Course Wrap-Up

Itthi Chatnuntawech
Nanoinformatics and Artificial Intelligence (NAI)
National Nanotechnology Center (NANOTEC)
National Science and Technology Development Agency (NSTDA)
September 28, 2022

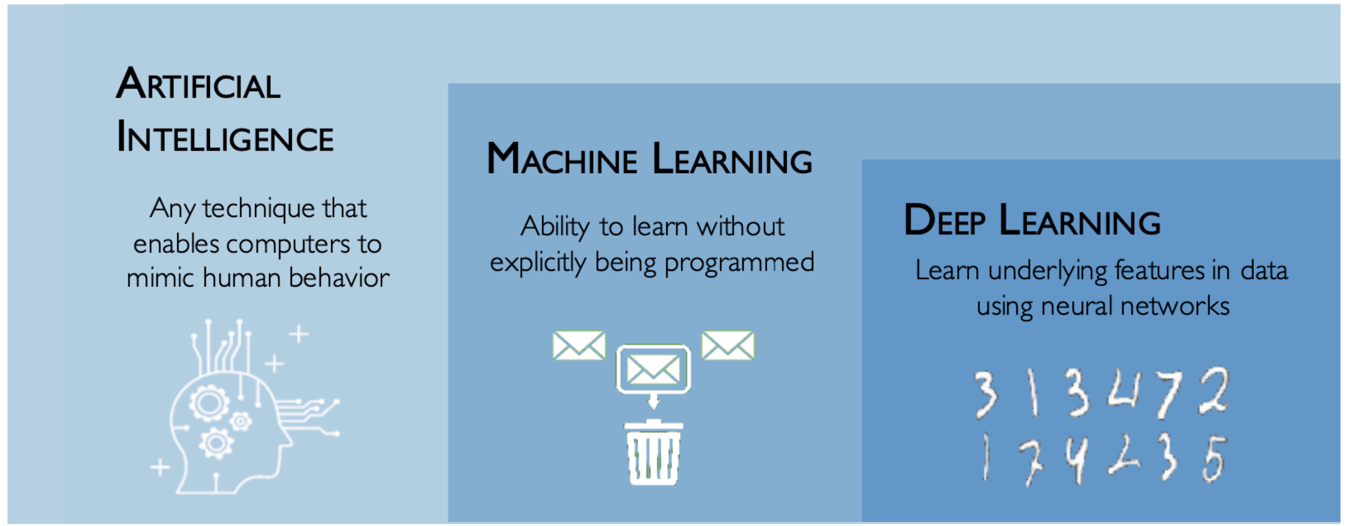
Outline

- Machine Learning Pitfalls
  - ML algorithms are not one-size-fits-all
  - Having more data doesn't always help
  - Exploiting spurious correlations
  - Training data don't capture real use cases
  - Having data leakage
  - Reporting a "non-informative" evaluation metric
  - Solving "irrelevant" problems
- A Way to Communicate with ML Practitioners
  - Be clear about what the inputs and outputs are
  - State the requirements and use-cases
  - Provide domain-/field-specific information
  - Summary of related papers
- Resources

Outline

- Machine Learning Pitfalls
  - ML algorithms are not one-size-fits-all
  - Having more data doesn't always help
  - Exploiting spurious correlations
  - Training data don't capture real use cases
  - Having data leakage
  - Reporting a "non-informative" evaluation metric
  - Solving "irrelevant" problems
- A Way to Communicate with ML Practitioners
  - Be clear about what the inputs and outputs are
  - State the requirements and use-cases
  - Provide domain-/field-specific information
  - Summary of related papers
- Resources

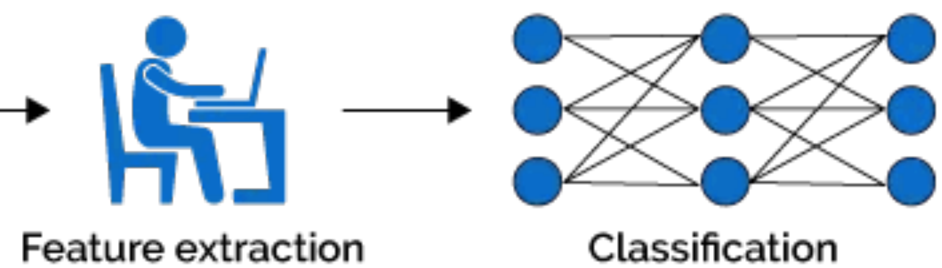
Pitfalls: ML Algorithms Are Not One-Size-Fits-All



Traditional Machine Learning

Input image

random forest, SVM, kNN, XGBoost, LGBM



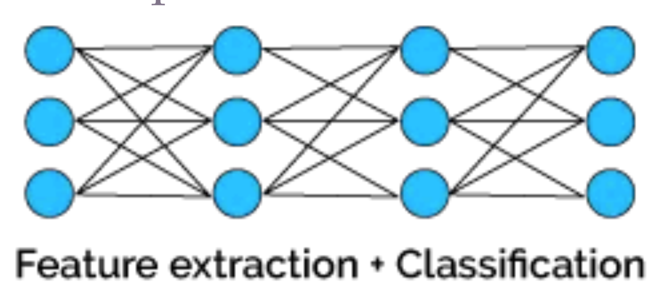
Output

watermelon

Deep Learning

Input image

deep neural network



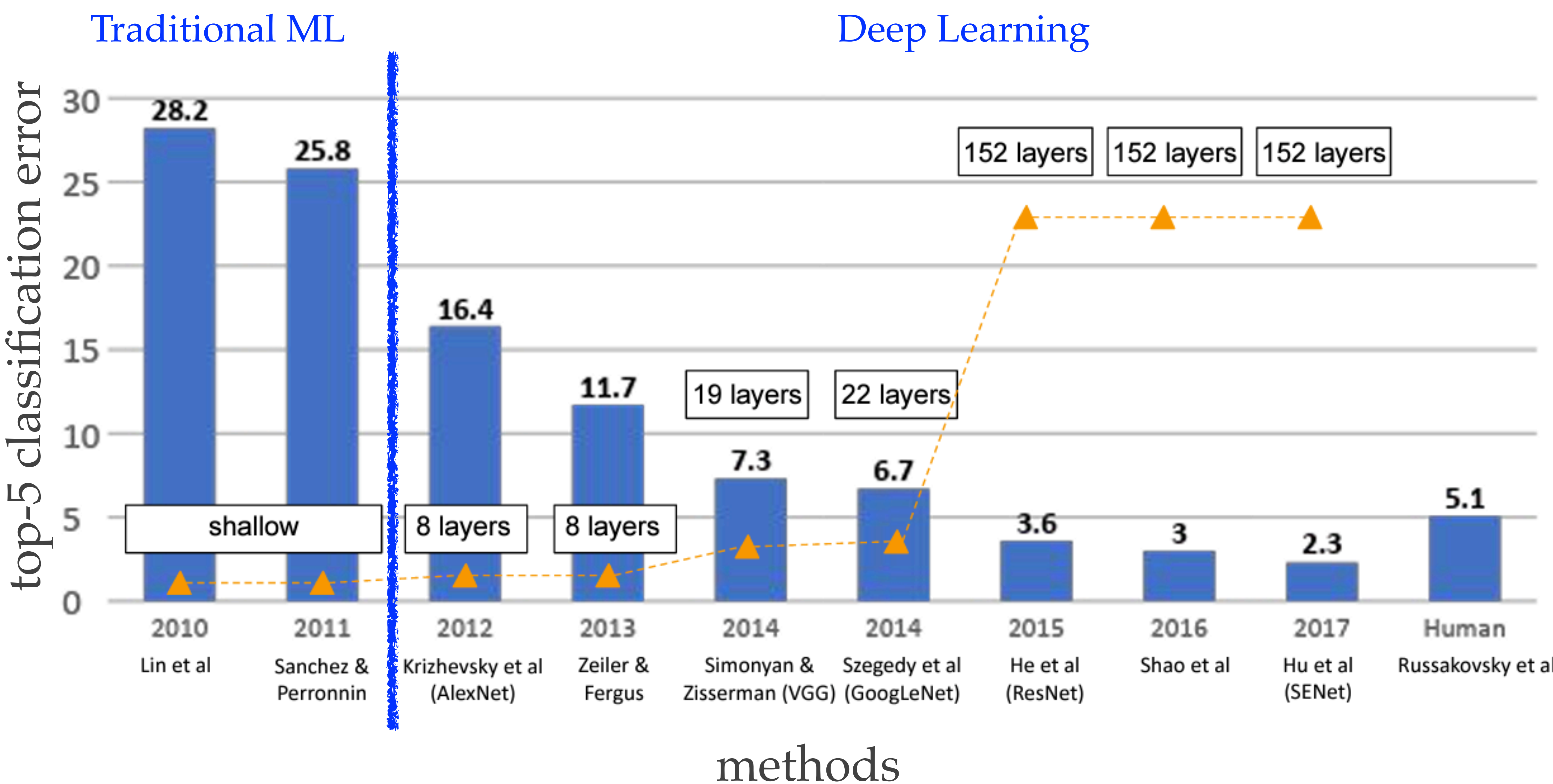
Output

watermelon

Some parts taken from: Azua Tech

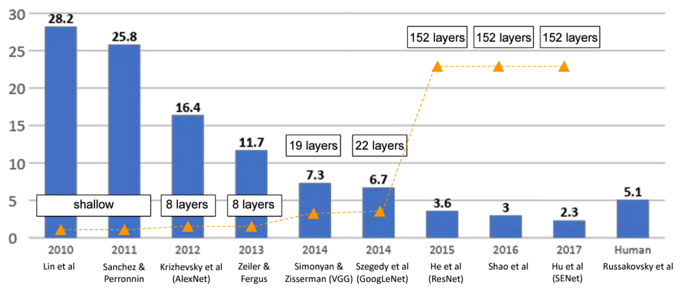
Pitfalls: ML Algorithms Are Not One-Size-Fits-All

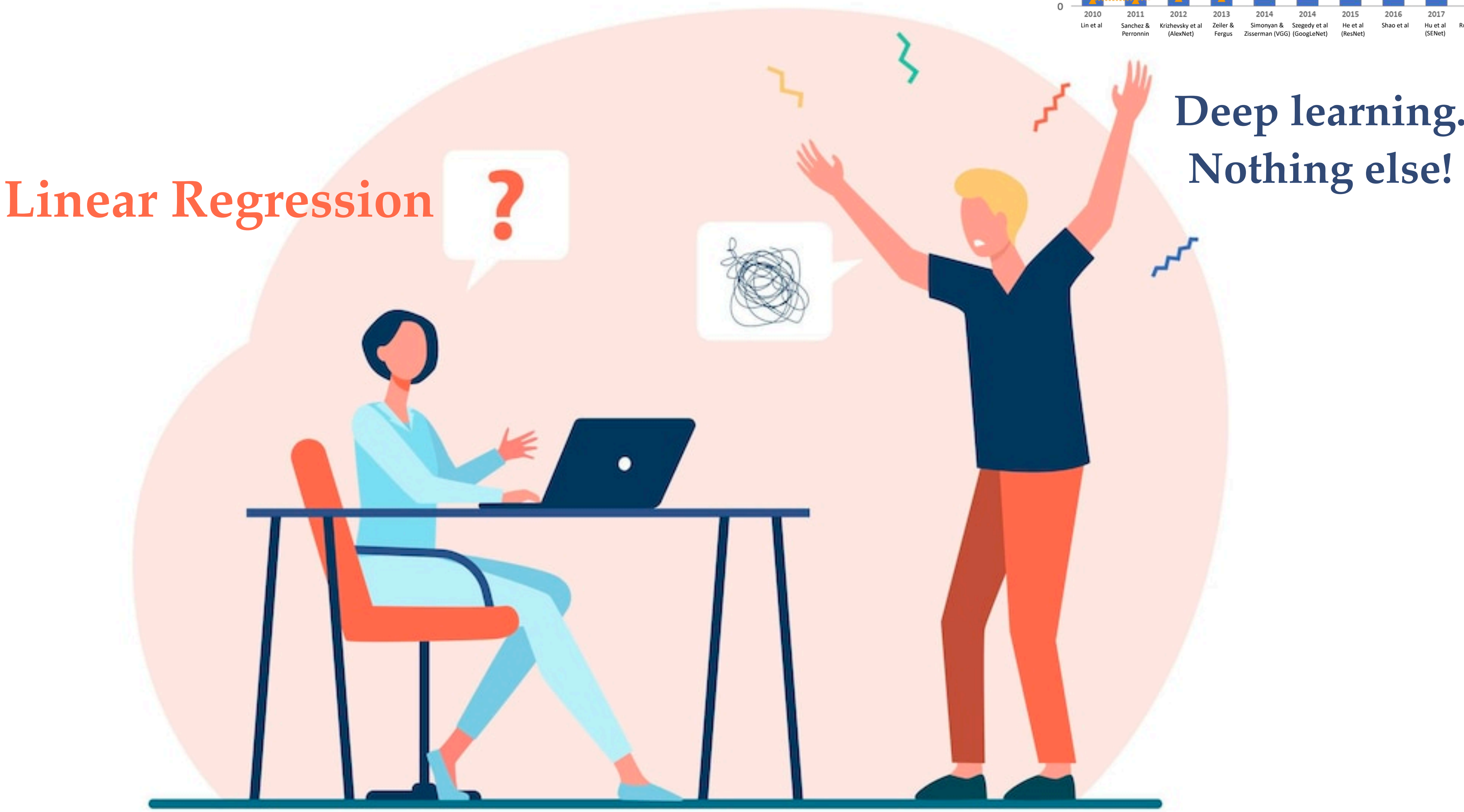
ImageNet Large Scale Visual Recognition Challenge (ILSVRC) winners



Stanford CS231n (Spring 2022): Lecture 6.

