Deep Learning for Astronomy: An introduction

A/Prof Truyen Tran
Tung Hoang

Deakin University

truyen.tran@deakin.edu.au
truyentran.github.io
@truyenoz
letdataspeak.blogspot.com
goo.gl/3jJ1O0

Ballarat, June 2017

Agenda

Machine learning basics

Deep learning

Applications in Astronomy

Machine learning settings

Supervised learning

(mostly machin

A → B

Will be quickly solved for "easy" problems (Andrew Ng)

Anywhere in between: semi-supervised learning, reinforcement learning, lifelong learning, meta-learning, few-shot learning, knowledge-based ML

Unsupervised learning

an)

$$\mathbf{v} \sim P_{model}(\mathbf{v})$$

$$(\mathbf{v}) \approx P_{data}(\mathbf{v})$$

Best tricks in machine learning

Best classifiers

- Deep Neural Networks
- XGBoost
- Random Forests

Choosing right priors

- Extensive feature engineering
- Model architecture
- Loss functions
- Hyper-parameter tuning

Managing uncertainty

- Data augmentation
- Ensemble methods
- Bayesian methods

Model reuse

- Domain adaptation
- Transfer learning
- Multitask learning

Feature ~~engineering~~ learning

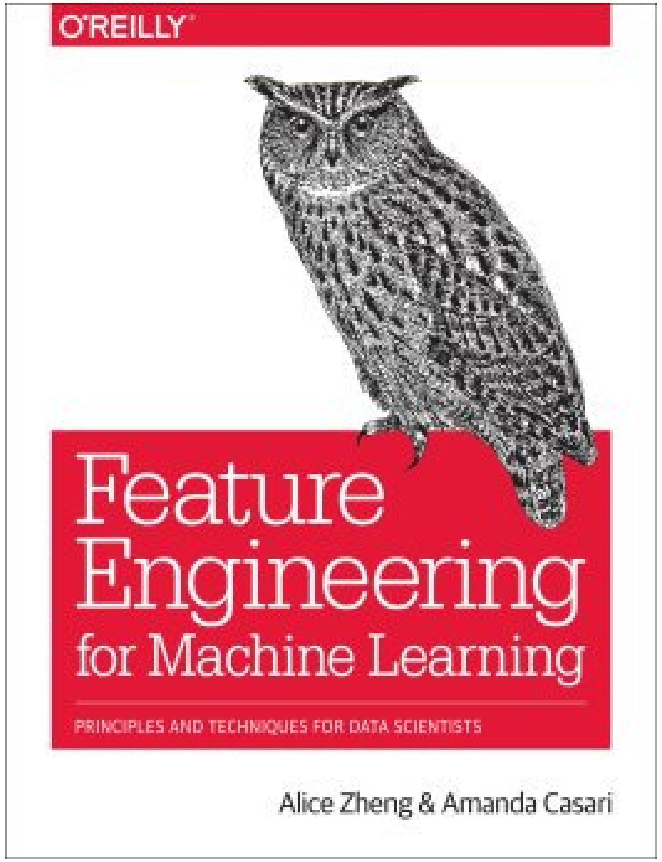
In typical machine learning projects, 80-90% effort is on <u>feature engineering</u>

- A right feature representation doesn't need fancy classifiers to work well.

Text : BOW, n-gram, POS, topics, stemming, tf-idf, etc.

Image: Histogram, SIFT, HOG, Filter banks, LBP, whitening, centring, color correction, denoising, etc.

Try yourself on Kaggle.com!



Why ML works?

Expressiveness

- Can represent the complexity of the world
- Can compute anything computable

Learnability

- Have mechanism to learn from the training signals

Generalizability

- Work on unseen data

What ML can do

Filling the slot

- In-domain (intrapolation), e.g., an alloy with a given set of characteristics
- Out-domain (extrapolation), e.g., weather/stock forecasting
- Classsification, recognition, identification
- Action, e.g., driving
- Mapping space, e.g., translation
- Replacing expensive simulations
- Novelty detection

Estimating semantics, e.g., concept/relation embedding

Assisting experiment designs

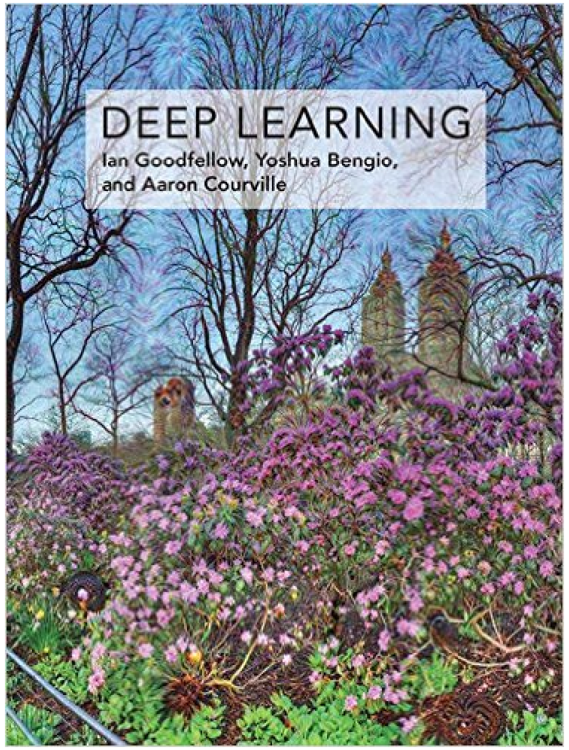
Finding unknown, causal relation, e.g., disease-gene

Predicting experiment results, e.g., alloys -> phase diagrams -> material characteristics

Deep learning

Deep learning page:

