Evaluating Machine Learning Models – A Beginner's Guide

Alice Zheng, Dato

September 15, 2015

My machine learning trajectory

Carnegie Mellon University

Applied machine learning
(Data science)

Microsoft Research

Shortage of experts
and good tools.

GraphLab

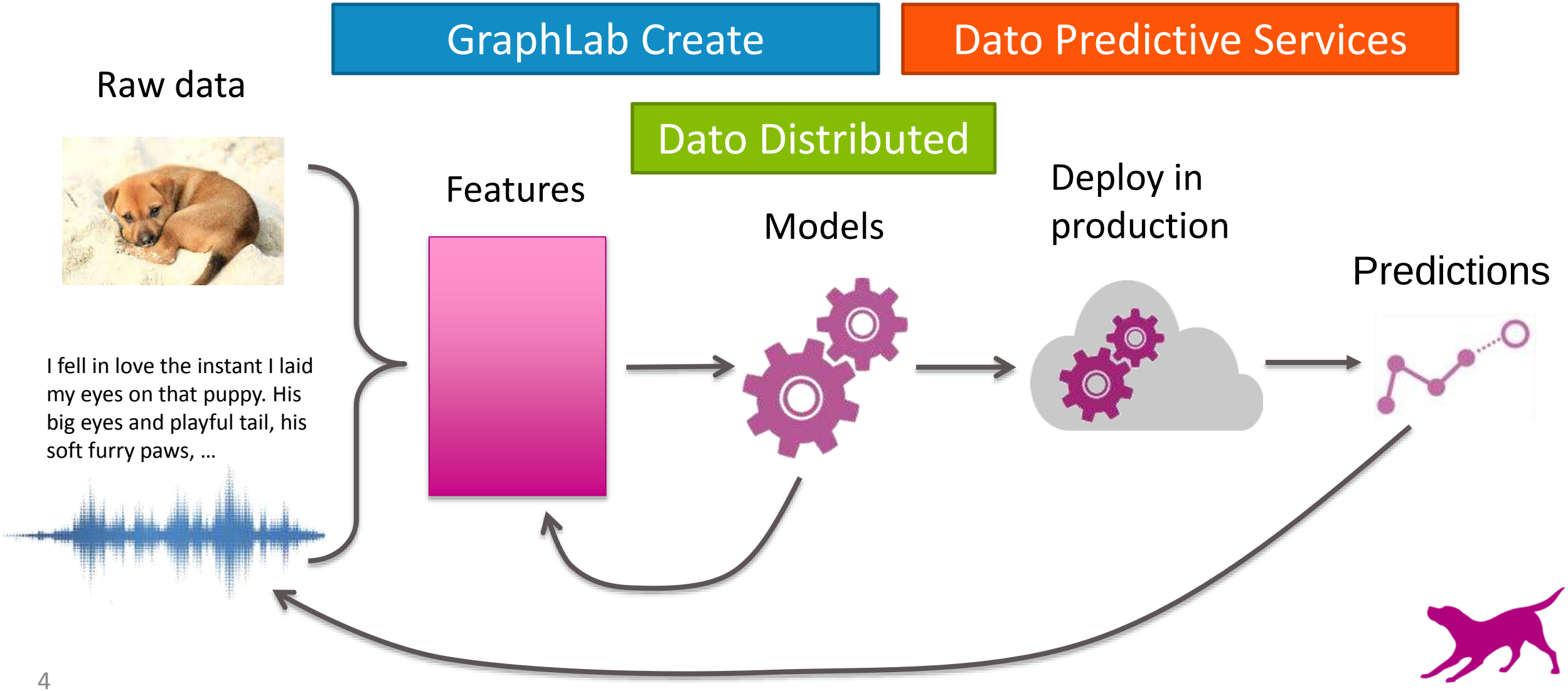
Dato

Build ML tools

Why machine learning?

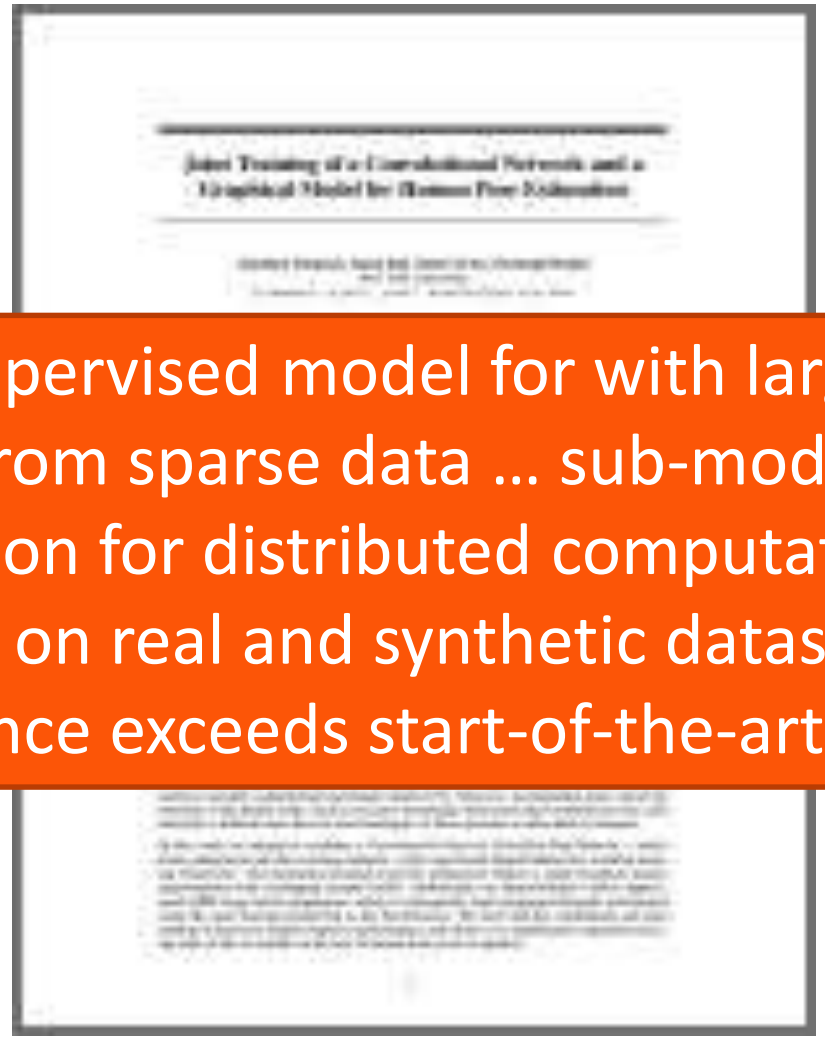
Model data.
Make predictions.
Build intelligent applications.

