Day 05

Wan Fang
Southern University of Science and Technology

Day 5

- 8:00 – 8:30 Preparation
- 8:30 – 10:10 Interim Review (< 20 mins x 5)
- 10:20 – 11:10 Lecture : AI Risk Management Framework
- 11:20 – 12:10 Exercise : AI Fairness + Google's What-If Tool

- 2:00 – 5:00 Exercise: Prototyping + Testing with data/model
- 5:00 – 6:00 Review of the day

AI Risk Management Framework

From National Institute of Standards and Technology (NIST), U.S. Department of Commerce

Second draft released on August 18th, 2022

https://www.nist.gov/itl/ai-risk-management-framework

Content

- Part I Motivation
  - Trustworthy and Responsible AI
  - Sources of AI Risks
  - Who should be involved?
  - Understanding Risk, Impact, and Harms
  - AI Risk and trustworthy
- Part 2 Framework Core

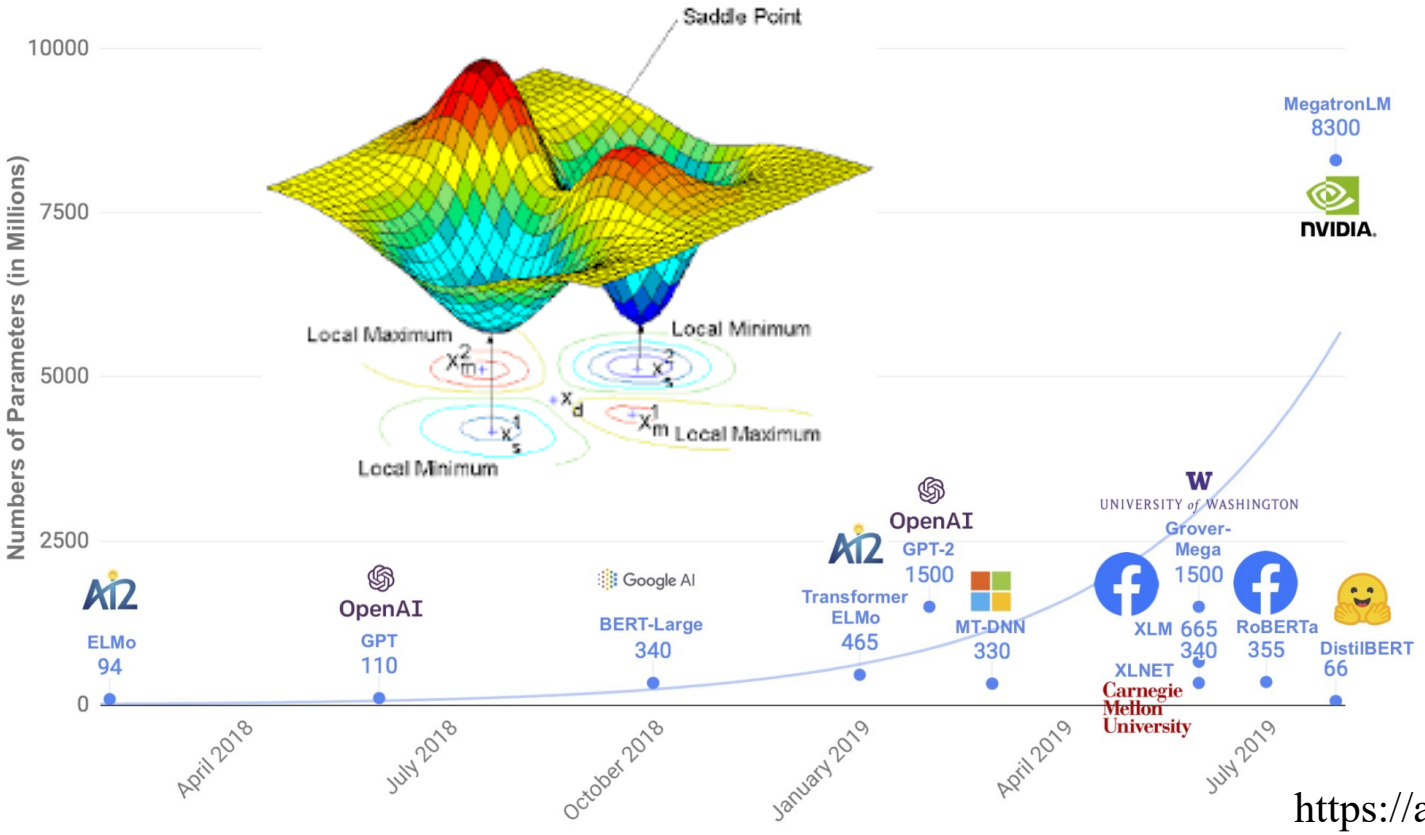
Motivation

- Managing AI risk towards **Trustworthy** and **Responsible AI**
  - **Trustworthy AI** is valid and reliable, safe, fair and bias is managed, secure and resilient, accountable and transparent, explainable and interpretable, and privacy-enhanced
  - **Responsible use and practice of AI systems** is a counterpart to AI system trustworthiness

- Risks to any software or information-based system apply to AI
  - including concerns related to cybersecurity, privacy, safety, and infrastructure

Motivation

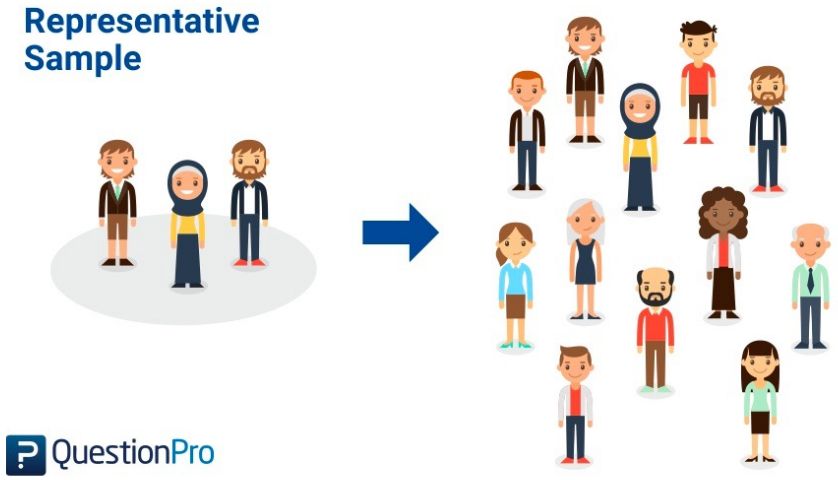
- New challenges:
  - A useful mathematical representation of the data interactions that drive the AI system's behavior is **not fully known**



https://aman.ai/papers/

Motivation

- Sources of Risks including
  - **Data used to train the AI system**
    - Data quality: inaccurate
    - Data not appropriate representation of the context