https://truyentran.github.io/deep.html



Keras

PYTORCH

TensorFlow

Yann LeCun

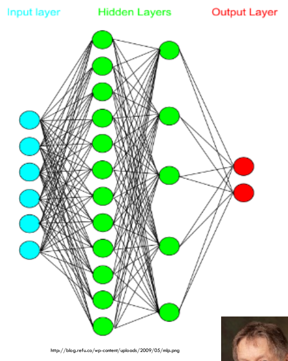
1988

Geoff Hinton

2006

Rosenblatt's perceptron

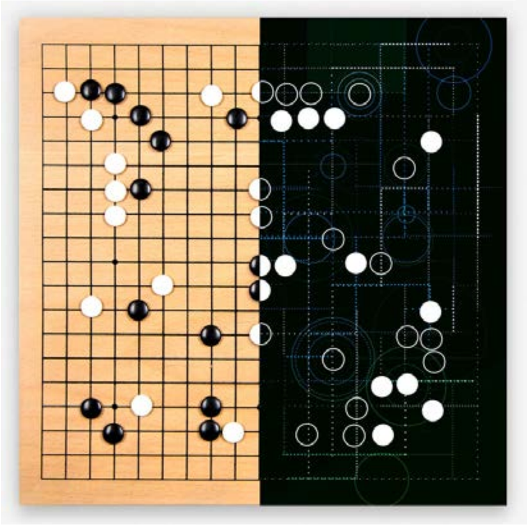
1958



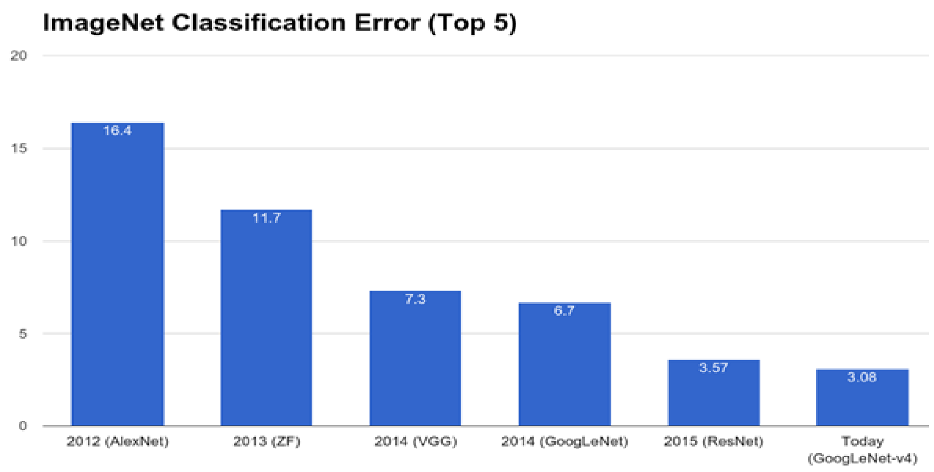
1986

2012

AlphaGo

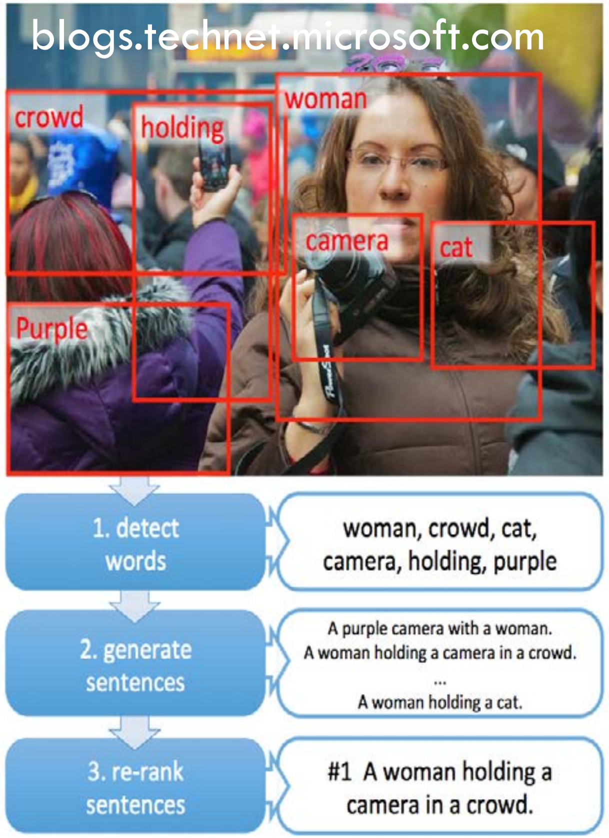


2016-2017



http://redcatlabs.com/2016-07-30_FifthElephant-DeepLearning-Workshop/#/

Deep learning in cognitive domains

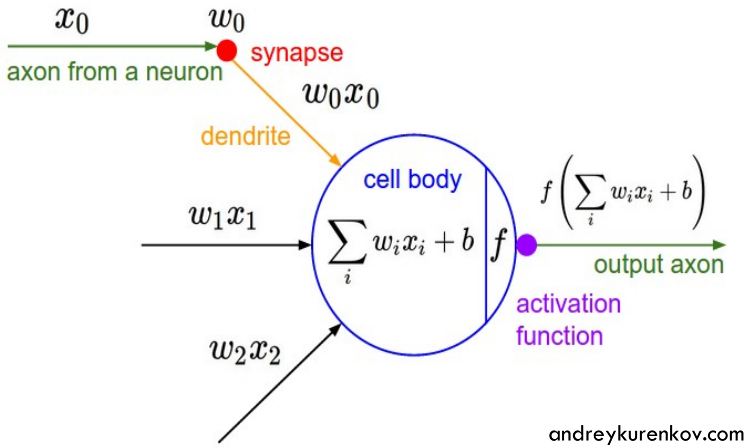


Where human can recognise, act or answer accurately within seconds

What is deep learning?

Quick answer: multilayer perceptrons (aka deep neural networks) of the 1980s rebranded in 2006

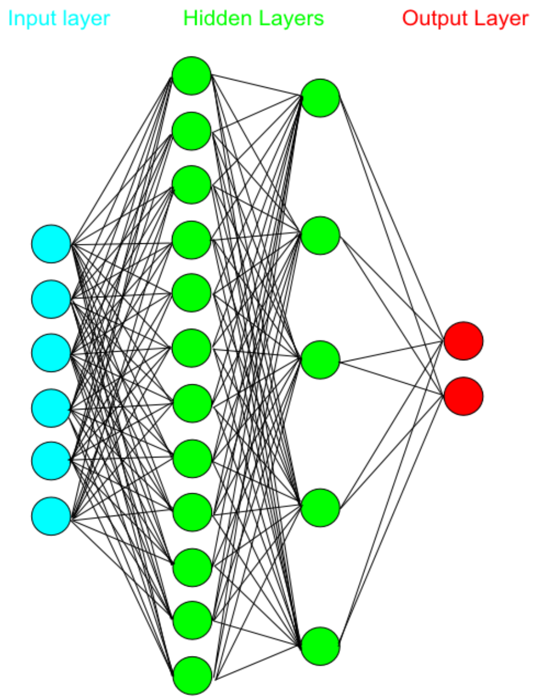
- Same backprop trick, as of 2017.
- Has a lot more hidden layers (100-1000X).
- Much bigger labelled datasets.
- Lots of new arts (dropout, batch-norm, Adam/RMSProp, skip-connections, Capsnet, external memory, GPU/TPU, etc.).
- Lots more people looking at lots of (new) things (VAE, GAN, meta-learning, continual learning, fast weights, etc.)



andreykurenkov.com

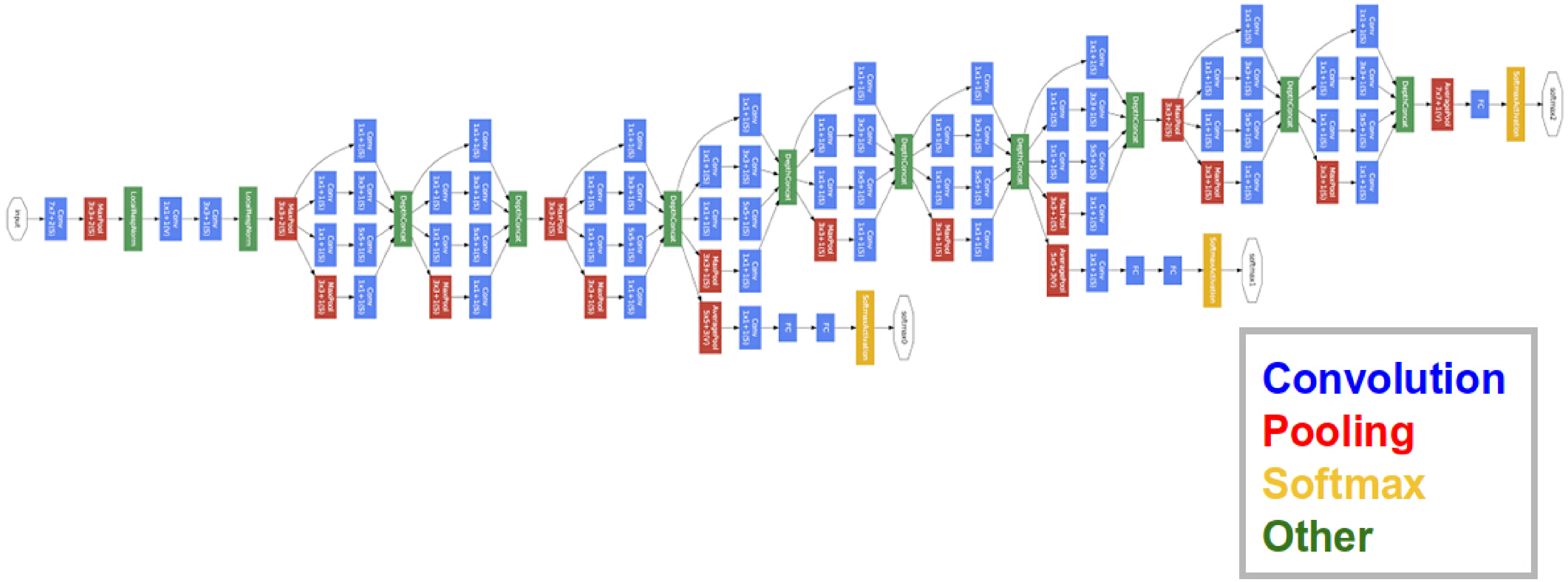
Much has changed

1986



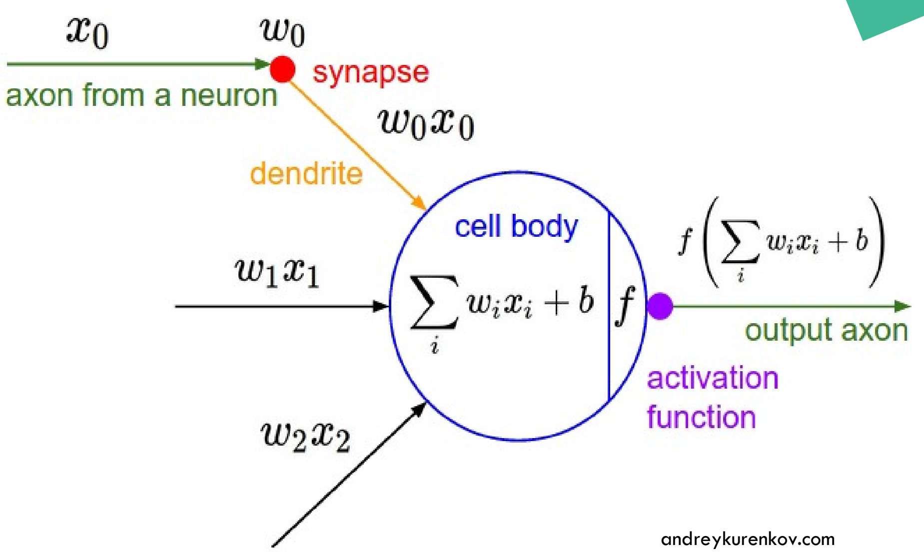
http://blog.refu.co/wp-content/uploads/2009/05/mlp.png