Machine learning pipeline



The ML Jargon Challenge

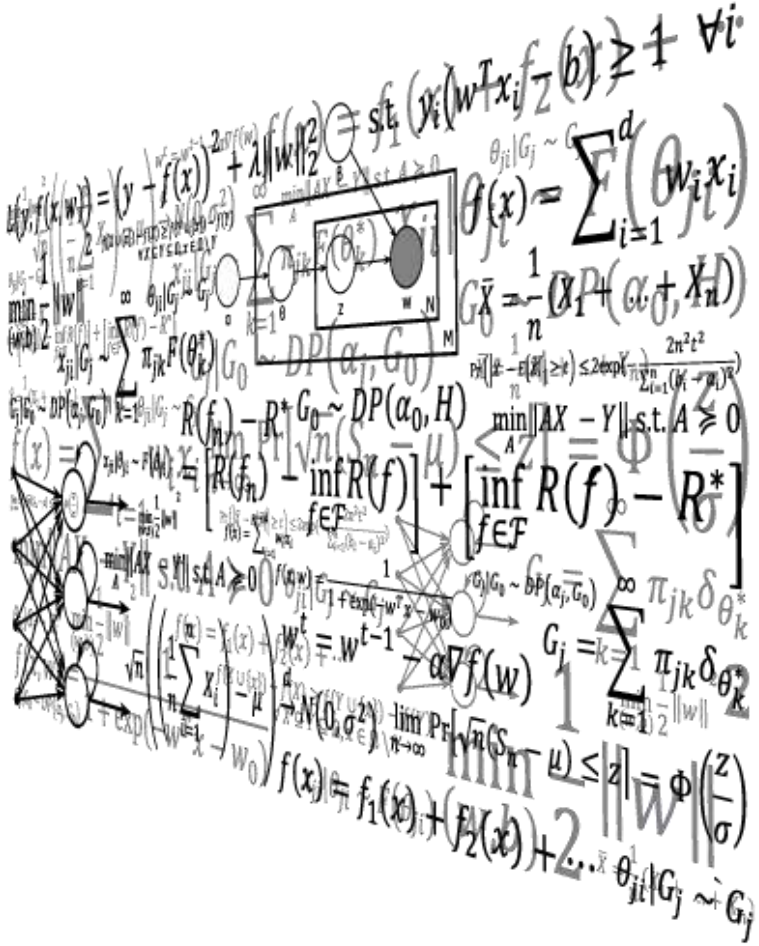
Typical machine learning paper



… semi-supervised model for with large-scale learning from sparse data … sub-modular optimization for distributed computation… evaluated on real and synthetic datasets… performance exceeds start-of-the-art methods

