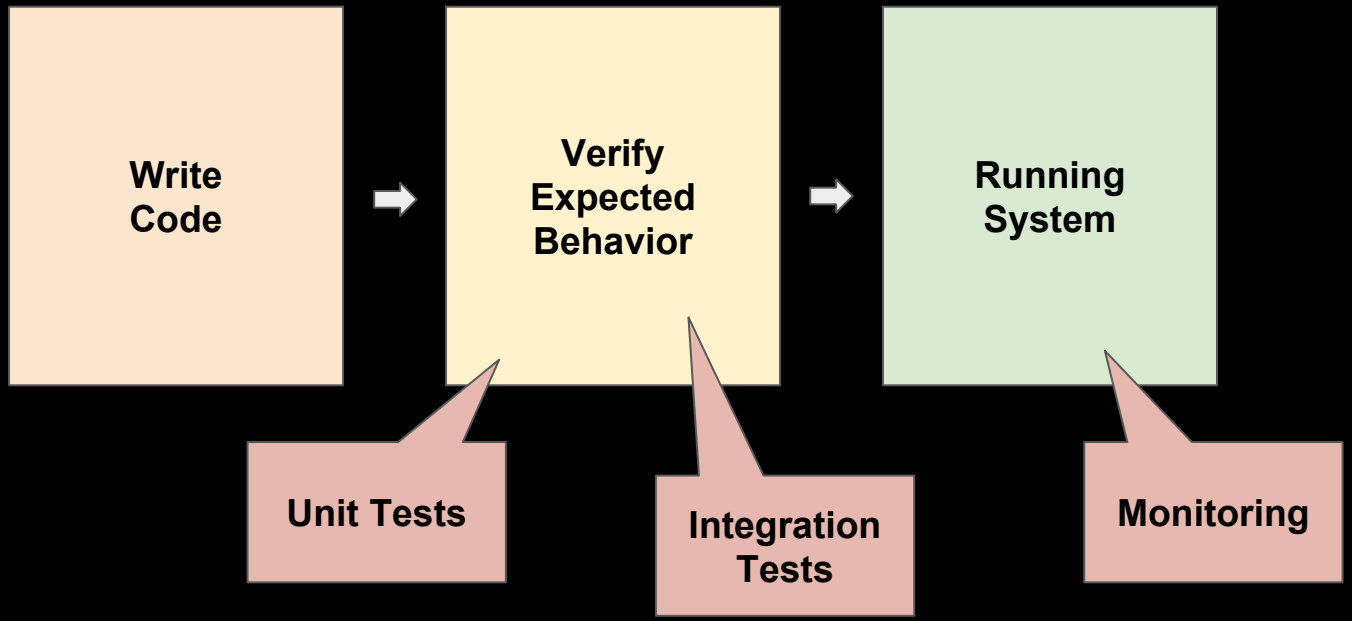
What's your ML Test Score?

A rubric for ML production systems

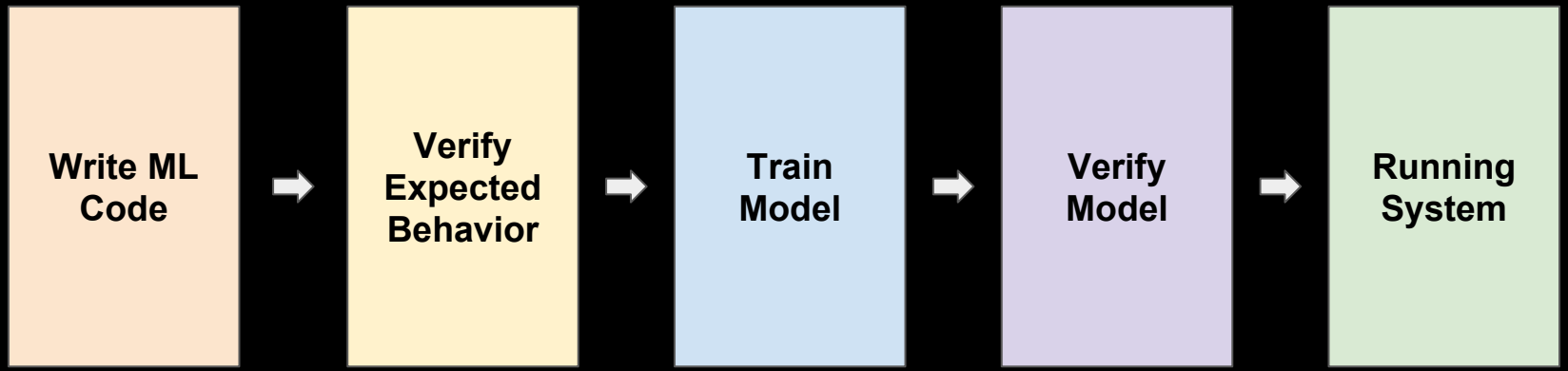
Eric Breck, Shanqing Cai, Eric Nielsen, Michael Salib and D. Sculley
{ebreck, cais, nielsene, msalib, dsculley}@google.com

