Machine Learning Frontiers

Review of what we’ve covered

Challenges and open questions

Exciting applications

Where to go from here?

Supervised Learning
K-Nearest Neighbors
Linear regression
Perceptron
Logistic regression
Linear Discriminant Analysis (LDA)
Quadratic Discriminant Analysis
Naïve Bayes
Classification and Regression Trees
Random Forests
Neural Networks, backpropagation, CNNs
Kernel Methods / Support Vector Machines

Ensemble methods
- Bagging, boosting, stacking

Regularization (ridge and lasso), feature selection, cost functions, and norms
Decision theory
Gradient descent and stochastic gradient descent

Performance Evaluation
Cross validation
Bootstrap sampling
Confusion Matrices
ROC curves
Precision/Recall/Error Types
Bias-variance tradeoff
Curse of Dimensionality

Unsupervised Learning
Clustering
- K-Means
- Gaussian mixture model
- Agglomerative clustering
- DBSCAN
- Spectral clustering

Density Estimation
- Gaussian mixture models

Dimensionality Reduction
- Principal Components Analysis, LDA

Markov Models
Markov chains
Hidden Markov Models
Markov reward processes
Markov Decision Processes

Reinforcement Learning
Dynamic Programming
Policy Evaluation
Policy Improvement
Policy Iteration
Value Iteration
Generalized policy iteration
Monte Carlo Control
Model free learning

Semi-supervises learning
Self-supervised learning

Topics we covered

Machine Learning is NOT magic

It's **one tool of many** to assist us in making better, faster decisions

It is not a substitute for critical thinking and domain expertise – they go together

Practical Considerations for Machine Learning

1. Let your **problem/question** drive your design choices
2. Set a reasonable goal and clear metric of success
3. Ask yourself if there are **non-ML approaches that would work?**
4. Develop an **end-to-end pipeline** as soon as you're able and keep it maintained (data preparation & preprocessing, analysis, and performance evaluation)
5. Start with the **simplest solution** you can and layer on complexity as needed

- **Features / representations** are often more important than algorithms
- **Experimental design** is often more important than algorithms

Adapted from Google: https://developers.google.com/machine-learning/guides/rules-of-ml

More advice

- ALWAYS look at your data – before you begin, the inputs/outputs, etc.
  - Check your distributions
  - Explore outliers to get insights on the model
- Report confidence intervals whenever possible (make sure your "better" model is not just a noisy aberration)
- When comparing supervised models, make sure your comparing on the same validation set
- Make sure you NEVER mix training and validation information

Adapted from https://www.unofficialgoogledatascience.com/2016/10/practical-advice-for-analysis-of-large.html

Challenges and Open Questions

Unsolved challenges in machine learning

Generalizing from **small numbers of examples** (few-shot learning)

Adapting to **new environments** and non-stationary problems

Transferring knowledge between tasks (transfer learning)

Combining **heterogeneous data sources** into a model

Interpretability for confidence in algorithms

Ethics, fairness, and privacy (asks the question: should we?)

Greenwald and Oertel, 2017, Future Directions in Machine Learning
Jeannette Wing, 2020, Ten Research Challenge Areas in Data Science

Types of Biases

- Historical bias
- **Representation bias**
- Measurement bias
- Evaluation bias
- Aggregation bias
- Population bias
- Simpson’s paradox
- Longitudinal data fallacy
- **Sampling bias**
- Behavioral bias
- Content production bias
- Linking bias
- Temporal bias
- Popularity bias
- Algorithmic bias
- User interaction bias
- Presentation bias
- Ranking bias
- Social bias
- Emergent bias
- Self-selection bias
- Omitted variable bias
- Cause-effect bias
- Observer bias
- Funding bias

Bold indicates those biases we’ve discussed through the semester – there are MANY more

Mehrabi, N., Morstatter, F., Saxena, N., Lerman, K. and Galstyan, A., 2019. A survey on bias and fairness in machine learning. arXiv preprint arXiv:1908.09635.

Ethical Considerations

- Privacy and surveillance
- Manipulation of behavior
- Transparency in machine learning systems
- Decision system biases
- Automation and employment
- Autonomous systems and responsibility
- Machine ethics (e.g. Isaac Asimov's 3 laws)
- Artificial moral agents
- Singularity

Müller, V.C., 2020. Ethics of artificial intelligence and robotics.

The Bitter Lesson (link,video)

Richard Sutton

"The biggest lesson that can be read from 70 years of AI research is that general methods that leverage computation are ultimately the most effective, and by a large margin."

Computation

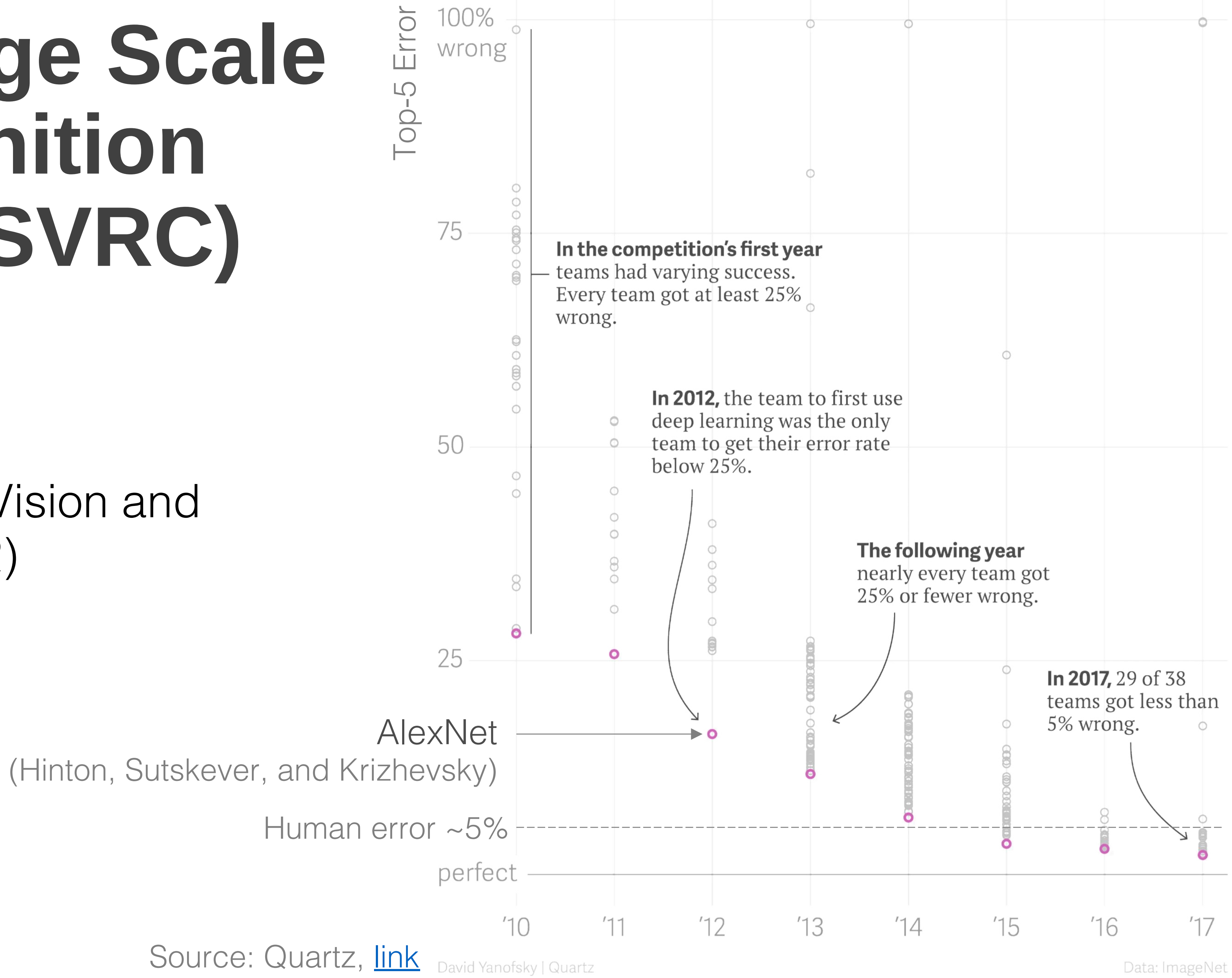
NVIDIA
Graphics Processing Units (GPU)
(link)

Google
Tensor Processing Unit (TPU)
(link)

ImageNet Large Scale Visual Recognition Challenge (ILSVRC)

Fei-Fei Li et al. 2010 (link)

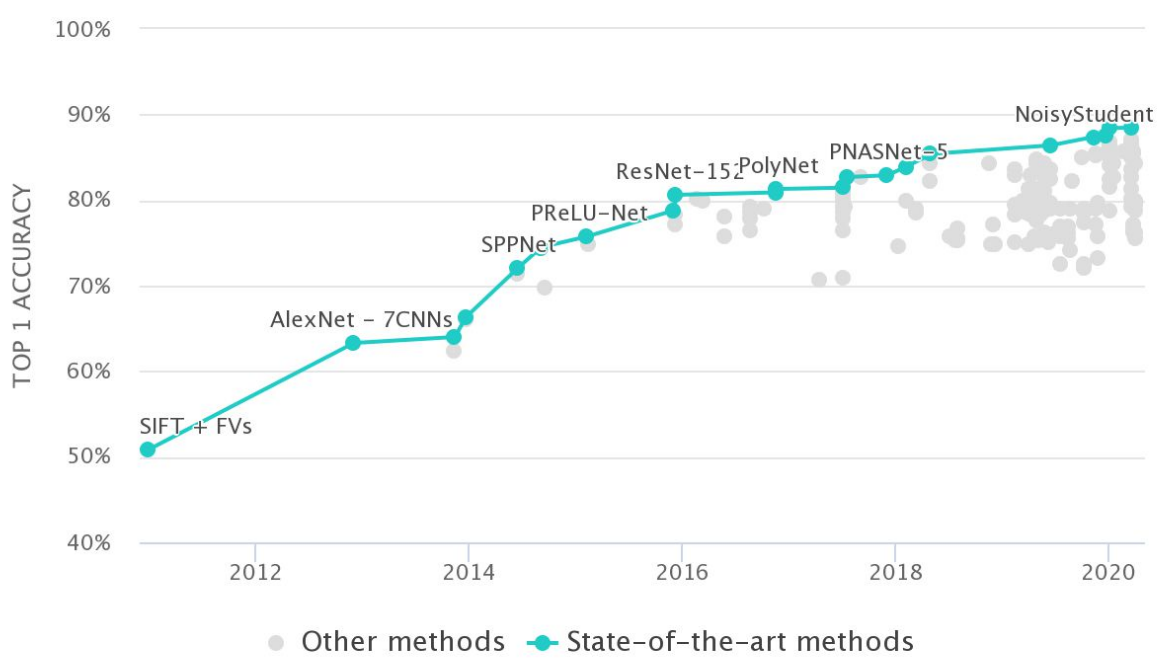
Competition at:
Conference on Computer Vision and Pattern Recognition (CVPR)



