Rm Palaniappan, *Alien Planet-X-9*
Viscosity, pencil colour and ink on handmade paper

Understanding and using the embedding spaces of large generative models

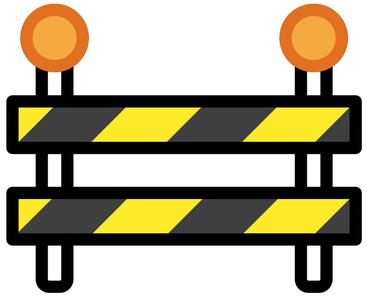
Anand D. Sarwate (Rutgers University)

21 November 2025

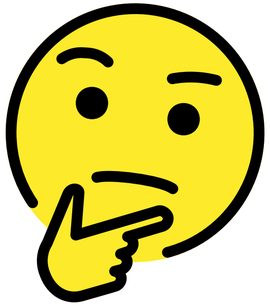
IEEE ITSOC Distinguished Lecture
University of Illinois Chicago
Chicago, IL, USA

Sorry to the theorists

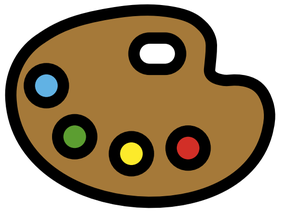
This talk is almost entirely empirical…



Not sure if this talk counts as information ***theory***.



I think there are lots of **interesting questions for theory** which this brings up!



There are a lot of questions in this space that need **creative ideas** to approach!

Thanks to my collaborators/coauthors!

Most of this is their work, obviously

Sinjini Banerjee (Rutgers)

Sutenay Choudhury (PNNL)

Tim Marrinan (PNNL)

Reilly Cannon (Michigan)

Ioana Dumitriu (UC San Diego)

Max Vargas

Tony Chiang (ARPA-H)

Andrew Engel (Ohio State)

Zhichao Wang (UC Berkeley)

Xin Li (Rutgers)

Papers:

[Science Advances (in press)] Vargas et al. https://arxiv.org/abs/2408.10437

[JSTSP 2025] Banerjee et al. https://doi.org/10.1109/JSTSP.2025.3583140
[ICLR 2024] Engel et al. https://openreview.net/forum?id=yKksu38BpM
[NeurIPS 2023] Wang et al. https://openreview.net/forum?id=gpqBGyKeKH

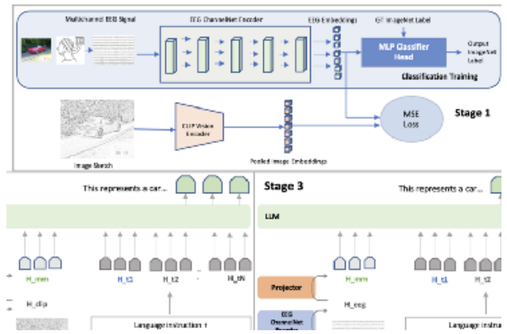
Motivation: AI for Science

An umbrella term for a lot of different ideas/approaches

Source: Wikipedia

Source: IBM

Source: Mishra et al.

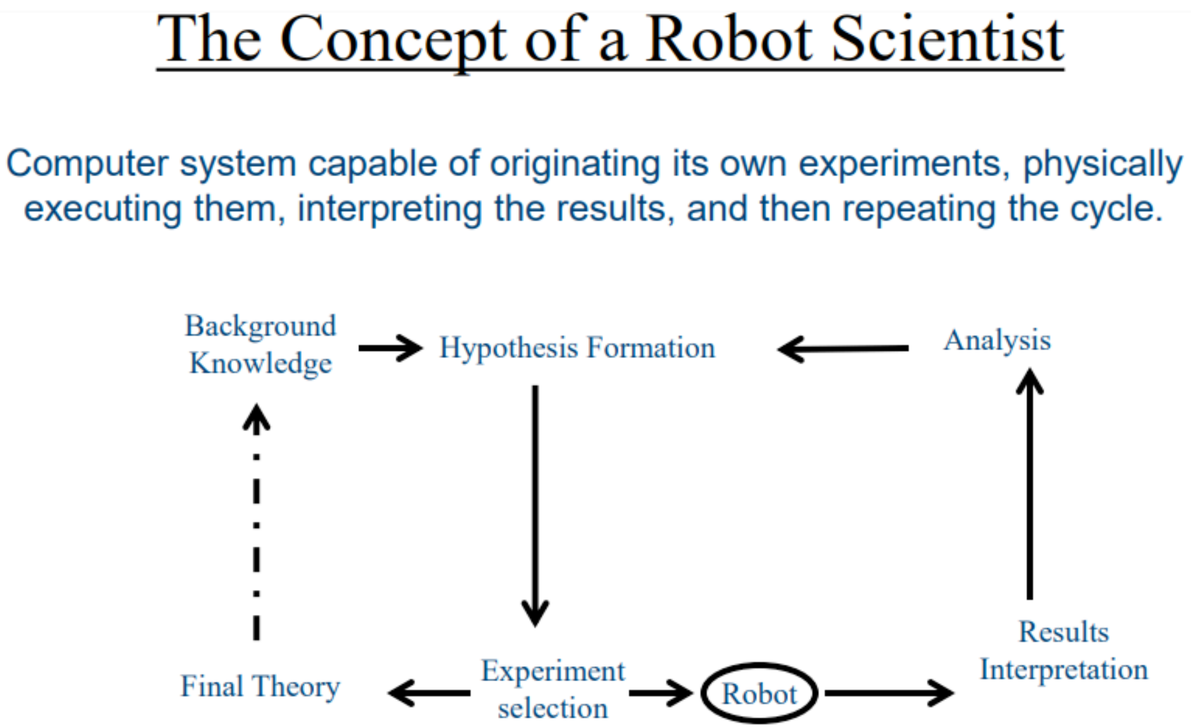


We often hear a lot about "**AI for Science**" but that can mean a lot of different things! Some examples:

- Using computer vision to do automated analysis of medical images.
- Use generative AI to build a "digital twin" of a complex system.
- Use LLM architectures to decode brain activity for assistive technology.
- Many more…

Generative AI for science sounds appealing

But is it feasible?

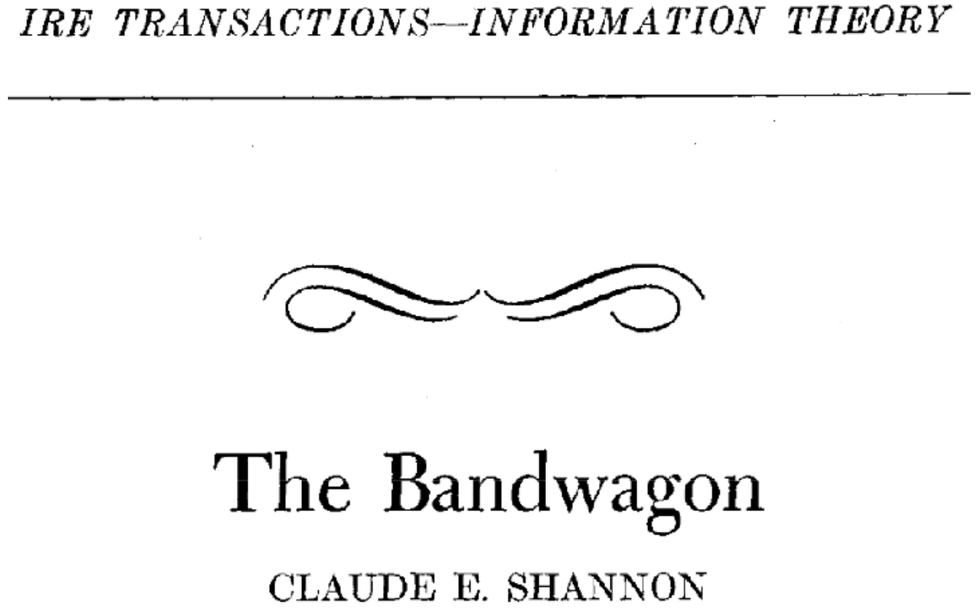


https://futuretech.mit.edu/news/ai-and-the-future-of-scientific-discovery

NY Times
14 May 2025

Your A.I. Radiologist Will Not Be With You Soon

Experts predicted that artificial intelligence would steal radiology jobs. But at the Mayo Clinic, the technology has been more friend than foe.



Shannon, 1956

GenAI models as measurement devices

Or “scientific instruments”

Fintan Mallory (2025): LLMs **during training** are “stochastic measuring devices.”

> *“Just as a chemist puts a thermometer in a glass of liquid to learn its temperature, machine learning researchers put neural networks into datasets to discover patterns within.”*

Maxim Raginsky (2025) ML/AI systems **during inference/generation** map a prompt/input to an internal state which can generate an inference or sample.

> *“[The question is]* what *they are measuring, both during training and during generation or inference.”*

Max Raginsky - https://realizable.substack.com/p/the-metrologic-of-machine-learning
Fintan Mallory - https://philarchive.org/rec/MALLLM

Do AIs extract the “useful” features?

Eventually, will they find “universal” features?

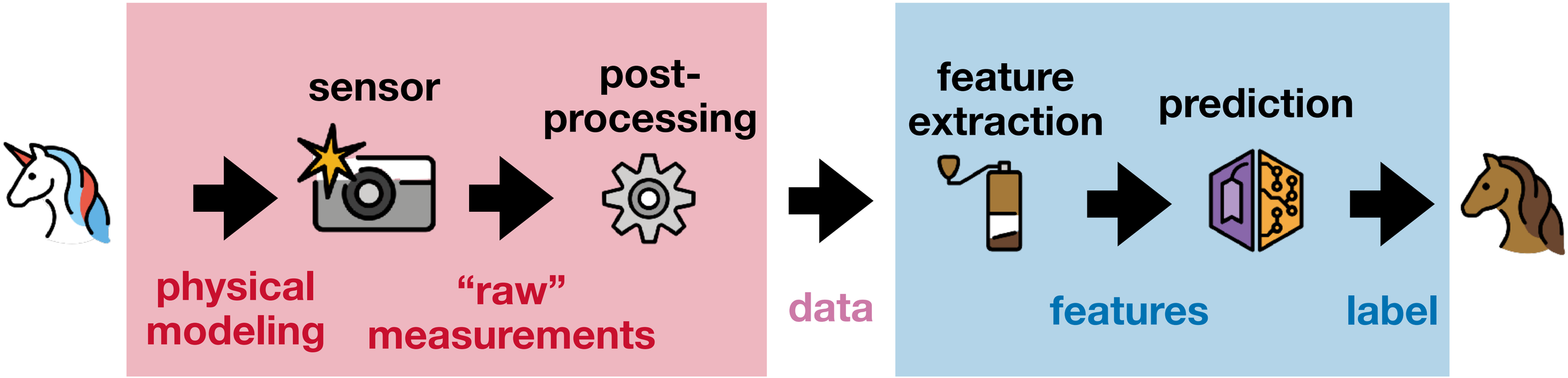
- The “unreasonable effectiveness” of generative AI methods suggests that the features carry rich “semantics”.

- Fine-tuning on a given task shows that the **features can be further processed** into inferences optimized for specific domains.

- Huh et al. (2024)’s “Platonic Representation Hypothesis” claims that scaling up creates “convergence [that] is driving toward a shared statistical model of reality.”

