What's next for ML & you

Emily Fox & Carlos Guestrin
Machine Learning Specialization
University of Washington

©2015 Emily Fox & Carlos Guestrin

Deploying an ML service

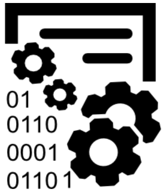
What is Production?

Serving live predictions

Deployment

Measuring quality of deployed models

Evaluation



Choosing between deployed models

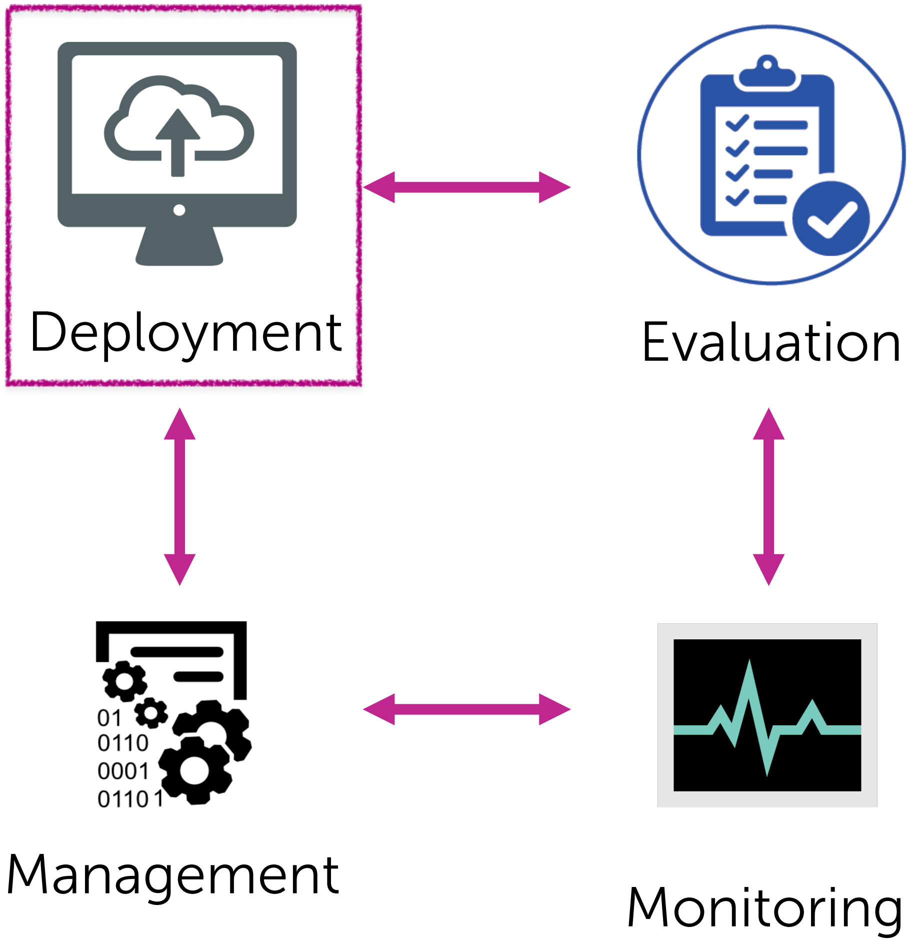
Management

Tracking model quality & operations

Monitoring

©2015 Emily Fox & Carlos Guestrin

Lifecycle of ML in Production



©2015 Emily Fox & Carlos Guestrin

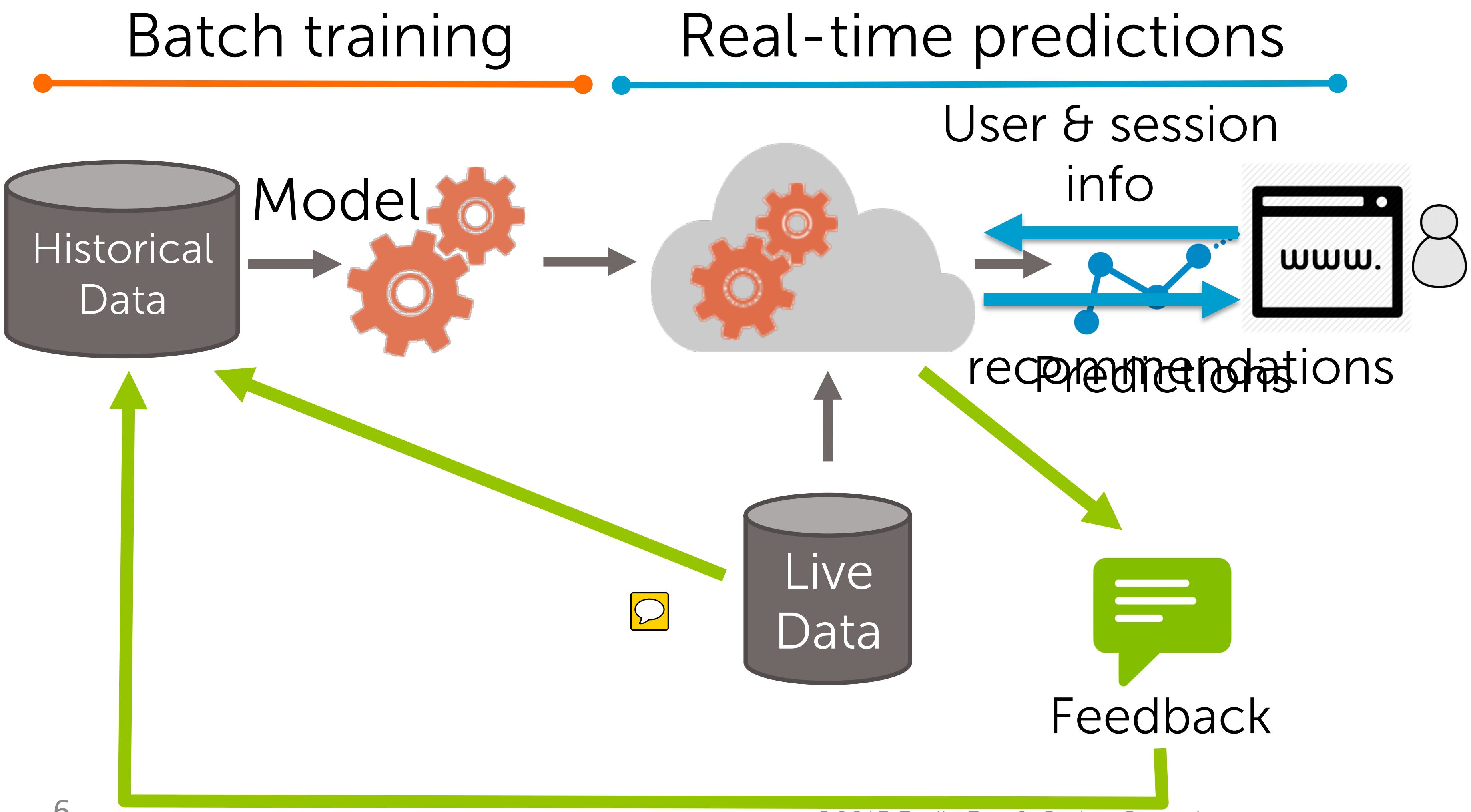
The Setup…

Suppose we are building a website with product recommendations, trained using user reviews.

