ECE444: Software Engineering

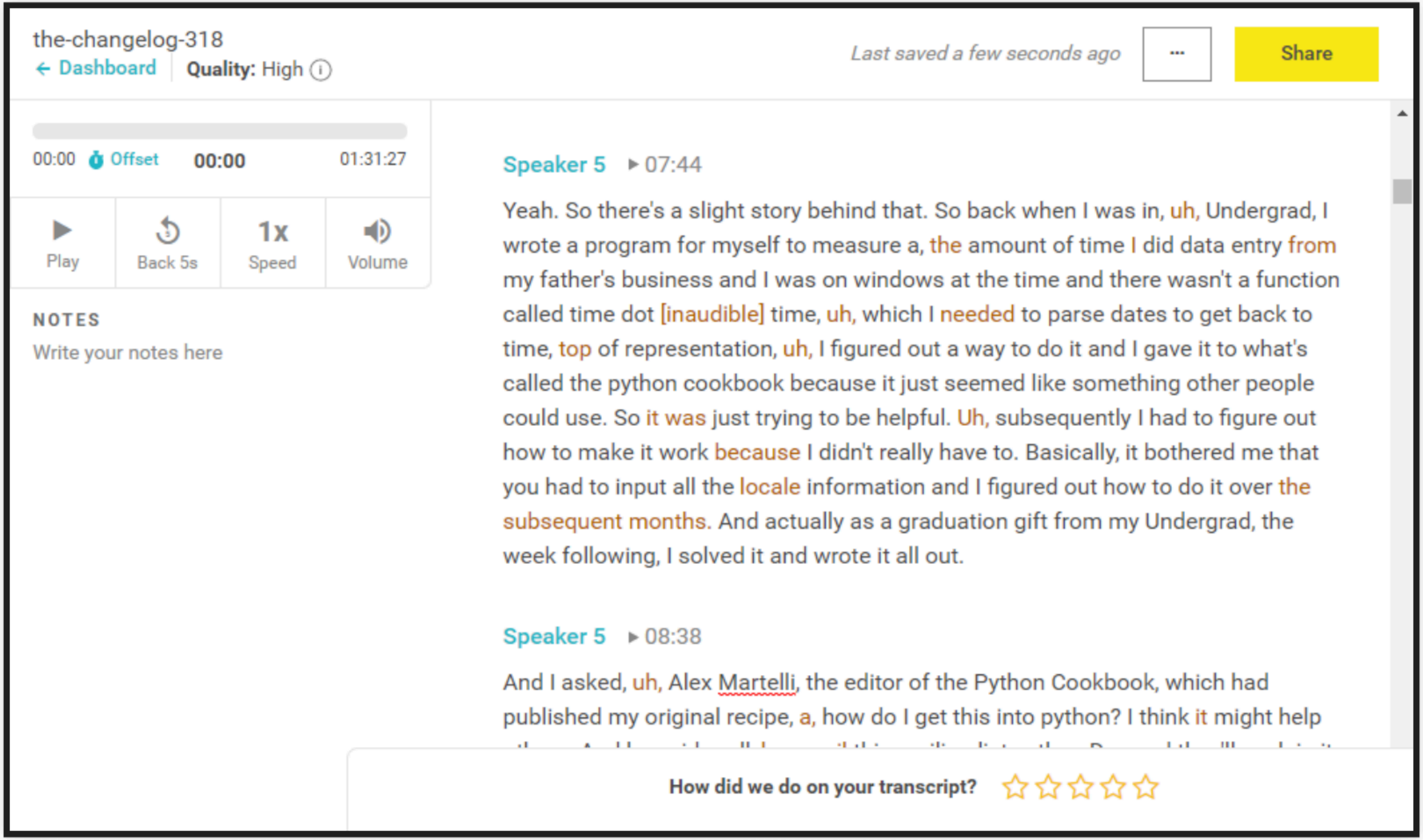
Software Engineering for AI/ML

Shurui Zhou

The Edward S. Rogers Sr. Department
of Electrical & Computer Engineering
UNIVERSITY OF TORONTO

Learning Goals

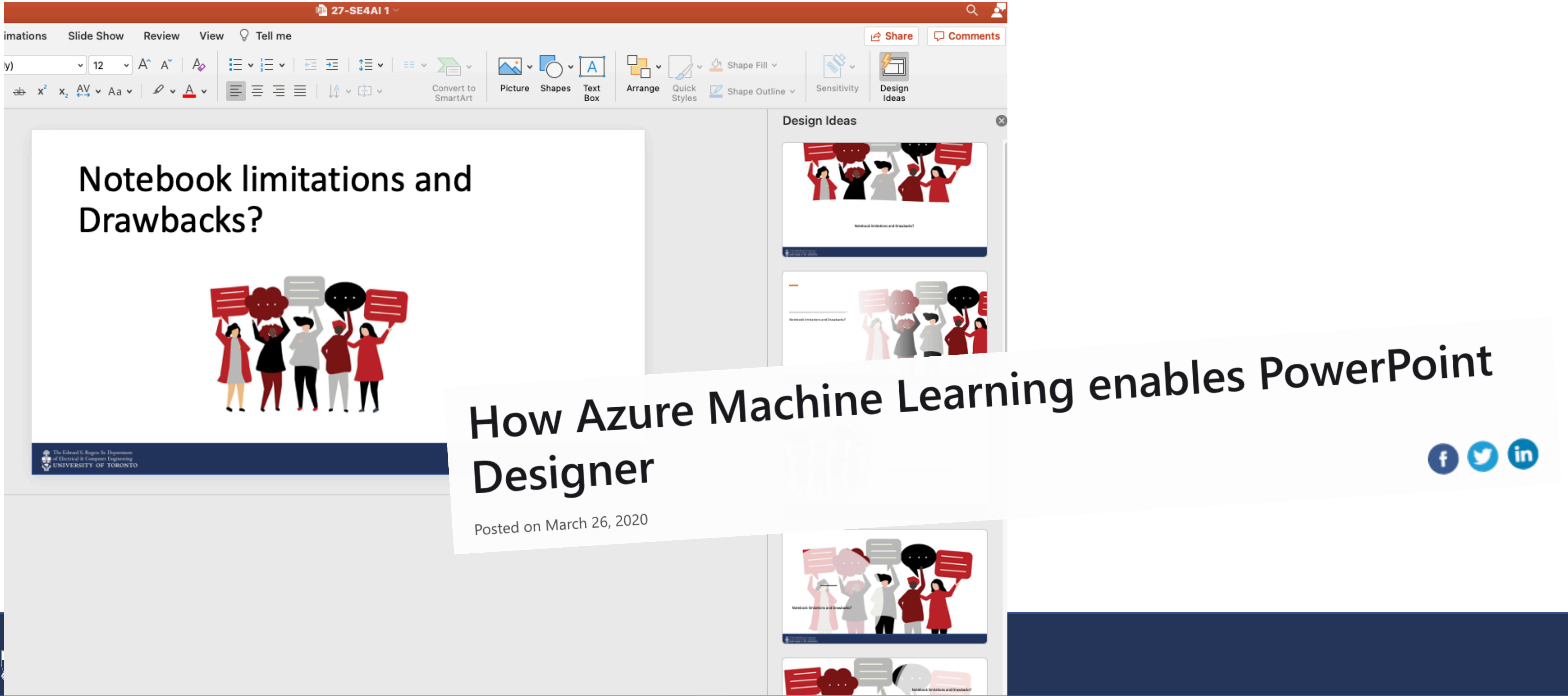
- Understand how AI components are parts of larger systems
- Illustrate the challenges in engineering an AI-enabled system beyond accuracy
- Explain the role of specifications and their lack in machine learning and the relationship to deductive and inductive reasoning
- Summarize the respective goals and challenges of software engineers vs data scientists



https://www.temi.com/

Microsoft PowerPoint

https://azure.microsoft.com/en-us/blog/how-azure-machine-learning-enables-powerpoint-designer/