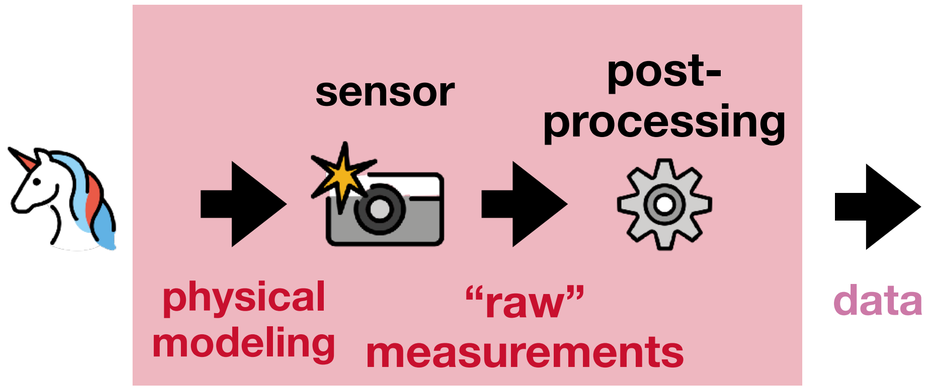
A typical setup for sensing/measurement

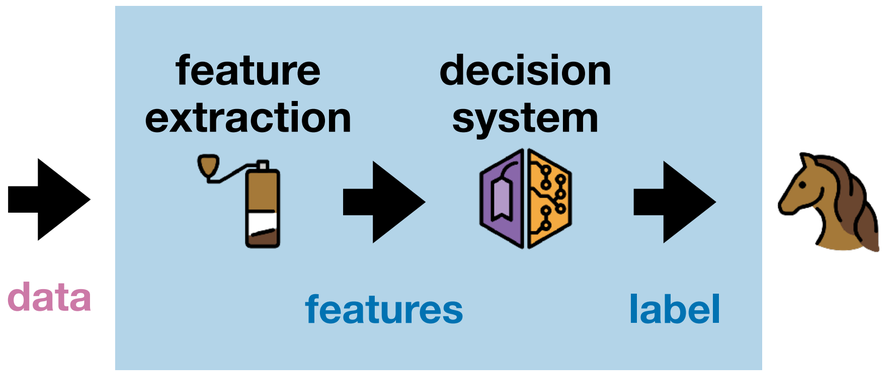
Image acquisition/formation and image processing/analysis



What is “AI as instrumentation”?

Putting neural networks into measurement devices





Data (images, time series, etc.) produced by a **scientific instrument** (camera/microscope/scanner) can be understood in terms of the **science** (physics, chemistry, biology).

We use the data in **analytics pipelines** for more complex tasks. This relies on assumptions:

- Data from the **same camera** is “consistent”.
- Data from **different cameras** are “consistent”.

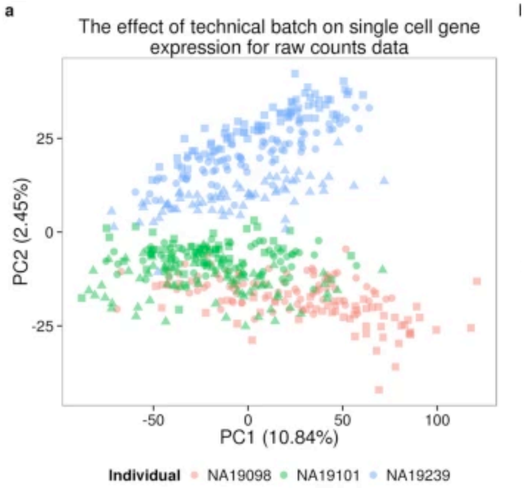
If we put AI “into the camera” will these be true?

Real scientific instruments are not consistent

A lot of work in calibration

Data are almost never consistent in the ways we assume.

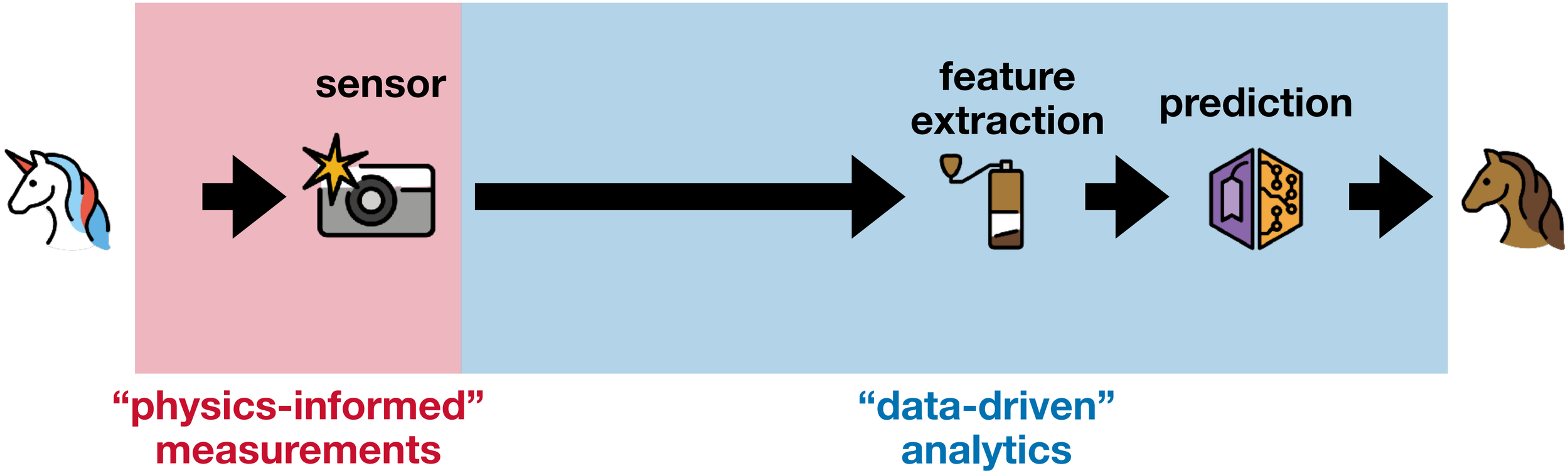
- Calibration issues
- “Batch effects” (c.f. DNA/RNA sequencing)
- Information forensics
- Sampling bias
- Etc…



Images: pngrepo.com, Tung et al. (2017), OpenMoji.org

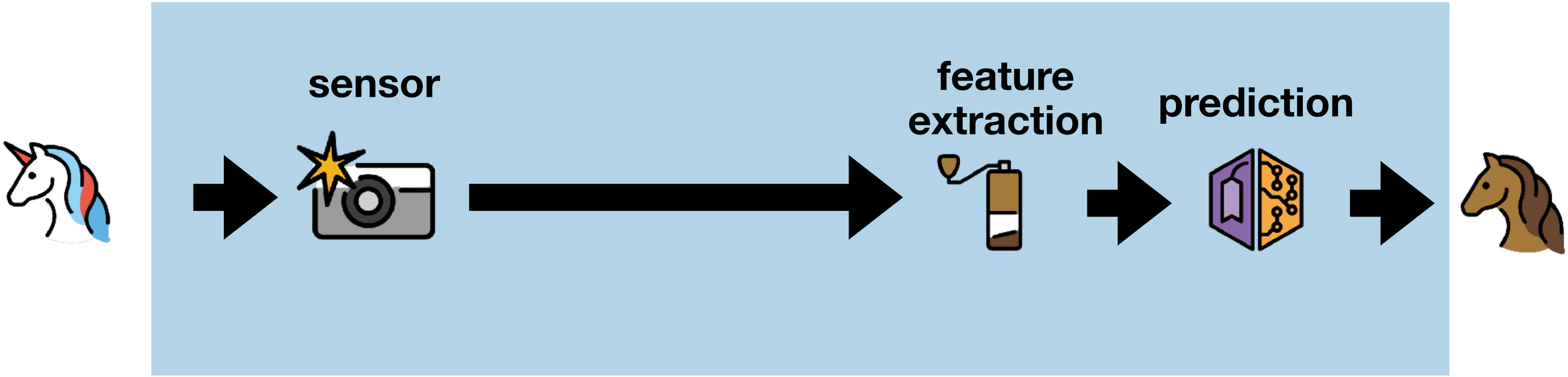
Generating features without “data”

Can an AI extract “all the useful features” from raw measurements?



A possible (and weird) future?

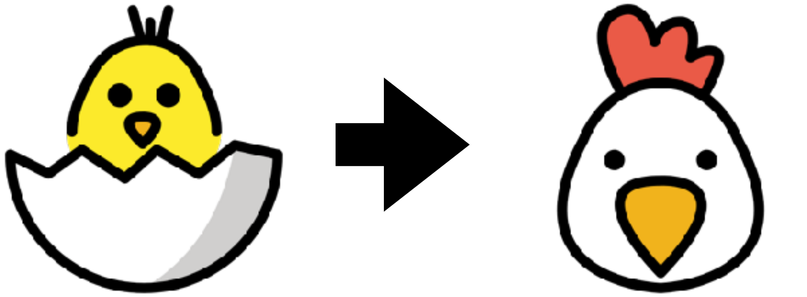
Who needs raw measurements?

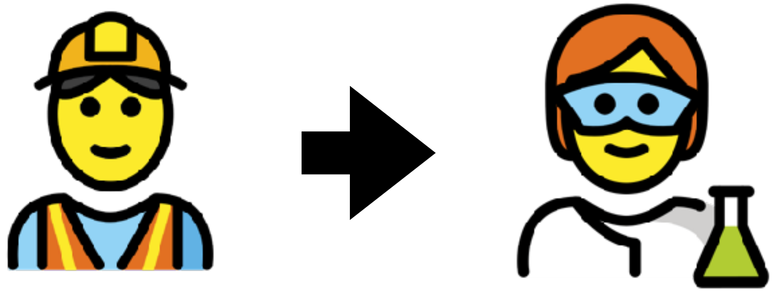


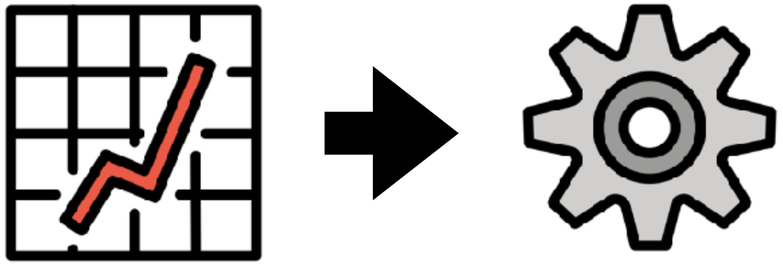
“data-driven”
scientific instrumentation

Research challenges on the theory side

Tracking the moving target of AI practice







At the end of the day "artificial neural nets" are just a bunch of computational signal processing primitives chained together and jointly optimized with stochastic gradient methods.

- Ben Recht (on argmin.net)

ML/AI frameworks are evolving very quickly.

➜ Theory often lags behind practice.

➜ IT, SP, control, etc. are still relevant!

We want a camera to just be a camera

The main question: when are models the “same”?

When are two AI models “effectively the same”?
We need ways to **compare models**:

- SGD means models trained on the same data with the same architecture will be different.
- Internals of different models (e.g. GPT vs. Llama) use different internal features/ embeddings.

