

From Black Box to a Known Quantity: How to Build Predictable, Reliable ML-based Services

{salim,villavieja}@google.com

SREcon 2019

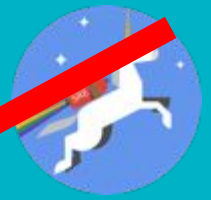
© 2018 Google LLC. All rights reserved.

We are not machine learning hackers/ninjas
We are not machine learning scientists



We are **experienced SREs** and we have collected production insights through a **large number of interviews (~40)** from teams using ML in production at Google over the last 15 years.

© 2019 Google LLC. All rights reserved.



From Black Box to a Known Quantity: How to Build Predictable, Reliable ML-based Services

{salim,villavieja}@google.com

SREcon 2019

© 2018 Google LLC. All rights reserved.



ML Best practices & monitoring

{salim,villavieja}@google.com

SREcon 2019

© 2018 Google LLC. All rights reserved.

Our goals



ML best practices

&

monitoring



© 2018 Google LLC. All rights reserved.

Find the mistake about ML



ML is easy

ML is new

ML is a black box, no need to know more

Train “one and done”

You rarely rollback

ML monitoring is like any other monitoring

More data is better

Learn all the patterns

Transparent to user

Compatibility is a no-op

© 2019 Google LLC. All rights reserved.

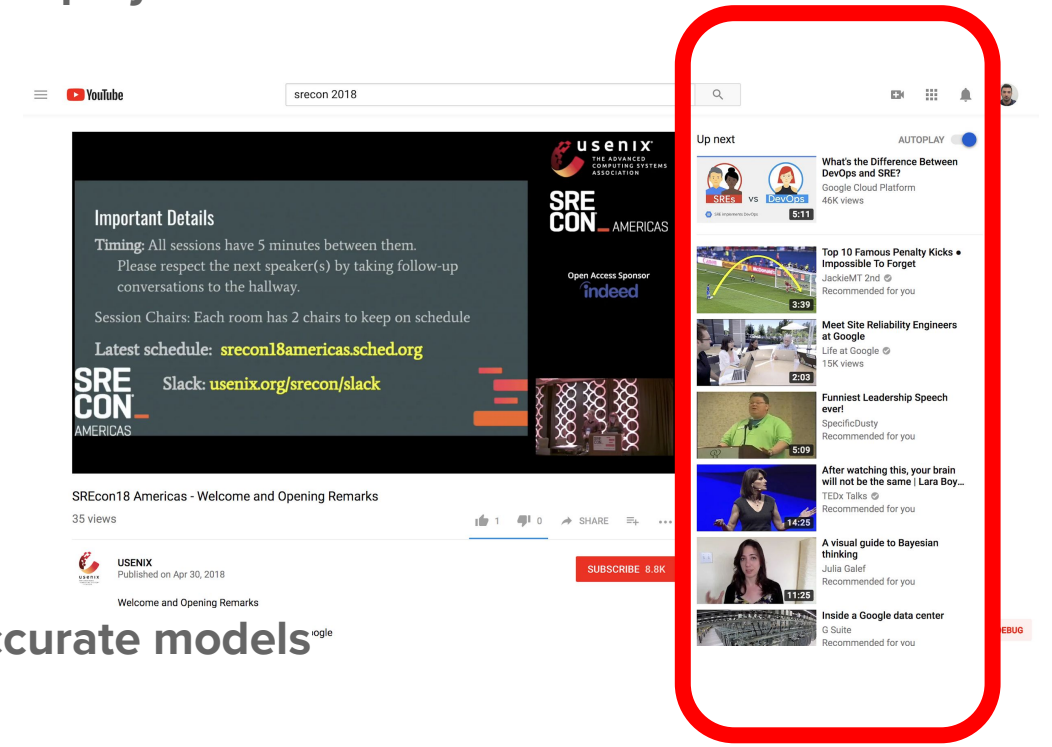


One Very Important ML model At Google

© 2019 Google LLC. All rights reserved.

Youtube ML for video recommendations

- Continuously training & Fast deployment
- Keep high accuracy
- World Wide input data
- Revenue facing
- More video time +
  More ad clicks
- Special events one day
- User can easily detect not accurate models



© 2019 Google LLC. All rights reserved.

But it's not that easy in production



GUARANTEE FRESHNESS

MULTIPLE DEVICES

MONITOR VIEW TIME/...