What it looks like to ML researchers

What it looks like to normal people



What it’s like in practice

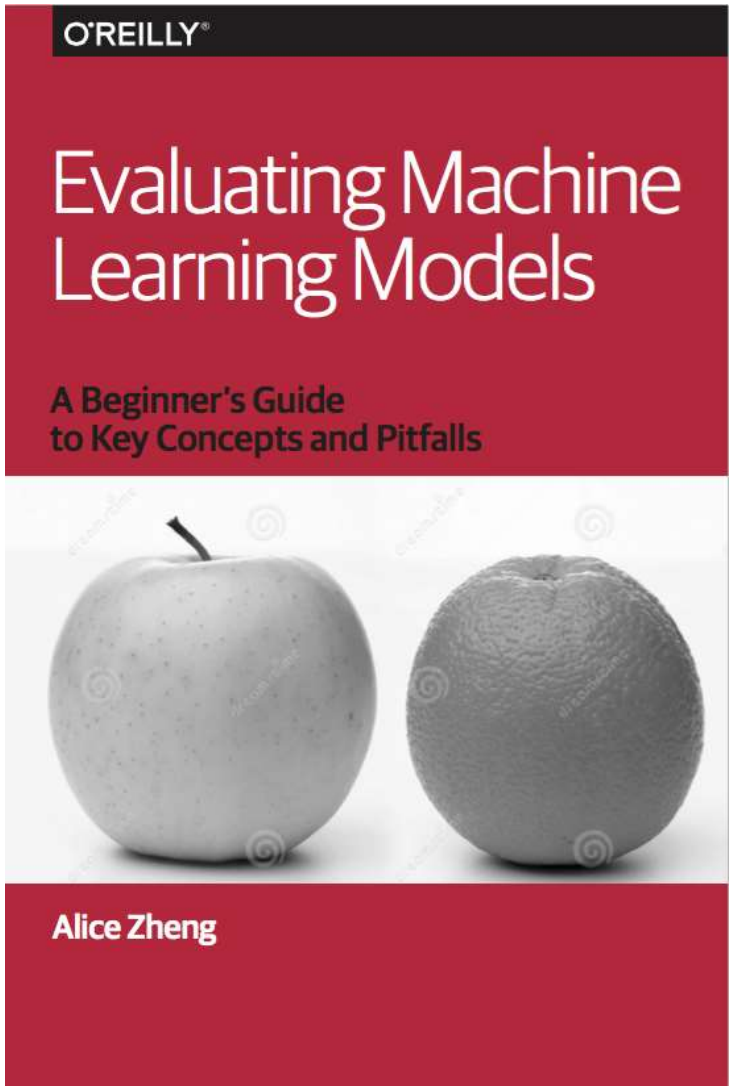
Brittle

Hard to tune

Doesn’t scale

Doesn’t solve *my* problem on *my* data

Achieve Machine Learning Zen



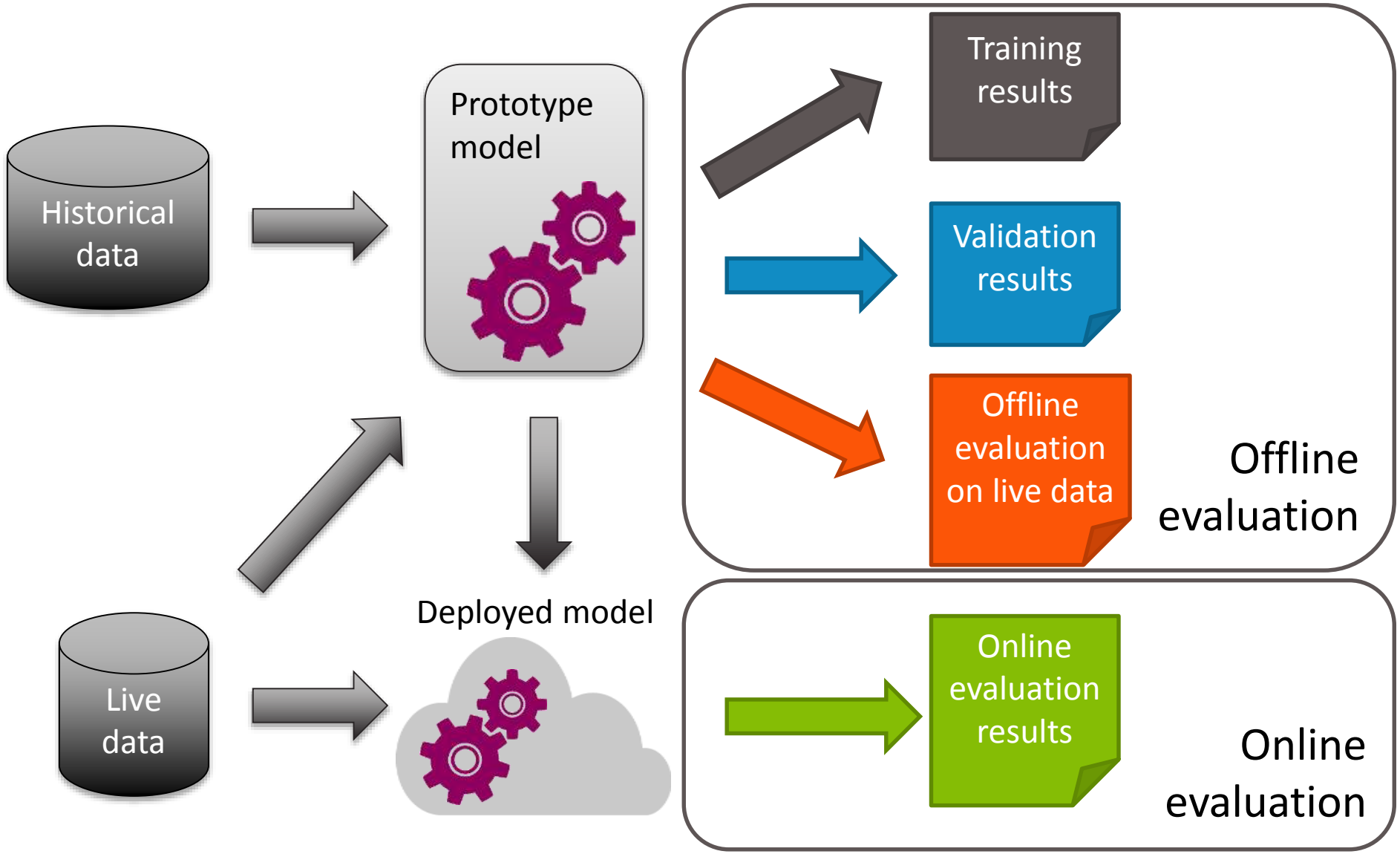
Why is evaluation important?

- So you know when you've succeeded
- So you know how much you've succeeded
- So you can decide when to stop
- So you can decide when to update the model

Basic questions for evaluation

- When to evaluate?
- What metric to use?
- On what data?

When to evaluate



Evaluation Metrics

Types of evaluation metric

- Training metric
- Validation metric
- Tracking metric
- Business metric

Example: recommender system

- Given data on which users liked which items, recommend other items to users
- Training metric
  - How well is it predicting the preference score?
  - Residual mean squared error: $(\text{actual} - \text{predicted})^2$
- Validation metric
  - Does it rank known preferences correctly?
  - Ranking loss

Example: recommender system

- Tracking metric
  - Does it rank items correctly, especially for top items?
  - Normalized Discounted Cumulative Gain (NDCG)
- Business metric
  - Does it increase the amount of time the user spends on the site/service?