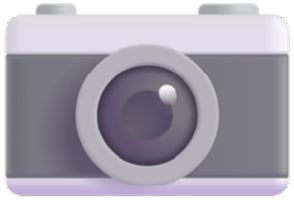
Question: How can basic techniques in signal processing and communications help?

iOS 8.3

iOS 18.4

HarmonyOS 4.0

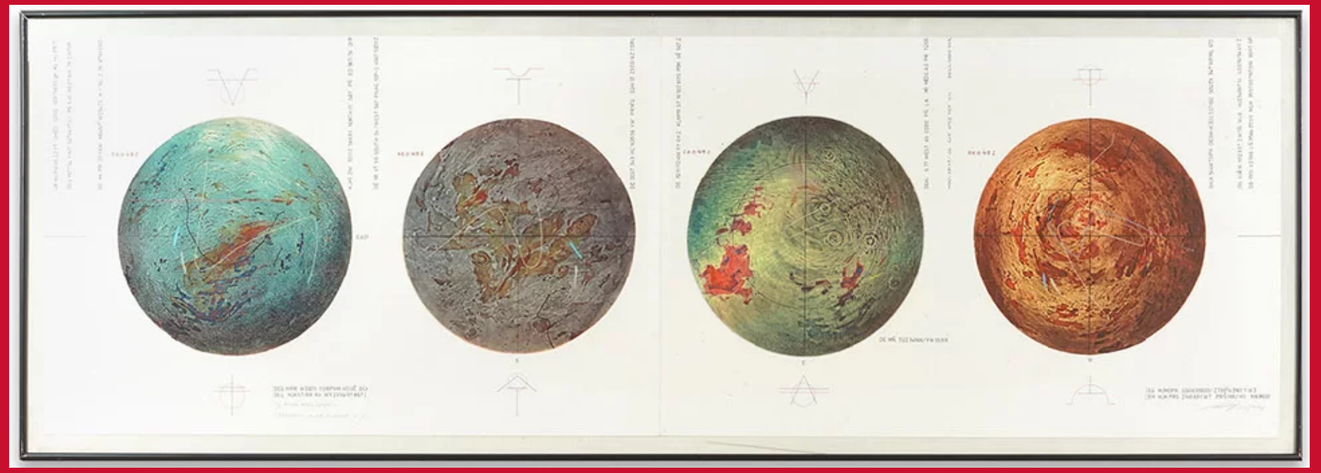
Samsung UI 7.0



MS 3D Fluent

SerenityOS

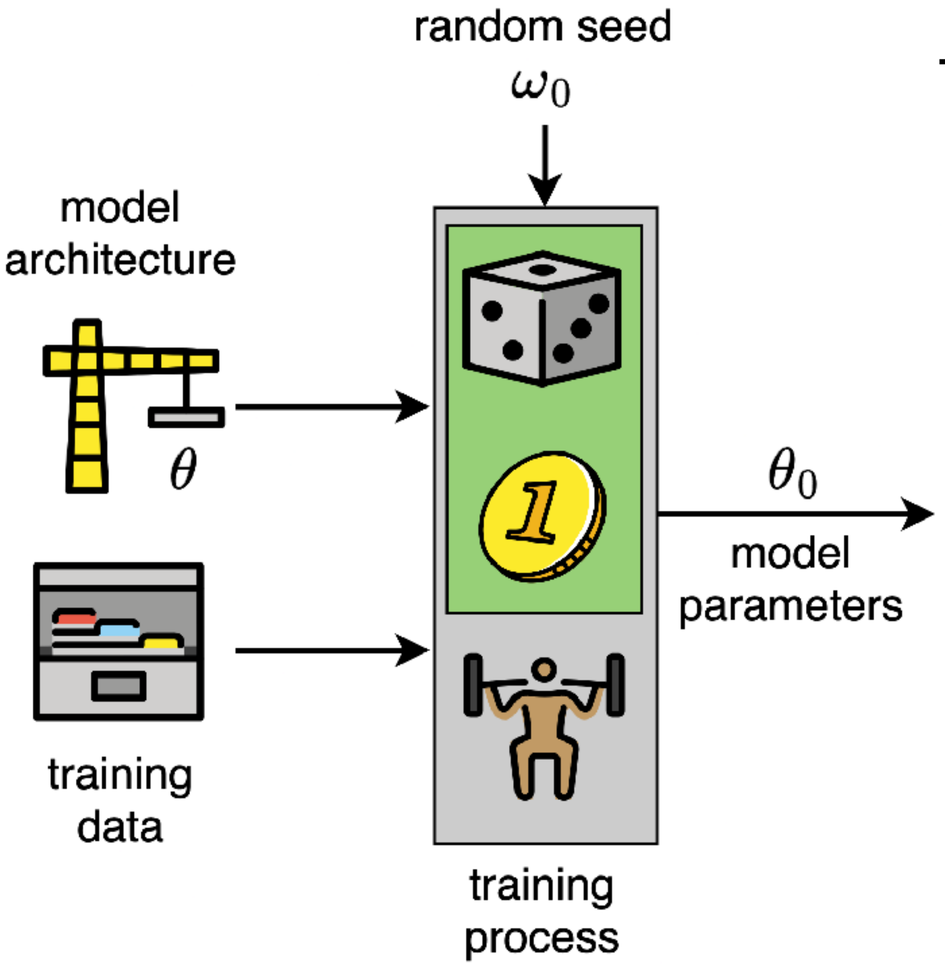
How should we understand model comparisons?



Rm Palaniappan, *Alien Planet-A*
Viscosity, pencil colour and ink on handmade paper

The standard statistical setup for modern ML

Machine learning as function-fitting



Traditional setup for estimating parameters/weights:

- Parameterized **set of functions/models** $\mathcal{F} = \{f_\theta(x) : \theta \in \Theta\}$.
- **Training data** used to estimate the parameters by **minimizing a loss function**.
- **Stochastic optimization** algorithm that does the actual minimization (e.g. stochastic gradient descent = SGD).

How should we characterize a model?

Drawing samples from the function space

For a **fixed training set**, **architecture**, and **training algorithm**, we can think of an ML/AI model as a **random sample from the function space**:

$$\mathscr{F} = \{f : f \text{ representable by the NN}\}$$

Examples:

- In classification, each $f : \mathscr{X} \to [L]$ labels input data points.
- In representation learning, each $f : \mathscr{X} \to \mathscr{R}$ maps inputs to representations/embeddings.

Some natural questions

Comparing models is not clear

If we have two different models we might have

$$\mathscr{F} = \{f : f \text{ representable by NN A}\}$$
$$\mathscr{G} = \{g : g \text{ representable by NN B}\}$$

Can we meaningfully compare these models?

- If $\mathscr{F} = \mathscr{G}$ we can use their outputs to do a comparison.
- If $\mathscr{F} \neq \mathscr{G}$ we need some way to do a comparison.

Can we use GenAI for generating data?

Are GenAI models “interchangeable”?

HarmonyOS 4.0

Samsung UI 7.0

Suppose we all have **AI scientific instruments** using GenAI. *Example:* two MRI machines using GenAI for image formation?

Can you collaborate with a lab which uses **a different model for the same task**?

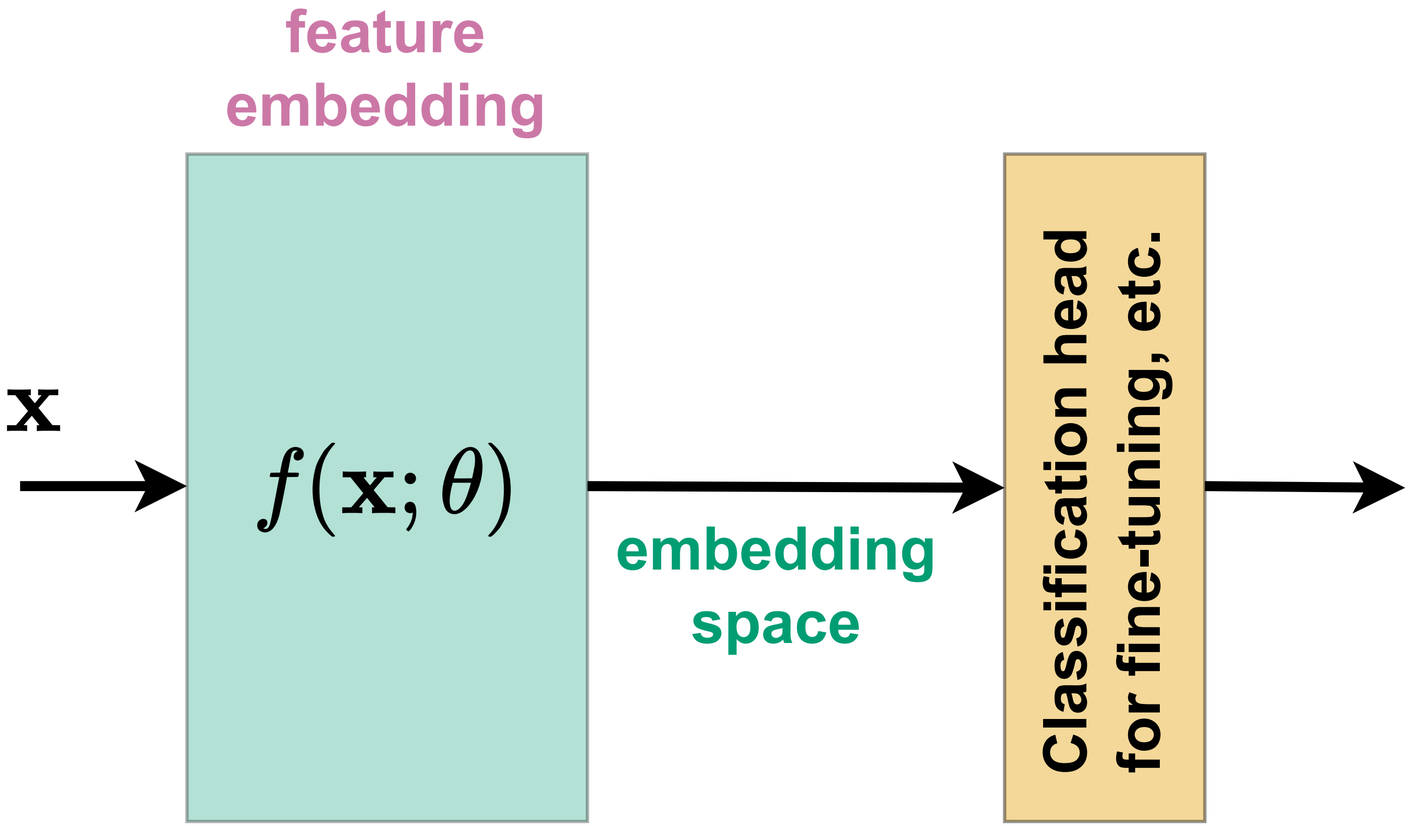
Are these models producing outputs that **“look the same?”**

Challenges:

- To the human eye, they are functionally similar.
- Can we quantitatively see if they are different?

Embedding spaces of large models

Splitting a model into a feature extraction and decision



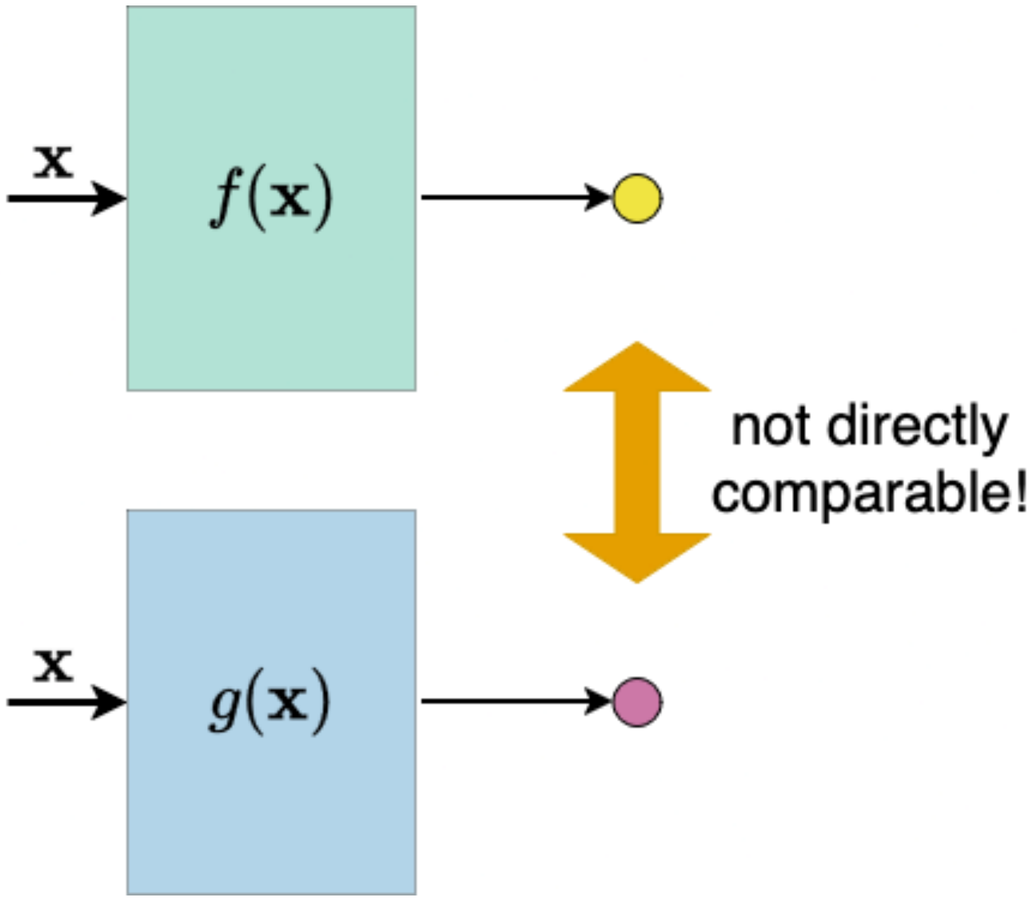
We can think of many models as having “feature embedding” stage followed by “downstream tasks.”

