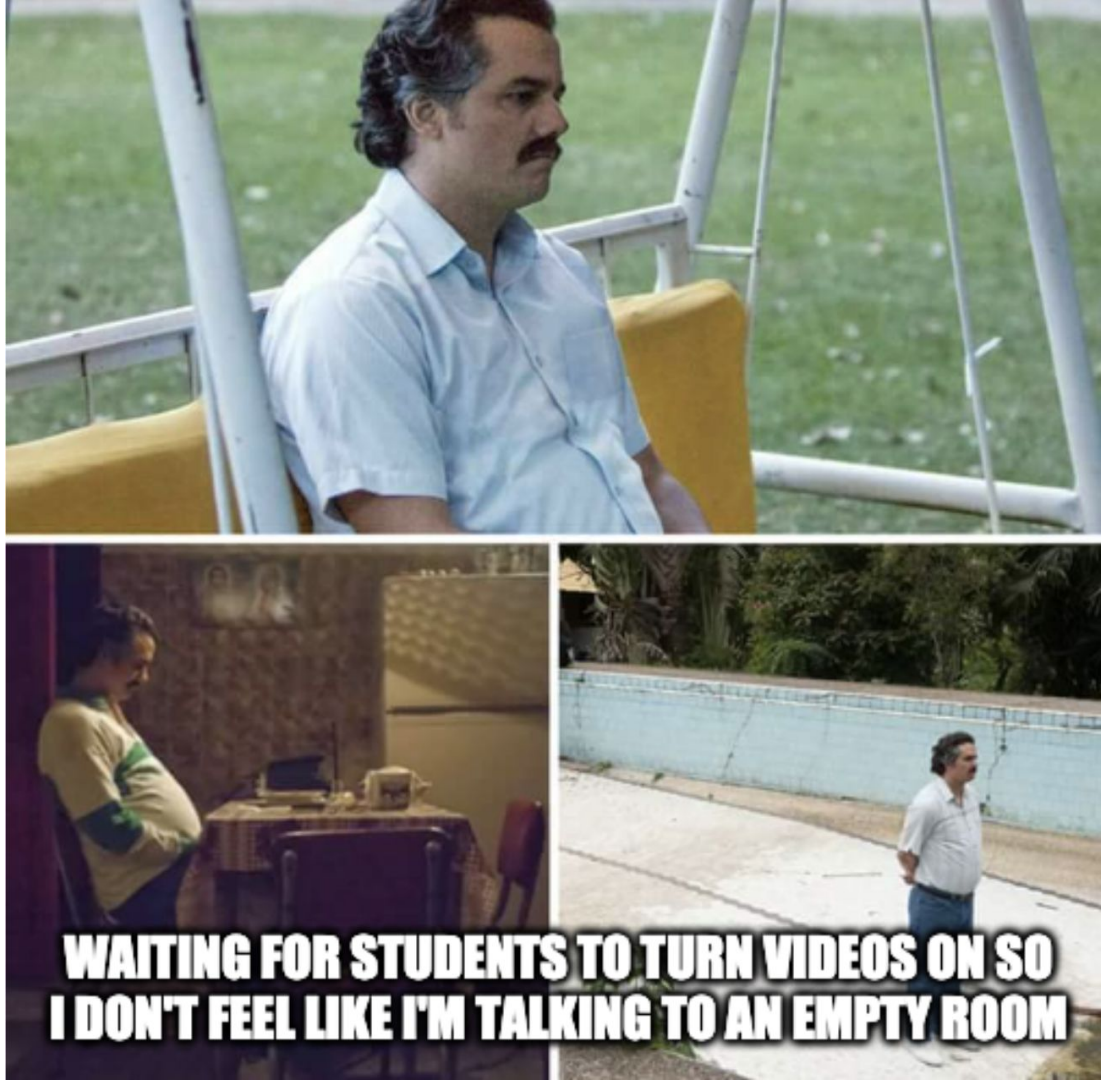
# Machine Learning Systems Design

Lecture 11: Continual Learning
Data Distribution Shifts on Streams

# Zoom etiquettes

We appreciate it
if you keep videos on!

- More visual feedback for us to adjust materials
- Better learning environment
- Better sense of who you're with in class!



