MOVING ML PROJECTS INTO PRODUCTION WITH INTERDISCIPL. TEAMS

Christian Kästner

Carnegie Mellon University

https://github.com/ckaestne/seai

CHRISTIAN KÄSTNER

Associate Professor

Director of SE PhD Prog.

@ Carnegie Mellon University

**

Background and interests:

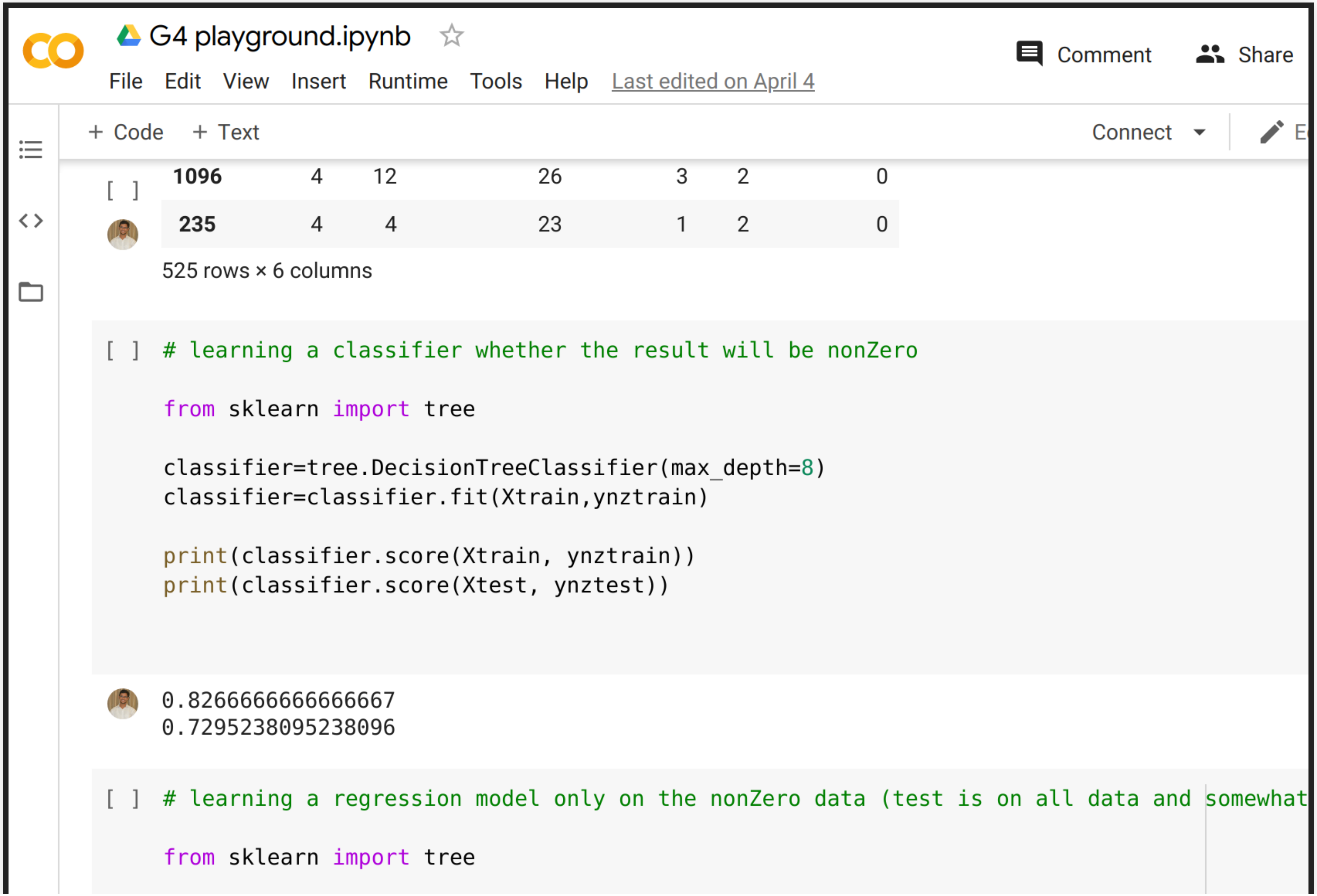
- Software Engineering
- Highly-Configurable Systems & Configuration Engineering
- Software Engineering for ML-Enabled Systems

BUILDING PRODUCTION SYSTEMS WITH MACHINE LEARNING

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

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

BEYOND BUILDING MODELS

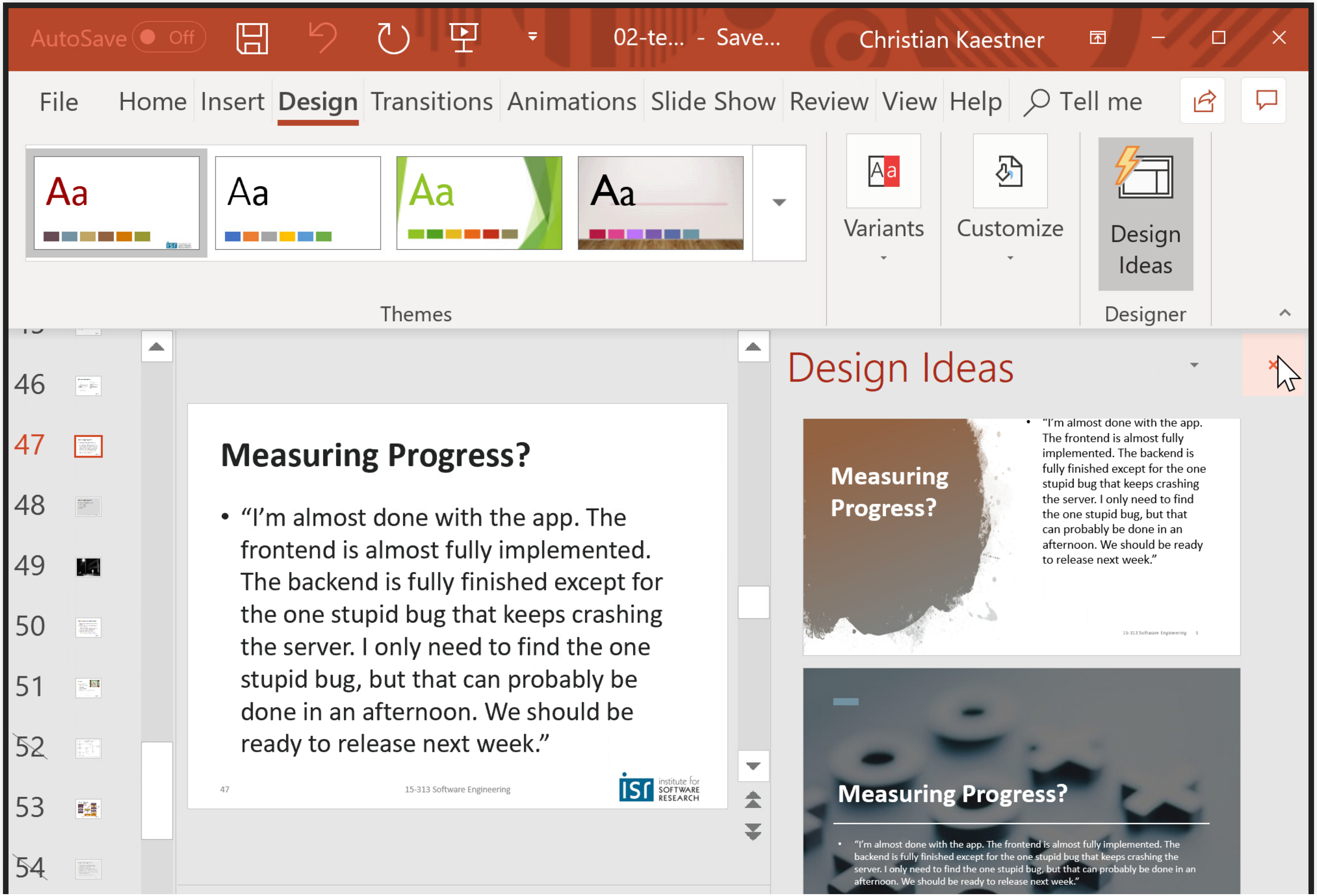


```
predictor=tree.DecisionTreeRegressor(max_depth=8)
predictor=predictor.fit(XnzTrain,YnzTrain)


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

```
0.9376379365613154
-2.437397740412892
```

PRODUCTION ML SYSTEMS



55

56

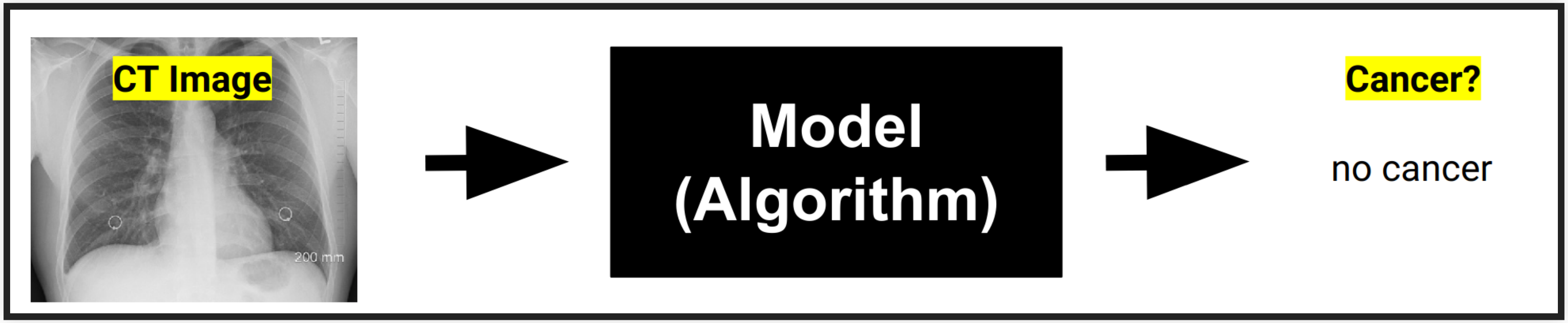
Tap to add notes

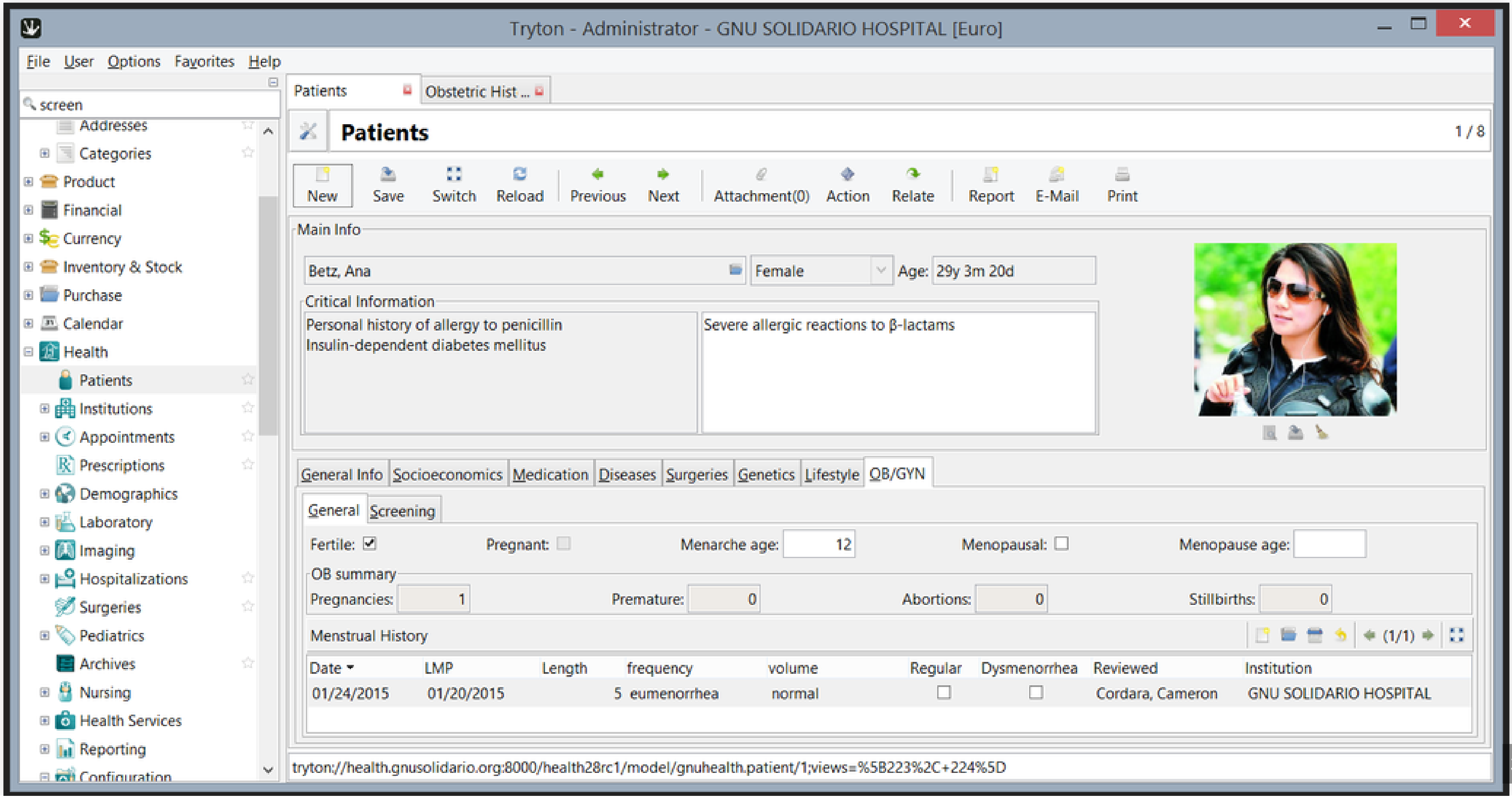
Slide 47 of 74

Notes

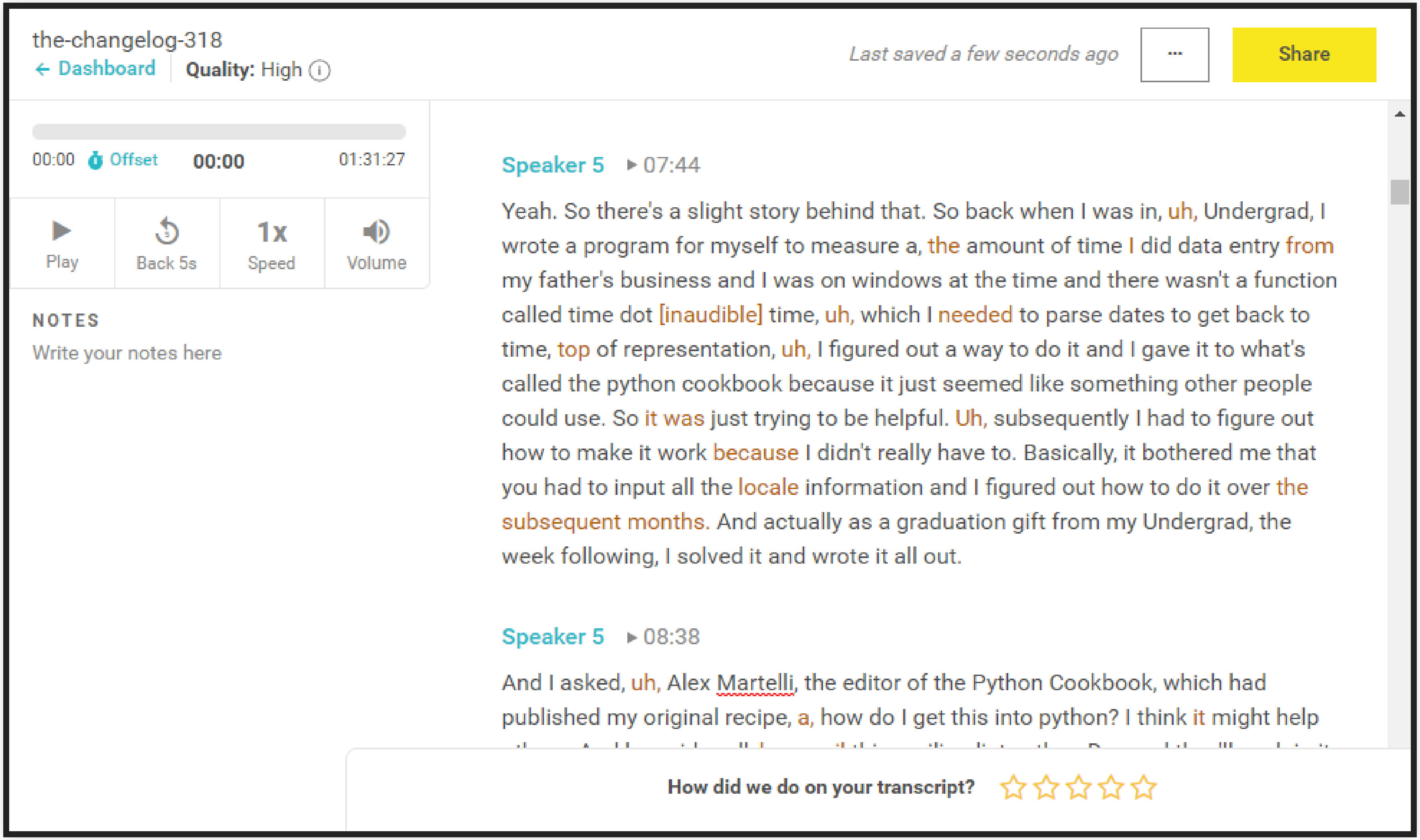
29%

PRODUCTION ML SYSTEMS

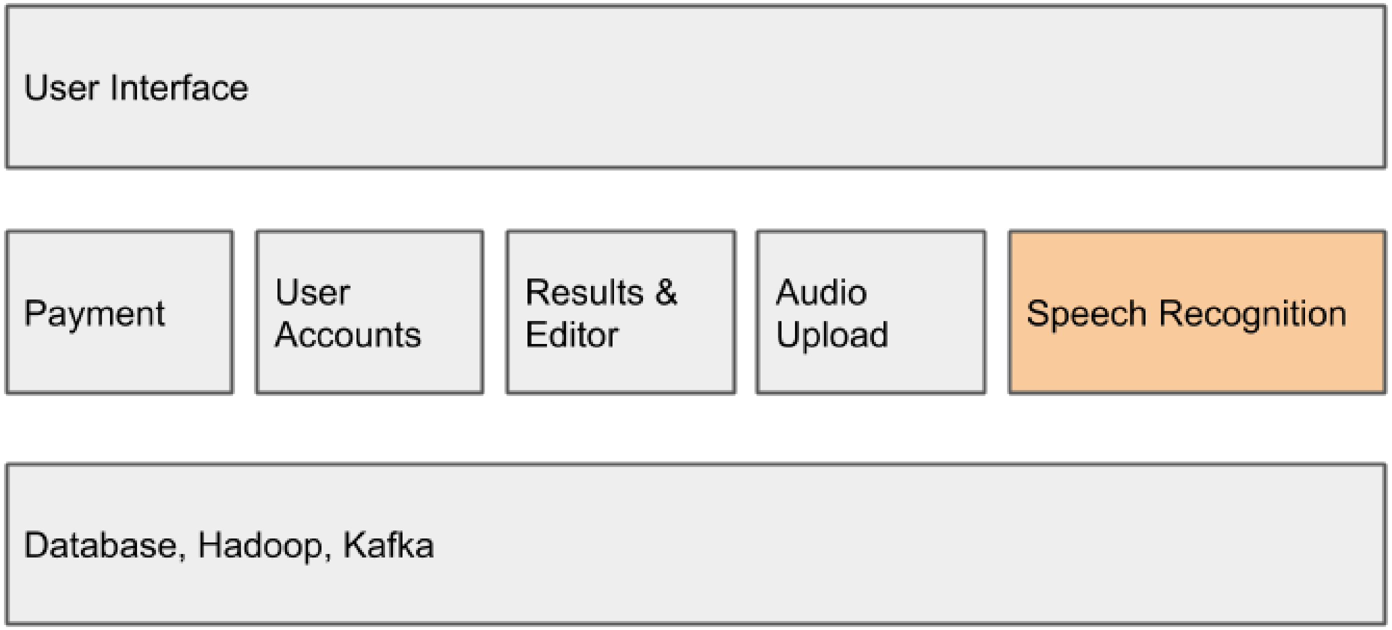

CT Image
200 mm
Model
(Algorithm)
Cancer?
no cancer

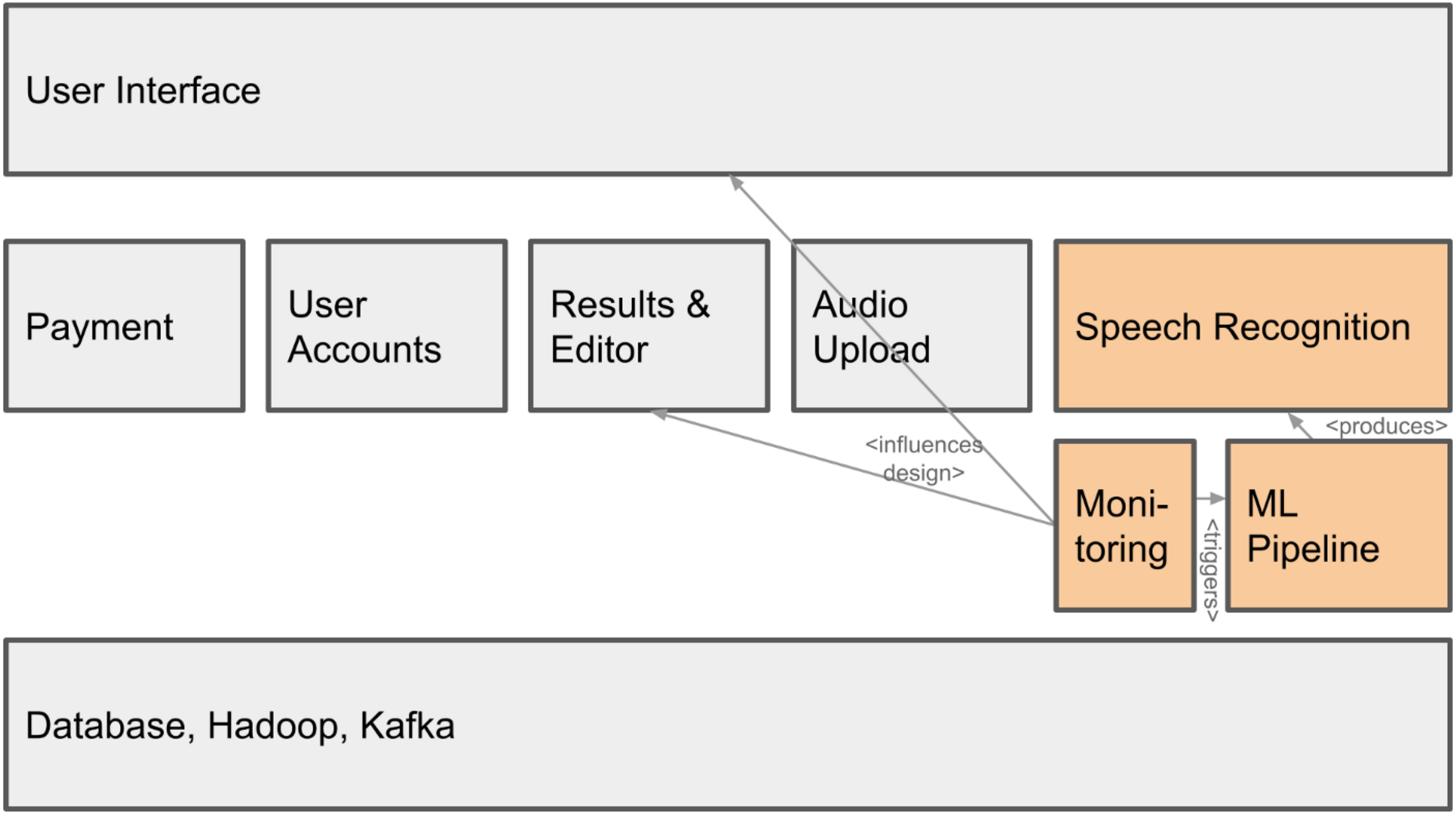

Tryton - Administrator - GNU SOLIDARIO HOSPITAL [Euro]
File User Options Favorites Help
screen
Addresses
Categories
Product
Financial
Currency
Inventory & Stock
Purchase
Calendar
Health
Patients
Institutions
Appointments
Prescriptions
Demographics
Laboratory
Imaging
Hospitalizations
Surgeries
Pediatrics
Archives
Nursing
Health Services
Reporting
Configuration
Patients
Obstetric Hist ...
Patients
1 / 8
New
Save
Switch
Reload
Previous
Next
Attachment(0)
Action
Relate
Report
E-Mail
Print
Main Info
Betz, Ana
Female
Age: 29y 3m 20d
Critical Information
Personal history of allergy to penicillin
Insulin-dependent diabetes mellitus
Severe allergic reactions to β-lactams
General Info
Socioeconomics
Medication
Diseases
Surgeries
Genetics
Lifestyle
OB/GYN
General
Screening
Fertile:
Pregnant:
Menarche age: 12
Menopausal:
Menopause age:
OB summary
Pregnancies: 1
Premature: 0
Abortions: 0
Stillbirths: 0
Menstrual History
(1/1)
Date
LMP
Length
frequency
volume
Regular
Dysmenorrhea
Reviewed
Institution
01/24/2015
01/20/2015
5
eumenorrhea
normal
Cordara, Cameron
GNU SOLIDARIO HOSPITAL
tryton://health.gnusolidario.org:8000/health28rc1/model/gnuhealth.patient/1;views=%5B223%2C+224%5D

PRODUCTION ML SYSTEMS



temi.com


User Interface
Payment
User Accounts
Results & Editor
Audio Upload
Speech Recognition
Database, Hadoop, Kafka


User Interface
Payment
User Accounts
Results & Editor
Audio Upload
Speech Recognition
<produces>
<influences design>
Moni-toring
ML Pipeline
<triggers>
Database, Hadoop, Kafka

FROM PROTOTYPE TO PRODUCTION

DZone > AI Zone > Top 10 Reasons Why 87% of Machine Learning Projects Fail

Top 10 Reasons Why 87% of Machine Learning Projects Fail

In this article, find out why 87% of machine learning projects fail.

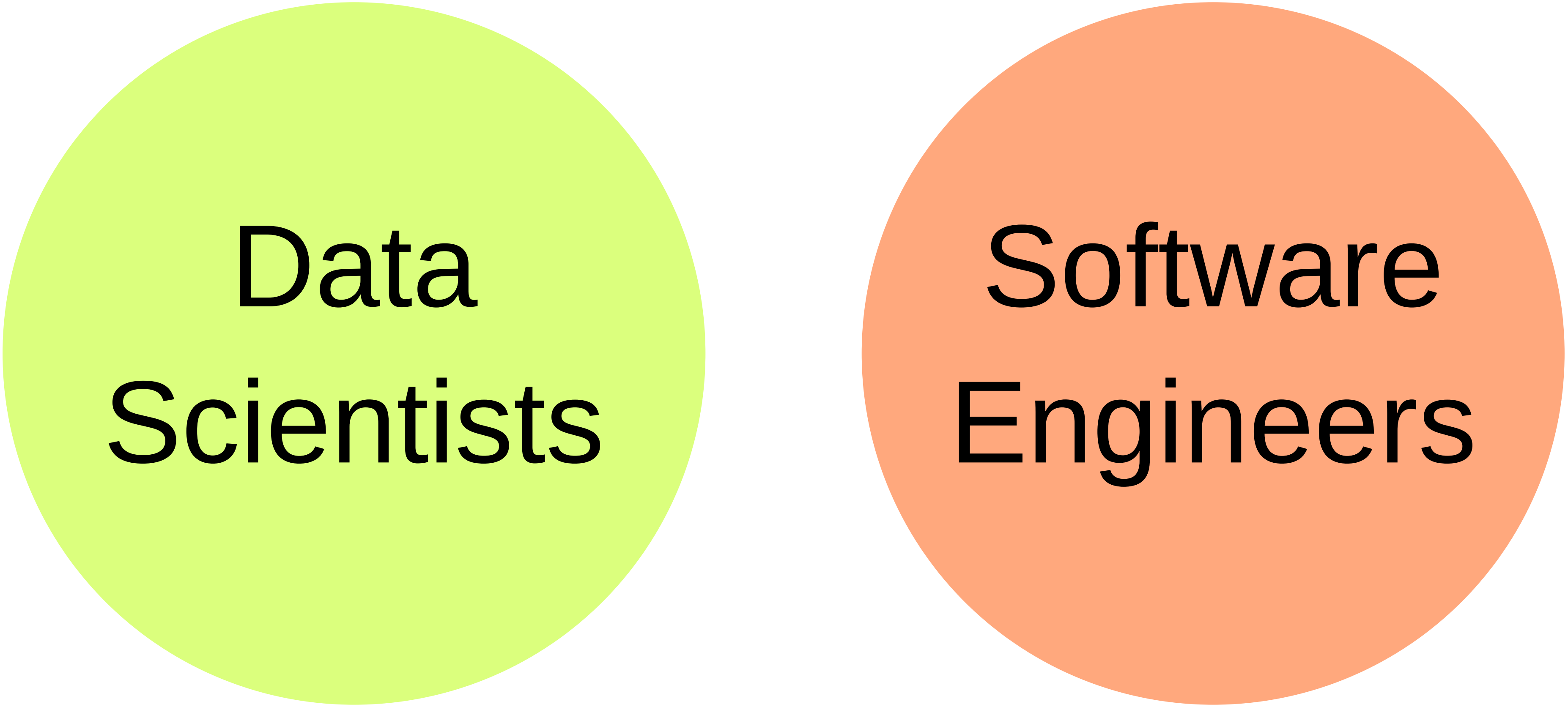
by Prajeen MV · Oct. 13, 20 · AI Zone · Opinion

Like (6) Comment (4) Save Tweet 9.51K Views

We see news about machine learning everywhere. Indeed, there is a lot of potential in machine learning. According to Gartner's predictions, "*Through 2020, 80% of AI projects will remain alchemy, run by wizards whose talents will not scale in the organization*" and Transform 2019 of VentureBeat predicted that *87% of AI projects will never make it into production.*

Why is it like that? Why do so many projects fail?

https://dzone.com/articles/top-10-reasons-why-87-of-the-machine-learning-proj



and domain experts + lawyers + operators + security experts + regulators + ...

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..."

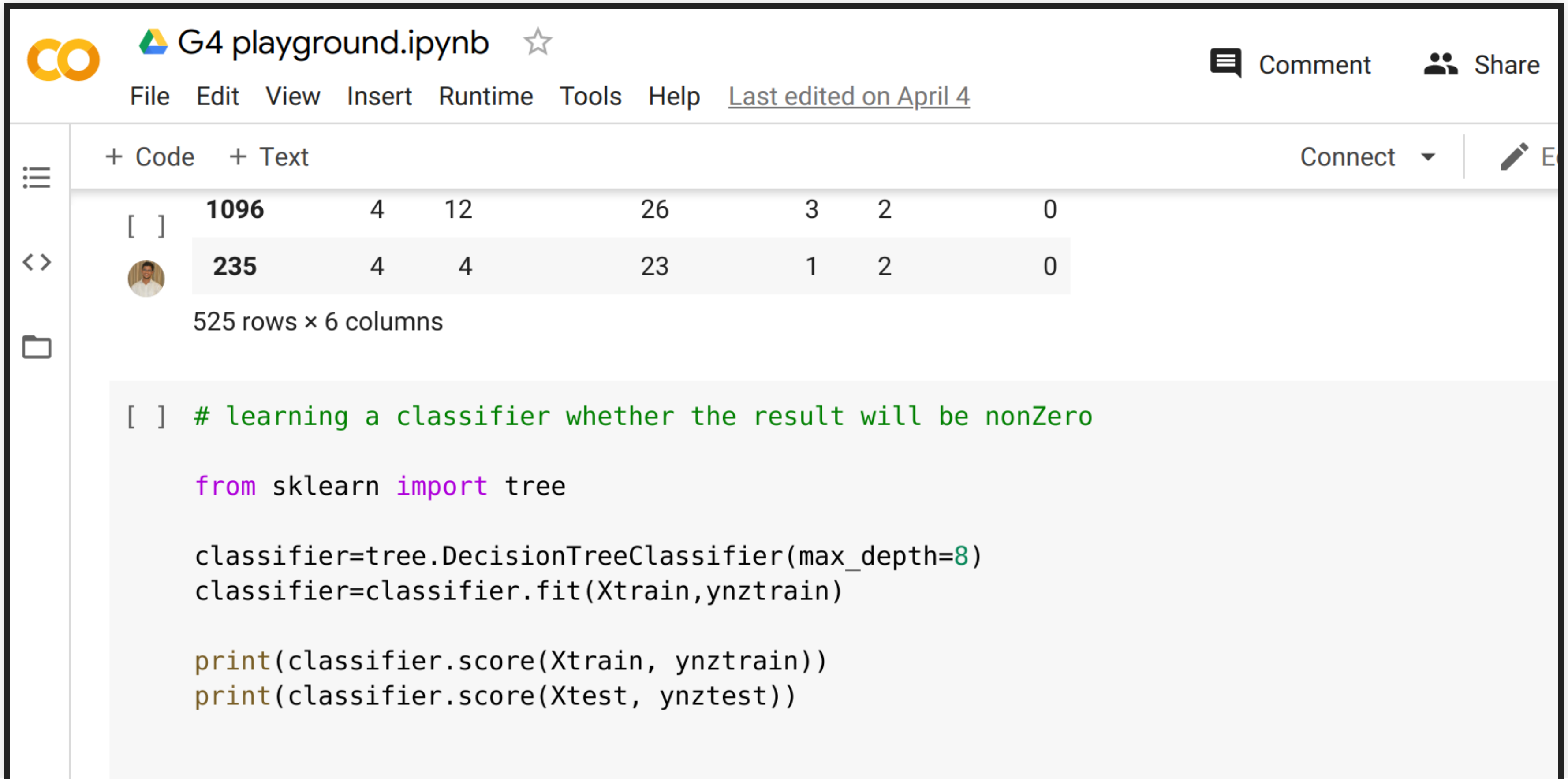
Mary Shaw. ed. Software Engineering for the 21st Century: A basis for rethinking the curriculum. 2005.

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))
```

