



Discovery



Diagnosis



Prognosis



Care

Deep learning for biomedicine



Truyen Tran
Deakin University

Seoul, Nov 2017



truyen.tran@deakin.edu.au



truyentran.github.io



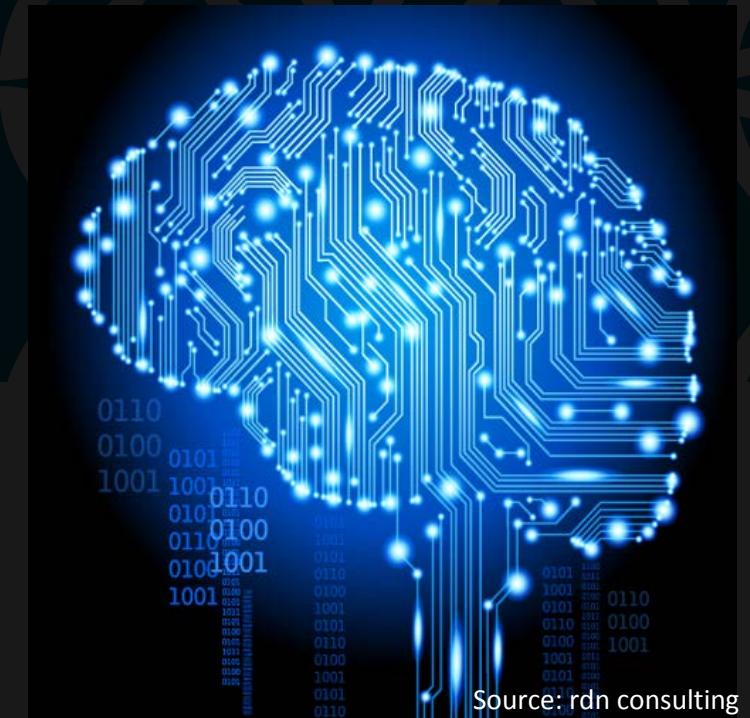
@truyenoz



letdataspeak.blogspot.com



goo.gl/3jJ100



The screenshot shows a web browser window with the URL [https://medium.c...](https://medium.com/). The page title is "Medium" and the main heading is "Andrew Ng Says Enough Papers, Let's Build AI Now!". Below the heading is a photograph of Andrew Ng standing at a whiteboard during a presentation at the Frontiers Conference in Santa Clara. The whiteboard has various handwritten notes and diagrams, including the words "FRONTIERS CONFERENCE", "SANTA CLARA", and "Data Mining". At the bottom of the page, there is a sidebar with a "Never miss a story from Synced" section, a "GET UPDATES" button, and two PDF files listed: "1611.09340.pdf" and "CB-Insights_Health....pdf". The date "12/11/2017" is also visible at the bottom.

Andrew Ng:

"AI is the new electricity"

"We have enough papers. **Stop publishing**, and start transforming people's lives with technology!"

We will quickly solve “easy” problems of the form:

$$A \rightarrow B$$

BUT ... Should we solve all problems of and for those Internet giants like Google, Facebook & Baidu?

Resources

Slides and references:

- <https://truyentran.github.io/acml17-tute.html>

Key survey paper (updated frequently):

- **Ching, Travers, et al. "Opportunities And Obstacles For Deep Learning In Biology And Medicine." *bioRxiv* (2017): 142760**

The Team



Agenda

Topic 1: Introduction (20 mins)

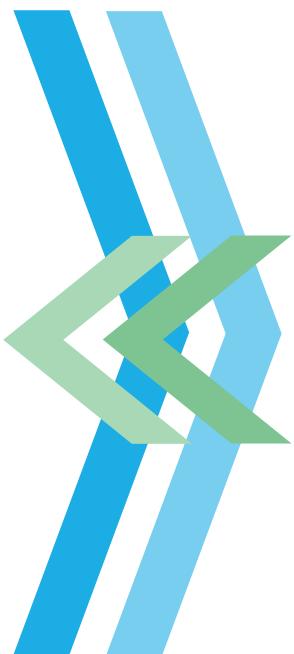
Topic 2: Brief review of deep learning (25 mins)

- Classic architectures
- Capsules
- Graphs
- Memory-augmented nets

Topic 3: Genomics (25 mins)

- Nanopore sequencing
- Genomics modelling

QA (10 mins)



Break (20 mins)

Topic 4: Biomedical imaging (15 mins)

- Cellular imaging
- Diagnostics imaging
- EEG/ECG

Topic 5: Healthcare (25 mins)

- Time-series of physio measures
- Trajectories prediction

Topic 6: Generative biomed (30 mins)

- Few-shot learning
- Generative models
- Drug design
- Future outlook

QA (10 mins)

Biomedicine is ripe for ML/DL/AI – or other way around

ML that matters

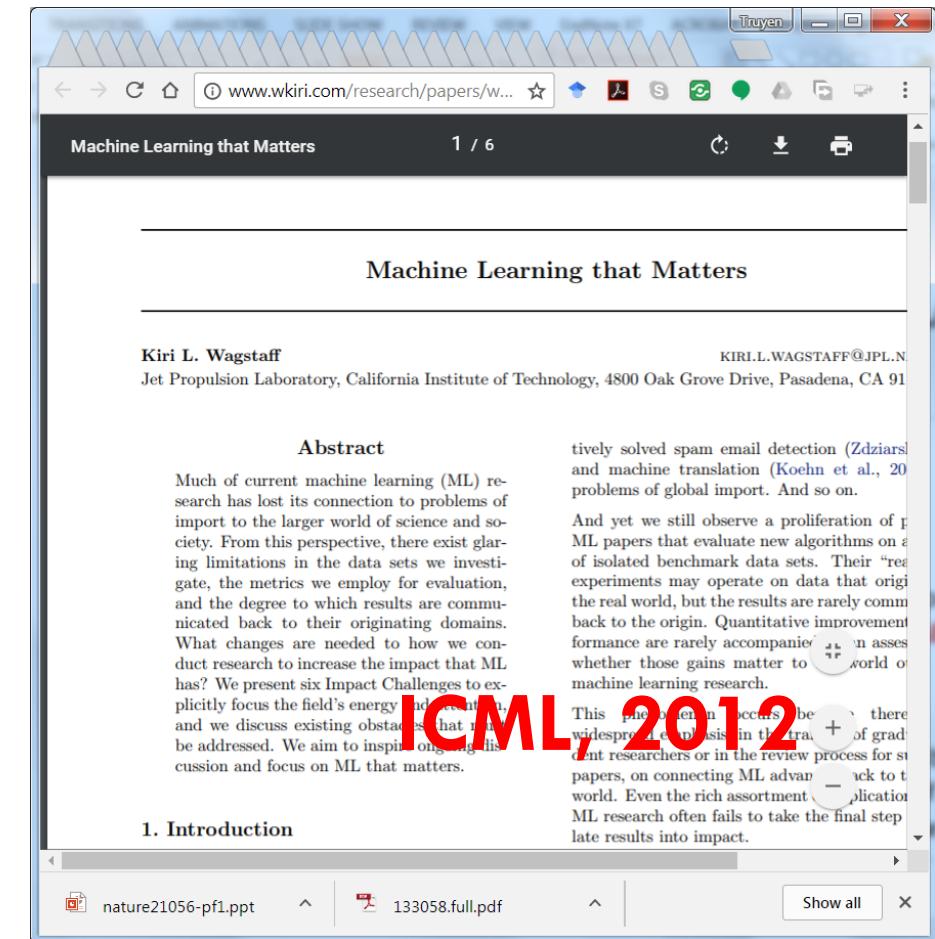
- E.g., huge successes in radiology with off-the-shelf CNNs
- Big business opportunities, e.g., IBM Watson for Health

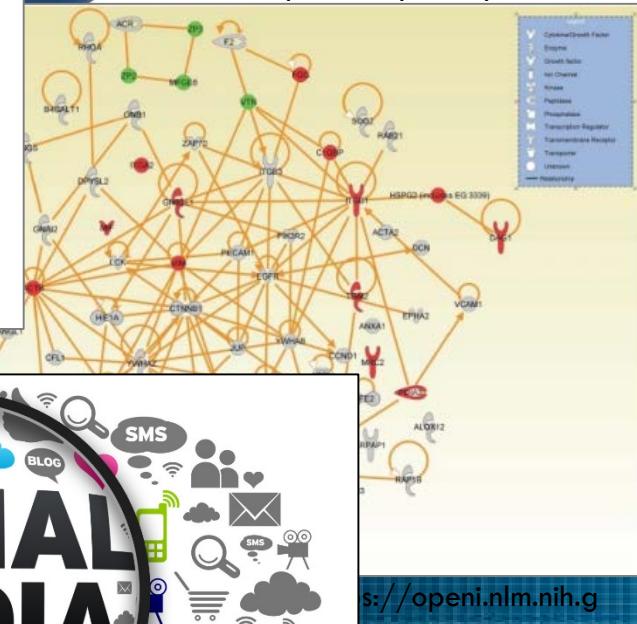
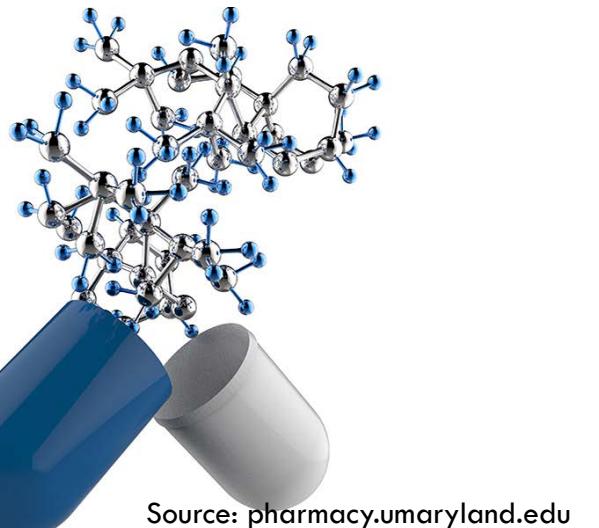
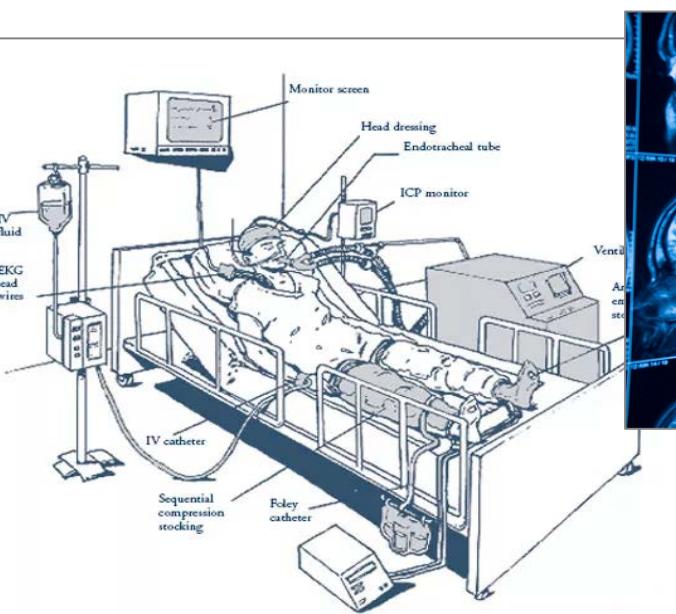
Excellent testbed for machine learning techniques

- Any modality: 2D-4D vision, time-series, 1D signals, sound, text, social network, graphs.
- **For DL, any neural architectures: CNN/CapsNet, RNN, Memory, DBN/VAE/GAN**
- An excellent escape from the UCI datasets!

Excellent sources of new problems

- Metric scale from nano-meter (atoms) to meters (human body and brain).
- Time scale from mini-seconds (ion channels) to 100 years.
- Complexity unimaginable (e.g., brain, DNA, cell networks).

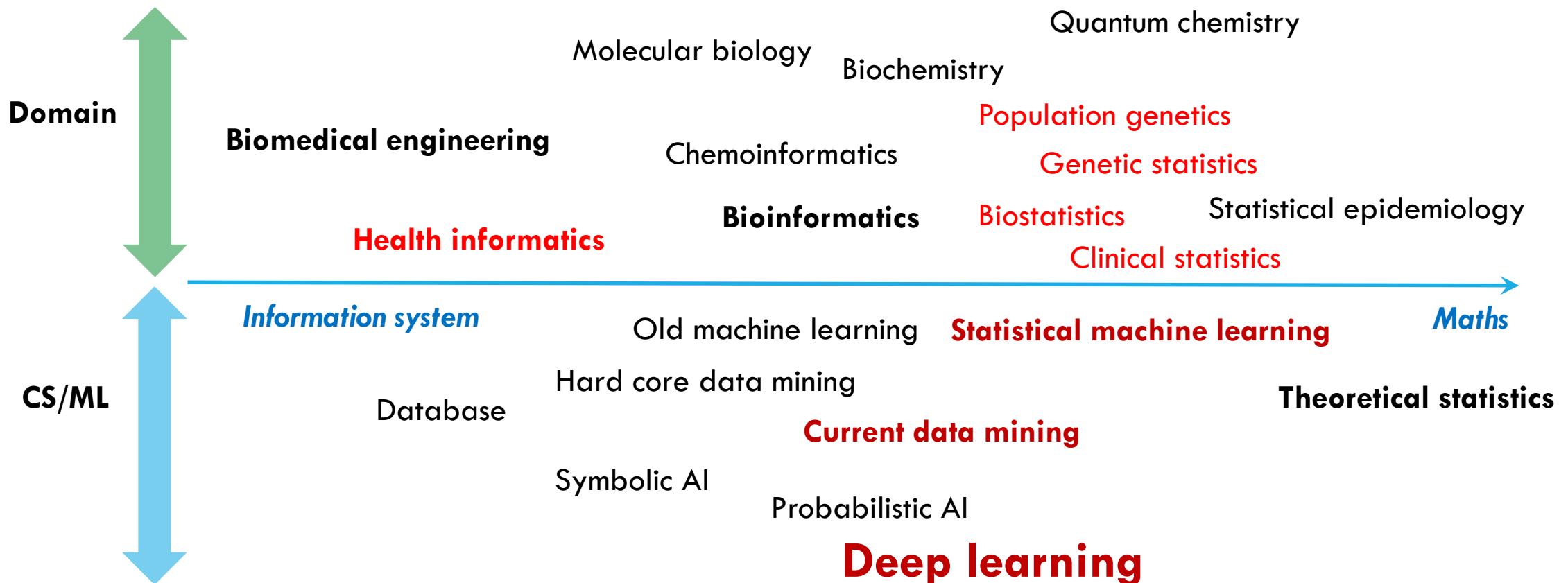




Big Rooms in Biomedicine



First thing first: Speak their languages



Using intuition and domain knowledge

Intuition is important.

- There are infinite number of hypotheses
- We need to search for some highly probable ones!

But it can be deadly wrong!

- A recently discharged patient can be readmitted right away (just like not treated).
- Good doctors can be associated with high rate of mortality and readmission.

Domain knowledge is critical

- Check the literature. Obey the laws. Follow protocols.
- Do the home work, e.g., pregnancy diabetes; women with prostate cancer; men with breast cancer.
- **Choose right neural architectures!**

But ... a lot of data can support any dumb tricks!

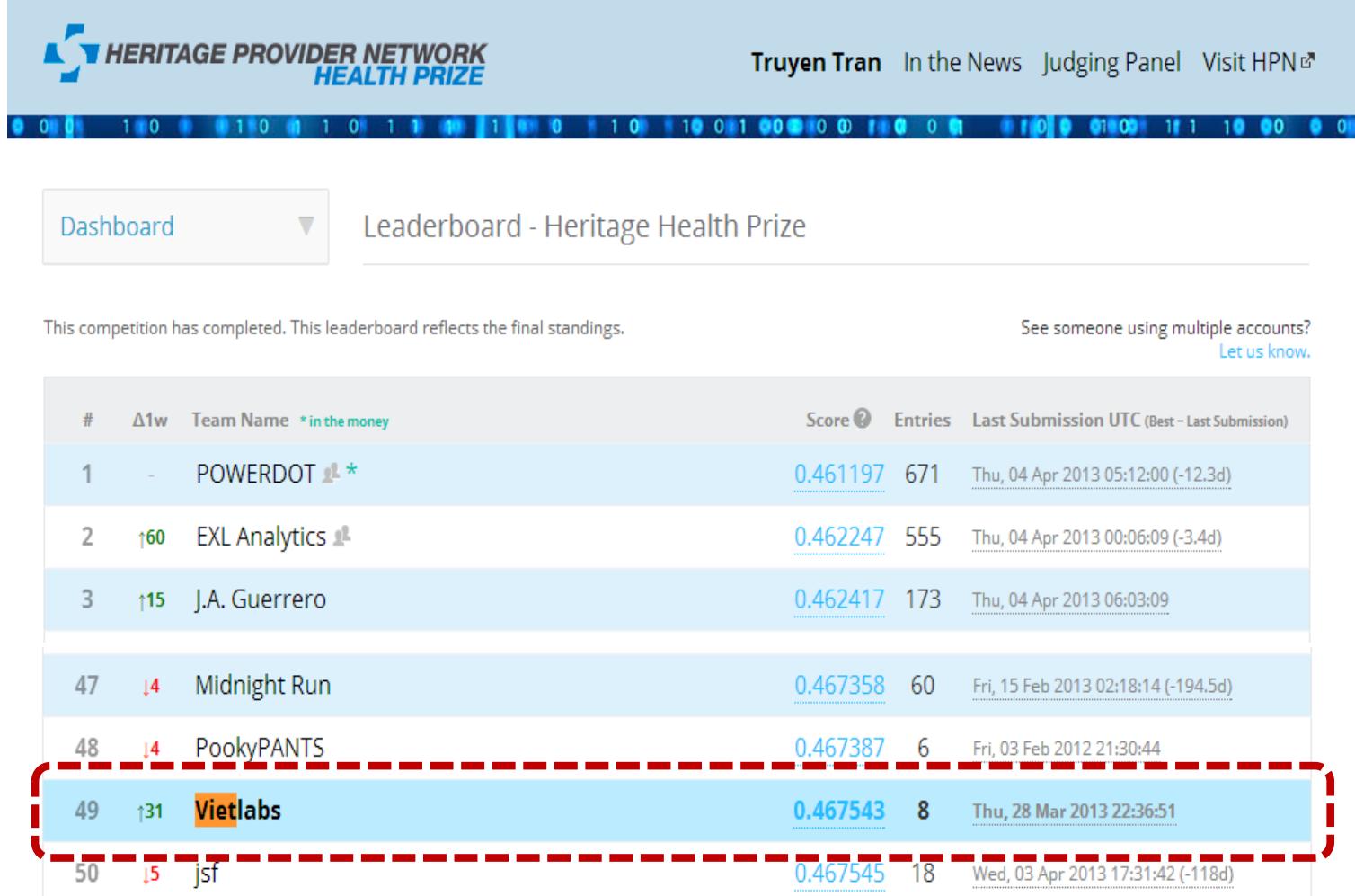
Big data mitigate the need of domain expertise

\$3M Prize, 3 years

170K patients, 4 years worth of data

Predict length-of-stay next year

Not deep learning yet (early 2013), but strong ensemble needed → suggesting dropout/batch-norm



The screenshot shows the final standings of the Heritage Health Prize competition. The table has columns for rank (#), change in weight (Δ1w), team name, *in the money status, score, entries, and last submission UTC. A red dashed box highlights the entry for Vietlabs.

| # | Δ1w | Team Name | *in the money | Score | Entries | Last Submission UTC (Best - Last Submission) |
|----|-----|---------------|---------------|----------|---------|--|
| 1 | - | POWERDOT | ⬆* | 0.461197 | 671 | Thu, 04 Apr 2013 05:12:00 (-12.3d) |
| 2 | ↑60 | EXL Analytics | ⬇ | 0.462247 | 555 | Thu, 04 Apr 2013 00:06:09 (-3.4d) |
| 3 | ↑15 | J.A. Guerrero | | 0.462417 | 173 | Thu, 04 Apr 2013 06:03:09 |
| 47 | ↓4 | Midnight Run | | 0.467358 | 60 | Fri, 15 Feb 2013 02:18:14 (-194.5d) |
| 48 | ↓4 | PookyPANTS | | 0.467387 | 6 | Fri, 03 Feb 2012 21:30:44 |
| 49 | ↑31 | Vietlabs | | 0.467543 | 8 | Thu, 28 Mar 2013 22:36:51 |
| 50 | ↓5 | jsf | | 0.467545 | 18 | Wed, 03 Apr 2013 17:31:42 (-118d) |

Let's be warned!

