SOFTWARE ENGINEERING FOR ML-ENABLED SYSTEMS

Christian Kästner

@p0nk

https://github.com/ckaestne/seai

CHRISTIAN KÄSTNER

@p0nk

Associate Professor @ CMU

Interests:

- Software Engineering
- Highly-Configurable Systems & Configuration Engineering
- Sustainability and Stress in Open Source
- Software Engineering for ML-Enabled Systems

SOFTWARE ENGINEERING FOR ML-ENABLED SYSTEMS

Building, operating, and maintaining software systems with machine-learned components

*with interdisciplinary collaborative teams of **data scientists** and **software engineers***

SE FOR ML-ENABLED SYSTEMS != BUILDING MODELS

G4 playground.ipynb

File Edit View Insert Runtime Tools Help Last edited on April 4

Comment Share

+ Code + Text Connect

1096	4	12	26	3	2	0
235	4	4	23	1	2	0

525 rows × 6 columns

```
# learning a classifier whether the result will be nonZero

from sklearn import tree

classifier=tree.DecisionTreeClassifier(max_depth=8)
classifier=classifier.fit(Xtrain,ynztrain)

print(classifier.score(Xtrain, ynztrain))
print(classifier.score(Xtest, ynztest))
```

```
0.8266666666666667
0.7295238095238096
```

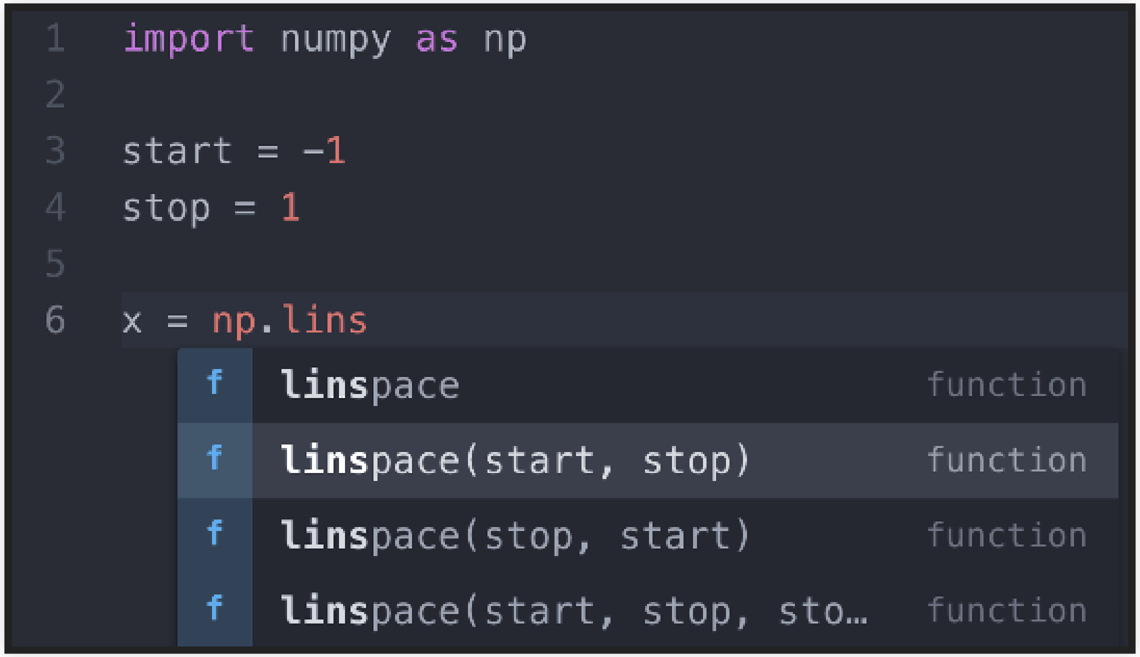
```
# learning a regression model only on the nonZero data (test is on all data and somewhat
```

```python
from sklearn import tree

predictor=tree.DecisionTreeRegressor(max_depth=8)
predictor=predictor.fit(XnzTrain,YnzTrain)


print(predictor.score(XnzTrain, YnzTrain))
print(predictor.score(Xtest, ytest))
```

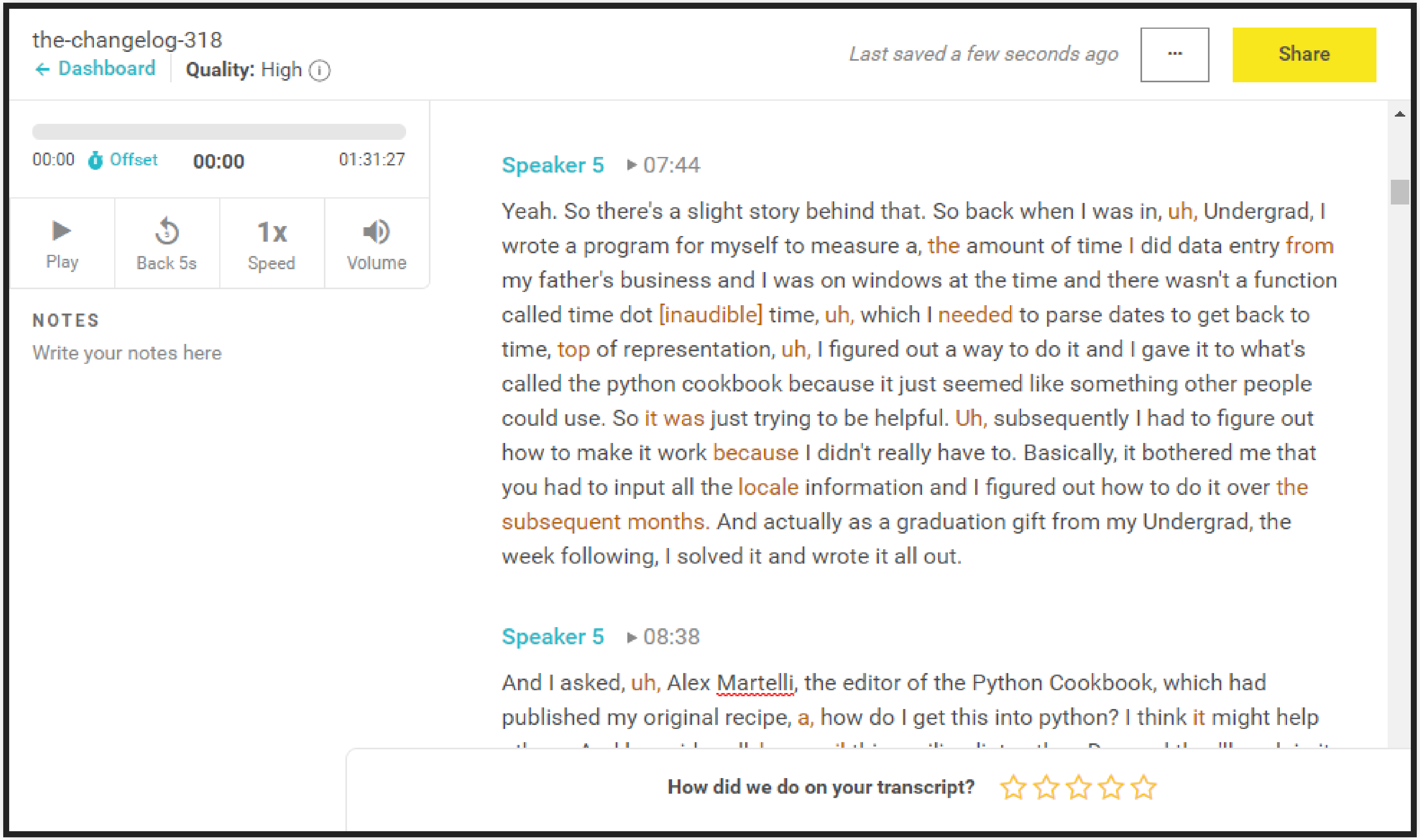
```
0.9376379365613154
-2.437397740412892
```

SE FOR ML-ENABLED SYSTEMS != CODING ML FRAMEWORKS

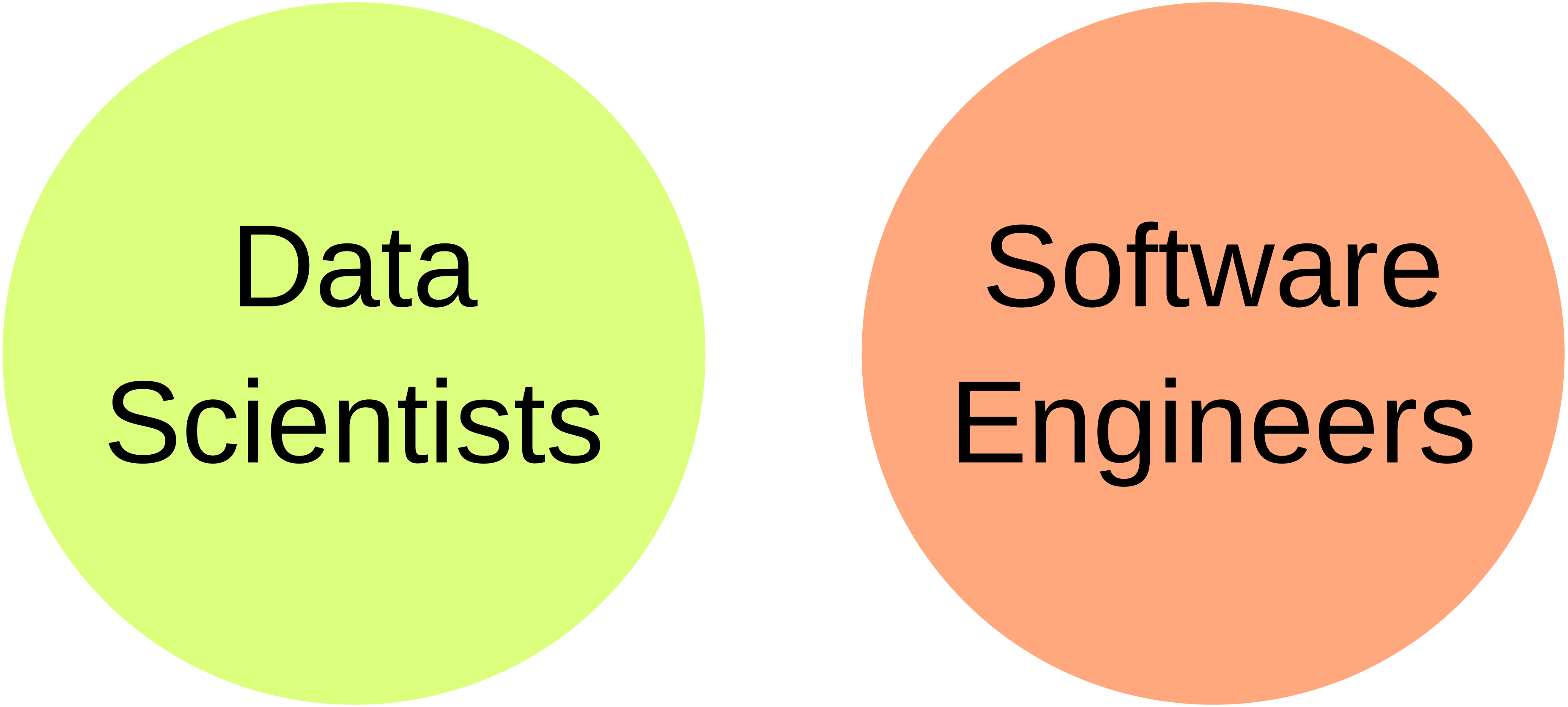
SE FOR ML-ENABLED SYSTEMS != ML FOR SE TOOLS



SE FOR ML-ENABLED (AI-ML-BASED, ML-INFUSED) SYSTEMS


the-changelog-318
Dashboard
Quality: High
Last saved a few seconds ago
...
Share
00:00
Offset
00:00
01:31:27
Play
Back 5s
1x
Speed
Volume
NOTES
Write your notes here
Speaker 5 07:44
Yeah. So there's a slight story behind that. So back when I was in, uh, Undergrad, I wrote a program for myself to measure a, the amount of time I did data entry from my father's business and I was on windows at the time and there wasn't a function called time dot [inaudible] time, uh, which I needed to parse dates to get back to time, top of representation, uh, I figured out a way to do it and I gave it to what's called the python cookbook because it just seemed like something other people could use. So it was just trying to be helpful. Uh, subsequently I had to figure out how to make it work because I didn't really have to. Basically, it bothered me that you had to input all the locale information and I figured out how to do it over the subsequent months. And actually as a graduation gift from my Undergrad, the week following, I solved it and wrote it all out.
Speaker 5 08:38
And I asked, uh, Alex Martelli, the editor of the Python Cookbook, which had published my original recipe, a, how do I get this into python? I think it might help
How did we do on your transcript?

temi.com


Data Scientists
Software Engineers

SOFTWARE ENGINEERING

> *Software engineering is the branch of computer science that creates practical, cost-effective solutions to computing and information processing problems, preferentially by applying scientific knowledge, developing software systems in the service of mankind.*

Engineering judgements under limited information and resources

A focus on design, tradeoffs, and the messiness of the real world

Many qualities of concern: cost, correctness, performance, scalability, security, maintainability, ...