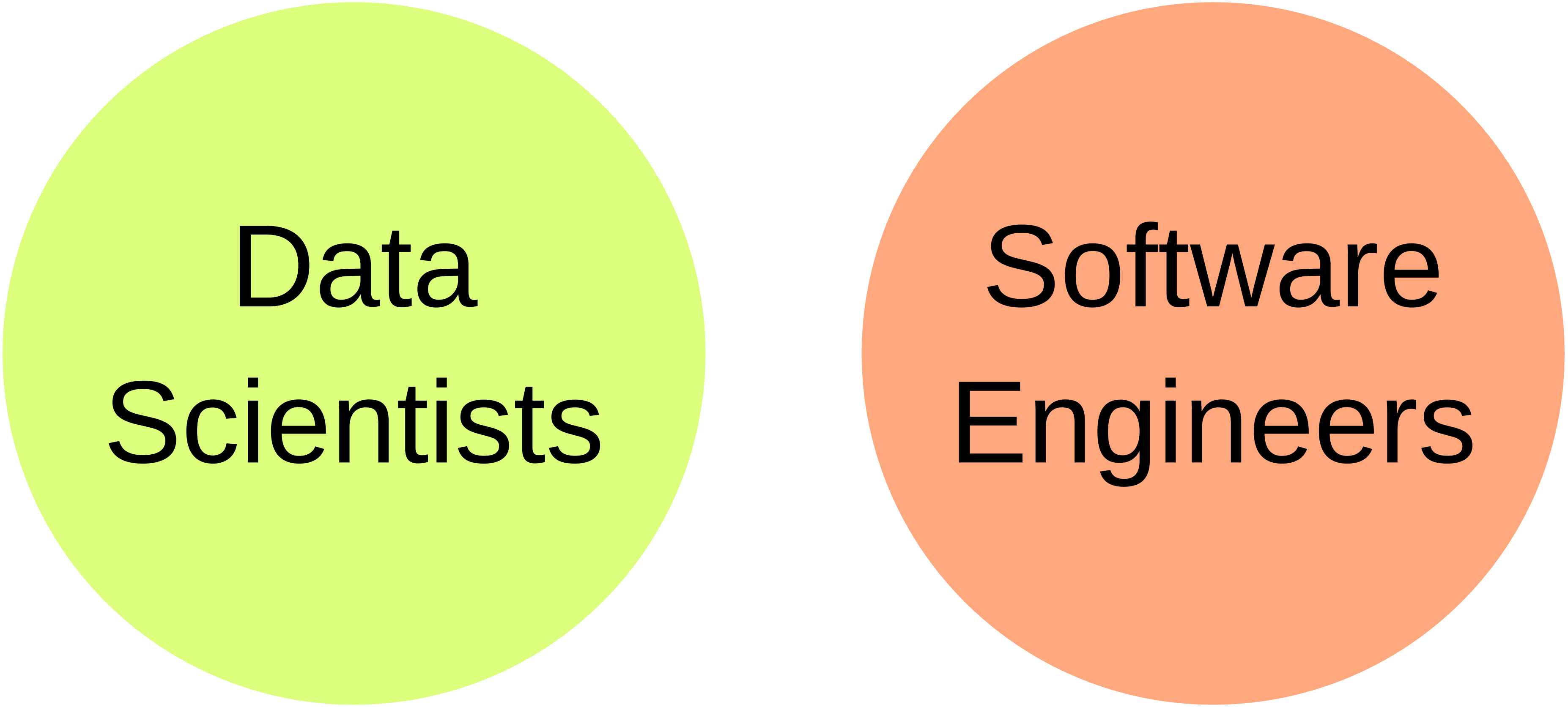
```
0.9376379365613154
-2.437397740412892
```

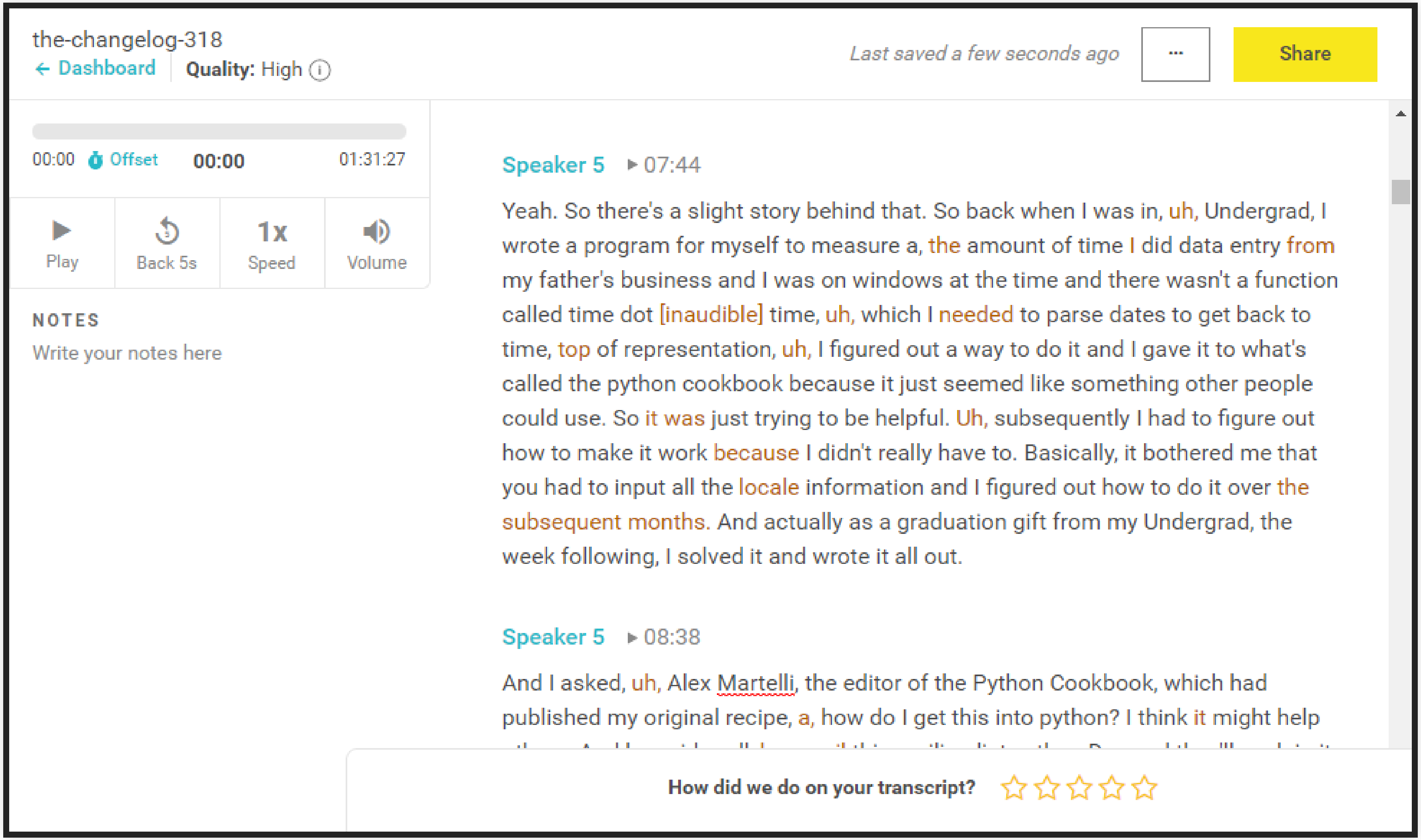
DATA SCIENTIST

- 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, ...

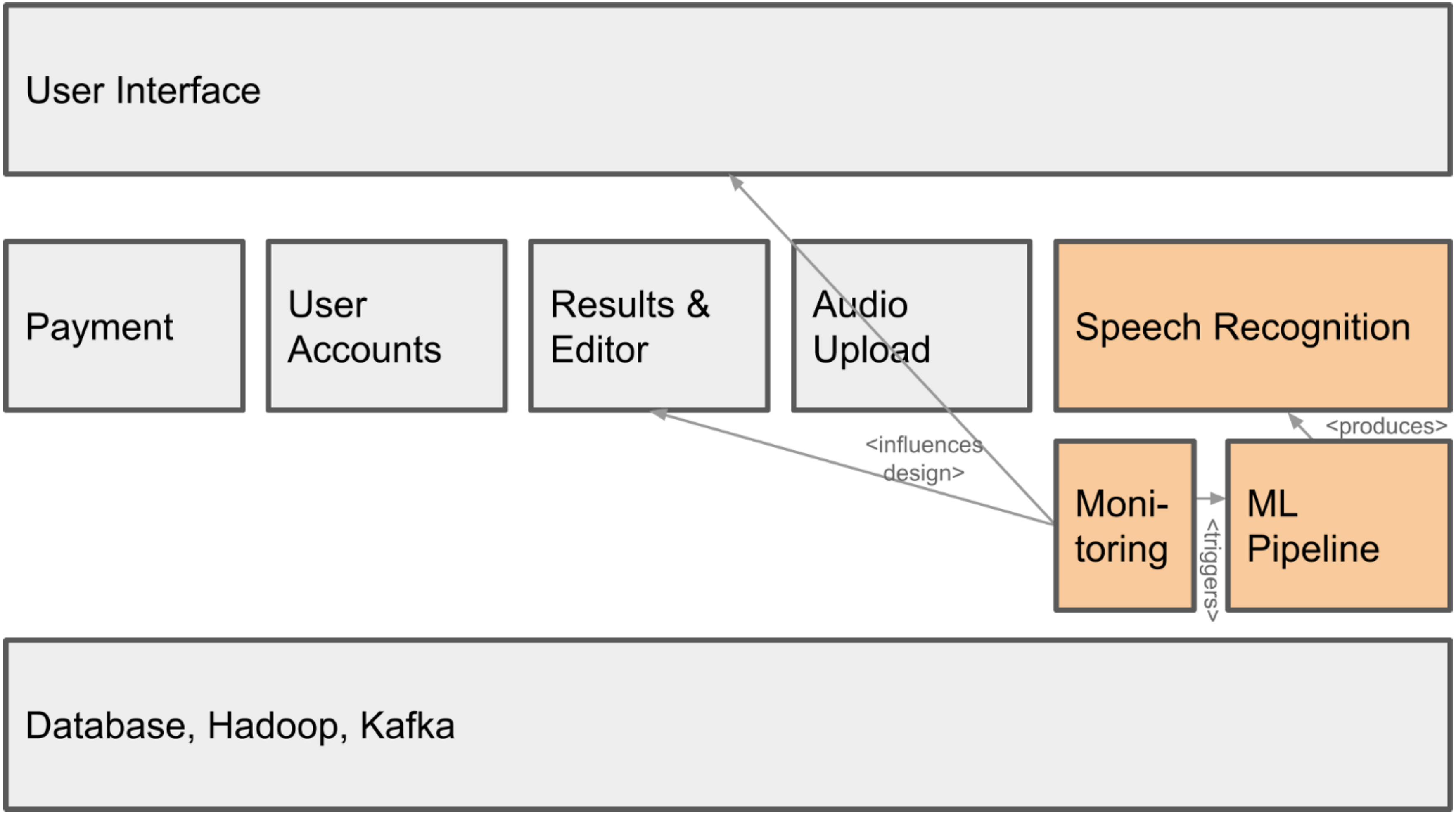
SOFTWARE ENGINEER

- 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?