Pitfalls: ML Algorithms Are Not One-Size-Fits-All





Modified version of an image by pch.vector on Freepik

Pitfalls: ML Algorithms Are Not One-Size-Fits-All

Table 1: Dataset properties. Notation: “RMSE” ~ root-mean-square error, “Acc.” ~ accuracy.

Tabular Data

	CA	AD	HE	JA	HI	AL	EP	YE	CO	YA	MI
#objects	20640	48842	65196	83733	98050	108000	500000	515345	581012	709877	1200192
#num. features	8	6	27	54	28	128	2000	90	54	699	136
#cat. features	0	8	0	0	0	0	0	0	0	0	0
metric	RMSE	Acc.	Acc.	Acc.	Acc.	Acc.	Acc.	RMSE	Acc.	RMSE	RMSE
#classes	–	2	100	4	2	1000	2	–	7	–	–

larger datasets

Table 4: Results for ensembles of GBDT and the main DL models. For each model-dataset pair, the metric value averaged over three ensembles is reported. See supplementary for standard deviations. Notation follows Table 3.

	CA ↓	AD ↑	HE ↑	JA ↑	HI ↑	AL ↑	EP ↑	YE ↓	CO ↑	YA ↓	MI ↓
Default hyperparameters											
XGBoost	0.462	**0.874**	0.348	0.711	0.717	0.924	0.8799	9.192	0.964	0.761	0.751
CatBoost	**0.428**	0.873	0.386	0.724	0.728	0.948	0.8893	8.885	0.910	0.749	0.744
FT-Transformer	0.454	0.860	**0.395**	**0.734**	**0.731**	**0.966**	**0.8969**	**8.727**	**0.973**	**0.747**	**0.742**

Deep learning

FT-Transformer allows building powerful ensembles out of the box.

Pitfalls: ML Algorithms Are Not One-Size-Fits-All

Table 1: Dataset properties. Notation: "RMSE" ~ root-mean-square error, "Acc." ~ accuracy.

Tabular Data

	CA	AD	HE	JA	HI	AL	EP	YE	CO	YA	MI
#objects	20640	48842	65196	83733	98050	108000	500000	515345	581012	709877	1200192
#num. features	8	6	27	54	28	128	2000	90	54	699	136
#cat. features	0	8	0	0	0	0	0	0	0	0	0
metric	RMSE	Acc.	Acc.	Acc.	Acc.	Acc.	Acc.	RMSE	Acc.	RMSE	RMSE
#classes	–	2	100	4	2	1000	2	–	7	–	–

larger datasets

Table 4: Results for ensembles of GBDT and the main DL models. For each model-dataset pair, the metric value averaged over three ensembles is reported. See supplementary for standard deviations. Notation follows Table 3.

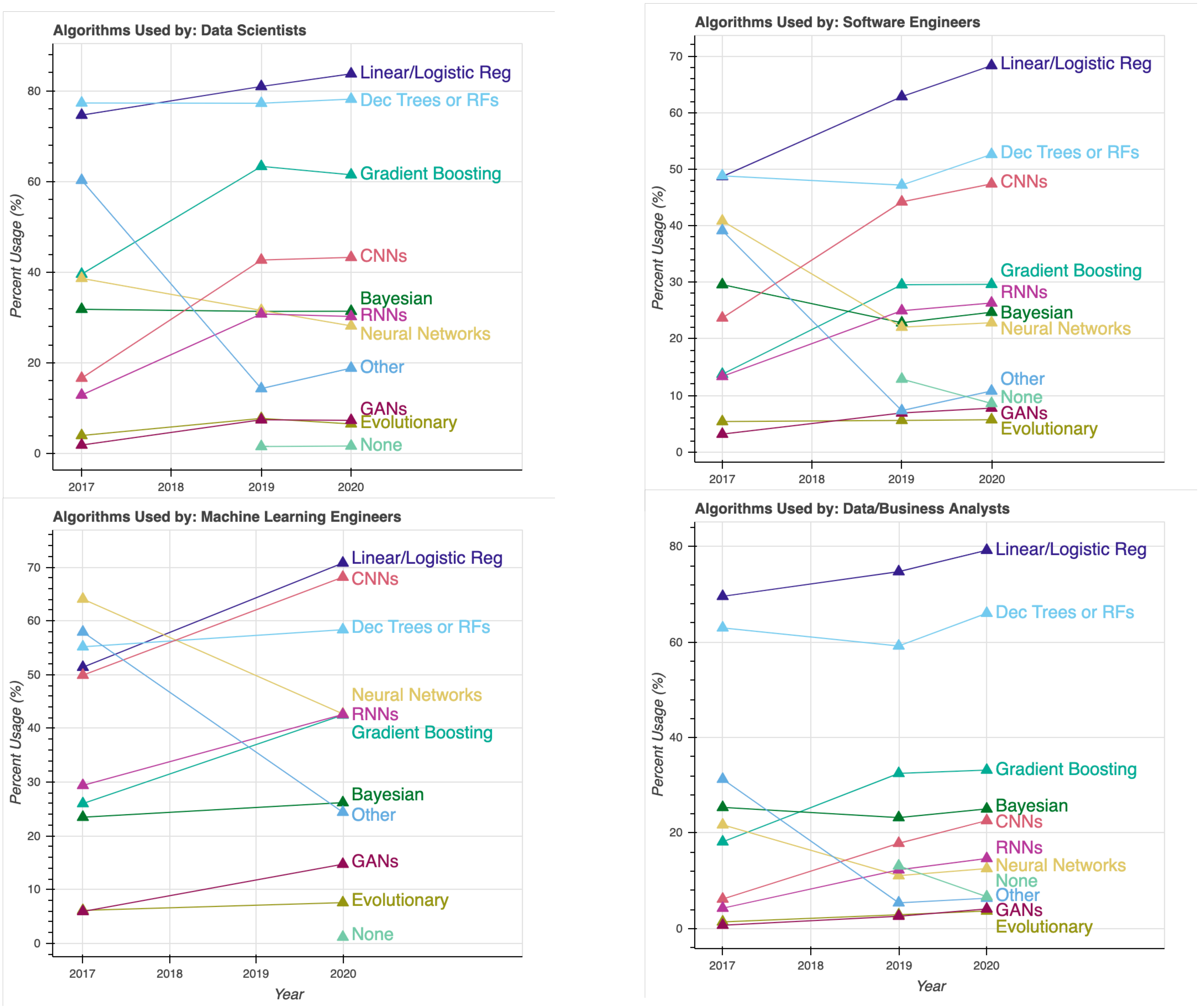
	CA ↓	AD ↑	HE ↑	JA ↑	HI ↑	AL ↑	EP ↑	YE ↓	CO ↑	YA ↓	MI ↓
Default hyperparameters											
XGBoost	0.462	**0.874**	0.348	0.711	0.717	0.924	0.8799	9.192	0.964	0.761	0.751
CatBoost	**0.428**	0.873	0.386	0.724	0.728	0.948	0.8893	8.885	0.910	0.749	0.744
FT-Transformer	0.454	0.860	**0.395**	**0.734**	**0.731**	**0.966**	**0.8969**	**8.727**	**0.973**	**0.747**	**0.742**
Tuned hyperparameters											
XGBoost	0.431	0.872	0.377	0.724	0.728	–	0.8861	8.819	0.969	**0.732**	0.742
CatBoost	**0.423**	**0.874**	0.388	0.727	0.729	–	0.8898	8.837	0.968	0.740	**0.741**
ResNet	0.478	0.857	0.398	0.734	0.731	0.966	0.8976	8.770	0.967	0.751	0.745
FT-Transformer	0.448	0.860	**0.398**	**0.739**	**0.731**	**0.967**	**0.8984**	**8.751**	**0.973**	0.747	0.743

Deep learning

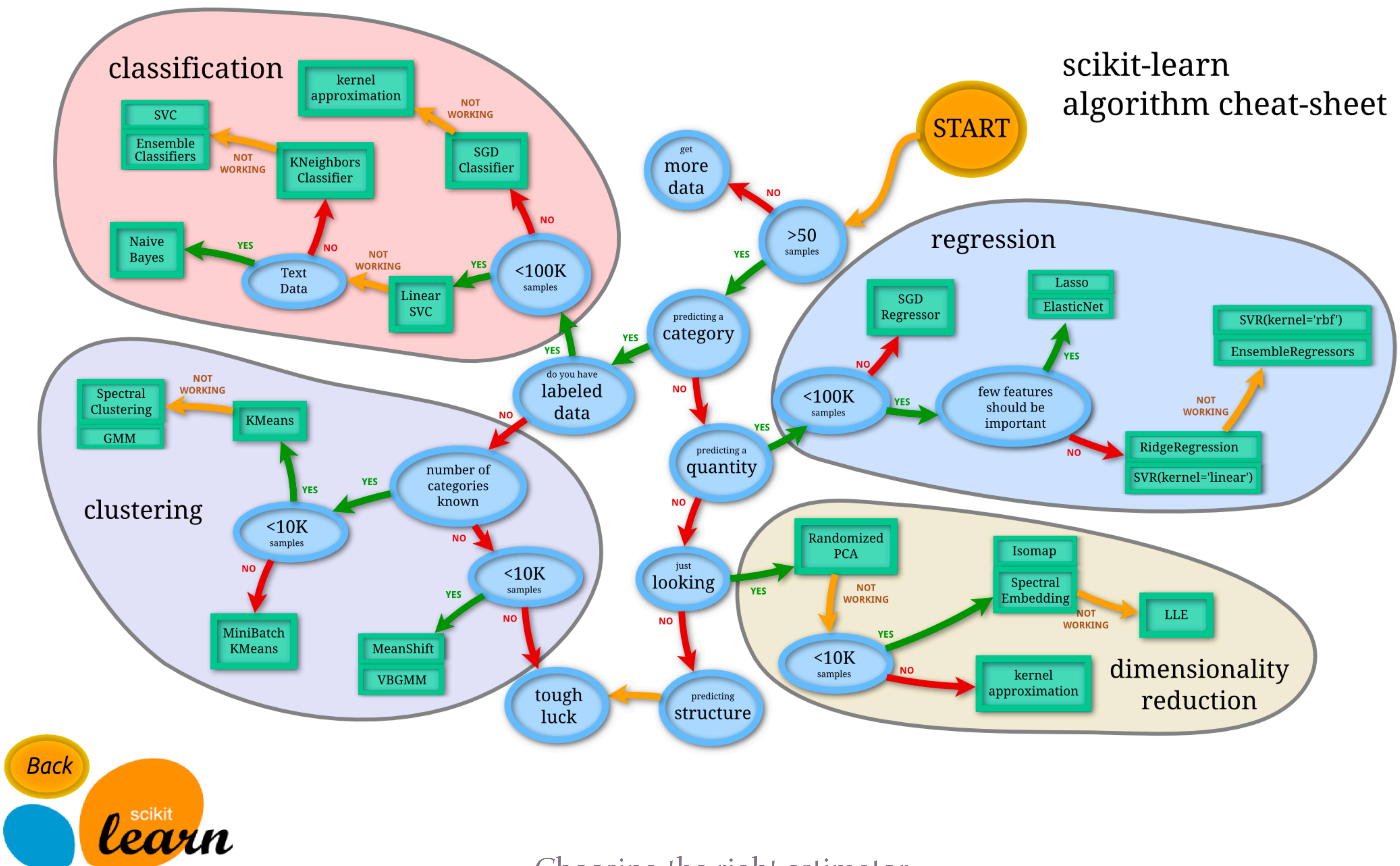
Deep learning

No single best method

Revisiting Deep Learning Models for Tabular Data


Algorithms Used by: Data Scientists
Percent Usage (%)
2017
2018
2019
2020
Linear/Logistic Reg
Dec Trees or RFs
Gradient Boosting
CNNs
Bayesian
RNNs
Neural Networks
Other
GANs
Evolutionary
None
Algorithms Used by: Software Engineers
Percent Usage (%)
Linear/Logistic Reg
Dec Trees or RFs
CNNs
Gradient Boosting
RNNs
Bayesian
Neural Networks
Other
None
GANs
Evolutionary
Algorithms Used by: Machine Learning Engineers
Percent Usage (%)
Year
Linear/Logistic Reg
CNNs
Dec Trees or RFs
Neural Networks
RNNs
Gradient Boosting
Bayesian
Other
GANs
Evolutionary
None
Algorithms Used by: Data/Business Analysts
Percent Usage (%)
Year
Linear/Logistic Reg
Dec Trees or RFs
Gradient Boosting
Bayesian
CNNs
RNNs
Neural Networks
None
Other
GANs
Evolutionary

