Introduction to machine learning

Thomas Stoeger, Northwestern University

Slides:

https://bit.ly/3bKMrhM

Slides: https://bit.ly/3bKMrhM

demystify

demystify

orientation

demystify

orientation

coding

demystify → understand conceptual possibilities

orientation

coding

demystify → understand conceptual possibilities

orientation → understand main steps

coding

demystify → understand conceptual possibilities

orientation → understand main steps

coding → have a well-extendable example

Slides: https://bit.ly/3bKMrhM

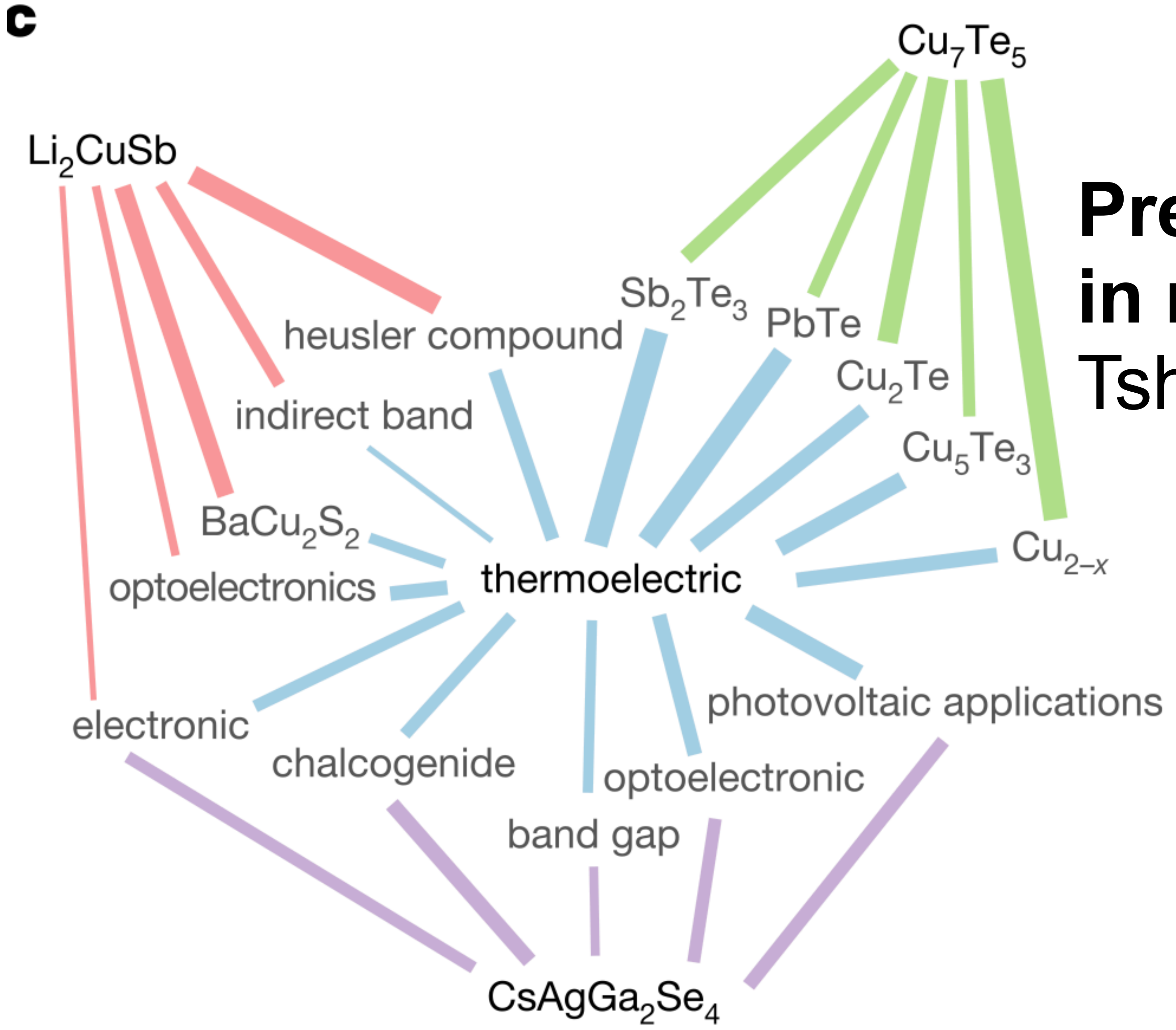
demystify → understand conceptual possibilities