2016

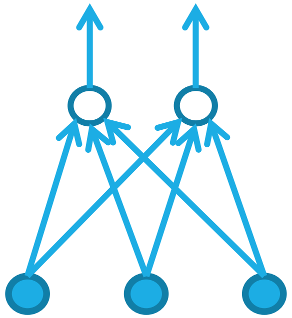


Deep learning as feature learning

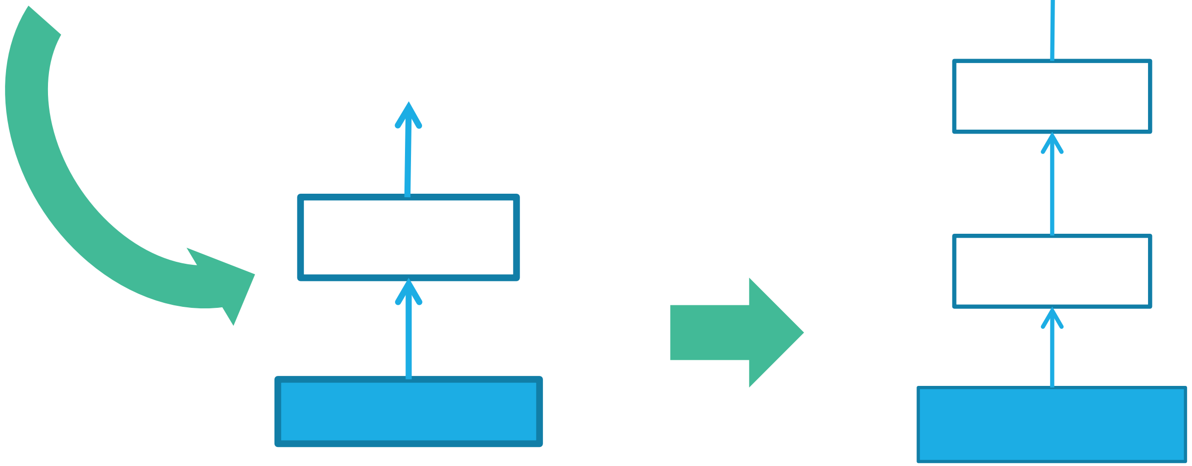
Integrate-and-fire neuron



andreykurenkov.com

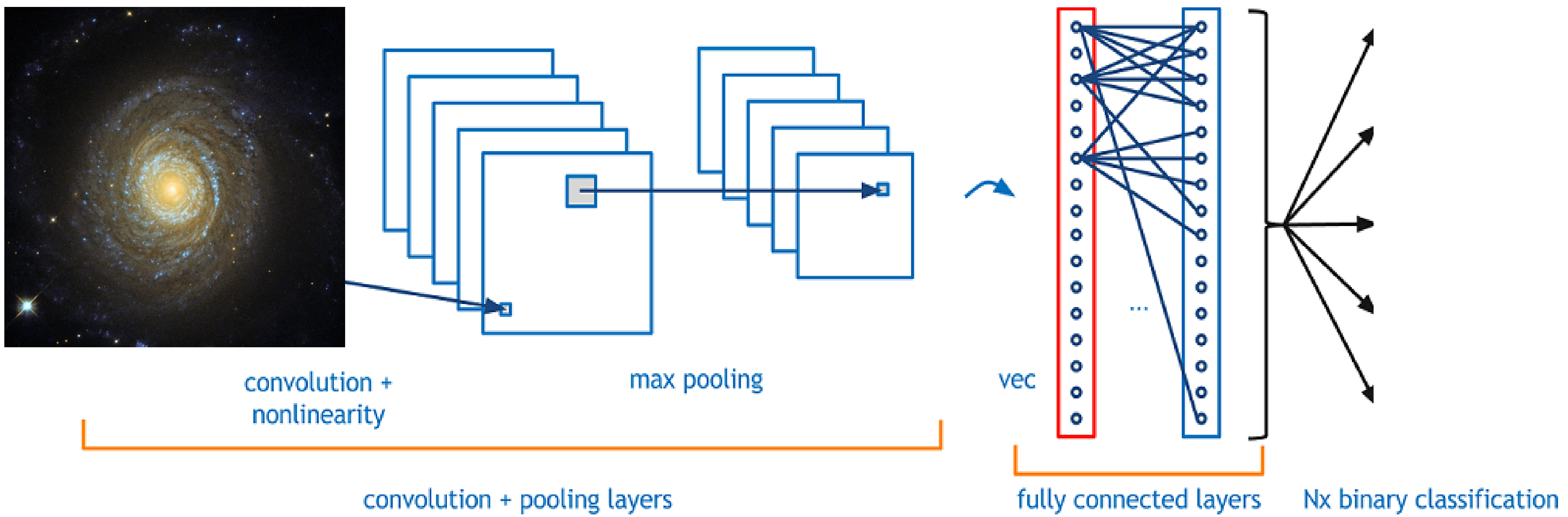


Feature detector



Block representation

Convolutional nets



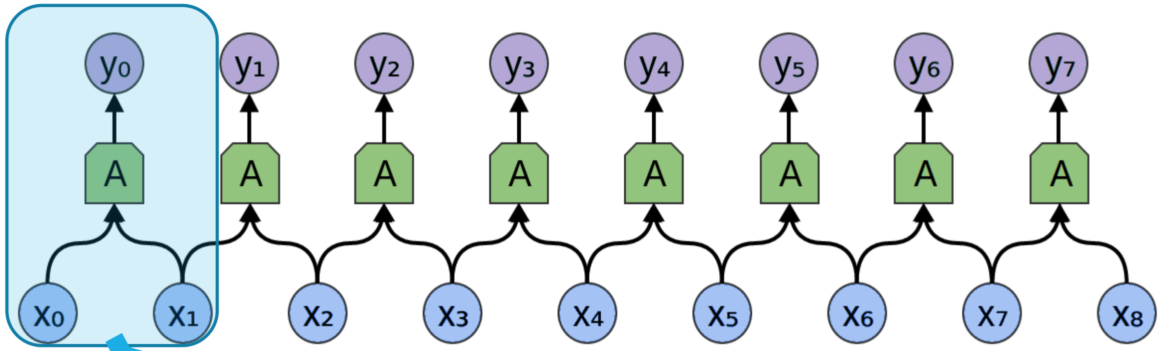
adeshpande3.github.io

Learnable convolution

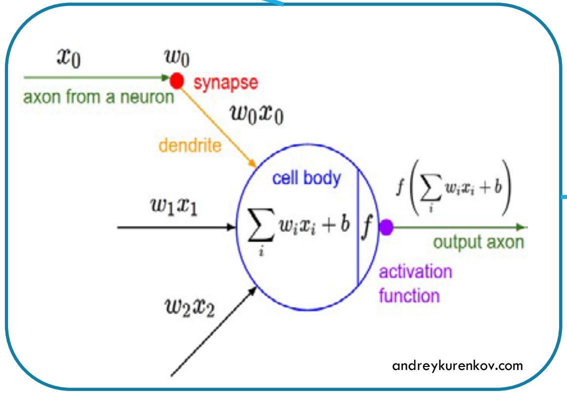
Learnable kernels

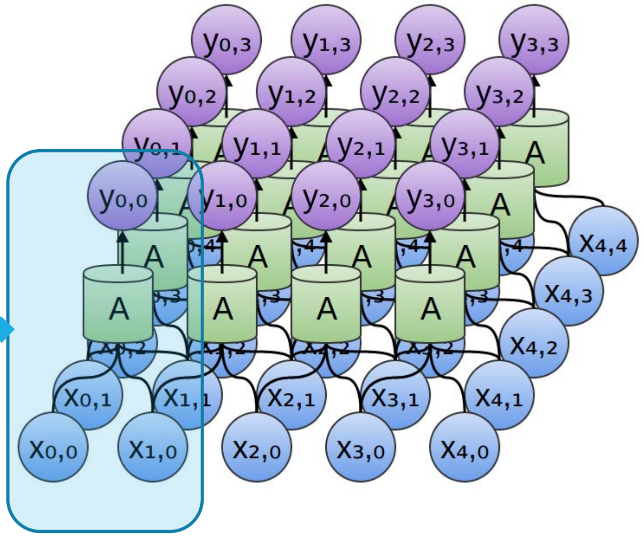
$$y_i = \sum_c \boxed{K(c)} x_{i+c}$$

$$y_{ij} = \sum_{c,d} \boxed{K(c,d)} x_{i+c,j+d}$$



http://colah.github.io/posts/2015-09-NN-Types-FP/

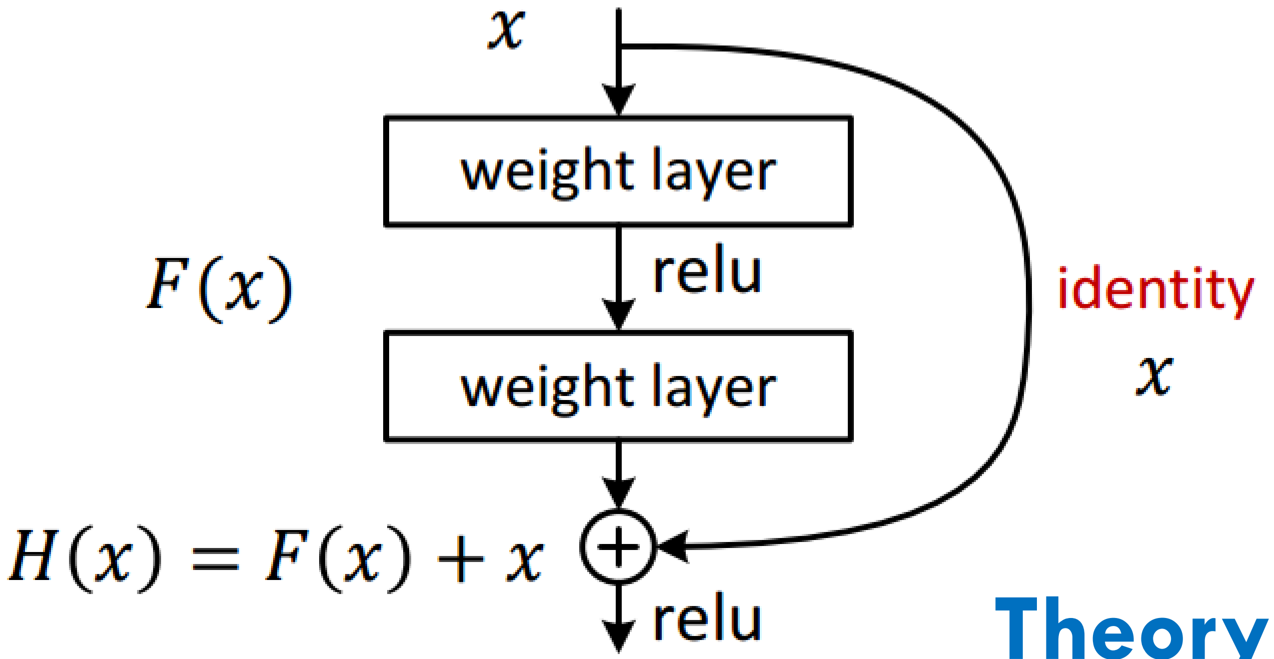




Feature detector,
often many

Skip-connections