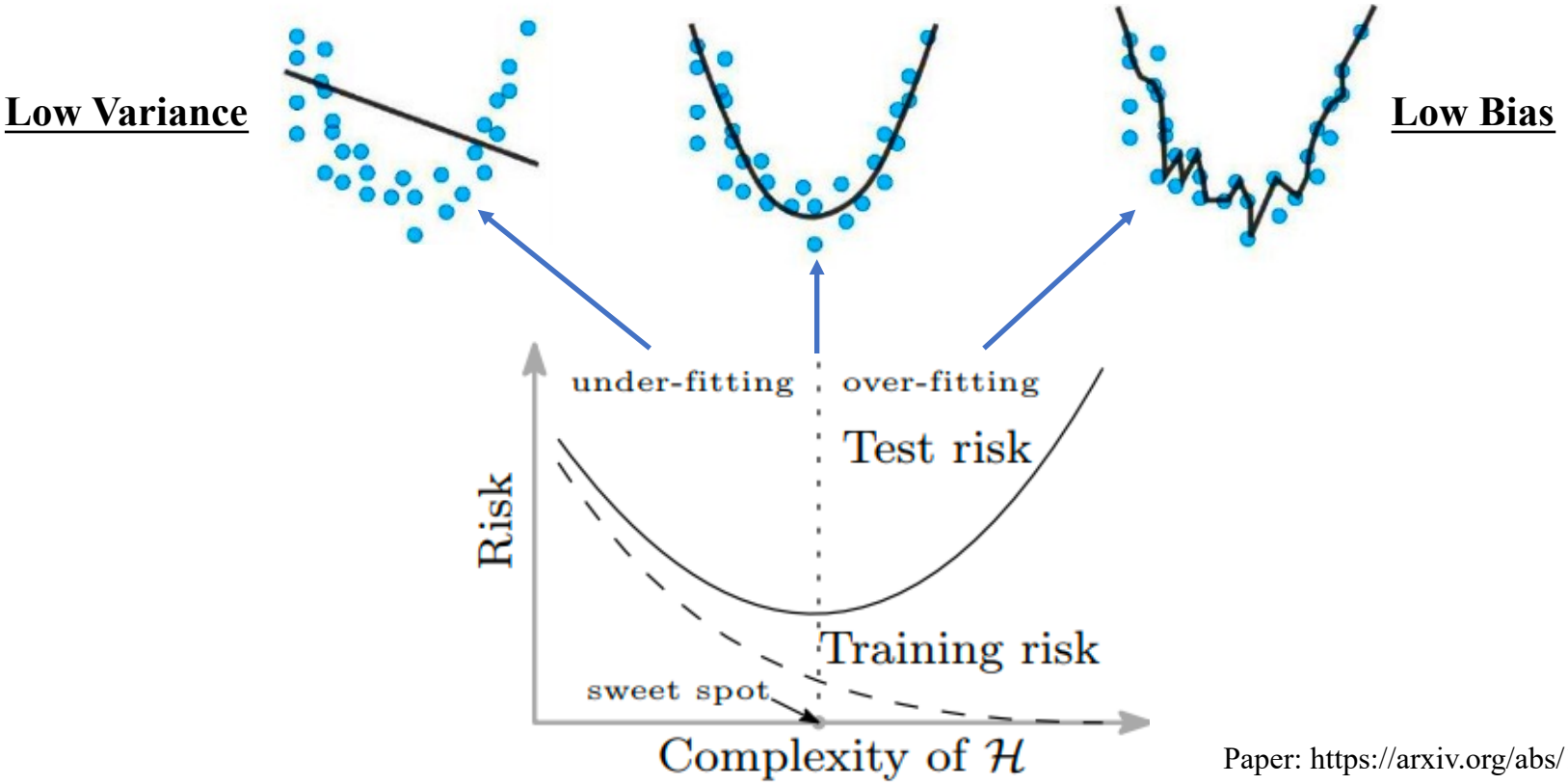


https://www.questionpro.com/blog/representative-sample/

L. Yang et al, doi: 10.1109/LRA.2020.2969932

Motivation

- Sources of Risks including
  - **AI system itself**
    - Overfitting/Underfitting, Instability

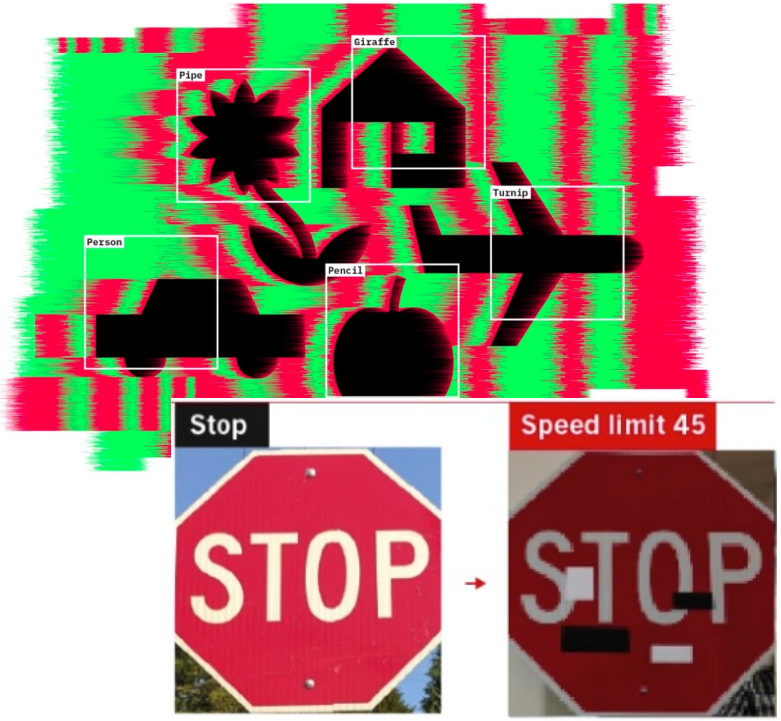


Paper: https://arxiv.org/abs/1812.11118

Motivation

- Sources of Risks including
  - **<u>Use of the AI system, or interaction of people with the AI system</u>**

