FROM MODELS TO AI-ENABLED SYSTEMS

Christian Kaestner

- Hulten, Geoff. "Building Intelligent Systems: A Guide to Machine Learning Engineering." (2018), Chapters 5 (Components of Intelligent Systems).
- 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." In Advances in neural information processing systems, pp. 2503-2511. 2015.

LEARNING GOALS

- Explain how machine learning fits into the larger picture of building and maintaining production systems
- Describe the typical components relating to AI in an AI-enabled system and typical design decisions to be made

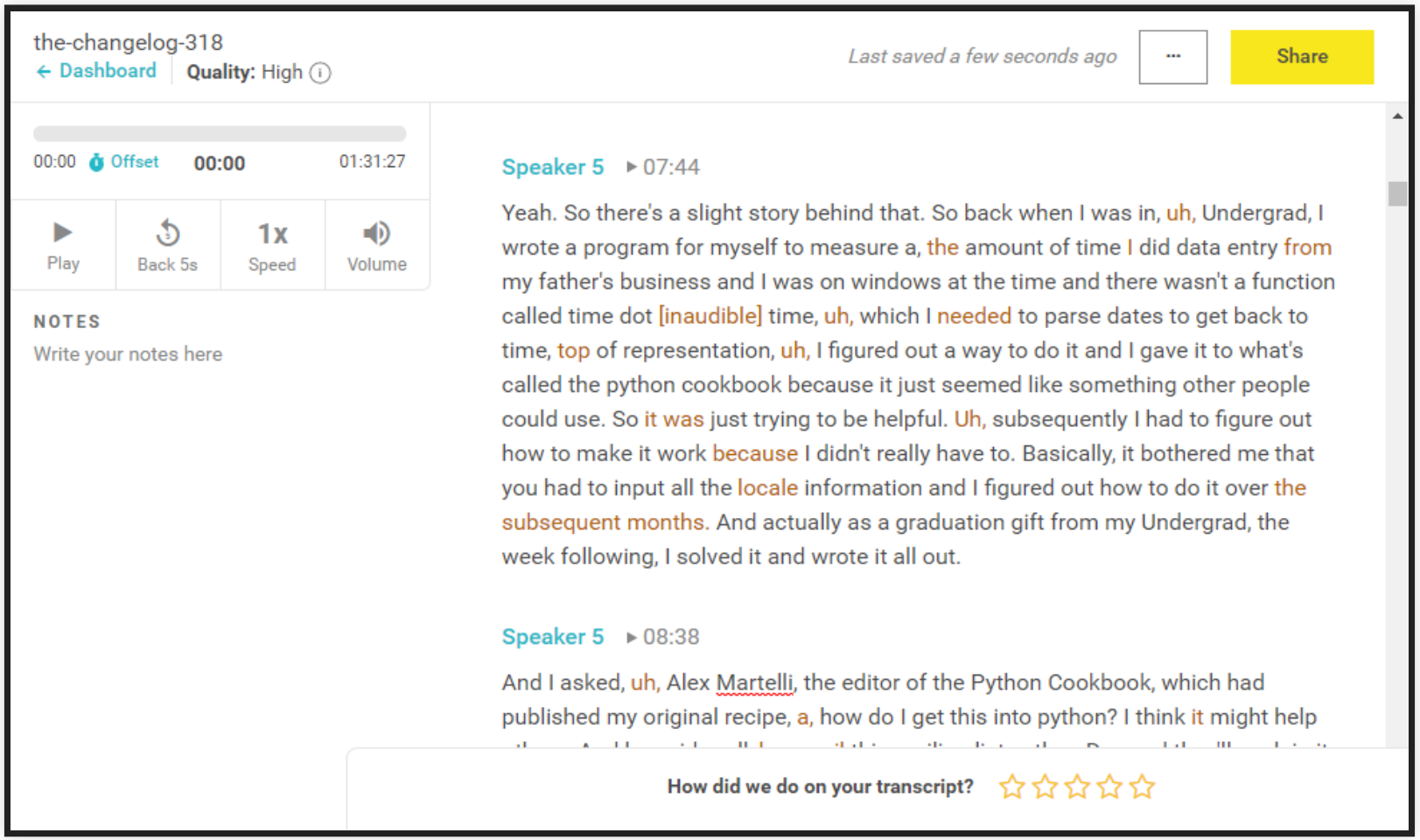
AI-ENABLED SYSTEMS

WHOLE SYSTEM PERSPECTIVE

- 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