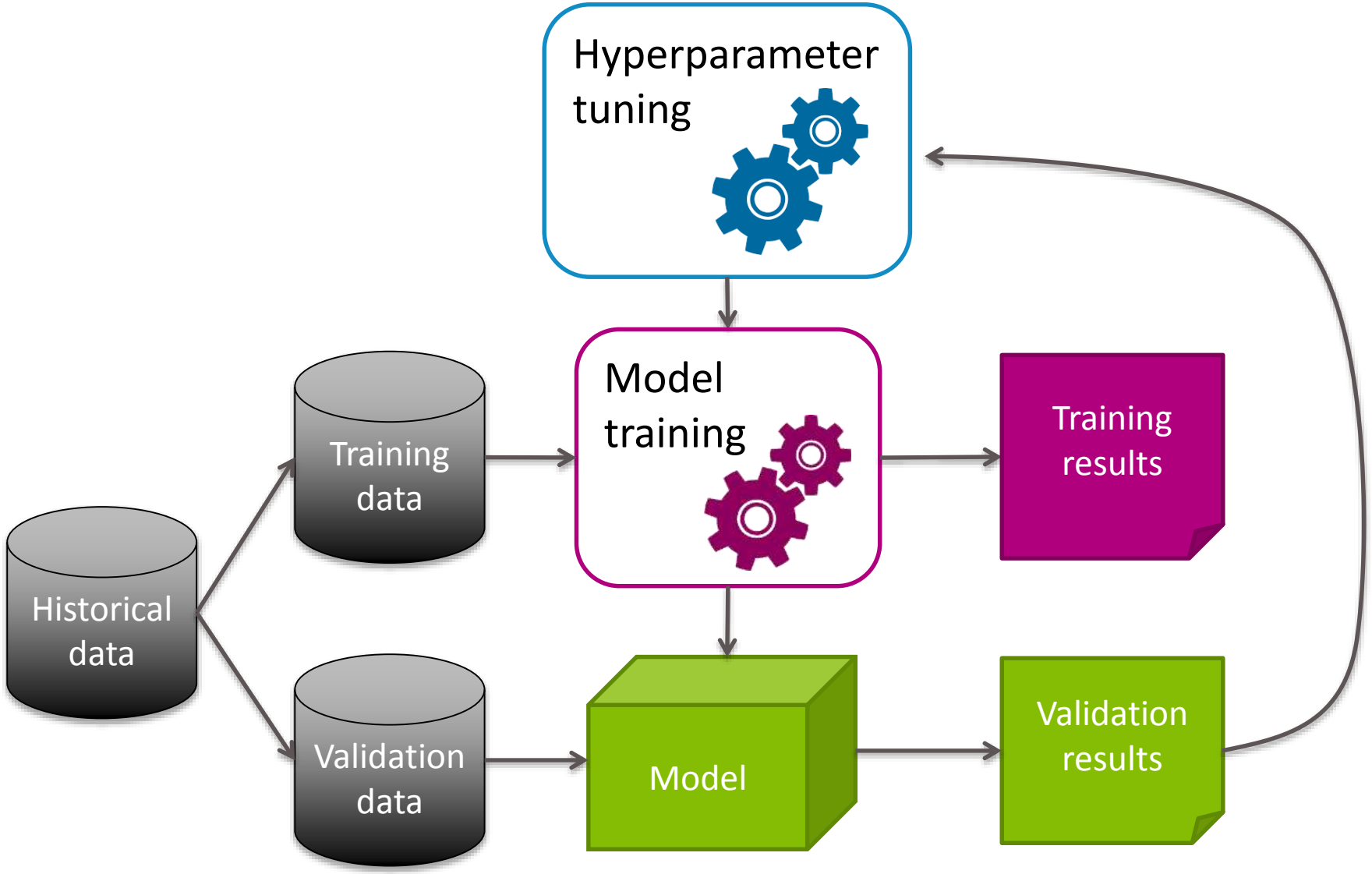
Dealing with metrics

- Many possible metrics at different stages
- Defining the right metric is an art
  - What's useful? What's feasible?
- Aligning the metrics will make everyone happier
  - Not always possible: cannot directly train model to optimize for user engagement

Okedokey Donkey

Model Selection and Tuning

Model Selection and Tuning



Key questions for model selection

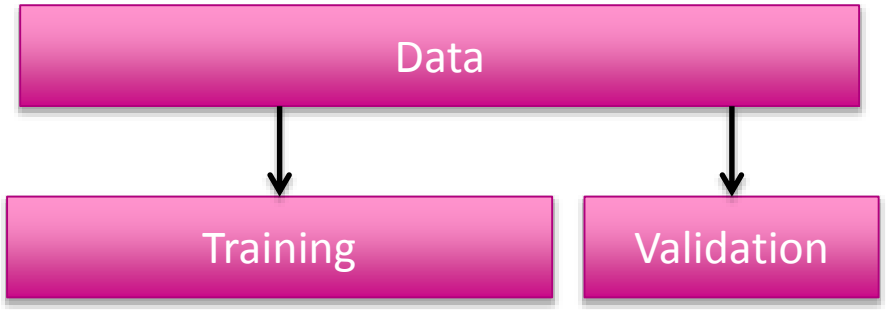
- What's validation?
- What's a hyperparameter and how do you tune it?