2011



2017

"He said later that the background information Watson provided, including medical journal articles, was helpful, giving him more confidence that using a specific chemotherapy was a sound idea.

But the system did not directly help him make that decision, nor did it tell him anything he didn't already know."

A STAT INVESTIGATION
IBM pitched its Watson supercomputer as a revolution in cancer care. It's nowhere close
By CASEY ROSS @byCaseyRoss and IKE SWETLITZ @ikeswetlitz
SEPTEMBER 5, 2017

What make biomedicine hard for deep learning?

Great diversity but may be small in size

High uncertainty, low-quality/missing data

Reusable models do not usually exist

Human doesn't know how to read biomedicine (Brendan Frey, U of Toronto)

Require deep thinking for a reasonable deep architecture

However, at the end of the day, we need only a few generic things:

- Vector → DNN (e.g., highway net) | Sequence → RNN (e.g., LSTM, GRU)
- Repeated motifs → CNN | Set → Attention
- Graphs → Conv graphs; Column Networks

Health data is exceptionally hard to model

Many diseases are not fully understood

- Cancers, mental health

Many treatments have little or unknown effects

Care processes are complex: Protocols, regulations, multiple stakeholders (patients, family, nurse, doctor, hospital, community, government, the public).

Data is (usually) small, biased, noisy, irregular/episodic, missing with external interventions

Data is not shared, due to privacy and ethical concerns

Predicting into future isn't like finding out what is there (e.g., classification)

Decision making is complex.

Doctors are rightfully sceptical of what is new.

Then the good news

Healthcare Remains The Hottest AI Category For Deals

April 12, 2017

f t in e
Artificial Intelligence Digital Health Funding & Dealflow

WHERE IS THIS
COMING FROM?
Start your free trial

Email

SIGN UP

RESEARCH BRIEFS

Industries Geographies Investments & Exits Infographics Reports Events Expert Intelligence

CONNECTED
DEVICES

\$3.8M SENTRIAN

Healthcare AI Investments Rise To Record Rates, Thanks To New Startups

August 8, 2017 f t in e
Artificial Intelligence Digital Health Early-Stage Expert Intelligence

WHERE IS THIS DATA
COMING FROM?
Start your free trial today

Email

H1'17 saw a record number of new startups entering the space, putting 2017 on track to far exceed 2016 numbers.

Healthcare is currently the **leading industry** to adopt and experiment with artificial intelligence-based solutions. Healthcare AI startups have raised over 300 deals in the last 5 years. A huge portion of this, 45%, can be attributed to new startups entering the space raising their first equity rounds.

Secure | https://www.cbinsights.com/research/ai-healthcare-startups-market-map-expert-research/

RESEARCH BRIEFS

Industries Geographies Investments & Exits Infographics Reports Events Expert Intelligence

SIGN UP FOR A FREE TRIAL

Search

Virtual Nurses, Drug Discovery, & More: 101 Artificial Intelligence Startups In Healthcare

September 5, 2017 f t in e
Artificial Intelligence Digital Health Expert Intelligence

WHERE IS THIS DATA
COMING FROM?
Start your free trial today

Email

Healthcare is the hottest sector for artificial intelligence, with startups applying AI to everything from genetic research to emergency room management to clinical trials.

Deals to healthcare-focused AI startups rocketed from fewer than 20 in 2012 to 88

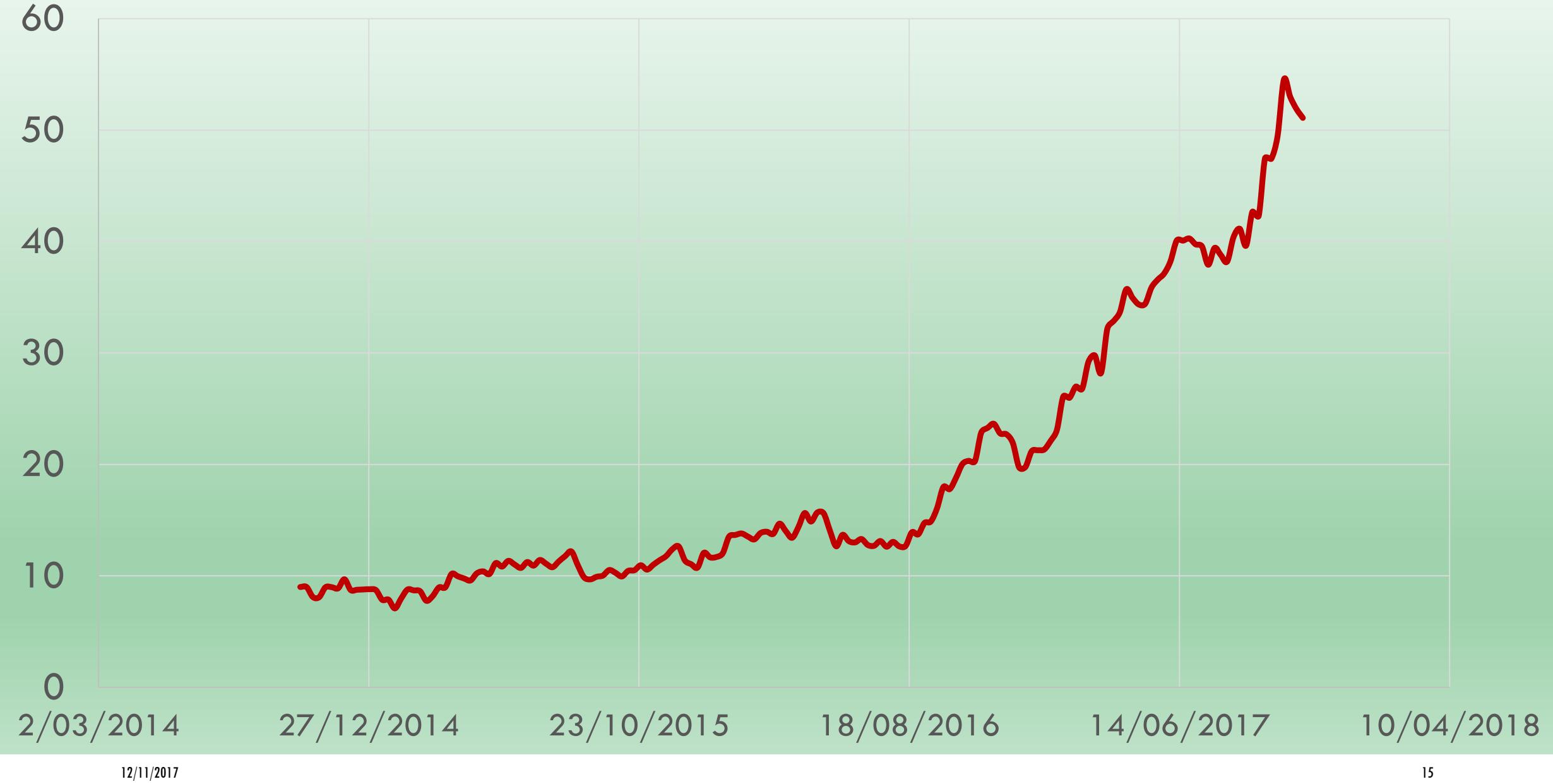
multiTimeline.csv SIT112_unit_detail_r...xls SIT112_teacher_re...xls DATA SCIENCE CO...xls eVALUATE-Full-Uni....pdf

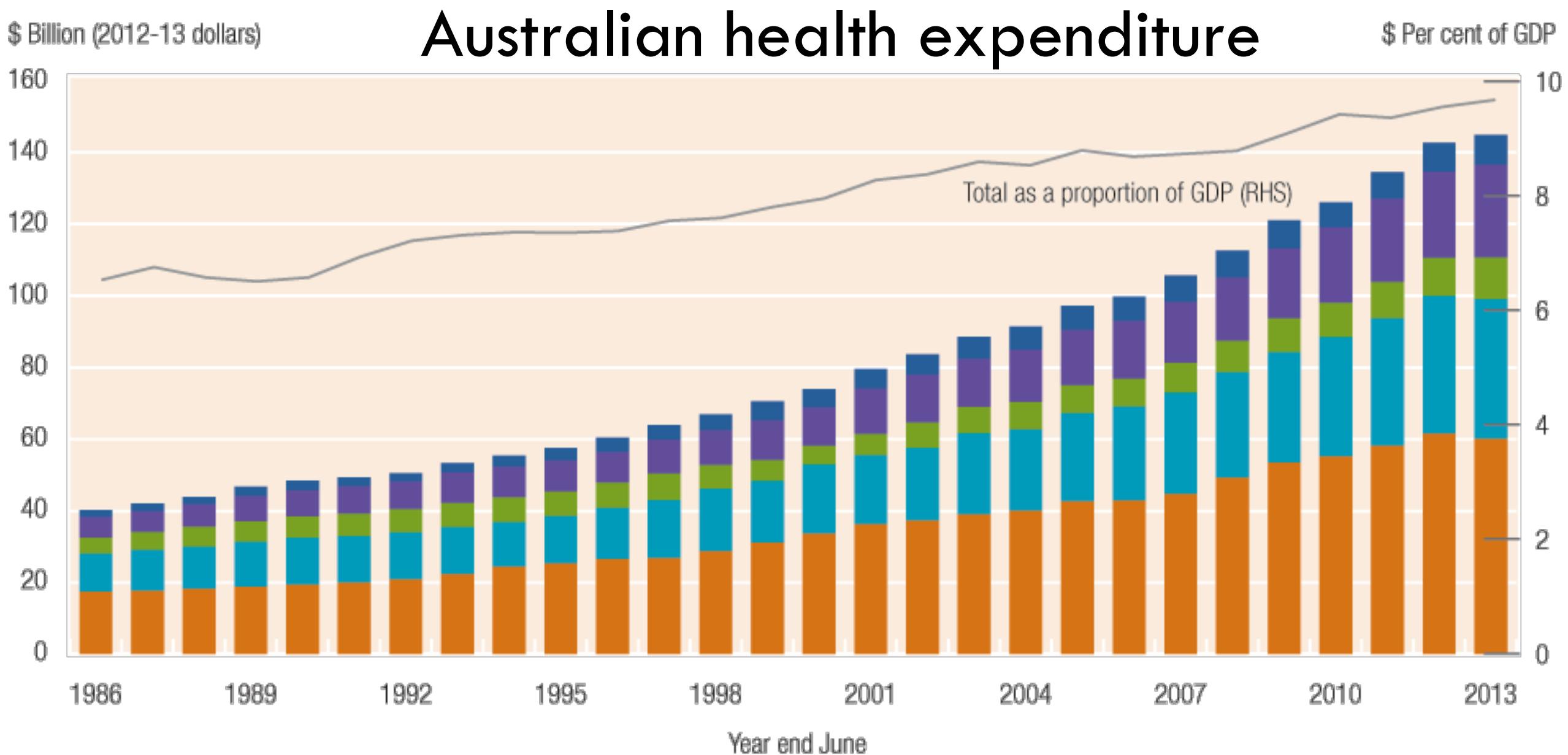
Show all

multiTimeline.csv SIT112_unit_detail_r...xls SIT112_teacher_re...xls DATA SCIENCE CO...xls eVALUATE-Full-Uni....pdf

Show all

Google Trends: "artificial intelligence" + healthcare





Australian
Government

State and territory
governments

Private health
insurance

Patients

Other non-government

It has just started.

Supervised learning
(mostly machine learning)



Will be quickly solved for
“easy” problems (Andrew Ng)

Unsupervised learning
(man)

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

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



Biologist
Bioinformatician



Physician
Health informatician



AI/ML/DL

How does deep learning work for biomedicine?



Discovery



Diagnosis



Prognosis



Efficiency

Agenda

Topic 1: Introduction (20 mins)

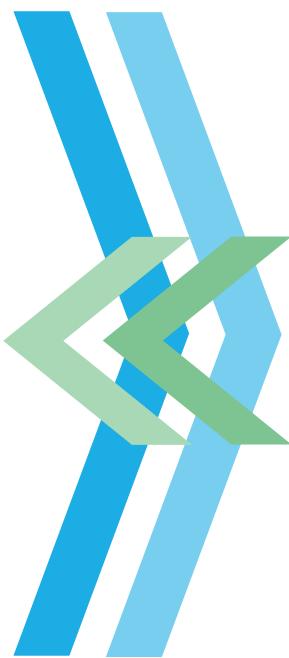
Topic 2: Brief review of deep learning (25 mins)

- Classic architectures
- Capsules
- Graphs
- Memory-augmented nets

Topic 3: Genomics (25 mins)

- Nanopore sequencing
- Genomics modelling

QA (10 mins)



Break (20 mins)

Topic 4: Biomedical imaging (15 mins)

- Cellular imaging
- Diagnostics imaging
- EEG/ECG

Topic 5: Healthcare (25 mins)

- Time-series of physio measures
- Trajectories prediction

Topic 6: Generative biomed (30 mins)

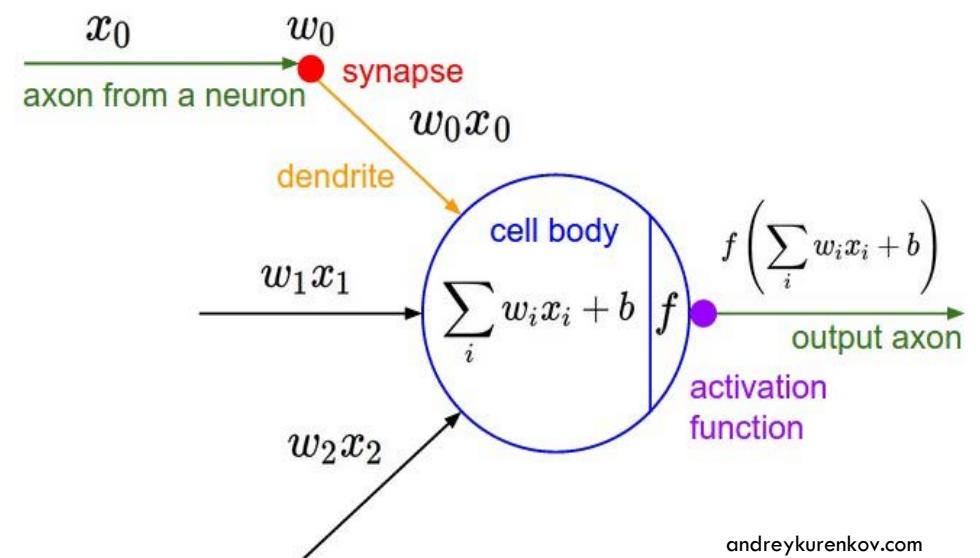
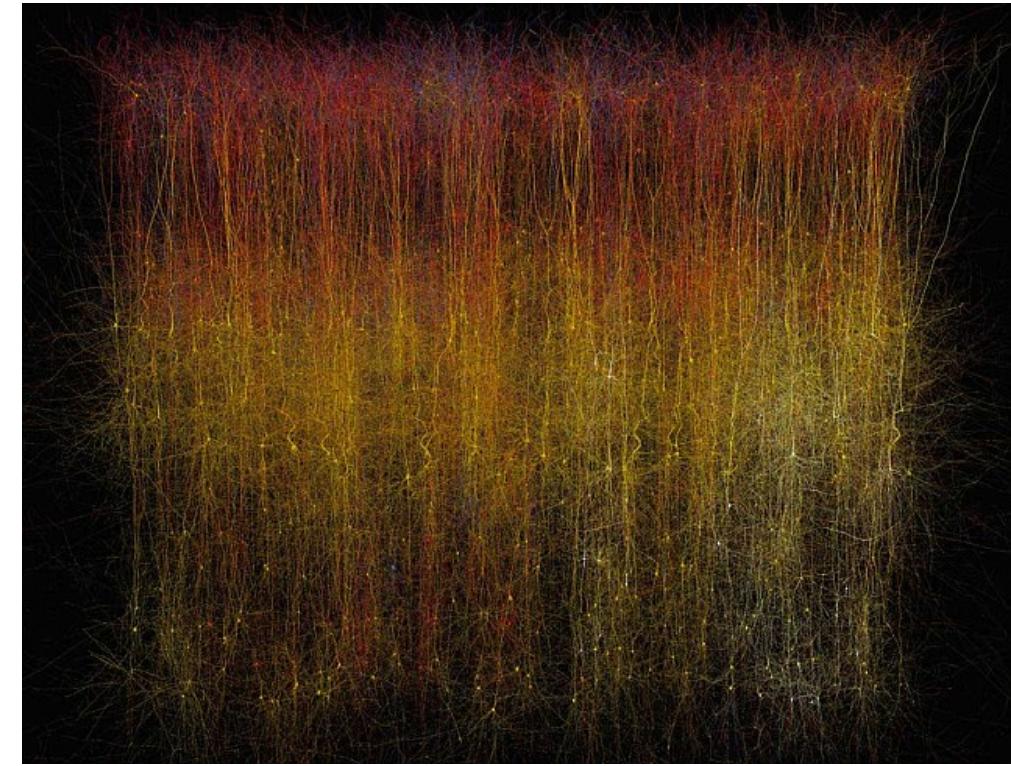
- Few-shot learning
- Generative models
- Drug design
- Future outlook

QA (10 mins)

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).
- Lots of new arts (batch-norm, Adam/RMSProp, skip-connections, Capsnet, external memory, GPU/TPU).
- Much bigger labelled datasets.
- Lots more people looking at lots of things



Feature learning

In typical machine learning projects, 80-90% effort is on feature engineering

- A right feature representation doesn't need much work. Simple linear methods often work well.