User Interface
Payment
User Accounts
Results & Editor
Audio Upload
Speech Recognition
<produces>
<influences design>
Moni-toring
ML Pipeline
<triggers>
Database, Hadoop, Kafka

PART 1:

HOW DOES MACHINE LEARNING CHANGE SOFTWARE ENGINEERING?

WHAT'S DIFFERENT?

- Missing specifications
- Environment is important (feedback loops, data drift)
- Nonlocal and nonmonotonic effects
- Data is central and BIG
- ...

MANAGING COMPLEXITY IN SOFTWARE

- **Abstraction**: Hide details & focus on high-level behaviors
- **Reuse**: Package into reusable libraries & APIs with well-defined *contracts*
- **Composition**: Build large components out of smaller ones

```
/**
 * 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
 *
 * Adjusted gross income must be positive;
 * returned deductions are not negative.
 */
float computeDeductions(float agi, Expenses expenses) {
  ...
}
```

DIVIDE AND CONQUER

- Human cognitive ability is limited
- Decomposition of software necessary to handle complexity
- Allows division of labor
- Deductive reasoning, using logic
- Testing each component against its specification

```
//@ requires x >= 0.0;
/*@ ensures JMLDouble.approximatelyEqualTo(x,
  @                                           \result * \result,
  @                                           eps);
  @*/
public static double sqrt(double x) {
    /*...*/
}
```

ML: MISSING SPECIFICATIONS

from deductive to inductive reasoning, from specs to examples

```
/**
  ????
*/
String transcribe(File audioFile);
```

```
/**
  ????
*/
Boolean predictRecidivism(int age,
                          List<Crime> priors,
                          Gender gender,
                          int timeServed,
                          ...);
```

```
/**
  ????
*/
Boolean hasCancer(byte[][] image);
```

> *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

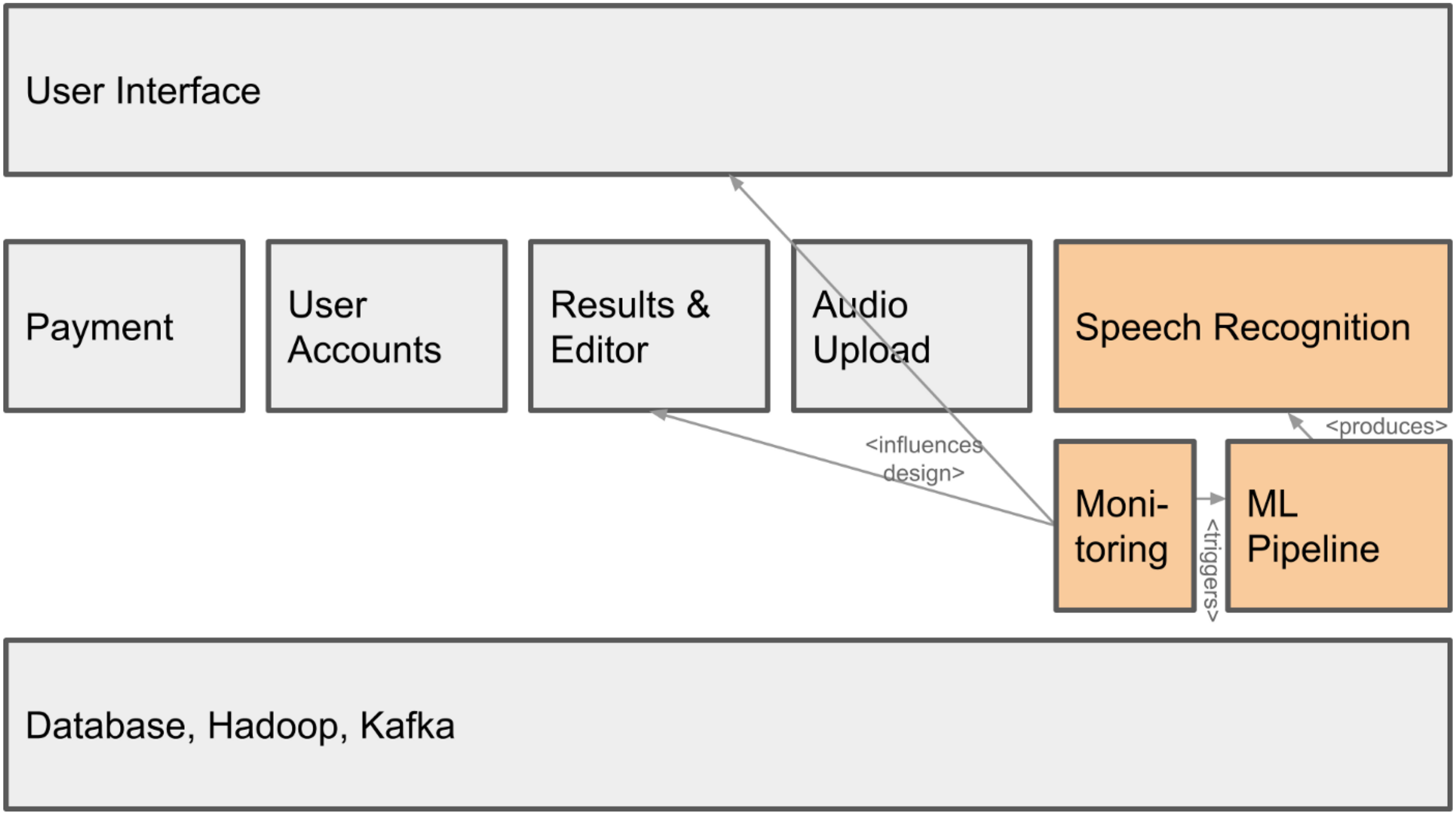
> *2nd law: "the rate of change of momentum of a body over time is directly proportional to the force applied, and occurs in the same direction as the applied force"* $\boldsymbol{F} = \frac{d\boldsymbol{p}}{dt}$

"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

CONSEQUENCE: ML AS UNRELIABLE COMPONENTS



SOFTWARE ENGINEERING REALITY

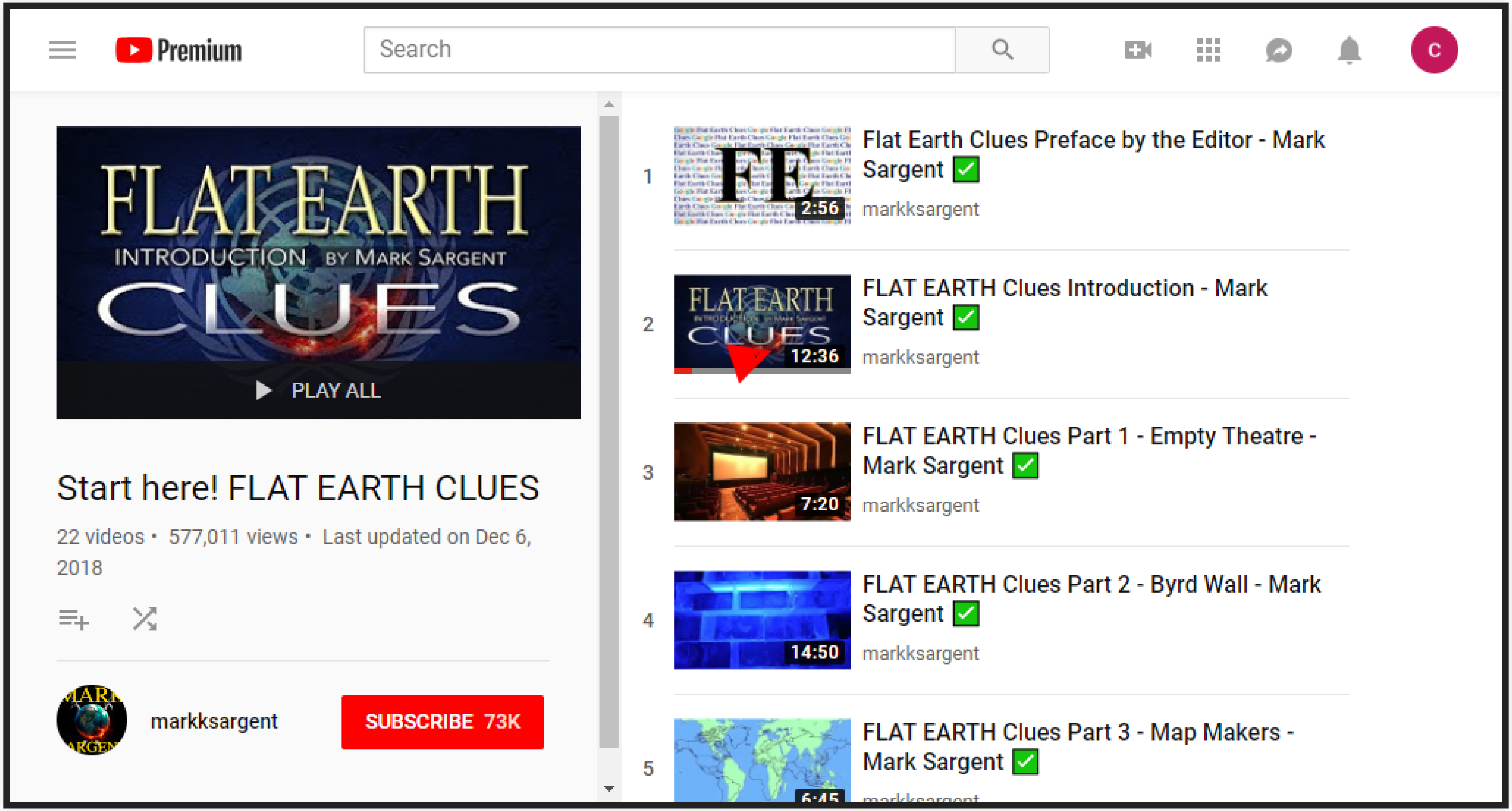
- Missing and weak specs very common
- Agile methods
- Communication over formal specifications
- Integration and system testing, not just unit testing
- Testing in production

- Safe systems from unreliable components
- Safety engineering, risk analysis, mitigation strategies

See also Christian Kaestner. "Machine learning is requirements engineering". Medium 2020.

ML: ENVIRONMENT IS IMPORTANT

(feedback loops, data drift, safety concerns)

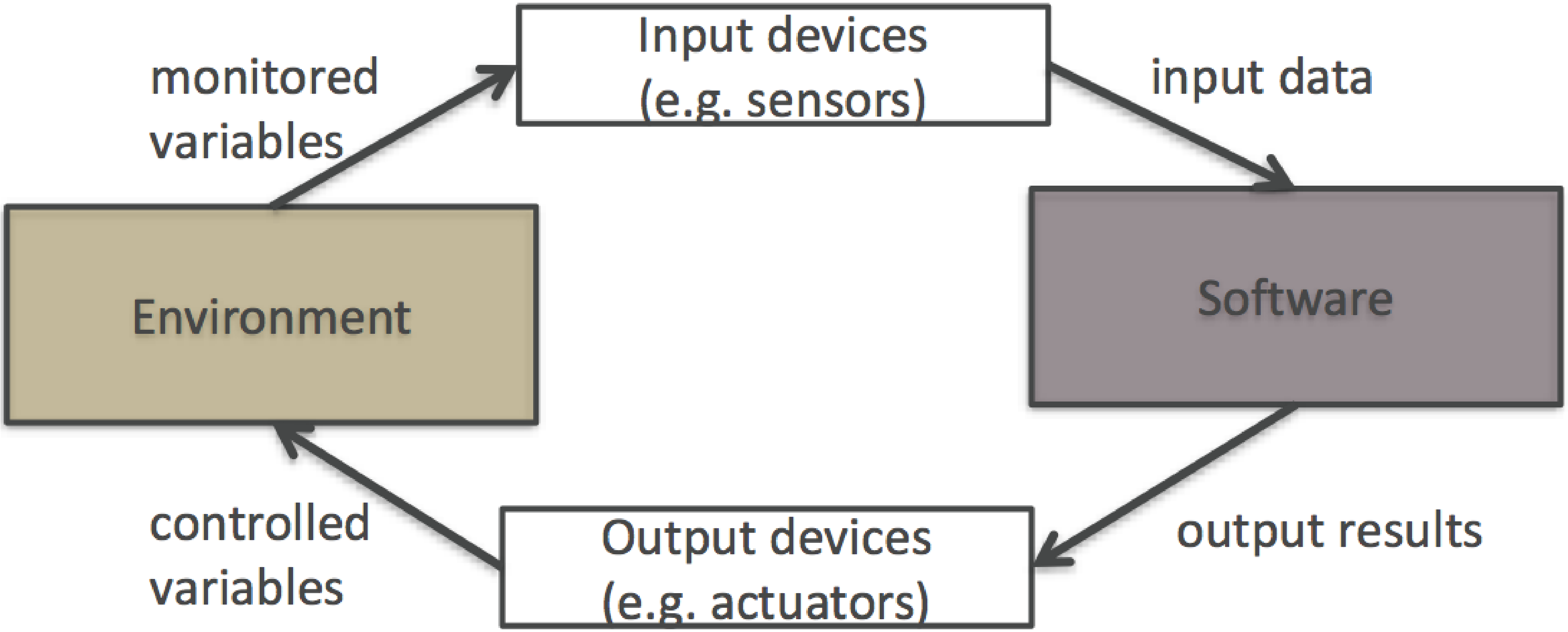


SOFTWARE ENGINEERING REALITY

(Lufthansa Flight 2904)

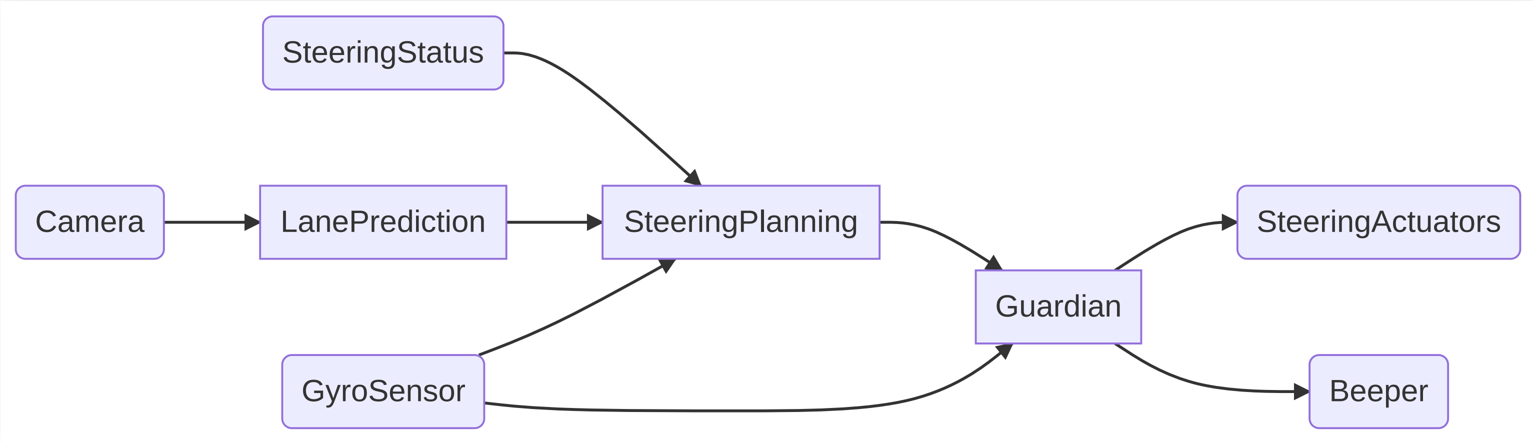
SOFTWARE ENGINEERING REALITY

- The environment is often important
- Most safety concerns stem from interactions between world and machine (Jackson ICSE 95)
- Requirements engineering is essential



ML: NONLOCAL AND NONMONOTONIC EFFECTS

multiple models in most systems



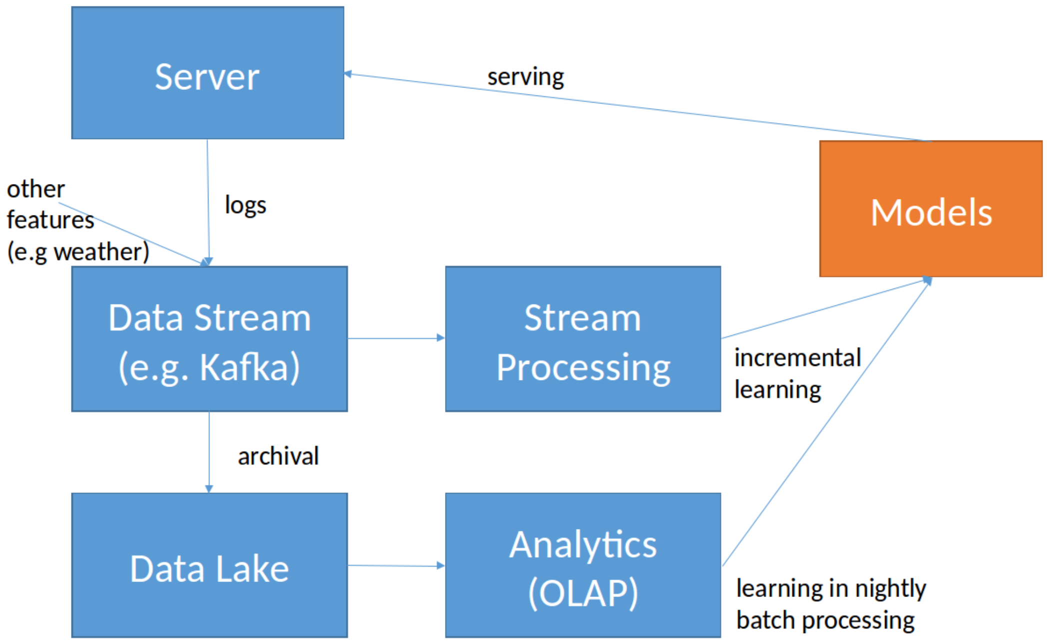
SOFTWARE ENGINEERING REALITY

- Subsystems and plugins may interact in unanticipated ways
- Feature interactions hard to predict
- Software design is important
- System testing is important

ML: DATA IS ESSENTIAL AND BIG

SOFTWARE ENGINEERING REALITY

- Software architecture and design for scalability
- Distributed systems
- Batch processing, stream processing, lambda architecture
- Databases, big data, cloud infrastructure
- Extensive work on data schema, versioning, and provenance



SO, WHAT'S DIFFERENT?

- Missing specifications
- Environment is important (feedback loops, data drift)
- Nonlocal and nonmonotonic effects
- Data is central and BIG
- ...