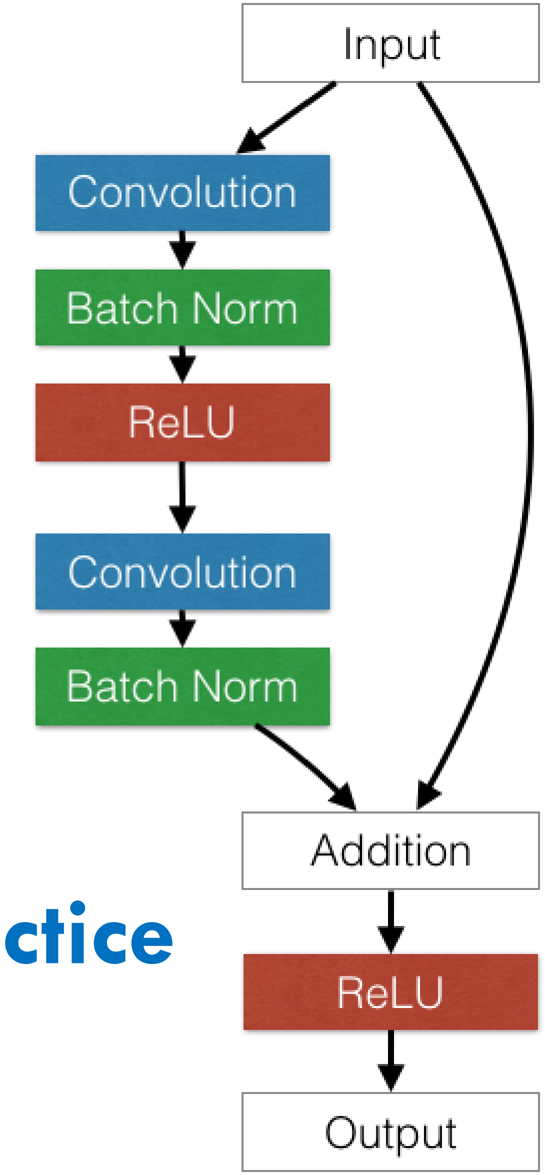
- Residual net



Theory

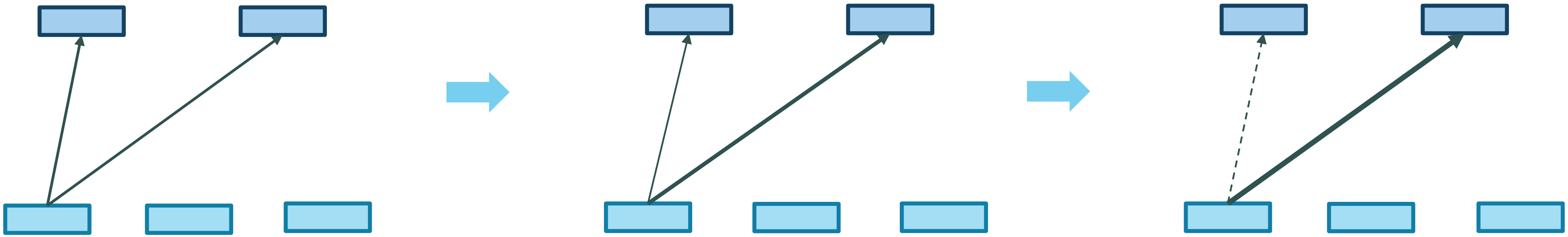
http://qiita.com/supersaiakujin/items/935bbc9610d0f87607e8

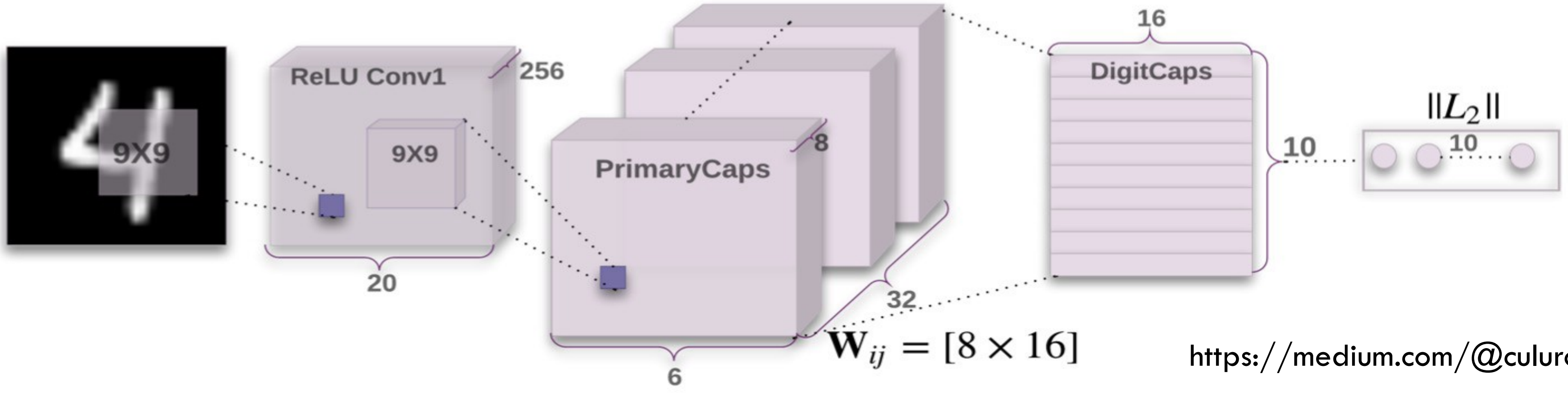
Practice



http://torch.ch/blog/2016/02/04/resnets.html

CapsNet (Hinton's group)





https://medium.com/@culurciello

Attention mechanisms

Need attention model to select or ignore certain inputs

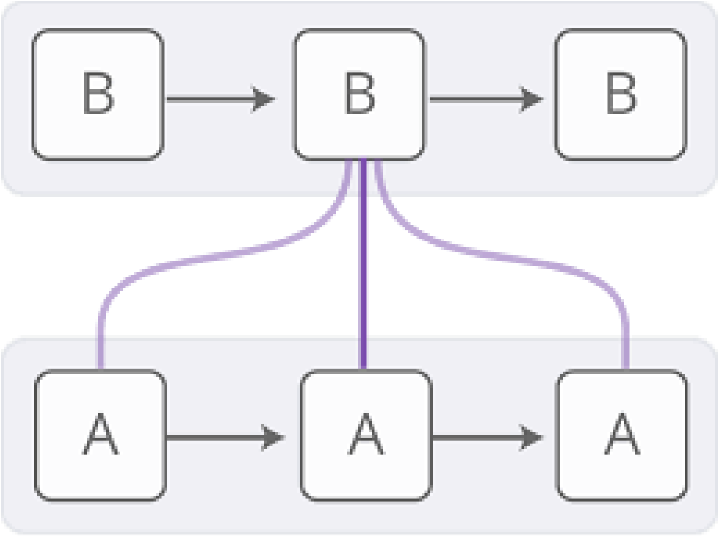
Human exercises great attention capability – the ability to filter out unimportant noises

- Foveating & saccadic eye movement

In life, events are not linear but interleaving.

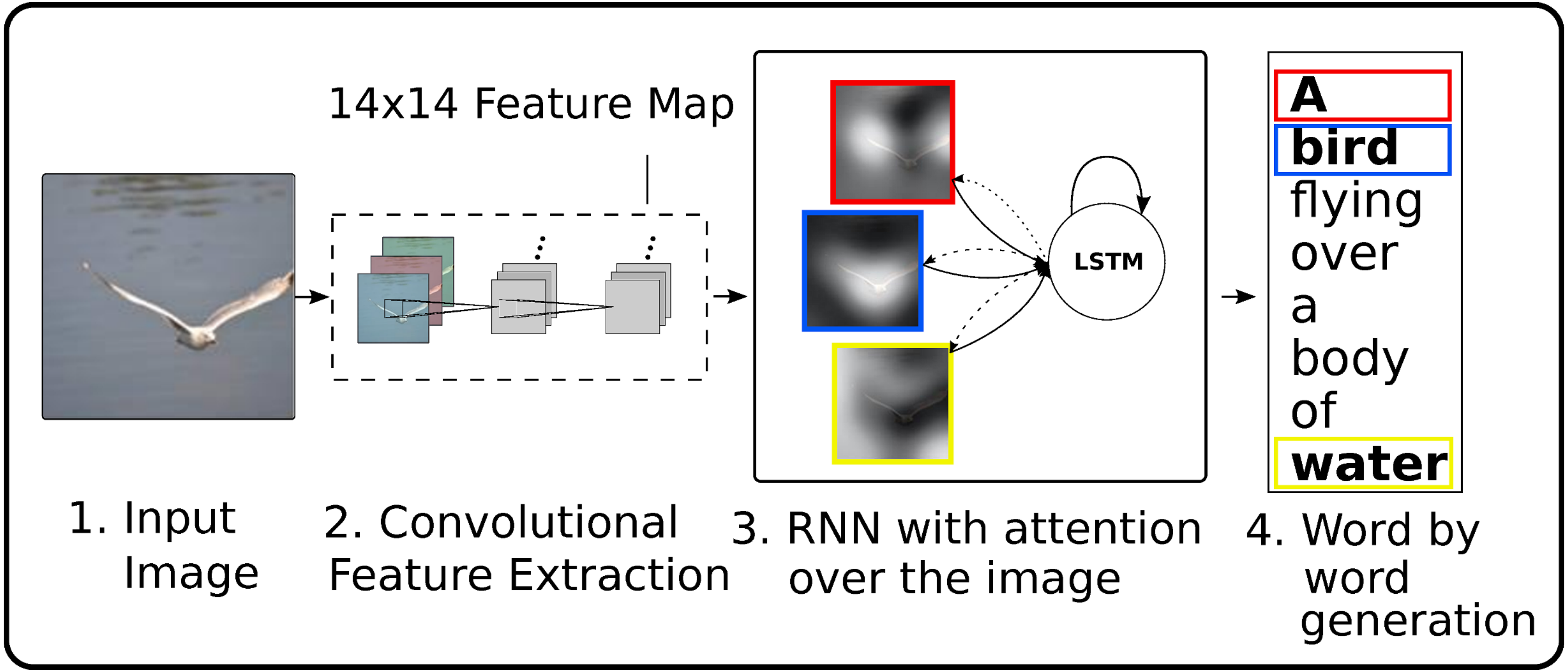
<u>Pooling</u> (as in CNN) is also a kind of attention