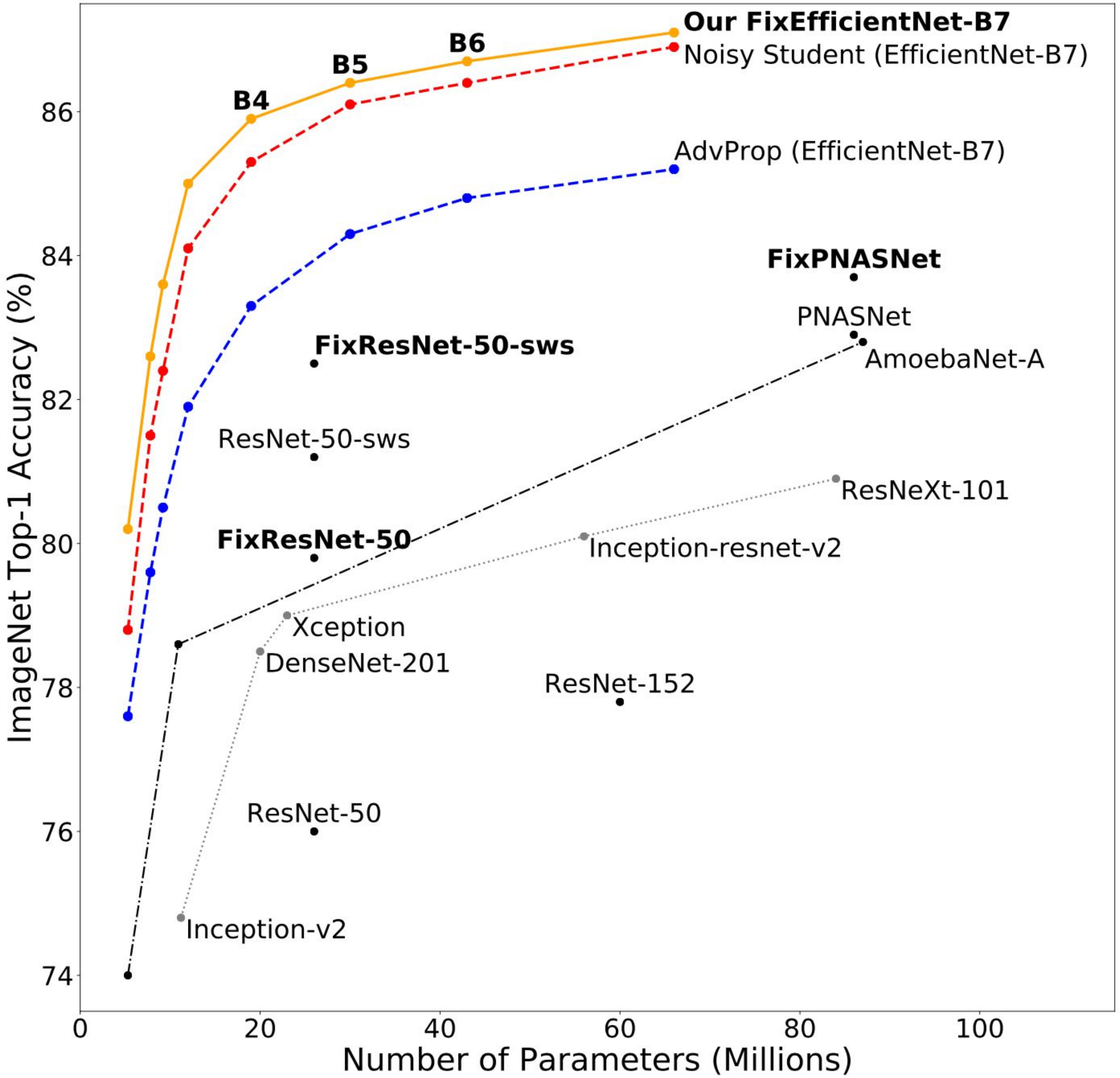
Source: Quartz, link

Latest Advances with Image Recognition

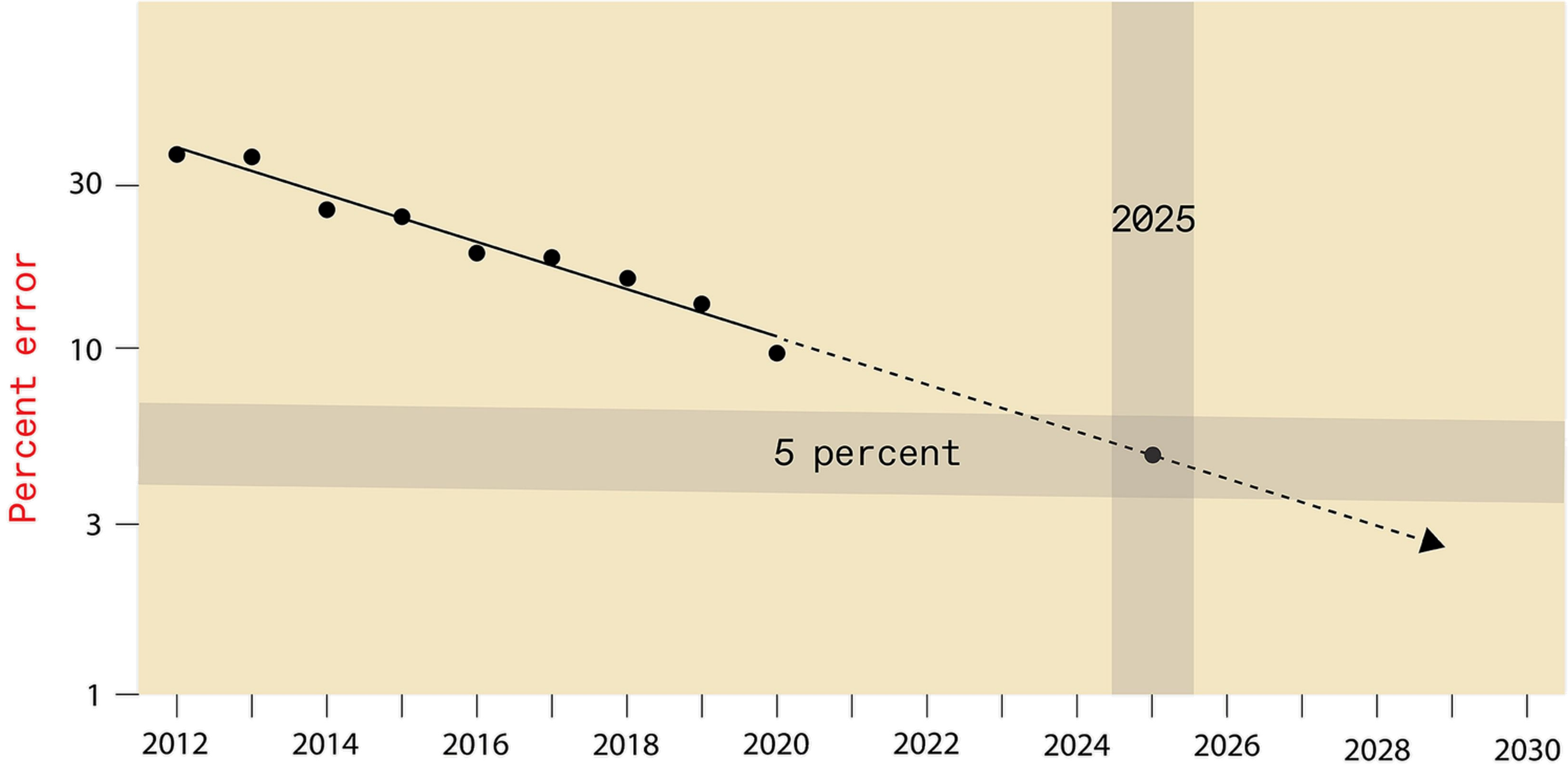
Touvron, H., Vedaldi, A., Douze, M. and Jégou, H., 2020. Fixing the train-test resolution discrepancy: FixEfficientNet. arXiv preprint arXiv:2003.08237.



Source: https://paperswithcode.com/sota/image-classification-on-imagenet

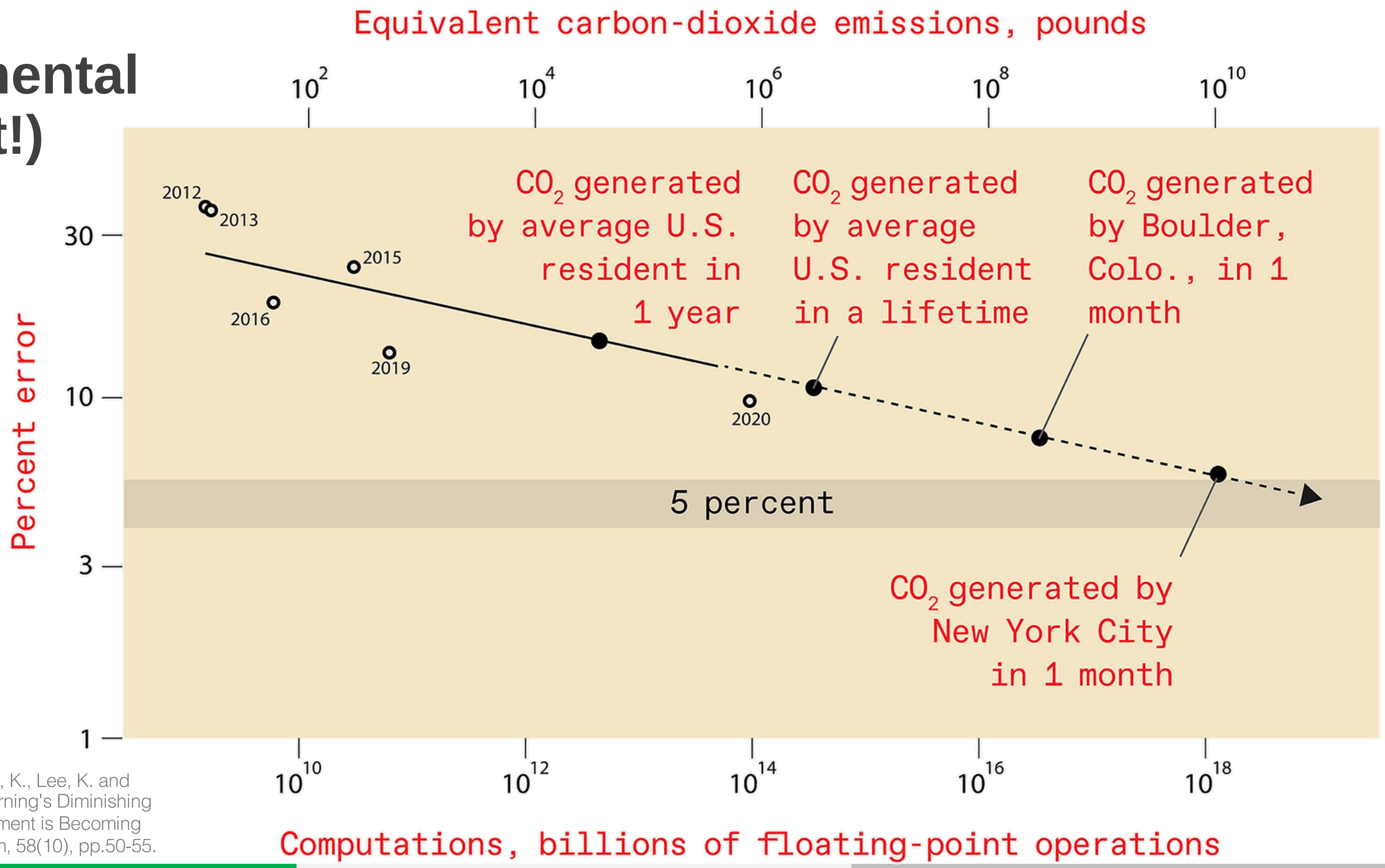


Trend in ImageNet performance



Thompson, N.C., Greenewald, K., Lee, K. and Manso, G.F., 2021. Deep Learning's Diminishing Returns: The Cost of Improvement is Becoming Unsustainable. IEEE Spectrum, 58(10), pp.50-55.