Not all new, but pushing the envelope in system complexity

MY VIEW

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

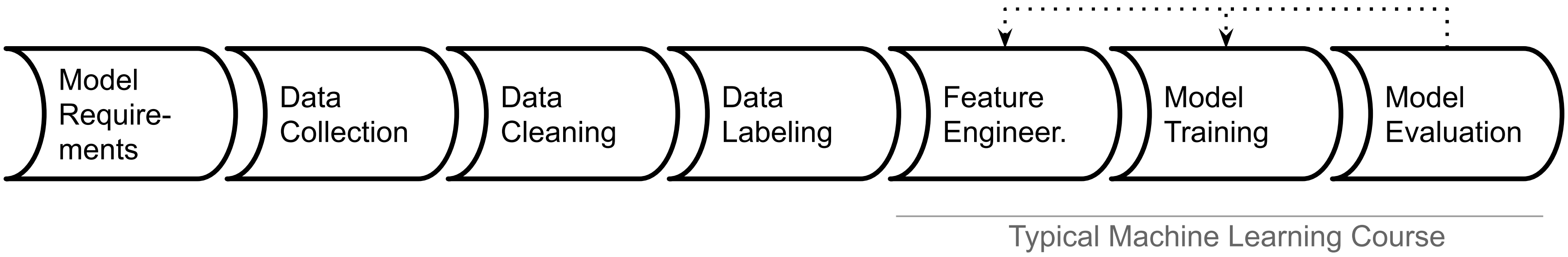
PART 2:

FROM MODEL TO SYSTEM

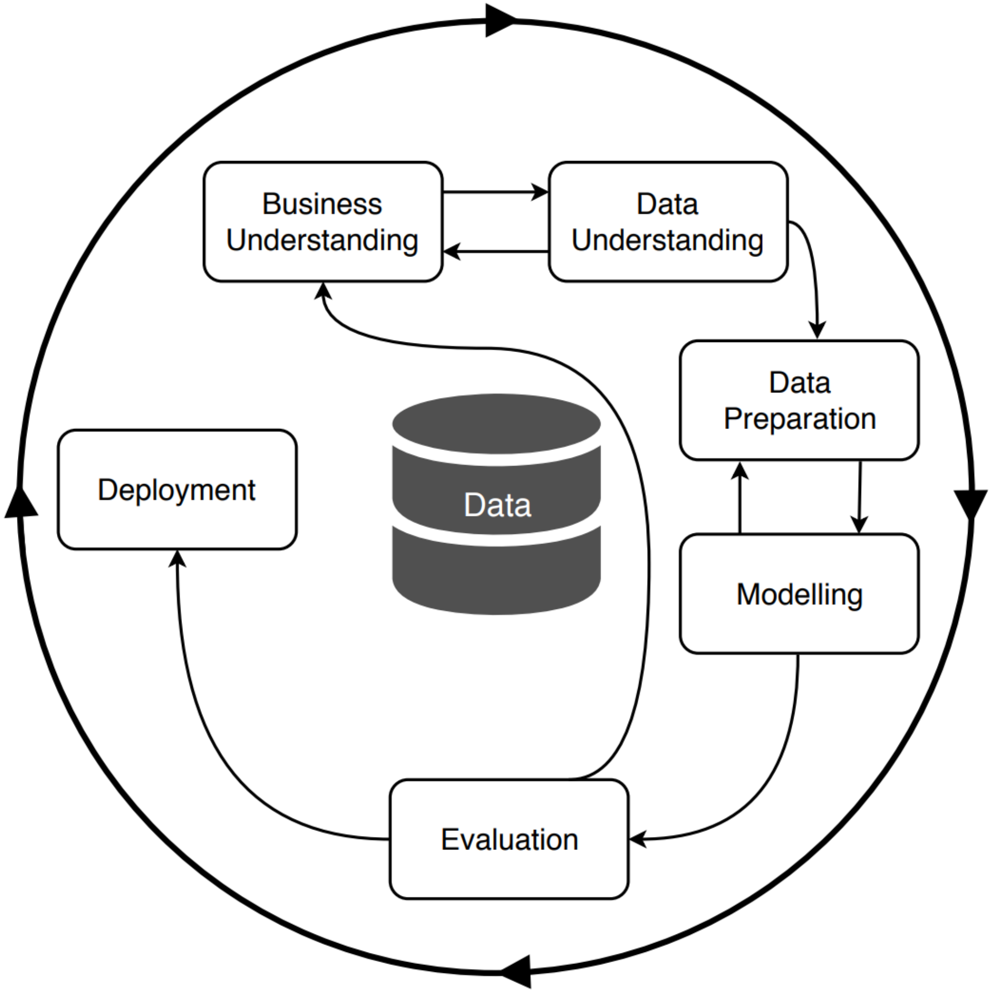
Illustrated with *Quality Assurance*

TRAINING A MODEL

(often in computational notebooks)

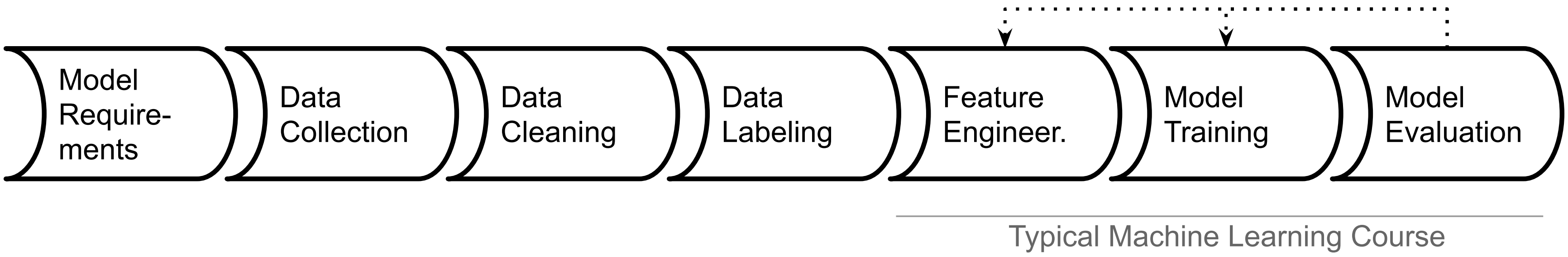


CRISP-DM PROCESS MODEL



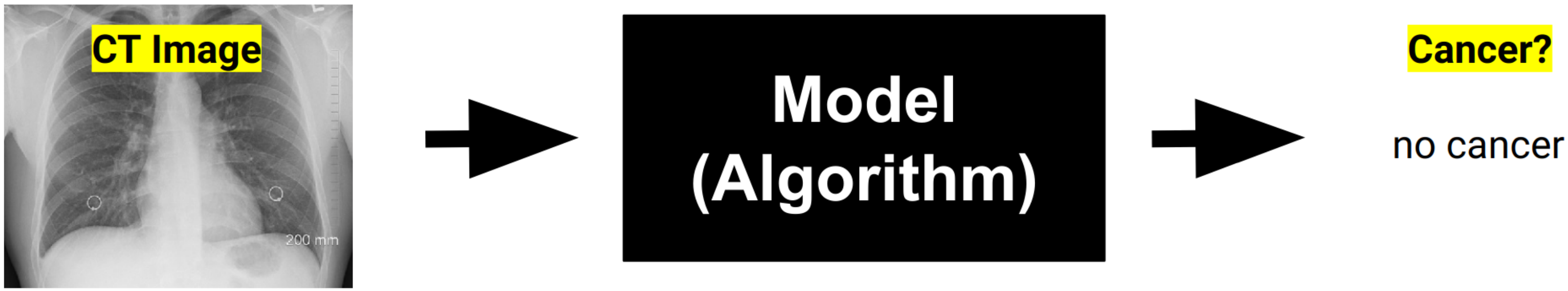
TRAINING A MODEL

(often in computational notebooks)



TRADITIONAL FOCUS: MODEL ACCURACY

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



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

from sklearn import tree

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

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

```
0.9376379365613154
-2.437397740412892
```

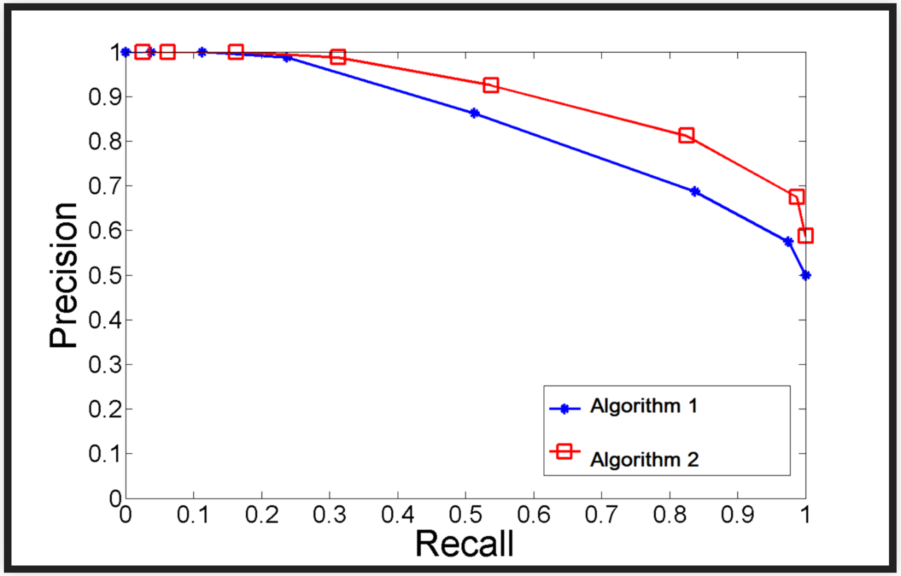
TRADITIONAL FOCUS: MODEL ACCURACY

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

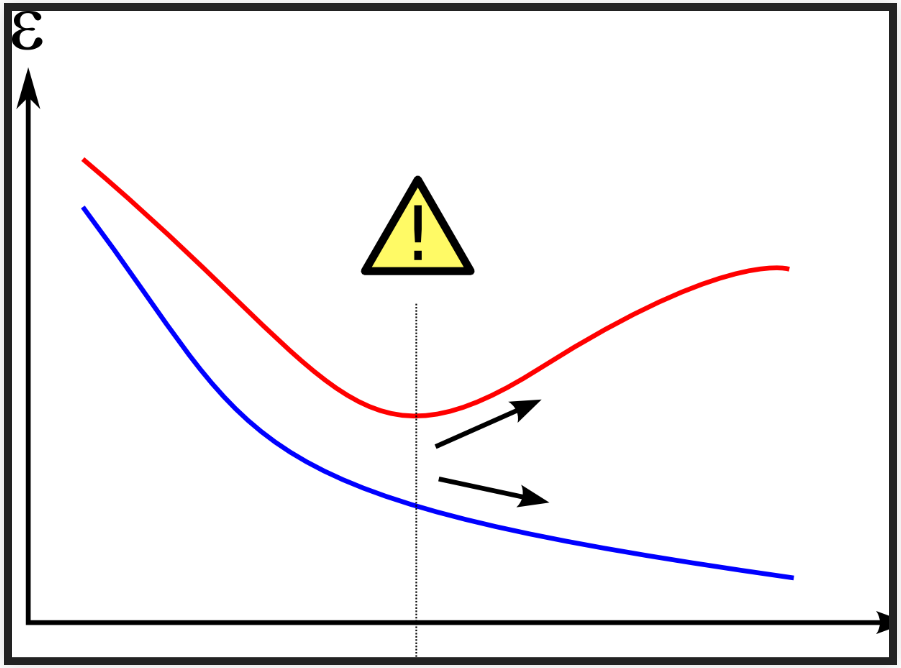
Accuary, Recall, Precision, F1-Score

MORE TRADITIONAL MODEL QUALITY DISCUSSIONS

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



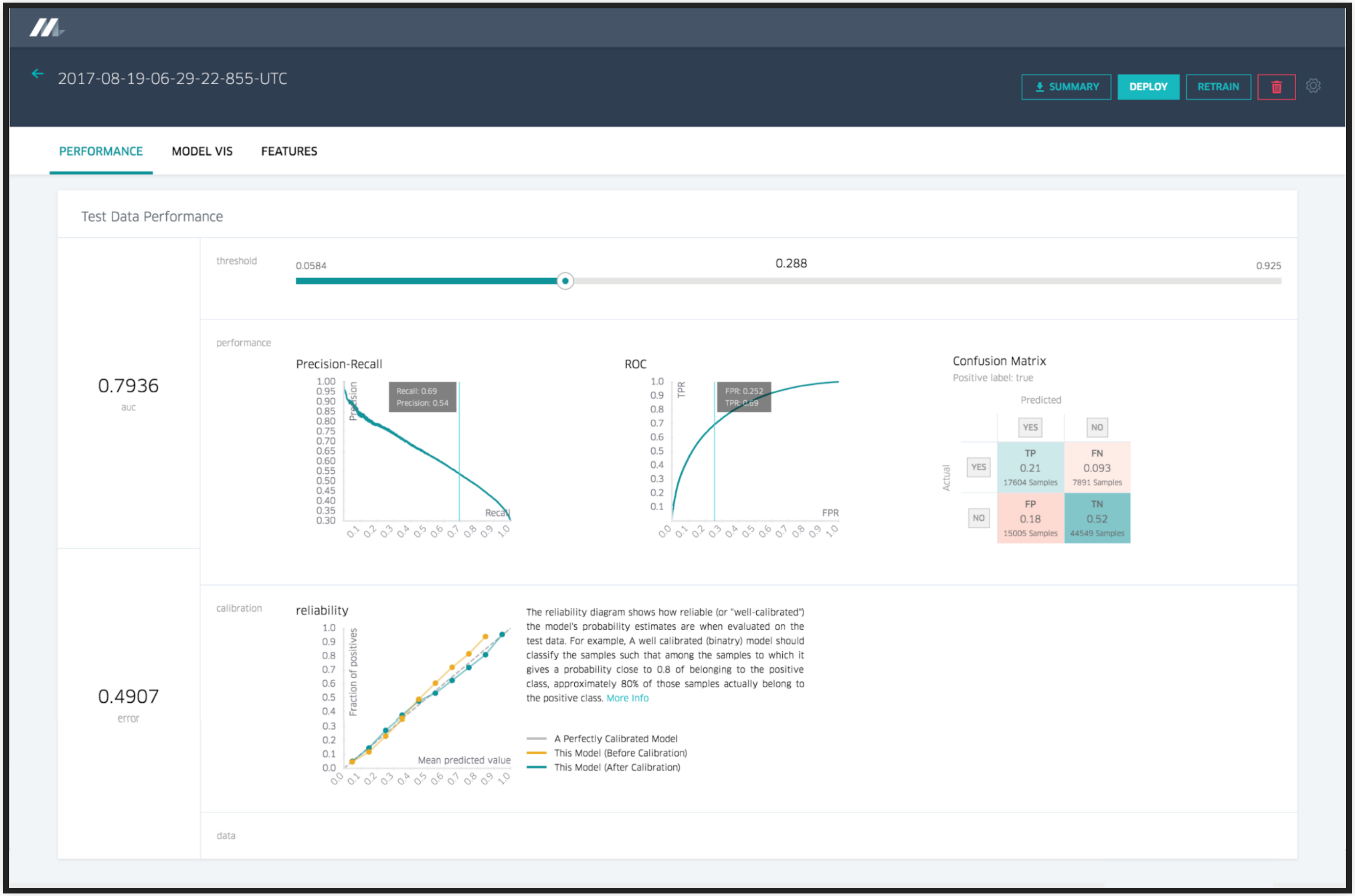
Generalization/overfitting (train/test/eval split, crossvalidation)



(CC SA 3.0 by Dake)

AUTOMATING MODEL EVALUATION

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

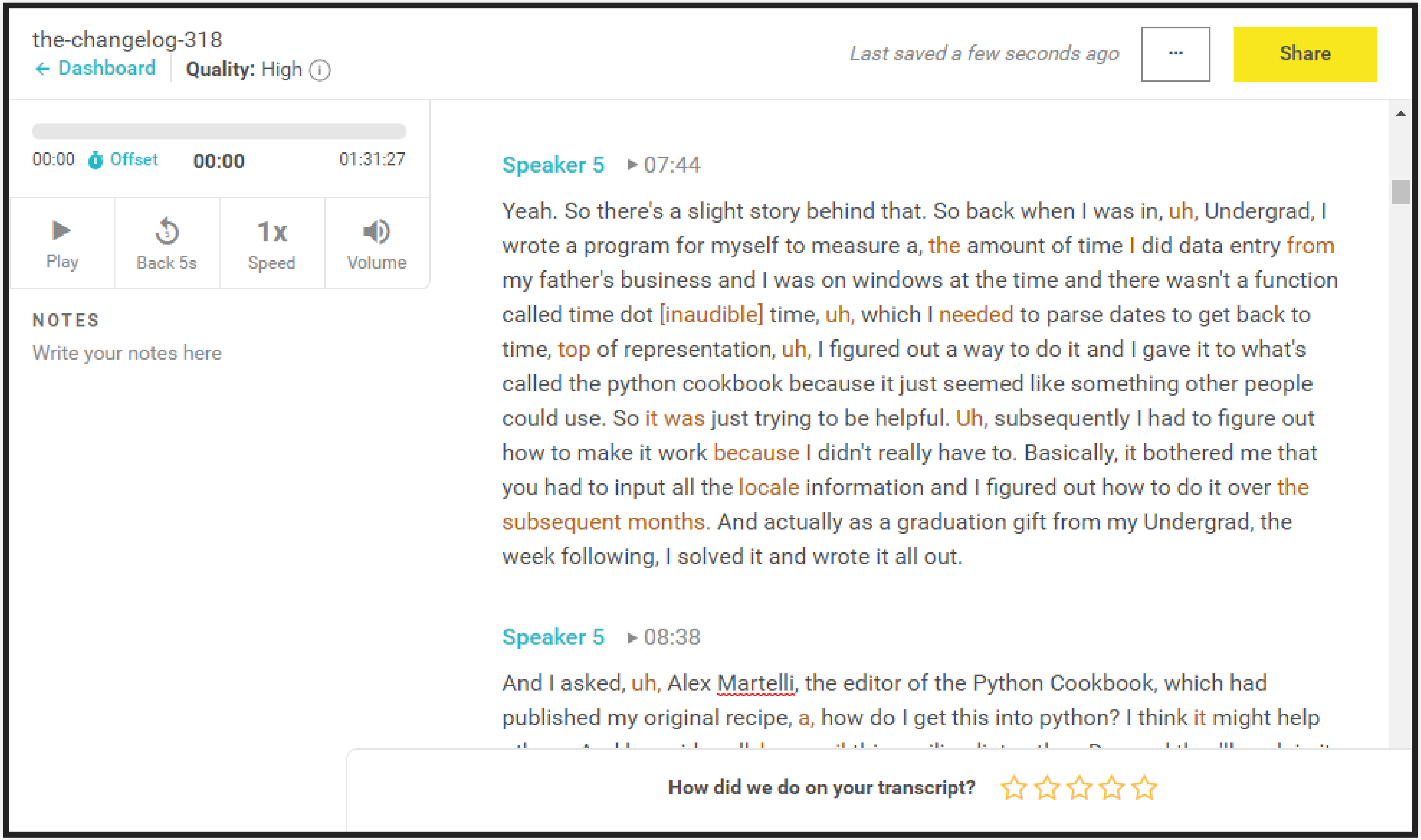


2017-08-19-06-29-22-855-UTC
SUMMARY
DEPLOY
RETRAIN
PERFORMANCE
MODEL VIS
FEATURES
Test Data Performance
threshold
0.0584
0.288
0.925
0.7936
auc
performance
Precision-Recall
Recall: 0.69
Precision: 0.54
Precision
Recall
ROC
FPR: 0.252
TPR: 0.69
TPR
FPR
Confusion Matrix
Positive label: true
Predicted
YES
NO
Actual
TP
0.21
17604 Samples
FN
0.093
7891 Samples
FP
0.18
15005 Samples
TN
0.52
44549 Samples
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)
0.4907
error
data

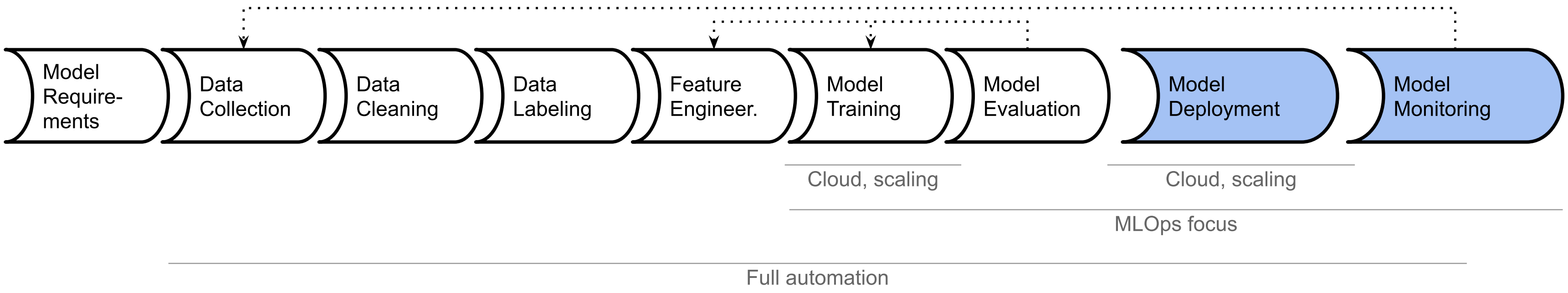
BEYOND ACCURACY:

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)

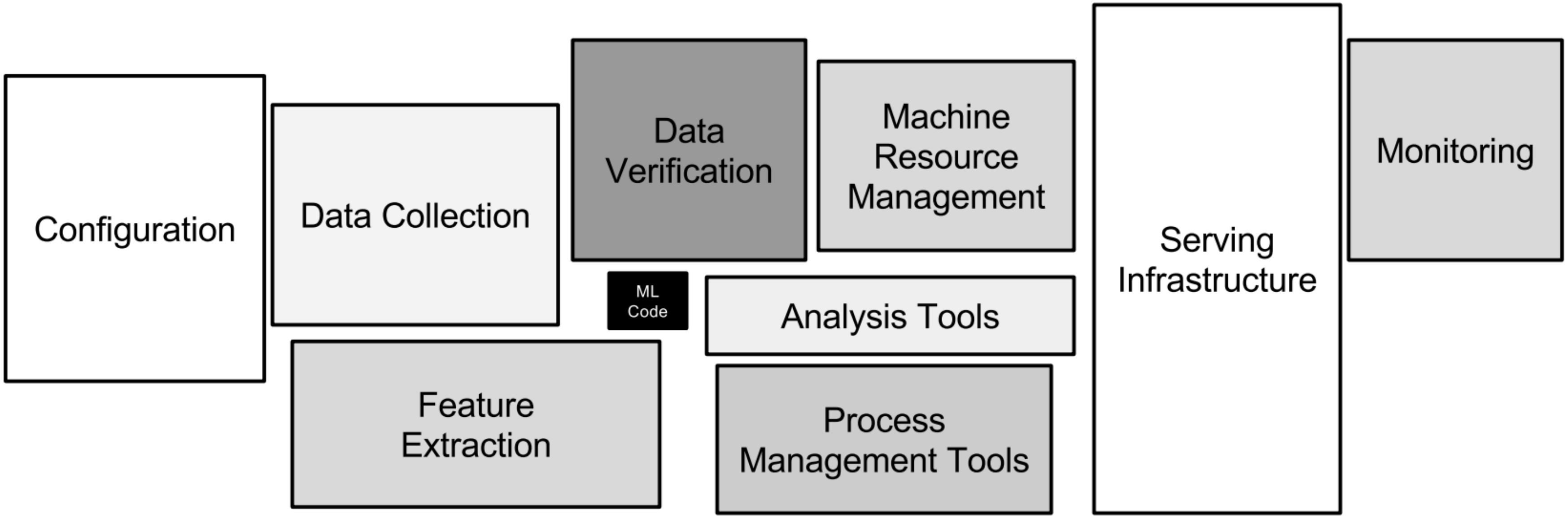

the-changelog-318
← Dashboard
Quality: High
Last saved a few seconds ago
···
Share
00:00
Offset
00:00
01:31:27
Play
Back 5s
1x
Speed
Volume
NOTES
Write your notes here
Speaker 5 ▶ 07:44
Yeah. So there's a slight story behind that. So back when I was in, uh, Undergrad, I wrote a program for myself to 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?