Routing (as in CapsNet) is another example.



http://distill.pub/2016/augmented-rnns/

Show, Attend and Tell



Show, Attend and Tell: Neural Image Caption Generation with Visual Attention, K. Xu , J. Ba, R. Kiros, K. Cho, A. Courville, R. Salakhutdinov, R. Zemel, Y. Bengio

Supervised deep learning: steps

Step 0: Collect LOTS of high-quality data

- Corollary: Spend LOTS of time, $$ and compute power

Step 1: Specify the **computational graph** Y = F(X; W)

Step 2: Specify the loss L(W; D) for data D = {(X1,Y1), (X2,Y2),… }

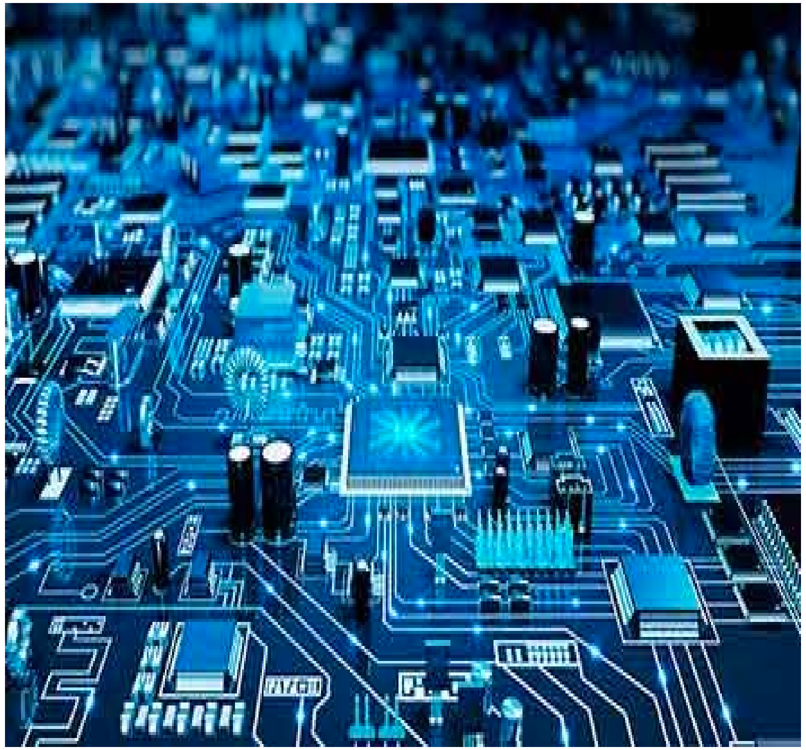
Step 3: Differentiate the loss w.r.t. W (now mostly automated)

Step 4: Optimize the loss (a lot of tools available)

Deep learning as new electronics (or LEGO?)

Analogies:

- Neuron as feature detector → SENSOR, FILTER
- Multiplicative gates → AND gate, Transistor, Resistor
- Attention mechanism → SWITCH gate
- Memory + forgetting → Capacitor + leakage
- Skip-connection → Short circuit
- Computational graph → Circuit
- Compositionality → Modular design



Relationships

- **Now**: Electronics redesigned to support tensors in deep learning
- Prediction: Deep learning helps to design faster electronics

Deep generative models

Many applications:

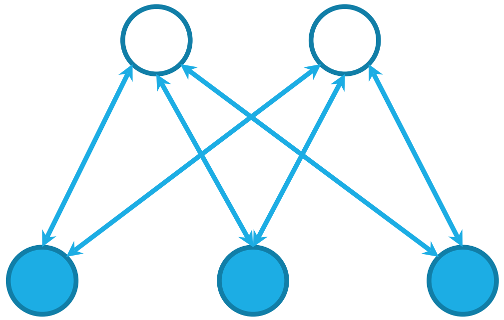
- Text to speech
- **Simulate data that are hard to obtain/share in real life (e.g., healthcare)**
- Generate meaningful sentences conditioned on some input (foreign language, image, video)
- Semi-supervised learning
- Planning

$$\mathbf{v} \sim P_{model}(\mathbf{v})$$

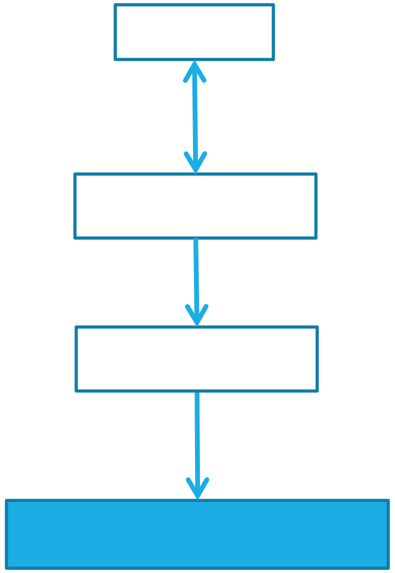
$$P_{model}(\mathbf{v}) \approx P_{data}(\mathbf{v})$$

A family: RBM → DBN → DBM

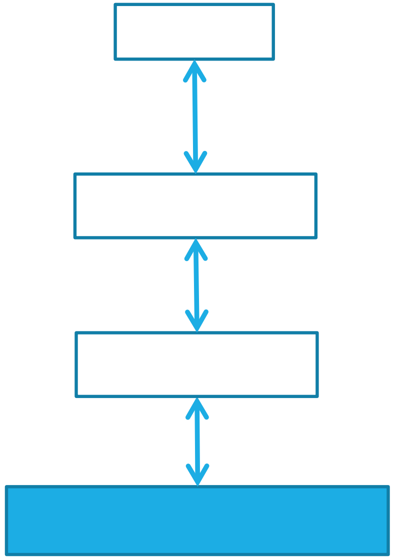
$$p(\mathbf{v},\mathbf{h};\psi) \propto \exp\left[-\underbrace{E(\mathbf{v},\mathbf{h};\psi)}_{energy}\right]$$



Restricted Boltzmann Machine
(~1994, 2001)



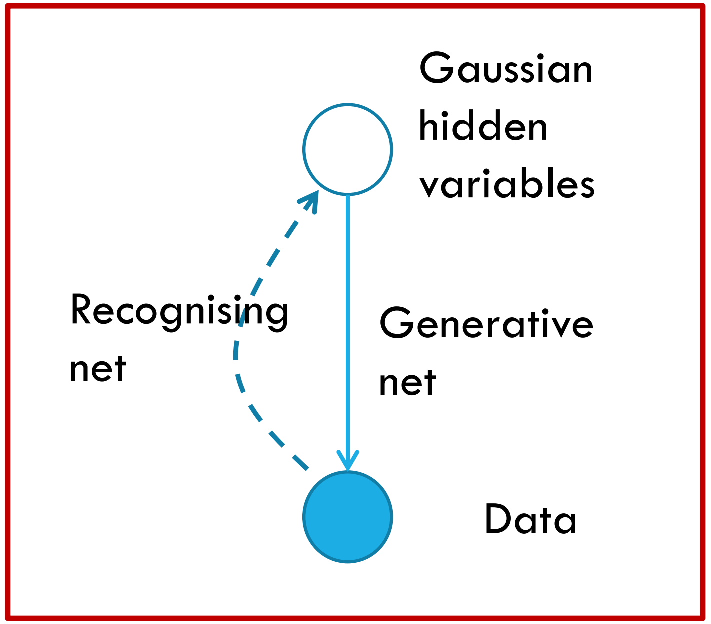
Deep Belief Net
(2006)

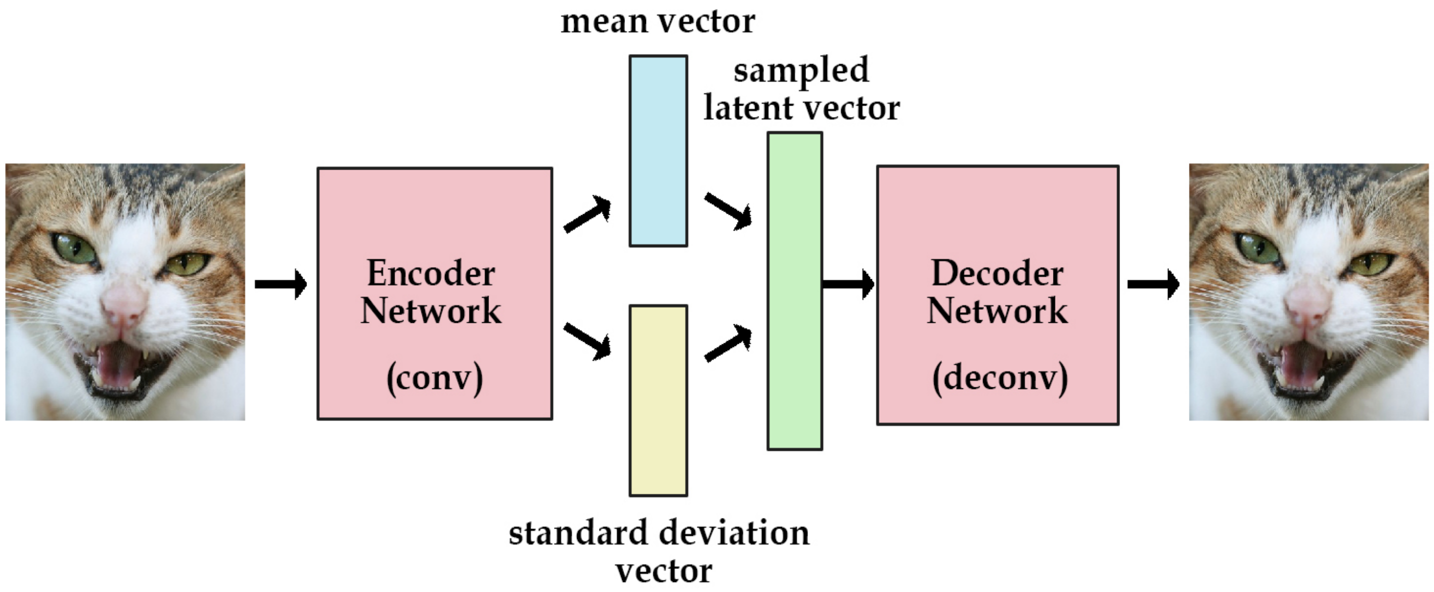


Deep Boltzmann Machine
(2009)