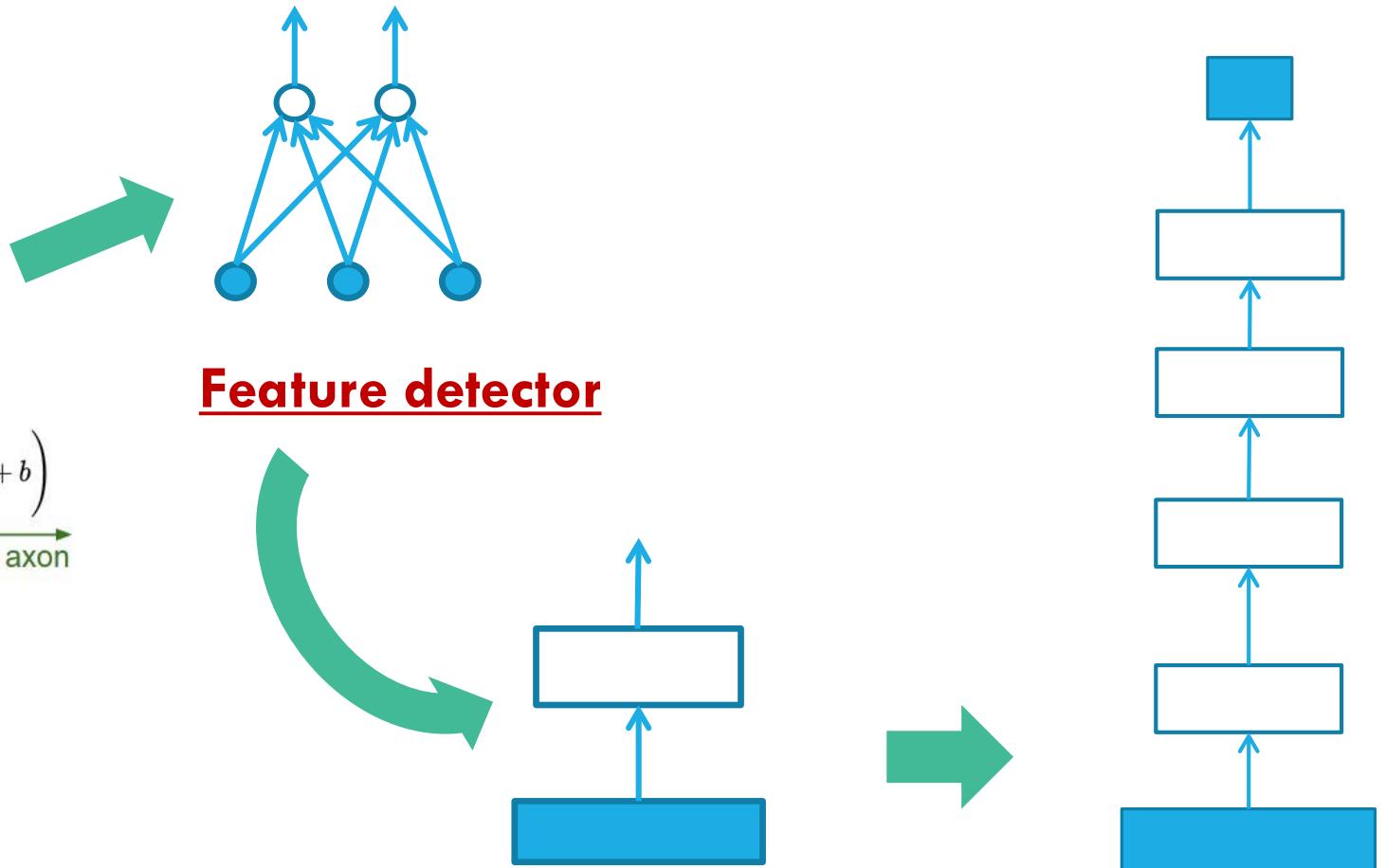
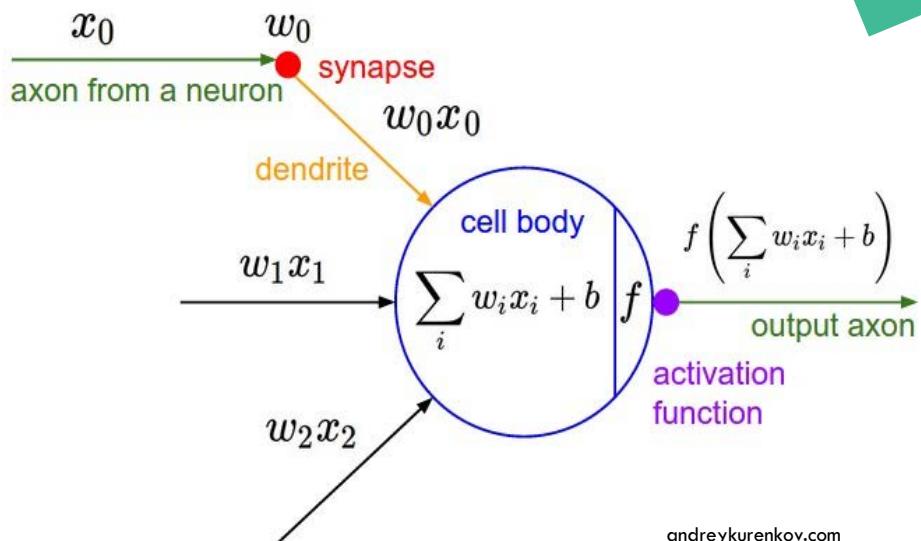
Text : BOW, n-gram, POS, topics, stemming, tf-idf, etc.

Software: token, LOC, API calls, #loops, developer reputation, team complexity, report readability, discussion length, etc.

Try yourself on Kaggle.com!

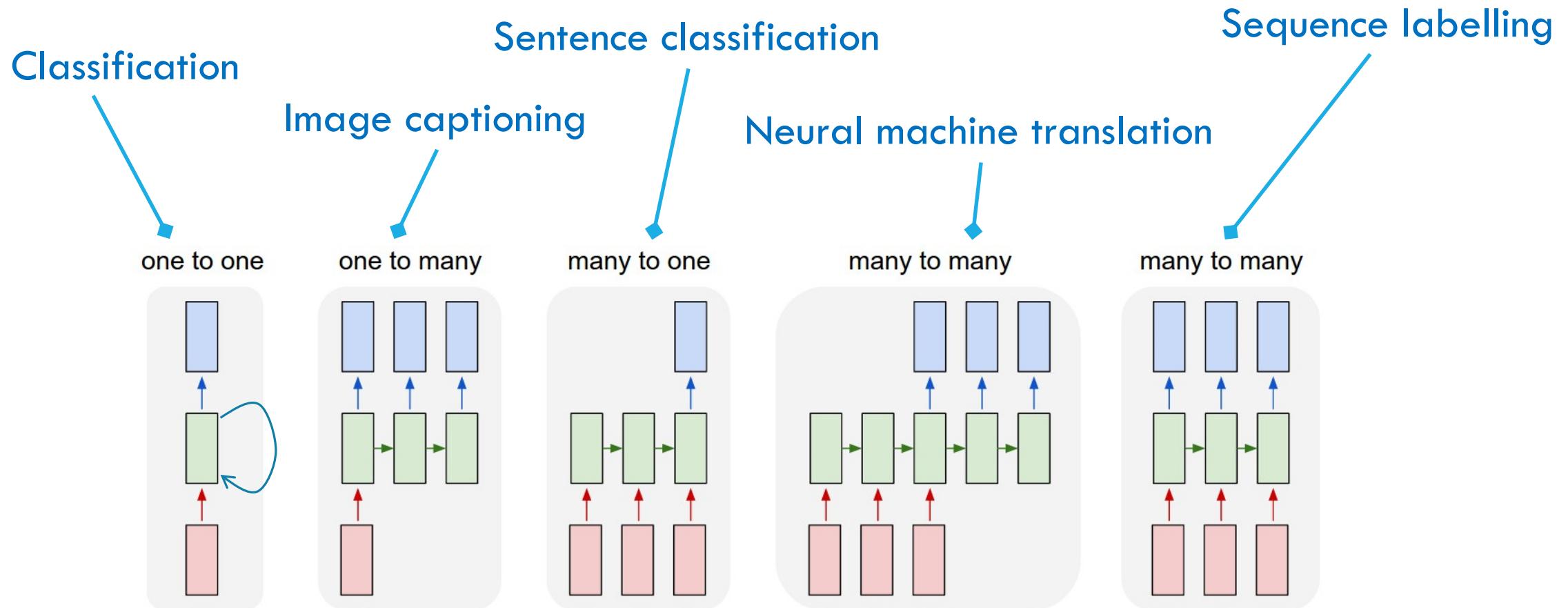
Deep learning as feature learning

Integrate-and-fire neuron



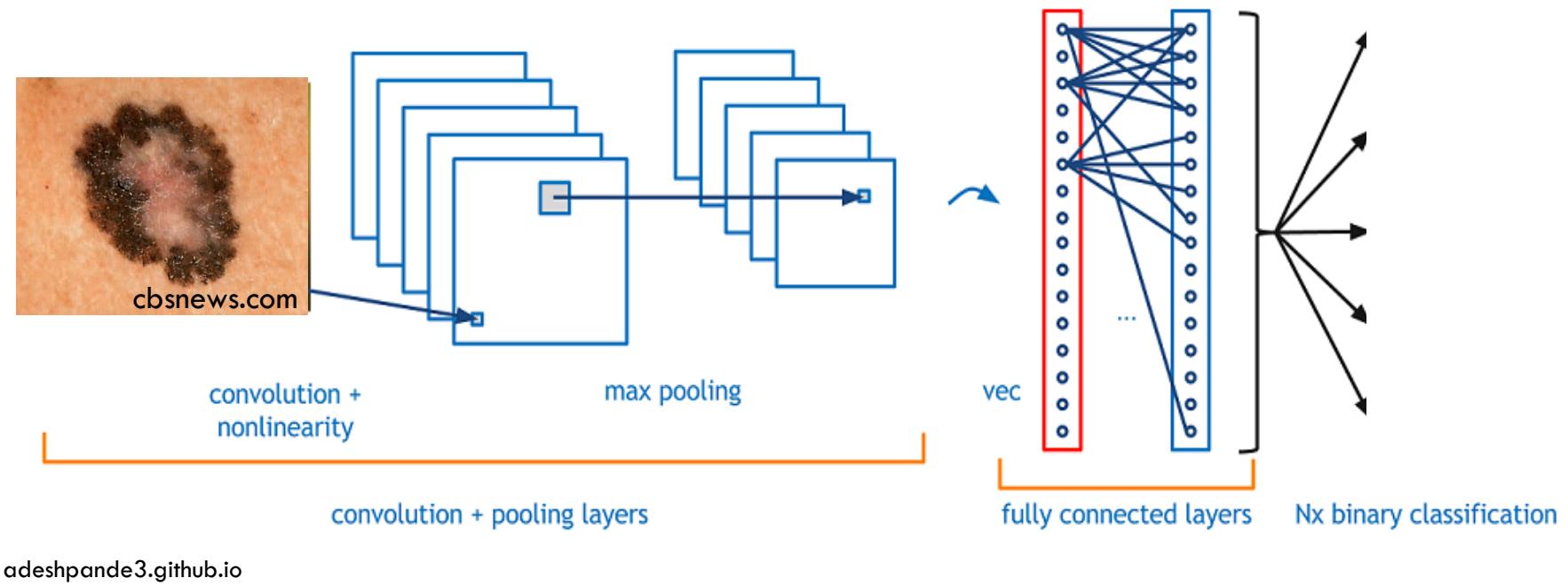
Block representation

Recurrent neural networks

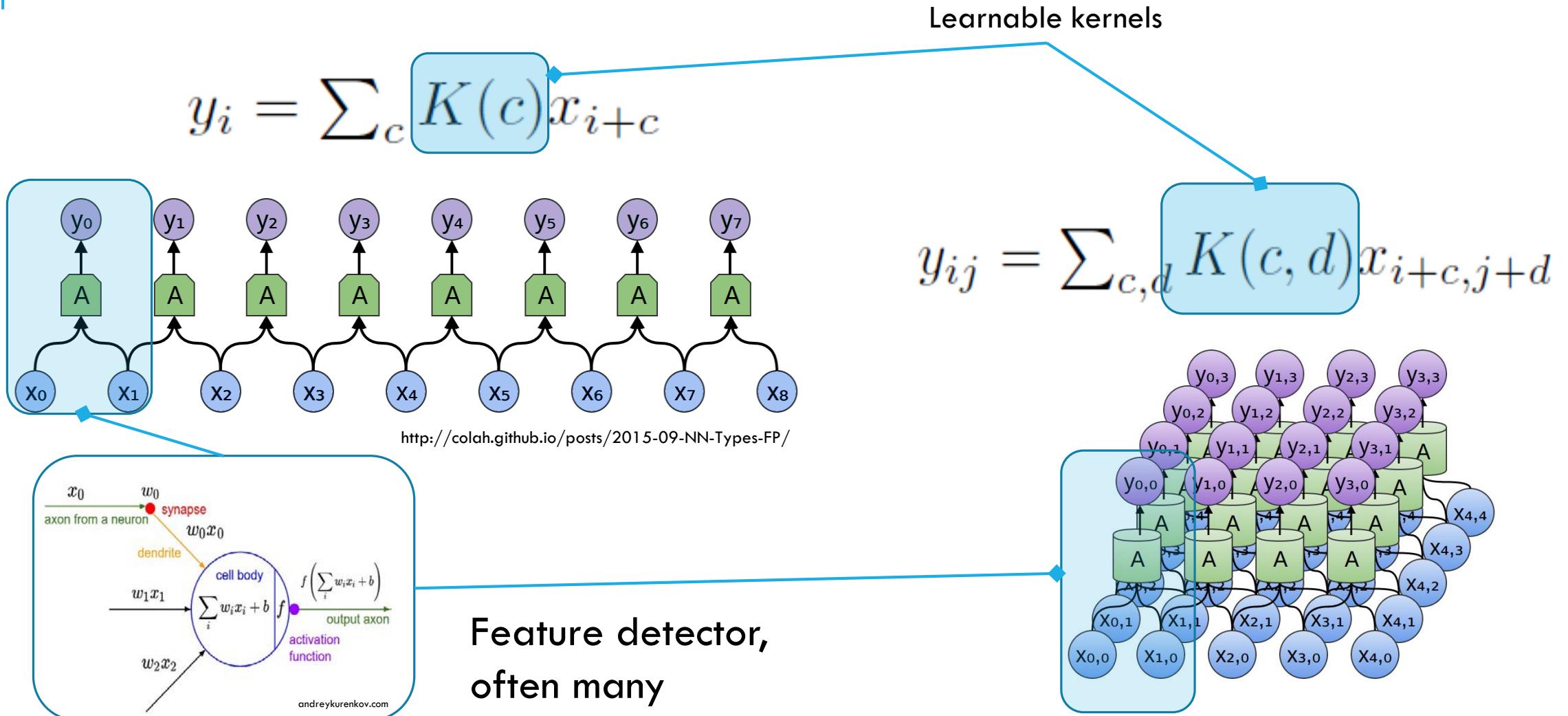


Source: <http://karpathy.github.io/assets/rnn/diags.jpeg>

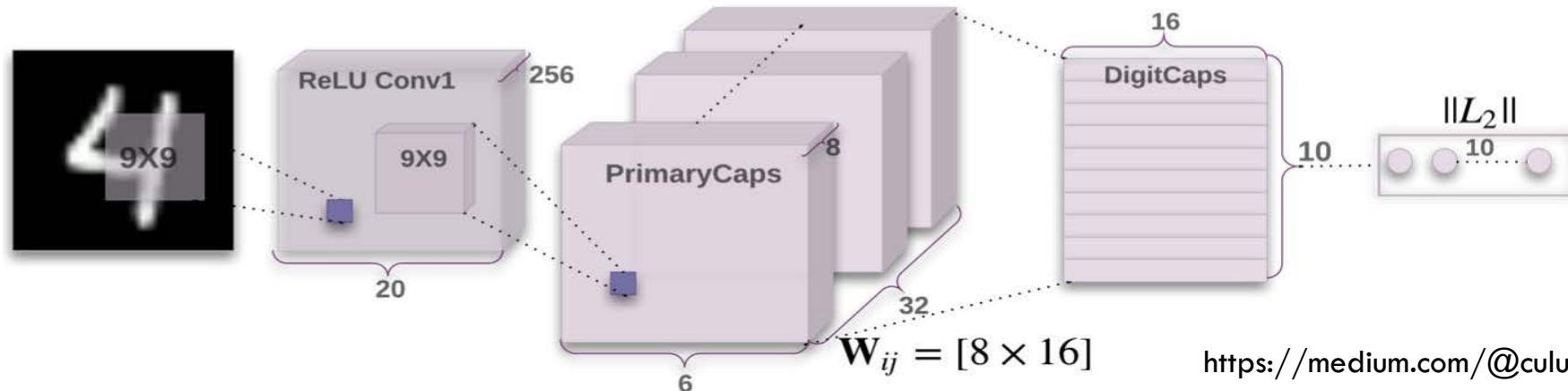
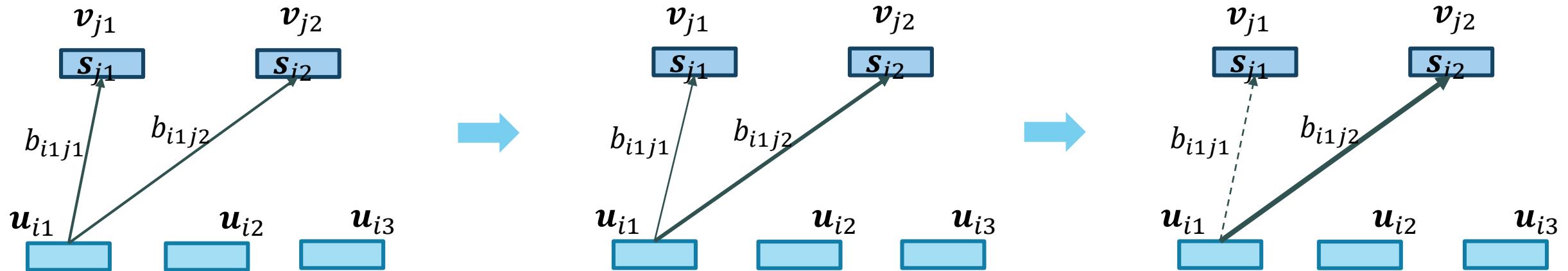
Convolutional nets



