GO DATA DRIVEN

PyData London 2022

Models-schmodels: why you should care about Data-Centric AI

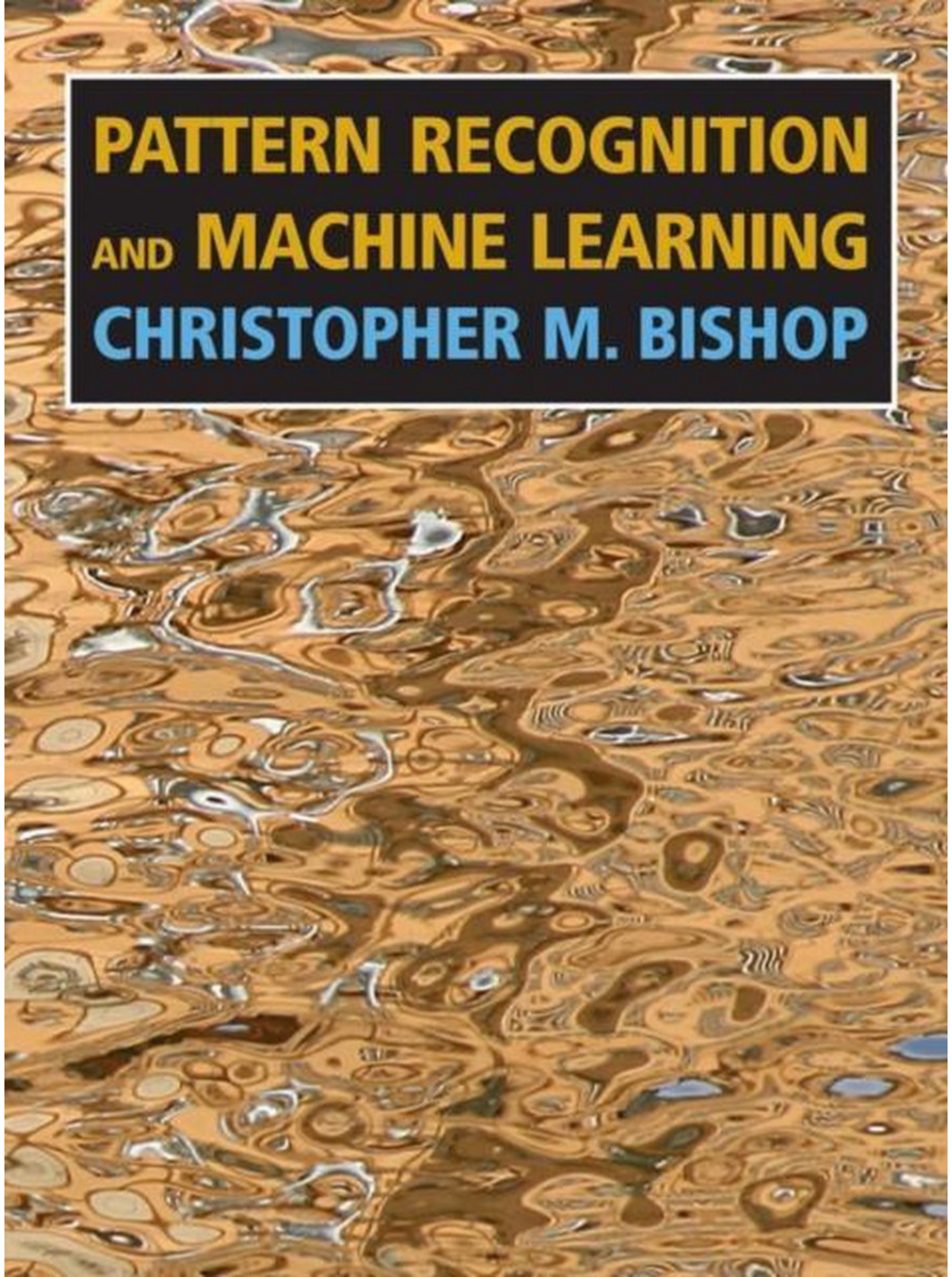
Marysia Winkels, PyData London 2022

GO
DATA
DRIVEN

The way I learned

data science

🎓 Study the algorithms



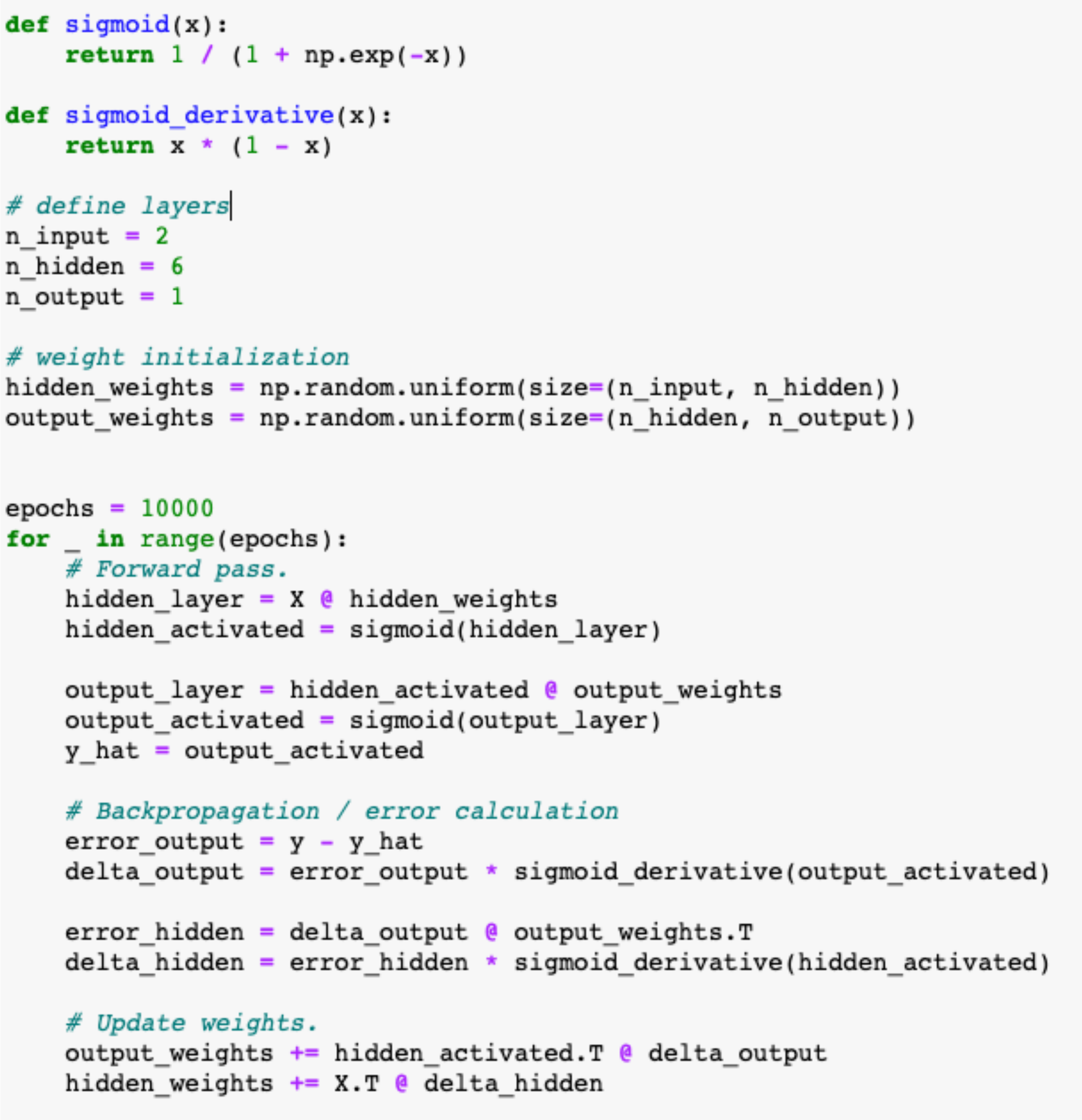
PATTERN RECOGNITION AND MACHINE LEARNING

CHRISTOPHER M. BISHOP

The way I learned

data science

- 🎓 Study the algorithms
- 🛠️ Implement the algorithms



```
def sigmoid(x):
    return 1 / (1 + np.exp(-x))

def sigmoid_derivative(x):
    return x * (1 - x)

# define layers
n_input = 2
n_hidden = 6
n_output = 1

# weight initialization
hidden_weights = np.random.uniform(size=(n_input, n_hidden))
output_weights = np.random.uniform(size=(n_hidden, n_output))


epochs = 10000
for _ in range(epochs):
    # Forward pass.
    hidden_layer = X @ hidden_weights
    hidden_activated = sigmoid(hidden_layer)

    output_layer = hidden_activated @ output_weights
    output_activated = sigmoid(output_layer)
    y_hat = output_activated

    # Backpropagation / error calculation
    error_output = y - y_hat
    delta_output = error_output * sigmoid_derivative(output_activated)

    error_hidden = delta_output @ output_weights.T
    delta_hidden = error_hidden * sigmoid_derivative(hidden_activated)

    # Update weights.
    output_weights += hidden_activated.T @ delta_output
    hidden_weights += X.T @ delta_hidden
```

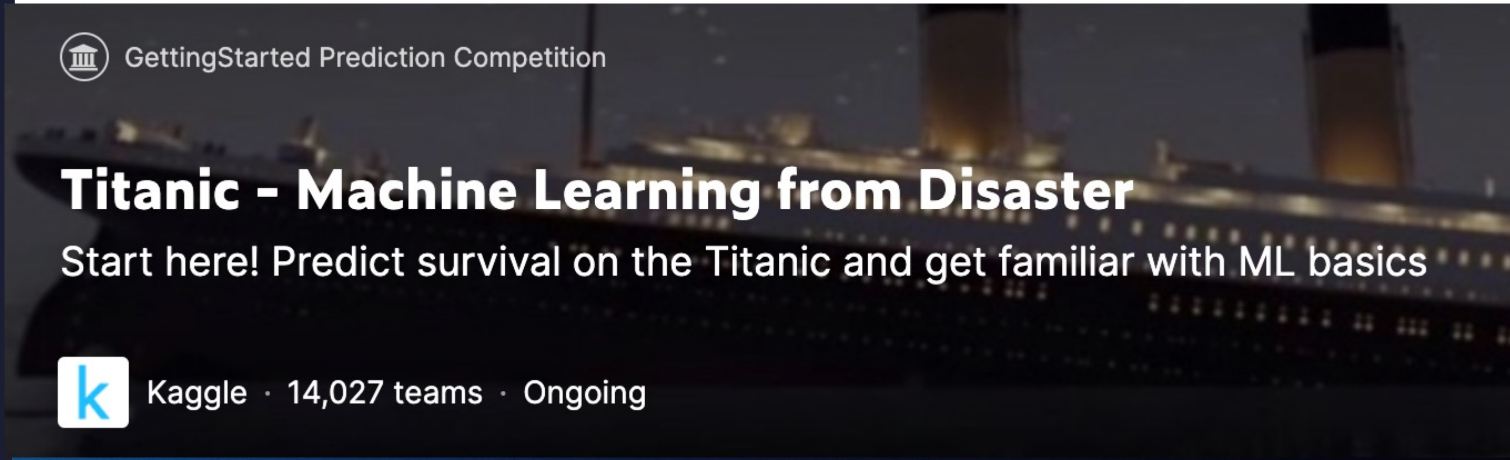
GO
DATA
DRIVEN

The way I learned

data science

- 🎓 Study the algorithms
- 🛠️ Implement the algorithms
- 💪 Practice on toy datasets

kaggle

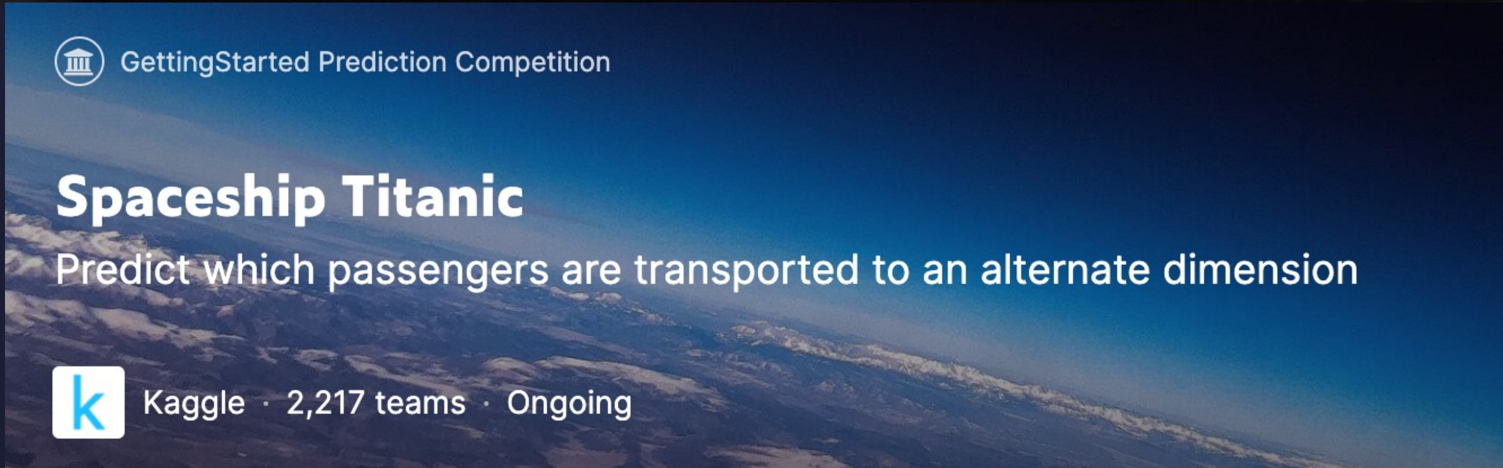


GettingStarted Prediction Competition

Titanic - Machine Learning from Disaster

Start here! Predict survival on the Titanic and get familiar with ML basics

Kaggle · 14,027 teams · Ongoing



GettingStarted Prediction Competition

Spaceship Titanic

Predict which passengers are transported to an alternate dimension

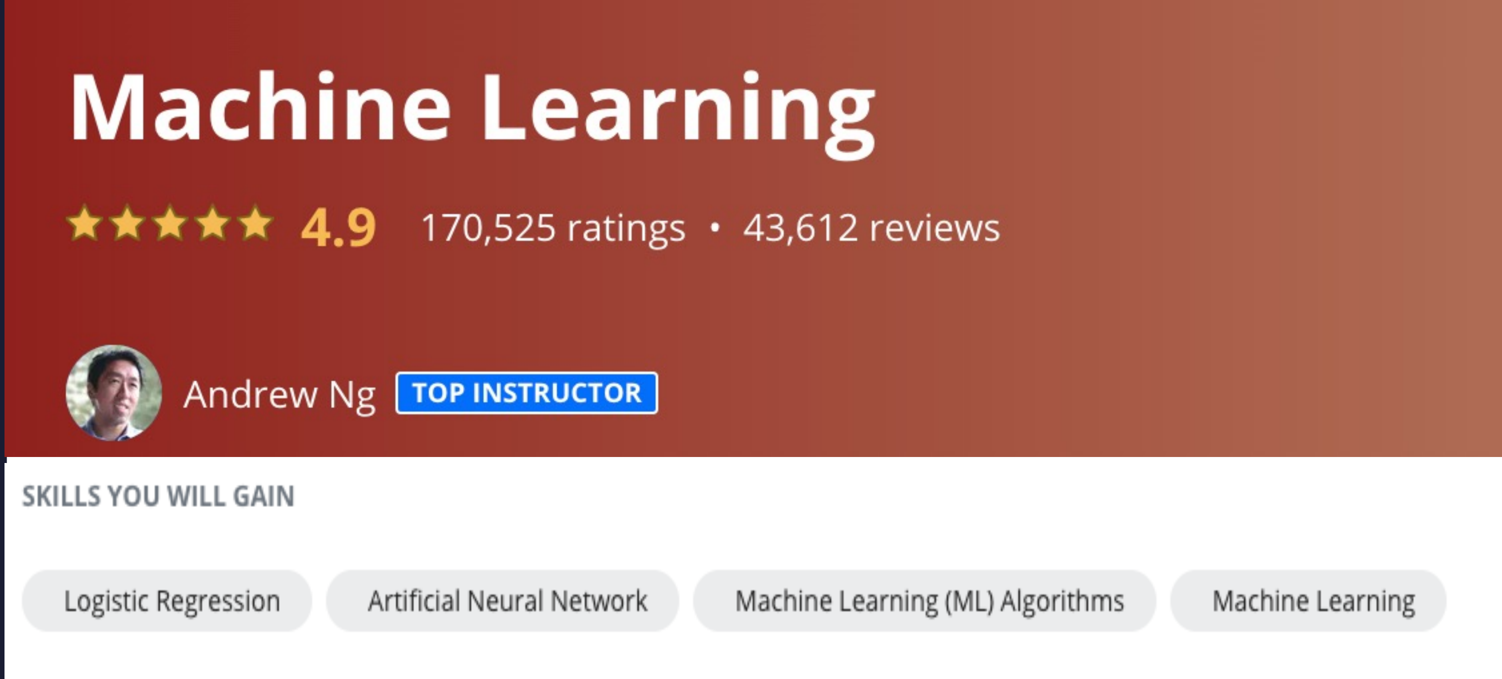
Kaggle · 2,217 teams · Ongoing

GO
DATA
DRIVEN

The way *many people* learn

data science

- 🎓 Study the algorithms
- 🛠️ Apply the algorithms
- 💪 Practice on toy datasets



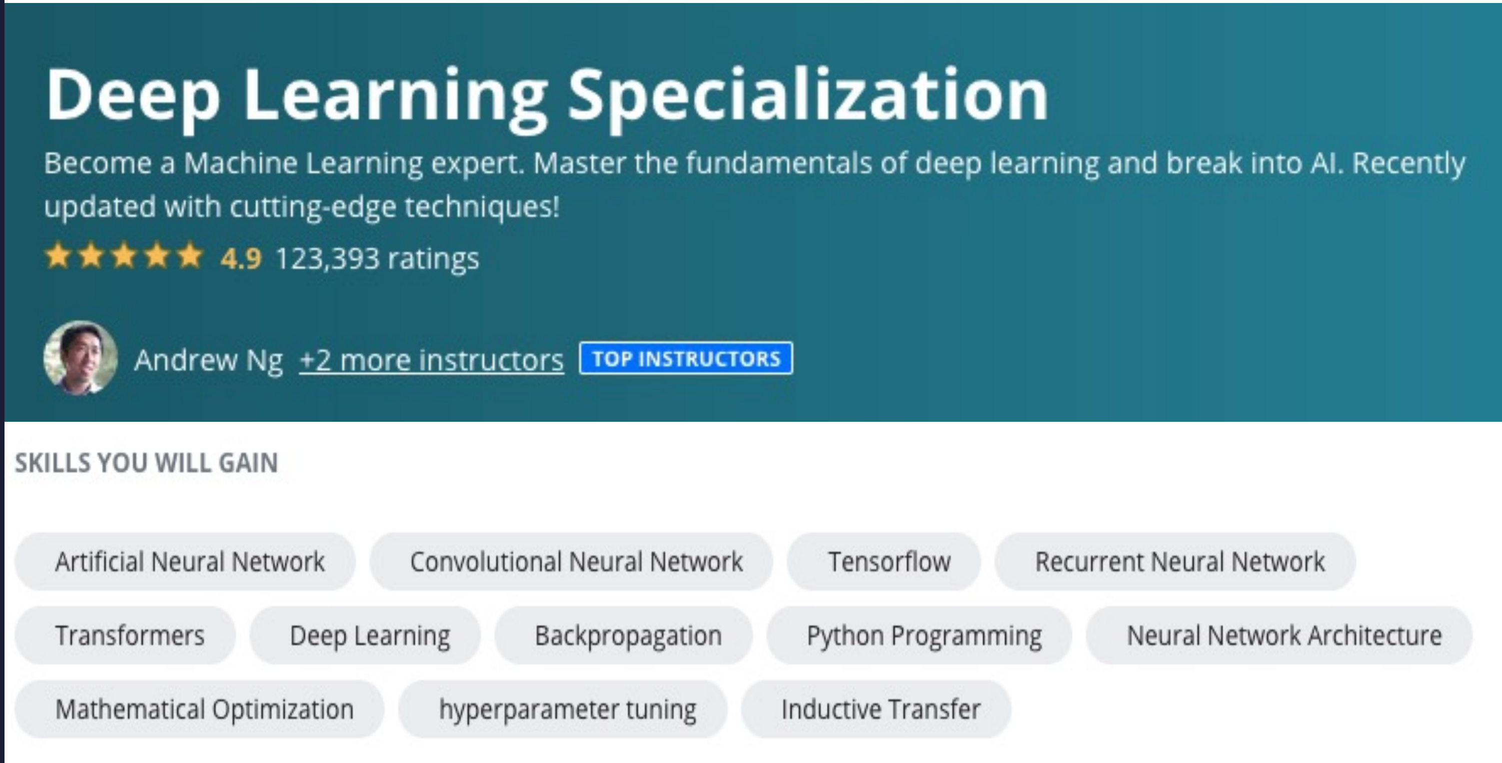
Machine Learning

★★★★★ 4.9 170,525 ratings • 43,612 reviews

Andrew Ng TOP INSTRUCTOR

SKILLS YOU WILL GAIN

Logistic Regression | Artificial Neural Network | Machine Learning (ML) Algorithms | Machine Learning



Deep Learning Specialization

Become a Machine Learning expert. Master the fundamentals of deep learning and break into AI. Recently updated with cutting-edge techniques!

★★★★★ 4.9 123,393 ratings

Andrew Ng +2 more instructors TOP INSTRUCTORS

SKILLS YOU WILL GAIN

Artificial Neural Network | Convolutional Neural Network | Tensorflow | Recurrent Neural Network | Transformers | Deep Learning | Backpropagation | Python Programming | Neural Network Architecture | Mathematical Optimization | hyperparameter tuning | Inductive Transfer

GO
DATA
DRIVEN

If machine learning is *20% modelling* and *80% data prep*…

If machine learning is *20% modelling* and *80% data prep*…

…. why is *data prep* not taught?