orientation → understand main steps

coding → have a well-extendable example

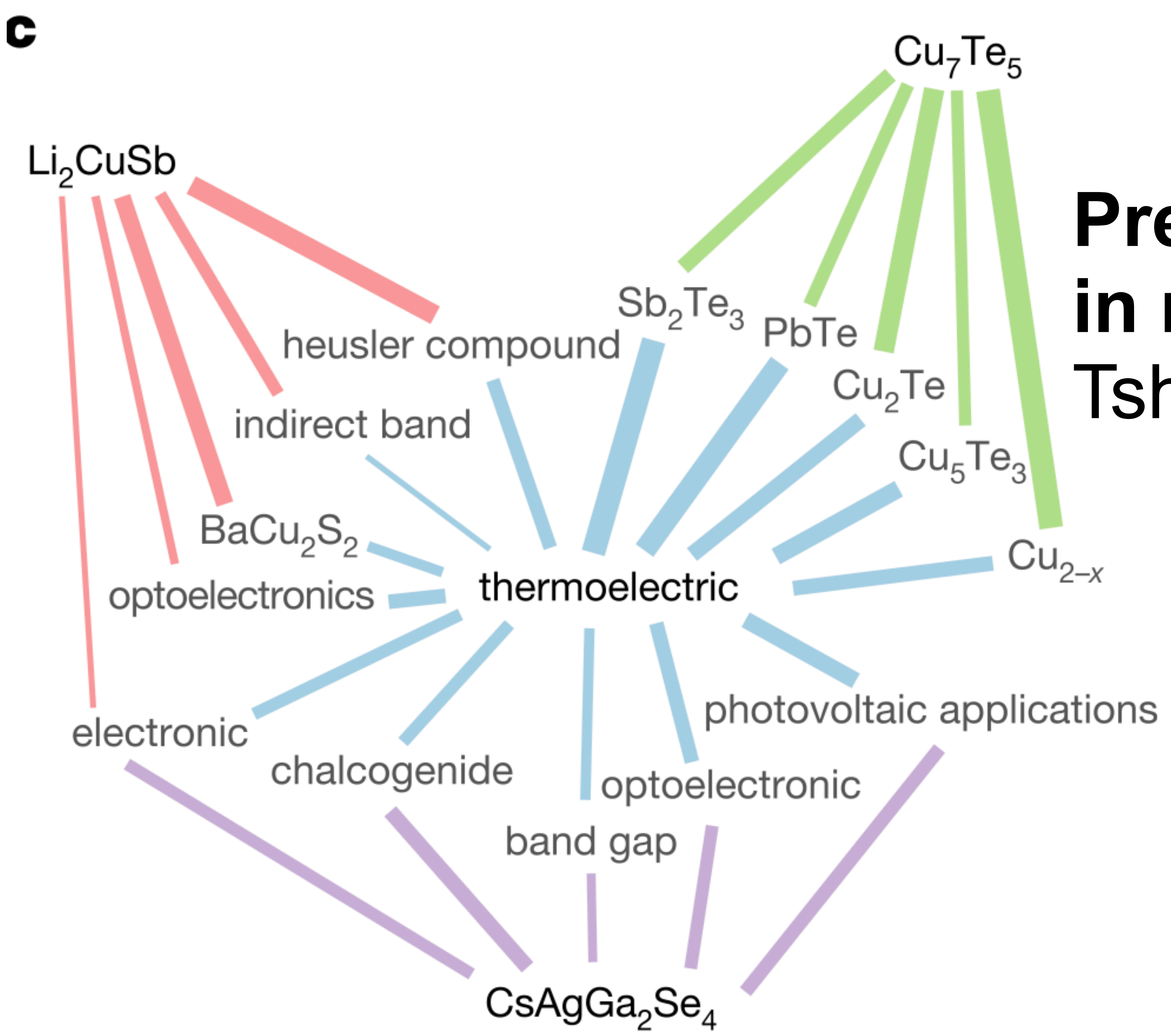
Slides: https://bit.ly/3bKMrhM

c

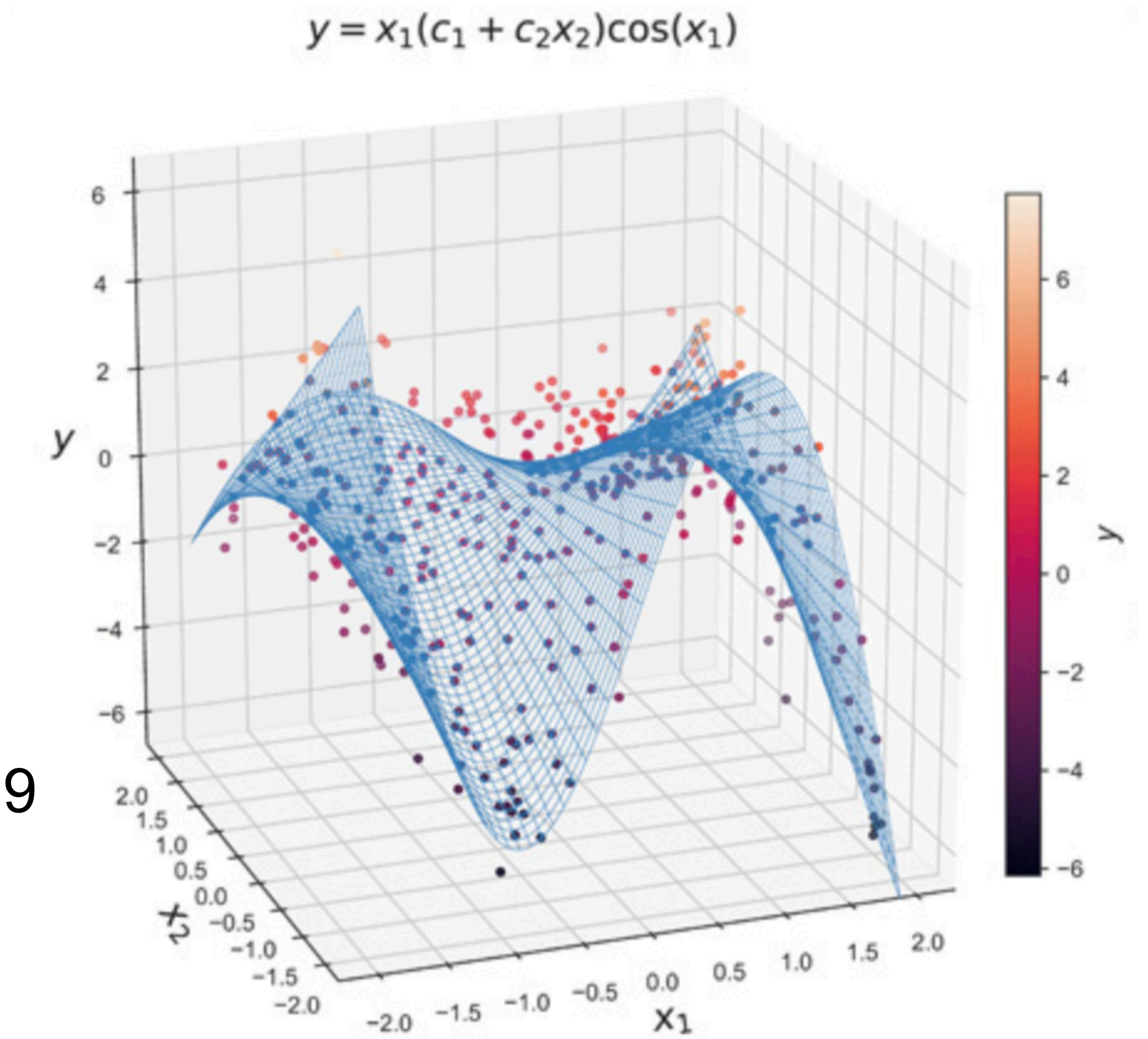


Predict future discoveries in material science

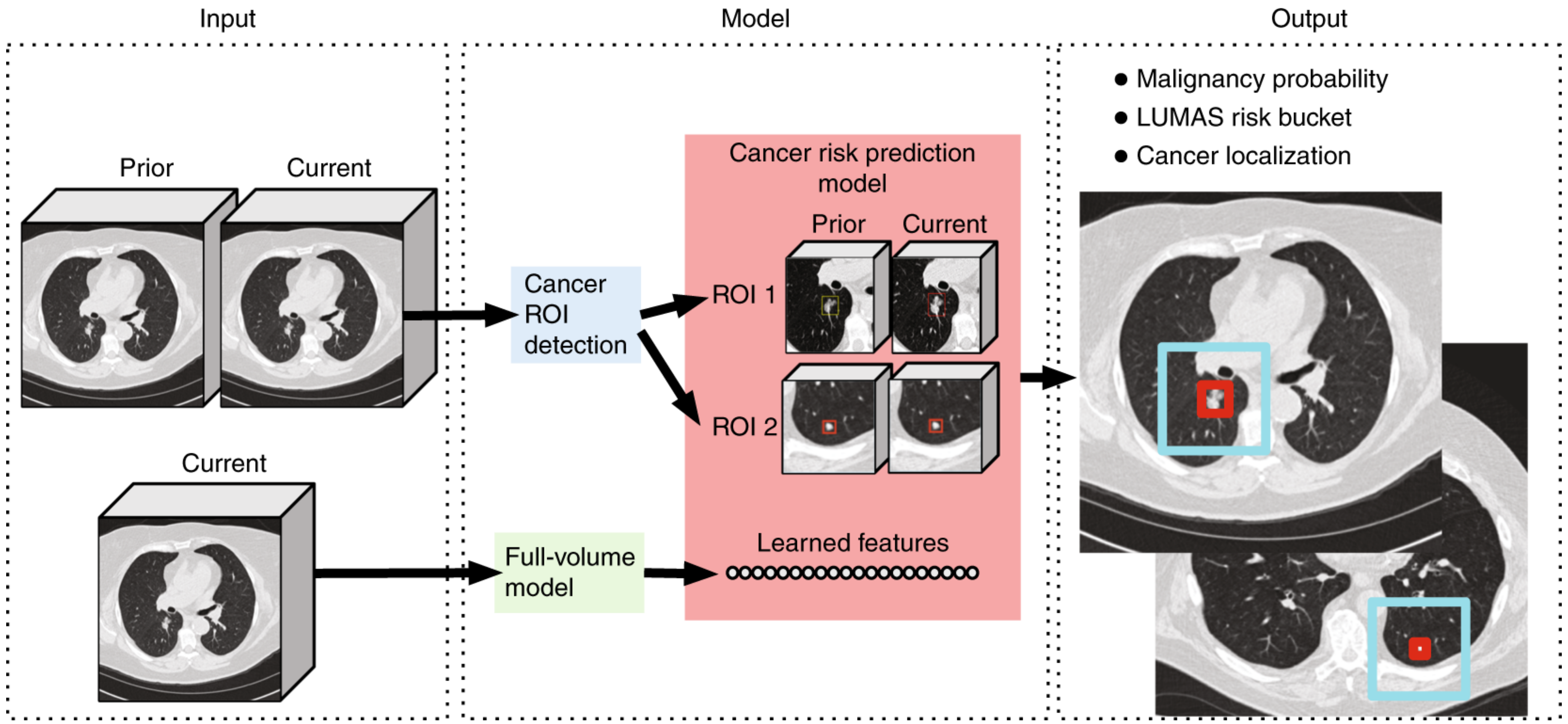
Tshitoyan et al., Nature, 2019



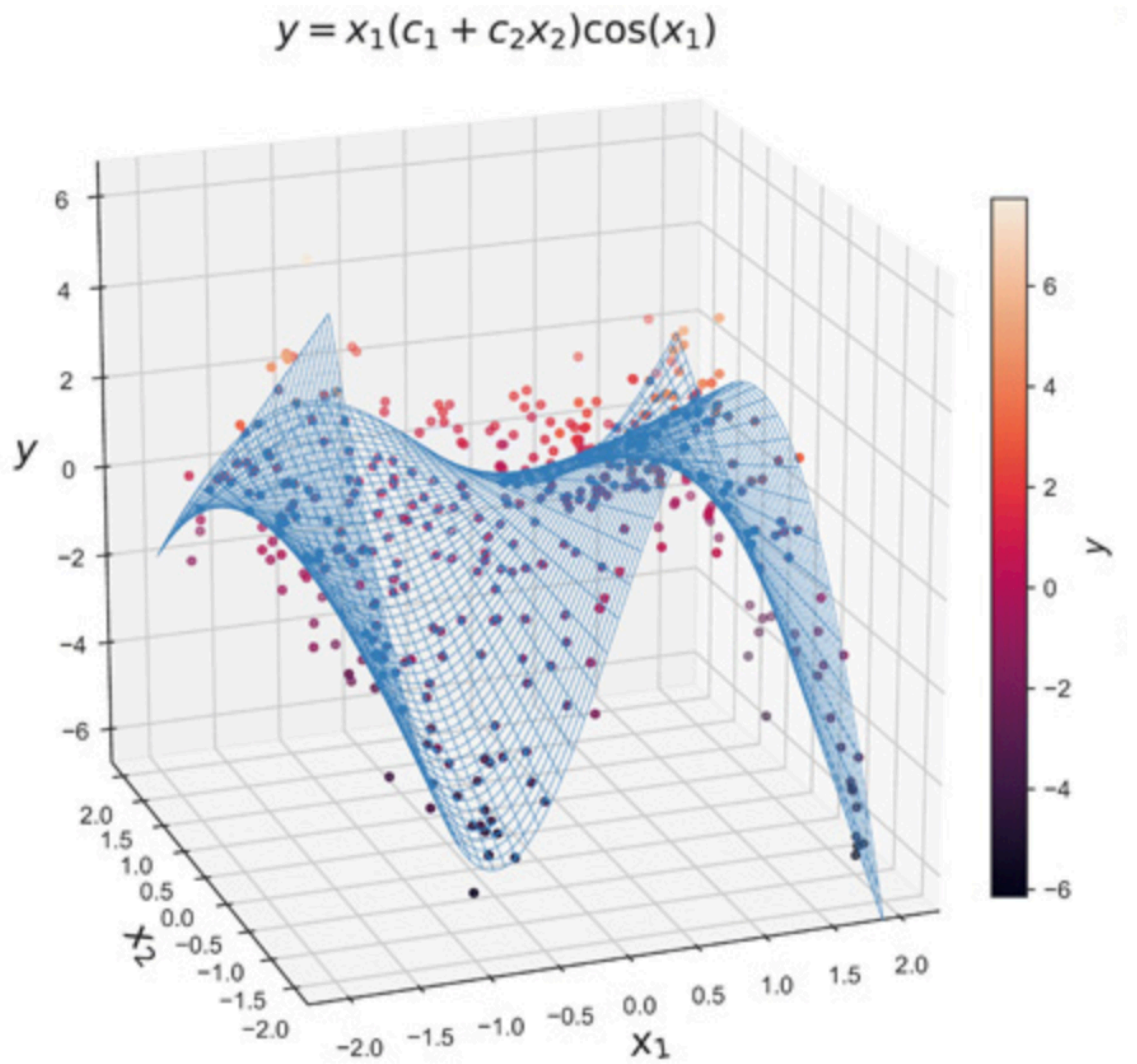
Predict future discoveries in material science
Tshitoyan et al., Nature, 2019

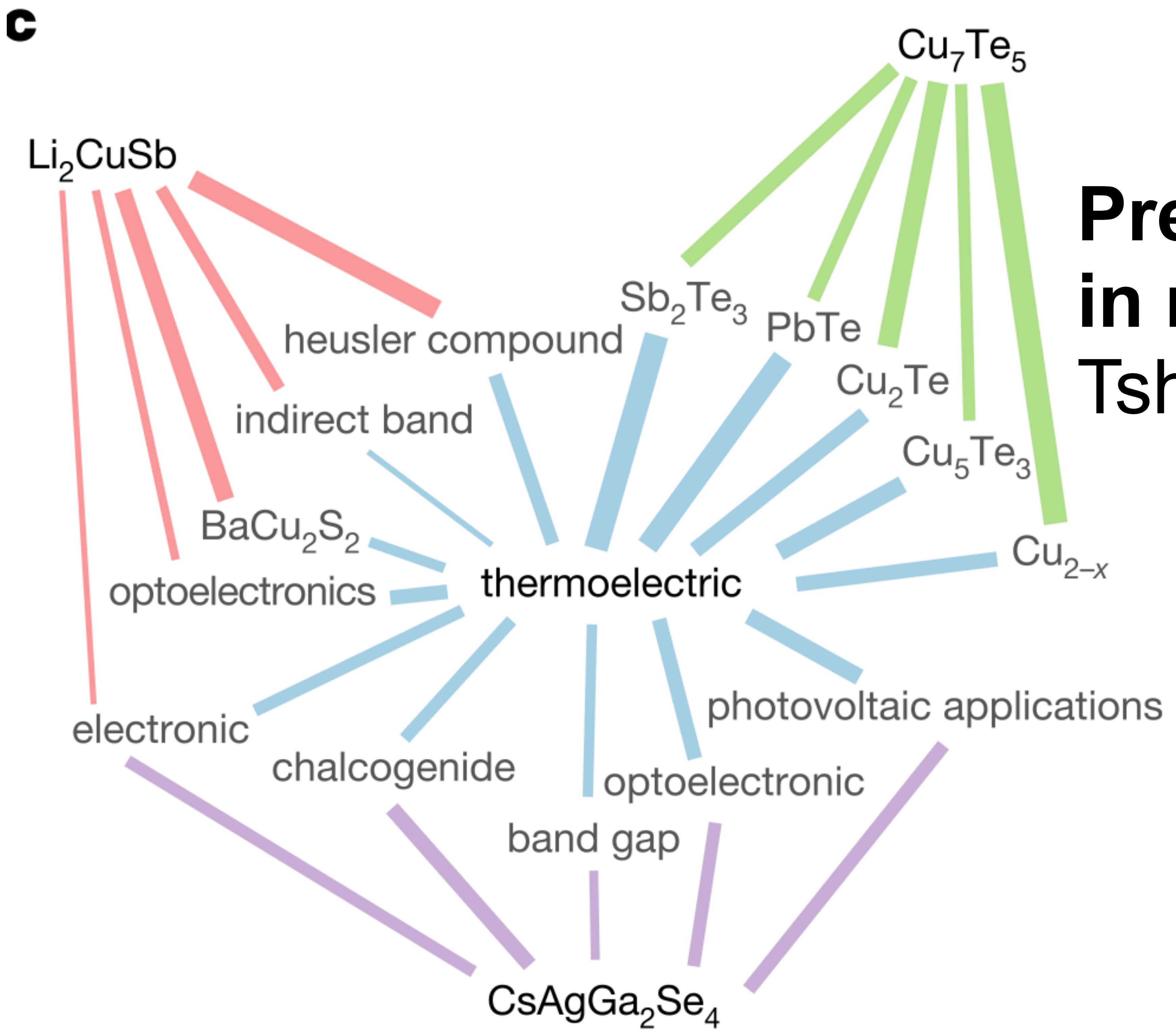


A Bayesian machine scientist to aid the solution of challenging scientific problems.
Guimera et al., Science Advances, 2020



