

Setting up ML Projects

Machine Learning Projects



IN CS, IT CAN BE HARD TO EXPLAIN
THE DIFFERENCE BETWEEN THE EASY
AND THE VIRTUALLY IMPOSSIBLE.

Machine Learning Projects

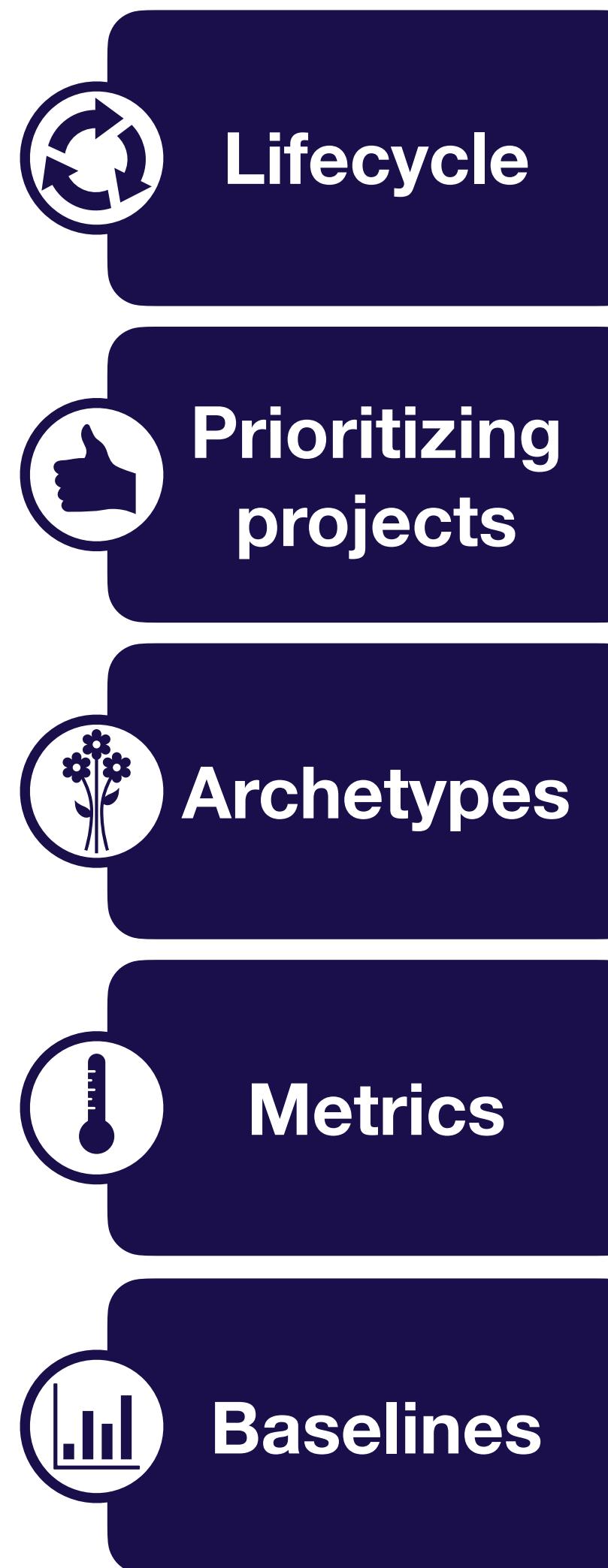
85% of AI projects fail¹

¹ Pactera Technologies

Why do so many projects fail?

- ML is still research - you shouldn't aim for 100% success rate
- But, many are doomed to fail:
- Technically infeasible or poorly scoped
- Never make the leap to production
- Unclear success criteria
- Poor team management

Module overview



- **How to think about all of the activities in an ML project**
- **Assessing the feasibility and impact of your projects**
- **The main categories of ML projects, and the implications for project management**
- **How to pick a single number to optimize**
- **How to know if your model is performing well**

Running case study - pose estimation

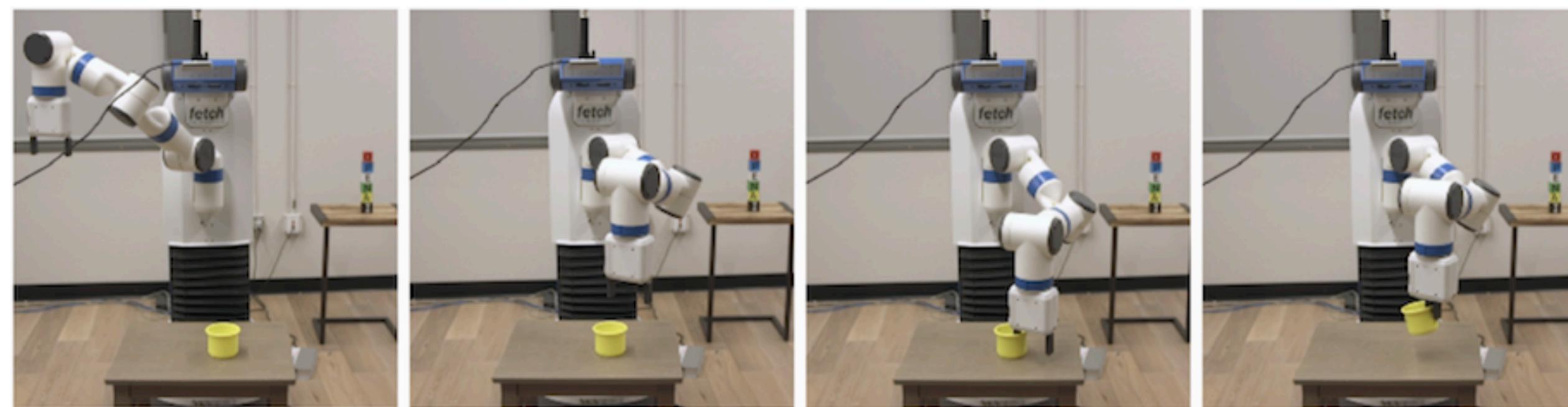
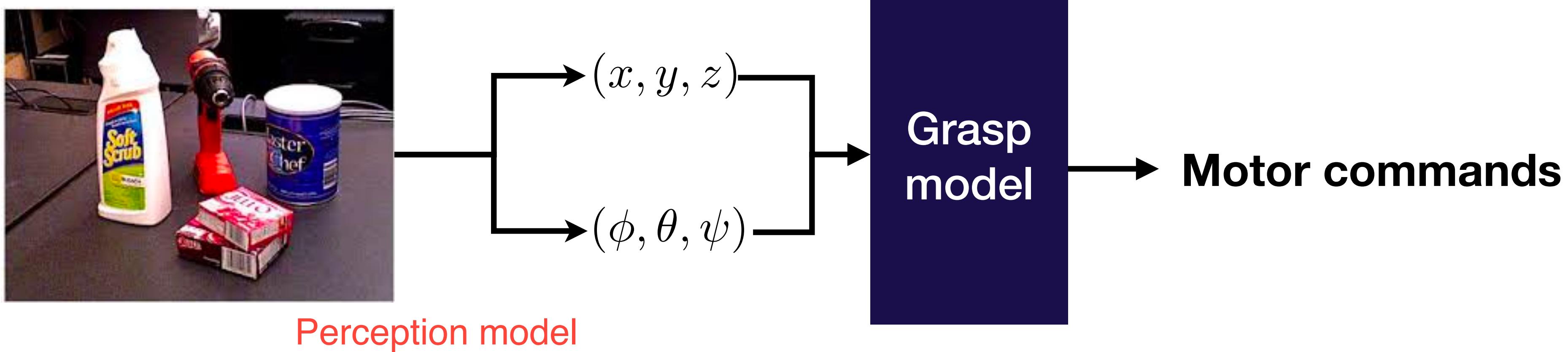


(x, y, z) **Position (L2 loss)**

(ϕ, θ, ψ) **Orientation (L2 loss)**

Xiang, Yu, et al. "PoseCNN: A Convolutional Neural Network for 6D Object Pose Estimation in Cluttered Scenes." *arXiv preprint arXiv:1711.00199* (2017).

Full Stack Robotics works on grasping



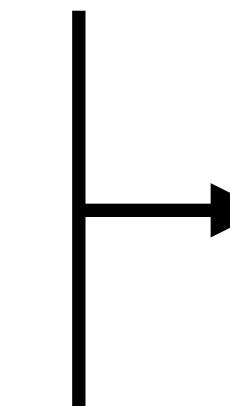
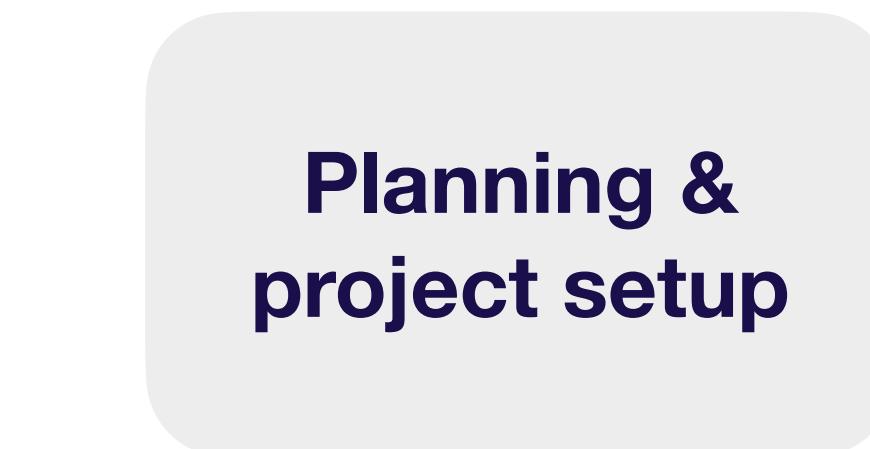
Module overview

- Lifecycle**
 - How to think about all of the activities in an ML project
- Prioritizing projects**
 - Assessing the feasibility and impact of your projects
- Archetypes**
 - The main categories of ML projects, and the implications for project management
- Metrics**
 - How to pick a single number to optimize
- Baselines**
 - How to know if your model is performing well

Lifecycle of a ML project

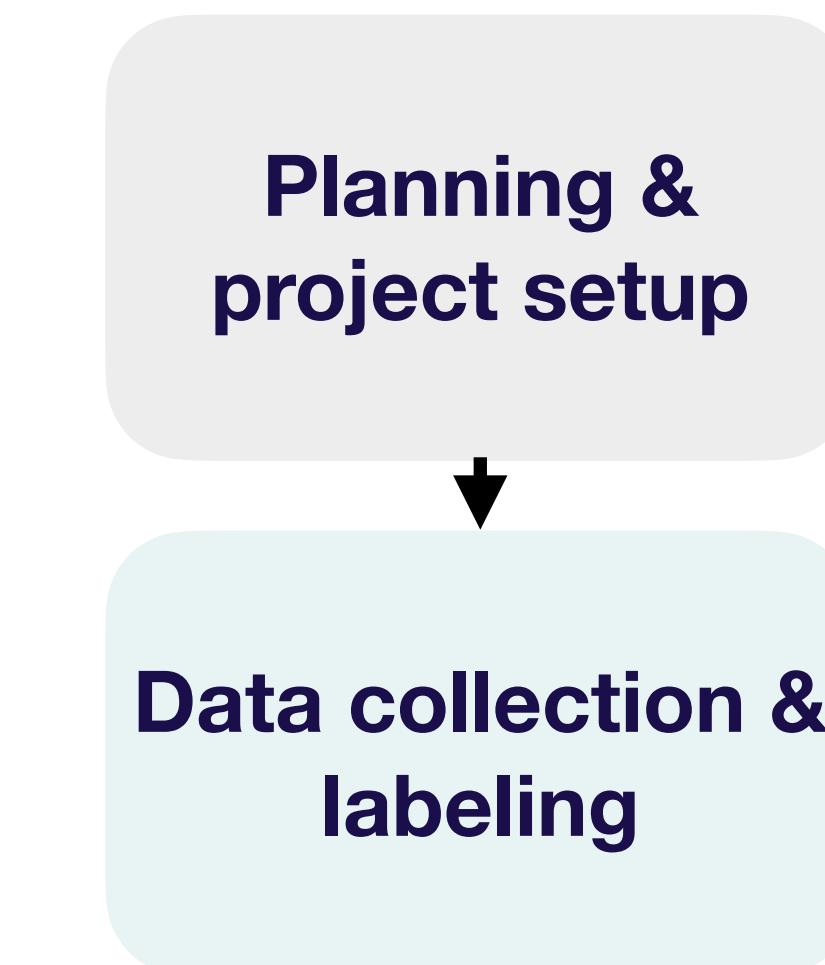
Planning &
project setup

Lifecycle of a ML project

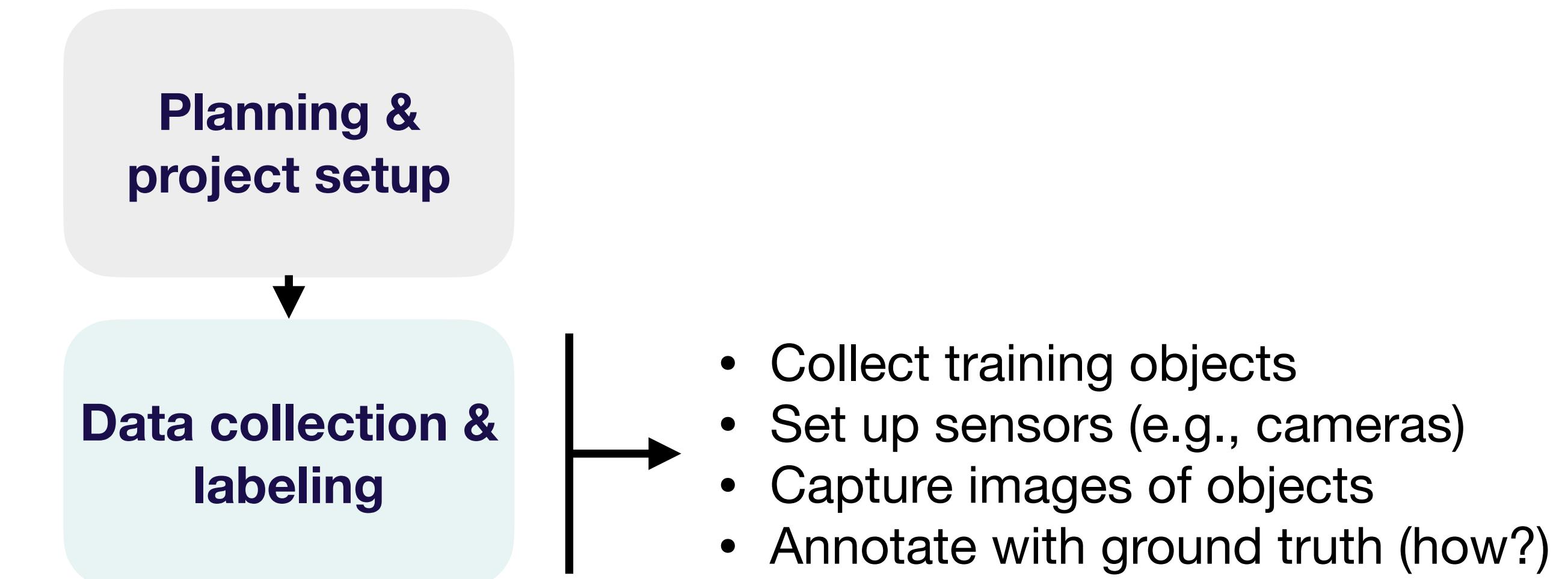


- Decide to work on pose estimation
- Determine requirements & goals
- Allocate resources
- Consider the ethical implications
- Etc.