Pitfalls: ML Algorithms Are Not One-Size-Fits-All

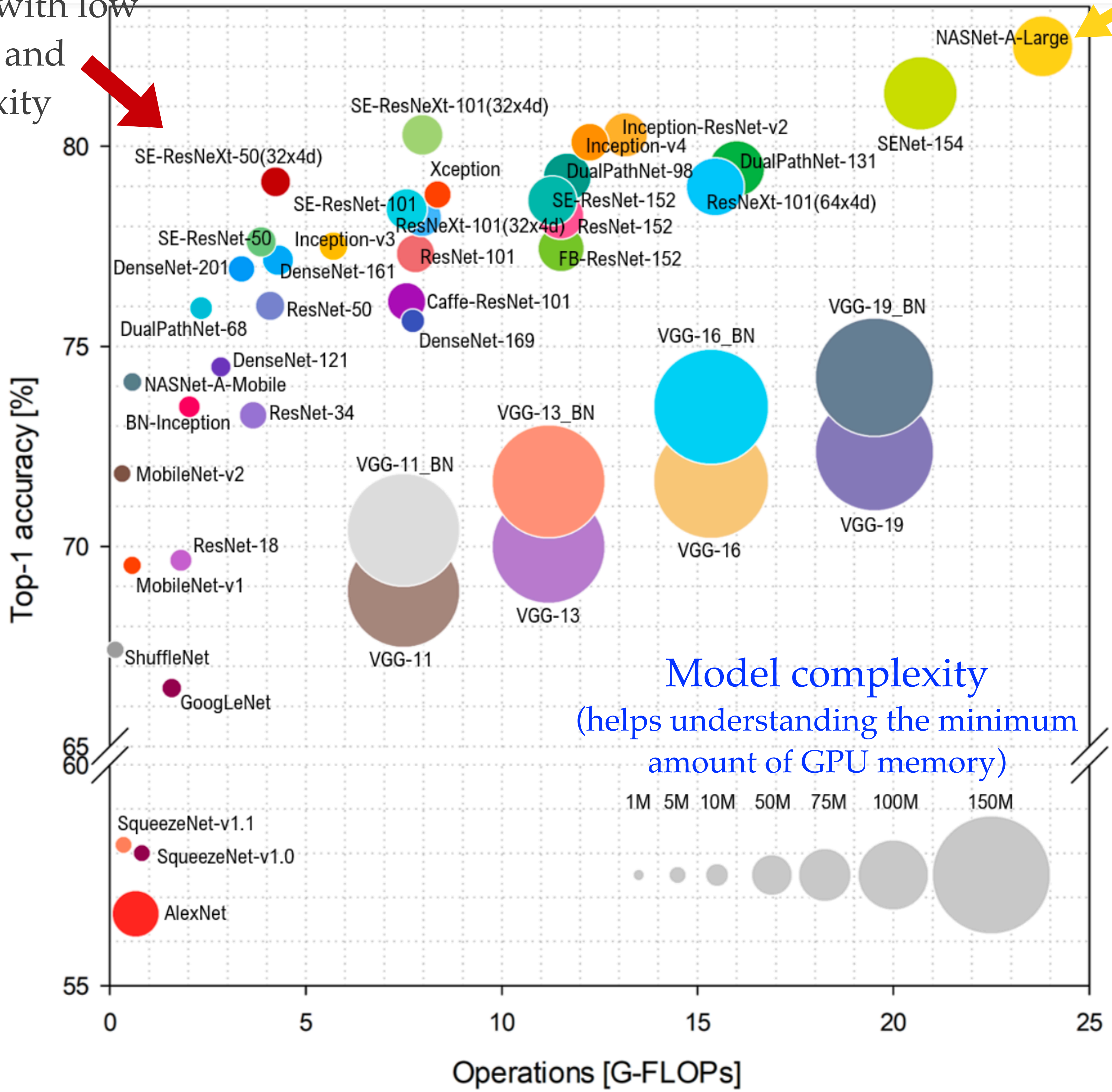


Choosing the right estimator

high accuracy with low computational and model complexity

highest accuracy & computational complexity

ImageNet-1k validation set



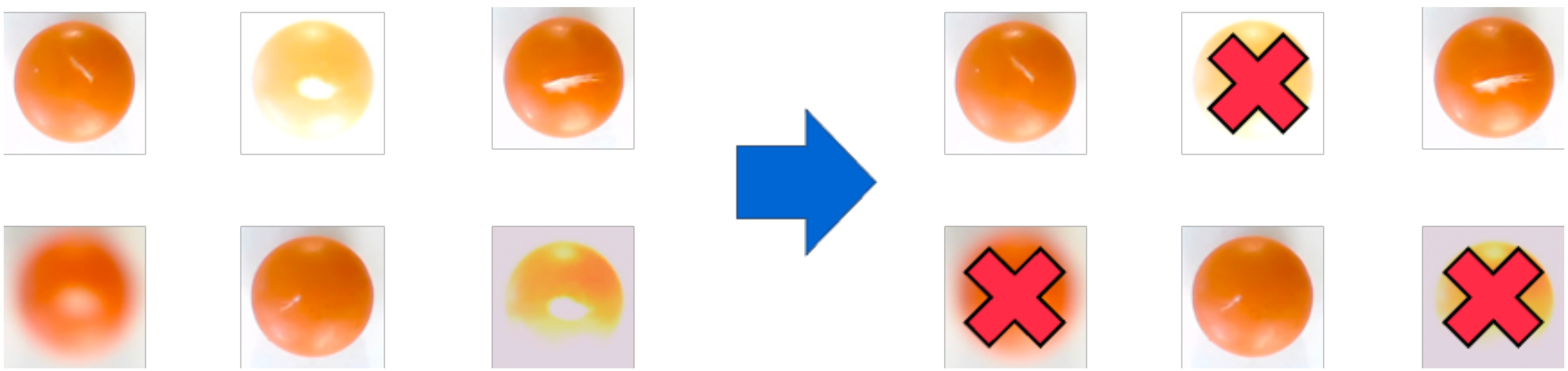
Model complexity (helps understanding the minimum amount of GPU memory)

Floating-point operations (FLOPs) required for a single forward pass (computational cost)

Bianco, Simone, et al. "Benchmark analysis of representative deep neural network architectures." IEEE access 6 (2018): 64270-64277.

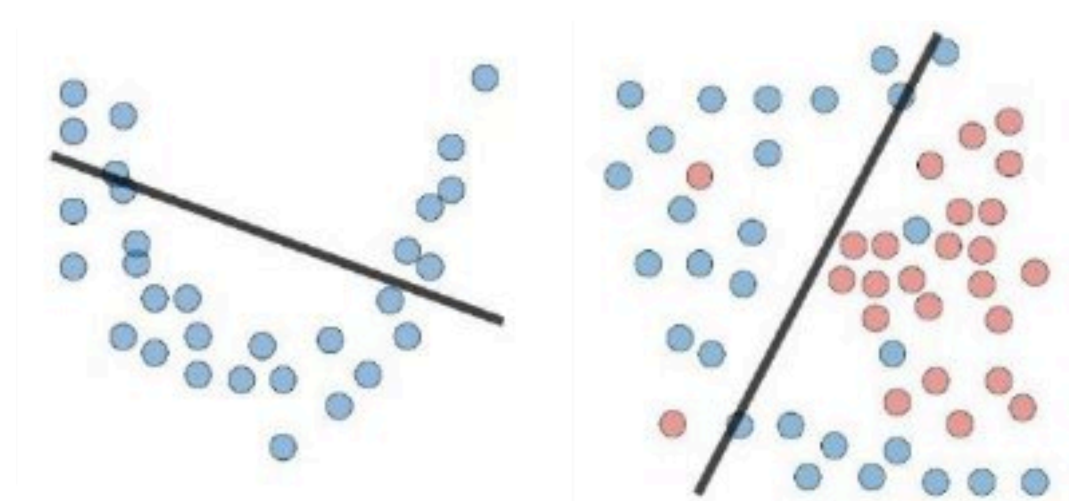
Pitfalls: Having More Data Doesn't Always Help