Fall Detection Devices

How fall detection is moving beyond the pendant

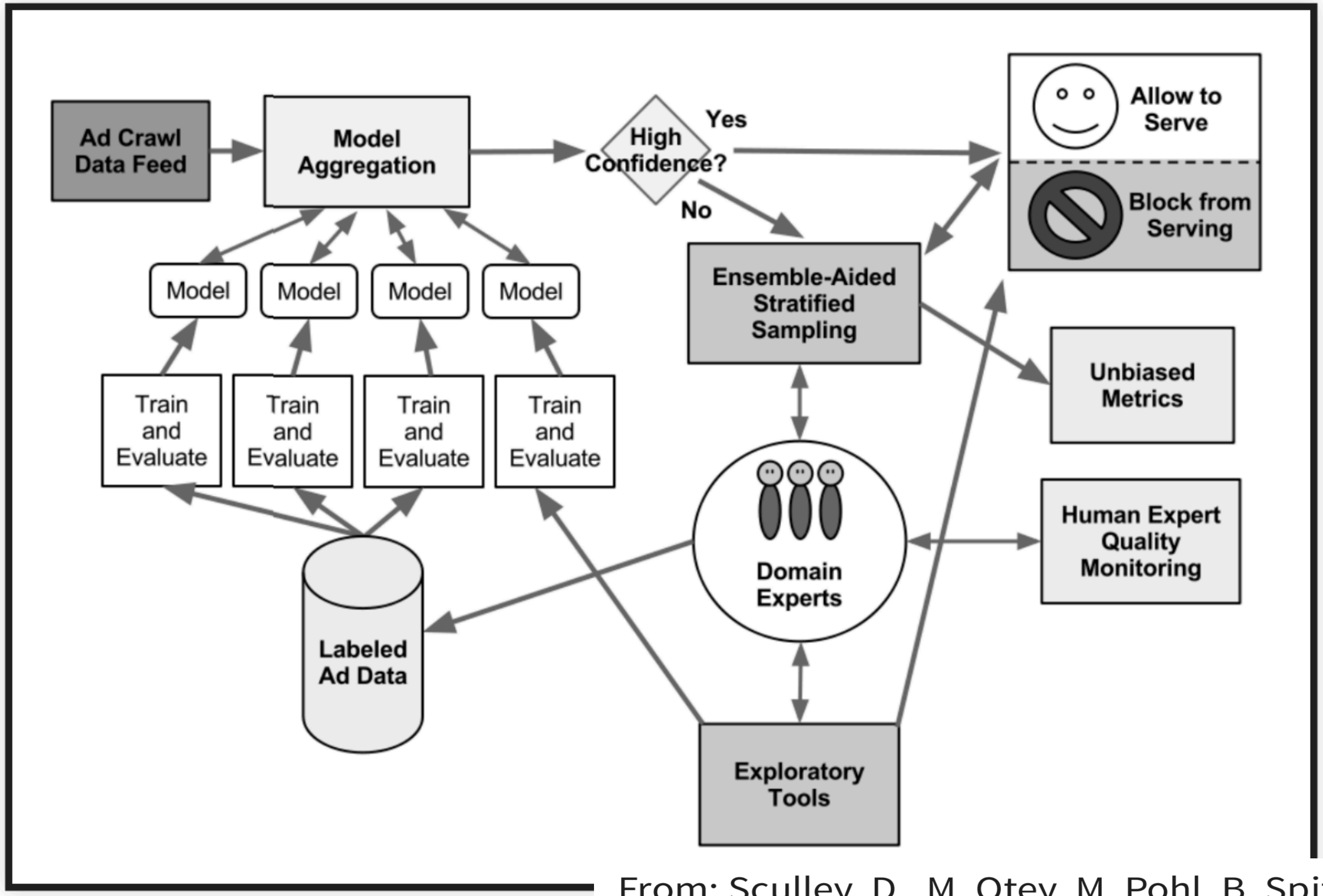
Digital health innovators look to the wrist, the ears and the wall for new ways to keep seniors safe.

By **Jonah Comstock** | April 19, 2019 | 04:22 pm

SHARE Share 519

https://www.mobihealthnews.com/content/how-fall-detection-moving-beyond-pendant

Google Add Fraud Detection



From: Sculley, D., M. Otey, M. Pohl, B. Spitznagel, J. Hainsworth, and Y. Zhou. Detecting Adversarial Advertisements in the Wild. In Proc. KDD, 2011.

Recidivism Detection

```
IF age between 18–20 and sex is male THEN predict arrest
ELSE IF age between 21–23 and 2–3 prior offenses THEN predict ar
ELSE IF more than three priors THEN predict arrest
ELSE predict no arrest
```

Read more: Julia Angwin, Jeff Larson, Surya Mattu and Lauren Kirchner. "Machine Bias." ProPublica 2016

Many more examples

- Product recommendations on Amazon
- Surge price calculation for Uber
- Inventory planning in Walmart
- Search for new oil fields by Shell
- Adaptive cruise control in a car
- Smart app suggestion in Android
- Fashion trends prediction with social media data
- Suggesting whom to talk to in a presidential campain
- Tracking and predicting infections in a pandemic
- Adaptively reacting to network issues by a cell phone provider
- Matching players in a computer game by skill
- ...

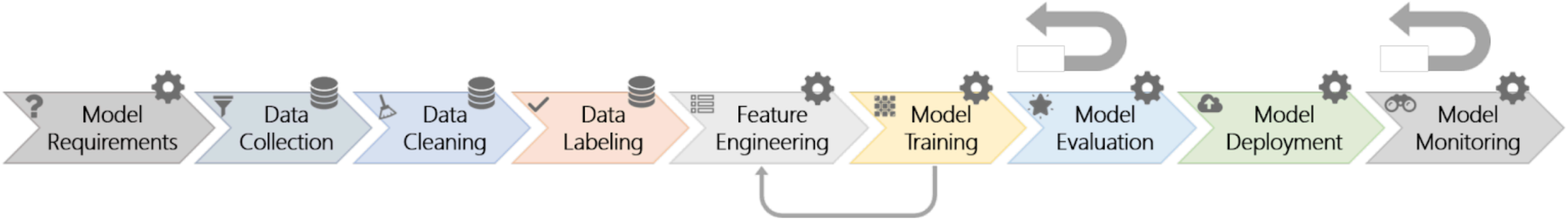
- Some for end users, some for employees, some for expert users
- Big and small components of a larger system