FILTERING SPAM/BAD VIDEOS

DEPLOYING EVERY *N* HOURS / DAYS / WEEKS

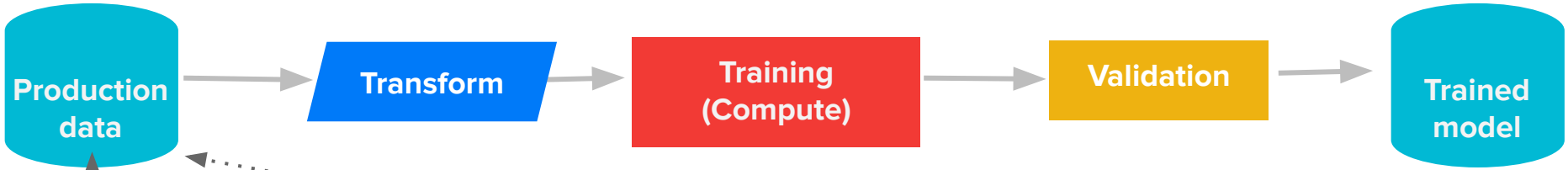
CONTINUOUSLY TRAINING MODELS

© 2019 Google LLC. All rights reserved.

Theoretical Machine Learning Pipeline



Training Offline: (effort spent 10%)



Deployment

Serving Online: (effort spent 90%)



User-facing request

Prediction Classification

© 2019 Google LLC. All rights reserved.

Theoretical Machine Learning Pipeline



DEPRECATED!

NOT

RELIABLE

Training Offline: (effort spent 10%)

Production data

Transform

Training (Compute)

Validation

Trained model

Deployment

Serving Online: (effort spent 90%)

User-facing request

Transform

Serving

Prediction / Classification

Serving model

© 2019 Google LLC. All rights reserved.

BEST PRACTICES: Why are they important



Part 1 : TRAINING & DATA QUALITY	RELIABLE
Part 2 : HARDWARE RESOURCES (GPU/TPU)	FAST
Part 3 : QUALIFICATION	PROD READY
Part 4 : BACKWARDS COMPATIBILITY/CONF.MANAGEMENT	EASY

© 2019 Google LLC. All rights reserved.



(re) Training

(not prototyping)

© 2019 Google LLC. All rights reserved.

Training



Integral part of the release process

Not coding, debugging, testing
Input data coming to the training pipeline can't be stopped

Production changes fast:

Model loss increases with time at a constant rate.

© 2019 Google LLC. All rights reserved.

Training



Production changes fast:

Model loss increases with time at a constant rate.

© 2019 Google LLC. All rights reserved.

Training



Production changes fast:

Model loss increases with time at a constant rate.

© 2019 Google LLC. All rights reserved.

Training



Production changes fast:

Model loss increases with time at a constant rate.

© 2019 Google LLC. All rights reserved.

Training



Production changes fast:

Model loss increases with time at a constant rate.

© 2019 Google LLC. All rights reserved.

Training



Production changes fast:

Model loss increases with time at a constant rate.

© 2019 Google LLC. All rights reserved.

Model Age for a popular Google Service



- ➔ Over 3 weeks, model never older than ***X***
- ➔ Updates continuously

Model Age for a popular Google Service



➔ Over 3 weeks, model never older than ***X***

➔ Updates continuously

Non-stop training

Models might need to evolve fast



© 2019 Google LLC. All rights reserved.

Training: Filtering is key



- **Good:** Train your model with ***all*** data, from oldest to newest
- **Bad:** We can't ***ALWAYS*** train on all production data. (Youtube 1.2 TB ML model)
- Production data has ***tons of duplicate information*** and needs to be filtered.
- **Filtering**: collapse duplicate values, to construct the model efficiently.
- **Data Imputation:** replacing missing data with substituted values

```
1, carlos, male,41, Spanish, 6.2, SRE, NULL, 80%
2, salim, male, 44, American, 5.8, SRE, +4, 90%
3, maria, female, 0, Norway, 6.0, SWE, +25, 60%
4, fep, agender, Spanish, 6.0, SWE, +5, 75%
5, maria, female, 0, Norway, 6.0, SWE, +25, 60%
```

© 2019 Google LLC. All rights reserved.