DEPLOYING AND UPDATING MODELS WITH PIPELINES



Automate each step -- test each step

ML ENGINEERING: BUILDING PIPELINES



(Nowadays, MLOps is shrinking most of these boxes)

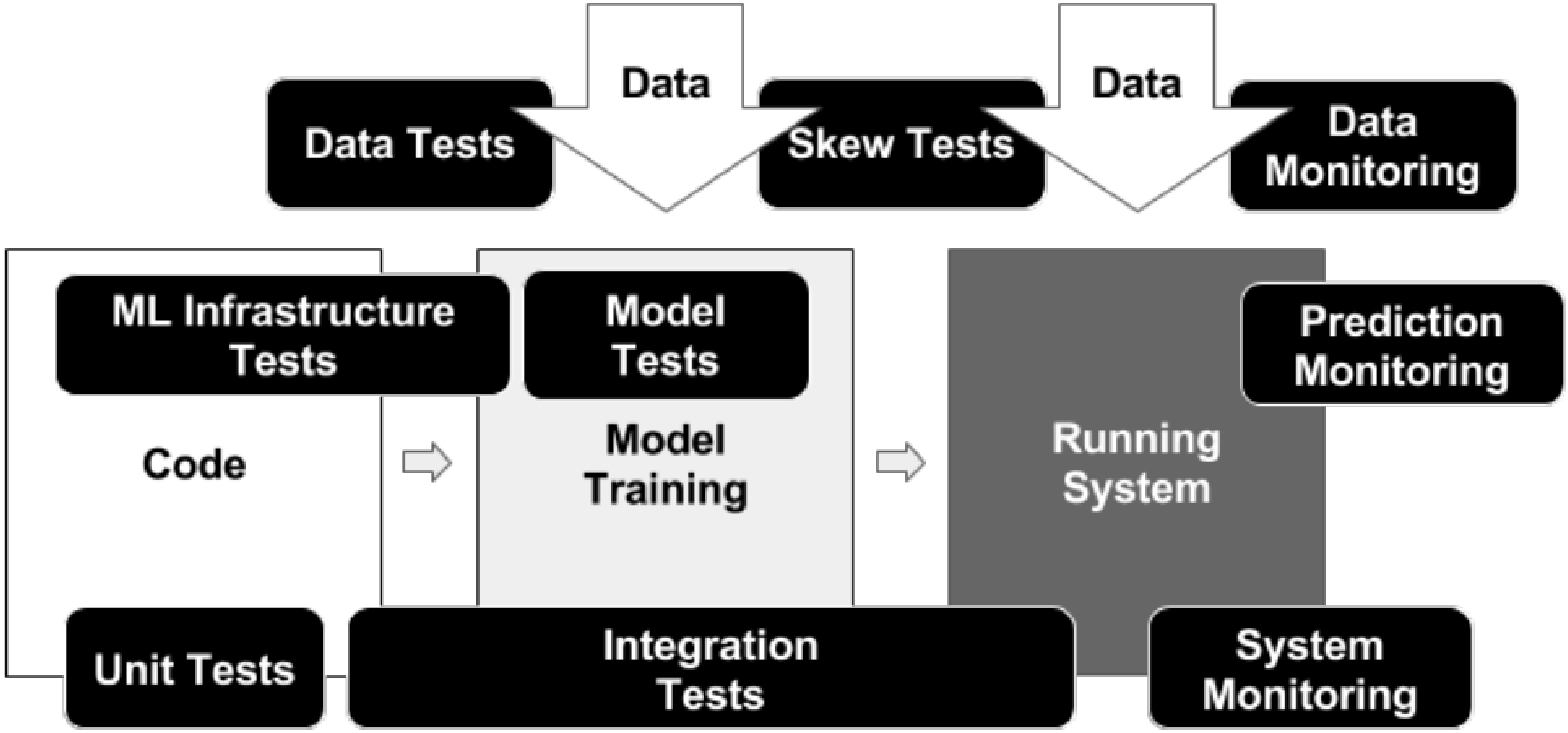
Source: Sculley, David, Gary Holt, Daniel Golovin, Eugene Davydov, Todd Phillips, Dietmar Ebner, Vinay Chaudhary, Michael Young, Jean-Francois Crespo, and Dan Dennison. "Hidden technical debt in machine learning systems." Advances in neural information processing systems 28 (2015): 2503-2511.

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
- ...

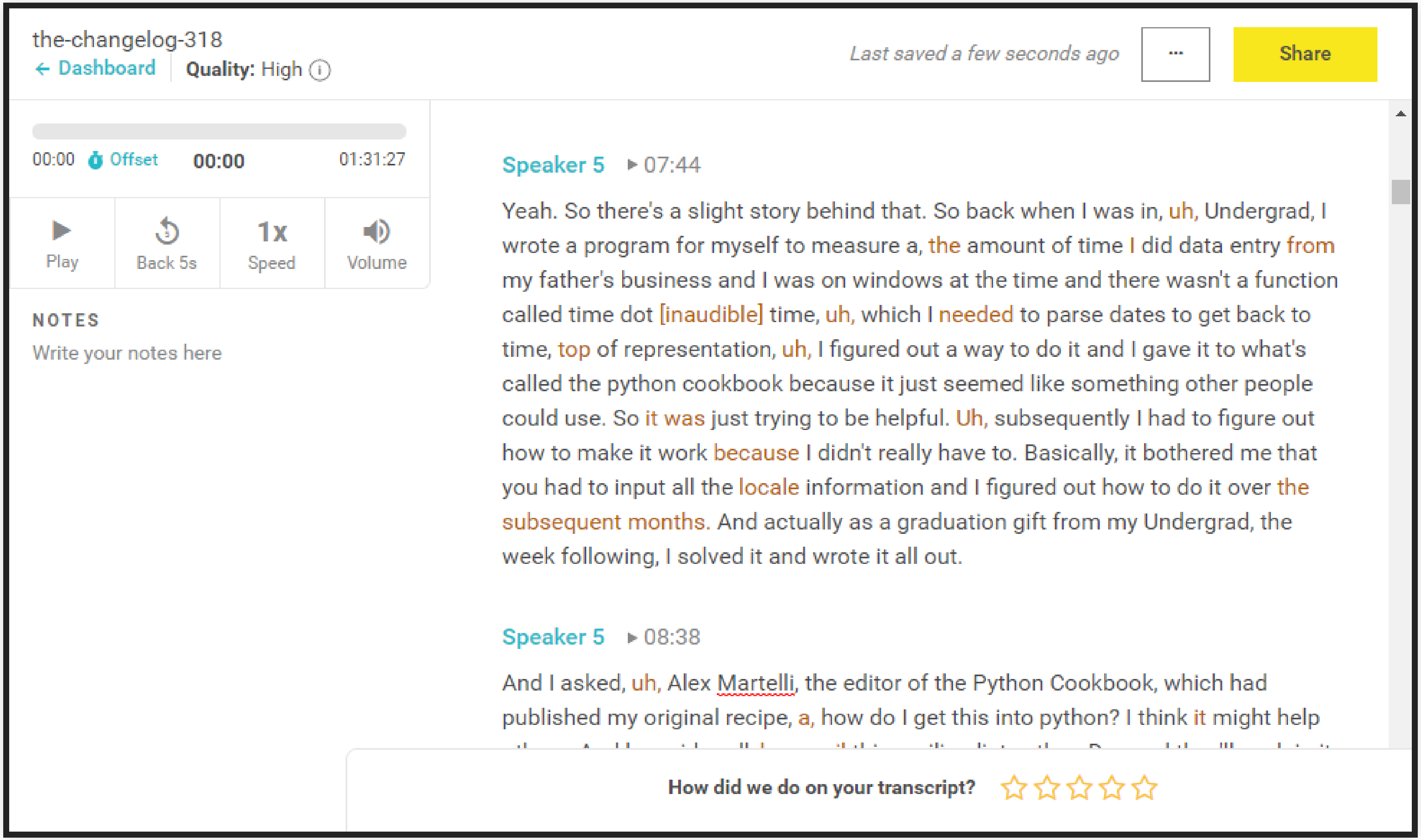
QUALITY ASSURANCE FOR THE ENTIRE PIPELINE



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)

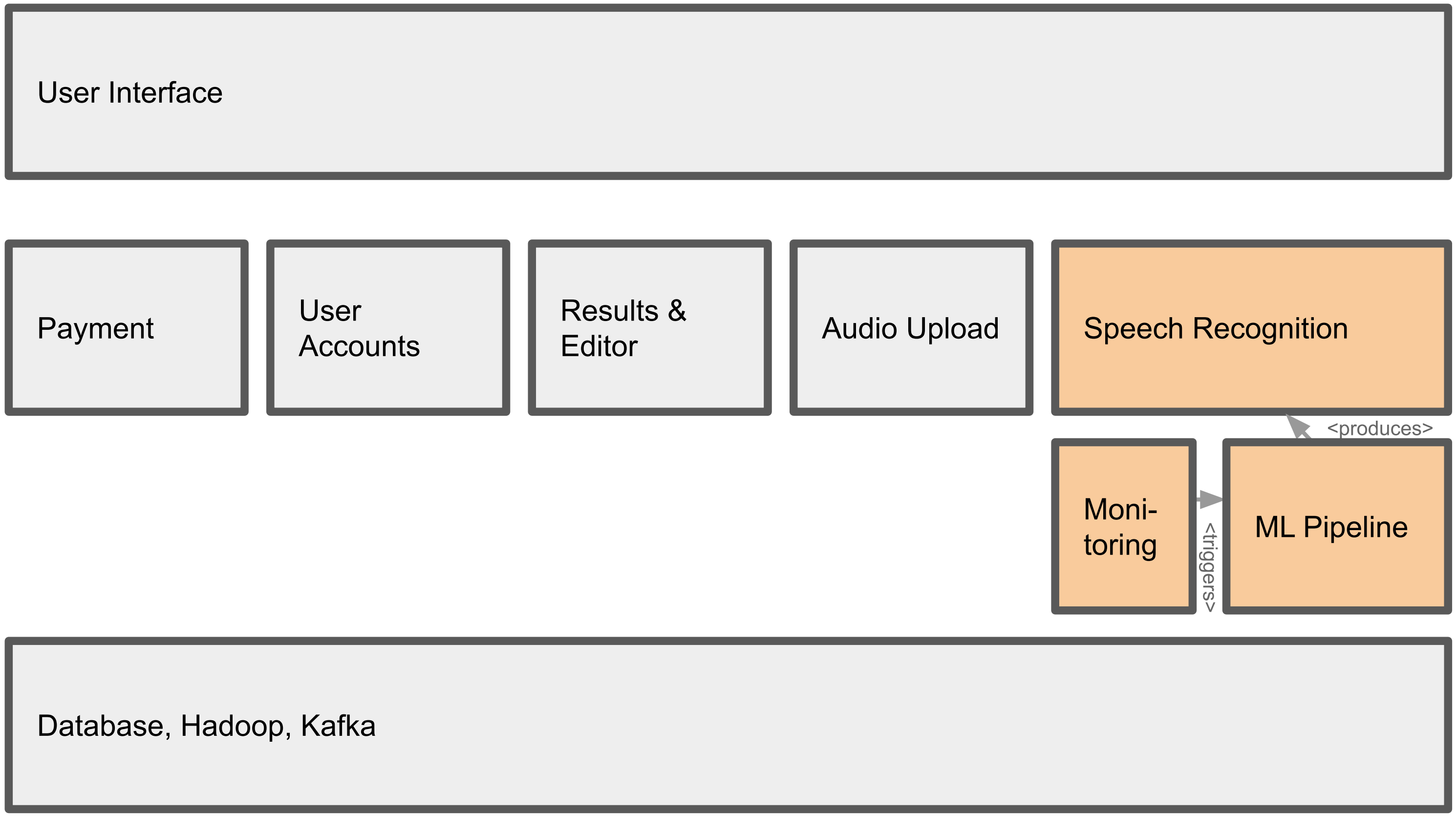
PIPELINE TESTING

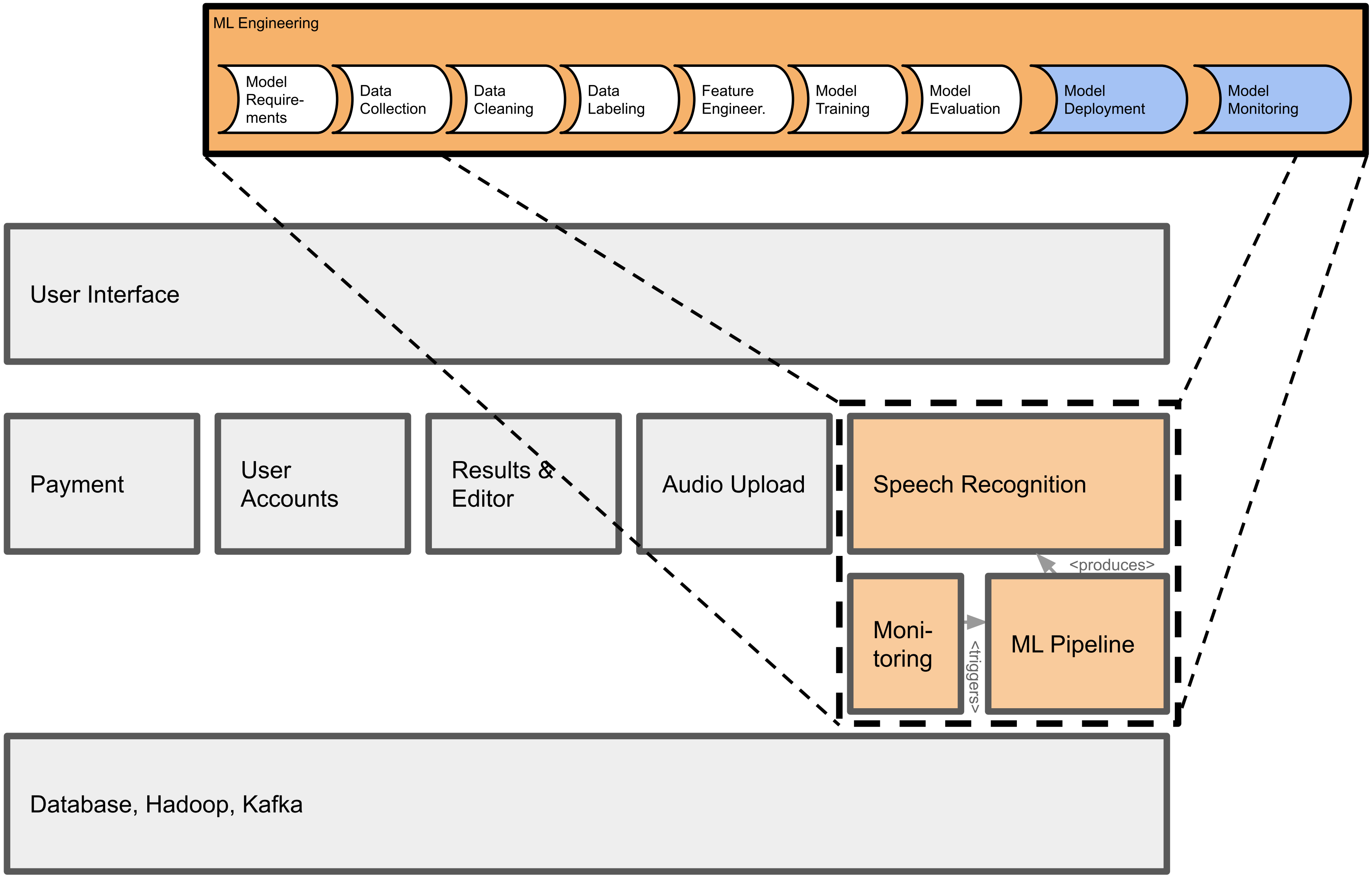
- Unit tests (e.g., data cleaning)
- End to end pipeline tests
- Testing with stubs, test error handling (e.g., test model redeployment after dropped connection)
- Test monitoring infrastructure (e.g., "fire drills")
- Chaos engineering


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?

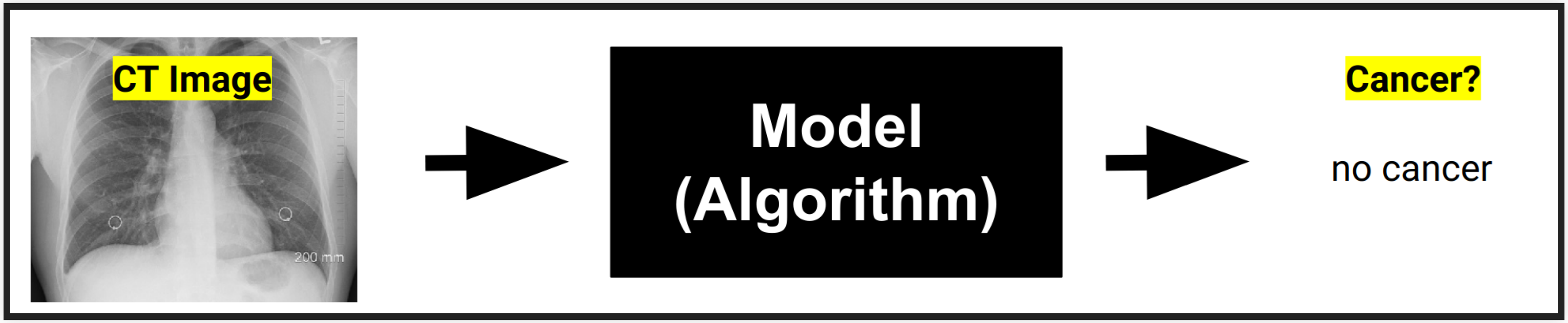
FOCUSING ON THE SYSTEM

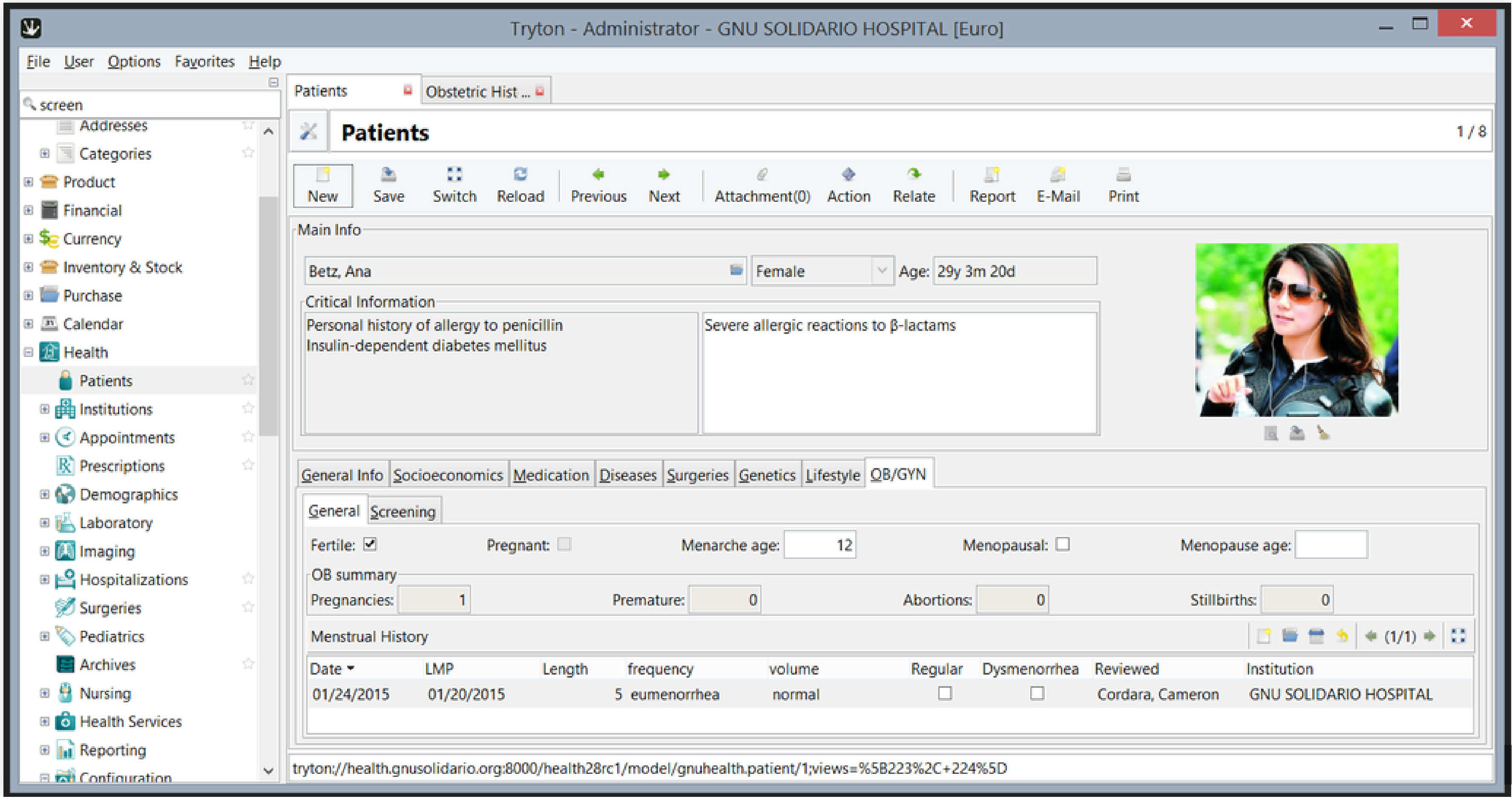
ML models are "just" one component


User Interface
Payment
User Accounts
Results & Editor
Audio Upload
Speech Recognition
<produces>
Moni-toring
<triggers>
ML Pipeline
Database, Hadoop, Kafka


ML Engineering
Model Require-ments
Data Collection
Data Cleaning
Data Labeling
Feature Engineer.
Model Training
Model Evaluation
Model Deployment
Model Monitoring
User Interface
Payment
User Accounts
Results & Editor
Audio Upload
Speech Recognition
<produces>
Moni-toring
<triggers>
ML Pipeline
Database, Hadoop, Kafka

DESIGNING THE RIGHT SYSTEM


CT Image
200 mm
Model
(Algorithm)
Cancer?
no cancer


Tryton - Administrator - GNU SOLIDARIO HOSPITAL [Euro]
File User Options Favorites Help
screen
Addresses
Categories
Product
Financial
Currency
Inventory & Stock
Purchase
Calendar
Health
Patients
Institutions
Appointments
Prescriptions
Demographics
Laboratory
Imaging
Hospitalizations
Surgeries
Pediatrics
Archives
Nursing
Health Services
Reporting
Configuration
Patients
Obstetric Hist ...
Patients
1 / 8
New
Save
Switch
Reload
Previous
Next
Attachment(0)
Action
Relate
Report
E-Mail
Print
Main Info
Betz, Ana
Female
Age: 29y 3m 20d
Critical Information
Personal history of allergy to penicillin
Insulin-dependent diabetes mellitus
Severe allergic reactions to β-lactams
General Info
Socioeconomics
Medication
Diseases
Surgeries
Genetics
Lifestyle
OB/GYN
General
Screening
Fertile:
Pregnant:
Menarche age: 12
Menopausal:
Menopause age:
OB summary
Pregnancies: 1
Premature: 0
Abortions: 0
Stillbirths: 0
Menstrual History
(1/1)
Date
LMP
Length
frequency
volume
Regular
Dysmenorrhea
Reviewed
Institution
01/24/2015
01/20/2015
5
eumenorrhea
normal
Cordara, Cameron
GNU SOLIDARIO HOSPITAL
tryton://health.gnusolidario.org:8000/health28rc1/model/gnuhealth.patient/1;views=%5B223%2C+224%5D

DESIGNING THE RIGHT SYSTEM

Radiology example:

Radiologists do not like systems that just automate the simple cases. They can do this themselves. They do not want to be replaced.