- 34.6M reviews
- 2.4M products
- 6.6M users

©2015 Emily Fox & Carlos Guestrin

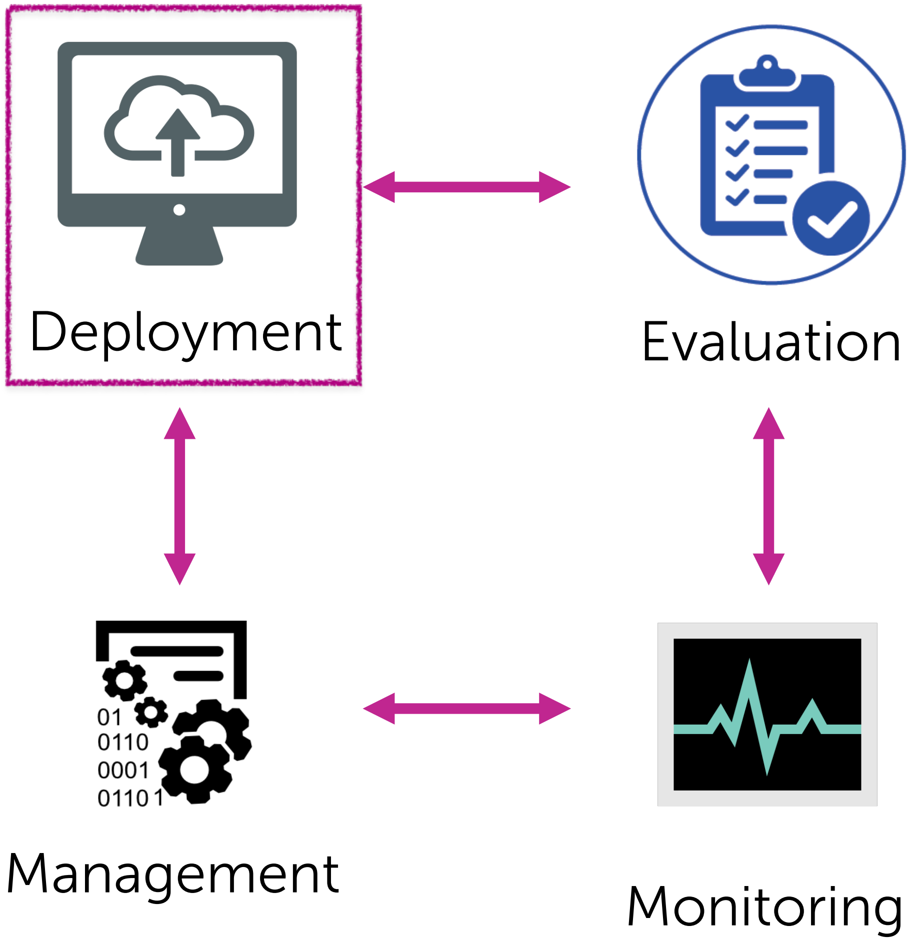
Deployment System



©2015 Emily Fox & Carlos Guestrin

What happens after (initial) deployment

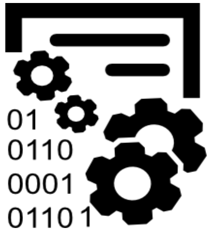
Lifecycle of ML in Production



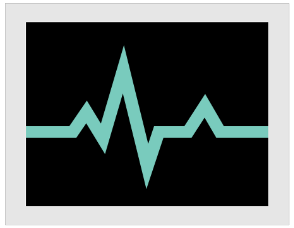
©2015 Emily Fox & Carlos Guestrin

After deployment

Evaluation



Management



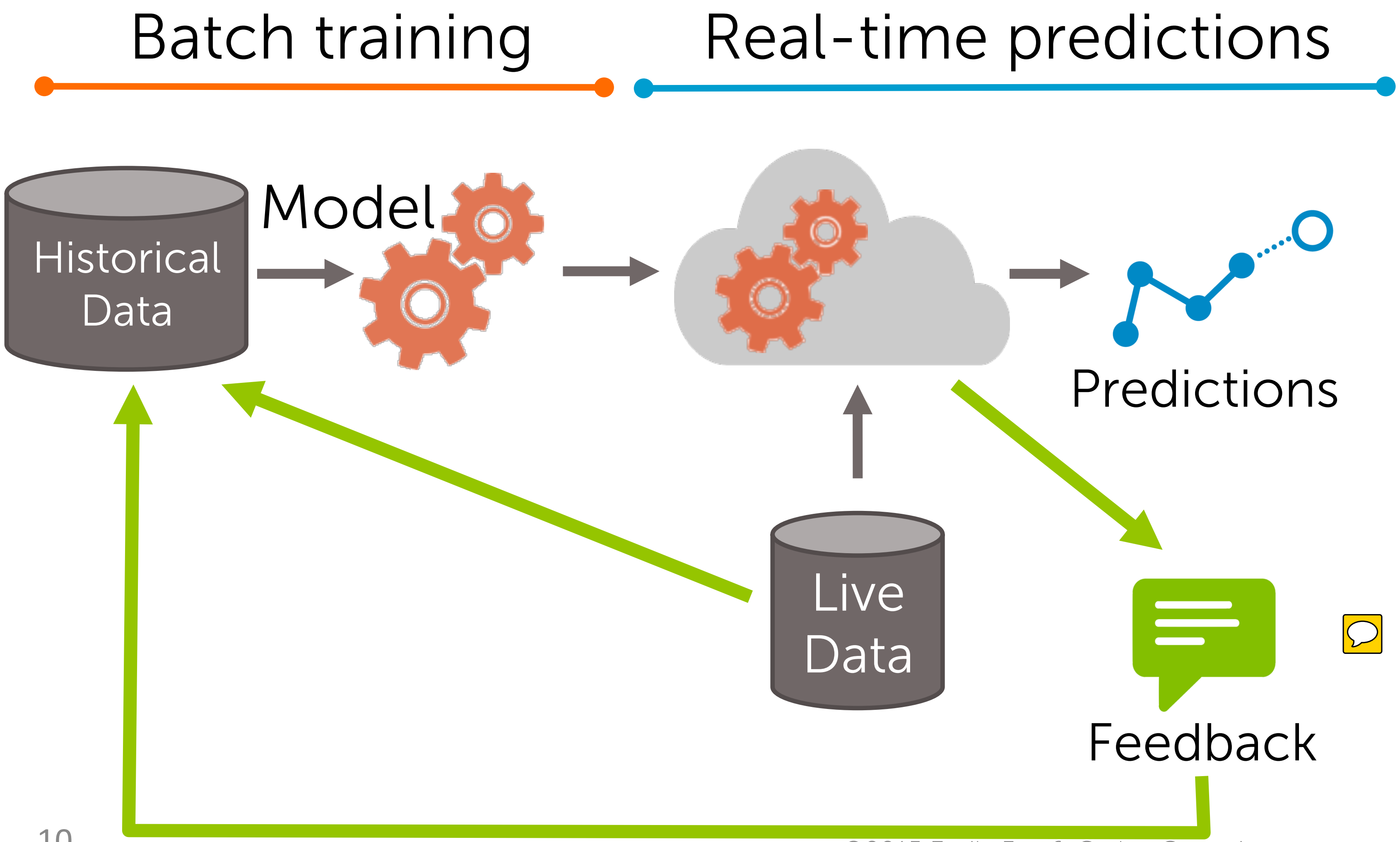
Monitoring

Evaluate and track metrics over time

React to feedback from deployed models

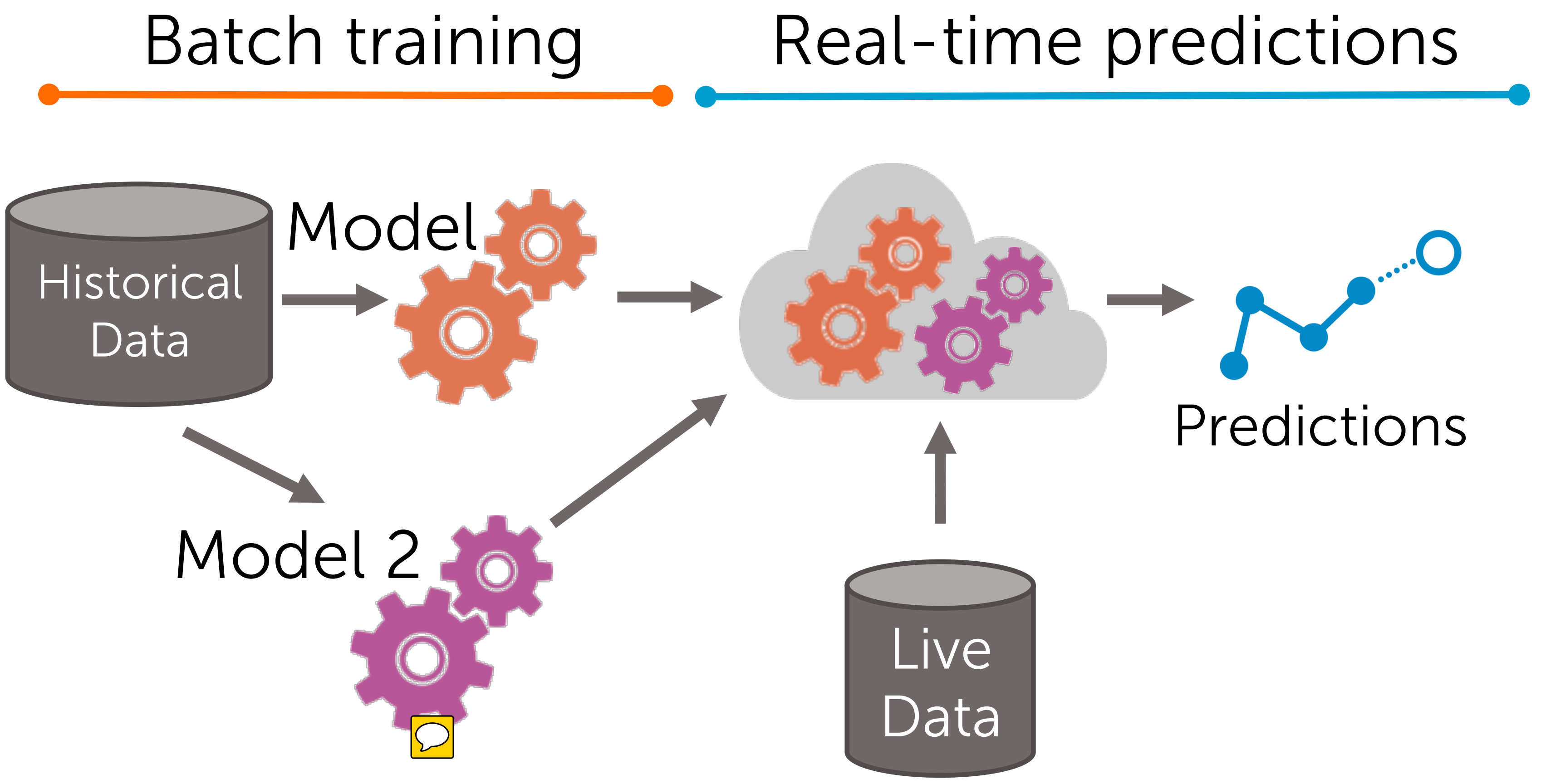
