SOFTWARE ARCHITECTURE OF AI-ENABLED SYSTEMS

Guest Lecture by Christian Kaestner

Required reading:

- Vogelsang, Andreas, and Markus Borg. "Requirements Engineering for Machine Learning: Perspectives from Data Scientists." In Proc. of the 6th International Workshop on Artificial Intelligence for Requirements Engineering (AIRE), 2019.

MACHINE LEARNING IN SOFTWARE SYSTEMS

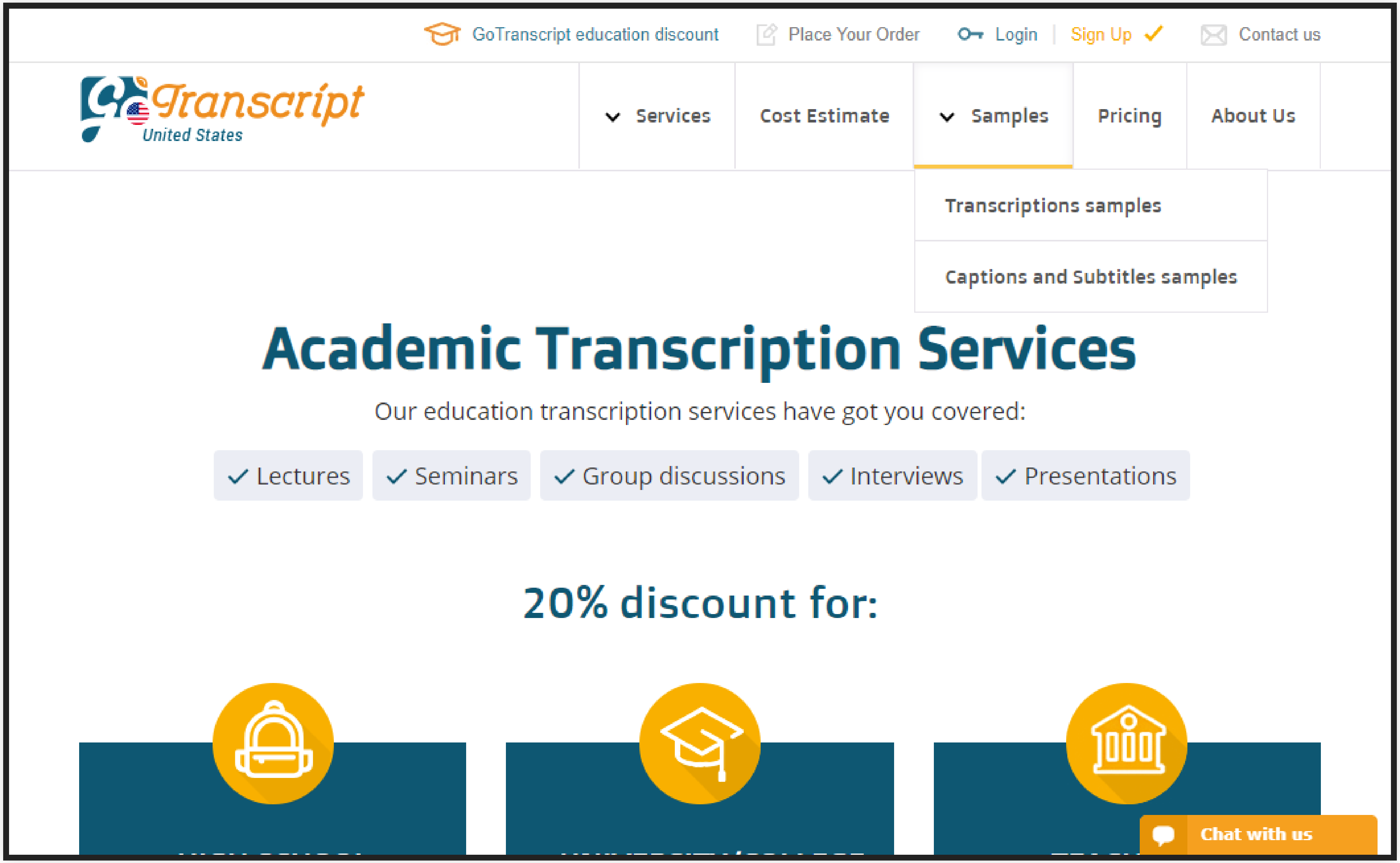
MACHINE LEARNING

Function making predictions for inputs

$$f(x_1, x_2, x_3) \rightarrow y$$

No specification, function learned by generalizing from example data (inductive reasoning)

RUNNING EXAMPLE: TRANSCRIPTION SERVICE



THE STARTUP IDEA

PhD research on domain-specific speech recognition, that can detect technical jargon

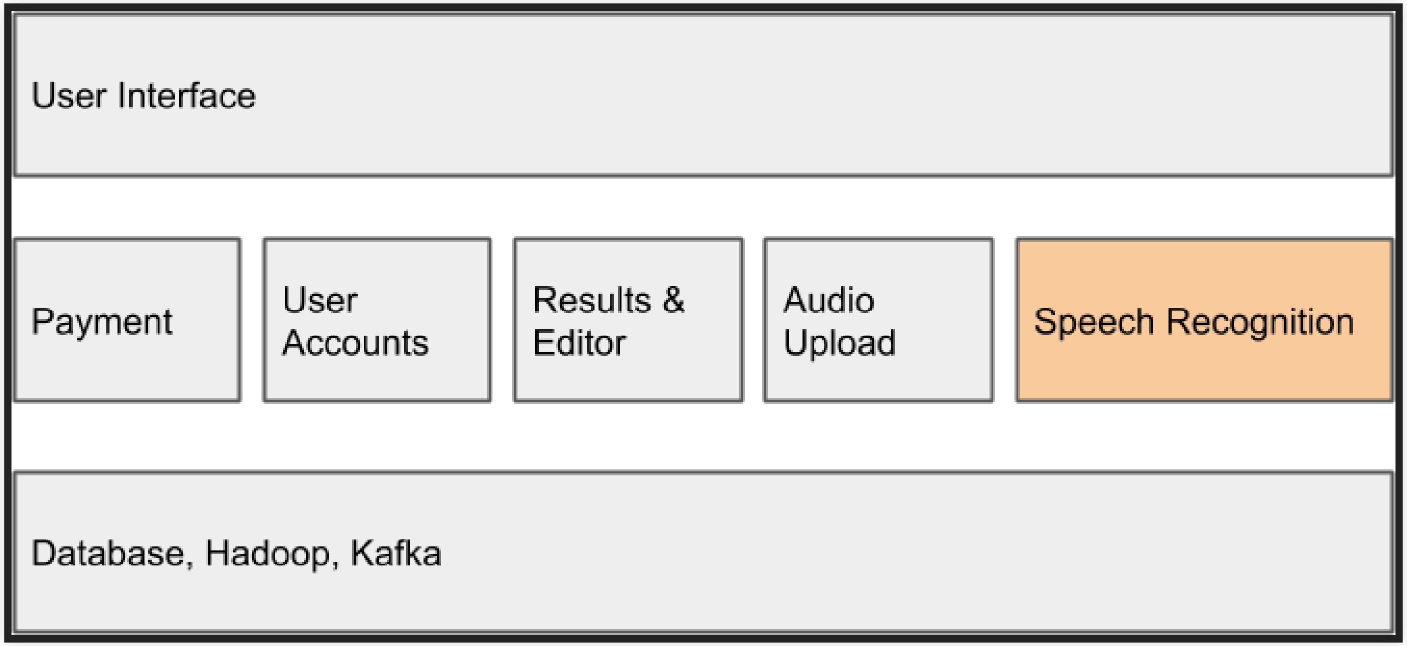
DNN trained on public PBS interviews + transfer learning on smaller manually annotated domain-specific corpus

Research has shown amazing accuracy for talks in medicine, poverty and inequality research, and talks at Ruby programming conferences; published at top conferences