let's look at a couple of examples

TEMI TRANSCRIPTION SERVICE

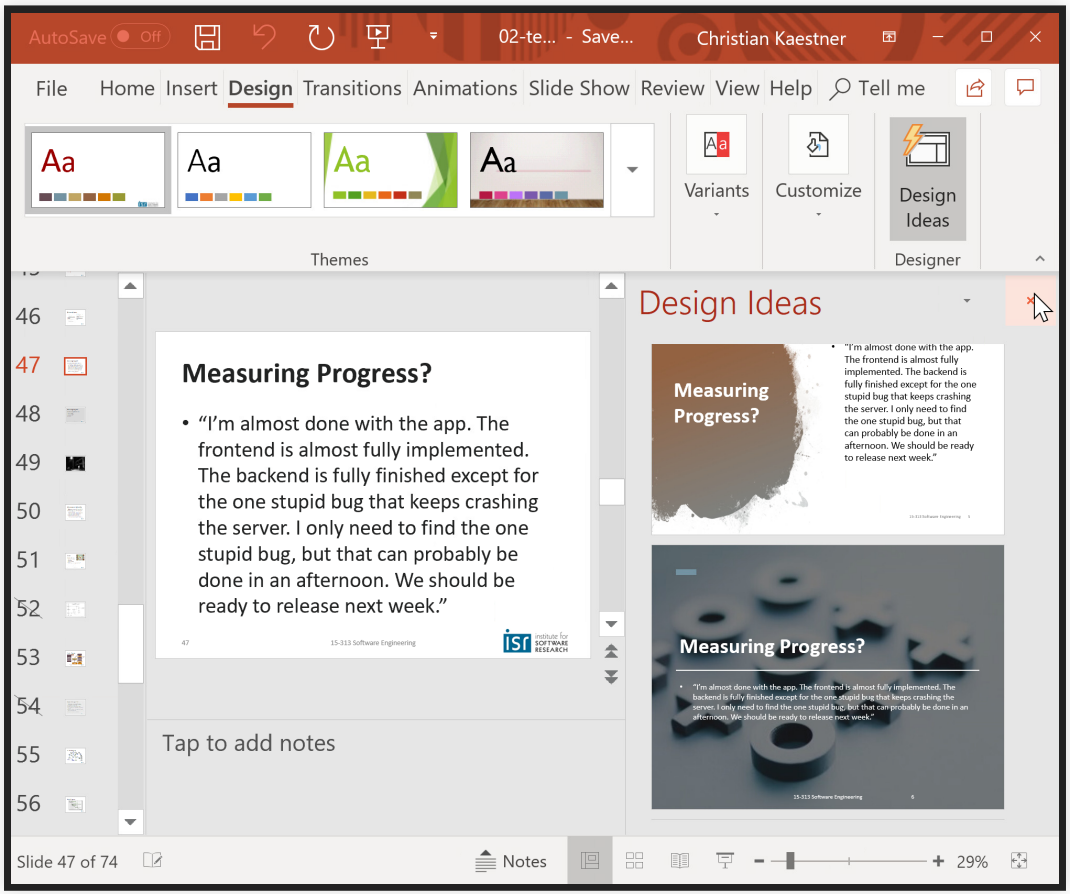


https://www.temi.com/

Speaker notes

A model is very central to this service. Product built around a model. Still, lots of nonmodel code for UI, storage of customer data, credit card processing, ...

MICROSOFT POWERPOINT



Read more: How Azure Machine Learning enables PowerPoint Designer, Azure Blog, March 2020

Speaker notes

Traditional application that uses machine learning in a few smaller places (more and more these days).

FALL DETECTION DEVICES

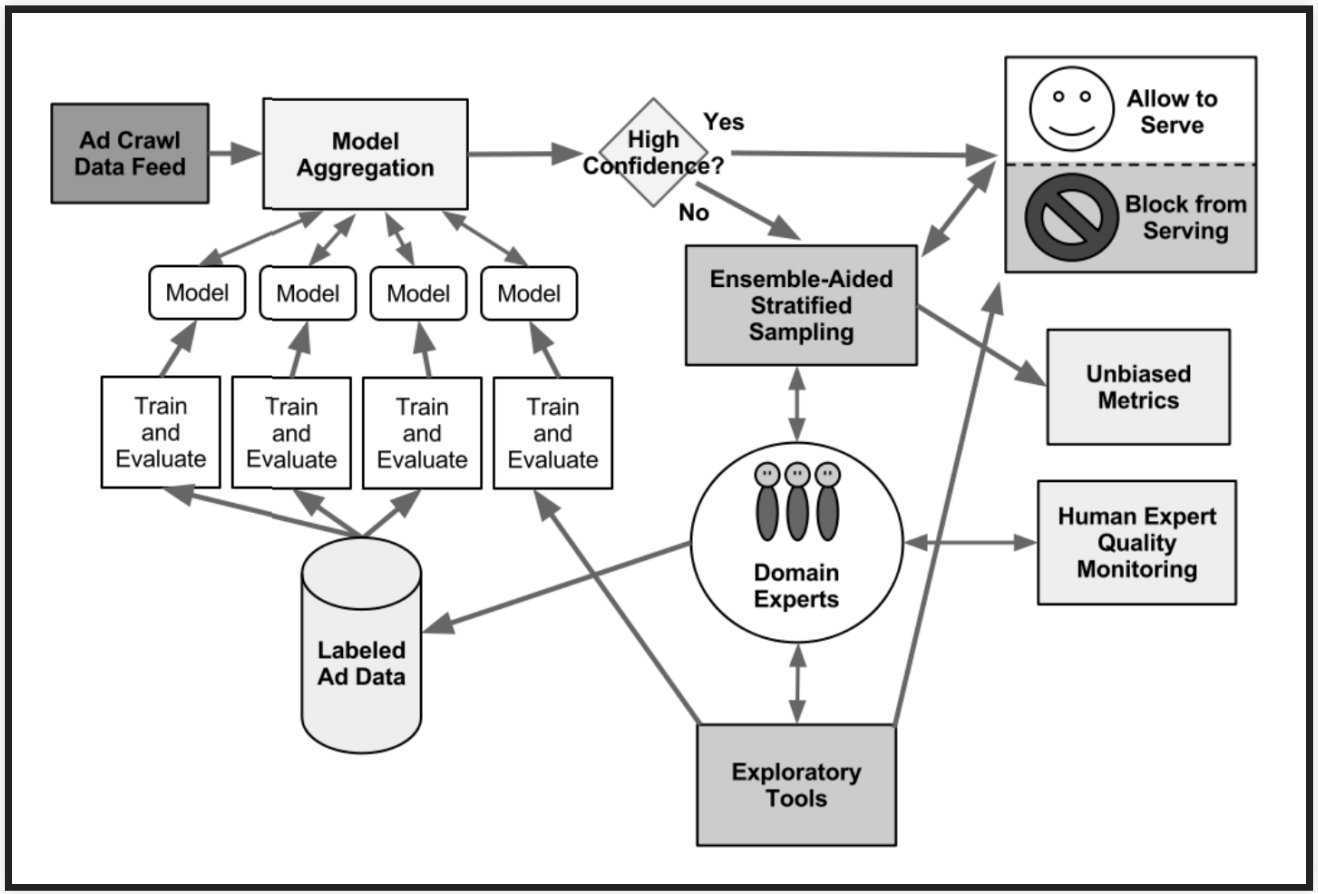
(various devices explored, including smart watches, hearing aids, and wall and floor sensors)