Fine-tuning works because the embeddings carry “a lot of information.”

Idea: can we compare models using their embedding spaces?

Comparing embedding spaces directly?

Generally this is a non-starter



Given two models with different architectures, how can we compare embedding spaces directly?

- Different **dimensions**.
- Different **compression strategies**
- Different **“semantics”**

Unlike with classification, we need to compare the outputs of the generative models.

Comparing two different models

In general, they may have different architectures

HarmonyOS 4.0

Samsung UI 7.0

Are these sets of functions “equivalent” in some way?

$$\mathcal{F} = \{f(x\,|\,\theta) : \theta \in \Theta\} \quad \text{vs.} \quad \mathcal{G} = \{g(x\,|\,\theta) : \theta \in \Theta\}$$

1. Focus on **performance**: two models with the same error are “effectively the same”.
2. Focus on **features**: come up with a mapping from one model to the other to show they are the same.
3. Focus on **approximations**: use proxies for each model which are more comparable.

Understanding generative AI output with embedding models

Rm Palaniappan, *Alien Planet-B*
Viscosity, pencil colour and ink on handmade paper

Comparing models using another model

Don’t compare embedding spaces directly: use the outputs

Our approach: use two (or more) GenAI models to generate outputs based on the same inputs/prompts.

Use a **different model to embed** the outputs. Now we can compare them!

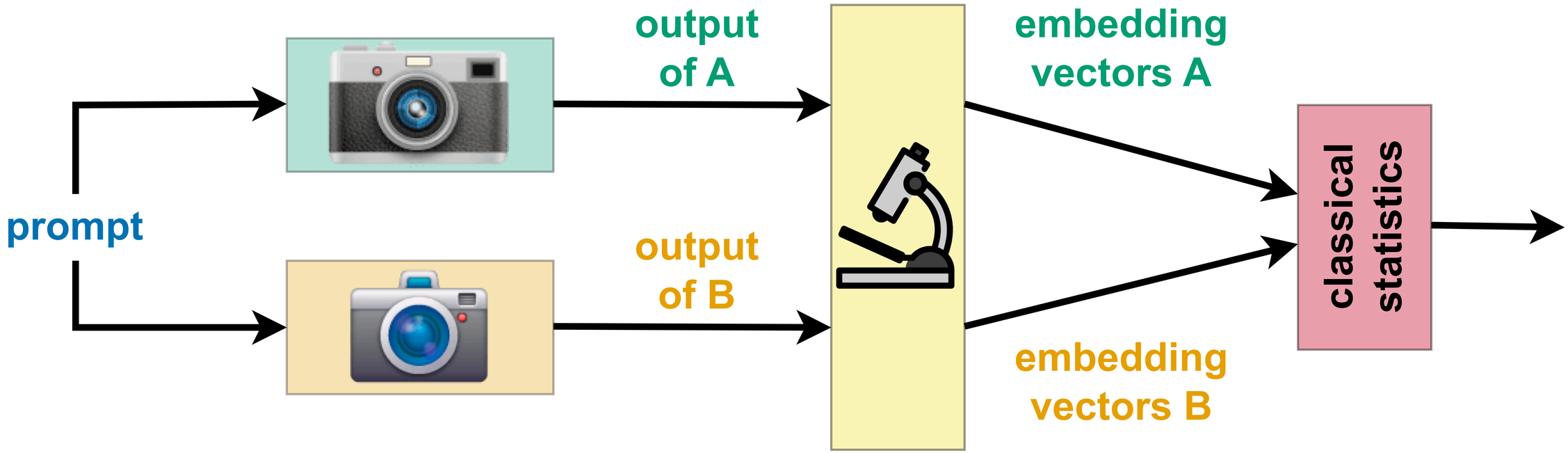
1. **Do different data sets have different/separable embeddings?**
   Example: does a vision model “see” differences between data sources?

2. **Are these separations human-interpretable?**
   Example: does text translation reduce variations?

3. **Do different GenAIs generate distinct-looking outputs?**
   Example: can an LLM separate other LLMs vs. real data?

A (relatively) cheap experiment

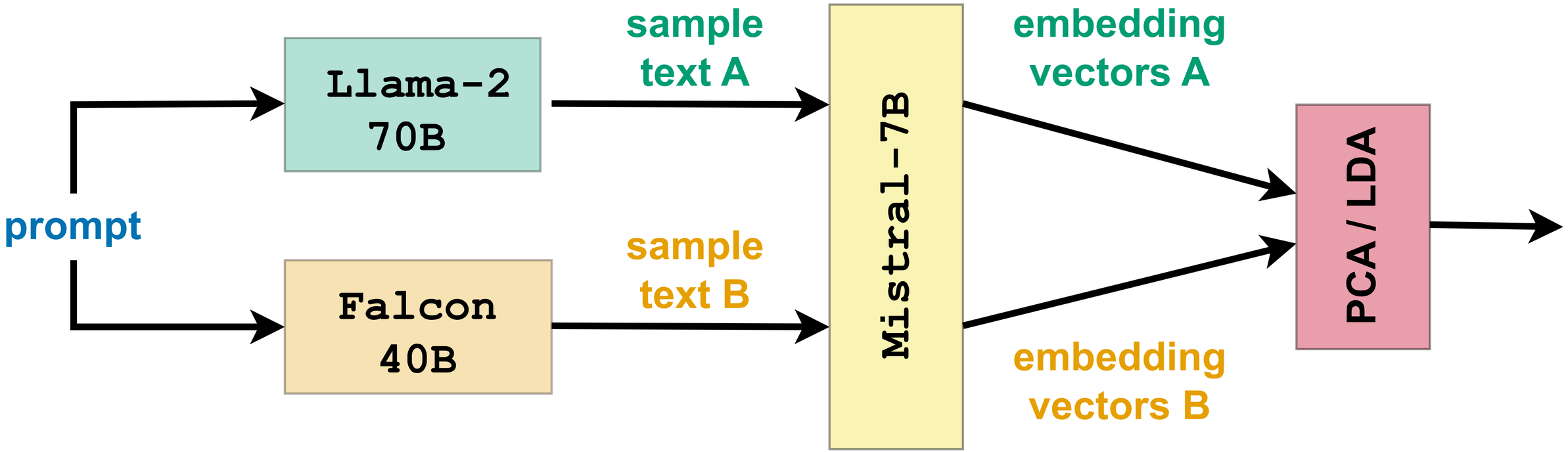
We work with pretrained models (open weights)



The **third AI model** acts as a “microscope” to compare the outputs of two different AI models.

A specific example for GenAI

Compare the outputs using a 3rd model for embedding



Using a large model as a “microscope”

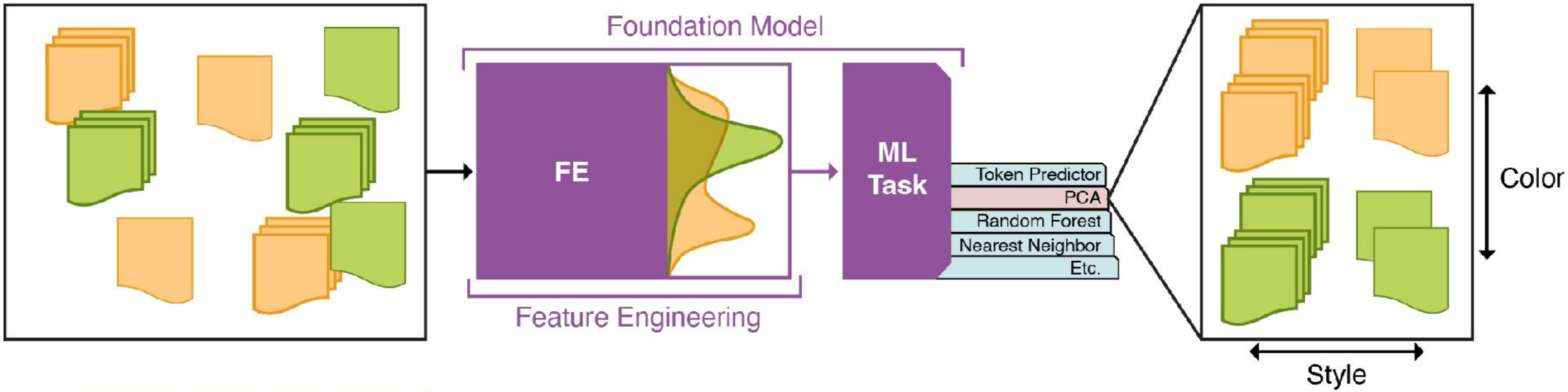
It takes one to know one

Idea: Use a large model to embed the outputs of the models we want to compare.

- **`Mistral-7B`**: LLM, transformer-based, 32 layers, 13b parameters per token and 32 token vocabulary. Embeddings from the final hidden layer of dimension 4,096.
- **`Multilingual-e5-large`**: extracts sentence embeddings from text in different languages to 1024-dimensional embedding vectors. 60M parameters, context window of 512 tokens and long text is truncated to fit within this window.
- **`Data Filtering Network`**: a CLIP model trained on 5B images that were filtered from an uncurated dataset of image-text pairs. It has 1B parameters and can be used to encode both text and images.

The generic approach

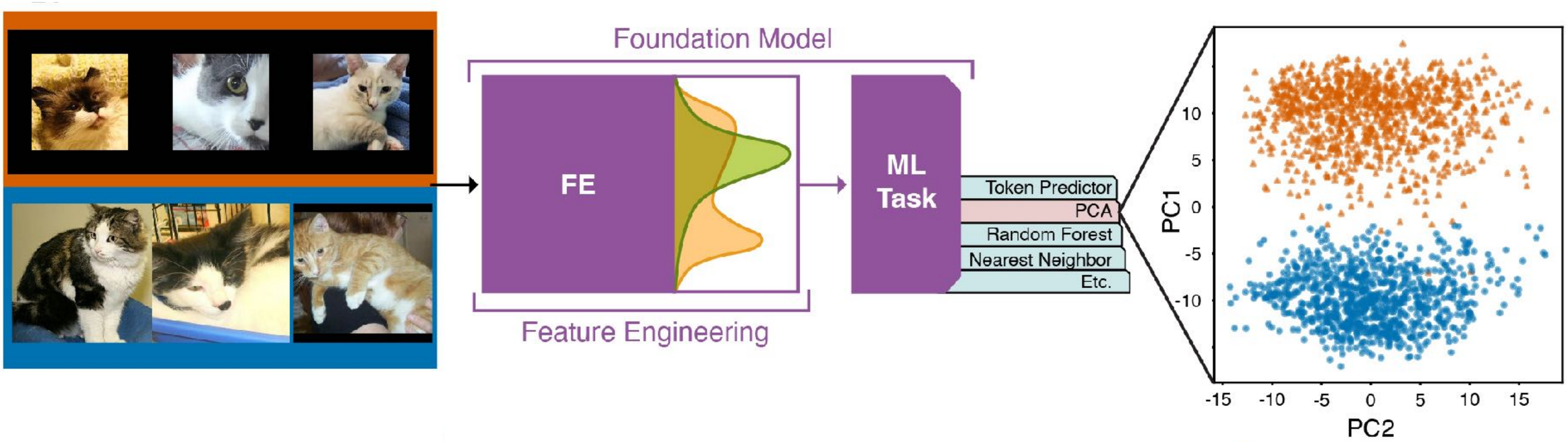
Using embeddings/features to reveal known differences



Use simple methods like PCA or LDA to analyze the variability between the embeddings from different classes. The principal directions are often interpretable.