"it depends..."

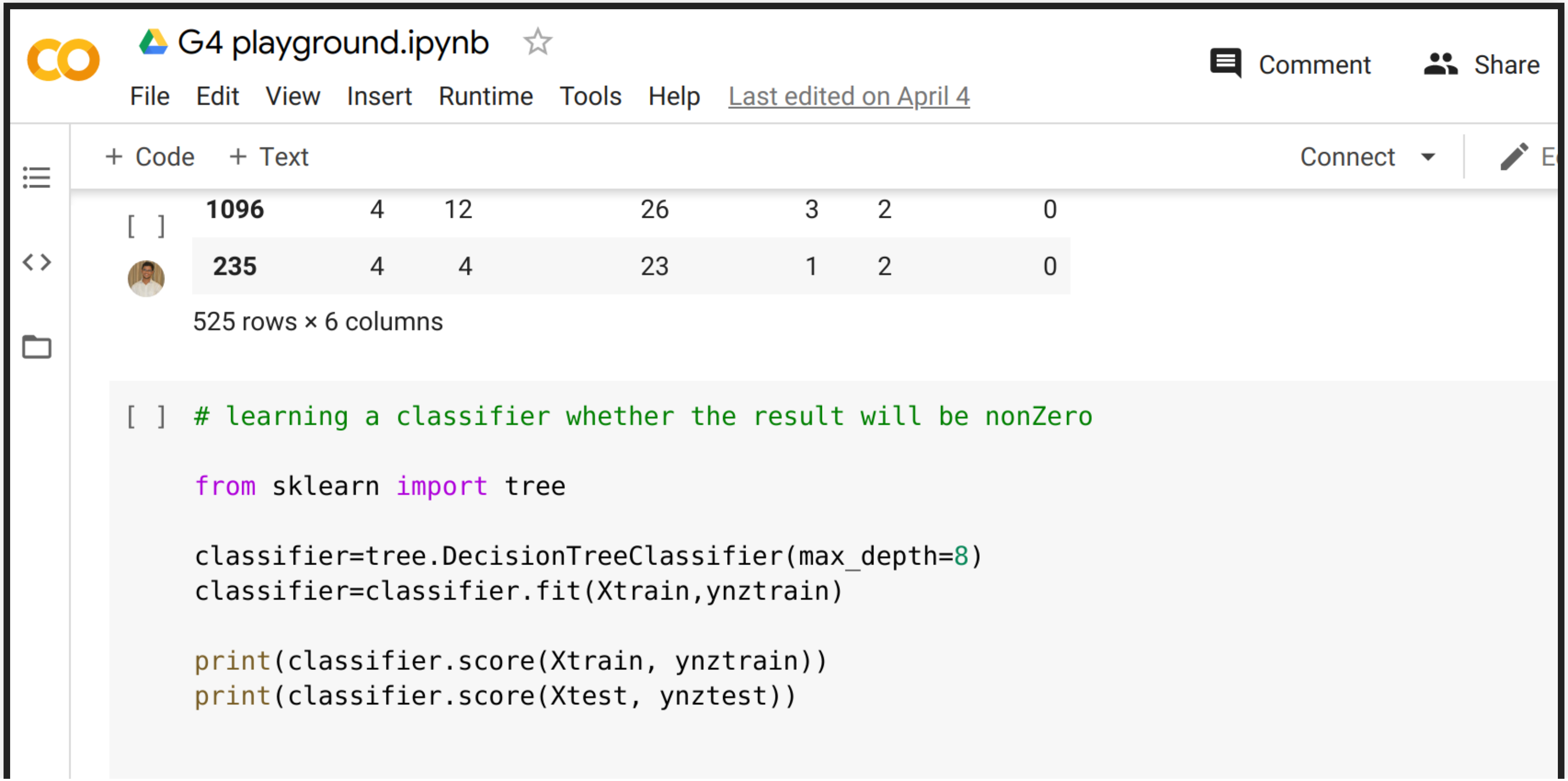
Mary Shaw. ed. Software Engineering for the 21st Century: A basis for rethinking the curriculum. 2005.

MOST ML COURSES/TALKS

Focus narrowly on modeling techniques or building models

Using notebooks, static datasets, evaluating accuracy

Little attention to software engineering aspects of building complete systems



```
0.8266666666666667
0.7295238095238096
```

```
[ ] # learning a regression model only on the nonZero data (test is on all data and somewhat

    from sklearn import tree

    predictor=tree.DecisionTreeRegressor(max_depth=8)
    predictor=predictor.fit(XnzTrain,YnzTrain)


    print(predictor.score(XnzTrain, YnzTrain))
    print(predictor.score(Xtest, ytest))
```

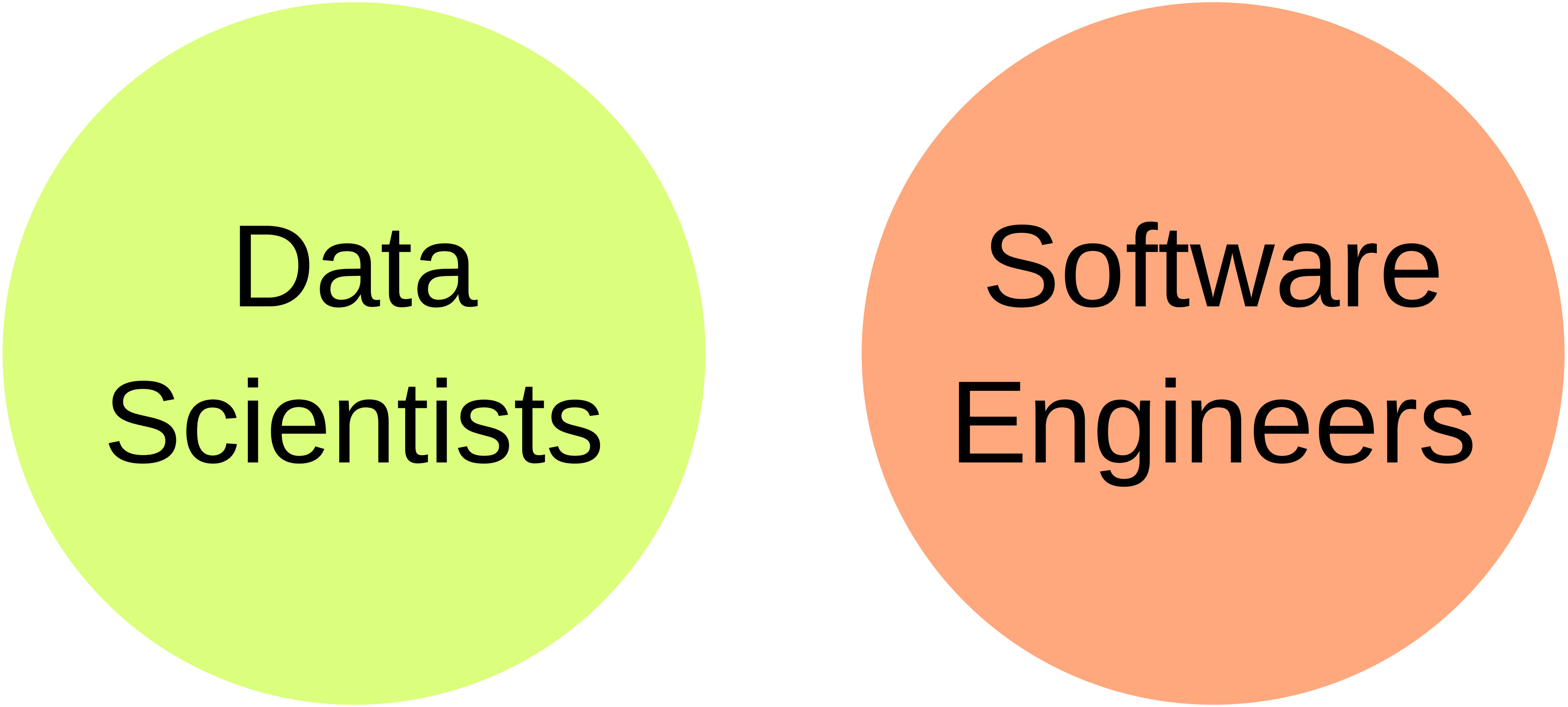
```
0.9376379365613154
-2.437397740412892
```

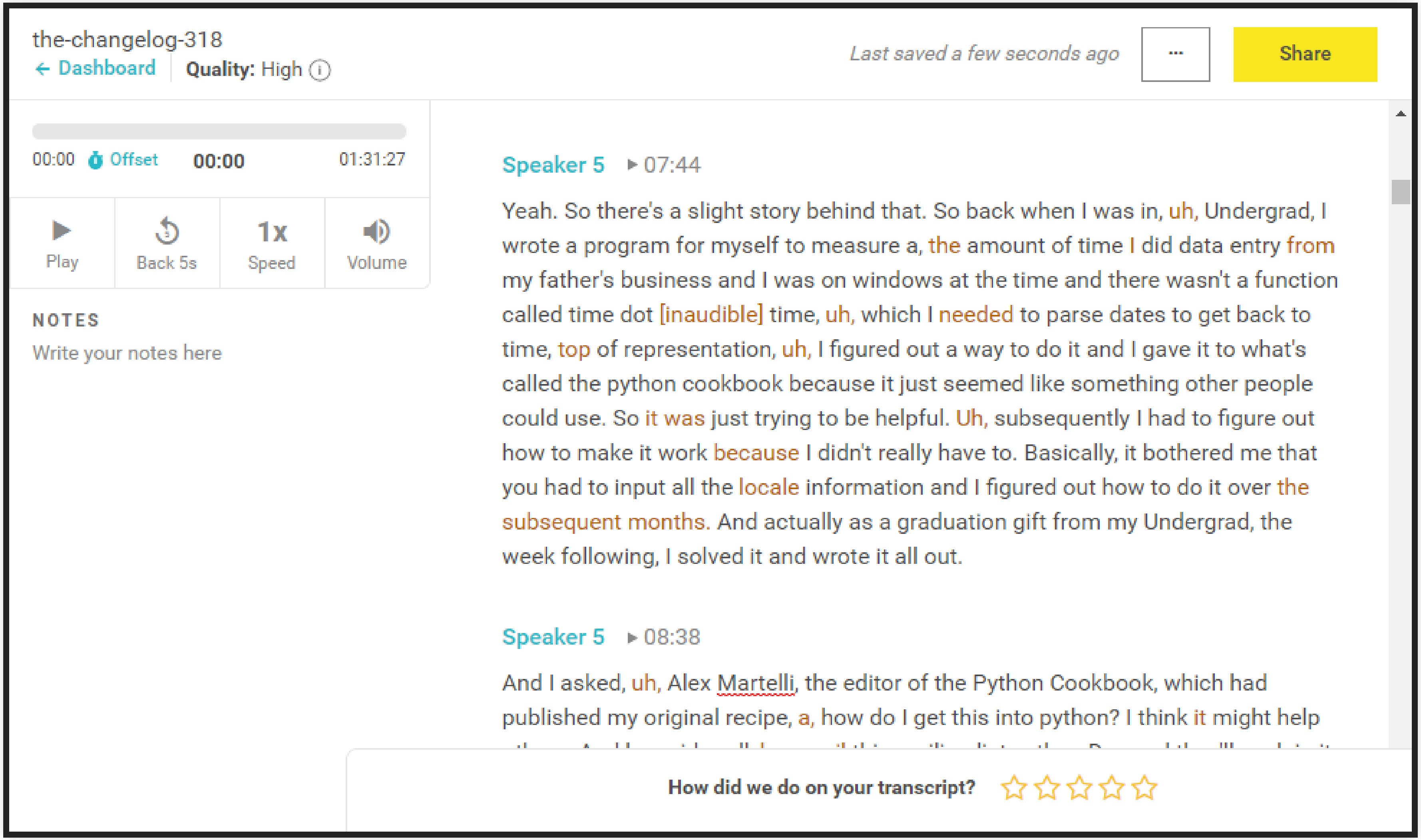
DATA SCIENTIST

- Often fixed dataset for training and evaluation (e.g., PBS interviews)
- Focused on accuracy
- Prototyping, often Jupyter notebooks or similar
- Expert in modeling techniques and feature engineering
- Model size, updateability, implementation stability typically does not matter
- Starting to worry about fairness, robustness, ...