Read more: How fall detection is moving beyond the pendant, MobiHealthNews, 2019

Speaker notes

Devices for older adults to detect falls and alert caretaker or emergency responders automatically or after interaction. Uses various inputs to detect falls.

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.

Speaker notes

See first homework assignment. System largely build around a model for a specific purpose but integrated into larger infrastructure.

RECIDIVISM PREDICTION

```
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

Speaker notes

The system is very narrowly built around a model, but has large societal implications.

LOGISTICS, ROUTE PLANNING

Speaker notes

Heavy AI (not just ML) integrated in large system approximating planning problems with many inputs, interfacing with many other systems.

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

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*

Leyla Acaroglu. "Tools for Systems Thinkers: The 6 Fundamental Concepts of Systems Thinking." Blogpost 2017

SYSTEM-LEVEL CHALLENGES FOR AI-ENABLED SYSTEMS

- Getting and updating data, concept drift, changing requirements
- Handling massive amounts of data
- Interactions with the real world, feedback loops
- Lack of modularity of AI components, lack of specifications, nonlocal effects
- Deployment and maintenance
- Versioning, debugging and incremental improvement
- Keeping training and operating cost manageable
- Interdisciplinary teams
- Setting system goals, balancing stakeholders and requirements
- ...

Examples?

ON TERMINOLOGY

- There is no standard term for referring to building systems with AI components
- "AI-Enabled Systems", "ML-Enabled Systems" or "ML-Infused Systems"
- SE4AI, SE4ML