Potential environmental (and cost!) impacts



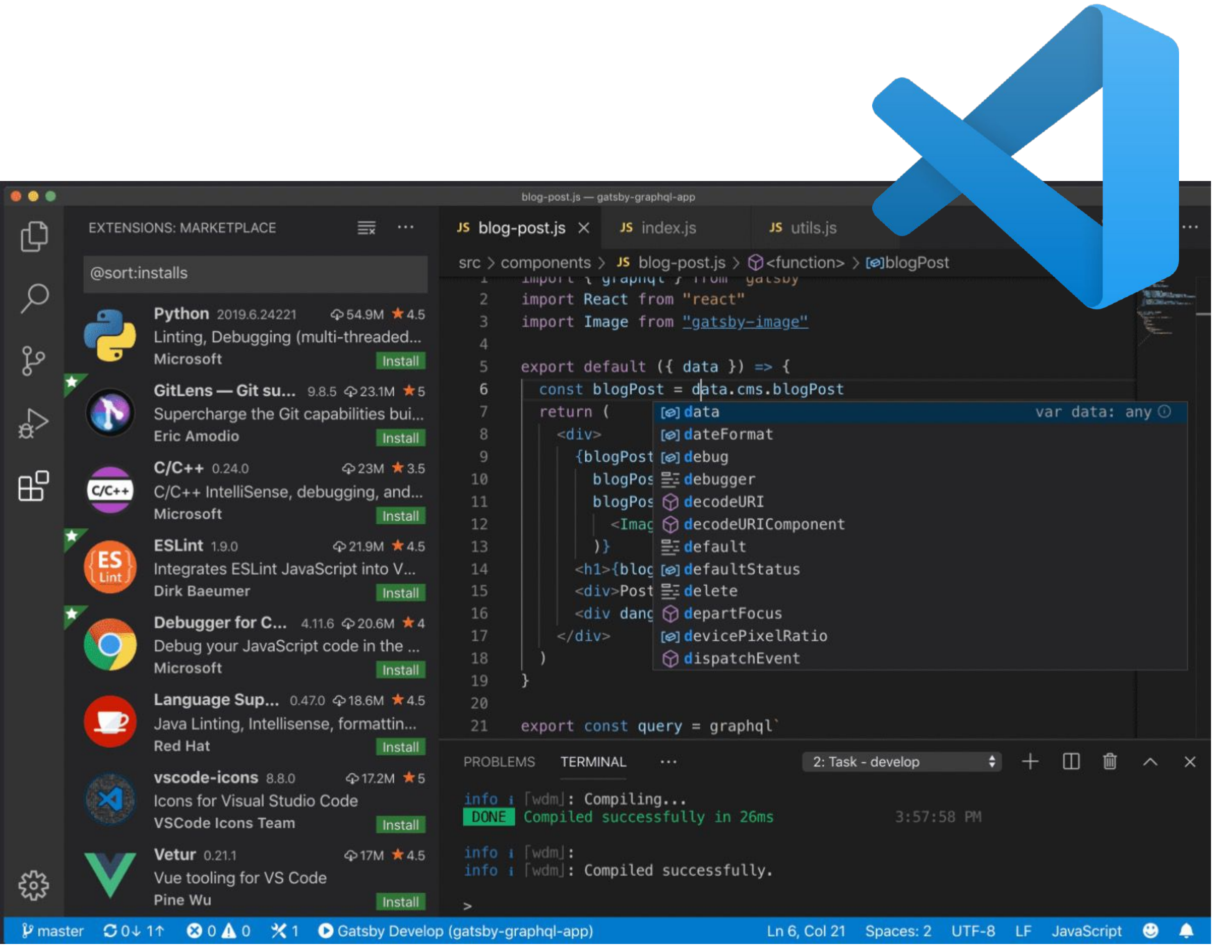
Thompson, N.C., Greenewald, K., Lee, K. and Manso, G.F., 2021. Deep Learning's Diminishing Returns: The Cost of Improvement is Becoming Unsustainable. IEEE Spectrum, 58(10), pp.50-55.

How to deal with increasing complexity?

More code
More machine learning experiments to keep track of

Tools: Development Environments

Use the best coding tools for the job: VS Code, Jupyter, PyCharm, Spyder, etc.



For additional practical considerations, explore Berkeley's [Full Stack Deep Learning]

Organization: Experiment Tracking

Track your experiments in an organized way. If it helps, use Comet.ml, Weights and Biases, MLFlow, Neptune, etc.

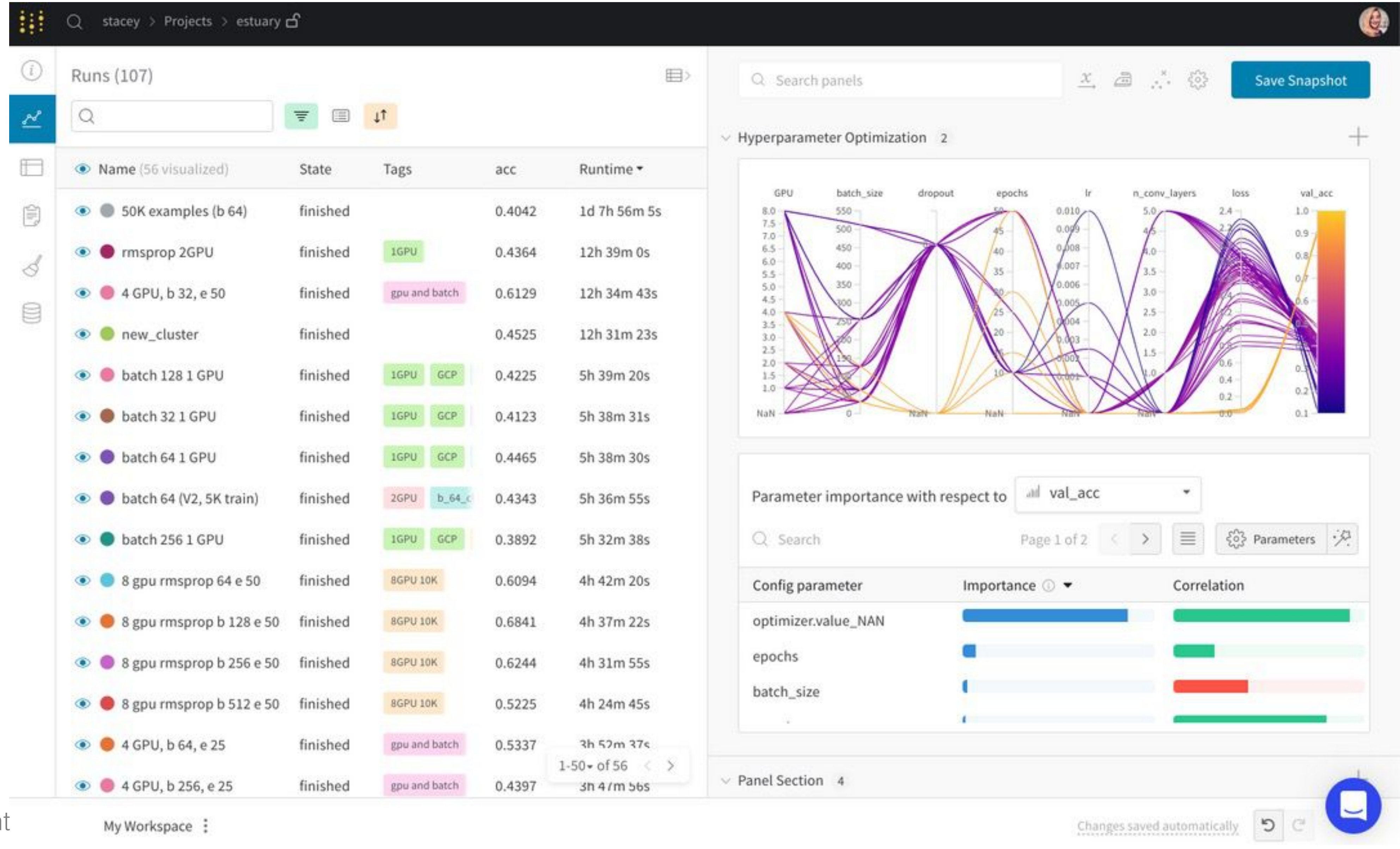
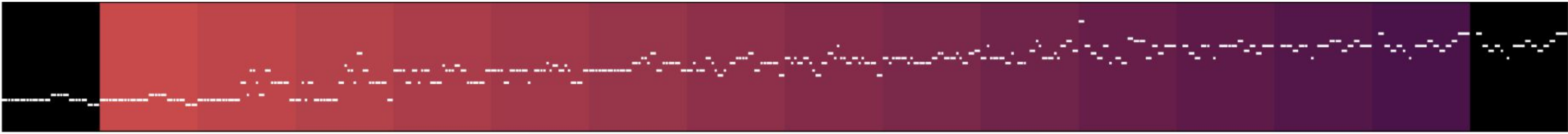


Image from:
https://github.com/wandb/client

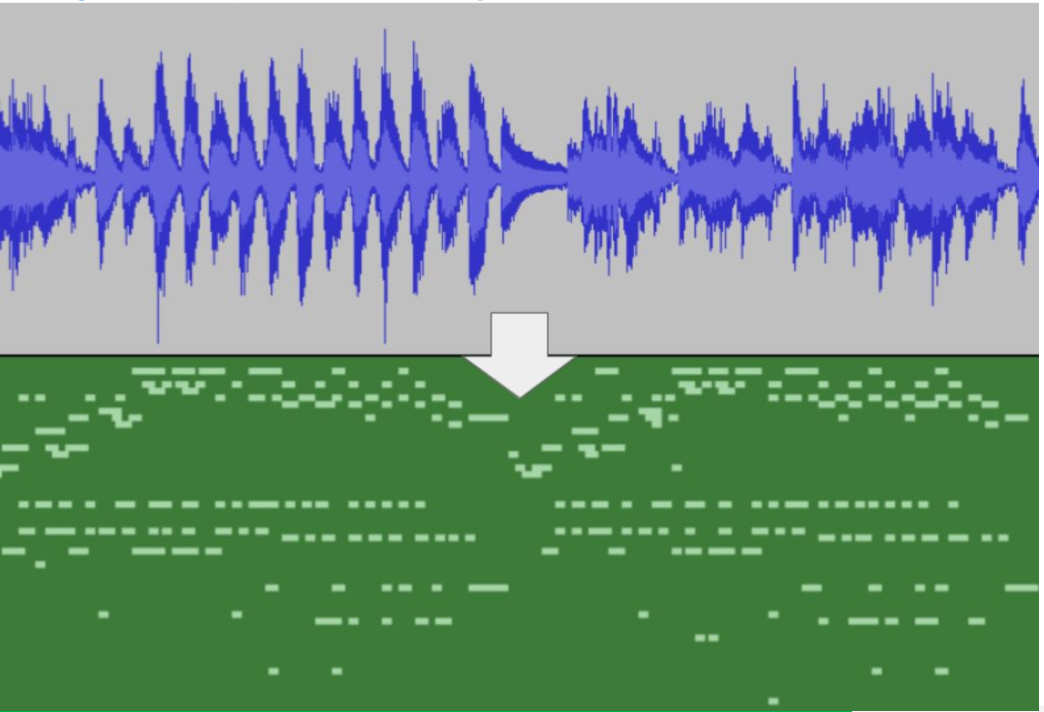
Machine Learning Applications

Music

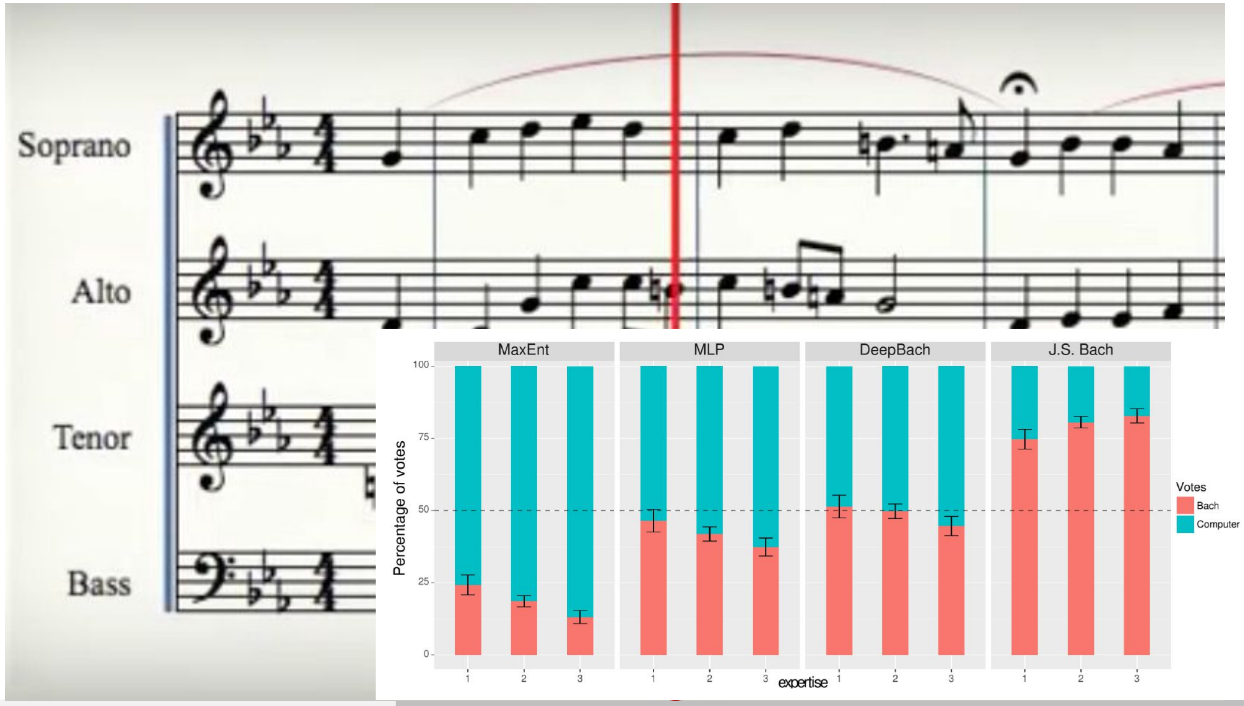
MusicVAE (Magenta):
Blending musical scores (link)
Magenta Project (from Google Brain / tensorflow)



