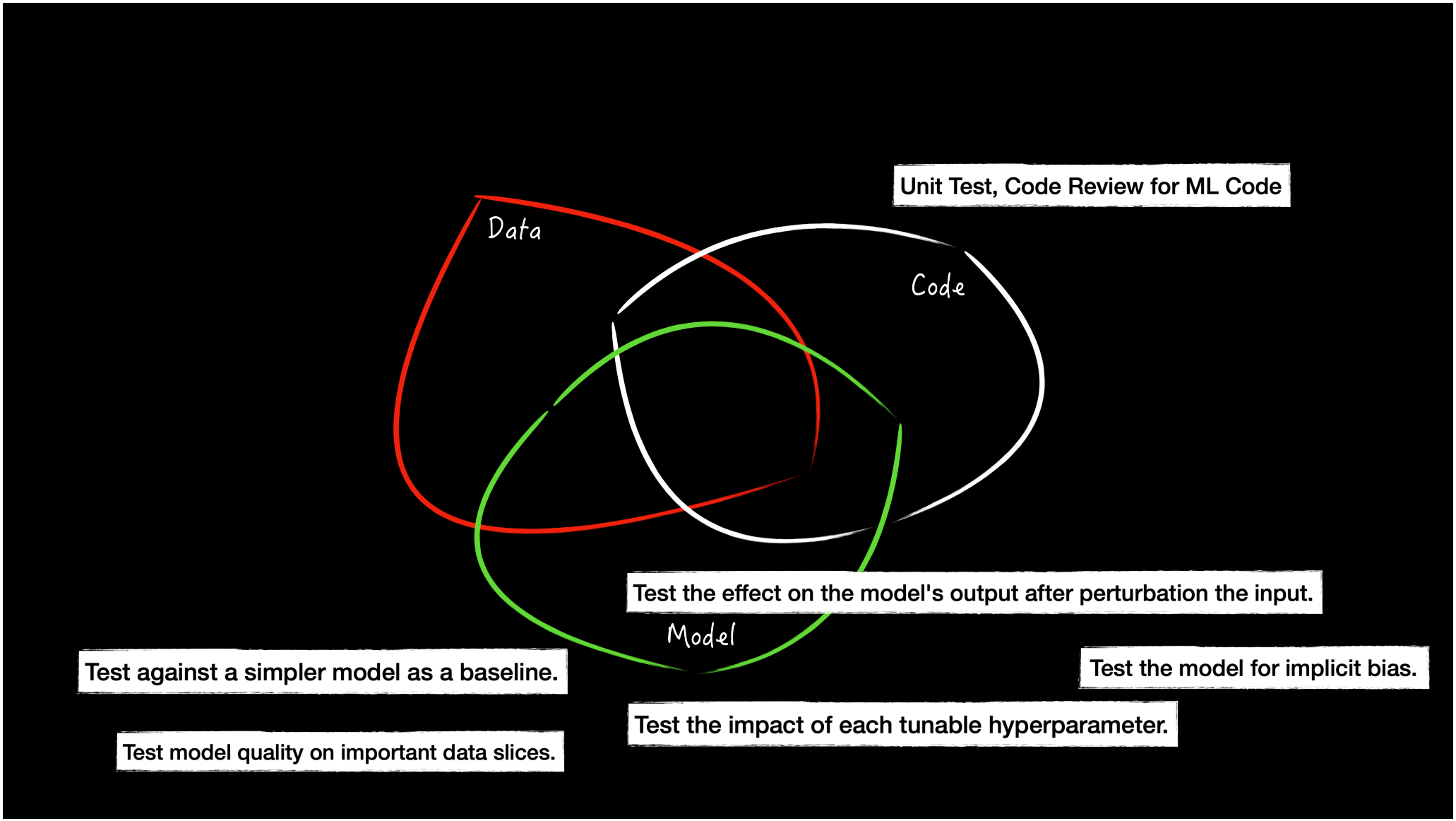
Quality Assessment

System Quality

Jin Guo
SOCS McGill University


Unit Test, Code Review for ML Code
Data
Code
Model
Test the effect on the model's output after perturbation the input.
Test against a simpler model as a baseline.
Test the model for implicit bias.
Test the impact of each tunable hyperparameter.
Test model quality on important data slices.

Data Slicing

A subset that is particularly relevant for the project/feature objectives.