Variational Autoencoder
(Kingma & Welling, 2013)

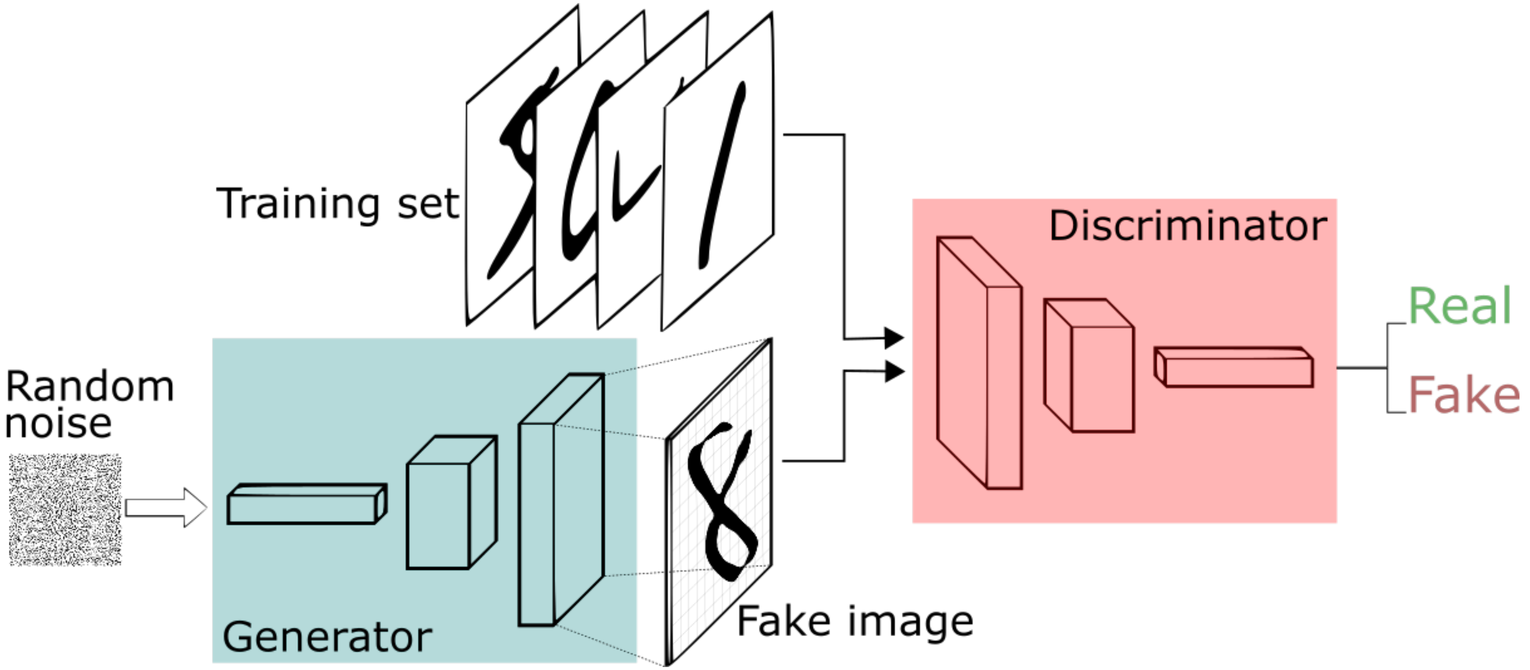
Two separate processes: generative (hidden → visible) versus recognition (visible → hidden)





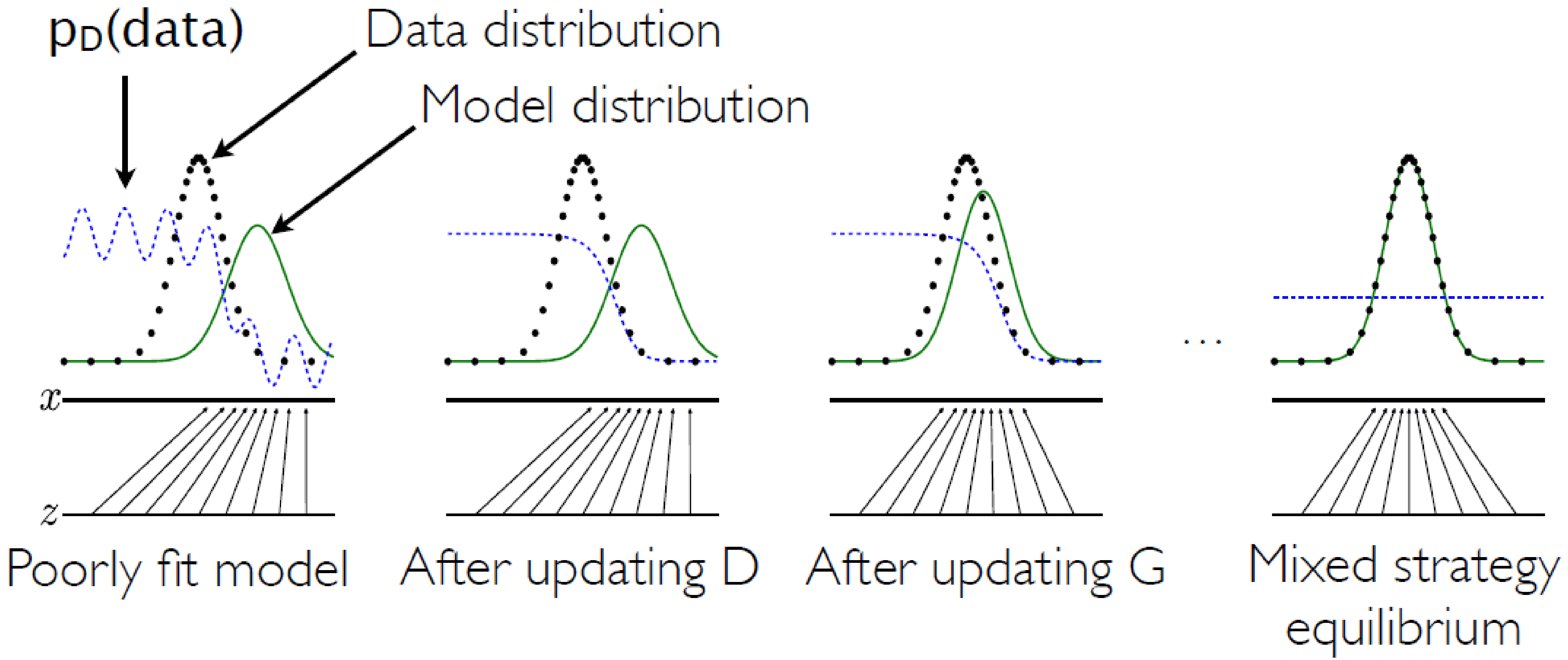
http://kvfrans.com/variational-autoencoders-explained/

Generative Adversarial Networks
(Goodfellow et al, NIPS 2014)