To be useful, a system must help in difficult cases with missing information. It needs to provide explanations.

Explanations do not just explain a single prediction, but also how the system works, what information it has access to, how it is calibrated, what limitations it has, ...

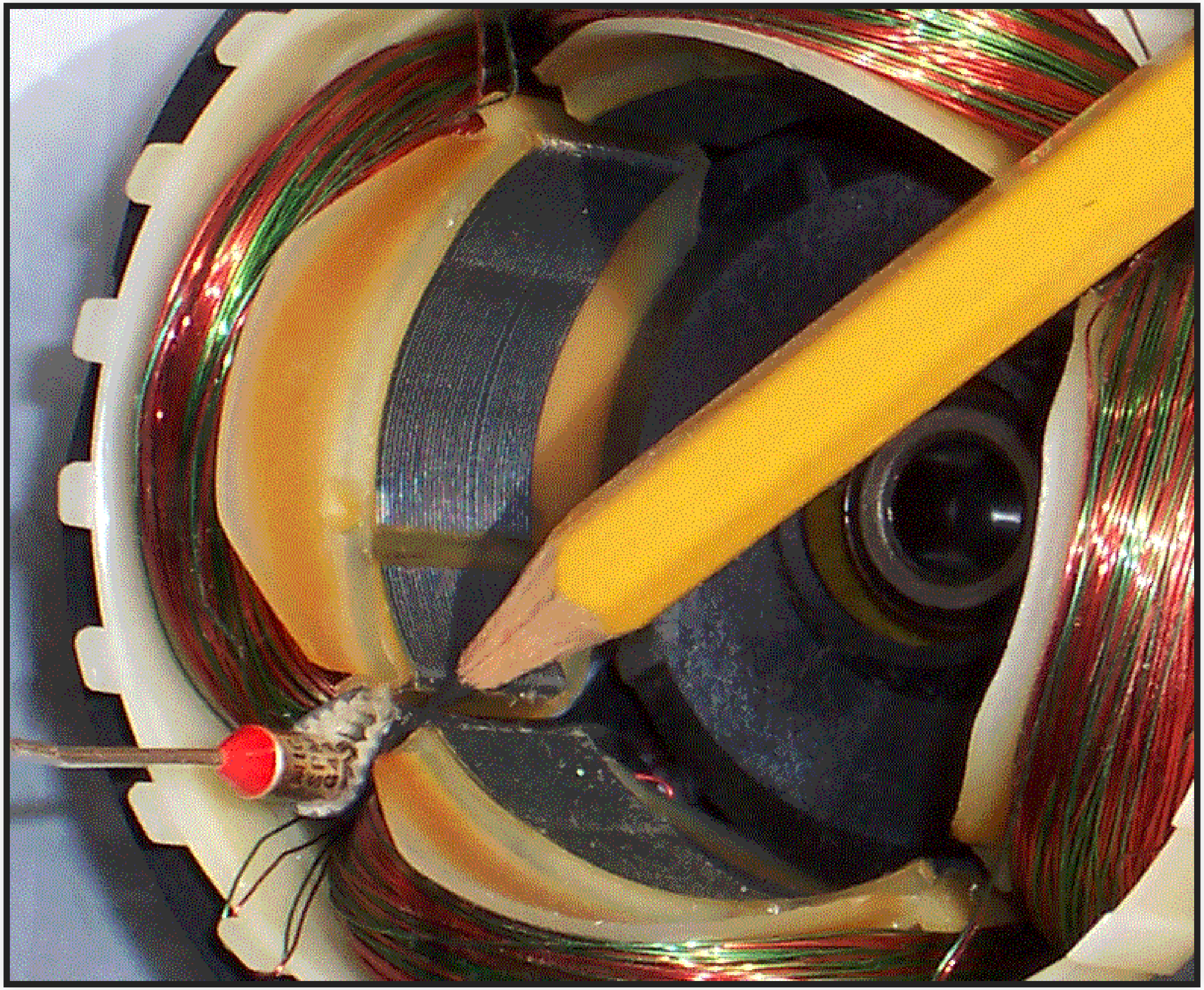
LIVING WITH MISTAKES

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 GOALS

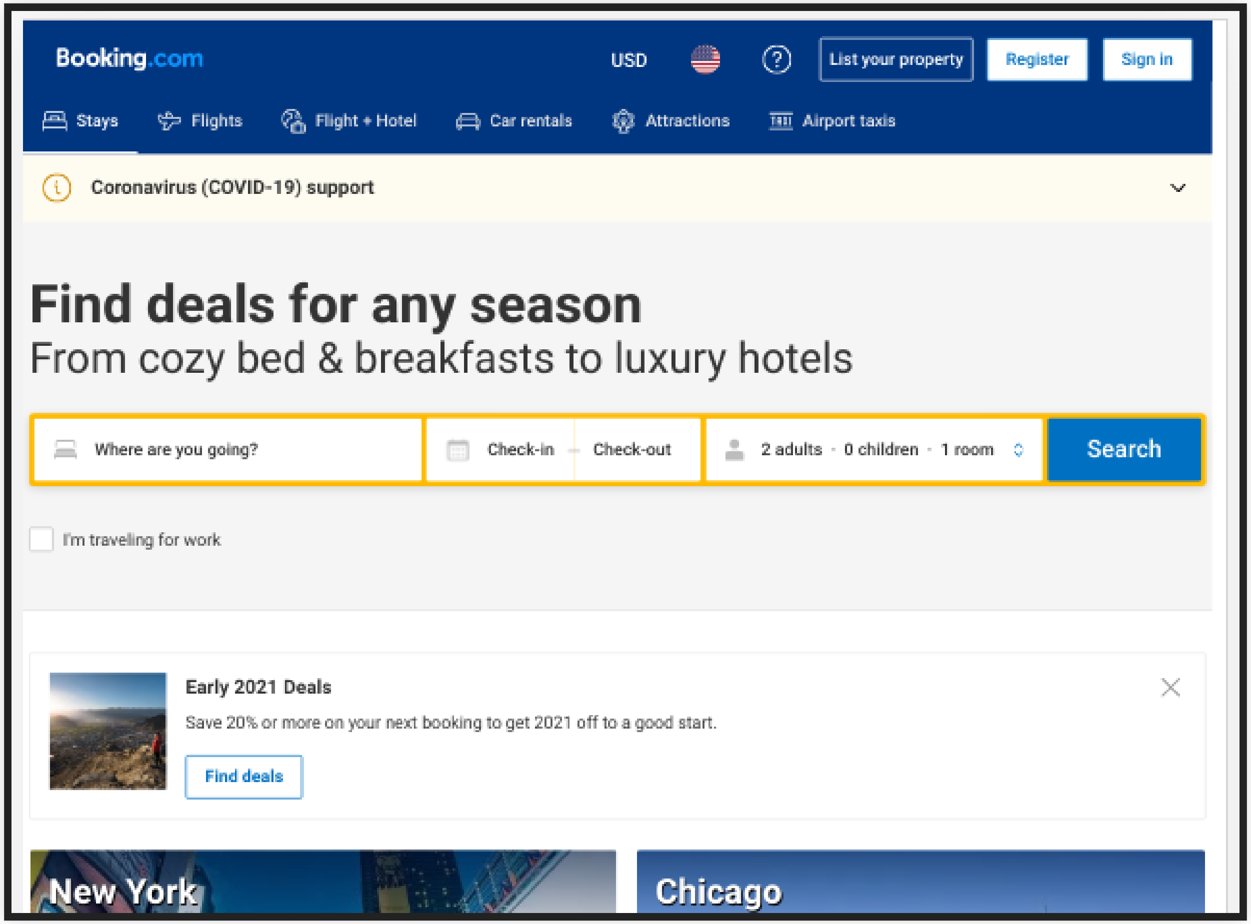
- **Accuracy**
- Fairness
- Low latency
- Low training cost

SYSTEM GOALS

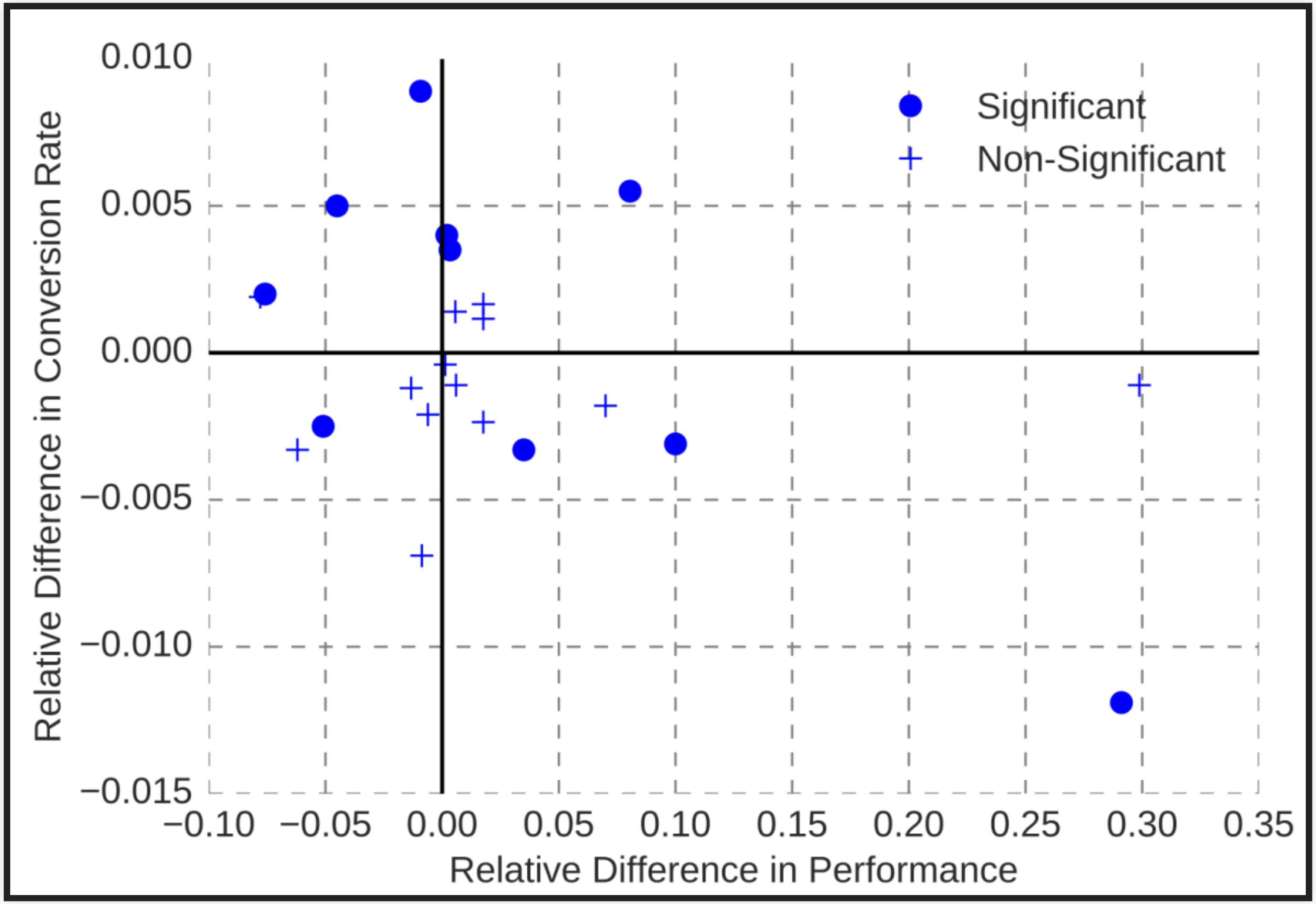
- **Maximizing sales**
- Maximizing community growth
- Retaining customers
- Maximizing engagement time

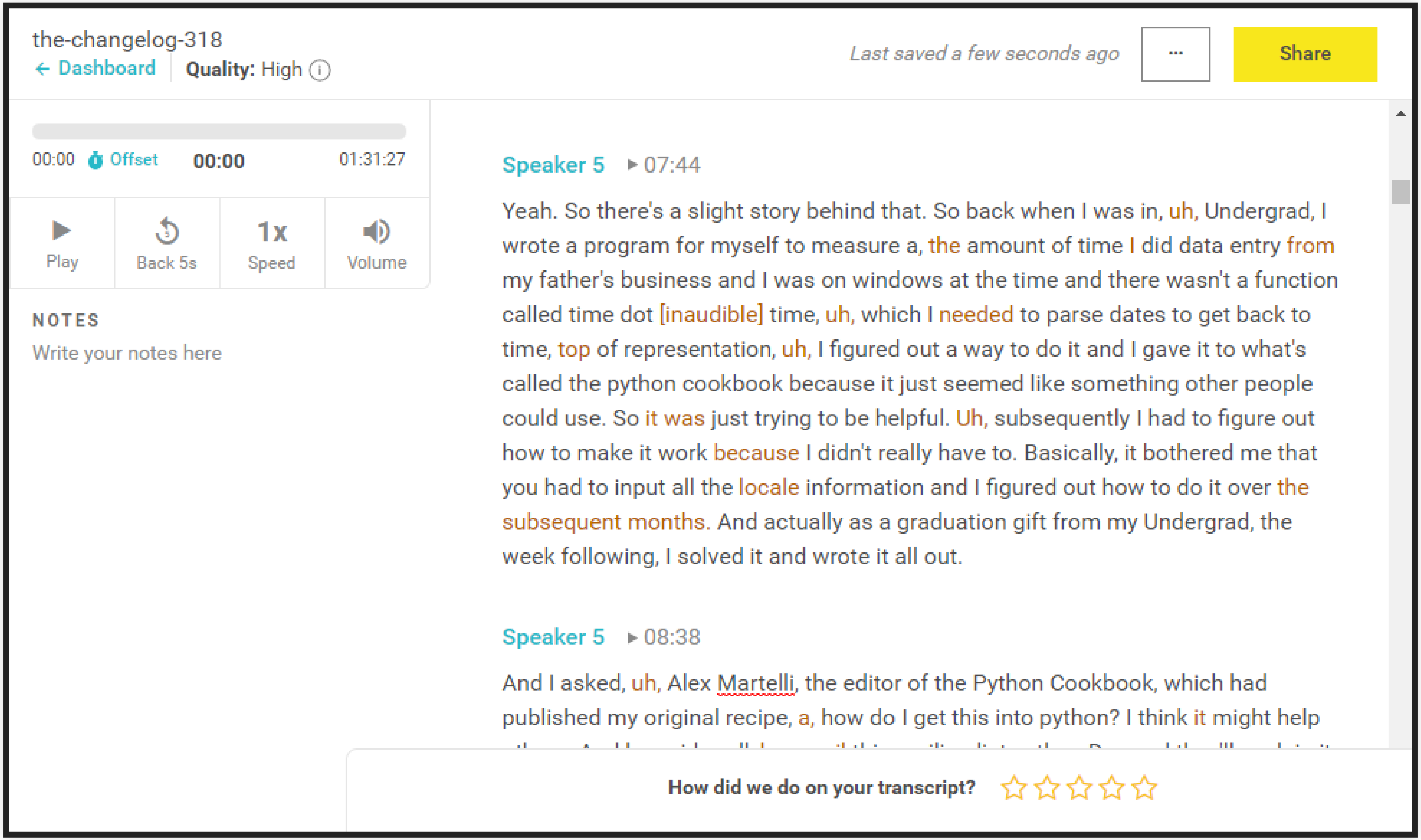
A better model will, hopefully, support system goals better

MODEL ACCURACY VS SYSTEM GOALS



MODEL ACCURACY VS SYSTEM GOALS




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

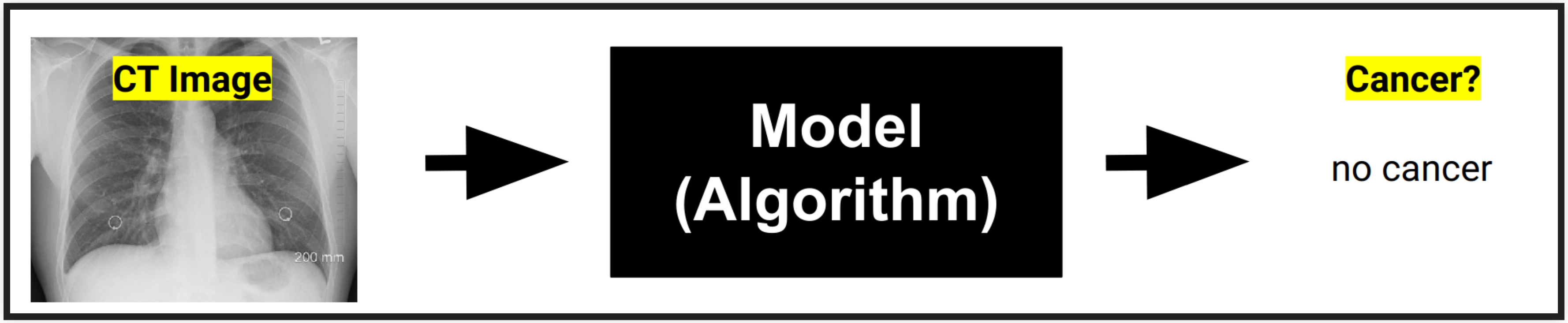
Can evaluate system goals, not just model accuracy

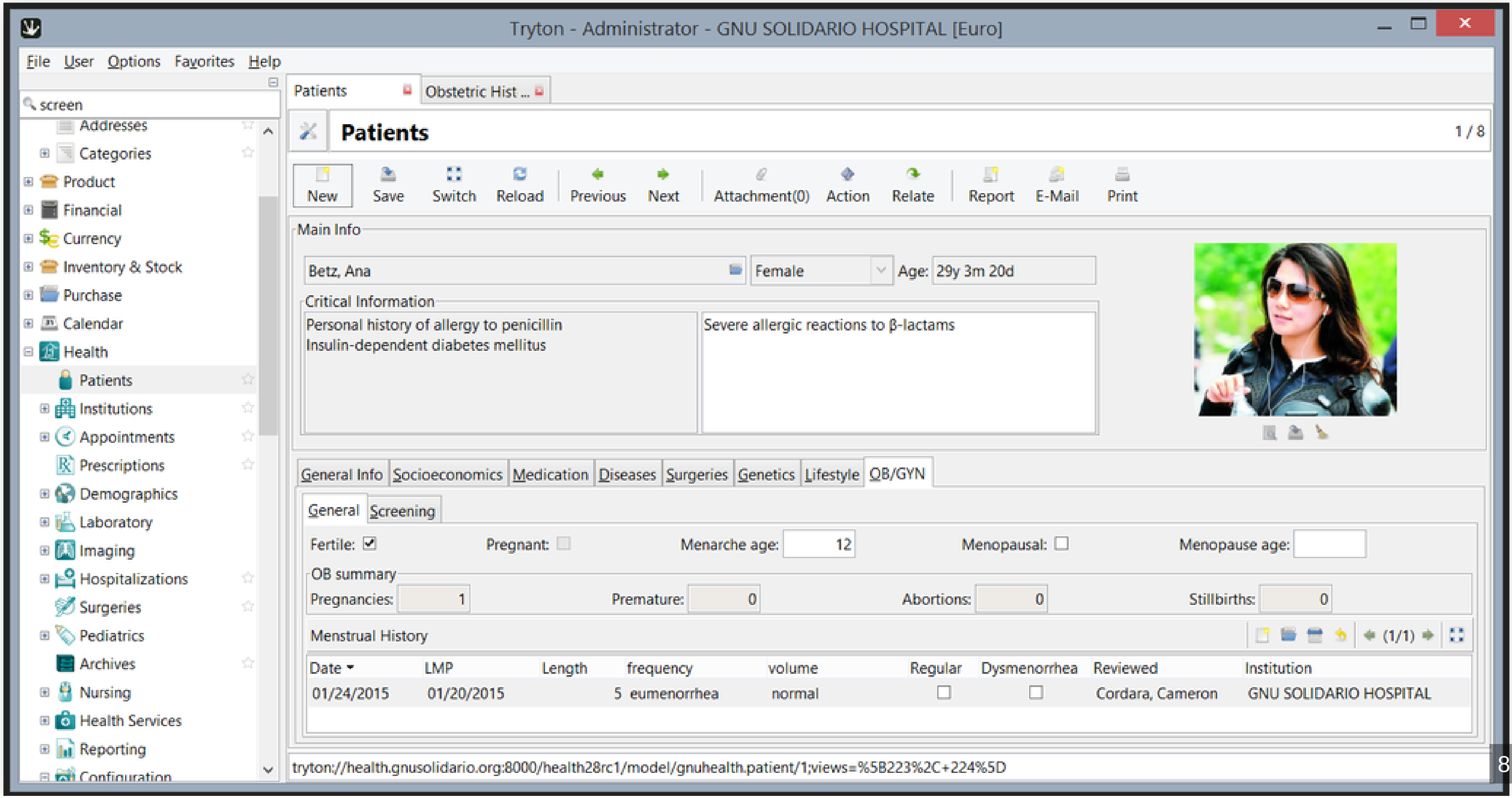
Monitoring performance over time, canary releases