Model validation

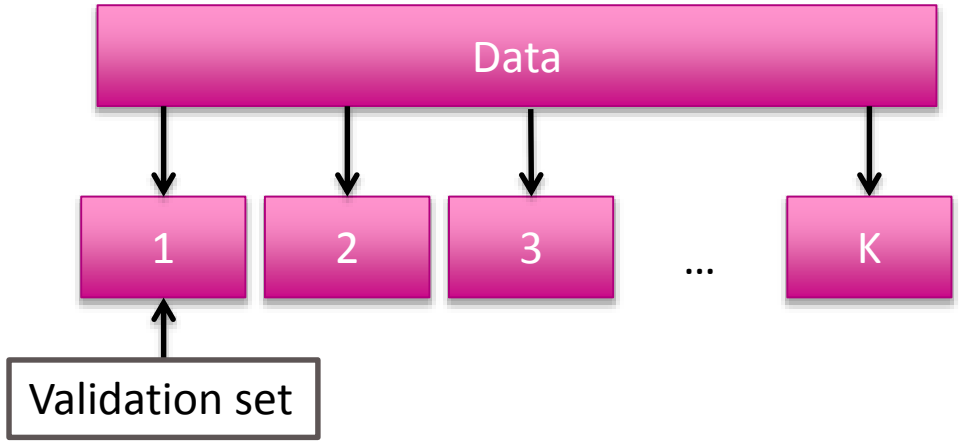
- Measure generalization error
  - How well the model works on new data
  - “New” data = data not used during training
- Train on one dataset, validate on another
- Where to find “new” data for validation?
  - Clever re-use of old data