Training: Data size

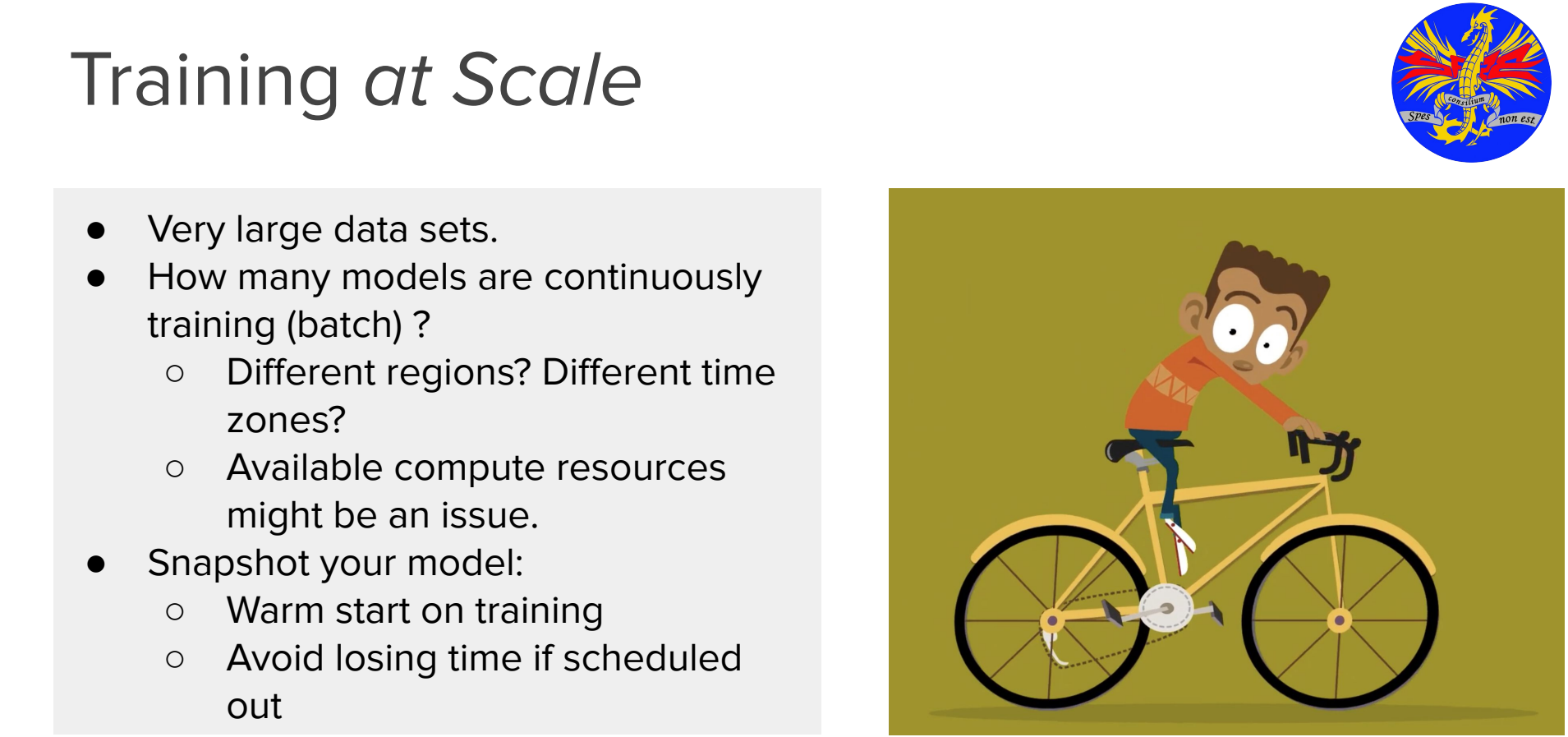
- Validation data is not the same as trained data
  - Trust that your high-accuracy model is correct with data not used during training.
    - 80-20/70-30 might vary depending on the model
    - Randomly selected set from the trained data
- Do not confuse with *qualification* (to be seen later)



Image source: https://vimeo.com/122534562
License: https://creativecommons.org/licenses/by/3.0/

© 2019 Google LLC. All rights reserved.

Training *at Scale*



- Very large data sets.
- How many models are continuously training (batch) ?
  - Different regions? Different time zones?
  - Available compute resources might be an issue.
- Snapshot your model:
  - Warm start on training
  - Avoid losing time if scheduled out

Image source: https://vimeo.com/122534562
License: https://creativecommons.org/licenses/by/3.0/

© 2019 Google LLC. All rights reserved.