Data scientists treat datasets as static

- 🧑‍🏫 What they learn in courses

coursera

education for evervone

Data scientists treat datasets as static

- 🧑‍🏫 What they learn in courses
- 🎲 It’s what most online competitions focus on

coursera
education for evervone

kaggle

Data scientists treat datasets as static

- 🧑‍🏫 What they learn in courses
- 🎲 It’s what most online competitions focus on
- 🎓 Because that’s what they do in academia

coursera
education for everyone

kaggle

IMAGENET

Data scientists treat datasets as static

- 🧑‍🏫 What they learn in courses
- 🎲 It’s what most online competitions focus on
- 🎓 Because that’s what they do in academia
- 🛠️It’s what most tools are being built for

coursera
education for evervone

kaggle

IMAGENET

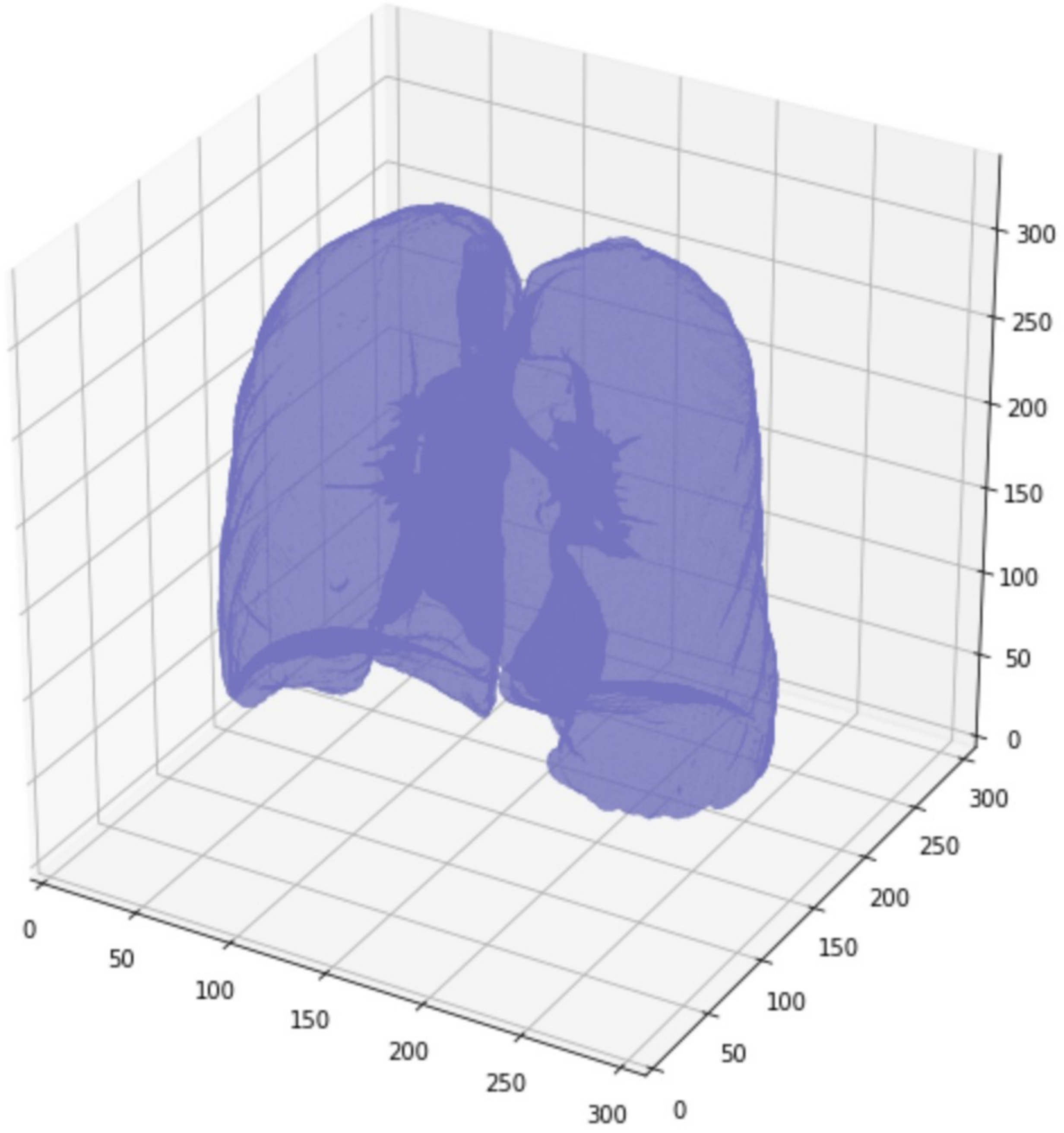
scikit
learn

GO
DATA
DRIVEN

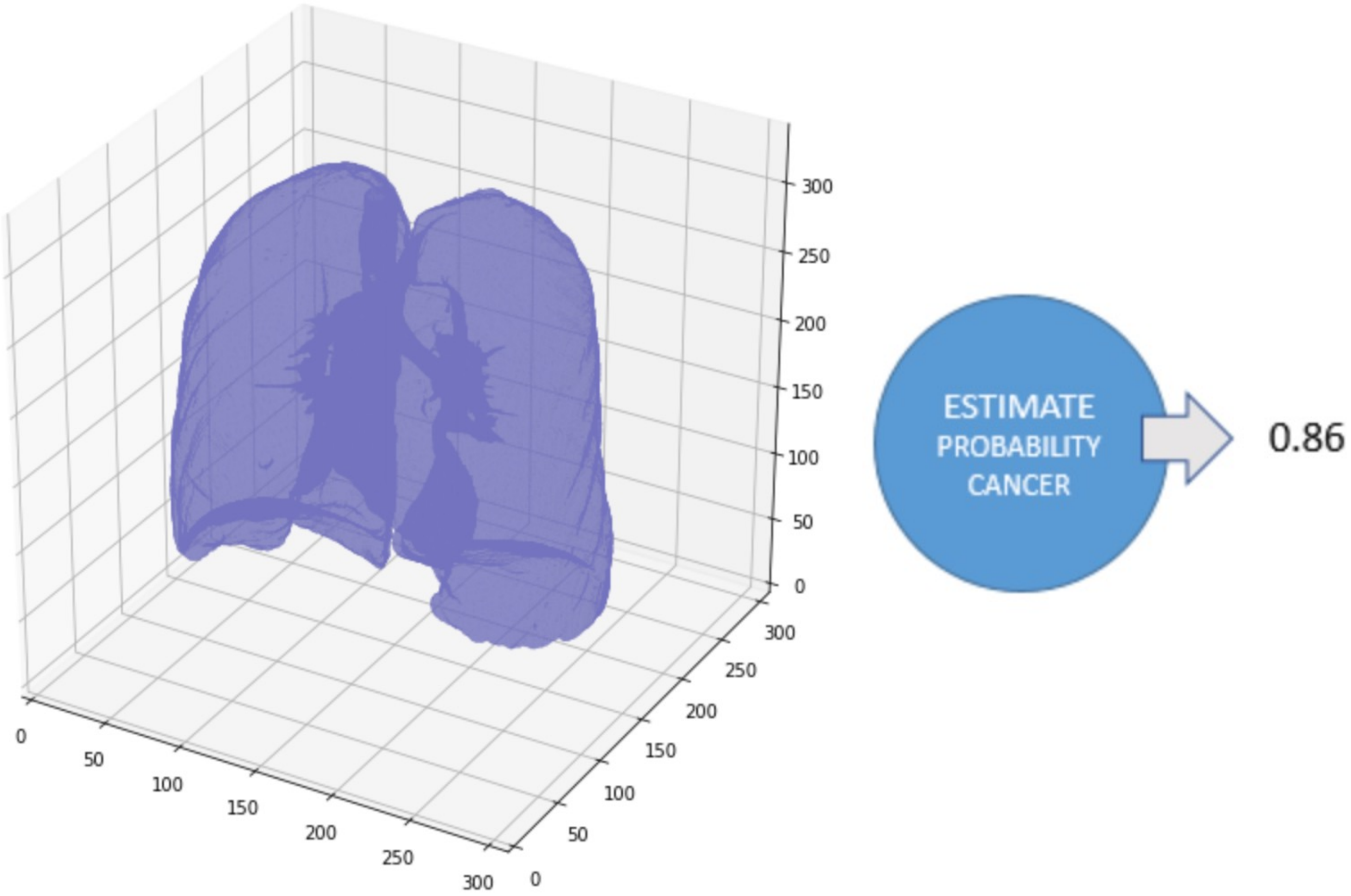
But datasets should not be *static*

GO
DATA
DRIVEN

Example: Data Science Bowl 2017



Example: Data Science Bowl 2017

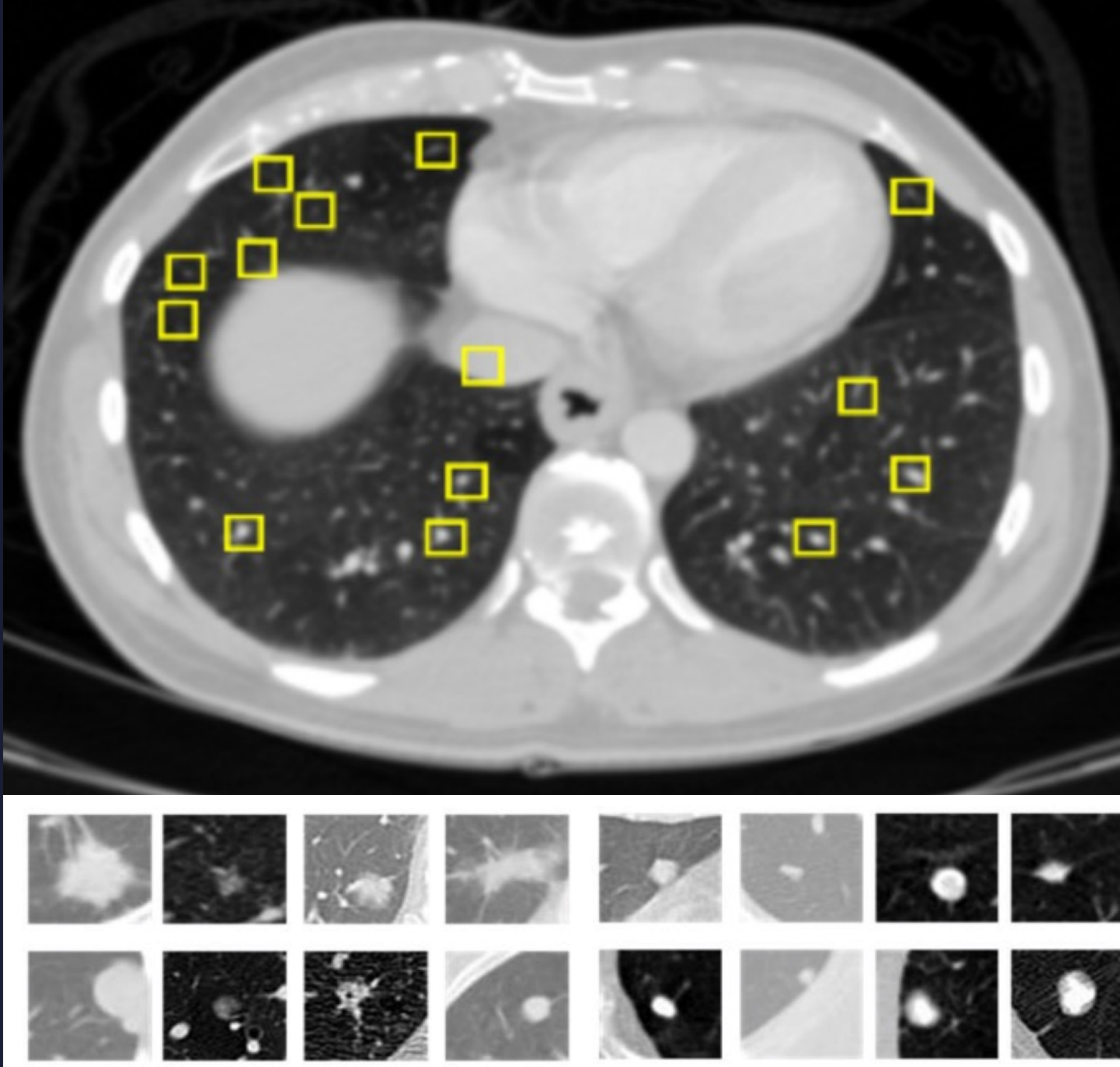


ESTIMATE PROBABILITY CANCER

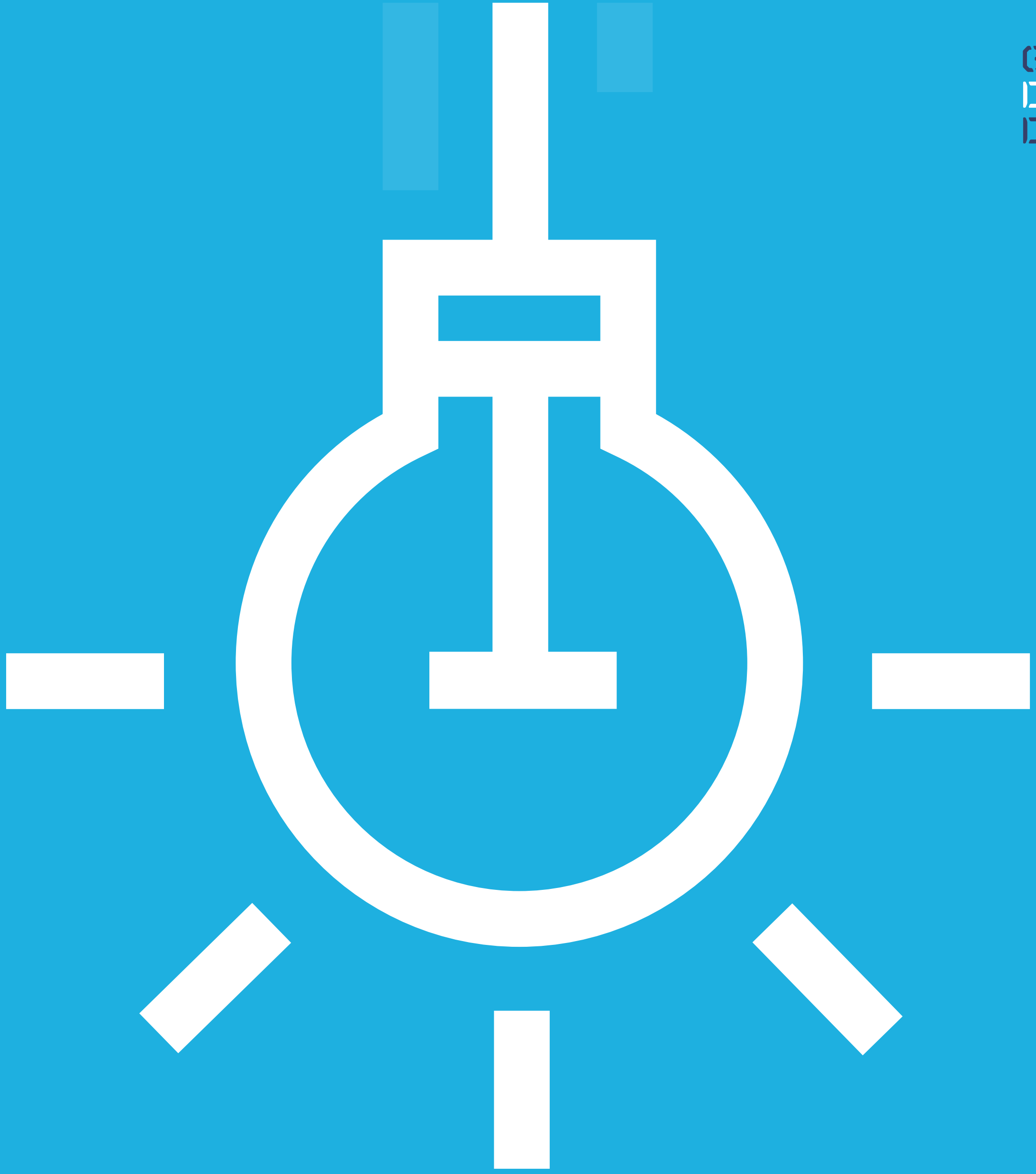
0.86

> *"For this solution, engineering the train set was an essential – if not **the** most essential – part."*
>
> *- Julian de Wit, 2nd place*

Example: Data Science Bowl 2017



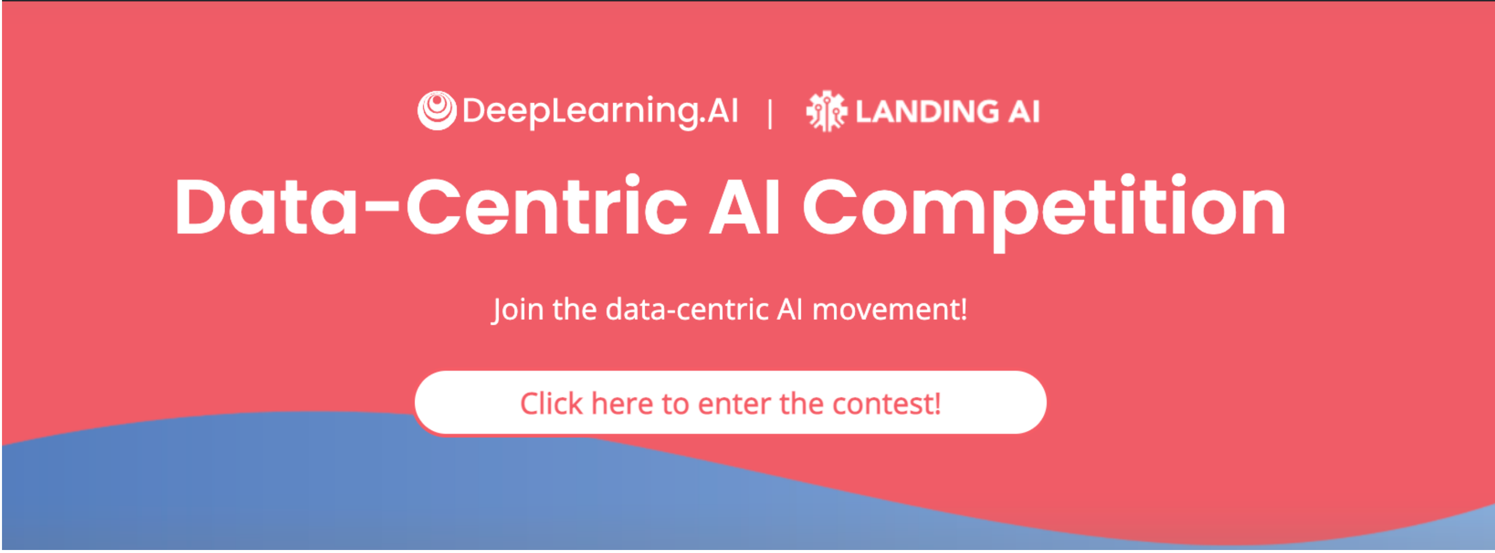
Data-Centric AI competition



Data-centric AI *is the discipline of systematically engineering the data used to build an* ***AI*** *system.*

Data-Centric AI competition

- Model is fixed (ResNet50)



DeepLearning.AI | LANDING AI

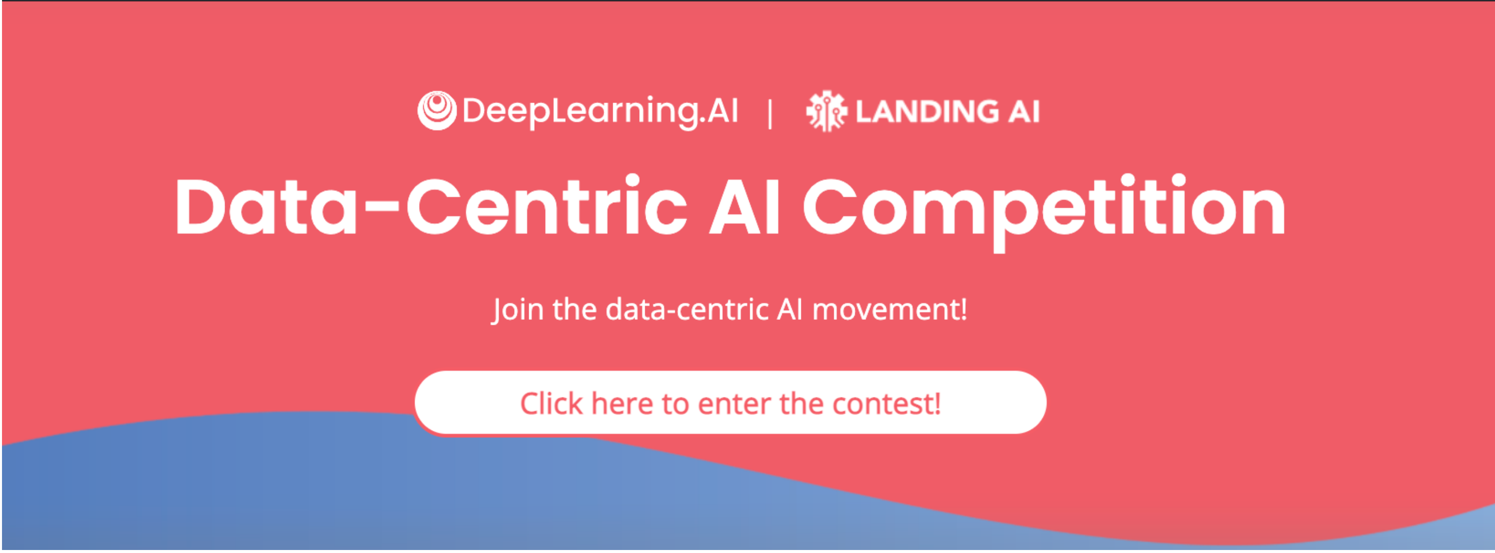
Data-Centric AI Competition

Join the data-centric AI movement!

GO
DATA
DRIVEN

Data-Centric AI competition

- Model is fixed (ResNet50)
- Roman numerals from 1 to 10

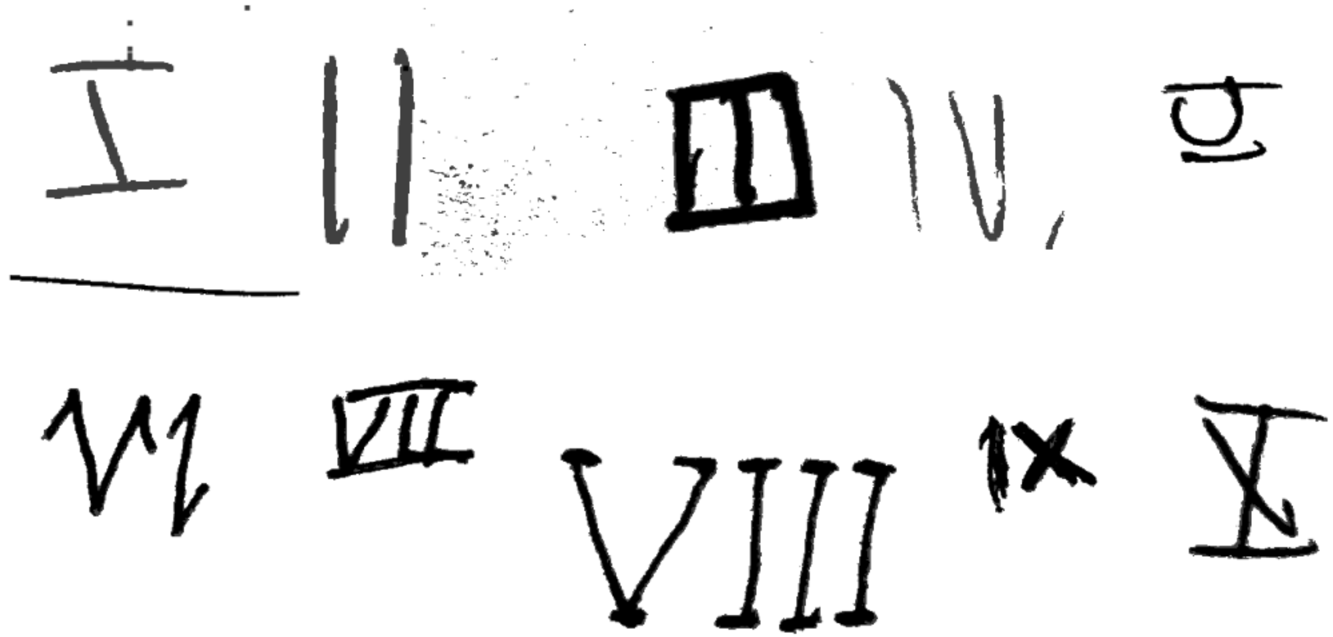


DeepLearning.AI | LANDING AI

Data-Centric AI Competition

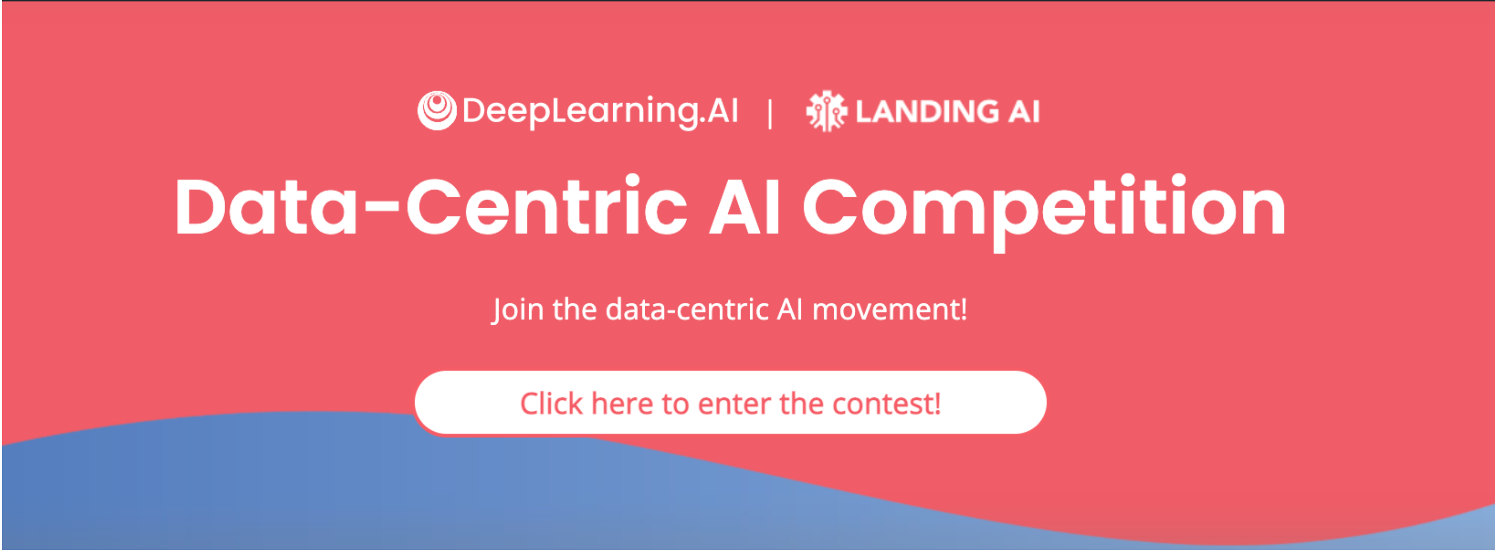
Join the data-centric AI movement!

Click here to enter the contest!