©2015 Emily Fox & Carlos Guestrin

Feedback loop for ML in production



©2015 Emily Fox & Carlos Guestrin

Learning new, alternative models



©2015 Emily Fox & Carlos Guestrin

Key questions

- When to update a model?
- How to choose between existing models?
- Answer: continuous evaluation and testing

©2015 Emily Fox & Carlos Guestrin

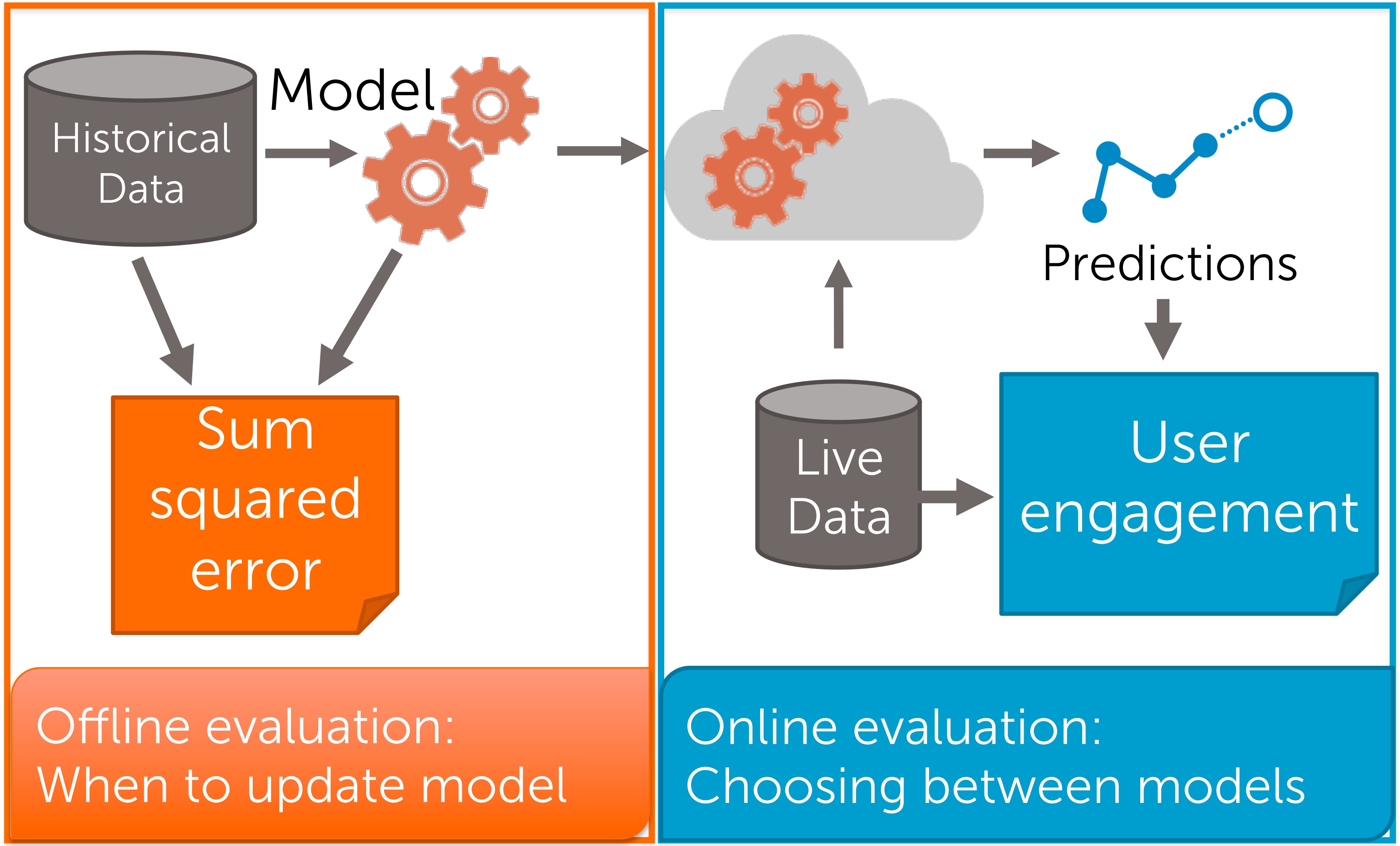
What is evaluation?

What data?

Which metric?

©2015 Emily Fox & Carlos Guestrin

Evaluating a recommender



©2015 Emily Fox & Carlos Guestrin

Updating ML models

Why update?