Input	Actual Output	Model Output
1	No	0.8 -> Yes
2	No	0.7 -> No
3	Yes	0.75 -> No
.....	.....	.....
14	Yes	0.4

Data Slicing

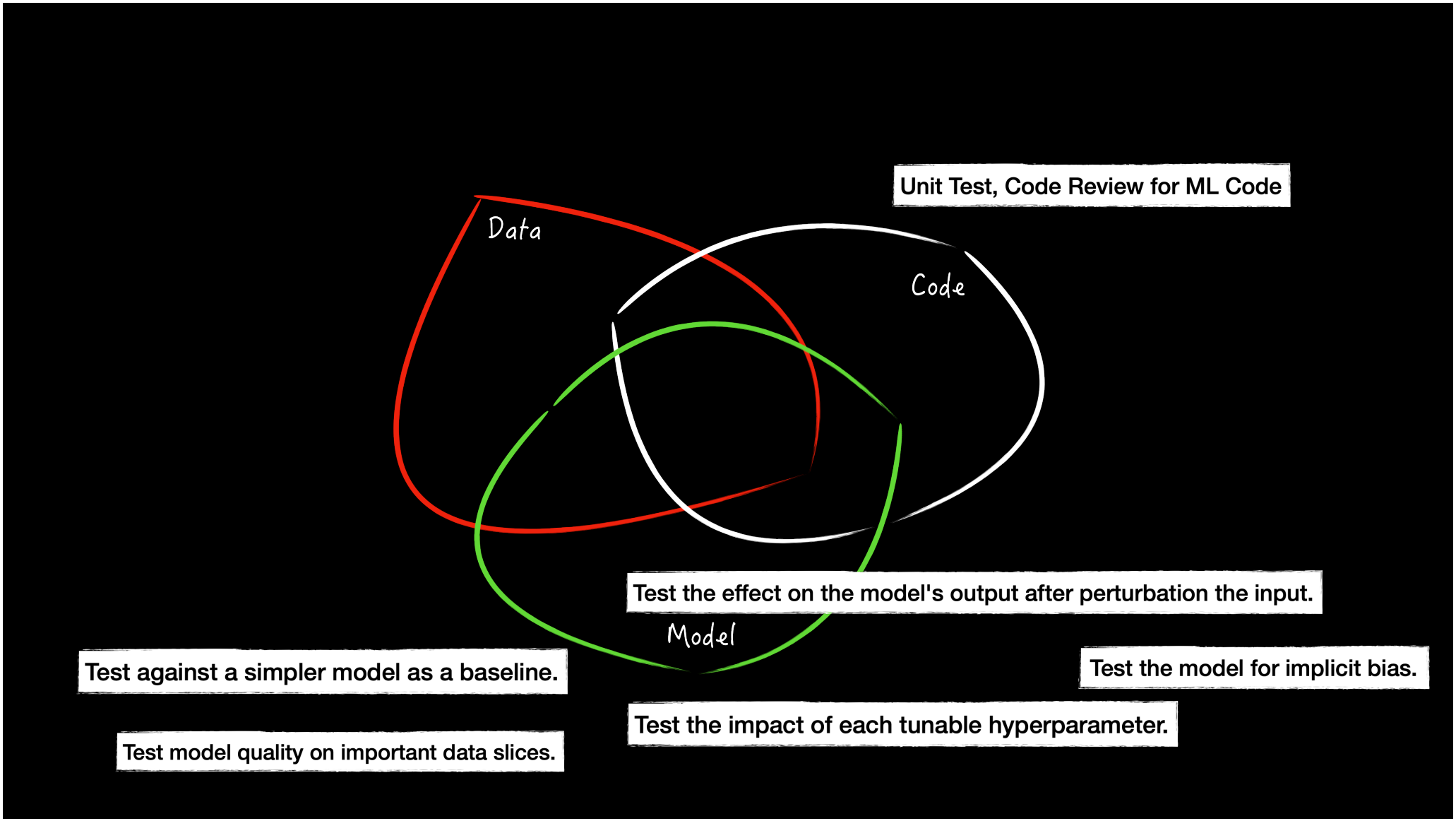
snorkel

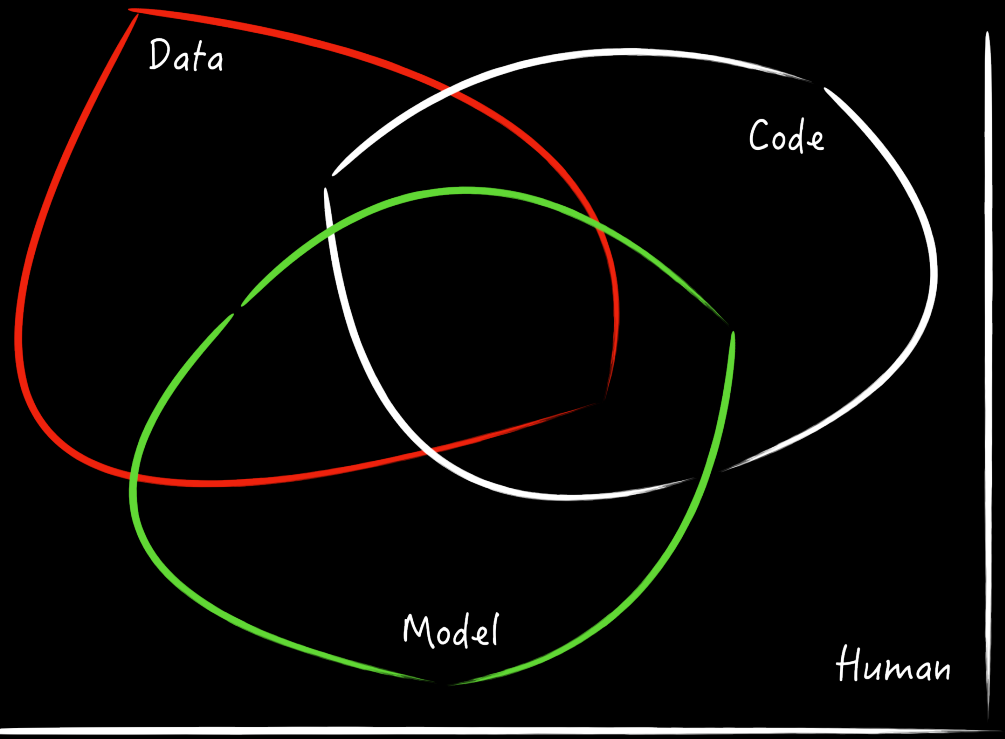
```python
from snorkel.slicing import slicing_function


@slicing_function()
def short_link(x):
    """Return whether text matches common pattern for shortened ".ly" links."""
    return int(bool(re.search(r"\w+\.ly", x.text)))


scorer.score_slices(
    S=S_test, golds=Y_test, preds=preds_test, probs=probs_test, as_dataframe=True
)
```

	F1
OVERALL	0.925000
SHORT_COMMENT	0.666667
KEYWORD_PLEASE	1.000000
REGEX_CHECK_OUT	1.000000
SHORT_LINK	0.500000