Data-Centric AI competition

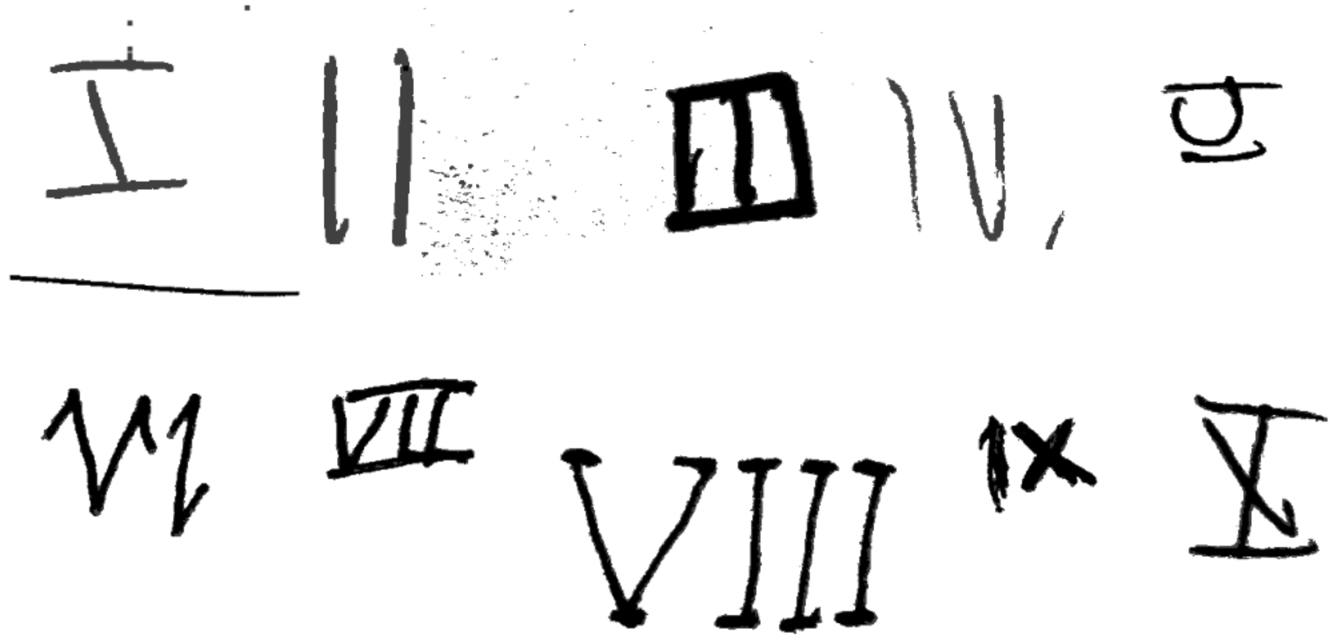
- Model is fixed (ResNet50)
- Roman numerals from 1 to 10
- 3K images in a train/validation set split



DeepLearning.AI | LANDING AI

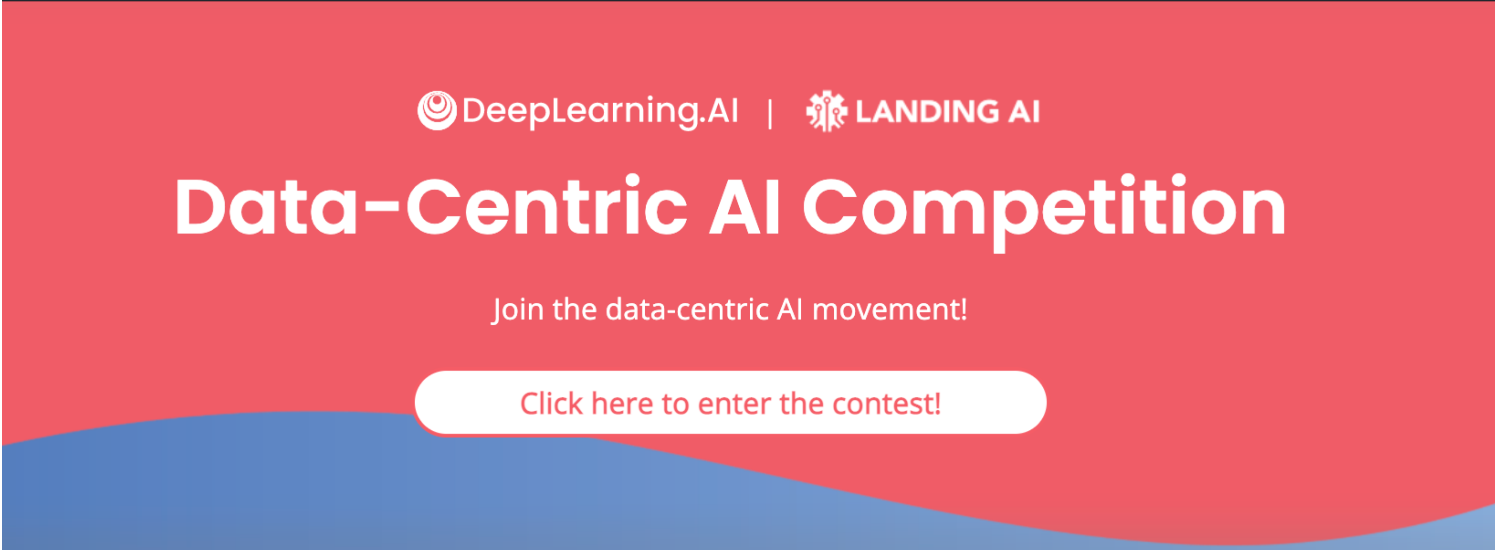
Data-Centric AI Competition

Join the data-centric AI movement!



Data-Centric AI competition

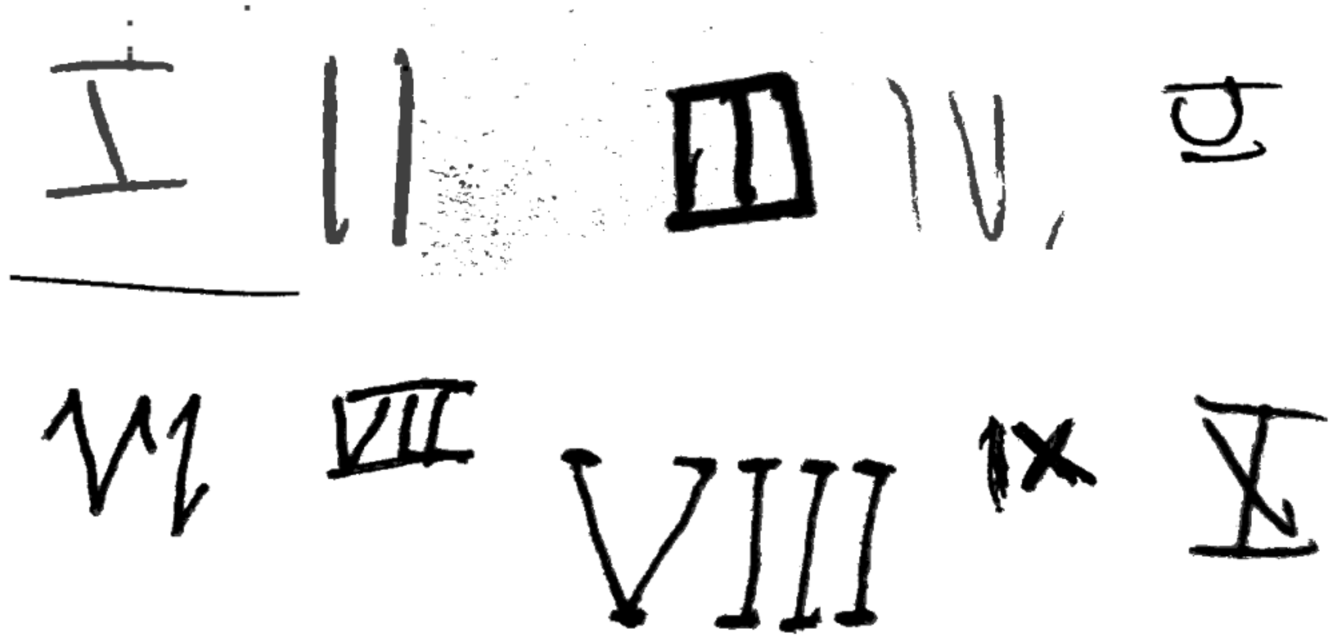
- Model is fixed (ResNet50)
- Roman numerals from 1 to 10
- 3K images in a train/validation set split
- Labelbook with to-be-expected examples of each class



DeepLearning.AI | LANDING AI

Data-Centric AI Competition

Join the data-centric AI movement!

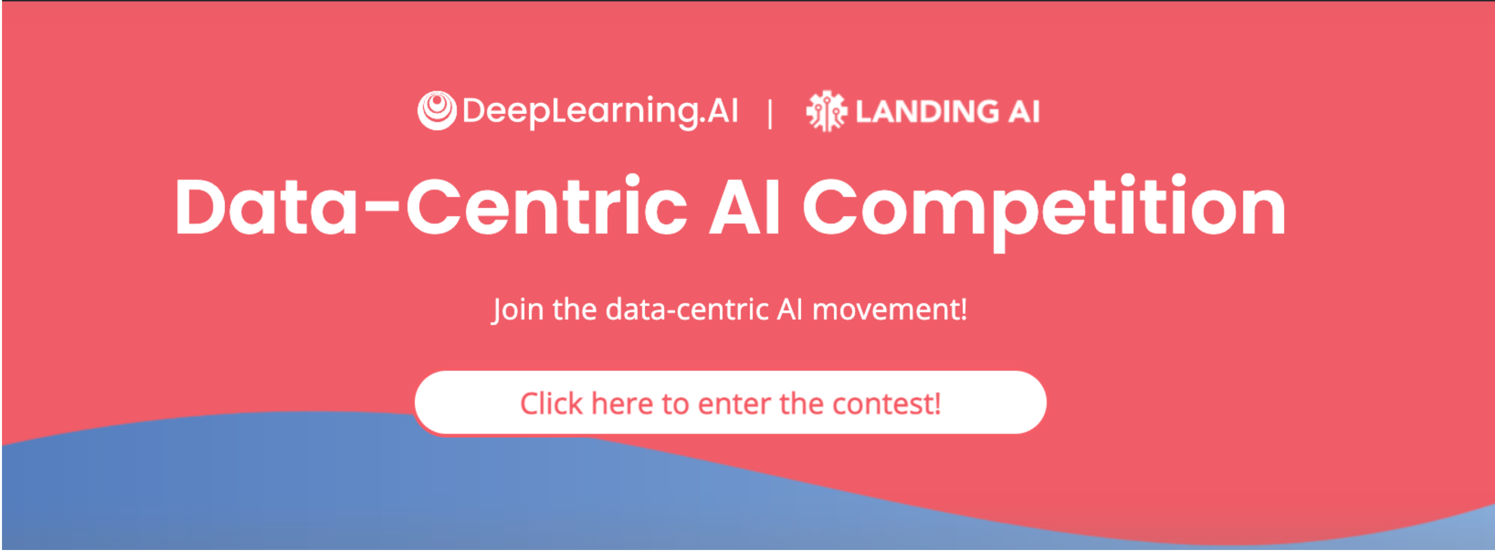


Data-Centric AI competition

- Model is fixed (ResNet50)
- Roman numerals from 1 to 10
- 3K images in a train/validation set split
- Labelbook with to-be-expected examples of each class

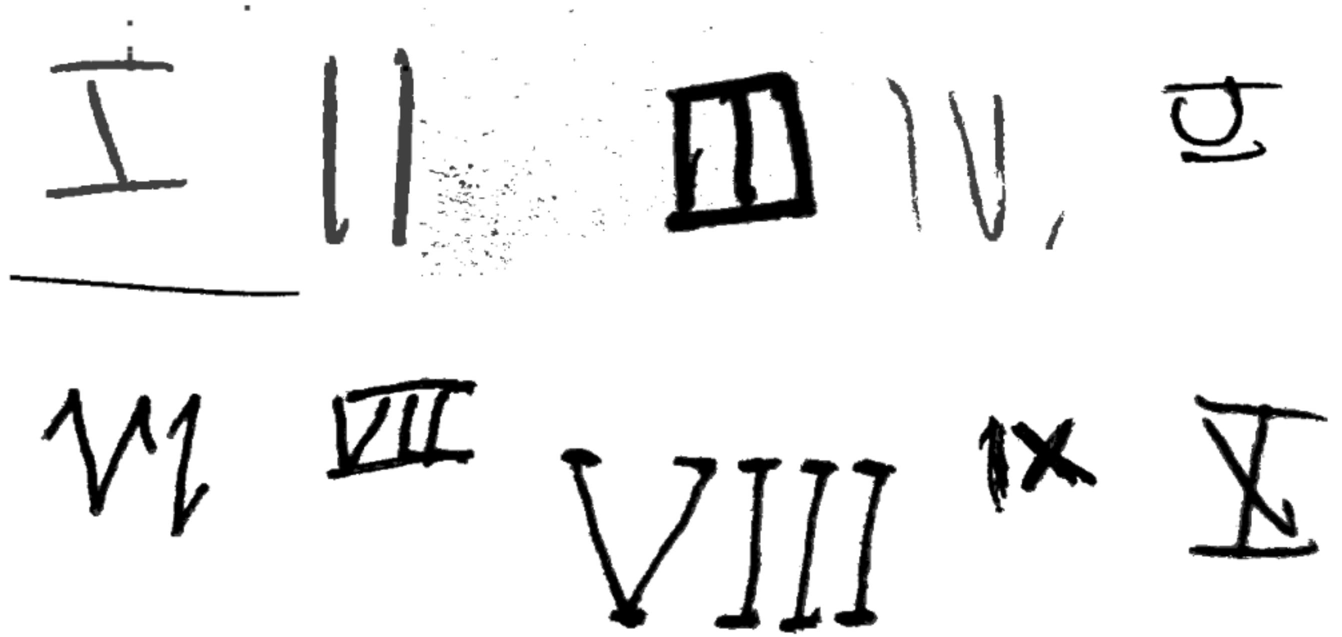
TASK

Enhance the dataset to a maximum of 10K images that maximizes the model accuracy on a hidden test set


DeepLearning.AI | LANDING AI

Data-Centric AI Competition

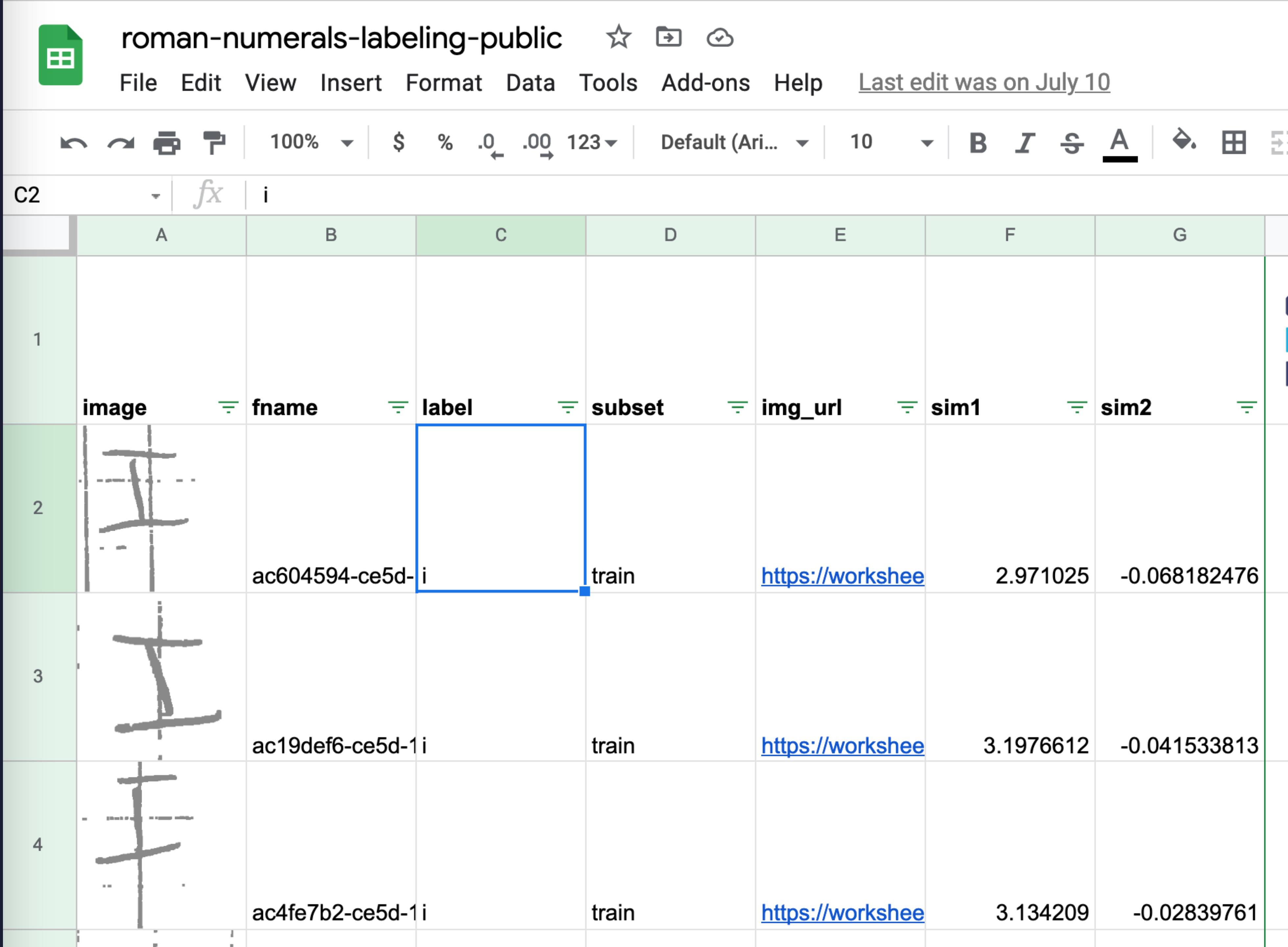
Join the data-centric AI movement!



OUR SOLUTION

Use low-tech tools to get started together

GO
DATA
DRIVEN



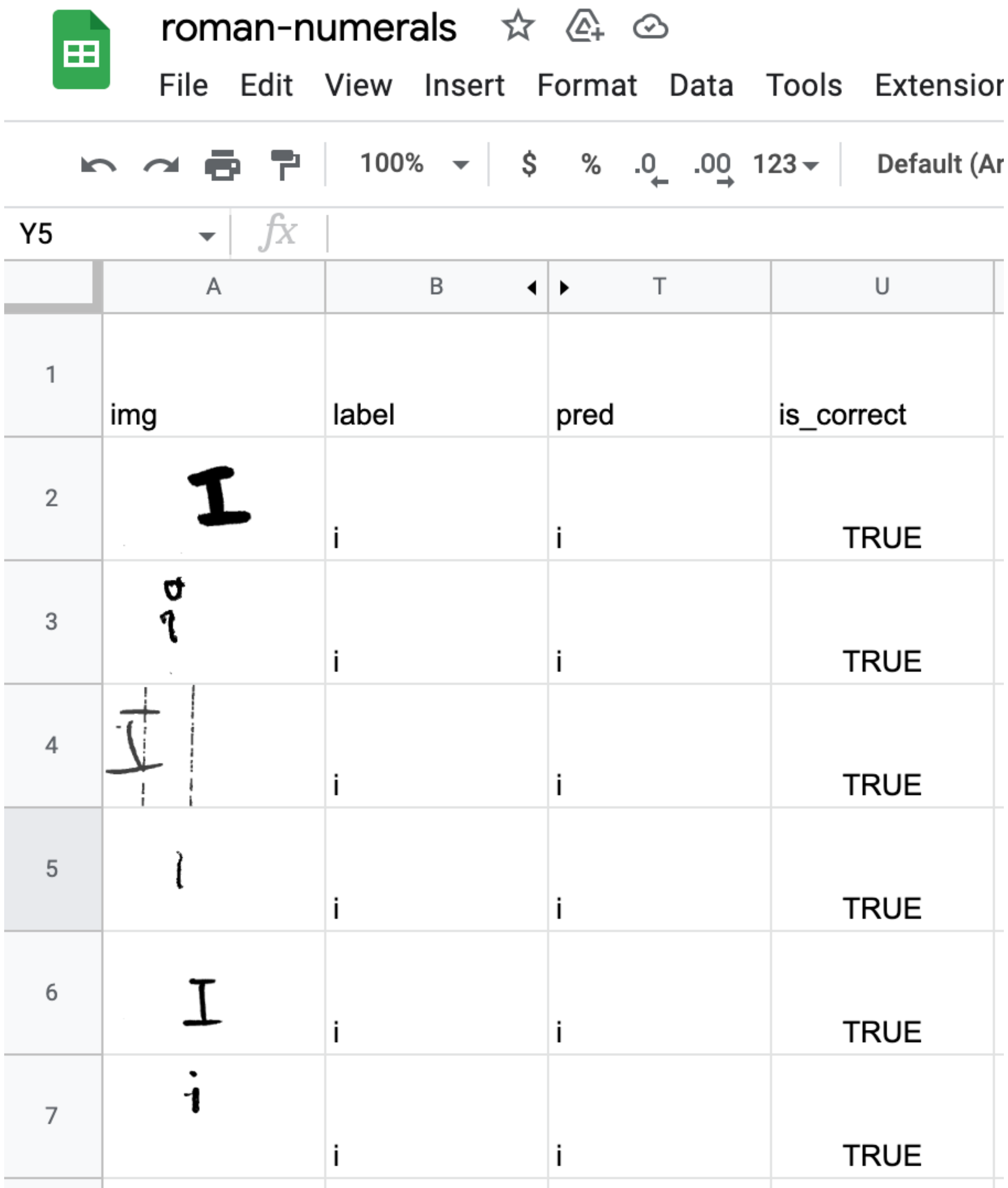
roman-numerals-labeling-public

C2 fx i

	A	B	C	D	E	F	G
1	image	fname	label	subset	img_url	sim1	sim2
2		ac604594-ce5d-	i	train	https://workshee	2.971025	-0.068182476
3		ac19def6-ce5d-1	i	train	https://workshee	3.1976612	-0.041533813
4		ac4fe7b2-ce5d-1	i	train	https://workshee	3.134209	-0.02839761

Improve label quality

1. Get predictions from baseline model



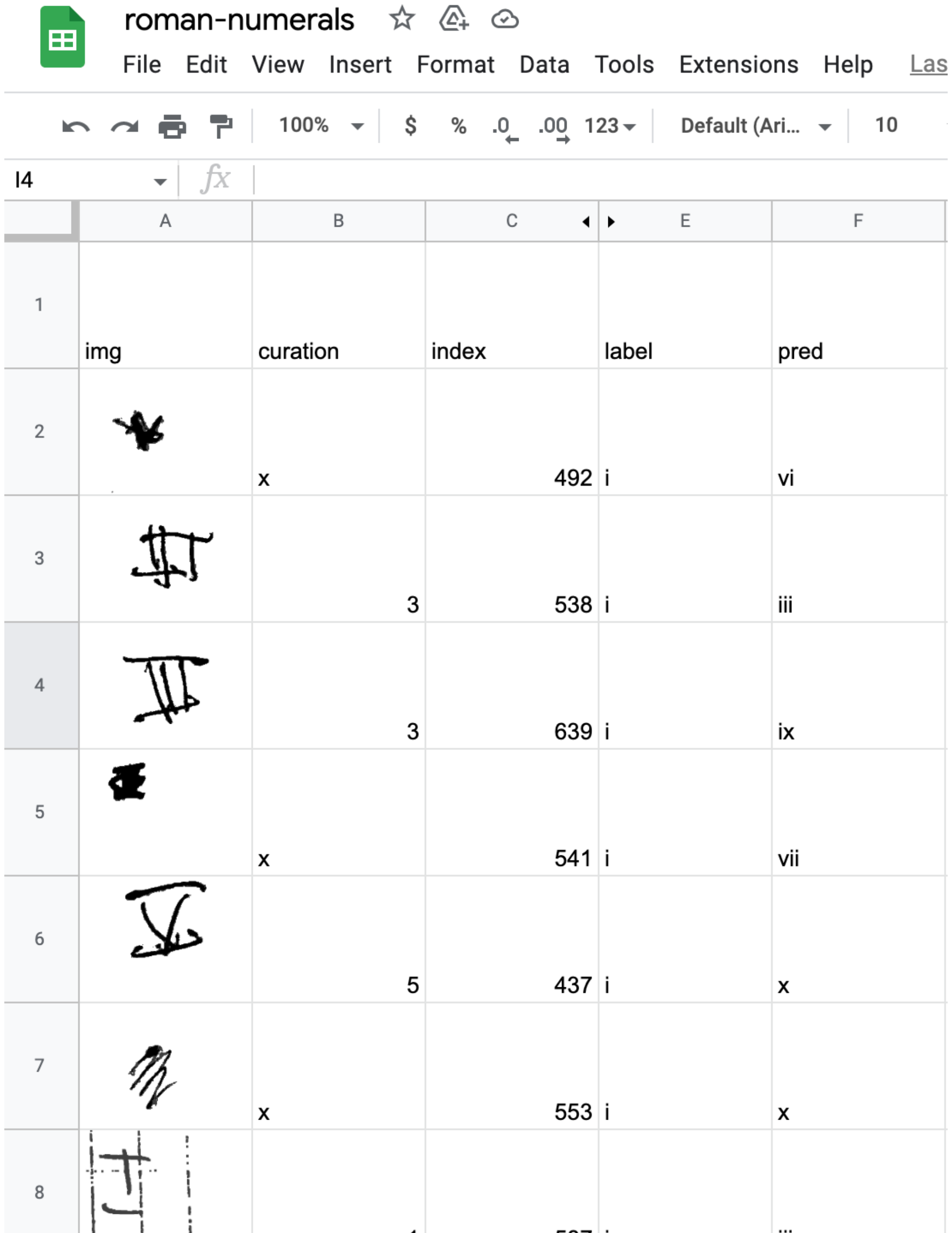
roman-numerals
File Edit View Insert Format Data Tools Extensior
100% $ % .0 .00 123 Default (Ar
Y5

	A	B	T	U
1	img	label	pred	is_correct
2		i	i	TRUE
3		i	i	TRUE
4		i	i	TRUE
5		i	i	TRUE
6		i	i	TRUE
7		i	i	TRUE