Low quality/irrelevant samples



Inconsistent labels

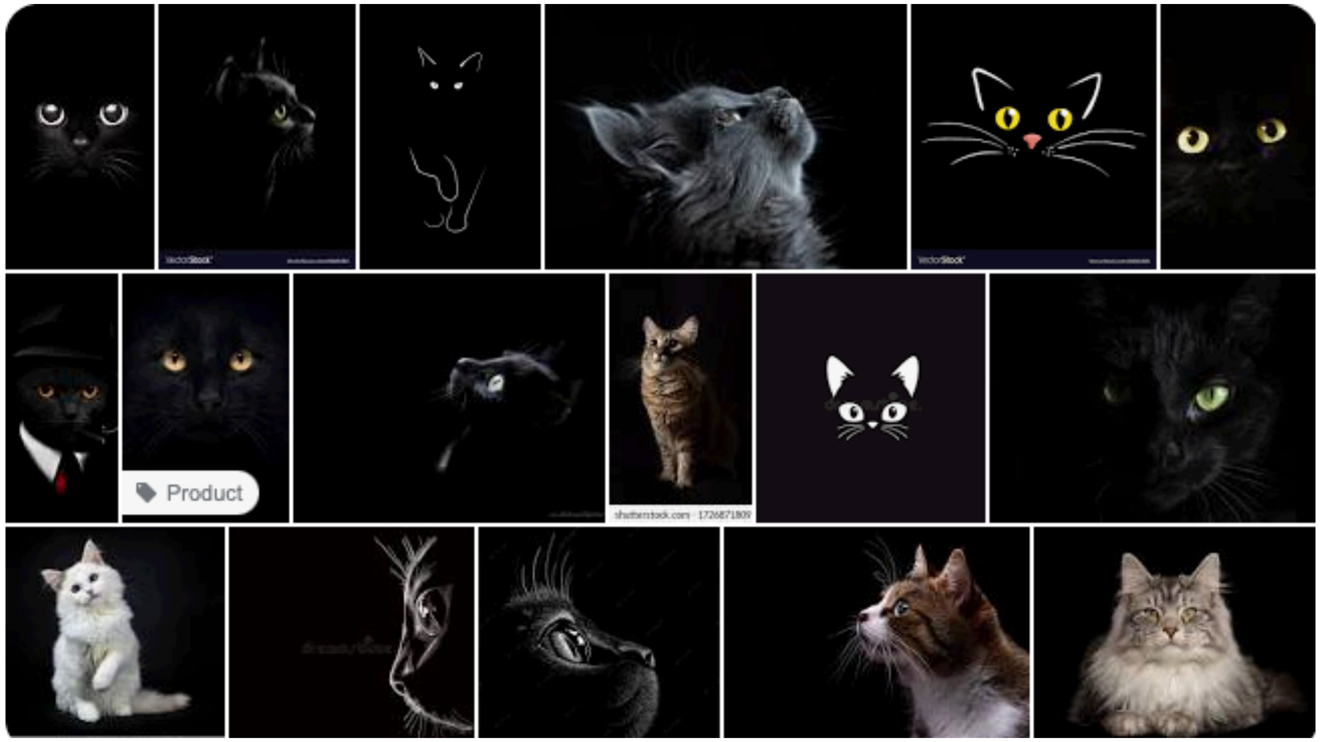
Underfitting



https://landing.ai/tips-for-a-data-centric-ai-approach/

https://i.pinimg.com/originals/72/e2/22/72e222c1542539754df1d914cb671bd7.png

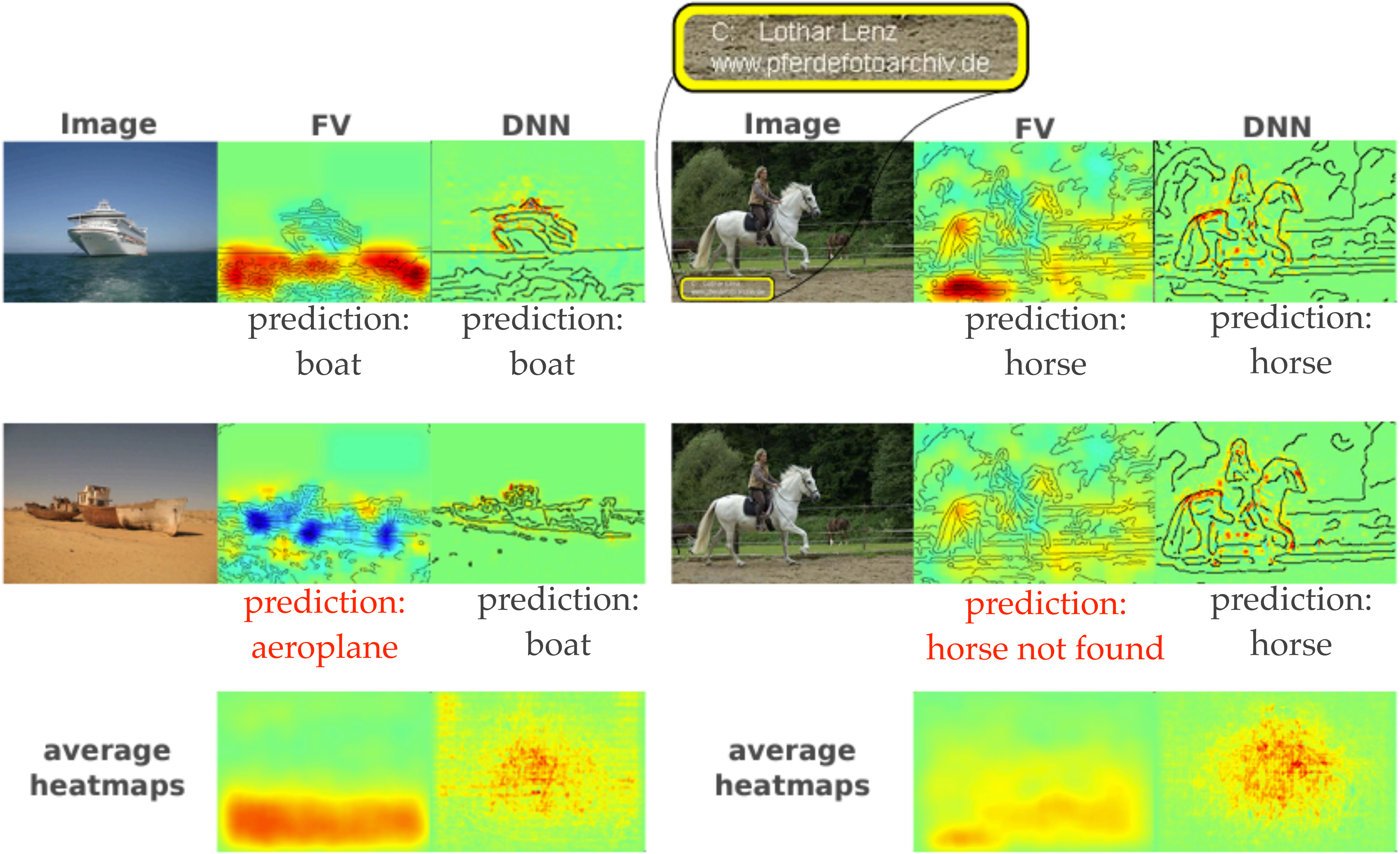
Pitfalls: Exploiting Spurious Correlations



Images obtained using Google Images

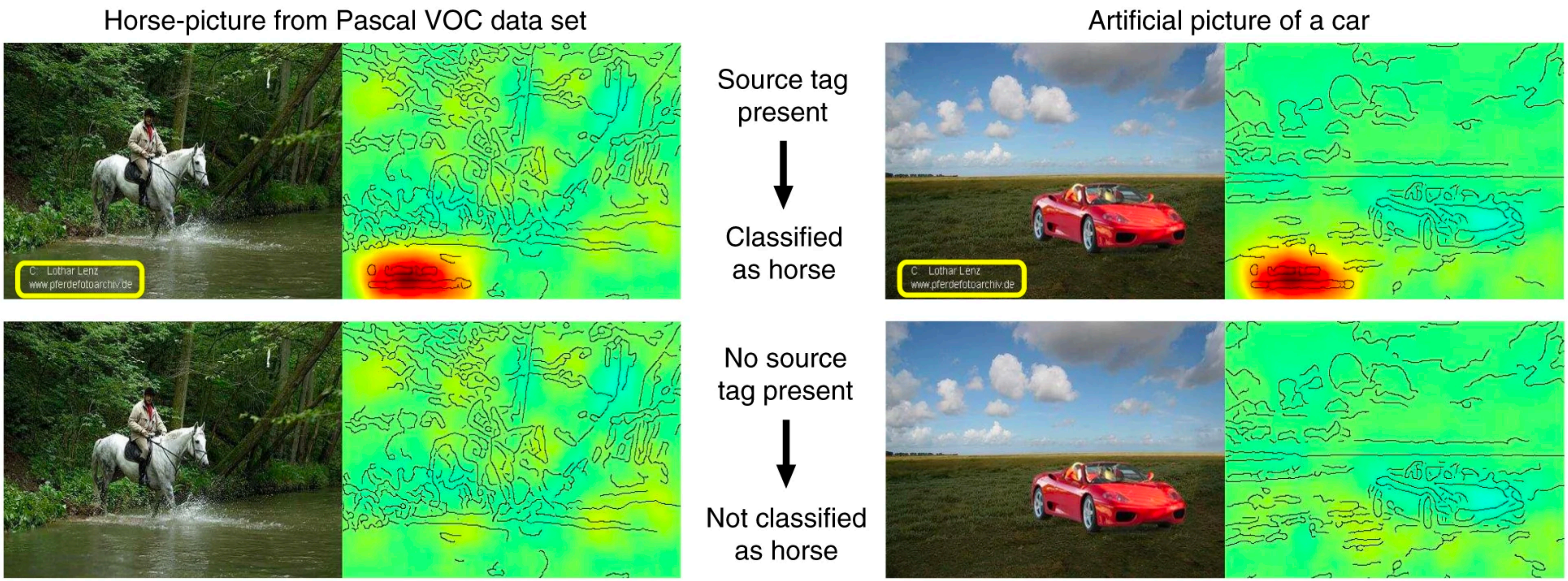
[Machine learning pitfalls]

Pitfalls: Exploiting Spurious Correlations



[Unmasking Clever Hans predictors and assessing what machines really learn]

Pitfalls: Exploiting Spurious Correlations

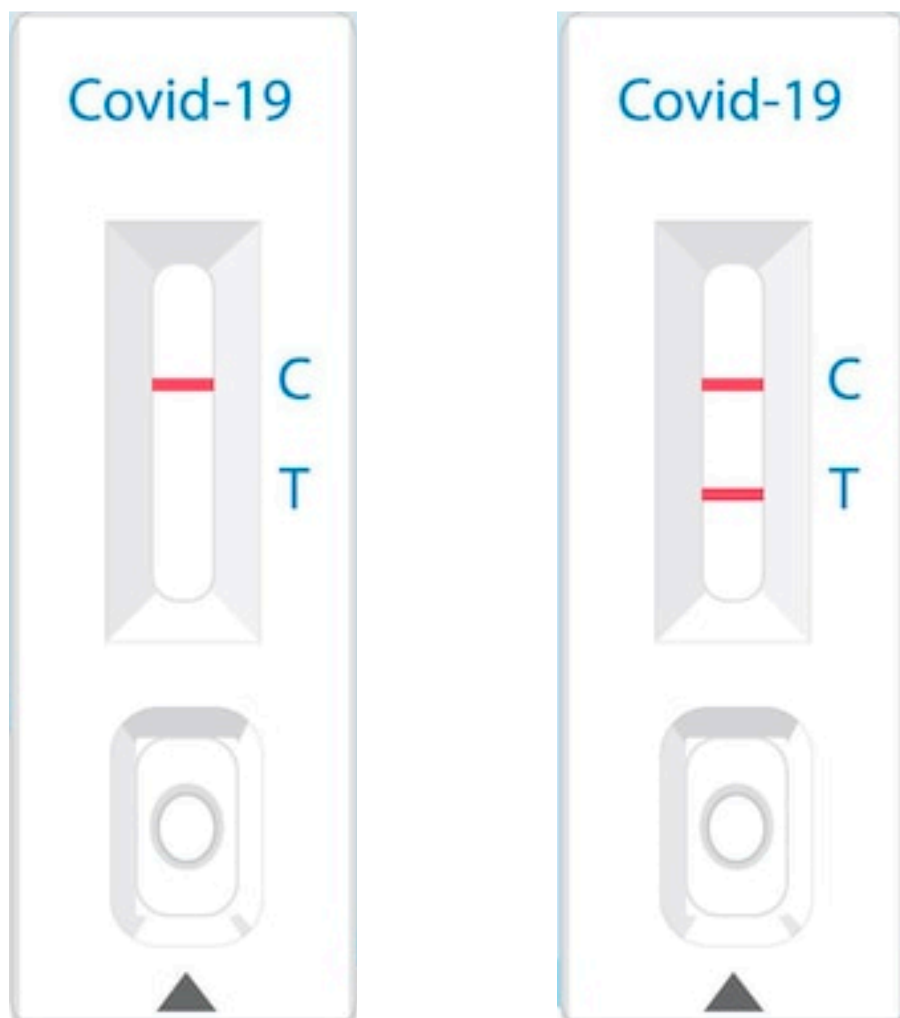


“The Fisher vector classifier trained on the PASCAL VOC 2007 data set focuses on a source tag present in about one-fifth of the horse figures. Removing the tag also removes the ability to classify the picture as a horse. Furthermore, inserting the tag on a car image changes the classification from car to horse.”

[Unmasking Clever Hans predictors and assessing what machines really learn]

Pitfalls: Training Data Don't Capture Real Use Cases

Training data

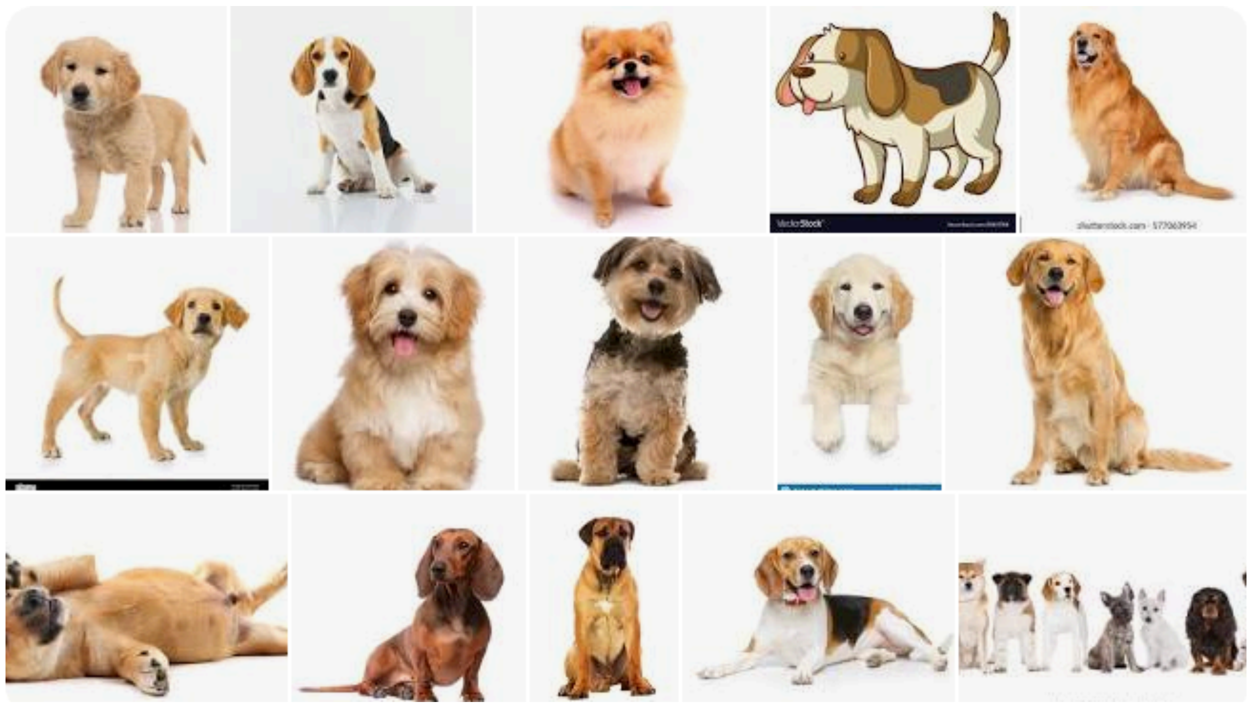


Test data

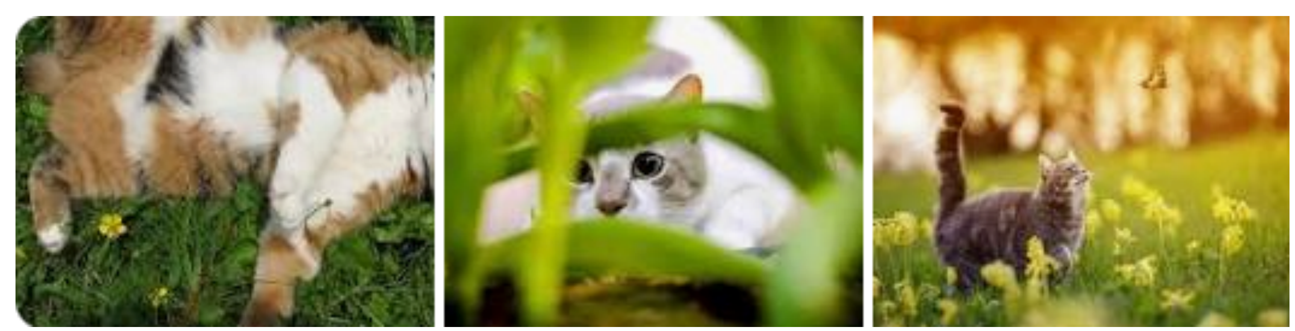
Images by brgfx and kjpargeter on Freepik