TEXTBLOB_POLARITY	0.727273

https://www.snorkel.org/use-cases/03-spam-data-slicing-tutorial

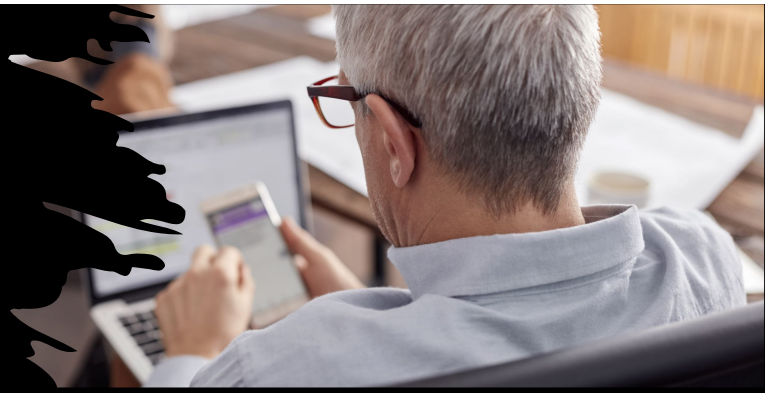
Vincent S. Chen, Sen Wu, Zhenzhen Weng, Alexander Ratner, Christopher Ré
"Slice-based Learning: A Programming Model for Residual Learning in Critical Data Slices"


Unit Test, Code Review for ML Code
Data
Code
Model
Test the effect on the model's output after perturbation the input.
Test against a simpler model as a baseline.
Test the model for implicit bias.
Test the impact of each tunable hyperparameter.
Test model quality on important data slices.



Test the relationship between offline proxy metrics and the actual online impact metrics.