Software Engineering and ML

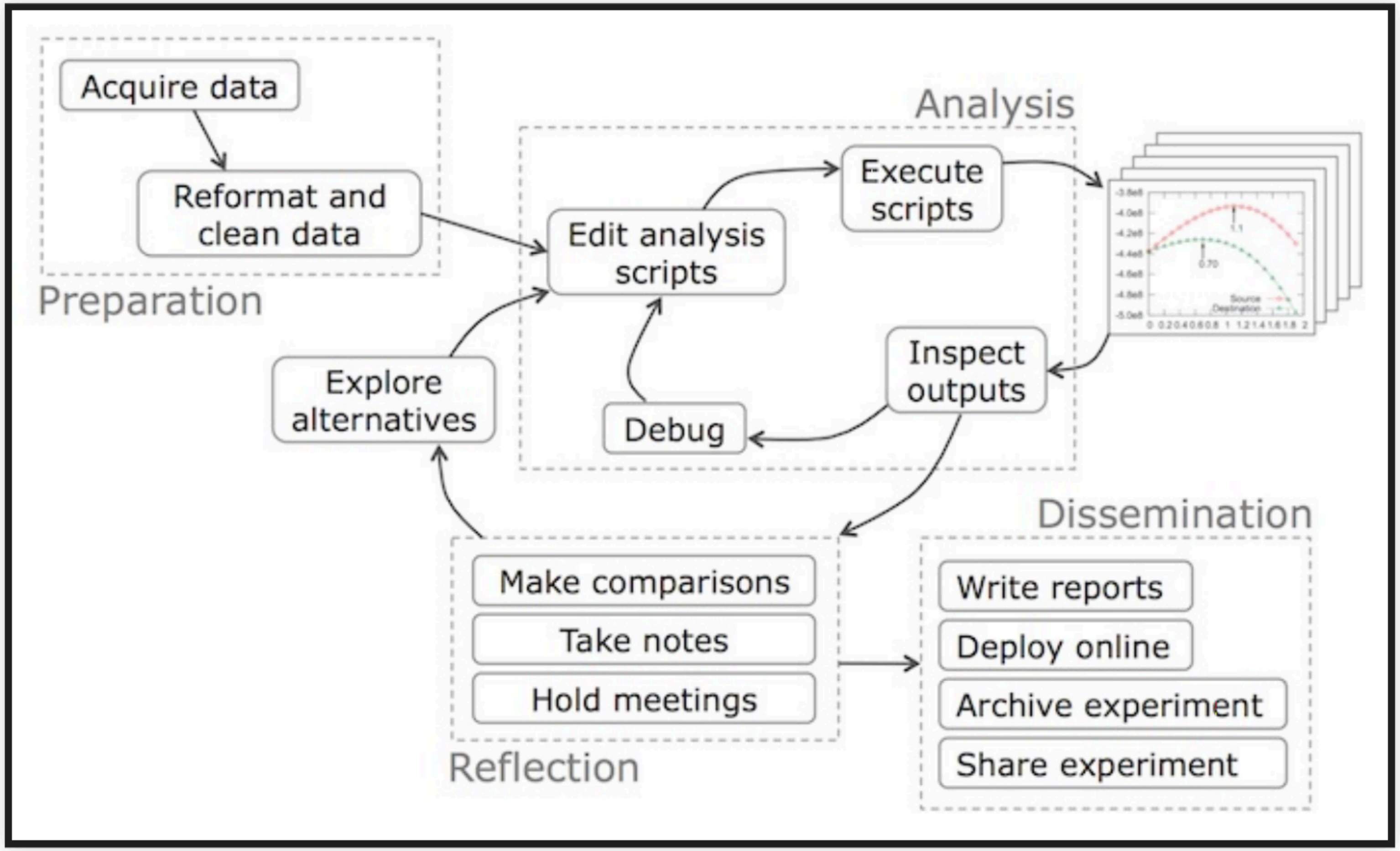
ML Development

- Observation
- Hypothesis
- Predict
- Test
- Reject or Refine Hypothesis

Microsoft's view of Software Engineering for ML

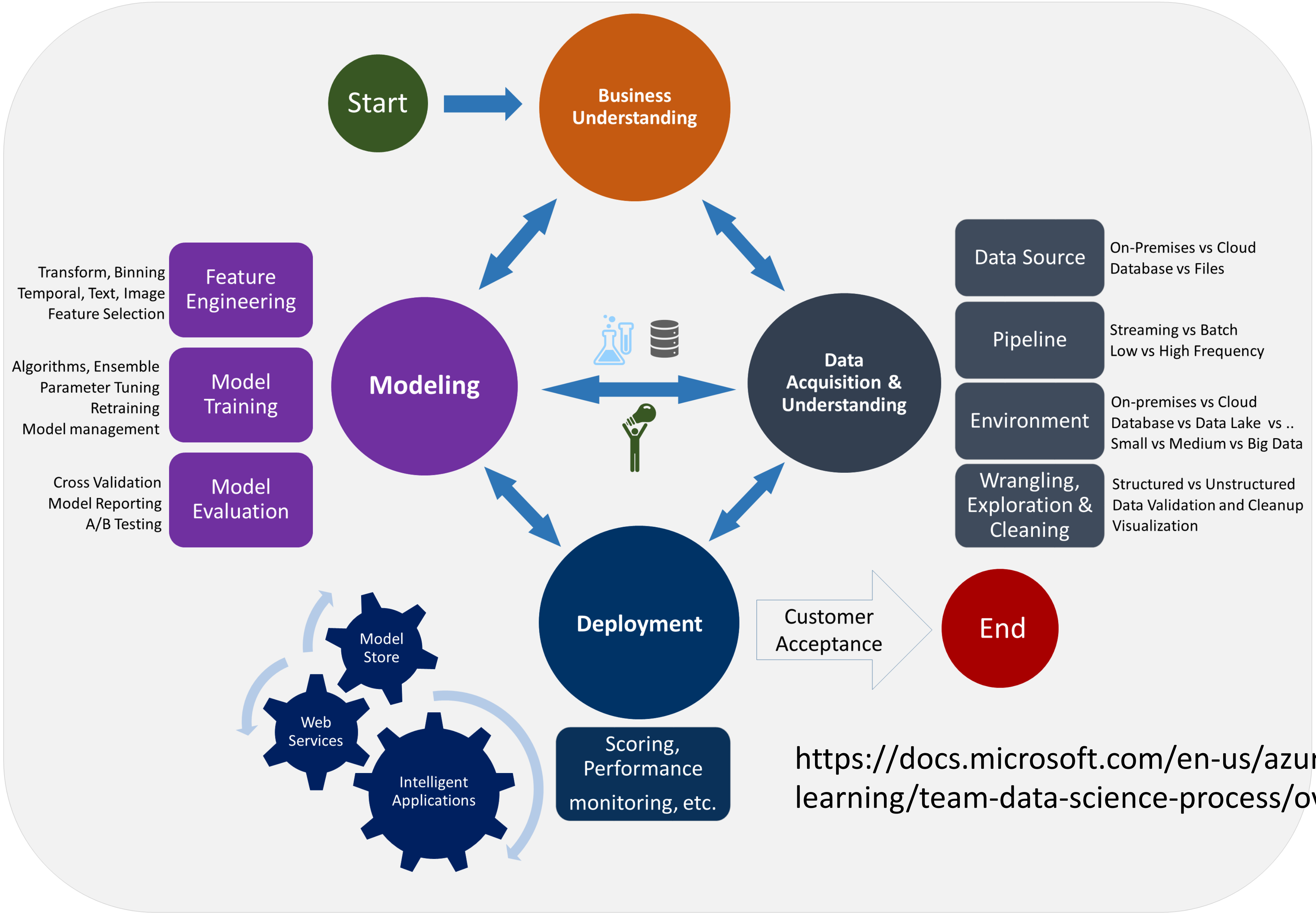


Data science is iterative and exploratory



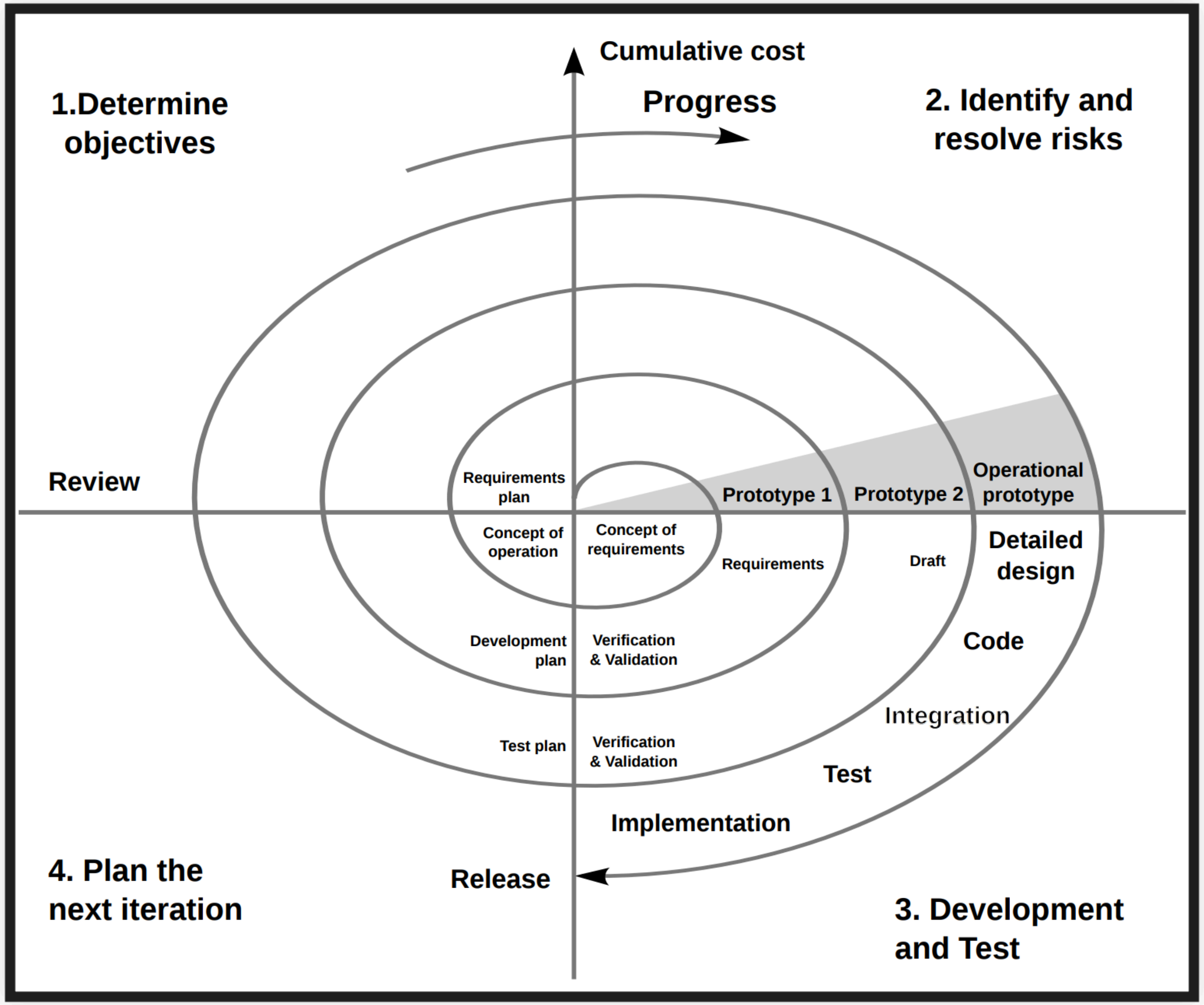
https://cacm.acm.org/blogs/blog-cacm/169199-data-science-workflow-overview-and-challenges/fulltext