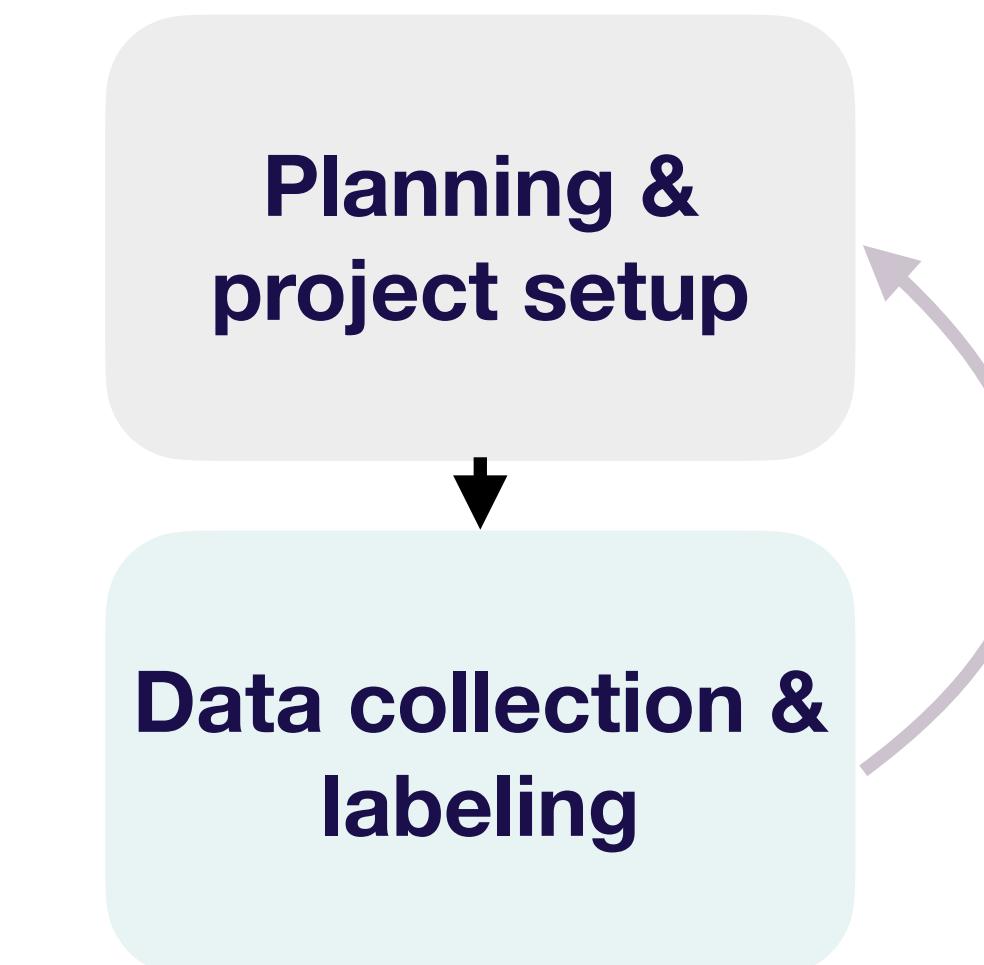
Lifecycle of a ML project



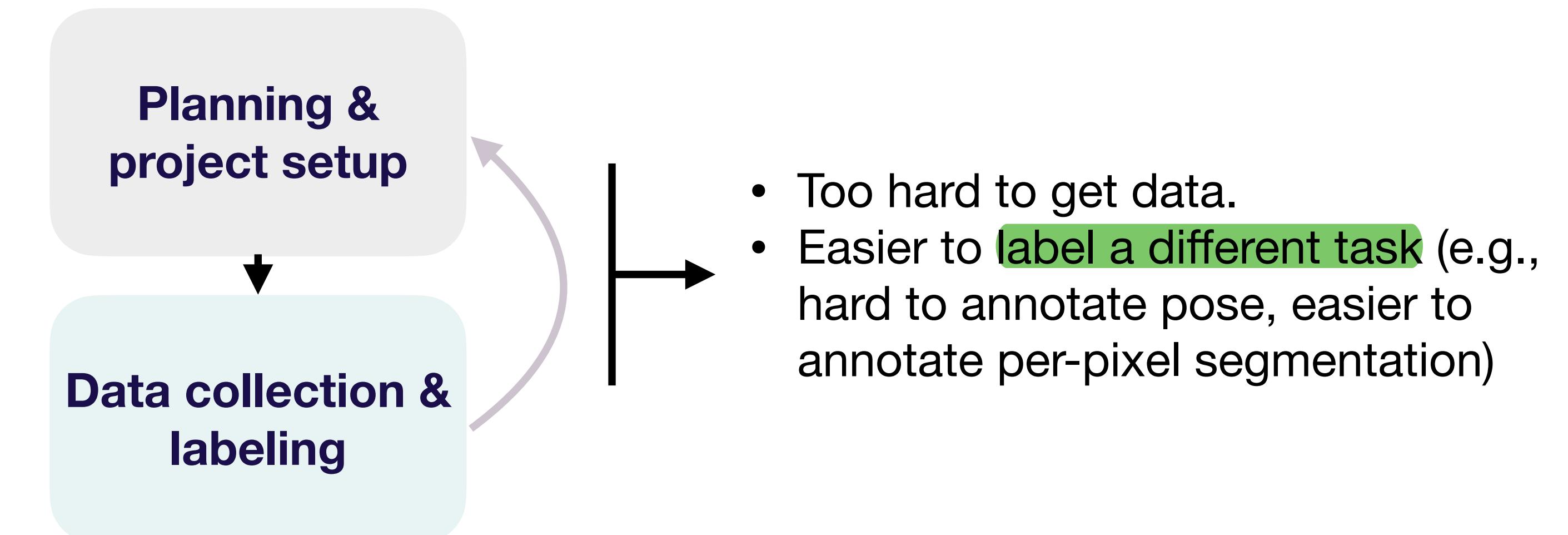
Lifecycle of a ML project



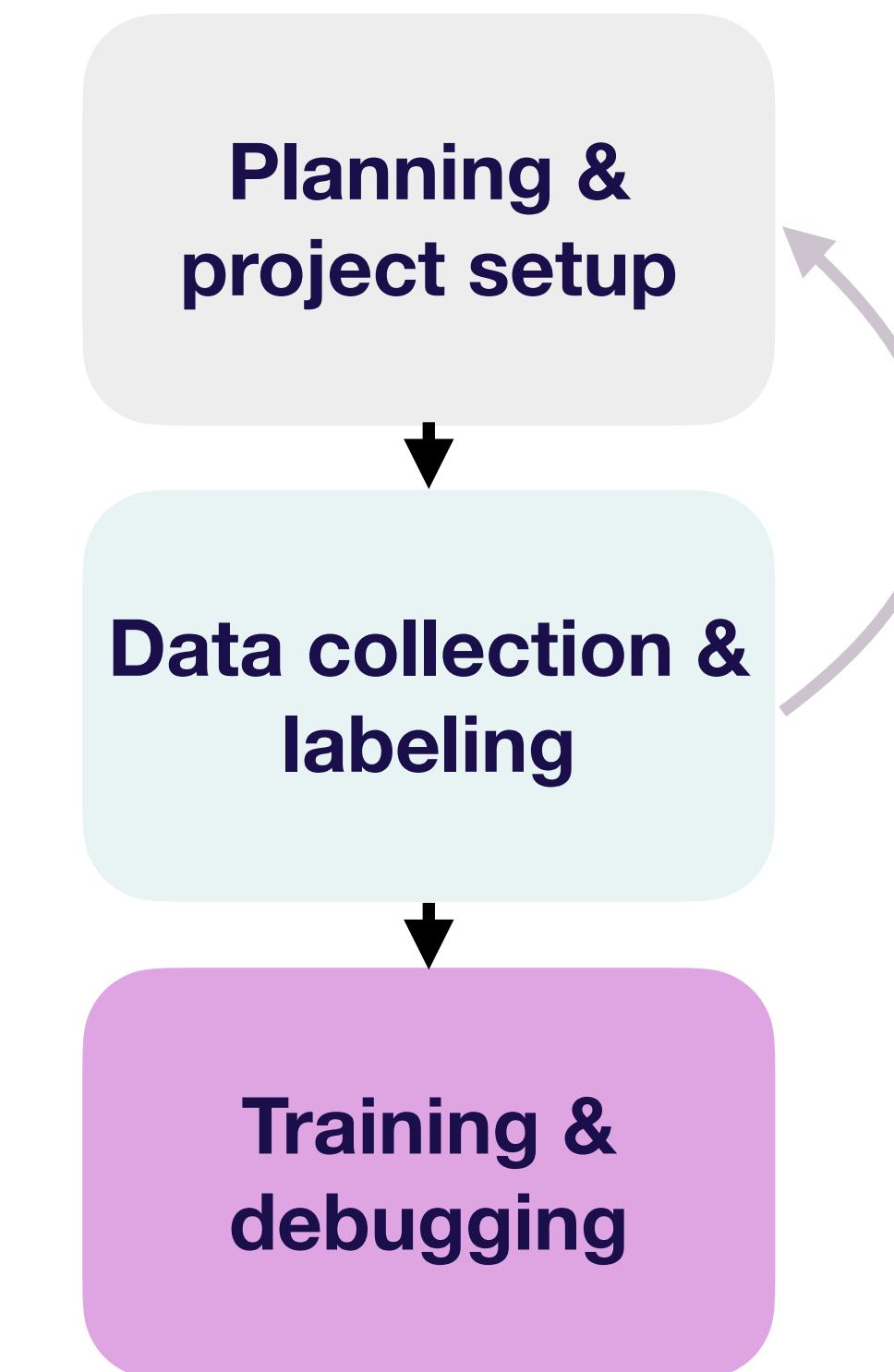
Lifecycle of a ML project



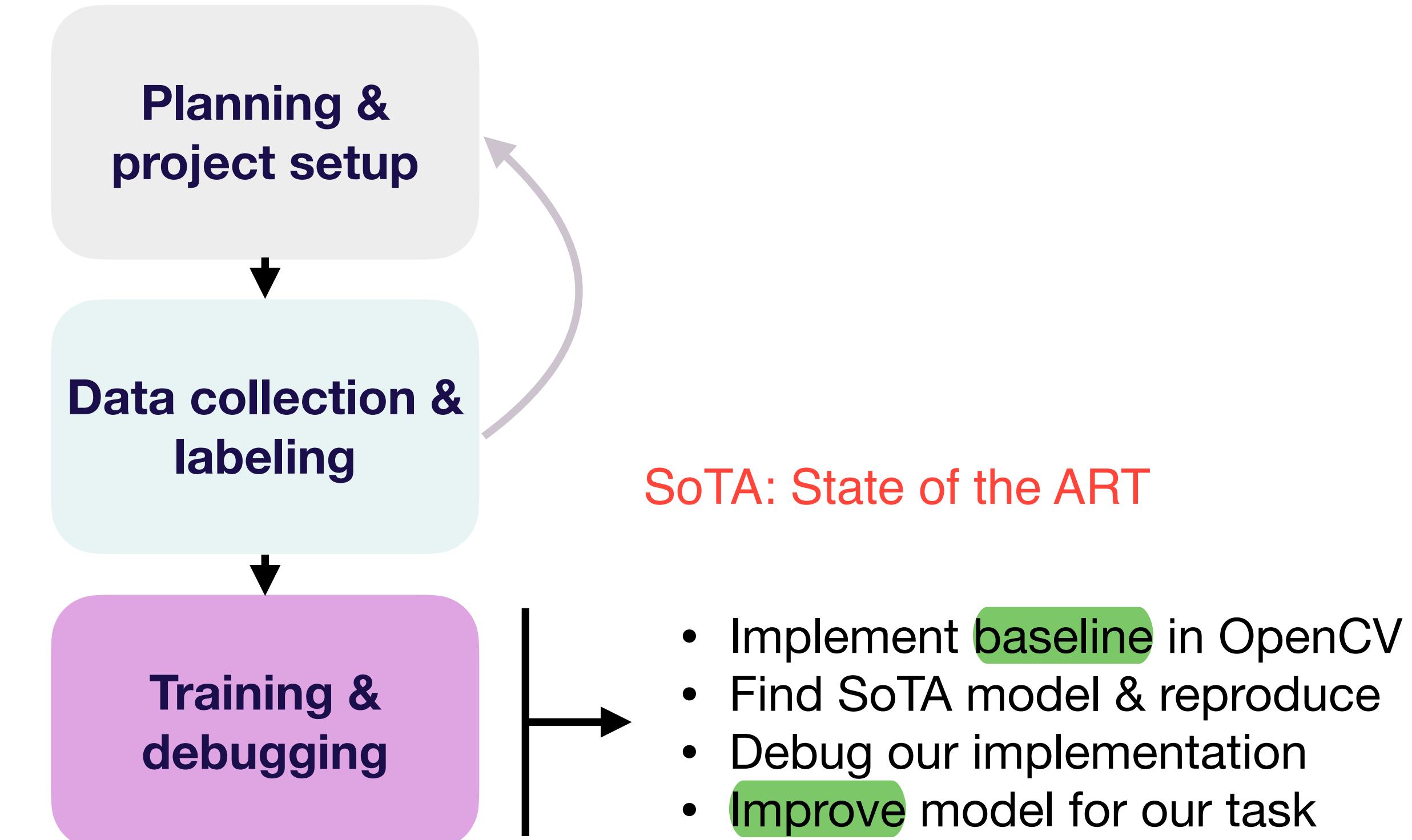
Lifecycle of a ML project



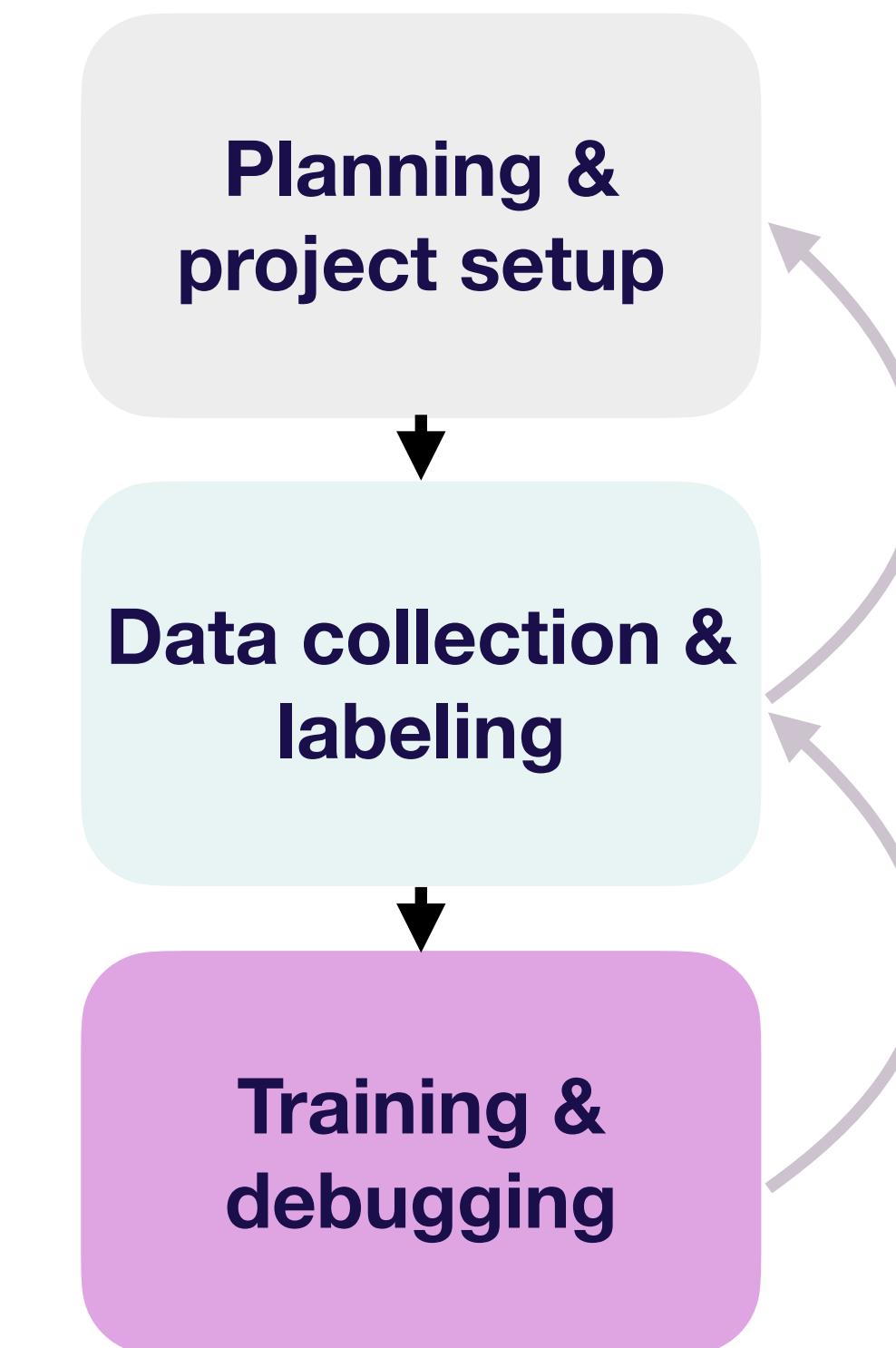
Lifecycle of a ML project



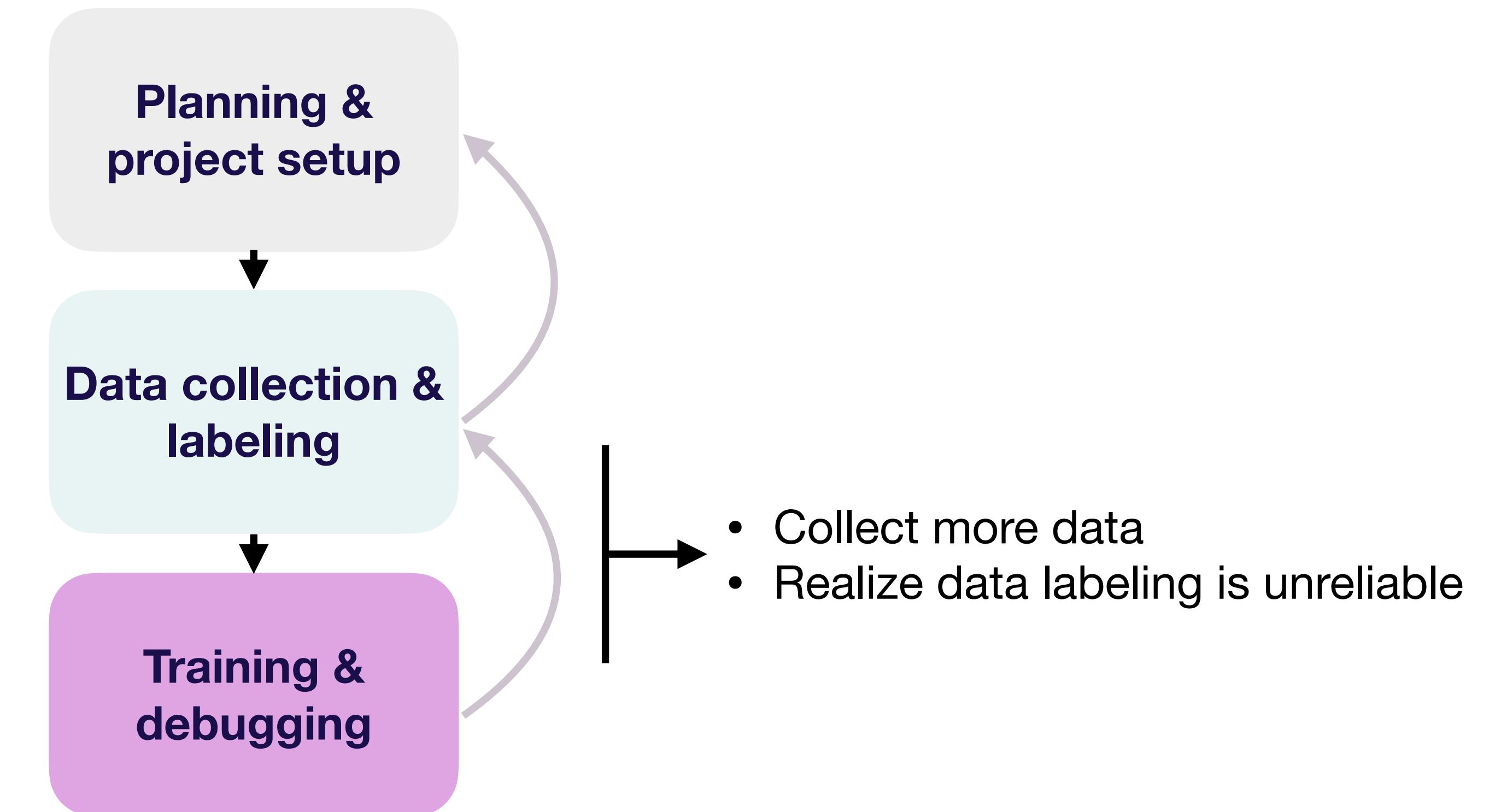
Lifecycle of a ML project



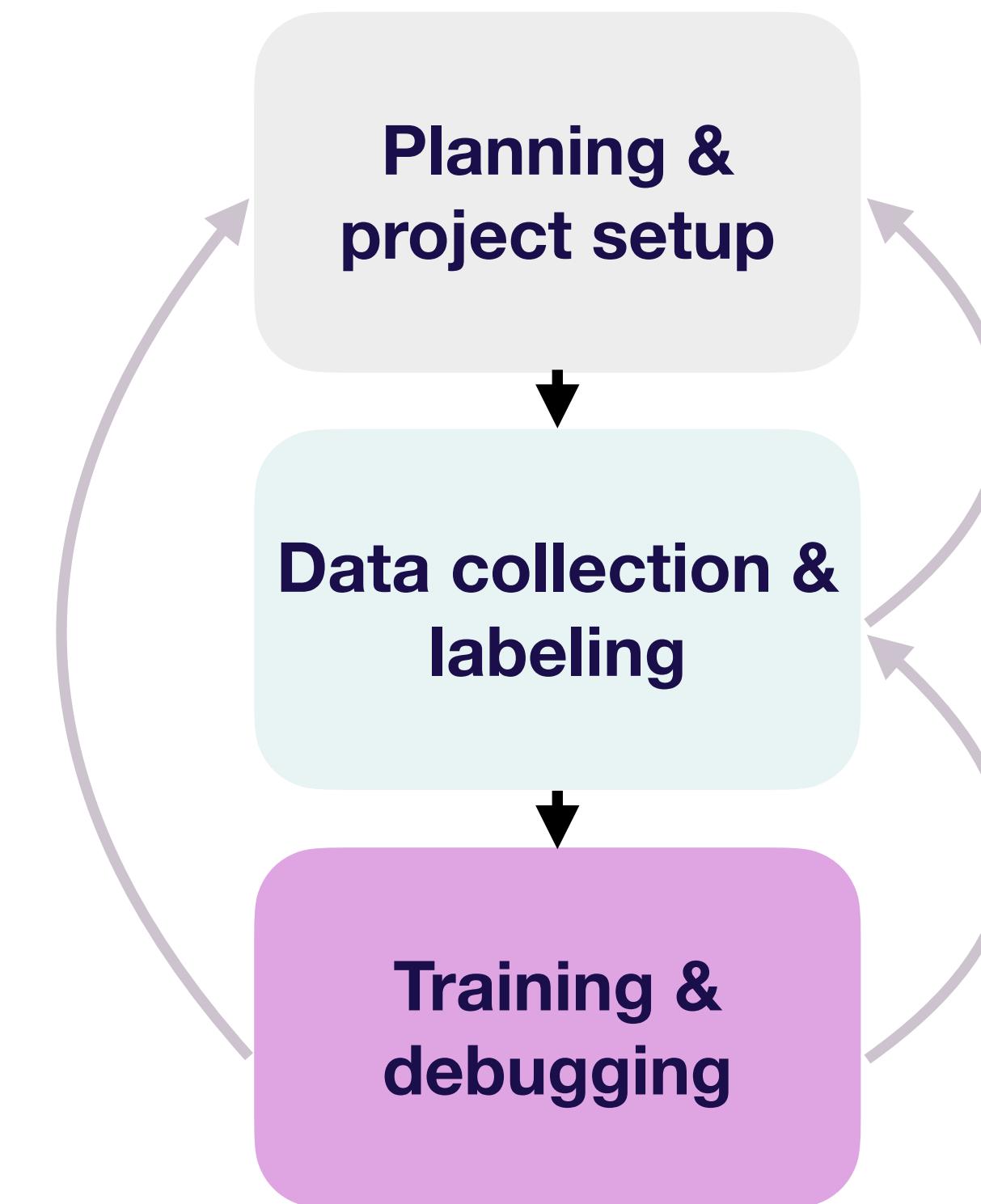
Lifecycle of a ML project



Lifecycle of a ML project

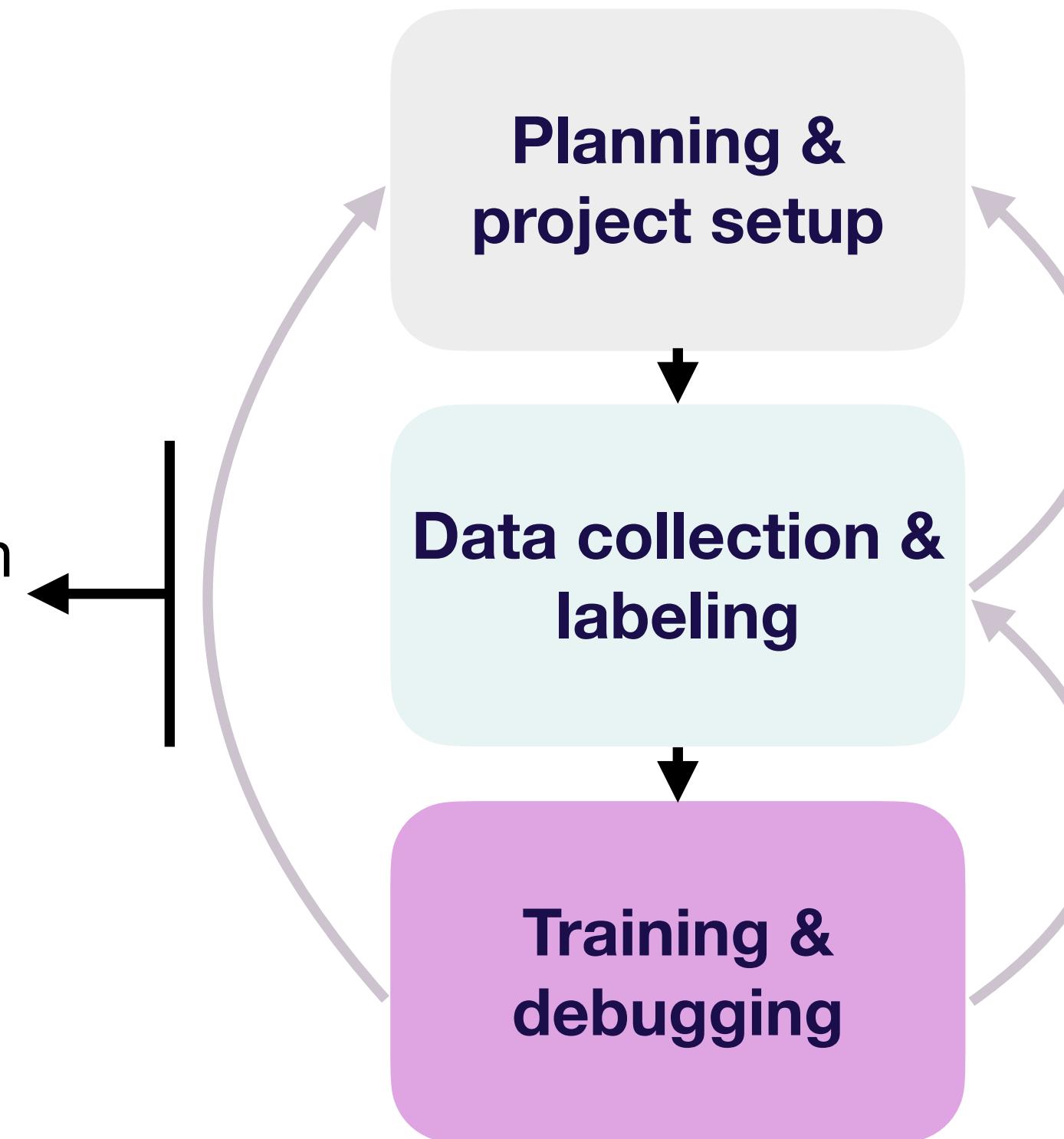


Lifecycle of a ML project

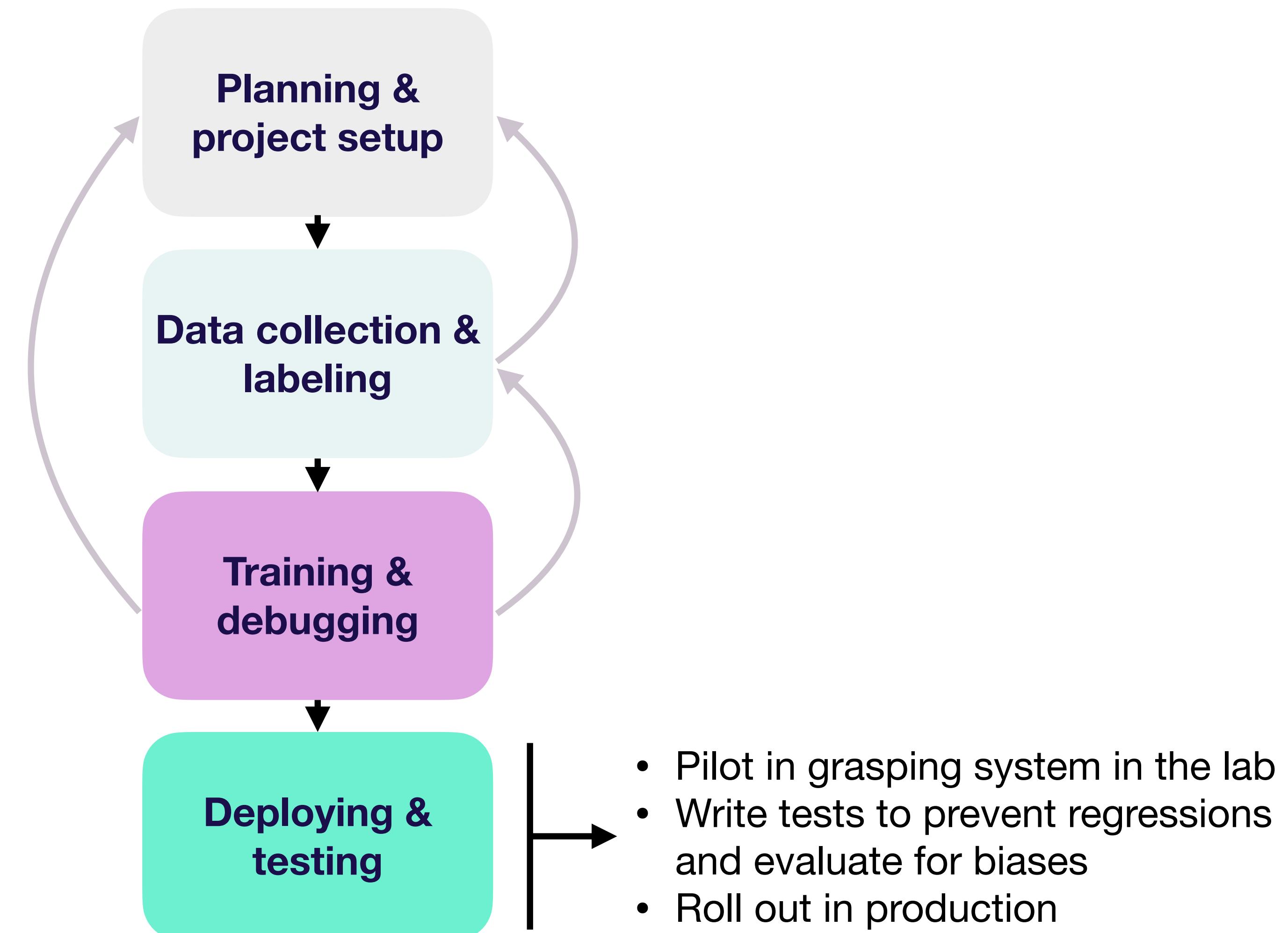


Lifecycle of a ML project

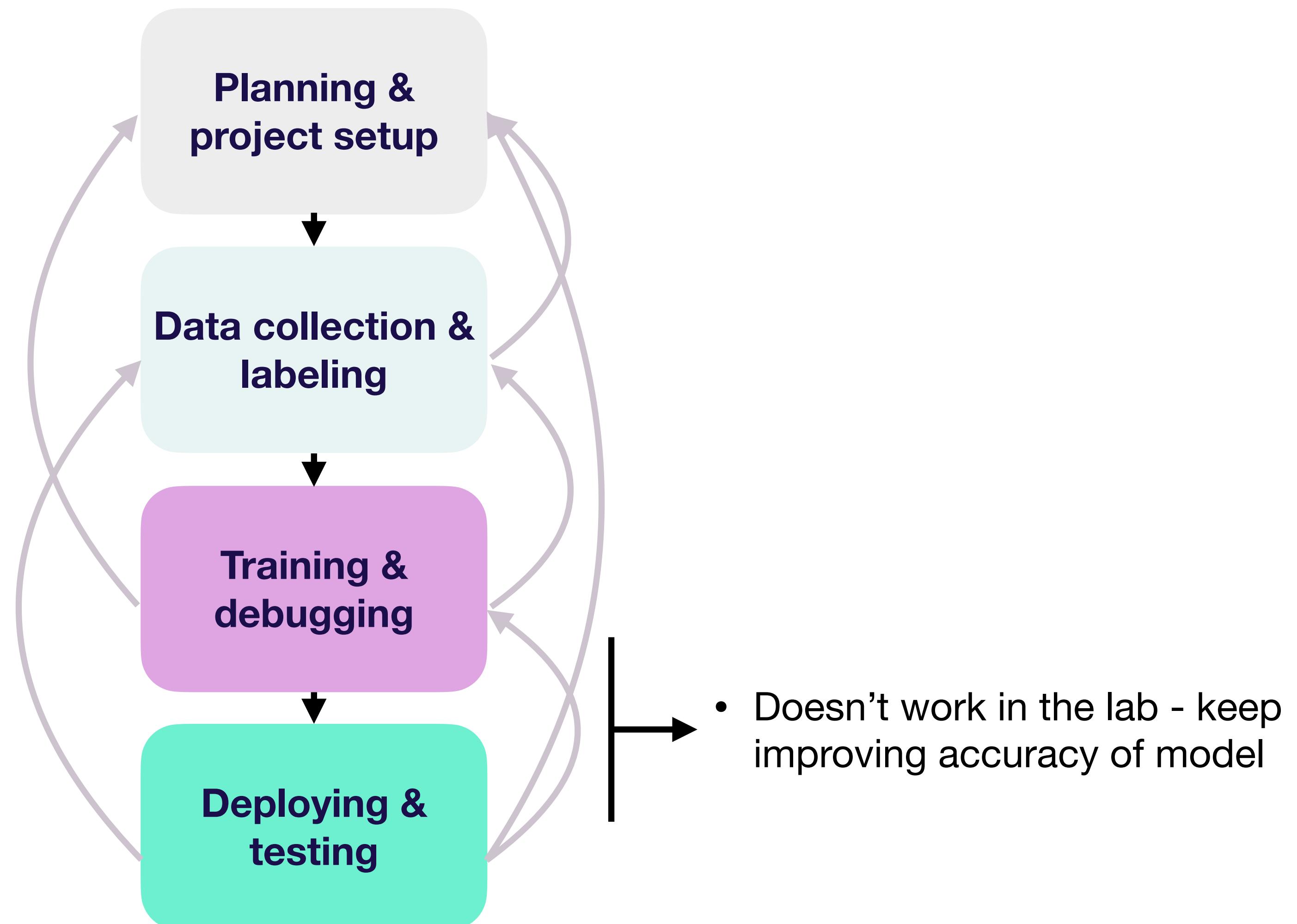
- Realize task is too hard
- Requirements trade off with each other - revisit which are most important