Telemetry Design



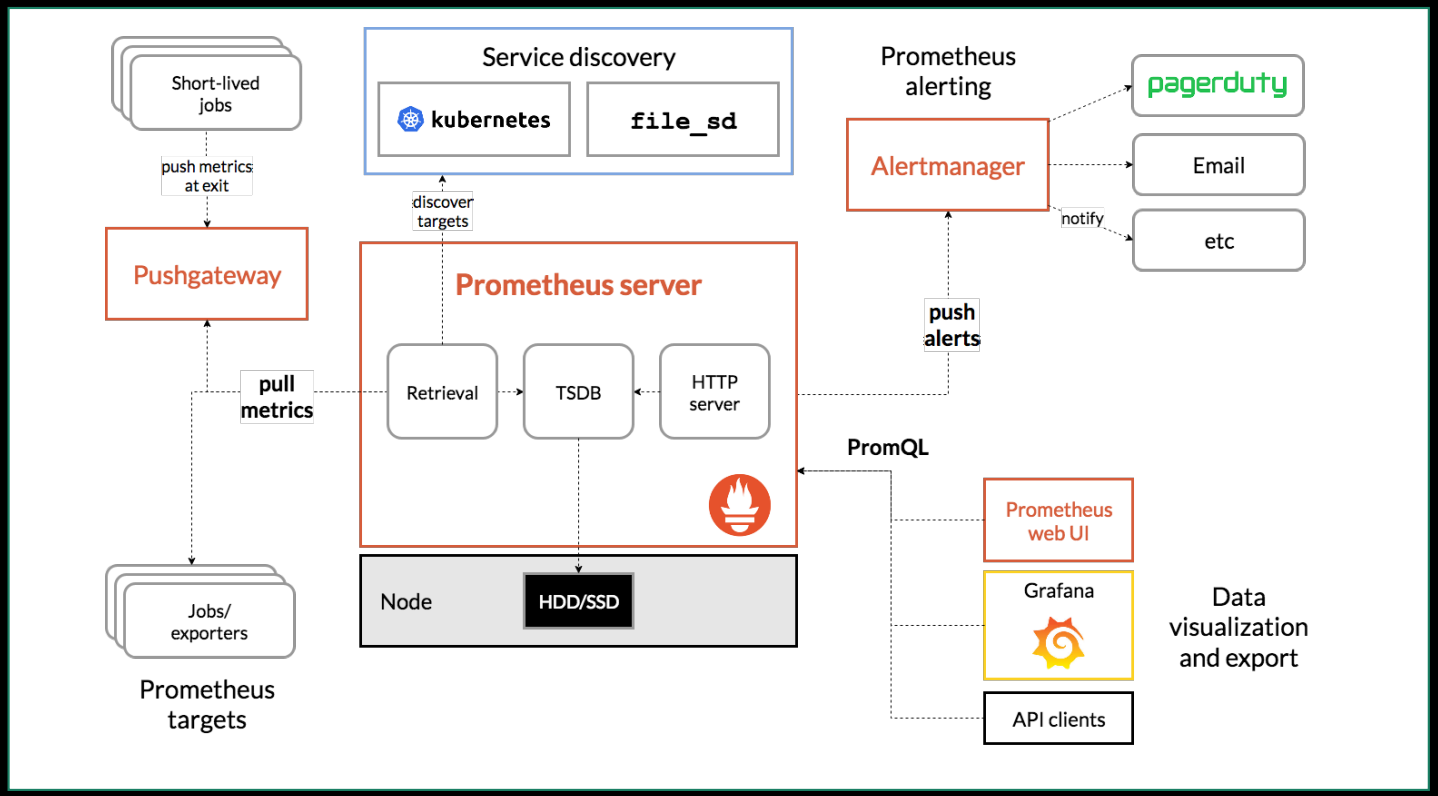
Responsible for collecting observations about how users are interacting with the system