Inappropriate use of narrow AI



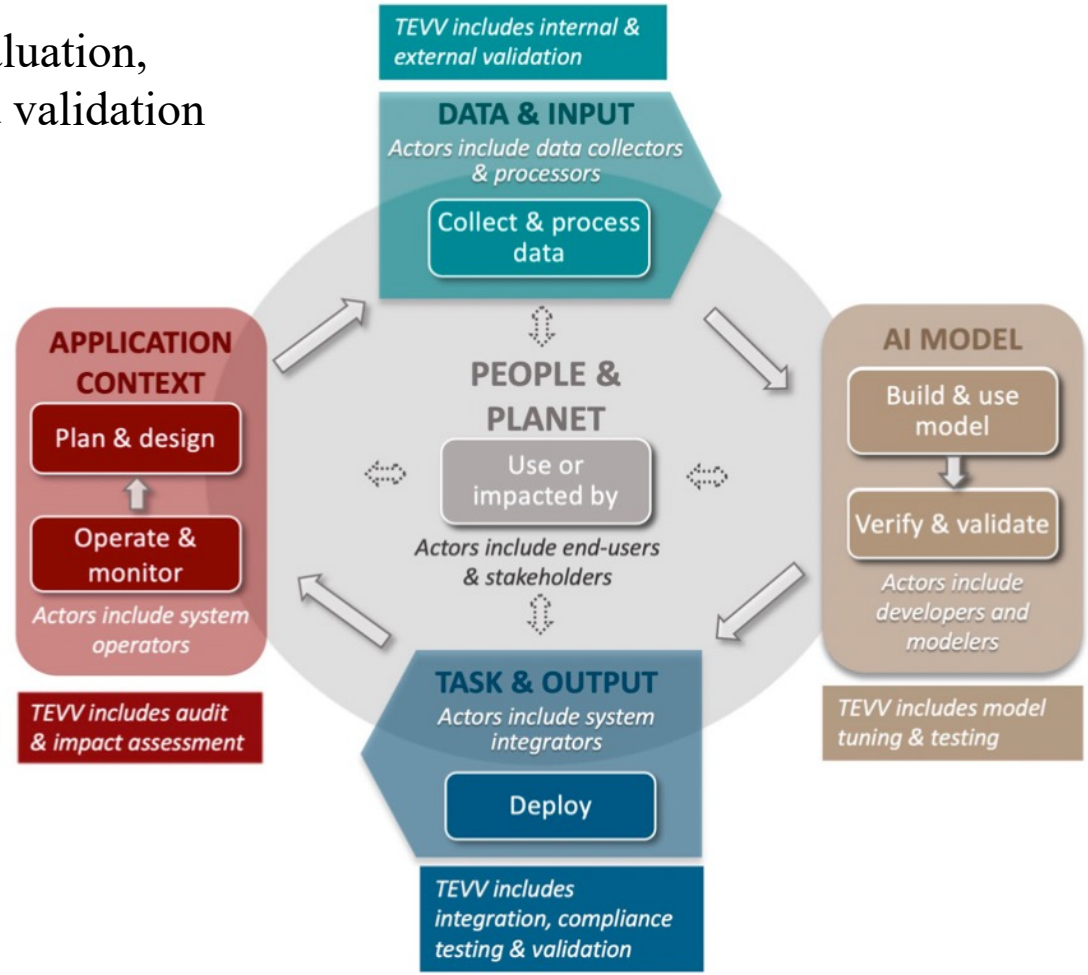
https://www.nature.com/articles/d41586-019-03013-5

Unintended, Malicious uses

Who should be involved?

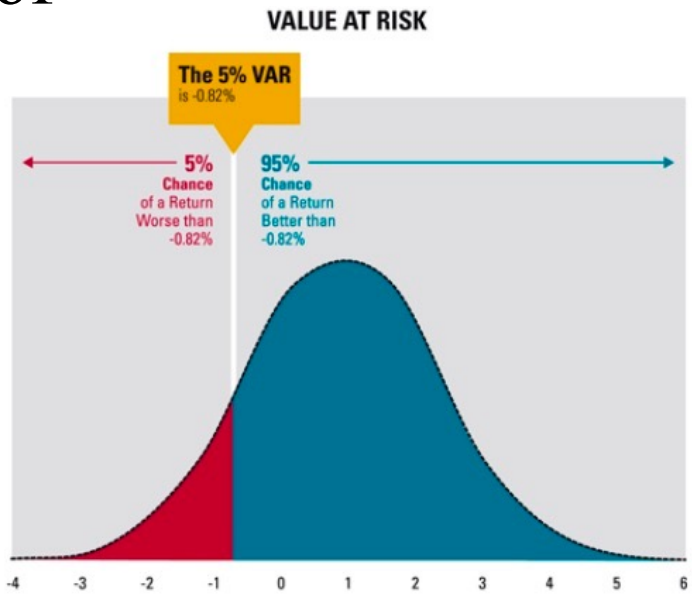
AI actors *who play an active role in the **AI system lifecycle**, including organizations and individuals that deploy or operate AI*

TEVV: test, evaluation, verification, and validation



Understanding Risk, Impacts, and Harms

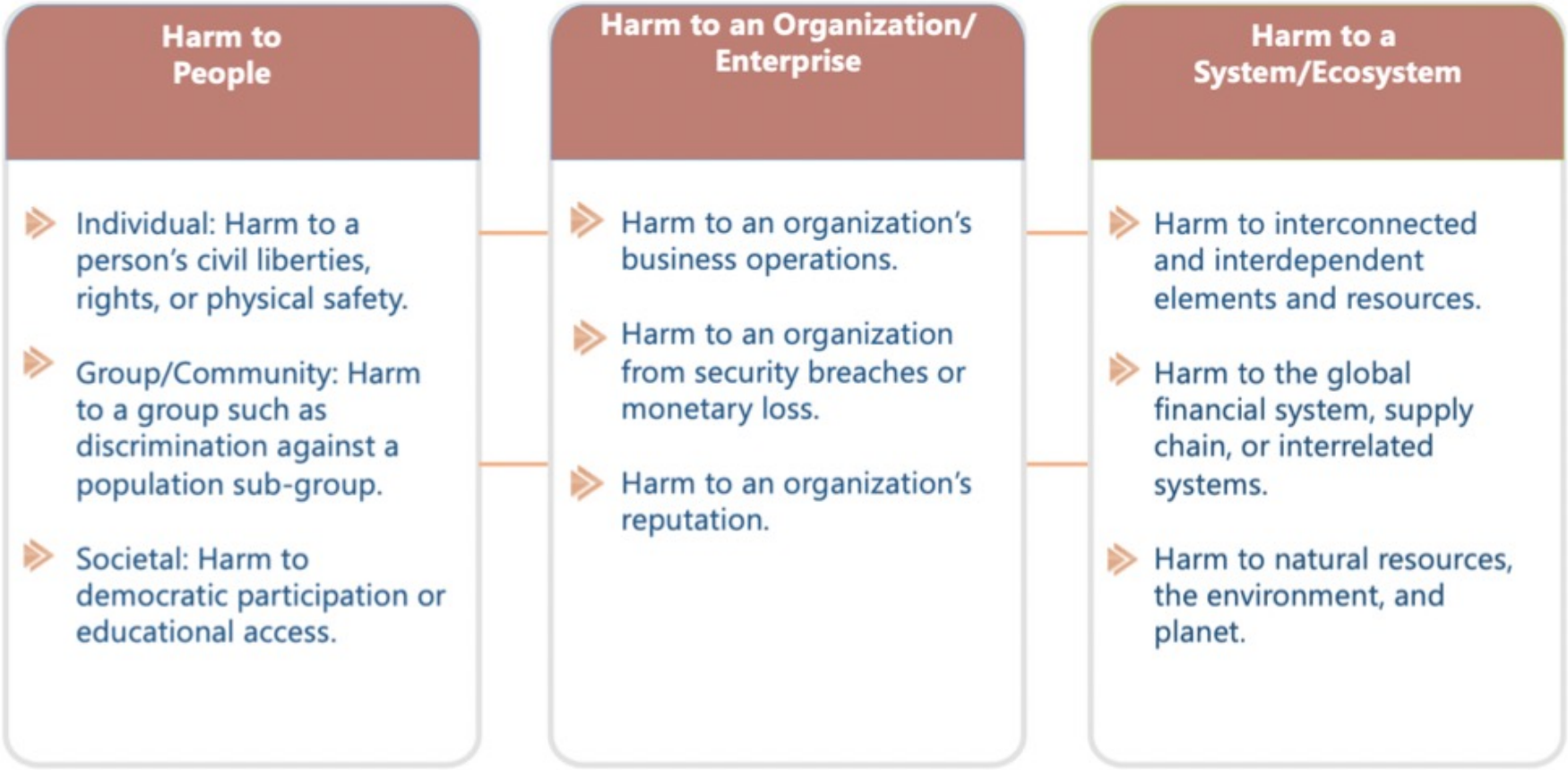
- Risk refers to the composite measure of
  - an event's probability of occurring
  - the magnitude of the consequences