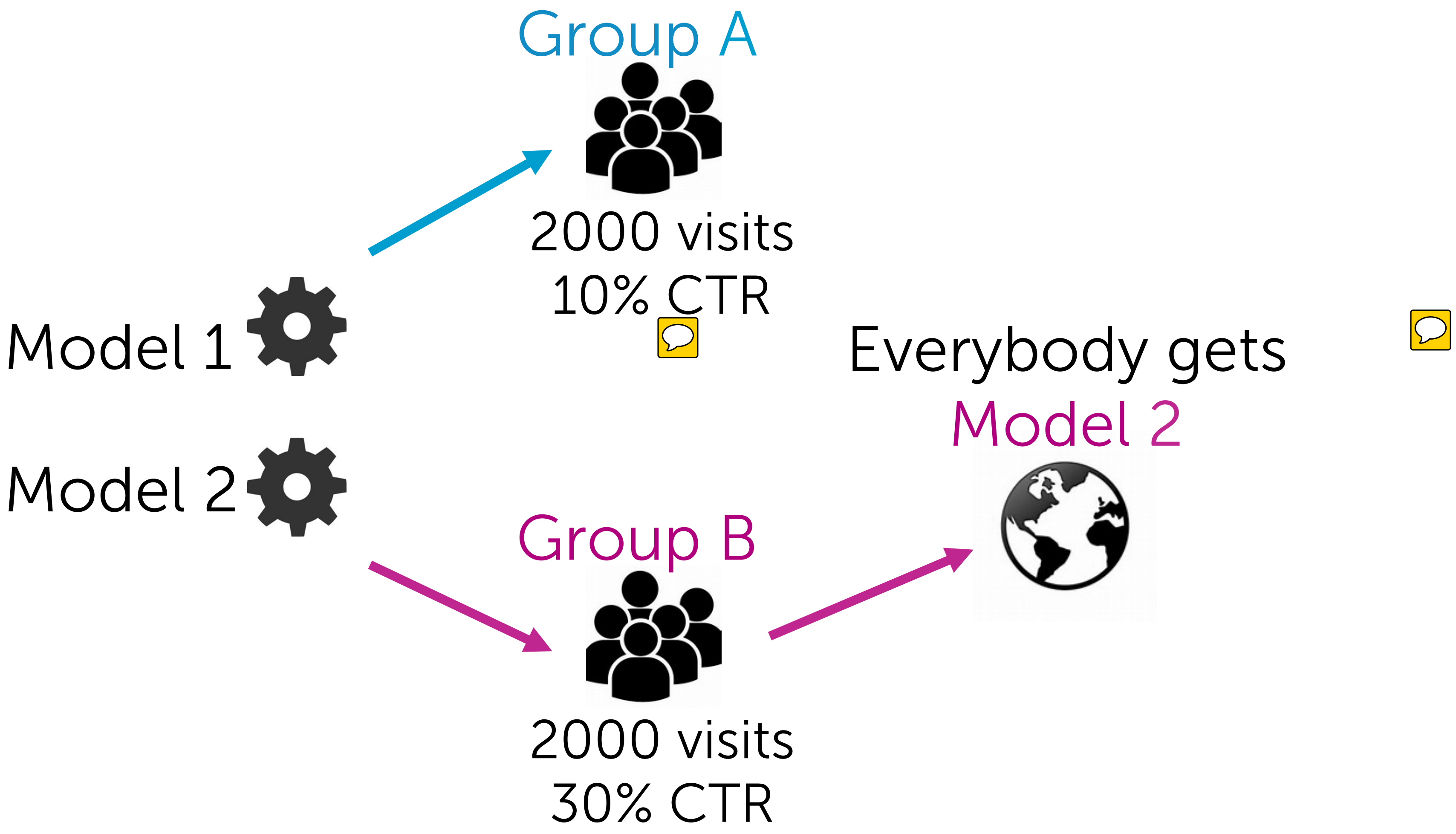
- Trends and user tastes change over time
- Model performance drops

When to update?

- Track statistics of data over time
- Monitor both offline & online metrics
- Update when offline metric diverges from online metrics or not achieving desired targets

©2015 Emily Fox & Carlos Guestrin

A/B Testing: Choosing between ML models



©2015 Emily Fox & Carlos Guestrin

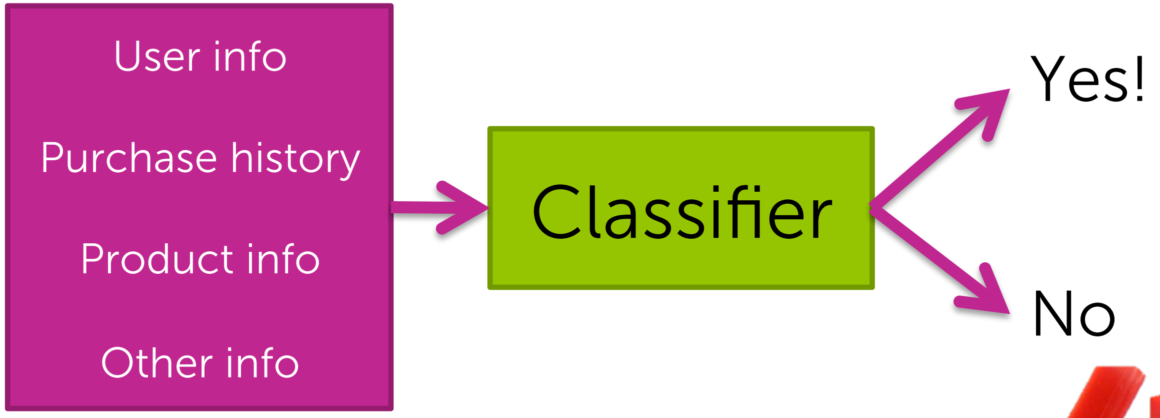
Other production considerations

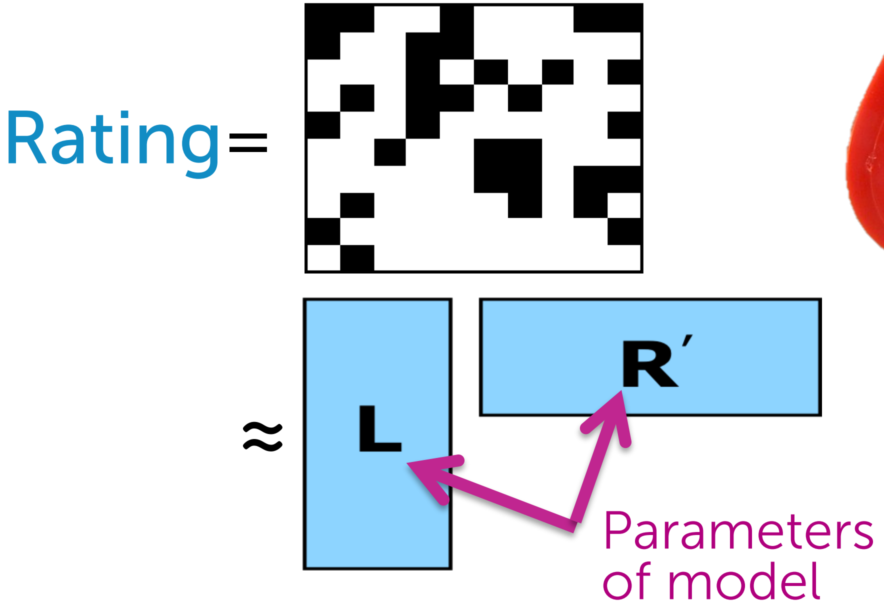
- A/B testing caveats
  - Also multi-armed bandits
- Versioning
- Provenance
- Dashboards
- Reports
- …

©2015 Emily Fox & Carlos Guestrin

Machine learning challenges

Open challenges: Model selection





©2015 Emily Fox & Carlos Guestrin

Open challenges: Feature engineering/representation

1	0	0	0	5	3	0	0	1	0	0	0	0

- Bag of word raw counts?
- Normalize?
- tf-idf? (which version???)
- Bigrams
- Trigrams
- …