Comparing different data sets

Embedding model can easily separate different sources



Data from different sources (image data sets) are easily separable in the embedding space using PCA.

Supervised learning using LDA

Different generative models and real data are distinguishable

Experiment	Classes	Avg. Test Acc. (%)	SD (%)
Cat Images	LSUN, All Models	98.8	0.38
	LSUN, DDPM, SDXL, Open-Dalle	97.5	0.67
	Cats & Dogs, LSUN, DDPM, SDXL, Open-Dalle	98.0	0.37
	LSUN, DDPM	92.7	1.42
	SDXL, Open-Dalle	90.1	1.63
	(LSUN, DDPM) vs (SDXL, Open-Dalle)	100	0.0
GenImage	ImageNet vs All 8 Models	98.3	0.07

Table S4: **LDA accuracies across image experiments.** LDA accuraces across 50 train/test splits for different subsets of data in each experiment. Standard deviation (SD) is indicated in the final column. For creating embeddings, we use the default (bolded) embedding model corresponding to the respective experiment listed in Table S1.

Pairwise comparisons for genAI data

Data generated from different models are distinguishable using LDA

	ADM	BigGAN	Glide	Midjourney	SDV4	SDV5	VQDM	Wukong
ADM	–							
BigGAN	100	–						
Glide	100	100	–					
Midjourney	100	100	98.66	–				
SDV4	100	100	98.17	97.28	–			
SDV5	100	100	97.97	96.02	94.85	–		
VQDM	100	100	98.26	99.00	98.88	98.61	–	
Wukong	100	100	99.25	98.78	97.60	96.55	96.07	–
ImageNet	100	100	> 99.99	99.95	99.99	99.88	99.86	99.75

Maybe this isn’t surprising…

Visually the images have different “styles”

DDPM

OpenDalle

SDXL

- Results seem to suggest that it’s “easy” to distinguish
  - real images from genAI images
  - between different genAIs
- We are using latent discriminant analysis (LDA) to train the detector. **Isn’t that cheating?**
- Are the principal component (PCs) more directly interpretable?

Applied to text translation

Embed articles in two languages and also translated articles

El grupo italiano Enel dio ayer el paso definitivo para adquirir los activos latinoamericanos de Endesa...

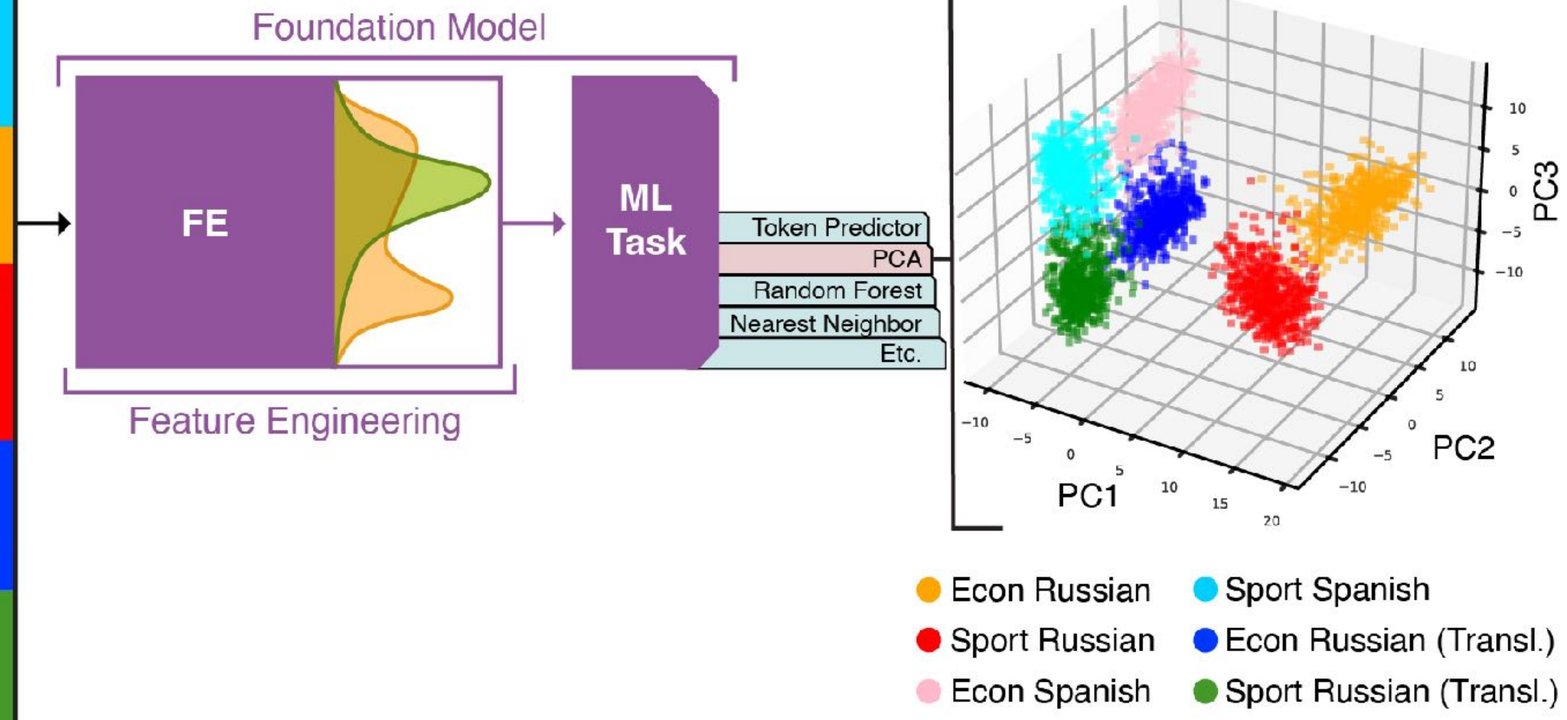
El Celta más irreconocible en mucho tiempo se fundió en Mendizorroza en un instante decisivo que viró en lágrimas...

Как известно, все реформы в РФ затеваются ради одной цели: “распила” бюджетных денег...

Победоносный для испанцев матч в Аликанте обслуживала бригада российских арбитров под руководством Алексея Еськова...

Como se sabe, todas las reformas en la Federación se inician con un solo objetivo: gastar dinero presupuestario...

El partido ganador para los españoles en Alicante fue atendido por una brigada de árbitros rusos bajo la dirección de Alexei Yeskov...

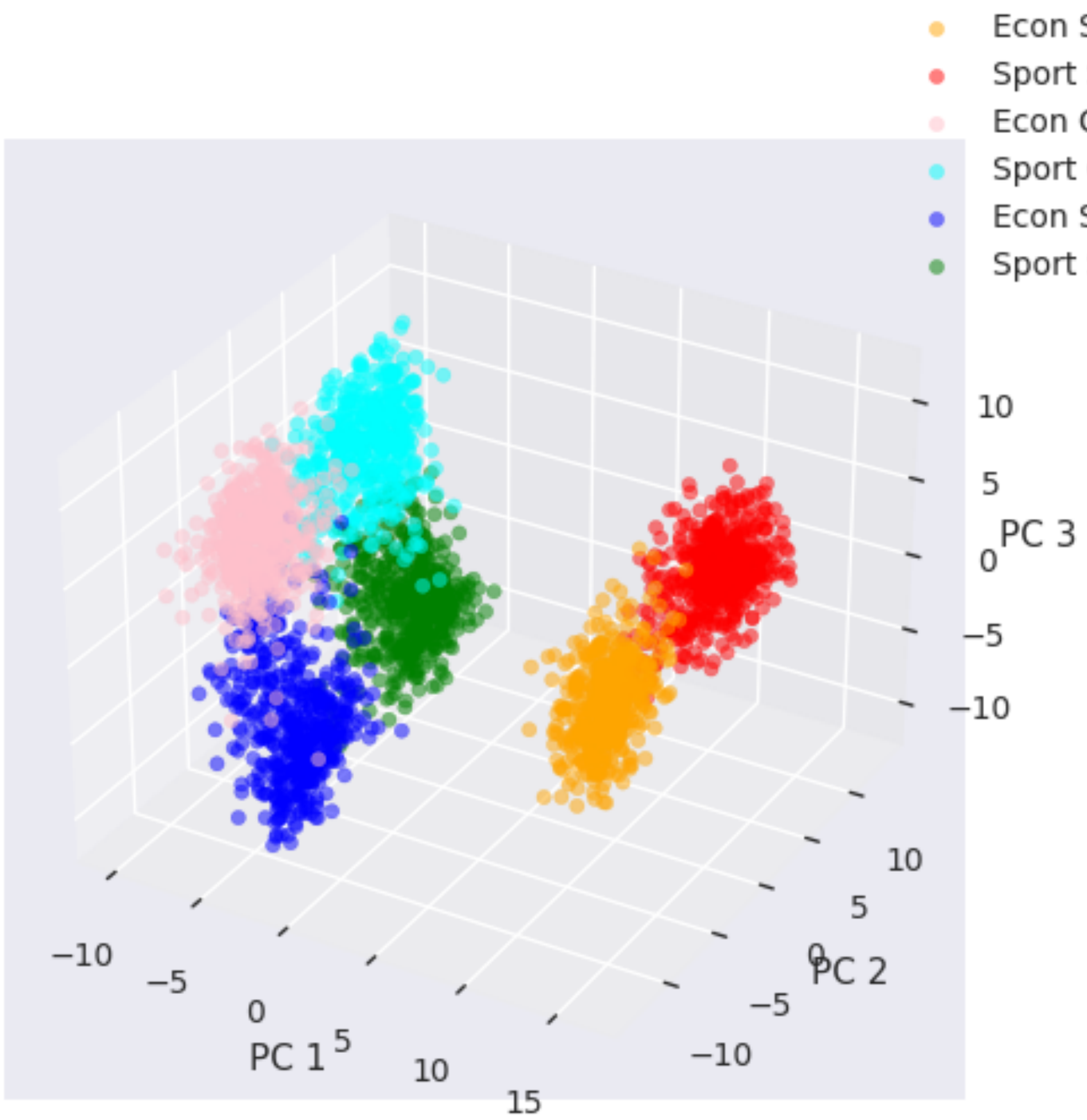


An experiment in translation

Interpreting principal components

The experiment:

1. Take news articles with **two different topics** and **two different languages**.
2. Use machine translation to “remove” the language difference.
3. Interpret the principal components:
    - PC1: language
    - PC2: subject
    - PC3: translated or original



Claim: PCs reflect interpretable features/known hidden labels.

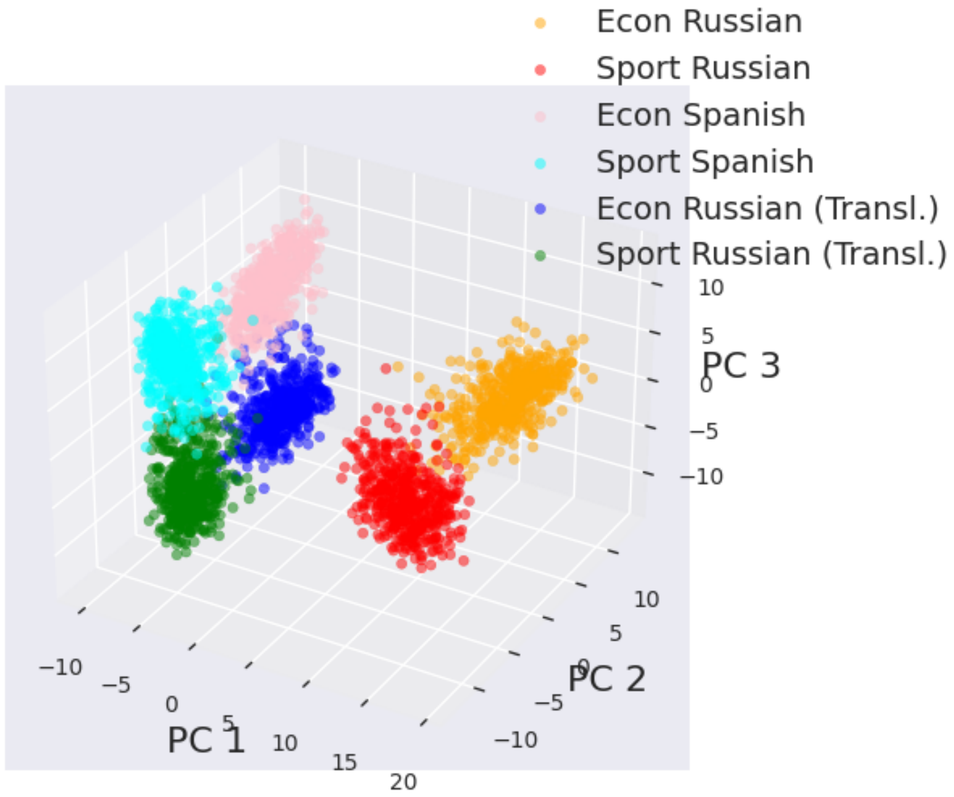
Used a ML translator to translate German to Spanish.