# Agenda

1. Continual Learning
2. Test in Production
3. Data Distribution Shifts on Streaming Data

Lecture note is on course website / syllabus

# Continual Learning

Kinbert Chou

# Model’s performance degrades in production

- Data distribution shifts
  - Sudden
  - Cyclic
  - Gradual

# From Monitoring to Continual Learning

- Monitoring: detect changing data distributions
- Continual learning: continually adapt models to changing data distributions

# Continual learning

- Set up infrastructure such that models can continuously learn from new data in production
- Stateful training

# Continual learning: use cases

- Rare events
  - Christmas/Black Friday/Prime Day shopping
  - Total Landscaping
- Continuous cold start (in-session adaptation)
  - New users
  - New devices
  - Users not logged in
  - Users rarely logged in

# Continual learning is especially good for

- Natural labels: e.g. user click -> good prediction
- Short feedback loops
- Examples:
  - RecSys
  - Ranking
  - Ads CTR prediction
  - eDiscovery

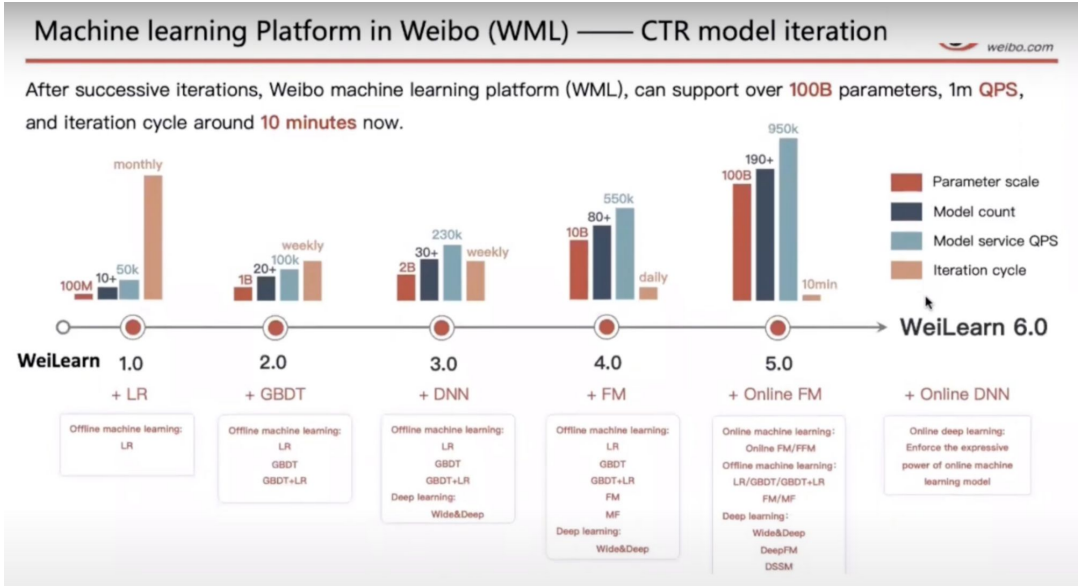
# How often to retrain your model is just a knob to turn

# How frequently should you update your models?

- Very few companies actually update models with each incoming sample
  - Catastrophic forgetting
  - Can get unnecessarily expensive*
- Update models with micro-batches

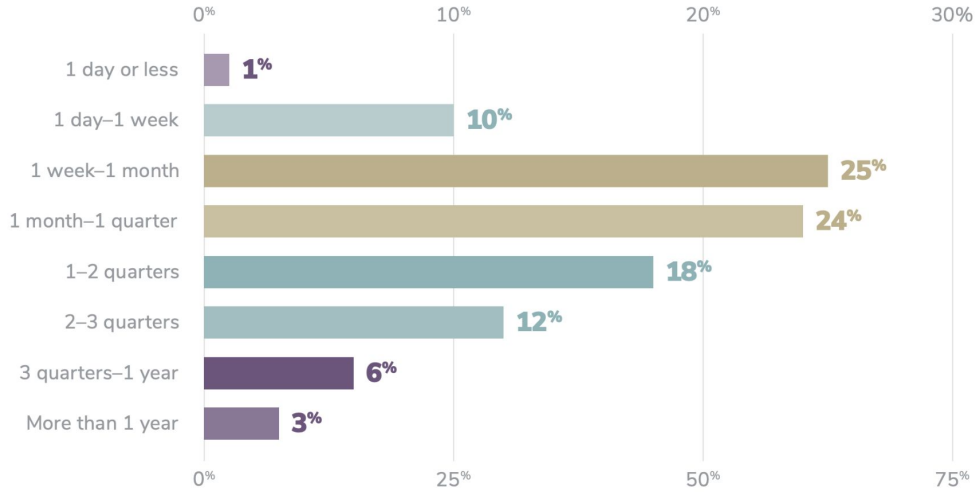
# Iteration cycle: minutes

- Alibaba: Singles Day sale
- Weibo
- Tiktok
- SheIN



# Iteration cycle: US

Only 11% of organizations can put a model into production within a week, and 64% take a month or longer