©2015 Emily Fox & Carlos Guestrin

Open challenges: Scaling

Data is getting big…

facebook

flickr

WIKIPEDIA
The Free Encyclopedia

yelp

Insta

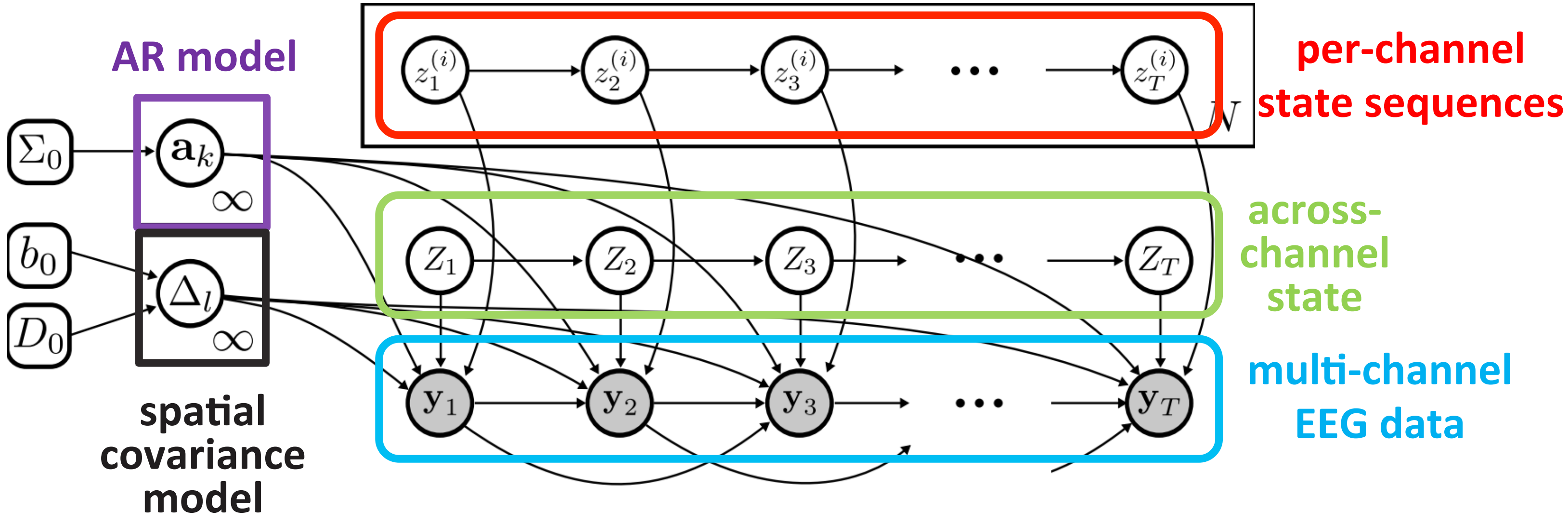
YouTube

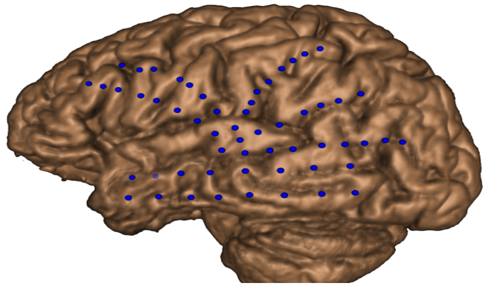
amazon

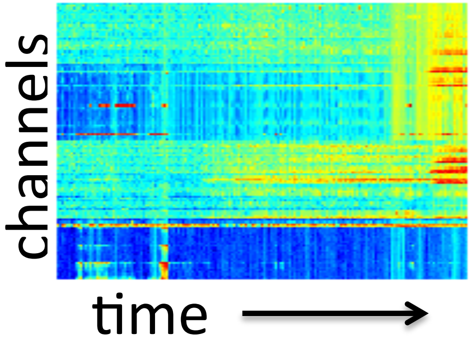
©2015 Emily Fox & Carlos Guestrin

Open challenges: Scaling

Concurrently, models are getting big…

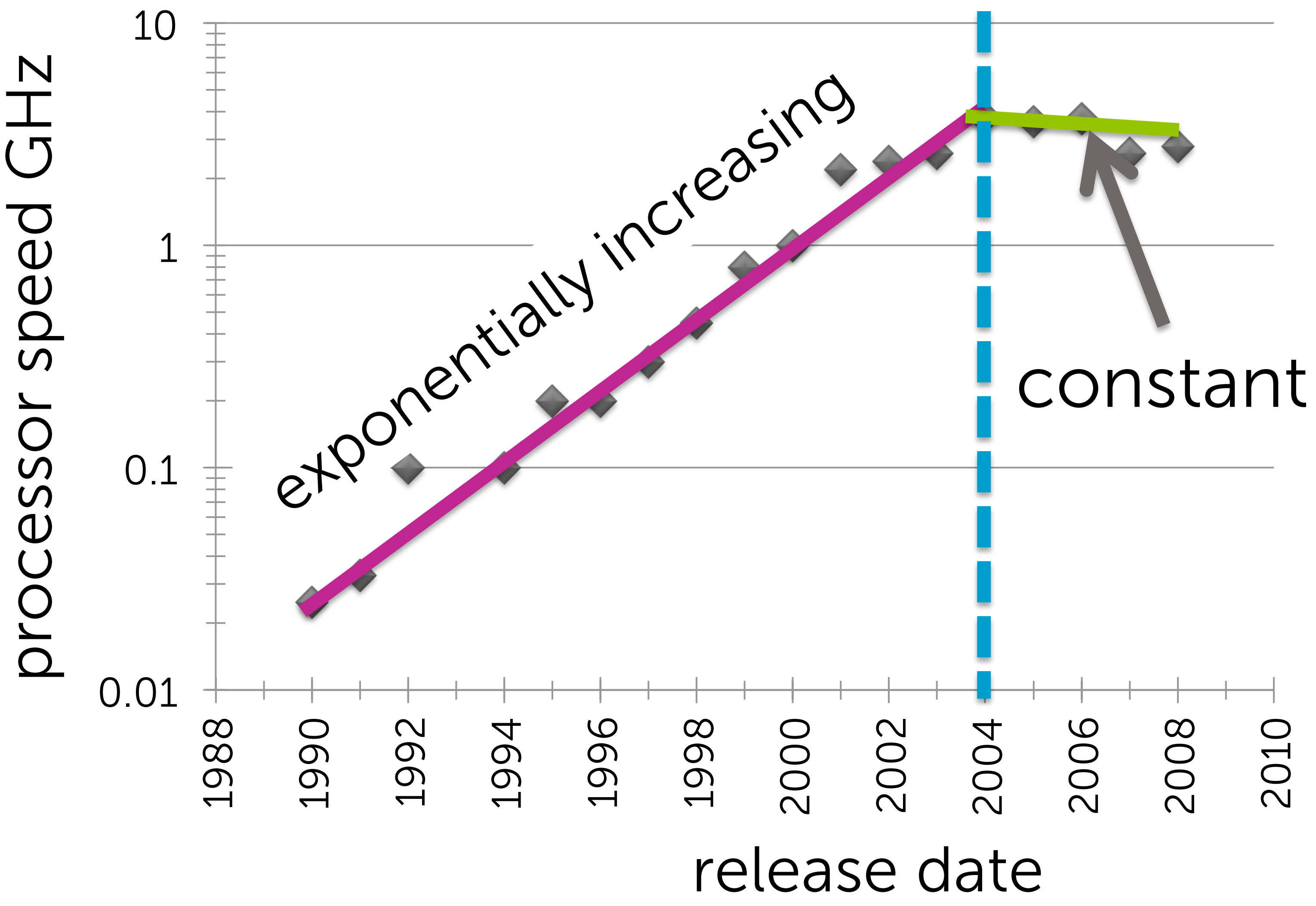






©2015 Emily Fox & Carlos Guestrin

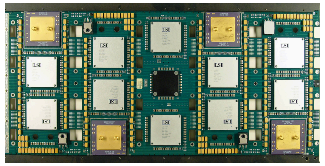
CPUs stopped getting faster…



©2015 Emily Fox & Carlos Guestrin

ML in the context of parallel architectures


GPUs


Multicore


Clusters

amazon web services™

Clouds

Supercomputers

But scalable ML in these systems is **hard**, especially in terms of:

1. Programmability
2. Data distribution
3. Failures

©2015 Emily Fox & Carlos Guestrin

What's ahead in this specialization

2. Regression

Case study: Predicting house prices

Models

- Linear regression
- Regularization:
  Ridge (L2), Lasso (L1)

Including many features:

- Square feet
- # bathrooms
- # bedrooms
- Lot size
- Year built
- …

©2015 Emily Fox & Carlos Guestrin

2. Regression

Case study: Predicting house prices

Algorithms

- Gradient descent
- Coordinate descent

$$
\begin{aligned}
&\mathrm{RSS}(w_0,w_1) = \\
&\quad (\$_{\text{house }1}-[w_0+w_1 \text{sq.ft.}_{\text{house }1}])^2 \\
&+ (\$_{\text{house }2}-[w_0+w_1 \text{sq.ft.}_{\text{house }2}])^2 \\
&+ (\$_{\text{house }3}-[w_0+w_1 \text{sq.ft.}_{\text{house }3}])^2 \\
&+ \ldots \text{ [include all houses]}
\end{aligned}
$$

$$\downarrow$$

$$\hat{w}$$

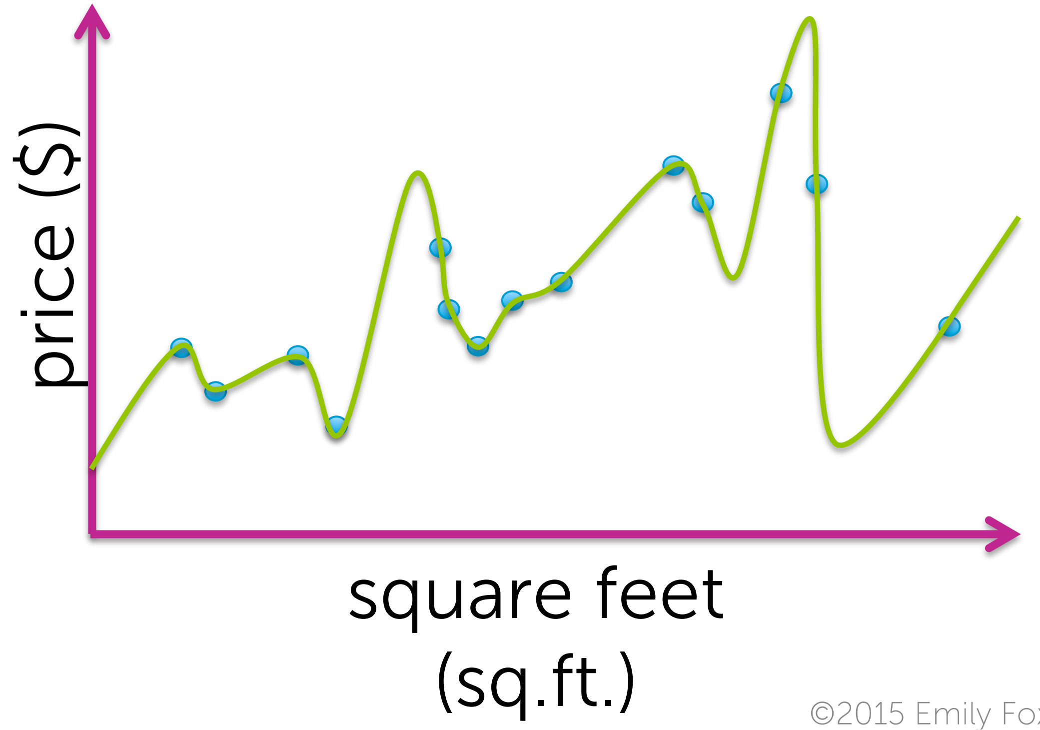
©2015 Emily Fox & Carlos Guestrin

2. Regression

Case study: Predicting house prices

Concepts

- Loss functions, bias-variance tradeoff, cross-validation, sparsity, overfitting, model selection



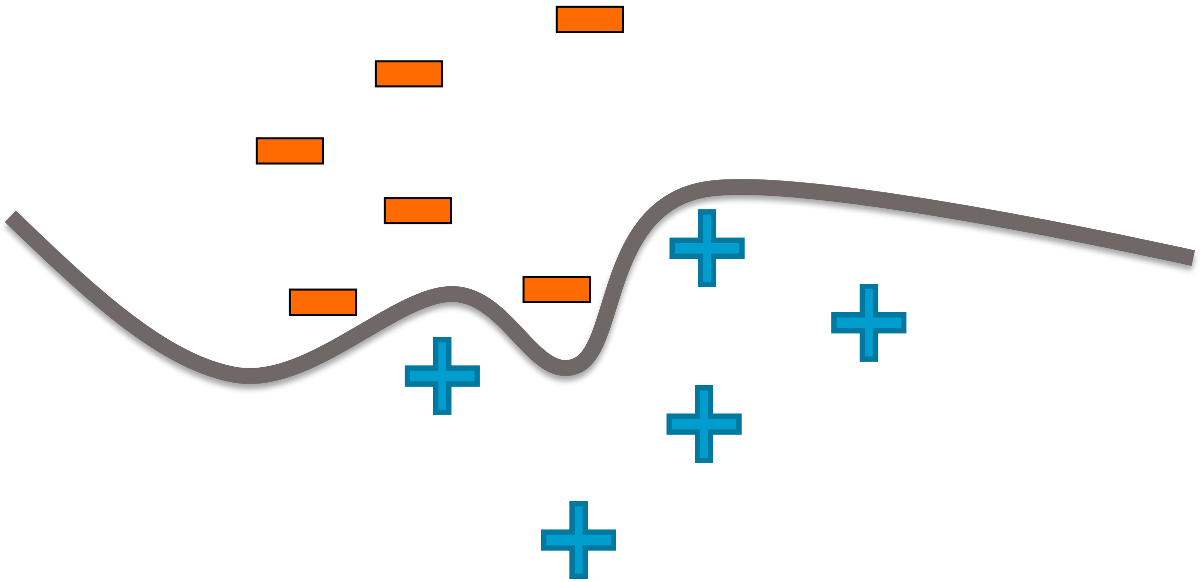
©2015 Emily Fox & Carlos Guestrin

3. Classification

Case study: Analyzing sentiment

Models

- Linear classifiers (logistic regression, SVMs, perceptron)
- Kernels
- Decision trees