Testing Code

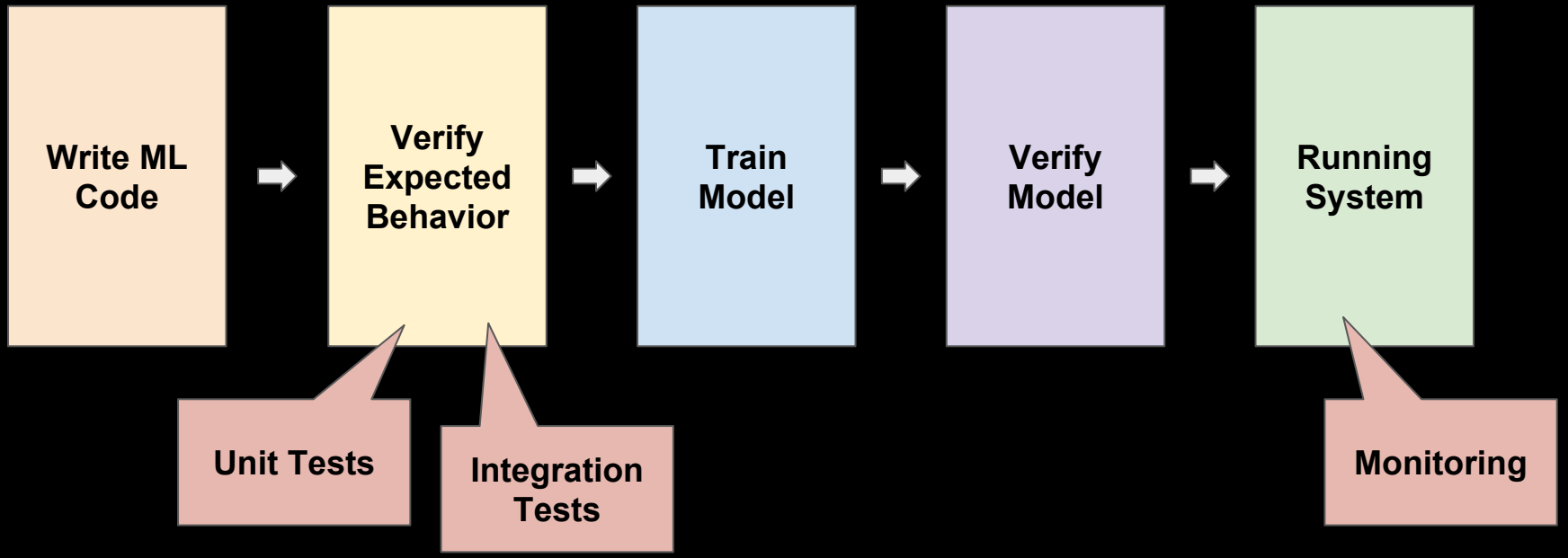
Testing Code



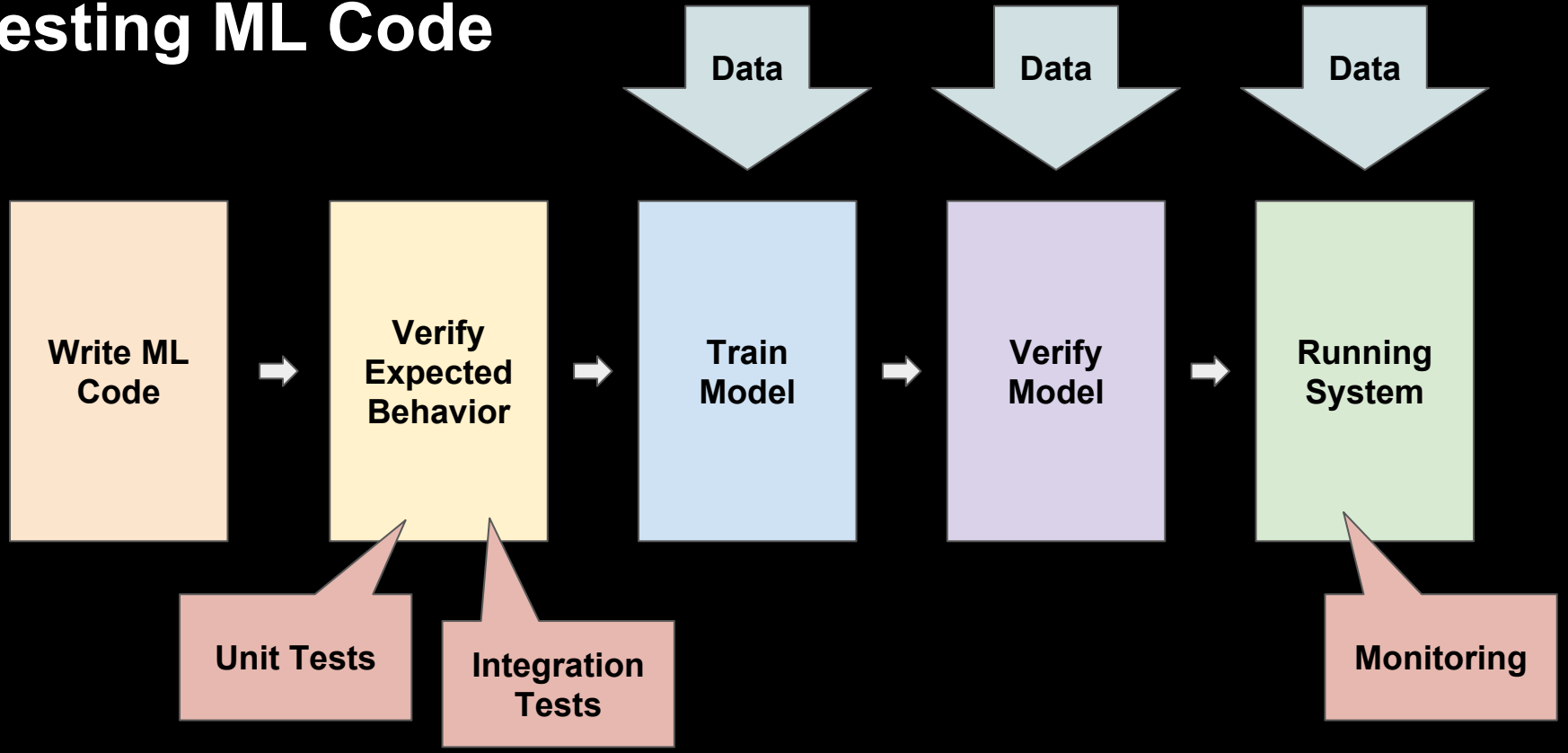