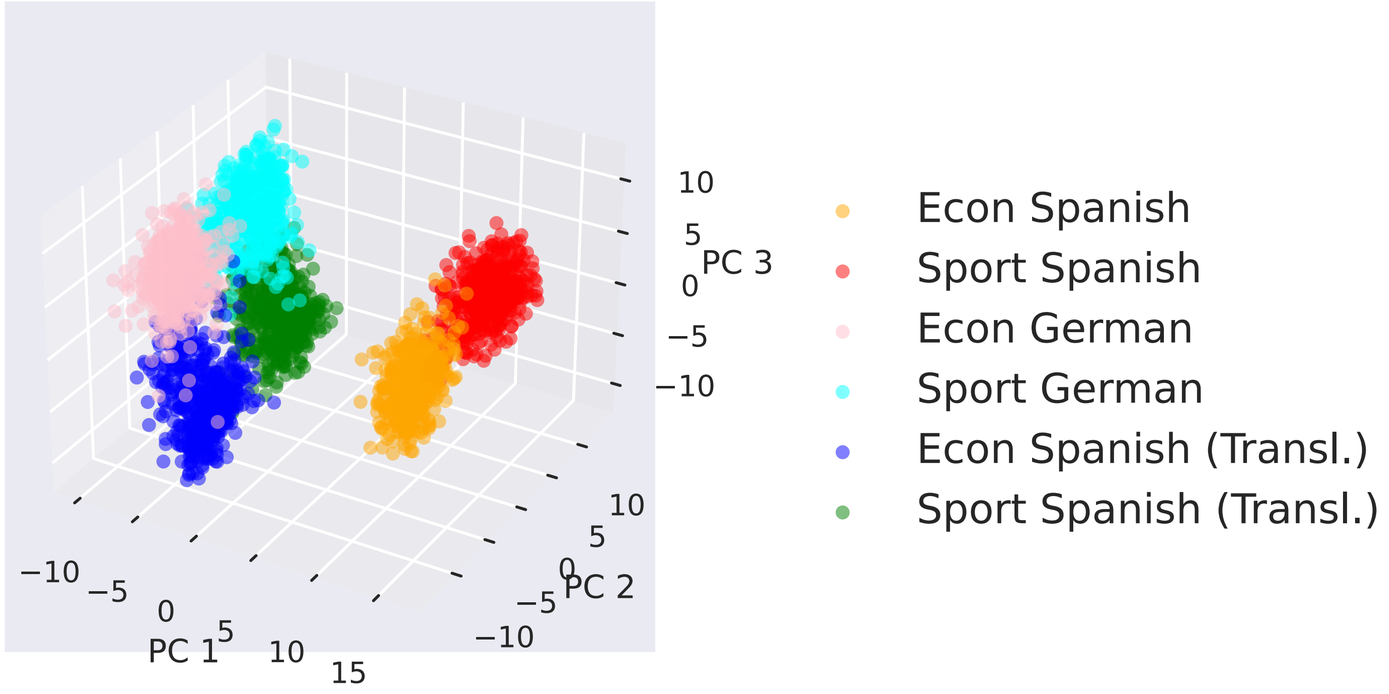
Translating news articles helps reduce the variation in one dimension (language).

But machine translation is also a visitble artifact (PC3)!

Replicable across other language pairs

It’s more than just special characters





Text generation using an LLM

Randomly sample (pre 2020) arXiv abstracts

1. Selected abstracts from

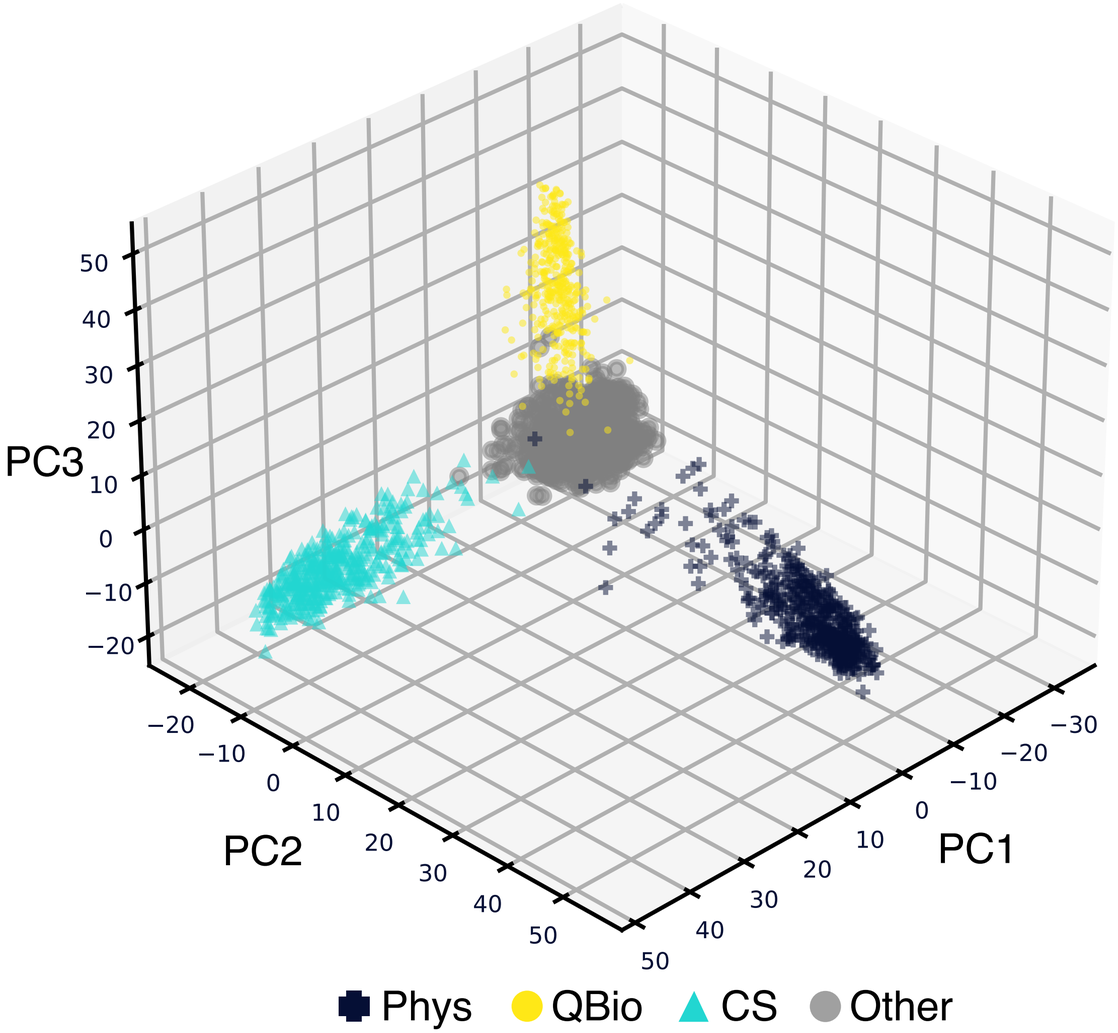
   high-energy particle physics (hep-ex),
   programming languages (cs.PL),
   quantitative biology—cell behavior (q-bio.CB),
   statistical methodology (stat.ME),
   quantitative finance—portfolio management (q-fin.PM).

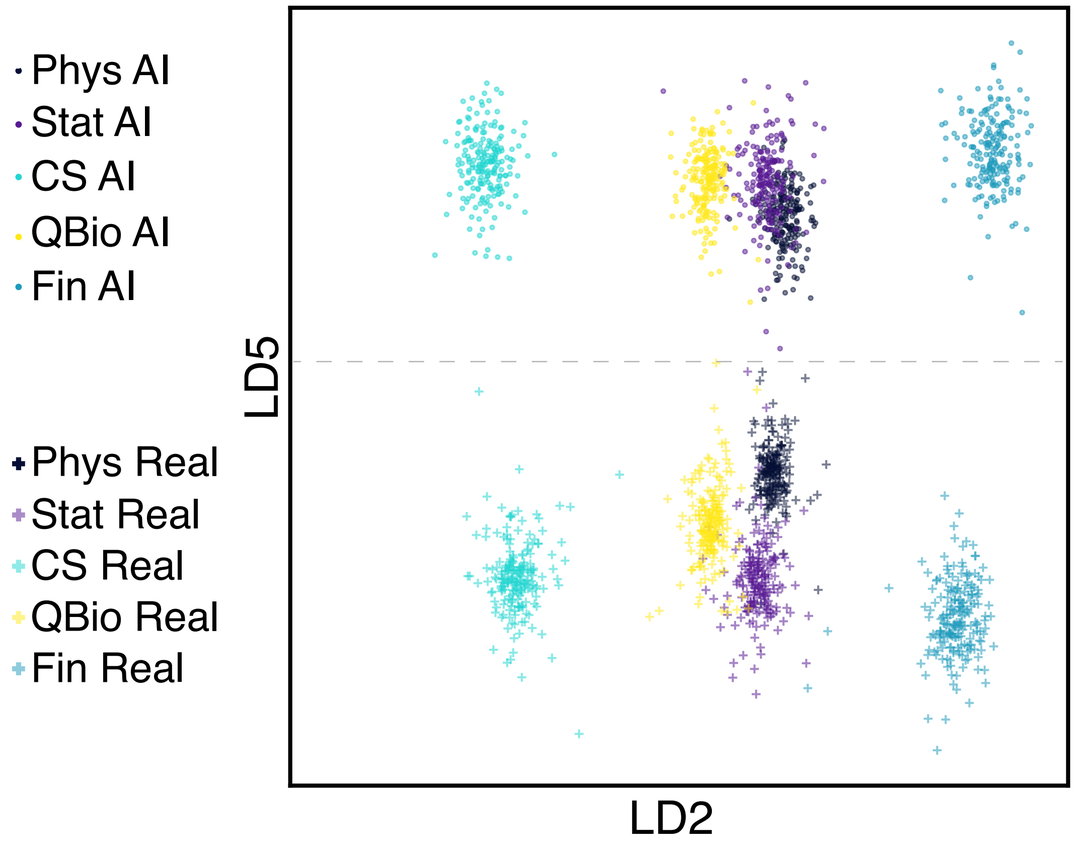
2. Used **Llama-2 70B** to generate, for each real paper, a synthetic abstract for a paper with the same title.