- sometimes AI engineering
- sometimes ML Systems Engineering (but often this refers to building distributed and scalable ML learning and data storage platforms)
- AIOps ~ using AI to make automated decisions in operations; DataOps ~ use of agile methods and automation in business data analytics; MLOps ~ technical infrastructure for operating AI-based products and on deploying updates
- Developers with Software Engineering and ML skills were often referred to as "unicorns" in earlier days

COMPONENTS OF AN AI-ENABLED SYSTEM

(Using Hulten's Terminology)

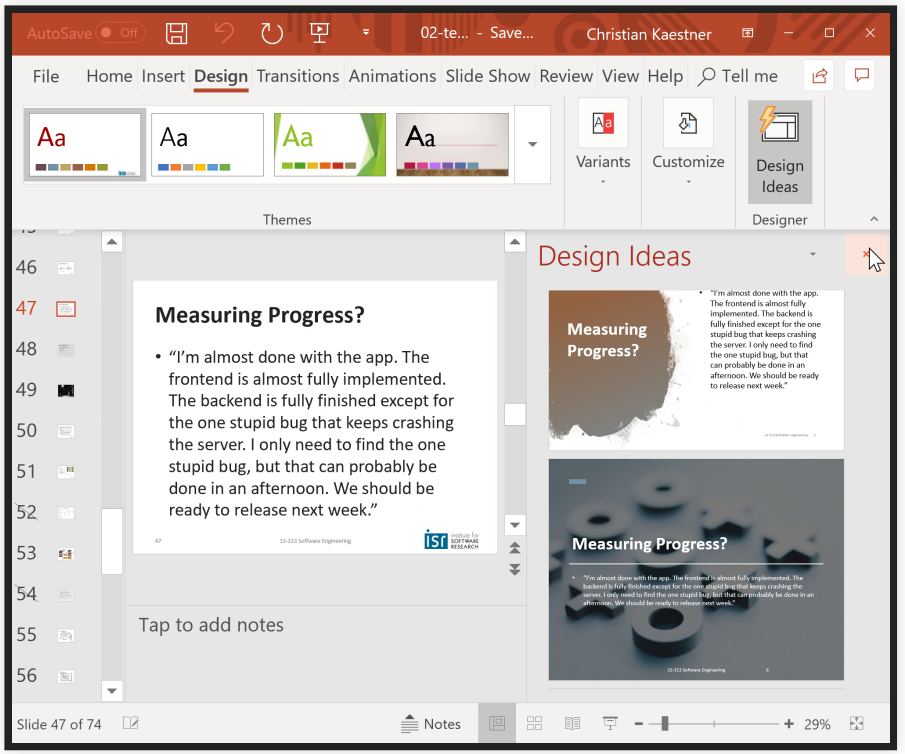
- Hulten, Geoff. "Building Intelligent Systems: A Guide to Machine Learning Engineering." (2018).

ELEMENTS OF AN INTELLIGENT SYSTEM

- **Meaningful objective**: goals, requirements, business case
- **Intelligent experience**: user interactions -- presenting model predictions to users; user interactions; eliciting feedback, telemetry
- **Intelligence implementation**: infrastructure -- learning and serving the model and collecting feedback (telemetry)
- **Intelligence creation**: learning and evaluating models
- **Orchestration**: operations -- maintaining and updating the system over time, debugging, countering abuse

DESIGN DECISIONS FOR EACH ELEMENT?

- Meaningful objective
- Intelligent experience / user interaction design
- Intelligence implementation / infrastructure
- Intelligence creation
- Orchestration / operations



USER INTERACTIONS (INTELLIGENT EXPERIENCES)

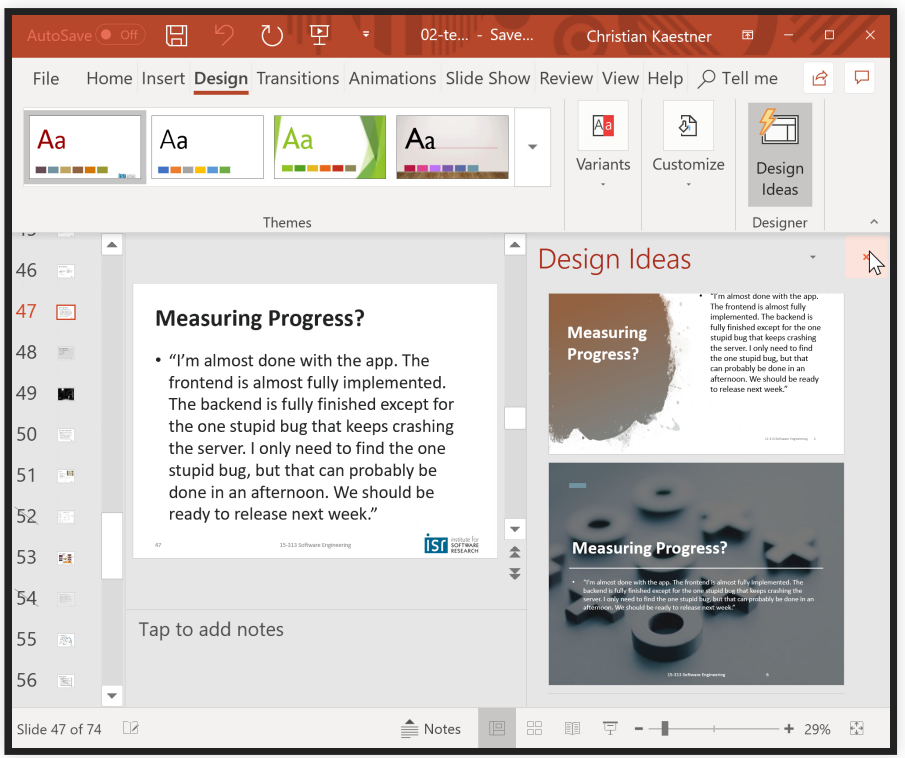
DESIGNING INTELLIGENT EXPERIENCES

- How to use the output of a model's prediction (for a goal)?
- Design considerations:
    - How to present prediction to a user? Suggestions or automatically take actions?