Summary: Data Quality on Training



CORRECT

Data imputation and data validation so that your models never receive unexpected inputs.

SNAPSHOTS	Train over previous models (resuming and rollback)
BIAS	Monitor amount of data from different sources. Features skews (train features diff from inference features)
ANOMALIES	Can't train with all SuperBowl day/New Years

COMPLETE

Missing inputs previously used

DATA RATIOS	Continent X pipeline stopped and the youtube recommendation models stops taking into account those videos.
AUTOMATION	Be ready to add fields on old data. Be ready to fix your data (spam data in trained models)

© 2019 Google LLC. All rights reserved.

Before Machine Learning Pipeline



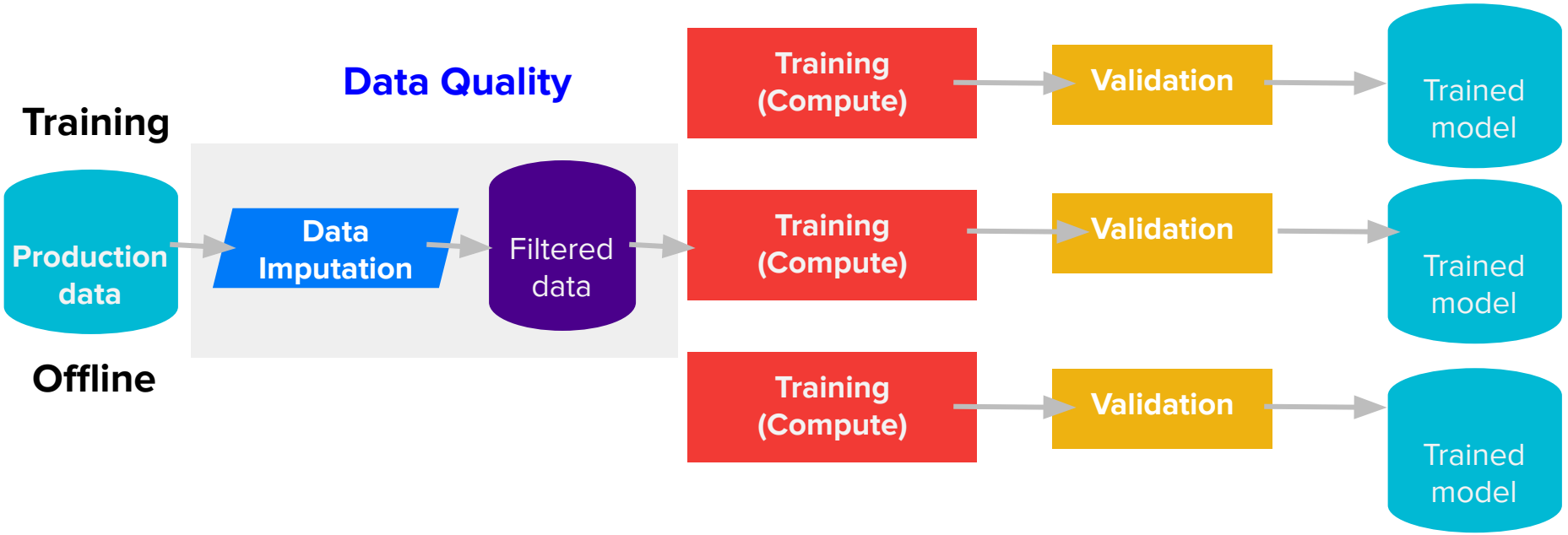
Training (time spent 10%)



Offline

© 2019 Google LLC. All rights reserved.

After Machine Learning Pipeline



© 2019 Google LLC. All rights reserved.



Resources

© 2019 Google LLC. All rights reserved.

Why hardware resources are important



- Two different & disjoint environments
  - Training
  - Serving/Inference
- Cost of Training resources grows **at a higher rate** than Production resources



© 2019 Google LLC. All rights reserved.



Qualification

© 2019 Google LLC. All rights reserved.