3. Use **Mistral-7B** for the embeddings (4096 dimensional).

LDA and PCA both show separation

Embeddings cluster by topic



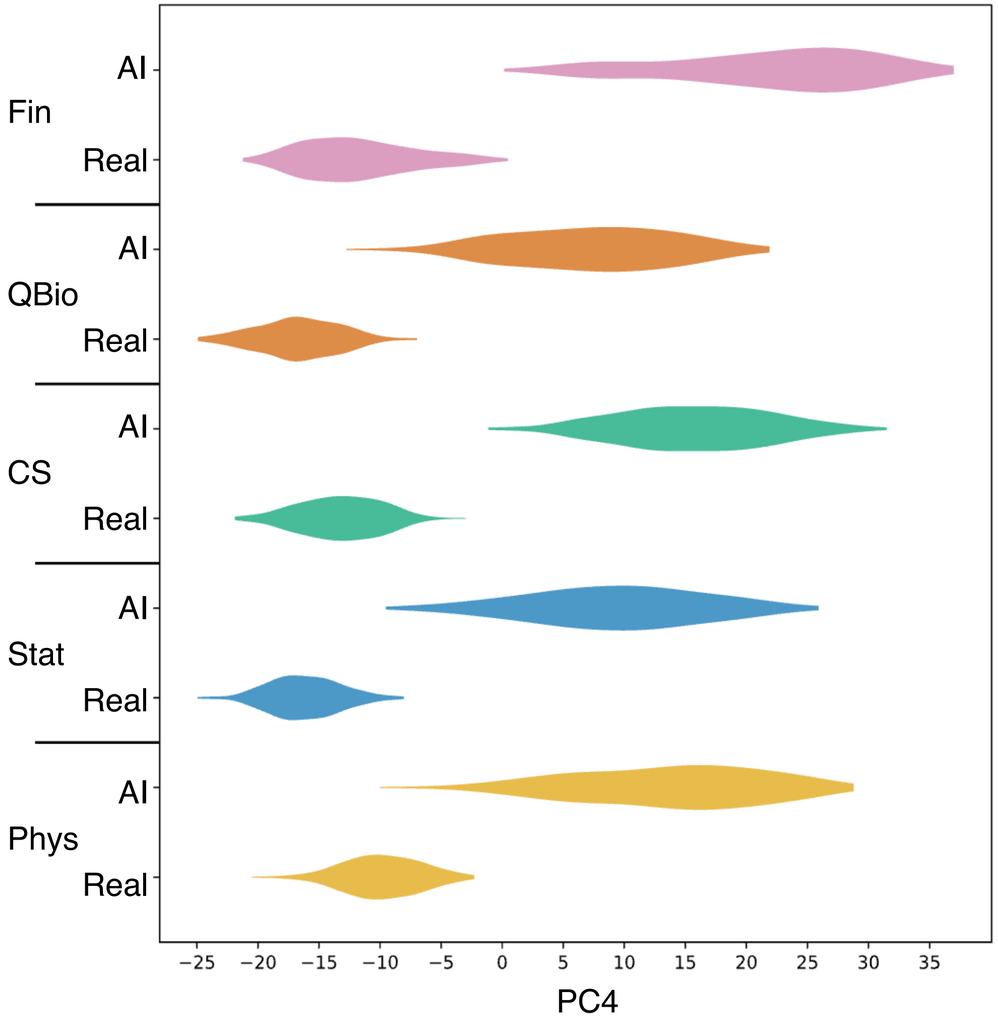


Test accuracy (10 classes): 99%

Strong differences between AI and real data

The 4th PC shows that AI-generated abstracts “look fake”

- AI-generated abstracts overclaim: “important”, “significant” and “valuable” are much more frequent.
- Consistent with evidence that LLMs often “flatter” queriers/users and “exaggerate” claims.
- While LDA uses labels, the PCA does not: these differences are solely from the embedding differences.



... The paper's findings have **important** implications for portfolio managers and investors who need to optimize their portfolios in the presence of assets with uncertain liquidation times. The proposed approach can help them to better manage risk and maximize returns in their portfolios. Overall, this research paper makes a **significant** contribution to the field of portfolio management by proposing a novel approach to portfolio optimization that ...

During chemotaxis, fibroblasts exhibit directional decision-making, which is critical for tissue repair and regeneration... These findings have **important** implications for our understanding of cell behavior and may lead to the development of new strategies for controlling cell migration in tissue engineering and regenerative medicine.

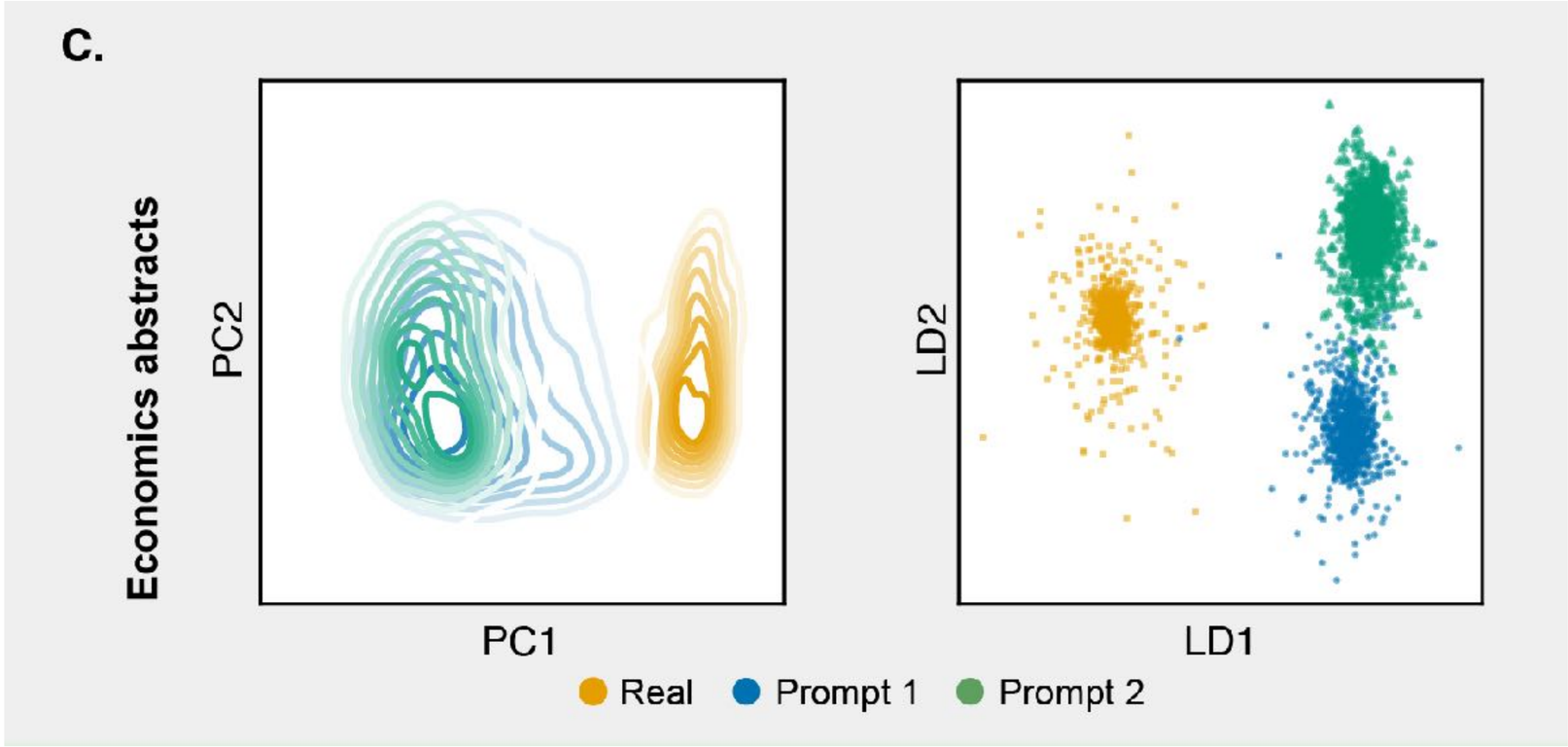
...approach **significant**ly improves upon existing verification techniques in terms of automation, consistency, and reusability, making it a **valuable** tool for programming language researchers and practitioners... Our work has **important** implications ...

... Our method has **important** implications for a wide range of applications, including data visualization, anomaly detection, and clustering... Overall, this paper makes a **significant** contribution to the field of statistical methodology by introducing a powerful and flexible nonparametric method for estimating directional HDRs. The method is easy to implement and can be applied to a wide range of data types, making it a **valuable** tool for researchers and practitioners in many fields.

The topological branching fractions of the tau lepton have been measured using data collected at the Large Electron-Positron Collider (LEP) at CERN... The results of this study provide **important** constraints on the properties of the tau lepton and have implications for the search for new physics at future high-energy colliders.

Differences in prompts are detectable

Prompt engineering may lead to prompt artifacts…



- Here, PCA cannot readily detect prompt differences but LDA can.
- Experiment shows that there is “signal” there to detect.
- Detection in an unsupervised way may be more challenging.

Some takeaways and ongoing work

Where can we go from here (requires more research)

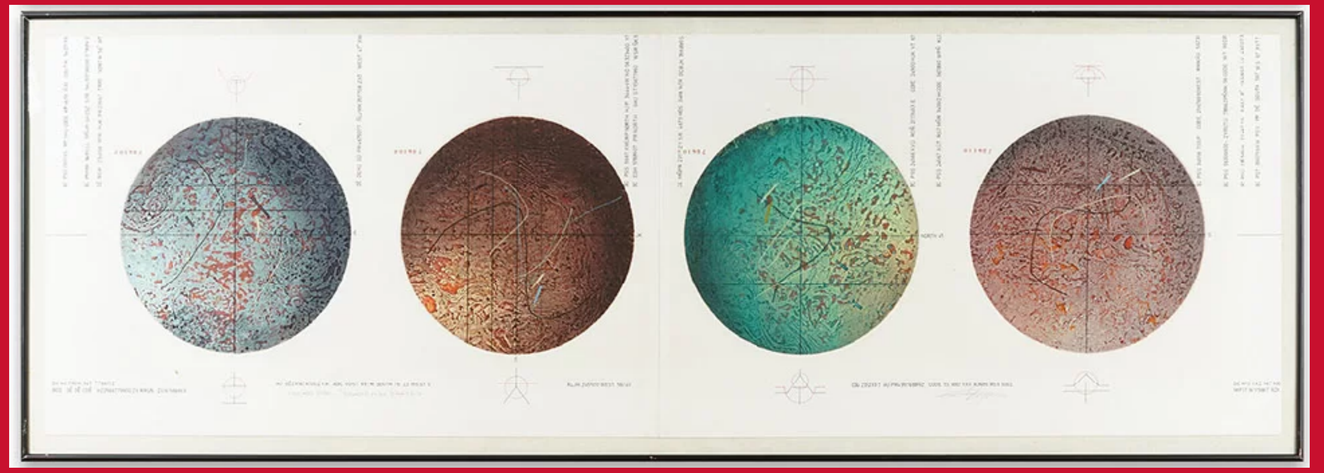
HarmonyOS 4.0

Samsung UI 7.0

Preliminary experiments show that the embedding spaces of large “foundation models” can separate data generated from different sources.

- Forensics applications: comparing models, detecting deepfakes, etc.
- “Model DNA”: fine-tuned or “lightly modified” models make minor modifications to the embeddings.
- Use post processing to “align” embeddings for calibration, ensembling, federated learning, etc.