Learnable convolution



Capsules (Hinton's group)



Graphs

Goal: representing a graph as a vector

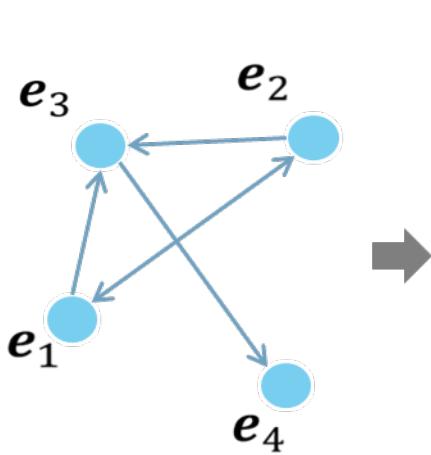
Many applications

- Drug molecules
- Object sub-graph in an image
- Dependency graph in software deliverable

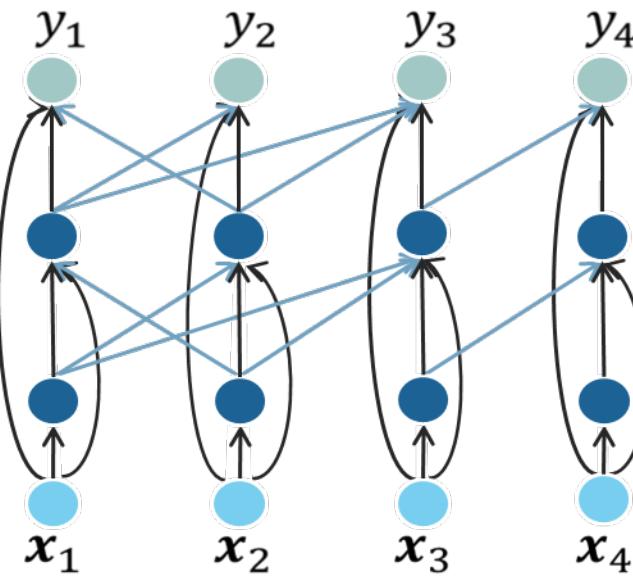
Recent works:

- Graph recurrent nets, column nets (Pham et al, 2017).
- Graph variational autoencoder (Kipf & Welling, 2016)
- Graph convolutional nets (LeCun, Welling and many others)

Column networks

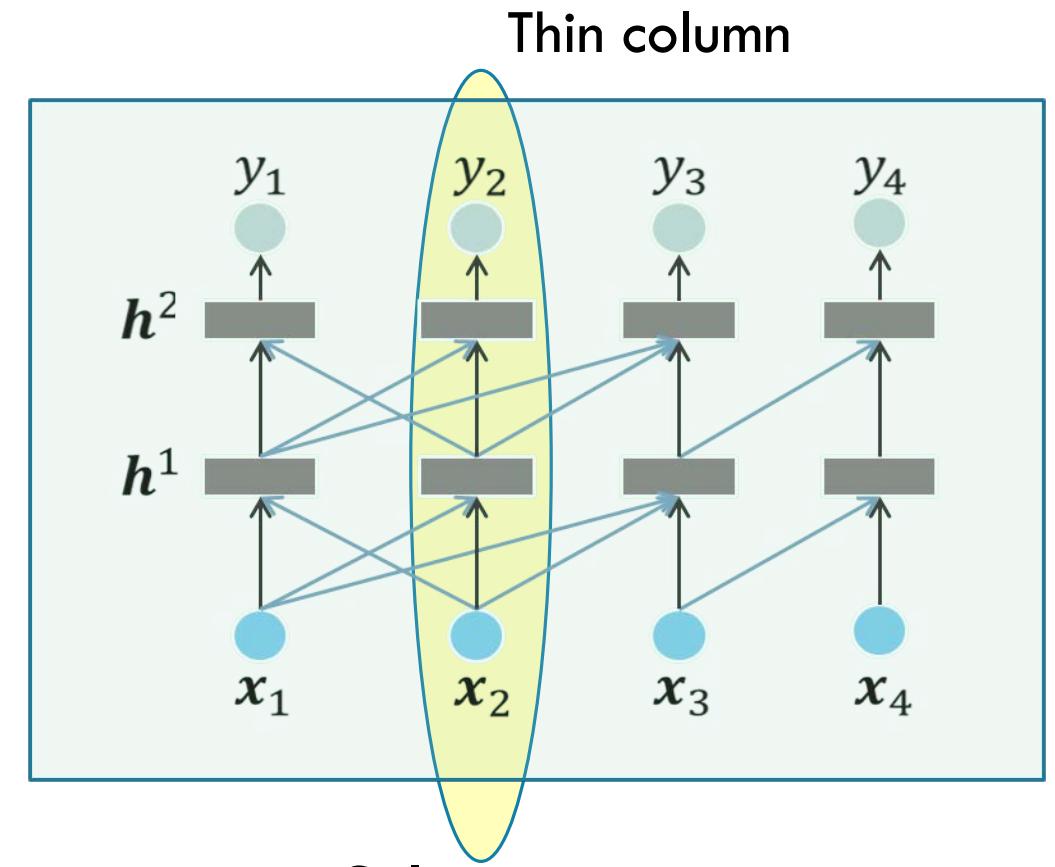


Relation graph



Stacked learning

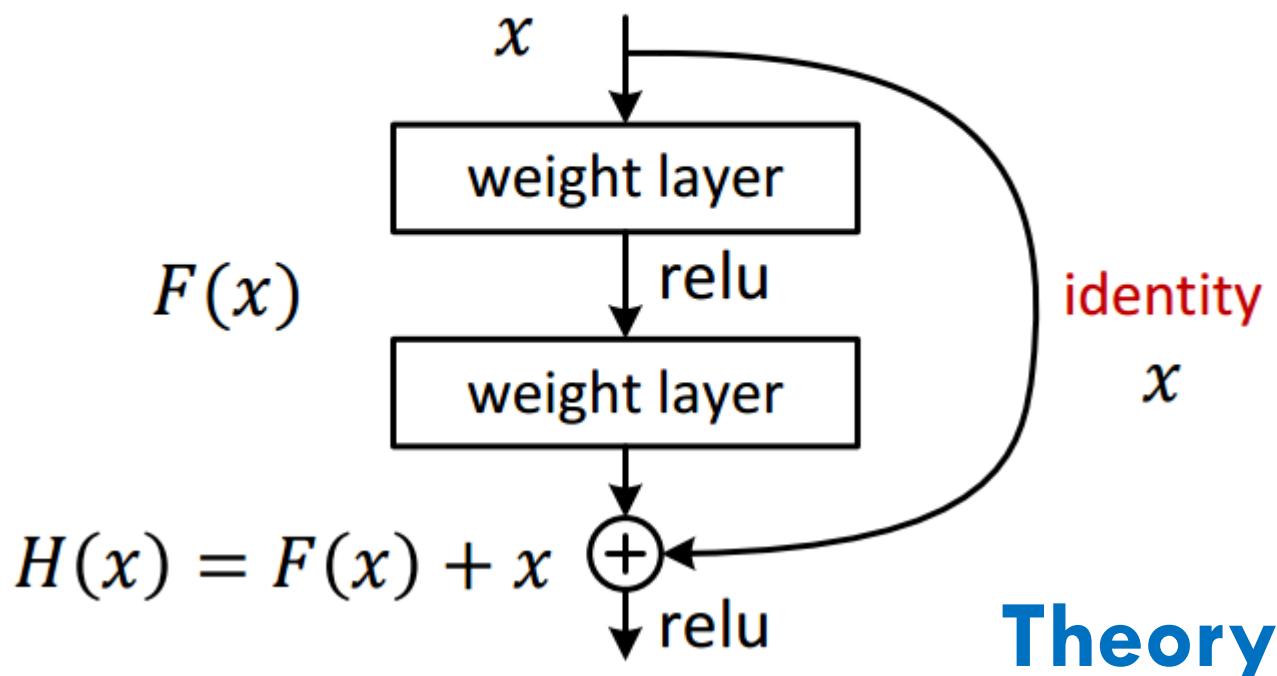
#REF: Pham, Trang, et al. "Column Networks for Collective Classification." AAAI. 2017.



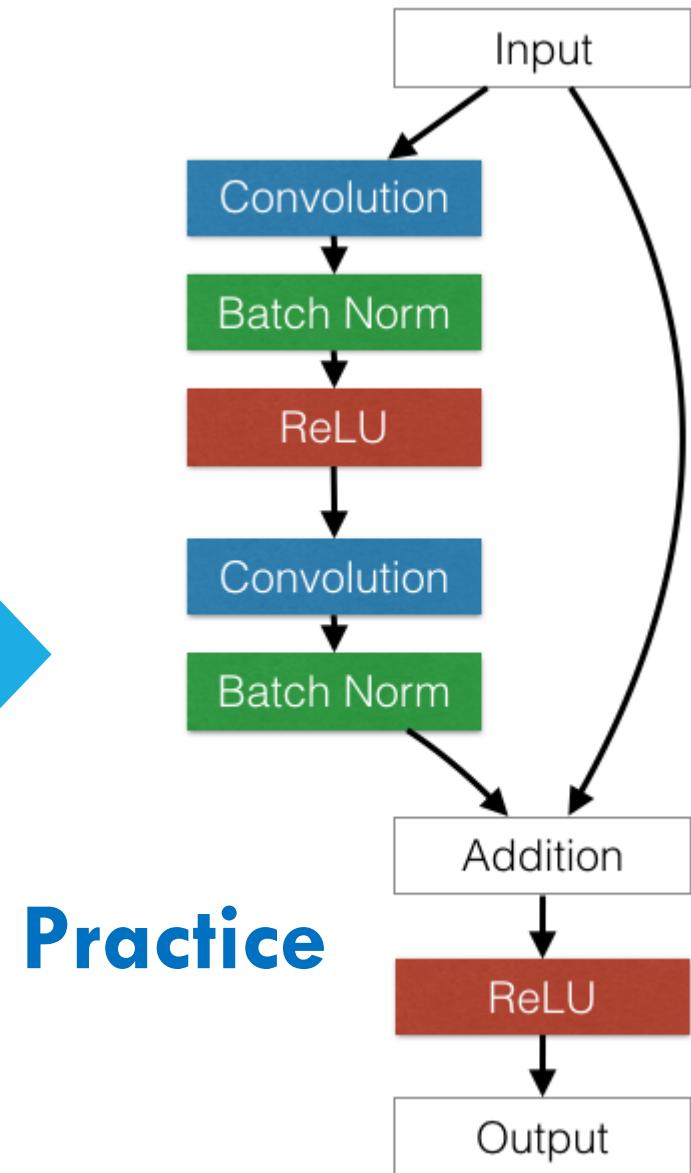
Column nets

Skip-connections

- Residual net



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



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

MANN: Memory-augmented neural networks

Long-term dependency

- E.g., outcome depends on the far past
- Memory is needed (e.g., as in LSTM)

Complex program requires multiple computational steps

- Each step can be selective (attentive) to certain memory cell

Operations: Encoding | Decoding | Retrieval

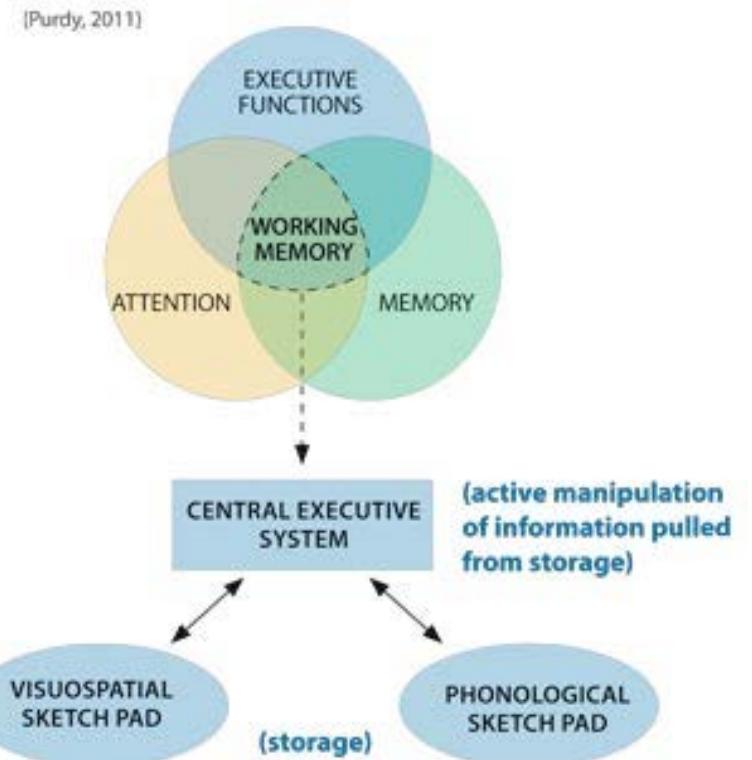
Memory types

Short-term/working (temporary storage)

Episodic (events happened at specific time)

Long-term/semantic (facts, objects, relations)

Procedural (sequence of actions)



Attention mechanisms

Need attention model to select or ignore certain inputs

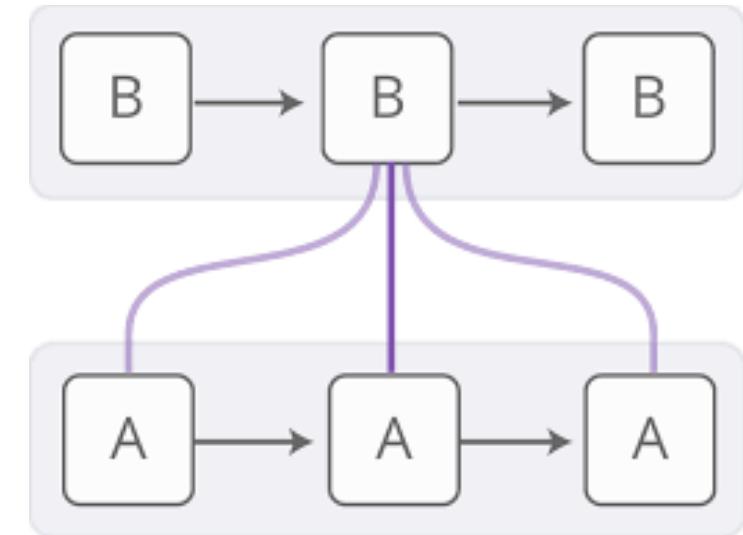
Human exercises great attention capability – the ability to filter out unimportant noises

- Foveating & saccadic eye movement

In life, events are not linear but interleaving.

Pooling (as in CNN) is also a kind of attention

Routing (as in CapsNet) is another example.