GO DATA DRIVEN

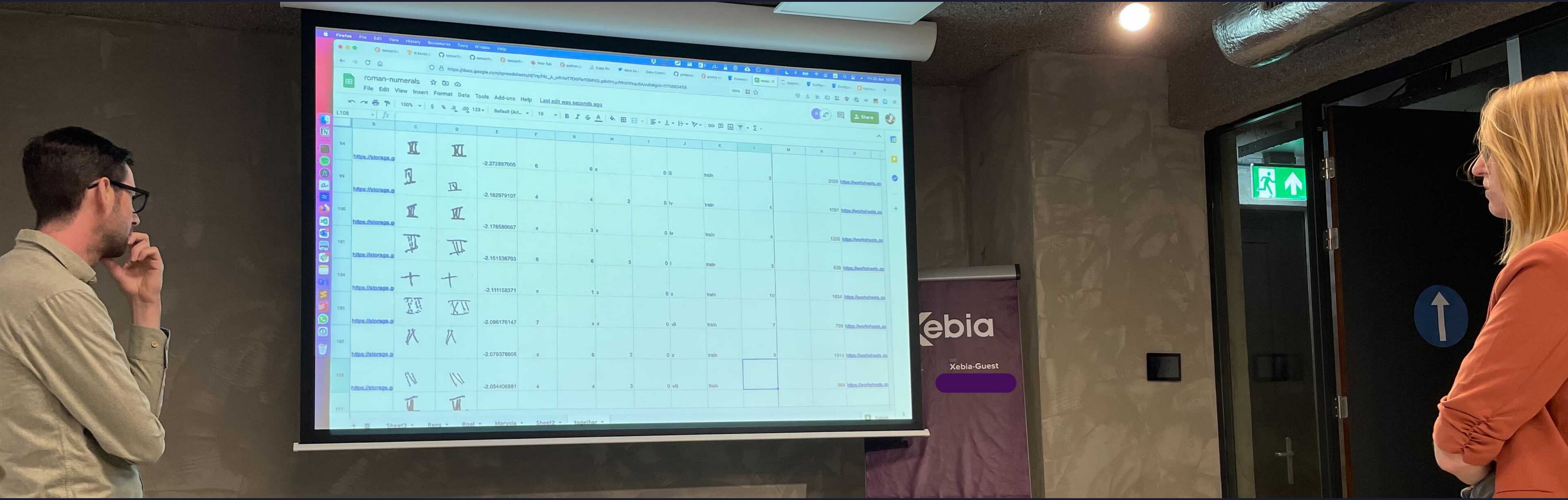
Improve label quality

1. Get predictions from baseline model
2. Focus on discrepancies between the model and the ground truth



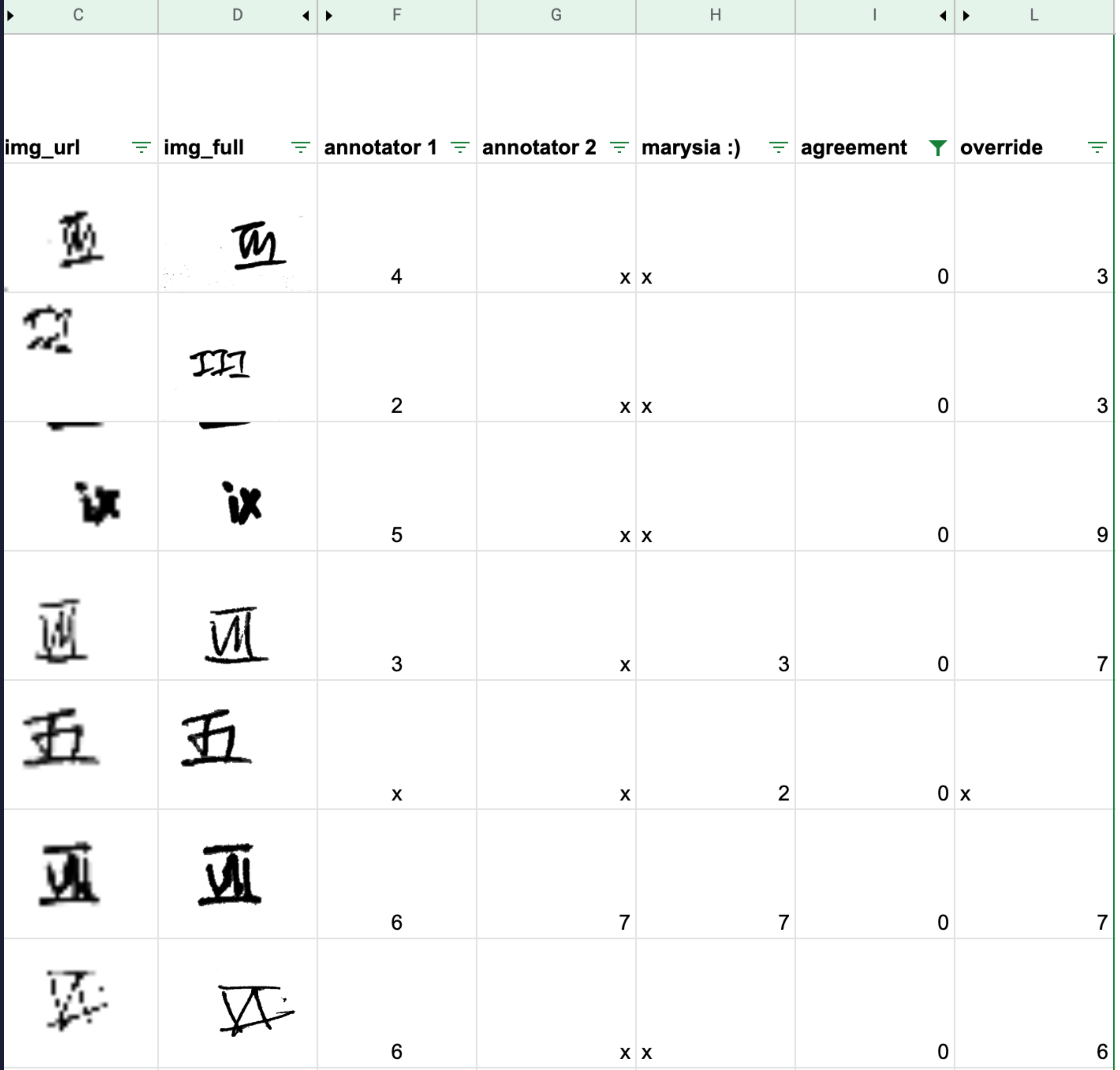
	A	B	C	E	F
1	img	curation	index	label	pred
2		x	492	i	vi
3		3	538	i	iii
4		3	639	i	ix
5		x	541	i	vii
6		5	437	i	x
7		x	553	i	x
8		1	[illegible]	i	iii

GO
DATA
DRIVEN


roman-numerals
Xebia
Xebia-Guest

Improve label quality

1. Get predictions from baseline model
2. Focus on discrepancies between the model and the ground truth
3. Individually annotate and create annotator consensus



img_url	img_full	annotator 1	annotator 2	marysia :)	agreement	override
		4	x	x	0	3
		2	x	x	0	3
		5	x	x	0	9
		3	x	3	0	7
		x	x	2	0	x
		6	7	7	0	7
		6	x	x	0	6

LESSON ONE

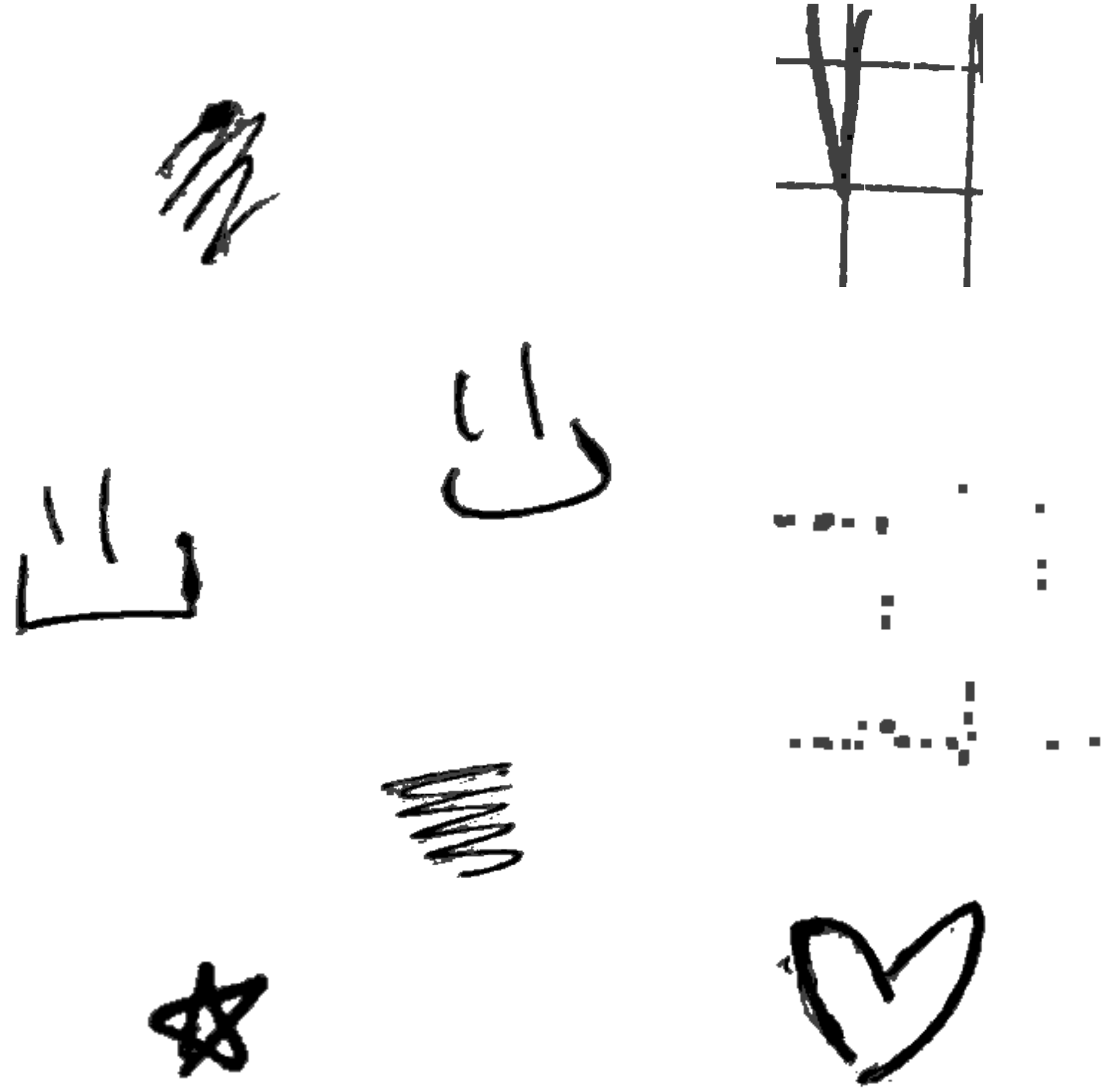
The labeling IS the learning

GO DATA DRIVEN

Lessons learnt

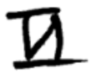
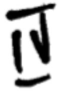
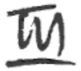
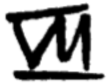
❌ Some data points simply needed to be removed

GO
DATA
DRIVEN



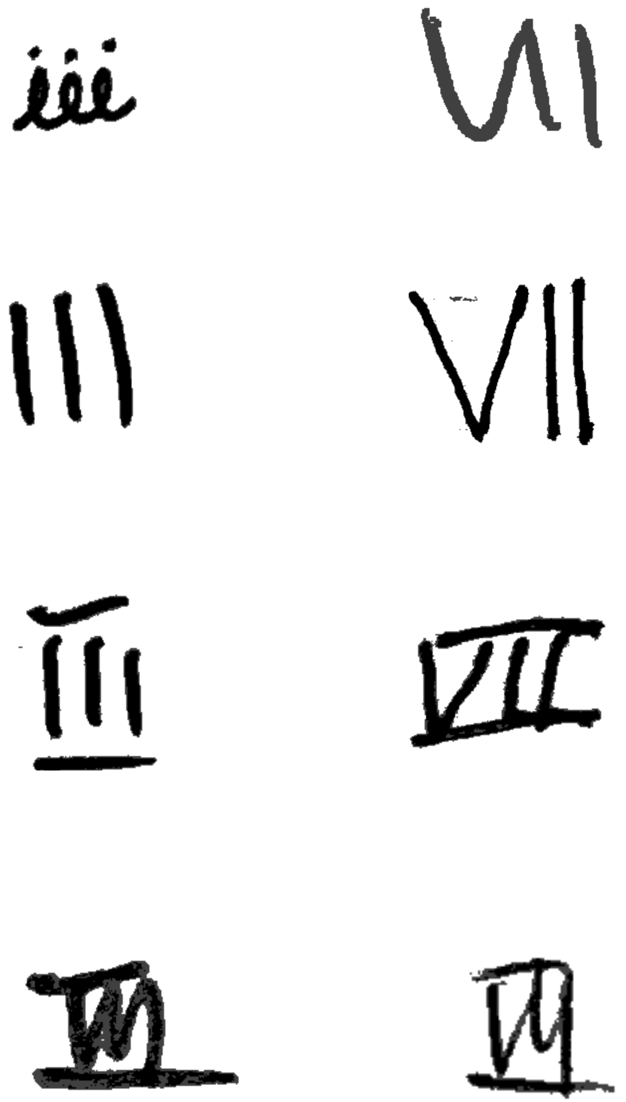
Lessons learnt

- ❌ Some data points simply needed to be removed
- 🙅 Lack of consensus between annotators was often about the same classes

	2	6	2
	4	2	2
	7	7	3
	7	3	7

Lessons learnt

- ❌ Some data points simply needed to be removed
- 🙅 Lack of consensus between annotators was often about the same classes
- ✍ Different styles of writing

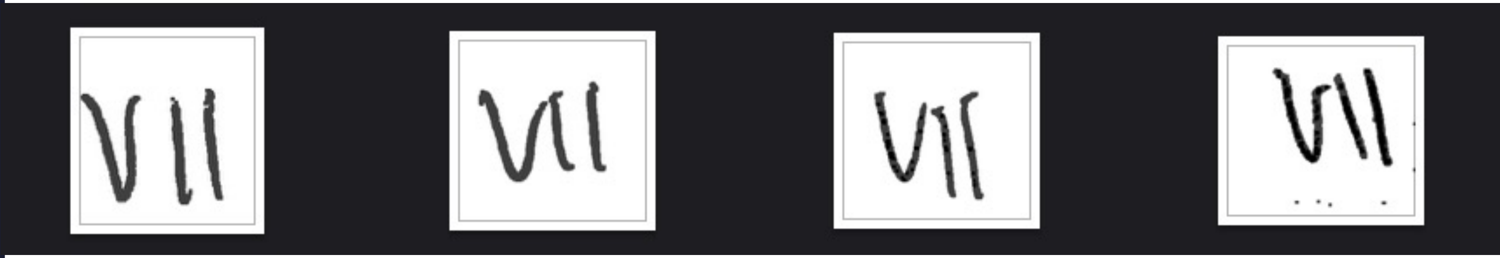


Lessons learnt

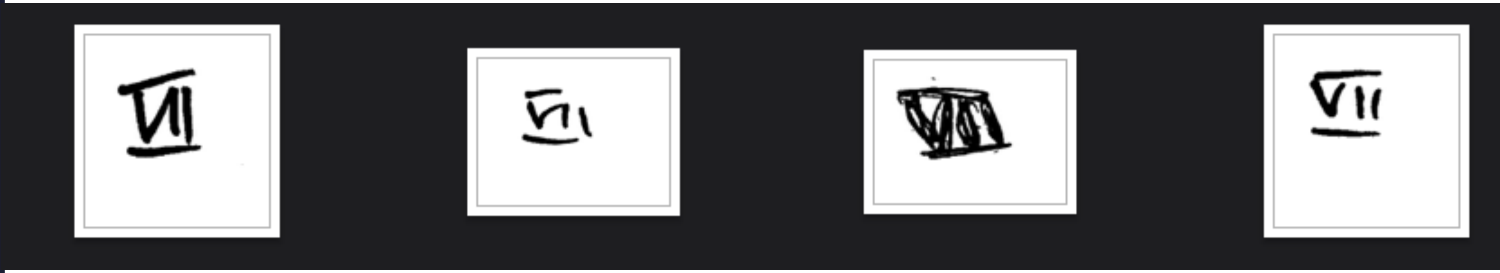
- ❌ Some data points simply needed to be removed
- 🙅 Lack of consensus between annotators was often about the same classes
- ✍️ Different styles of writing
- ⁉️ Difference in train/validation..?? set?!

GO DATA DRIVEN

Sample from training set



Sample from validation set



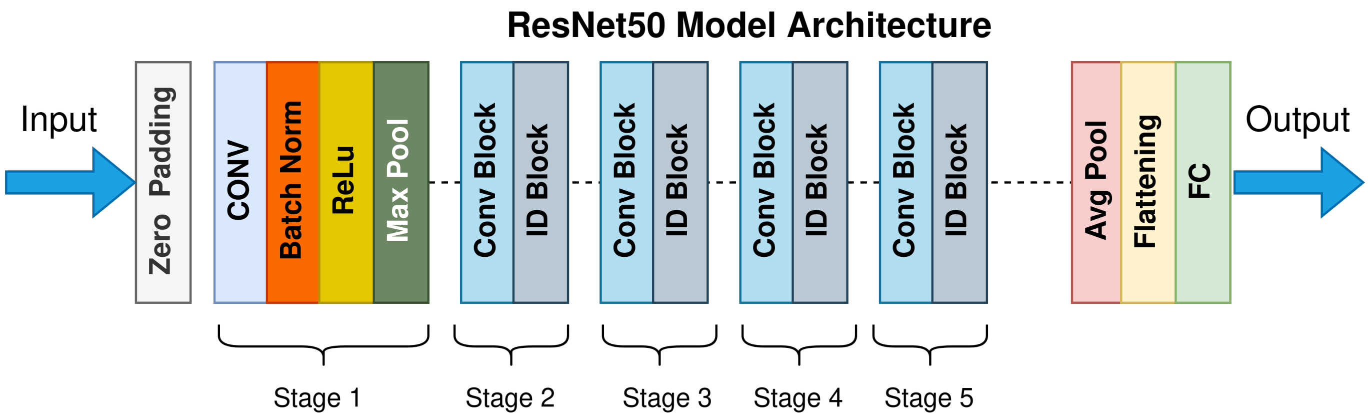
OUR SOLUTION

Use embeddings to get a sense of typicality and style

GO DATA DRIVEN

Visualize the data

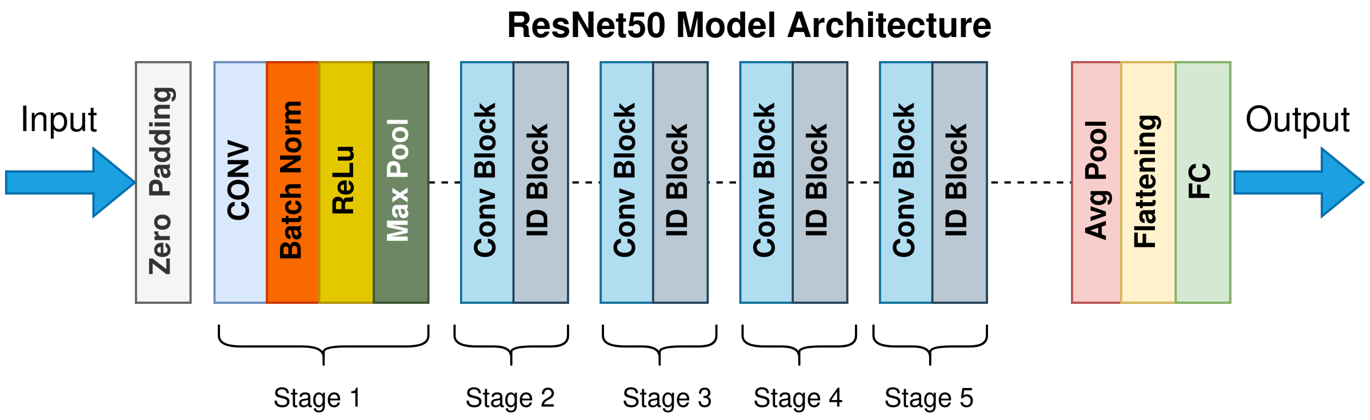
1. Pass all the data through the network to obtain the embeddings

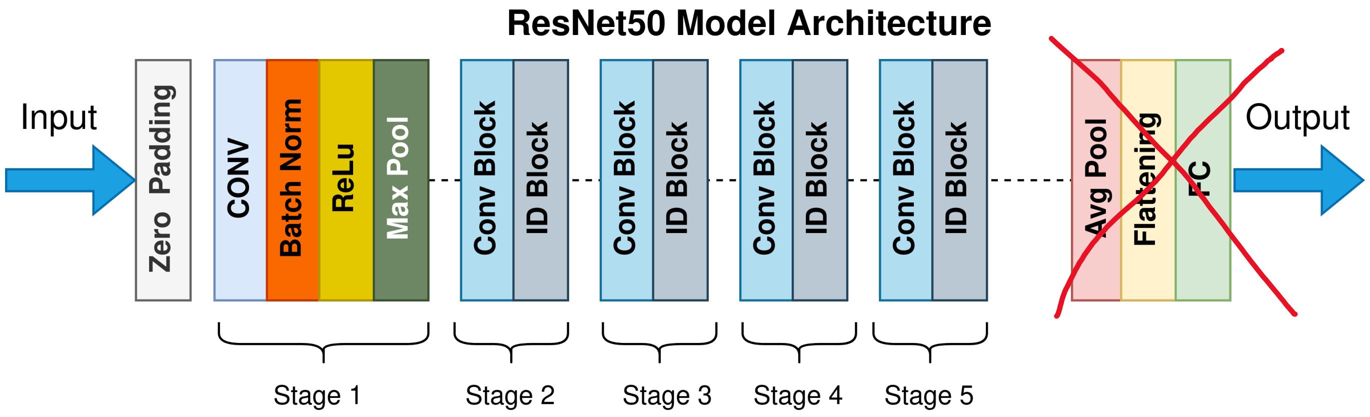


GO DATA DRIVEN

Visualize the data

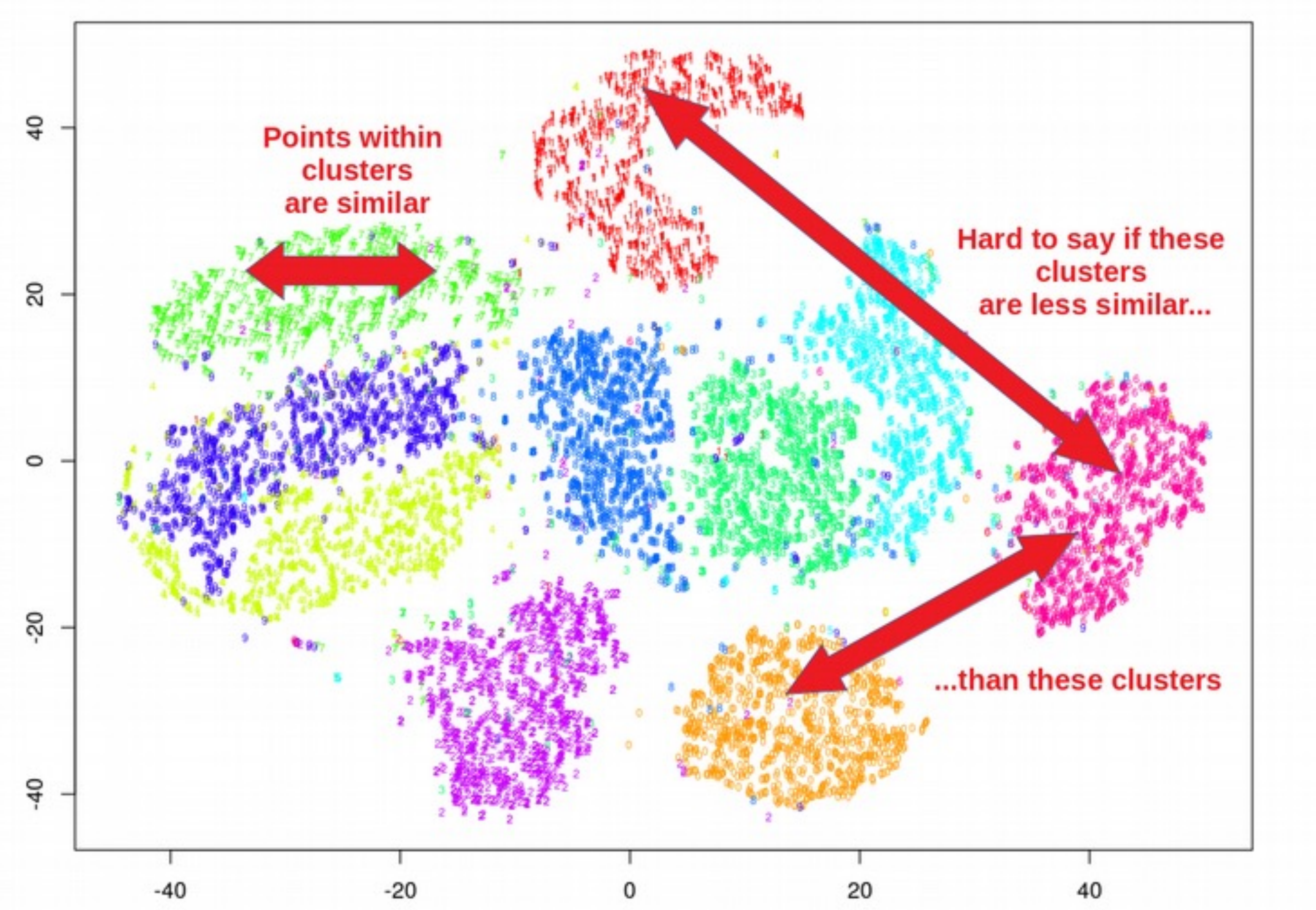
1. Pass all the data through the network to obtain the embeddings