©2015 Emily Fox & Carlos Guestrin

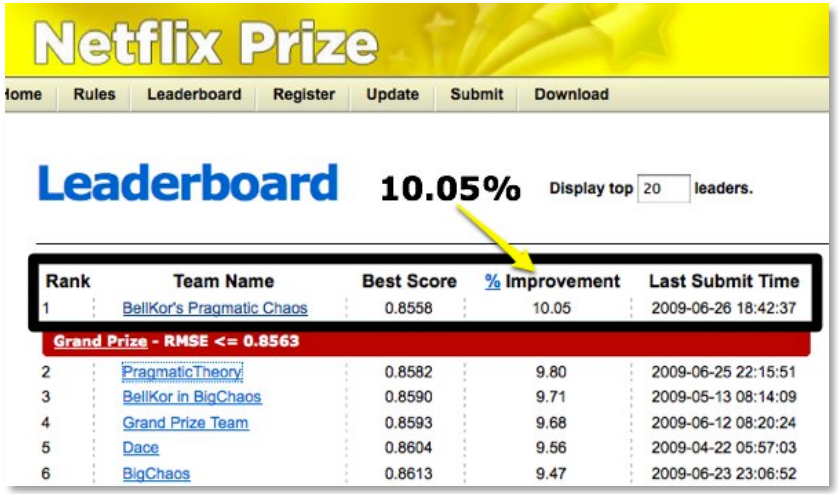
3. Classification

Case study: Analyzing sentiment

Algorithms

- Stochastic gradient descent
- Boosting

Squeezing last bit of accuracy by blending models



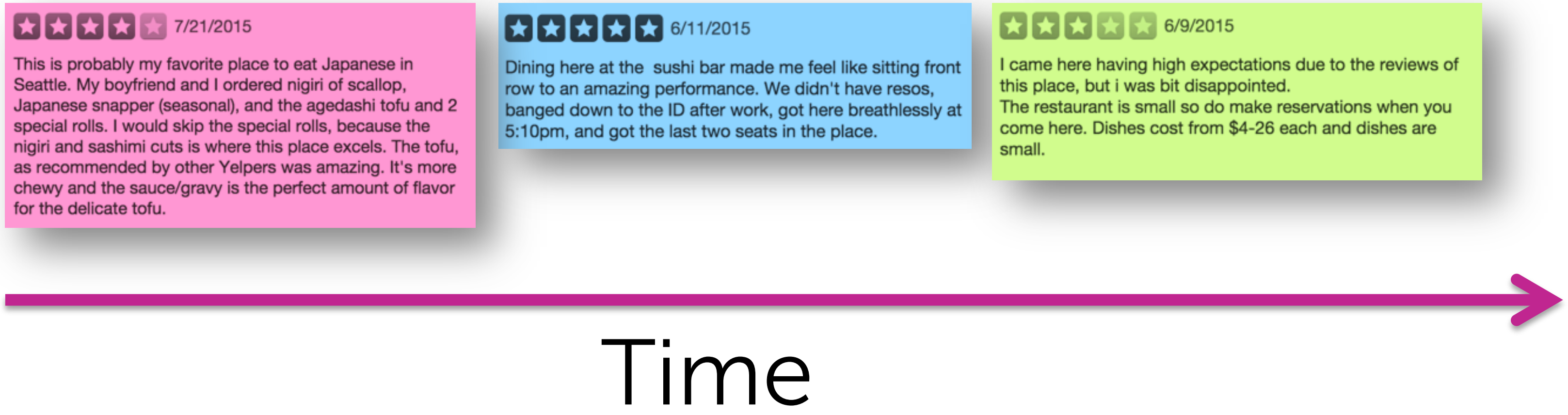
©2015 Emily Fox & Carlos Guestrin

3. Classification

Case study: Analyzing sentiment

Concepts

- Decision boundaries, MLE, ensemble methods, random forests, CART, online learning



©2015 Emily Fox & Carlos Guestrin

4. Clustering & Retrieval

Case study: Finding documents

Models

- Nearest neighbors
- Clustering, mixtures of Gaussians
- Latent Dirichlet allocation (LDA)

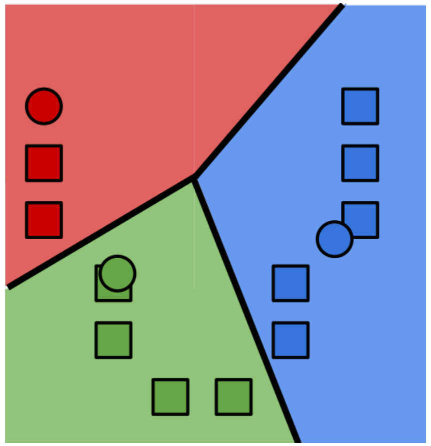
©2015 Emily Fox & Carlos Guestrin

4. Clustering & Retrieval

Case study: Finding documents

Algorithms

- KD-trees, locality-sensitive hashing (LSH)
- K-means
- Expectation-maximization (EM)



©2015 Emily Fox & Carlos Guestrin

4. Clustering & Retrieval

Case study: Finding documents

Concepts

- Distance metrics, approximation algorithms, hashing, sampling algorithms, scaling up with map-reduce