Predicting lung cancer
Ardila et al., Nature Medicine, 2019





Predict future discoveries in material science
Tshitoyan et al., Nature, 2019

A Bayesian machine scientist to aid the solution of challenging scientific problems.
Guimera et al., Science Advances, 2020

have something predicted
(e.g.: interest in outcome, or missing data)

have something predicted
(e.g.: interest in outcome, or missing data)

understand relationships behind a phenomenon
(e.g.: what informs on outcomes)

have something predicted
(e.g.: interest in outcome, or missing data)

understand relationships behind a phenomenon
(e.g.: what informs on outcomes)

how well do some observations explain others
(e.g.: apply across distinct sets of observations)

have something predicted
(e.g.: interest in outcome, or missing data)

understand relationships behind a phenomenon
(e.g.: what informs on outcomes)

how well do some observations explain others
(e.g.: apply across distinct sets of observations)

understand where something is missing
(e.g.: what can not be predicted)

have something predicted
(e.g.: interest in outcome, or missing data)

understand relationships behind a phenomenon
(e.g.: what informs on outcomes)

how well do some observations explain others
(e.g.: apply across distinct sets of observations)

understand where something is missing
(e.g.: what can not be predicted)

have something predicted
(e.g.: interest in outcome, or missing data)

understand relationships behind a phenomenon
(e.g.: what informs on outcomes)

how well do some observations explain others
(e.g.: apply across distinct sets of observations)

understand where something is missing
(e.g.: what can not be predicted)

have something predicted
(e.g.: interest in outcome, or missing data)

understand relationships behind a phenomenon
(e.g.: what informs on outcomes)

how well do some observations explain others
(e.g.: apply across distinct sets of observations)

understand where something is missing
(e.g.: what can not be predicted)

understand where something is missing

example: Salganik et al., PNAS, 2020:

How well are life outcomes of children understood?

understand where something is missing

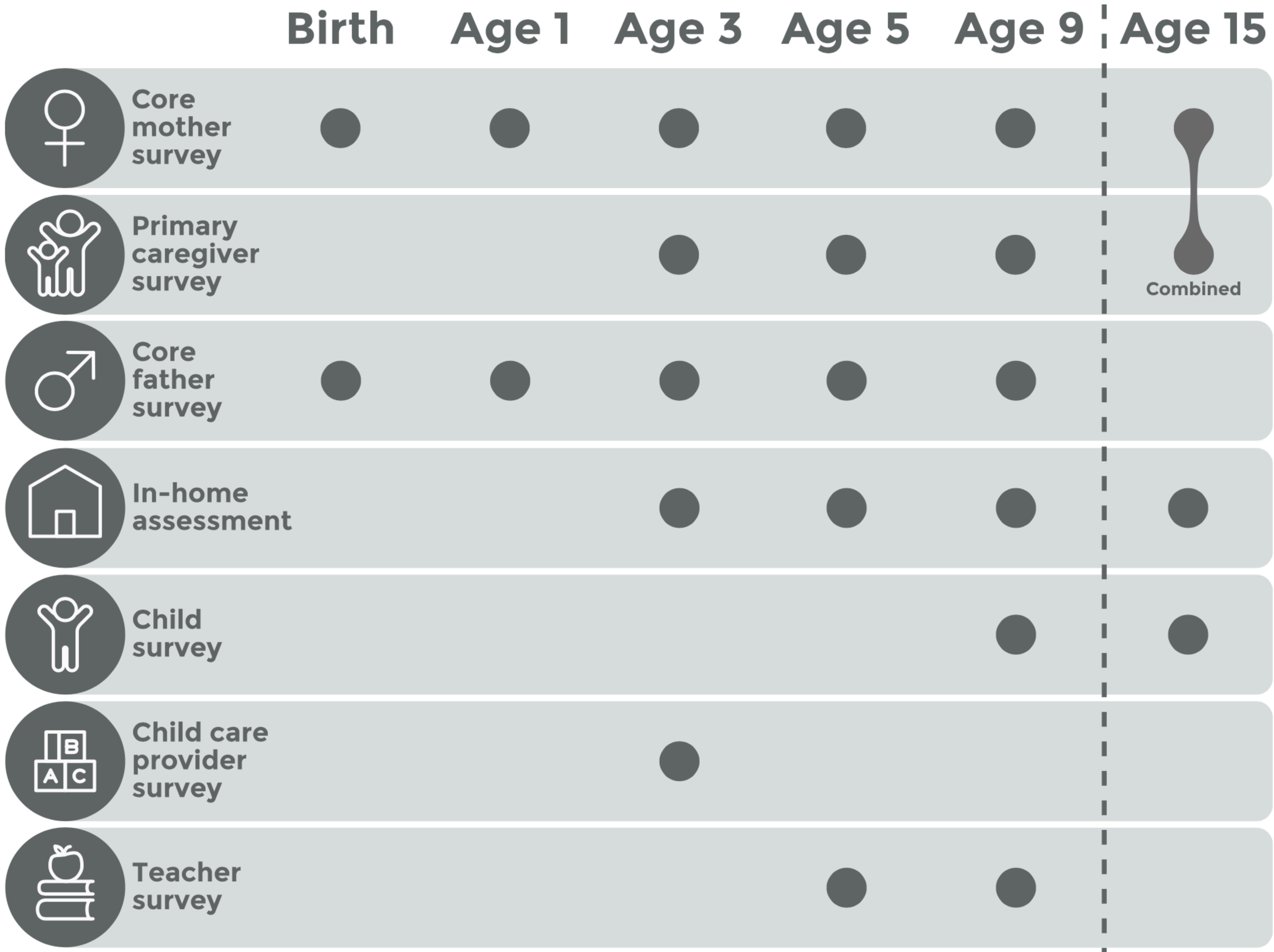
example: Salganik et al., PNAS, 2020:

How well are life outcomes of children understood?

	Birth	Age 1	Age 3	Age 5	Age 9
Core mother survey	●	●	●	●	●
Primary caregiver survey			●	●	●
Core father survey	●	●	●	●	●
In-home assessment			●	●	●
Child survey					●
Child care provider survey			●		
Teacher survey				●	●

understand where something is missing

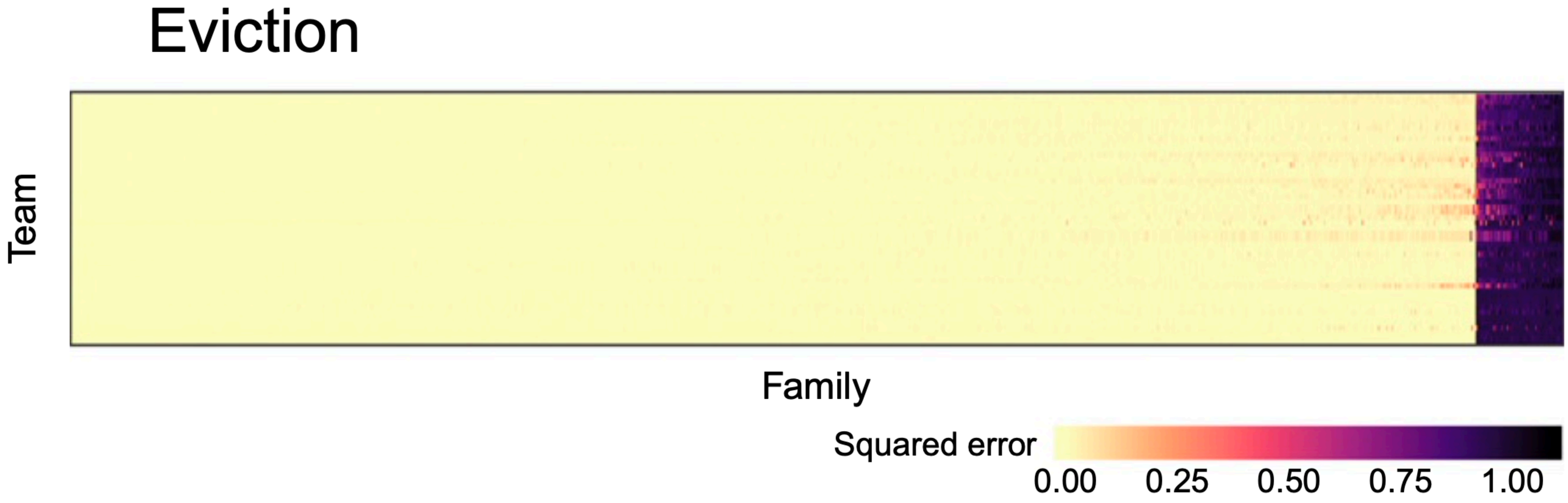
example: Salganik et al., PNAS, 2020:
How well are life outcomes of children understood?



understand where something is missing

example: Salganik et al., PNAS, 2020:

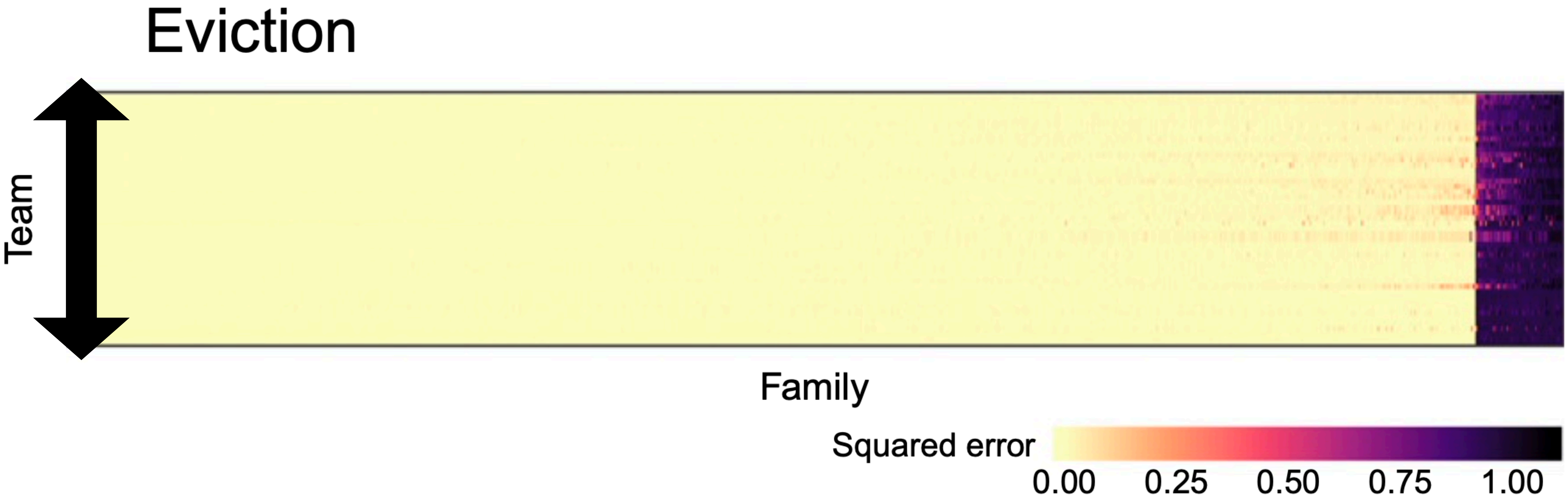
How well are life outcomes of children understood?



understand where something is missing

example: Salganik et al., PNAS, 2020:

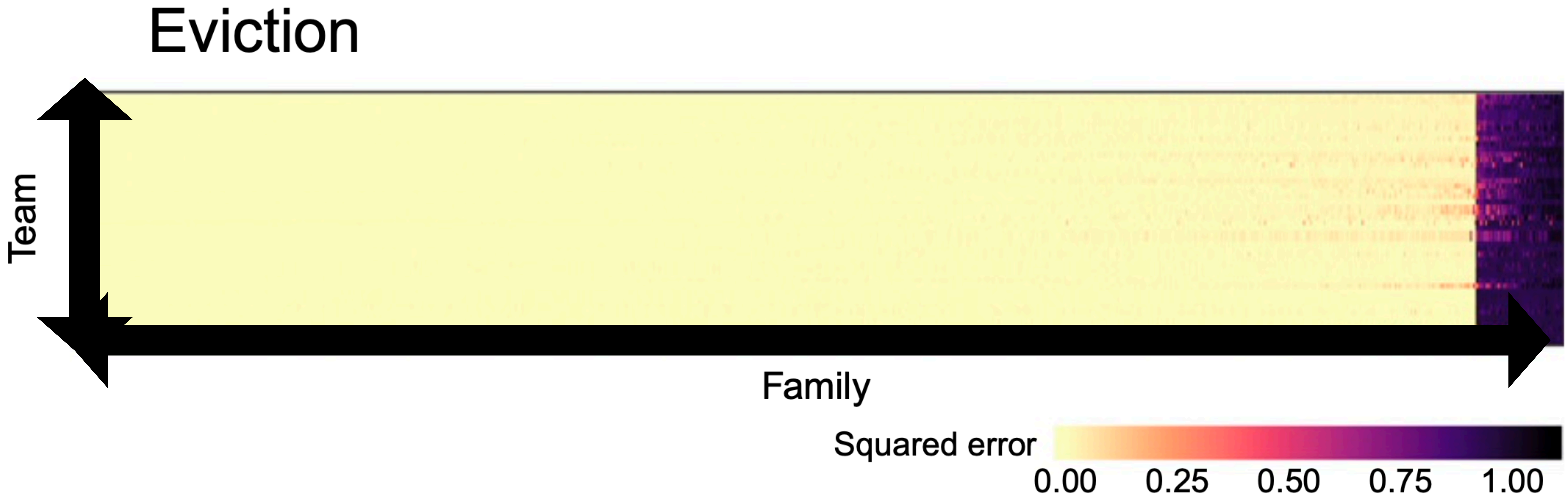
How well are life outcomes of children understood?



understand where something is missing

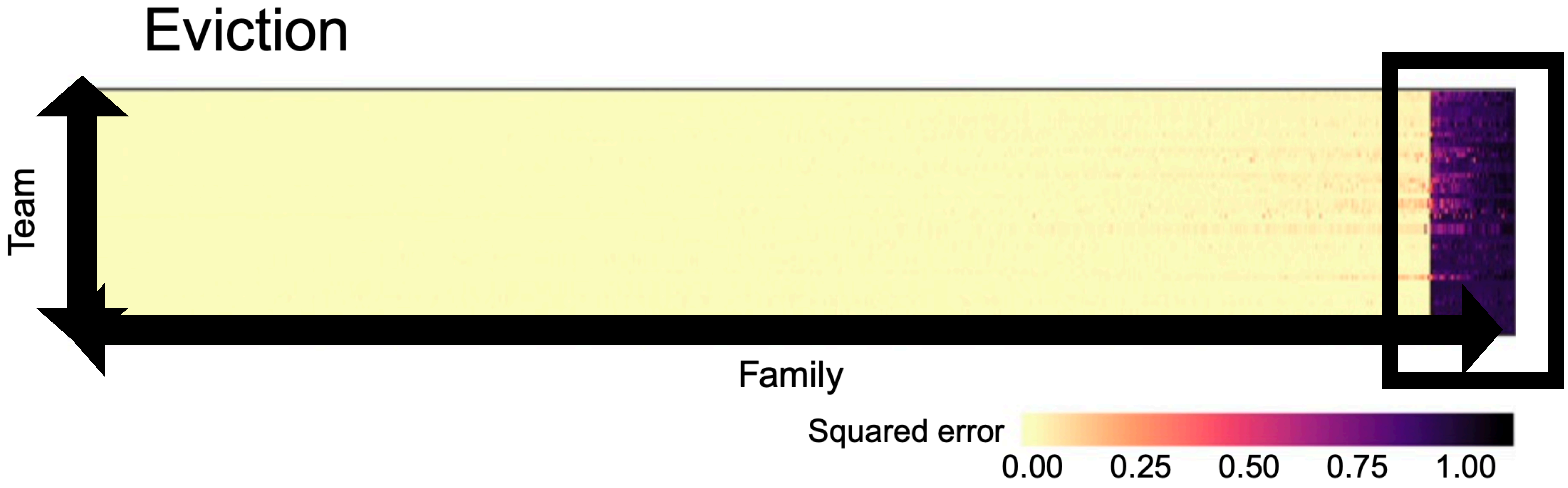
example: Salganik et al., PNAS, 2020:

How well are life outcomes of children understood?



understand where something is missing

example: Salganik et al., PNAS, 2020:
How well are life outcomes of children understood?



have something predicted
(e.g.: __________)

understand relationships behind a phenomenon
(e.g.: __________)

how well do some observations explain others
(e.g.: __________)

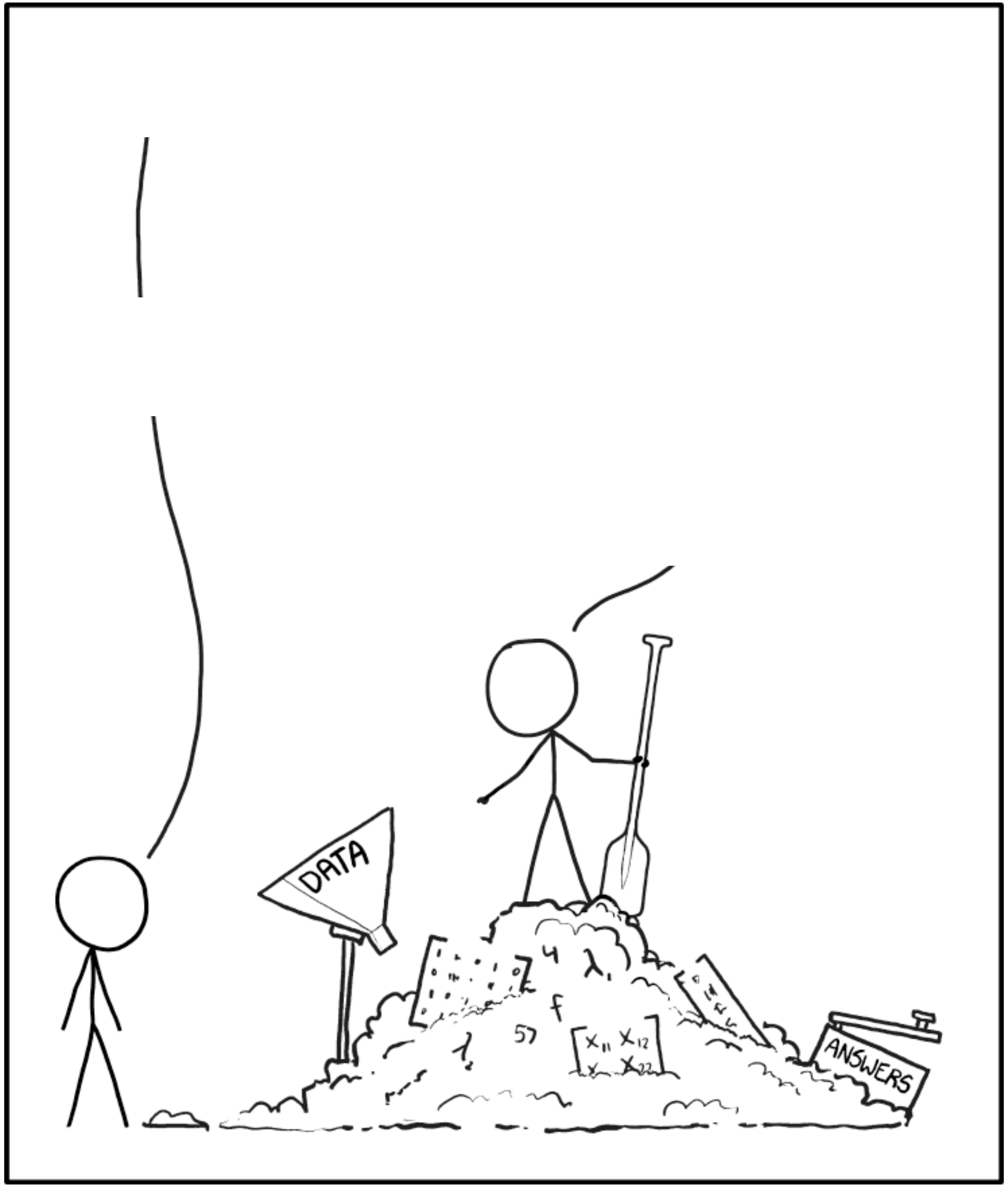
understand where something is missing
(e.g.: __________)