    - How to effectively influence the user's behavior toward the system's goal?
    - How to minimize the consequences of flawed predictions?
    - How to collect data to continue to learn from users and mistakes?
- Balancing at least three outcomes:
    - Achieving goals
    - Protection from mistakes
    - Collecting data for training

DESIGNING INTELLIGENT EXPERIENCES

- How to use the output of a model's prediction (for a goal)?
- Design considerations:
  - How to present prediction to a user? Suggestions or automatically take actions?
  - How to effectively influence the user's behavior toward the system's goal?
  - How to minimize the consequences of flawed predictions?
  - How to collect data to continue to learn from users and mistakes?

Automatic slide design:



DESIGNING INTELLIGENT EXPERIENCES

- How to use the output of a model's prediction (for a goal)?
- Design considerations:
  - How to present prediction to a user? Suggestions or automatically take actions?
  - How to effectively influence the user's behavior toward the system's goal?
  - How to minimize the consequences of flawed predictions?
  - How to collect data to continue to learn from users and mistakes?

Fall detection:

FORCEFULNESS

- Forceful (hard to ignore or stop):
  - Automate an action
  - Interrupt the user and ask for confirmation before they can continue
- Passive experience:
  - Prompt that does not require immediate answer
  - Icon or information box making suggestion

Examples?

When to chose which?

MODES OF INTERACTION

- Automate
- Prompting
- Organizing information
- Annotate
- Hybrids

Examples?

Speaker notes

Lots of examples in Hulten's book, Chapter 8

FREQUENCY

- Interact whenever a new prediction is available