Onsets and Frames:
Automated transcription (link)
Magenta Project (from Google Brain / tensorflow)

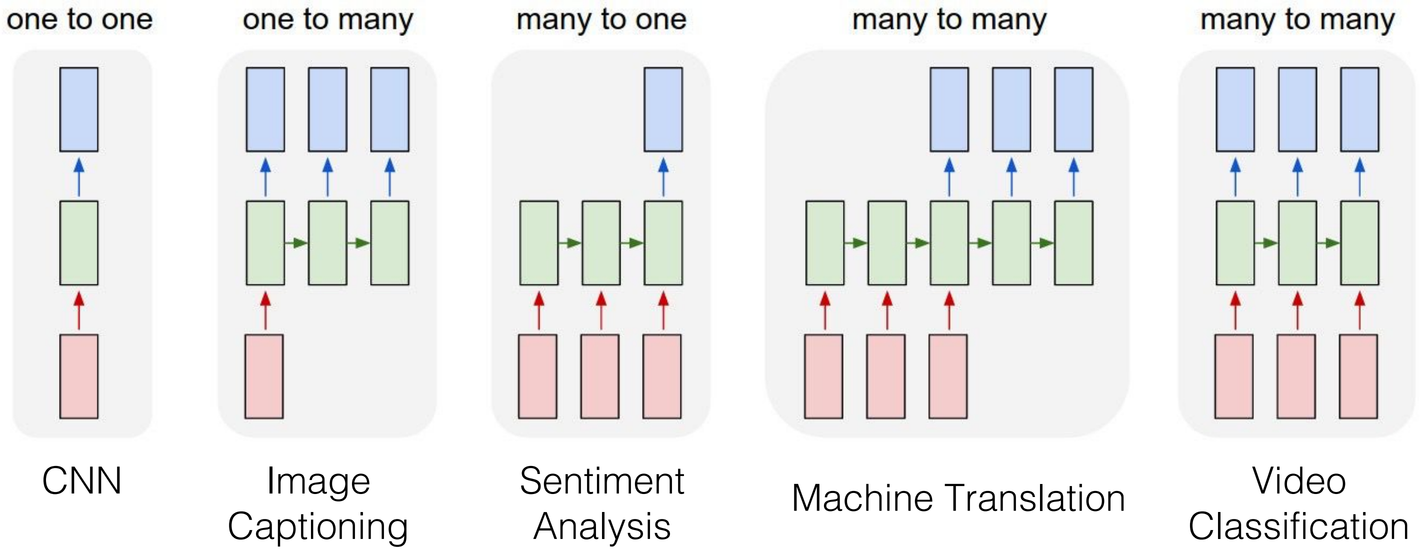


Deep Bach:
Automated harmonization (link)
Paper available here: https://arxiv.org/abs/1612.01010



Natural Language Processing

Recurrent Neural Networks



Example architecture: Long short-term memory (LSTM)

Drawbacks: Sequential data processing (slow)

Transformers

Do not process sequentially, enables parallelization

Use an attention mechanism
E.g. a way of indicating which words in a sentence are most relevant to a specific word

Examples: BERT, Transformer XL

Image source: http://karpathy.github.io/2015/05/21/rnn-effectiveness/ Resource on Transformers: https://arxiv.org/pdf/1706.03762.pdf

Computer Vision

Handwriting completion:
…with Neural Networks ([link](#))
Carter et al. 2016, our Experiments in Handwriting with a Neural Network (Google Brain)

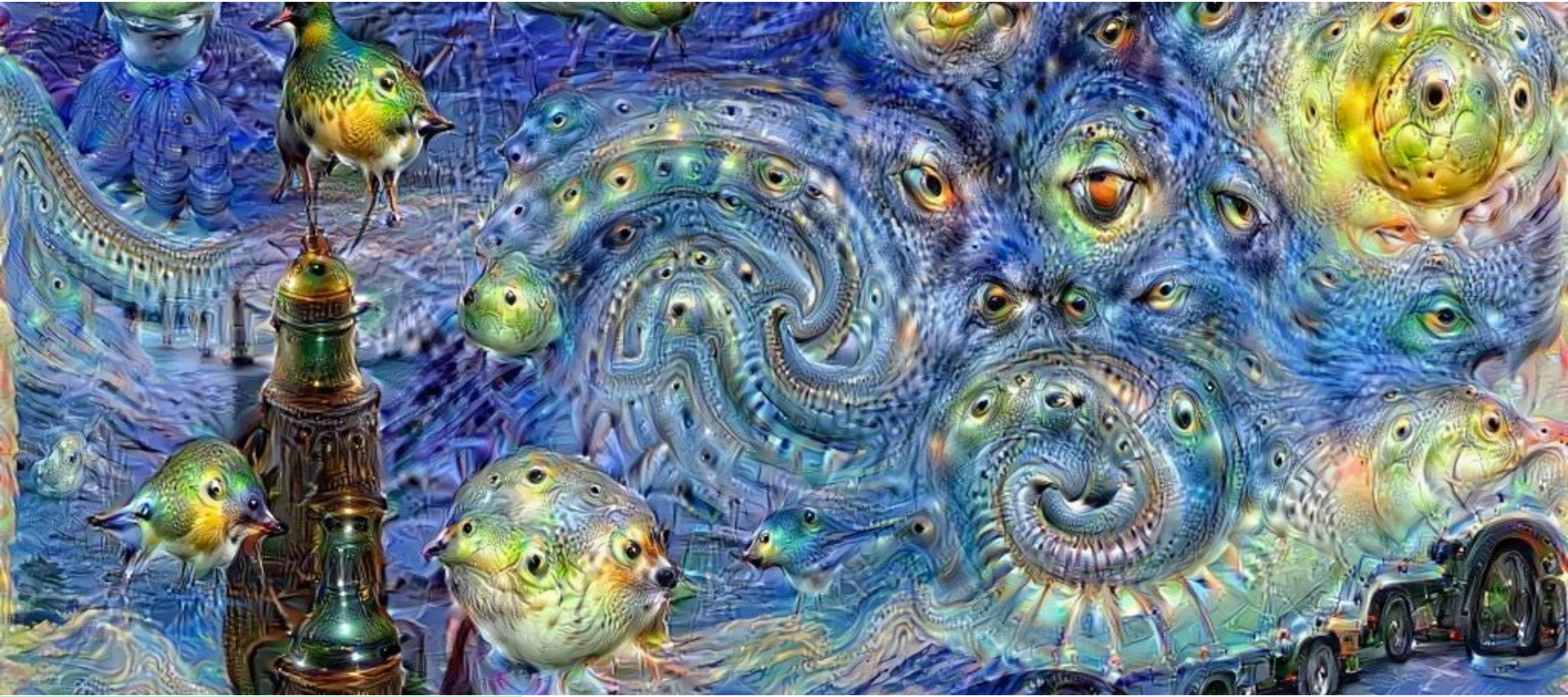
SketchRNN:
Automated sketching ([link](#))
Magenta Project (from Google Brain / tensorflow)

Computer Vision & Visual Arts

Deep Dream: Style Transfer and Abstract Art (link)

Originally developed by Alexander Mordvintsev from Google

There are now websites for DIY deep art: https://deepart.io/



Computer Vision

Style Transfer
(link)

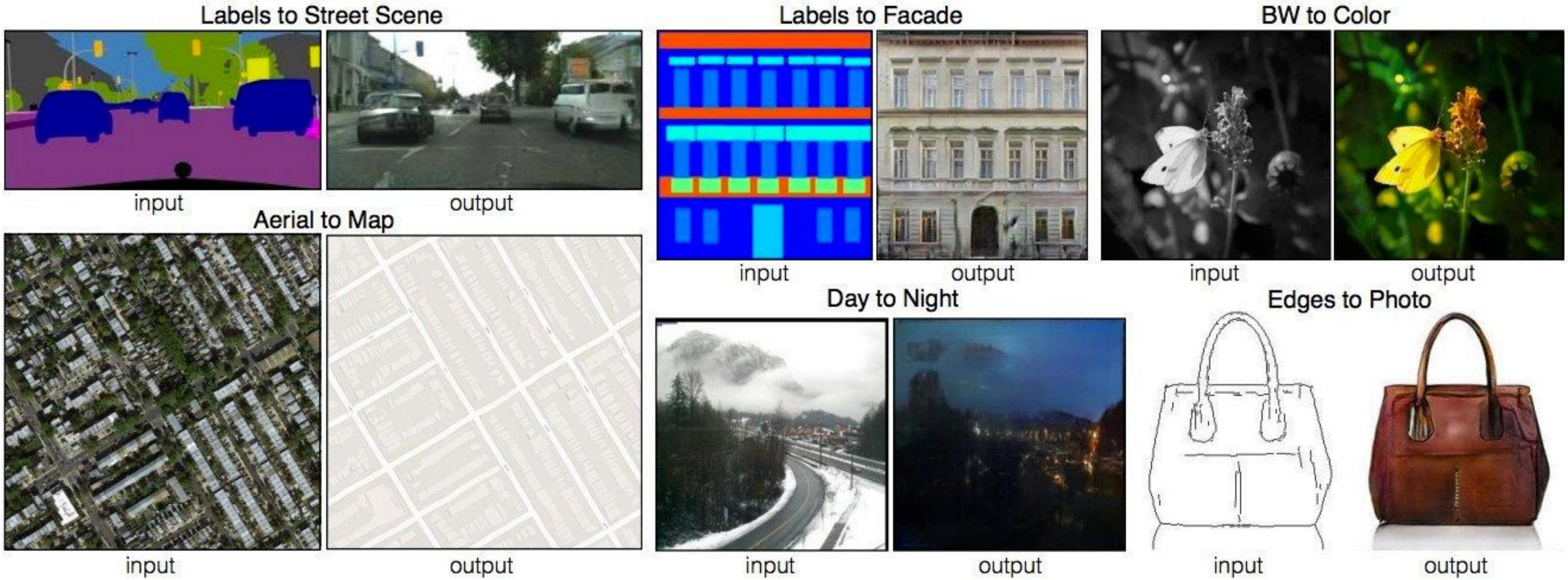
Dumoulin et al. 2016, A learned representation for artistic style

There are now websites for DIY deep art: https://deepart.io/ and where you can purchase it, like this one:

Computer Vision

Image-to-Image Translation (link)

Isola et al. 2017, Image-to-image translation with conditional adversarial networks



Computer Vision

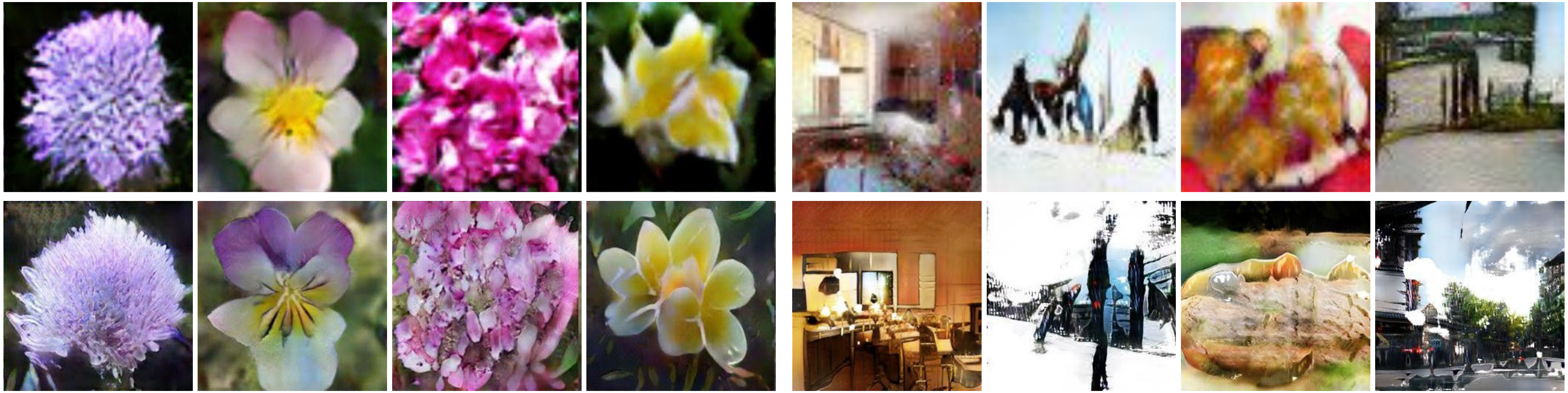
StackGAN:
Image Synthesis from Text Descriptions (link)