Visualize the data

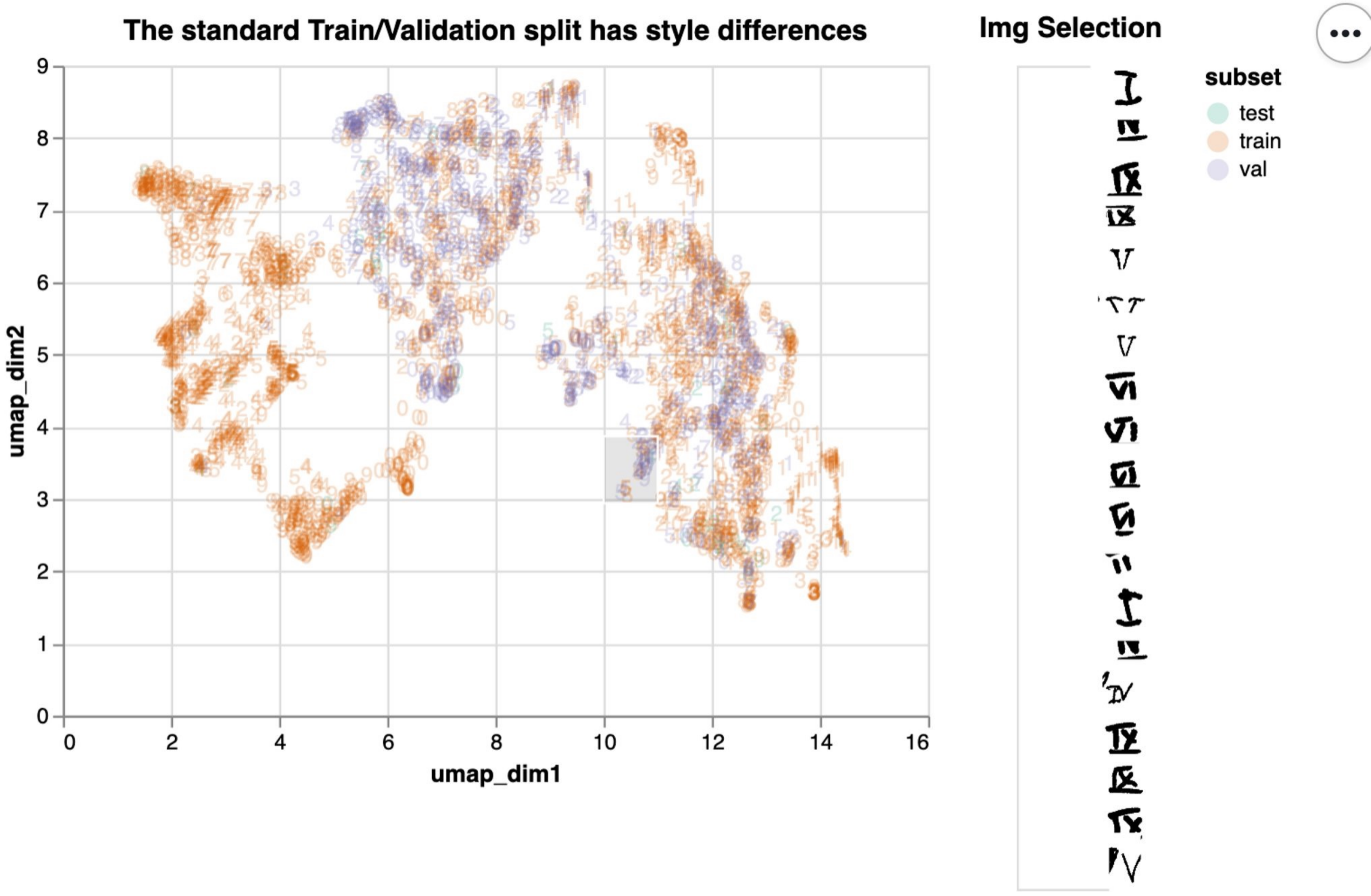
1. Pass all the data through the network to obtain the embeddings
2. Perform UMAP

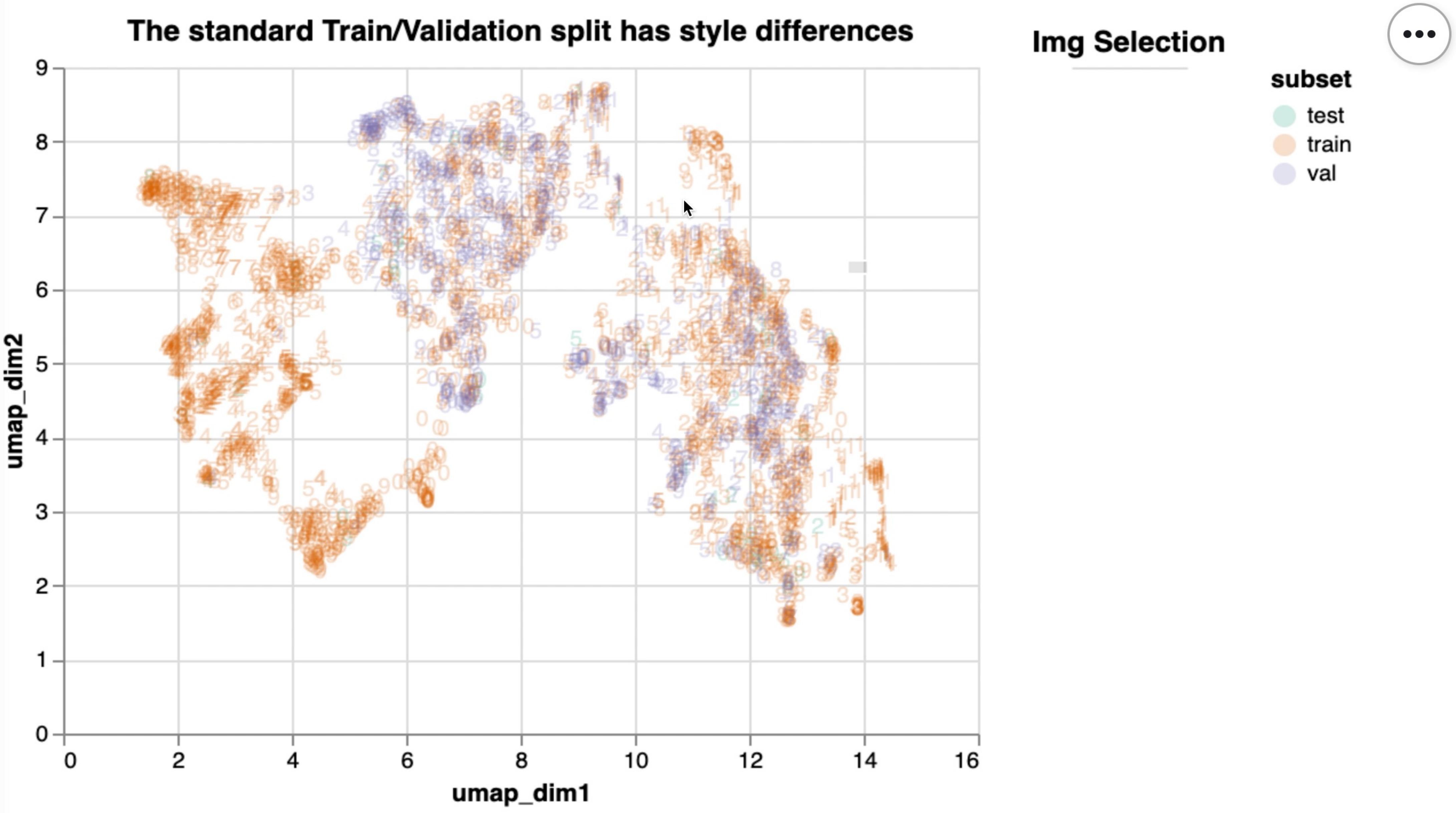


GO DATA DRIVEN

Visualize the data

1. Pass all the data through the network to obtain the embeddings
2. Perform UMAP
3. Visualize using interactive library

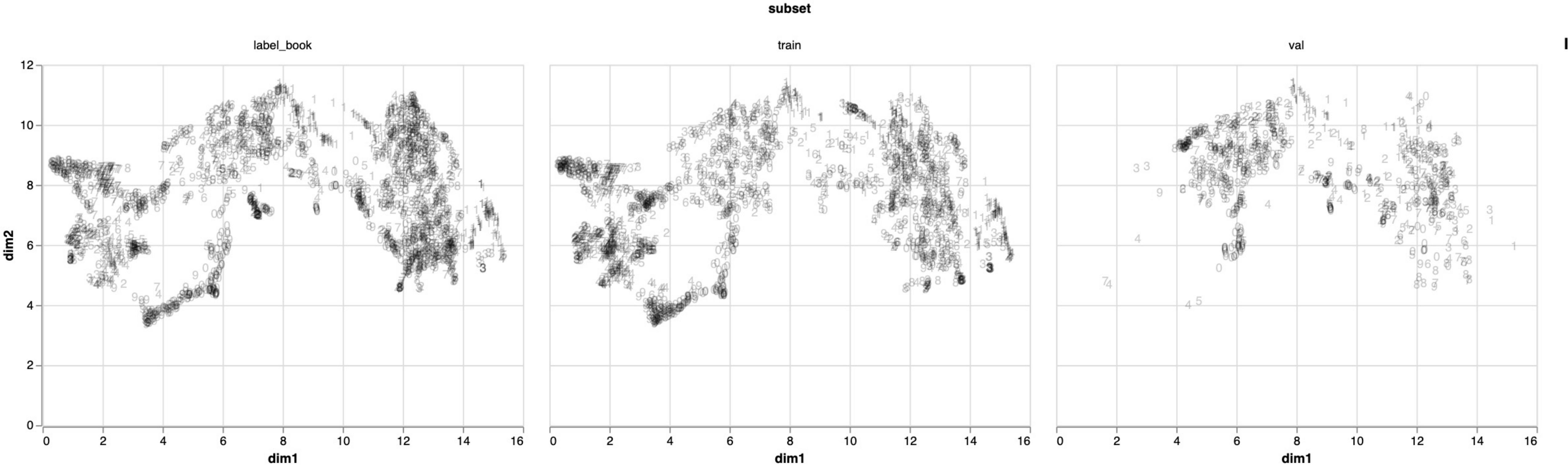




The standard Train/Validation split has style differences

Img Selection

Label book / Train / Validation / style differences.



GO DATA DRIVEN

LESSON TWO

Don’t be afraid to rebalance the *train/test* split

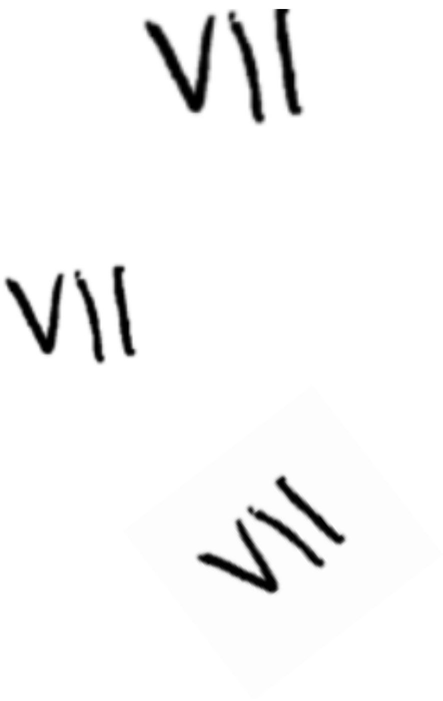
GO
DATA
DRIVEN

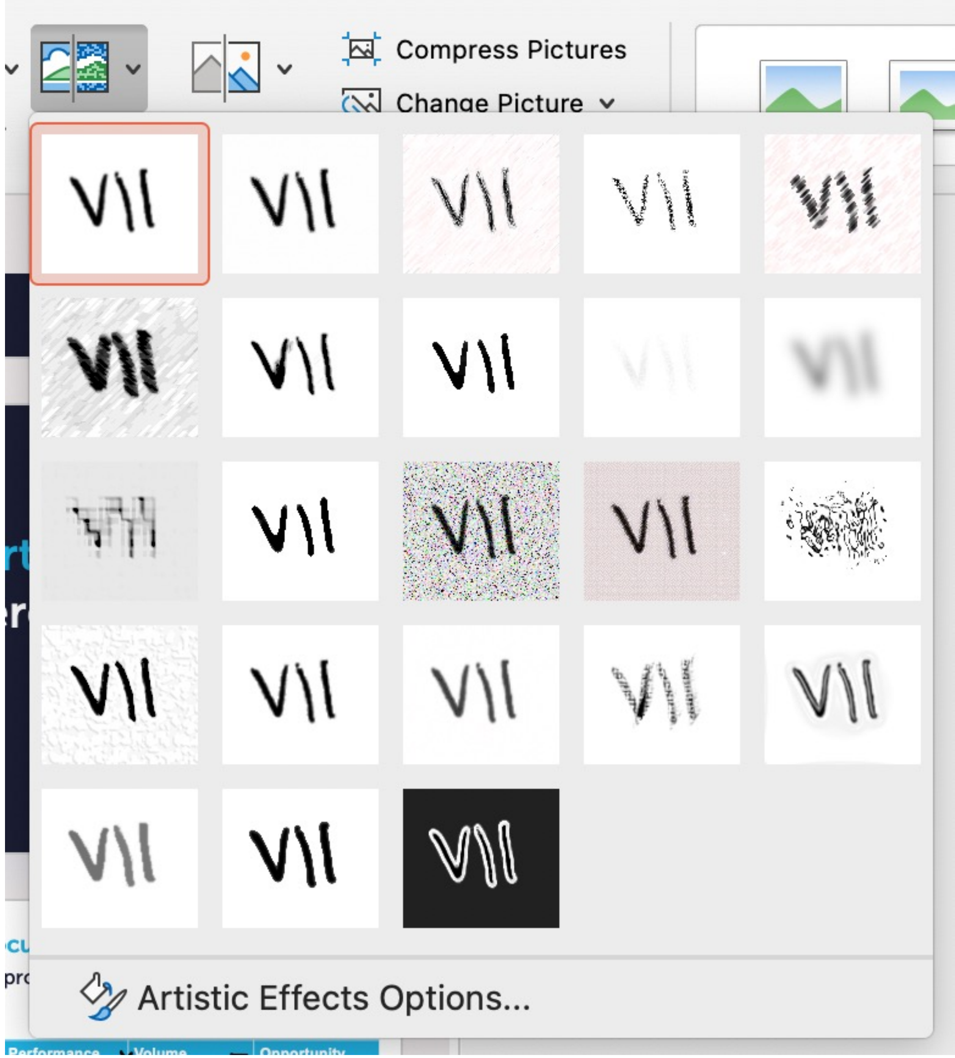
OUR SOLUTION

Use data augmentation to enrich your dataset

Augment your dataset

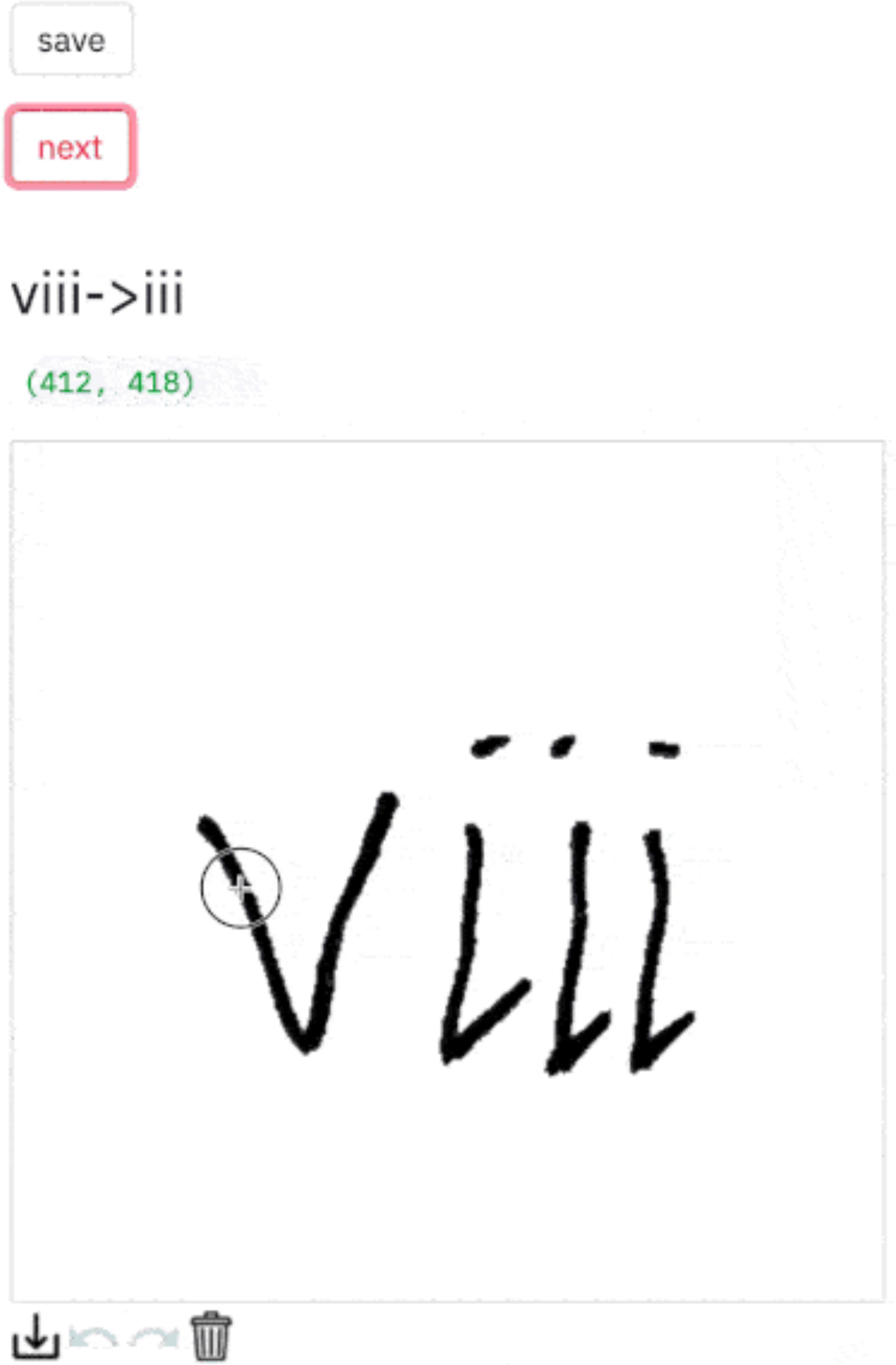
🎨 Transform existing data points to create augmented versions





Augment your dataset

- 🎨 Transform existing data points to create augmented versions
- 🧐 Create counterfactuals



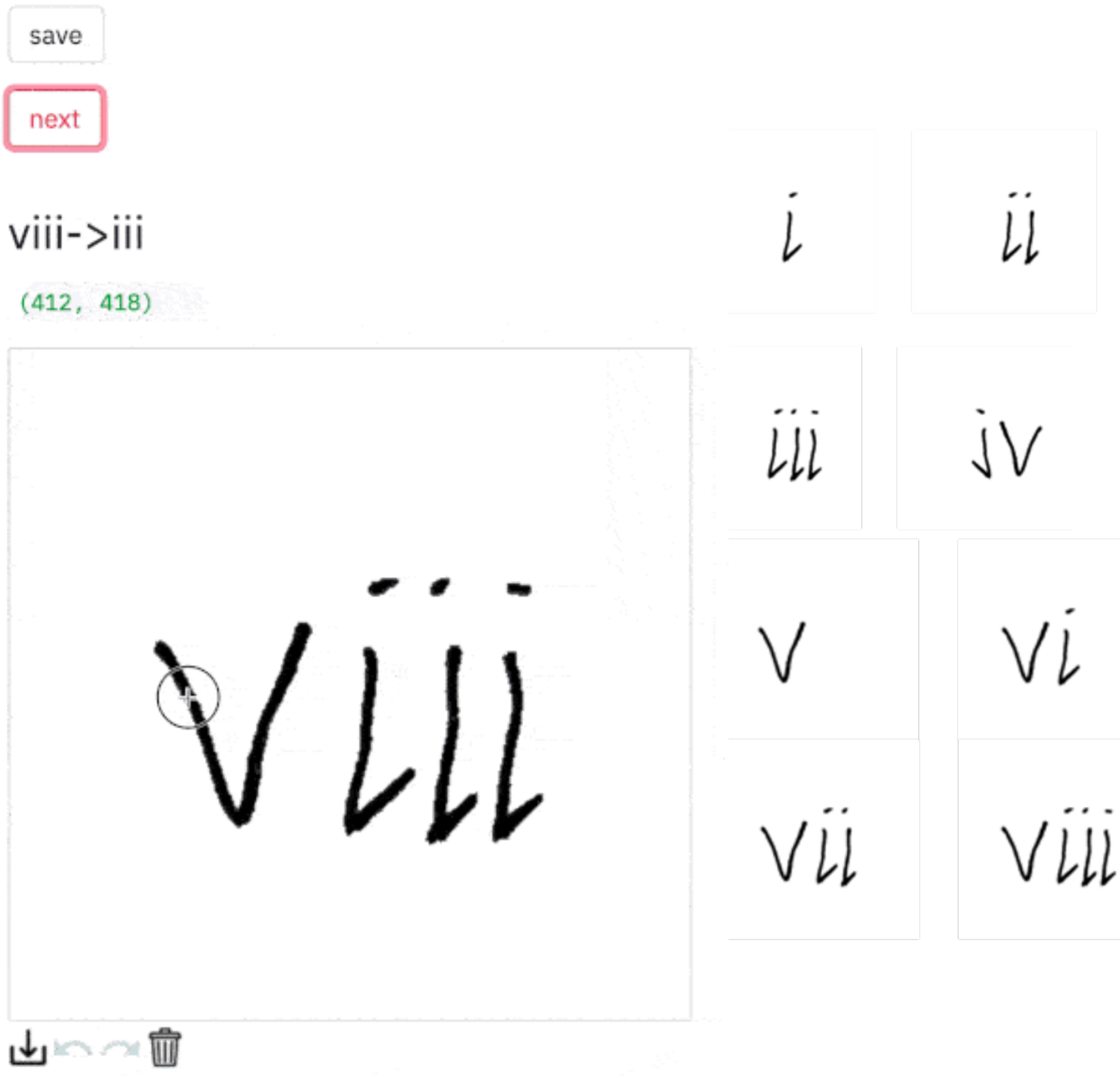
viii->iii

```
(412, 418)
```

GO
DATA
DRIVEN

Augment your dataset

1. Transform existing data points to create augmented versions
2. Create counterfactuals



LESSON THREE

Make it easy to quickly iterate over datasets

GO DATA DRIVEN

DeepLearning.AI | LANDING AI

Data-centric AI Competition

WINNER ANNOUNCEMENT

MOST INNOVATIVE

GoDataDriven

Roel Bertens

Marysia Winkels

Rens Dimmendaal

DeepLearning.AI | LANDING AI

Data-centric AI Competition

Winner Announcement

BEST PERFORMANCE

Innotescus

Divakar Roy

Shashank Deshpande

Chris Anderson

Rob Walsh

Synaptic-AnN

Asfandyar Azhar

Nidhish Shah

MOST INNOVATIVE

Mohammad Motamedi

Johnson Kuan

GoDataDriven

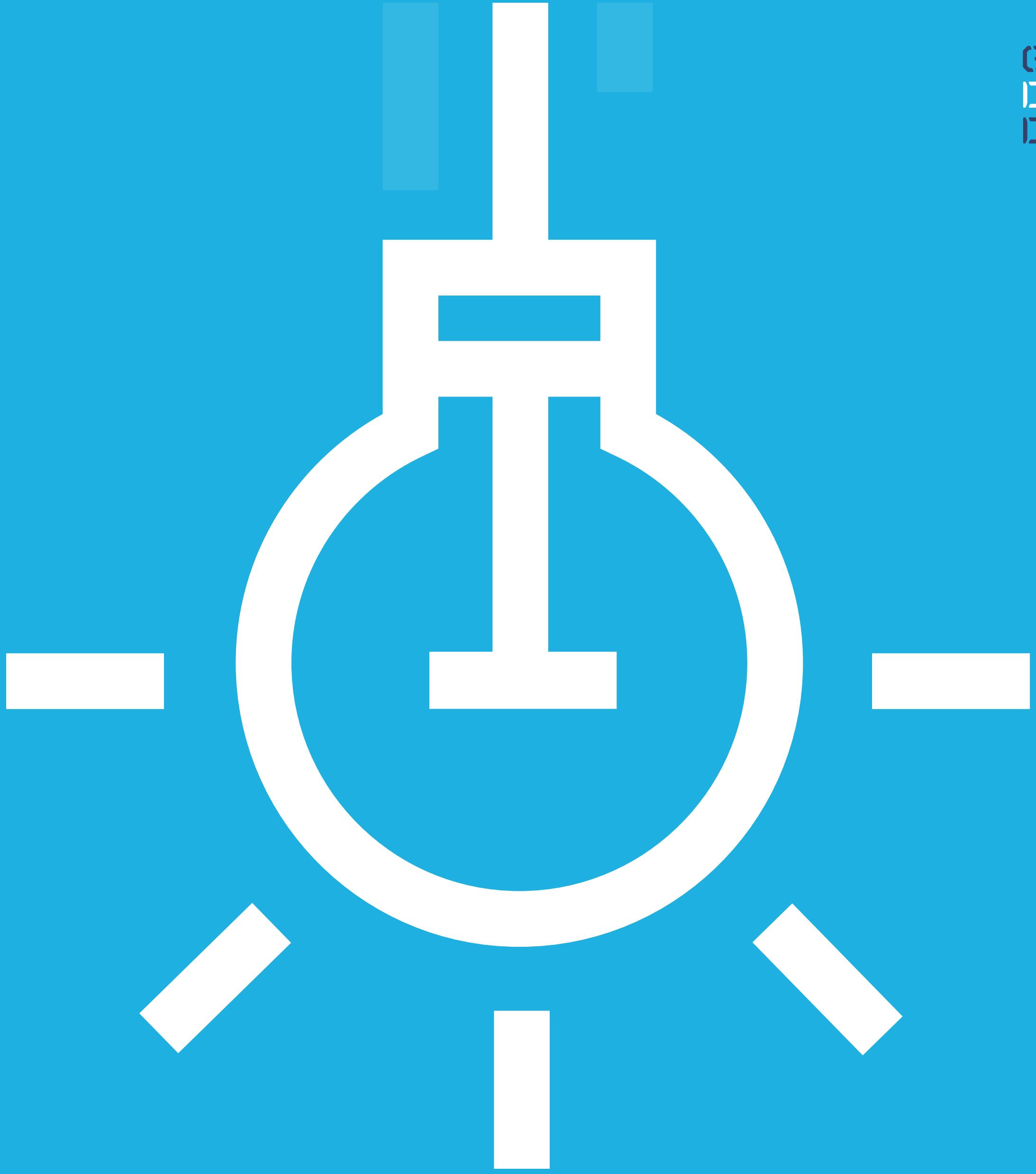
Roel Bertens

Marysia Winkels

Rens Dimmendaal

DCAI competition

What did others do?