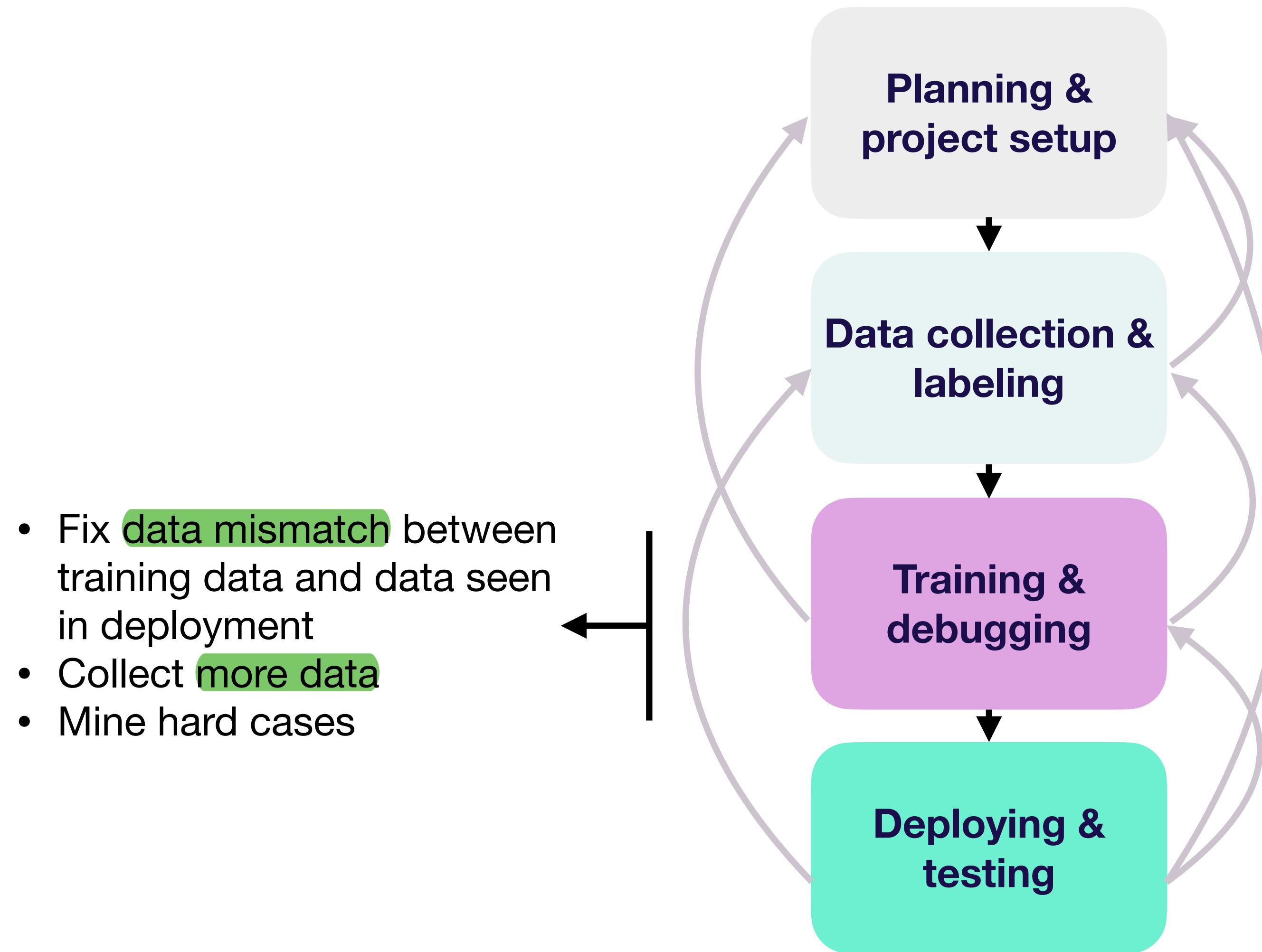
Lifecycle of a ML project



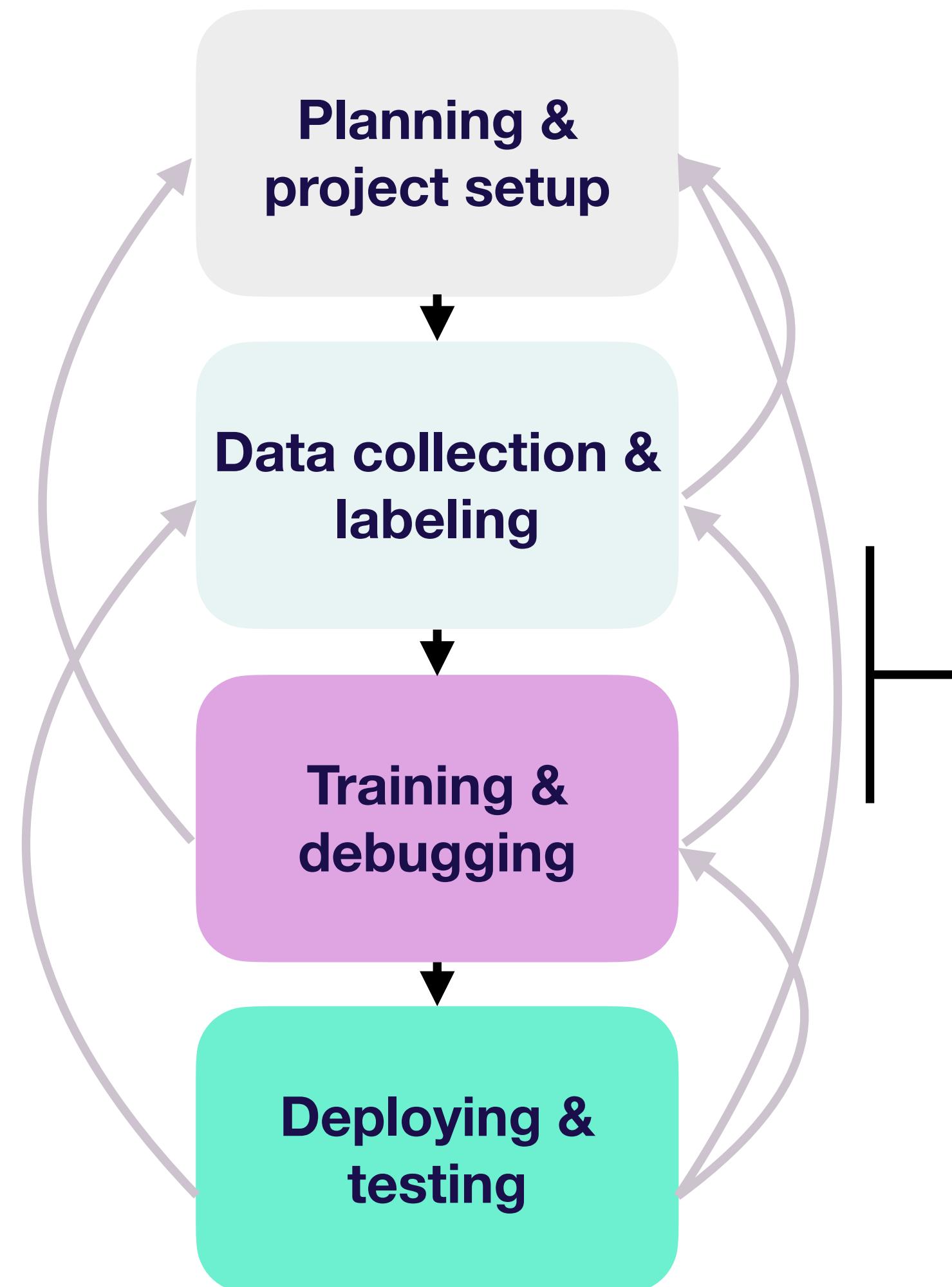
Lifecycle of a ML project



Lifecycle of a ML project

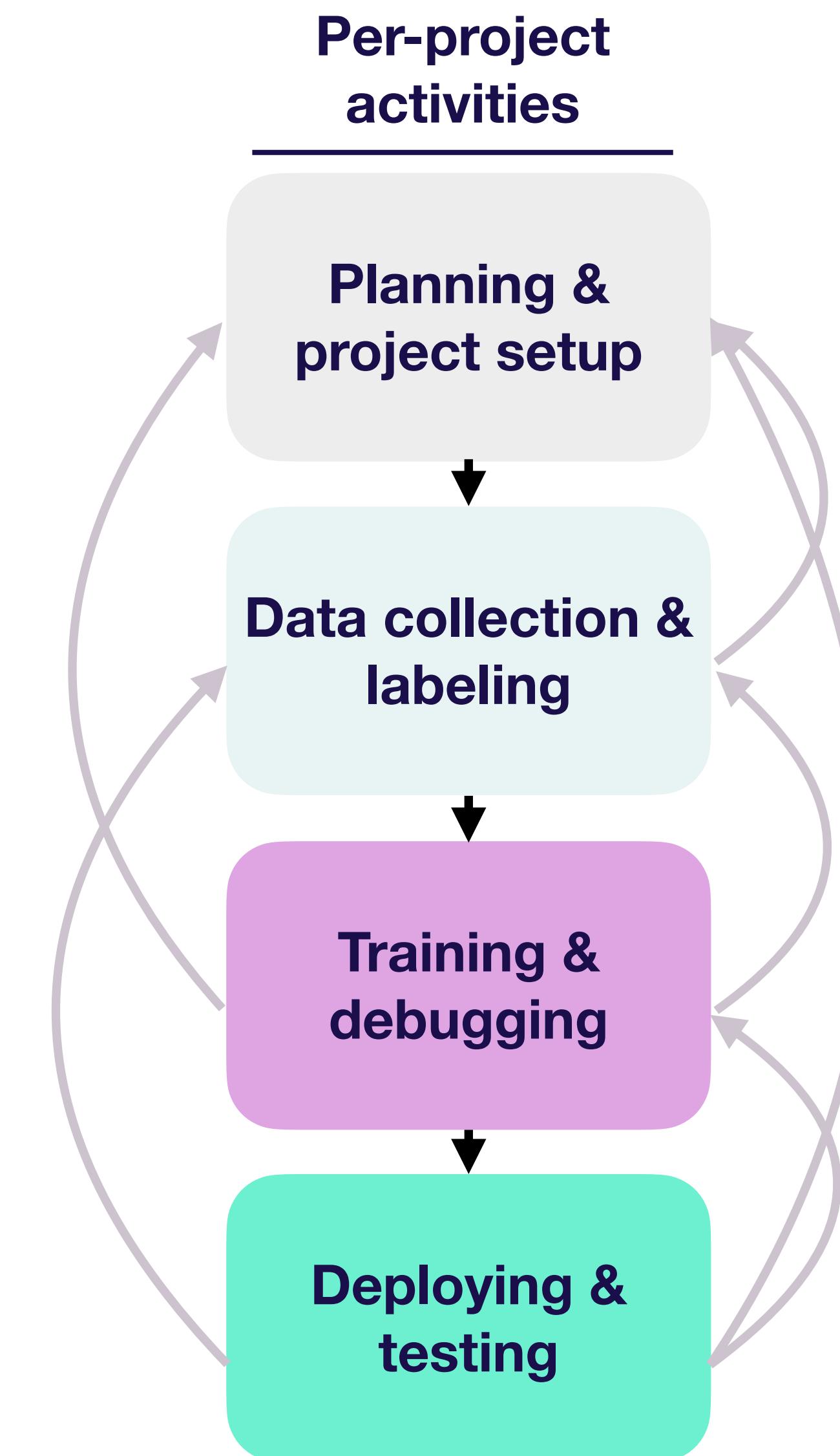


Lifecycle of a ML project

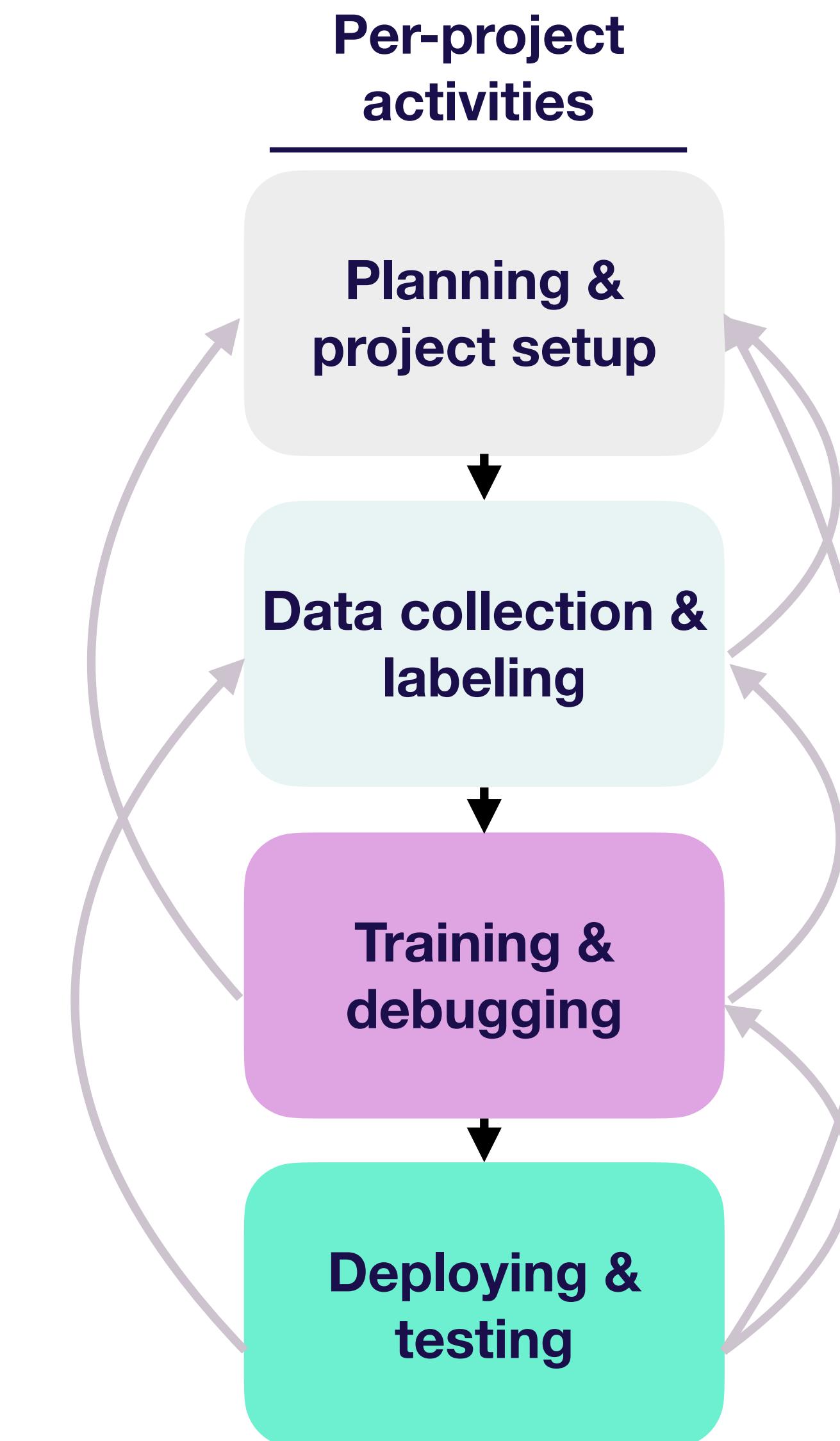


- The metric you picked doesn't actually drive downstream user behavior. **Revisit the metric.**
- Performance in the real world isn't great - revisit requirements (e.g., do we need to **be faster or more accurate?**)

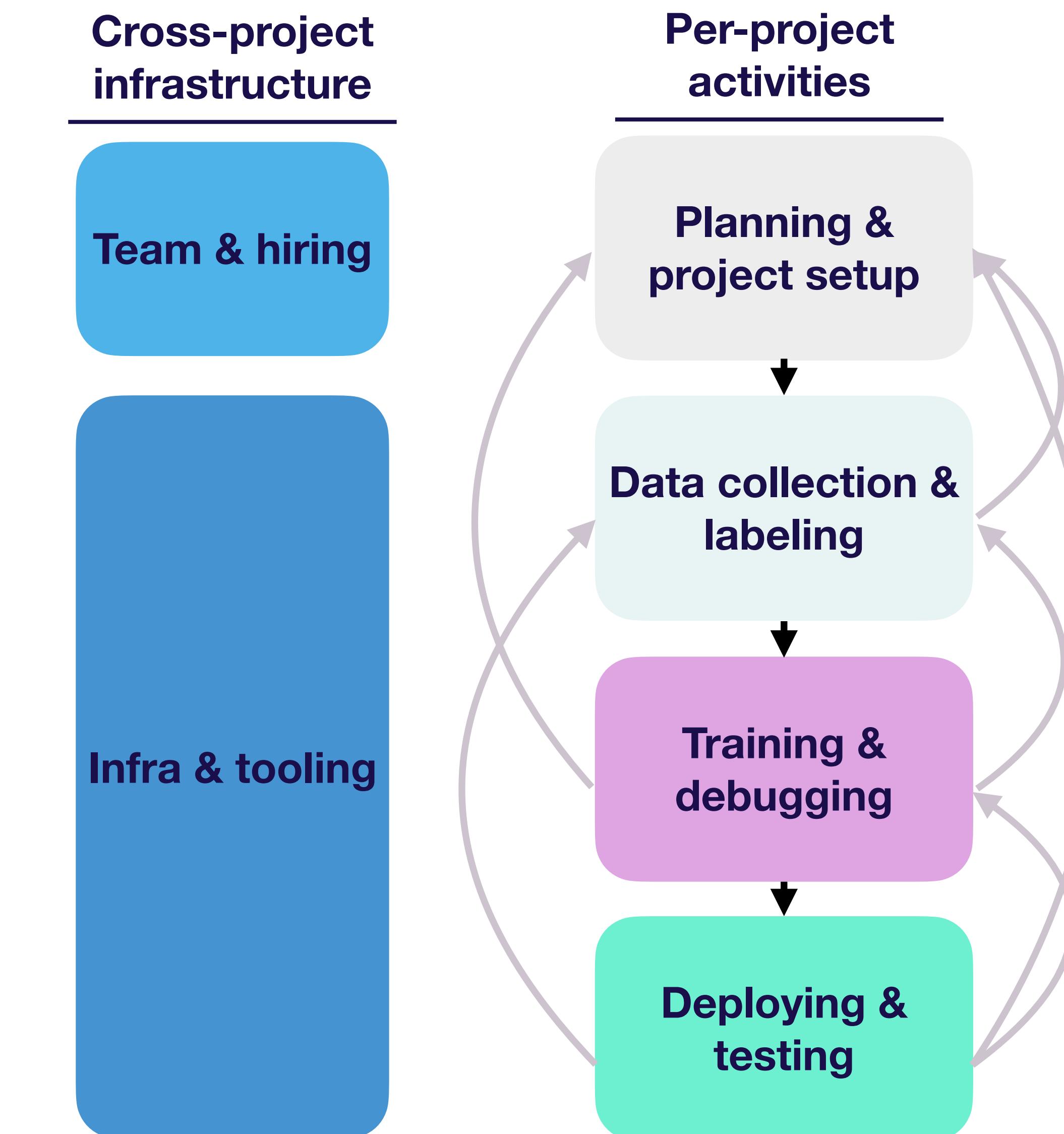
Lifecycle of a ML project



Lifecycle of a ML project



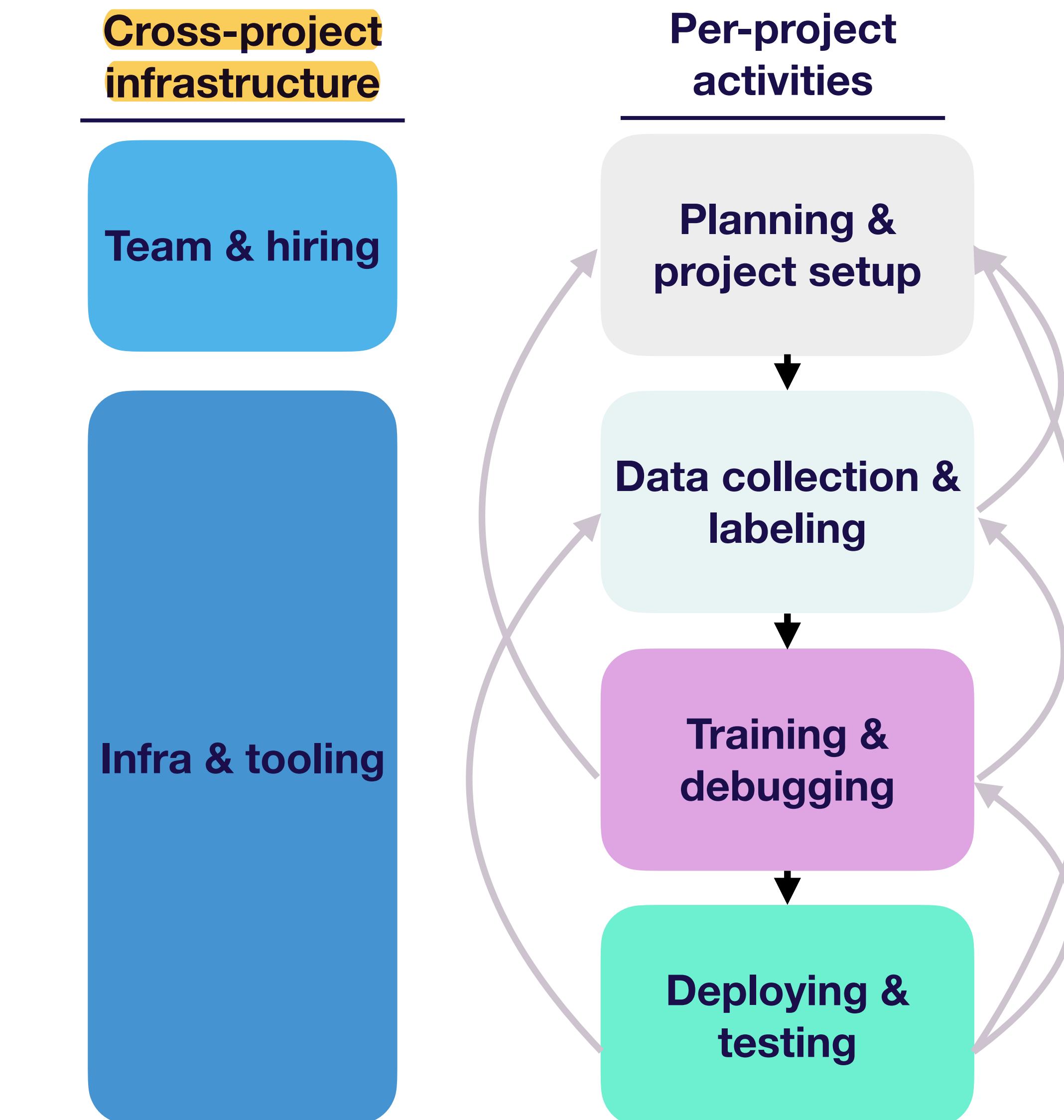
Lifecycle of a ML project



What else do you need to know?

- Understand state of the art in your domain
 - Understand what's possible
 - Know what to try next
- We will introduce most promising research areas

Lifecycle of a ML project



Questions?

Module overview

-  **Lifecycle**
 - How to think about all of the activities in an ML project
-  **Prioritizing projects**
 - **Assessing the feasibility and impact of your projects**
-  **Archetypes**
 - The main categories of ML projects, and the implications for project management
-  **Metrics**
 - How to pick a single number to optimize
-  **Baselines**
 - How to know if your model is performing well

Key points for prioritizing projects

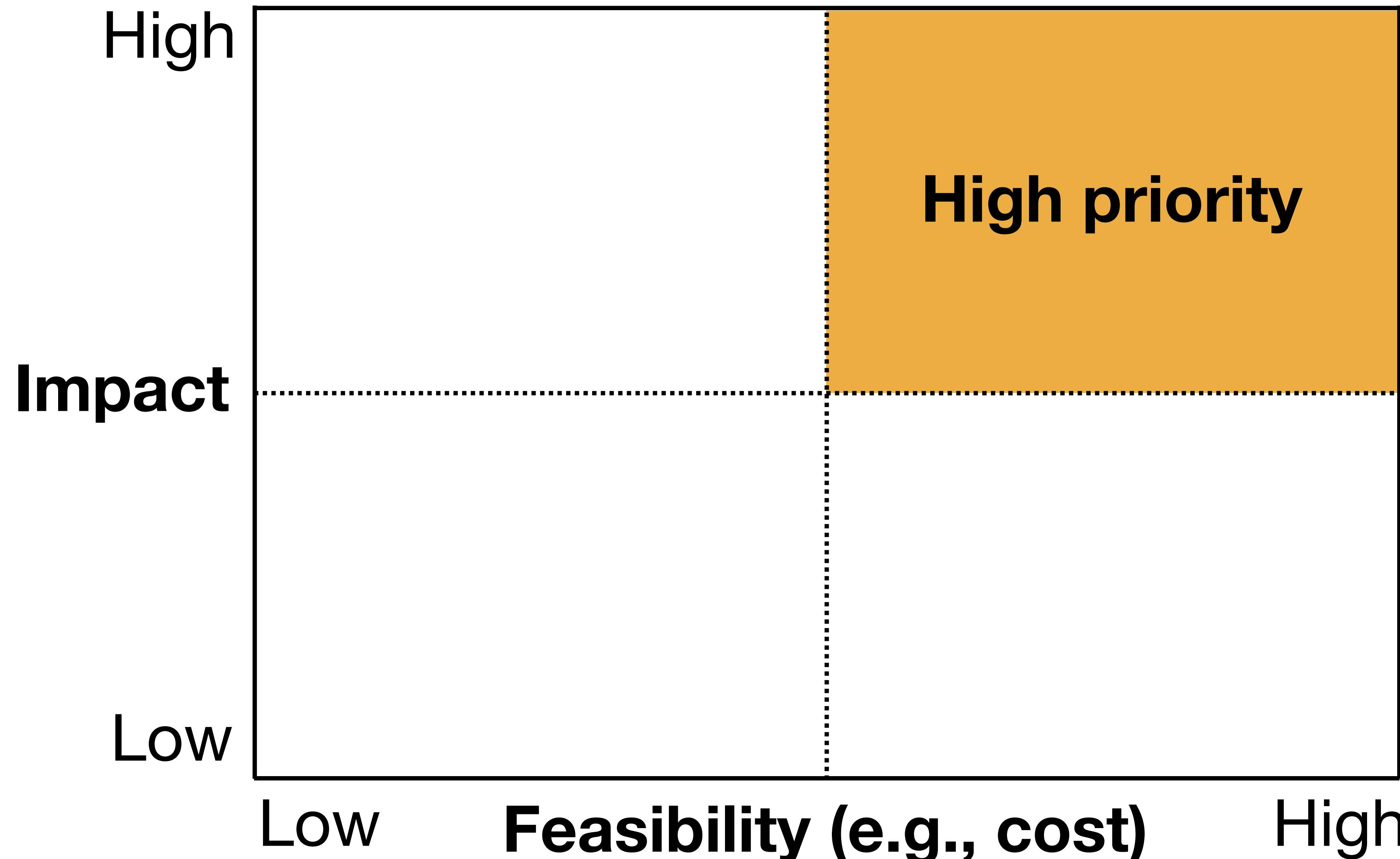
A. High-impact ML problems

- Friction in your product
- Complex parts of your pipeline
- Places where cheap prediction is valuable
- What else are people doing?

B. Cost of ML projects is driven by data availability. Also consider accuracy requirements and intrinsic difficulty of the problem



A (general) framework for prioritizing projects



Mental models for high-impact ML projects

1. Where can you take advantage of cheap prediction?
2. Where is there friction in your product?
3. Where can you automate complicated manual processes?
4. What are other people doing?

What does ML make economically feasible?

The economics of AI (Agrawal, Gans, Goldfarb)

- AI reduces cost of prediction
- Prediction is **central** for decision making
- **Cheap** prediction means
 - Prediction will be everywhere
 - Even in problems where it was too expensive before (e.g., for most people, hiring a driver)
- **Implication:** Look for projects where cheap prediction will have a huge business impact

Prediction Machines: The Simple Economics of Artificial Intelligence (Agrawal, Gans, Goldfarb)

What does your product need?

*“Discover Weekly **removed the friction of chasing everything down yourself and instead brought the music to you in a neat little package every Monday morning.**”*

NOTED

Three Principles for Designing ML-Powered Products

October 2019

What is ML good at?

Software 2.0
(Andrej Karpathy)

Andrej Karpathy

@karpathy

Following

Gradient descent can write code better than you. I'm sorry.

1:56 PM - 4 Aug 2017

358 Retweets 1,183 Likes

72 358 1.2K

Software 2.0 (Andrej Karpathy): <https://medium.com/@karpathy/software-2-0-a64152b37c35>

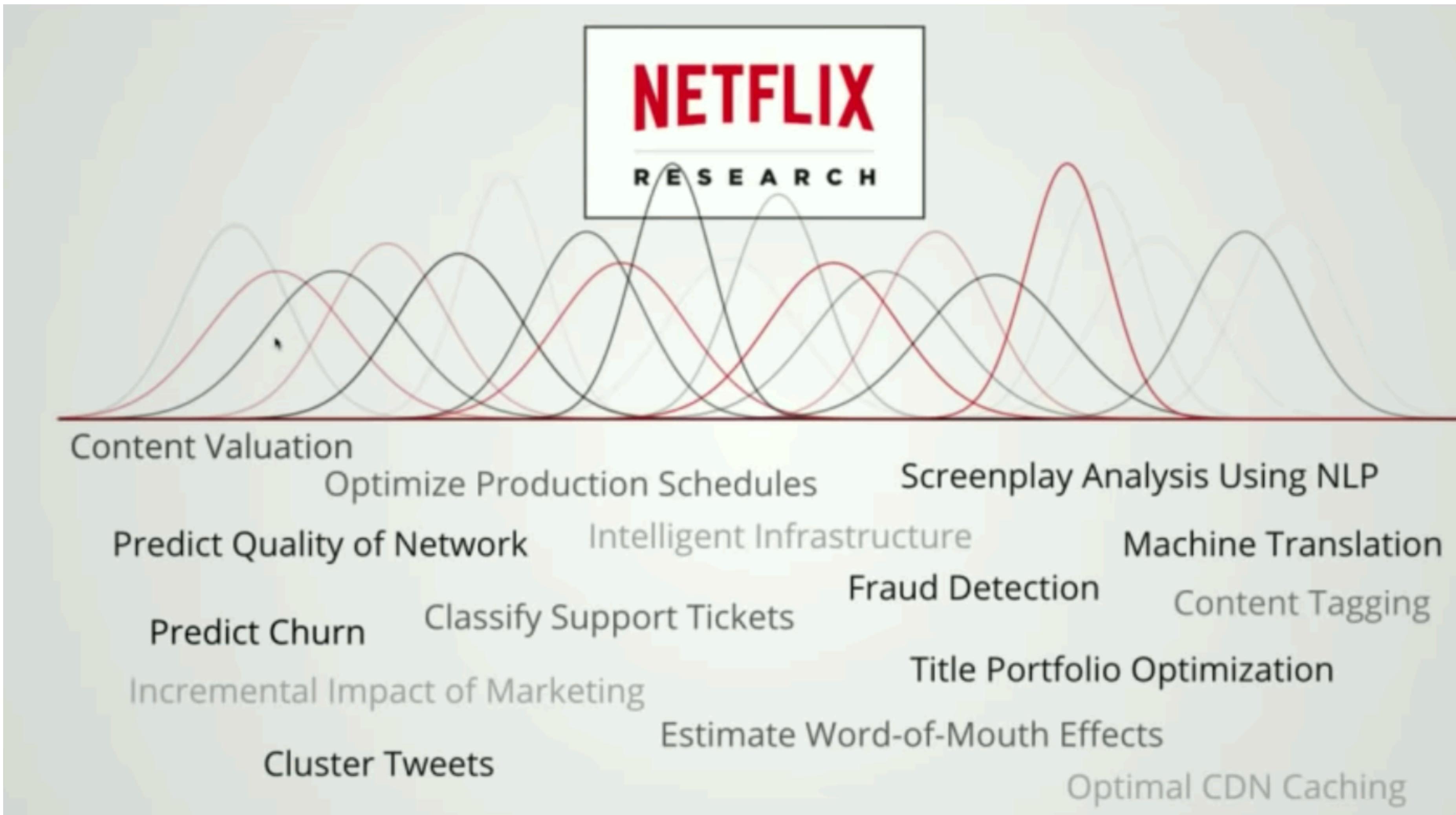
What is ML good at?