Model Qualification

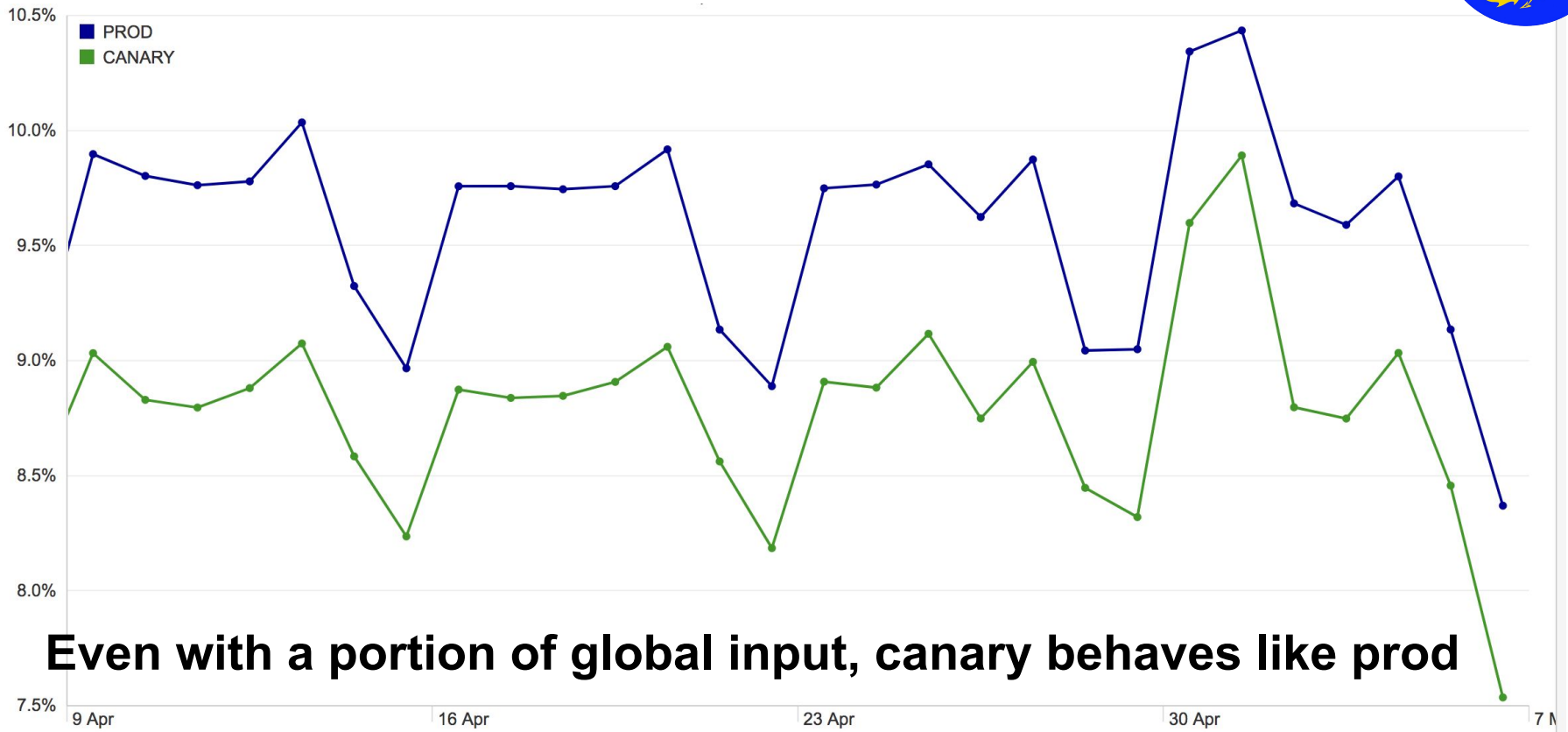


- Models are qualified with a separate input data
  - How is this data chosen? (previous or same prod day)
- **Models are tested with the same production binary.**
- Or we have an A/B testing scenario
  - Same production code/release
  - Dynamically decide % predictions to each model



© 2019 Google LLC. All rights reserved.

Canary is a must



Even with a portion of global input, canary behaves like prod

© 2019 Google LLC. All rights reserved.

Model Qualification



The model is signed post qualification.

- ❏ Some providers allow to register models for versioning
- ❏ Signature specifies type of model, input/output data

Only allow *signed* models in production.





© 2019 Google LLC. All rights reserved.



Backwards Compatibility

© 2019 Google LLC. All rights reserved.

Backwards Compatibility



- Input data changes:
  - New fields, new values, null/empty values not contemplated.
- API changes:
  - Tensorflow API changes frequently, Incompatible model
  - "The model was completely valid and healthy as configured, it was simply not configured for the type of traffic it would receive"
- Fallback mechanisms, when rollback not an option:
  - Most teams do not have a non-ML fallback mechanisms
  - What happens if you run out of quota/capacity.
- **Old models need to be deprecated, they might not be reusable**
  - New inputs (labels) deployed
  - Signing the models helps on this. However, we're not able to ask the model compatibility?