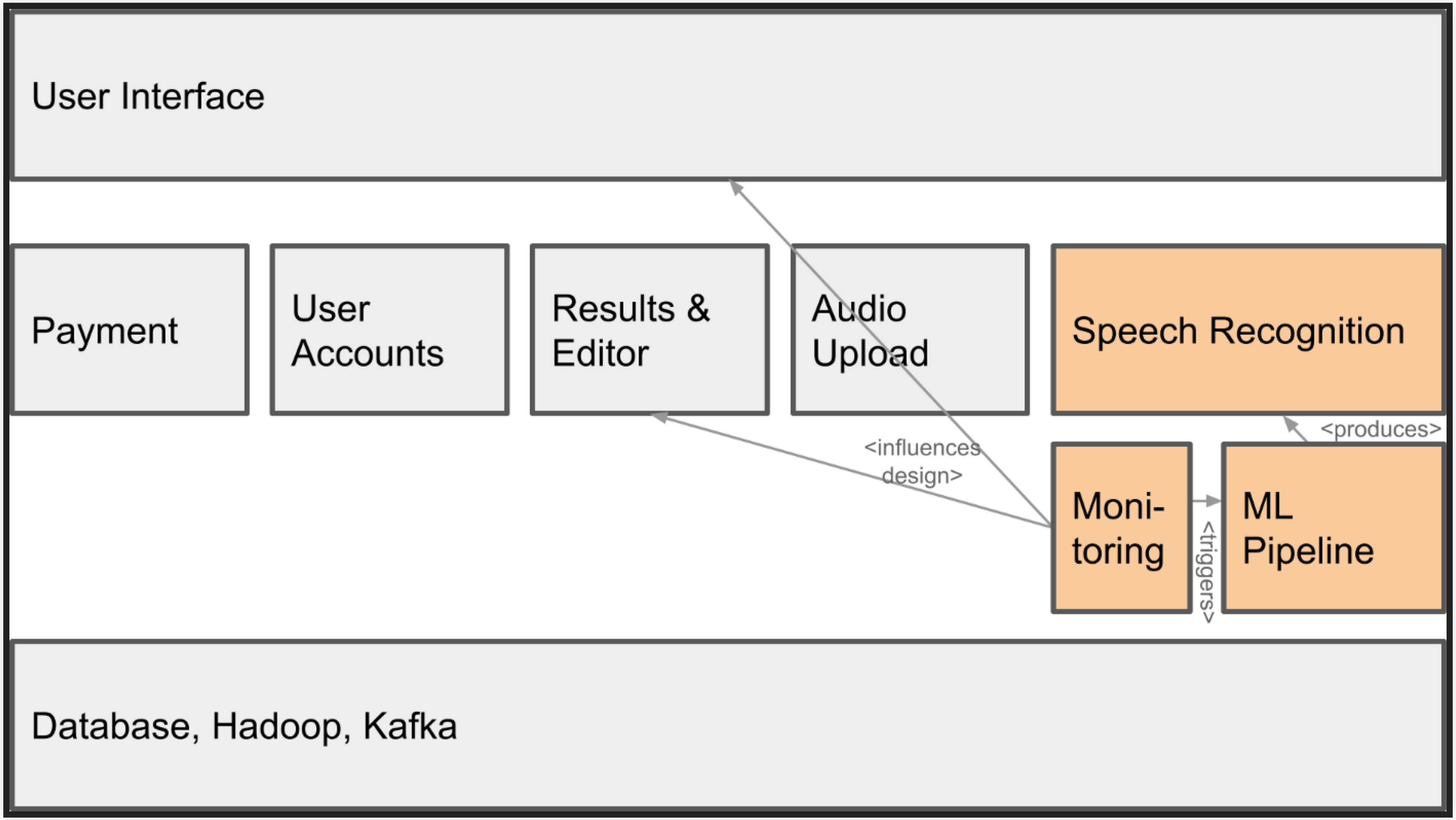
Idea: Let's commercialize the software and sell to academics and conference organizers

WHAT QUALITIES ARE IMPORTANT FOR A GOOD COMMERCIAL TRANSCRIPTION PRODUCT?

ML IN A PRODUCTION SYSTEM



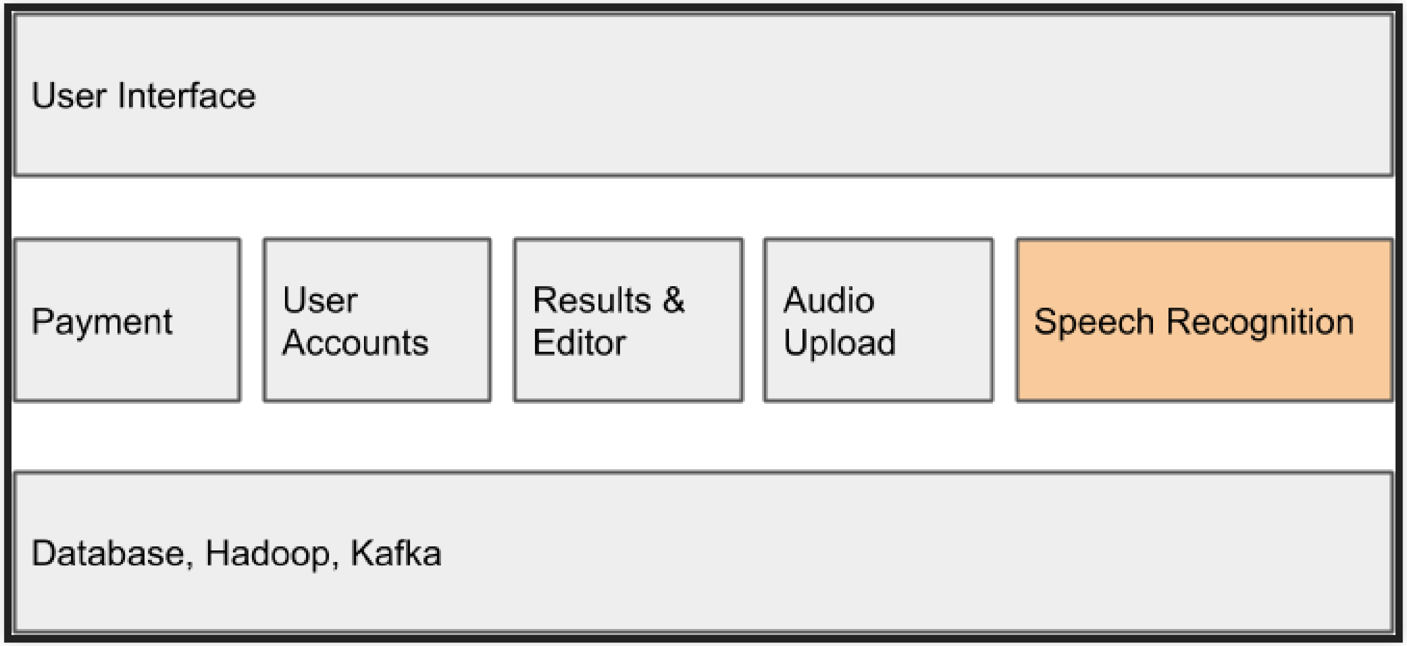
ML IN A PRODUCTION SYSTEM



ACCURACY, CORRECTNESS, AND OTHER QUALITIES

TRADITIONAL ML FOCUS: MODEL ACCURACY

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



TRADITIONAL SE FOCUS: FUNCTIONAL CORRECTNESS

Given a specification, do outputs match inputs?