Testing ML Code



Testing ML Code

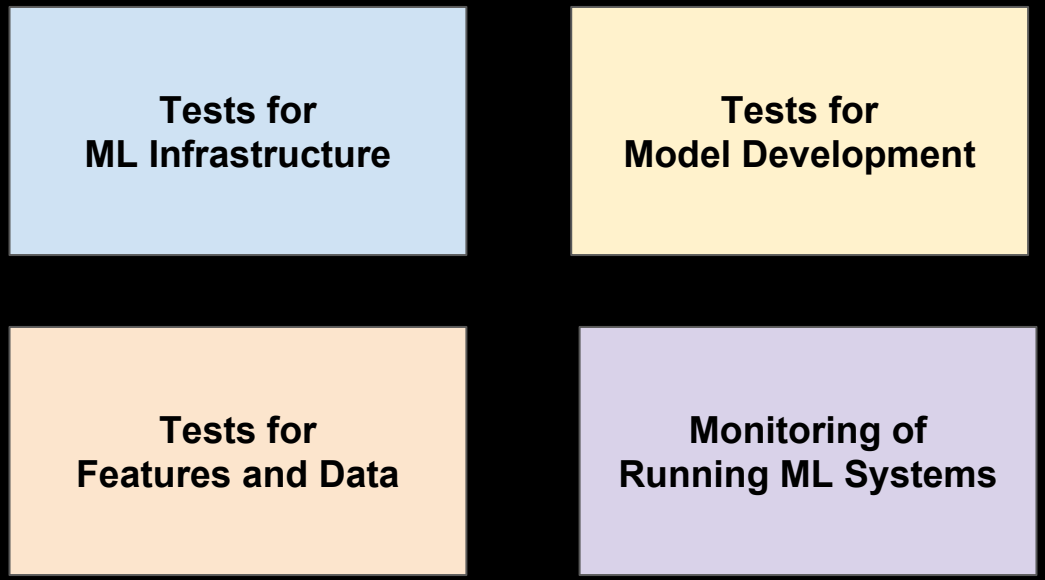


Testing ML Code



Okay, so what should we test?