Zhang et al. 2017, StackGAN: Text to Photo-realistic Image Synthesis with Stacked Generative Adversarial Networks

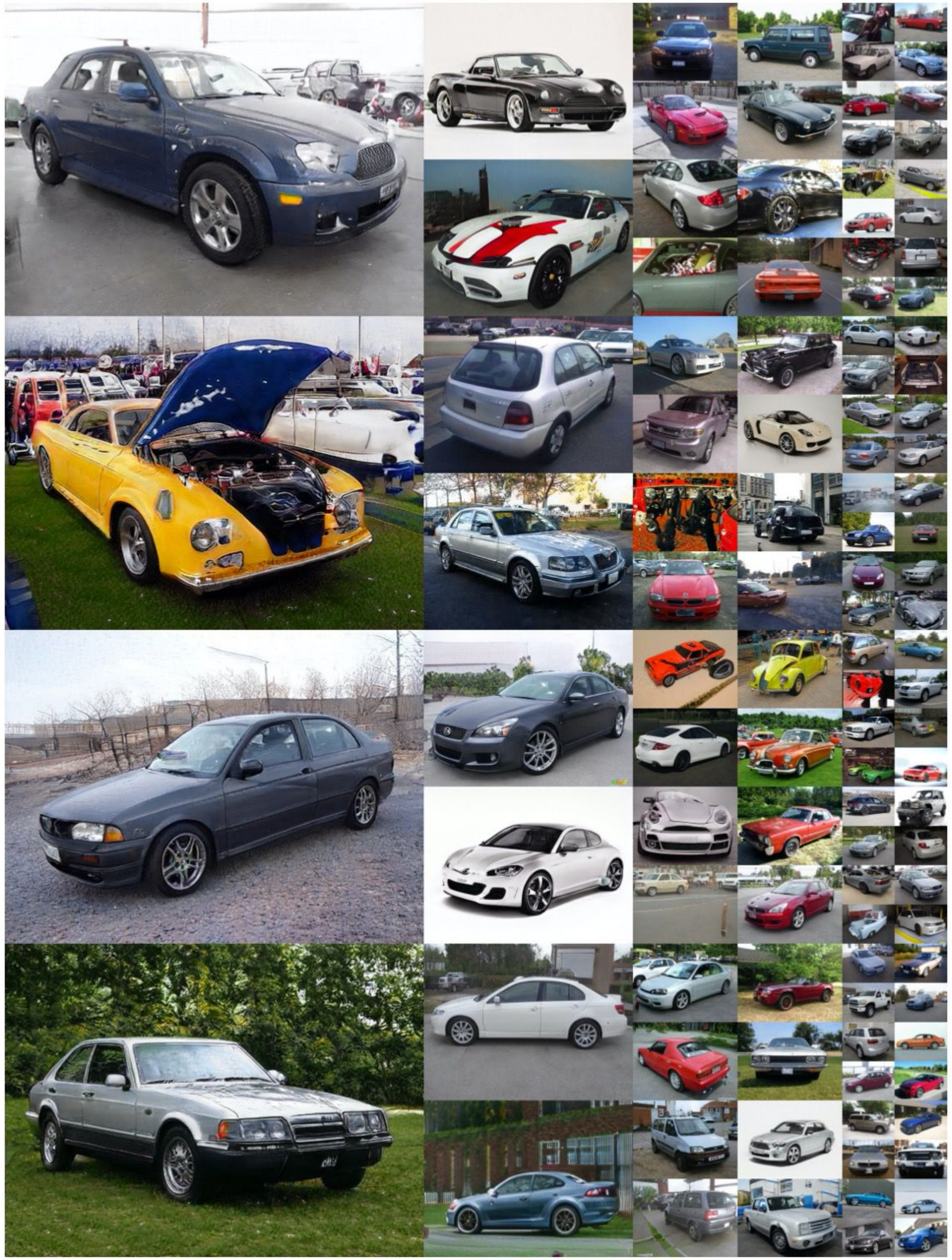
Text description	This flower has a lot of small purple petals in a dome-like configuration	This flower is pink, white, and yellow in color, and has petals that are striped	This flower has petals that are dark pink with white edges and pink stamen	This flower is white and yellow in color, with petals that are wavy and smooth	A picture of a very clean living room	A group of people on skis stand in the snow	Eggs fruit candy nuts and meat served on white dish	A street sign on a stoplight pole in the middle of a day

64x64 GAN-INT-CLS

256x256 StackGAN



Karras, T., Laine, S. and Aila, T., 2019. A style-based generator architecture for generative adversarial networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (pp. 4401-4410).



Synthetic images

Face Synthesis

Image Synthesis (link)

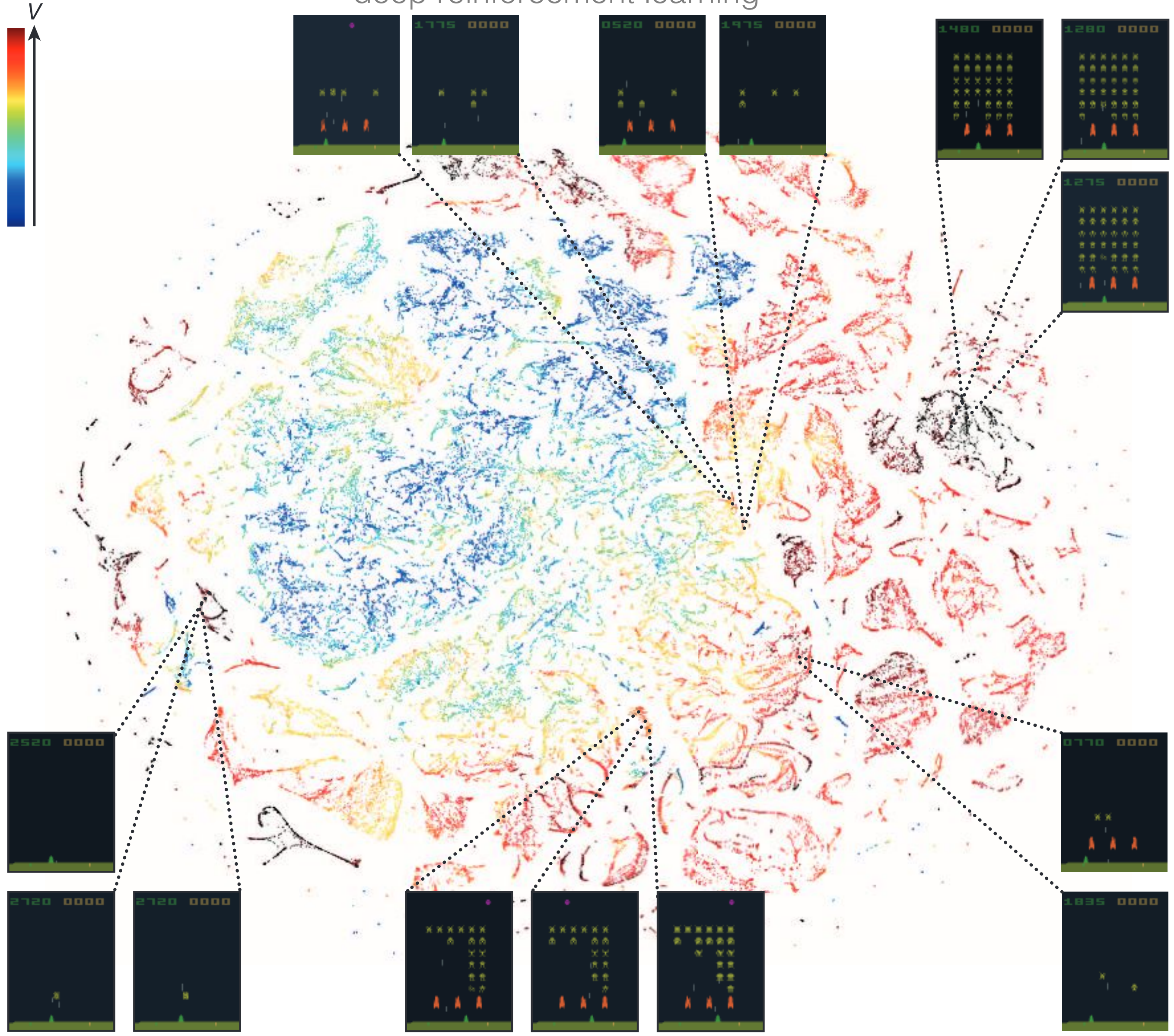
Karras et al. 2018, NVIDIA: Progressive growing of GANS for improved quality, stability, and variation

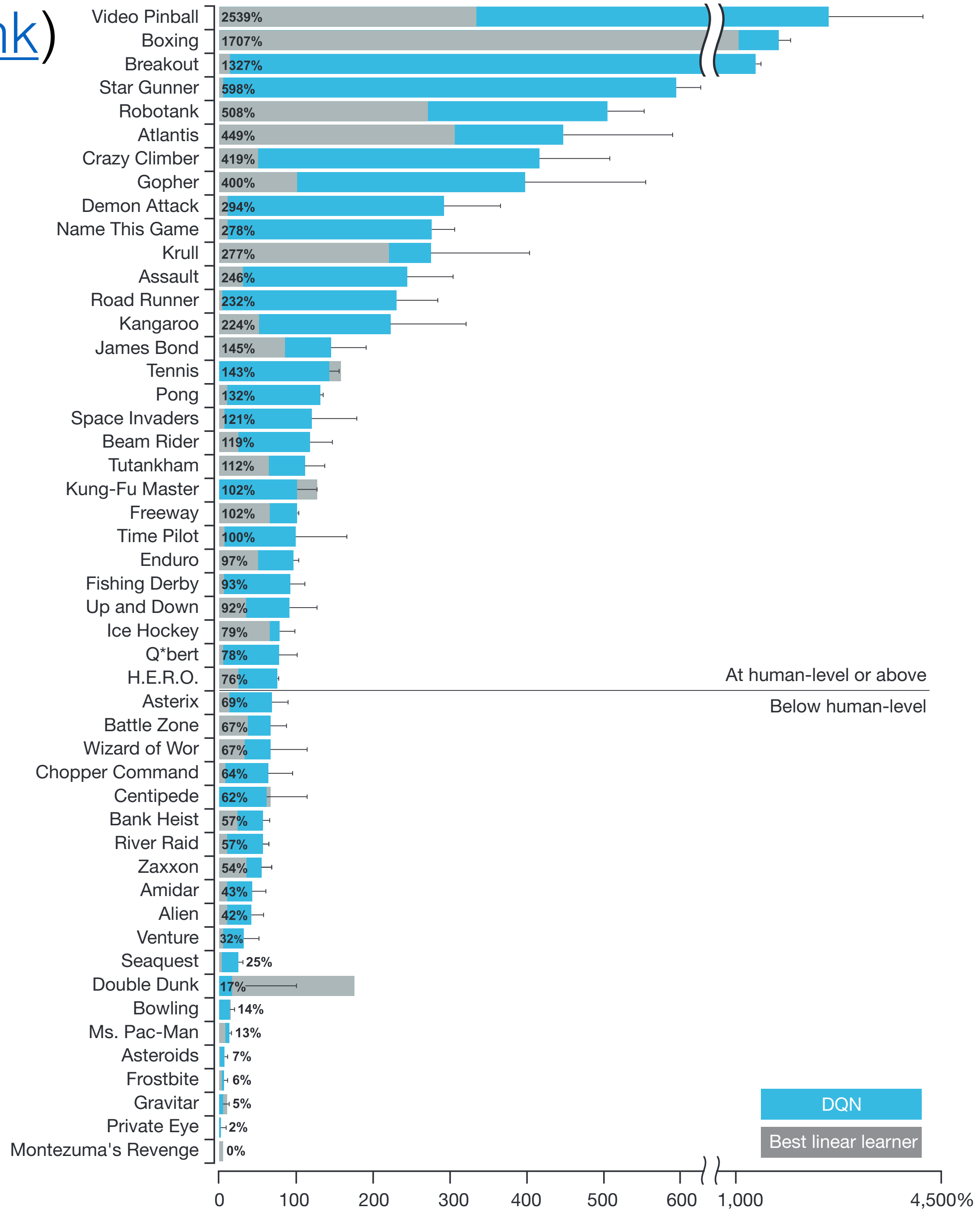
These images are all synthetic

Games

Reinforcement Learning for Atari (link)

Mnih et al. 2015 (DeepMind), Human-level control through deep reinforcement learning