SOFTWARE ENGINEER

- Builds a product
- Concerned about cost, performance, stability, release time
- Identify quality through customer satisfaction
- Must scale solution, handle large amounts of data
- Plan for mistakes and safeguards
- Maintain, evolve, and extend the product over long periods
- Consider requirements for security, safety, fairness


Data Scientists
Software Engineers


the-changelog-318
← Dashboard
Quality: High
Last saved a few seconds ago
···
Share
00:00
Offset
00:00
01:31:27
Play
Back 5s
1x
Speed
Volume
NOTES
Write your notes here
Speaker 5 ▸ 07:44
Yeah. So there's a slight story behind that. So back when I was in, uh, Undergrad, I wrote a program for myself to measure a, the amount of time I did data entry from my father's business and I was on windows at the time and there wasn't a function called time dot [inaudible] time, uh, which I needed to parse dates to get back to time, top of representation, uh, I figured out a way to do it and I gave it to what's called the python cookbook because it just seemed like something other people could use. So it was just trying to be helpful. Uh, subsequently I had to figure out how to make it work because I didn't really have to. Basically, it bothered me that you had to input all the locale information and I figured out how to do it over the subsequent months. And actually as a graduation gift from my Undergrad, the week following, I solved it and wrote it all out.
Speaker 5 ▸ 08:38
And I asked, uh, Alex Martelli, the editor of the Python Cookbook, which had published my original recipe, a, how do I get this into python? I think it might help
How did we do on your transcript?

A SOFTWARE ENGINEERING PERSPECTIVE ON ML

WHAT'S DIFFERENT?

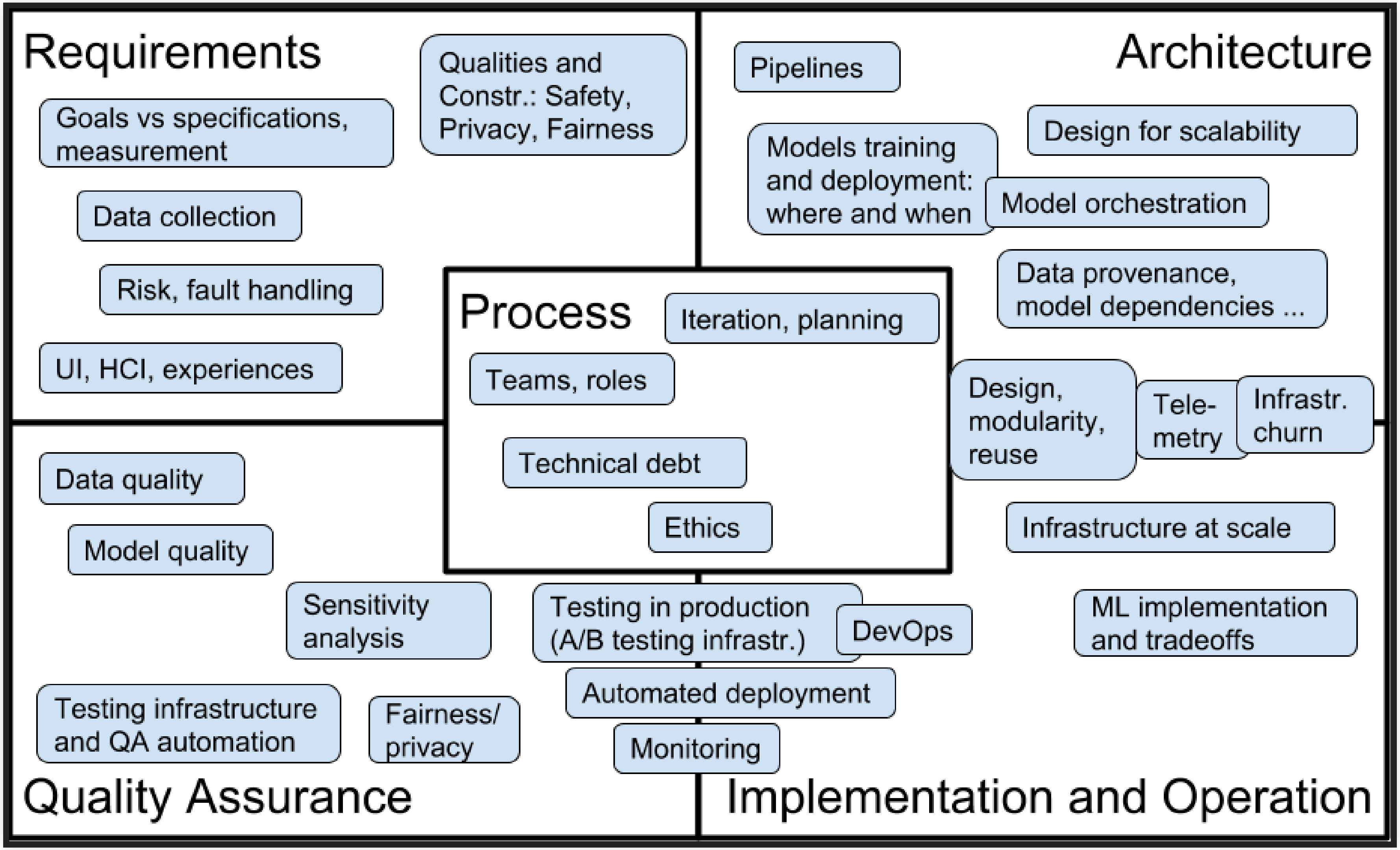
- Missing specifications
- Environment is important (feedback loops, data drift)
- Nonlocal and nonmonotonic effects
- Testing in production
- Data management, versioning, and provenance

REALLY DIFFERENT?

- Missing specifications -- *implicit, vague specs very common; safe systems from unreliable components* ("ML is requirements engineering")
- Environment is important -- *the world vs the machine* (paper)
- Nonlocal and nonmonotonic effects -- *feature interactions, system testing*
- Testing in production -- *continuous deployment, A/B testing*
- Data management, versioning, and provenance -- *stream processing, event sourcing, data modeling*

EXAMPLES OF SOFTWARE ENGINEERING CONCERNS

- How to build robust AI pipelines and facilitate regular model updates?
- How to deploy and update models in production?
- How to evaluate data and model quality in production?
- How to deal with mistakes that the model makes and manage associated risk?
- How to trade off between various qualities, including learning cost, inference time, updatability, and interpretability?
- How to design a system that scales to large amounts of data?
- How to version models and data?
- How to manage interdisciplinary teams with data scientists, software engineers, and operators?



https://github.com/ckaestne/seai

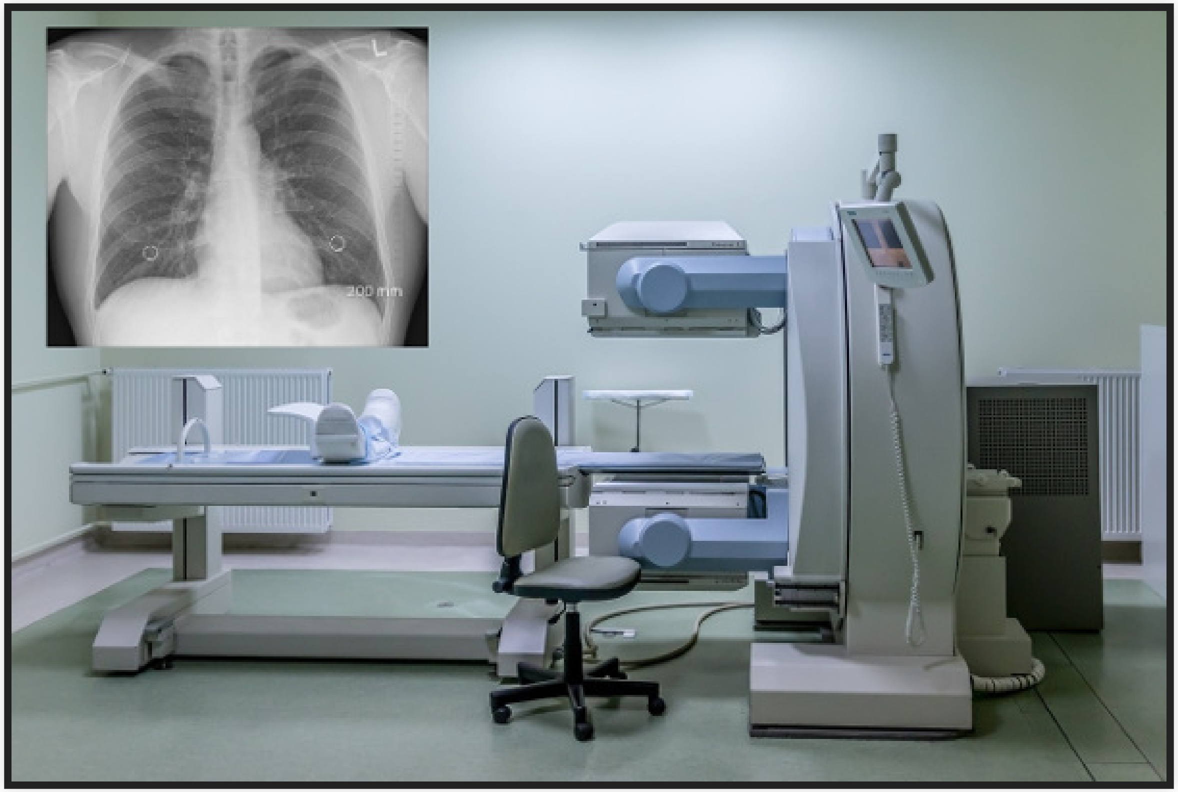
MY VIEW

While developers of simple traditional systems may get away with poor practices, most developers of ML-enabled systems will not.

QUALITY ASSURANCE FOR ML-ENABLED SYSTEMS

TRADITIONAL FOCUS: MODEL ACCURACY

- Train and evaluate model on fixed labled data set
- Compare prediction with labels



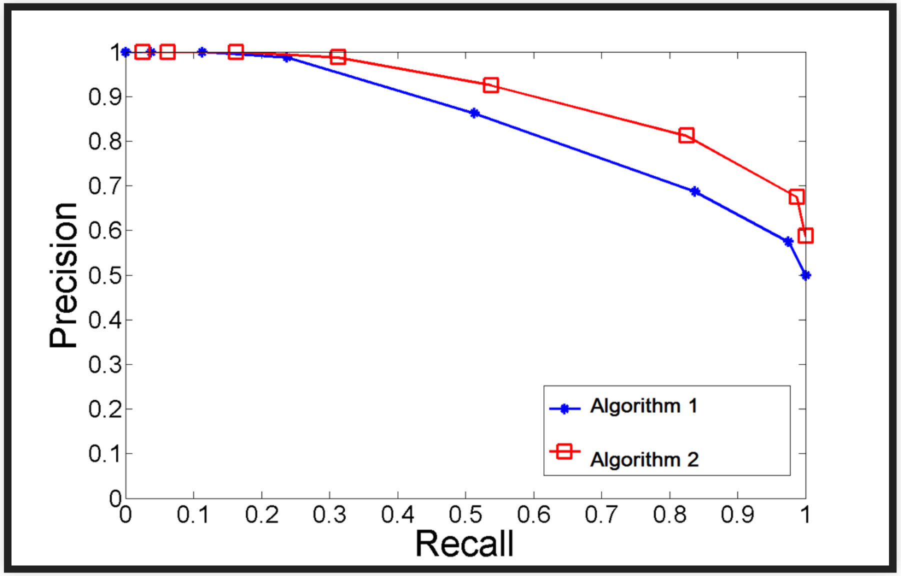
TRADITIONAL FOCUS: MODEL ACCURACY

	Actually A	Actually not A
AI predicts A	True Positive (TP)	False Positive (FP)
AI predicts not A	False Negative (FN)	True Negative (TN)

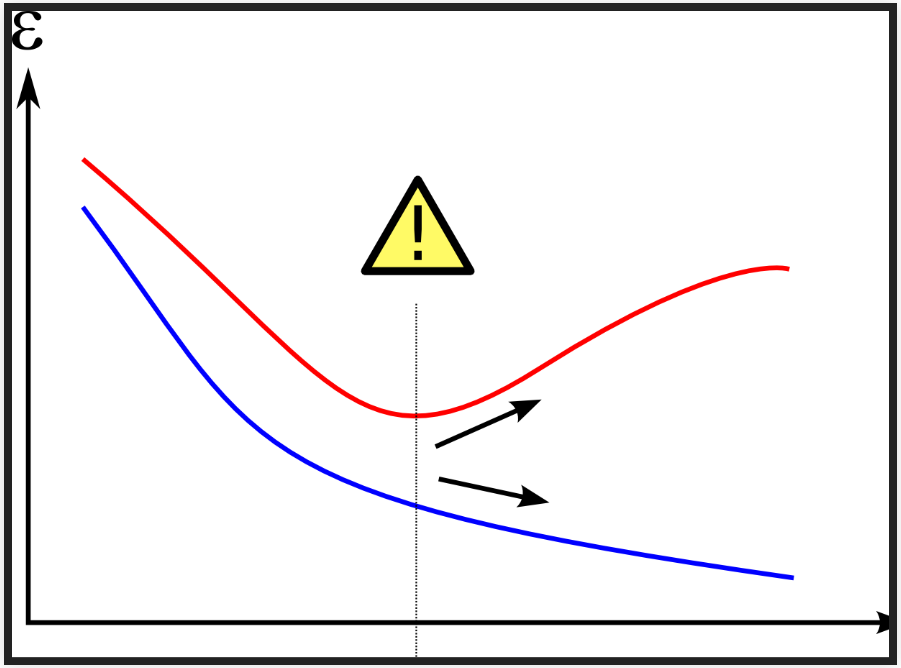
Accuary, Recall, Precision, F1-Score

MORE TRADITIONAL MODEL QUALITY DISCUSSIONS

Many model quality metrics (recall, MAPE, ROC, log loss, top-k, ...)