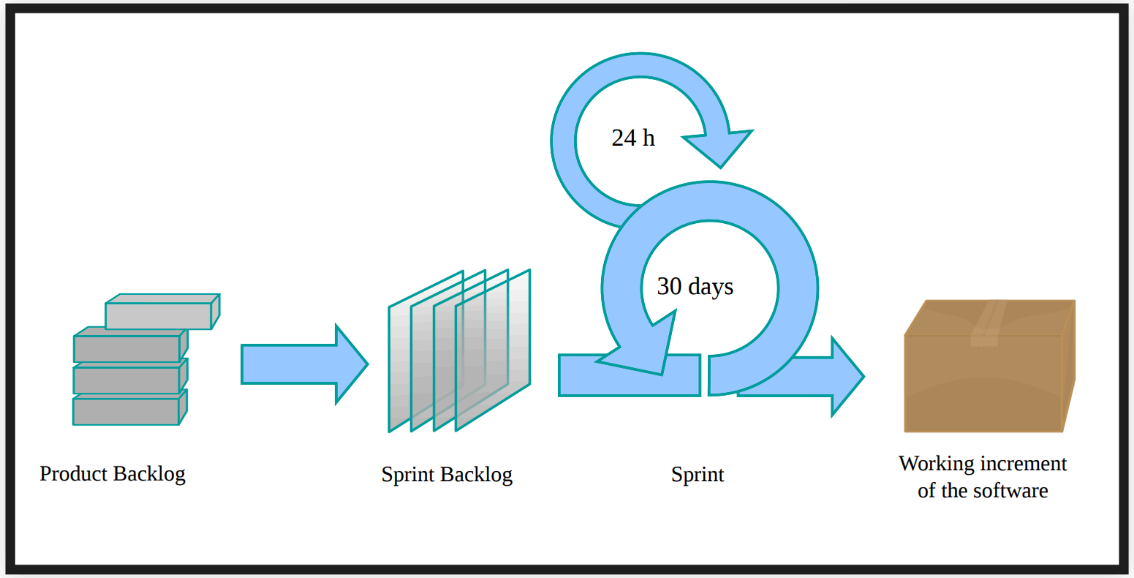
Data Science Lifecycle



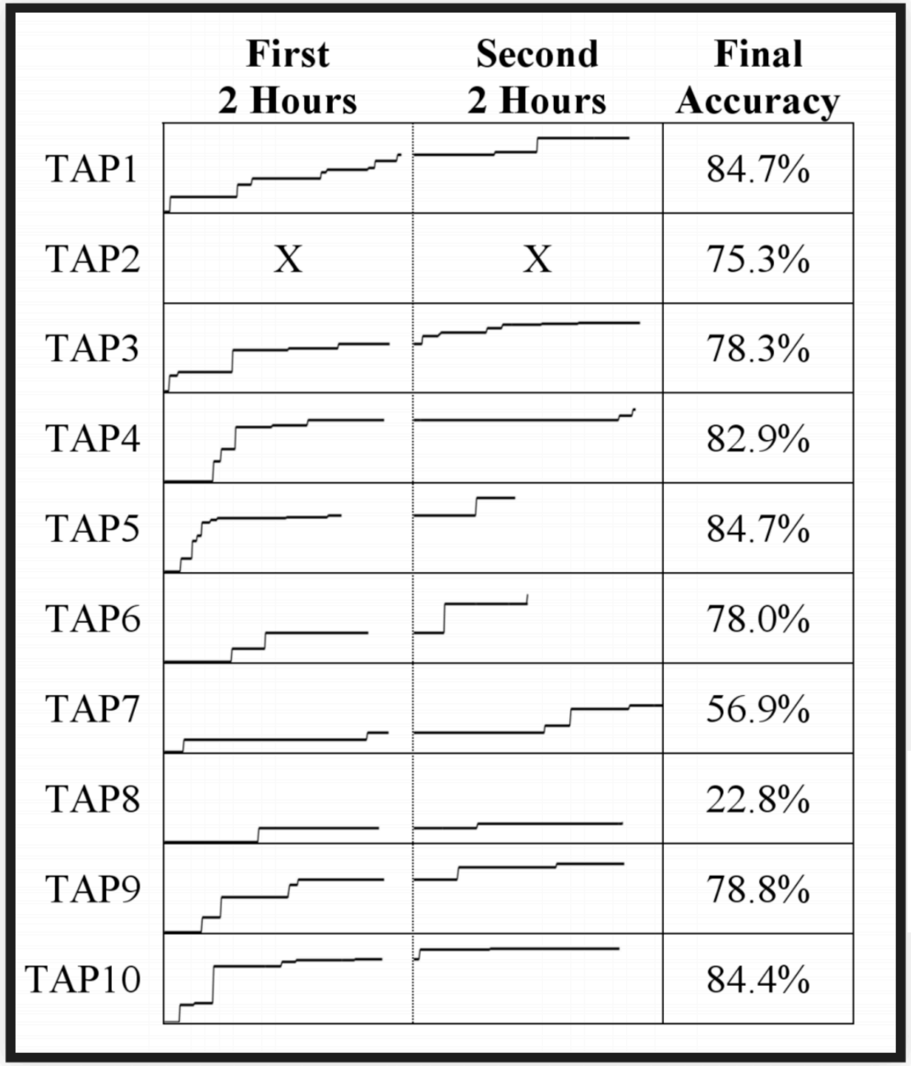
https://docs.microsoft.com/en-us/azure/machine-learning/team-data-science-process/overview

Similar to Spiral Process or Agile?





Data science is iterative and exploratory



	First 2 Hours	Second 2 Hours	Final Accuracy
TAP1			84.7%
TAP2	X	X	75.3%
TAP3			78.3%
TAP4			82.9%
TAP5			84.7%
TAP6			78.0%
TAP7			56.9%
TAP8			22.8%
TAP9			78.8%
TAP10			84.4%

Source: Patel, Kayur, James Fogarty, James A. Landay, and Beverly Harrison. "Investigating statistical machine learning as a tool for software development." In Proc. CHI, 2008.

Data science is iterative and exploratory

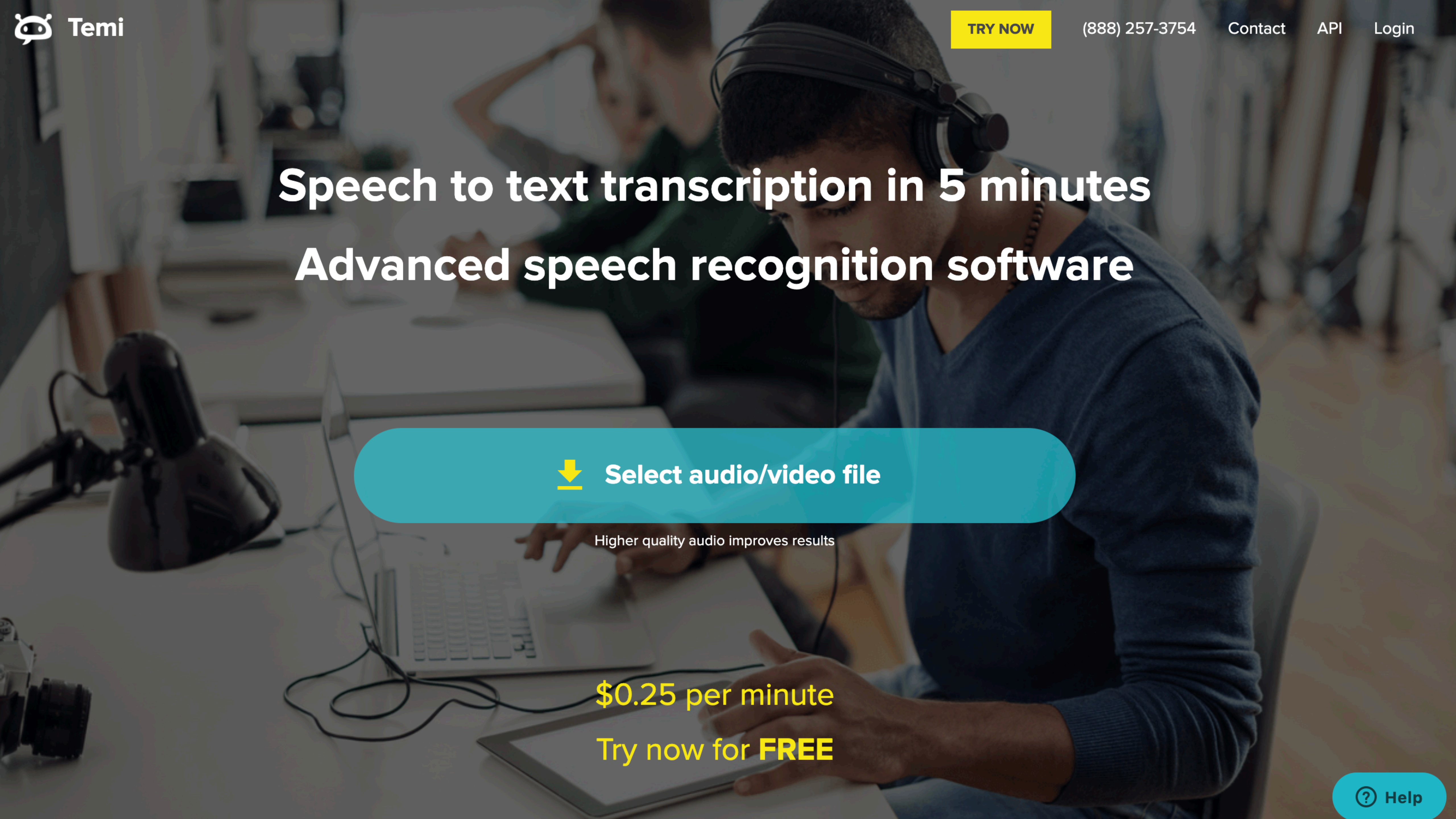
- Science mindset: start with rough goal, no clear specification, unclear whether possible
- Heuristics and experience to guide the process
- Try and error, refine iteratively, hypothesis testing
- Go back to data collection and cleaning if needed, revise goals

Share experience?