Software 2.0 (Andrej Karpathy)

- *Software 1.0* = traditional programs with explicit instructions (python / c++ / etc)
- Software 2.0 = humans specify goals, and algorithm searches for a program that works
- 2.0 programmers work with datasets, which get compiled via optimization
- Why? Works better, more general, computational advantages
- **Implication:** look for complicated rule-based software where we can learn the rules instead of programming them

Software 2.0 (Andrej Karpathy): <https://medium.com/@karpathy/software-2-0-a64152b37c35>

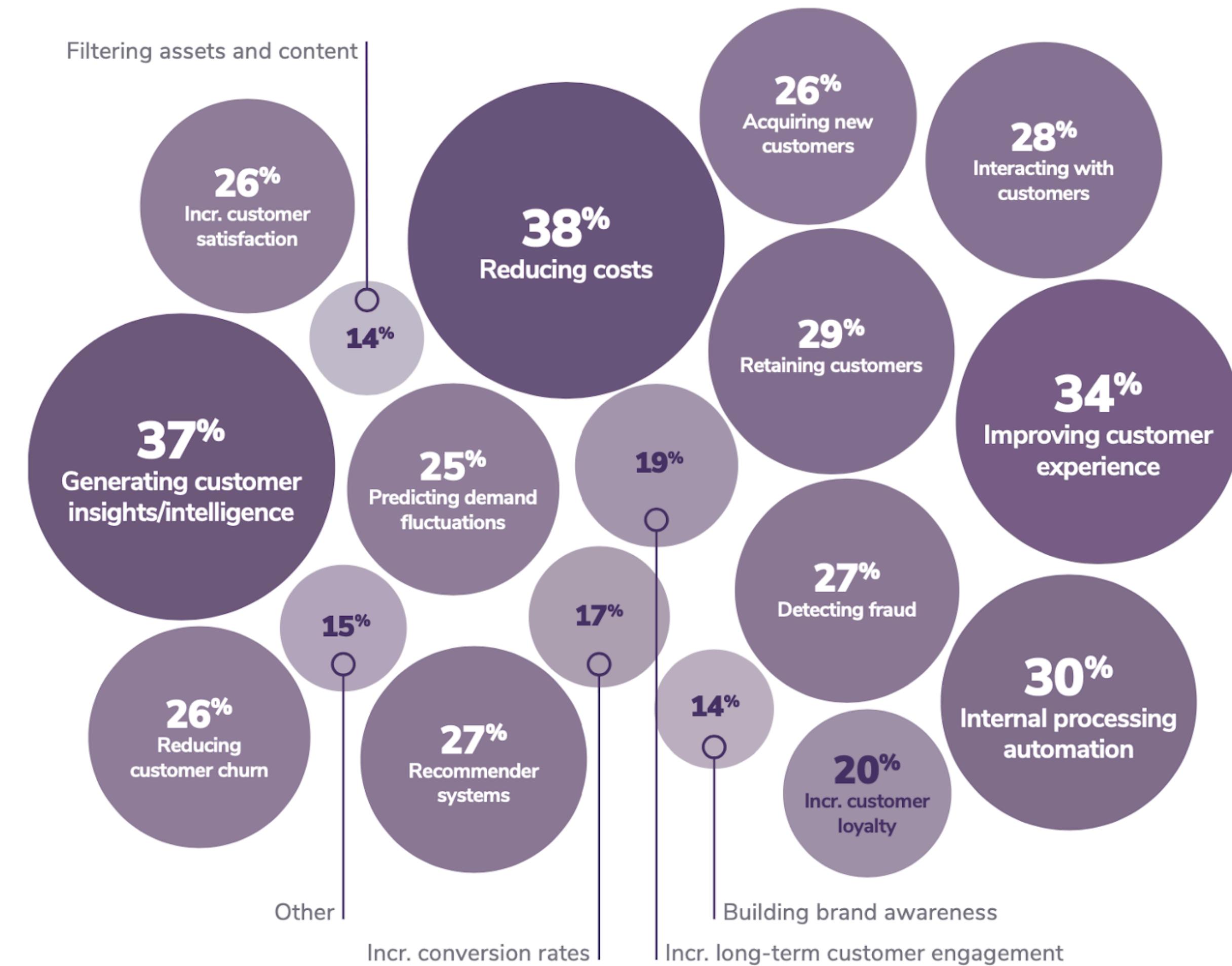
What are other people doing?



[Human-Centric Machine Learning Infrastructure @Netflix](#) (Ville Tuulos, InfoQ 2019)

What are other people doing?

Machine learning use case frequency



2020 state of enterprise machine learning (Algorithmia, 2020)

What are other people doing?

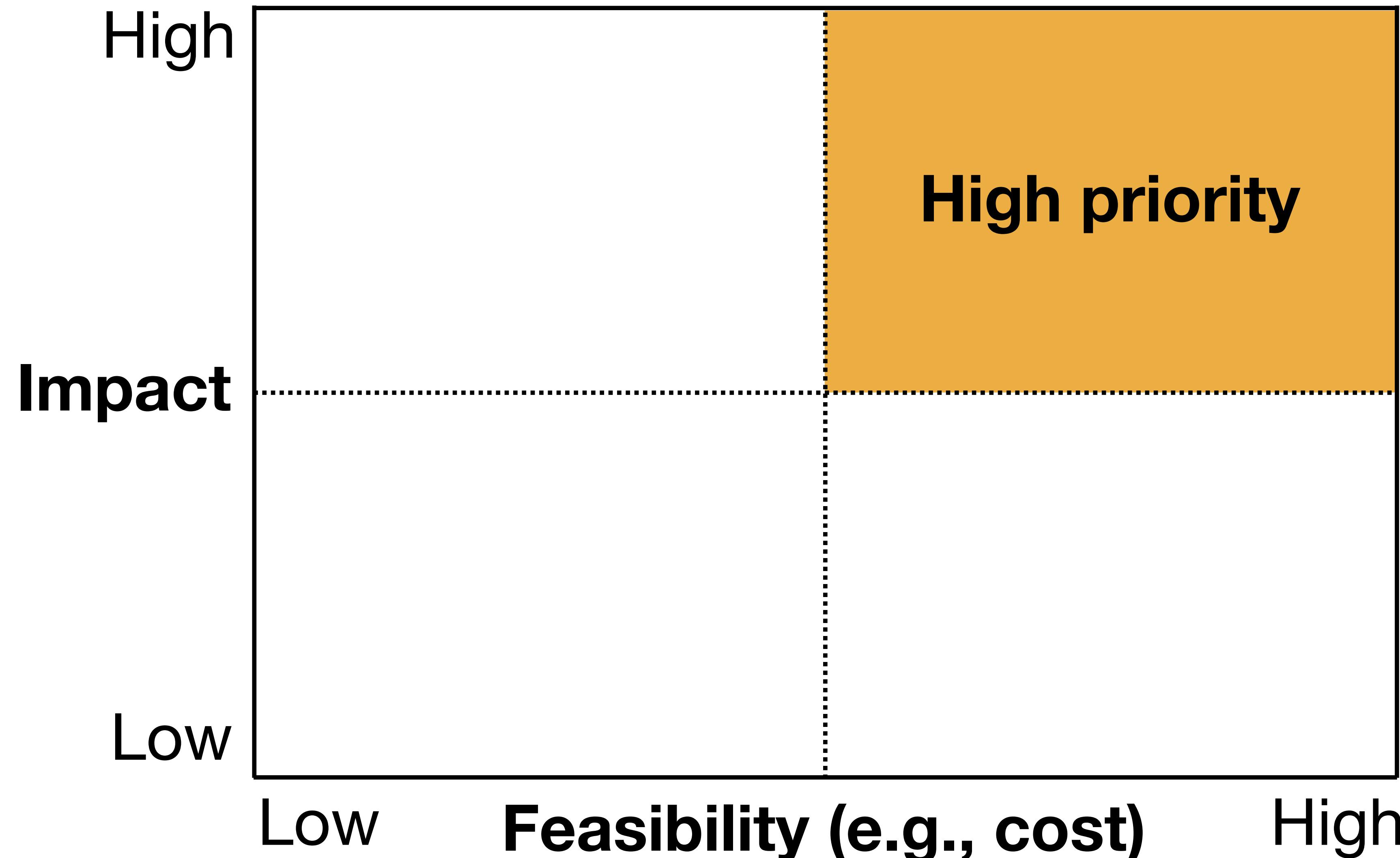
- Papers from Google, Facebook, Nvidia, Netflix, etc
- Blog posts from top earlier-stage companies (Uber, Lyft, Spotify, Stripe, etc)

Case studies

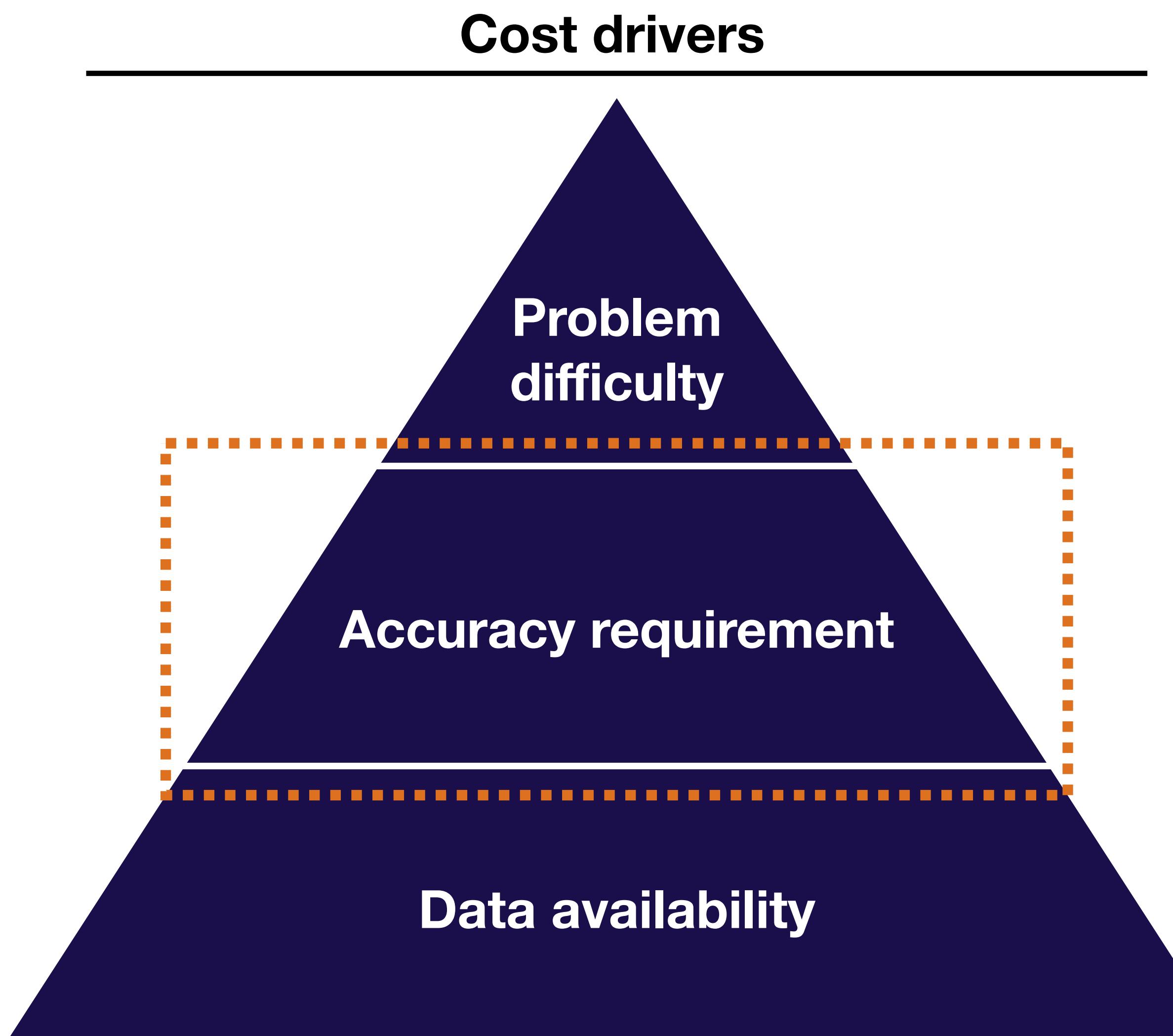
- [Using Machine Learning to Predict Value of Homes On Airbnb](#) (Robert Chang, Airbnb Engineering & Data Science, 2017)
- [Using Machine Learning to Improve Streaming Quality at Netflix](#) (Chaitanya Ekanadham, Netflix Technology Blog, 2018)
- [150 Successful Machine Learning Models: 6 Lessons Learned at Booking.com](#) (Bernardi et al., KDD, 2019)Asdf
- [How we grew from 0 to 4 million women on our fashion app, with a vertical machine learning approach](#) (Gabriel Aldamiz, HackerNoon, 2018)
- [Machine Learning-Powered Search Ranking of Airbnb Experiences](#) (Mihajlo Grbovic, Airbnb Engineering & Data Science, 2019)
- [From shallow to deep learning in fraud](#) (Hao Yi Ong, Lyft Engineering, 2018)
- [Space, Time and Groceries](#) (Jeremy Stanley, Tech at Instacart, 2017)
- [Creating a Modern OCR Pipeline Using Computer Vision and Deep Learning](#) (Brad Neuberg, Dropbox Engineering, 2017)
- [Scaling Machine Learning at Uber with Michelangelo](#) (Jeremy Hermann and Mike Del Balso, Uber Engineering, 2019)
- [Spotify's Discover Weekly: How machine learning finds your new music](#) (Umesh .A Bhat, 2017)

Credit for compiling this list: Chip Huyen ([Machine Learning Systems Design Lecture 2 note](#))

A (general) framework for prioritizing projects



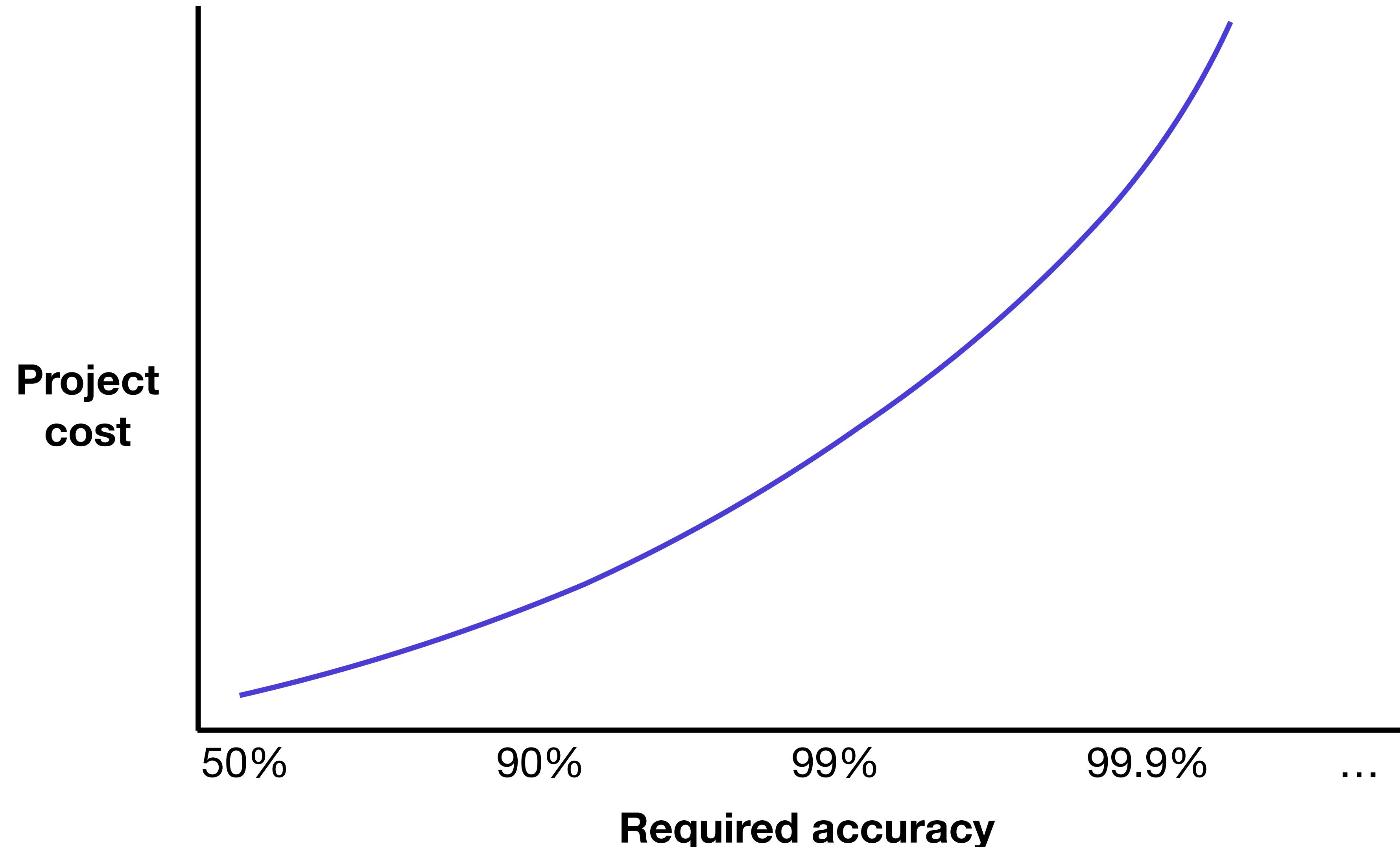
Assessing feasibility of ML projects



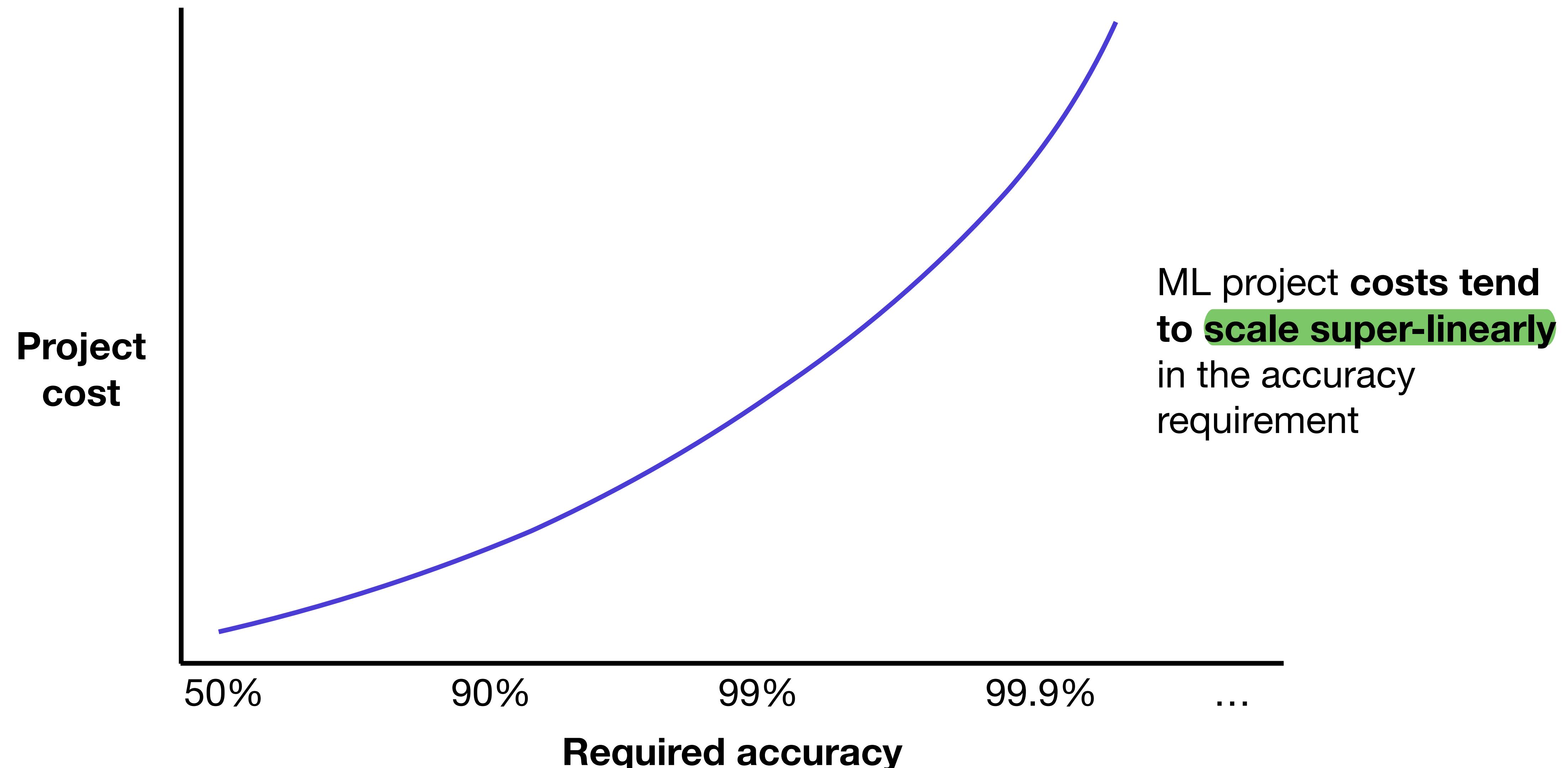
Main considerations

- Is the problem well-defined?
 - Good published work on similar problems?
(newer problems mean more risk & more technical effort)
 - Compute requirements?
 - Can a human do it?
-
- How **costly** are wrong predictions?
 - How **frequently** does the system need to be **right** to be useful?
 - Ethical implications?
-
- How hard is it to **acquire** data?
 - How expensive is data **labeling**?
 - How **much** data will be needed?
 - How **stable** is the data?
 - Data **security** requirements?

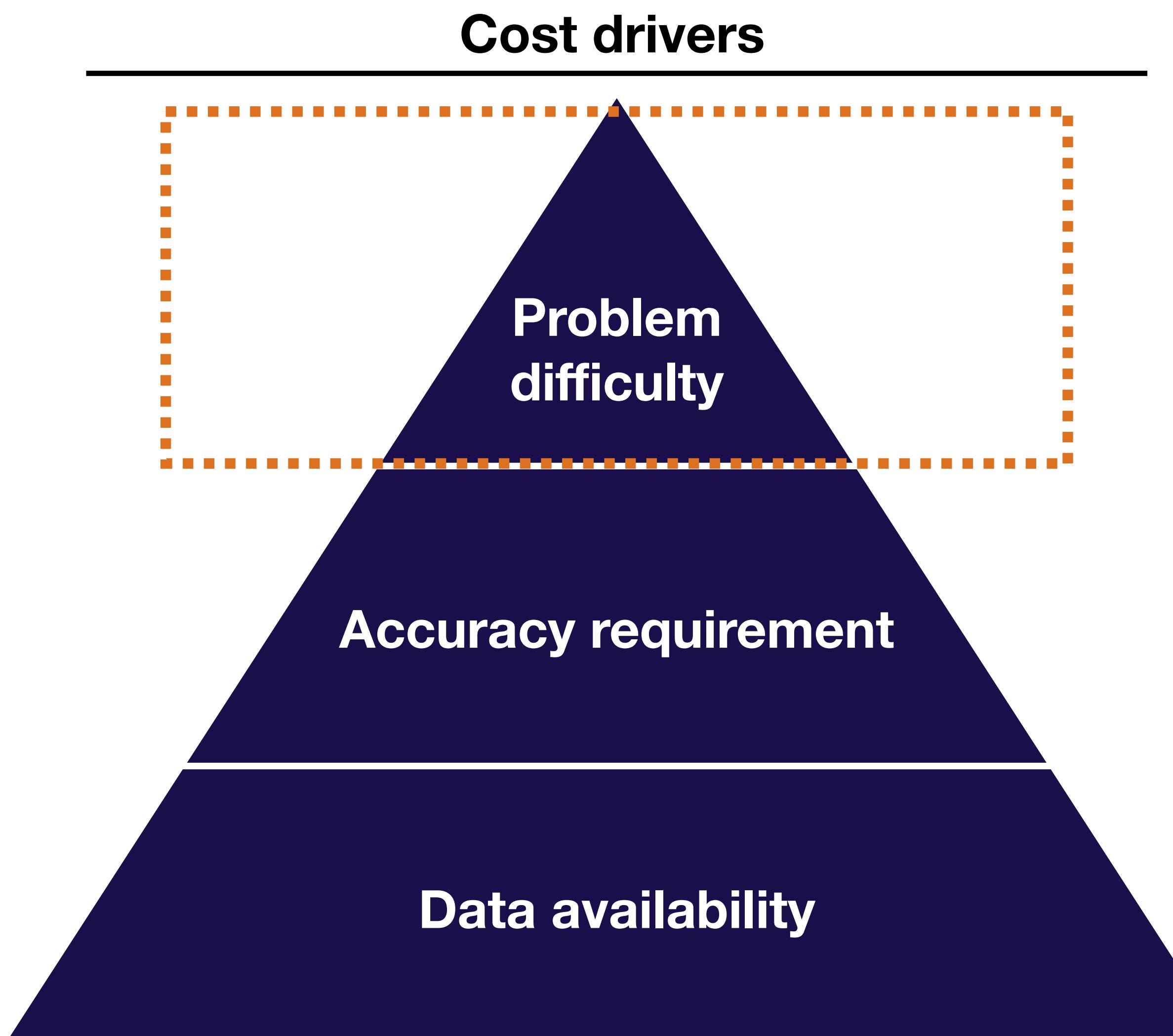
Why are accuracy requirements so important?



Why are accuracy requirements so important?



Assessing feasibility of ML projects



Main considerations

- Is the problem well-defined?
 - Good published work on similar problems? (newer problems mean more risk & more technical effort)
 - Compute requirements?
 - Can a human do it?
-
- How costly are wrong predictions?
 - How frequently does the system need to be right to be useful?
 - Ethical implications?
-
- How hard is it to acquire data?
 - How expensive is data labeling?
 - How much data will be needed?
 - How stable is the data?
 - Data security requirements?

What's still hard in machine learning?

"It may be a hundred years before a computer beats humans at Go -- maybe even longer," said Dr. Piet Hut, an astrophysicist at the Institute for Advanced Study in Princeton, N.J., and a fan of the game. "If a reasonably intelligent person learned to play Go, in a few months he could beat all existing computer programs. You don't have to be a Kasparov."

New York Times, July 1997

What's still hard in machine learning?



What's still hard in machine learning?



Andrew Ng

@AndrewYNg

Following

Pretty much anything that a normal person can do in <1 sec, we can now automate with AI.

Examples

- Recognize content of images
- Understand speech
- Translate speech
- Grasp objects
- etc.

Counter-examples?

- Understand humor / sarcasm
- In-hand robotic manipulation
- Generalize to new scenarios
- etc.