Generalization/overfitting (train/test/eval split, crossvalidation)



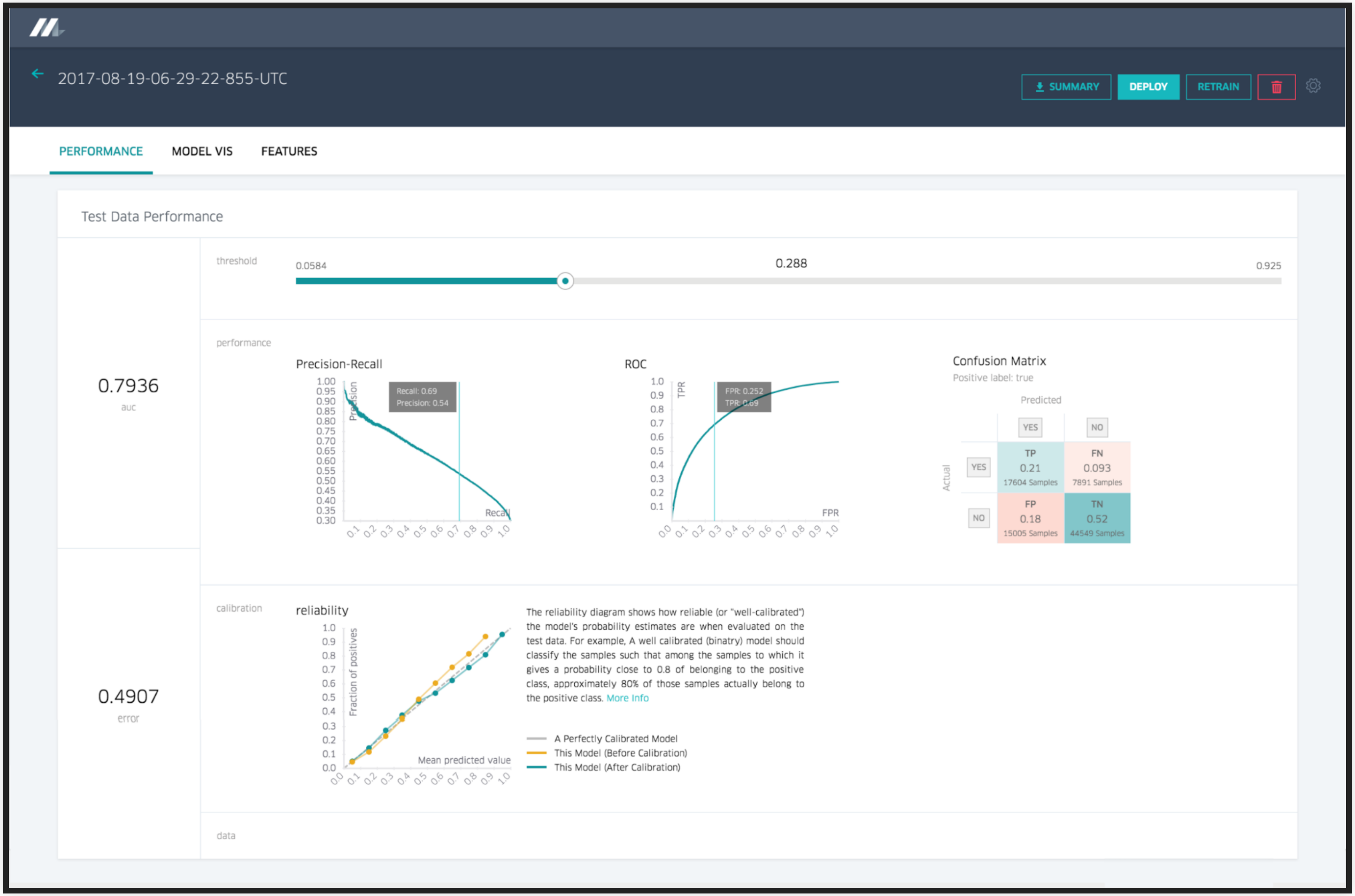
NOT ALL MISTAKES ARE EQUAL

- False positives vs false negatives (e.g., cancer detection)
- Fairness across subpopulations

- Learn from black-box testing:
    - Equivalence classes
    - Boundary conditions
    - Critical test cases ("call mom")
    - Combinatorial testing
    - Fuzzing

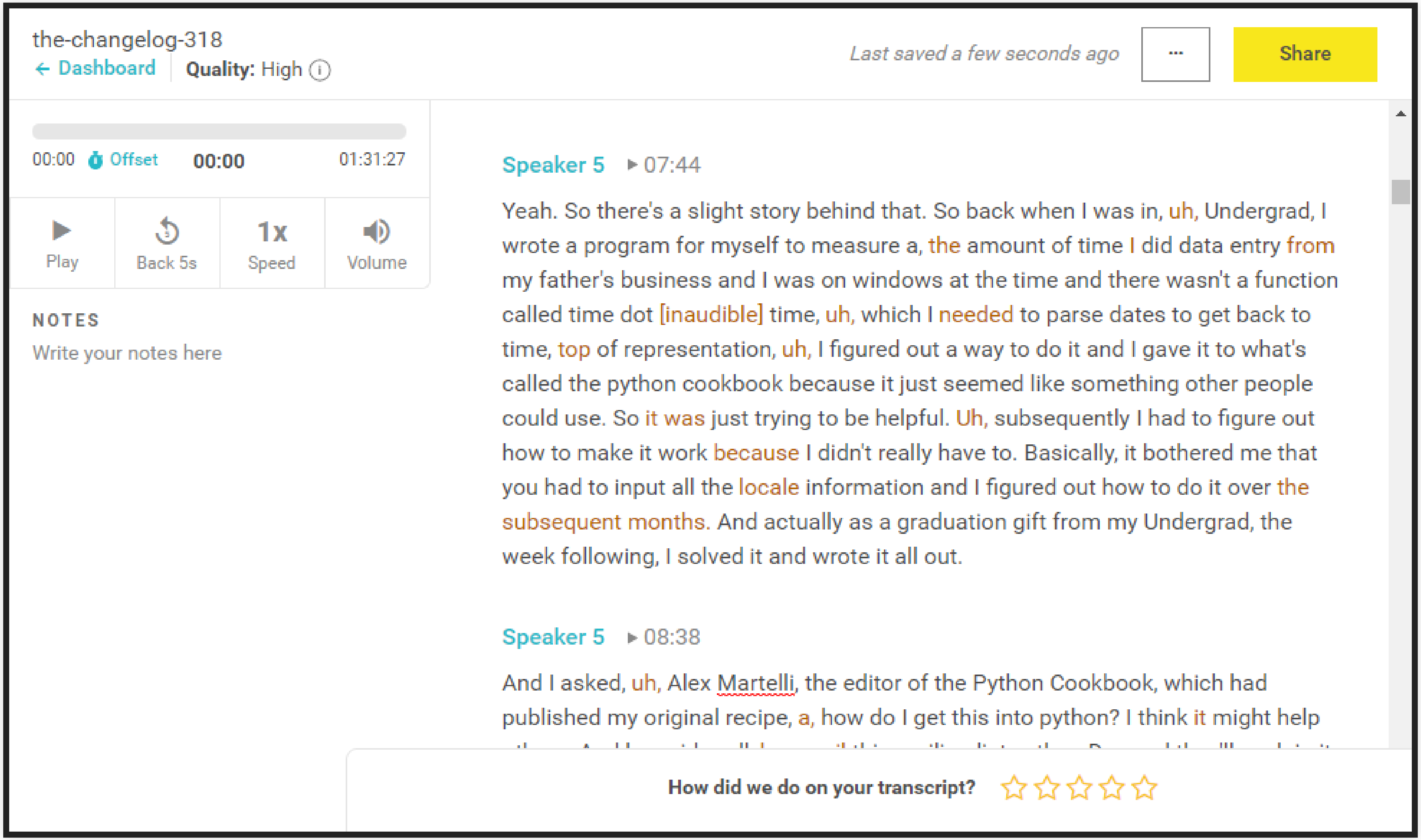
AUTOMATING MODEL EVALUATION

- Continuous integration, automated measurement, tracking of results
- Data and model versioning, provenance


2017-08-19-06-29-22-855-UTC
SUMMARY
DEPLOY
RETRAIN
PERFORMANCE
MODEL VIS
FEATURES
Test Data Performance
threshold
0.0584
0.288
0.925
0.7936
auc
performance
Precision-Recall
Recall: 0.69
Precision: 0.54
Precision
Recall
ROC
FPR: 0.252
TPR: 0.69
TPR
FPR
Confusion Matrix
Positive label: true
Predicted
YES
NO
Actual
TP 0.21 17604 Samples
FN 0.093 7891 Samples
FP 0.18 15005 Samples
TN 0.52 44549 Samples
0.4907
error
calibration
reliability
Fraction of positives
Mean predicted value
The reliability diagram shows how reliable (or "well-calibrated") the model's probability estimates are when evaluated on the test data. For example, A well calibrated (binatry) model should classify the samples such that among the samples to which it gives a probability close to 0.8 of belonging to the positive class, approximately 80% of those samples actually belong to the positive class. More Info
A Perfectly Calibrated Model
This Model (Before Calibration)
This Model (After Calibration)
data

QUALITY CONCERNS FOR ML-ENABLED SYSTEMS

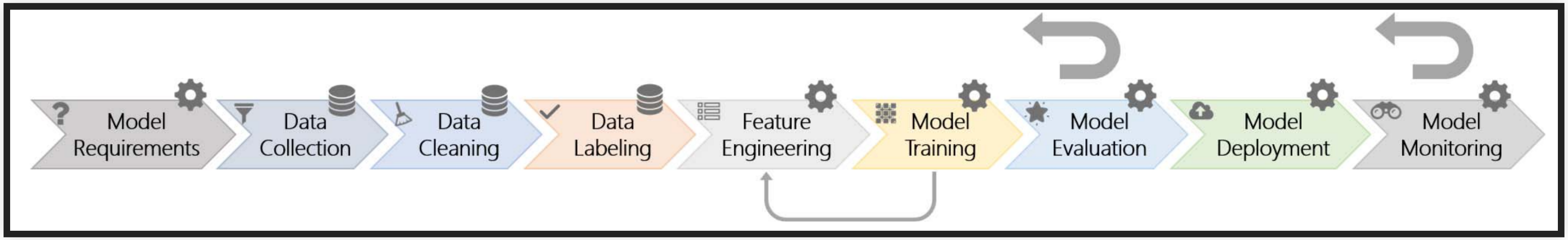
- Learning time, cost and scalability
- Update cost, incremental learning
- Inference cost
- Size of models learned
- Amount of training data needed
- Fairness
- Robustness
- Safety, security, privacy
- Explainability, reproducibility
- Time to market
- Overall operating cost (cost per prediction)


the-changelog-318
← Dashboard
Quality: High
Last saved a few seconds ago
···
Share
00:00
Offset
00:00
01:31:27
Play
Back 5s
1x
Speed
Volume
NOTES
Write your notes here
Speaker 5
07:44
Yeah. So there's a slight story behind that. So back when I was in, uh, Undergrad, I wrote a program for myself to measure a, the amount of time I did data entry from my father's business and I was on windows at the time and there wasn't a function called time dot [inaudible] time, uh, which I needed to parse dates to get back to time, top of representation, uh, I figured out a way to do it and I gave it to what's called the python cookbook because it just seemed like something other people could use. So it was just trying to be helpful. Uh, subsequently I had to figure out how to make it work because I didn't really have to. Basically, it bothered me that you had to input all the locale information and I figured out how to do it over the subsequent months. And actually as a graduation gift from my Undergrad, the week following, I solved it and wrote it all out.
Speaker 5
08:38
And I asked, uh, Alex Martelli, the editor of the Python Cookbook, which had published my original recipe, a, how do I get this into python? I think it might help
How did we do on your transcript?