demystify → understand conceptual possibilities

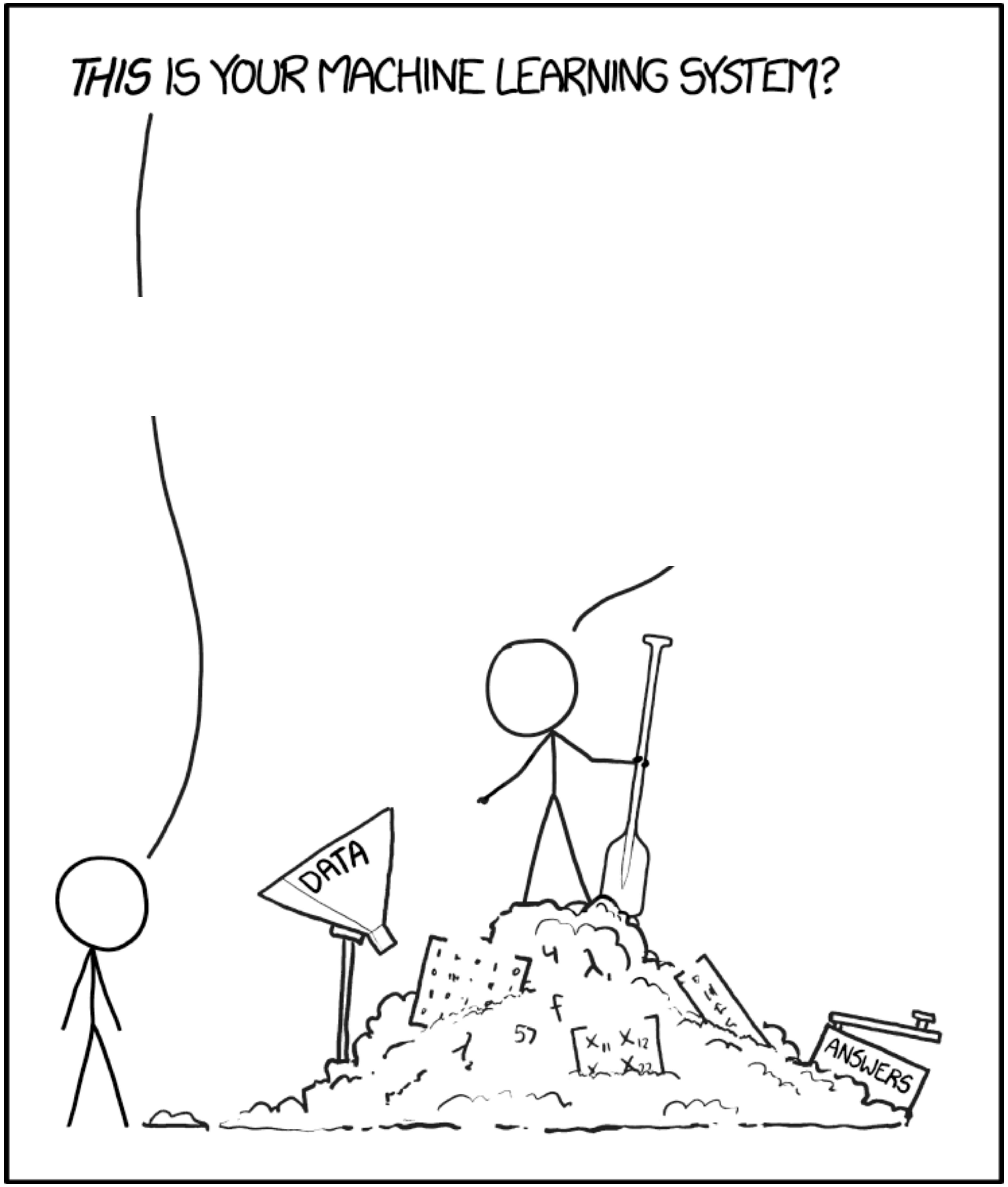
Slides: https://bit.ly/3bKMrhM

demystify → understand conceptual possibilities

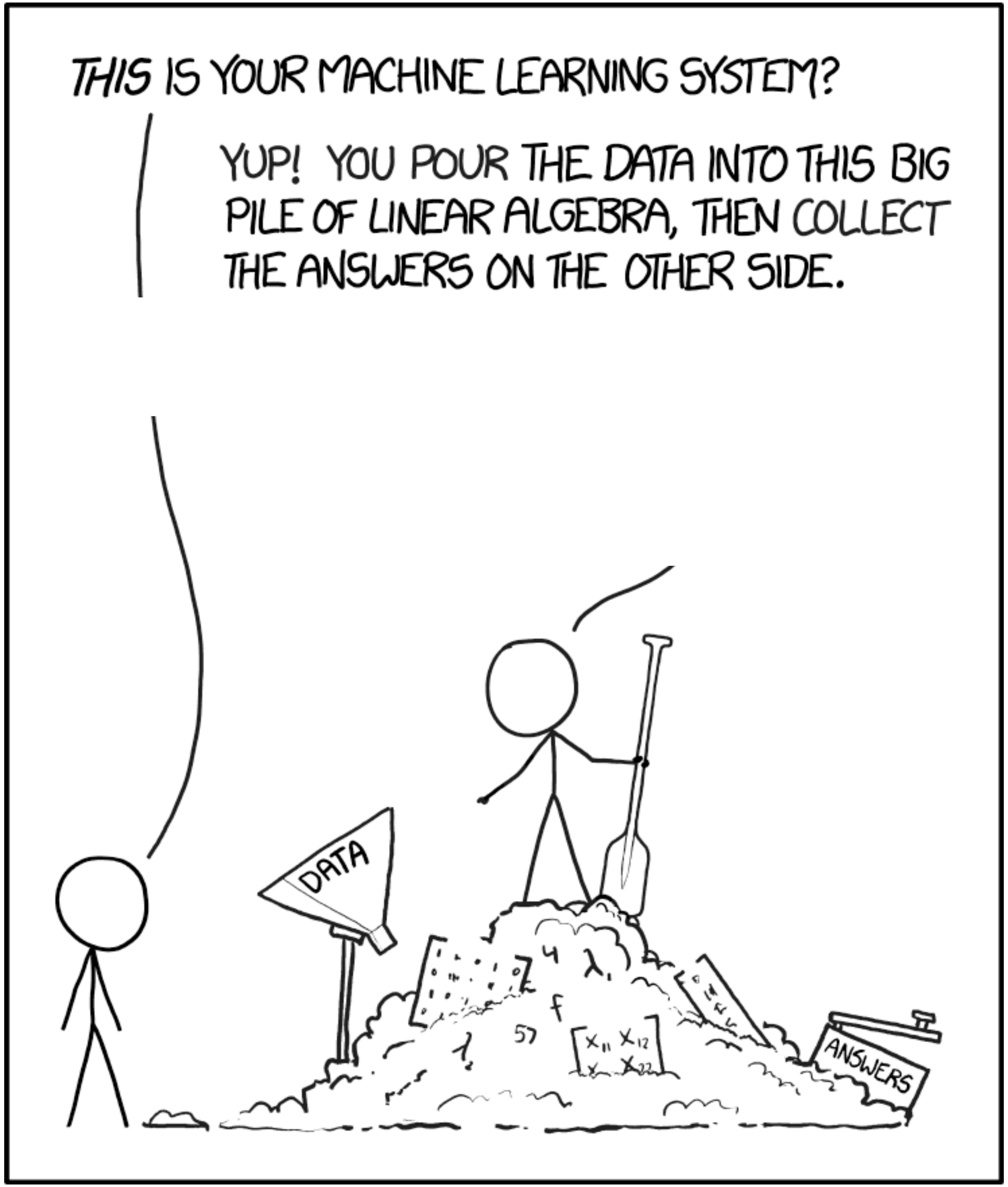
orientation → understand main steps



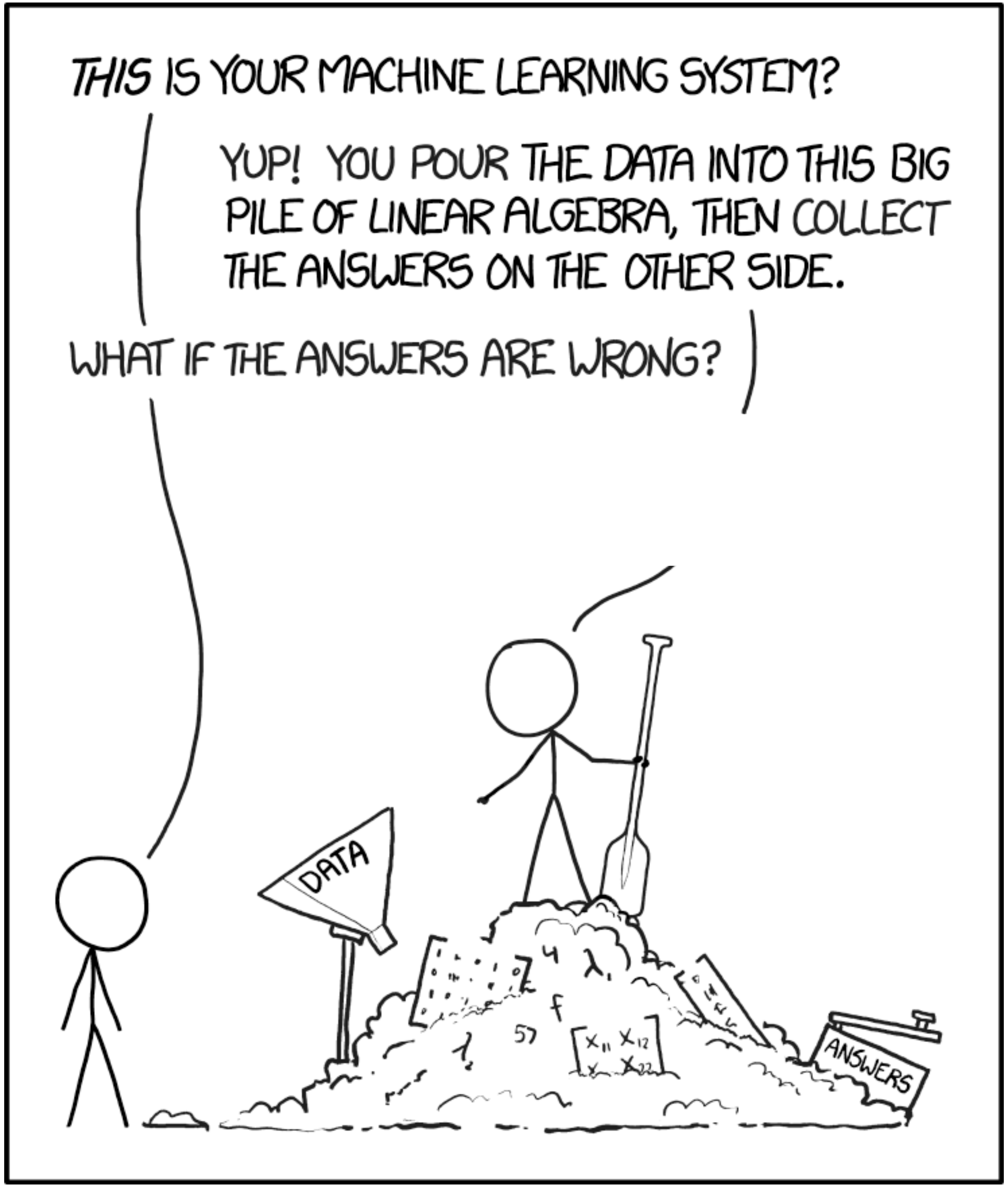
xkcd.com



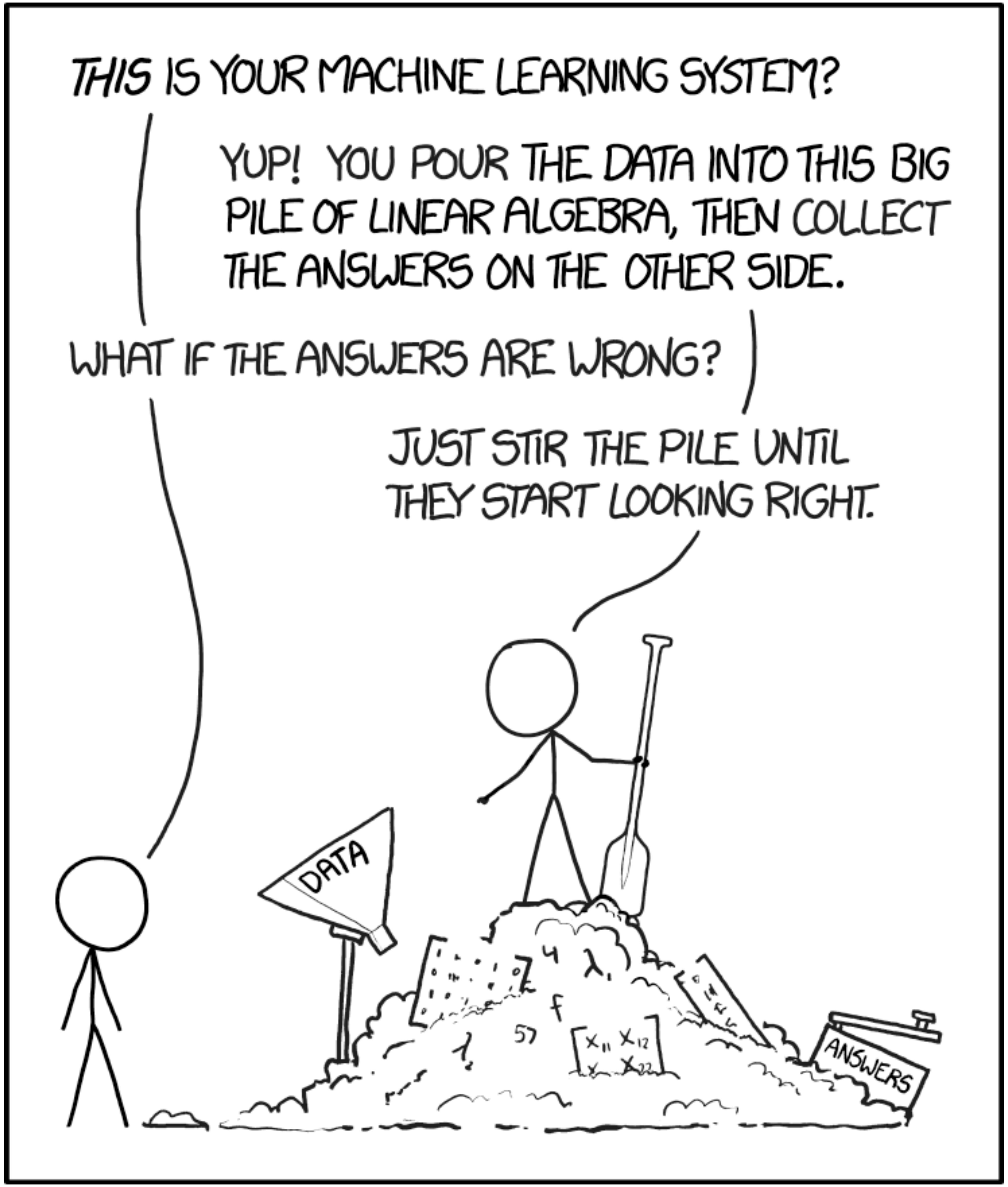
xkcd.com


THIS IS YOUR MACHINE LEARNING SYSTEM?
YUP! YOU POUR THE DATA INTO THIS BIG PILE OF LINEAR ALGEBRA, THEN COLLECT THE ANSWERS ON THE OTHER SIDE.
DATA
ANSWERS