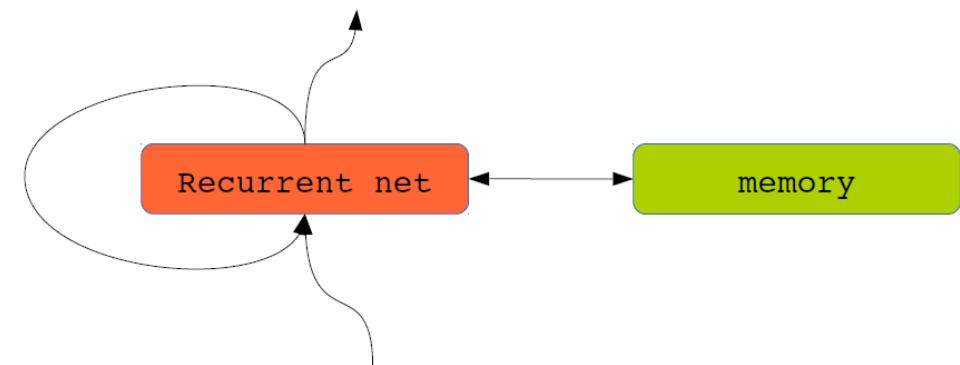
<http://distill.pub/2016/augmented-rnns/>

MANN: examples

Memory networks of Facebook: (Weston et al, Facebook, 2015); (Sukhbaatar et al, 2015) – associative memory

Dynamic memory networks of MetaMind: (Kumar et al, 2015) – episodic memory

Neural Turing machine and Differential Neural Computer of DeepMind (Graves et al. 2014, 2016) -- tape



(LeCun, 2015)

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 $D = \{(X_1, Y_1), (X_2, Y_2), \dots\}$

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

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

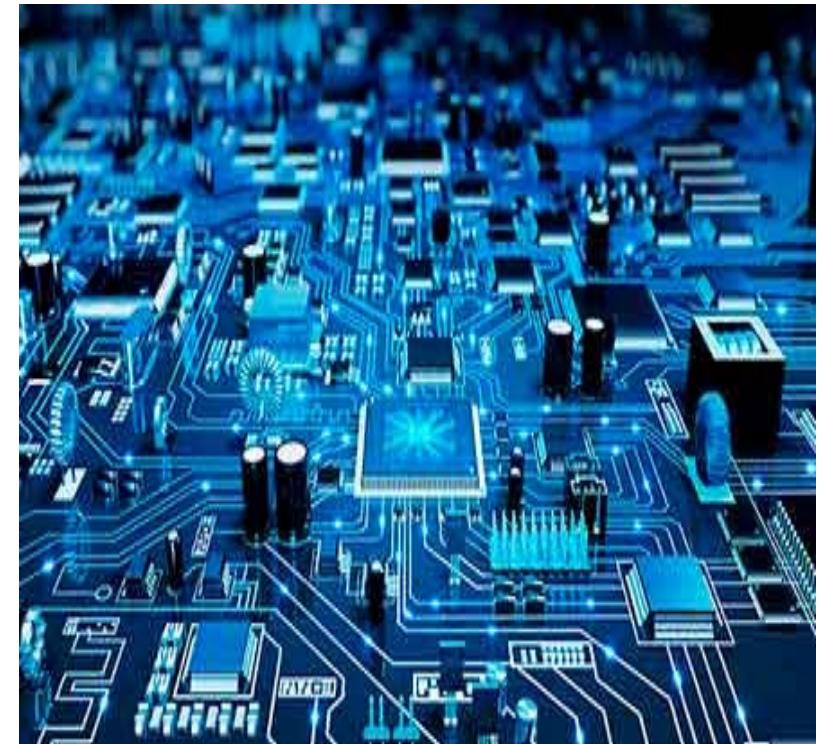
Deep learning as new electronics

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



ICLR18 submissions

~1000 submissions

Doubles the
number submitted
to ICLR17



Agenda

Topic 1: Introduction (20 mins)

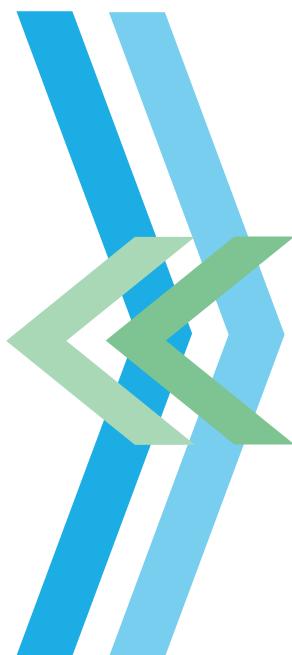
Topic 2: Brief review of deep learning (25 mins)

- Classic architectures
- Capsules
- Graphs
- Memory-augmented nets

Topic 3: Genomics (25 mins)

- Nanopore sequencing
- Genomics modelling

QA (10 mins)



Break (20 mins)

Topic 4: Biomedical imaging (15 mins)

- Cellular imaging
- Diagnostics imaging
- EEG/ECG

Topic 5: Healthcare (25 mins)

- Time-series of physio measures
- Trajectories prediction

Topic 6: Generative biomed (30 mins)

- Few-shot learning
- Generative models
- Drug design
- Future outlook

QA (10 mins)

Human genome

3 billion base-pairs (characters), 20K genes, 98% non-coding regions

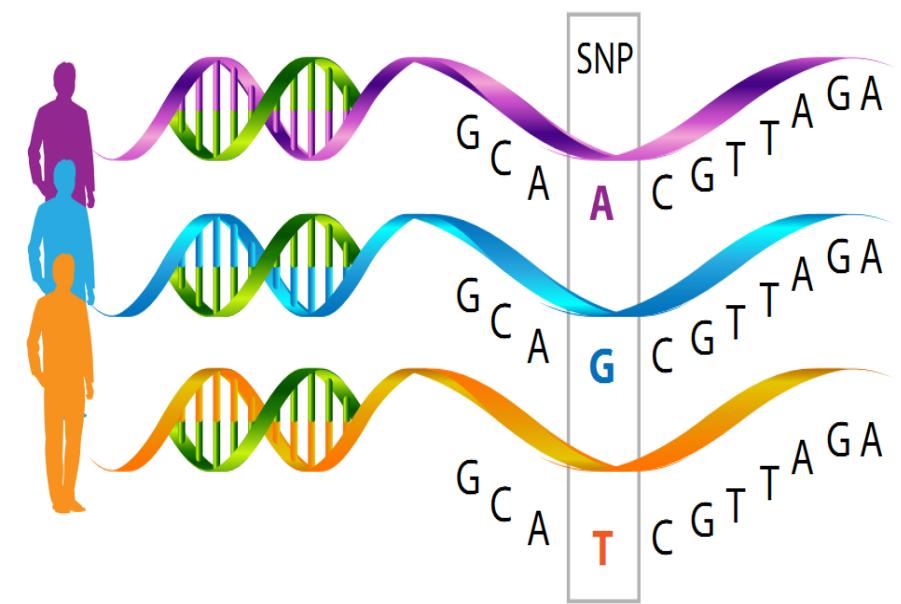
Any two random persons share 99.9% genome

The 0.1% difference is thought to account for all variations between us

- Appearance: Height (80% heritable), BMI, hair, skin colors
- IQ, education levels
- Genetic disorders such as cancers, bipolar, schizophrenia, autism, diabetes, etc.

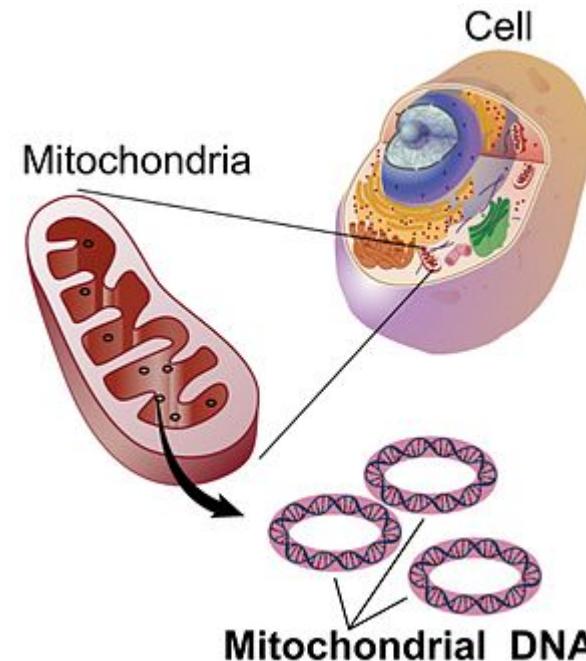
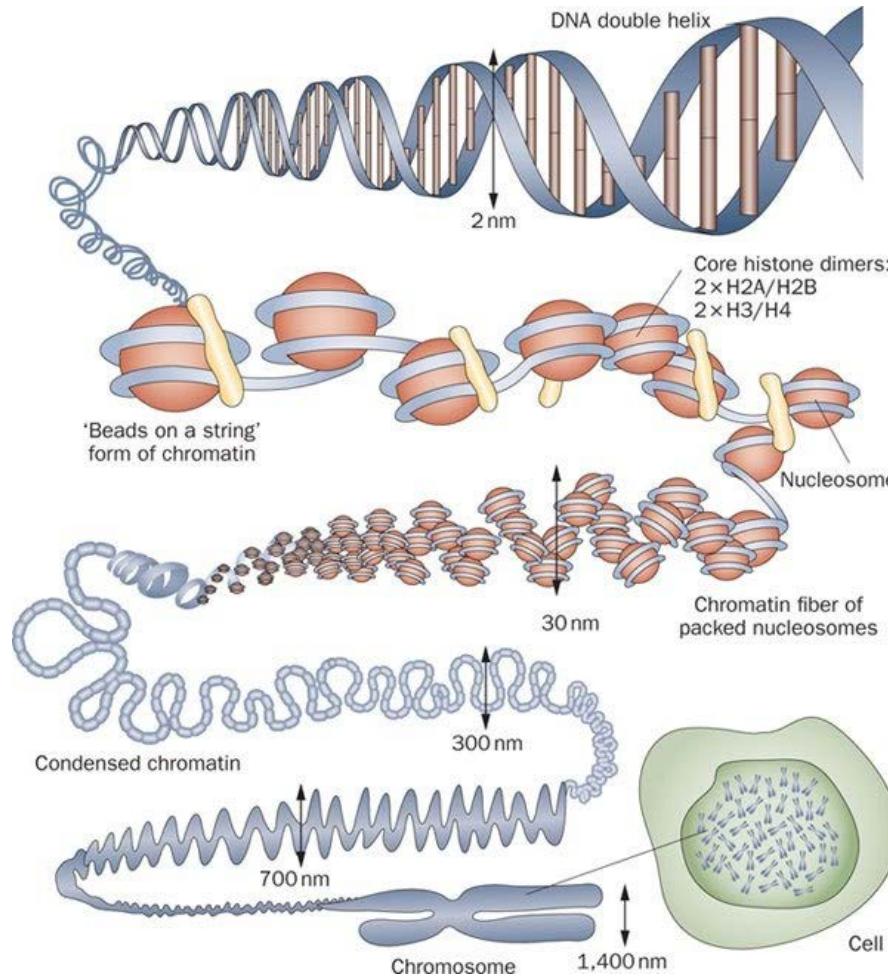
Any two random persons share about 60% variations (SNV/SNP)

As we age, there are small mutations within our cells

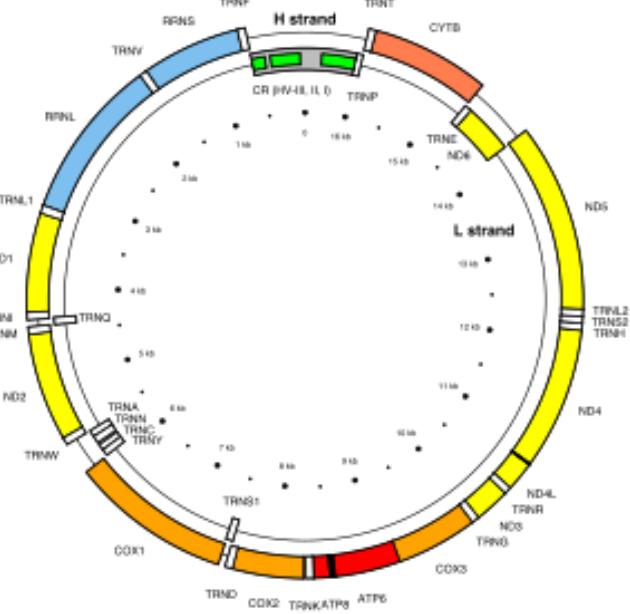


<https://neuroendoimmune.files.wordpress.com>

The cell, nuclear DNA & MtDNA



MtDNA ring



Sequencing

The first step is to read (sequence) the DNA/MtDNA, and represent the information as string of characters (A,C,G,T) in computer.

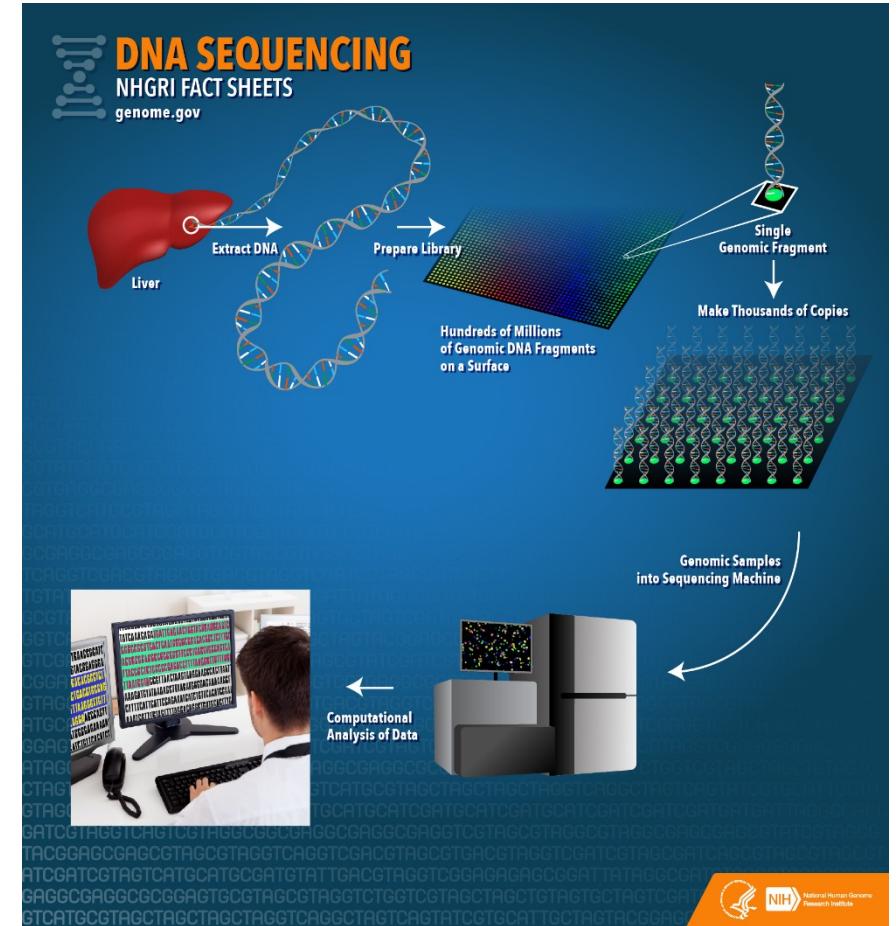
The most popular technique these days read short sequences (hundreds of characters), and align.

Each position is read typically at least 30 times to get enough confidence → Huge storage!!!

String alignment is then the key to final sequence → Need super-computer to do this fast.

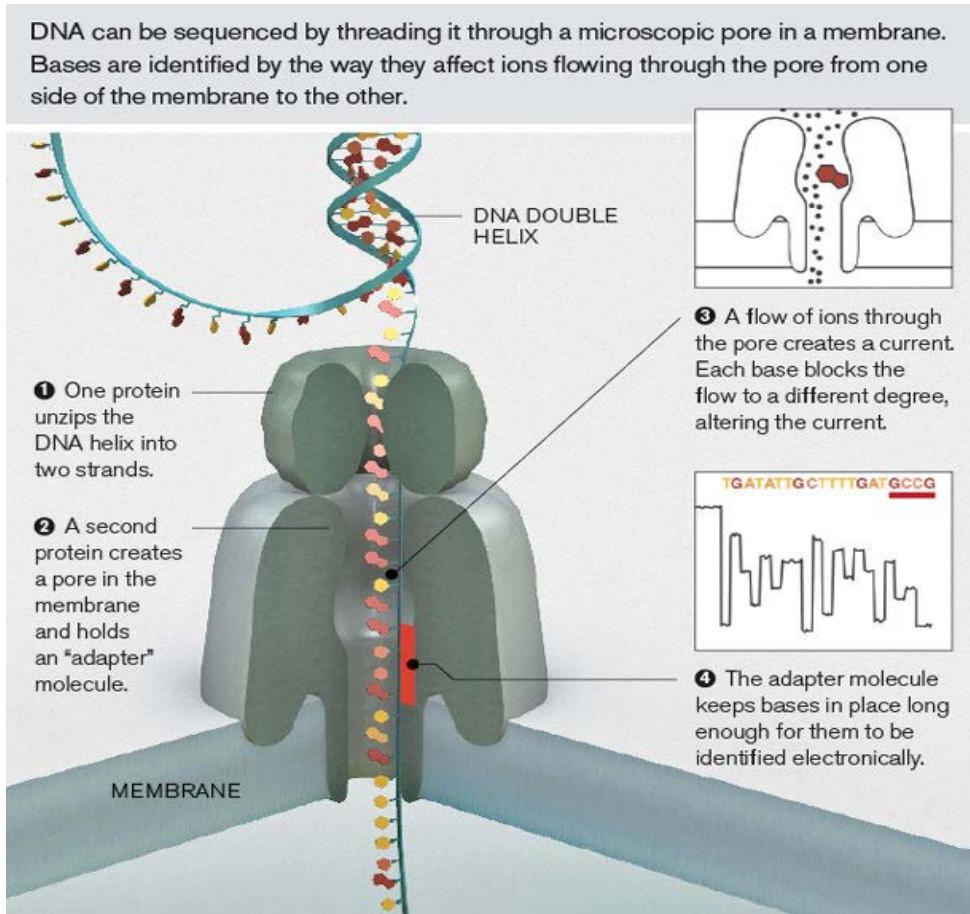
A DNA sequence is compared against the reference genome. Only the difference (0.1%) need to be stored.

- This does not usually apply for MtDNA, as each cell has as many as 500 MtDNAs, they are slightly different! More different as we age.



Source: <https://www.genome.gov>

The latest: nanopore sequencing (electrical signals → A|C|G|T)



Continuous segmentation & labelling

Deep architectures for nanopore sequencing

Aimed at real time recognition

The setting is similar to speech recognition!

- → The early days used HMMs. Now LSTMs.