6.8

INFRASTRUCTURE QUALITY

THINK OF PIPELINES, NOT MODELS, NOT NOTEBOOKS



Many steps: Data collection, data cleaning, labeling, feature engineering, training, evaluation, deployment, monitoring

Automate each step -- test each step

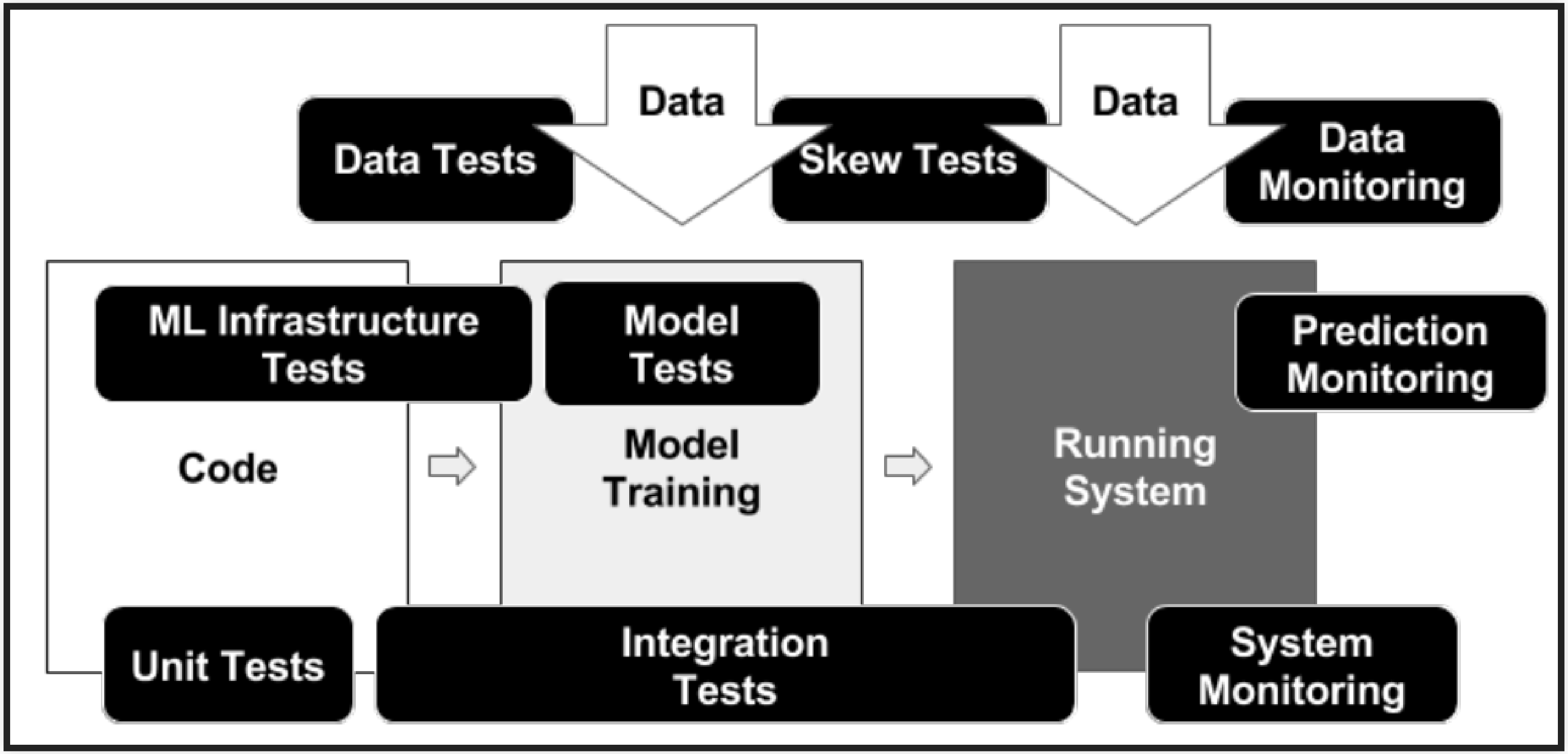
Graphic: Amershi, Saleema, Andrew Begel, Christian Bird, Robert DeLine, Harald Gall, Ece Kamar, Nachiappan Nagappan, Besmira Nushi, and Thomas Zimmermann. "Software engineering for machine learning: A case study." In 2019 IEEE/ACM 41st International Conference on Software Engineering: Software Engineering in Practice (ICSE-SEIP), pp. 291-300. IEEE, 2019.

POSSIBLE MISTAKES IN ML PIPELINES

Danger of "silent" mistakes in many phases:

- Dropped data after format changes
- Failure to push updated model into production
- Incorrect feature extraction
- Use of stale dataset, wrong data source
- Data source no longer available (e.g web API)
- Telemetry server overloaded
- Negative feedback (telemtr.) no longer sent from app
- Use of old model learning code, stale hyperparameter
- Data format changes between ML pipeline steps
- ...

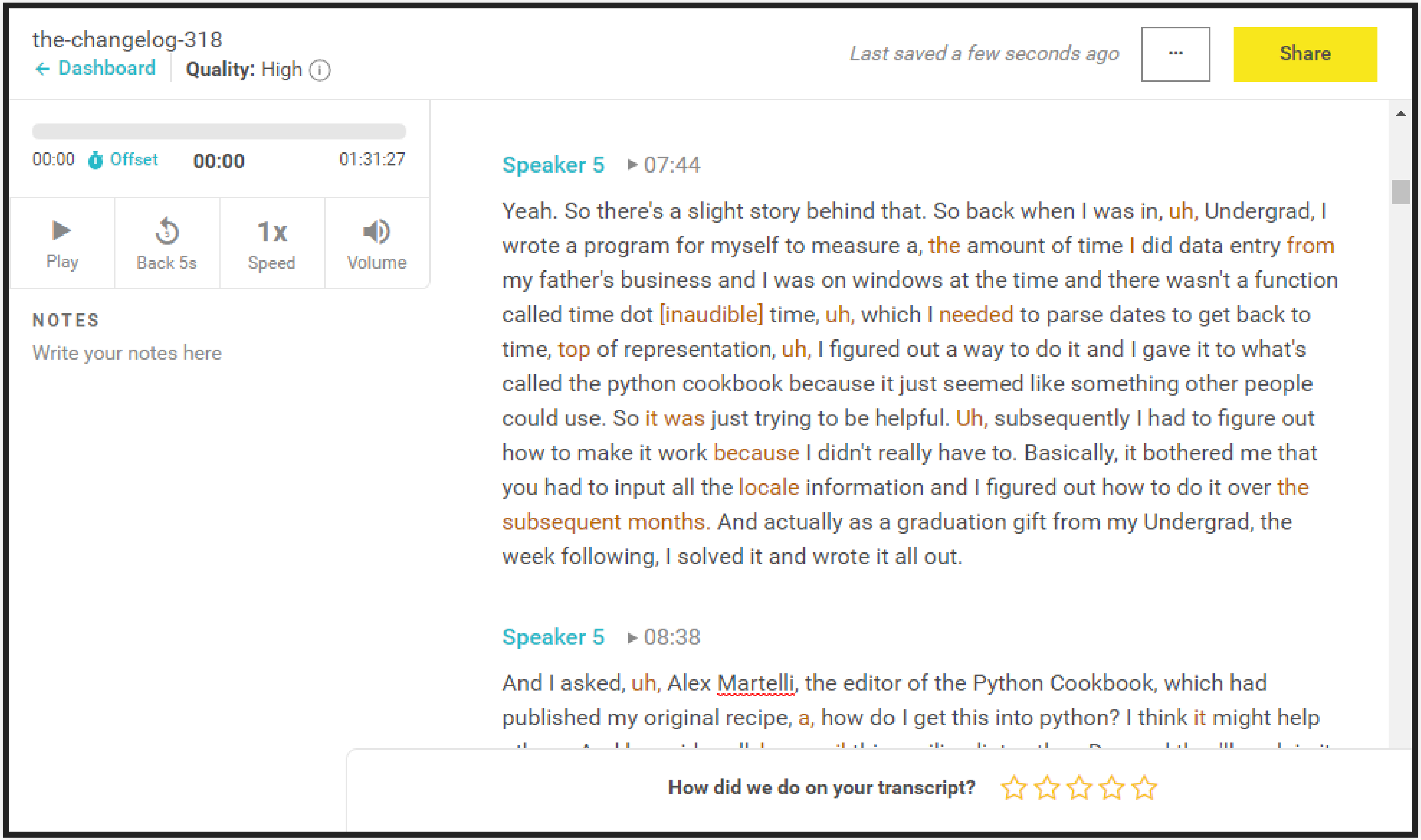
QUALITY ASSURANCE FOR THE ENTIRE PIPELINE



Source: Eric Breck, Shanqing Cai, Eric Nielsen, Michael Salib, D. Sculley. [The ML Test Score: A Rubric for ML Production Readiness and Technical Debt Reduction](). Proceedings of IEEE Big Data (2017)

PIPELINE TESTING

- Unit tests (e.g., data cleaning)
- End to end pipeline tests
- Testing with stubs, test error handling (e.g., test model redeployment after dropped connection)
- Test monitoring infrastructure (e.g., "fire drills")


the-changelog-318
Dashboard
Quality: High
Last saved a few seconds ago
···
Share
00:00
Offset
00:00
01:31:27
Play
Back 5s
1x
Speed
Volume
NOTES
Write your notes here
Speaker 5
07:44
Yeah. So there's a slight story behind that. So back when I was in, uh, Undergrad, I wrote a program for myself to measure a, the amount of time I did data entry from my father's business and I was on windows at the time and there wasn't a function called time dot [inaudible] time, uh, which I needed to parse dates to get back to time, top of representation, uh, I figured out a way to do it and I gave it to what's called the python cookbook because it just seemed like something other people could use. So it was just trying to be helpful. Uh, subsequently I had to figure out how to make it work because I didn't really have to. Basically, it bothered me that you had to input all the locale information and I figured out how to do it over the subsequent months. And actually as a graduation gift from my Undergrad, the week following, I solved it and wrote it all out.
Speaker 5
08:38
And I asked, uh, Alex Martelli, the editor of the Python Cookbook, which had published my original recipe, a, how do I get this into python? I think it might help
How did we do on your transcript?

7.6

THINKING OF THE ENTIRE SYSTEM

ML models are "just" one component

LIVING WITH MISTAKES

The smart toaster may occasionally burn my toast, but it should not burn down my kitchen.

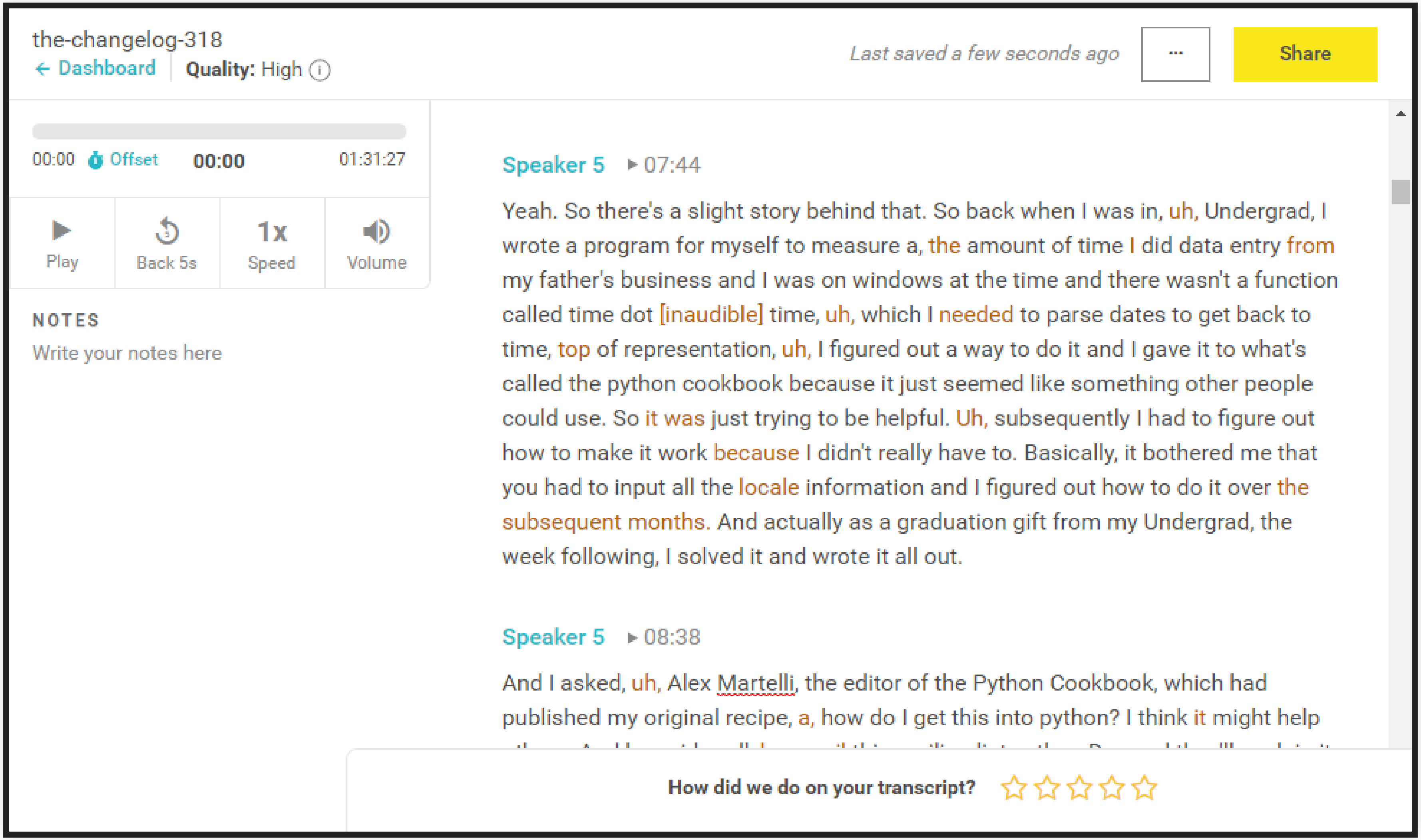
Speaker notes

A smart toaster may occasionally burn the toast, but it should never burn down the kitchen. The latter can be achieved without relying on perfect accuarcy of a smart component, just stop it when it's overheating.