xkcd.com


THIS IS YOUR MACHINE LEARNING SYSTEM?
YUP! YOU POUR THE DATA INTO THIS BIG PILE OF LINEAR ALGEBRA, THEN COLLECT THE ANSWERS ON THE OTHER SIDE.
WHAT IF THE ANSWERS ARE WRONG?
DATA
ANSWERS

xkcd.com


THIS IS YOUR MACHINE LEARNING SYSTEM?
YUP! YOU POUR THE DATA INTO THIS BIG PILE OF LINEAR ALGEBRA, THEN COLLECT THE ANSWERS ON THE OTHER SIDE.
WHAT IF THE ANSWERS ARE WRONG?
JUST STIR THE PILE UNTIL THEY START LOOKING RIGHT.
DATA
ANSWERS

xkcd.com

Two tricks to avoid misleading models:

Machine learning

Two tricks to avoid misleading models:

Data cleaning

Machine learning

Two tricks to avoid misleading models:

Data cleaning

Separation of training- and test-set

Machine learning

Two tricks to avoid misleading models:

recommended start:
1) Broman et Woo, 2017: Data Organization in spreadsheets;
2) "tidy data"

Data cleaning

Separation of training- and test-set

Machine learning

Two tricks to avoid misleading models:

recommended start:
1) Broman et Woo, 2017: Data Organization in spreadsheets;
2) “tidy data”

Data cleaning

e.g.: different experimental series, or subsample data

Separation of training- and test-set

Machine learning

Machine learning is based on **features**.

features ➞ model ➞ inference

e.g.: measurements

Machine learning is based on **features**.

features → model → inference

e.g.: measurements

- generally, qualitatively diverse features are better

Machine learning is based on **features**.

features → model → inference

e.g.: measurements

- generally, qualitatively diverse features are better

- generally, number of independent features should be smaller than observations

Feature engineering may help or be needed.

Feature engineering may help or be needed.

Dimensionality reduction: e.g.: if too many features (PCA etc.)

Feature engineering may help or be needed.

Dimensionality reduction: e.g.: if too many features (PCA etc.)

Normalize: e.g.: “to compare apples to oranges” (z-scoring etc.)

Feature engineering may help or be needed.

Dimensionality reduction: e.g.: if too many features (PCA etc.)

Normalize: e.g.: “to compare apples to oranges” (z-scoring etc.)

Encode: e.g.: transform categorial data to numbers (one-hot encoding etc.)

Feature engineering may help or be needed.

Dimensionality reduction: e.g.: if too many features (PCA etc.)

Normalize: e.g.: “to compare apples to oranges” (z-scoring etc.)

Encode: e.g.: transform categorial data to numbers (one-hot encoding etc.)

Note: feature engineering may also do harm.

Feature engineering may help or be needed.

Dimensionality reduction: e.g.: if too many features (PCA etc.)

Normalize: e.g.: “to compare apples to oranges” (z-scoring etc.)

Encode: e.g.: transform categorial data to numbers (one-hot encoding etc.)

Note: feature engineering may also do harm.

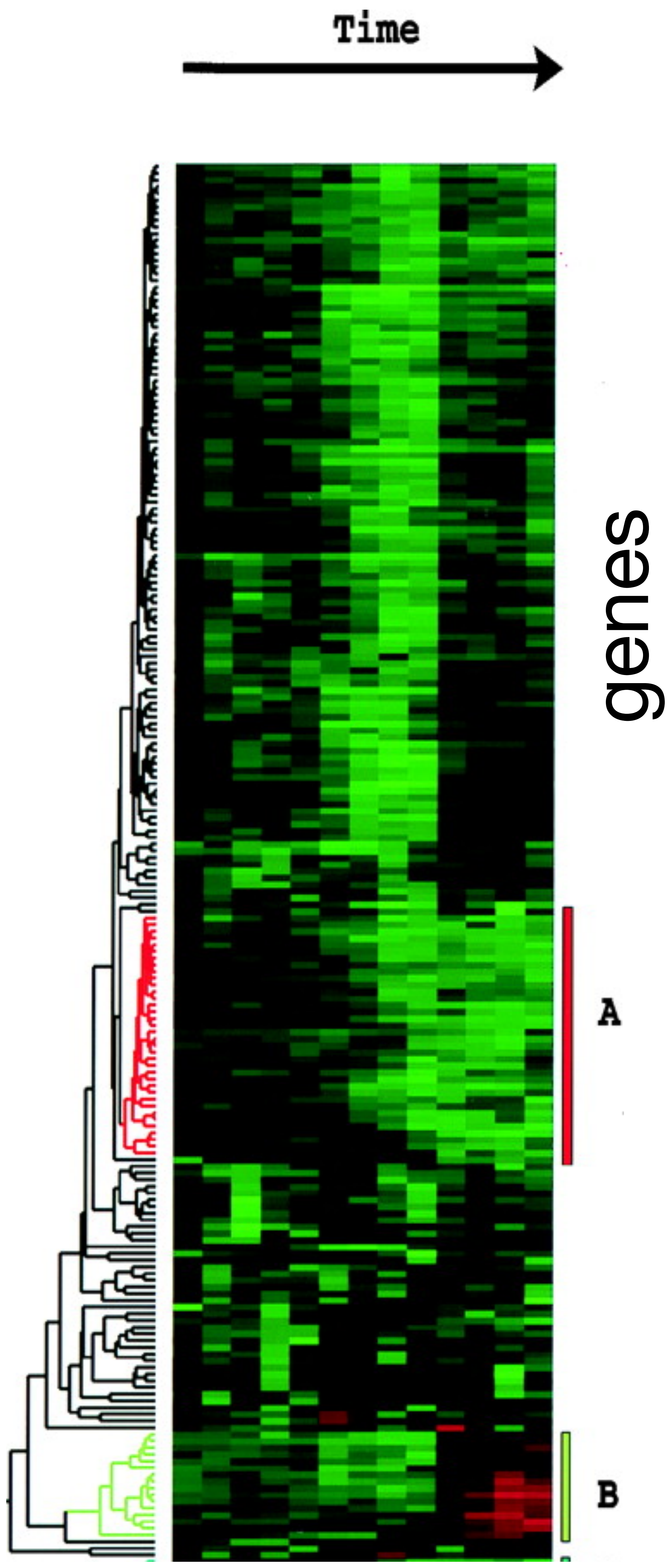
A nice guide to feature engineering:
https://towardsdatascience.com/feature-engineering-for-machine-learning-3a5e293a5114

Which machine learning approach should I follow?

Supervised and unsupervised approaches form the two major families of machine learning algorithms.

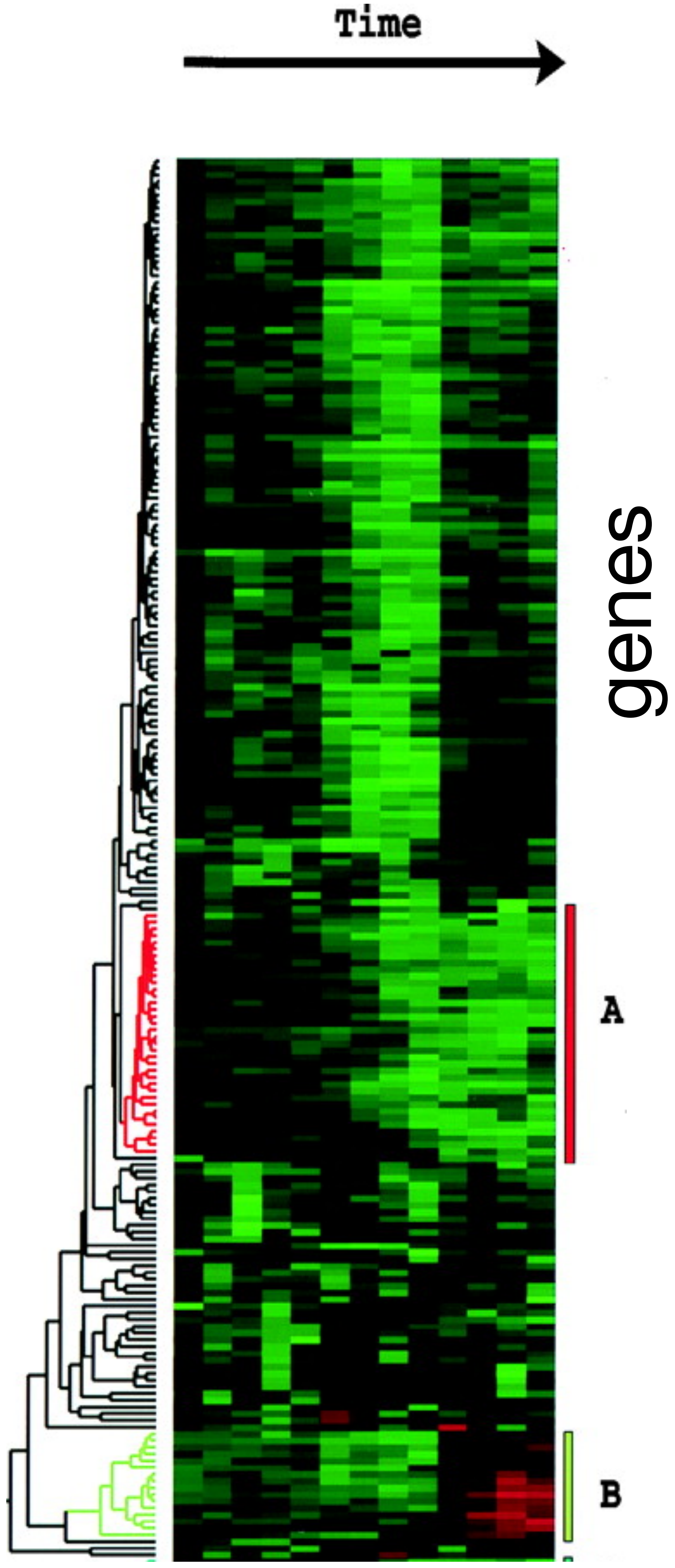
Supervised and unsupervised approaches form the two major families of machine learning algorithms.