An ML Test Rubric



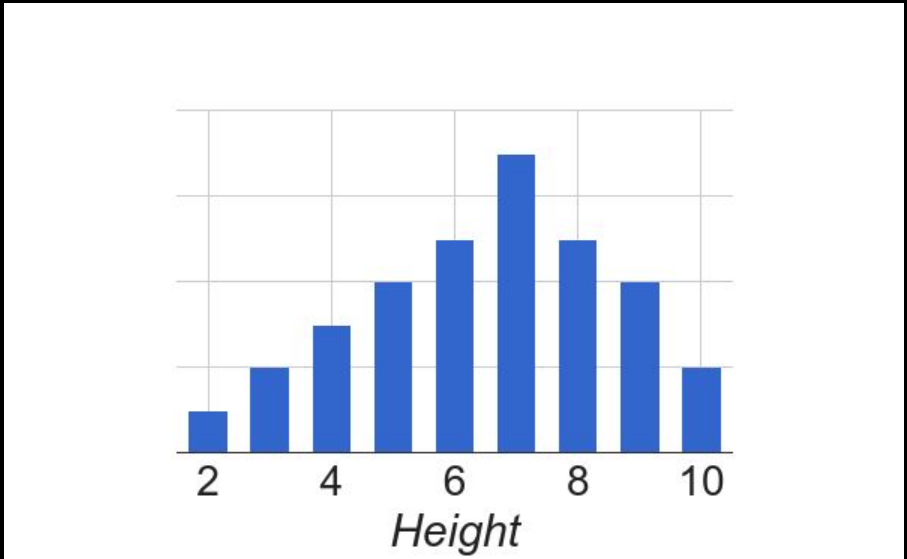
A Few Examples

Test that the distribution of each feature matches your expectation

Tests for
Features and Data

Test that the distribution of each feature matches your expectation

Tests for Features and Data

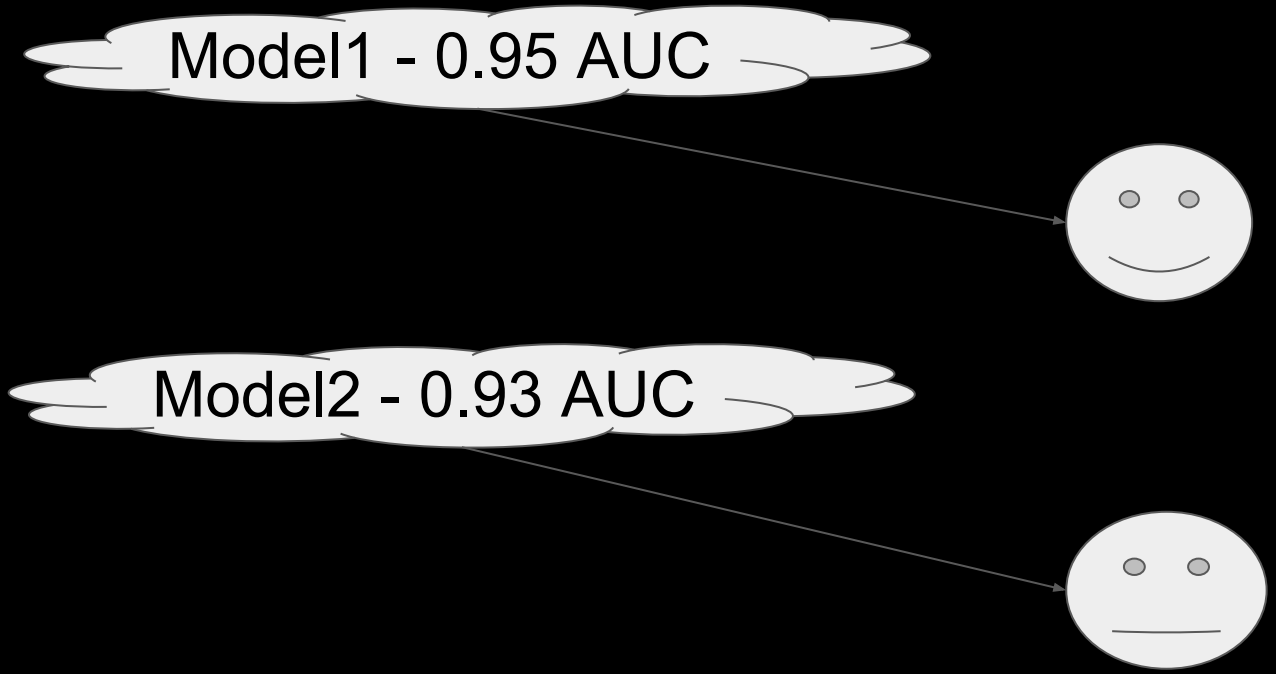


Test the relationship between offline and online metrics

Tests for
Model Development

Test the relationship between offline and online metrics

Tests for Model Development



Test models via a canary process before serving in production

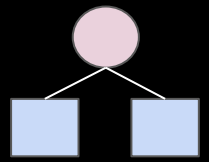
Tests for ML Infrastructure

Test models via a canary process before serving in production

Tests for