We will briefly review the latest:

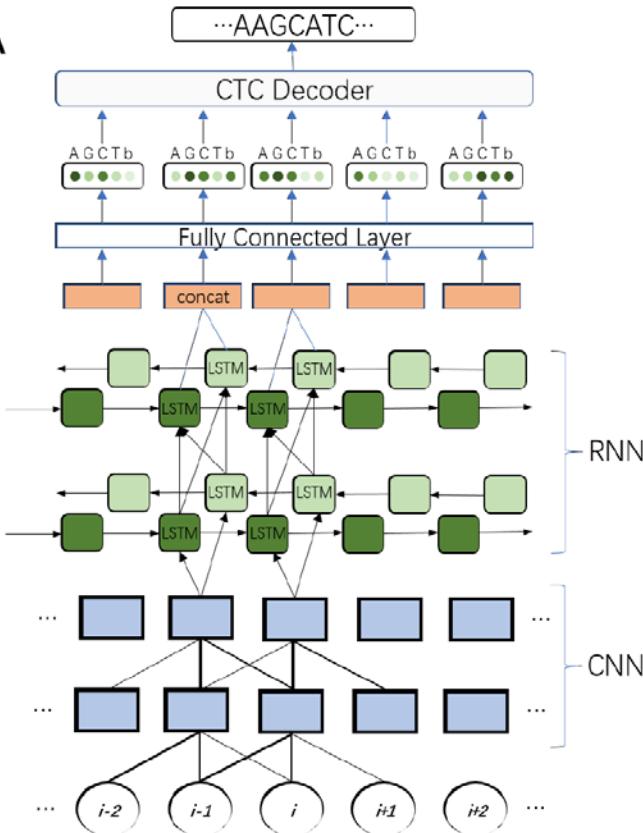
- **Chiron** (Teng et al., August 2017, UQ, Australia)

Other GRU/LSTM variants

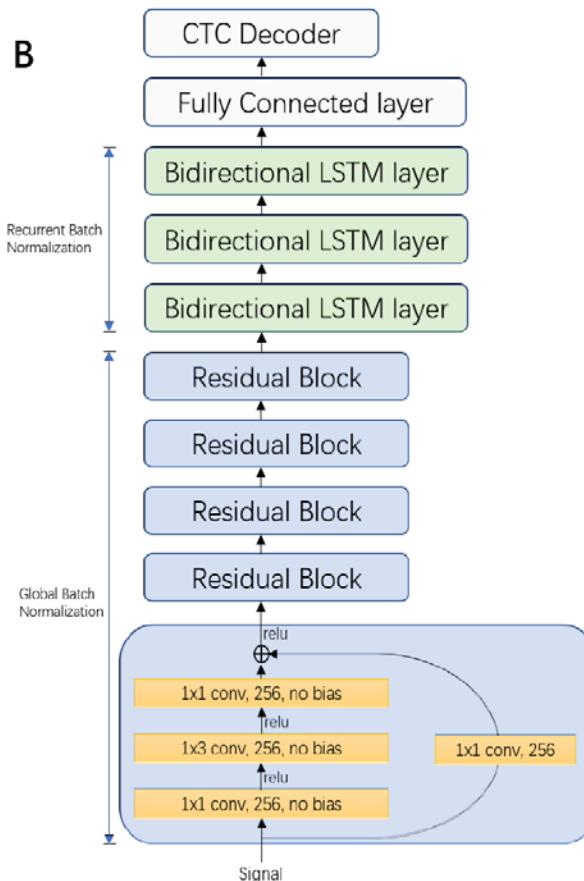
- Nanonet (Oxford Nanopore Technologies, 2016)
- BasecRAWller (Stoiber & Brown, May 2017)
- **DeepNano** (Boza et al., June 2017, Comenius University in Bratislava, Slovakia)

Chiron

A



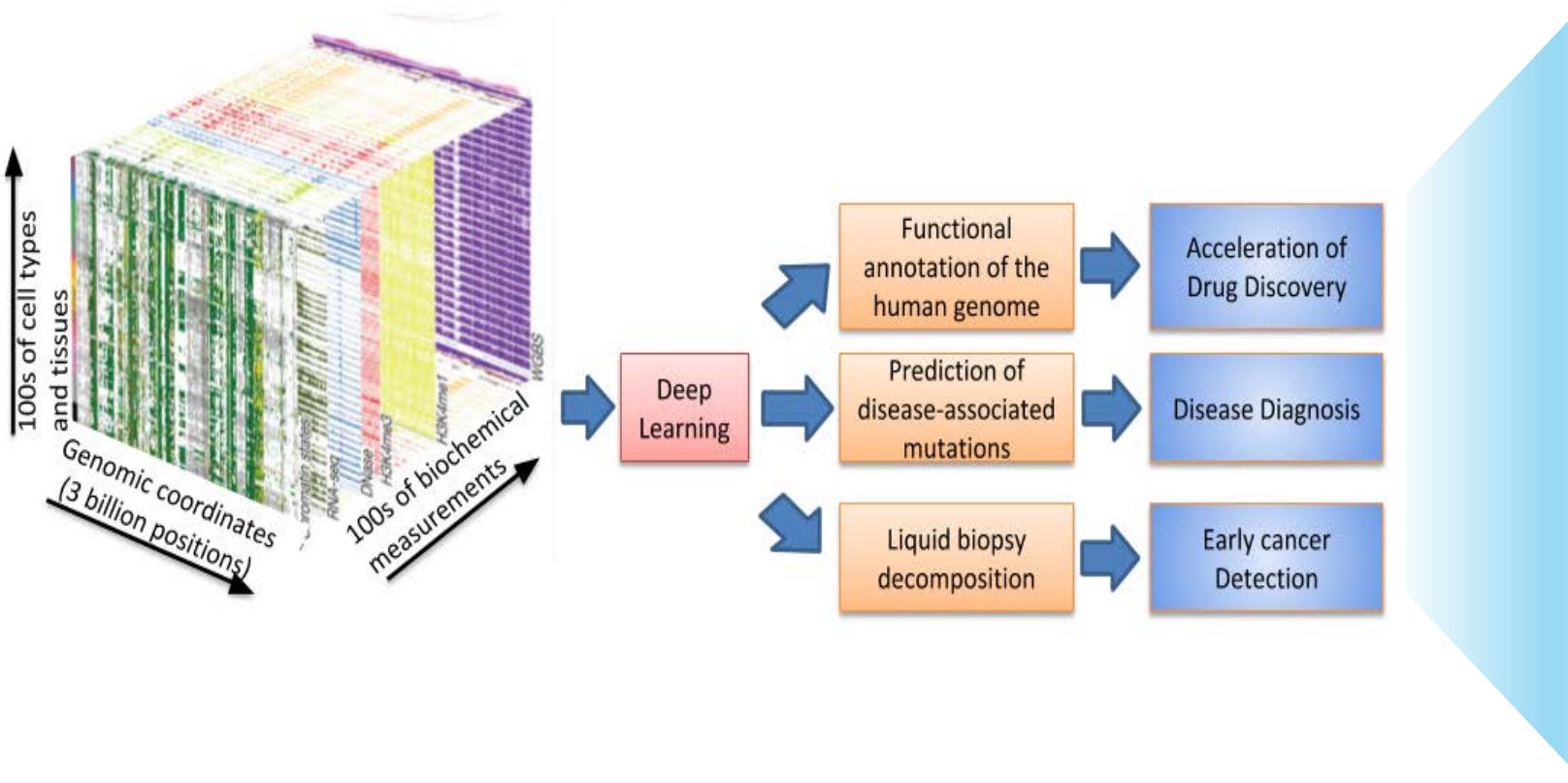
B



| Dataset | Basecaller | Identity Rate |
|------------------------|--------------|------------------------|
| Lambda | Metricchor | 0.8650 (-0.0246) |
| | Albacore | 0.8896 |
| | BasecRAWller | 0.8154 (-0.0742) |
| | Chiron | 0.8776 (-0.012) |
| <i>E. coli</i> | Metricchor | 0.8864 (-0.0193) |
| | Albacore | 0.901 (-0.0047) |
| | BasecRAWller | 0.8254 (-0.0803) |
| | Chiron | 0.9057 |
| <i>M. tuberculosis</i> | Metricchor | 0.8802 (-0.0117) |
| | Albacore | 0.8919 |
| | BasecRAWller | 0.8241 (-0.0678) |
| | Chiron | 0.8851 (-0.0068) |
| Human | Metricchor | 0.794 (-0.0446) |
| | Albacore | 0.8386 |
| | BasecRAWller | 0.8149 (-0.0237) |
| | Chiron | 0.8154 (-0.0232) |

#REF: Teng, Haotien, et al. "Chiron: Translating nanopore raw signal directly into nucleotide sequence using deep learning." *bioRxiv* (2017): 179531.

Opportunities for Deep Learning in Genomics



Genetic diagnostics
Refining drug targets
Pharmaceutical development
Personalized medicine
Better health insurance
Synthetic biology

Some AI problems

DNA is a book, easy to read (costs less than \$1K to sequence), extreme difficult to comprehend.

- It has 3B characters (A,C,T,G), 46 volumes (chromosomes), 20K chapters.
- The longest book has less than 10M characters, 13 volumes ("A la recherche du temps perdu" (In Search of Lost Time), by Marcel Proust, 2012) – as recognized by Guinness World Records.

Short sequences (100 chars) are predictive of protein binding, also gene start/end.

Proteins are big 3D graphs interacting with the 1D-2D strings (DNA, RNA), and other proteins & drugs (which are graphs themselves).

Long chains of influence, from SNP to cell, tissue and organ functions.

Viruses can be generated/edited on computer, hence discrete sequence generation problem.

Filling the genotypes → phenotypes gap

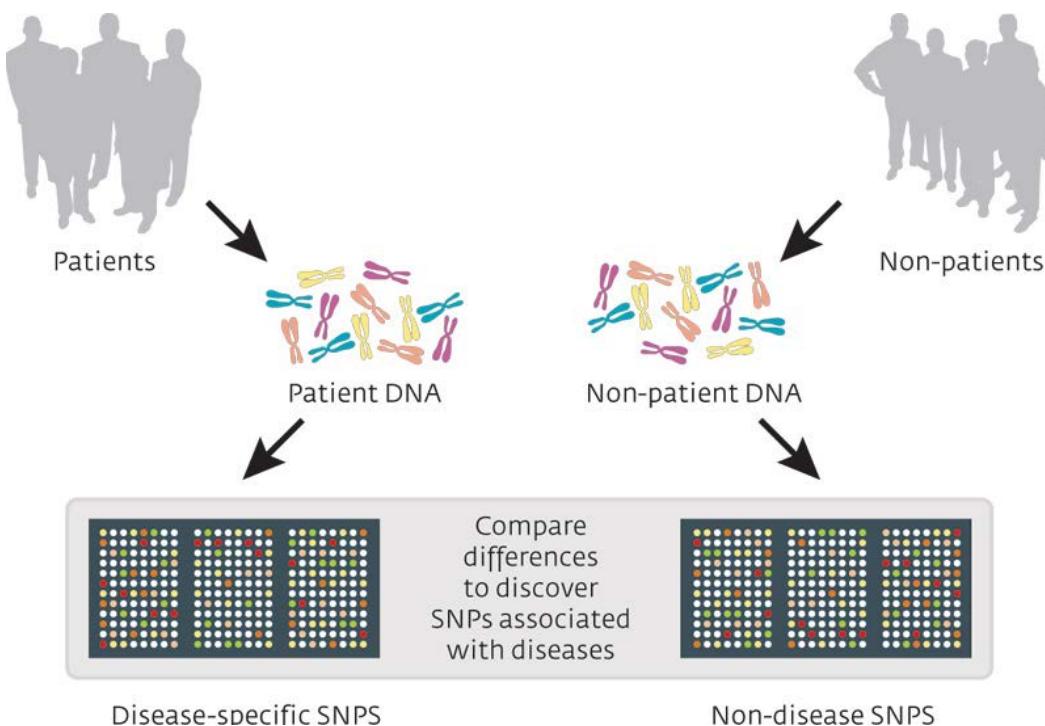
Ultimate goals:

- Estimating explained variance in inheritability
- Discover risk factors
- Predicting individual phenotypes: Height, Glucose, BMI, IQ, Edu, Mental, Cancers...

Some paths under investigation

- Predicting the bio of the cells, DNA + MtDNA, and more
- Statistical modeling of genetic architectures, e.g., Bayesian, mixed linear models, Gaussian Processes.
- Motif modeling with DNA/RNA/protein, e.g., predict binding sites
- Developing data-efficient techniques for genomics
- Integrating multimodalities

GWAS: Genome-Wide Association Study



Setting:

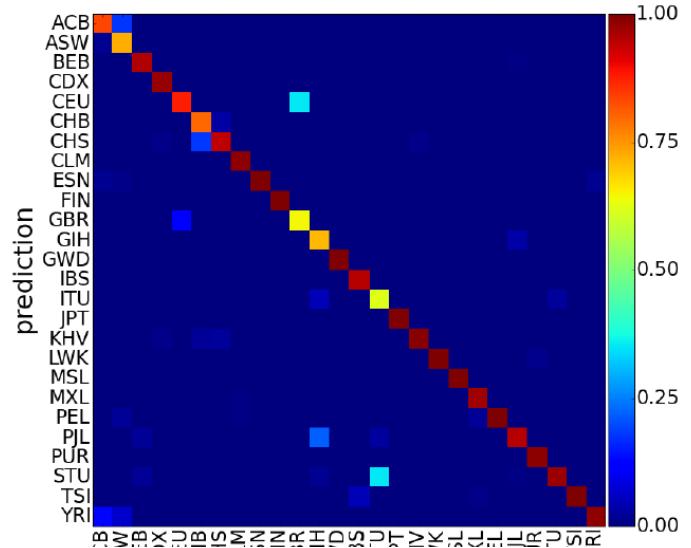
- For each DNA, only differences from a reference genome are recorded.
- The differences are SNPs, one per dimension.

Problems

- Very high dimensional (typically **hundreds of thousands**), low sample size (typically **hundreds**)
- Missing/unreliable data
- Typically very weak association
- Combating the False Discovery Rate (FDR) due to multiple parallel hypotheses: Individual p -value must be extremely small, e.g. **5×10^{-8}**

Diet networks for GWAS

#REF: Romero, Adriana, et al. "Diet Networks: Thin Parameters for Fat Genomic." *arXiv preprint arXiv:1611.09340* (2016).



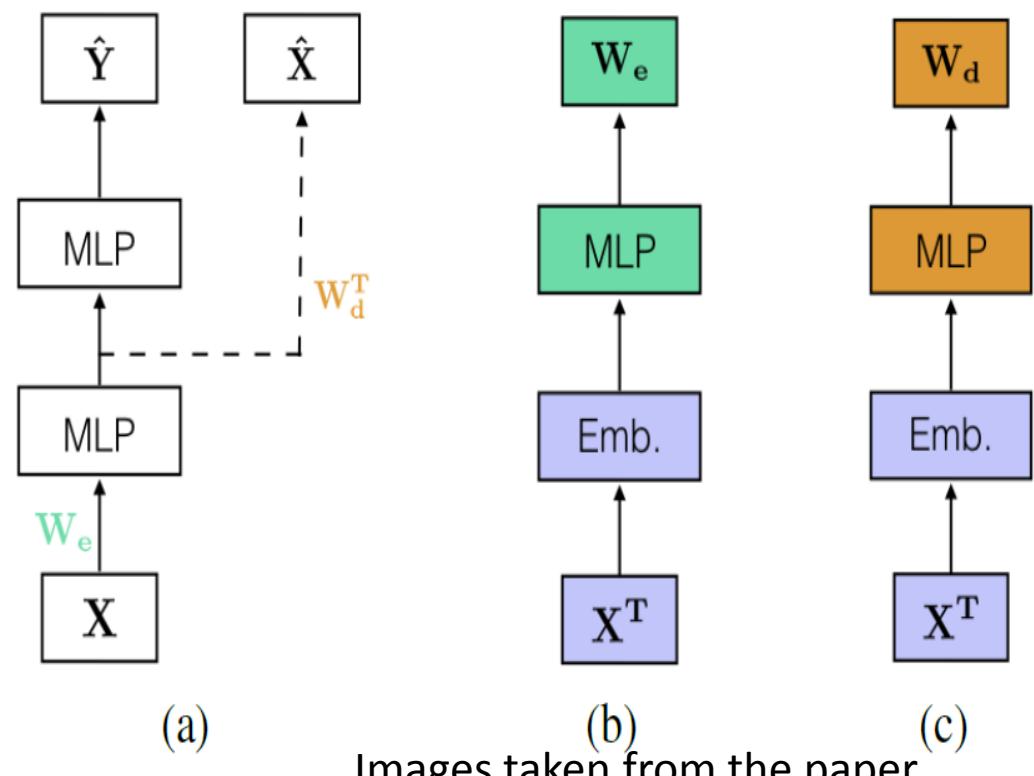
Use a “hypernet” to generate the main net.

Features are embedded (not data instance).

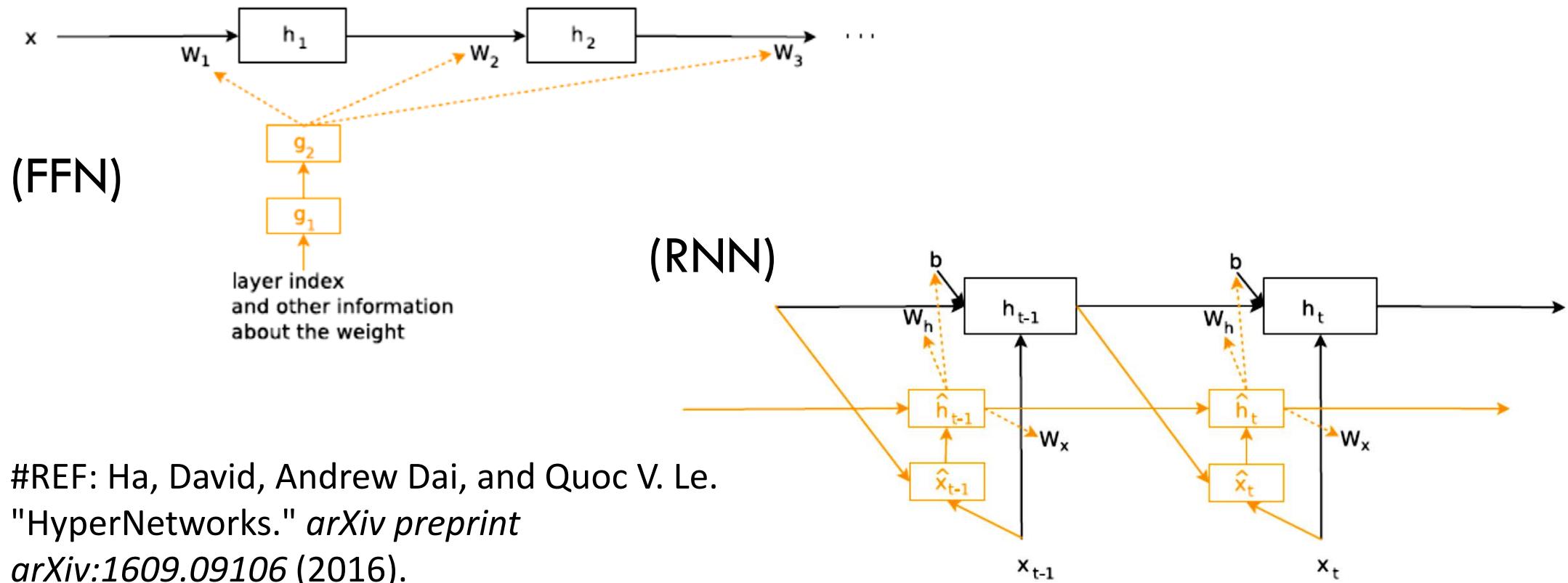
Unsupervised autoencoder as regularizer.

Works well on country prediction on the 1000 Genomes Project dataset.

- But this is a relatively easy problem. PCA, even random subspace can do quite well!



HyperNetworks: Network to generate networks



#REF: Ha, David, Andrew Dai, and Quoc V. Le.
"HyperNetworks." *arXiv preprint arXiv:1609.09106* (2016).

GWAS: Challenges

We are detecting rare events!!!

Results hard to replicate across studies.

- Model stability?

SNP → phenotypes seem impossible.

If it is (e.g., race prediction), little insights can be drawn upon.