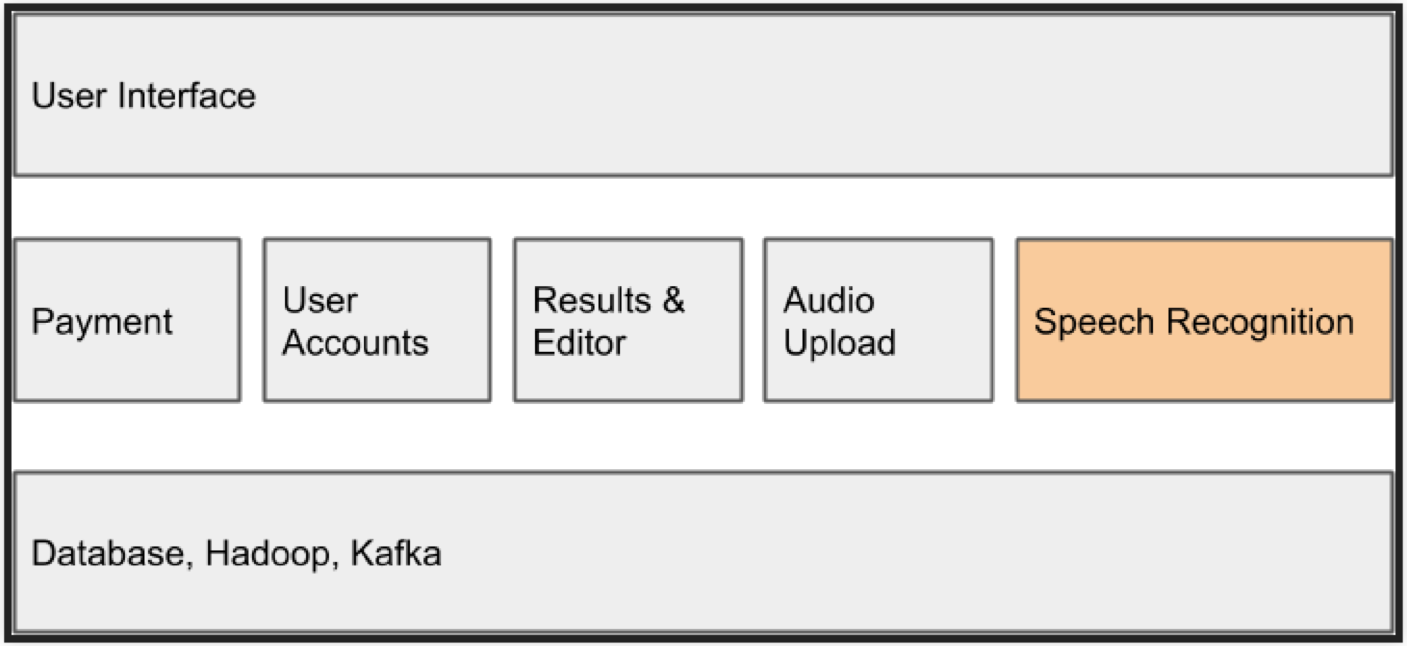
```
/**
 * compute deductions based on provided adjusted
 * gross income and expenses in customer data.
 *
 * see tax code 26 U.S. Code A.1.B, PART VI
 */
float computeDeductions(float agi, Expenses expenses);
```

Each mismatch is considered a bug, should to be fixed*.

(*=not every bug is economical to fix, may accept some known bugs)

NO SPECIFICATION!

We use ML precisely because we do not have a specification (too complex, rules unknown)



We are usually okay with some wrong predictions

> *All models are approximations. Assumptions, whether implied or clearly stated, are never exactly true.* ***All models are wrong, but some models are useful****. So the question you need to ask is not "Is the model true?" (it never is) but "Is the model good enough for this particular application?"*
> *-- George Box*

See also https://en.wikipedia.org/wiki/All_models_are_wrong

NON-ML EXAMPLE: NEWTON'S LAWS OF MOTION

> *2nd law: "the rate of change of momentum of a body over time is directly proportional to the force applied, and occurs in the same direction as the applied force"* $\boldsymbol{F} = \frac{d\boldsymbol{p}}{dt}$

"Newton's laws were verified by experiment and observation for over 200 years, and they are excellent approximations at the scales and speeds of everyday life."

Do not generalize for very small scales, very high speeds, or in very strong gravitational fields. Do not explain semiconductor, GPS errors, superconductivity, ... Those require general relativity and quantum field theory.

Further readings: https://en.wikipedia.org/wiki/Newton%27s_laws_of_motion

LIMITATIONS OF OFFLINE MODEL EVALUATION

- Training and test data drawn from the same population
    - **i.i.d.: independent and identically distributed**
- Is the population representative of production data?
- If not or only partially or not anymore: Does the model generalize beyond training data?

TESTING IN PRODUCTION

Tweet

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)

DEPLOYING ML MODELS

ACCESSIBILITY: LIVE SUBTITLES

WHERE TO DEPLOY THE TRANSCRIPTION MODEL?

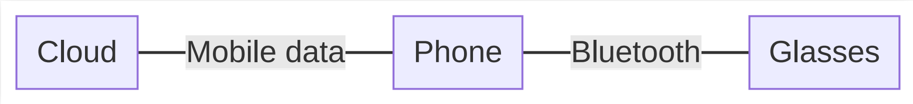
Cloud —Mobile data— Phone —Bluetooth— Glasses

WHERE TO DEPLOY THE TRANSCRIPTION MODEL?

Cloud —Mobile data— Phone —Bluetooth— Glasses

Which qualities and tradeoffs to consider?

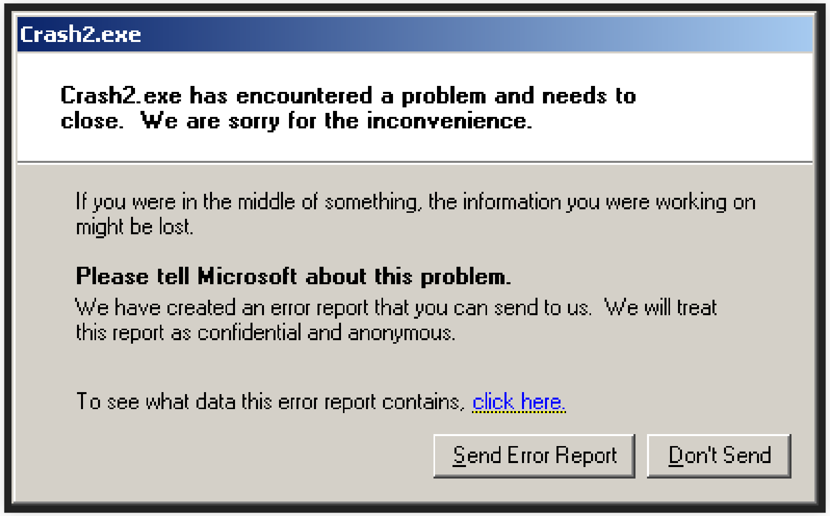
WHERE TO DEPLOY THE TRANSCRIPTION MODEL?