ML Infrastructure

Monday

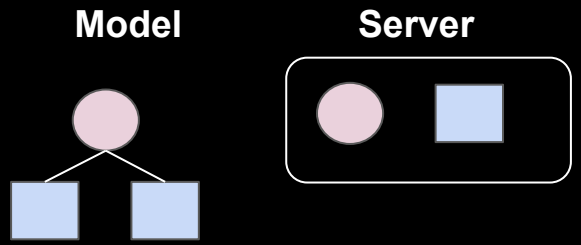
Model



Test models via a canary process before serving in production

Tests for
ML Infrastructure

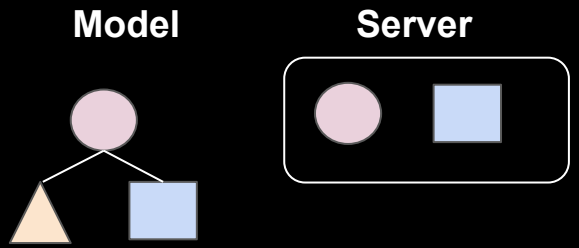
Monday



Test models via a canary process before serving in production

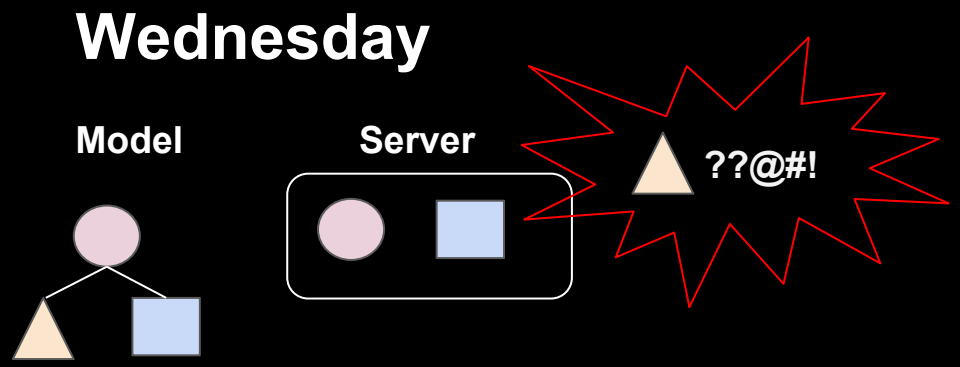
Tests for
ML Infrastructure

Wednesday



Test models via a canary process before serving in production

Tests for
ML Infrastructure

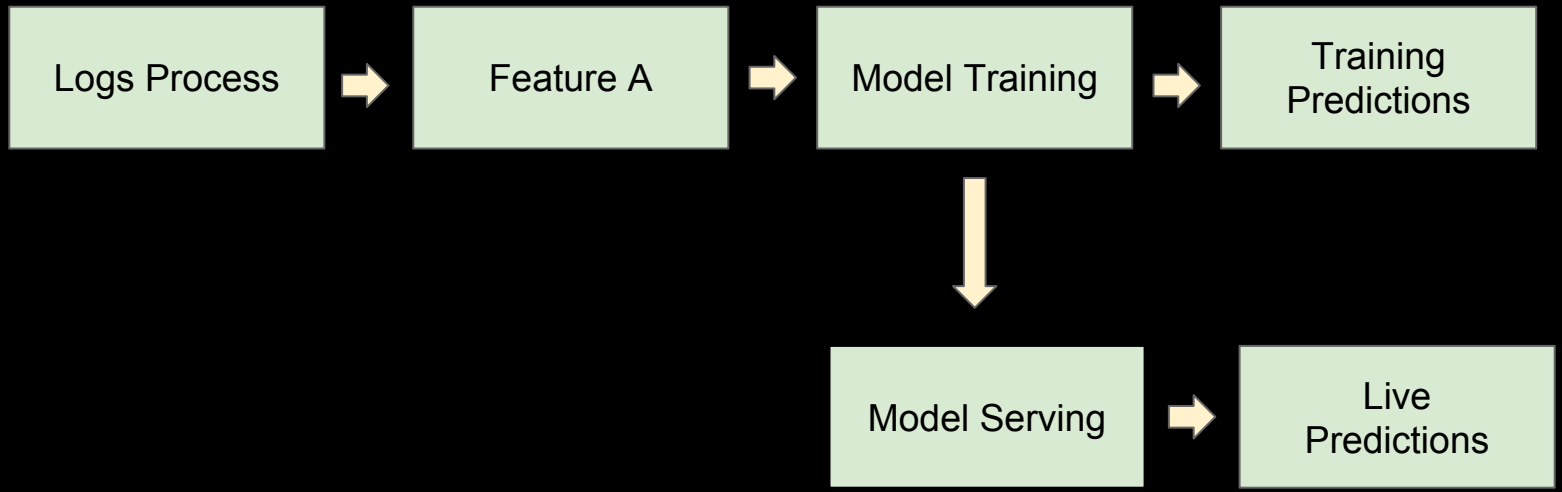


Test that training and serving features compute the same values

Monitoring of Running ML Systems

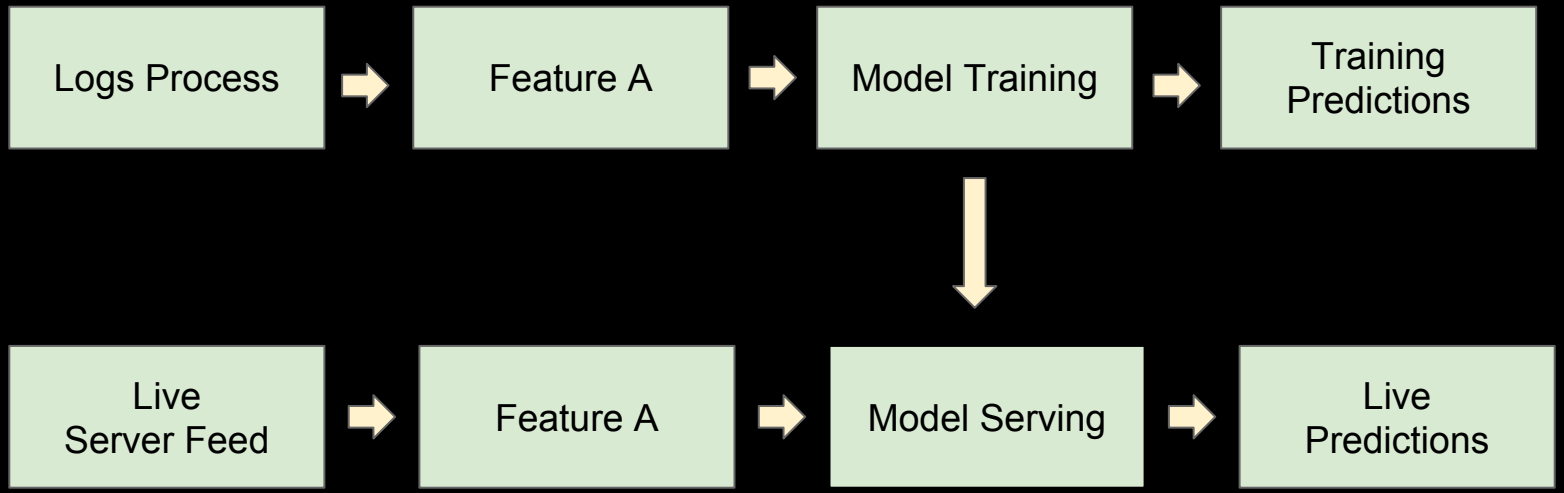
Test that training and serving features compute the same values

