAN ORGANIZATIONS BALANCE THE NEED
FOR COLLECTING DIVERSE DATA WITH THE
ETHICAL CONCERNS AROUND PRIVACY, BIAS, AND
CONSENT?





HOW CAN TEAMS ENSURE THE REPRODUCIBILITY
OF ML EXPERIMENTS AND RESULTS WHEN
DEALING WITH LARGE AND DYNAMIC DATASETS?





HAVE YOU FACED CHALLENGES RELATED TO DATA
QUALITY AND MANAGEMENT IN YOUR ML
PROJECTS. HOW WERE THESE CHALLENGES
ADDRESSED?