Pitfalls: Training Data Don't Capture Real Use Cases

Training data

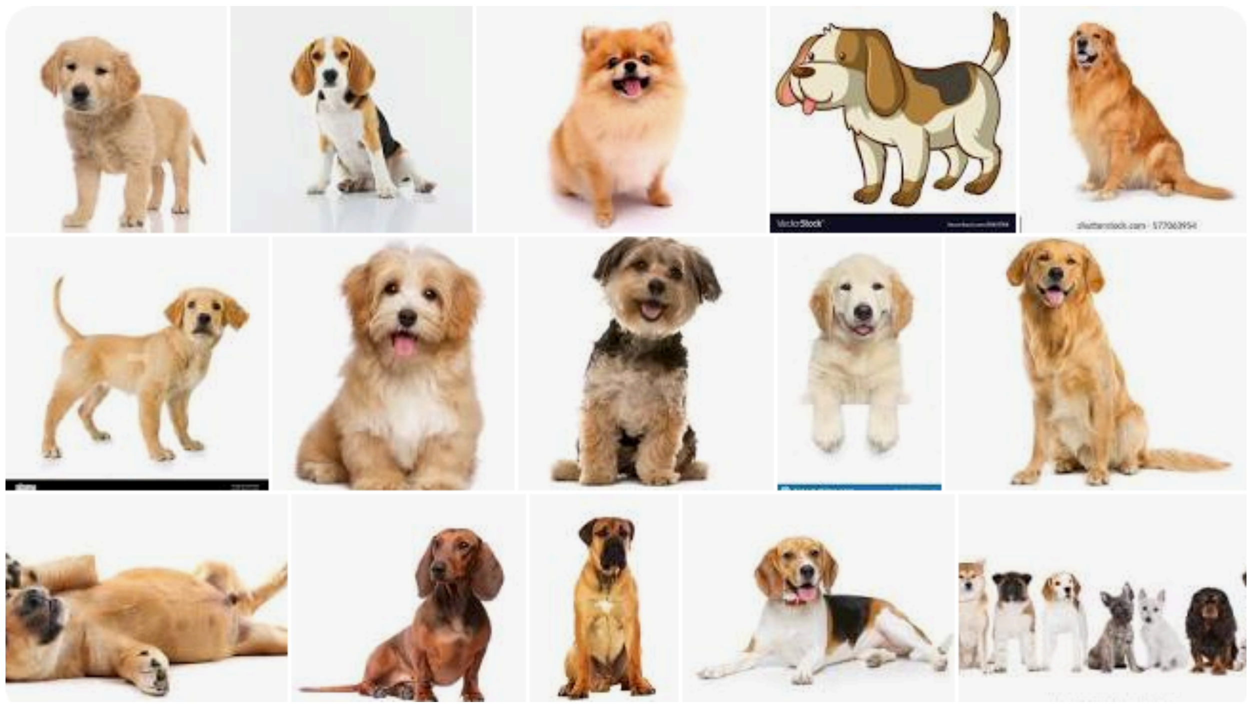


Test data

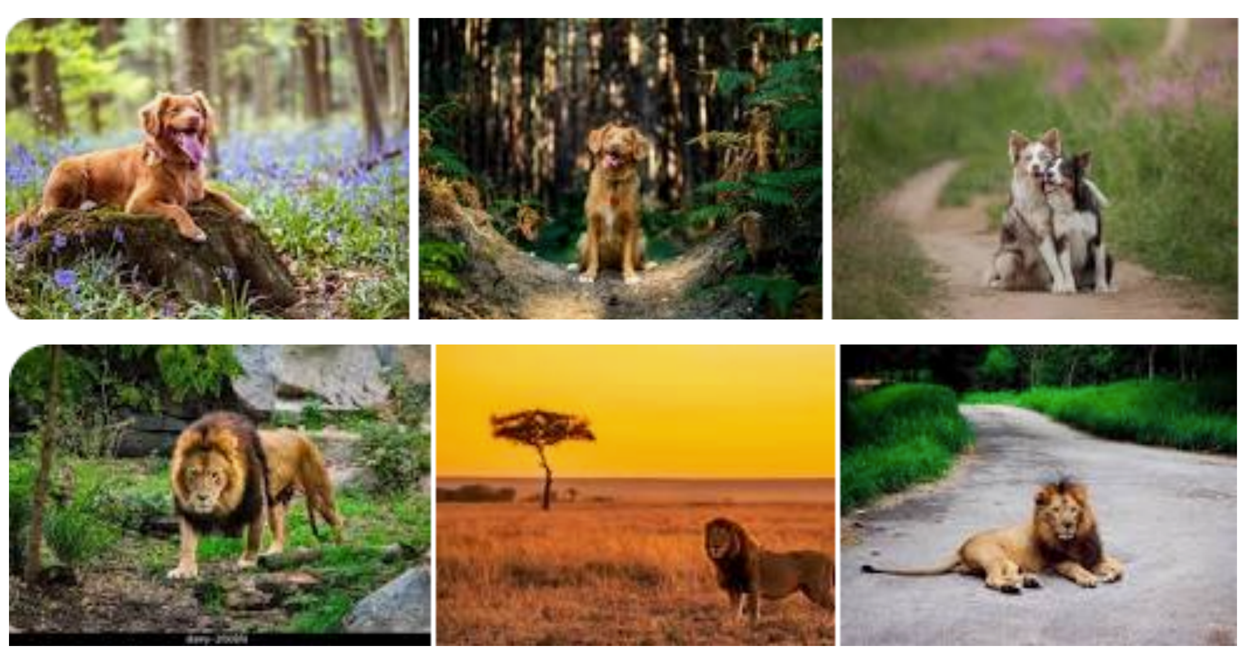


Pitfalls: Training Data Don't Capture Real Use Cases

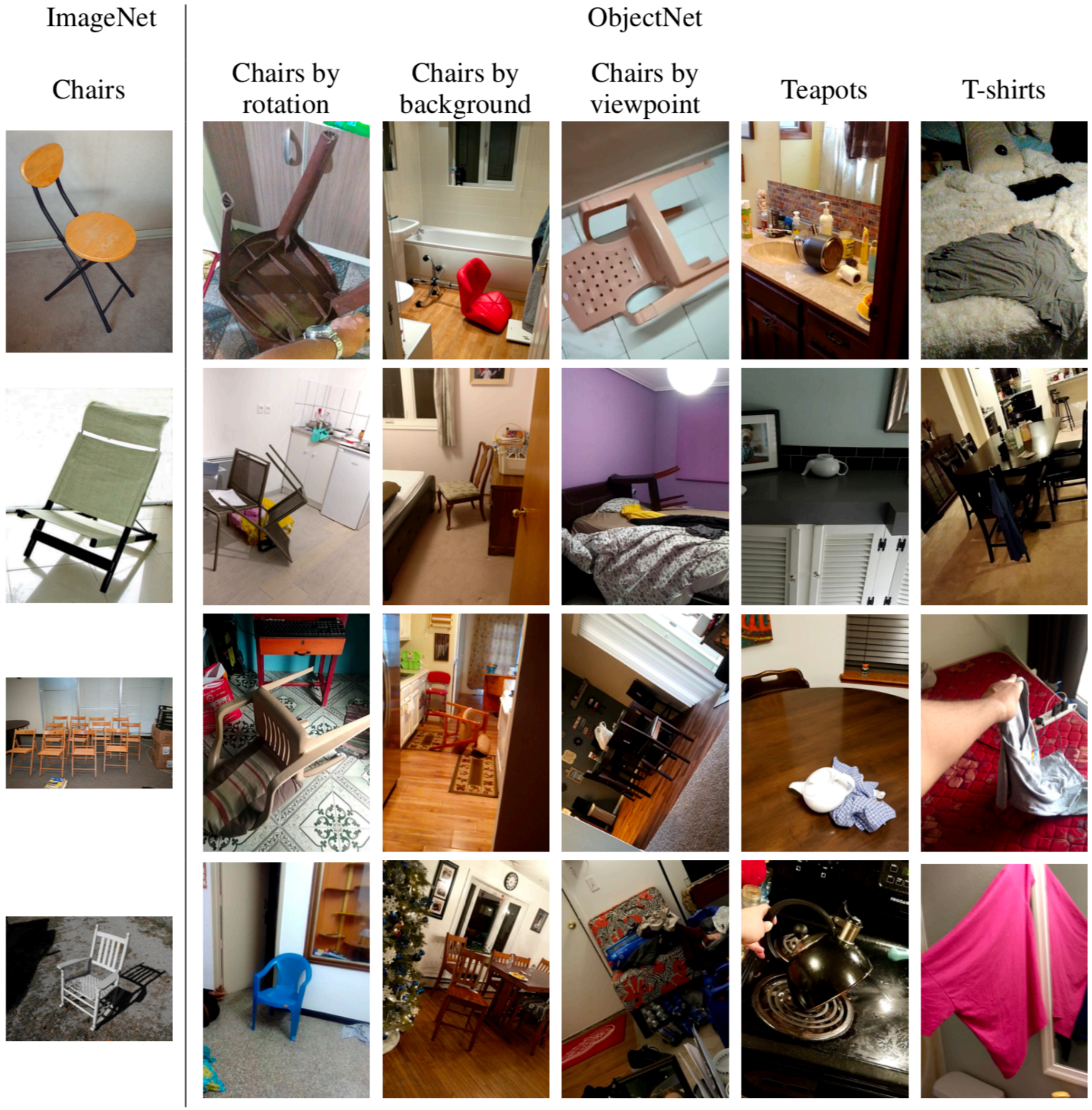
Training data



Test data

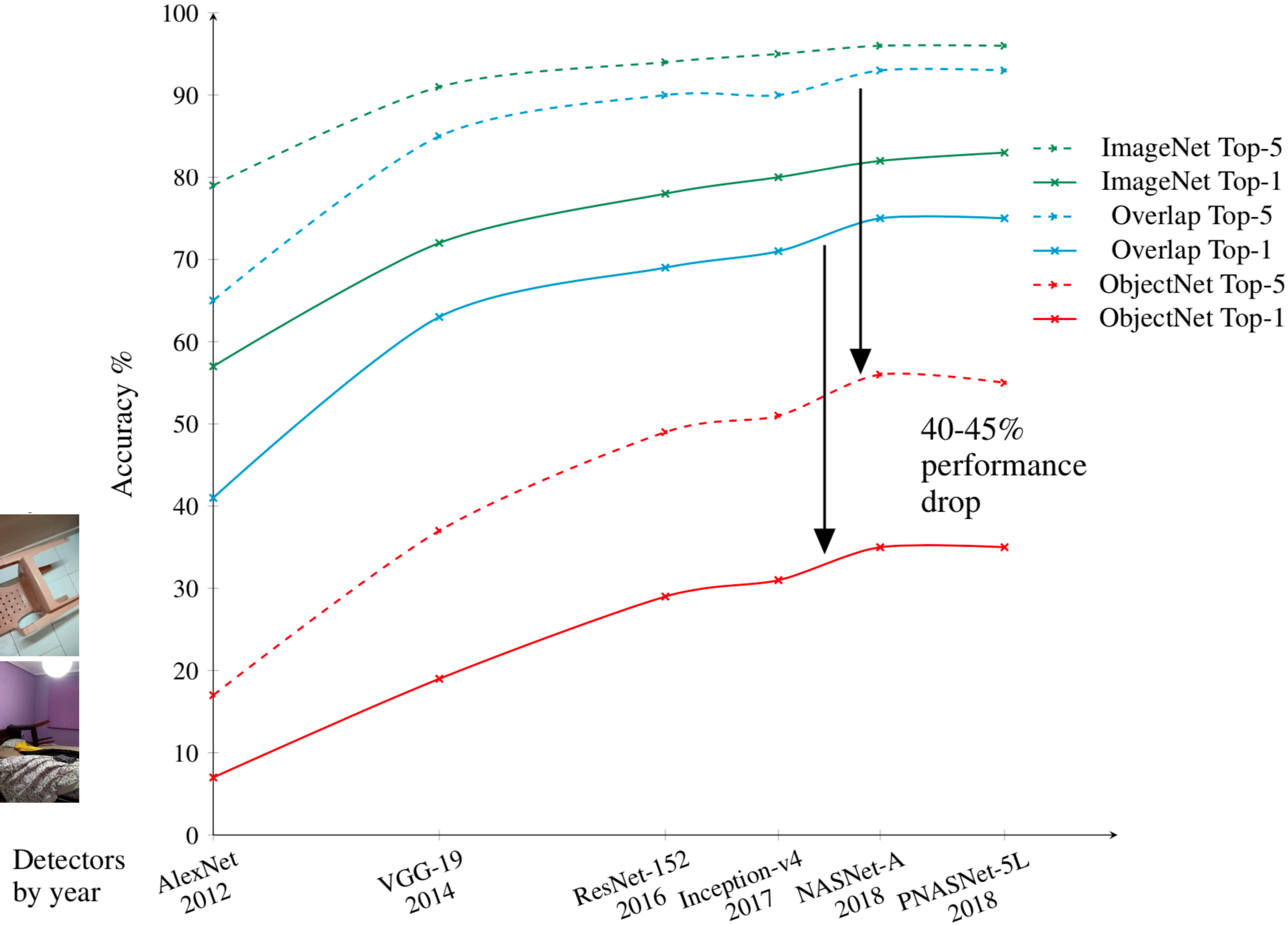


Pitfalls: Training Data Don't Capture Real Use Cases



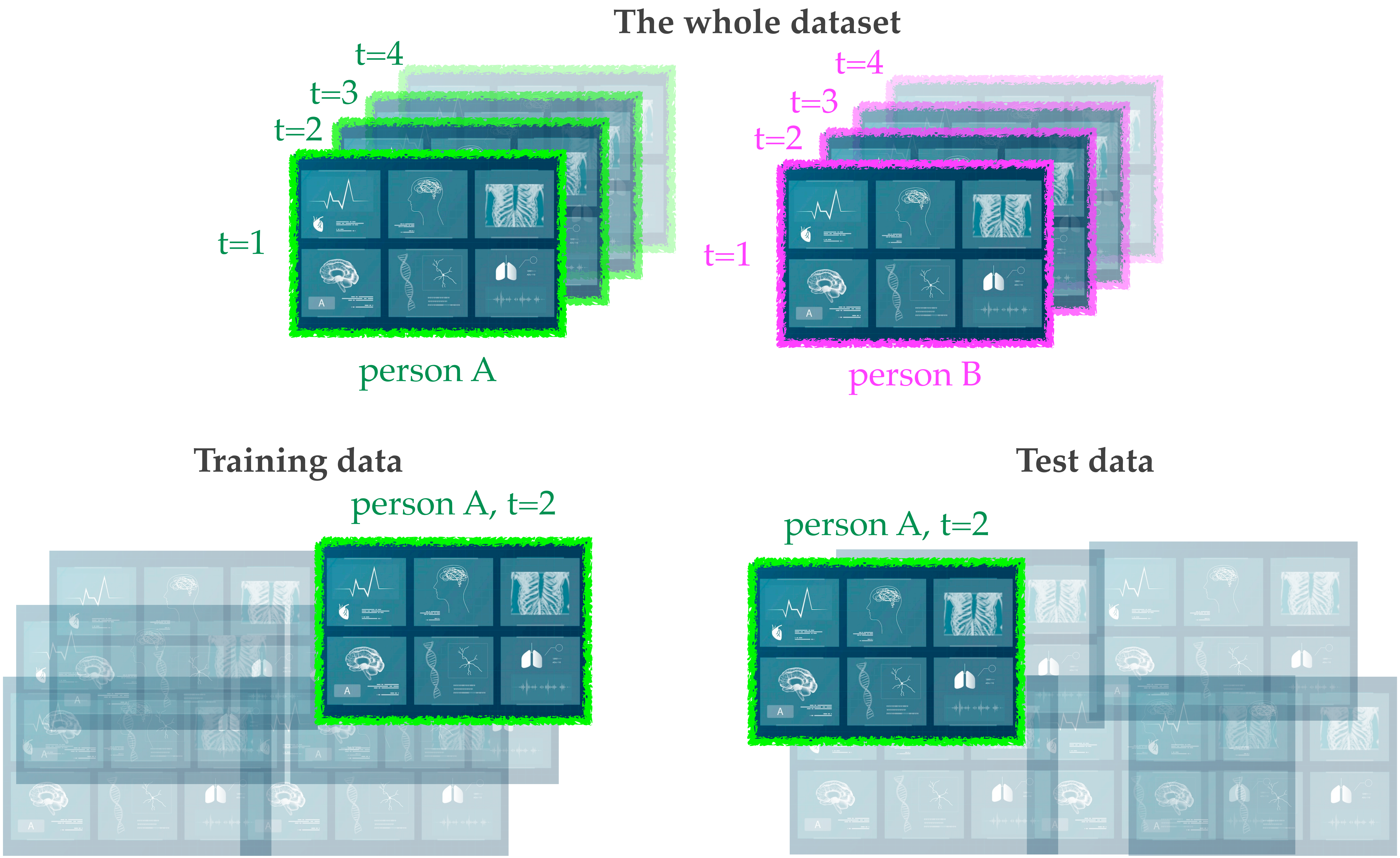
https://objectnet.dev/

Pitfalls: Training Data Don't Capture Real Use Cases



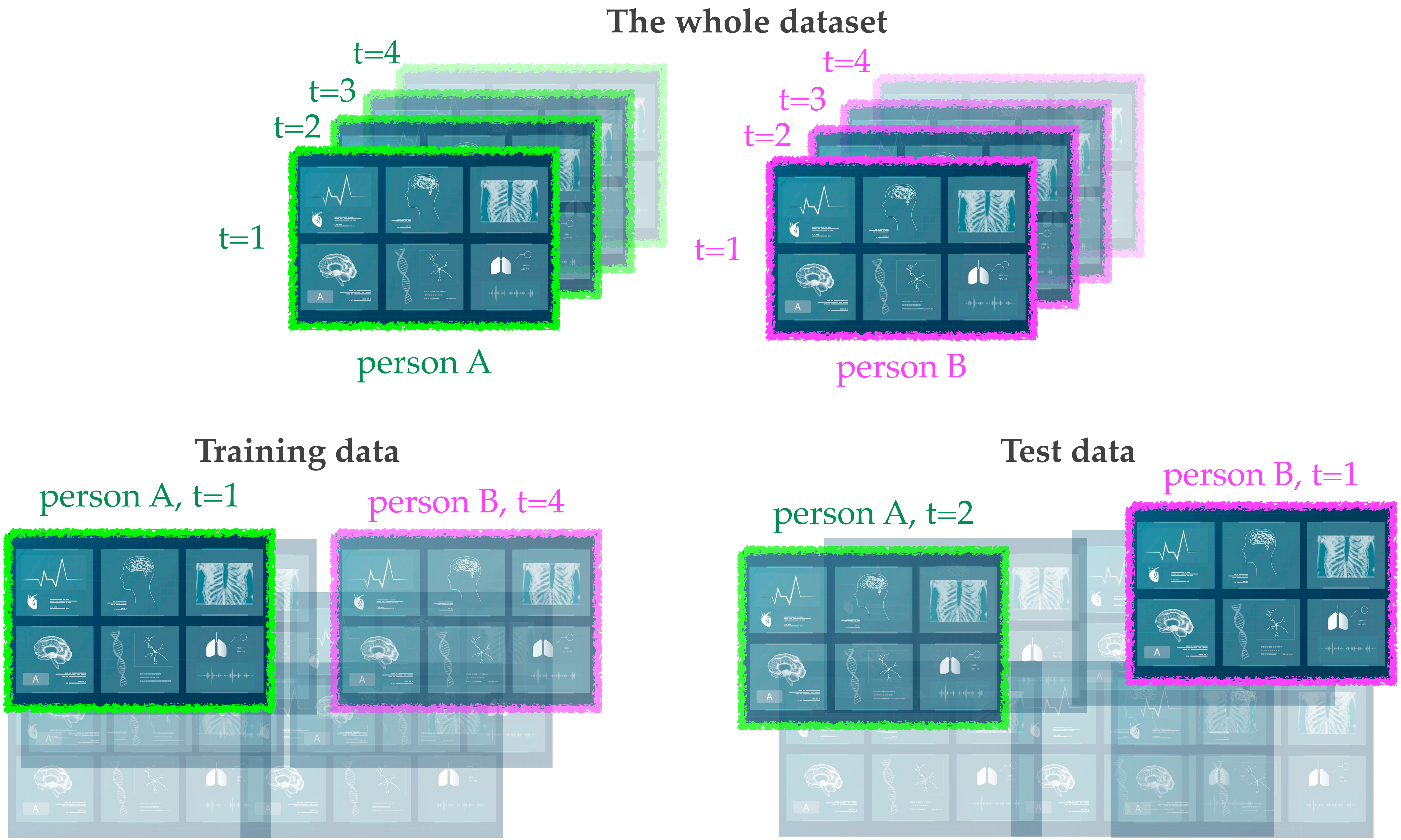
https://objectnet.dev/

Pitfalls: Having Data Leakage



Some of the training and test samples are identical

Pitfalls: Having Data Leakage

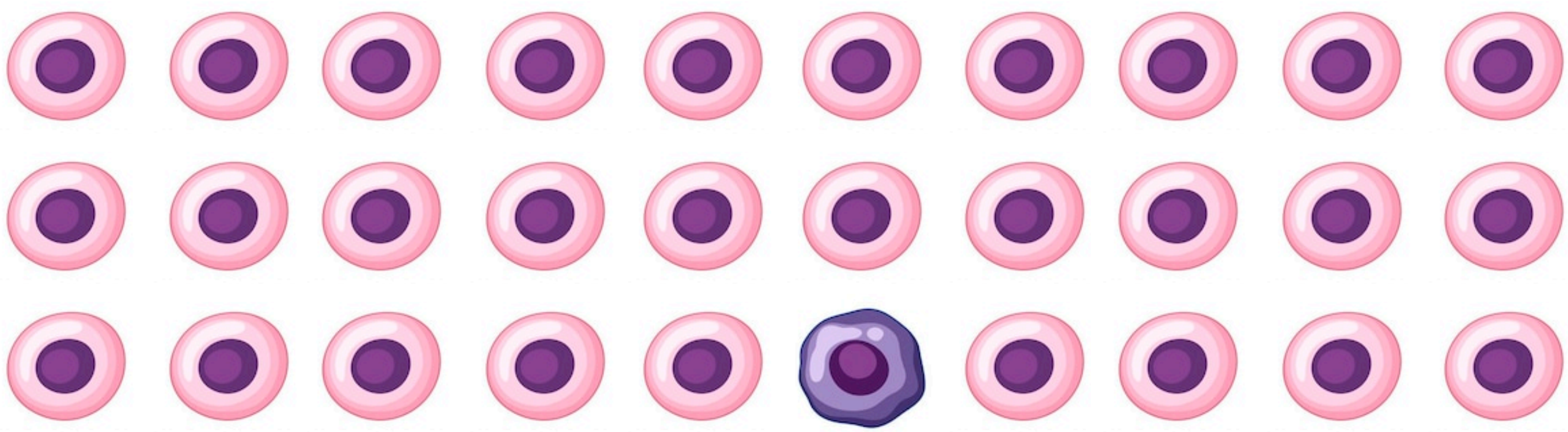


Images from the same person appear in both the training and test data (with different t)

Pitfalls: Reporting a “Non-informative” Evaluation Metric

Normal cell   Cancer cell

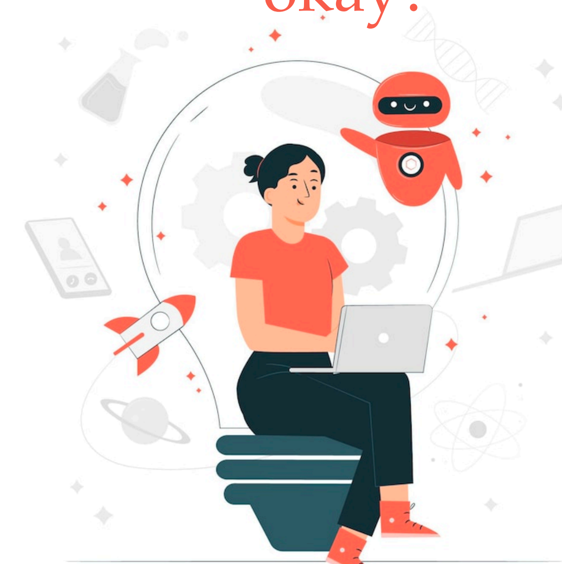
Test data



We got only one test sample wrong, and achieved the classification accuracy of **96.67%**!

This is much higher than the accuracy of a random guess!

Just say “normal cell”, okay?



Images by brgfx, pct.vector and storyset on Freepik

Pitfalls: Solving “Irrelevant” Problems

MIT Technology Review

OPINION

Too many AI researchers think real-world problems are not relevant

The community’s hyperfocus on novel methods ignores what's really important.

By Hannah Kerner

August 18, 2020

“Many studies applying machine learning to viticulture aim to optimize grape yields, but winemakers “want the right levels of sugar and acid, not just lots of big watery berries,” says Drake Whitcraft of Whitcraft Winery in California.”

“For example, most studies applying deep learning to echocardiogram analysis try to surpass a physician’s ability to predict disease.
But predicting normal heart function would actually save cardiologists more time by identifying patients who do not need their expertise.”