- Risk management processes address negative impacts

- This framework offers approaches to
  - minimize anticipated negative impacts
  - identify opportunities to maximize positive impacts

Examples of potential harms related to AI systems



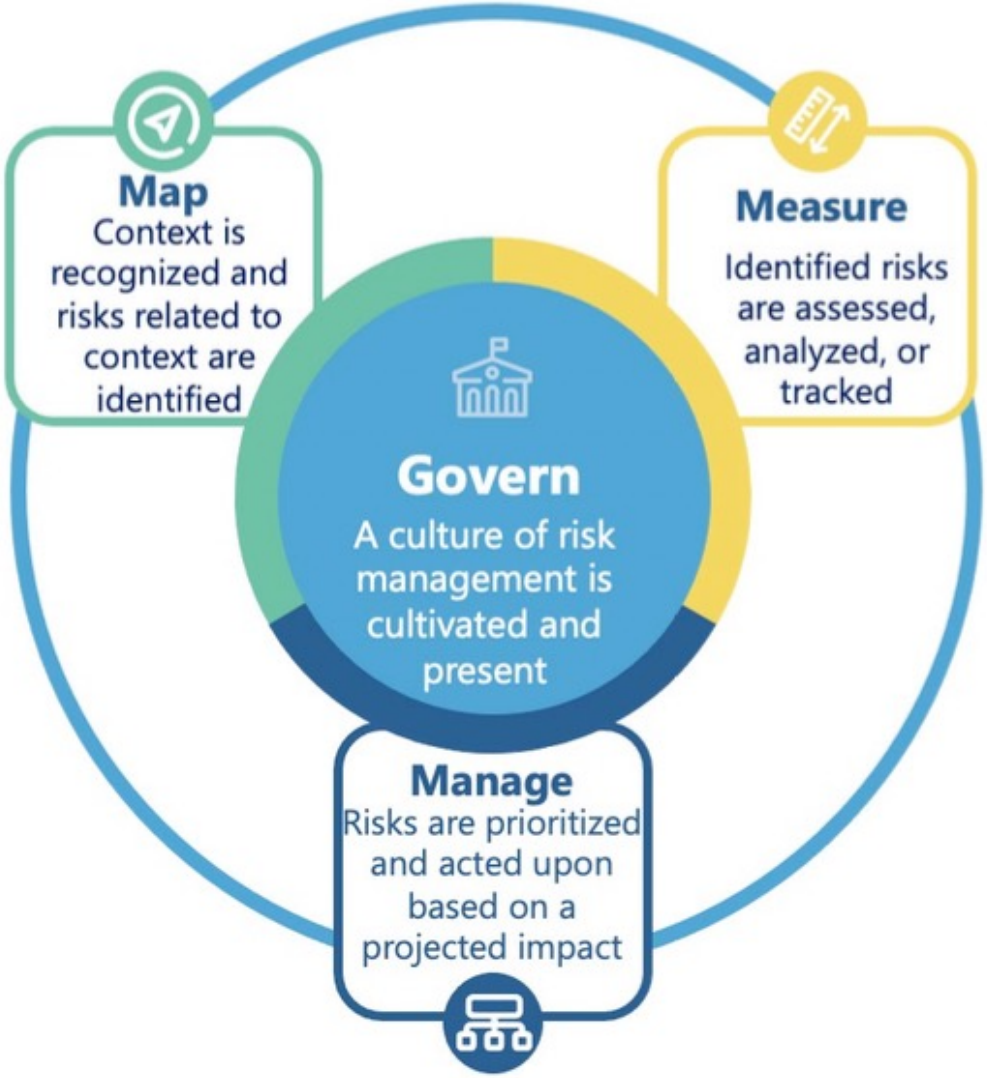
AI Risk and Trustworthiness

- These characteristics are tie to
  - human social and organizational behavior
  - the datasets used by AI systems
  - the decisions made by those who build them
  - the interactions with the humans who provide insight from and oversight of such systems
- Trustworthiness characteristics are **interrelated**
  - Tradeoffs are always involved
  - Highly secure but unfair, accurate but opaque and uninterpretable, and inaccurate but secure, privacy-enhanced, and transparent

Definition of AI trustworthy characteristics

Characteristics	Definition
Reliability	ability of an item to perform as required, without failure, for a given time interval, under given conditions
Robustness	ability of an AI system to maintain its level of performance under a variety of circumstances
Safety	should not, under defined conditions, cause physical or psychological harm or lead to a state in which human life, health, property, or the environment is endangered
Fairness	concerns for equality and equity but can be complex and difficult to define
Security	maintain confidentiality, integrity, and availability through protection mechanisms that prevent unauthorized access and use
…	…
Privacy	norms and practices that help to safeguard human autonomy, identity, and dignity

AI RMF Core


Map
Context is recognized and risks related to context are identified
Measure
Identified risks are assessed, analyzed, or tracked
Govern
A culture of risk management is cultivated and present
Manage
Risks are prioritized and acted upon based on a projected impact

MAP-1: Context is estabished and understood. +

MAP-2: Classification of the AI system is performed. −

🔗 MAP 2.1: The specific task, and methods used to implement the task, that the AI system will support is defined (e.g., classifiers, generative models, recommenders, etc.).

🔗 MAP 2.2: Information is documented about the system's knowledge limits, and how output will be utilized and overseen by humans.

🔗 MAP 2.3: Scientific integrity and TEVV considerations are identified and documented including related to experimental design, data collection and selection (e.g., availability, representativeness, suitability), and construct validation.

MAP-3: AI capabilities, targeted usage, goals, and expected benefits and costs compared with the status quo are understood. +

MAP-4: Risks and benefits are mapped for third-party software and data. +

MAP-5: Impacts to individuals, groups, communities, organizational, or society are assessed. +

Source

Exercise: What if tool from Google

"Diagnostic tool lets users try on five different types of fairness."