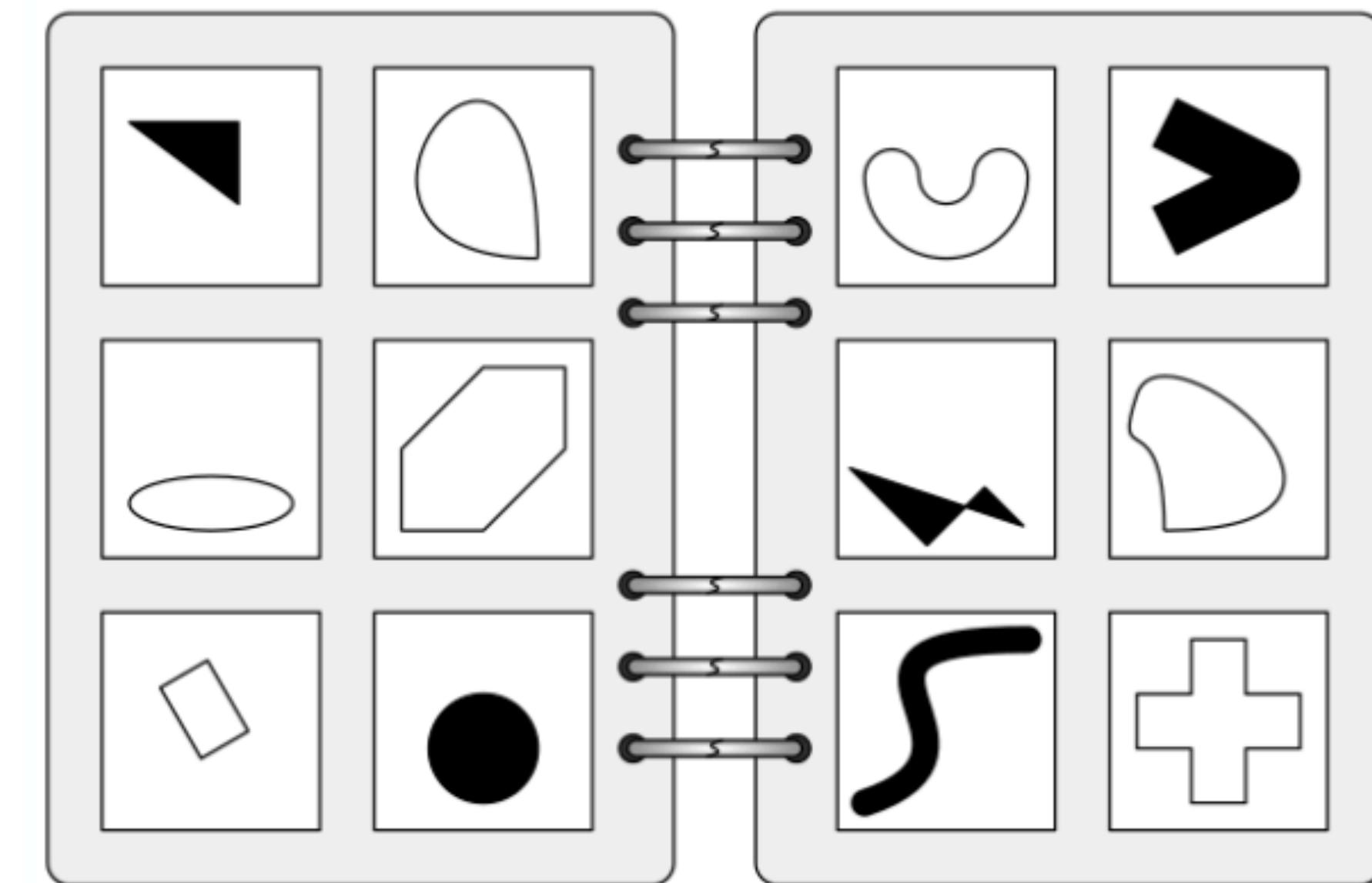
What's still hard in machine learning?

- Unsupervised learning
- Reinforcement learning popular in academic but not in industry
- Both are showing promise in limited domains where tons of data and compute are available

What's still hard in supervised learning?

- Answering questions
- Summarizing text
- Predicting video
- Building 3D models
- Real-world speech recognition
- Resisting adversarial examples
- Doing math
- Solving word puzzles
- Bongard problems
- Etc

Example of a Bongard Problem



What types of problems are hard?

	Instances	Examples
Output is complex	<ul style="list-style-type: none">• High-dimensional output• Ambiguous output	<ul style="list-style-type: none">• 3D reconstruction• Video prediction• Dialog systems• Open-ended recommender systems
Reliability is required	<ul style="list-style-type: none">• High precision is required• Robustness is required	<ul style="list-style-type: none">• Failing safely out-of-distribution• Robustness to adversarial attacks• High-precision pose estimation
Generalization is required	<ul style="list-style-type: none">• Out of distribution data• Reasoning, planning, causality	<ul style="list-style-type: none">• Self-driving: edge cases• Self-driving: control• Small data



Why is FSR focusing on pose estimation?

Impact	Feasibility
<ul style="list-style-type: none">• FSR's goal is grasping - requires reliable pose estimation• Traditional robotics pipeline uses hand-designed heuristics & online optimization<ul style="list-style-type: none">• Slow• Brittle• Great candidate for Software 2.0!	<ul style="list-style-type: none">• Data availability<ul style="list-style-type: none">• Easy to collect data• Labeling data could be a challenge, but can instrument lab with sensors• Accuracy requirement<ul style="list-style-type: none">• Require high accuracy to grasp an object: $<0.5\text{cm}$• However, low cost of failure - picks per hour important, not % successes• Problem difficulty<ul style="list-style-type: none">• Similar published results exist but need to adapt to our objects and robot

How to run a ML feasibility assessment

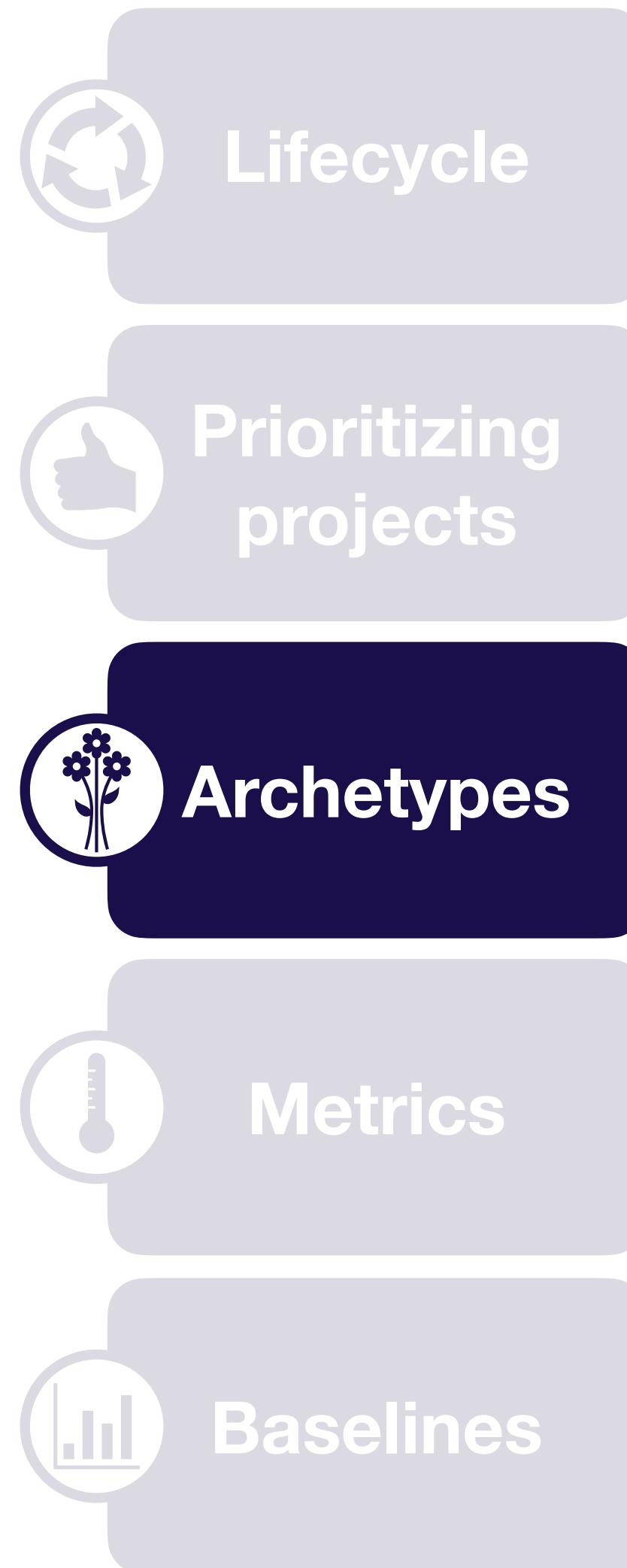
- A. Are you sure you need ML at all?
- B. Put in the work up-front to define success criteria with all of the stakeholders
- C. Consider the ethics of using ML
- D. Do a literature review
- E. Try to rapidly build a labeled benchmark dataset
- F. Build a *minimal* viable product (e.g., manual rules)
- G. Are you sure you need ML at all?



Questions?



Module overview

- How to think about all of the activities in an ML project
 - Assessing the feasibility and impact of your projects
 - **The main categories of ML projects, and the implications for project management**
 - How to pick a single number to optimize
 - How to know if your model is performing well
- 



Machine learning product archetypes

Software 2.0

Examples

- Improve code completion in an IDE
- Build a customized recommendation system
- Build a better video game AI



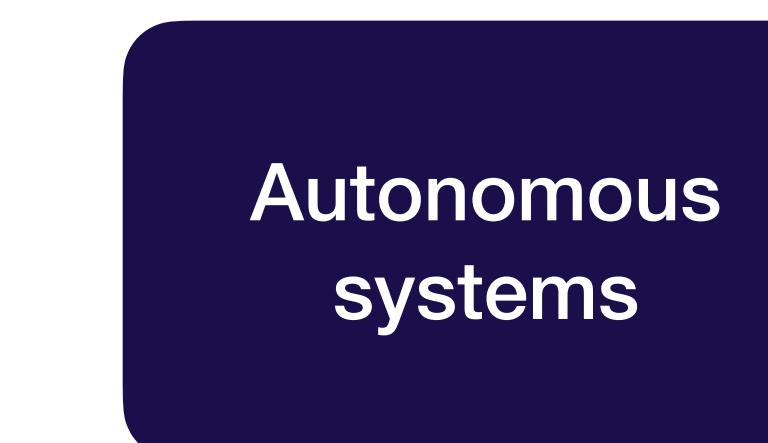
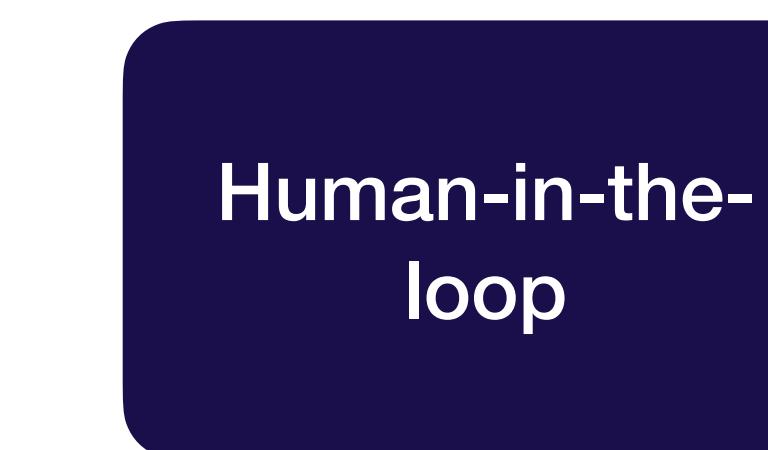
Machine learning product archetypes



Examples

- Improve code completion in an IDE
 - Build a customized recommendation system
 - Build a better video game AI
-
- Turn sketches into slides
 - Email auto-completion
 - Help a radiologist do their job faster

Machine learning product archetypes



Examples

- Improve code completion in an IDE
 - Build a customized recommendation system
 - Build a better video game AI
-
- Turn sketches into slides
 - Email auto-completion
 - Help a radiologist do their job faster
-
- Full self-driving
 - Automated customer support
 - Automated website design

Machine learning product archetypes

Software 2.0

Key questions

- Do your models truly improve performance?
- Does performance improvement generate business value?
- Do performance improvements lead to a data flywheel?



Machine learning product archetypes

Software 2.0

Human-in-the-loop

Key questions

- Do your models truly improve performance?
 - Does performance improvement generate business value?
 - Do performance improvements lead to a data flywheel?
-
- How good does the system need to be to be useful?
 - How can you collect enough data to make it that good?



Machine learning product archetypes

Software 2.0

-
- Key questions**
- Do your models truly improve performance?
 - Does performance improvement generate business value?
 - Do performance improvements lead to a data flywheel?

Human-in-the-loop

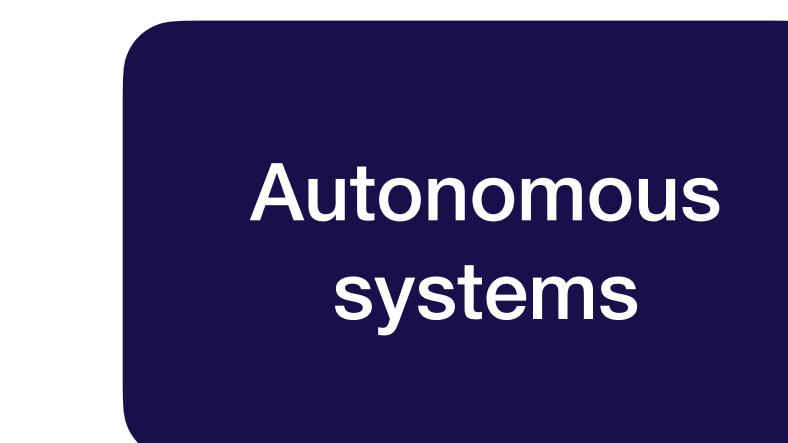
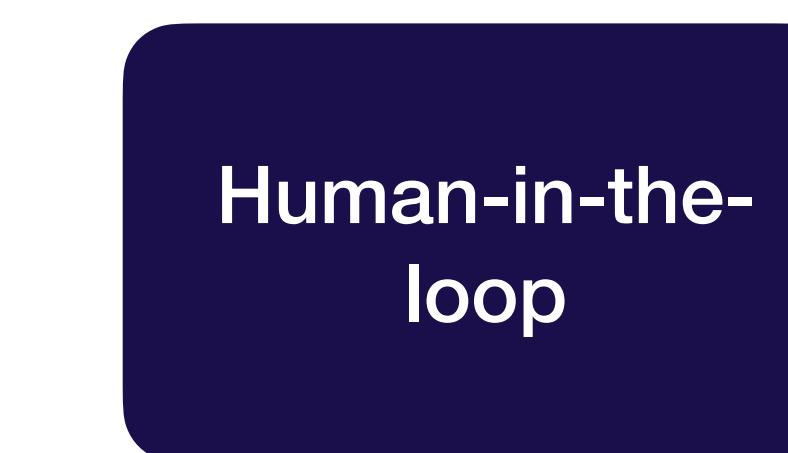
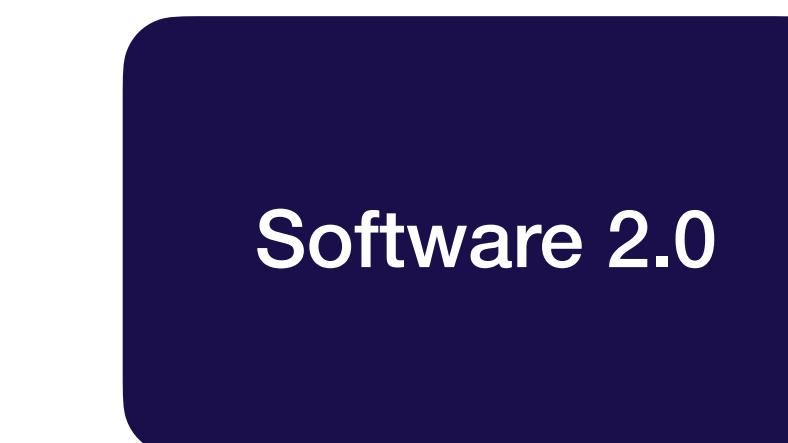
- How good does the system need to be to be useful?
- How can you collect enough data to make it that good?

Autonomous systems

- What is an acceptable failure rate for the system?
- How can you guarantee that it won't exceed that failure rate?
- How inexpensively can you label data from the system?



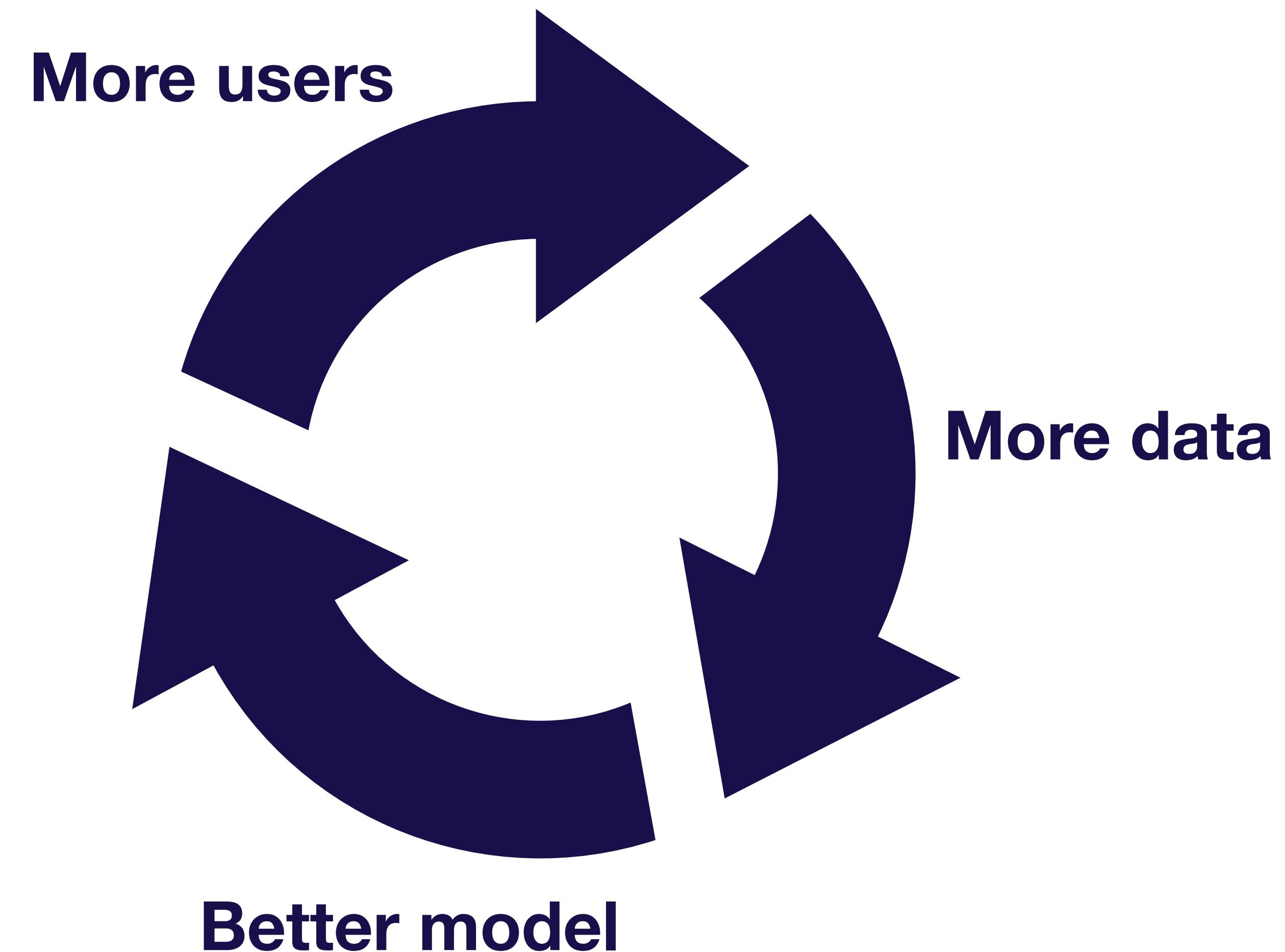
Machine learning product archetypes



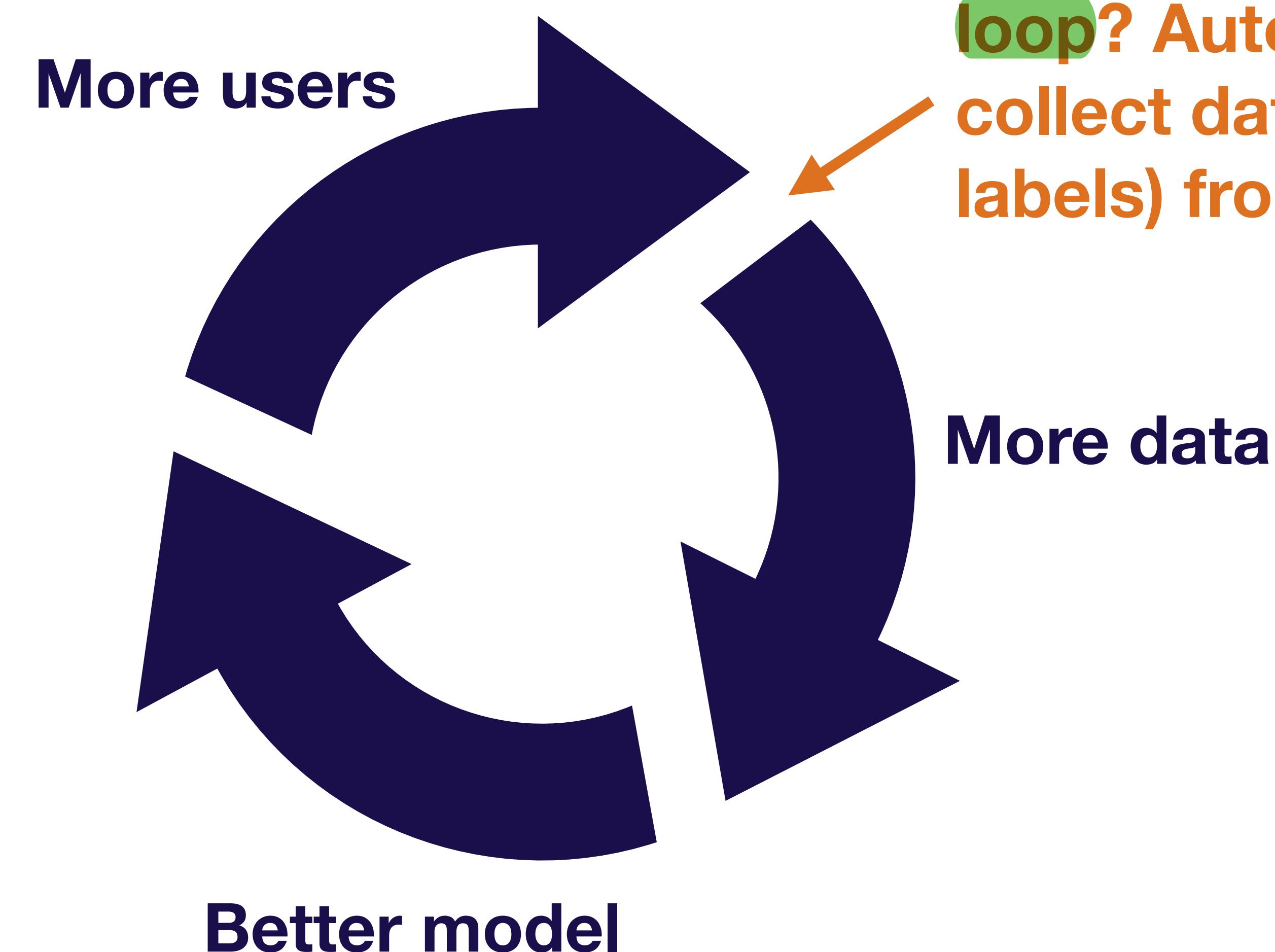
Key questions

- Do your models truly improve performance?
 - Does performance improvement generate business value?
 - Do performance improvements lead to a data flywheel?
-
- How good does the system need to be to be useful?
 - How can you collect enough data to make it that good?
-
- What is an acceptable failure rate for the system?
 - How can you guarantee that it won't exceed that failure rate?
 - How inexpensively can you label data from the system?

Data flywheels

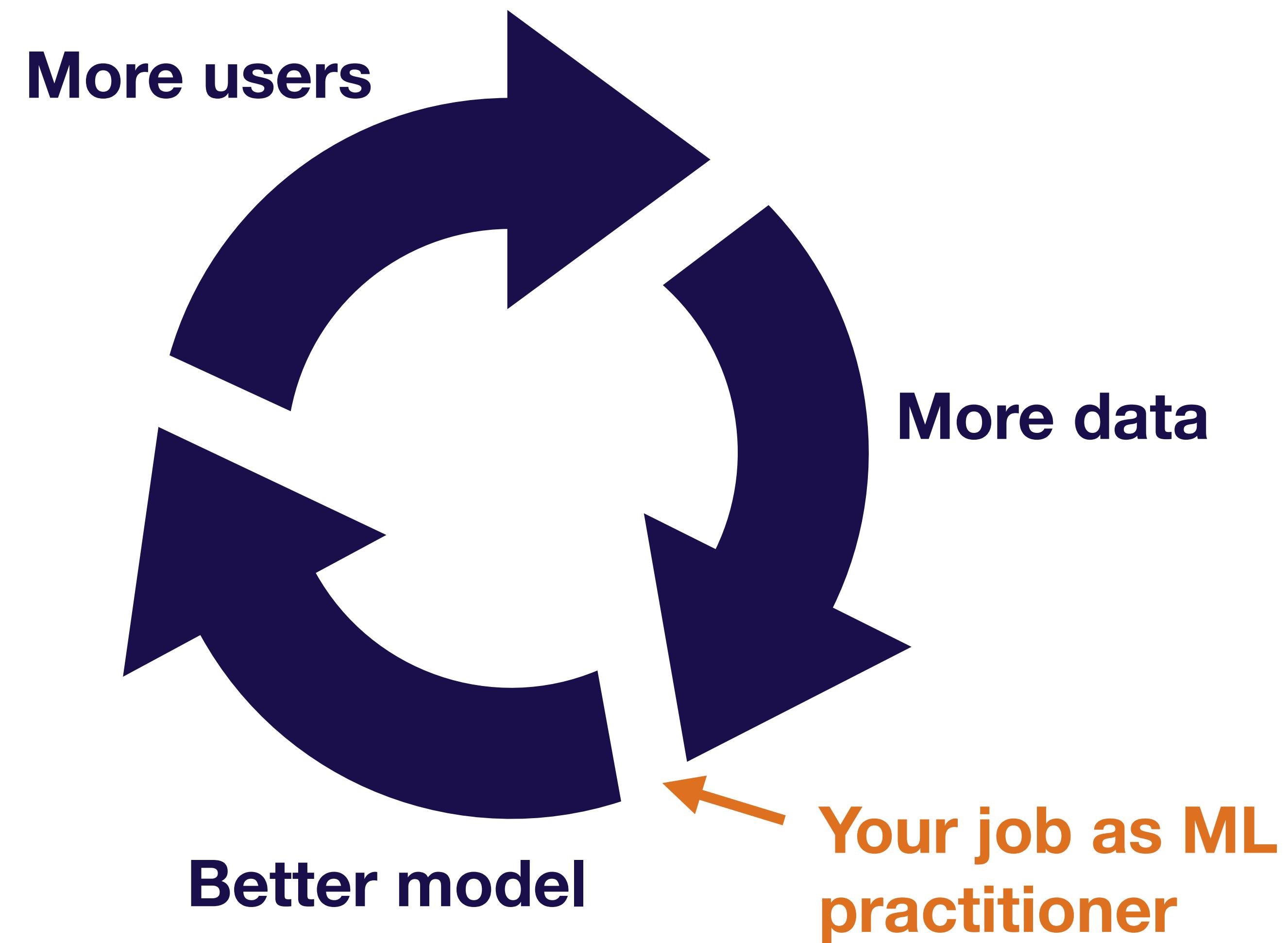


Data flywheels

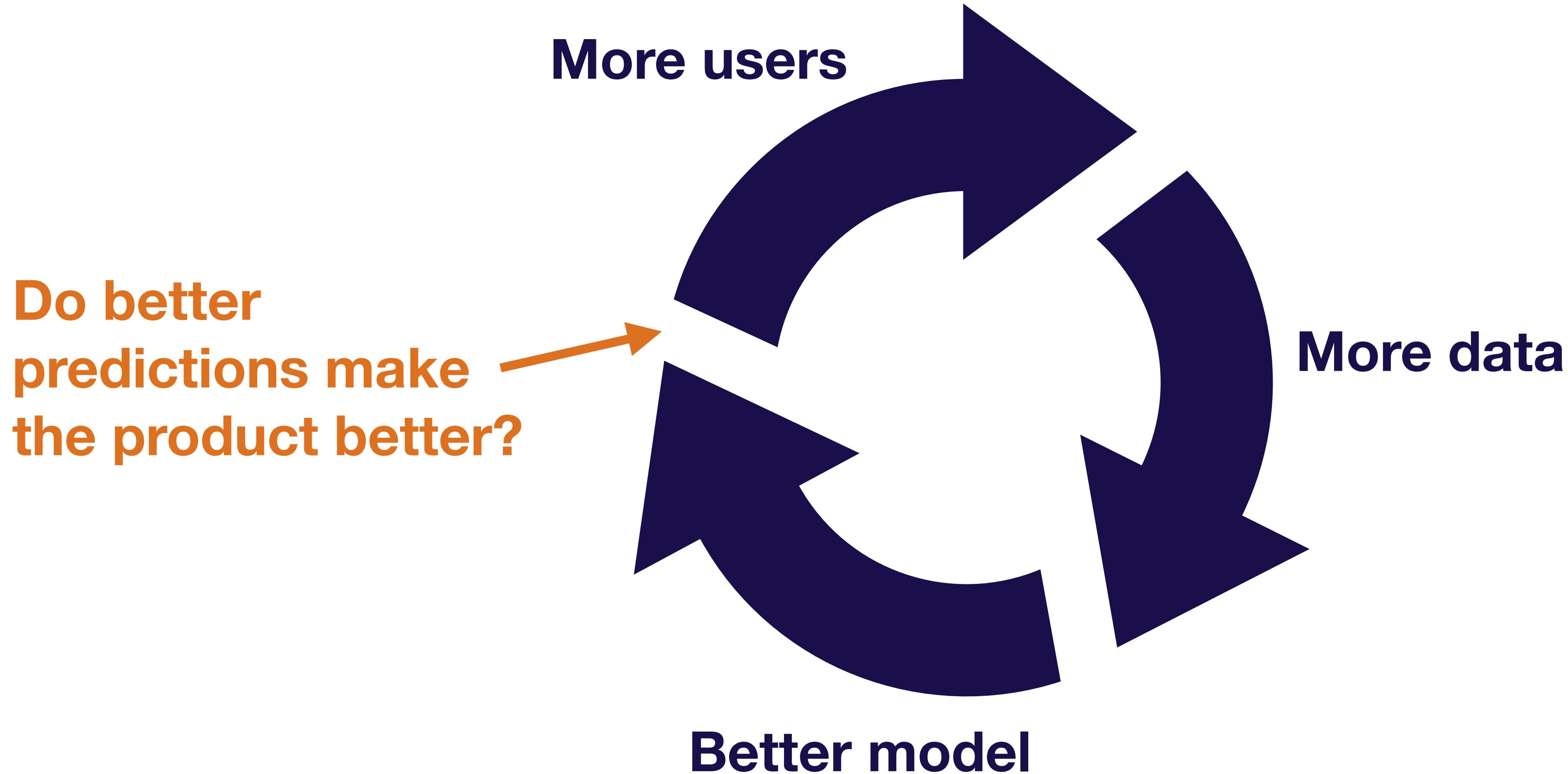


Do you have a **data loop**? Automatically collect data (and ideally labels) from your users