© 2019 Google LLC. All rights reserved.

Three important (config) questions



- Do code and models go together? Are they deployed in the same package?
- How do we verify model and code compatibility?
  - *Always* push through canary
- What API version is this model for?

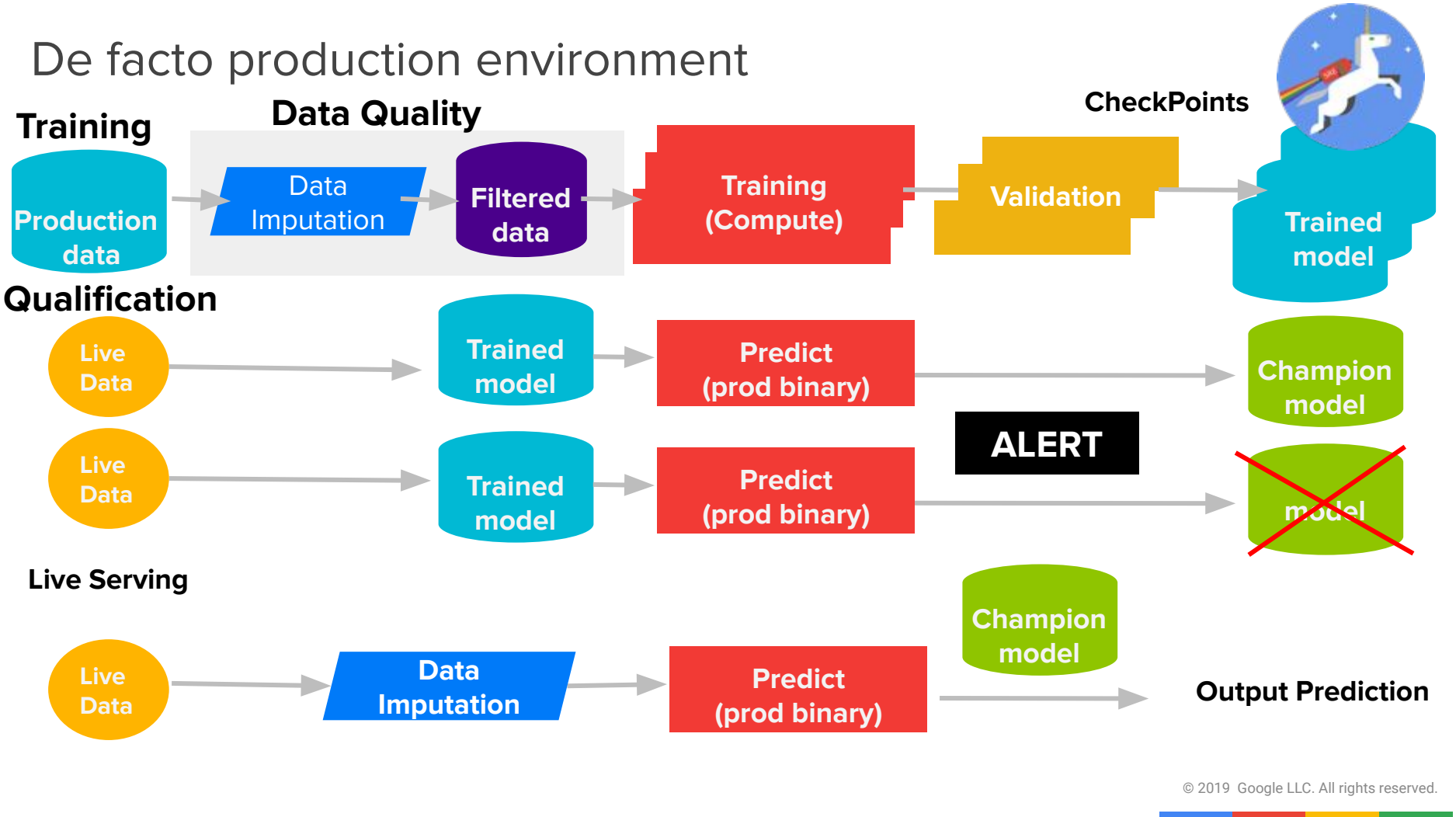
© 2019 Google LLC. All rights reserved.