Machine learning pitfalls

Outline

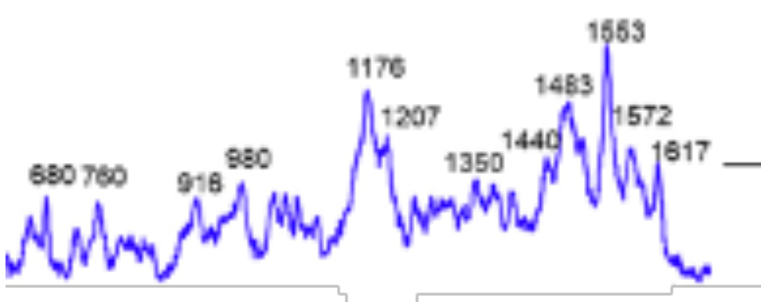
- Machine Learning Pitfalls
  - ML algorithms are not one-size-fits-all
  - Having more data doesn't always help
  - Exploiting spurious correlations
  - Training data don't capture real use cases
  - Having data leakage
  - Reporting a "non-informative" evaluation metric
  - Solving "irrelevant" problems
- A Way to Communicate with ML Practitioners
  - Be clear about what the inputs and outputs are
  - State the requirements and use-cases
  - Provide domain-/field-specific information
  - Summary of related papers
- Resources

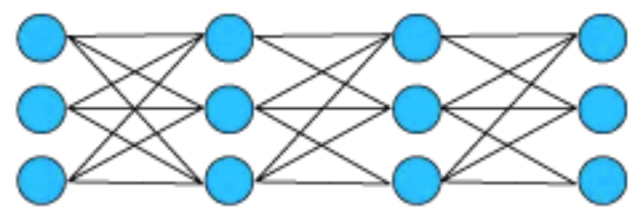
A Way to Communicate with ML Practitioners

- Be clear about what the inputs and outputs are

Spectrum classification

Input: a list of numbers





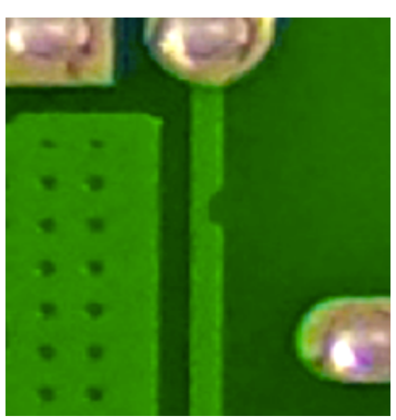
$$f_W(x)$$

Output

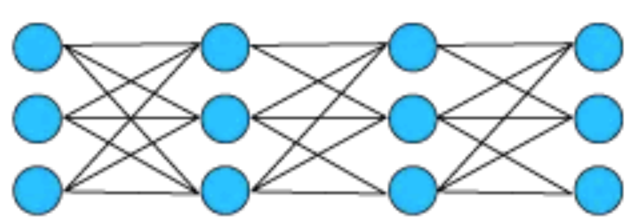
Category

Object detection

Input: an image



Huang, Weibo, and Peng Wei. "A PCB dataset for defects detection and classification." arXiv preprint arXiv:1901.08204 (2019).

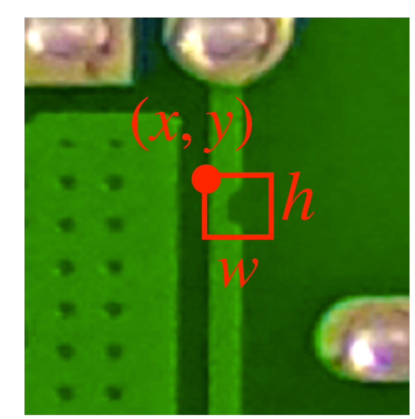


$$f_W(x)$$

Output: rectangular box(es) and the corresponding class(es)

$$\begin{bmatrix} x \\ y \\ w \\ h \end{bmatrix}$$

corresponding class



A Way to Communicate with ML Practitioners

- Be clear about what the inputs and outputs are

Molecule classification

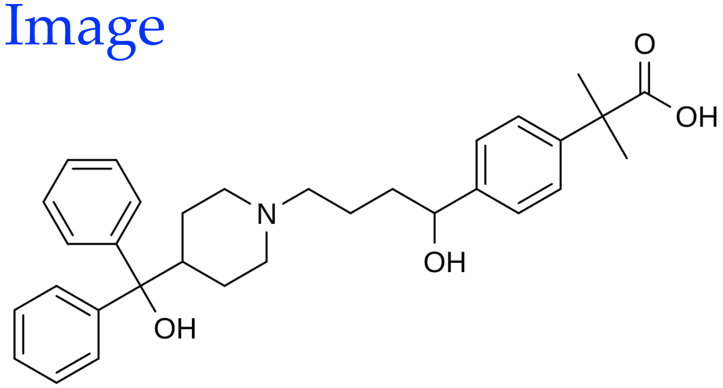
Input

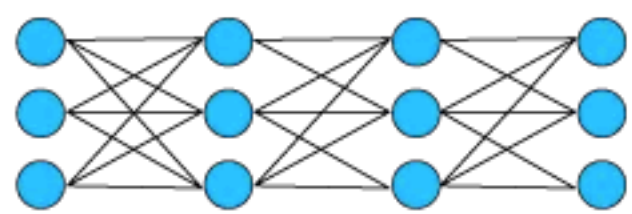
Output

String of characters

$C_{32}H_{39}NO_4$

Image

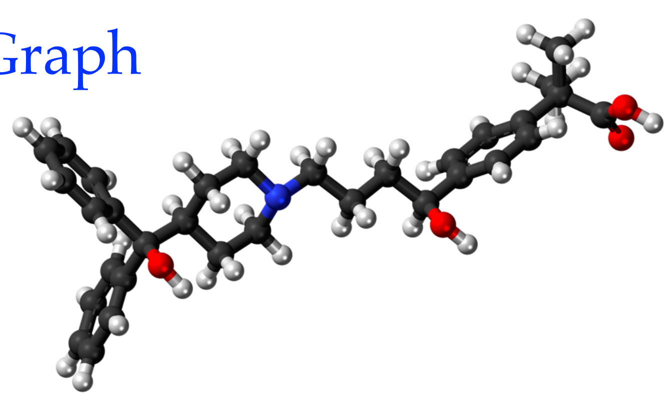




$f_W(x)$

2nd-generation antihistamine drug

Graph



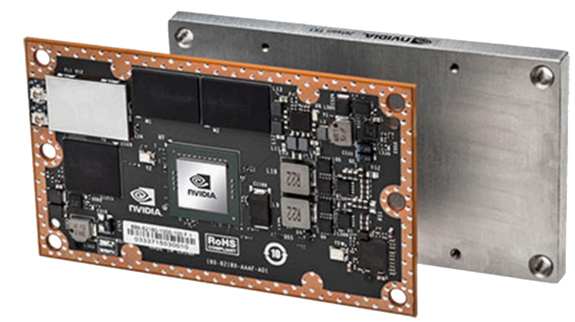
A Way to Communicate with ML Practitioners

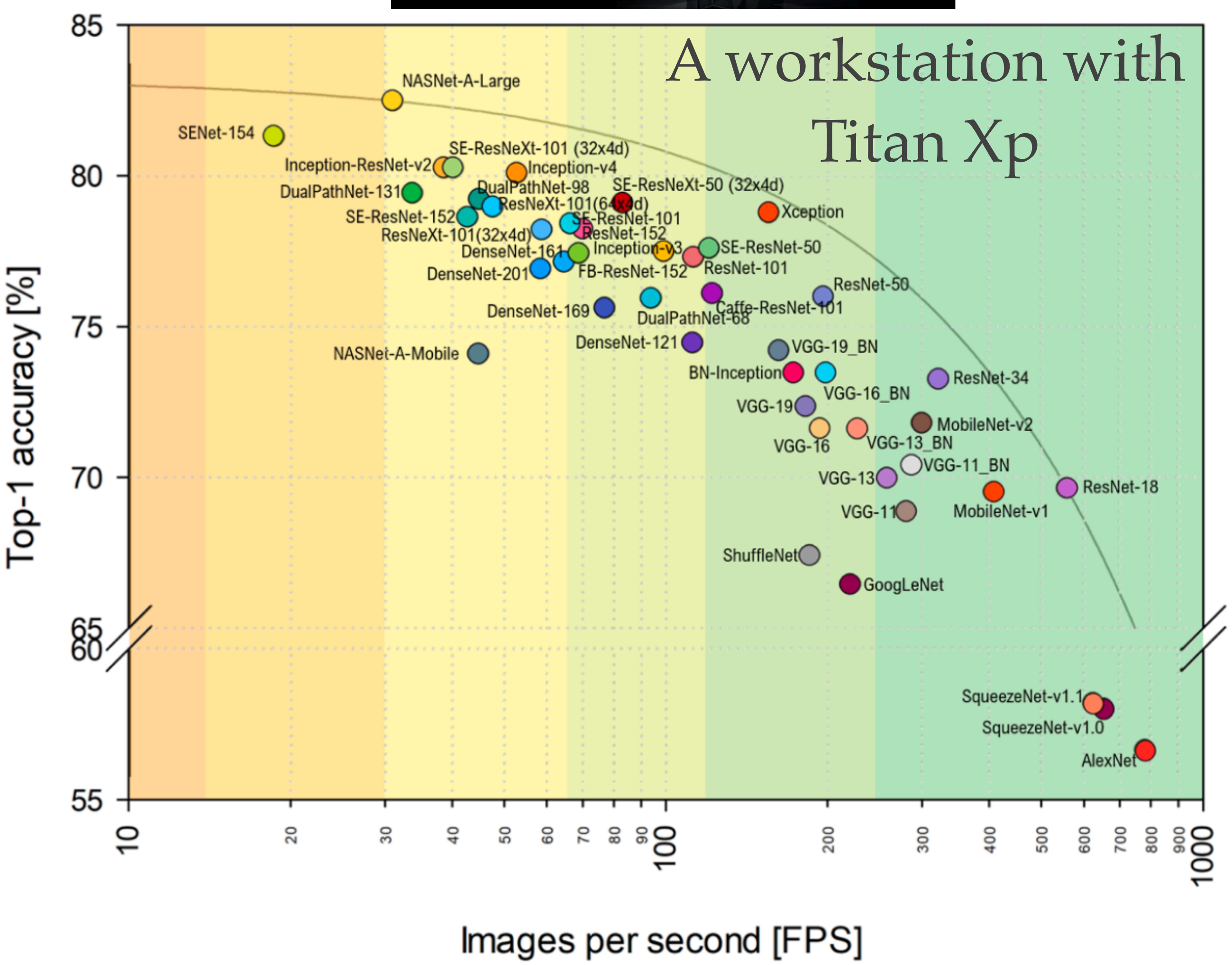
- State the requirements and use-cases
  - Computational resources: computer servers, workstations, IoT devices, smartphones
  - Model deployment
    - Web/mobile applications

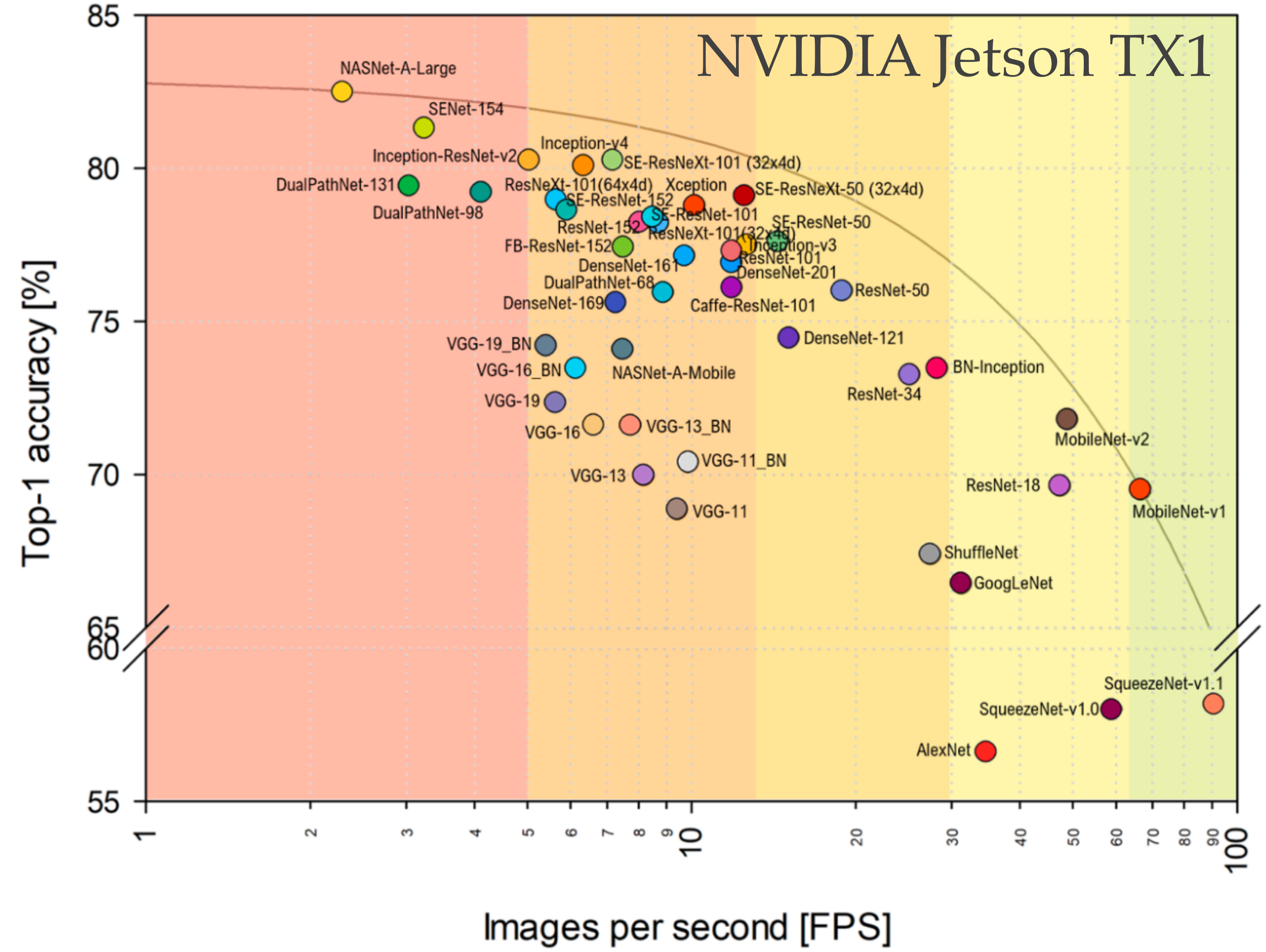
  - Required processing speed
    - How fast do you need it to run? How many samples per second? Real-time?