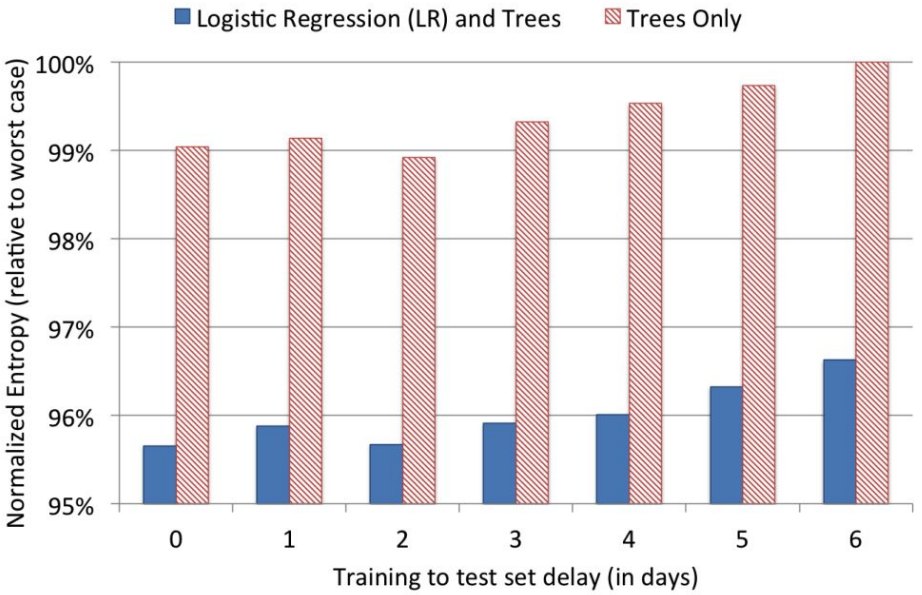



Algorithmia's State of Enterprise Machine Learning (2021)

# Quantify the value of data freshness

1. How much model's performance changes if switch from retraining monthly to weekly to daily to hourly?
   a. FB: CTR loss can be reduced ~1%

      going from training weekly to daily




Practical Lessons from Predicting Clicks on Ads at Facebook (2014)

# Quantify the value of data freshness

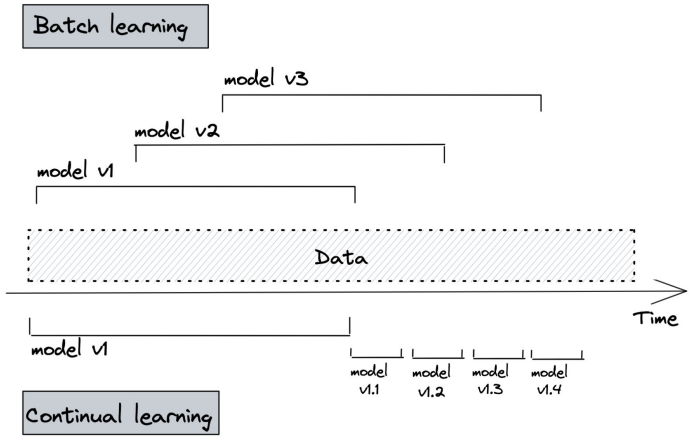
1. How much model's performance changes if switch from retraining monthly to weekly to daily to hourly?
2. How would retention change if you can do in-session adaptation?

# Quantify the value of data freshness

1. How much model's performance changes if switch from retraining monthly to weekly to daily to hourly?
2. How would retention change if you can do in-session adaptation?
3. Model iteration vs. data iteration

# Quantify cloud bill savings

- Train model incrementally each day on fresh data
- Faster convergence → less compute needed




GRUBHUB™ Going from monthly training to daily training gives

**45x cost savings** and **+20% metrics increase**



Online Learning for Recommendations at Grubhub (Alex Egg, 2021)