Case Study:
The Transcription Service Startup

Transcription Services

- Take audio or video files and produce text.
  - Used by academics to analyze interview text
  - Podcast show notes
  - Subtitles for videos
- State of the art: Manual transcription, often mechanical turk (1.5 $/min)

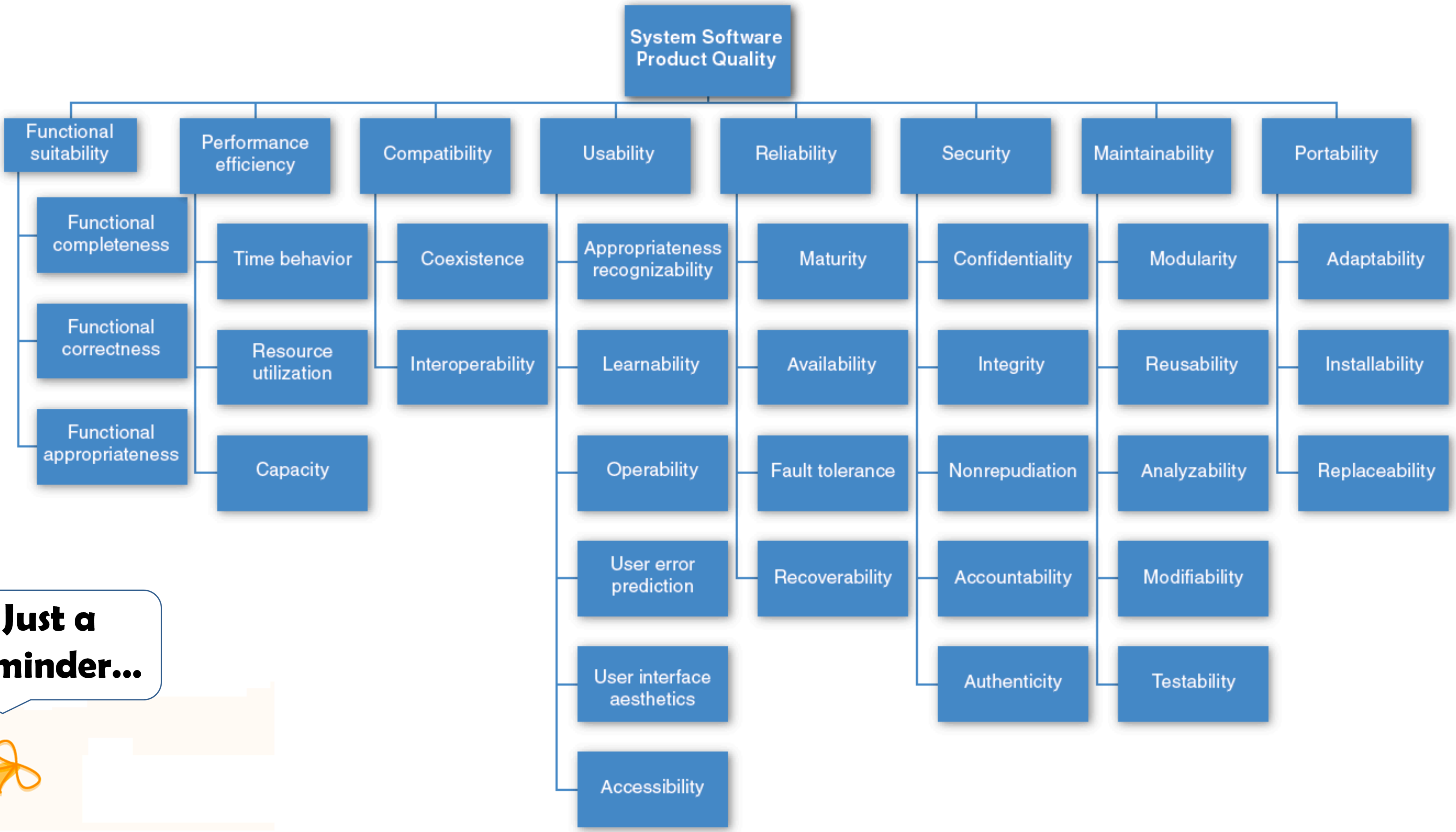

Temi
TRY NOW
(888) 257-3754
Contact
API
Login
Speech to text transcription in 5 minutes
Advanced speech recognition software
Select audio/video file
Higher quality audio improves results
$0.25 per minute
Try now for FREE
Help

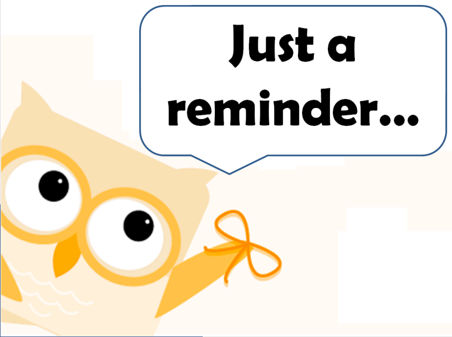
The startup idea

- 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

Likely Challenges?

Quality Attributes





Qualities of Interest?

Quality of Interests

- Quality is about more than the absence of defects
- Quality in use (effectiveness, efficiency, satisfaction, freedom of risk, ...)
- Product quality (functional correctness and completeness, performance efficiency, compatibility, usability, dependability, scalability, security, maintainability, portability, ...)
- Process quality (manageability, evolvability, predictability, ...)

- "Quality is never an accident; it is always the result of high intention, sincere effort, intelligent direction and skillful execution; it represents the wise choice of many alternatives." (many attributions)

Examples for Discussion

- What does correctness or accuracy really mean? What accuracy do customers care about?
- How can we see how well we are doing in practice? How much feedback are customers going to give us before they leave?
- Can we estimate how good our transcriptions are? How are we doing for different customers or different topics?
- How to present results to the customers (including confidence)?
- When customers complain about poor transcriptions, how to prioritize and what to do?

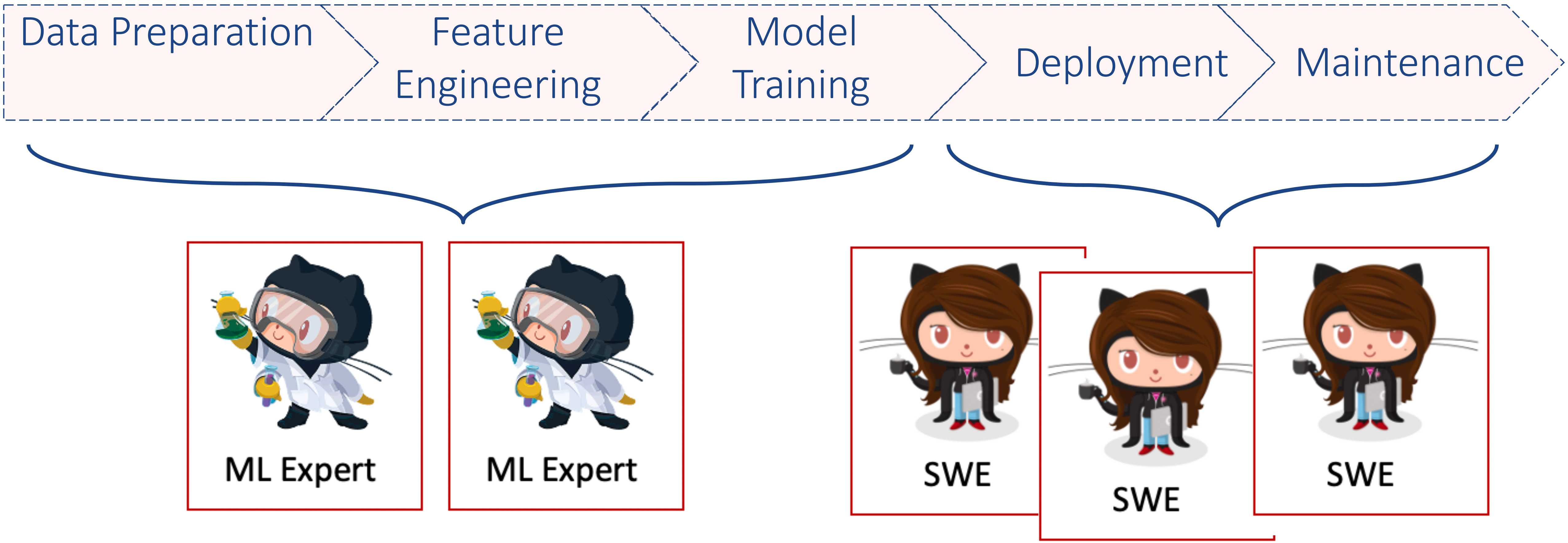
- What are unacceptable mistakes and how can they be avoided? Is there a safety risk?
- Can we cope with an influx of customers?
- Will transcribing the same audio twice produce the same result? Does it matter?
- How can we debug and fix problems? How quickly?