Monitoring of Running ML Systems



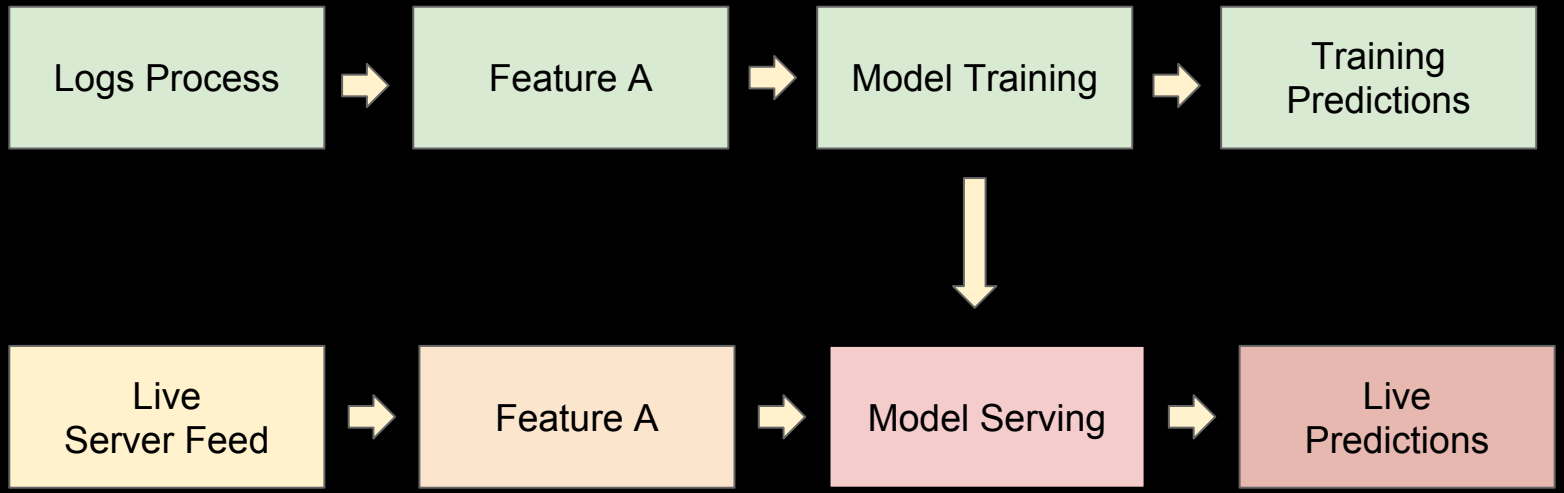
Test that training and serving features compute the same values