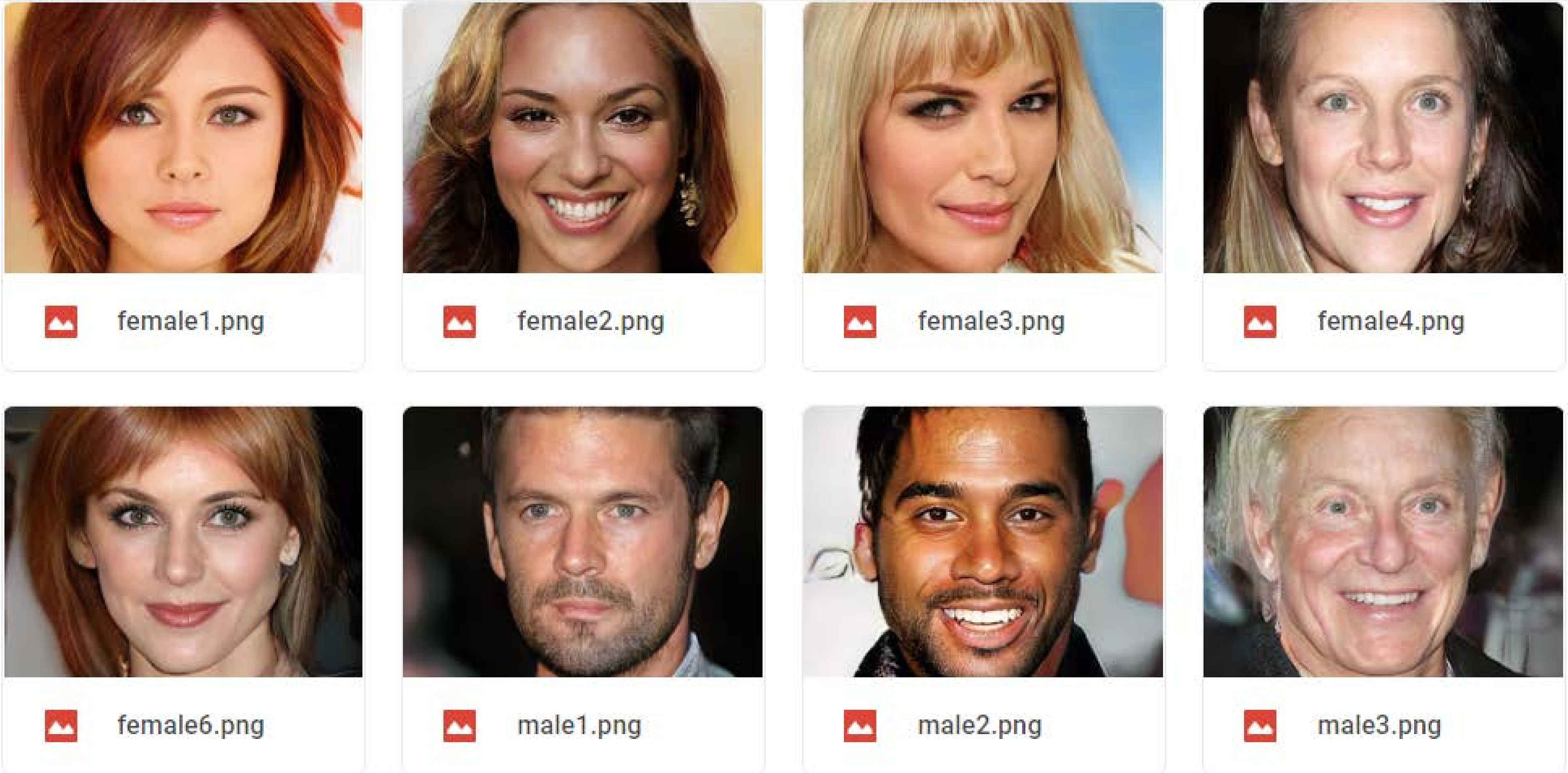


GAN architecture. Source: DL4J

GAN: implicit density models
(Adapted from Goodfellow's, NIPS 2014)



Progressive GAN: Generated images



Karras, T., Aila, T., Laine, S., & Lehtinen, J. (2017). Progressive growing of gans for improved quality, stability, and variation. *arXiv preprint arXiv:1710.10196.*

Why DL works: theory

Expressiveness

- Can represent the complexity of the world → Feedforward nets are universal function approximator
- Can compute anything computable → Recurrent nets are Turing-complete

Learnability

- Have mechanism to learn from the training signals → Neural nets are highly trainable

Generalizability

- Work on unseen data → Deep nets systems work in the wild (Self-driving cars, Google Translate/Voice, AlphaGo)

Why DL works: practice

Strong/flexible priors (80-90% gain):

- Have good features → Feature engineering (Feature learning)
- Respect data structure → HMM, CRF, MRF, Bayesian nets (FFN, RNN, CNN)
- Theoretically motivated model structures, regularisation & sparsity → SVM, compressed sensing (Architecture engineering + hidden norm)
- Respect the manifold assumption, class/region separation → Metric + semi-supervised learning (Sesame net)
- Disentangle factors of variation → PCA, ICA, FA (RBM, DBN, DBM, DDAE, VAE, GAN, multiplicative neuron)

Uncertainty quantification (1-5% gain):

- Leverage Bayesian, ensemble → RF, GBM (Dropout, batch-norm, Bayesian neural nets)

Sharing statistical strength (1-10% gain):

- Encourage model reuse → transfer learning, domain adaption, multitask learning, lifelong learning (Column Bundle, Deep CCA, HyperNet, fast weight)

Two major views of “depth” in DL

[2006-2012] Learning layered representations, from raw data to abstracted goal (DBN, DBM, SDAE, GSN).

- Typically 2-3 layers.
- High hope for unsupervised learning. A conference set up for this: **ICLR**, starting in 2013.

[1991-1997] & [2012-2016] Learning using multiple steps, from data to goal (LSTM/GRU, NTM/DNC, N2N Mem, HWN, CLN).

- Reach hundreds if not thousands layers.
- Learning as credit-assignment.
- Supervised learning won.
- Unsupervised learning took a detour (VAE, GAN, NADE/MADE).

Today’s view: Differentiable programming.

When does deep learning work?

Lots of data (e.g., millions)

Strong, clean training signals (e.g., when human can provide correct labels – cognitive domains).

- Andrew Ng of Baidu: When humans do well within sub-second.

Data structures are well-defined (e.g., image, speech, NLP, video)

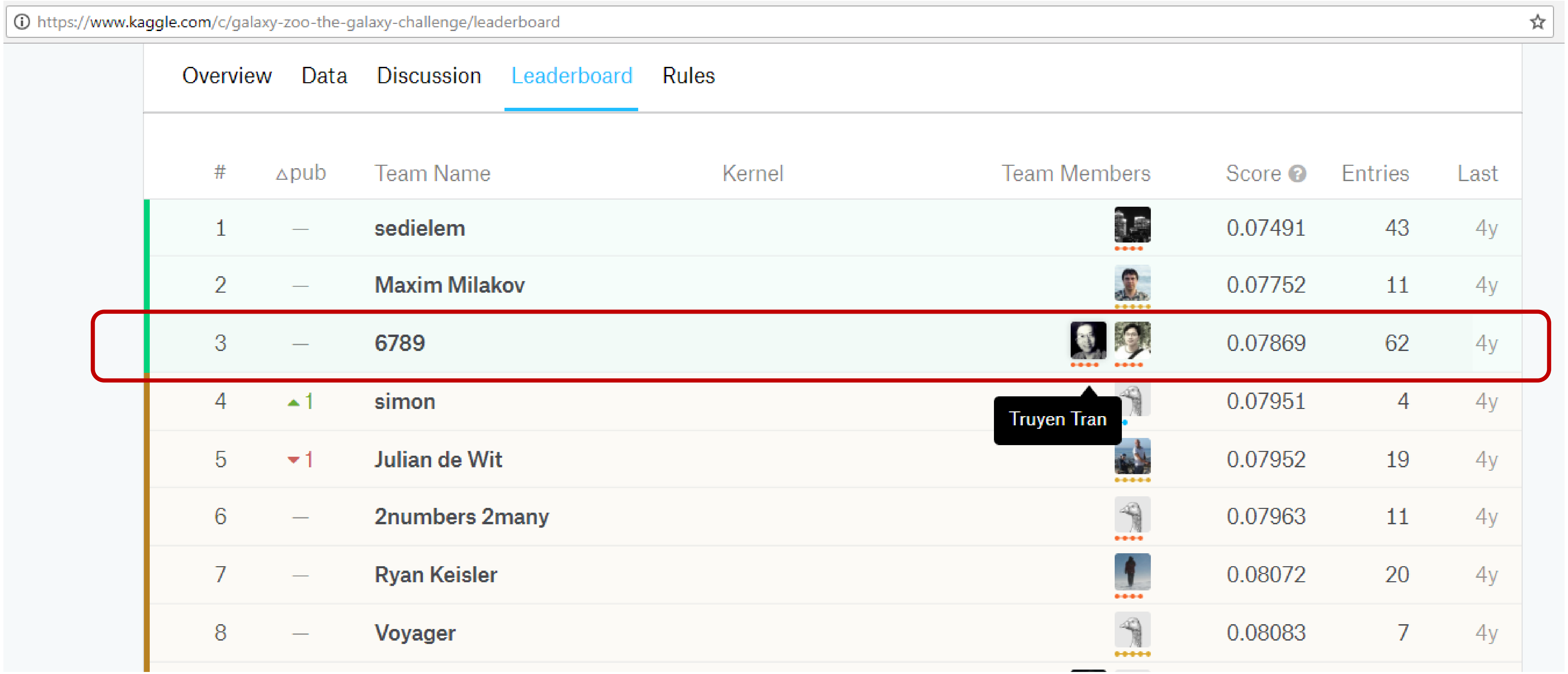
Data is compositional (luckily, most data are like this)

The more primitive (raw) the data, the more benefit of using deep learning.

https://www.kaggle.com/c/galaxy-zoo-the-galaxy-challenge

Applications in astrophysics

Galaxy Zoo challenge: Categorization (joint work with Tu Nguyen)



https://www.kaggle.com/c/galaxy-zoo-the-galaxy-challenge/leaderboard

Overview | Data | Discussion | Leaderboard | Rules

#	△pub	Team Name	Kernel	Team Members	Score	Entries	Last
1	—	sedielem			0.07491	43	4y
2	—	Maxim Milakov			0.07752	11	4y
3	—	6789			0.07869	62	4y
4	▲1	simon			0.07951	4	4y
5	▼1	Julian de Wit			0.07952	19	4y
6	—	2numbers 2many			0.07963	11	4y
7	—	Ryan Keisler			0.08072	20	4y
8	—	Voyager			0.08083	7	4y

Truyen Tran

Our solution

Reduce data variances

- Pre-processing: cropping and down-sampling
- Augmentation: rotation, flipping, zooming, translation

Right “prior” architecture: CNN

- OverFeat for feature extraction & prediction
- MLP on top to improve further

Ensemble methods

- Simple averaging of many models

Sermanet, Pierre, et al. "Overfeat: Integrated recognition, localization and detection using convolutional networks." *arXiv preprint arXiv:1312.6229* (2013).