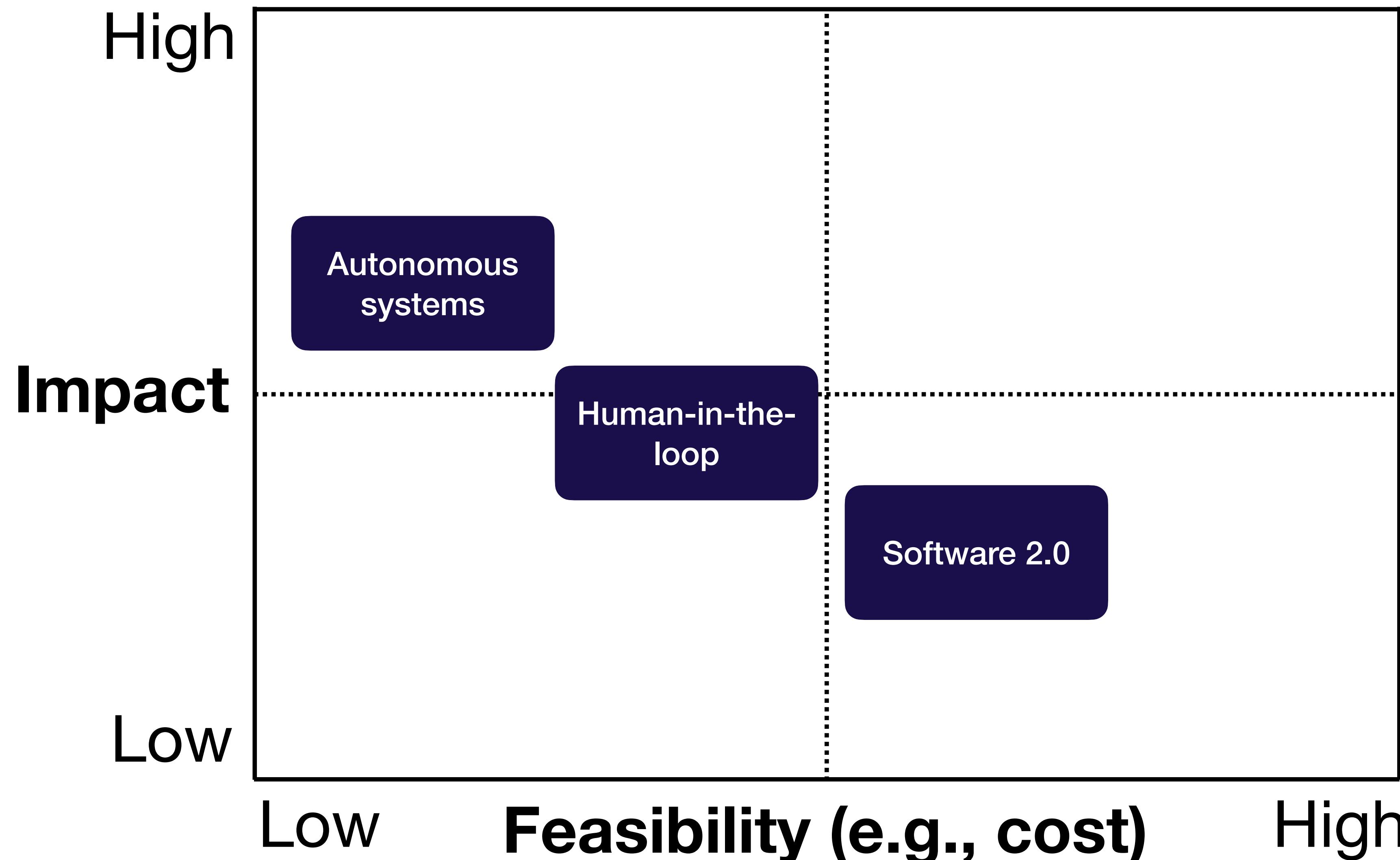
Data flywheels



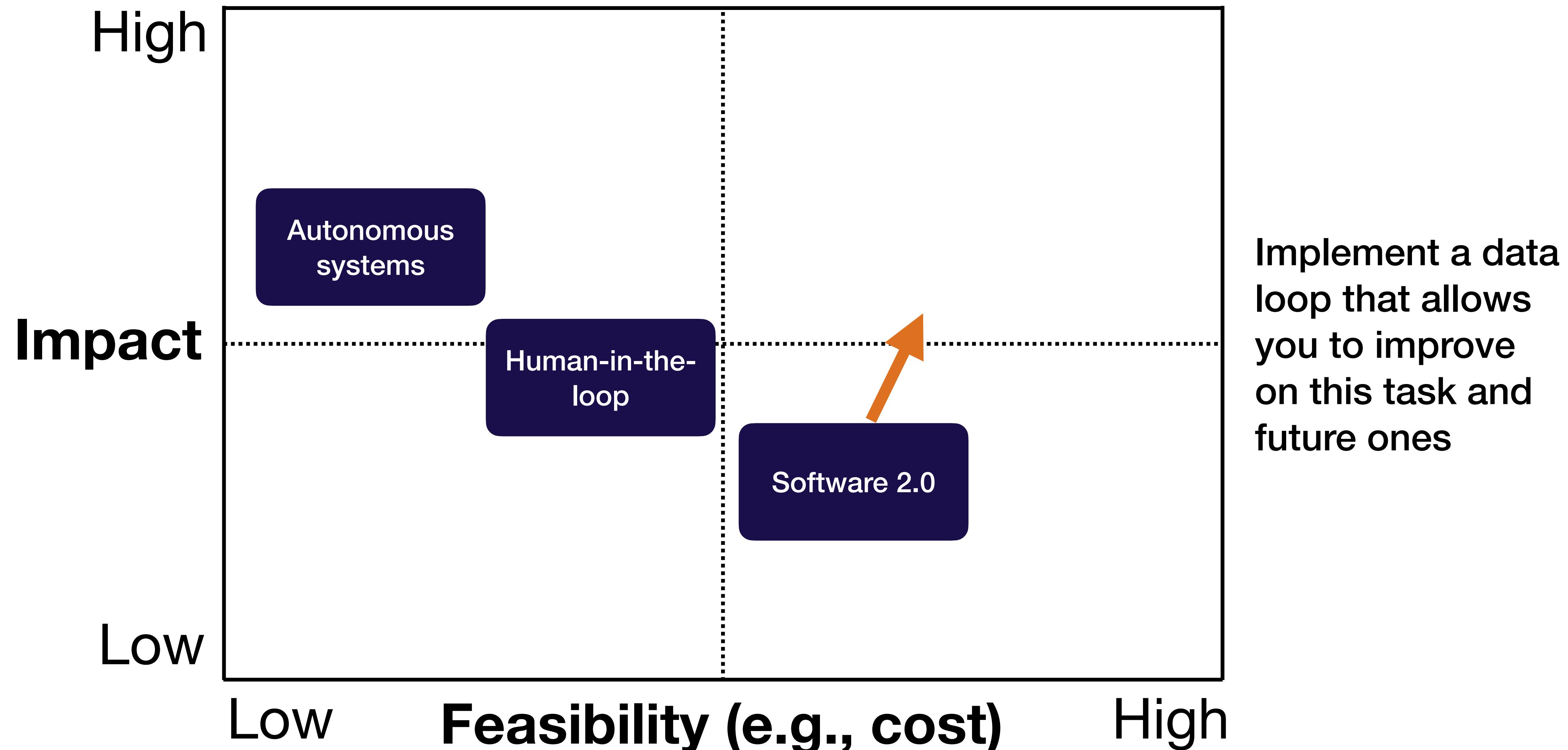
Data flywheels



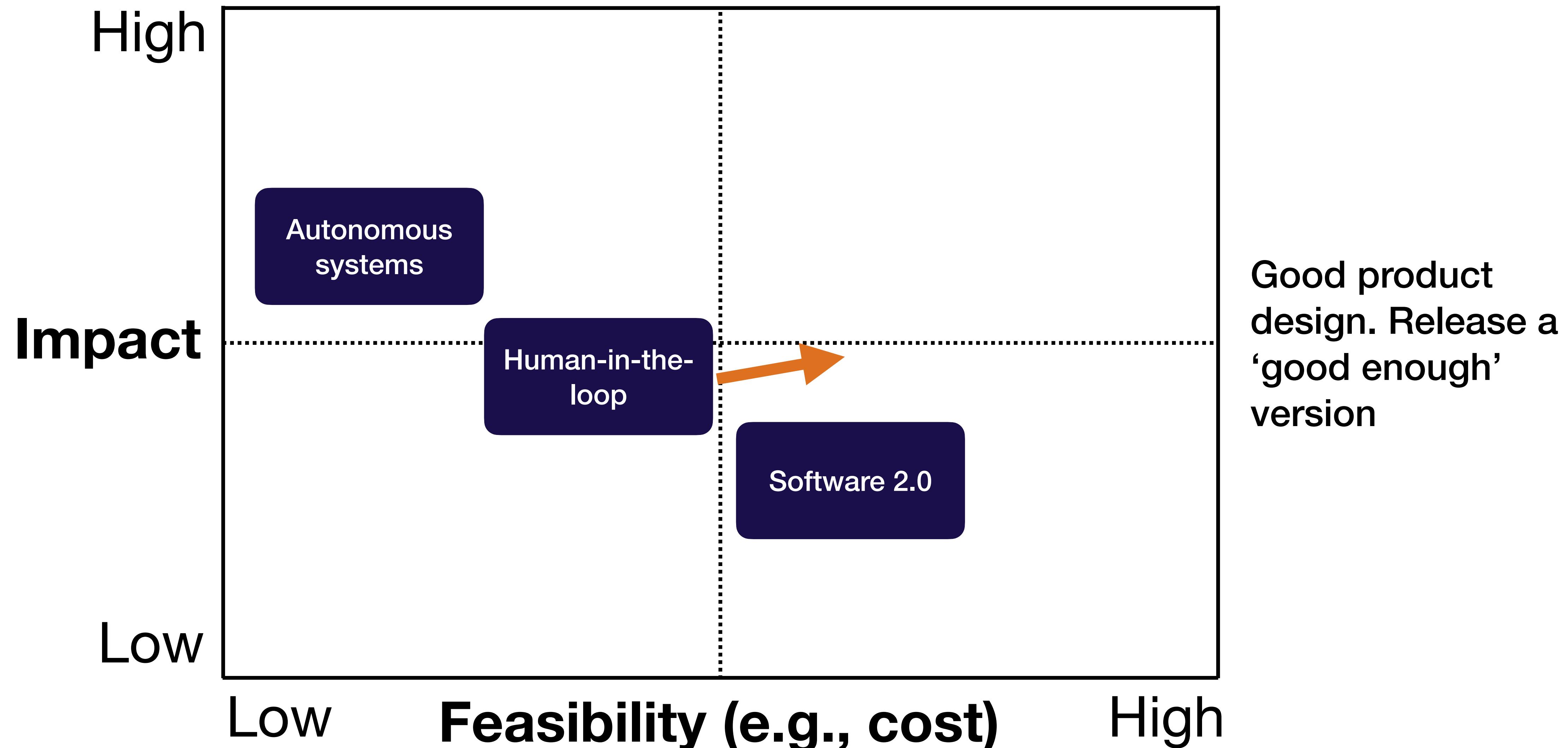
Machine learning project archetypes



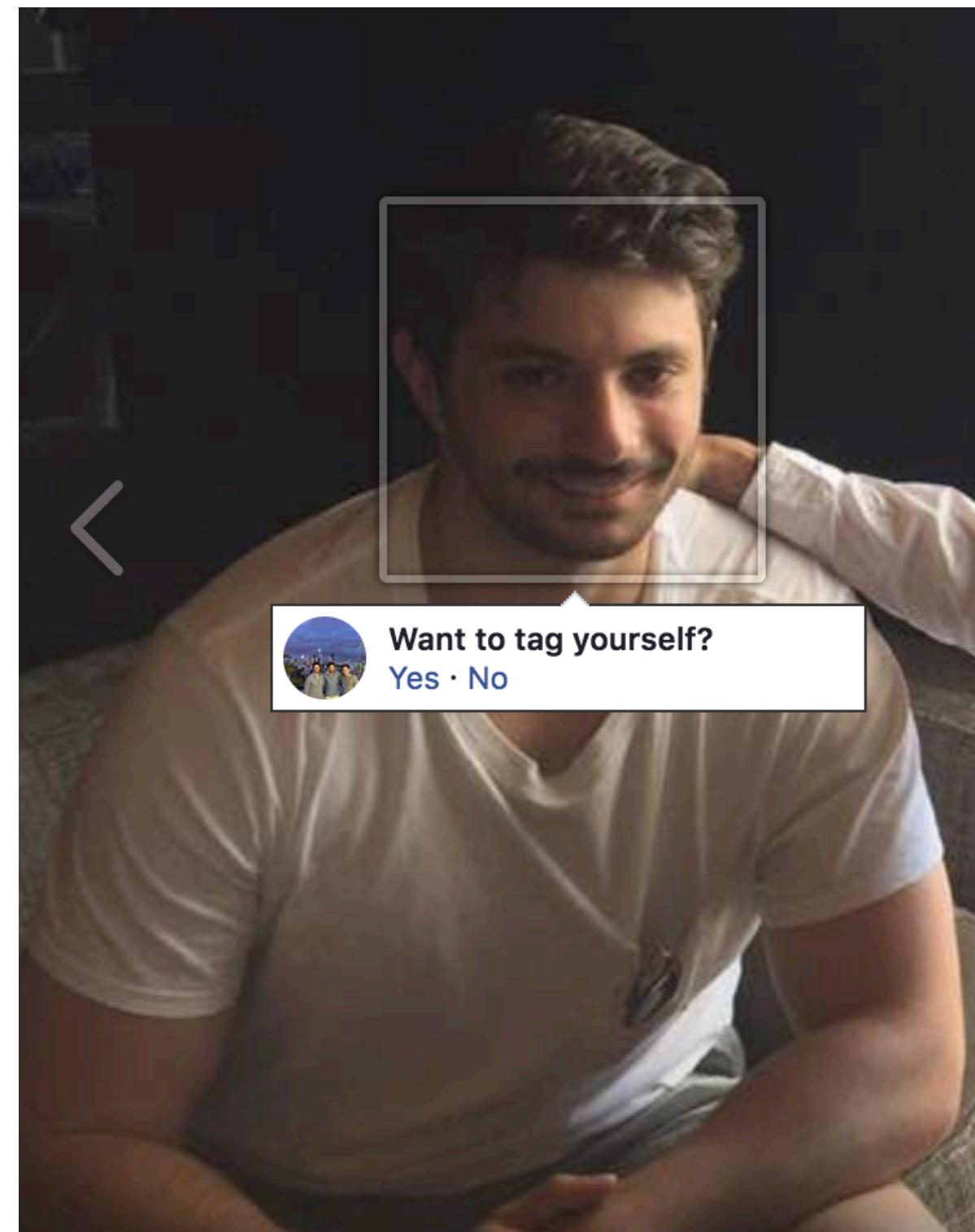
Machine learning product archetypes



Machine learning product archetypes



Product design can reduce need for accuracy



Grammarly

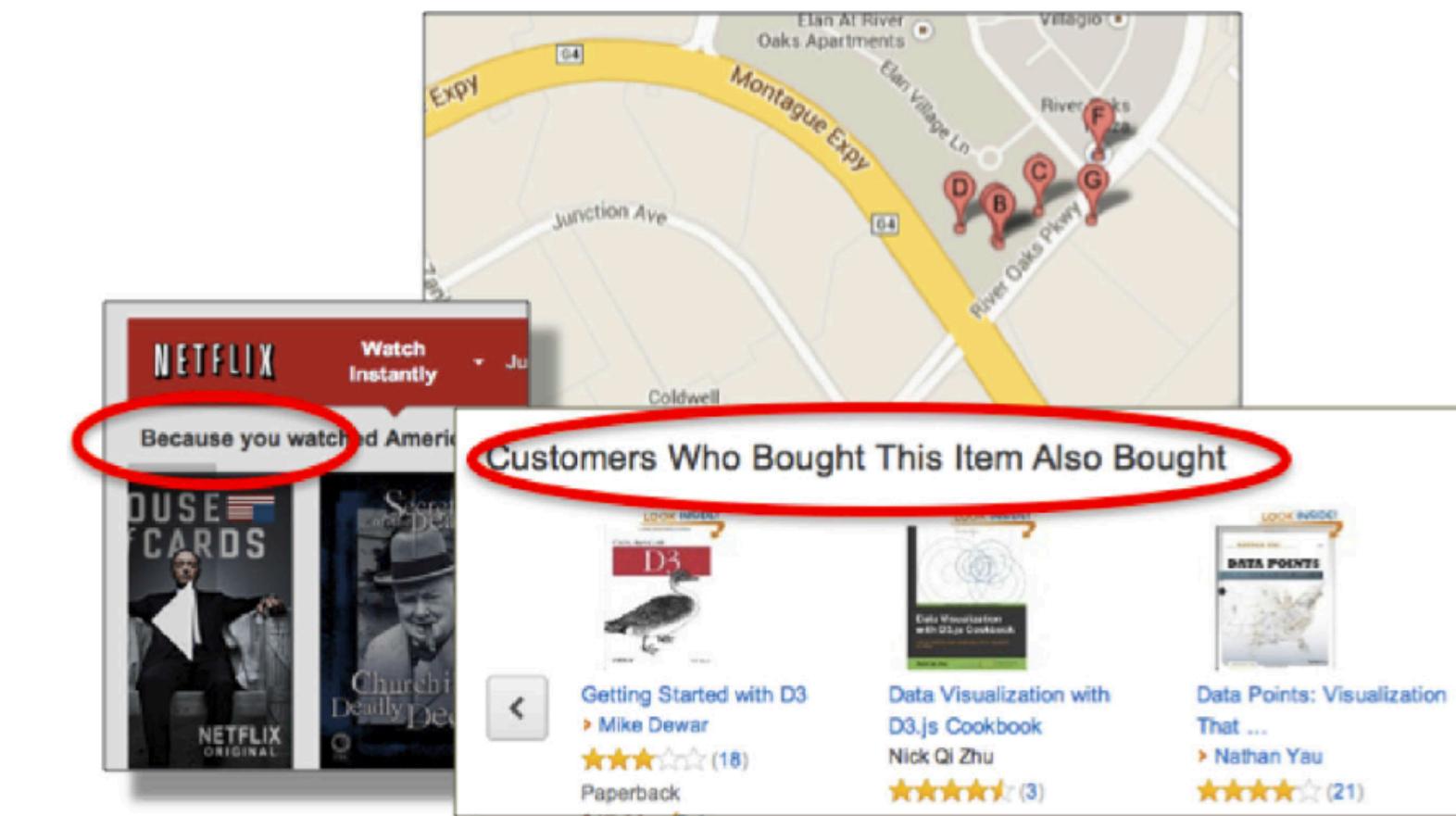
It can iterate with you in a tight feedback loop, proposing additional relevant issues to consider. It also exposes users to (potentially) new knowledge that they can apply themselves in the future without need for system critiquing.

the (or a (

The noun phrase *(potentially) new knowledge* seems to be missing a determiner before it. Consider adding an article.

An article (*a*, *an*, or *the*) is a type of determiner. Possessive adjectives (*my*, *his*, *our*), possessive nouns (*Joe's*, *mother's*), and quantifiers (*each*, *every*) are also determiners. Single countable nouns usually require a determiner.

Incorrect: I left book on table.
Correct: I left *a* book on *the* table.
Correct: I left *the* book on *a* table.



See “Designing Collaborative AI” (Ben Reinhardt and Belmer Negrillo):
https://medium.com/@Ben_Reinhardt/designing-collaborative-ai-5c1e8dbc8810



Apple's ML product design guidelines

- What role does ML play in your app?
 - Critical or complementary?
 - Private or public?
 - Proactive or retroactive?
 - Visible or invisible?
 - Dynamic or static?

<https://developer.apple.com/design/human-interface-guidelines/machine-learning/overview/introduction/>

Apple's ML product design guidelines

- How can you learn from your users?
 - Explicit feedback (e.g., “suggest less pop music”)
 - Implicit feedback (e.g., “like this song”)
 - Calibration during setup (e.g., scan your face for FaceID)
 - Corrections (e.g., fix mistakes model made)

<https://developer.apple.com/design/human-interface-guidelines/machine-learning/overview/introduction/>

Apple's ML product design guidelines

- How should your app handle mistakes?
 - Limitations (let your users know where you expect model to perform well)
 - Corrections (let your users succeed even if the model fails)
 - Attributions (help users understand where suggestions come from)
 - Confidence (help users gauge quality of results)

<https://developer.apple.com/design/human-interface-guidelines/machine-learning/overview/introduction/>

Machine learning product design

Guidelines for Human-AI Interaction

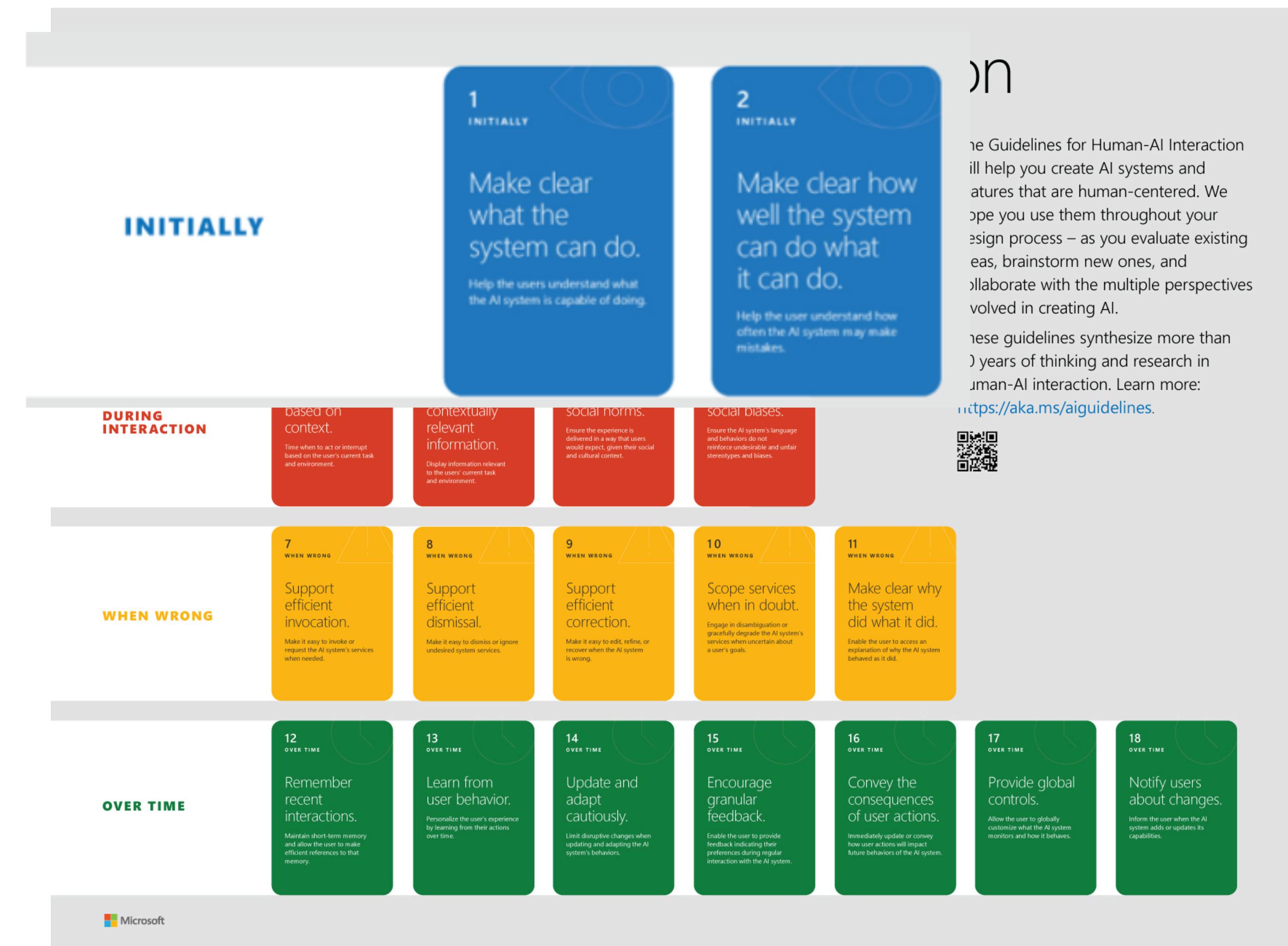


Guidelines for Human-AI Interaction

<https://www.microsoft.com/en-us/research/project/guidelines-for-human-ai-interaction/>