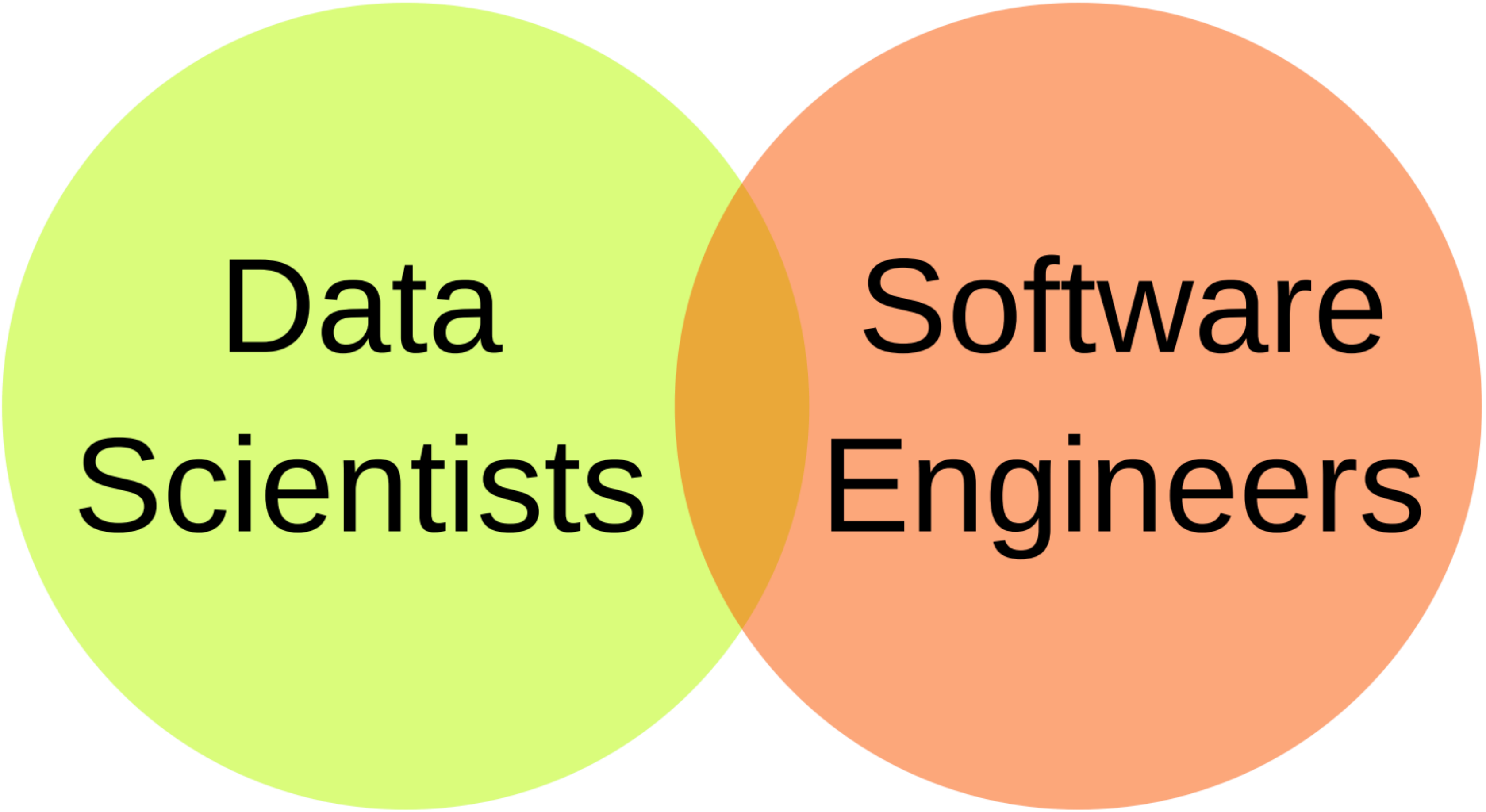
Examples for Discussion 2

- With more customers, transcriptions are taking longer and longer -- what can we do?
- Transcriptions sometimes crash. What to do?
- How do we achieve high availability?
- How can we see that everything is going fine and page somebody if it is not?
- We improve our entity detection model but somehow system behavior degrades... Why?
- Tensorflow update; does our infrastructure still work?
- Once somewhat successful, how to handle large amounts of data per day?
- Buy more machines or move to the cloud?

- Models are continuously improved. When to deploy? Can we roll back?
- Can we offer live transcription as an app? As a web service?
- Can we get better the longer a person talks? Should we then go back and reanalyze the beginning? Will this benefit the next upload as well?

Challenges


Data Preparation
Feature Engineering
Model Training
Deployment
Maintenance
ML Expert
ML Expert
SWE
SWE
SWE


Data
Scientists
Software
Engineers

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

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
- Detect and handle mistakes, preferably automatically
- Maintain, evolve, and extend the product over long periods
- Consider requirements for security, safety, fairness

Likely Collaboration Challenges?

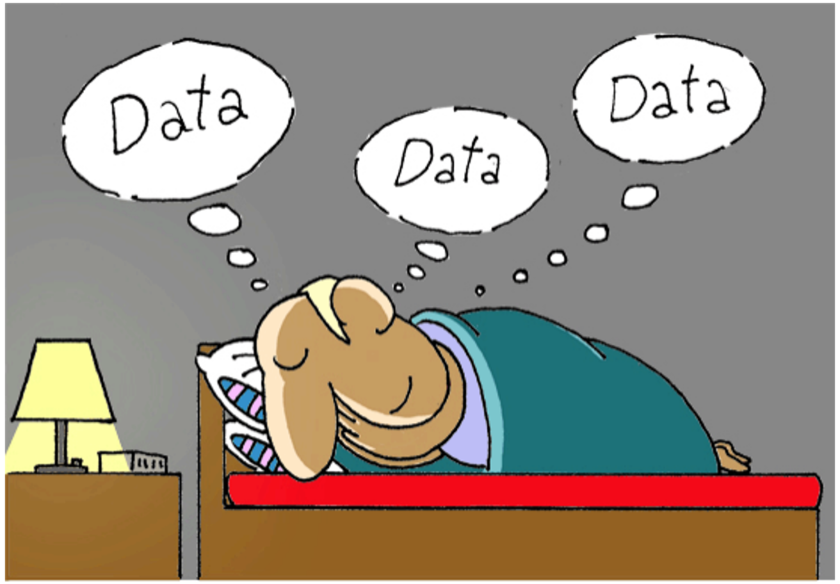
2016 IEEE/ACM 38th IEEE International Conference on Software Engineering

The Emerging Role of Data Scientists on Software Development Teams

Miryung Kim
UCLA
Los Angeles, CA, USA
miryung@cs.ucla.edu

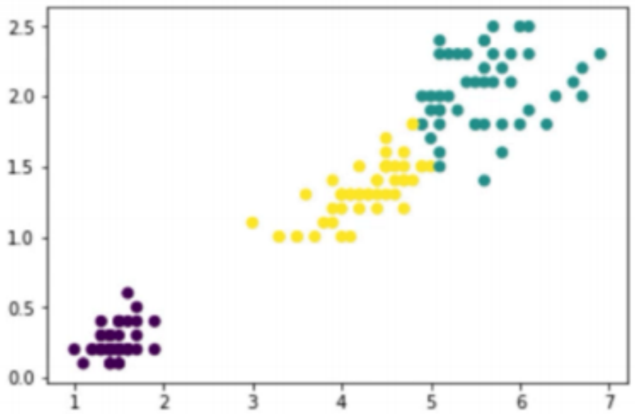
Thomas Zimmermann, Robert DeLine, Andrew Begel
Microsoft Research
Redmond, WA, USA
{tzimmer, rdeline, andrew.begel}@microsoft.com

ML Expert



Computational Notebooks

```
In [1]: import matplotlib.pyplot as plt
        from sklearn.cluster import KMeans
        from sklearn import datasets

In [2]: data = datasets.load_iris().data[:,2:4]
        petal_length, petal_width = data[:,0], data[:,1]

In [3]: print("Average petal length: %.3f" % (sum(petal_length) / len(petal_length),))

        Average petal length: 3.758

In [4]: clusters = KMeans(n_clusters=3).fit(data).labels_

In [5]: plt.scatter(petal_length, petal_width, c=clusters)

Out[5]: <matplotlib.collections.PathCollection at 0x124e294e0>
```




```
In [ ]: |
```