What's an ML pipeline like?

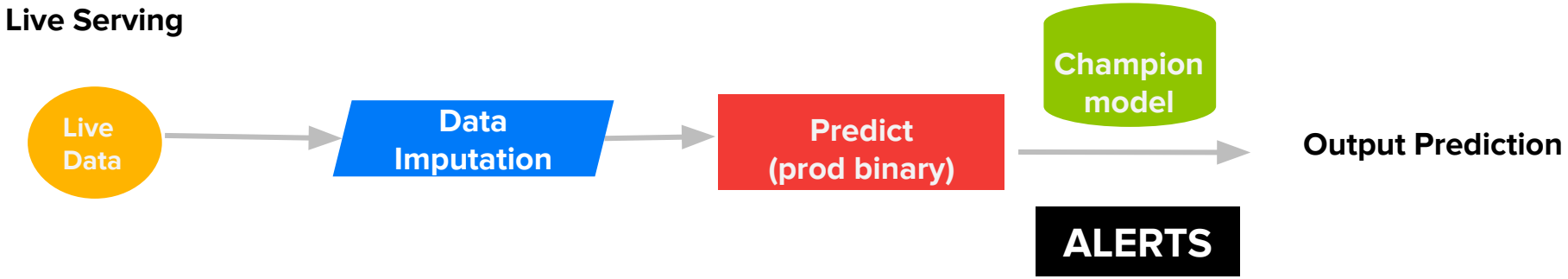
© 2019 Google LLC. All rights reserved.

De facto production environment



© 2019 Google LLC. All rights reserved.

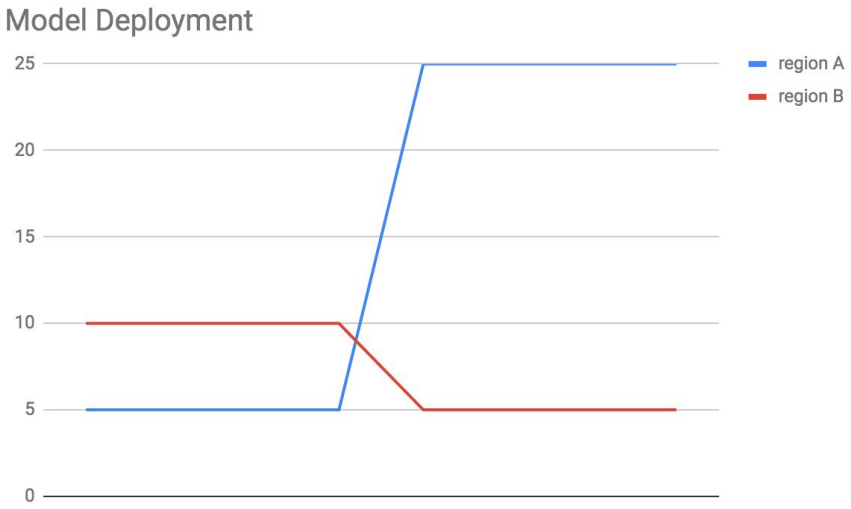
Monitoring: From SLIs to Alerting



- ❑ Monitor training phase as well as live serving
- ❑ Monitor Prediction latency
  - ❑ Consider TF/RPC overhead
- ❑ Monitor Model aging
  - ❑ Accuracy loss through model aging

© 2019 Google LLC. All rights reserved.

What goes wrong when you don't have alerting?



How can you identify this change in behavior?

Alerting must be domain-specific



© 2019 Google LLC. All rights reserved.

Training



Production changes fast:

Model loss increases with time at a constant rate.

© 2019 Google LLC. All rights reserved.

Conclusions



Train continuously

Add filtering

Data imputation

Stamp new models...

... Deprecate old models

Use domain-specific alerting

© 2018 Google LLC. All rights reserved.

The Future of ML in Production



- Open source available training data sets
  - Already anonymized + No need to delete user data
- Models as a Service
  - Credit card/Image recognition/Text to Speech as unique APIs

© 2019 Google LLC. All rights reserved.

With thanks to



very many of our colleagues across Alphabet: DeepMind, Google, YouTube

And to our colleagues at
Clarifai
ThoughtWorks

© 2018 Google LLC. All rights reserved.



That’s all.

Questions? Comments?

{salim,villavieja}@google.com

© 2019 Google LLC. All rights reserved.