Machine learning product design



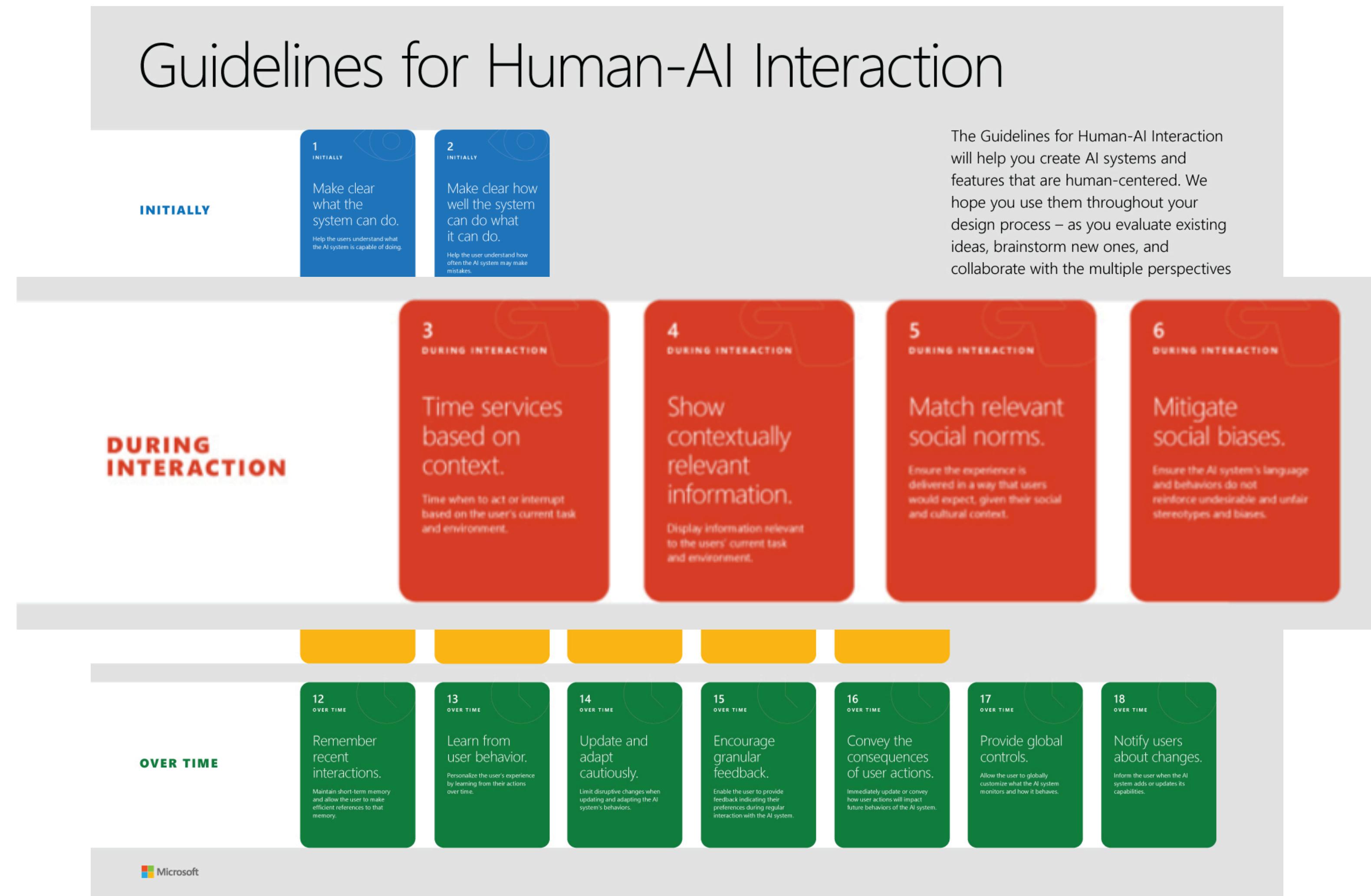
Guidelines for Human-AI Interaction

<https://www.microsoft.com/en-us/research/project/guidelines-for-human-ai-interaction/>



Machine learning product design

Guidelines for Human-AI Interaction



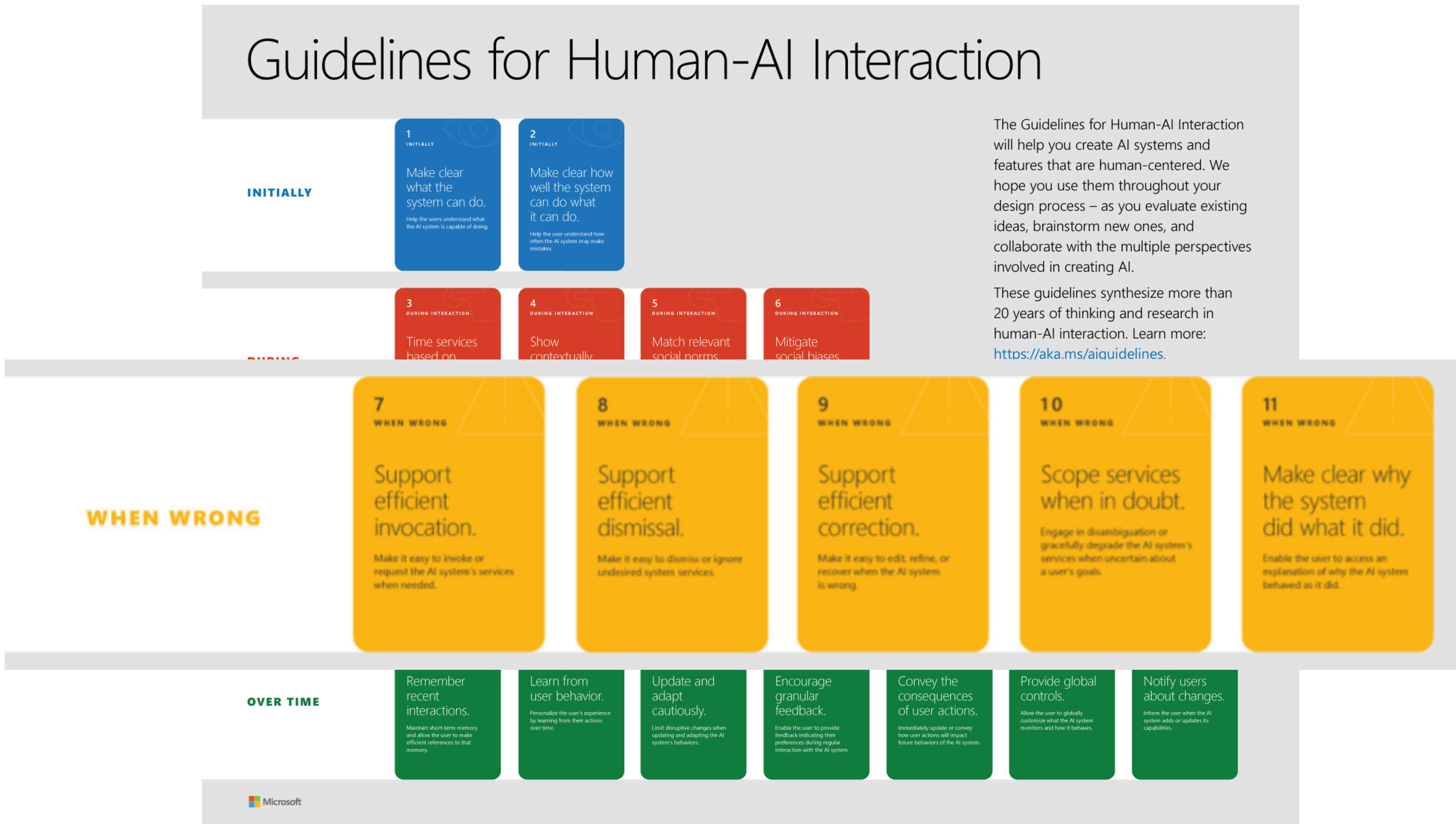
Guidelines for Human-AI Interaction

<https://www.microsoft.com/en-us/research/project/guidelines-for-human-ai-interaction/>



Machine learning product design

Guidelines for Human-AI Interaction



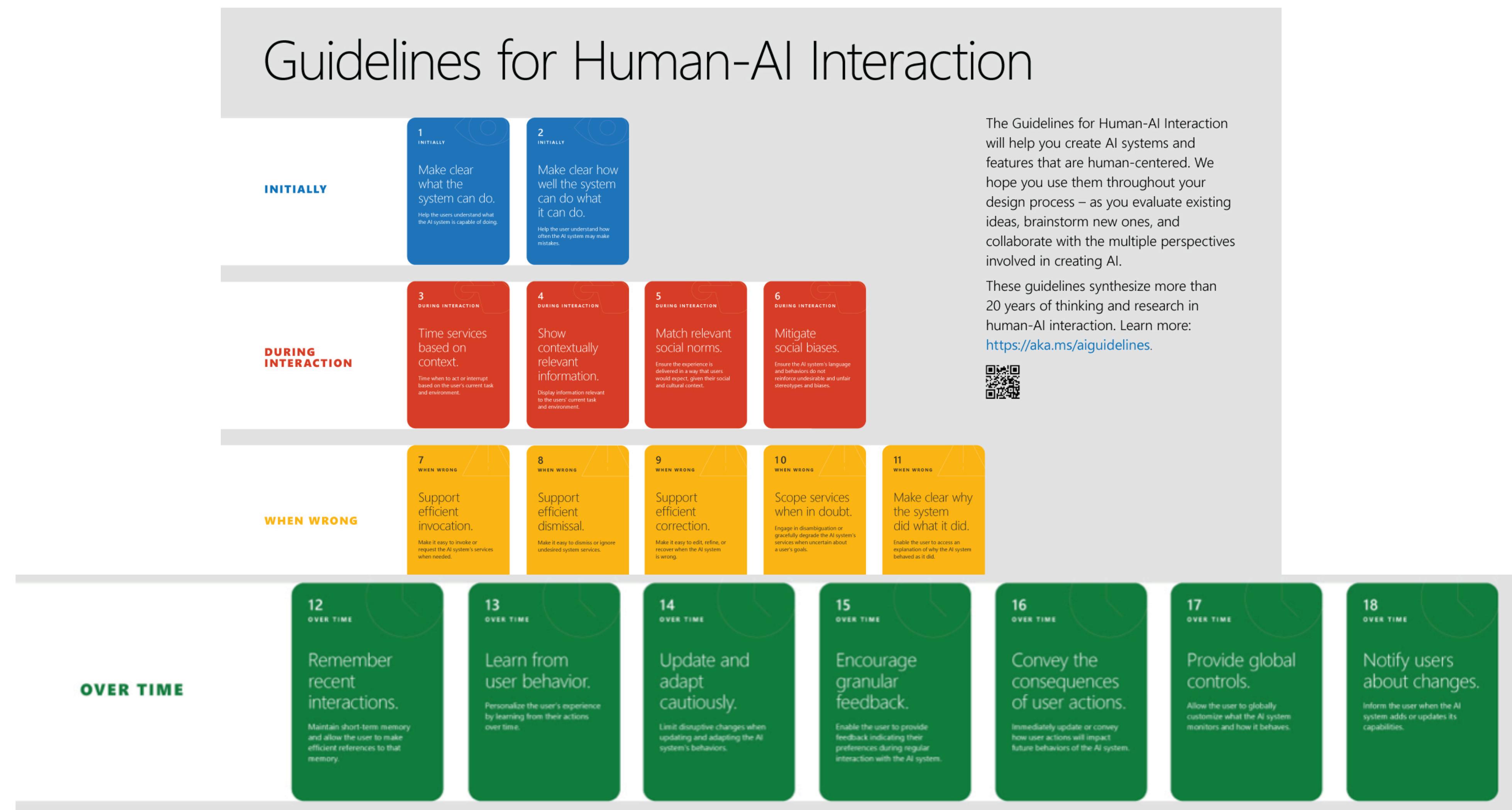
Guidelines for Human-AI Interaction

<https://www.microsoft.com/en-us/research/project/guidelines-for-human-ai-interaction/>



Machine learning product design

Guidelines for Human-AI Interaction

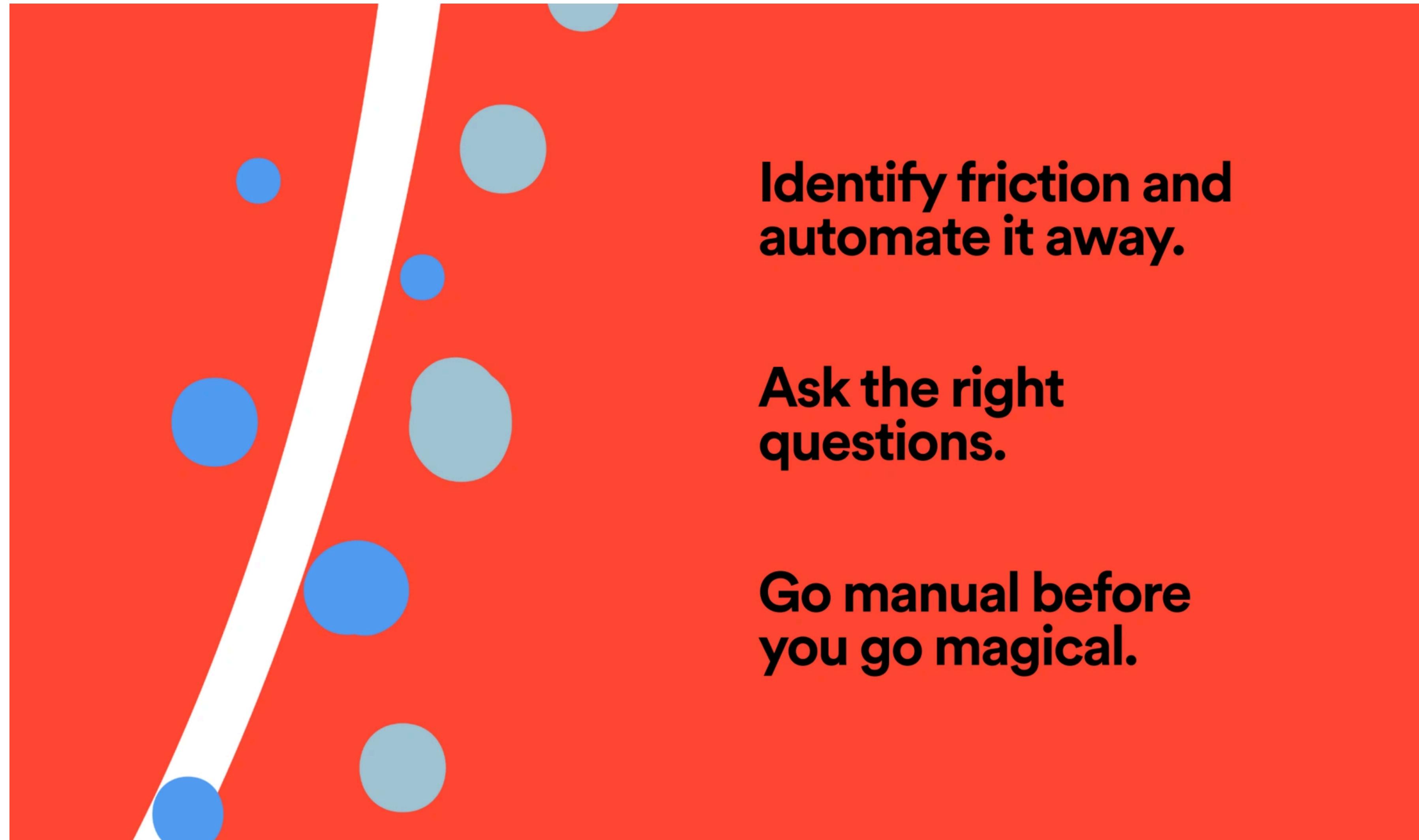


Guidelines for Human-AI Interaction

<https://www.microsoft.com/en-us/research/project/guidelines-for-human-ai-interaction/>

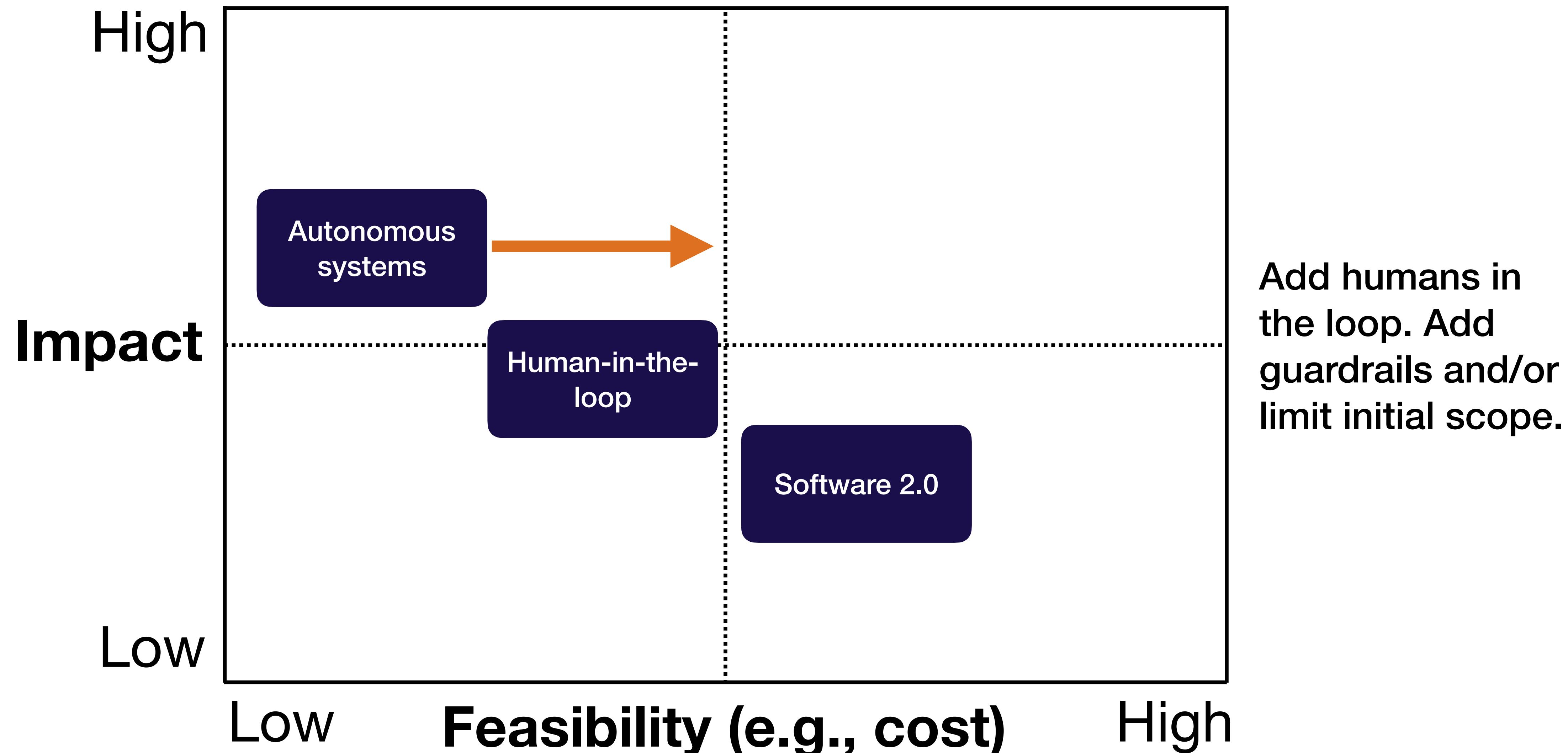


Product design can reduce need for accuracy



<https://spotify.design/article/three-principles-for-designing-ml-powered-products>

Machine learning product archetypes



Questions?



Module overview

- How to think about all of the activities in an ML project
 - Assessing the feasibility and impact of your projects
 - The main categories of ML projects, and the implications for project management
 - **How to pick a single number to optimize**
 - How to know if your model is performing well
-

Key points for choosing a metric

- A. The real world is messy; you usually care about lots of metrics
- B. However, ML systems work best when optimizing a single number
- C. As a result, you need to pick a formula for combining metrics
- D. This formula can and will change!

Review of accuracy, precision, and recall

Confusion matrix

		Predicted:	
		NO	YES
n=100	Predicted:		
	NO	5	5
Actual: NO	YES	5	10
Actual: YES	NO	45	45
YES	YES	90	50
		50	50

Review of accuracy, precision, and recall

Confusion matrix

		n=100	
		Predicted: NO	Predicted: YES
		Actual: NO	Actual: YES
Actual: NO	5	5	10
Actual: YES	45	45	90
	50	50	

Accuracy

50%

Correct

Total



Review of accuracy, precision, and recall

Confusion matrix

n=100	Predicted: NO	Predicted: YES	
Actual: NO	5	5	10
Actual: YES	45	45	90
	50	50	

Precision

true positives

true positives + false positives

90%



Review of accuracy, precision, and recall

Confusion matrix

n=100	Predicted: NO	Predicted: YES	
Actual: NO	5	5	10
Actual: YES	45	45	90
	50	50	

Recall

50%

true positives

Actual YES



Why choose a single metric?

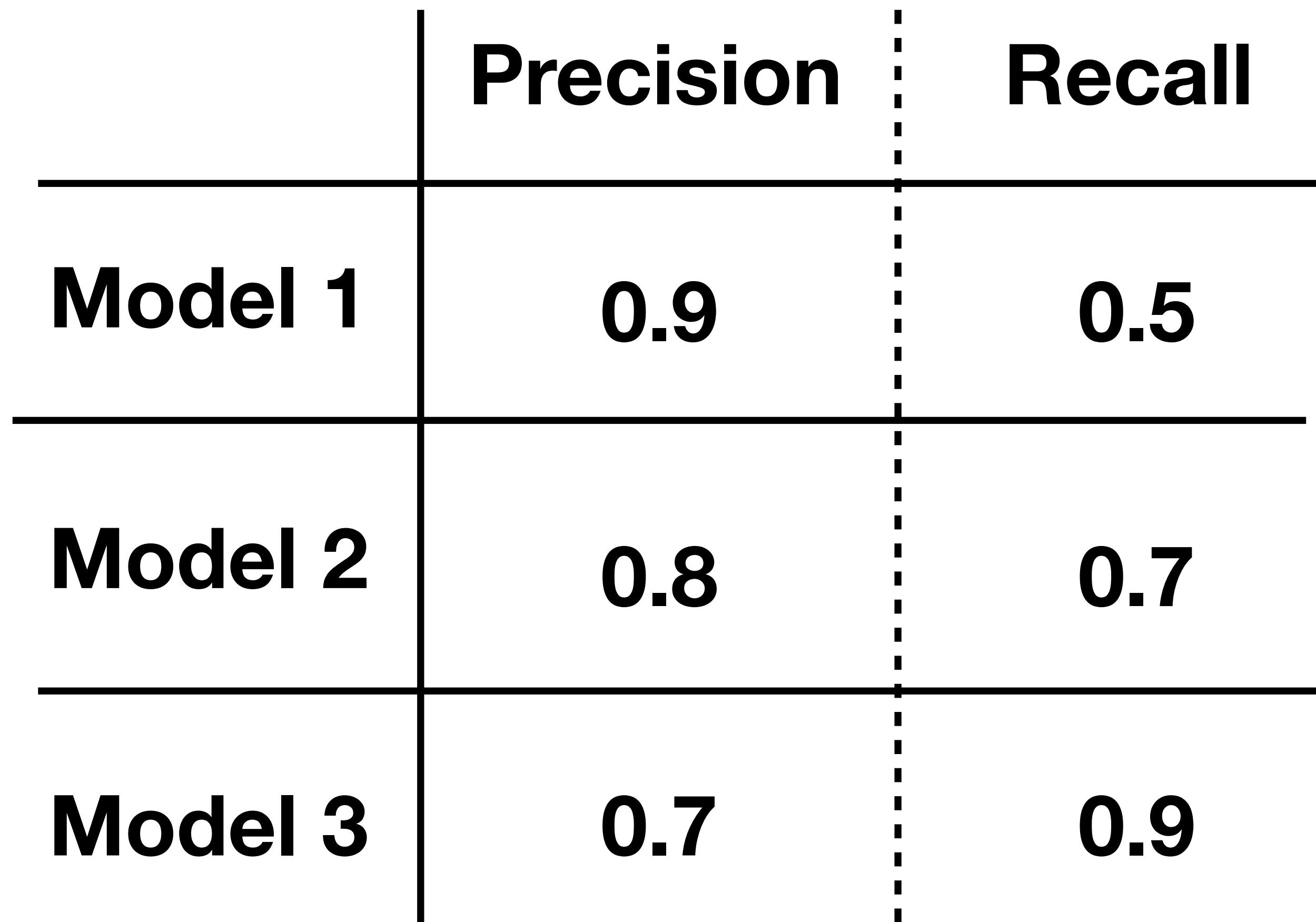


Which is best?

How to combine metrics

- Simple average / weighted average

Combining precision and recall



Combining precision and recall

	Precision	Recall	$(p + r) / 2$
Model 1	0.9	0.5	0.7
Model 2	0.8	0.7	0.75
Model 3	0.7	0.9	0.8

Combining precision and recall

	Precision	Recall	$(p + r) / 2$
Model 1	0.9	0.5	0.7
Model 2	0.8	0.7	0.75
Model 3	0.7	0.9	0.8

How to combine metrics

- Simple average / weighted average

How to combine metrics

- Simple average / weighted average
- Threshold n-1 metrics, evaluate the nth

Thresholding metrics

Choosing which metrics to threshold

- Domain judgment (e.g., which metrics can you engineer around?)
- Which metrics are least sensitive to model choice?
- Which metrics are closest to desirable values?

Choosing threshold values

- Domain judgment (e.g., what is an acceptable tolerance downstream? What performance is achievable?)
- How well does the baseline model do?
- How important is this metric right now?