# Quantify the value of fast iteration

1. How many more experiments can you run if model changes can be deployed automatically ASAP?

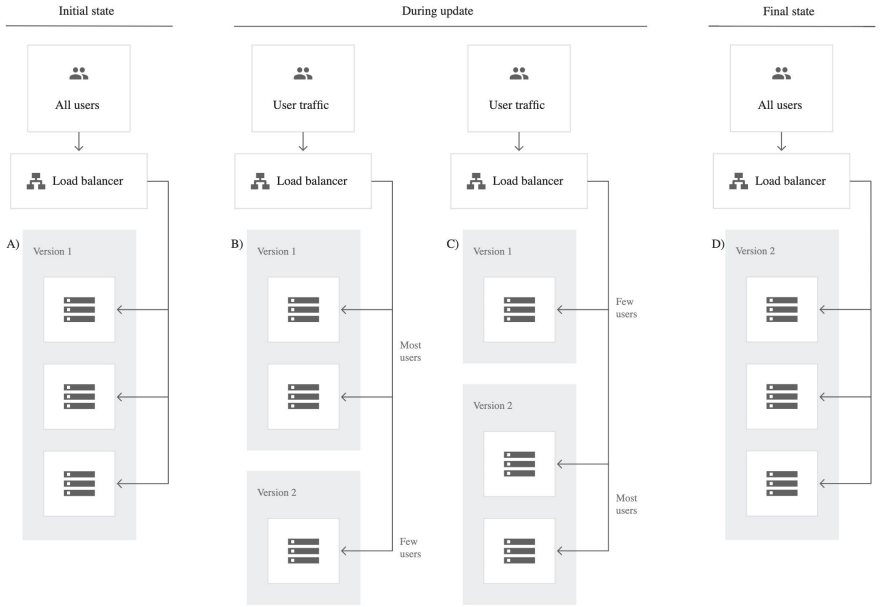
# Learning schedule != evaluating schedule

- Evaluated after a certain period of time
  - Offline evaluation (sanity check)
  - Online evaluation: canary analysis, A/B testing, bandits

# Test in Production

# Canary testing

- New model alongside existing system
- Some traffic is routed to new model
- Slowly increase the traffic to new model
  - E.g. roll out to Vietnam first, then Asia, then rest of the world

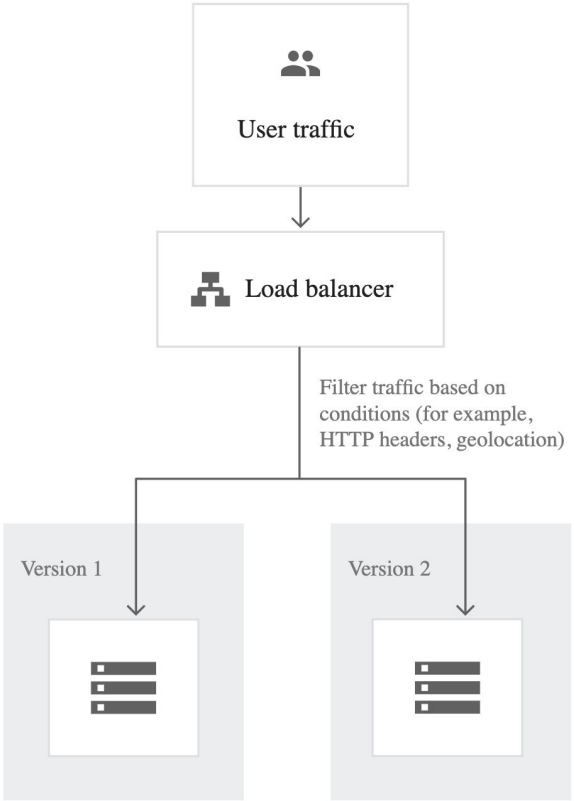





Application deployment and testing strategies (Google)

# A/B testing

- New model alongside existing system
- A percentage of traffic is routed to new model based on routing rules
- Control target audience & monitor any statistically significant differences in user behavior
- Can have more than 2 versions