- Interact when prediction changes significantly
- Hard limit on interaction frequency (e.g., max 1 prediction per hour)
- Interact based on anticipated user reaction; adaptive
- Interaction explicitly initiated by user

Examples?

Consider notification fatigue vs missed opportunities to help vs learnability

Speaker notes

Examples: Interact frequently during navigation or giving fitness instructions (whenever things change); fewer predictions after many ignored ones

FACTORS TO CONSIDER

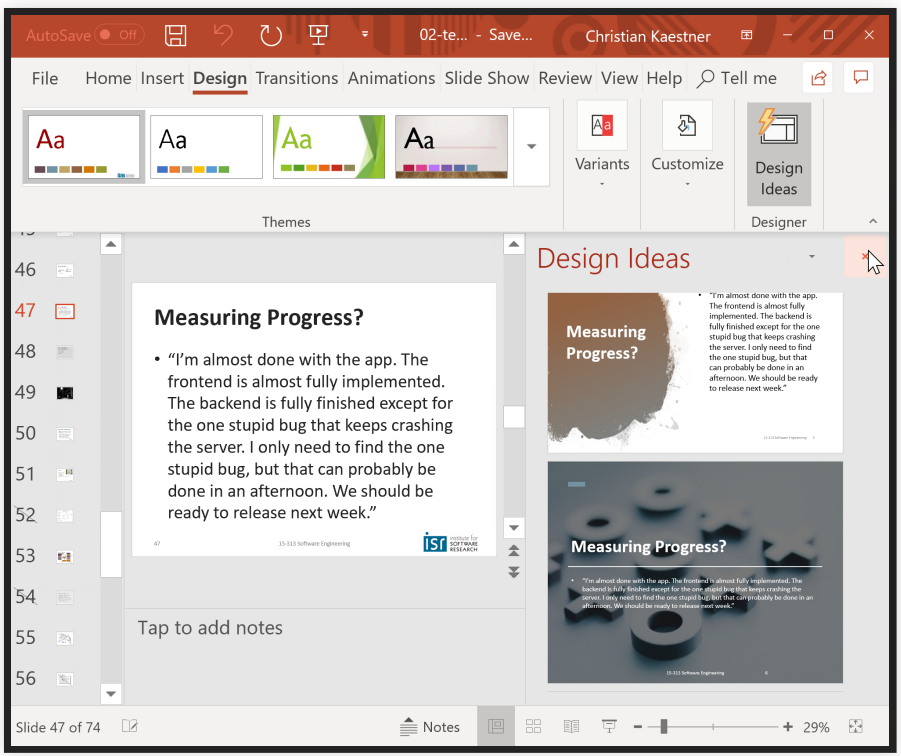
When designing an intelligent experience consider:

- Forcefulness: How strongly to encourage taking an action (or even automate it)?
- Frequency: How often to interact with the user?
- Value: How much does a user (think to) benefit from the prediction?
- Cost: What is the damage of a wrong prediction?
- Model quality: How often is the prediction wrong?

FACTORS IN CASE STUDIES

Consider: forcefulness, frequency, value, cost, model quality

Automatic slide design:



Fall detection:

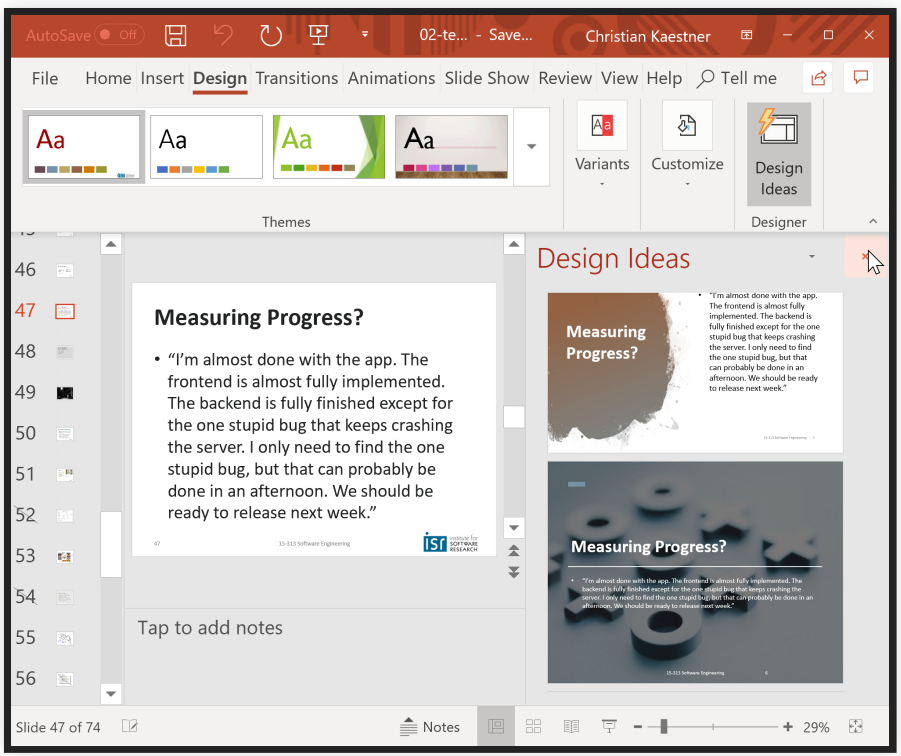
FEEDBACK (TELEMETRY)

- To design good interactions we need to know how we are doing...
- How many predictions are ignored?
- How many actions are reversed?
- How often does the user ask for extra predictions?
- How much value do users get out of predictions?
- How much are we supporting the system's goals?
- How much cost are wrong predictions causing for users/the system's goals?
- Are mistakes focused on specific kinds of inputs?

INITIAL TELEMETRY IDEAS?

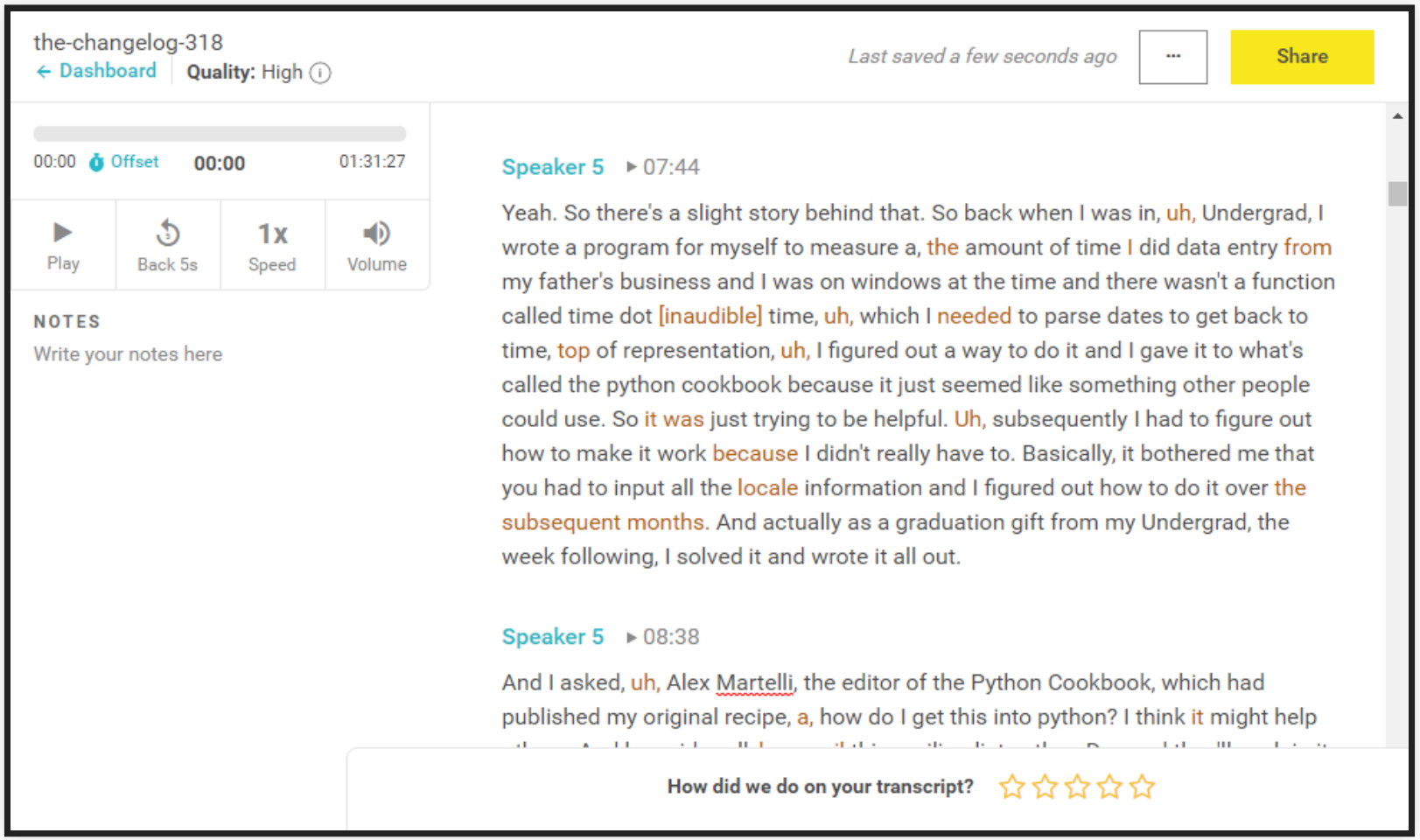
Identify: usage, mistakes, cost of mistakes, benefits to user, benefits to goals

Automatic slide design:



Fall detection:

OUTLOOK: TELEMETRY DESIGN



More on this later...