Combining precision and recall

	Precision	Recall	$(p + r) / 2$
Model 1	0.9	0.5	0.7
Model 2	0.8	0.7	0.75
Model 3	0.7	0.9	0.8

Combining precision and recall

	Precision	Recall	$(p + r) / 2$	$p @ (r > 0.6)$
Model 1	0.9	0.5	0.7	0.0
Model 2	0.8	0.7	0.75	0.8
Model 3	0.7	0.9	0.8	0.7

Combining precision and recall

	Precision	Recall	$(p + r) / 2$	$p @ (r > 0.6)$
Model 1	0.9	0.5	0.7	0.0
Model 2	0.8	0.7	0.75	0.8
Model 3	0.7	0.9	0.8	0.7

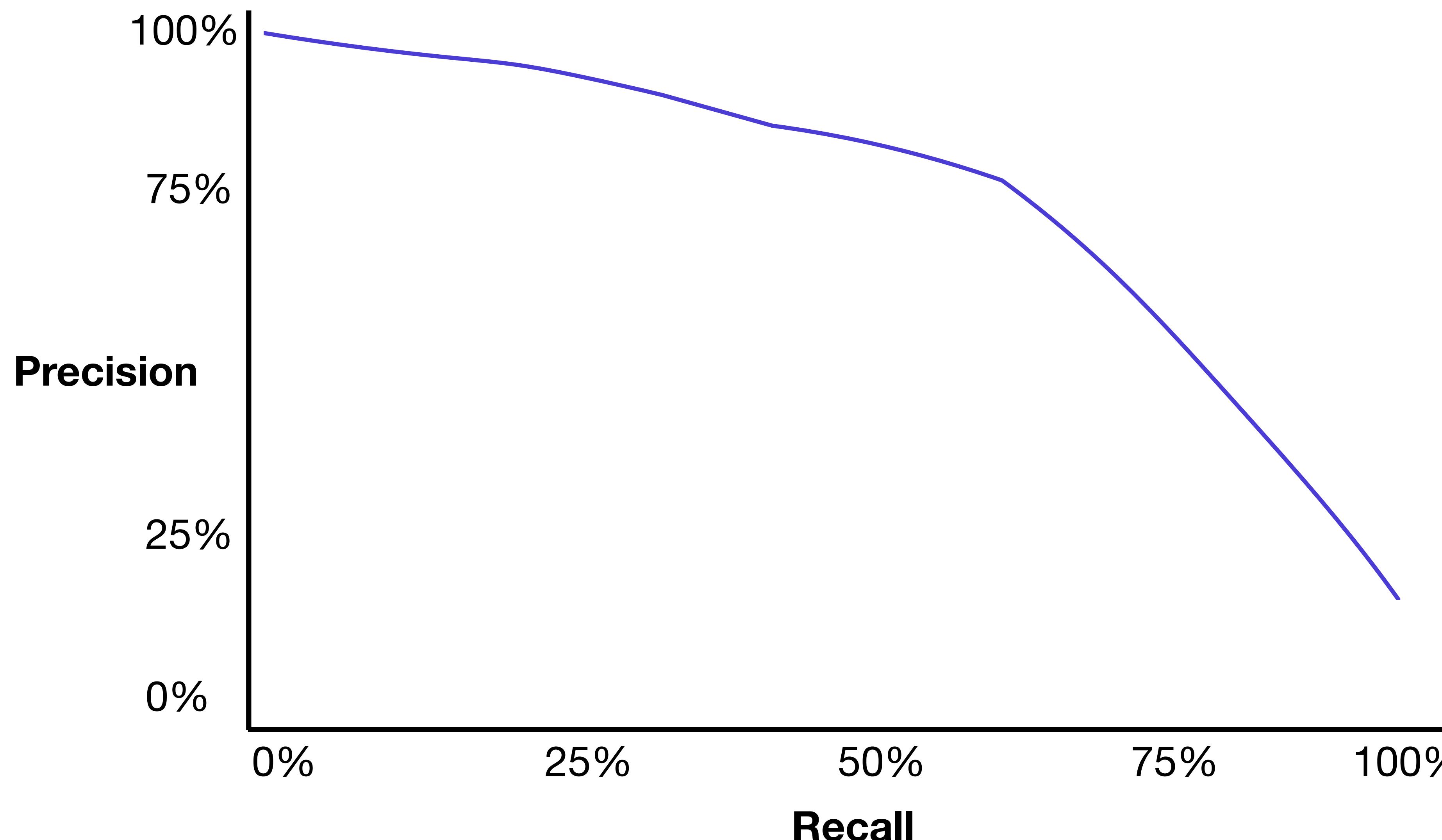
How to combine metrics

- Simple average / weighted average
- Threshold n-1 metrics, evaluate the nth

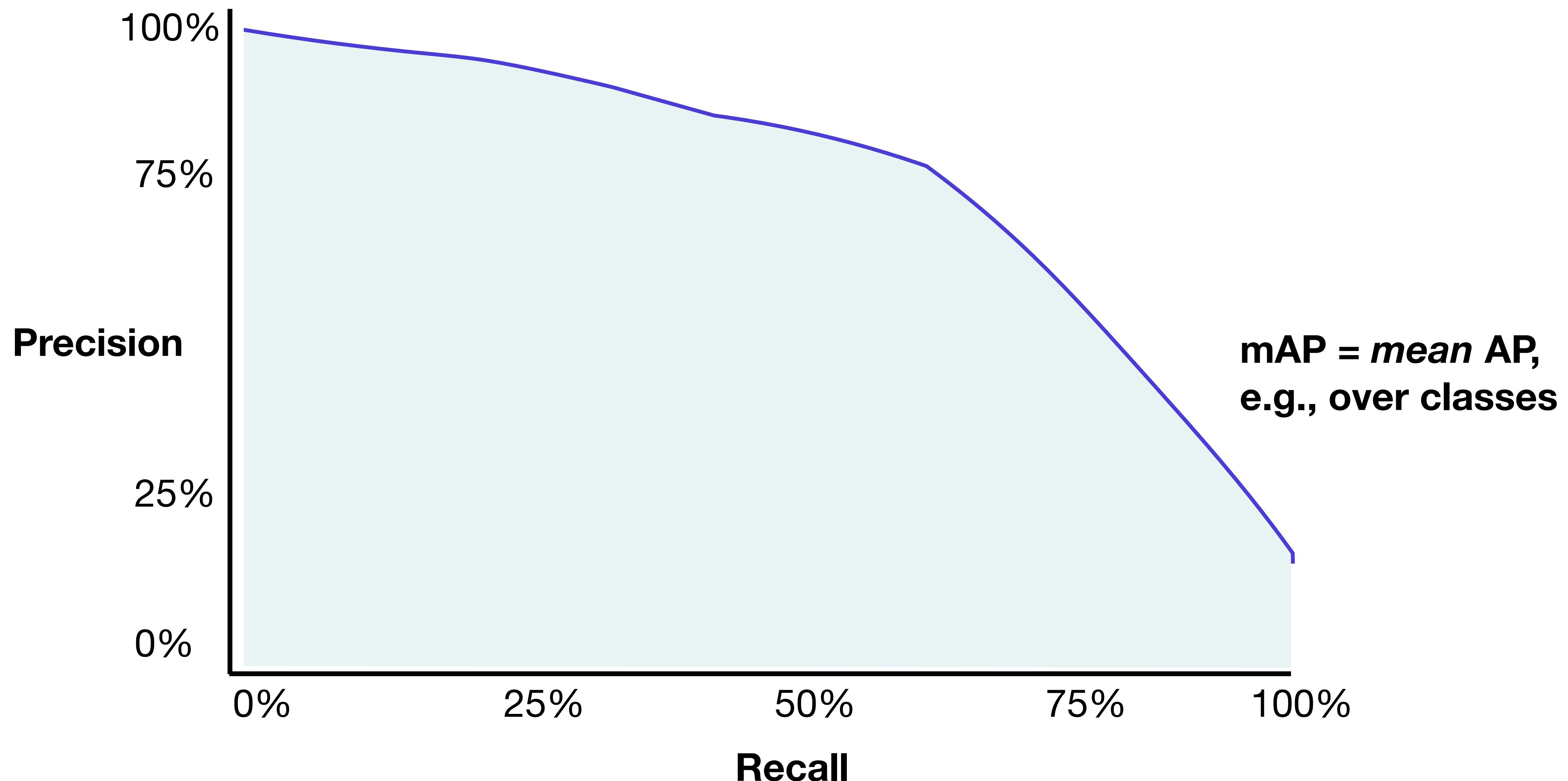
How to combine metrics

- Simple average / weighted average
- Threshold n-1 metrics, evaluate the nth
- More complex / domain-specific formula

Domain-specific metrics: mAP



Domain-specific metrics: mAP



Combining precision and recall

	Precision	Recall	$(p + r) / 2$	$p @ (r > 0.6)$
Model 1	0.9	0.5	0.7	0.0
Model 2	0.8	0.7	0.75	0.8
Model 3	0.7	0.9	0.8	0.7

Combining precision and recall

	Precision	Recall	$(p + r) / 2$	$p @ (r > 0.6)$	mAP
Model 1	0.9	0.5	0.7	0.0	0.7
Model 2	0.8	0.7	0.75	0.8	0.6
Model 3	0.7	0.9	0.8	0.7	0.6

Combining precision and recall

	Precision	Recall	$(p + r) / 2$	$p @ (r > 0.6)$	mAP
Model 1	0.9	0.5	0.7	0.0	0.7
Model 2	0.8	0.7	0.75	0.8	0.6
Model 3	0.7	0.9	0.8	0.7	0.6

Example: choosing a metric for pose estimation



(x, y, z) **Position (L2 loss)**

(ϕ, θ, ψ) **Orientation (L2 loss)**

t

Prediction time

Xiang, Yu, et al. "PoseCNN: A Convolutional Neural Network for 6D Object Pose Estimation in Cluttered Scenes." *arXiv preprint arXiv:1711.00199* (2017).

Example: choosing a metric for pose estimation

- **Enumerate requirements**
 - Downstream goal is real-time robotic grasping
 - Position error must be <1cm, not sure exactly how precise is needed
 - Angular error <5 degrees
 - Must run in 100ms to work in real-time

Example: choosing a metric for pose estimation

- Enumerate requirements
- **Evaluate current performance**
- Train a few models

Example: choosing a metric for pose estimation

- Enumerate requirements
- Evaluate current performance
- **Compare current performance to requirements**
 - Position error between 0.75 and 1.25cm (depending on hyperparameters)
 - All angular errors around 60 degrees
 - Inference time ~300ms

Example: choosing a metric for pose estimation

- Enumerate requirements
- Evaluate current performance
- **Compare current performance to requirements**
 - Prioritize angular error
 - Threshold position error at 1cm
 - Ignore run time for now

Example: choosing a metric for pose estimation

- Enumerate requirements
- Evaluate current performance
- Compare current performance to requirements
- **Revisit metric as your numbers improve**

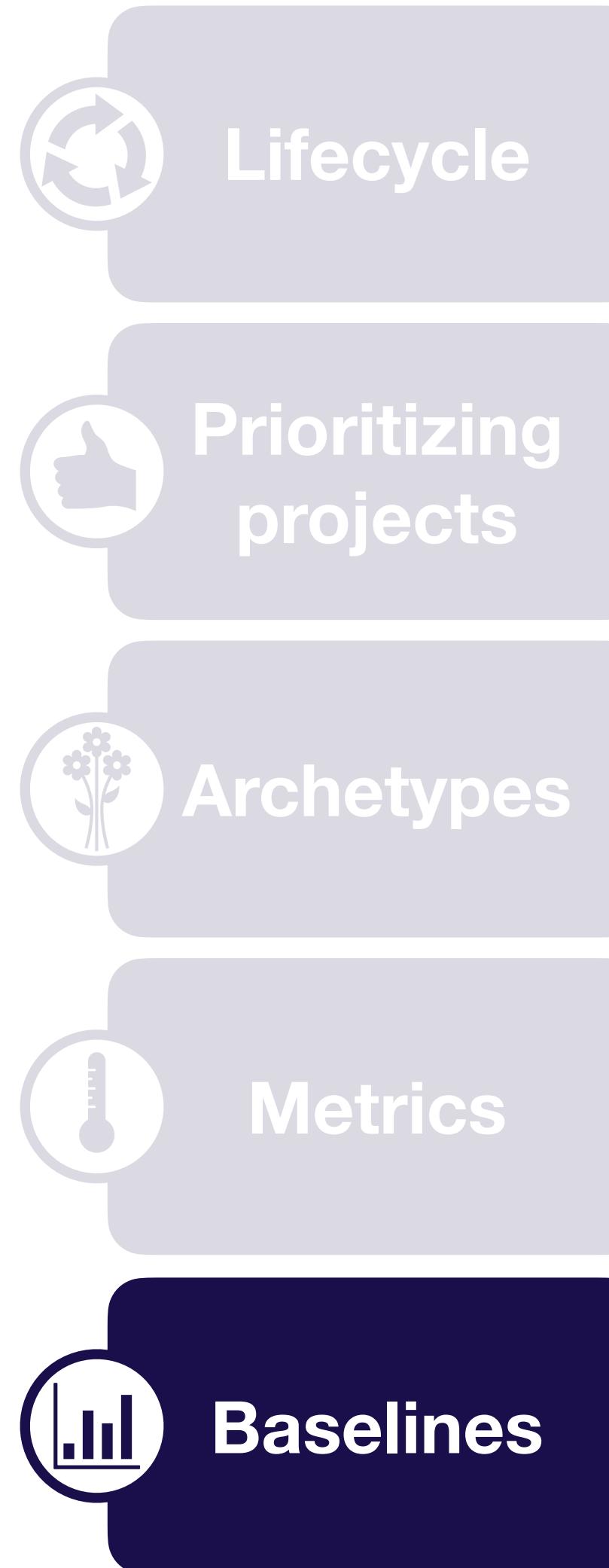
Key points for choosing a metric

- A. The real world is messy; you usually care about lots of metrics
- B. However, ML systems work best when optimizing a single number
- C. As a result, you need to pick a formula for combining metrics
- D. This formula can and will change!

Questions?



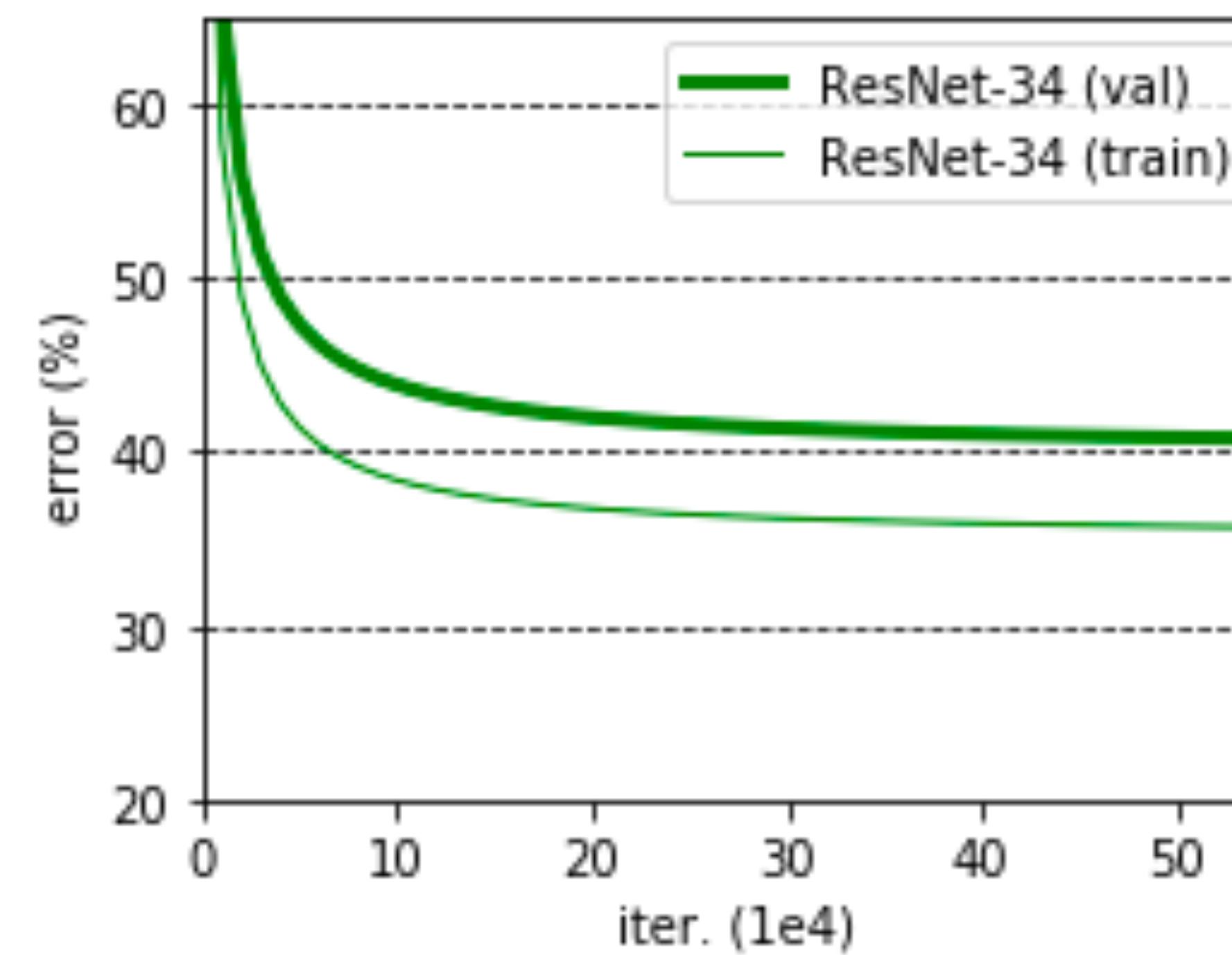
Module overview

- How to think about all of the activities in an ML project
 - Assessing the feasibility and impact of your projects
 - The main categories of ML projects, and the implications for project management
 - How to pick a single number to optimize
 - **How to know if your model is performing well**
- 

Key points for choosing baselines

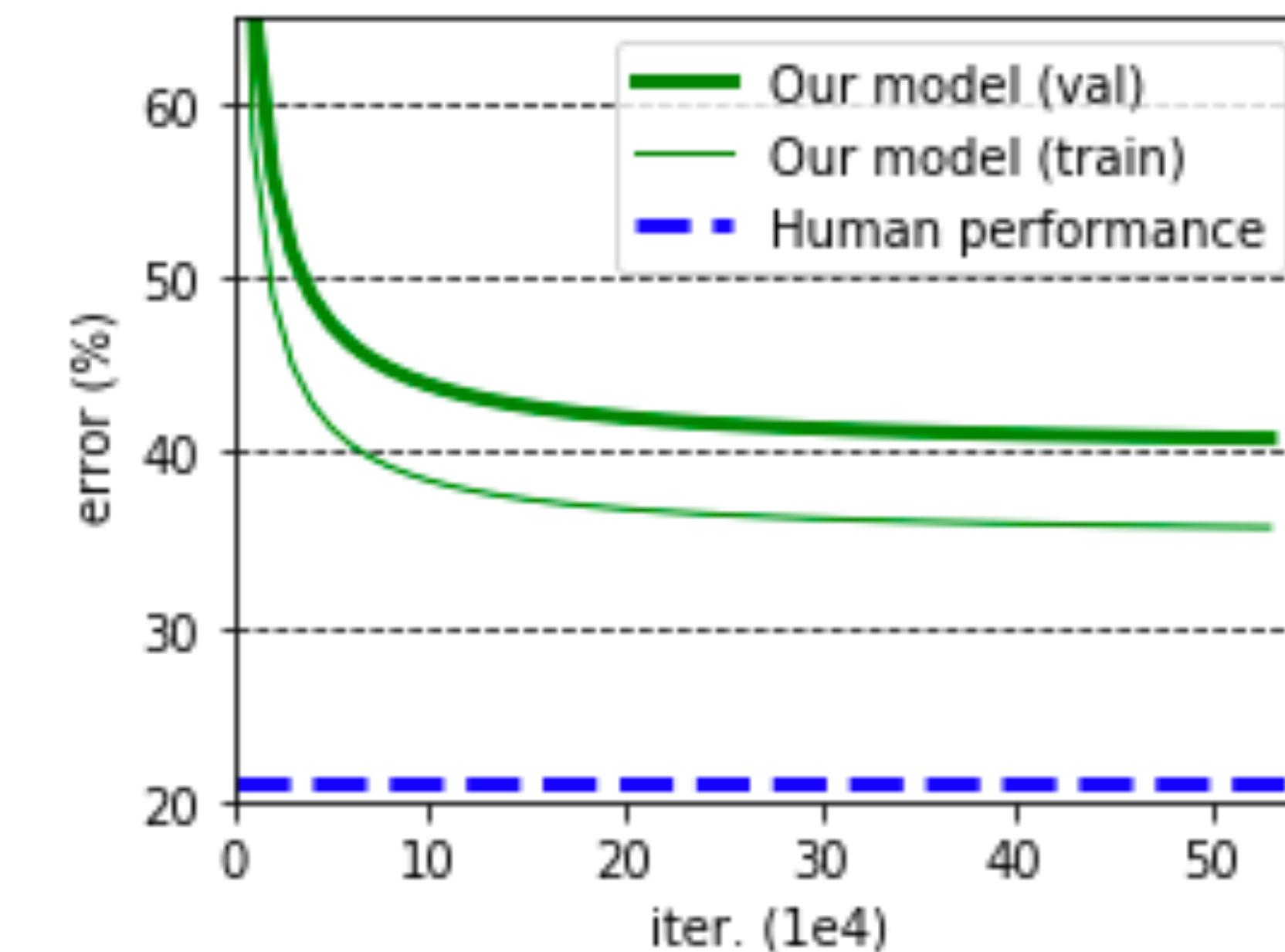
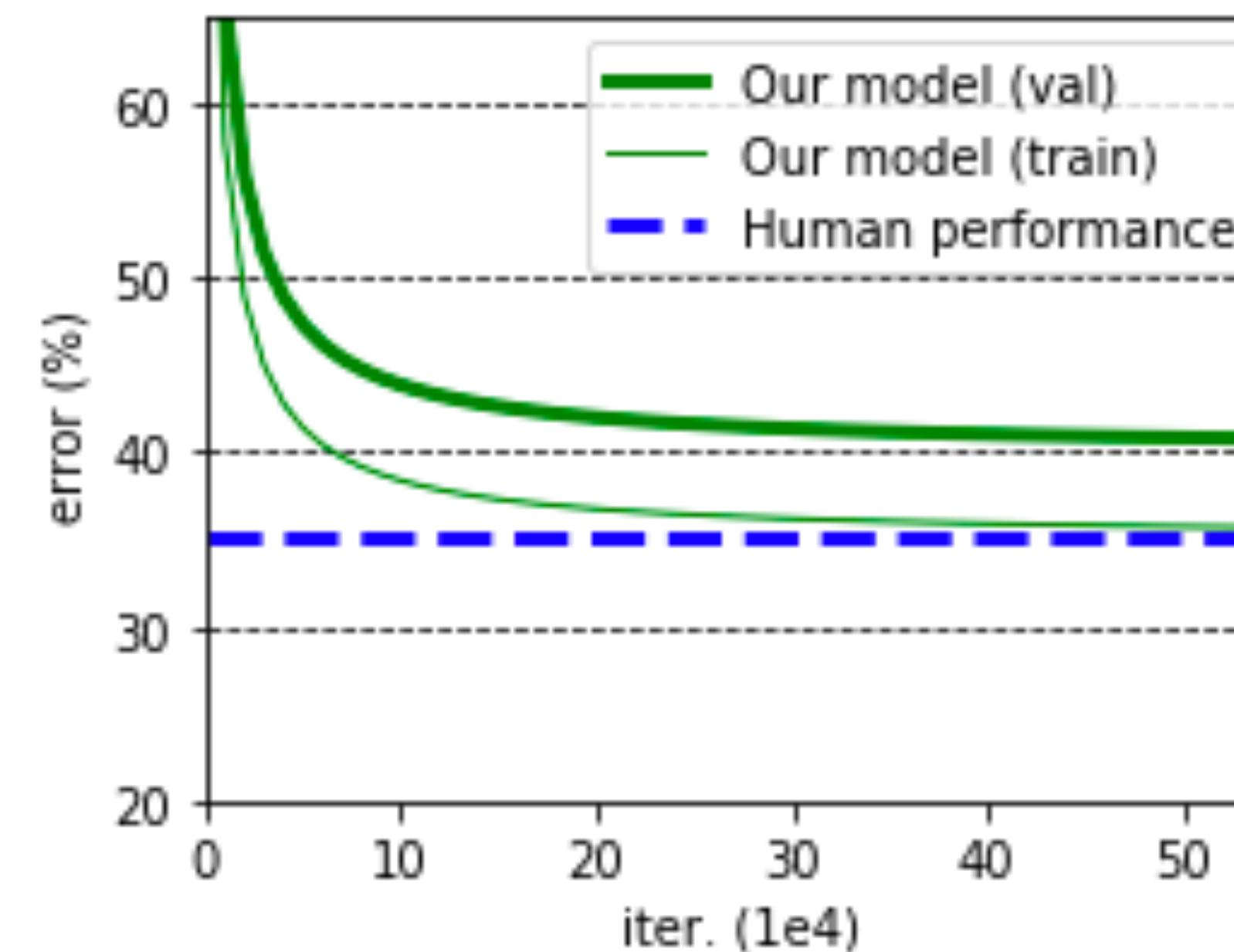
- A. Baselines give you a lower bound on expected model performance
- B. The tighter the lower bound, the more useful the baseline (e.g., published results, carefully tuned pipelines, & human baselines are better)

Why are baselines important?



Why are baselines important?

Same model, different baseline → different next steps



Where to look for baselines

External
baselines

- Business / engineering requirements

Where to look for baselines

External
baselines

- Business / engineering requirements
- Published results

→ **Make sure comparison
is fair!**

Where to look for baselines

External
baselines

- Business / engineering requirements
- Published results

Internal
baselines

- Scripted baselines (e.g., OpenCV, rules-based)

Where to look for baselines

External
baselines

- Business / engineering requirements
- Published results

Internal
baselines

- Scripted baselines (e.g., OpenCV, rules-based)
- Simple ML baselines (e.g., bag of words, linear regression)

Where to look for baselines

External
baselines

- Business / engineering requirements
- Published results

Internal
baselines

- Scripted baselines (e.g., OpenCV, rules-based)
- Simple ML baselines (e.g., bag of words, linear regression)
- Human performance

How to create good human baselines

Quality of baseline

Low

Random people (e.g., Amazon Turk)

Ensemble of random people

Domain experts (e.g., doctors)

Deep domain experts (e.g., specialists)

Mixture of experts

Ease of data collection

High

Low

How to create good human baselines

- Highest quality that allows more data to be labeled easily
- More specialized domains need more skilled labelers
- Find cases where model performs worse and concentrate data collection there

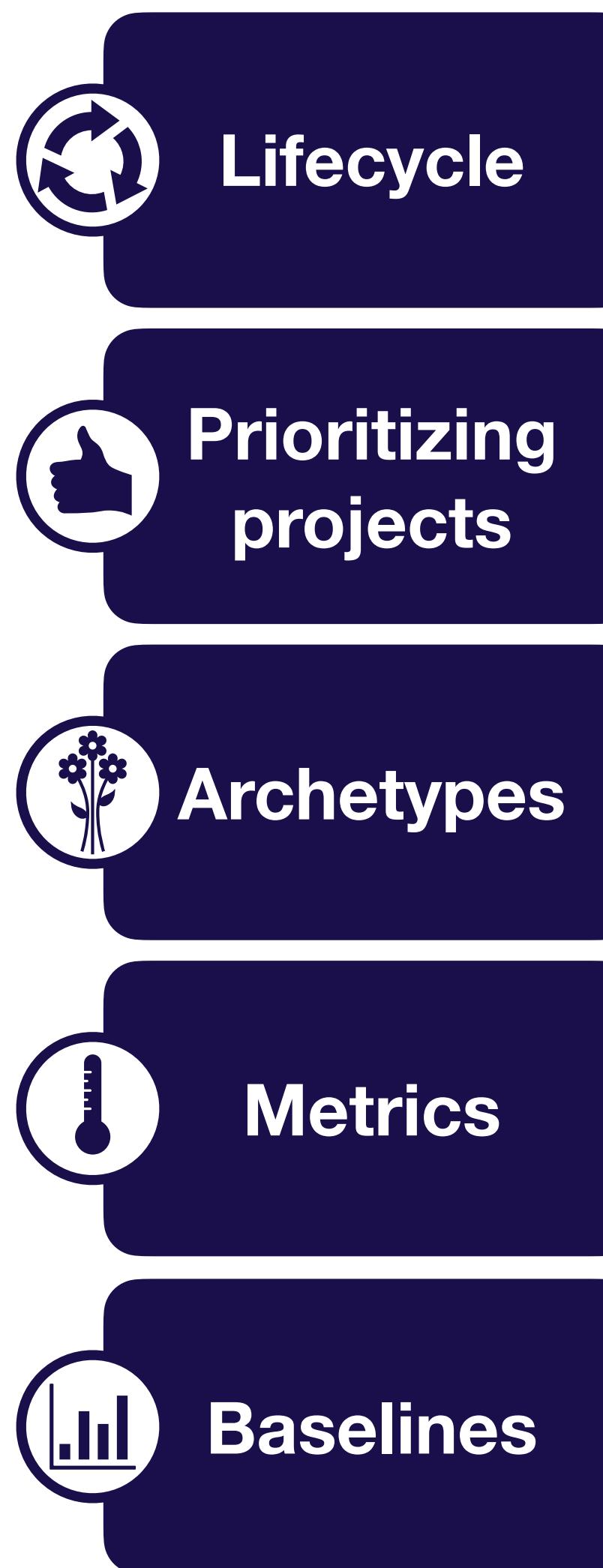
More on labeling in data lecture!

Key points for choosing baselines

- A. Baselines give you a lower bound on expected model performance
- B. The tighter the lower bound, the more useful the baseline (e.g., published results, carefully tuned pipelines, human baselines are better)

Questions?

Conclusion



- ML projects are iterative. Deploy something fast to begin the cycle.
- Choose projects that are high impact with low cost of wrong predictions
- The secret sauce to making projects work well is to build automated data flywheels
- In the real world you care about many things, but you should always have just one you're working on
- Good baselines help you invest your effort the right way

Where to go to learn more

- Andrew Ng’s “Machine Learning Yearning”
- Andrej Karpathy’s “Software 2.0”
- Agrawal’s “The Economics of AI”
- Chip Huyen’s “Introduction to Machine Learning Systems Design”
- Apple’s “Human Interface Guidelines for Machine Learning”
- Google’s “Rules of Machine Learning”

Thank you!