unsupervised:
e.g.: clustergram



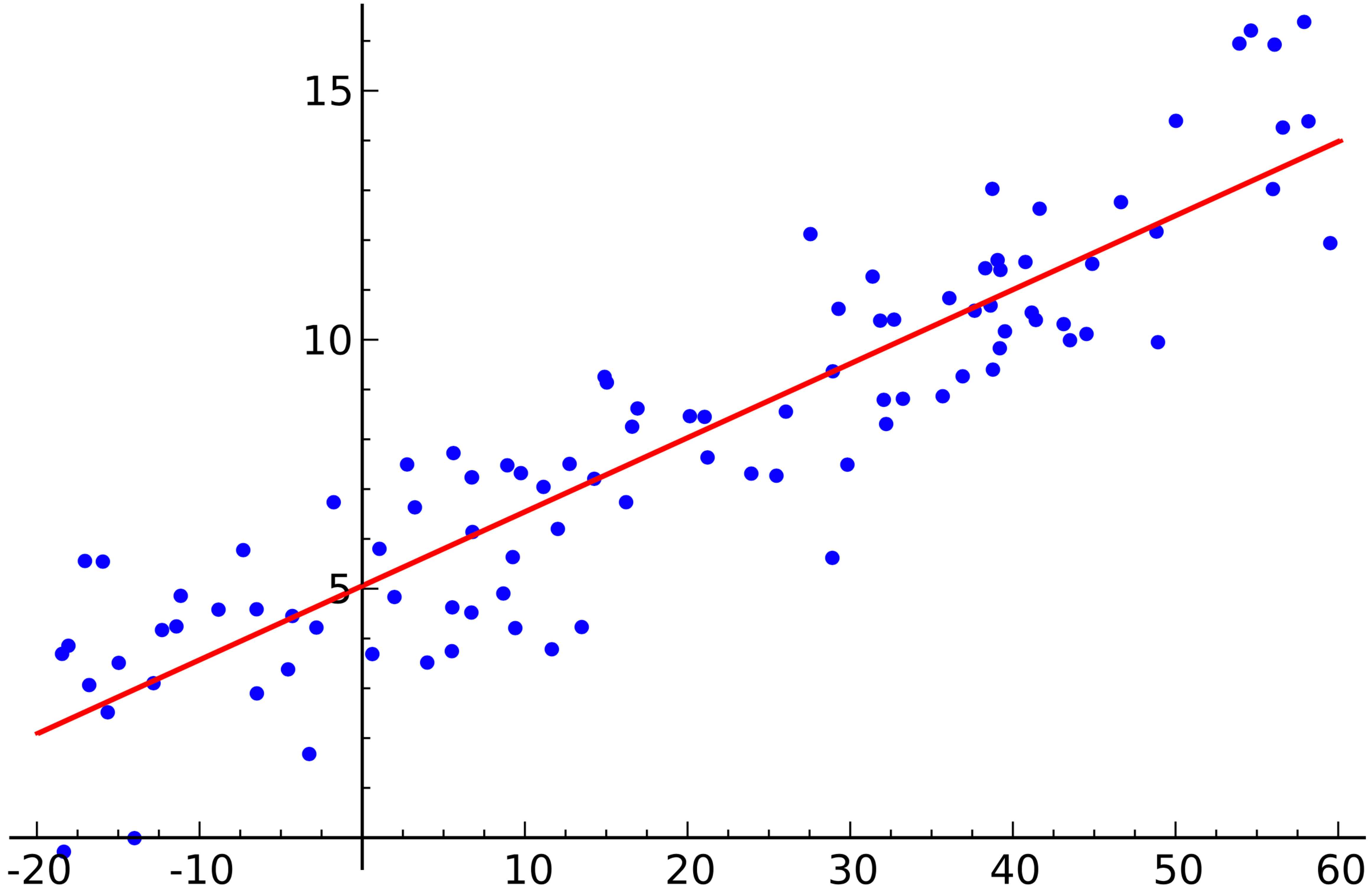
Eisen et al. 1998

Supervised and unsupervised approaches form the two major families of machine learning algorithms.



unsupervised:
e.g.: clustergram

Eisen et al. 1998

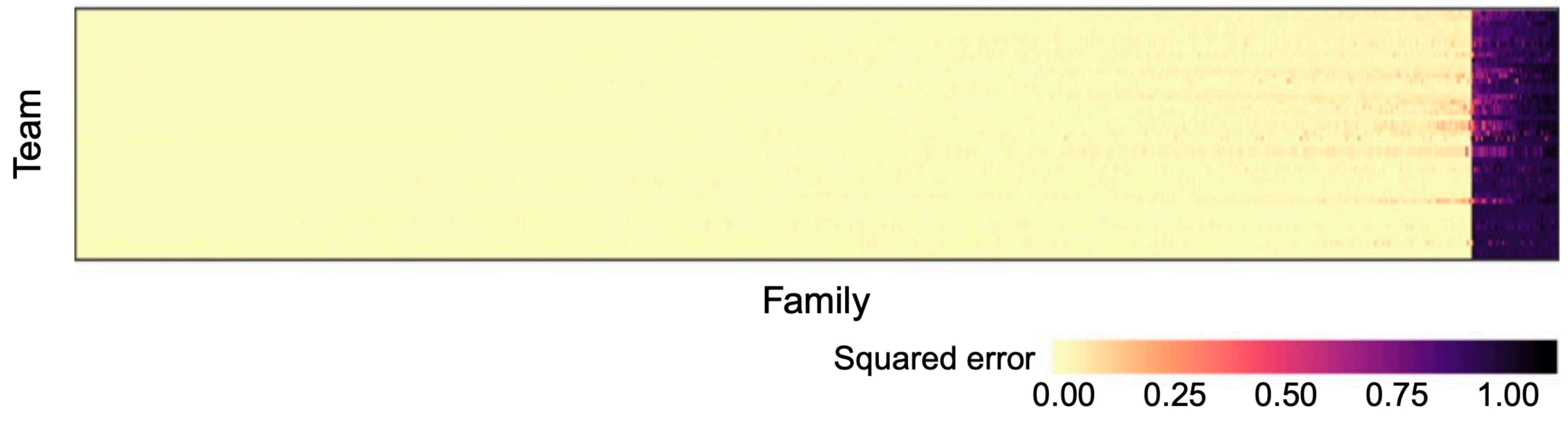


supervised:
e.g.: linear regression

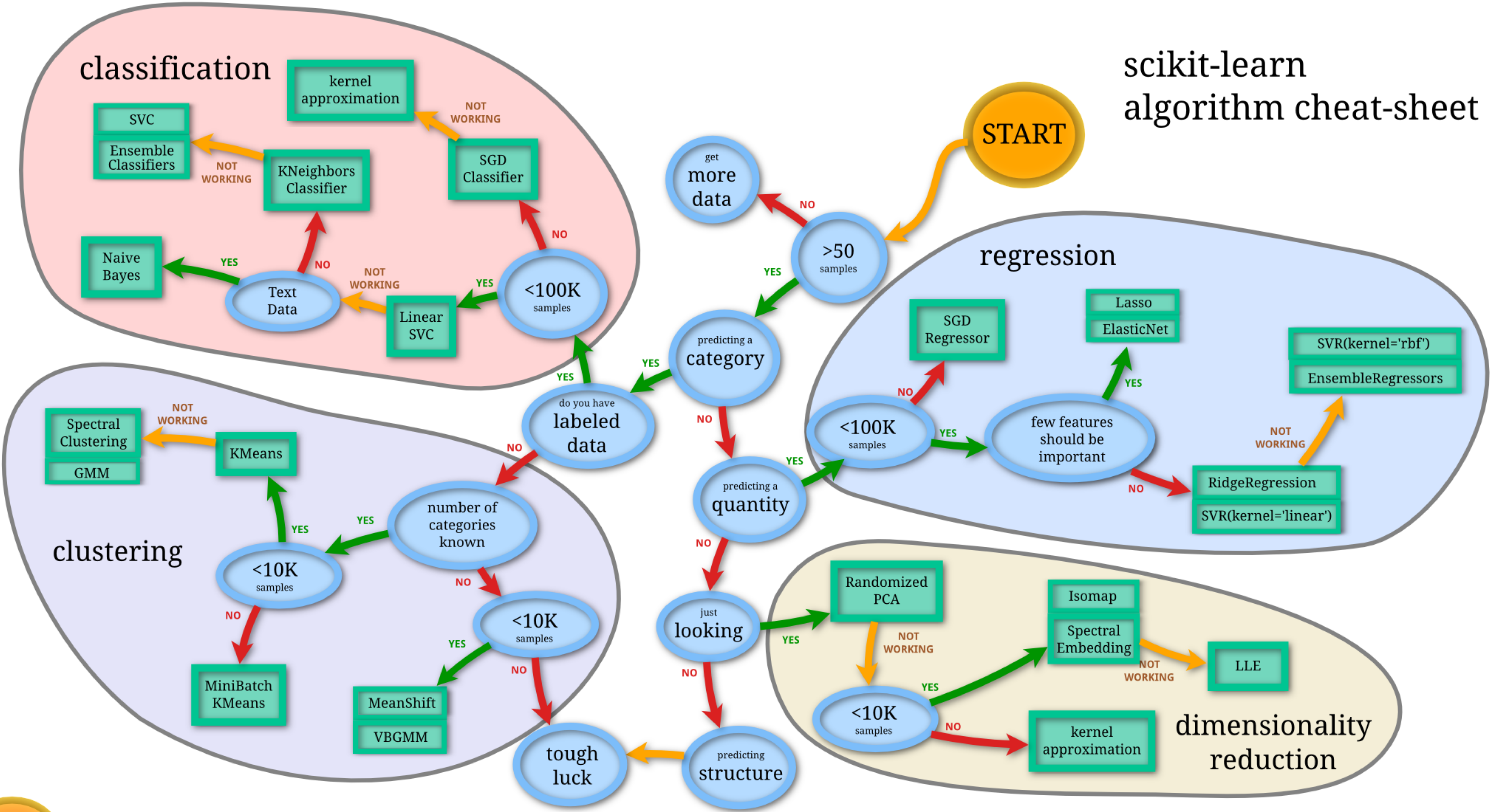
Which specific machine learning approach should I follow?

Which <u>specific</u> machine learning approach should I follow?

*It may not matter. Remember the following?



Which specific machine learning approach should I follow?



Back

https://scikit-learn.org/stable/tutorial/machine_learning_map/index.html

Note: above webpage is interactive, revealing more details on distinct approaches

Automated machine learning

nice summary:
https://heartbeat.fritz.ai/automl-the-next-wave-of-machine-learning-5494baac615f

Automated machine learning

- Reduces human time spent

nice summary:
https://heartbeat.fritz.ai/automl-the-next-wave-of-machine-learning-5494baac615f

Automated machine learning

- Reduces human time spent
- Reduces human bias

nice summary:
https://heartbeat.fritz.ai/automl-the-next-wave-of-machine-learning-5494baac615f

Automated machine learning

- Reduces human time spent
- Reduces human bias
- Superior to >95% of human data scientists

nice summary:
https://heartbeat.fritz.ai/automl-the-next-wave-of-machine-learning-5494baac615f

Automated machine learning

- Reduces human time spent
- Reduces human bias
- Superior to >95% of human data scientists
- Quickly evolving field

nice summary:
https://heartbeat.fritz.ai/automl-the-next-wave-of-machine-learning-5494baac615f

Automated machine learning

- Reduces human time spent
- Reduces human bias
- Superior to >95% of human data scientists
- Quickly evolving field
- Good start is auto-sklearn

nice summary:
https://heartbeat.fritz.ai/automl-the-next-wave-of-machine-learning-5494baac615f

demystify → understand conceptual possibilities

orientation → understand main steps

demystify → understand conceptual possibilities

orientation → understand main steps

coding → have a well-extendable example

Slides: https://bit.ly/3bKMrhM

NOTEBOOKS_ EDUCATORS HELP SIGN IN SIGN UP FOR FREE

Beth1126

PDA Python DataFrames and Graphs Webinar 2020

cloned from **Beth526/PDA-Python-Webinar**

0

72

FORK THIS PROJECT

Make a fork of this project and run your own experiments.