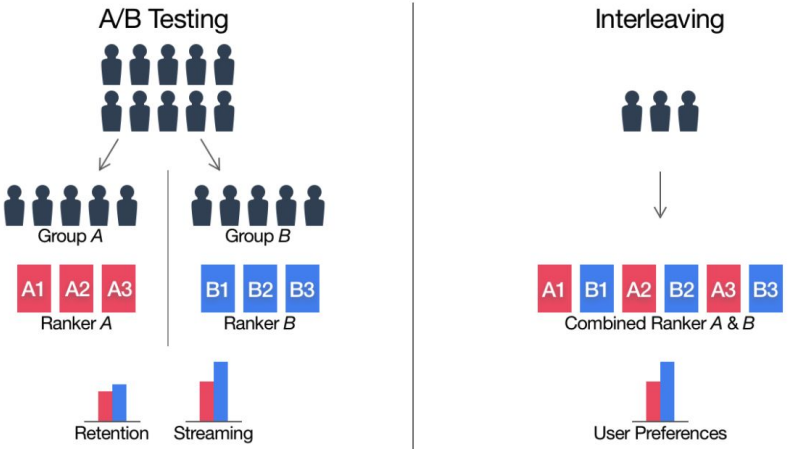




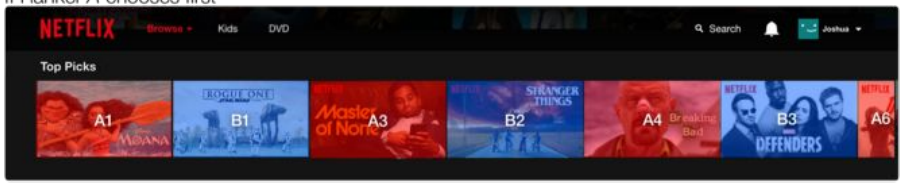

Application deployment and testing strategies (Google)

# Interleaved experiments

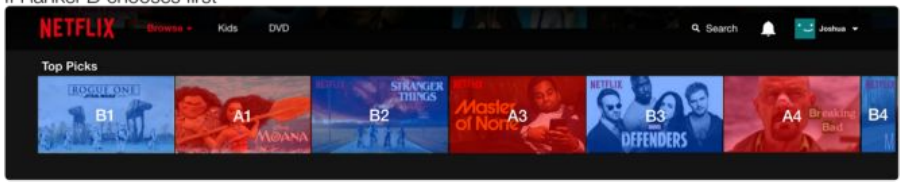
- Especially useful for ranking/recsys
- Take recommendations from both model A & B
- Mix them together and show them to users
- See which recommendations are clicked on




If Ranker *A* chooses first

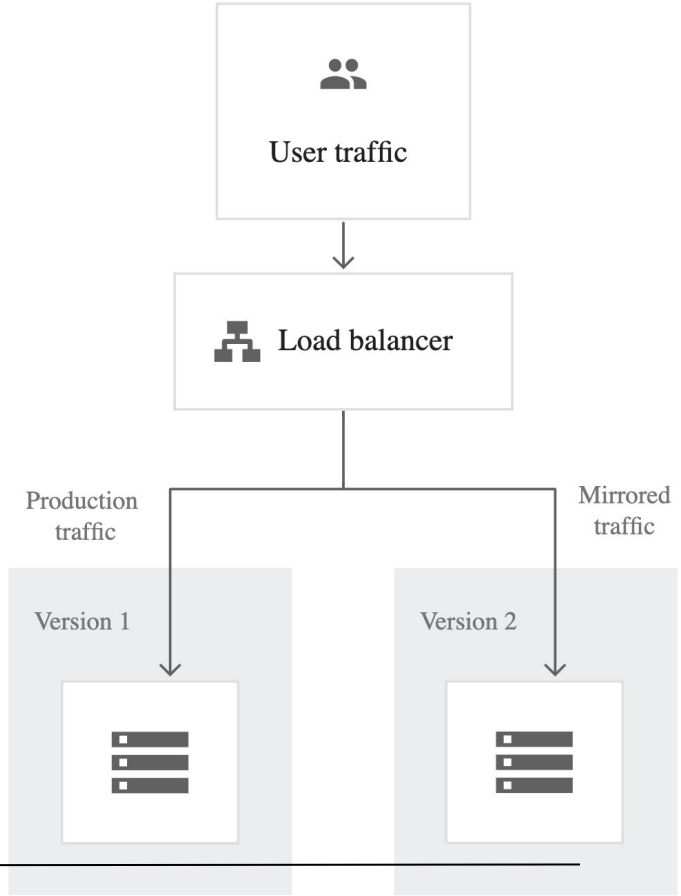



If Ranker *B* chooses first



# Shadow testing

- New model in parallel with existing system
- New model's predictions are logged, but not show to users
- Switch to new model when results are satisfactory




Application deployment and testing strategies (Google)

# Test internally first

- Use features even as they’re in development
- Share internally before externally

⚠️ You and your coworkers are not typical users ⚠️

# Distribution Shifts on Streaming Data

Shreya Shankar

# Machine Learning Systems Design

Next class: Experimental Tracking & Versioning