Games & Reinforcement Learning

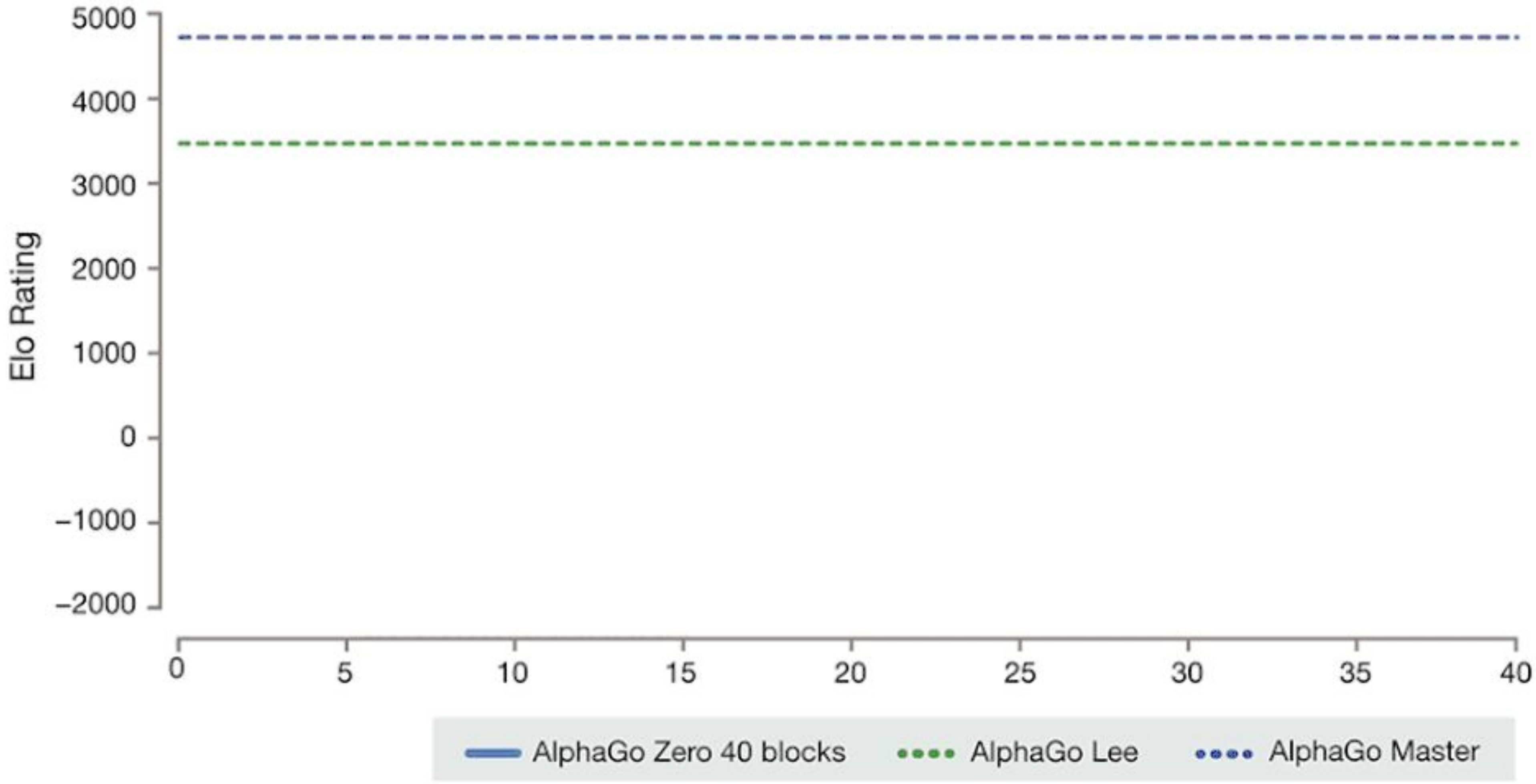
AlphaZero (link)

Silver et al. 2017 (DeepMind), Mastering Chess and Shogi by Self-Play with a General Reinforcement Learning Algorithm

Learning Go starting from random play

Mastered in 24 hours

Did the same with Chess



AlphaStar: Learning to Play StarCraft II

Vinyals, O., Babuschkin, I., Czarnecki, W.M., Mathieu, M., Dudzik, A., Chung, J., Choi, D.H., Powell, R., Ewalds, T., Georgiev, P. and Oh, J., 2019. Grandmaster level in StarCraft II using multi-agent reinforcement learning. Nature, 575(7782), pp.350-354.

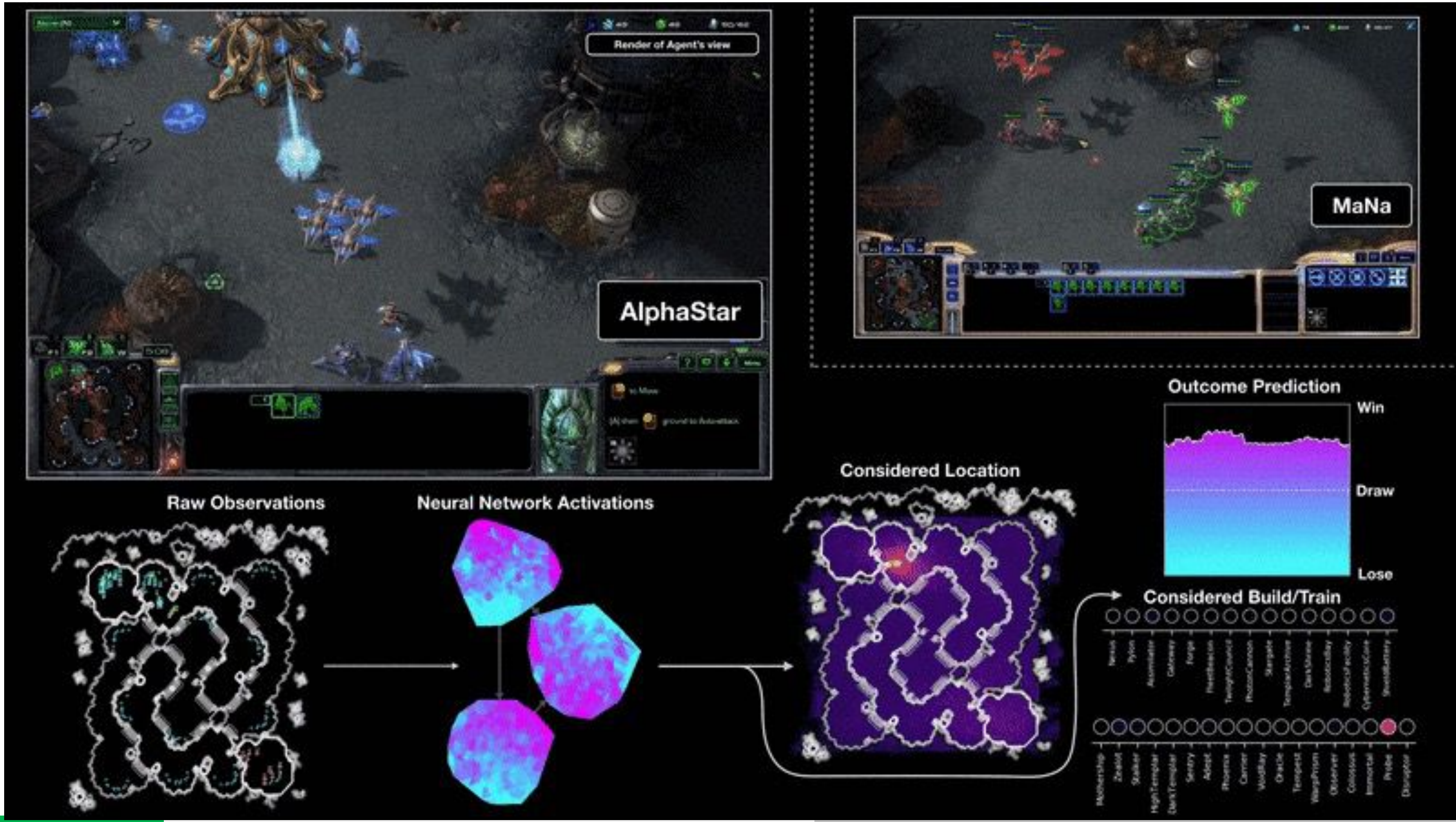


Image from https://deepmind.com/blog/article/alphastar-mastering-real-time-strategy-game-starcraft-ii

AlphaStar: Learning to Play StarCraft II

Vinyals, O., Babuschkin, I., Czarnecki, W.M., Mathieu, M., Dudzik, A., Chung, J., Choi, D.H., Powell, R., Ewalds, T., Georgiev, P. and Oh, J., 2019. Grandmaster level in StarCraft II using multi-agent reinforcement learning. Nature, 575(7782), pp.350-354.

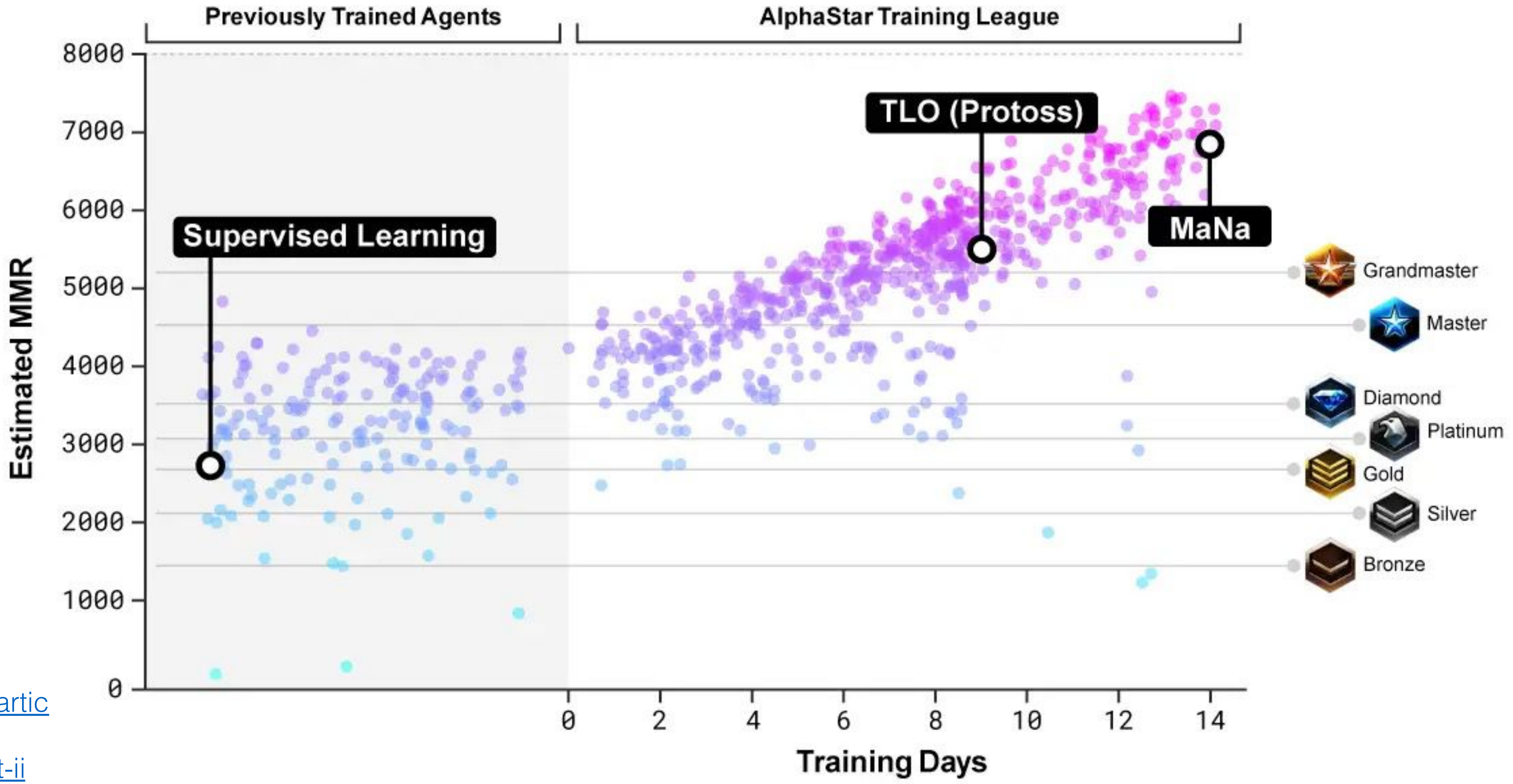


Image from https://deepmind.com/blog/article/alphastar-mastering-real-time-strategy-game-starcraft-ii