Finding and debugging common mistakes

Experimentation with A/B tests

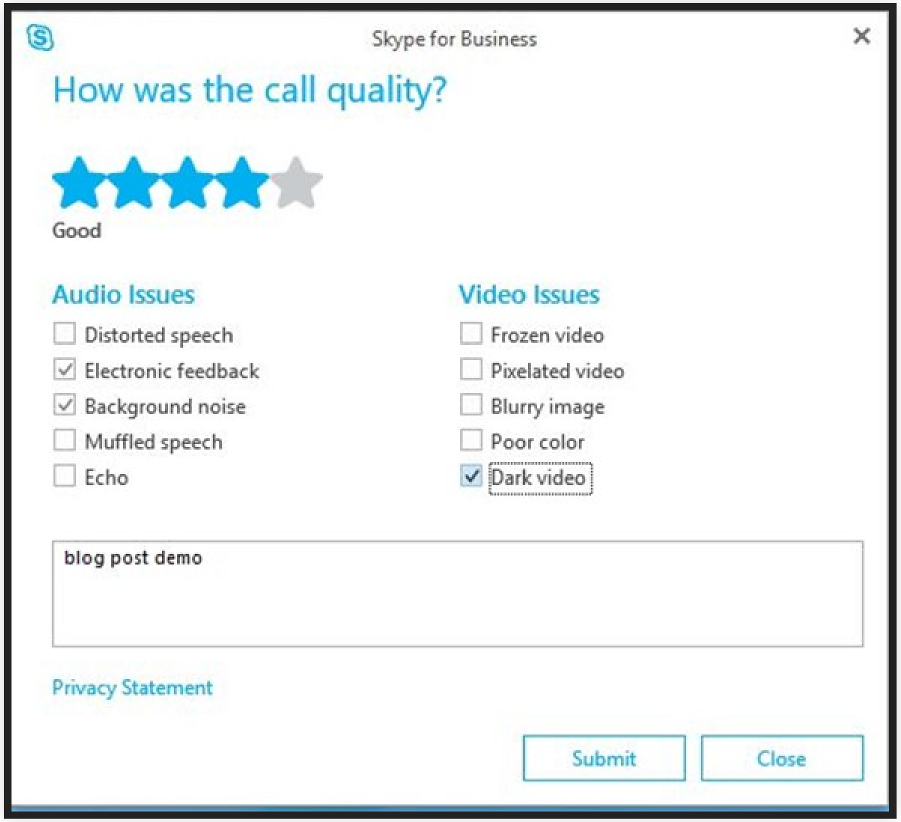
MONITORING MODEL/SYSTEM QUALITY IN PRODUCTION?

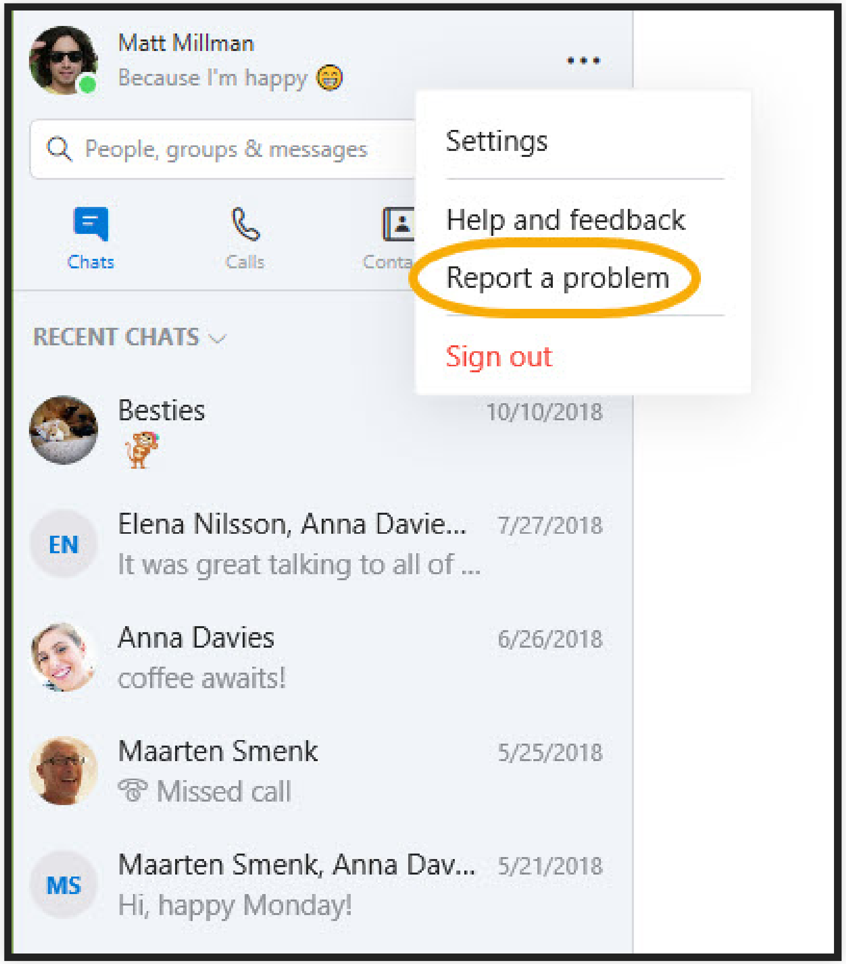

CT Image
200 mm
Model
(Algorithm)
Cancer?
no cancer


Tryton - Administrator - GNU SOLIDARIO HOSPITAL [Euro]
File User Options Favorites Help
screen
Addresses
Categories
Product
Financial
Currency
Inventory & Stock
Purchase
Calendar
Health
Patients
Institutions
Appointments
Prescriptions
Demographics
Laboratory
Imaging
Hospitalizations
Surgeries
Pediatrics
Archives
Nursing
Health Services
Reporting
Configuration
Patients
Obstetric Hist ...
Patients
1 / 8
New
Save
Switch
Reload
Previous
Next
Attachment(0)
Action
Relate
Report
E-Mail
Print
Main Info
Betz, Ana
Female
Age: 29y 3m 20d
Critical Information
Personal history of allergy to penicillin
Insulin-dependent diabetes mellitus
Severe allergic reactions to β-lactams
General Info
Socioeconomics
Medication
Diseases
Surgeries
Genetics
Lifestyle
OB/GYN
General
Screening
Fertile:
Pregnant:
Menarche age: 12
Menopausal:
Menopause age:
OB summary
Pregnancies: 1
Premature: 0
Abortions: 0
Stillbirths: 0
Menstrual History
(1/1)
Date
LMP
Length
frequency
volume
Regular
Dysmenorrhea
Reviewed
Institution
01/24/2015
01/20/2015
5
eumenorrhea
normal
Cordara, Cameron
GNU SOLIDARIO HOSPITAL
tryton://health.gnusolidario.org:8000/health28rc1/model/gnuhealth.patient/1;views=%5B223%2C+224%5D

KEY DESIGN CHALLENGE: TELEMETRY

- Identify model mistakes in production (“what would have been the right prediction?”)
  - 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?
- Measure system goals in production ("conversion rate")


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

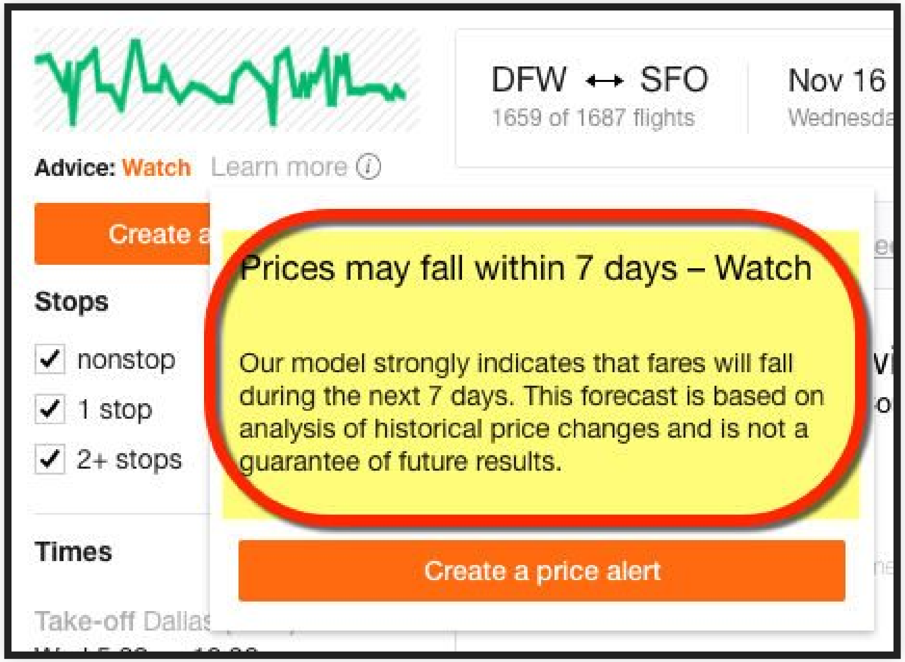

Matt Millman
Because I'm happy
People, groups & messages
Chats
Calls
Settings
Help and feedback
Report a problem
Sign out
RECENT CHATS
Besties
10/10/2018
Elena Nilsson, Anna Davie...
7/27/2018
It was great talking to all of ...
Anna Davies
6/26/2018
coffee awaits!
Maarten Smenk
5/25/2018
Missed call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

Speaker notes

Expect only sparse feedback and expect negative feedback over-proportionally

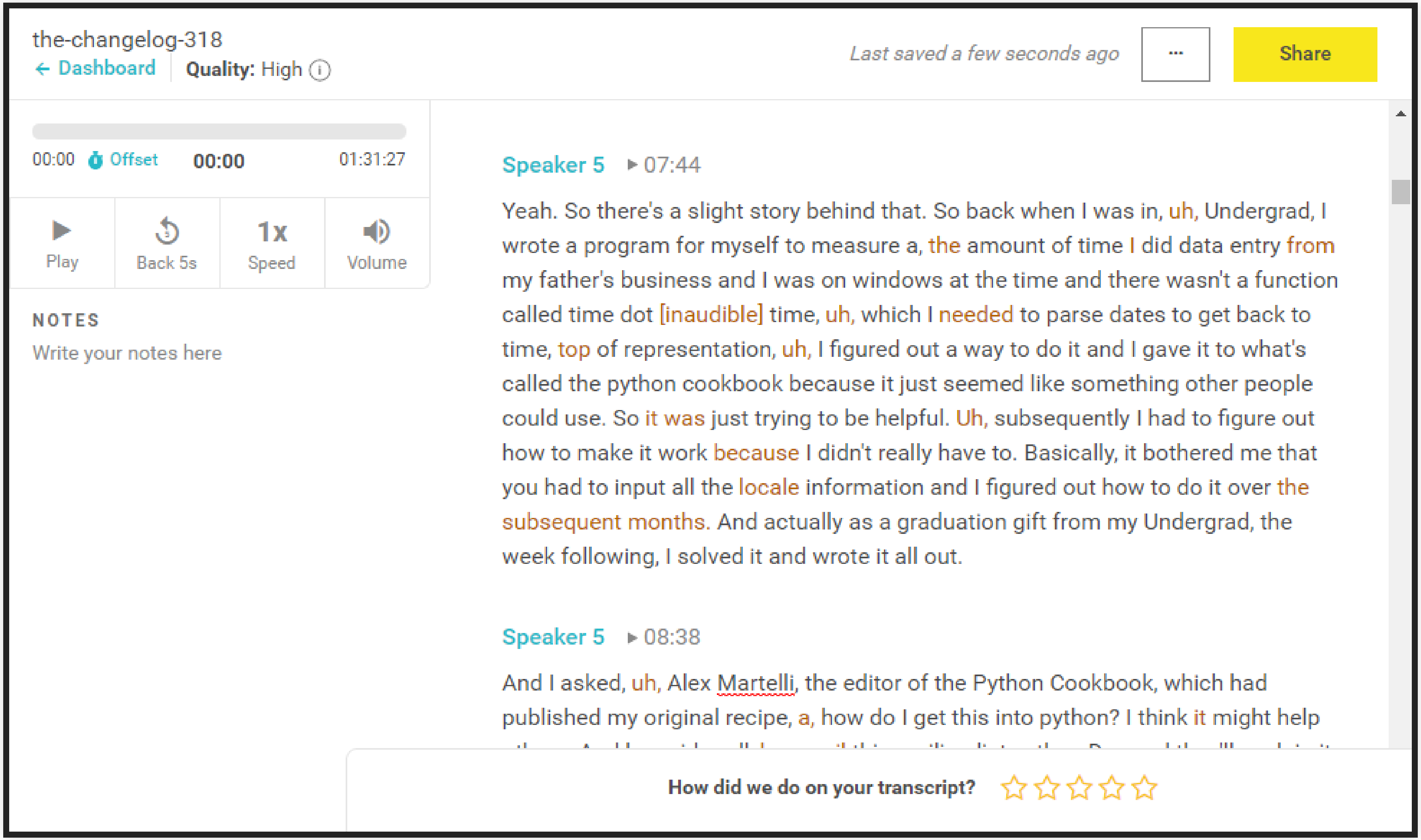
MANUALLY LABEL PRODUCTION SAMPLES

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

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

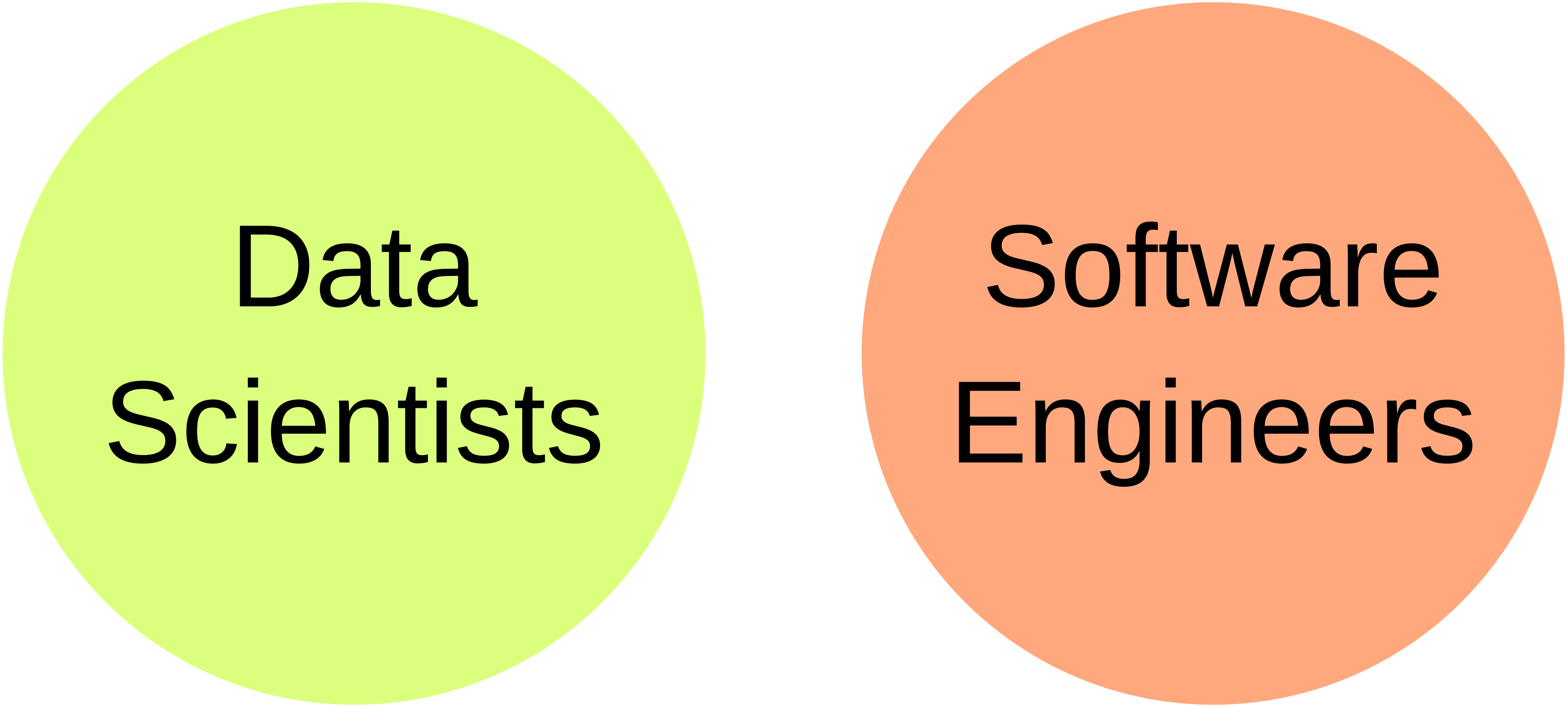
By Olivia Tambini 21 days ago Digital Home

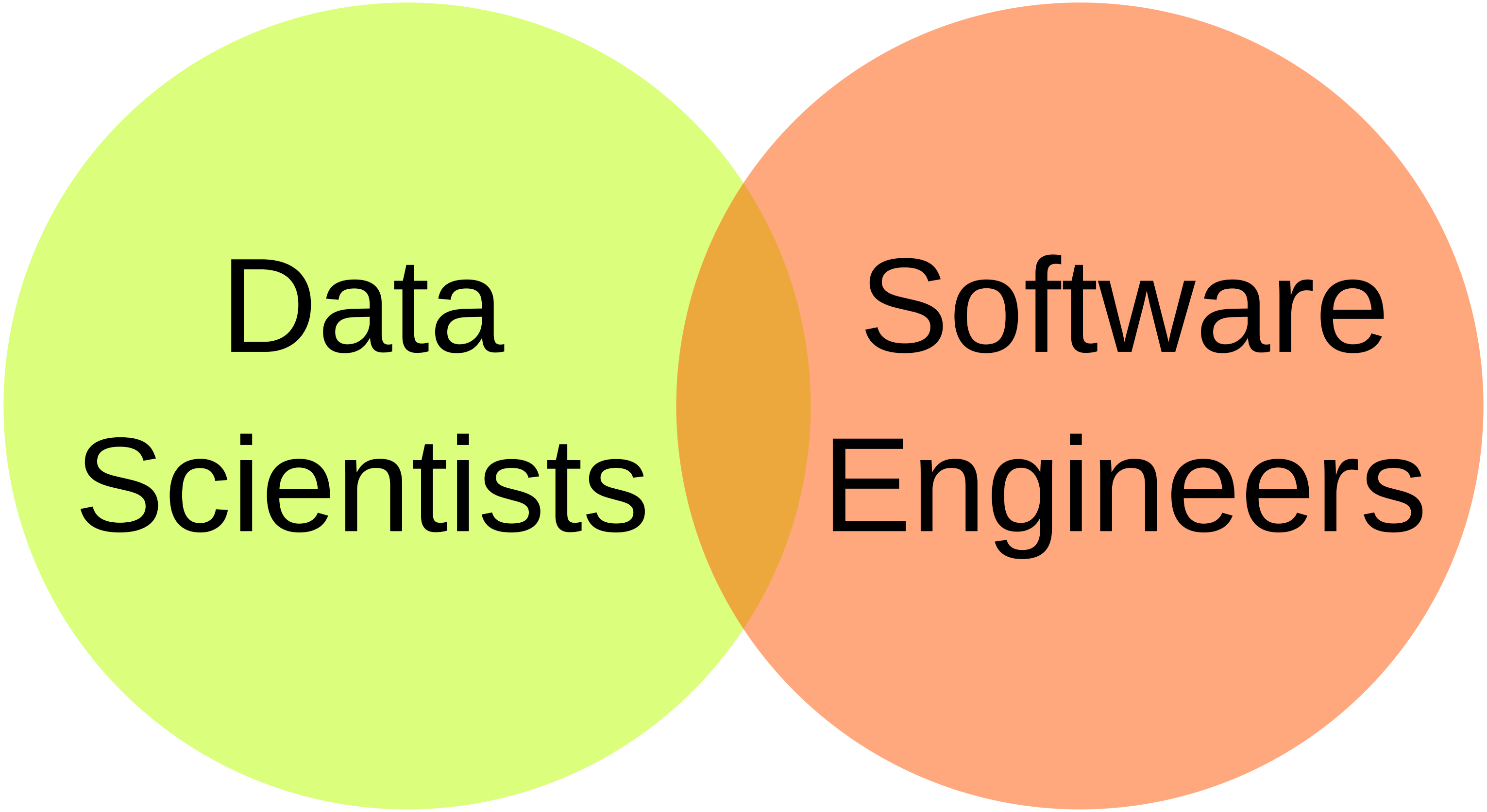
A letter from Amazon reveals all

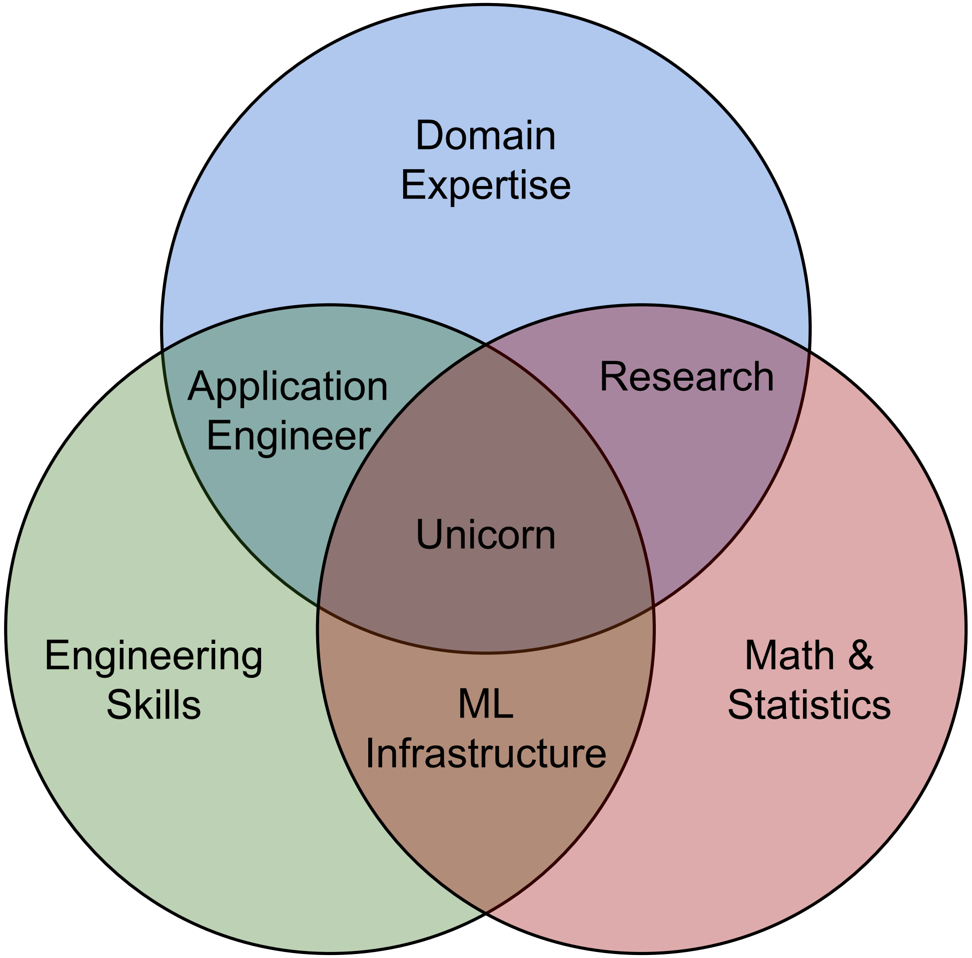
RECAP: FROM MODEL TO SYSTEM

- Plan the entire system, not just a model
- Requirements engineering + UX for the system is important
- Identify relevant qualities beyond accuracy, plan and test models accordingly
- Design for telemetry