Network architecture

Layer	1	2	3	4	5	6	7	8	9
Stage	conv+max	conv+max	conv	conv	conv	conv+max	full	full	full
# channels	48	96	192	192	384	384	2048	2048	37
Filter size	5×5	5×5	3×3	3×3	3×3	3×3	-	-	-
Conv. stride	1×1	1×1	1×1	1×1	1×1	1×1	-	-	-
Pooling size	3×3	2×2	-	-	-	3×3	-	-	-
Pooling stride	3×3	2×2	-	-	-	3×3	-	-	-
Zero-padding size	-	-	-	-	-	-	-	-	-
Spatial input size	120×120	39×39	18×18	16×16	14×14	12×12	4×4	1×1	1×1

Model: https://github.com/tund/kaggle-galaxy-zoo/blob/master/report/gz_report.pdf?raw=true

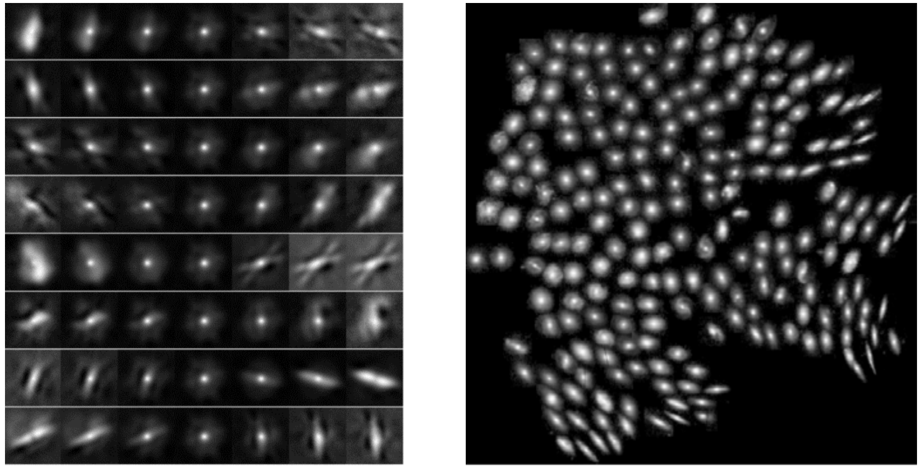
Code: https://github.com/tund/kaggle-galaxy-zoo

Deep generative models for astronomical imaging

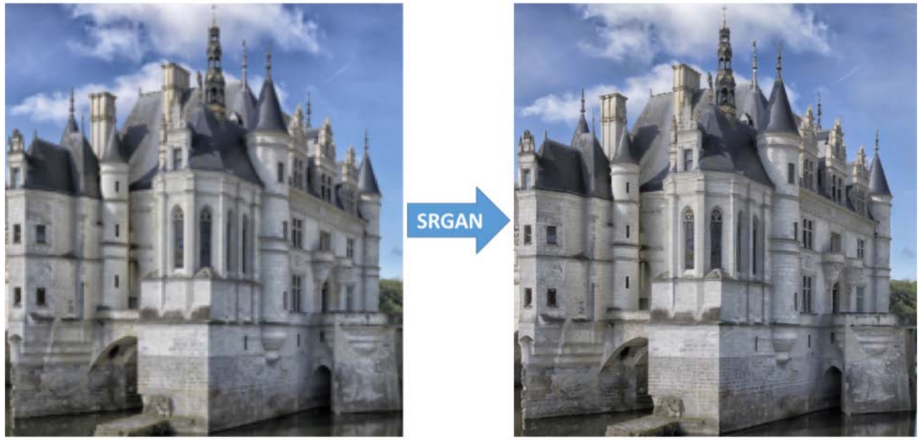
DGM achieved excellent results on various tasks

- Image generation (GAN [1], VAE[2], SAGAN[3])
- Image super resolution (SRGAN [4])
- Image denoising
- Image inpainting

SAGAN: self attention GAN
SRGAN: super resolution GAN



Source: Regier et al, ICML'15



LG Image Generated Image

Source: TFLayer

Regier, J., Miller, A., McAuliffe, J., Adams, R., Hoffman, M., Lang, D., Schlegel, D. and Prabhat, M., 2015, June. Celeste: Variational inference for a generative model of astronomical images. In *International Conference on Machine Learning* (pp. 2095-2103).

DGM for cosmology

DGM can be used to speed up/replace complex experiments/computation:

- Fast Cosmic Web Simulations with Generative Adversarial Networks. Rodriguez et al.
- Enabling Dark Energy Science with Deep Generative Models of Galaxy Images. Ravanbakhsh et al.

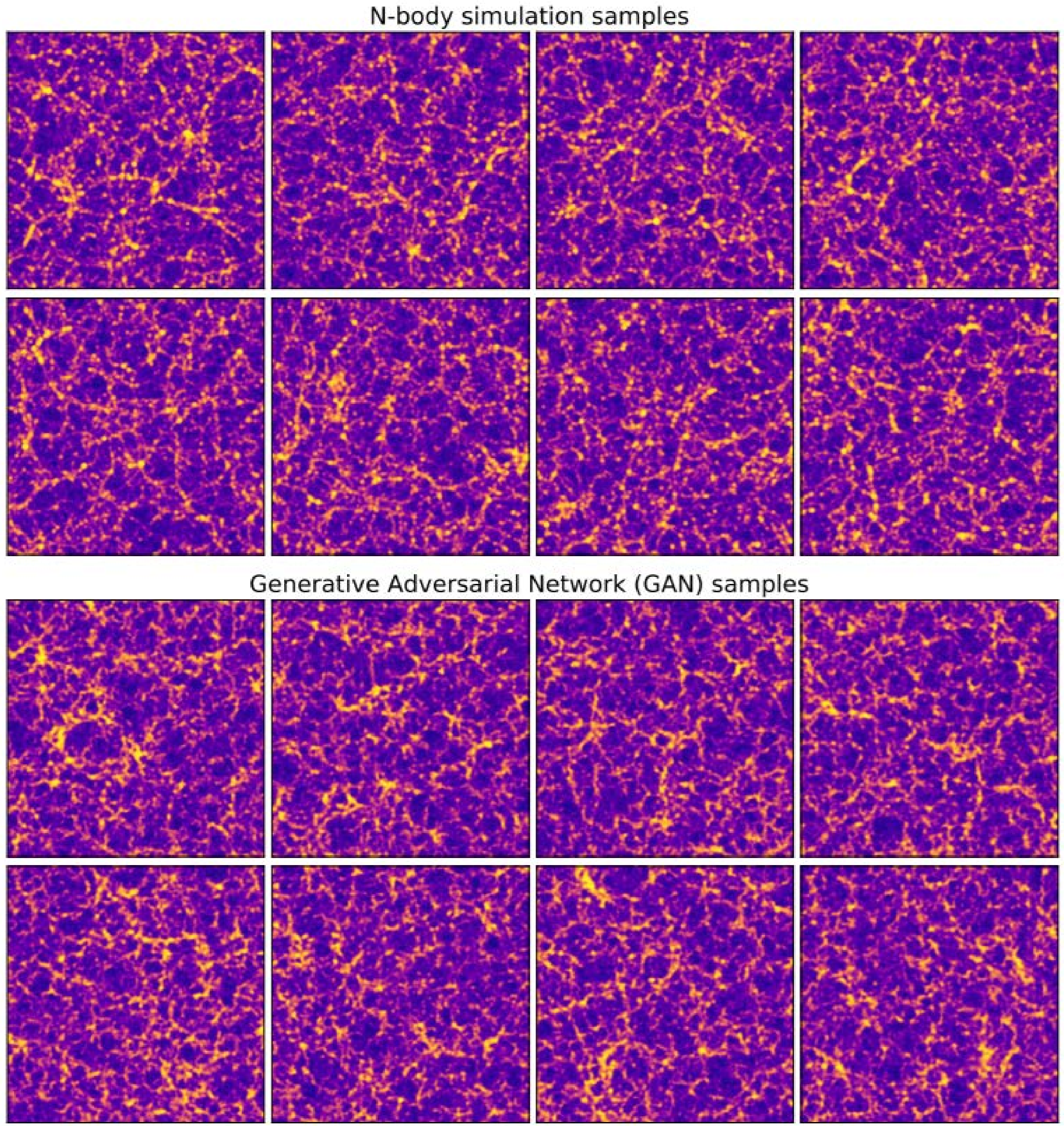


Figure 1: Samples from N-body simulations (top two rows) and from our GAN model (bottom two rows) for a box size of 500 Mpc. Note that the transformation in Equation 3.1 with $k = 20$ was applied to the images shown above for better readability.

References

[1] Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A. and Bengio, Y., 2014. Generative adversarial nets. In Advances in neural information processing systems (pp. 2672-2680).

[2] Kingma, D.P. and Welling, M., 2013. Auto-encoding variational bayes. arXiv preprint arXiv:1312.6114.

[3] Zhang, H., Goodfellow, I., Metaxas, D. and Odena, A., 2018. Self-Attention Generative Adversarial Networks. arXiv preprint arXiv:1805.08318.

[4] Ledig, C., Theis, L., Huszár, F., Caballero, J., Cunningham, A., Acosta, A., Aitken, A., Tejani, A., Totz, J., Wang, Z. and Shi, W., 2017. Photo-realistic single image super-resolution using a generative adversarial network. *arXiv preprint*.

[5] Regier, J., Miller, A., McAuliffe, J., Adams, R., Hoffman, M., Lang, D., Schlegel, D. and Prabhat, M., 2015, June. Celeste: Variational inference for a generative model of astronomical images. In *International Conference on Machine Learning* (pp. 2095-2103).

[6] Andres C Rodriguez, Tomasz Kacprzak, Aurelien Lucchi, Adam Amara, Raphael Sgier, Janis Fluri, Thomas Hofmann, Alexandre Réfrégier. Fast Cosmic Web Simulations with Generative Adversarial Networks. arXiv preprint arXiv:1801.09070v2

[7] Ravanbakhsh, S., Lanusse, F., Mandelbaum, R., Schneider, J.G. and Poczos, B., 2017. Enabling Dark Energy Science with Deep Generative Models of Galaxy Images. In *AAAI* (pp. 1488-1494).

[8] Abeer Alsaiari, Manu Mathew Thomas, Ridhi Rustagi. Image Denoising Using a Generative Adversarial Network

[9] Yang, Q., Yan, P., Zhang, Y., Yu, H., Shi, Y., Mou, X., Kalra, M.K., Zhang, Y., Sun, L. and Wang, G., 2018. Low dose CT image denoising using a generative adversarial network with Wasserstein distance and perceptual loss. *IEEE transactions on medical imaging*.

[10] Yu, J., Lin, Z., Yang, J., Shen, X., Lu, X. and Huang, T.S., 2018, January. Generative Image Inpainting with Contextual Attention. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (pp. 5505-5514).