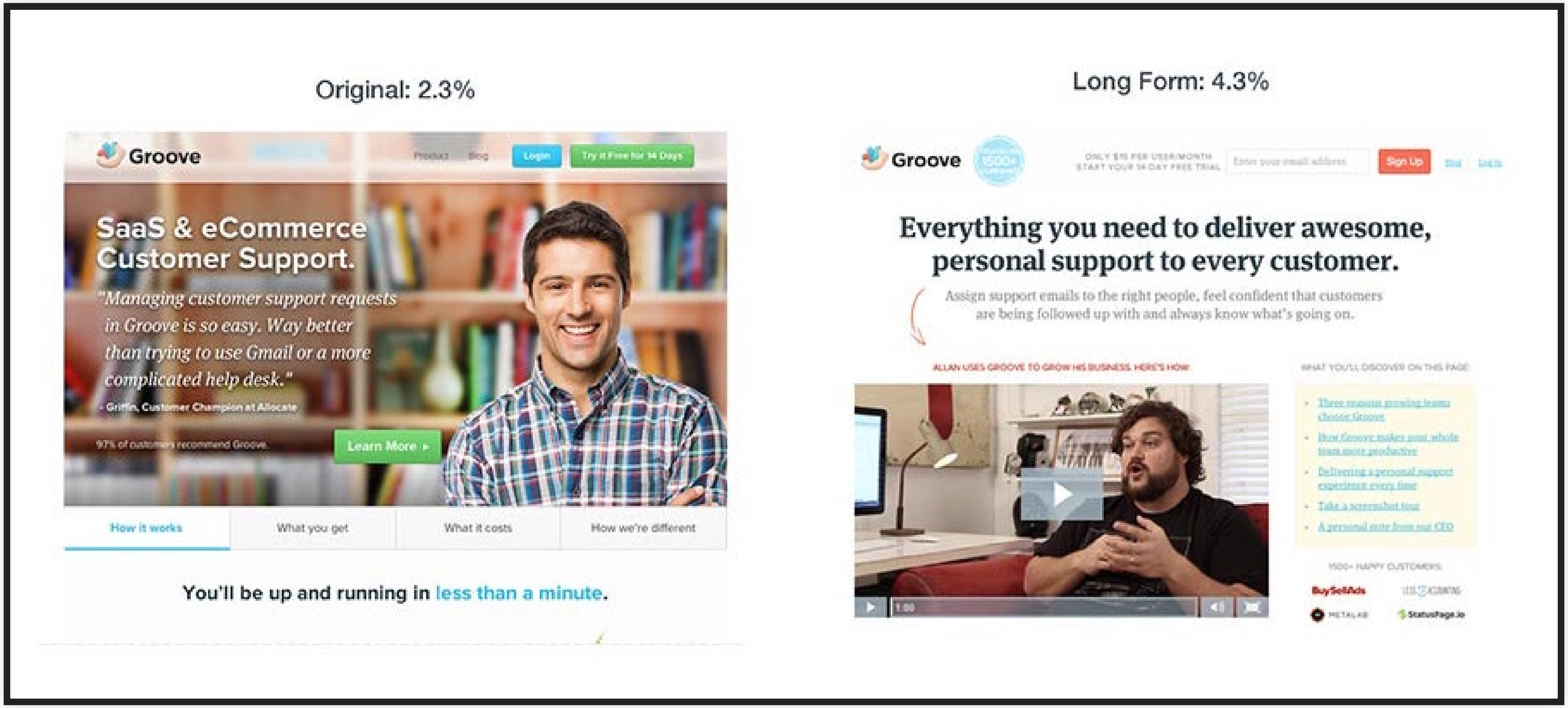
- Amount of data, bandwidth, bandwidth cost
- Latency
- Energy/battery cost
- Available memory, CPU capacity
- Ability to debug
- Offline functioning
- Privacy, security
- Accuracy
- Frequency of model updates

TELEMETRY DESIGN

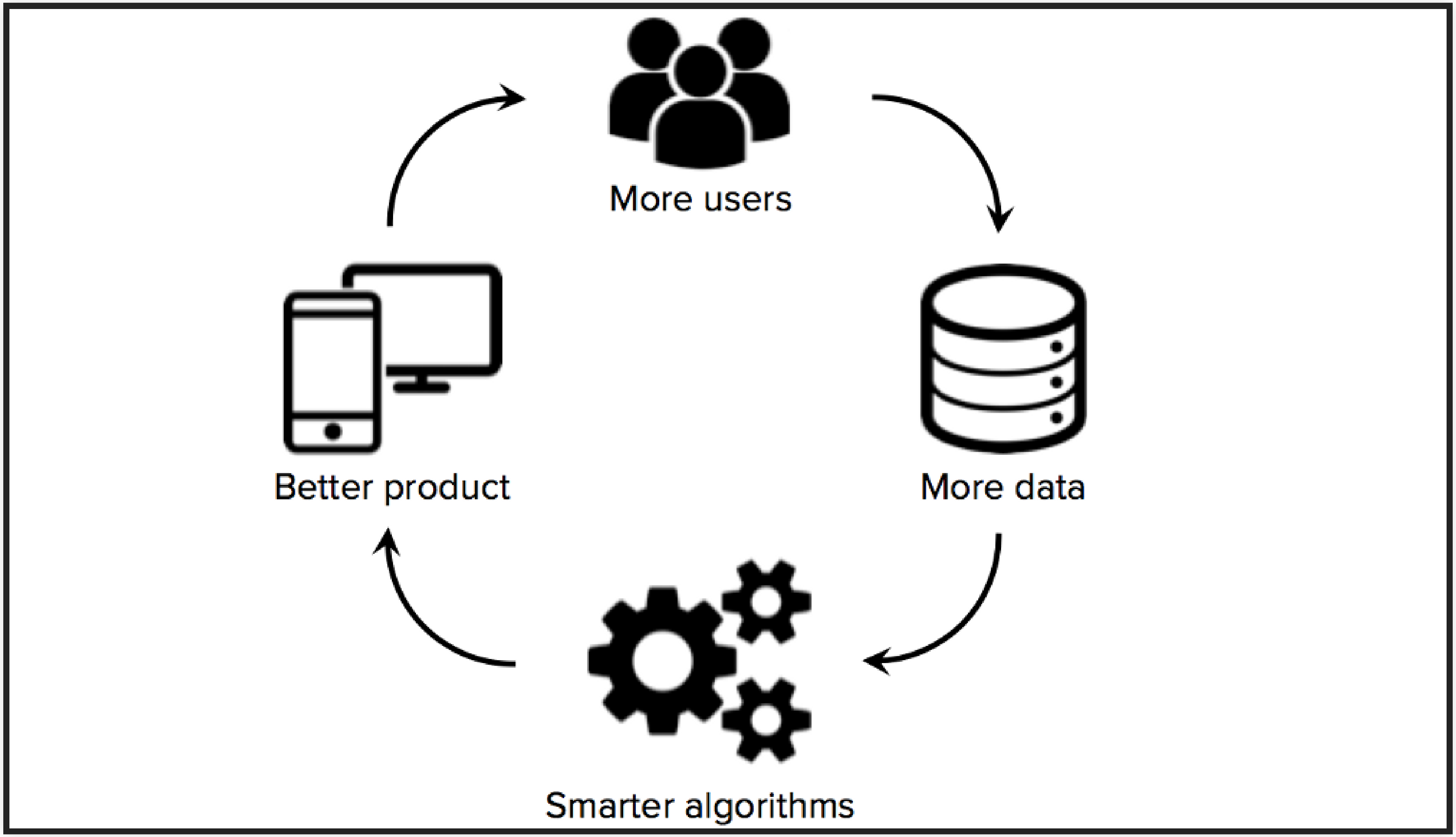
GOALS 1: EVALUATE MODEL AND SYSTEM QUALITY IN PRODUCTION



GOAL 2: EXPERIMENTING IN PRODUCTION



GOAL 3: GATHER MORE TRAINING DATA



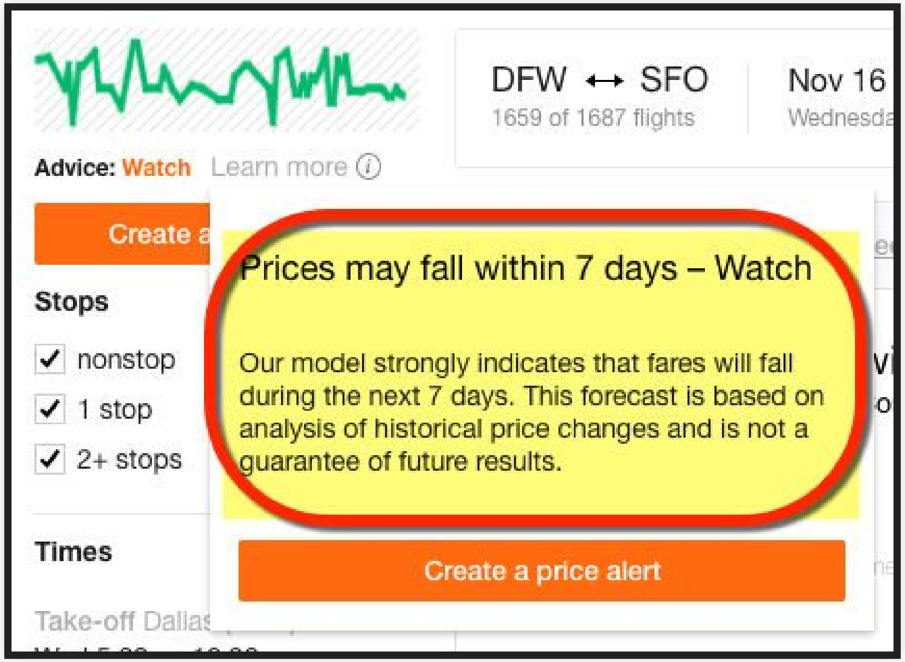
graphic by CBInsights

DISCUSSION: WAS THE TRANSCRIPTION ANY GOOD?