Monitoring of Running ML Systems



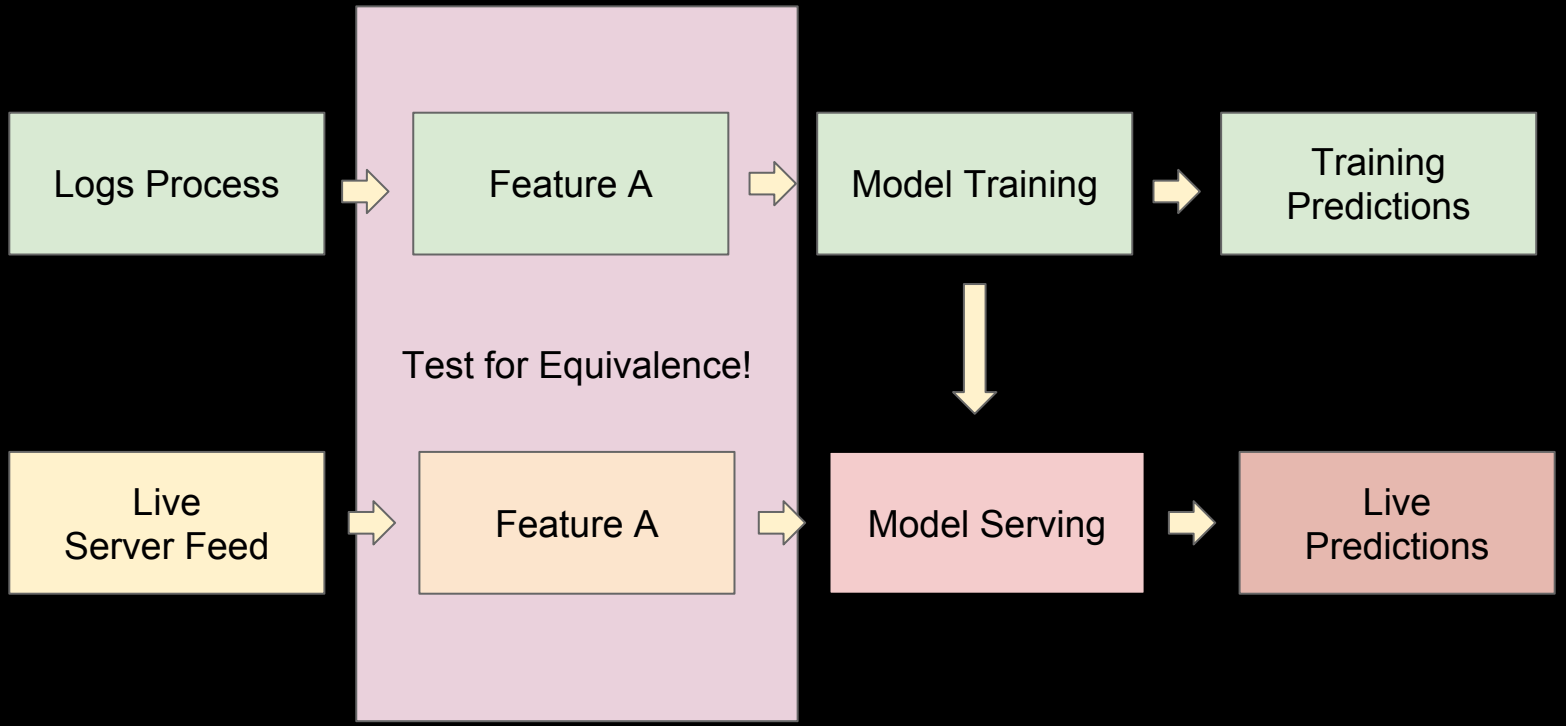
Test that training and serving features compute the same values