The pathway is deep and complex

- Room for deep learning?

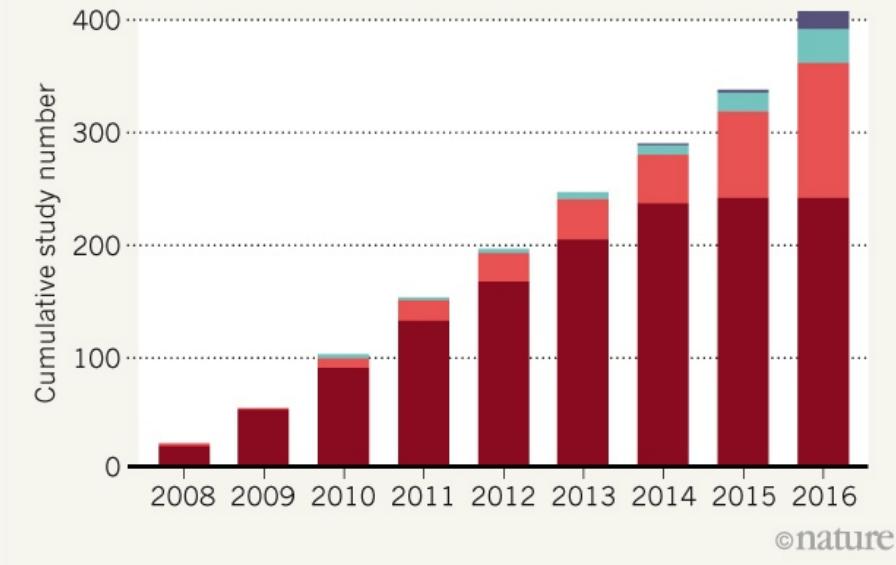
Room for structured models

- SNP annotations
- Spatial relationships
- Evolutionary trees

THE GENOME-WIDE TIDE

Large genome-wide association studies that involve more than 10,000 people are growing in number every year — and their sample sizes are increasing.

Sample sizes: ■ More than 200,000 ■ 100,000–199,999
■ 50,000–99,999 ■ 10,000–49,999

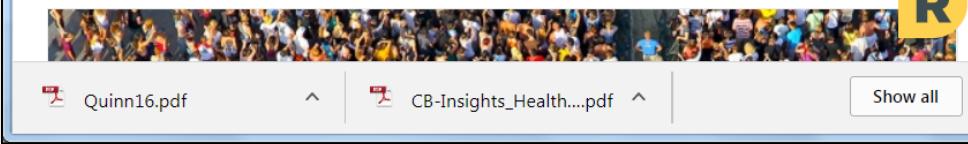


New concerns raised over value of genome-wide disease studies

Large analyses dredge up 'peripheral' genetic associations that offer little biological insight, researchers say.

Ewen Callaway

15 June 2017



Rooms for deep learning

Bridge the genotype-phenotype gap

- Incorporating HUGE amount of data
- Modelling the multiple layers of complex biological processes in between.
- Starting from the DNA and its immediate functions, e.g., **protein binding, gene start, alternative splicing, SNP annotations**.

Deep learning has shown to work well in cognitive domains, where human can perform in less than a second.

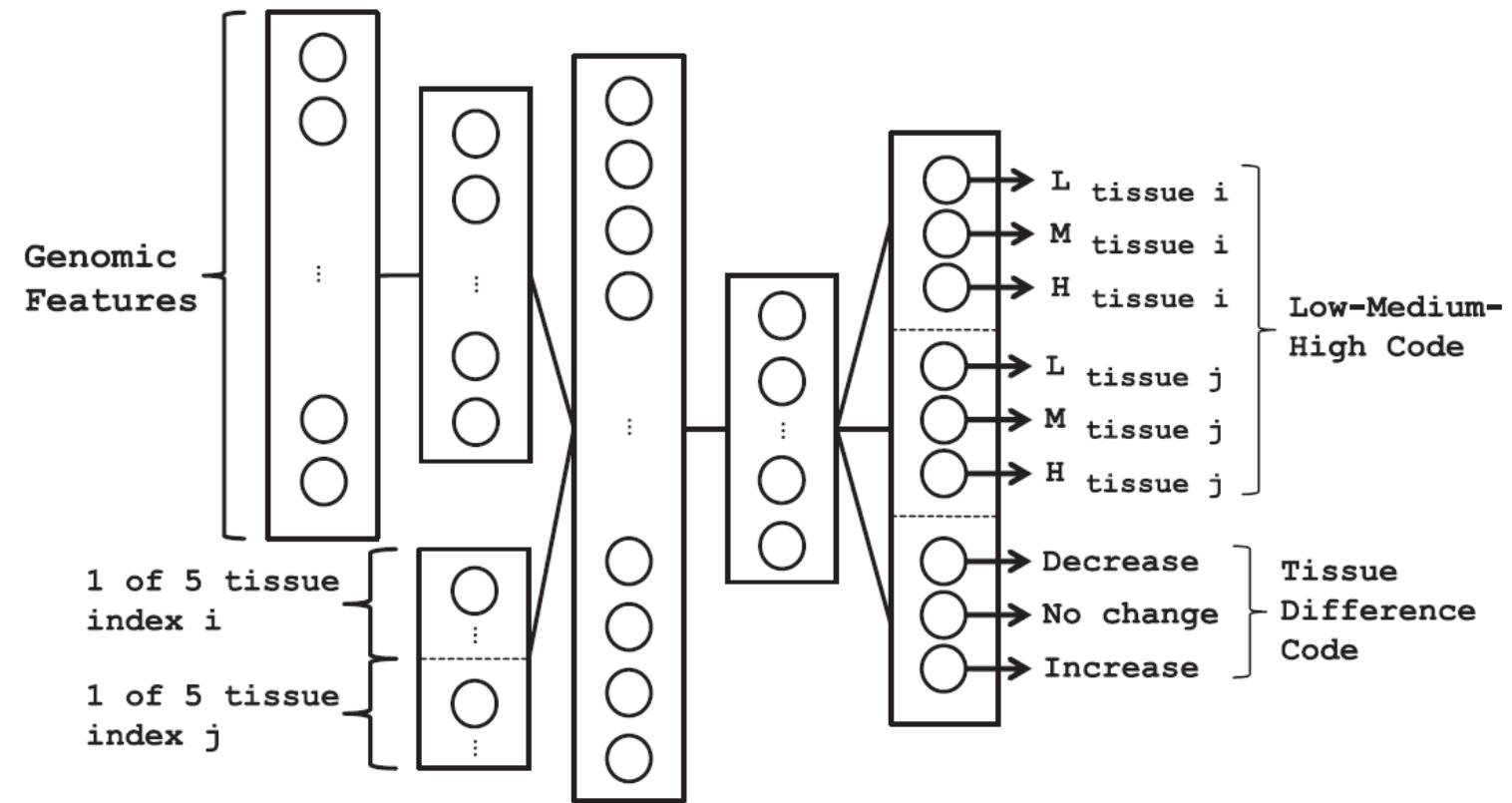
- We need to be super-human to bridge the gap.

New models for 2% of coding part, as well as 98% non-coding (probably having regulatory functions)

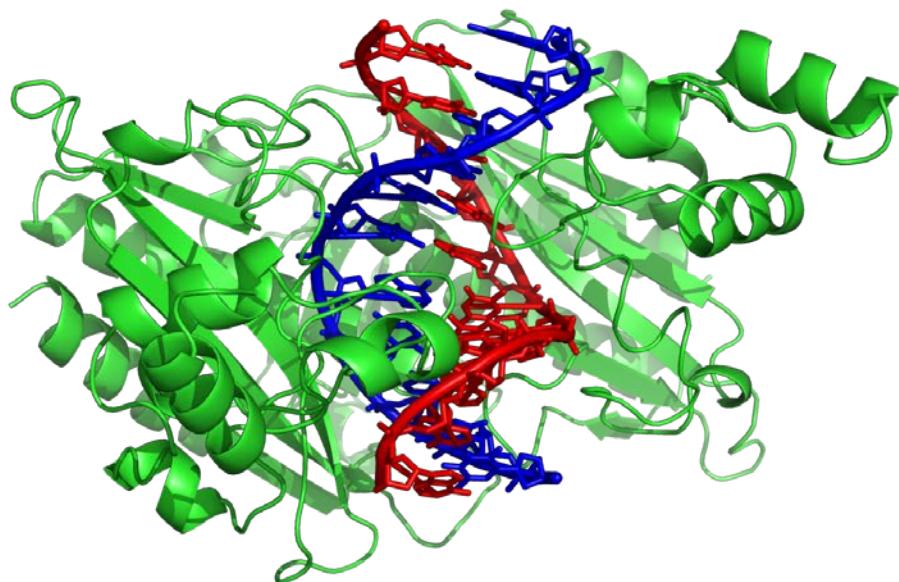
Incorporating biological understanding into model, not the black-box.

Use of feedforward nets: Tissue-regulated splicing code

#REF: Leung, Michael KK, et al.
"Deep learning of the tissue-regulated splicing code." *Bioinformatics* 30.12 (2014): i121-i129.



Use of CNNs: Discovery of DNA motifs

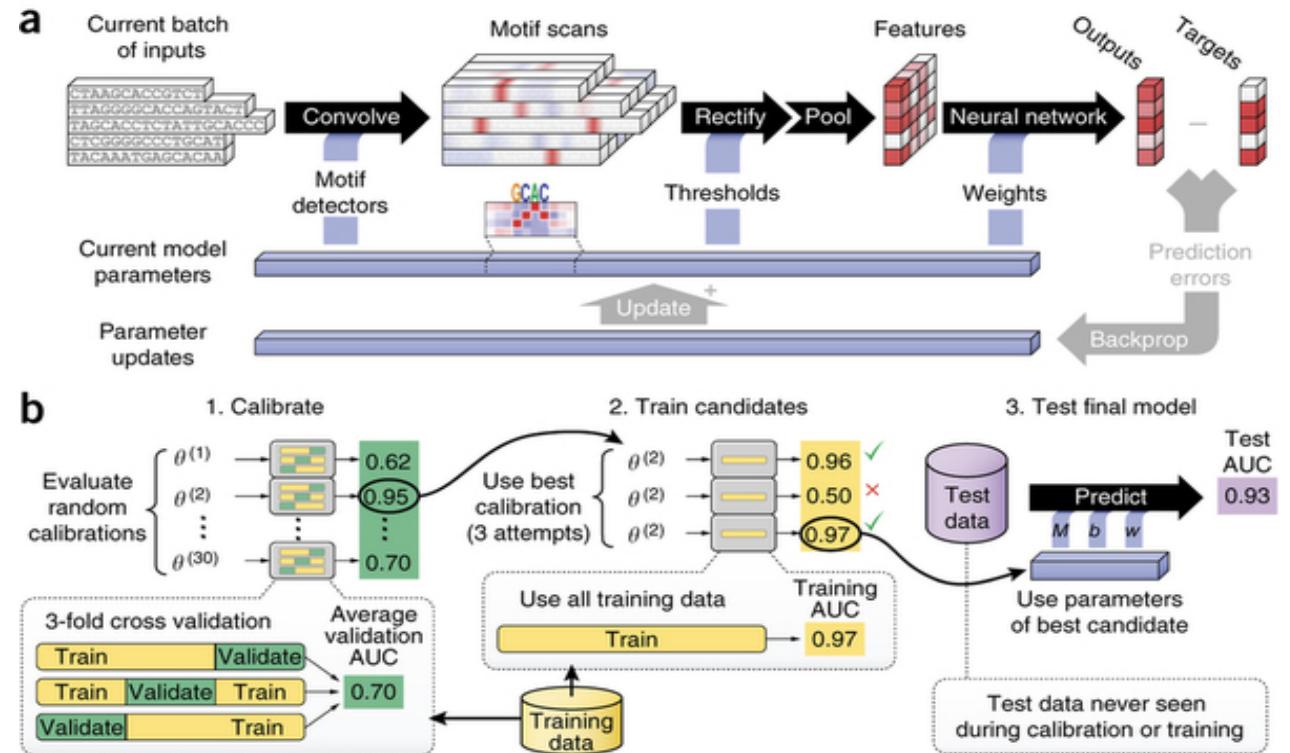


The restriction enzyme EcoRV (green)

Source: wikipedia.org/wiki/DNA-binding_protein

12/11/2017

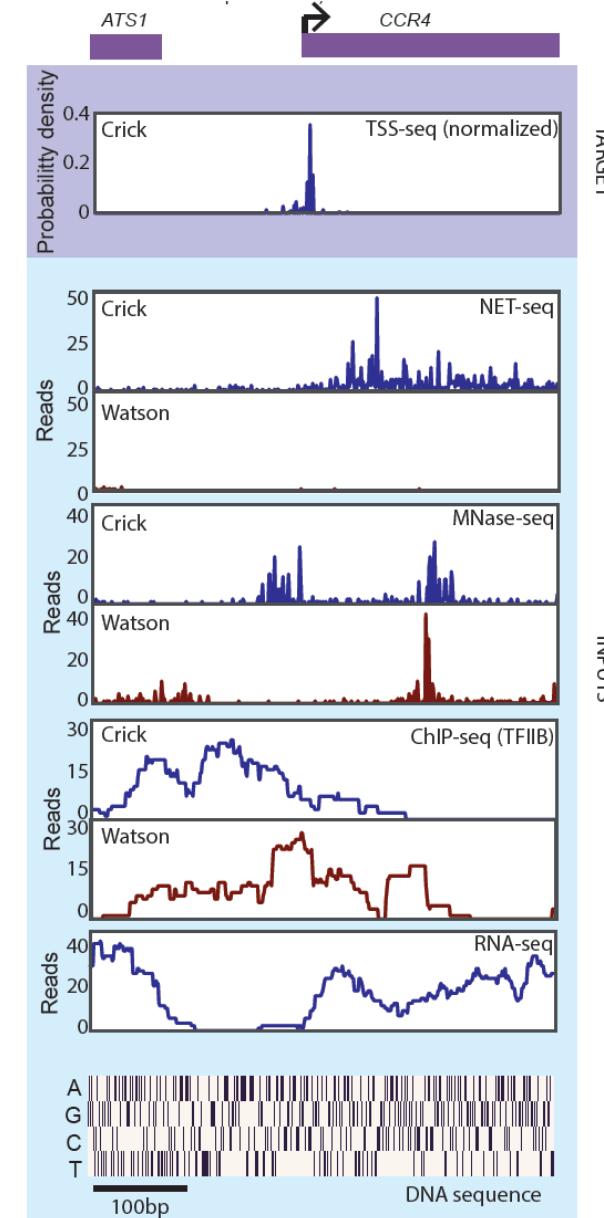
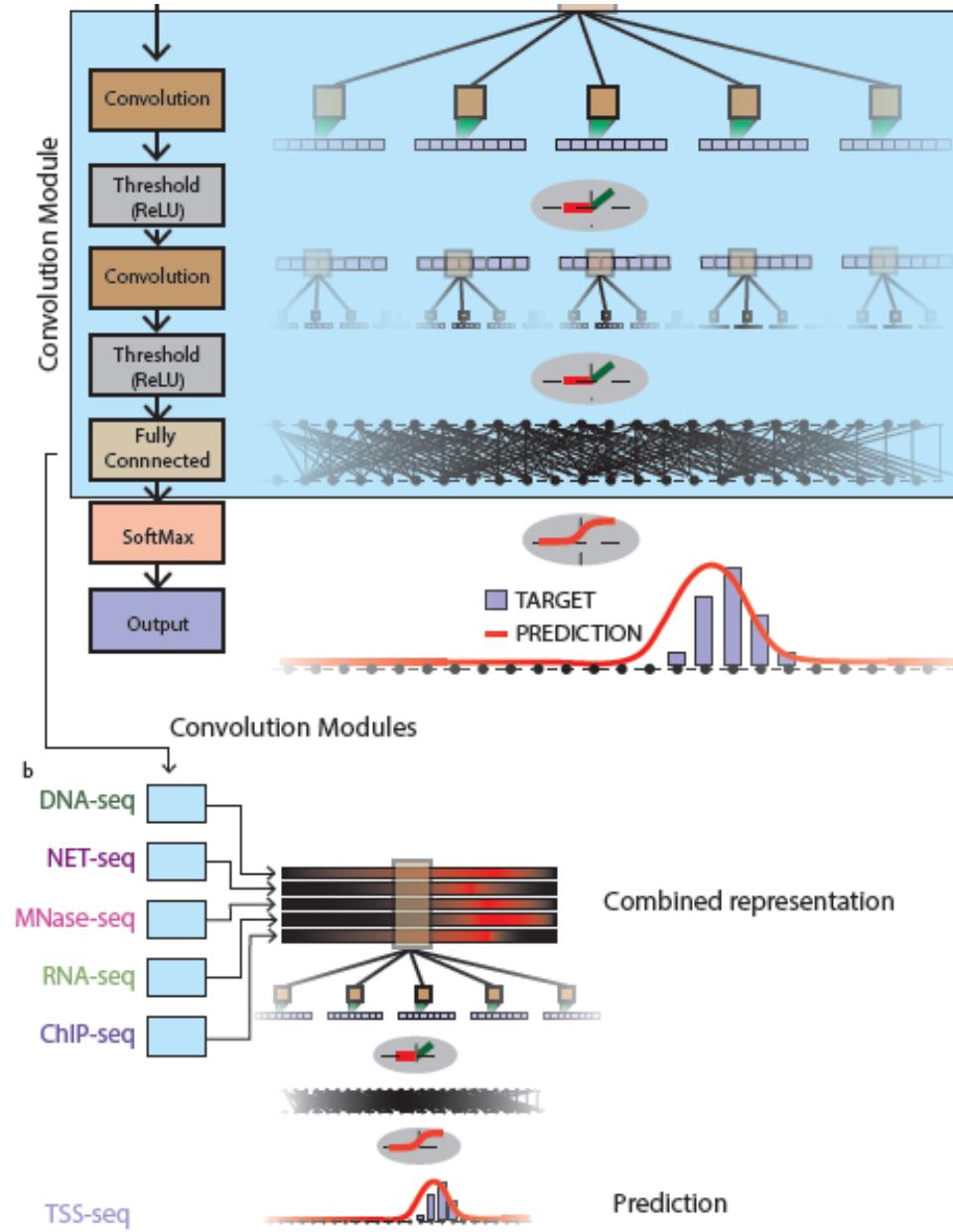
DeepBind (Alipanahi et al, Nature Biotech 2015)