- Gather feedback without being intrusive (i.e., labeling outcomes), without harming user experience
- What data can we collect to evaluate our transcription service?
    - Evaluate business goals
    - Evaluate system quality
    - Evaluate model quality

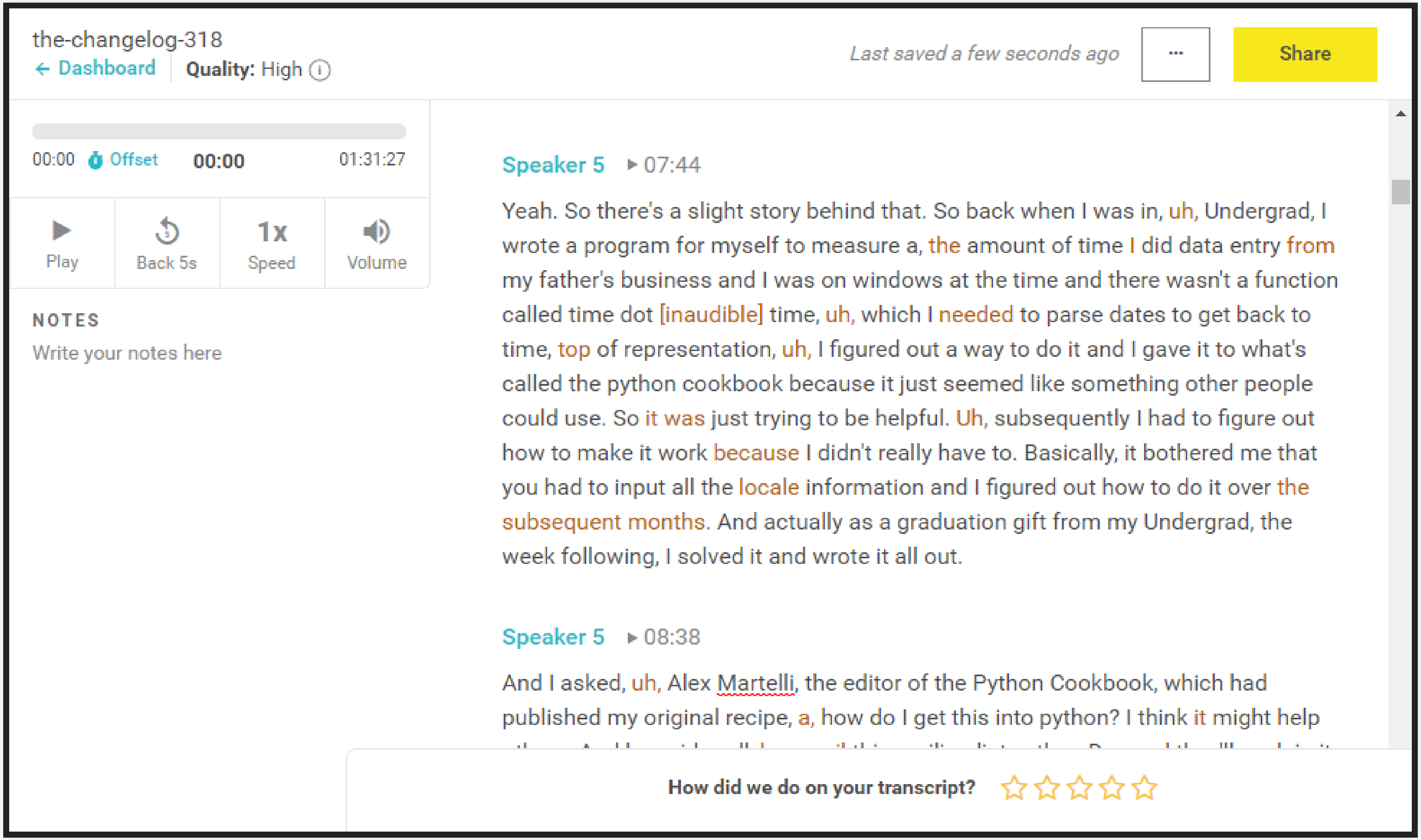
TYPICAL TELEMETRY STRATEGIES

- Wait and see
- Ask users
- Manual/crowd-source labeling, shadow execution
- Allow users to complain
- Observe user reaction