Methods for simulating new data

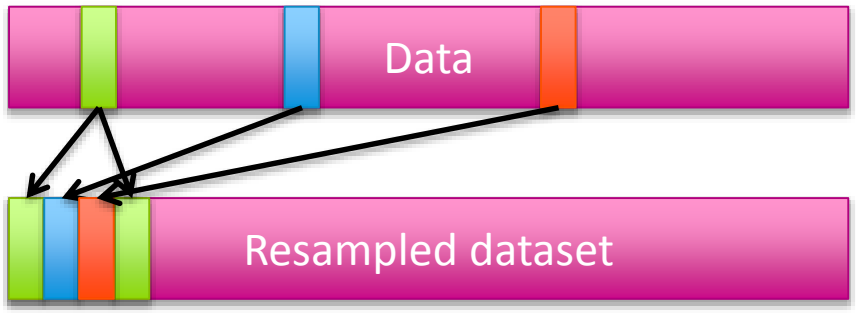
Hold-out validation



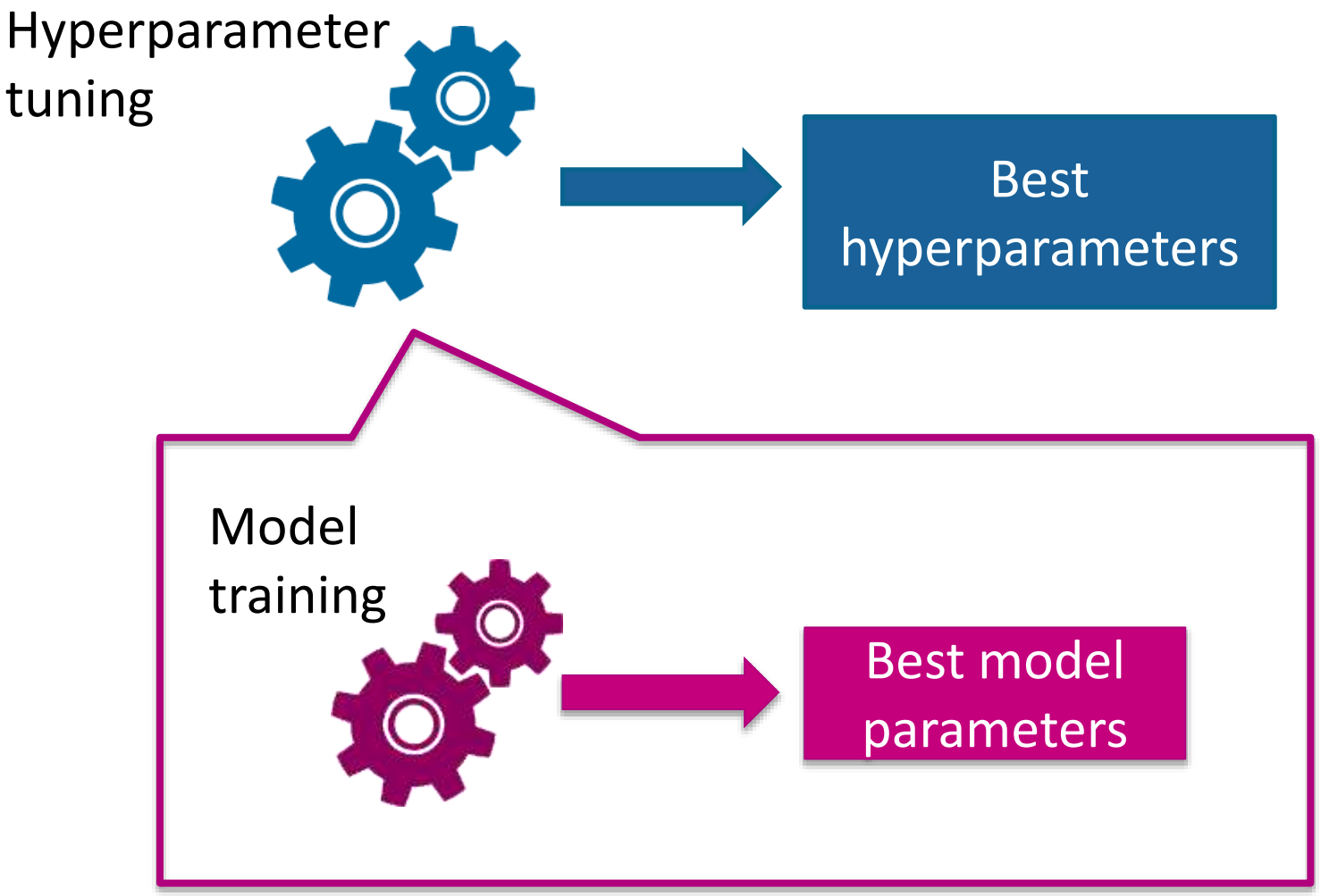
K-fold cross validation



Bootstrap resampling

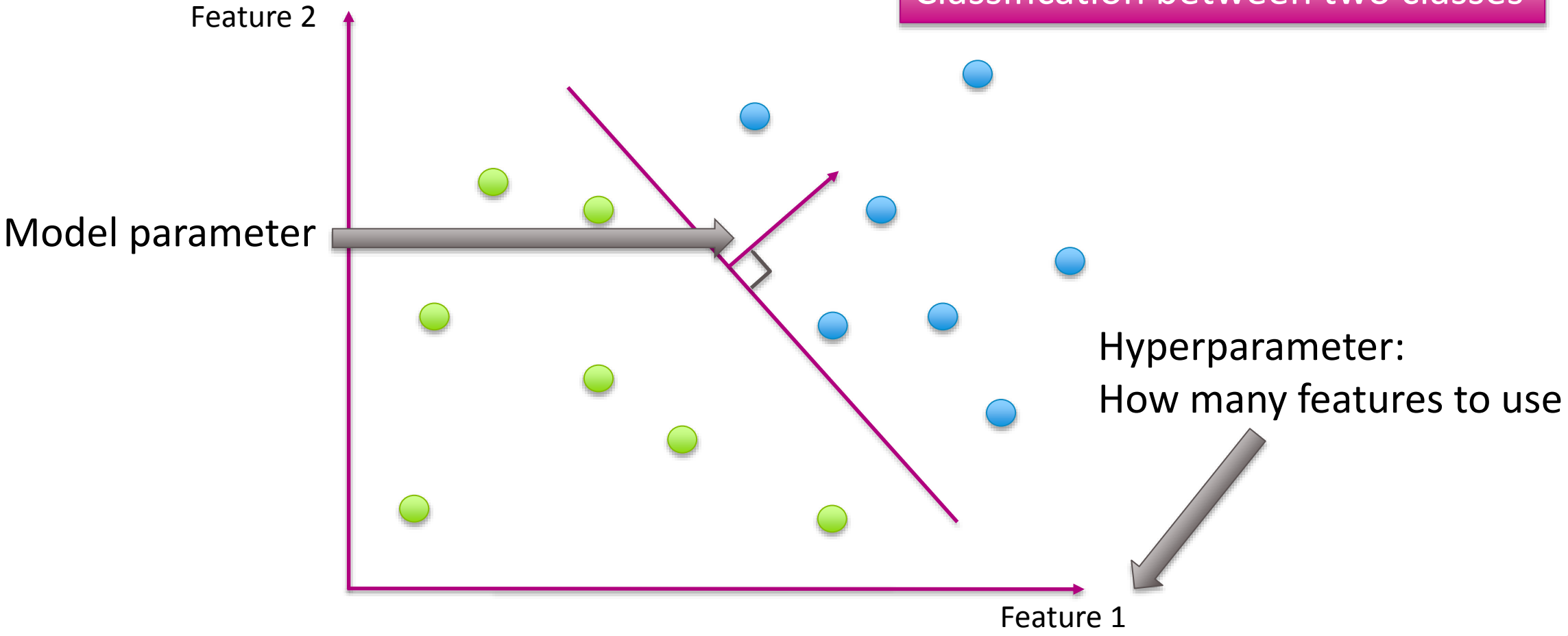


Hyperparameter tuning vs. model training



Hyperparameters != model parameters

Classification between two classes