A UIC Postdoctoral Association webinar by a postdoc who is learning python

Last updated: April 23rd, 2020

Python DataFrames and Basic Graphs

Python is a popular programming language that can be used for analysis and graphing of large datasets. There are a lot of **free resouces online** to learn python:

- https://www.codecademy.com
- https://www.datacamp.com
- https://docs.python.org/3.8/tutorial/index.html
- Coursera courses such as "Programming for Everybody (Getting Started with Python)" from University of Michigan
- Linkedin Learning courses (free with UIC login) such as "Learning Python"

To use this Jupyter Notebook you *do not have to download anything*, but a way to download and use Python and many of the Python packages is by downloading the **Anaconda platform**:

- https://www.anaconda.com/distribution/

In this short walk-through we are going to focus on practical Python packages for data analysis called **Pandas** and **Seaborn**. Packages add functionality to basic Python. Here are links to the

Remember? Last week's lecture from Bethany Baumann.

https://notebooks.ai/Beth1126/pda-python-dataframes-and-graphs-webinar-2020-54351877)

NOTEBOOKS_

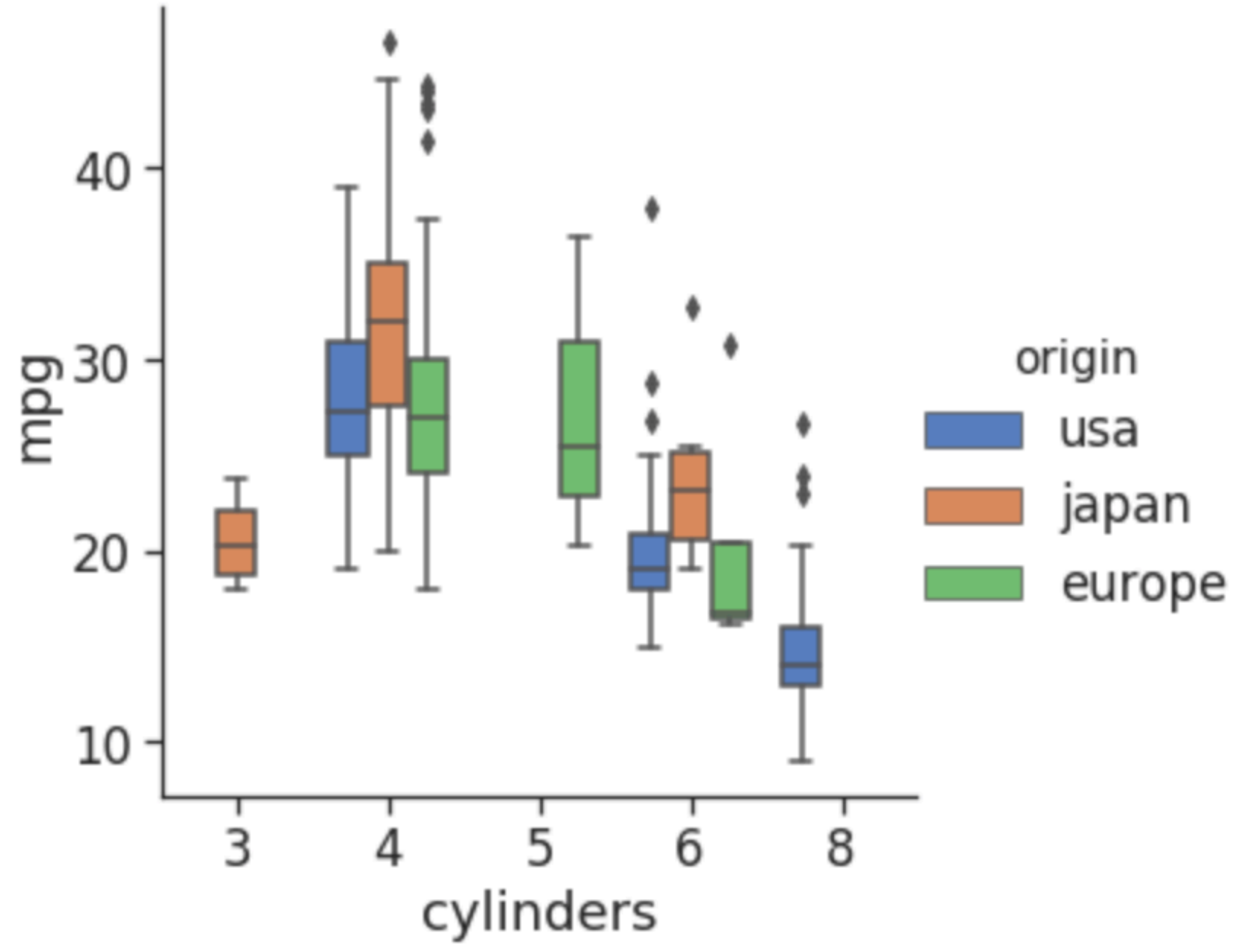
SIGN IN

SIGN UP FOR FREE

0

72

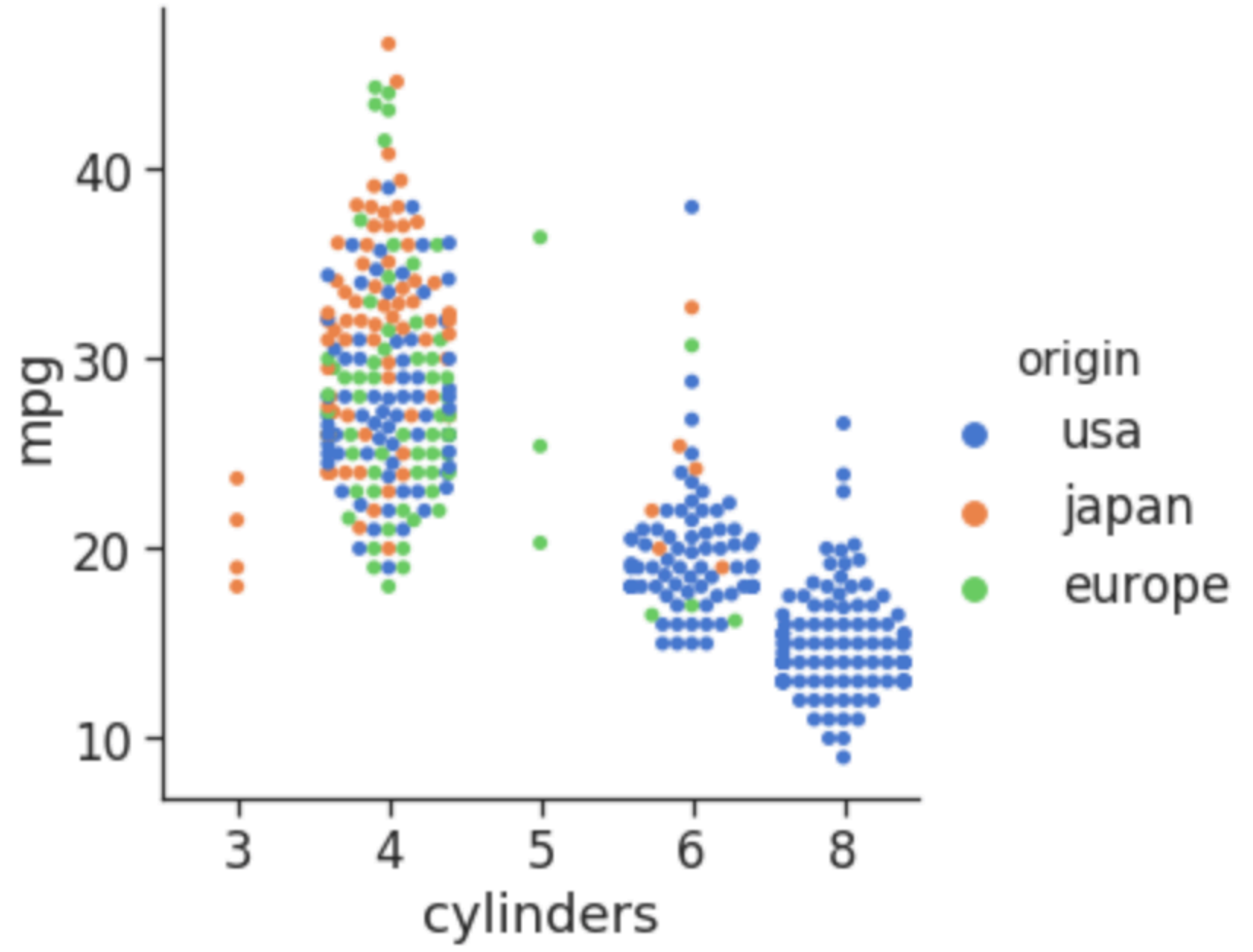
```
<seaborn.axisgrid.FacetGrid at 0x7f65293f9640>
```




```
sns.catplot(x = 'cylinders', y = "mpg", hue = "origin", data = df, kind = 'swarm')
```

OUTPUT

```
<seaborn.axisgrid.FacetGrid at 0x7f65272f42e0>
```



FORK THIS PROJECT

Make a fork of this project and run your own experiments.

Remember? Last week's lecture from Bethany Baumann.

https://notebooks.ai/Beth1126/pda-python-dataframes-and-graphs-webinar-2020-54351877)

mpg	cylinders	displacement	horsepower	weight	acceleration	model_year	origin	name
18.0	8	307.0	130.0	3504	12.0	70	usa	chevrolet chevelle malibu
15.0	8	350.0	165.0	3693	11.5	70	usa	buick skylark 320
18.0	8	318.0	150.0	3436	11.0	70	usa	plymouth satellite
16.0	8	304.0	150.0	3433	12.0	70	usa	amc rebel sst
17.0	8	302.0	140.0	3449	10.5	70	usa	ford torino

to predict: miles per gallon

mpg	cylinders	displacement	horsepower	weight	acceleration	model_year	origin	name
18.0	8	307.0	130.0	3504	12.0	70	usa	chevrolet chevelle malibu
15.0	8	350.0	165.0	3693	11.5	70	usa	buick skylark 320
18.0	8	318.0	150.0	3436	11.0	70	usa	plymouth satellite
16.0	8	304.0	150.0	3433	12.0	70	usa	amc rebel sst
17.0	8	302.0	140.0	3449	10.5	70	usa	ford torino

to predict: miles per gallon

Features

mpg	cylinders	displacement	horsepower	weight	acceleration	model_year	origin	name
18.0	8	307.0	130.0	3504	12.0	70	usa	chevrolet chevelle malibu
15.0	8	350.0	165.0	3693	11.5	70	usa	buick skylark 320
18.0	8	318.0	150.0	3436	11.0	70	usa	plymouth satellite
16.0	8	304.0	150.0	3433	12.0	70	usa	amc rebel sst
17.0	8	302.0	140.0	3449	10.5	70	usa	ford torino

to predict: miles per gallon

Features

ignore

mpg	cylinders	displacement	horsepower	weight	acceleration	model_year	origin	name
18.0	8	307.0	130.0	3504	12.0	70	usa	chevrolet chevelle malibu
15.0	8	350.0	165.0	3693	11.5	70	usa	buick skylark 320
18.0	8	318.0	150.0	3436	11.0	70	usa	plymouth satellite
16.0	8	304.0	150.0	3433	12.0	70	usa	amc rebel sst
17.0	8	302.0	140.0	3449	10.5	70	usa	ford torino

to predict: miles per gallon

Features

ignore

mpg	cylinders	displacement	horsepower	weight	acceleration	model_year	origin	name
18.0	8	307.0	130.0	3504	12.0	70	usa	chevrolet chevelle malibu
15.0	8	350.0	165.0	3693	11.5	70	usa	buick skylark 320
18.0	8	318.0	150.0	3436	11.0	70	usa	plymouth satellite
16.0	8	304.0	150.0	3433	12.0	70	usa	amc rebel sst
17.0	8	302.0	140.0	3449	10.5	70	usa	ford torino

https://notebooks.ai/tstoeger/predict-miles-per-gallon-1342dcb1/