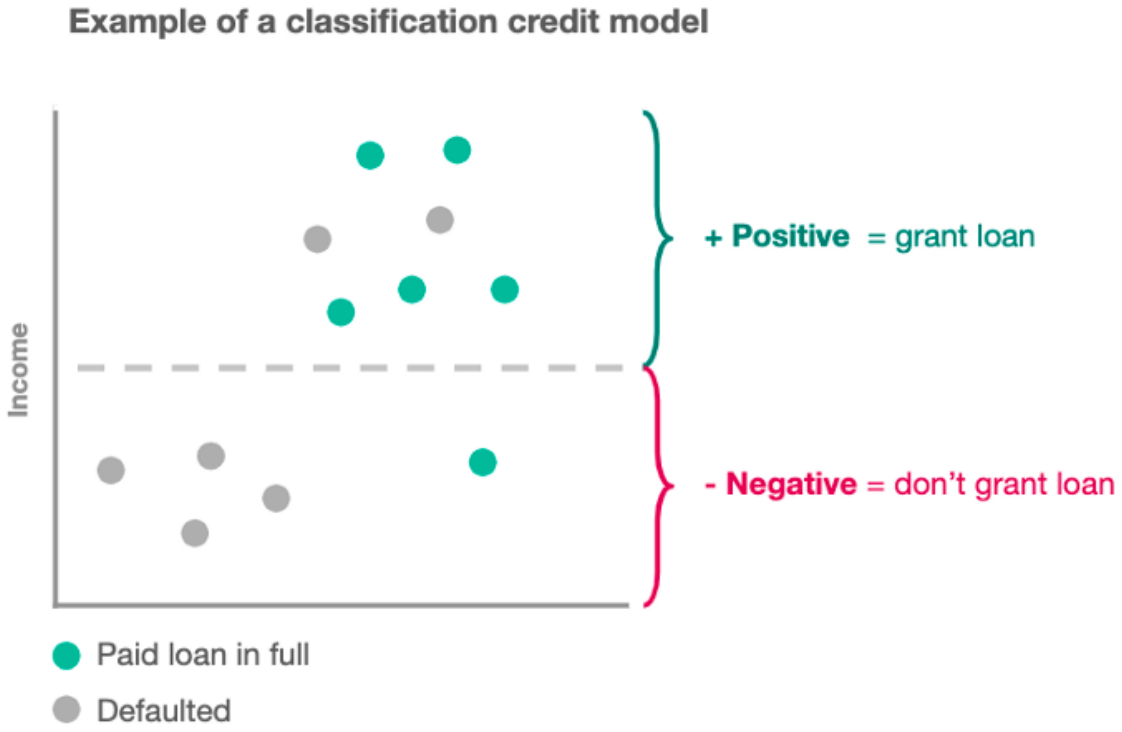
Web Demos: https://pair-code.github.io/what-if-tool/explore/#web

Types of fairness	Description
Group unaware	Disregard the different slices/groups
Group thresholds	Optimize a separate threshold for each slice based on the specified cost ratio.
Demographic parity	Similar percentages of datapoints from each slice are predicted as positive classifications.
Equal opportunity	Among those datapoints with the positive ground truth label, there is a similar percentage of positive predictions in each slice.
Equal accuracy	There is a similar percentage of correct predictions in each slice.

https://pair-code.github.io/what-if-tool/ai-fairness.html

Exercise: What if tool from Google

Types of fairness
Group unaware
Group thresholds
Demographic parity
Equal opportunity
Equal accuracy



Set a single income threshold in our training set to decide who gets a loan in the future (the dotted line)

Source

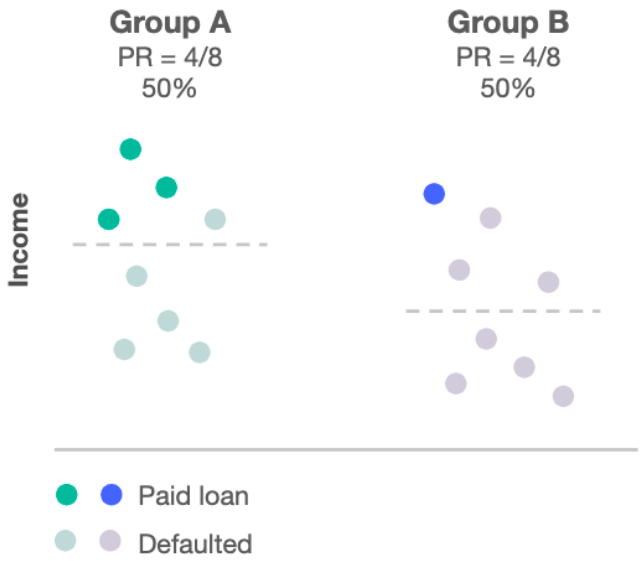
Exercise: What if tool from Google

Types of fairness
Group unaware
Group thresholds
Demographic parity
Equal opportunity
Equal accuracy

Source

Exercise: What if tool from Google

Types of fairness
Group unaware
Group thresholds
Demographic parity
Equal opportunity
Equal accuracy



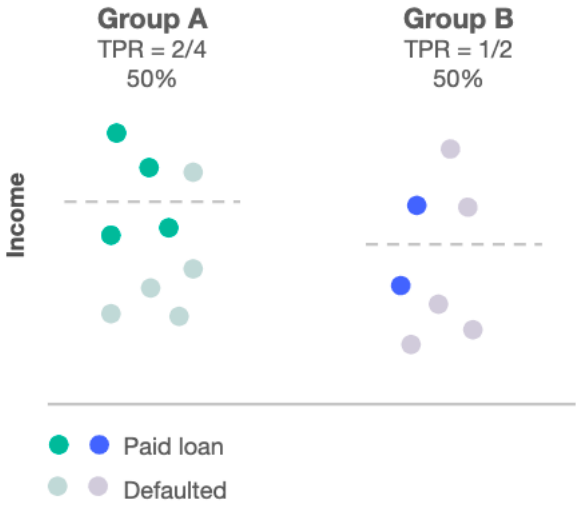
When to use Demographic Parity

- We want to change the state of our current world to improve it (e.g.: we want to see more minority groups getting to the top)
- We are aware of historical biases may have affected the quality of our data (e.g.: ML solution trained to hire software engineers, where nearly no women was hired before)
- We have a plan in place to support the unprivileged group

Source

Exercise: What if tool from Google

Types of fairness
Group unaware
Group thresholds
Demographic parity
Equal opportunity
Equal accuracy