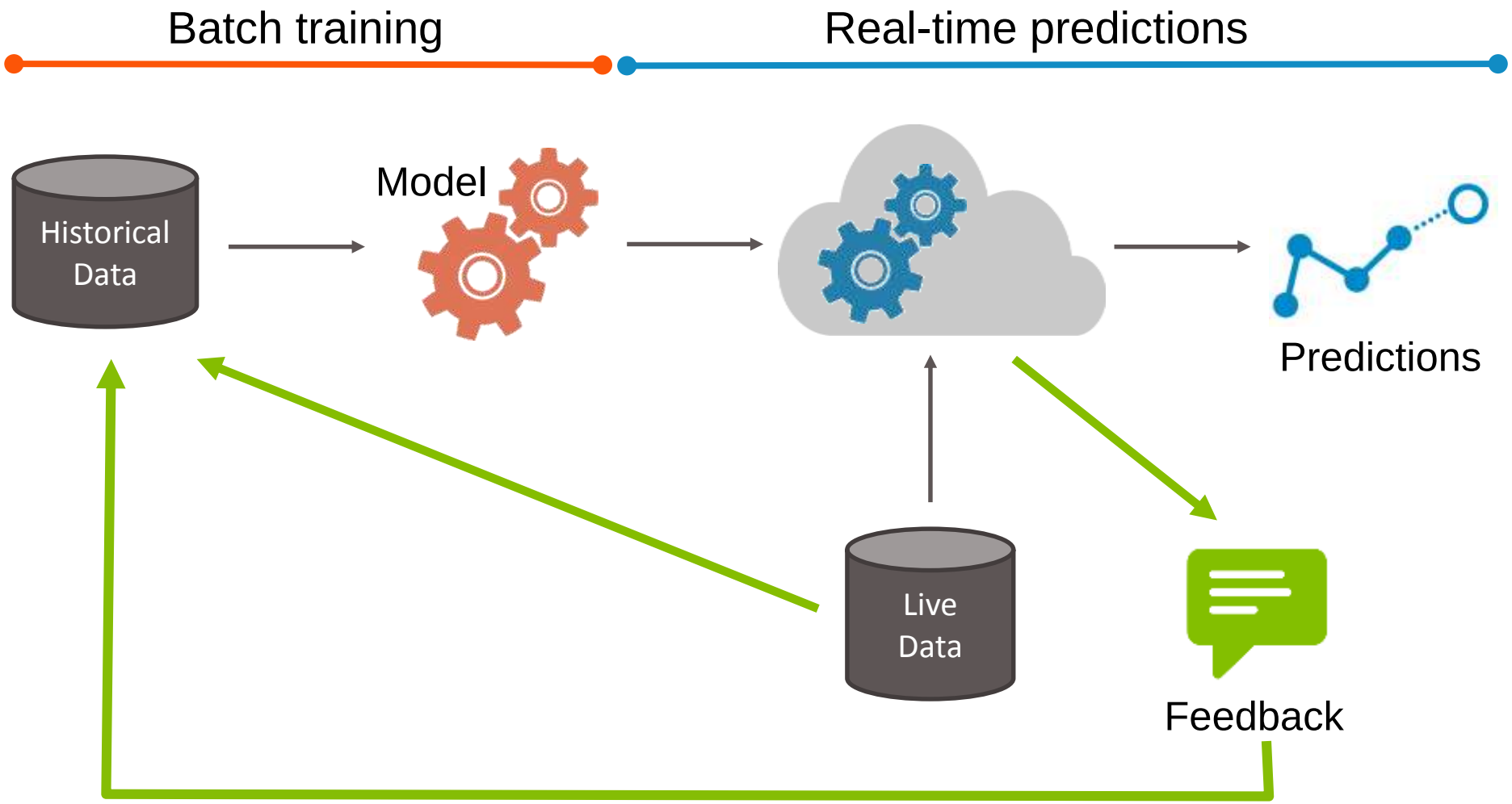


Why is hyperparameter tuning hard?

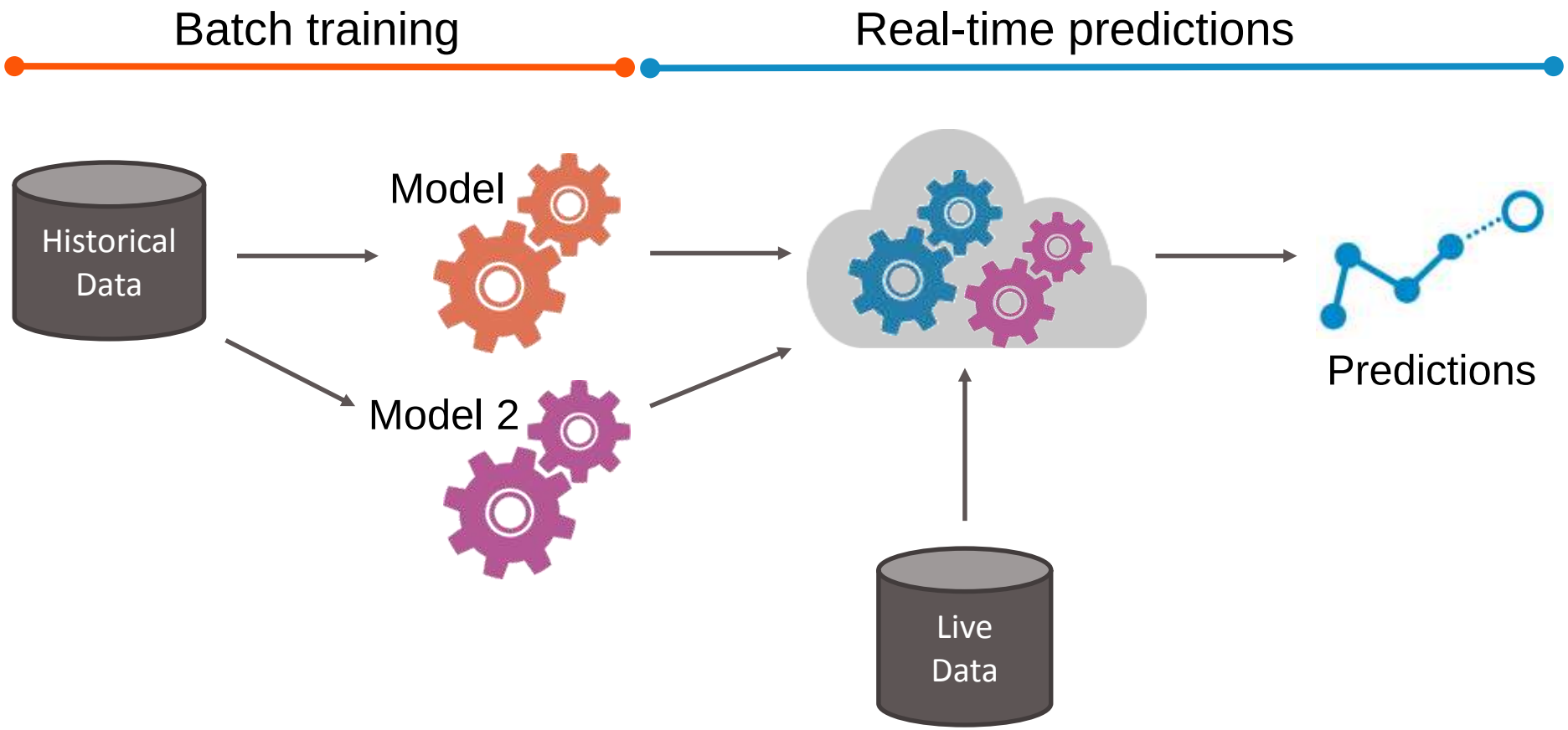
- Involves model training as a sub-process
  - Can’t optimize directly
- Methods:
  - Grid search
  - Random search
  - Smart search
    - Gaussian processes/Bayesian optimization
    - Random forests
    - Derivative-free optimization
    - Genetic algorithms