A Way to Communicate with ML Practitioners

- State the requirements and use-cases







The desired throughput can be used to pick a model on a specific platform

Bianco, Simone, et al. "Benchmark analysis of representative deep neural network architectures." IEEE access 6 (2018): 64270-64277.

A Way to Communicate with ML Practitioners

- State the requirements and use-cases
  - Computational resources: computer servers, workstations, embedded systems, IoT devices, smartphones
  - Model deployment
    - Web/mobile applications

  - Required processing speed
    - How fast do you need it to run? How many samples per second? Real-time?
  - Number of samples (ideally with labels) that you can prepare to train a model

Please also be realistic and be aware of ML limitations

A Way to Communicate with ML Practitioners

- Only after the ML practitioners get the big picture (task, inputs, outputs) along with the requirements correctly, you can start providing domain-/field-specific information
  - What could be a good feature for this task?
    - While there are methods that can extract feature automatically, having good/relevant handcrafted features is extremely helpful, especially when it is difficult to prepare large amount of data to train a model

A Way to Communicate with ML Practitioners

- Only after the ML practitioners get the big picture (task, inputs, outputs) along with the requirements correctly, you can start providing domain-/field-specific information
  - What are meaningful evaluation metrics for your application?
    - Classification accuracy, false positive, false negative, ROC, AUC, sensitivity, specificity, f1-score, MSE, MAE, MAPE

Highly imbalanced data

Ex. Two-class classification (5 samples from class 1 and 95 samples from class 2)

A "model" that always classifies any sample as class 2 already achieves 95% accuracy

Is classification accuracy a good metric?

ATK

Which one do you consider worse?
1. Diagnose a healthy person as "positive"
2. Diagnose a person with COVID-19 as "negative"

A Way to Communicate with ML Practitioners

- If there are papers that tackle related / similar tasks, you can provide the ML practitioners with the references
  - It would be even better if you also provide a brief summary of the papers, possibly by answering some of these questions
    - What is the task?
    - What are the inputs and outputs? What are the features?
    - What machine learning methods do they use?
    - How many samples do they use for training and testing the method(s)?
      - How do they split the data? If they need to pick hyper-parameters, do they use validation data?
    - What are the evaluation metrics that they use? Do they make sense?
    - How many times do they run their experiments?
    - What are the limitations of their methods? Frequently, many papers are able to report high performance because they restrict the use-cases

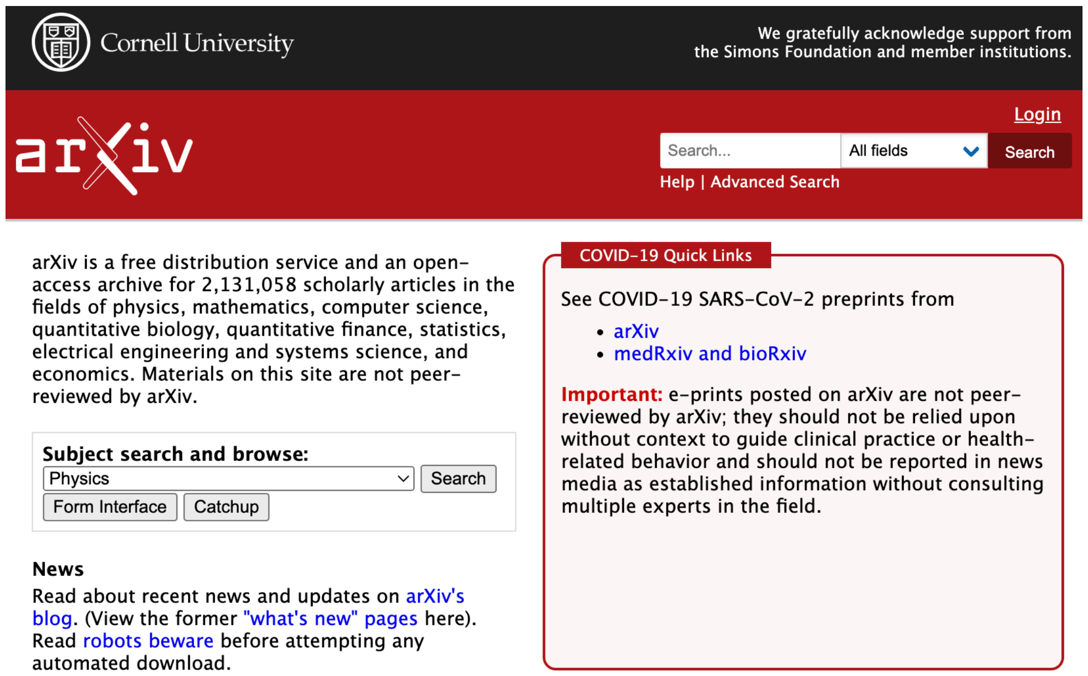
Outline

- Machine Learning Pitfalls
  - ML algorithms are not one-size-fits-all
  - Having more data doesn't always help
  - Exploiting spurious correlations
  - Training data don't capture real use cases
  - Having data leakage
  - Reporting a "non-informative" evaluation metric
  - Solving "irrelevant" problems
- A Way to Communicate with ML Practitioners
  - Be clear about what the inputs and outputs are
  - State the requirements and use-cases
  - Provide domain-/field-specific information
  - Summary of related papers
- Resources

Resources

Conferences

arXiv



Resources

CVPR 2016 open access

These CVPR 2016 papers are the Open Access versions, provided by the Computer Vision Foundation. Except for the watermark, they are identical to the accepted versions; the final published version of the proceedings is available on IEEE Xplore.

This material is presented to ensure timely dissemination of scholarly and technical work. Copyright and all rights therein are retained by authors or by other copyright holders. All persons copying this information are expected to adhere to the terms and constraints invoked by each author's copyright.

Search

Deep Residual Learning for Image Recognition

Kaiming He, Xiangyu Zhang, Shaoqing Ren, Jian Sun; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016, pp. 770-778

Cited **134,083** times (as of September 2022)

Attention is All you Need

Part of Advances in Neural Information Processing Systems 30 (NIPS 2017)

Bibtex Metadata Paper Reviews

Authors

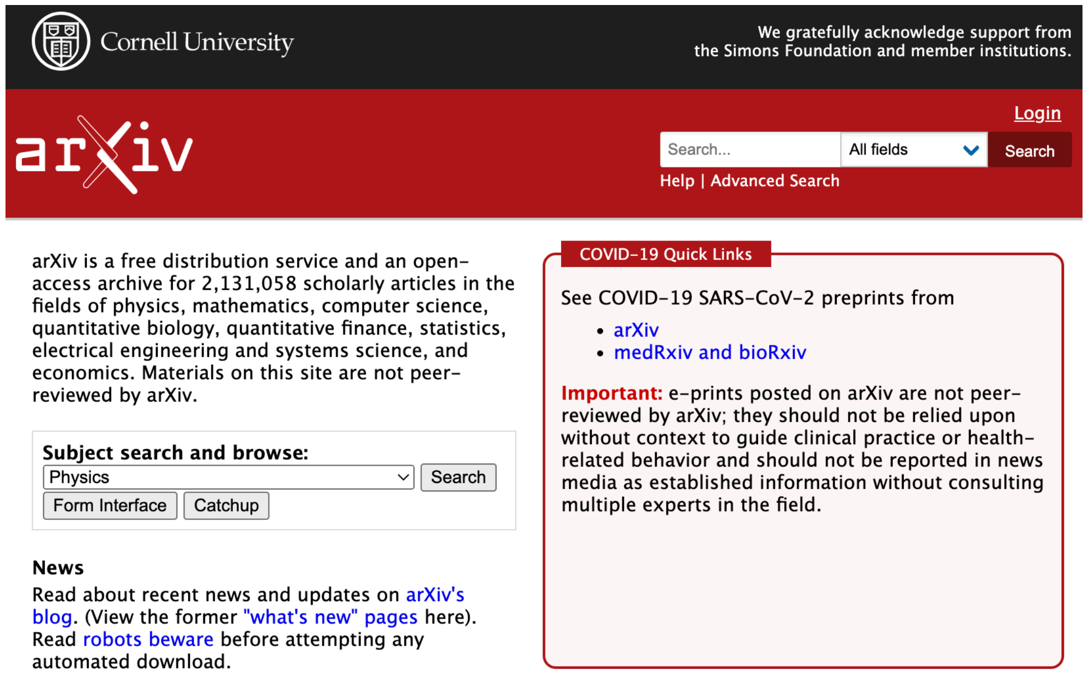
Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, Illia Polosukhin

Cited **52,002** times (as of September 2022)

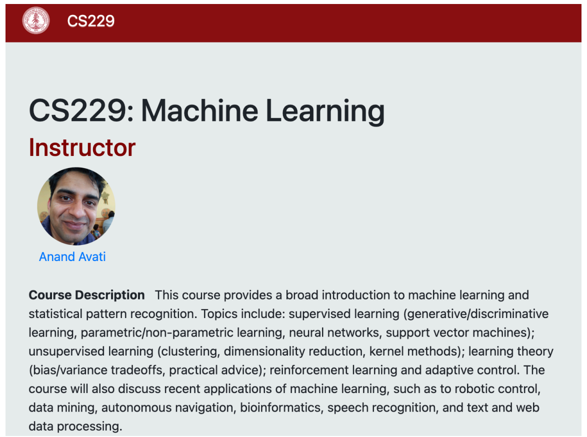
Resources

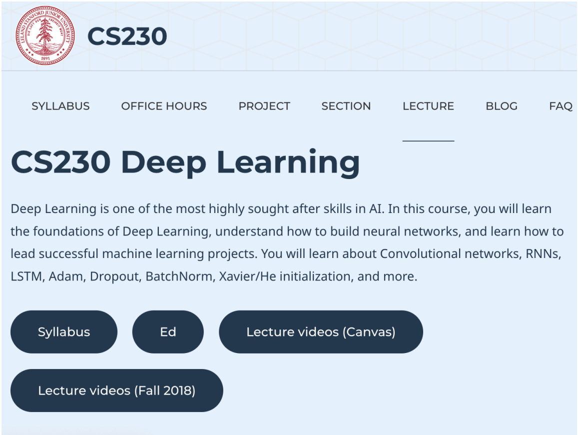
Conferences

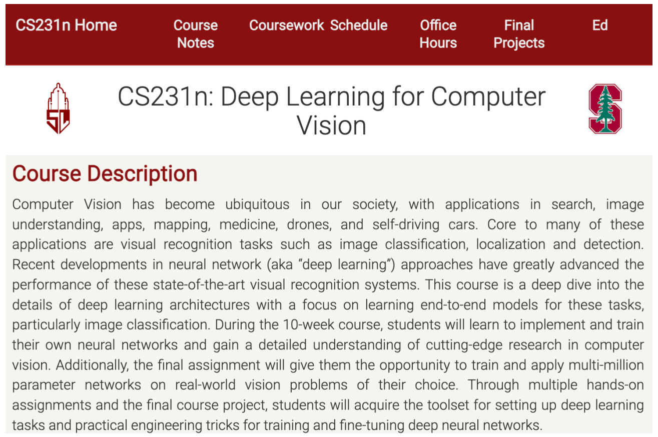
arXiv

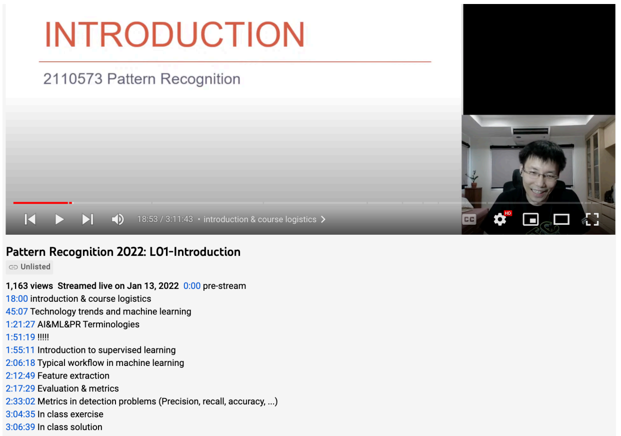


Courses









Textbooks

Journals

Tutorials

Blog posts

YouTube videos

Outline

- Machine Learning Pitfalls
  - ML algorithms are not one-size-fits-all
  - Having more data doesn't always help
  - Exploiting spurious correlations
  - Training data don't capture real use cases
  - Having data leakage
  - Reporting a "non-informative" evaluation metric
  - Solving "irrelevant" problems
- A Way to Communicate with ML Practitioners
  - Be clear about what the inputs and outputs are
  - State the requirements and use-cases
  - Provide domain-/field-specific information
  - Summary of related papers
- Resources

Q&A