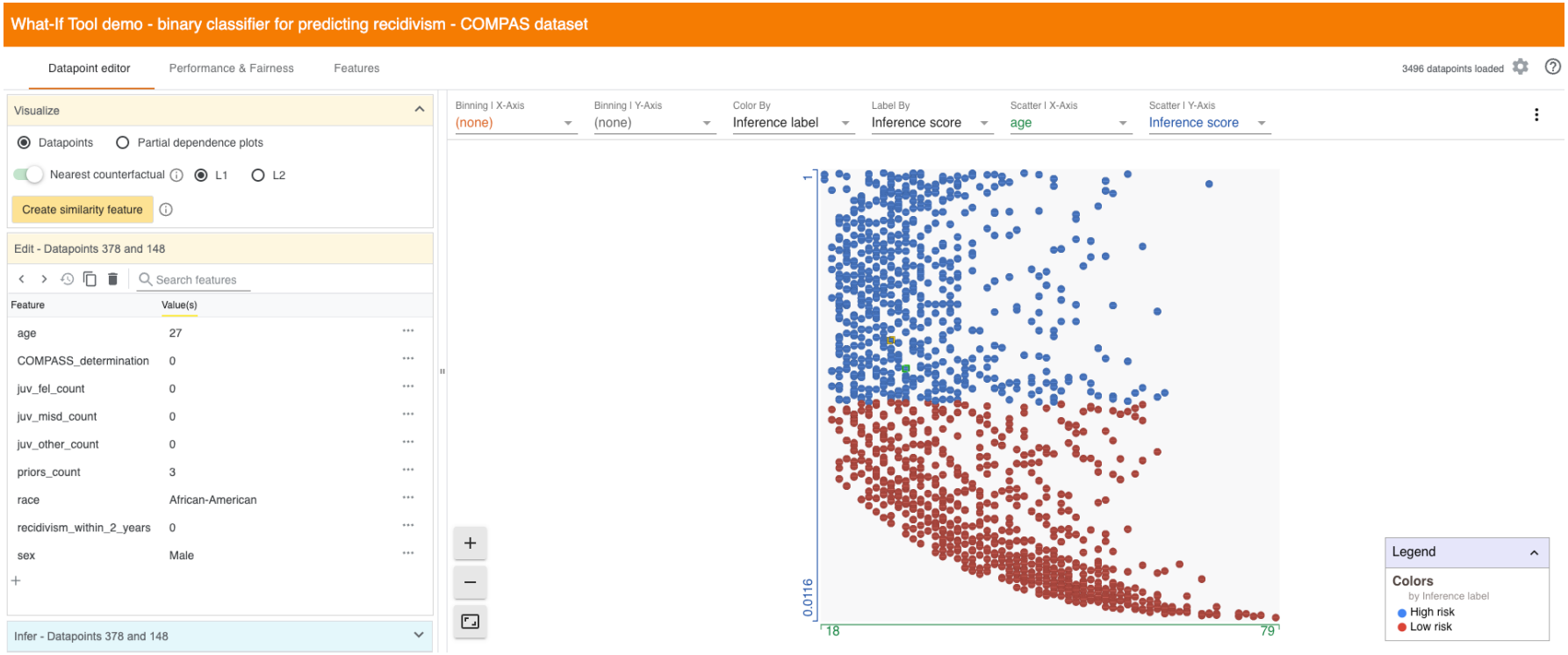
When to use Demographic Parity

- There is a strong emphasis on predicting the positive outcome correctly (e.g.: we need to be very good at detecting a fraudulent transaction)
- Introducing False Positives are not costly to the user nor the company (e.g.: wrongly notifying a customer about fraudulent activity will not be necessarily expensive to the customer nor the bank sending the alert)

Source

Investigate fairness on recidivism classification:

- COMPAS (Correctional Offender Management Profiling for Alternative Sanctions) is a popular commercial algorithm used by judges and parole officers for scoring criminal defendant's likelihood of reoffending (recidivism).
- Website: https://pair-code.github.io/what-if-tool/demos/compas.html



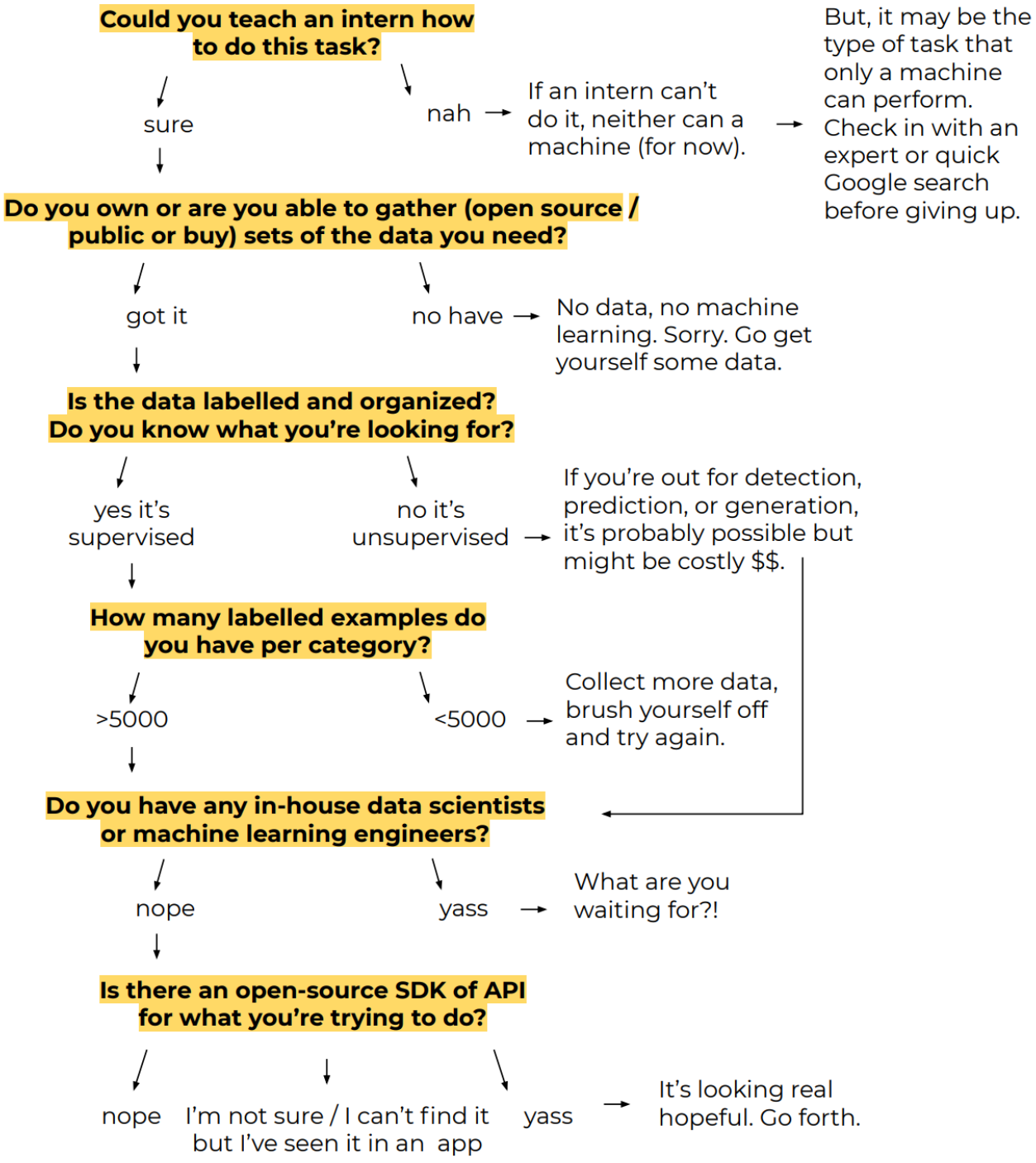
Feature Value	Count	Threshold	False Positives (%)	False Negatives (%)	Accuracy (%)	F1
African-American	1904	0.49	14.8	12.9	72.3	0.80
Caucasian	1111	0.49	7.6	18.9	73.5	0.64
Hispanic	305	0.49	3.6	16.1	80.3	0.66
Other	157	0.49	3.8	10.8	85.4	0.47
Asian	11	0.49	0.0	9.1	90.9	0.67
Native American	8	0.49	0.0	12.5	87.5	0.93