Plan for mistakes: User interaction, undo, safeguards

MODEL ACCURACY VS SYSTEM GOALS

- System goals are supported by AI components, e.g.,
    - maximizing sales
    - minimizing loss
    - maximizing community growth
    - retaining customers
    - maximizing engagement time
- A better model will support system goals better
    - more accurate
    - faster answers
    - fewer bad mistakes
    - more explainable
    - easier to evolve


the-changelog-318
Dashboard
Quality: High
Last saved a few seconds ago
···
Share
00:00
Offset
00:00
01:31:27
Play
Back 5s
1x
Speed
Volume
NOTES
Write your notes here
Speaker 5
07:44
Yeah. So there's a slight story behind that. So back when I was in, uh, Undergrad, I wrote a program for myself to measure a, the amount of time I did data entry from my father's business and I was on windows at the time and there wasn't a function called time dot [inaudible] time, uh, which I needed to parse dates to get back to time, top of representation, uh, I figured out a way to do it and I gave it to what's called the python cookbook because it just seemed like something other people could use. So it was just trying to be helpful. Uh, subsequently I had to figure out how to make it work because I didn't really have to. Basically, it bothered me that you had to input all the locale information and I figured out how to do it over the subsequent months. And actually as a graduation gift from my Undergrad, the week following, I solved it and wrote it all out.
Speaker 5
08:38
And I asked, uh, Alex Martelli, the editor of the Python Cookbook, which had published my original recipe, a, how do I get this into python? I think it might help
How did we do on your transcript?

TESTING IN PRODUCTION

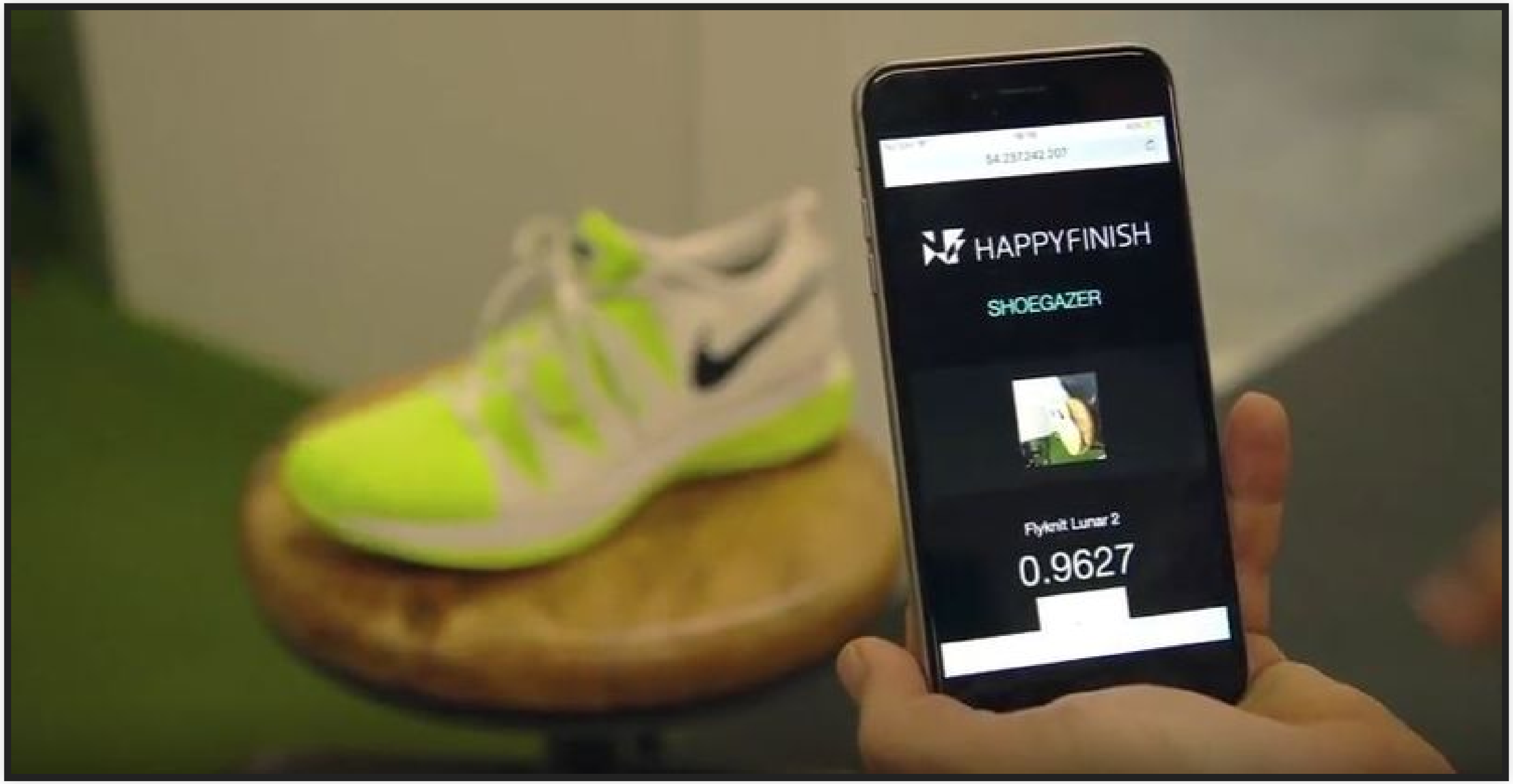
Production data = ultimate unseen data

Focus on system goals, not model accuracy

Monitoring performance over time, canary releases

Finding and debugging common mistakes

Experimentation with A/B tests



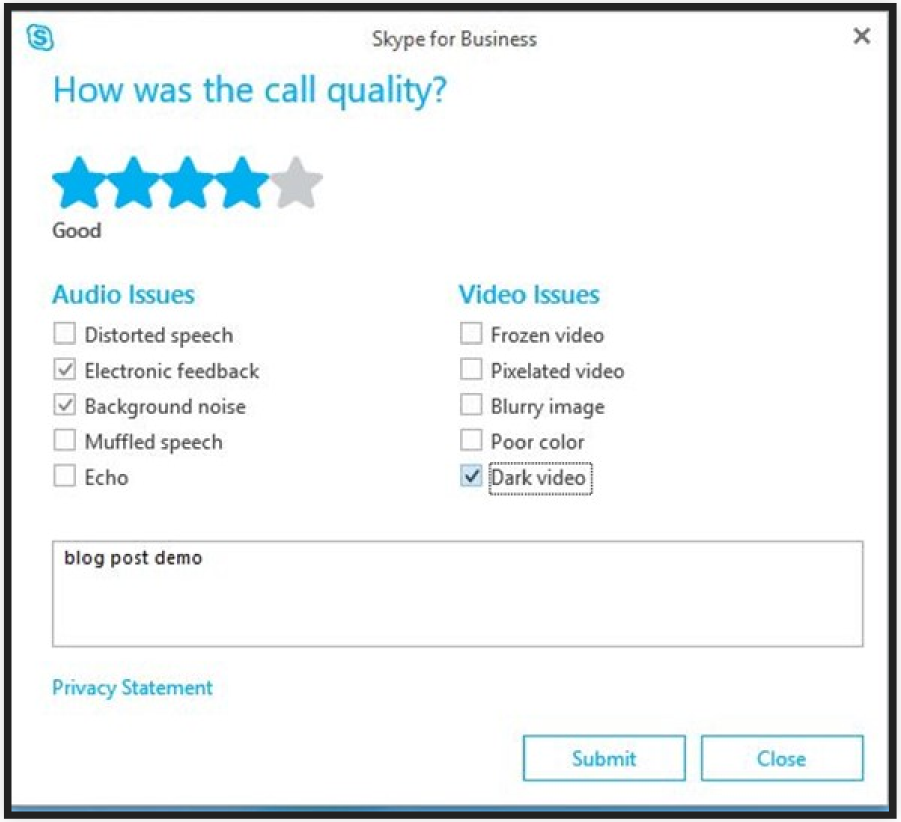
Source: https://www.trendhunter.com/trends/shoegazer

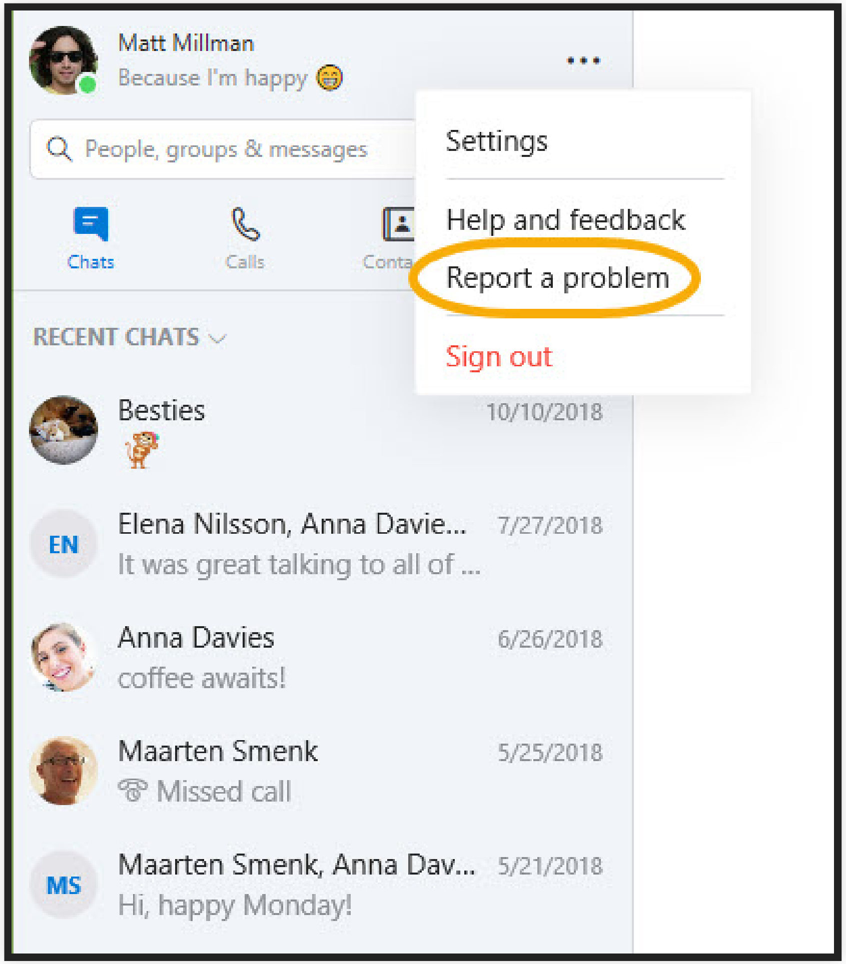
KEY DESIGN CHALLENGE: TELEMETRY

- Identify mistakes in production (“what would have been the right prediction?”)
- Many challenges:
  - How can we identify mistakes? Both false positives and false negatives?
  - How can we collect feedback without being intrusive (e.g., asking users about every interactions)?
  - How much data are we collecting? Can we manage telemetry at scale? How to sample properly?
  - How do we isolate telemetry for specific AI components and versions?

TELEMETRY DESIGN EXAMPLES

- Was there actually cancer in a scan?
- Did we identify the right soccer player?
- Did we correctly identify tanks?
- Was a Youtube recommendation good?
- Was the ranking of search results good?
- Was the weather prediction good?
- Was the translation correct?
- Did the self-driving car break at the right moment?