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

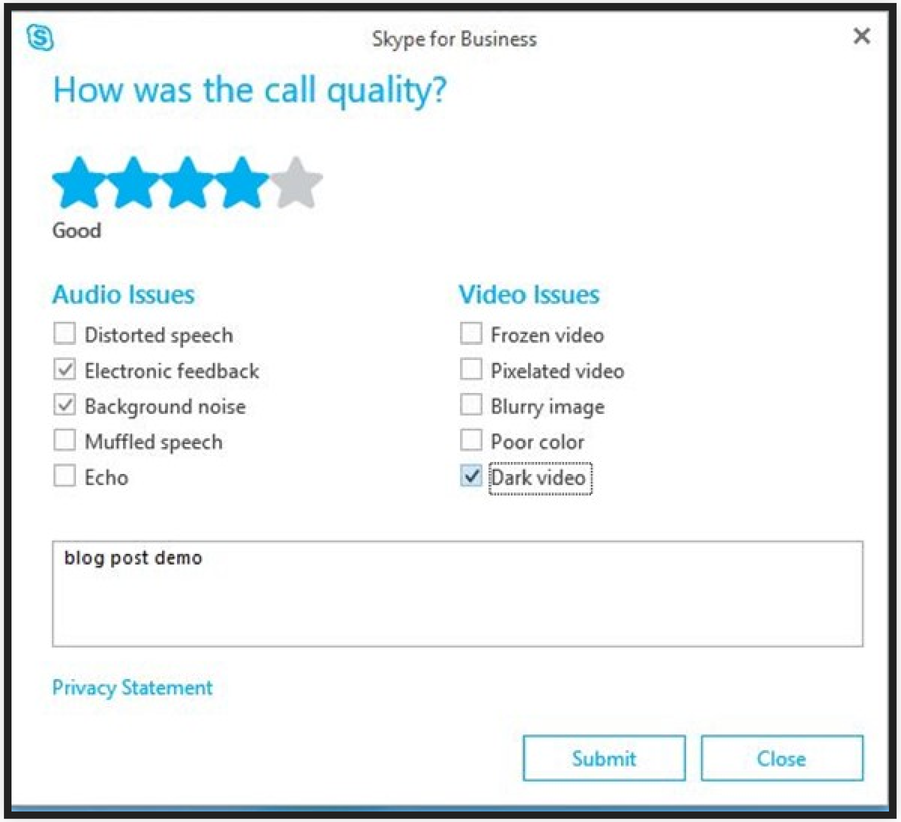
Speaker notes

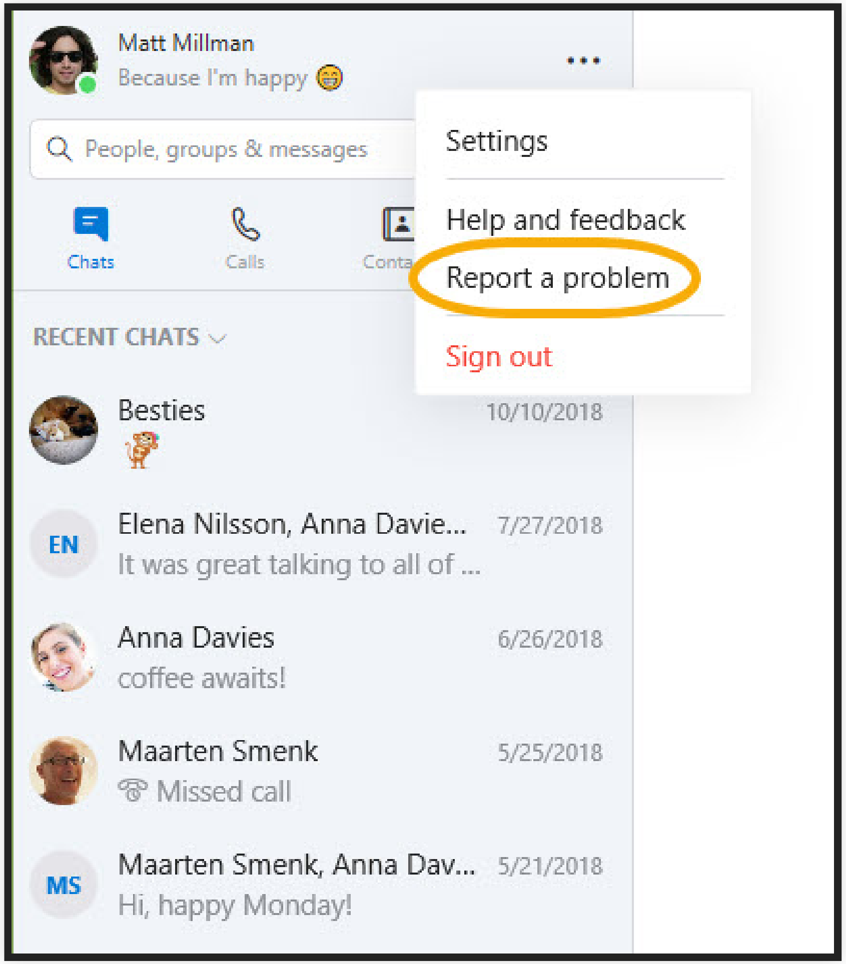
Can just wait 7 days to see actual outcome for all predictions


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


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
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!