PART 3: INTERDISCIPLINARY TEAMS


Data
Scientists
Software
Engineers


Data
Scientists
Software
Engineers



Based on Ryan Orban. Bridging the Gap Between Data Science & Engineer: Building High-Performance Teams. 2016

T-SHAPED PEOPLE

Broad-range generalist + Deep expertise

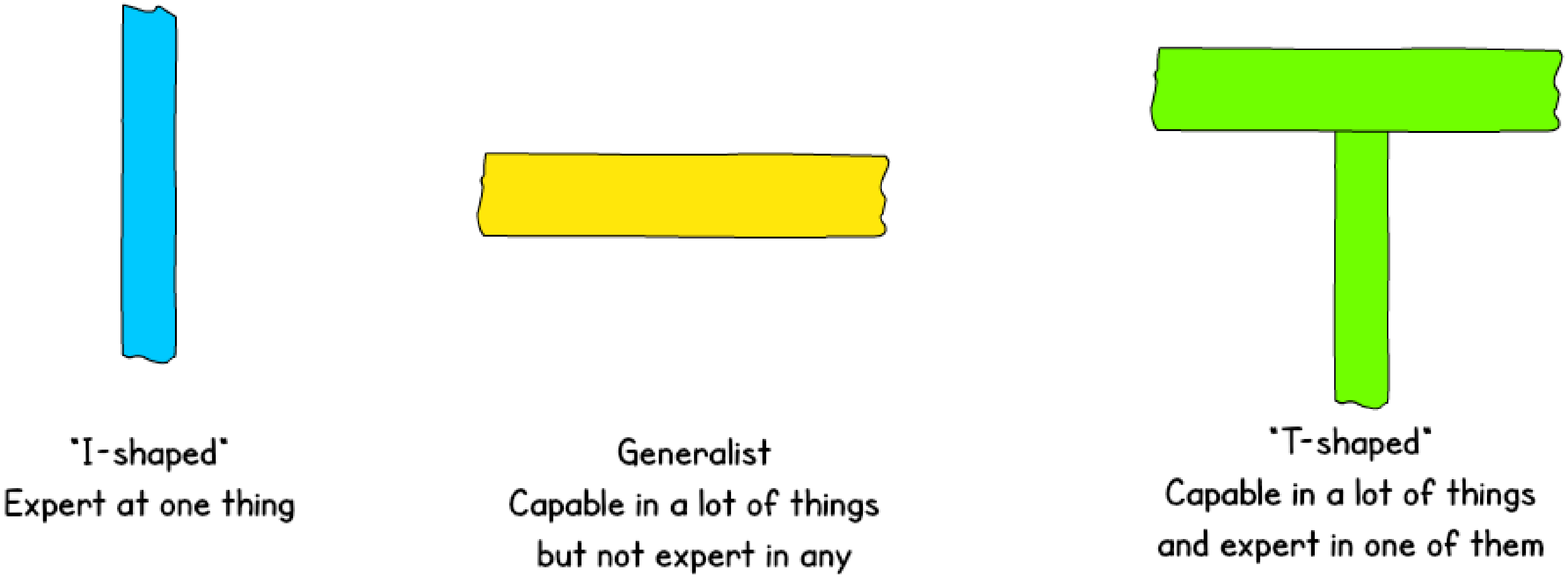
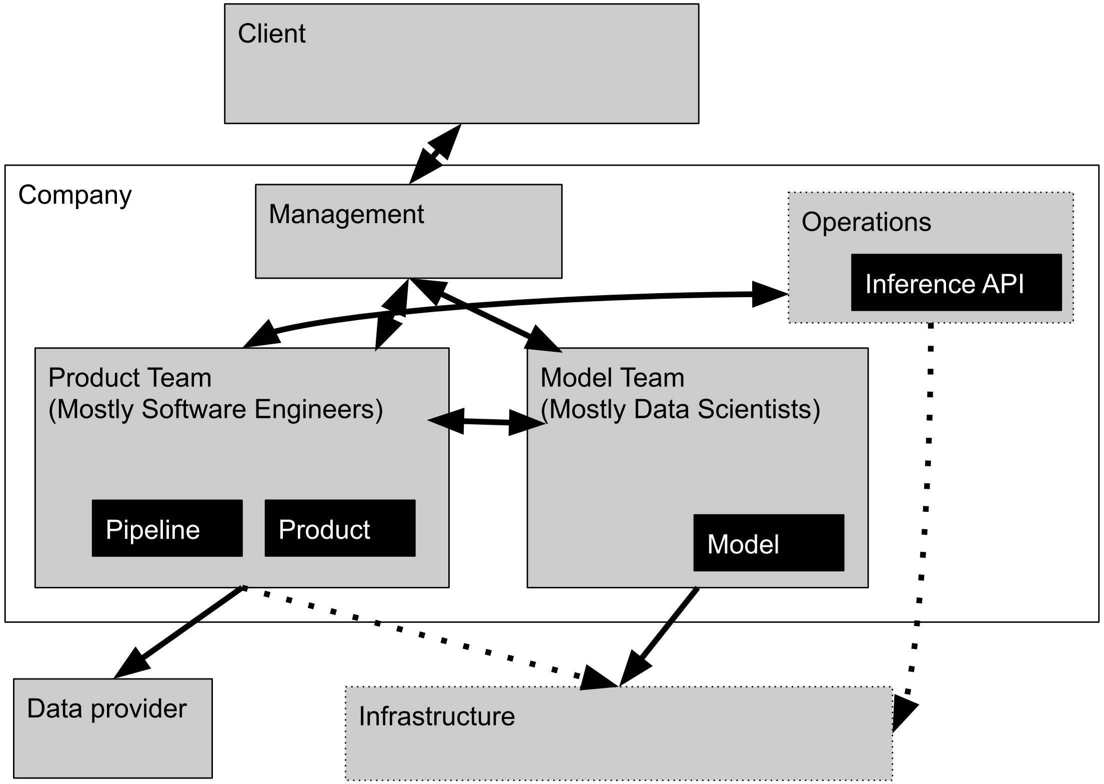


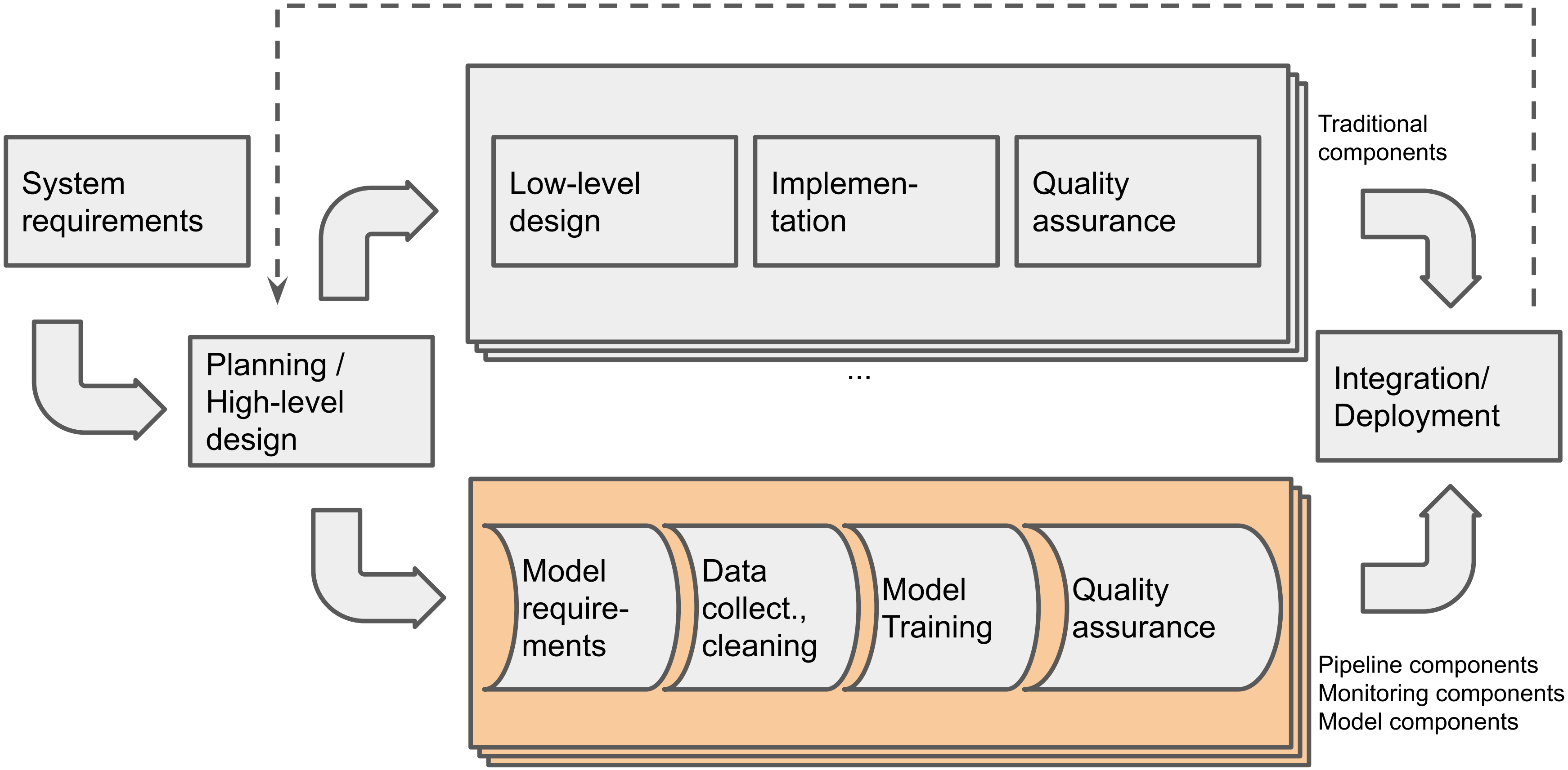
Figure: Jason Yip. Why T-shaped people?. 2018

SILOING IS BAD

- *We do not have clean interfaces between ML and non-ML components*
- Divide and conquer and information hiding *on hard mode*
- Foster collaboration among teams, mix teams, avoid silos

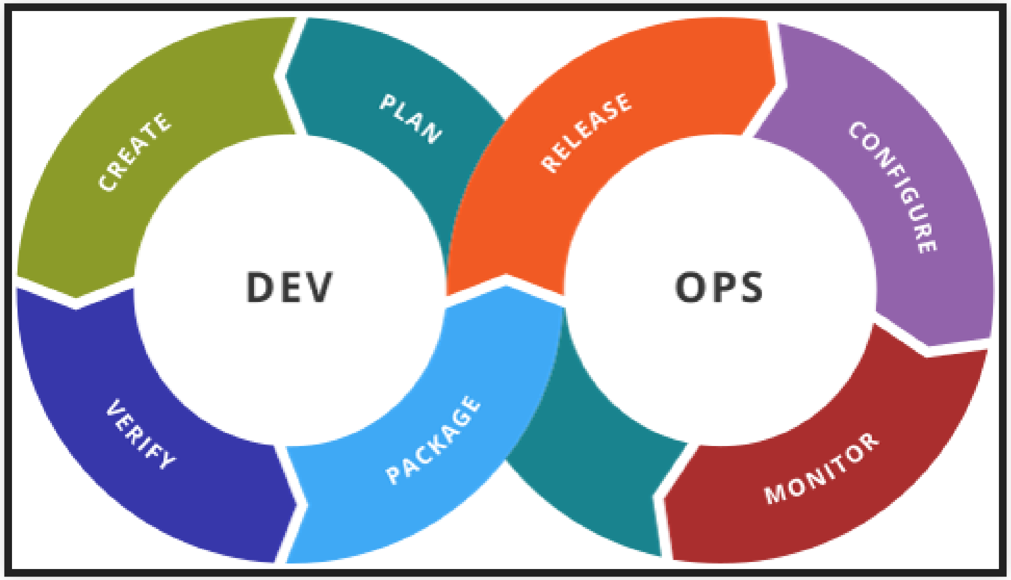


MODEL FIRST OR SYSTEM FIRST?



More details: Christian Kaestner. On the process for building software with ML components. Medium 2020

LET'S LEARN FROM DEVOPS



Distinct roles and expertise, but joint responsibilities, joint tooling

TOWARD BETTER ML-SYSTEMS ENGINEERING

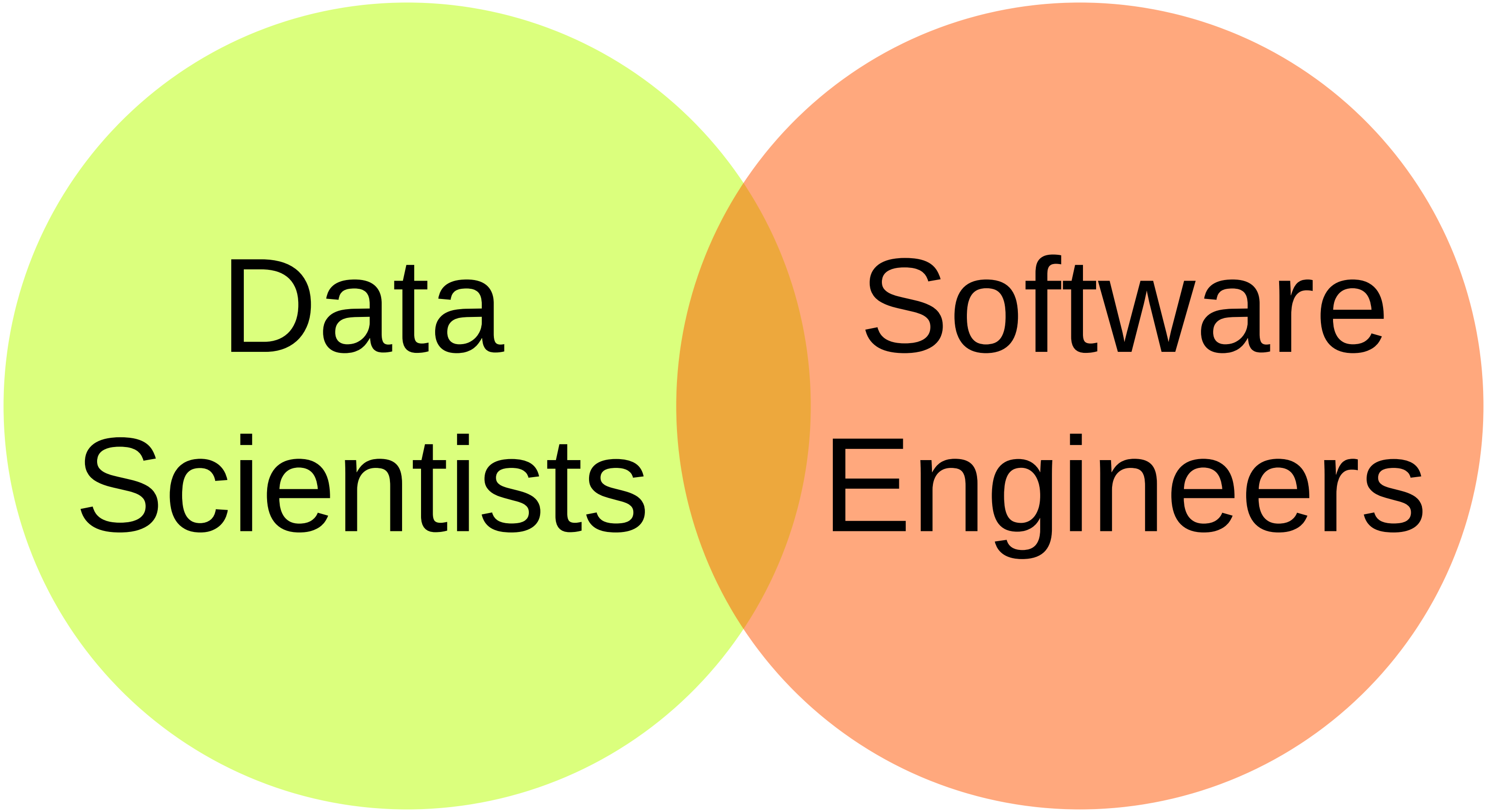
Interdisciplinary teams, split expertise, but joint responsibilities

Joint vocabulary and tools

Foster system thinking

Awareness of production quality concerns

Perform risk + hazard analysis


Data
Scientists
Software
Engineers

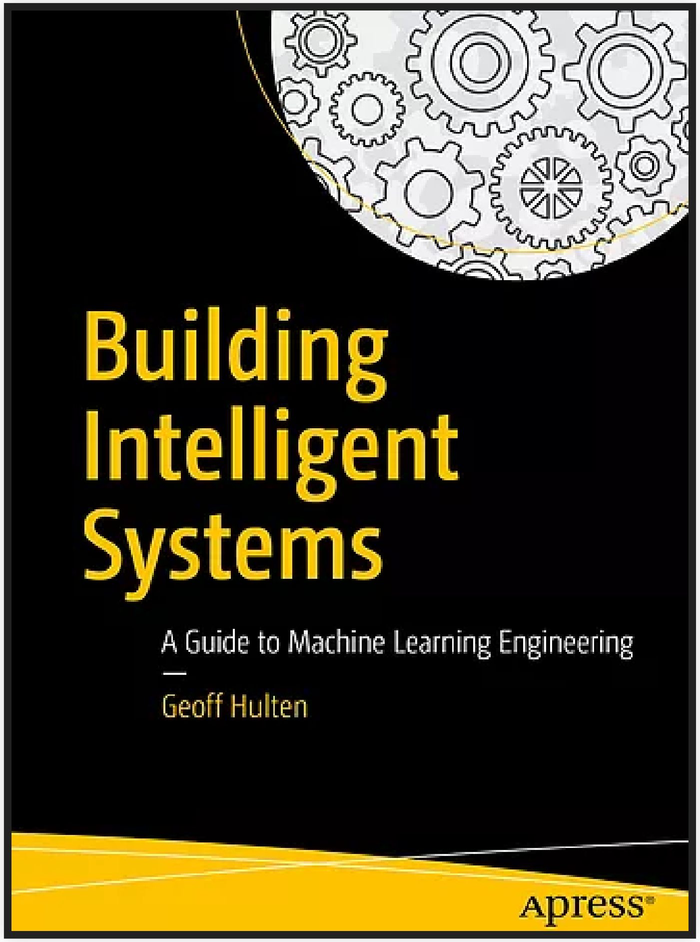
READINGS

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

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

Essays and book chapters:
https://ckaestne.medium.com/

Best book on the topic out there:



MOVING MACHINE LEARNING PROJECTS INTO PRODUCTION WITH INTERDISCIPLINARY TEAMS

- Building, operating, and maintaining systems with ML component
- Data scientists and software engineers have different expertise, both needed
- Need to consider entire system, not just model, e.g. in testing:
    - Model accuracy, blackbox testing, test automation
    - Testing and automating the entire ML pipeline
    - Understanding and testing system qualities
    - Design for mistakes
    - Testing in production with telemetry
- Interdisciplinary teams, T-shaped people, and joint vocabulary

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