Skype for Business
How was the call quality?
Good
Audio Issues
Distorted speech
Electronic feedback
Background noise
Muffled speech
Echo
Video Issues
Frozen video
Pixelated video
Blurry image
Poor color
Dark video
blog post demo
Privacy Statement
Submit
Close

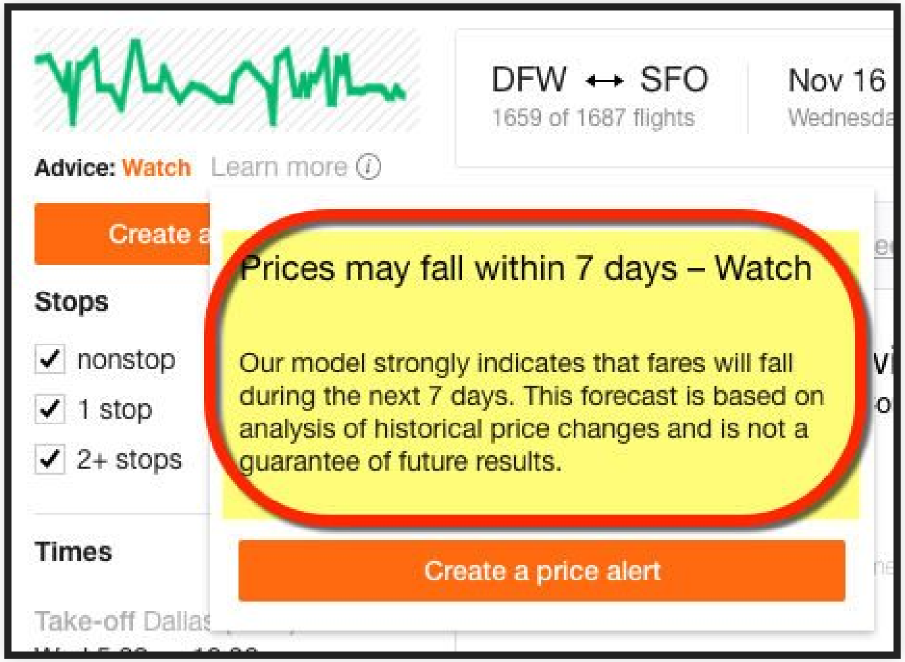

Matt Millman
Because I'm happy
People, groups & messages
Chats
Calls
Conta
Settings
Help and feedback
Report a problem
Sign out
RECENT CHATS
Besties
10/10/2018
EN
Elena Nilsson, Anna Davie...
7/27/2018
It was great talking to all of ...
Anna Davies
6/26/2018
coffee awaits!
Maarten Smenk
5/25/2018
Missed call
MS
Maarten Smenk, Anna Dav...
5/21/2018
Hi, happy Monday!

Speaker notes

Expect only sparse feedback and expect negative feedback over-proportionally

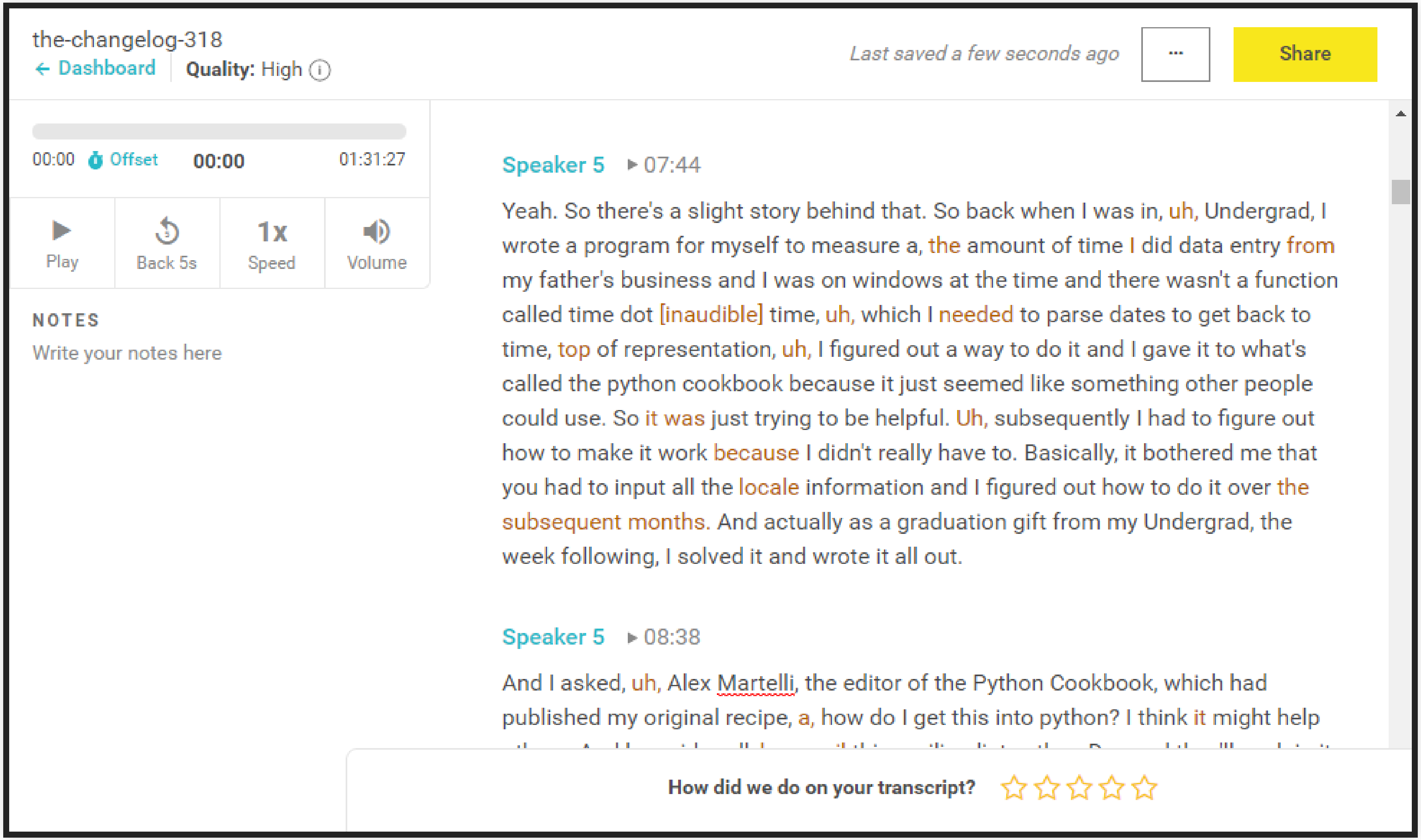
MANUALLY LABEL PRODUCTION SAMPLES

amazon
mechanical turk


DFW ↔ SFO
1659 of 1687 flights
Nov 16
Advice: Watch
Learn more
Stops
nonstop
1 stop
2+ stops
Times
Prices may fall within 7 days – Watch
Our model strongly indicates that fares will fall during the next 7 days. This forecast is based on analysis of historical price changes and is not a guarantee of future results.
Create a price alert

Speaker notes

Can just wait 7 days to see actual outcome for all predictions



the-changelog-318
← Dashboard
Quality: High
Last saved a few seconds ago
···
Share
00:00
Offset
00:00
01:31:27
Play
Back 5s
1x
Speed
Volume
NOTES
Write your notes here
Speaker 5
07:44
Yeah. So there's a slight story behind that. So back when I was in, uh, Undergrad, I wrote a program for myself to measure a, the amount of time I did data entry from my father's business and I was on windows at the time and there wasn't a function called time dot [inaudible] time, uh, which I needed to parse dates to get back to time, top of representation, uh, I figured out a way to do it and I gave it to what's called the python cookbook because it just seemed like something other people could use. So it was just trying to be helpful. Uh, subsequently I had to figure out how to make it work because I didn't really have to. Basically, it bothered me that you had to input all the locale information and I figured out how to do it over the subsequent months. And actually as a graduation gift from my Undergrad, the week following, I solved it and wrote it all out.
Speaker 5
08:38
And I asked, uh, Alex Martelli, the editor of the Python Cookbook, which had published my original recipe, a, how do I get this into python? I think it might help
How did we do on your transcript?

Speaker notes

Clever UI design allows users to edit transcripts. UI already highlights low-confidence words, can observe changes in editor (UI design encourages use of editor). In addition 5 star rating for telemetry.

MEASURING MODEL QUALITY WITH TELEMETRY

- Telemetry can provide insights for correctness
  - sometimes very accurate labels for real unseen data
  - sometimes only mistakes
  - sometimes indicates severity of mistakes
  - sometimes delayed
  - often just samples, may be hard to catch rare events
  - often just weak proxies for correctness
- Often sufficient to approximate precision/recall or other measures
- Mismatch to (static) evaluation set may indicate stale or unrepresentative test data
- Trend analysis can provide insights even for inaccurate proxy measures

MONITORING MODEL QUALITY IN PRODUCTION

- Watch for jumps after releases
    - roll back after negative jump
- Watch for slow degradation
    - Stale models, data drift, feedback loops, adversaries
- Debug common or important problems
    - Mistakes uniform across populations?
    - Challenging problems -> refine training, add regression tests