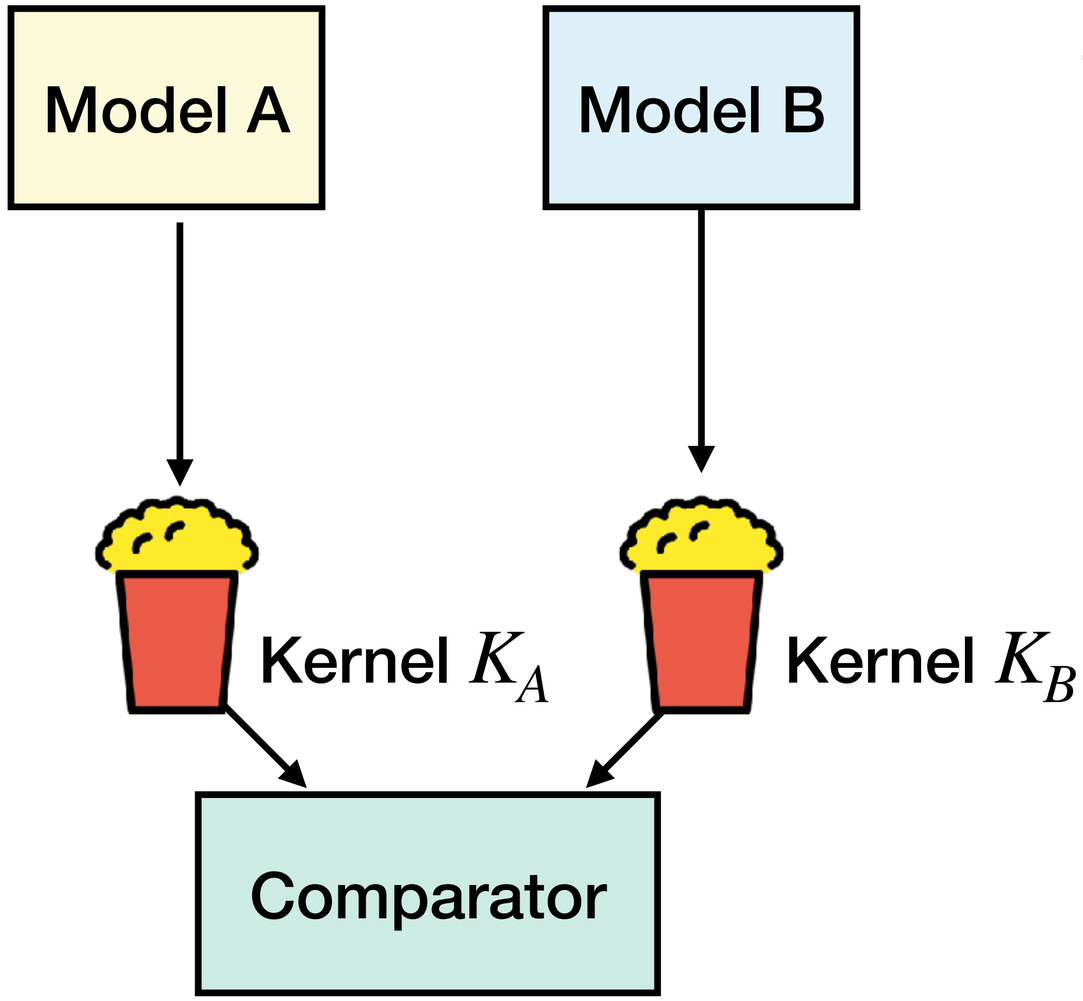
The wider world of embeddings



Rm Palaniappan, *Alien Planet-D*
Viscosity, pencil colour and ink on handmade paper

How else might we compare embeddings?

Lots of different good ideas out there!

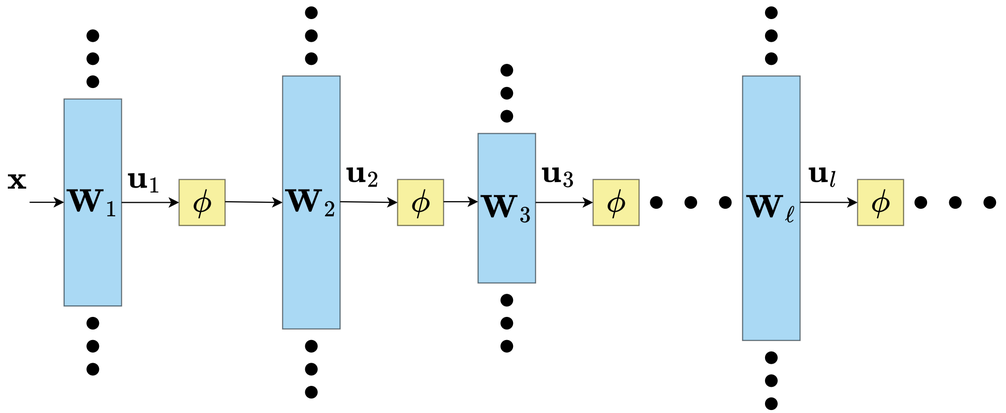


There are many other ways to approach the problem of model comparisons:

- For discriminative models, focus has often been on which model does better on benchmarks.
  Models can be very different but have similar error rates!
- Significant work on using *kernel models* to approximate DNNs: can then compare the kernels for model comparison.
  Models can be very different.

Approximating the NN with a kernel machine

Not practical, but perhaps informative?



$\approx$

kGLM

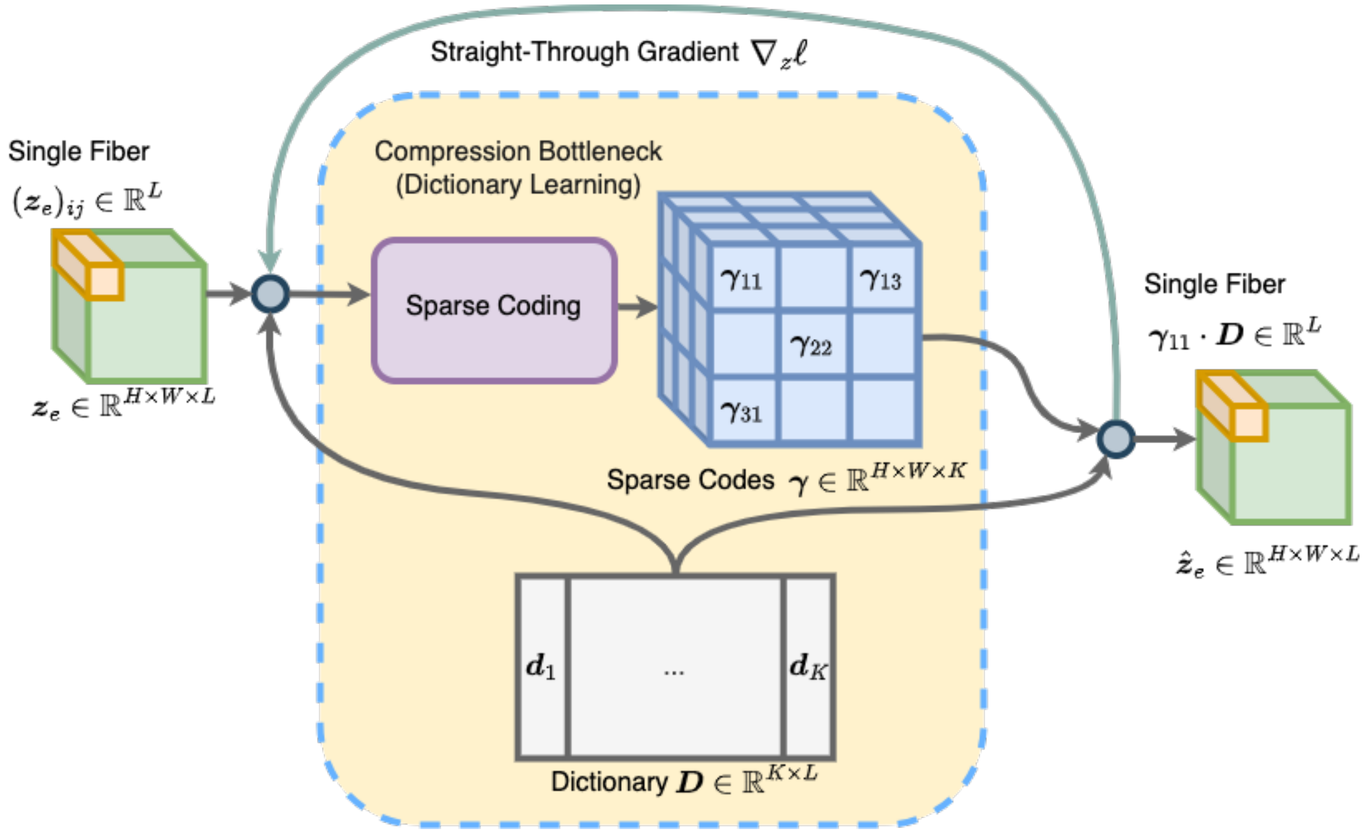
Suppose we compute some kernel function $\mathbf{K}$ associated to the model and fit a **surrogate model**.

One example: the **neural tangent kernel** (Jacot et al. 2018) can be used for comparisons and explainability (Engel et al. 2024).

Another approach is to look at the differences between kernels constructed from a reference set of samples (Jalali et al. 2025).

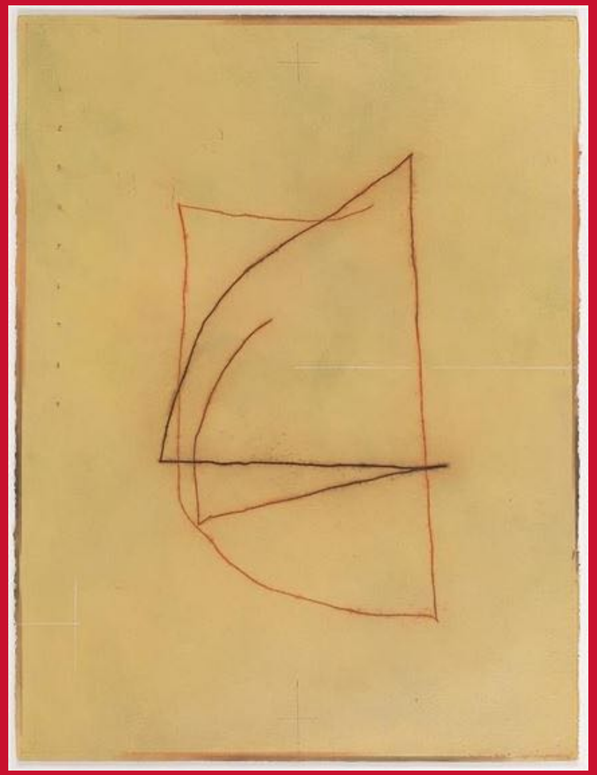
Ongoing work: structuring embedding spaces

What does the space of embeddings look like?



- Latent spaces for generative models are often vector quantized to create tokens used in transformers.
- Quantization is only one way to compress the latent space.
- Can we impose a different structure on latent/ embeddings during training? (Li and S. 2025).

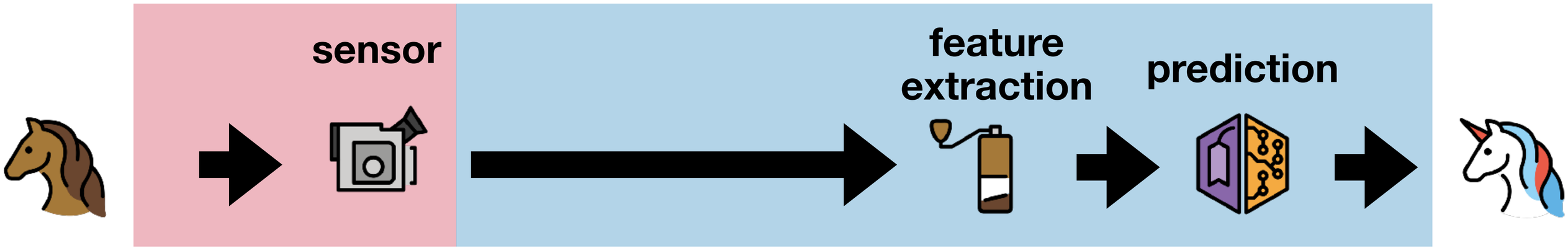
Some final remarks



Rm Palaniappan, *Intense Talk*
Mixed media on paper pasted on mount board

Back to the original question

What does any of this mean for “AI for Science”?



ML/AI models are not nearly as reliable, consistent, and interpretable as measurement devices: much more is needed before we throw away “raw” measurements.

We need more ways to **compare models directly** instead of only focusing on **performance on benchmarks**.

Where is this all going?

Maybe some strange new worlds

Developing a good set of techniques for model comparisons requires thinking from several different directions:

- **Theory:** can we instead compare surrogate models like “faithful” NTK representations (Engel et al. 2024)?
- **Experiment:** can we do these comparisons cheaply (e.g. using academic-level resources)?
- **Application:** how do we use model comparisons in forensics, process engineering, ensembling, and beyond?

Thank you!