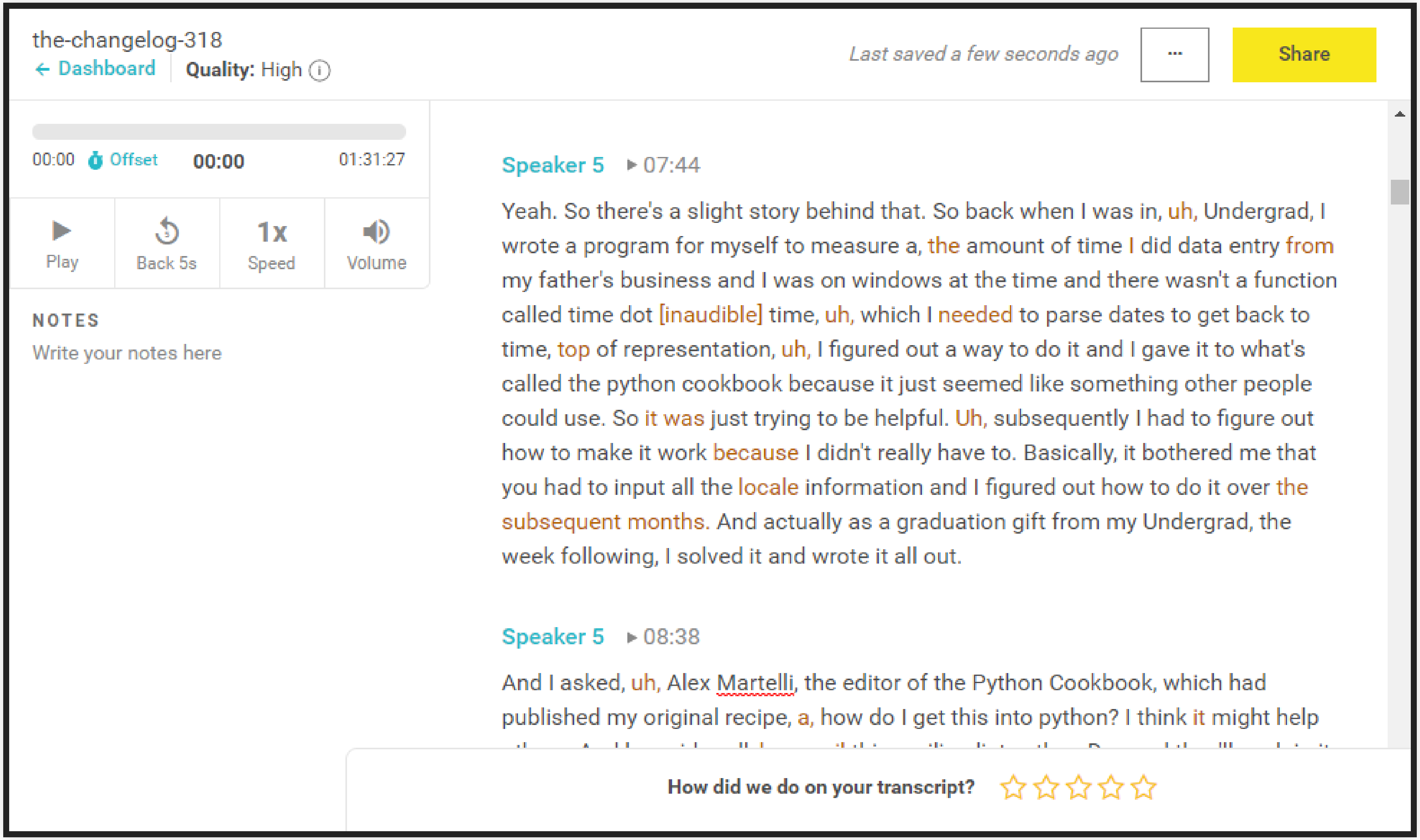
Speaker notes

Expect only sparse feedback and expect negative feedback over-proportionally

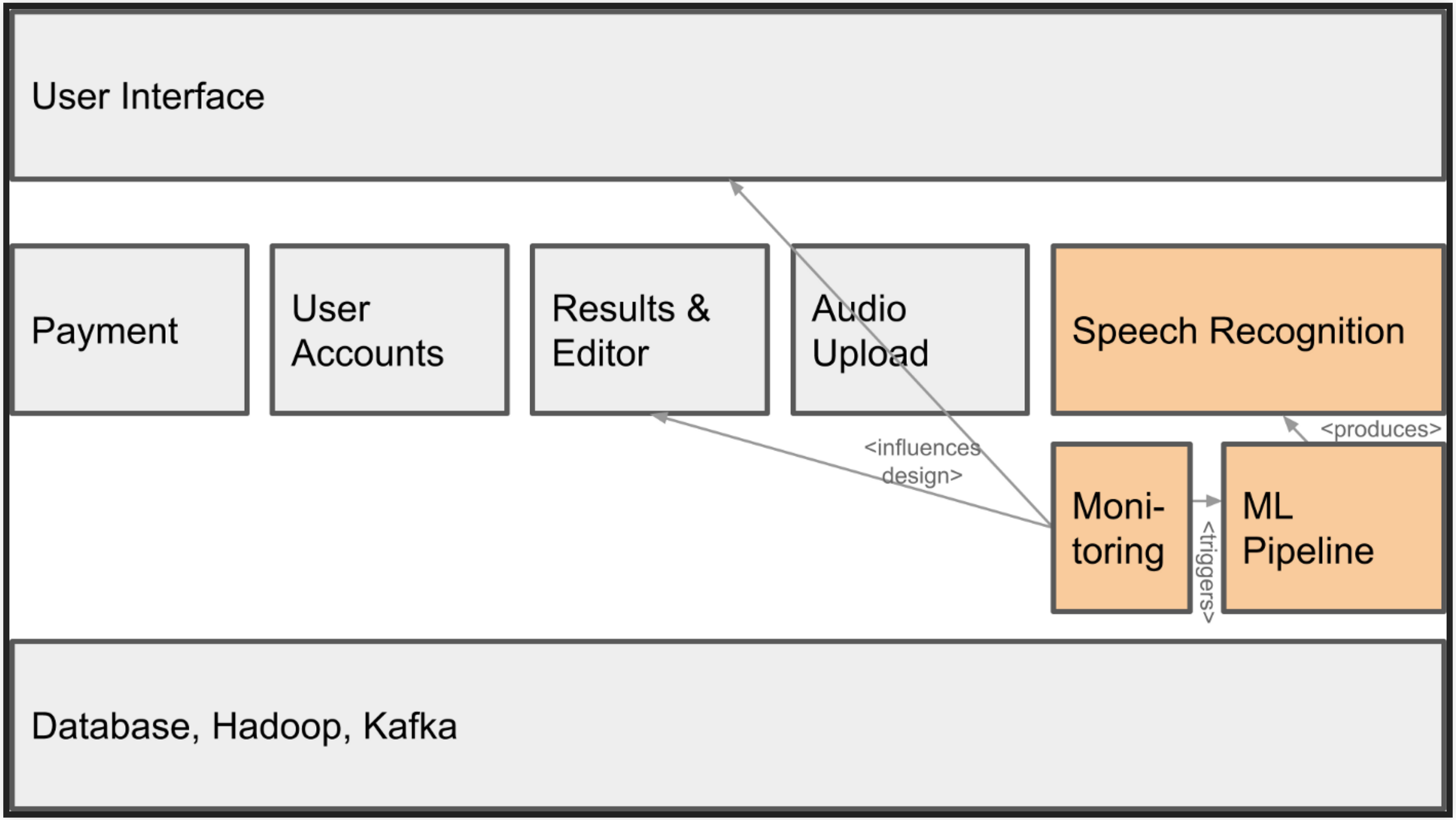
MANUALLY LABEL PRODUCTION SAMPLES

Similar to labeling learning and testing data, have human annotators

CLEVER UI DESIGN: TRANSCRIPTION SERVICE



ML IN A PRODUCTION SYSTEM

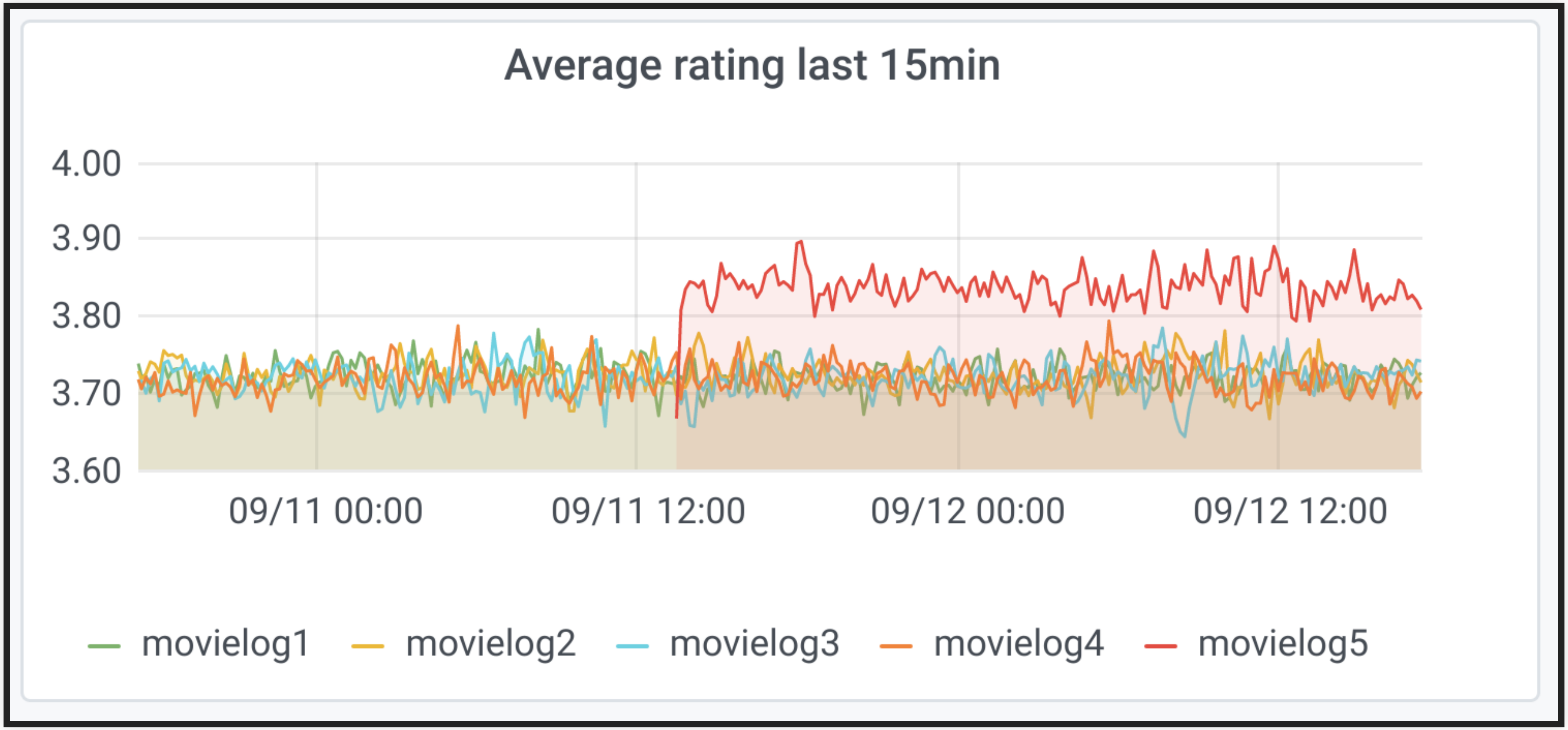


DISCUSSION 2: GOOGLE TAGGING UPLOADED PHOTOS WITH FRIENDS' NAMES

- Gather feedback without being intrusive (i.e., labeling outcomes), without harming user experience
- What data can we collect to evaluate our transcription service?
    - Evaluate business goals
    - Evaluate system quality
    - Evaluate model quality

MONITORING MODEL QUALITY IN PRODUCTION

- Monitor model quality together with other quality attributes (e.g., uptime, response time, load)
- Set up automatic alerts when model quality drops
- 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
    - Monitor characteristics of requests
    - Mistakes uniform across populations?
    - Challenging problems -> refine training, add regression tests



Average rating last 15min
4.00
3.90
3.80
3.70
3.60
09/11 00:00
09/11 12:00
09/12 00:00
09/12 12:00
movielog1
movielog2
movielog3
movielog4
movielog5