1	0	0	0	5	3	0	0	1	0	0	0	0

1*3
+
5*2
= 13

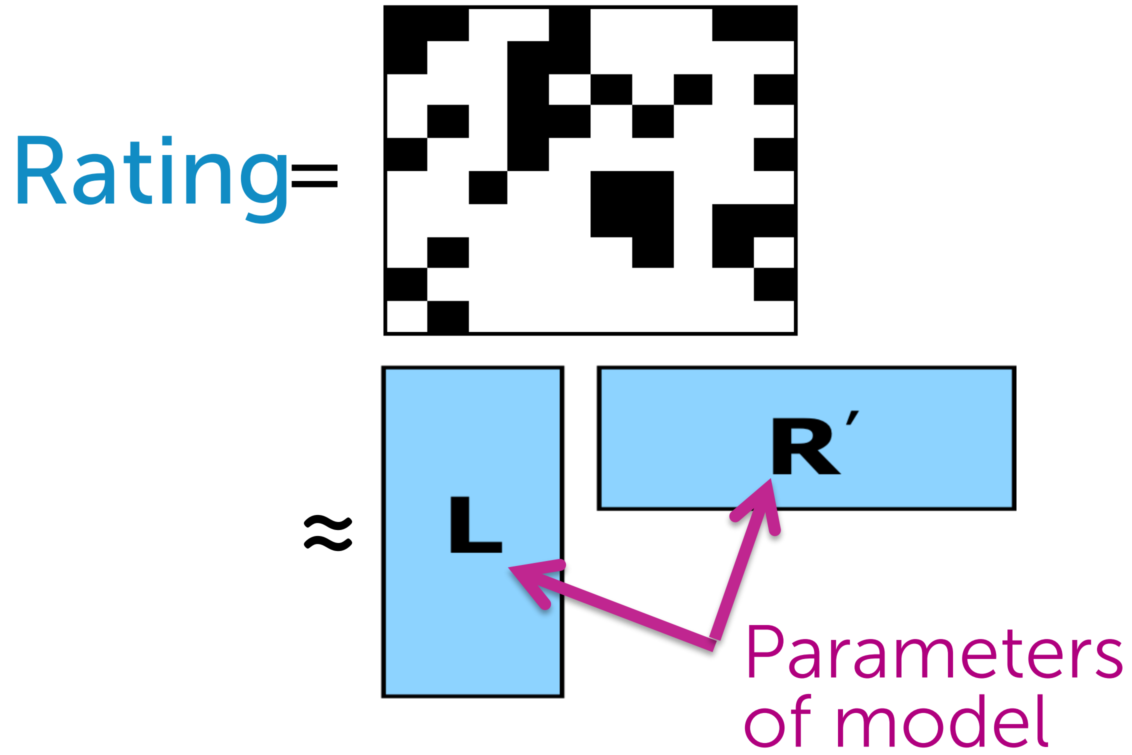
3	0	0	0	2	0	0	1	0	1	0	0	0

©2015 Emily Fox & Carlos Guestrin

5. Recommender Systems & Dimensionality Reduction *Case study: Recommending Products*

Models

- Collaborative filtering
- Matrix factorization
- PCA



©2015 Emily Fox & Carlos Guestrin

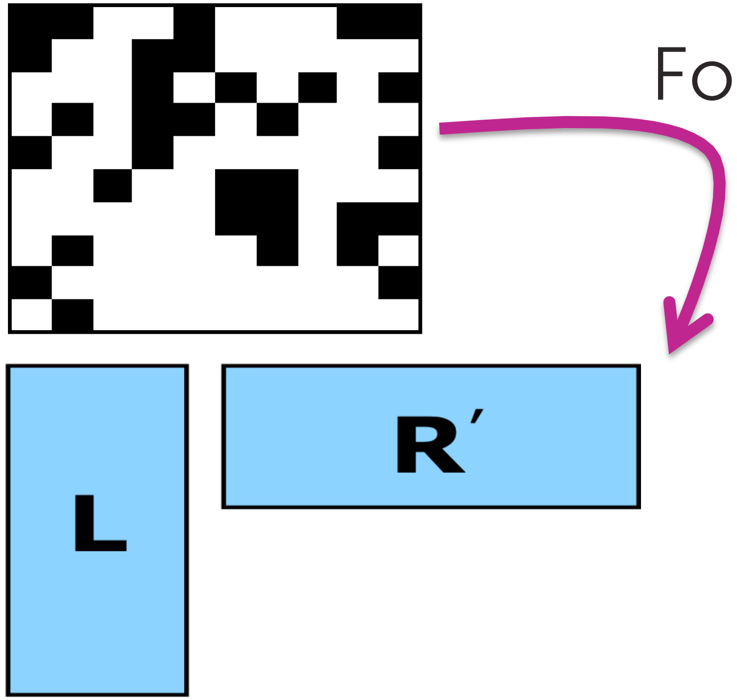
5. Matrix Factorization & Dimensionality Reduction

Case study: Recommending Products

Algorithms

- Coordinate descent
- Eigen decomposition
- SVD

Rating =



≈

Form estimates $\hat{L}_u$ and $\hat{R}_v$

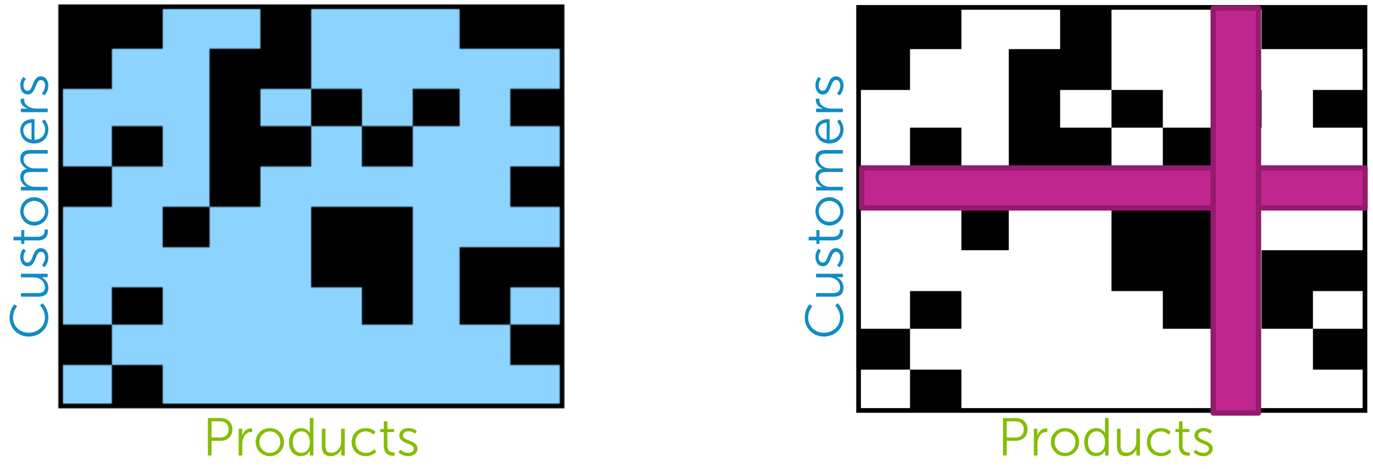
©2015 Emily Fox & Carlos Guestrin

5. Matrix Factorization & Dimensionality Reduction

Case study: Recommending Products

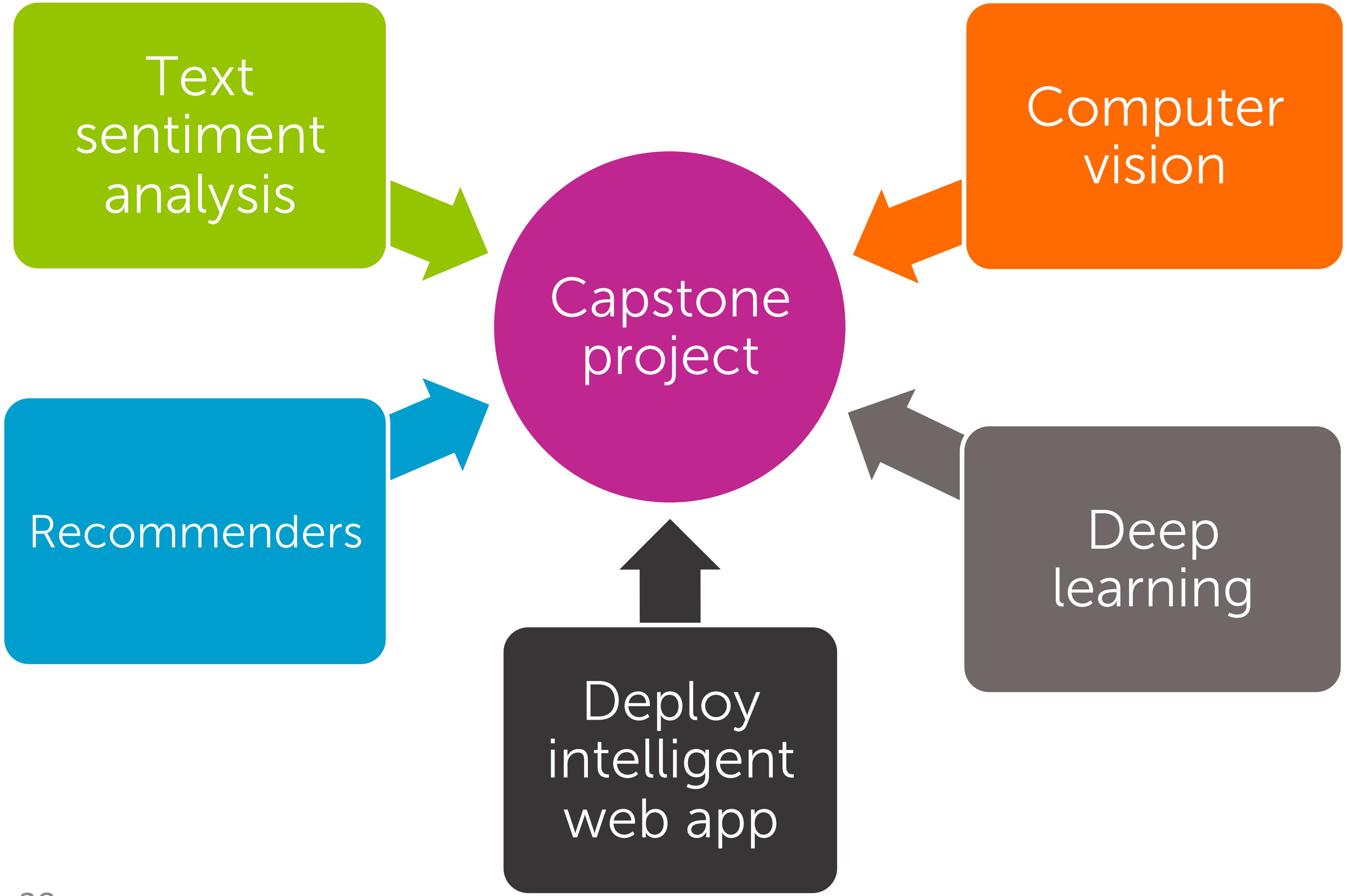
Concepts

- Matrix completion, eigenvalues, random projections, cold-start problem, diversity, scaling up



©2015 Emily Fox & Carlos Guestrin

6. Capstone: *Build and deploy an intelligent application with deep learning*



©2015 Emily Fox & Carlos Guestrin