Lecture
AI Meets Design

Wan Fang
Southern University of Science and Technology

Tool:
Assessing feasibility for idea selection

Use this flowchart to quickly assess how feasible / viable your AI idea is



Activity:
Framing your task for concept development

Google HCML team speak from experience when they say:
"Find experts who can be the best possible teachers for your machine learner—people with domain expertise relevant to whatever predictions you're trying to make. We recommend that you actually hire a handful of them, or as a fallback, transform someone on your team into the role. We call these folks "content specialists" on our team."

The strength of machine learning is that we don't have to program the rules explicitly. At this stage of the process, it is helpful to think about them and try construct a logic based on how we humans perform the task.

1. Start with the classic exercise: describe the way a human expert would perform the task or answer the question.

   If you were to ask 10 people, would they agree on the method (for the most part)? If some do it better or differently - what can we learn from their approach?

   Especially if what you're predicting is (highly) subjective, spend extra time on this step.

2. Imagine you're onboarding a new person for this job. What do they need to understand? What assumptions would you want them to make? How would you respond so they improve over time?

3. What's the nature of the task? Can you box it as an clustering, classification, or regression problem? Refer back to the crash course in the beginning of this toolkit to find the vocabulary. Knowing this will help you understand the task as well as communicate with your tech team.

In the example of Spotify's Discover Weekly, **the human expert** would be a music lover on the hunt for new music.

Do you have data of past well-executed and completed tasks? This could be used as an initial training data set.

Tip:
Draw a diagram of the current workflow including IFTT statements and data required to make decisions.

Activity:
Plotting your model for concept development

By plotting a simple flowchart, we can begin forming a rough idea of the inputs, outputs, and logic required for our model to create value. We're also surfacing our assumptions and unknowns in the process.

1. **Objective** - What is the question we're trying to answer and asking the machine?
   Output - How is the machine's answer presented and interpreted?

2. **Features** - What data points do you need or are important factors in answering the question?
   Input - Which data sets does that data reside in? What data will the model be trained on? What data does the user input?
   + Draw connections between the assumed features and data sets they reside in.

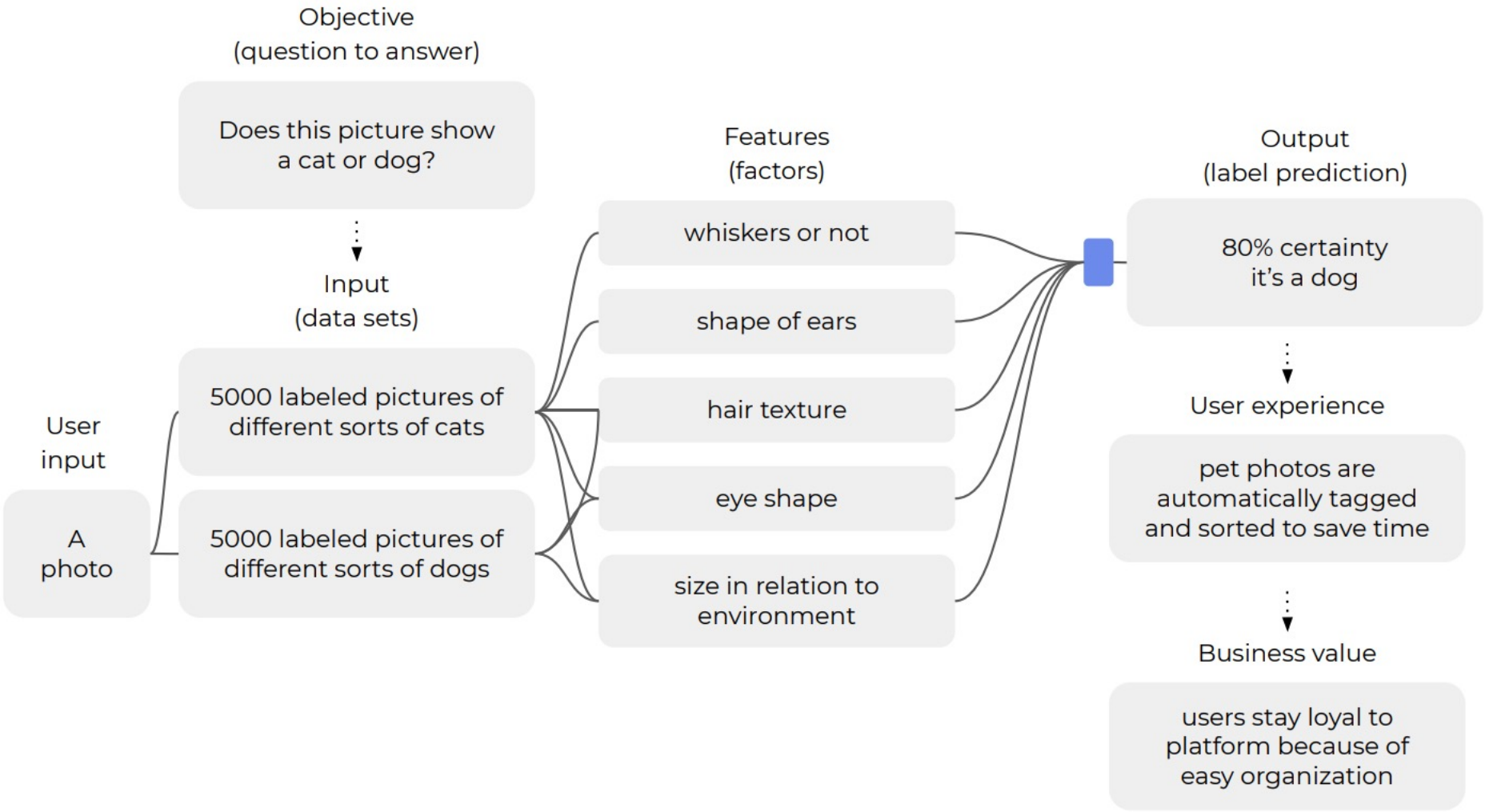
3. **User experience** - How does the outcome get presented to and help the user?
   Business value - How does the solution return value to the organization?