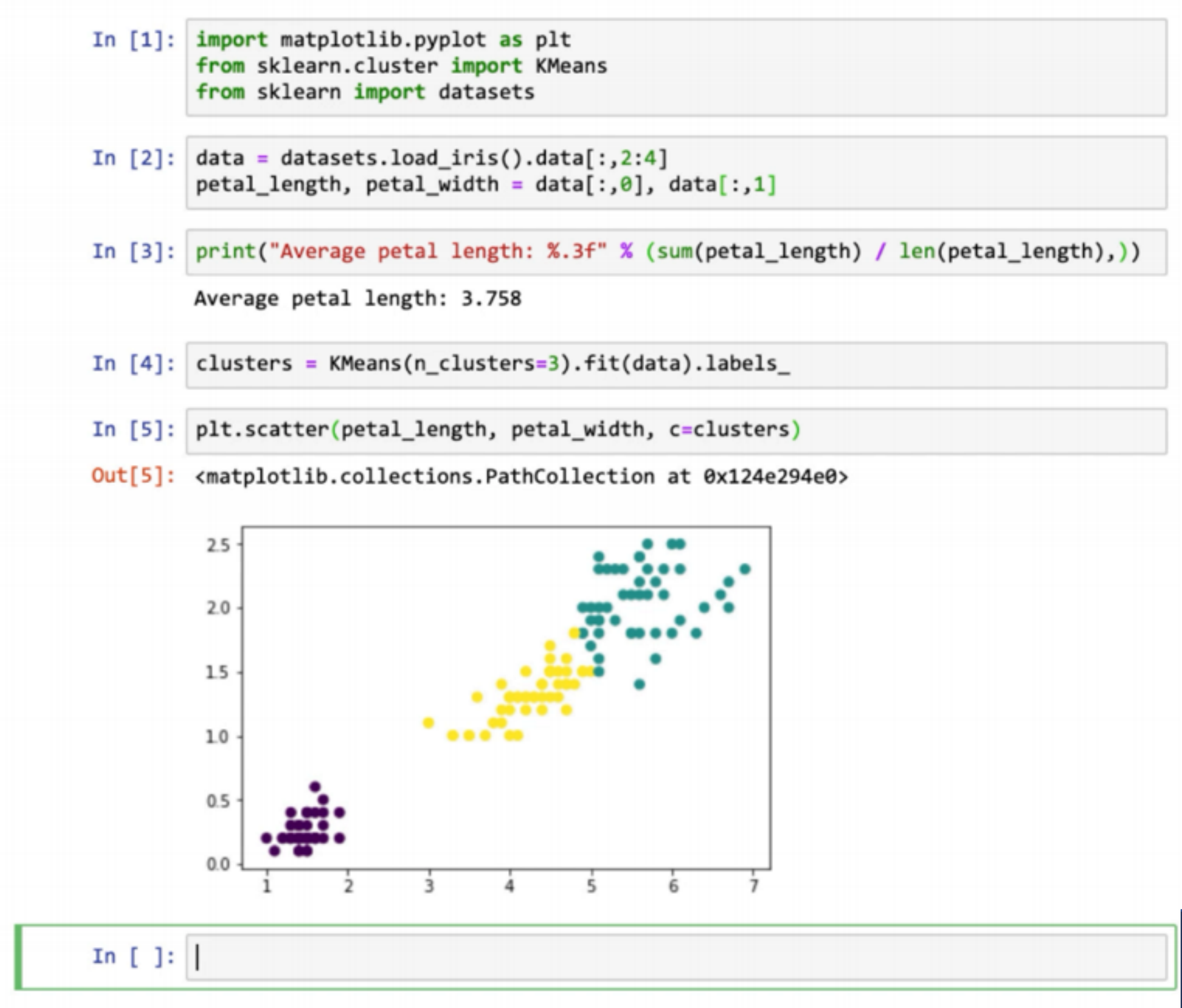
https://andrewhead.info/assets/pdf/notebook-gathering-slides.pdf

Computational Notebooks

- Quick feedback, similar to REPL **Read-Eval-Print-Loop**
- Visual feedback including figures and tables
- Incremental computation: reexecuting individual cells
- Quick and easy: copy paste, no abstraction needed
- Easy to share: document includes text, code, and results

Notebook limitations and Drawbacks?

Problem



1 WEEK LATER

1. How did I produce this result?
2. Didn't I have a better version of this?
3. What can I get rid of?

Problem

Poor code quality (Exploration)

- Buggy code (lack of testing)
- Duplicate code
- Tangled & Scattered code
- Unused code
- Lack of documentation

[Chattopadhyay et al. CHI'20, Head et al. CHI'19, Kery et al. CHI'19, Kery et al. VL/HCC'18]

Problem

Poor code quality
(Exploration)

Buggy code (lack of testing)
Duplicate code
Tangled & Scattered code
Unused code
Lack of documentation

[Chattopadhyay et al. CHI'20, Head et al. CHI'19, Kery et al. CHI'19, Kery et al. VL/HCC'18]

Reproducibility

"A startup's ML models were so disorganized it was causing serious problems as his team tried to build on each other's work and share it with clients. Even the original author sometimes couldn't train the same model and get similar results!" [1]

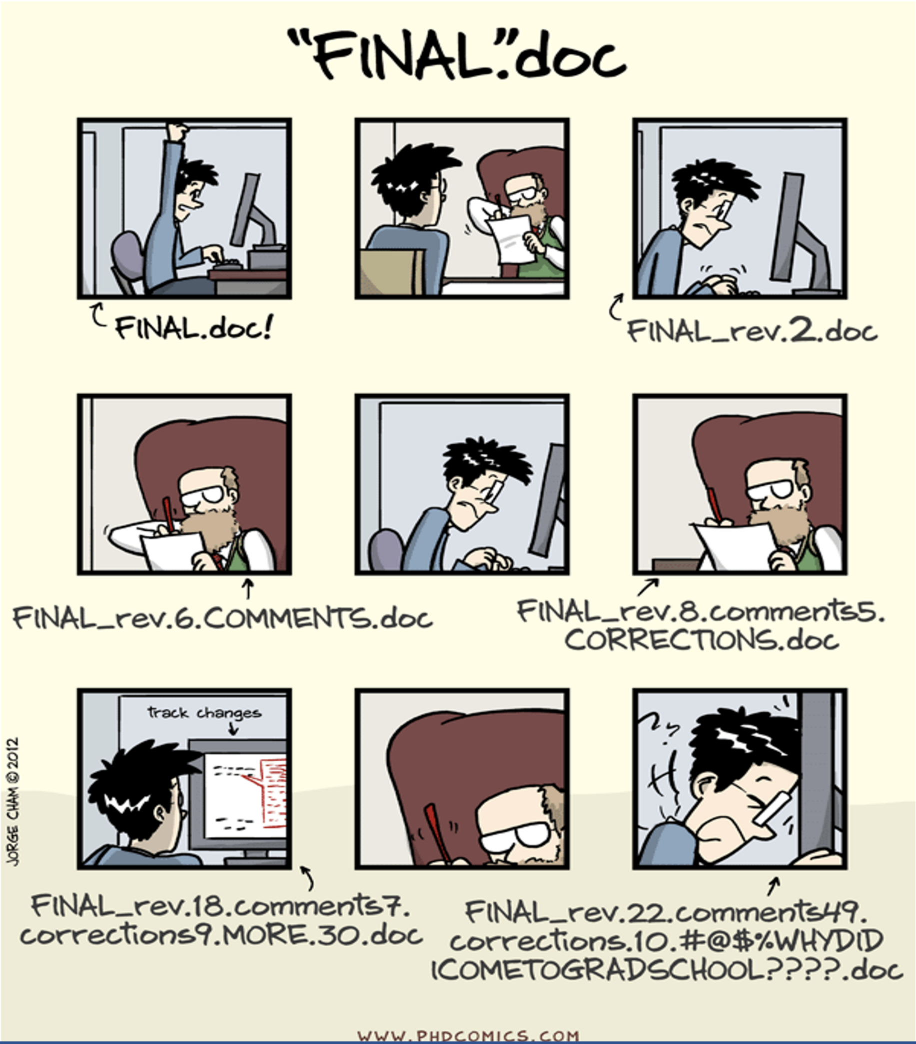
[1] The Machine Learning Reproducibility Crisis, https://petewarden.com/2018/03/19/the-machine-learning-reproducibility-crisis/