Synaptic-AnN

Best performance

1. Manual data cleaning
2. Manual data generation
3. Auto data generation
4. Distribution and style replication
5. Filtering by vote

GO DATA DRIVEN

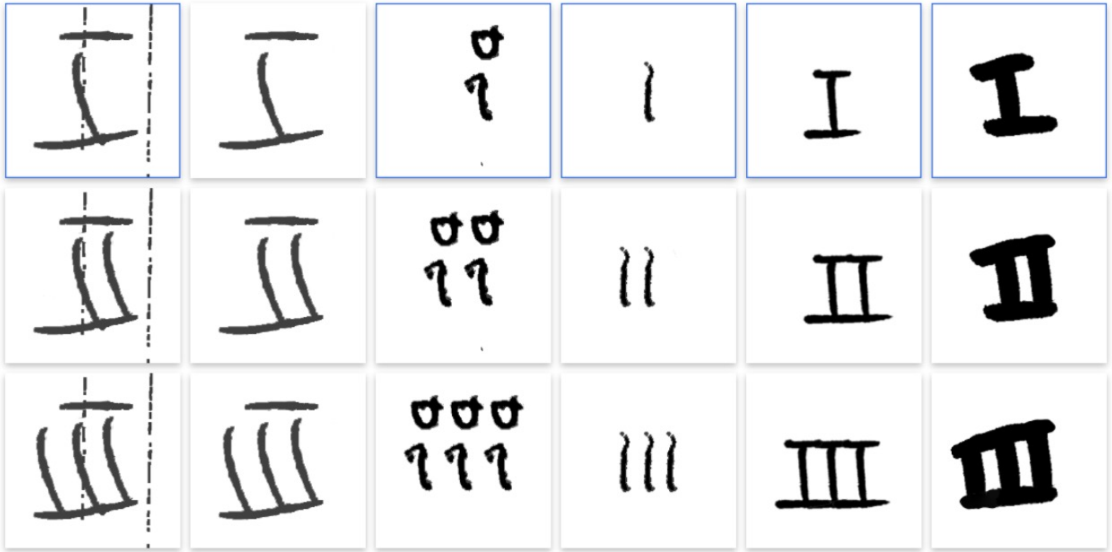
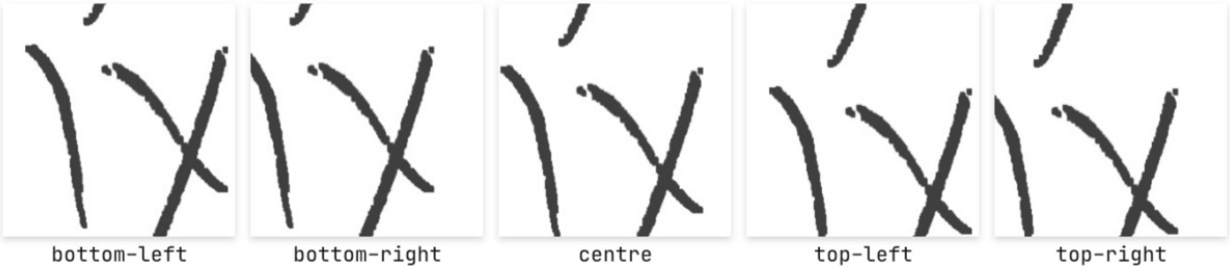


Figure 2: Style replication applied on class I of the label book — images bordered in blue are the original label book images.

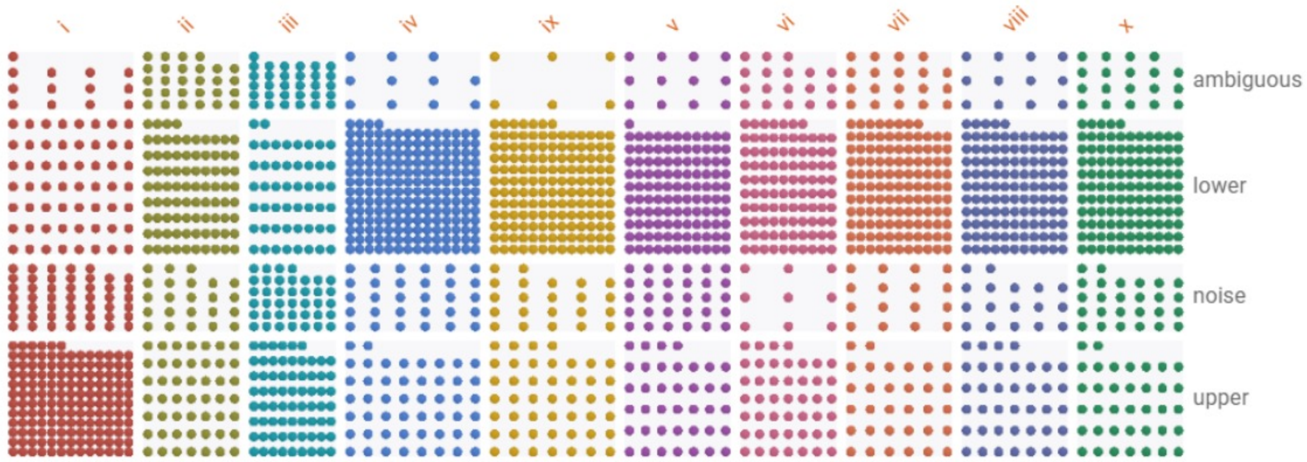


Read about it at www.deeplearning.ai/data-centric-ai-competition-synaptic-ann/

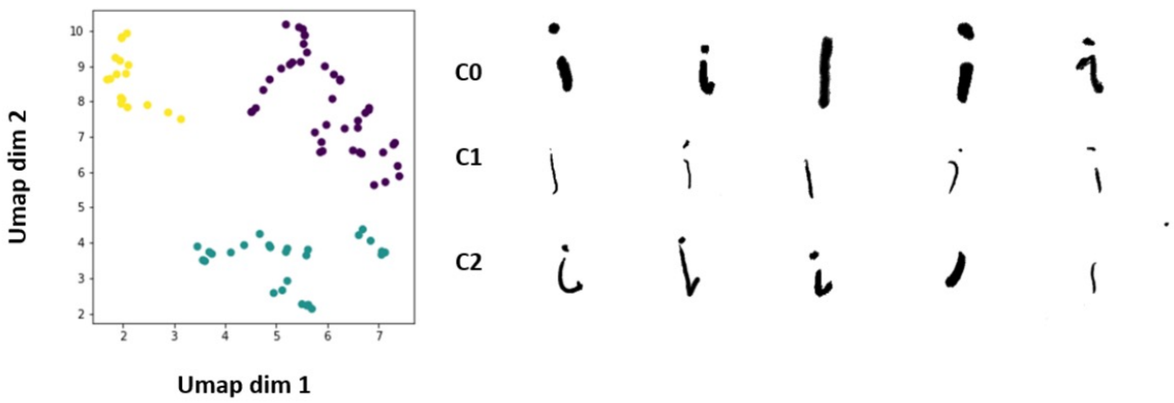
Innotescus

Best performance

1. Data cleaning
2. Rebalancing train & test dataset
3. Rebalancing subclasses using embeddings
4. Rebalancing edge cases with hard examples



Imbalance between lower and uppercase numerals (Innotescus chart)

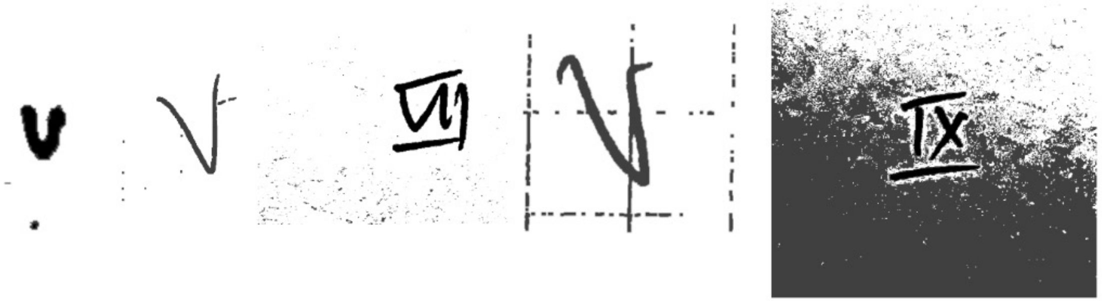


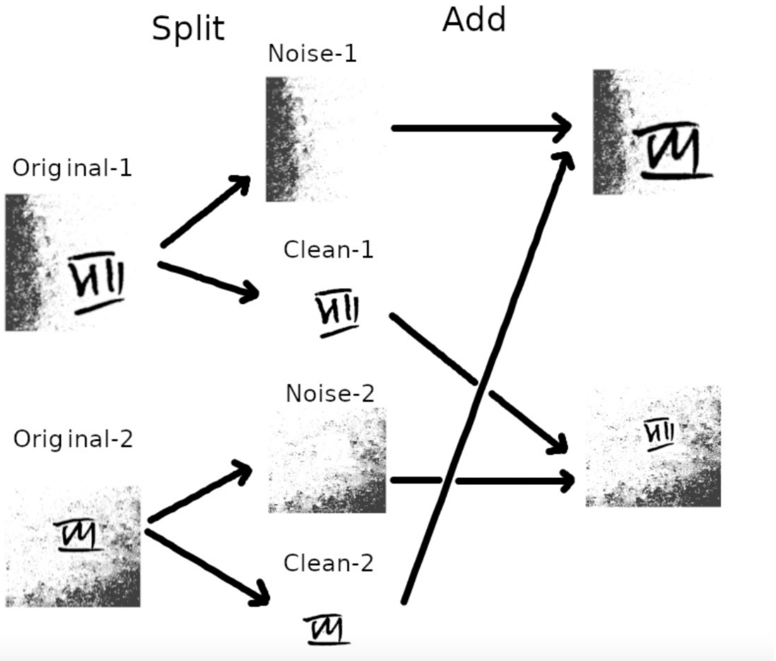
Read about it at https://www.deeplearning.ai/data-centric-ai-competition-innotescus/

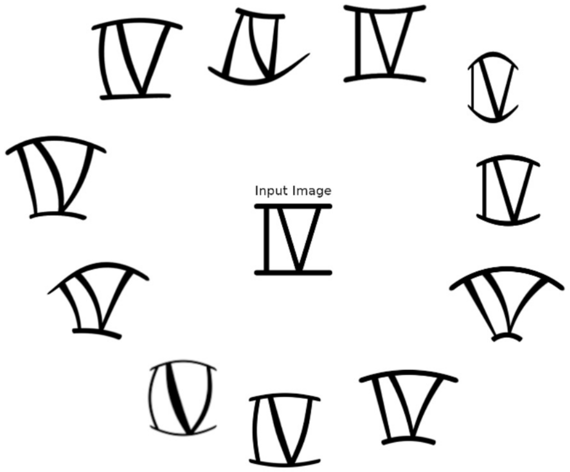
Divakar Roy

Best performance

1. Separate noise
2. Camera distortion onto skewed grid
3. Overlay into canvas
4. Data quality assessment and cleaning up





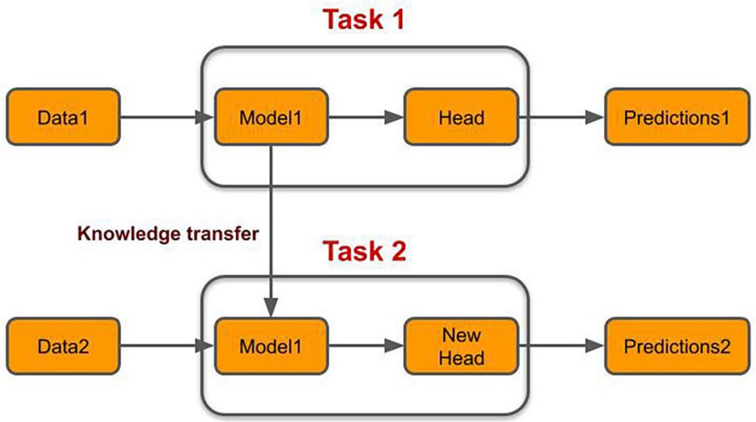


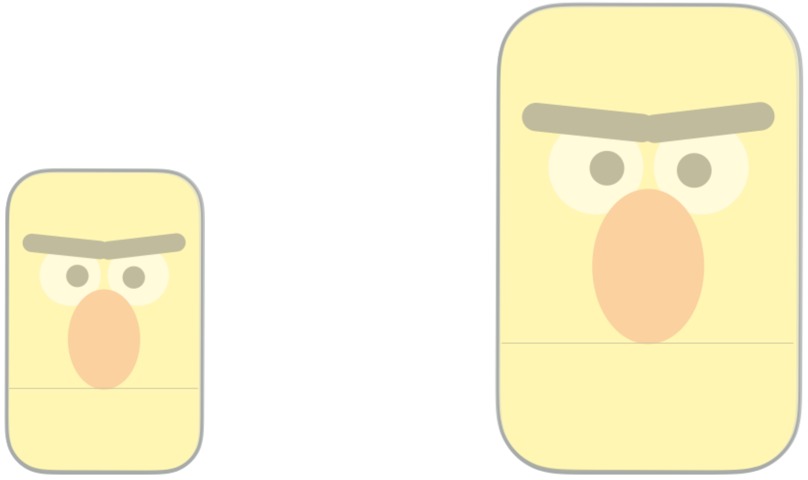
Read about it at https://www.deeplearning.ai/data-centric-ai-competition-divakar-roy/

The *why* of Data-Centric AI

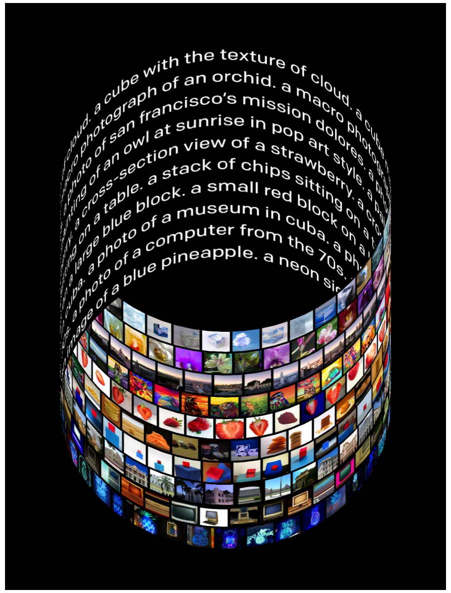
The *why* of data-centric AI

- 🧱 Foundation models & transfer learning





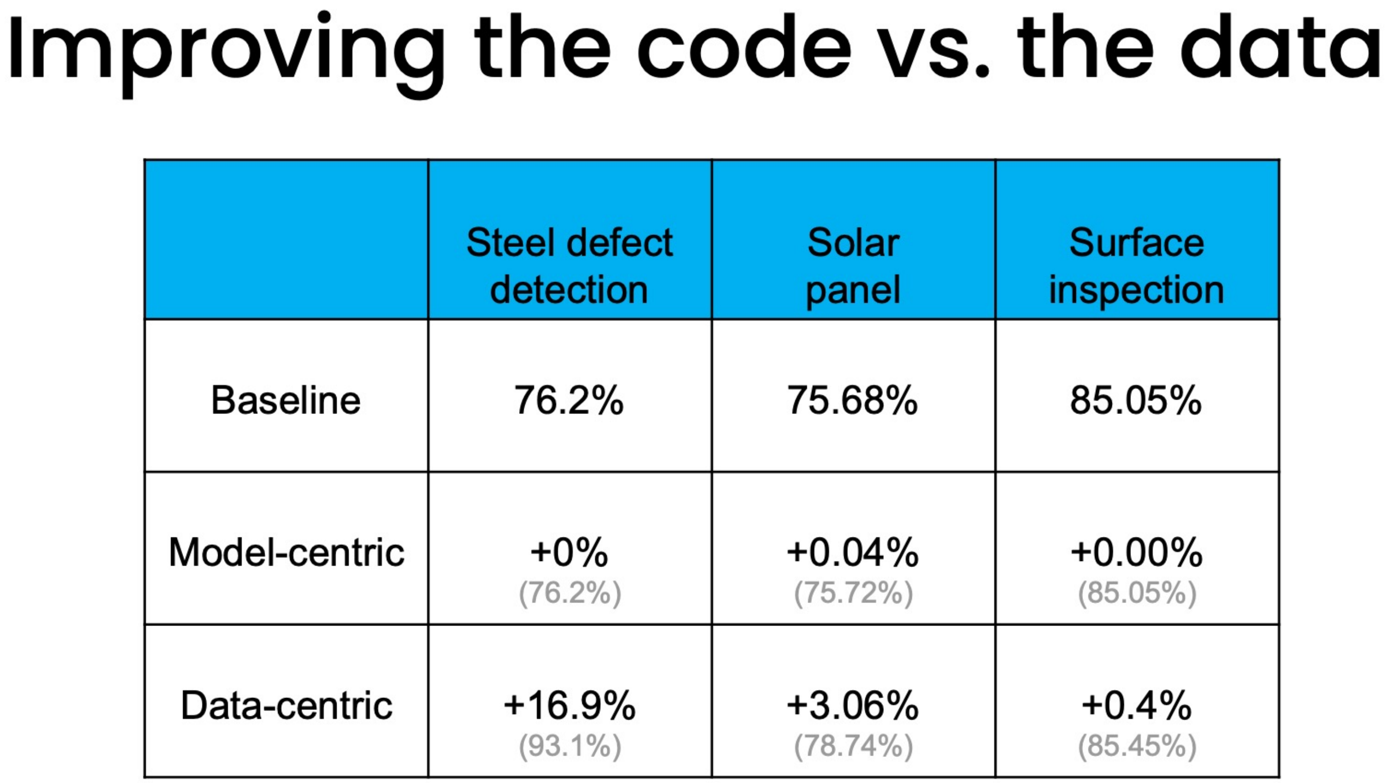
BERT$_{BASE}$ BERT$_{LARGE}$


a cube with the texture of cloud. a
photograph of an orchid. a macro photo
of san francisco's mission dolores
of an owl at sunrise in pop art style
cross-section view of a strawberry
a stack of chips sitting on a
table. a small red block on a
blue block. a museum in cuba
a photo of a computer from the 70s
a photo of a blue pineapple. a neon

The competitive advantage of data scientists lies in everything that surrounds the model.

The *why* of data-centric AI

- 🧱 Foundation models & transfer learning
- 💪 Improve performance



Improving the code vs. the data

	Steel defect detection	Solar panel	Surface inspection
Baseline	76.2%	75.68%	85.05%
Model-centric	+0% (76.2%)	+0.04% (75.72%)	+0.00% (85.05%)
Data-centric	+16.9% (93.1%)	+3.06% (78.74%)	+0.4% (85.45%)

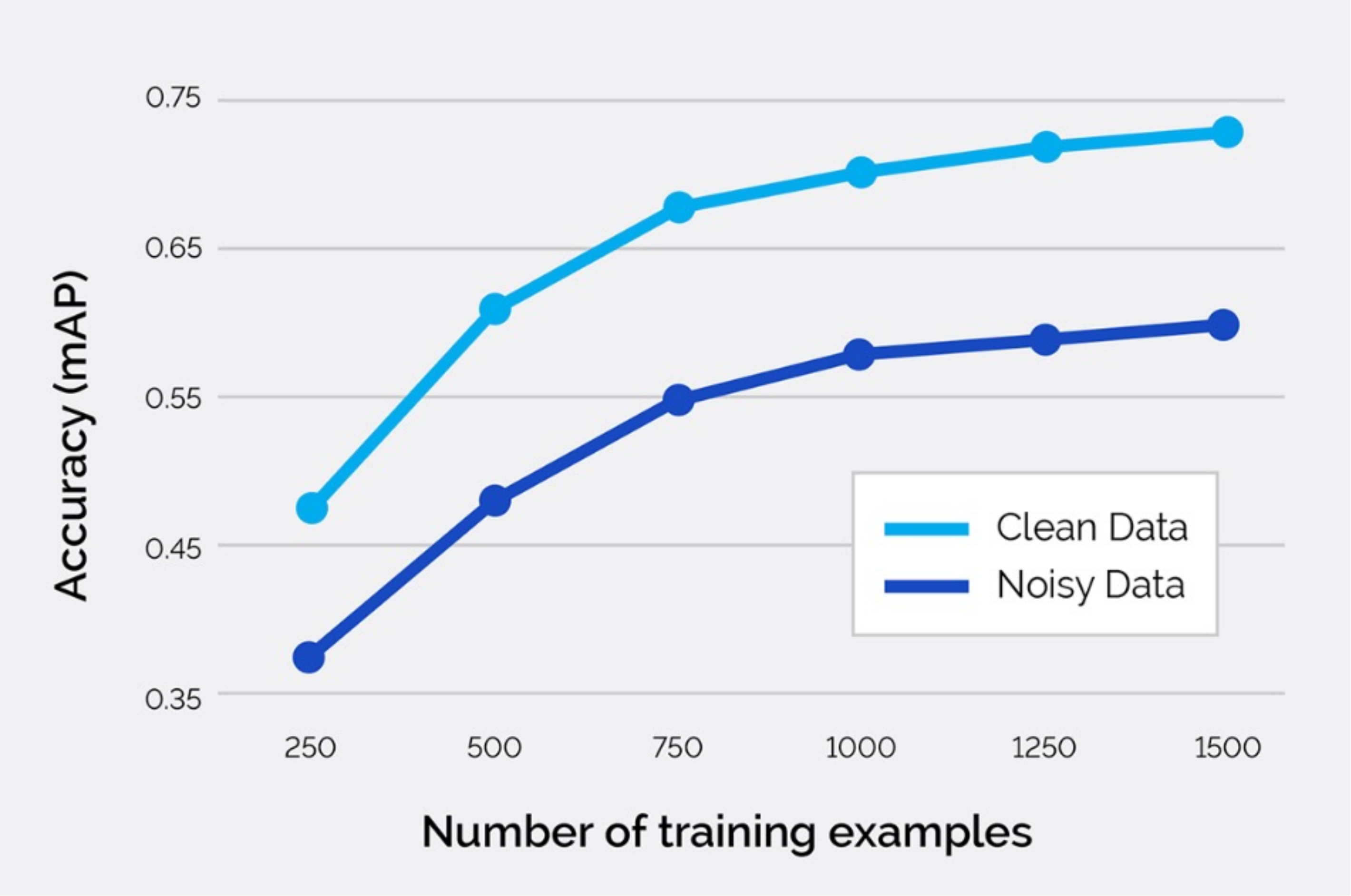
GO
DATA
DRIVEN

Credit:

Andrew Ng, MLOps: From Model-centric to Data-centric AI

The *why* of data-centric AI

- 🧱 Foundation models & transfer learning
- 💪 Improve performance



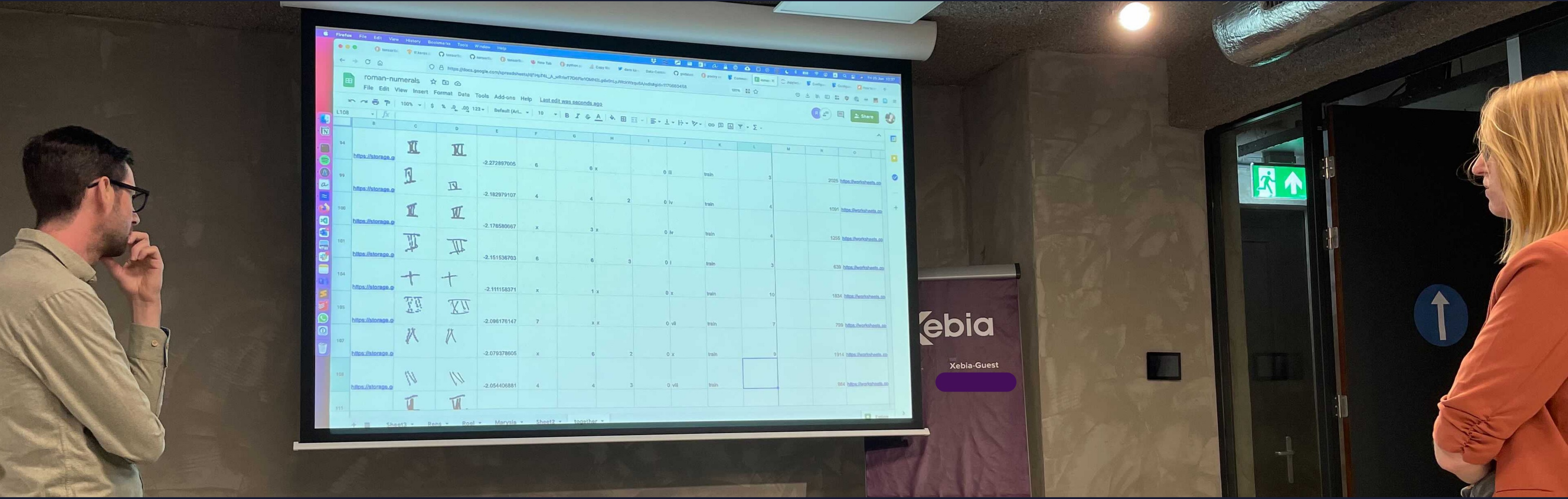
Credit: Landing.ai

The *why* of data-centric AI

- 🧱 Foundation models & transfer learning
- 💪 Improve performance

Big data problems can also small data problems

GO DATA DRIVEN

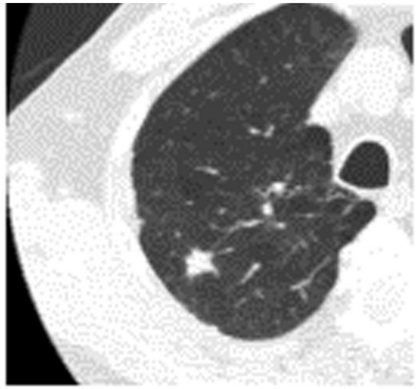

roman-numerals
Xebia
Xebia-Guest

The *why* of data-centric AI

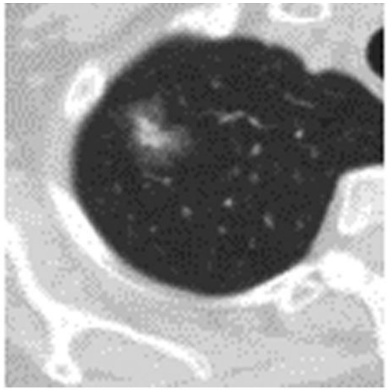
- 🧱 Foundation models & transfer learning
- 💪 Improve performance
- 🤝 Enables better collaboration

GO
DATA
DRIVEN

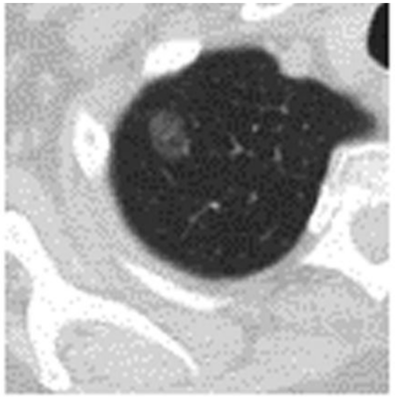
Nodule type classification



Solid nodule



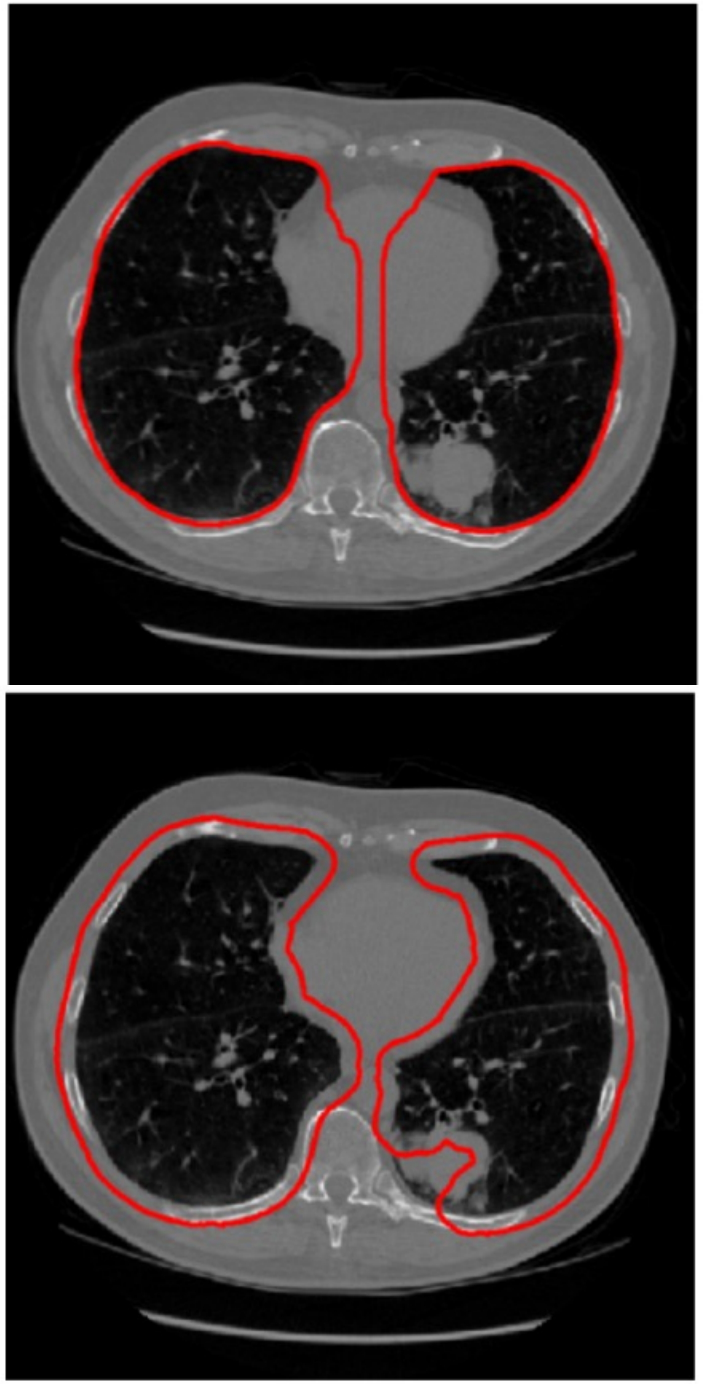
Part-solid nodule



Ground-glass nodule

The *why* of data-centric AI

- 🧱 Foundation models & transfer learning
- 💪 Improve performance
- 🤝 Enables better collaboration

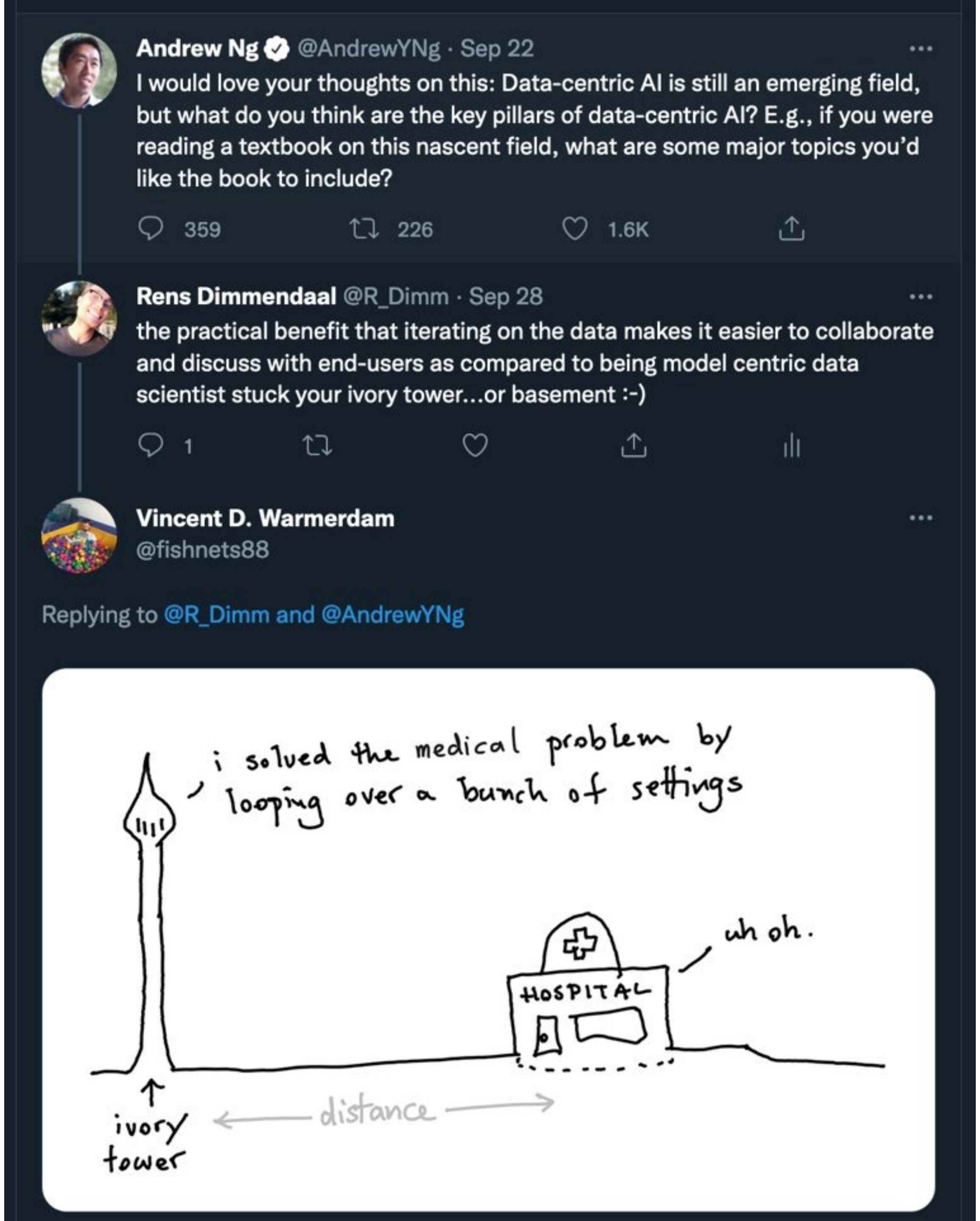


GO
DATA
DRIVEN

The *why* of data-centric AI

- 🧱 Foundation models & transfer learning
- 💪 Improve performance
- 🤝 Enables better collaboration

GO
DATA
DRIVEN



Andrew Ng @AndrewYNg · Sep 22

I would love your thoughts on this: Data-centric AI is still an emerging field, but what do you think are the key pillars of data-centric AI? E.g., if you were reading a textbook on this nascent field, what are some major topics you'd like the book to include?

359 226 1.6K

Rens Dimmendaal @R_Dimm · Sep 28

the practical benefit that iterating on the data makes it easier to collaborate and discuss with end-users as compared to being model centric data scientist stuck your ivory tower...or basement :-)

1

Vincent D. Warmerdam
@fishnets88

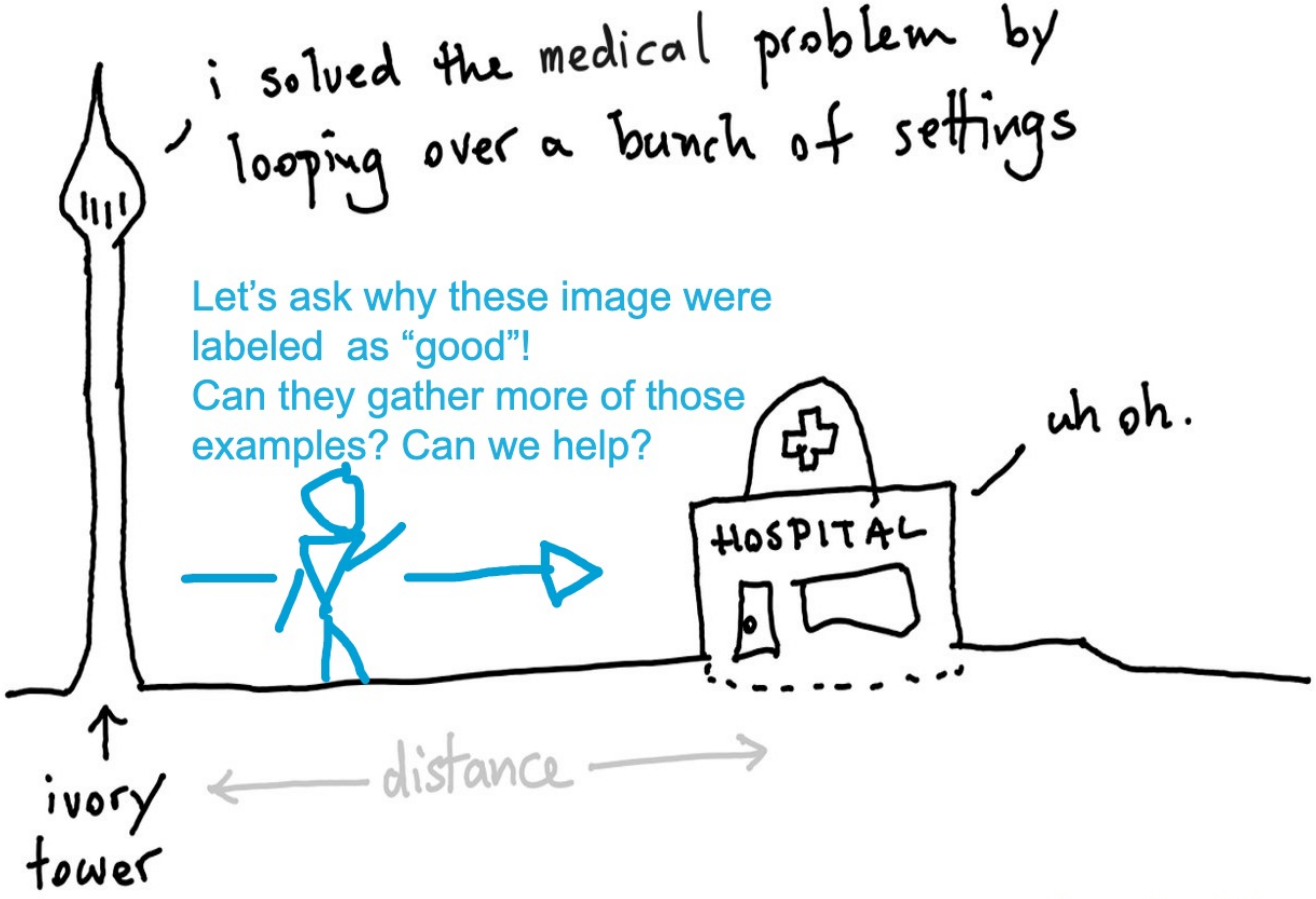
Replying to @R_Dimm and @AndrewYNg

i solved the medical problem by looping over a bunch of settings
uh oh.
HOSPITAL
ivory tower
distance

The *why* of data-centric AI

- 🧱 Foundation models & transfer learning
- 💪 Improve performance
- 🤝 Enables better collaboration

Data centric scientist



Let’s ask why these image were labeled as “good”! Can they gather more of those examples? Can we help?

i solved the medical problem by looping over a bunch of settings
HOSPITAL
uh oh.
ivory tower
distance

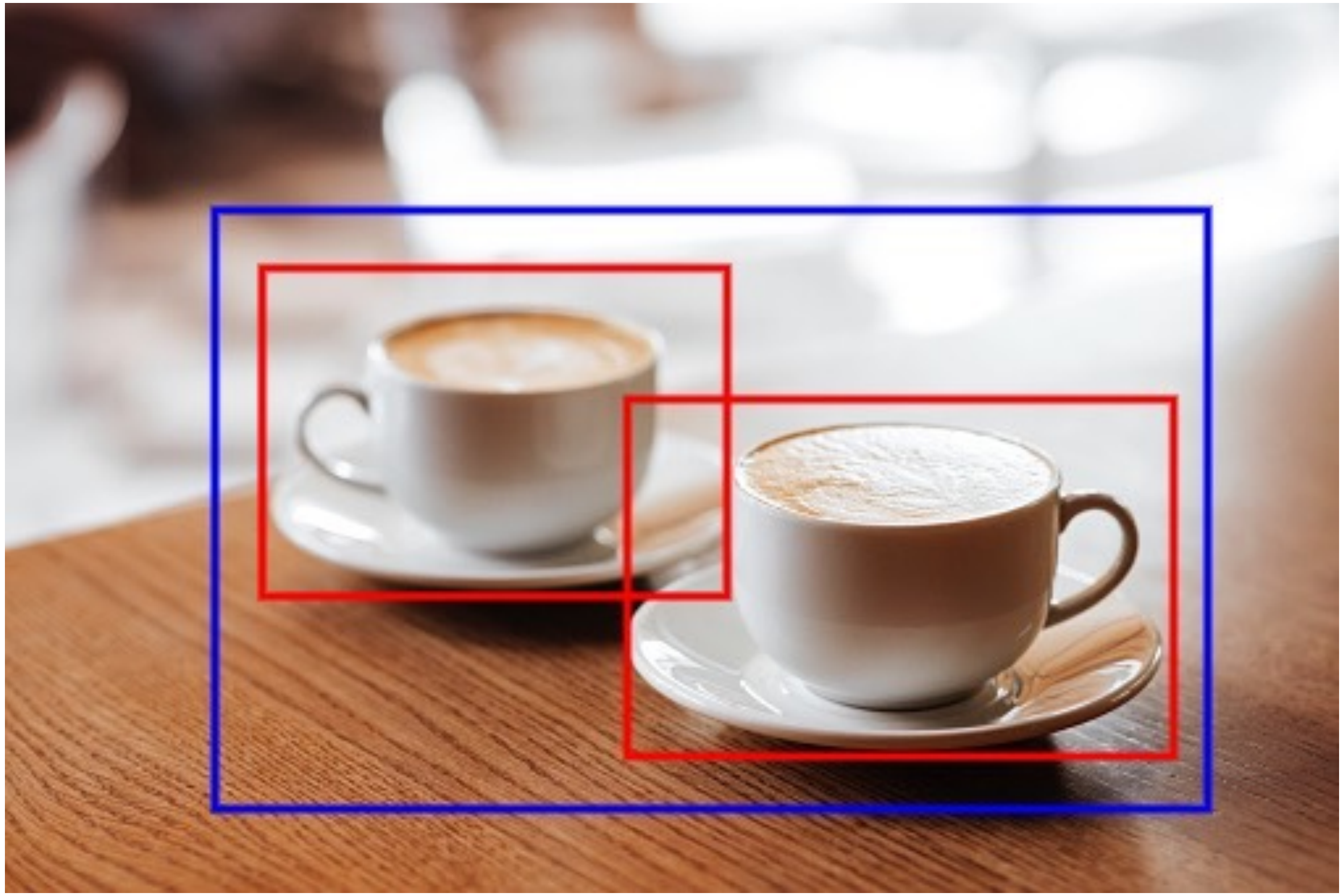
Source: Vincent Warmerdam
& Rens’s bad powerpoint skills

*“The focus has to shift from **big data to good data**. Having 50 thoughtfully engineered examples can be sufficient to explain to the neural network what you want it to learn.”*

- Andrew Ng

Data quality

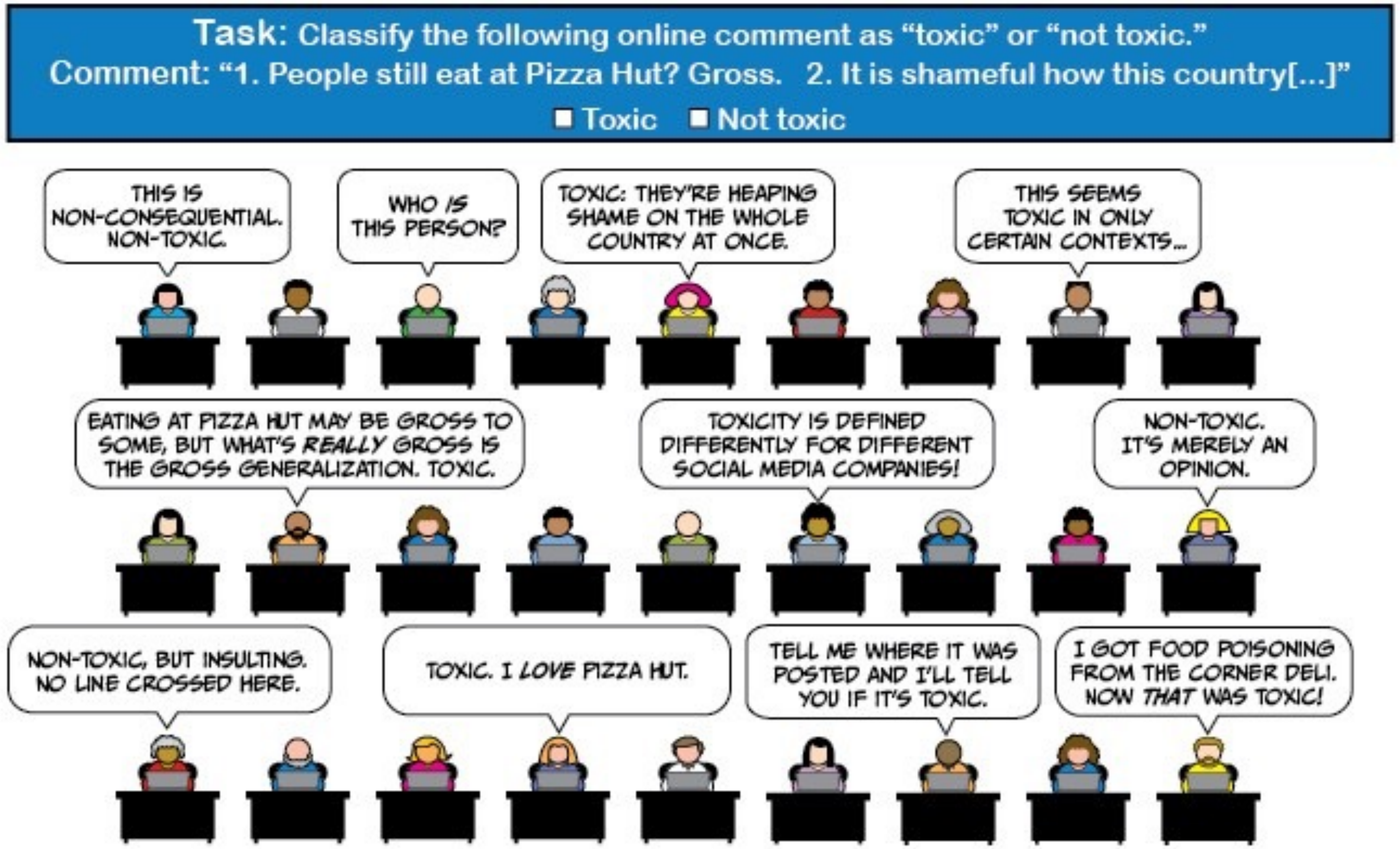
- Consistent data labels



GO DATA DRIVEN

Data quality

- Consistent data labels



Task: Classify the following online comment as "toxic" or "not toxic."
Comment: "1. People still eat at Pizza Hut? Gross. 2. It is shameful how this country[...]"
□ Toxic □ Not toxic

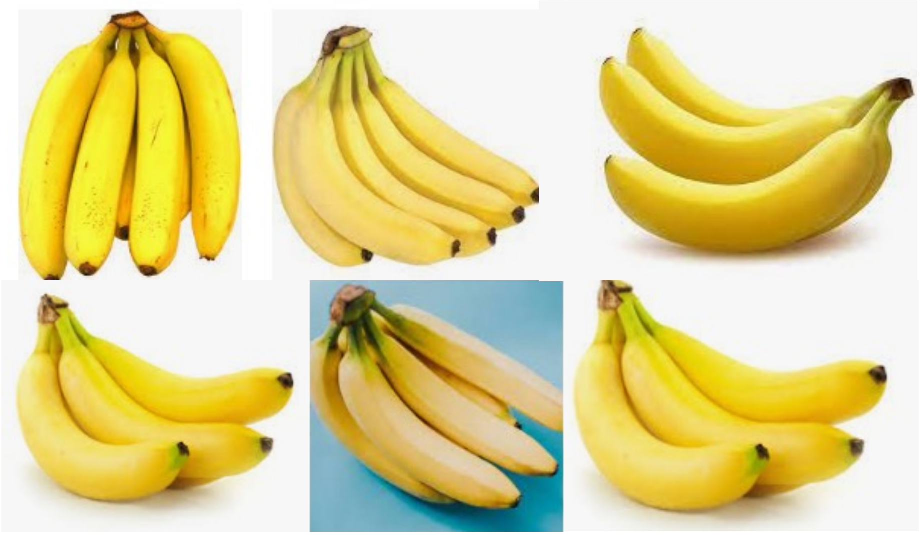
THIS IS NON-CONSEQUENTIAL. NON-TOXIC.
WHO IS THIS PERSON?
TOXIC: THEY'RE HEAPING SHAME ON THE WHOLE COUNTRY AT ONCE.
THIS SEEMS TOXIC IN ONLY CERTAIN CONTEXTS...
EATING AT PIZZA HUT MAY BE GROSS TO SOME, BUT WHAT'S *REALLY* GROSS IS THE GROSS GENERALIZATION. TOXIC.
TOXICITY IS DEFINED DIFFERENTLY FOR DIFFERENT SOCIAL MEDIA COMPANIES!
NON-TOXIC. IT'S MERELY AN OPINION.
NON-TOXIC, BUT INSULTING. NO LINE CROSSED HERE.
TOXIC. I *LOVE* PIZZA HUT.
TELL ME WHERE IT WAS POSTED AND I'LL TELL YOU IF IT'S TOXIC.
I GOT FOOD POISONING FROM THE CORNER DELI. NOW *THAT* WAS TOXIC!