DETECTING DRIFT

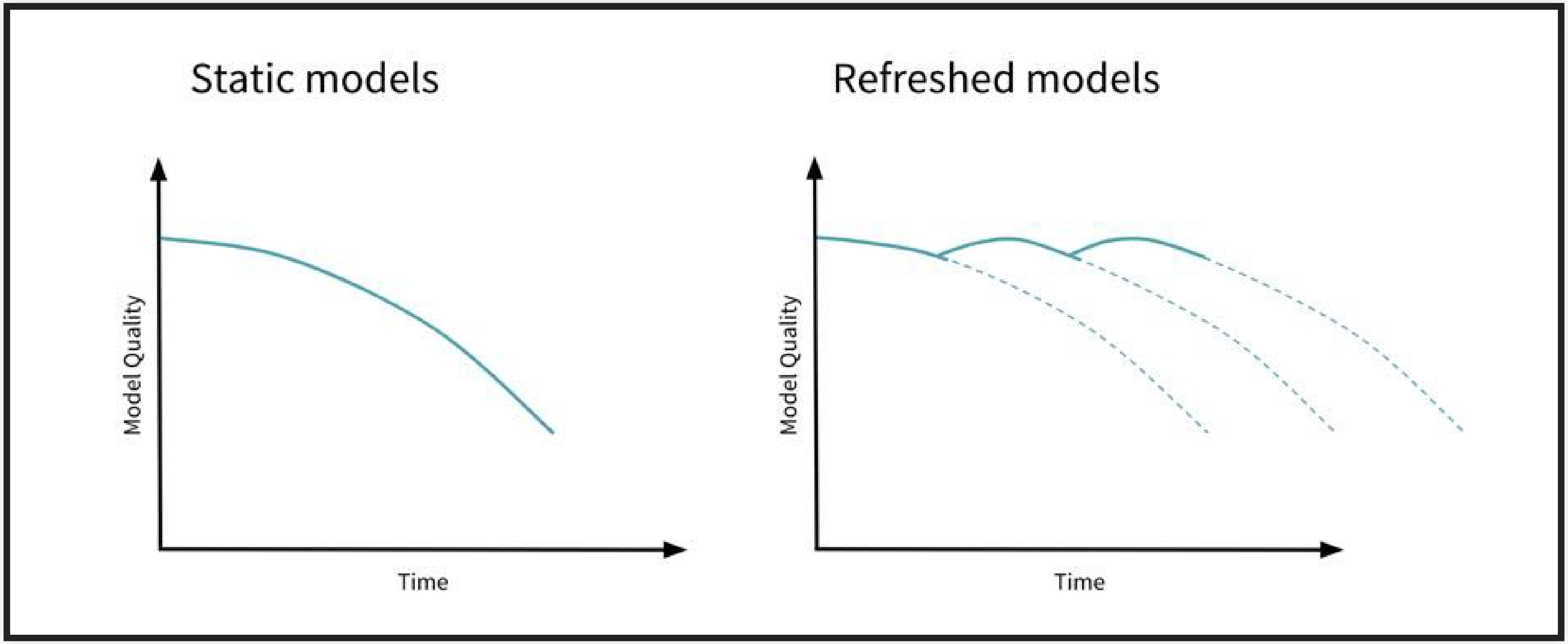
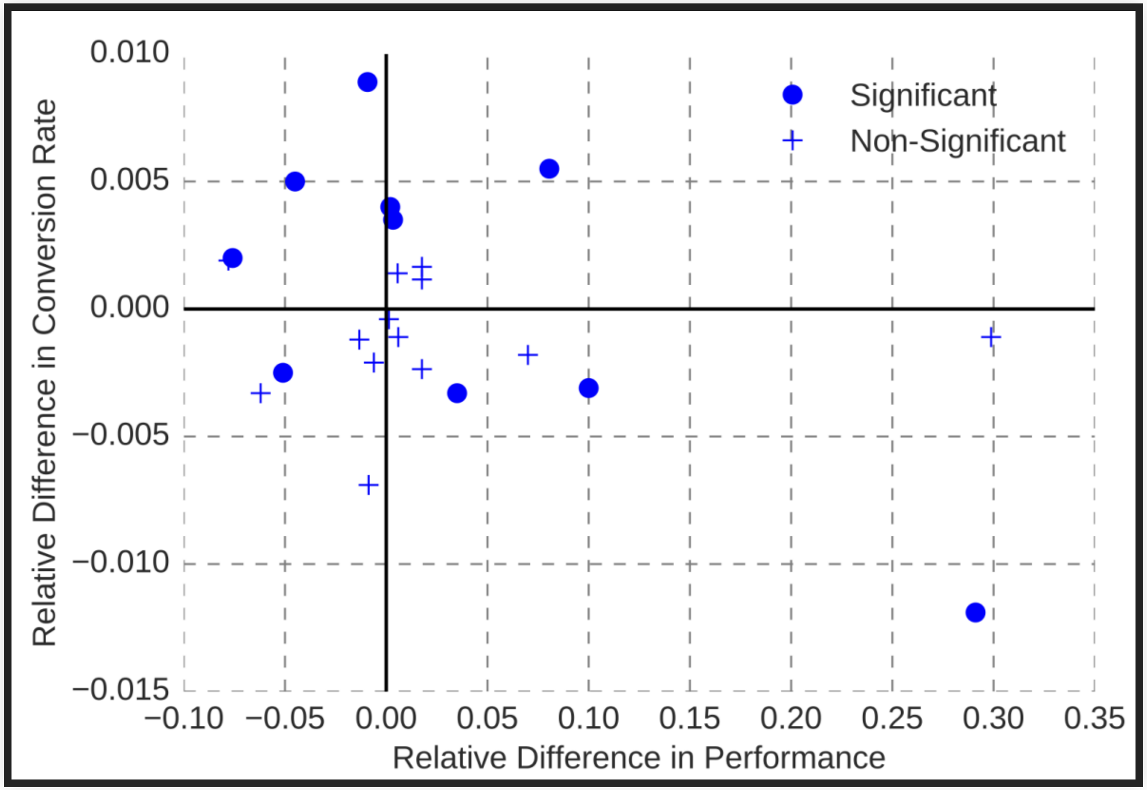


Image source: Joel Thomas and Clemens Mewald. Productionizing Machine Learning: From Deployment to Drift Detection. Databricks Blog, 2019

MODEL QUALITY VS SYSTEM QUALITY



Possible causes?

Bernardi et al. "150 successful machine learning models: 6 lessons learned at Booking.com." In Proc KDD, 2019.

Speaker notes

hypothesized

- model value saturated, little more value to be expected
- segment saturation: only very few users benefit from further improvement
- overoptimization on proxy metrics not real target metrics
- uncanny valley effect from "creepy AIs"

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

ENGINEERING CHALLENGES FOR TELEMETRY

- Data volume and operating cost
    - e.g., record "all AR live translations"?
    - reduce data through sampling
    - reduce data through summarization (e.g., extracted features rather than raw data; extraction client vs server side)
- Adaptive targeting
- Biased sampling
- Rare events
- Privacy
- Offline deployments?

EXERCISE: DESIGN TELEMETRY IN PRODUCTION

Discuss: Quality measure, telemetry, operationalization, cost, privacy, rare events

Google: Tagging uploaded photos with friends' names

SUMMARY

- Machine learning is a component of a larger system
- It brings new concerns, qualities, and design options
- Telemetry design is key for ML systems in production
- Many qualities and tradeoffs to consider

FURTHER POINTERS

- Full lecture (with videos, readings, assignments): https://ckaestne.github.io/seai/
- Annotated bibliography: https://github.com/ckaestne/seaibib
- Hulten, Geoff. Building Intelligent Systems: A Guide to Machine Learning Engineering. Apress. 2018
- Yokoyama, Haruki. "Machine learning system architectural pattern for improving operational stability." In 2019 IEEE International Conference on Software Architecture Companion (ICSA-C), pp. 267-274. IEEE, 2019.
- Hazelwood, Kim, Sarah Bird, David Brooks, Soumith Chintala, Utku Diril, Dmytro Dzhulgakov, Mohamed Fawzy et al. "Applied machine learning at facebook: A datacenter infrastructure perspective." In 2018 IEEE International Symposium on High Performance Computer Architecture (HPCA), pp. 620-629. IEEE, 2018.