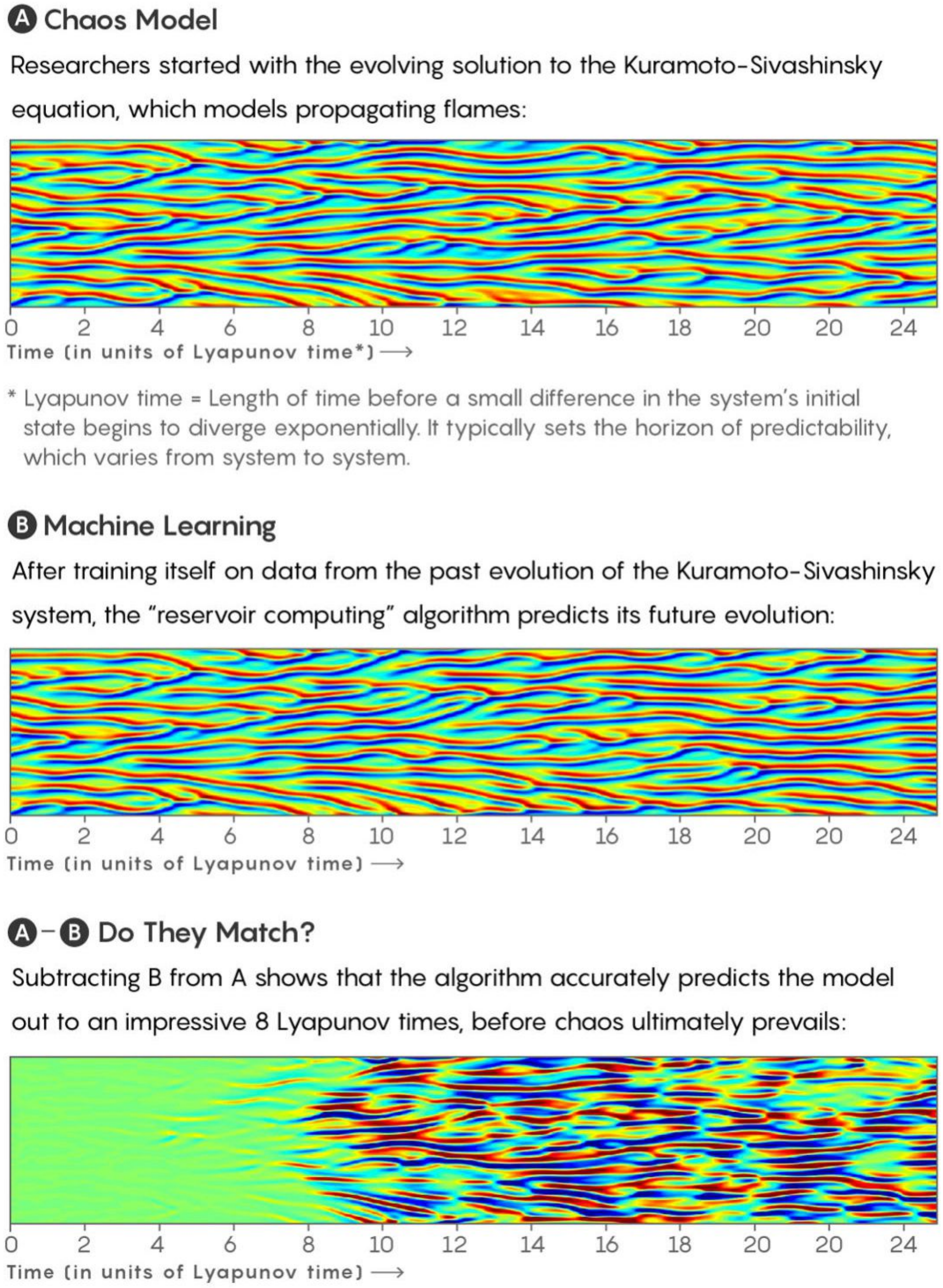
Other applications

Forecasting Chaos (link)

Pathak et al. 2017, Model-Free Prediction of Large Spatiotemporally Chaotic Systems from Data: A Reservoir Computing Approach

Figure (right) from Quanta Magazine (link)

- Weather prediction
- Heart attack prediction (monitoring cardiac arrhythmias)
- Monitoring neuronal firing patterns for signs of neuron spikes
- Predicting solar flares



Other applications

Self-driving cars ([link]())

Lee et al. 2017, DESIRE: Distant Future Prediction in Dynamic Scenes with Interacting Agents

Automated medical diagnostics

Drug development & chemical synthesis

Brain-computer interfaces

3D Shape detection and tracking

Energy demand forecasting

Open source frameworks

Tensorflow (link)
Framework for implementing graphical models, such as neural networks

PyTorch (link)
Framework for implementing graphical models, such as neural networks

OpenAI (link)
OpenAi Gym is a toolkit for developing and comparing reinforcement learning algorithms

Where to go from here?

Courses at Duke

ECE 585: Signal Detection and Extraction Theory
ECE 586: Vector Space Methods with Applications
ECE 588: Image & Video Processing
ECE 684: Natural Language Processing
ECE 685D: Deep Learning
CompSci 527: Computer Vision
Math 412: Topological Data Analysis
Math 465/CompSci 445: Introduction to High Dimensional Data Analysis
Stat 601: Bayesian and Modern Statistical Methods
Stat 623: Statistical Decision Theory
Stat 571: Advanced Probabilistic Machine Learning
ECE 590: Special Topics courses on machine learning and artificial intelligence topics

Courses on foundational concepts: Probability, Statistics, Linear Algebra, Mathematics, Programming and Software Development

Staying up-to-date

Data Elixir

ISSUE 202 *Oct 2nd 2018*

— In the News —

One Small Step for the Web...

Tim Berners-Lee announced a new platform that gives users complete control over their data. The platform works on the existing web but within its ecosystem, **you decide where you store your data.** Partly because of the people involved, this project is worth paying attention to.

inrupt.com

AI Could Provide Moment-by-Moment Nursing for a Hospital's Sickest Patients

ICUs are one of the most expensive parts of the medical system. The sickest patients often receive round-the-clock care and are typically attached to multiple machines and monitors. There's also vast amounts of data, which makes ICUs a prime target for AI disruption. In this article, the founders of Autonomous Healthcare explore the problems and opportunities.

ieee.org

— Sponsored Link —

Try Mode Studio: a complete toolkit for every analyst

Mode Studio combines a SQL editor, Python & R notebooks, and a visualization builder in one platform. And it's free forever. Connect data from anywhere and analyze with the best language for the job, without having to jump between tools. Build custom visualizations or use our out-of-the-box charts. Share your analysis with a click—every report lives at a URL.

modeanalytics.com

Data Science Roundup

Scaling Knowledge. 5 Tips for Better DS Writing. Infrastructure @ Stitch Fix. New Data o...

Scaling Knowledge. 5 Tips for Better DS Writing. Infrastructure @ Stitch Fix. New Data on DS Jobs. [DSR #152]

By Tristan Handy • Issue #152 • View online

Special thanks to Domino Data Lab for sponsoring this week's Roundup. I've actually jumped on the "Sponsored Content" bandwagon—you'll see a post from Domino below with a "Sponsored" callout in the title. We'll only accept sponsorships like this from organizations we think highly of and—of course—will always prioritize your reading experience.

Enjoy this week's issue :)

- Tristan

❤️ Want to support us? Forward this email to three friends!

🚀 Forwarded this from a friend? Sign up to the Data Science Roundup here.

This Week's Most Useful Posts

Scaling Knowledge

Data scientists, analysts, and engineers are ultimately employed by

Kaggle Newsletter

kaggle

Hey there!

This month's newsletter is sure to inspire those new to Kaggle (like me 👋) and experts alike. See a motivational **interview with our first-ever Kernels Grandmaster**, Martin Henze (AKA "Heads or Tails") as he shares his tips and tricks for getting to the next level. you're still getting your feet wet, start with **this video discussion with our own Meg Risdal**, who breaks down the basic steps to beginning a project.

Speaking of videos... Did you tune in for the **43-minute livestream** of Siraj Raval competing in our **Taxi Duration Challenge**?

Read on for more of the month's data science notes, highlights and competitions. And, oh — we're hiring, too!

Have a great one,
Rachel

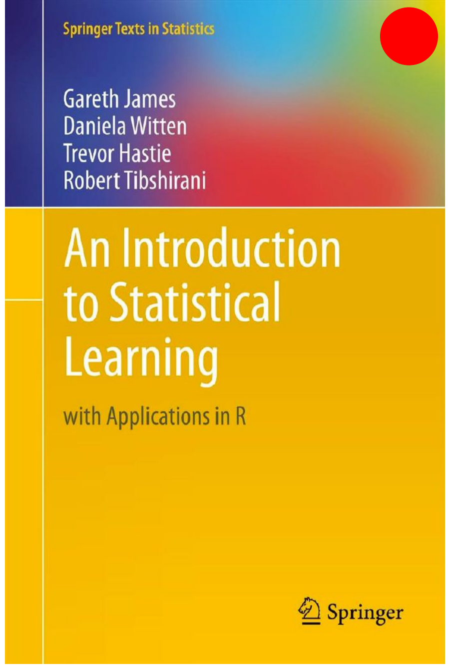
TALKING DATA SCIENCE

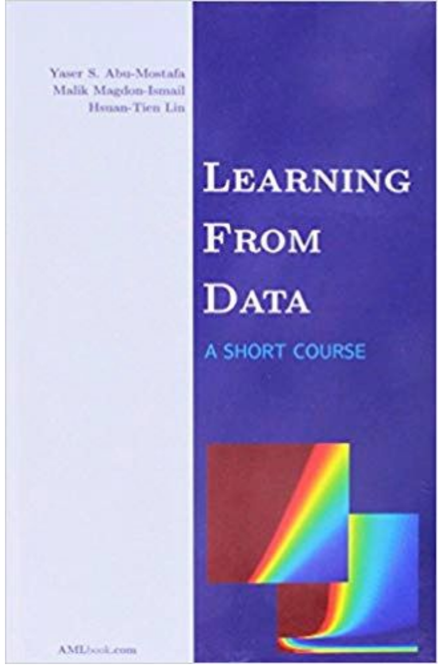
Profiling Top Kagglers: Martin Henze, World's First Kernels Grandmaster