AI answers (mostly) in probabilities with a confidence level. Formulate your output as a **probability**.

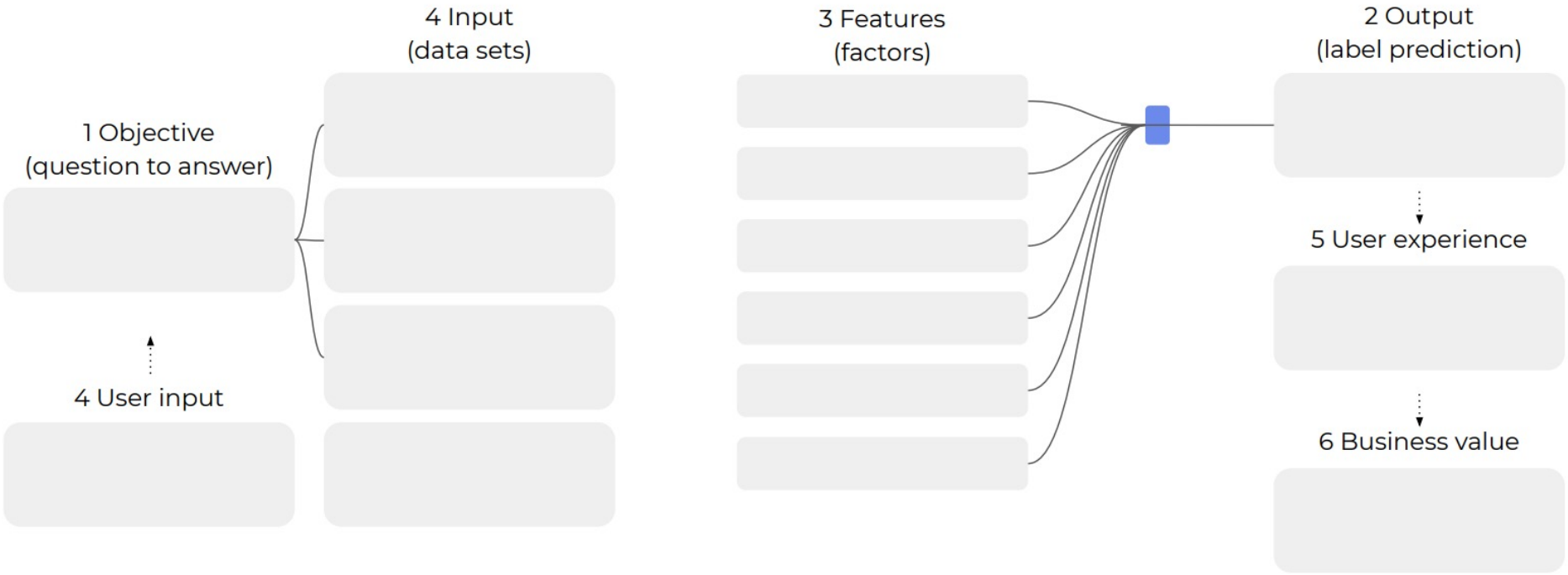
Do you know which features go into the answer? Think about the variables and patterns humans look at when performing this task or answering this question.

Do you have this **data to input**? If not, how do you acquire it?

Activity:
Plotting your model for concept development



Worksheet:
Plotting your model



1 Objective
What is the question we're trying to answer and asking the machine?

2 Output
How is the machine's answer presented and interpreted?

Formulate your output as a probability.

3 Features
What data points do you need or are important factors in answering the question?

Do you know which features go into the answer? Think about the variables and patterns humans look at when performing this task or answering this question.

4 Input
Which data sets does that data reside in? What data will the model be trained on? What data does the user input?

Do you have this data to input? If not, how do you acquire it?

+ Connect
Draw connections between the assumed features and data sets they reside in.

5 User experience
How does the outcome get presented to and help the user?

6 Business value
How does the solution return value to the organization?

Prototyping + testing

You're with a handful of ideas and it's time to get more in-depth with your user research. Through prototyping and testing, you (in)validate your AI ideas and their design and implementation specs.

Do users want and need your solution? Are they open to adoption? Are they willing to share data and invest themselves into training the model (if necessary)? How can we test rather than just ask? How can we prototype the experience of adaptive intelligent systems?

In this chapter you will find:

User research & feedback
to know what to inquire about in addition to the usual

Prototyping & testing
to explore how to prototype and test AI applications

Activity:
User research & feedback for assessing desirability

1

Assuming you did initial user research to inform your concepts so far, now it's time to go out and (in)validate your value proposition in more detail. First assess your need as you do for any problem, asking:

- What problem does it solve or opportunity does it tap into?
- Who benefits and in what scenario?
- How pressing is the problem? For how many?
- What do they gain from the new solution? How and how much better is it than the current solution? What other advantages do they see?

Activity:
User research & feedback for assessing desirability

Iterate on your value proposition statement based on your learnings and get ready to prototype for deeper insights.

2

Once you've validated that this is indeed a problem worth solving, gather insights about your users' perspective on the AI aspects of your concept(s).

Mental models
What are their notions about having an intelligent, adaptive system work for them? Are they willing to adopt it? How important is transparency? Depending on how visible your AI elements are, this might be more or less important.

Defining success and failure
How accurate must the model be to offer user value? How high are the costs of mistakes? What would best vs worst behavior look like?

Machine teaching
What does the user need to invest to get value out of the system? Are they willing to share the data your model needs? Are they willing to provide the necessary feedback and teach the model?

Ethical & experiential concerns
What concerns do they have? Do major ethical concerns arise? Unintended consequences, edge cases, and extreme users?

Activity:
Prototyping & testing
for assessing desirability

1

Prototype

To test desirability, opt to simulate the experience without building the model and observing the responses.

Testing the concept offering can be done with product / service posters or app marketplace.

Common prototyping techniques for AI are:
Role playing
Wizard of Oz
Personalized wireframes.

Where possible, gather and use real-life personal data in your prototypes rather than placeholder content.

Provotypes (prototypes that provoke) can also be a great way to build an understanding of your users' needs.

"Fake it till you make it. If forced to choose, it's leaps-and-bounds more useful to prototype your UX with a user's real content than it is to test with real ML models - as it affords you genuine insights into the way people will derive value and utility from your (theoretical) product."

by Google Clips' team on UX of AI

Activity:
Prototyping & testing
for assessing desirability

2

Testing

Do user testing as usual and observe users' behavior. Ask them to think out loud as they're interacting with your artefact.

Keep in mind that while testing is important to understand your user, working with adaptive systems requires the designer to sacrifice a certain level of control over the final user experience exactly because it will adapt to each user and over time.

3

Analysis & selection

Analyse and synthesize your findings. Based on all your findings, decide which idea(s) (if any) to move forward with.

It can help to revisit some of the activities in idea selection phase and reconsider feasibility, viability, desirability, and responsibility.

Day 02
AI Meets Design II

Thank you~

Wan Fang
Southern University of Science and Technology