- Monitoring system works correctly

- Understand the impact on users

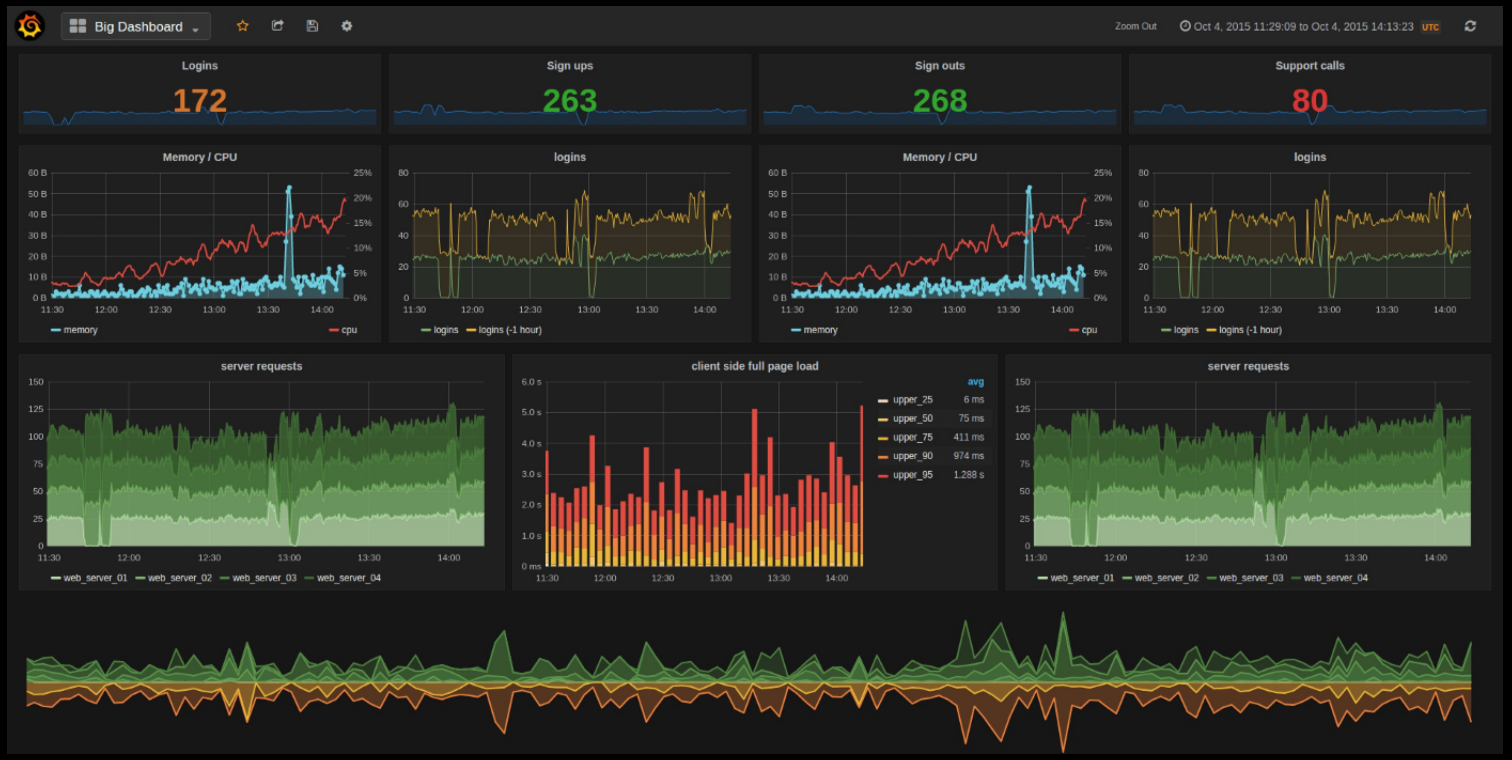
- Gather new training data

Monitoring and alerting



https://prometheus.io/docs/introduction/overview/

Monitoring and alerting



https://medium.com/@pacroy/application-telemetry-with-prometheus-sap-blogs-c4b5b6239d28ke

Understand Impact on Users

To determine if users are getting positive or negative outcomes and if the system is achieving its goals.

- Which experiences do users receive and how often do they receive them?
- What actions do users take in each experience?
- What experiences tend to drive users to look for help or to undo or revert their actions?
- What is the average time between users encountering a specific experience and leaving the application?
- Do users who interact more with the intelligent part of the system tend to be more or less engaged (or profitable) over time?

Activity

- Group 1: Amazon: Shopping app feature that detects the shoe brand from photos;
- Group 2: Google: Tagging uploaded photos with friends' names;
- Group 3: Spotify: Recommended personalized playlists;
- Group 4: Microsoft: Code completion recommendation in IDE.

- What information should the telemetry system capture?
- How are you going to use the information to identify and debug potential problems?
- How costly is it to collect the data? How do you plan to manage the cost?
- Any challenges/risks for the your telemetry design?

On next Tuesday:

Continuous Delivery for ML

Intro to Human-AI Interaction