Data Science Books

Additional Resources Available at:
https://kylebradbury.org/datascience.html

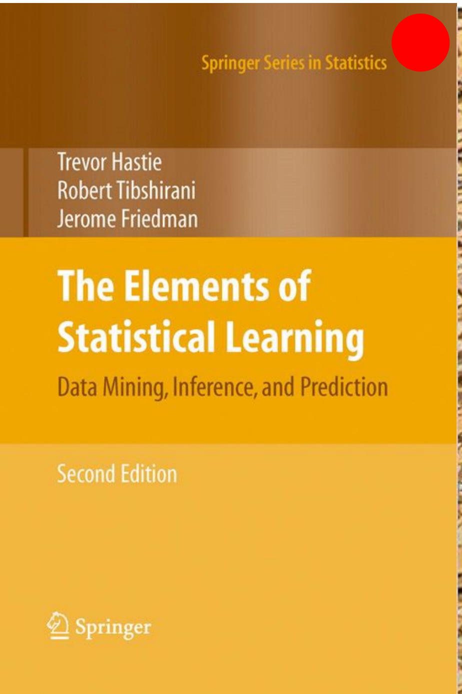
Introductory Texts

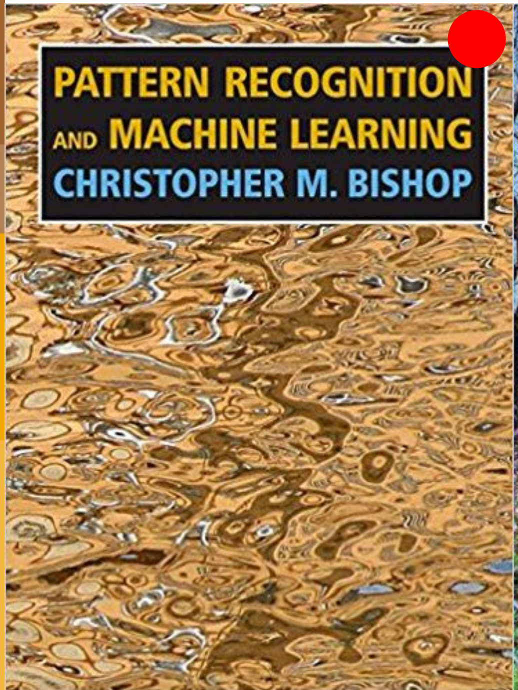


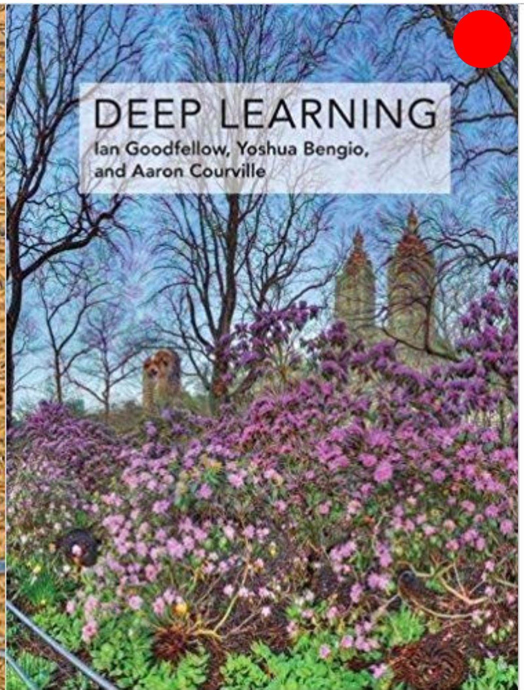


Free online

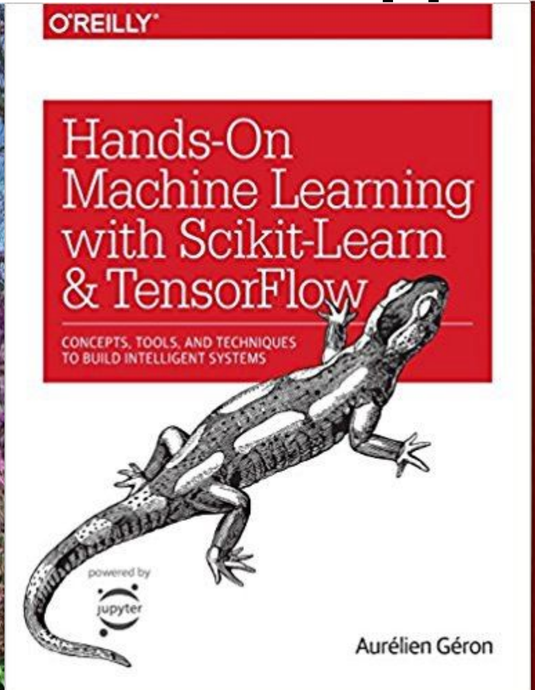
Advanced Texts

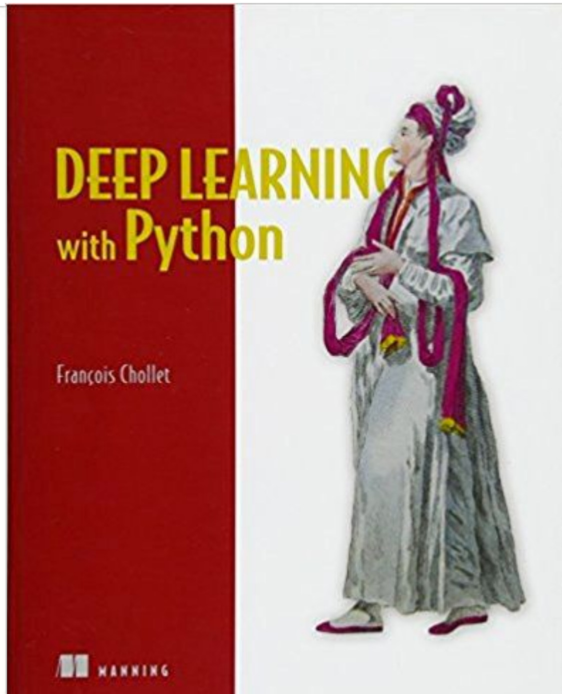






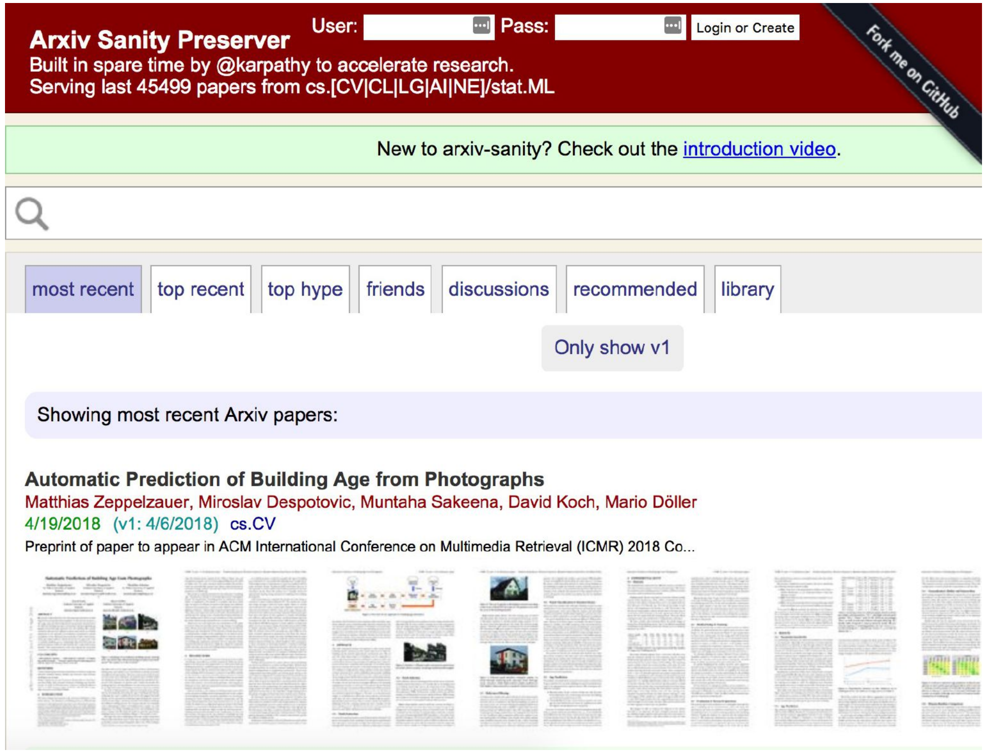
Applied Texts





Staying up-to-date

Arxiv (link) (Arxiv sanity-preserver)



Blogs:

Google Research (link)
Microsoft Research (link)
DeepMind Research (link)
Kaggle (link)

Conferences:

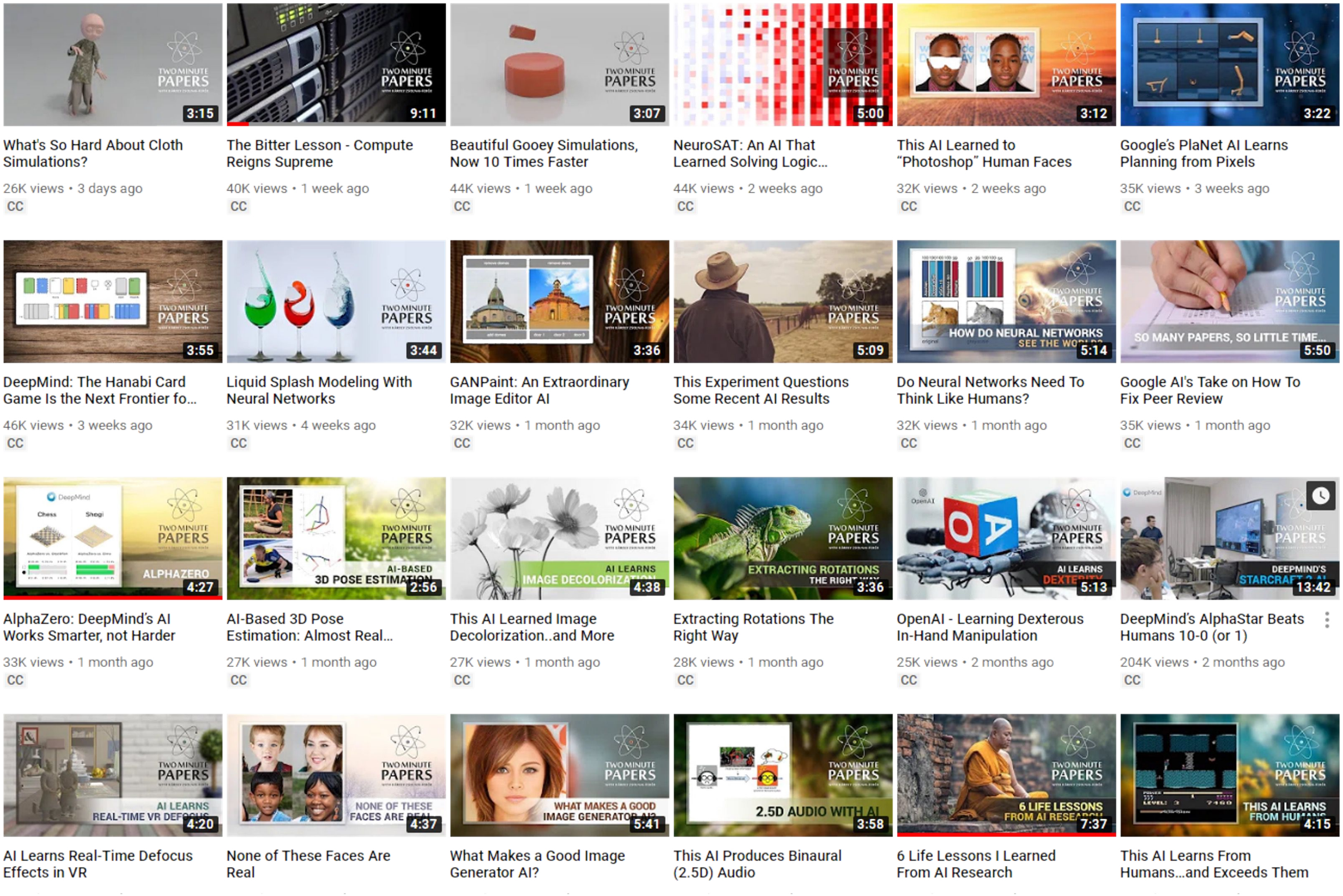
ICML: International Conference on Machine Learning (link)
NeurIPS (formerly NIPS): Neural Information Processing Systems (link)
ICLR: International Conference on Learning Representations (link)
CVPR: IEEE Conference on Computer Vision and Pattern Recognition (link)
SIGKDD: ACM International Conference on Knowledge Discovery & Data Mining (link)

TWO MINUTE PAPERS

by Károly Zsolnai-Fehér

Summaries of latest ML research

(link)



People to know & read about in Machine Learning

Jeffrey Hinton
Google & U. of Toronto

Backpropagation for neural nets. Won ImageNet Competition in 2012.


Image: Toronto Star

Andrew Ng
Baidu & Stanford

Founder of Google Brain, co-founder of Coursera


Image: Twitter

Andrej Karpathy
Tesla (formerly OpenAI)

CNNs for computer vision and natural language processing


Image: Stanford

Yann LeCunn
Facebook & NYU

Creator of the LeNet, Optical Character Recognition, and CNNs


Image: Facebook Research

Yoshua Bengio
Université de Montréal

Deep learning


Image: Université de Montréal

Ian Goodfellow
Google Brain

Creator of generative adversarial networks


Image: Michael Dukakis Institute

Fei-Fei Li
Stanford Artificial Intelligence Lab

Creator of ImageNet, computer vision


Image: Twitter

David Silver
DeepMind

Deep reinforcement learning


Image: Business Insider UK

Parting Thoughts

Hinton's Hints

Hinton is now "**deeply suspicious**" of backpropagation

"…'Science progresses **one funeral at a time**.' The future depends on some graduate student who is deeply suspicious of everything I have said."

"…I suspect that means getting rid of back-propagation. I don't think it's how the brain works," he said. "We clearly **don't need all the labeled data**."

Interview with Axios (link)

Parting advice

Baby steps: start small and with a simple model, add complexity as needed

Be prepared to iterate

Treat your data as a precious resource

Consider the consequences of your choices for others

No approach is perfect

Educating the **mind** without educating the **heart** is no education at all.

Aristotle