A SYSTEMS VIEW ON SAFETY

THE SMART TOASTER

the toaster may (occasionally) burn my toast, but should never burn down my kitchen

MAKING THE SMART TOASTER SAFE

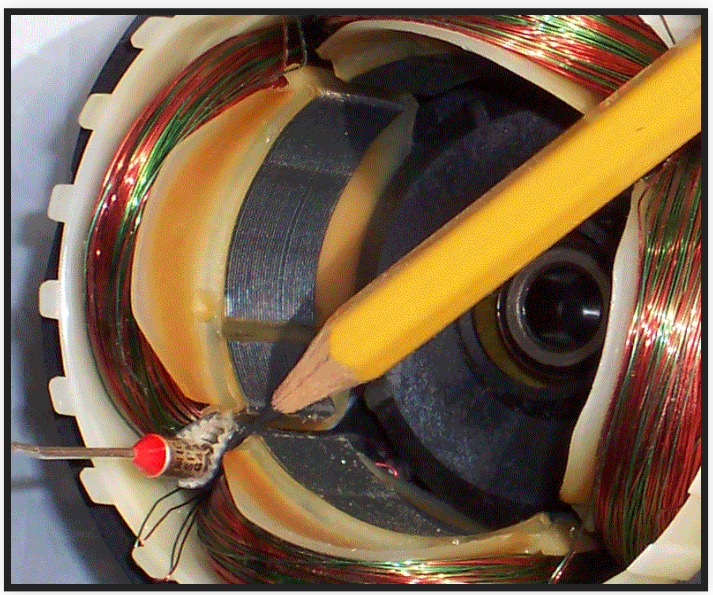
Assume classification model:

$$\text{continueToasting}(\text{camera}_{\text{initial}}, \text{camera}_{\text{now}}, \text{temperatureReading}, \text{userPref}) \rightarrow \text{Boolean}$$

How to assure the toaster does not overhead?

SAFEGUARDS / GUARDRAILS

- Hard constraints overrule model
  - `heat = (temperatureReading < MAX) && continueToasting(...)`
- External hardware or software failsafe mechanisms
  - outside the model, external observer, e.g., thermal fuses



(Image CC BY-SA 4.0, C J Cowie)

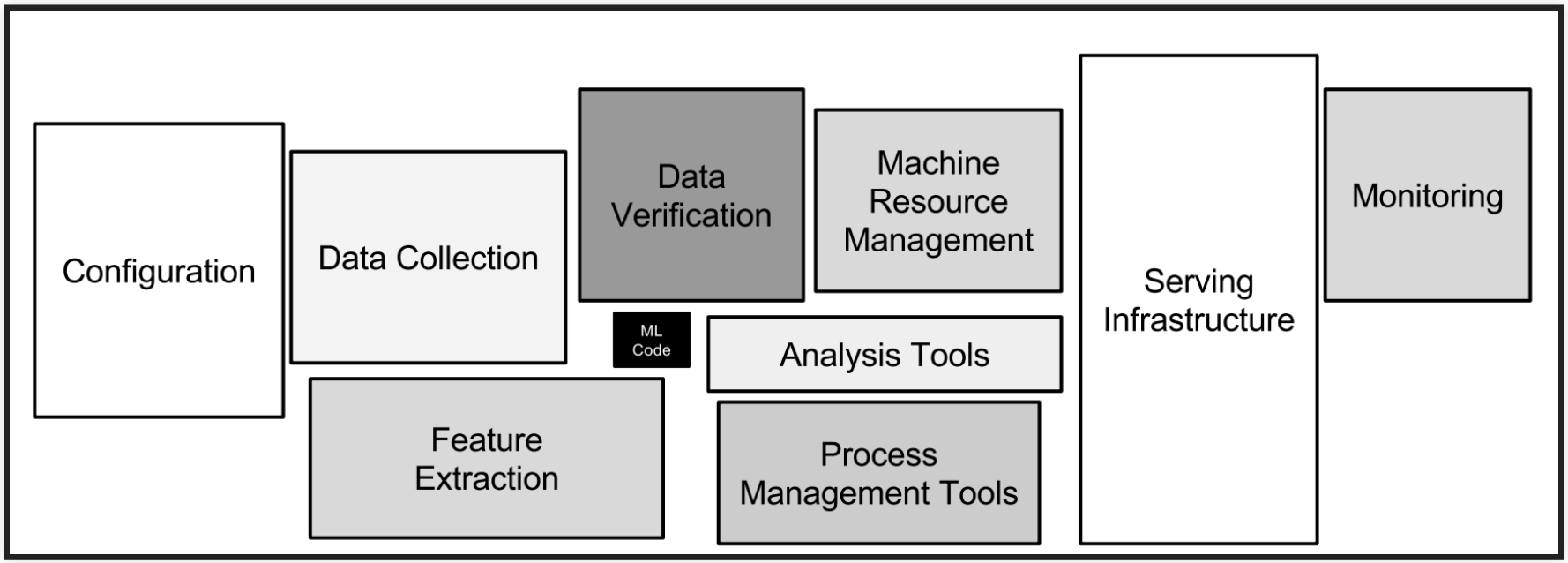
OTHER STRATEGIES

- Improve the model, more data, more testing
- Adjusting interaction models, e.g., involving users, confirmations
- Better hardware
- ...

In all cases, look beyond model accuracy at the entire system

A SYSTEM VIEW ON INTELLIGENCE INFRASTRUCTURE

INFRASTRUCTURE FOR ML COMPONENTS



This was 2015; many of those boxes are getting increasingly standardized these days.

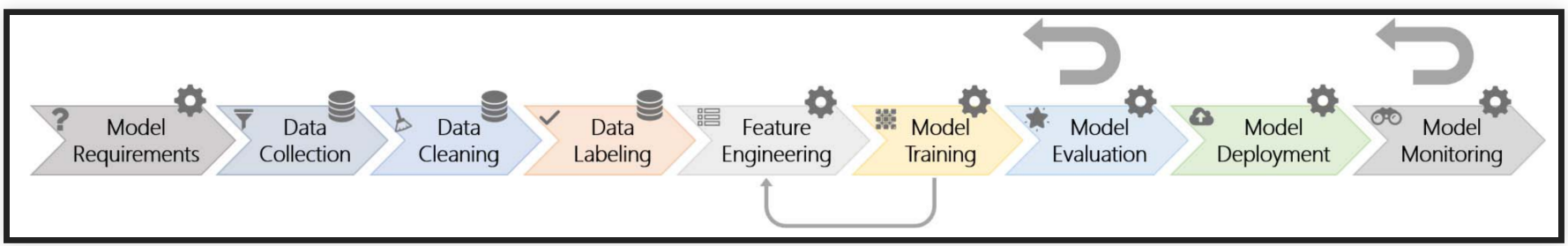
Graphic from Sculley, et al. "Hidden technical debt in machine learning systems." In Proc NIPS, 2015.

Speaker notes

Even for a single ML component and it's pipeline, there is a lot of infrastructure to build and serve the model.

THINKING IN PIPELINES OVER MODELS

- In production systems, models need to be deployed and updated
- Consider the entire pipeline, not just the model
  - Quality assurance, reproduciblity, repeatability, debugging
  - Modifiability, agility
  - Training cost and scalability
  - Data availability, data wrangling cost
  - Telemetry