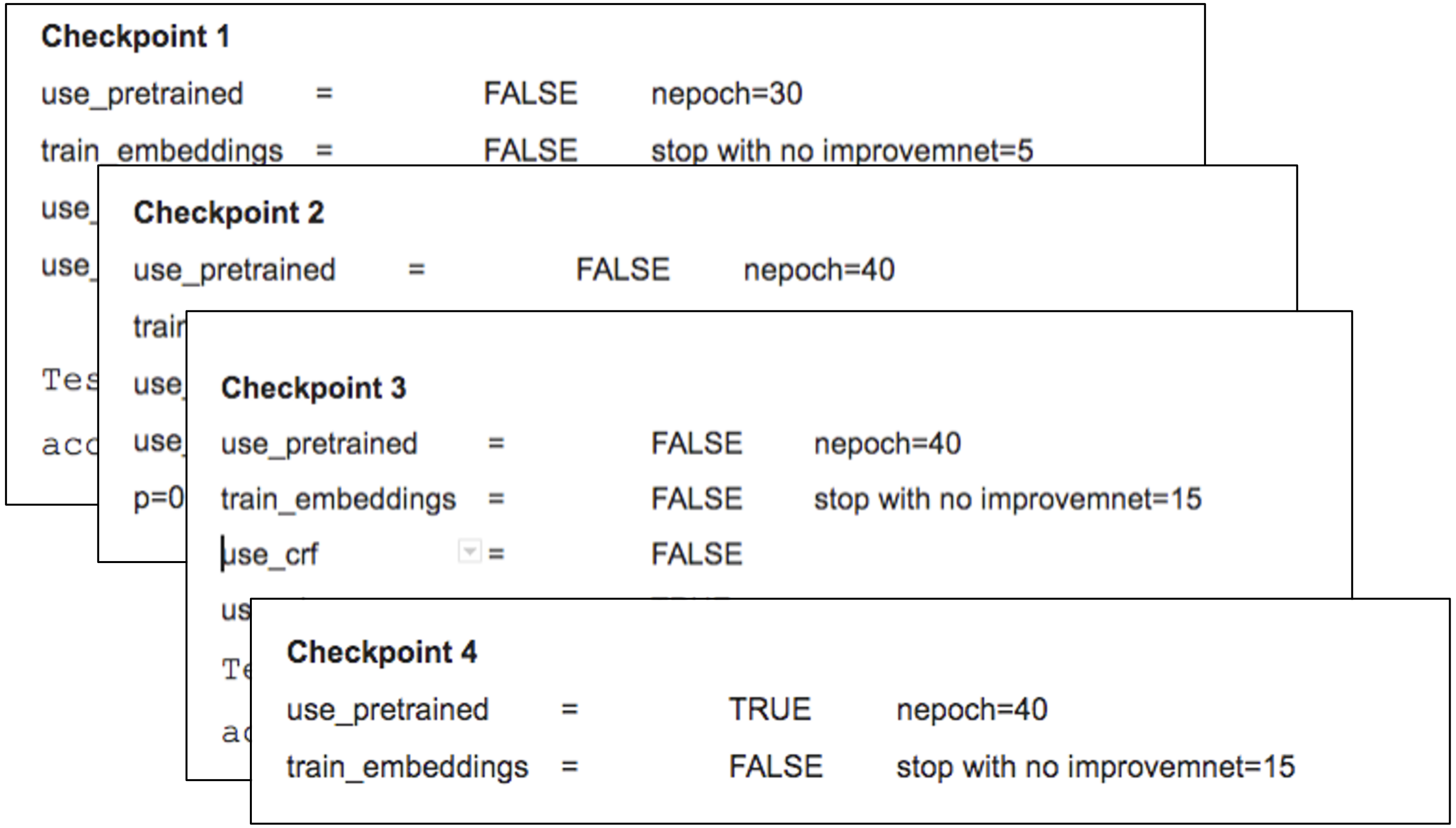
Reproducibility

Machine Learning: it's time to embrace version control [DataOps]

September 2018 · 9 minute read

Current Practices





Practices in SE don’t meet the needs

- " If you were to map this onto a traditional git workflow, what you would get is thousands of orphaned branches with one or two commits. Which isn't really useful, because none of our UIs are built for tracking thousands of branches, along with the results of those experiments."

https://www.reddit.com/r/MachineLearning/comments/9gakdd/ml_people_are_bad_at_version_control_d/

Machine Learning:
The High Interest Credit Card of Technical Debt

Andrew Ng
@AndrewYNg

1/The rise of Software Engineering required inventing processes like version control, code review, agile, to help teams w
Enginee
split trai

12:59 PM ·

1.1K Retwe

2/I'm also seeing many AI teams use new processes that haven't been formalized or named yet, ranging from how we write product requirement docs to how we version data anc
develop

3/Have you seen an idea for organizing ML projects that you'd like to share with others? If so please reply to this tweet!

AI 2020 = Software Engineering 1970s

How does Machine Learning Change Software Development Practices?

Zhiyuan Wan, Xin Xia, David Lo and Gail C. Murphy

Specifications

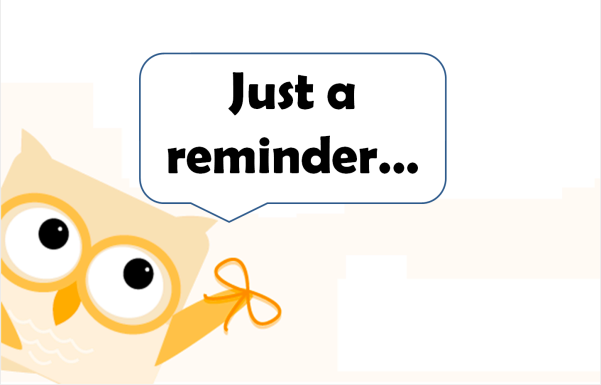
- Textual
- Assertions
- Formal specifications

```
/*@ requires amount >= 0;
    ensures balance == \old(balance)-amount &&
            \result == balance;
@*/
public int debit(int amount) {
    ...
}
```

- JML (Java modeling language specification)

```
/**
  * Calls the <code>read(byte[], int, int)</code> overloaded [..
  * @param buf The buffer to read bytes into
  * @return The value retured from <code>in.read(byte[], int, in
  * @exception IOException If an error occurs
  */
public int read(byte[] buf) throws IOException
{
    return read(buf, 0, buf.length);
}
```

- Textual specification with JavaDoc



Specification in ML?

```
/**
  ????
*/
List<Product> suggestedPurchases(List<Product> pastPurchases);
```

Specification in ML?

- Usually clear specifications do not exist -- we use machine learning exactly because we do not know the specifications
- Can define correctness for some data, but not general rules; sometimes can only determine correctness after the fact
- Learning for tasks for which we cannot write specifications
    - Too complex
    - Rules unknown
- AI will learn rules/specifications, often not in a human-readable form, but are those the right ones?

- Often *goals* used instead --> maximize a specific objective

From Models to AI-Based Systems

Whole System Perspectives

- A model is just one component of a larger system
- Also pipeline to build the model
- Also infrastructure to deploy, update, and serve the model
- Integrating the model with the rest of the system functionality
- User interaction design, dealing with mistakes
- Overall system goals vs model goals

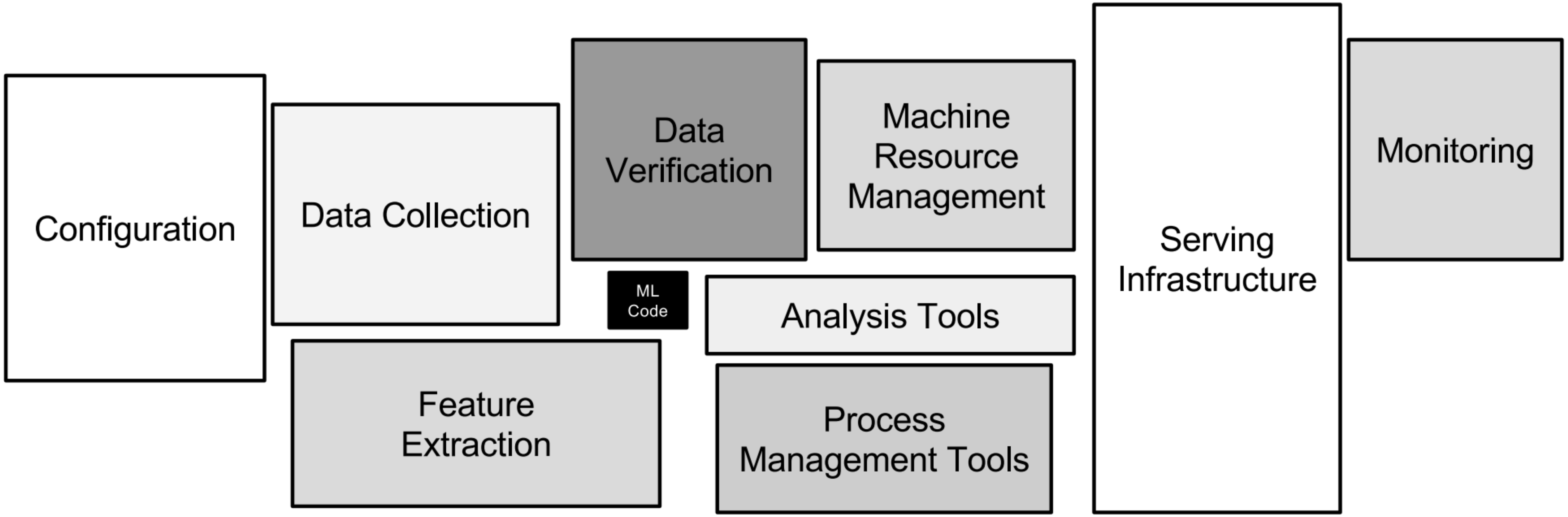


Figure 1: Only a small fraction of real-world ML systems is composed of the ML code, as shown by the small black box in the middle. The required surrounding infrastructure is vast and complex.

Hidden Technical Debt in Machine Learning Systems

Thinking about systems

- Holistic approach, looking at the larger picture, involving all stakeholders
- Looking at relationships and interactions among components and environments
  - Everything is interconnected
  - Combining parts creates something new with emergent behavior
  - Understand dynamics, be aware of feedback loops, actions have effects
- Understand how humans interact with the system

> *A system is a set of inter-related components that work together in a particular environment to perform whatever functions are required to achieve the system's objective -- Donella Meadows*