Online Evaluations

ML in production - 101



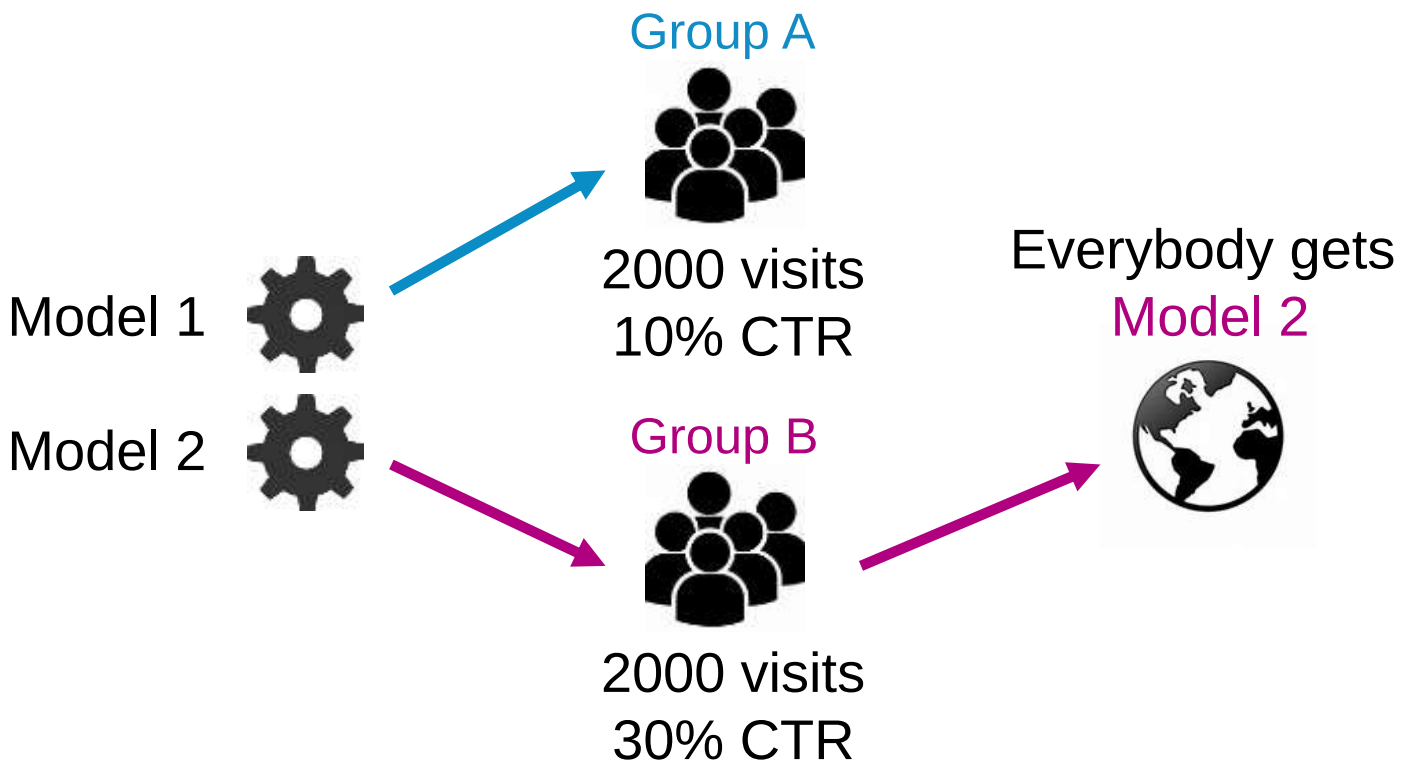
ML in production - 101



Why evaluate models online?

- Track real performance of model over time
- Decide which model to use when

Choosing between ML models



Strategy 1: A/B testing—select the best model and use it all the time

Choosing between ML models

A statistician walks into a casino…

Multi-armed bandits

Pay-off $1:$1000

Play this 5% of the time

Pay-off $1:$200

Play this 85% of the time

Pay-off $1:$500

Play this 10% of the time

Choosing between ML models

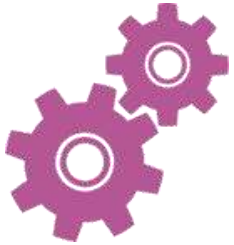
A statistician walks into an ML production environment

Model 1

Pay-off $1:$1000

Use this 5% of the time (Exploration)

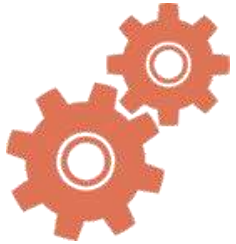
Model 2



Pay-off $1:$200

Use this 85% of the time (Exploitation)

Model 3



Pay-off $1:$500

Use this 10% of the time (Exploration)

MAB vs. A/B testing

Why MAB?

- Continuous optimization, “set and forget”
- Maximize overall reward

Why A/B test?

- Simple to understand
- Single winner
- *Tricky* to do right

That’s not all, folks!

Read the details

- Blog posts: http://blog.dato.com/topic/machine-learning-primer
- Report: http://oreil.ly/1L7dS4a

- Dato is hiring! jobs@dato.com

alicez@dato.com @RainyData