Credit: Michael Bernstein

Data quality

- Consistent data labels
- Complete and representative data

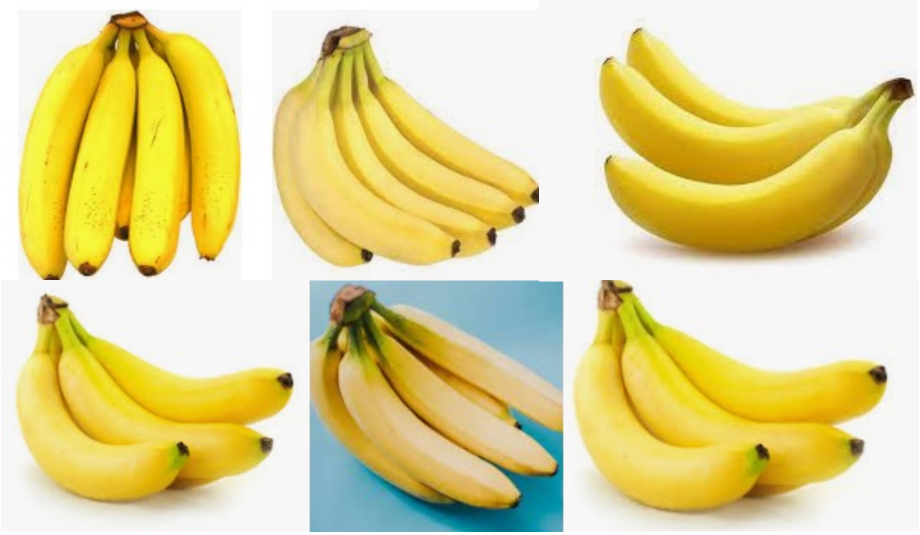
Train set



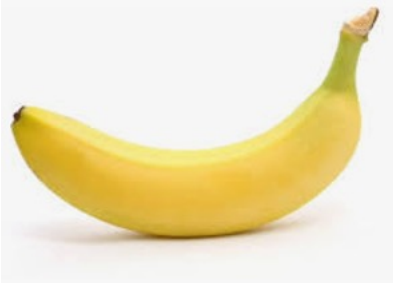
Data quality

- Consistent data labels
- Complete and representative data

Train set

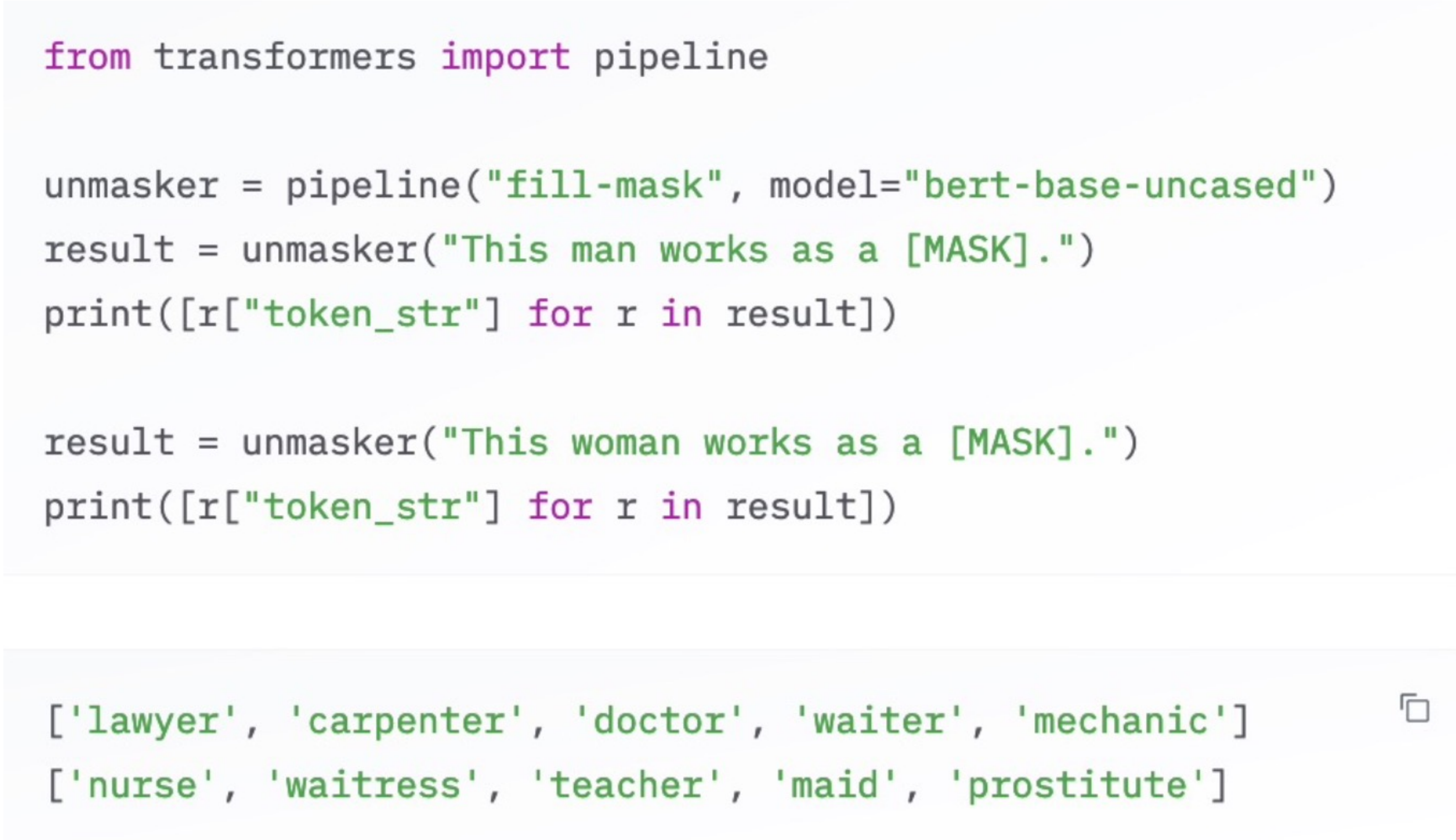


Real world example



Data quality

- Consistent data labels
- Complete and representative data
- Unbiased data

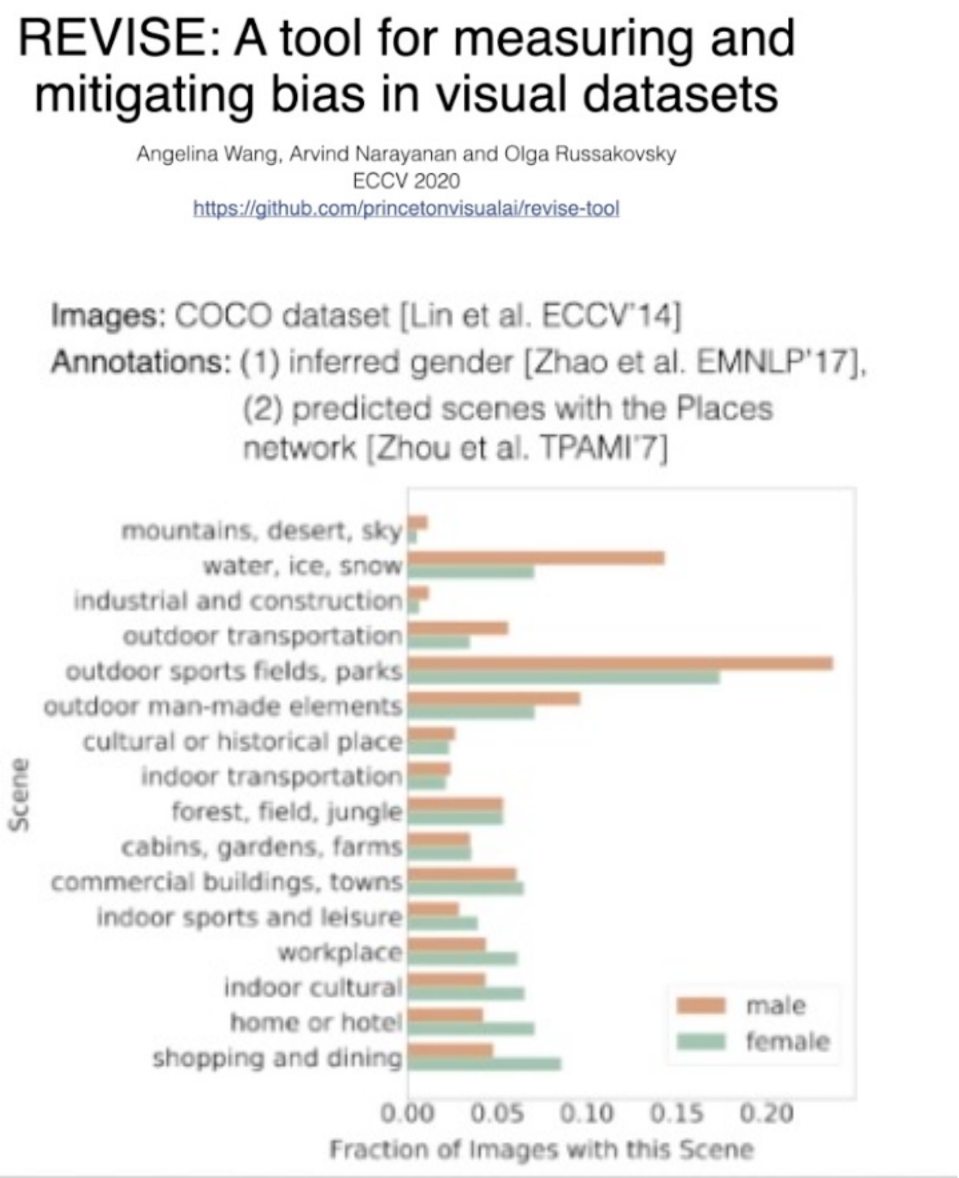


```
from transformers import pipeline

unmasker = pipeline("fill-mask", model="bert-base-uncased")
result = unmasker("This man works as a [MASK].")
print([r["token_str"] for r in result])

result = unmasker("This woman works as a [MASK].")
print([r["token_str"] for r in result])
```

```
['lawyer', 'carpenter', 'doctor', 'waiter', 'mechanic']
['nurse', 'waitress', 'teacher', 'maid', 'prostitute']
```

Data quality

- Consistent data labels
- Complete and representative data
- Unbiased data



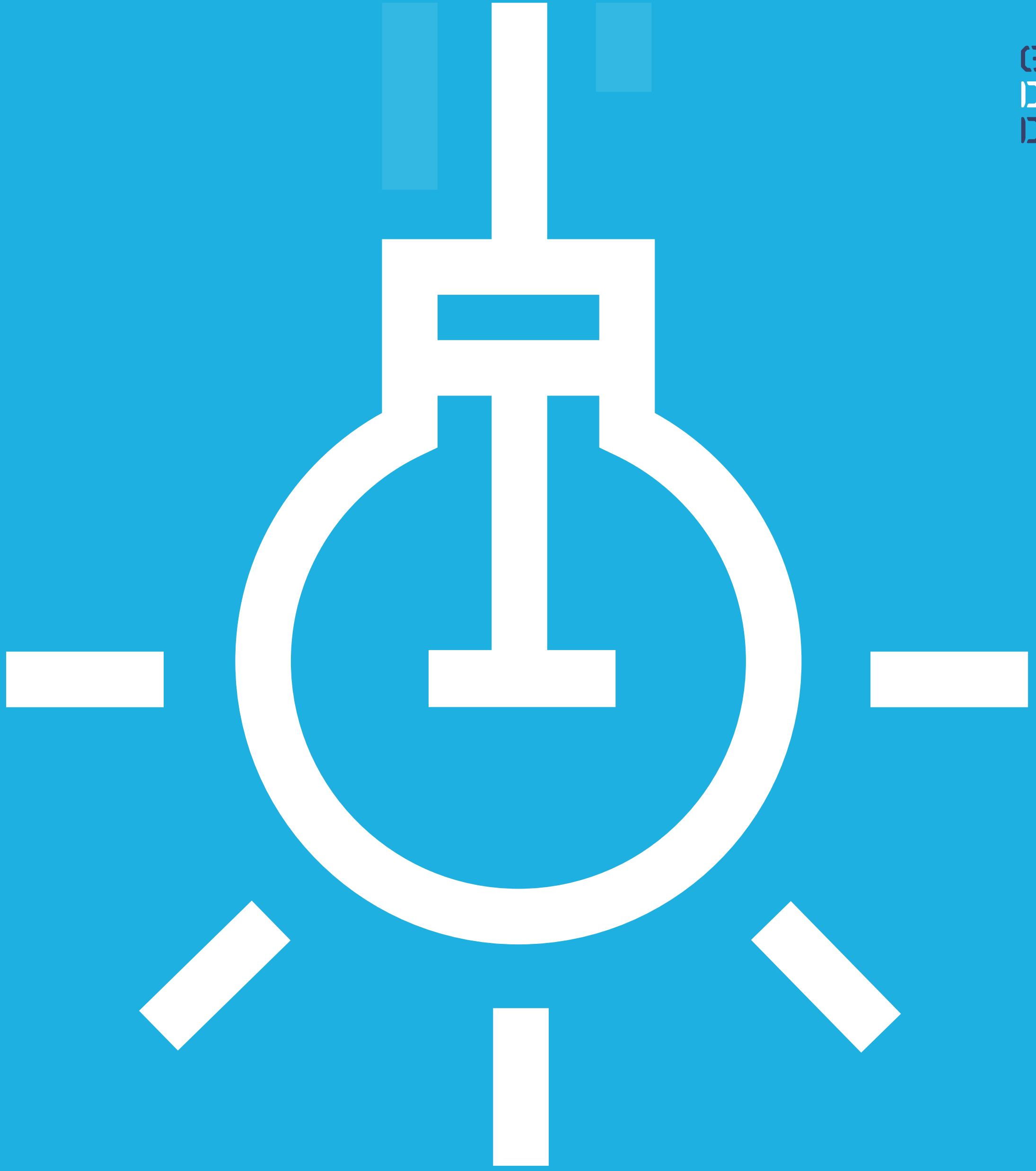
REVISE: A tool for measuring and mitigating bias in visual datasets

Angelina Wang, Arvind Narayanan and Olga Russakovsky
ECCV 2020
https://github.com/princetonvisualai/revise-tool

Images: COCO dataset [Lin et al. ECCV'14]
Annotations: (1) inferred gender [Zhao et al. EMNLP'17],
(2) predicted scenes with the Places network [Zhou et al. TPAMI'7]

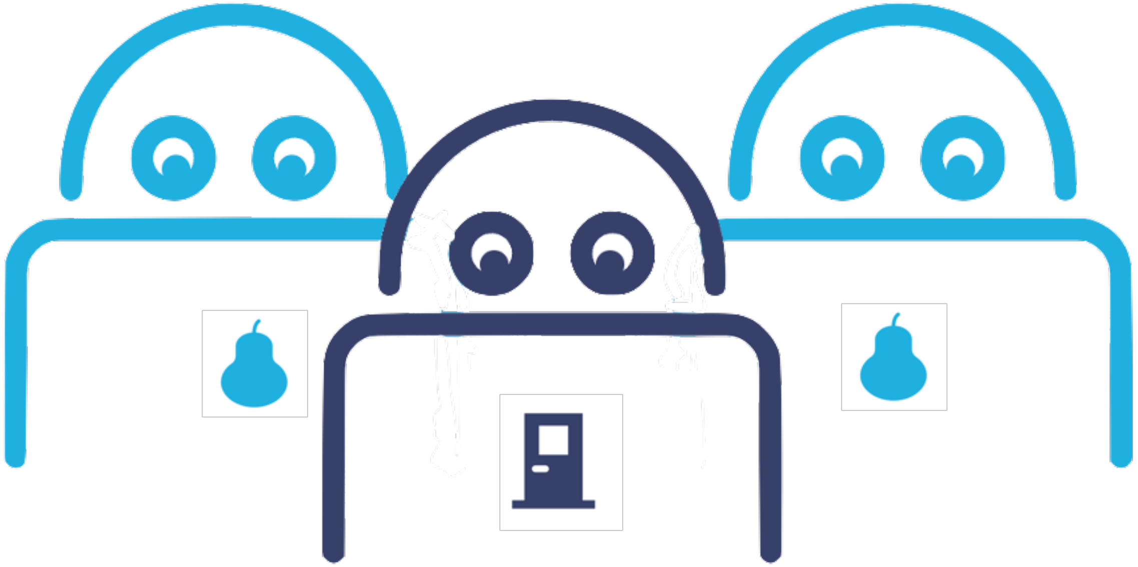
GO
DATA
DRIVEN

How do we get to Data-Centric AI?



Development of Data-Centric AI

1. Individuals take an interest



Development of Data-Centric AI

1. Individuals take an interest
2. Adopted by many as best practice

GO
DATA
DRIVEN

dc ai DATA-CENTRIC AI

NEURIPS DATA-CENTRIC AI WORKSHOP

DeepLearning.AI | LANDING AI

Data-Centric AI Competition

Join the data-centric AI movement!

The Future of Data-Centric AI

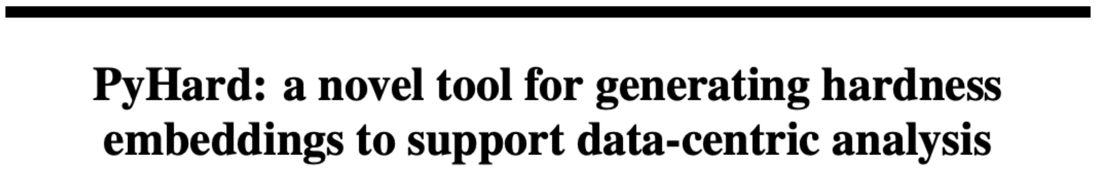
August 2, 2022 | Workshop

August 3-4, 2022 | Sessions

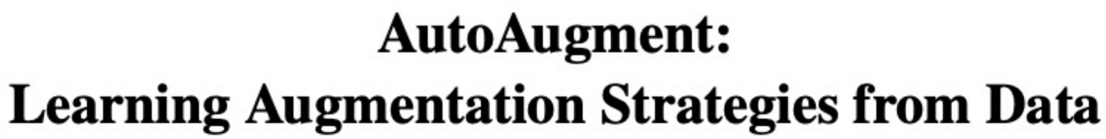
***Data-centric AI** is the discipline of systematically engineering the data used to build an **AI** system.*

Development of Data-Centric AI

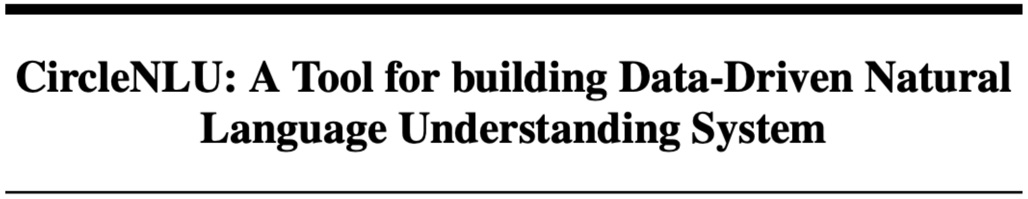
1. Individuals take an interest
2. Adopted by many as best practice
3. Systematic tools are developed



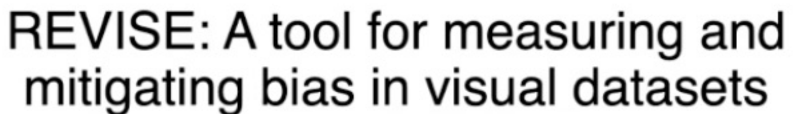
PyHard: a novel tool for generating hardness embeddings to support data-centric analysis



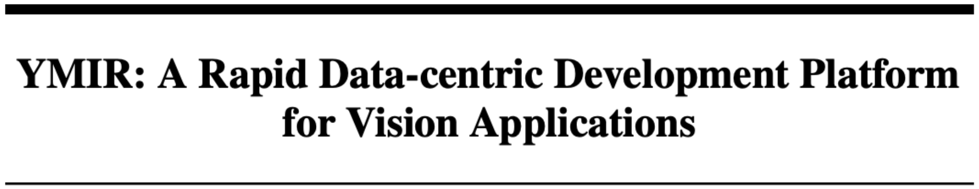
AutoAugment: Learning Augmentation Strategies from Data



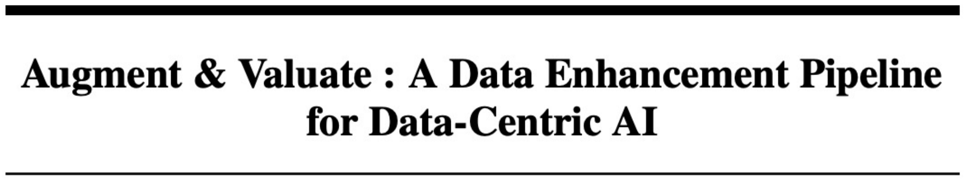
CircleNLU: A Tool for building Data-Driven Natural Language Understanding System



REVISE: A tool for measuring and mitigating bias in visual datasets



YMIR: A Rapid Data-centric Development Platform for Vision Applications



Augment & Valuate : A Data Enhancement Pipeline for Data-Centric AI

Development of Data-Centric AI

1. Individuals take an interest
2. Adopted by many as best practice
3. Systematic tools are developed

LANDING AI

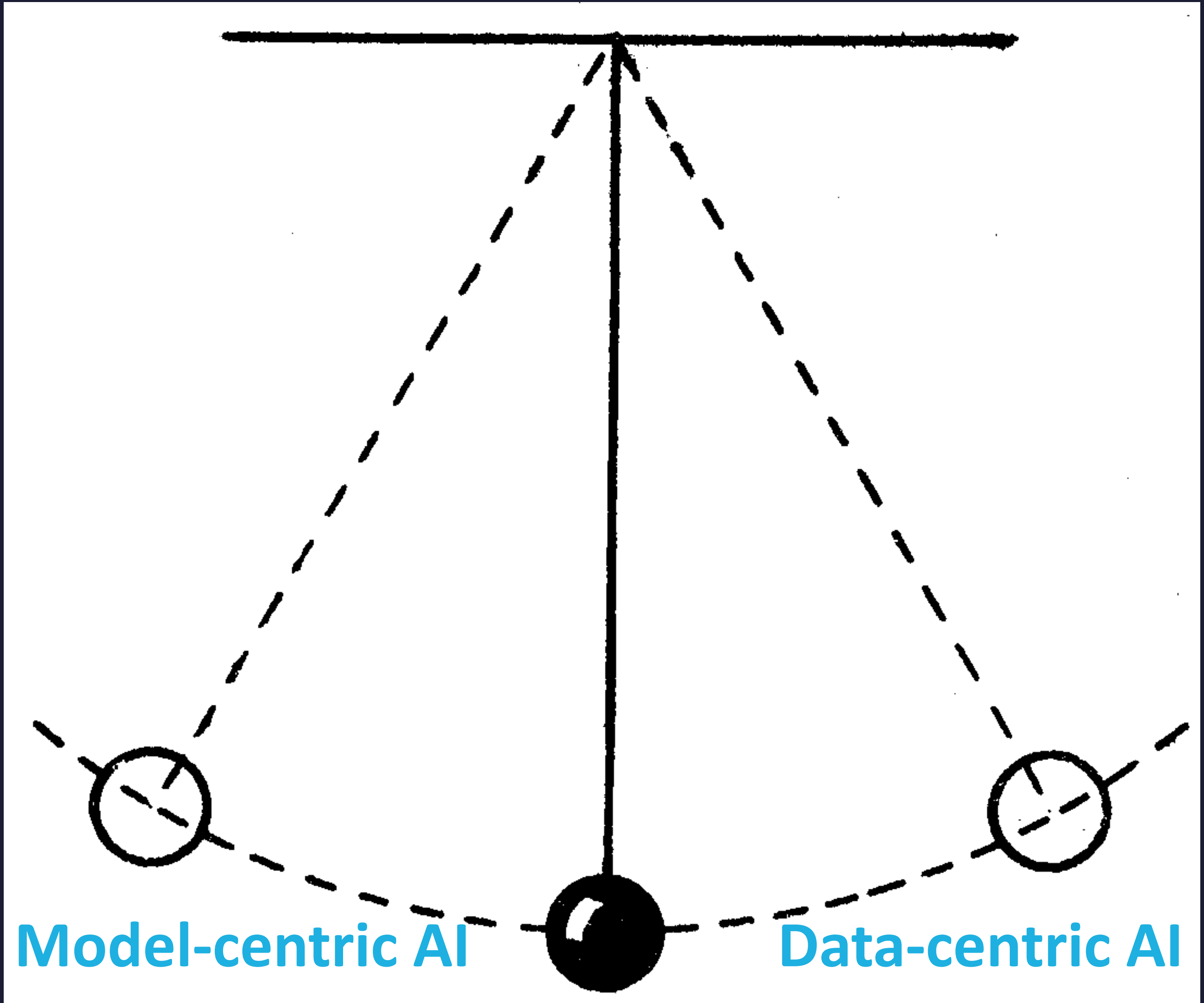
snorkel

HoloClean

Cleanlab

Albumentations

“But I *like* building models!”



Model-centric AI
Data-centric AI

Thank you!

GO
DATA
DRIVEN

PyData
London 2022