Monitoring of Running ML Systems

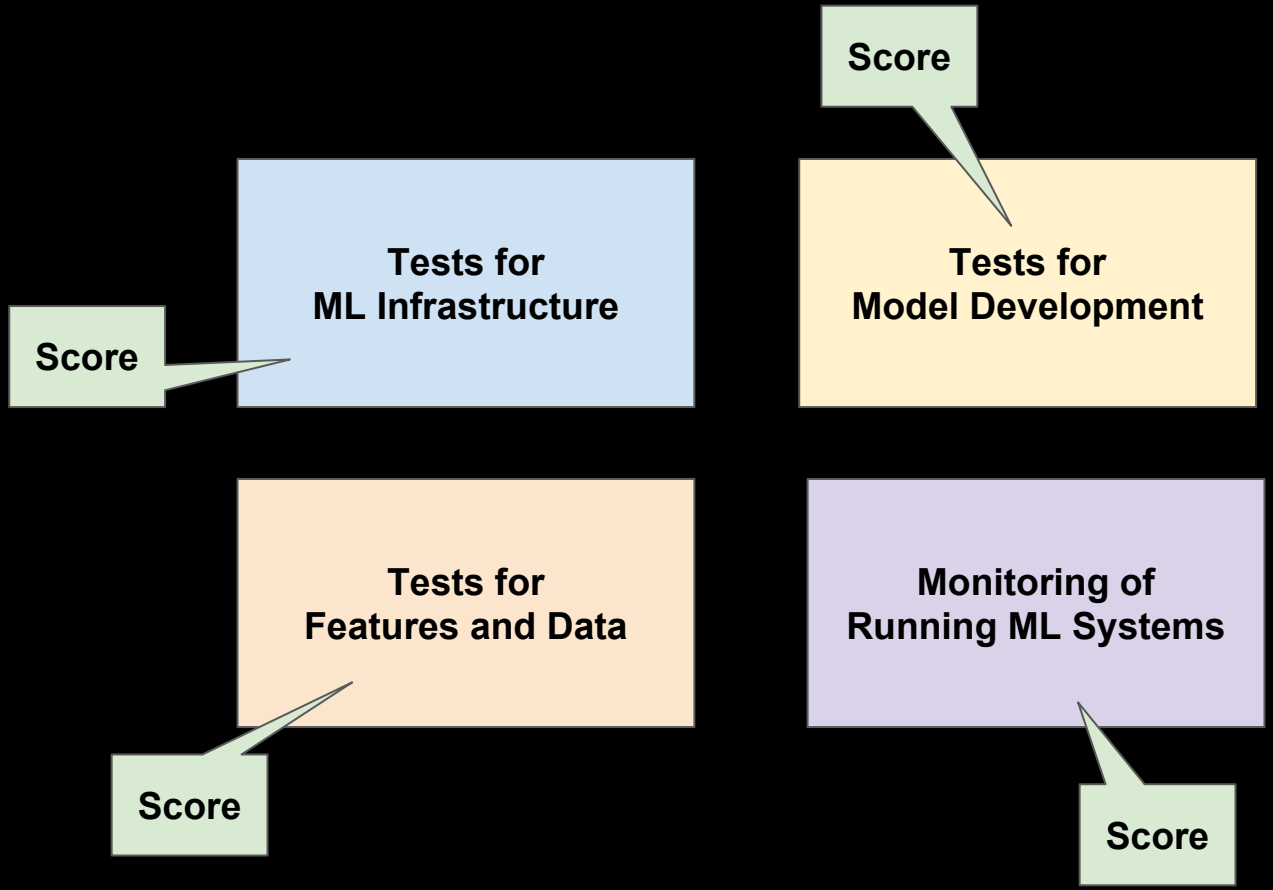


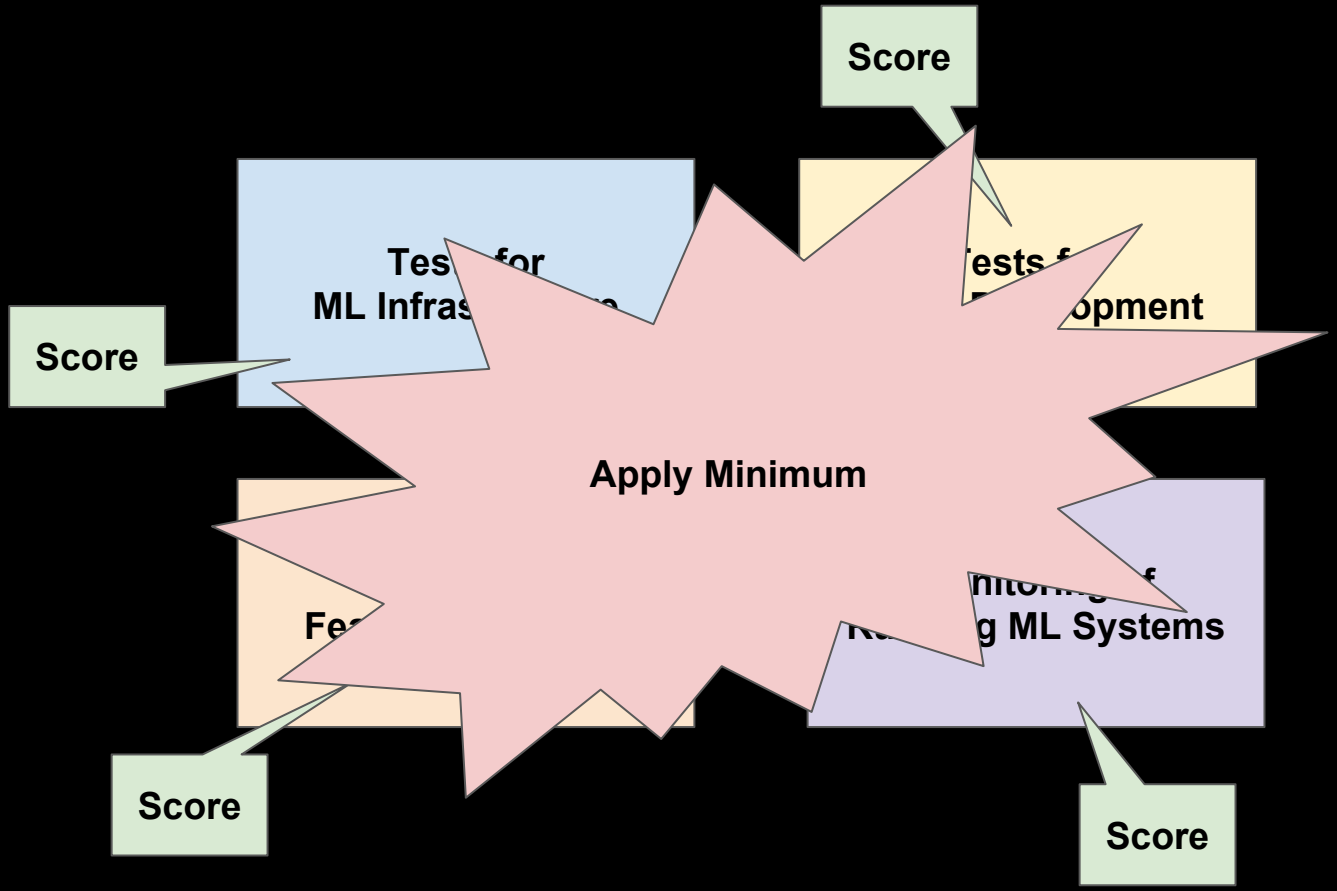
Test that training and serving features compute the same values

Monitoring of Running ML Systems



What’s my score?


Score
Tests for
ML Infrastructure
Score
Tests for
Model Development
Tests for
Features and Data
Score
Monitoring of
Running ML Systems
Score


Score
Score
Apply Minimum
Score
Score
ML Systems

Score	Interpretation
0	Research project
1-2	Not totally untested
3-4	A first pass at basic productionization
5-6	Reasonably tested
7-10	Appropriate for mission-critical systems.
12+	Exceptional levels of automated testing and monitoring.

Questions?

Thank You!

dsculley@google.com