ENGINEERING CHALLENGES FOR TELEMETRY

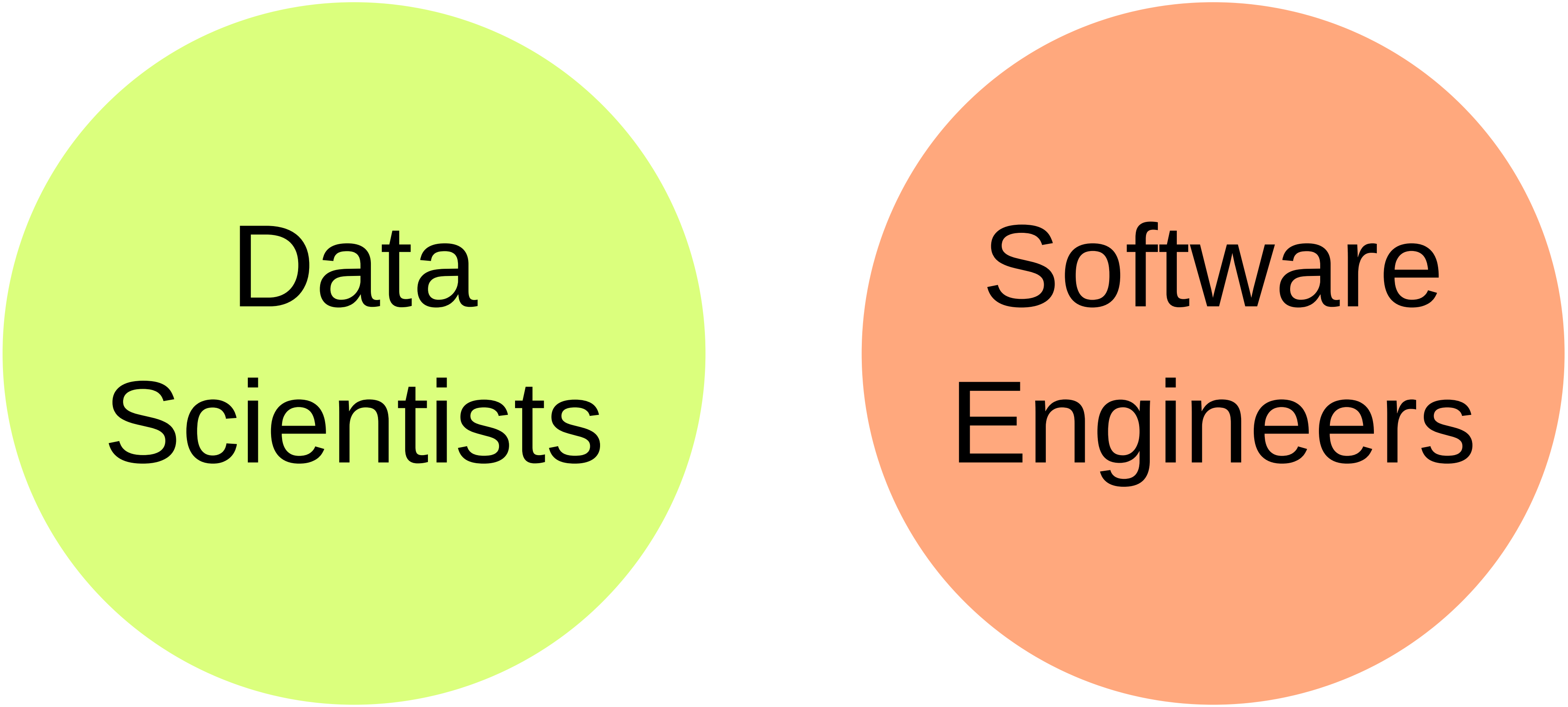
techradar

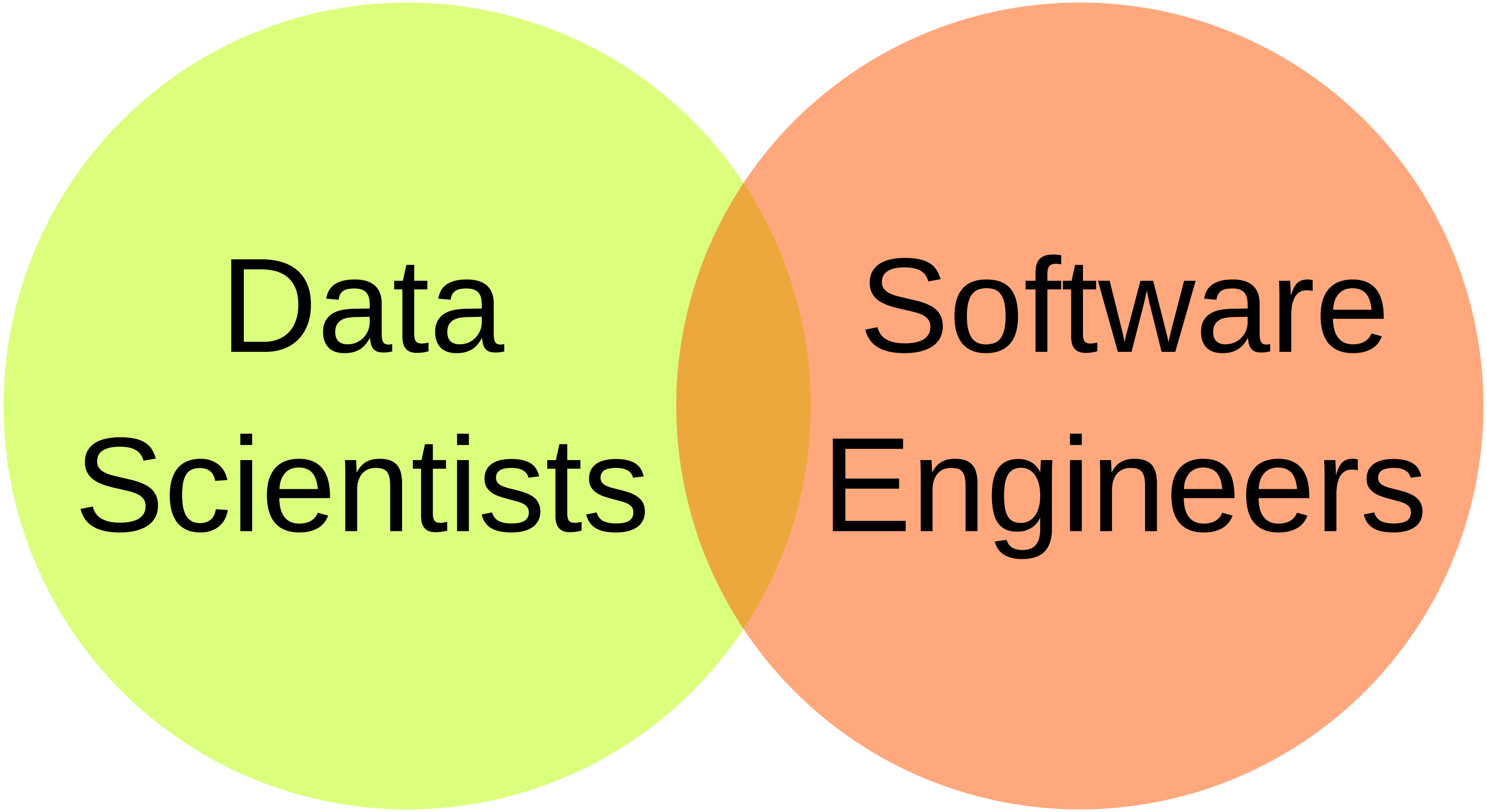
TRENDING Buying Guides Note 10 Best Laptops iOS 13 Best Phones

Amazon Alexa stores voice recordings for as long as it likes (and shares them too)

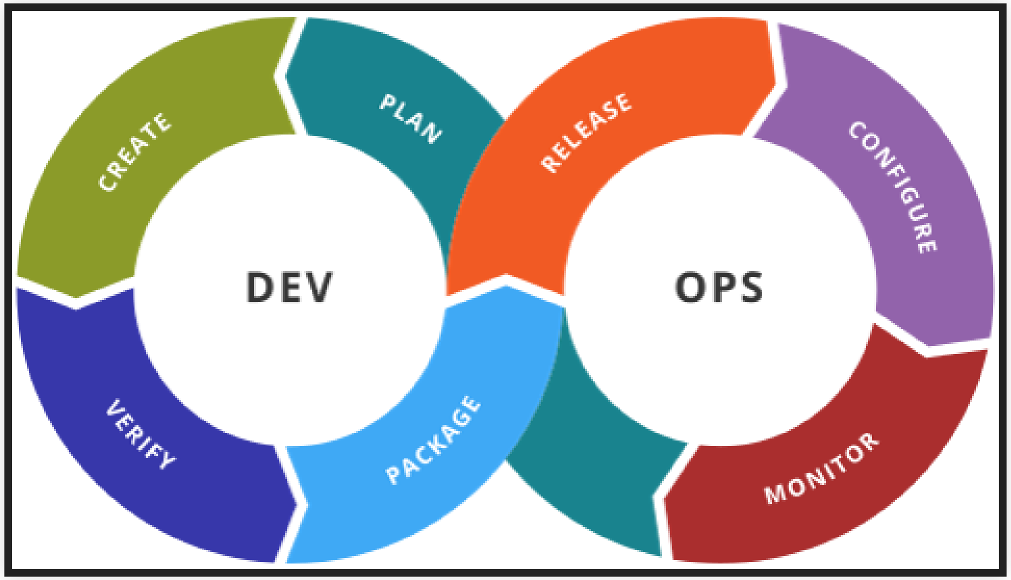
By Olivia Tambini 21 days ago Digital Home

A letter from Amazon reveals all


Data Scientists
Software Engineers


Data
Scientists
Software
Engineers

LET'S LEARN FROM DEVOPS



Distinct roles and expertise, but joint responsibilities, joint tooling

TOWARD BETTER ML-SYSTEMS ENGINEERING

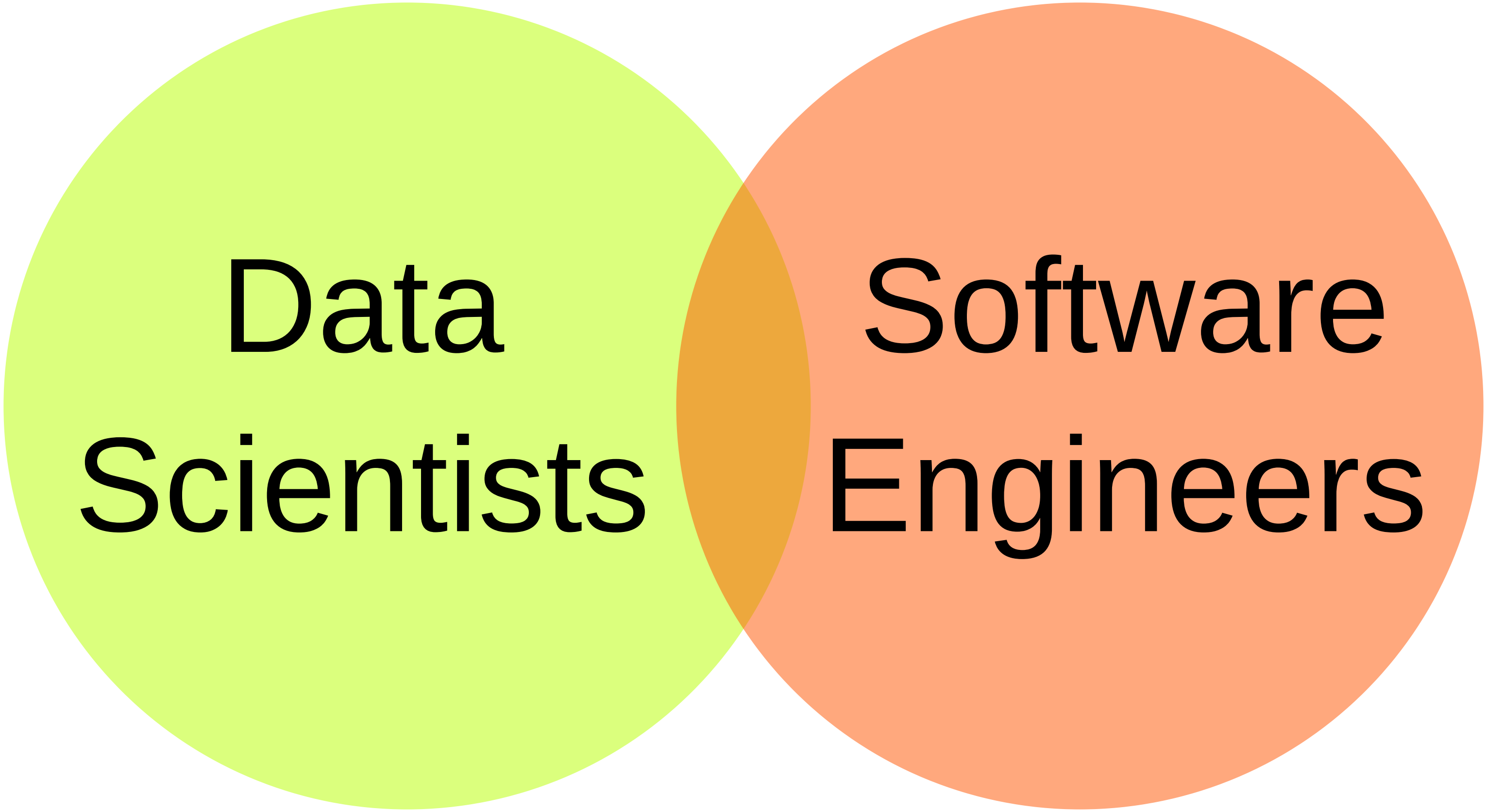
Interdisciplinary teams, split expertise, but joint responsibilities

Joint vocabulary and tools

Foster system thinking

Awareness of production quality concerns

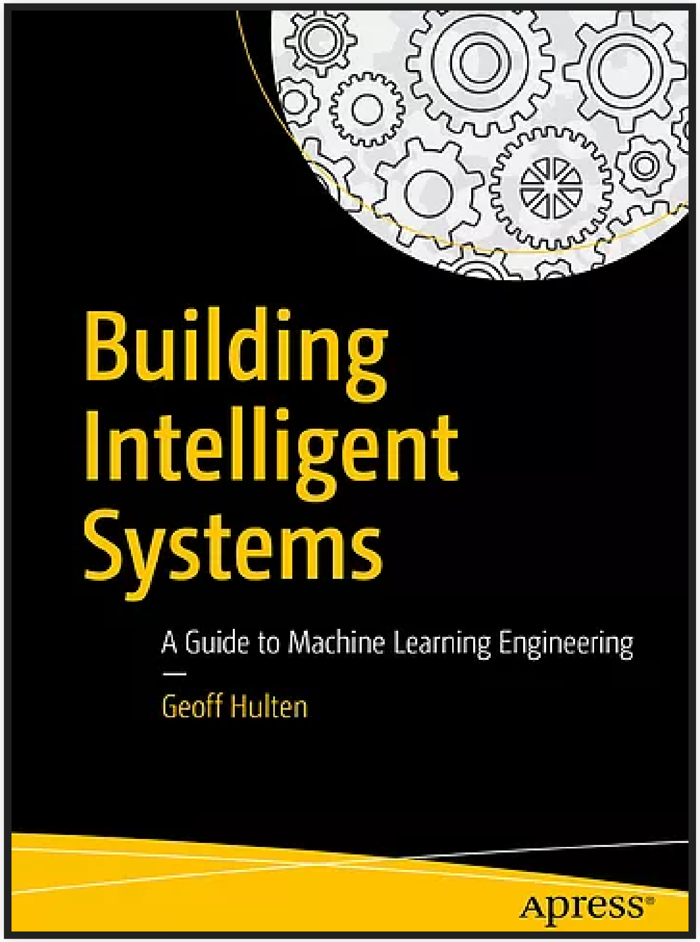
Perform risk + hazard analysis


Data
Scientists
Software
Engineers

READINGS

All lecture material: https://github.com/ckaestne/seai

Annotated bibliography: https://github.com/ckaestne/seaibib



SUMMARY: SOFTWARE ENGINEERING FOR ML-ENABLED SYSTEMS

- Building, operating, and maintaining systems with ML component
- Data scientists and software engineers have different expertise, both needed
- Quality assurance beyond model accuracy
  - Blackbox testing, test automation
  - Testing the entire ML pipeline
  - Consider whole system
  - Testing in production with telemetry
- Interdisciplinary teams, joint vocabulary, and awareness

kaestner@cs.cmu.edu -- @p0nk -- https://github.com/ckaestne/seai/