- Reported as one of the key challenges in production machine learning



- Graphic: Amershi et al. "Software engineering for machine learning: A case study." In Proc ICSE-SEIP, 2019.
- Key challenge claim: O'Leary and Uchida. "Common problems with Creating Machine Learning Pipelines from Existing Code." Proc. MLSys, 2020.

SYSTEM QUALITIES VS MODEL ACCURACY

SYSTEMS HAVE GOALS

... selling stuff, increasing engagement, encouraging responsible behavior

Model predictions support those goals

more next lecture

MORE ACCURATE PREDICTIONS MAY NOT BE THAT IMPORTANT

- "Good enough" may be good enough
- Prediction critical for system success or just an gimmick?
- Better predictions may come at excessive costs
    - need way more data, much longer training times
    - privacy concerns
- Better user interface ("experience") may mitigate many problems
    - e.g. explain decisions to users
- Use only high-confidence predictions?

BEYOND MODEL QUALITY

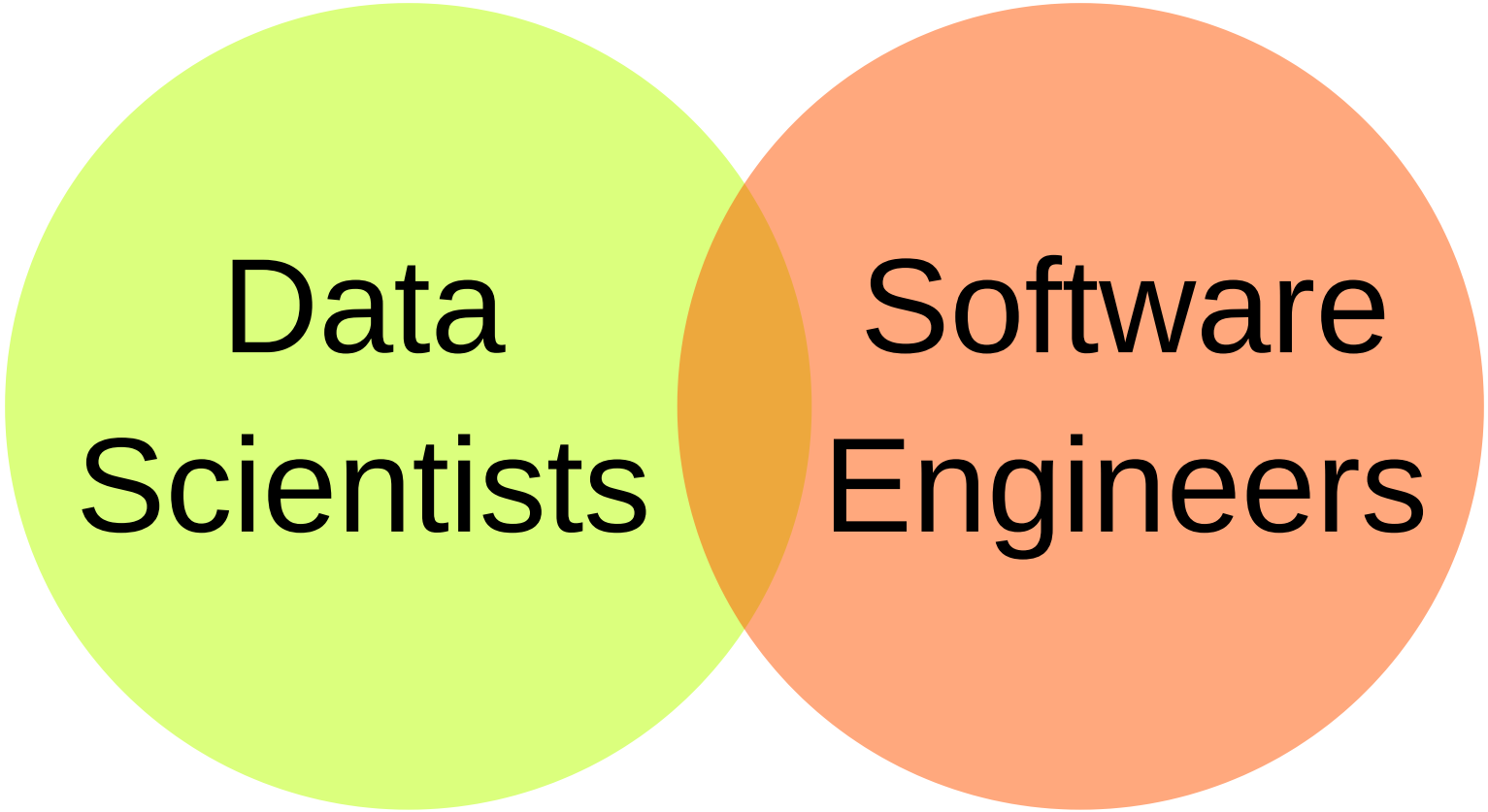
Many other aspects of a model's quality may matter when operating a system

Examples?

(more later)

Speaker notes

Learning time, inference time, incremental learning, explainability, model size, kinds of mistakes, fairness, privacy, security, robustness, reproducibility, maintainability


Data
Scientists
Software
Engineers

SUMMARY

- Production AI-enabled systems require a *whole system perspective*, beyond just the model
- Components: Objectives, user interface, infrastructure, AI component, and operations
- Large design space for user interface (intelligent experience): forcefulness, frequency, telemetry
- Quality at a system level: safety beyond the model, beyond accuracy
- Elevating the infrastructure: Thinking in pipelines, not models