<http://www.nature.com/nbt/journal/v33/n8/full/nbt.3300.html>

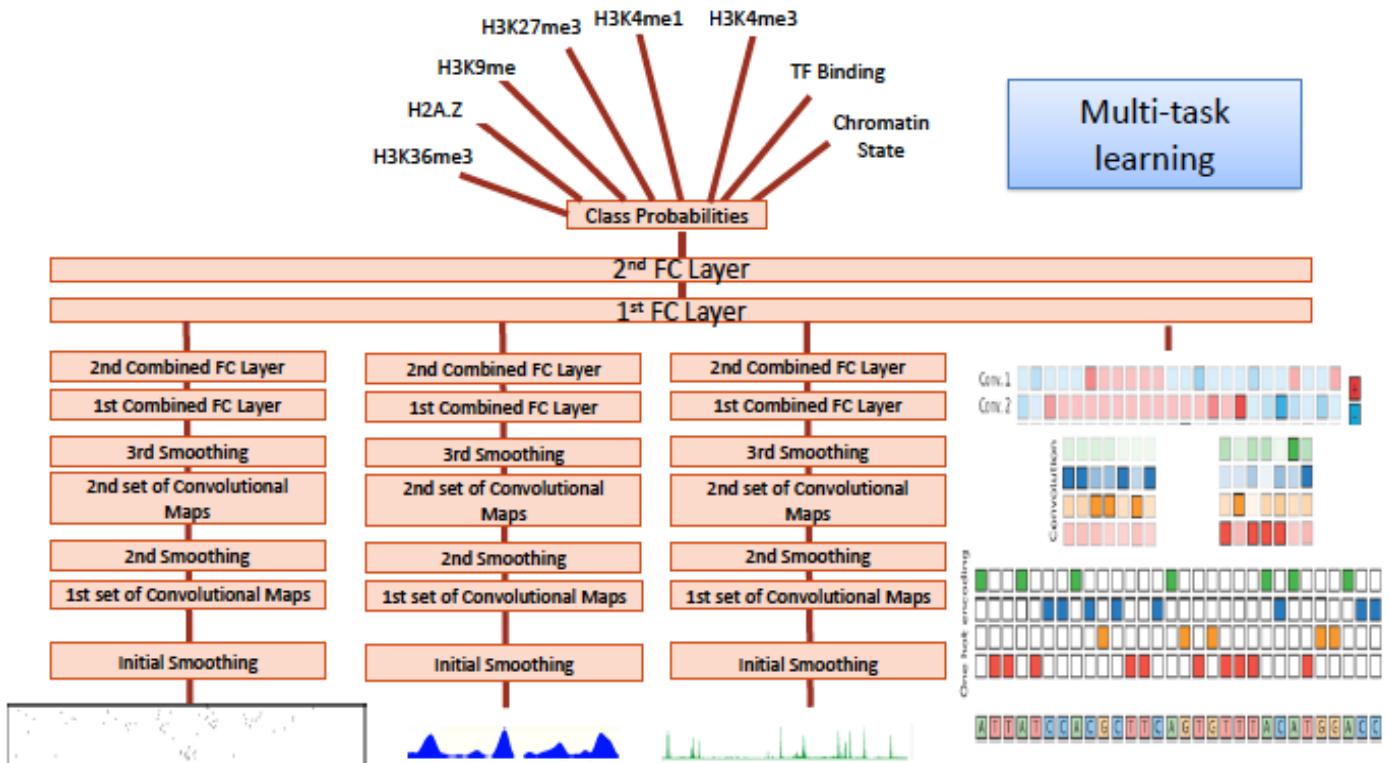
Use of CNNs: FIDDLE

#REF: Eser, Umut, and L. Stirling
 Churchman. "FIDDLE: An integrative deep learning framework for functional genomic data inference." *bioRxiv* (2016): 081380.

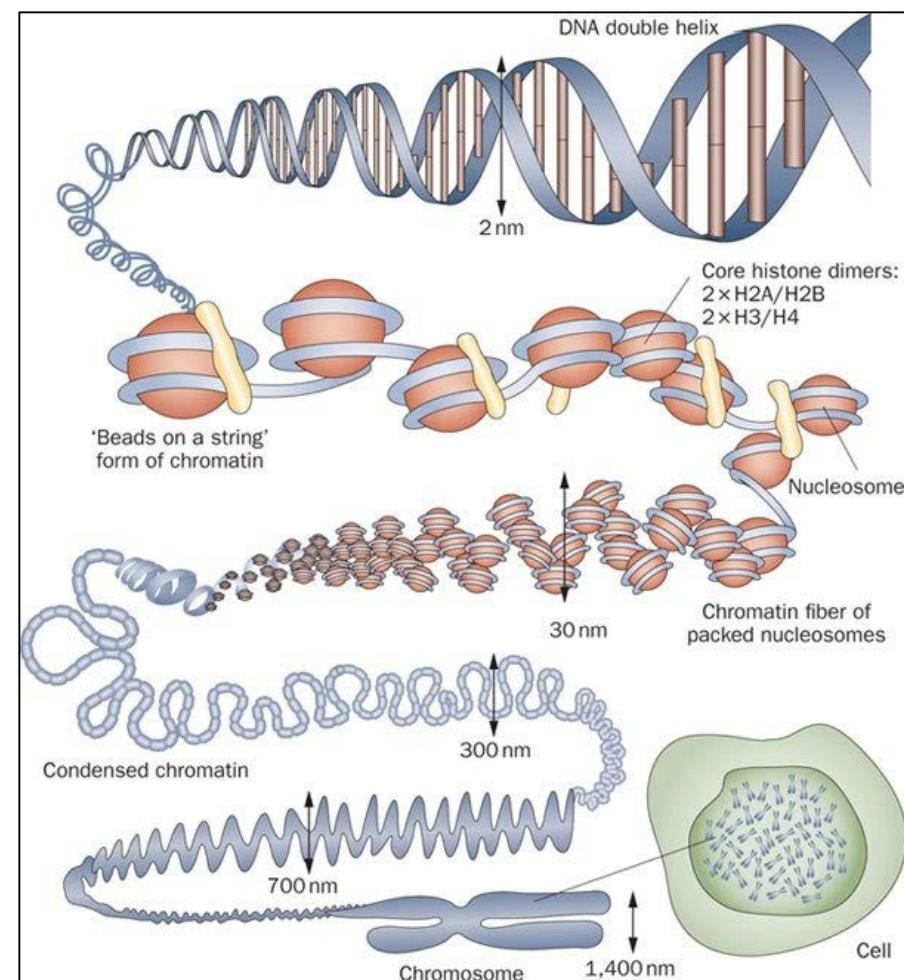


THE CHROMPUTER

Integrating multiple inputs (1D, 2D signals, sequence)
to simultaneously predict multiple outputs



Chromatins

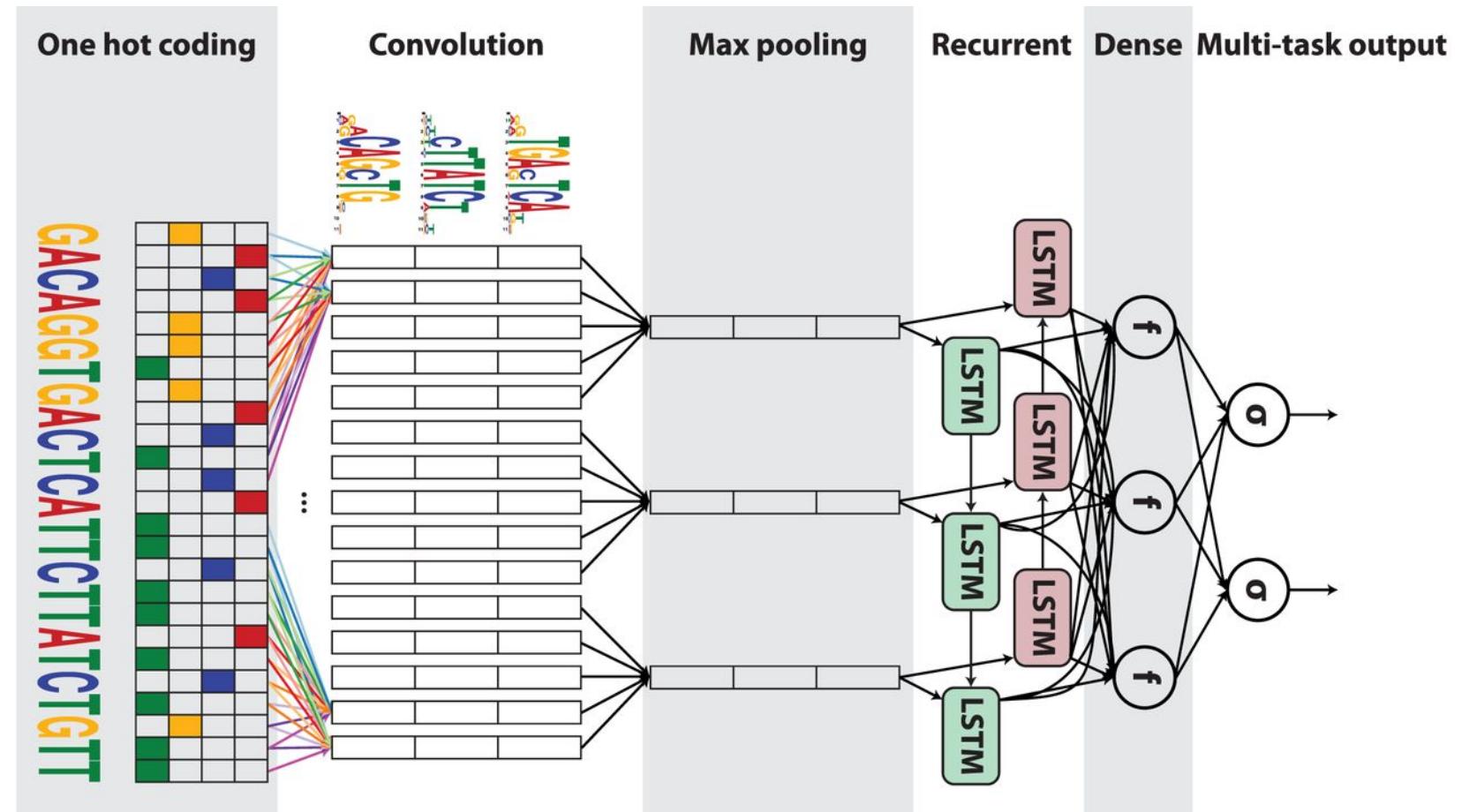


Source: <https://simons.berkeley.edu/sites/default/files/docs/4575/2016-kundaje-simonsinstitute-deeplearning.pdf>

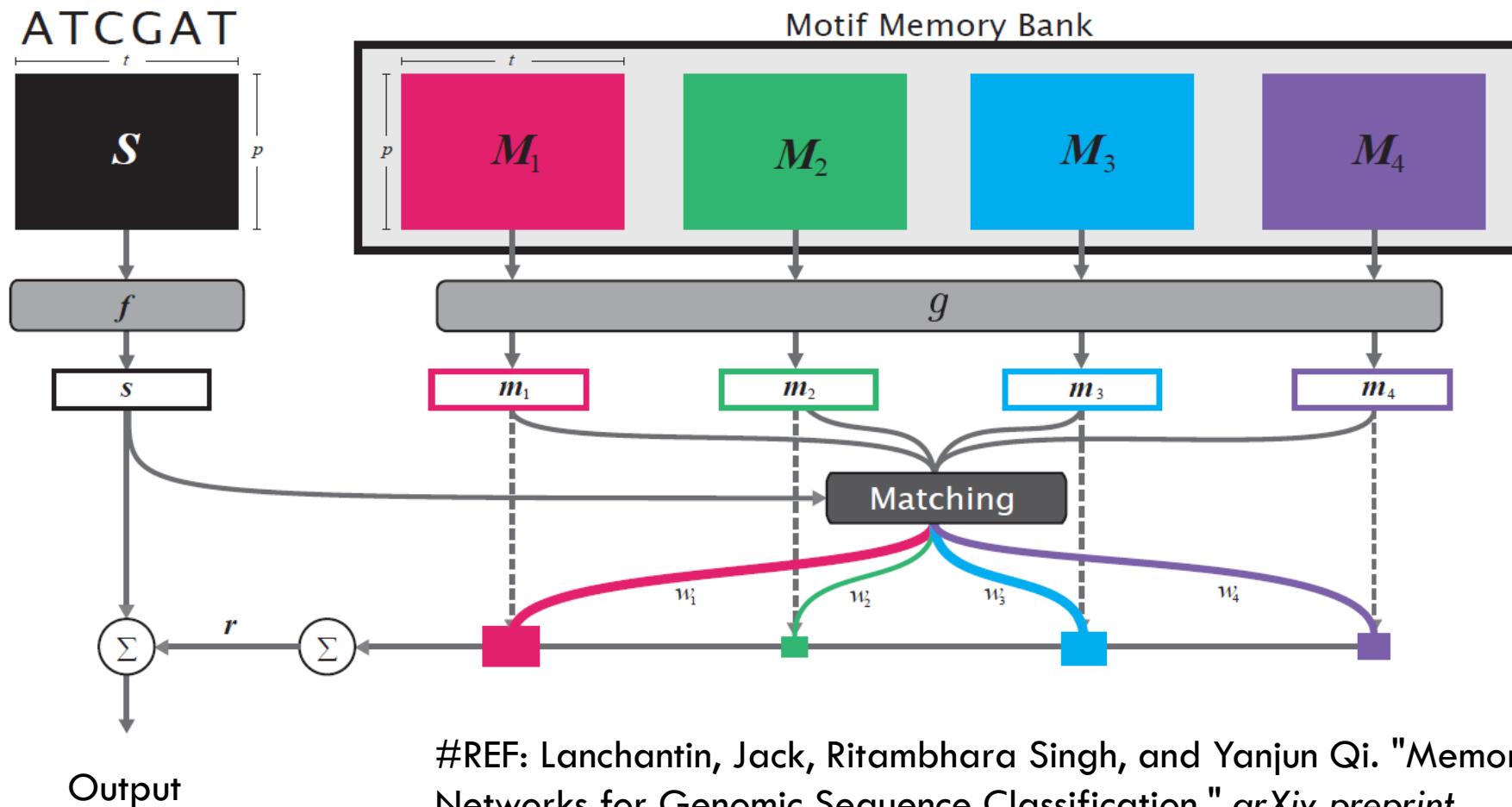
<https://qph.ec.quoracdn.net>

User of CNN+RNNs: DanQ

#REF: Quang, Daniel, and Xiaohui Xie.
"DanQ: a hybrid convolutional and recurrent deep neural network for quantifying the function of DNA sequences." *Nucleic acids research* 44.11 (2016): e107-e107.

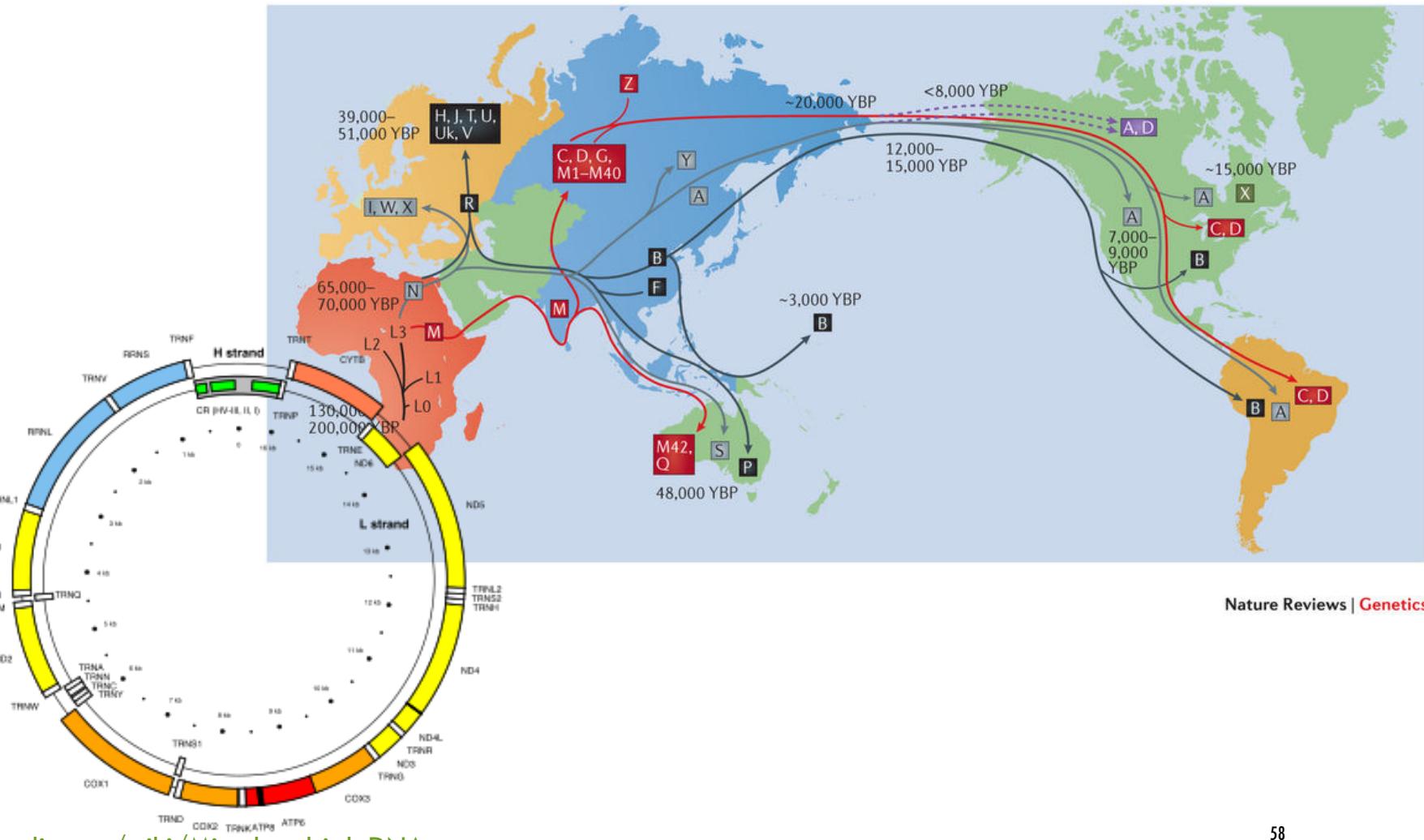
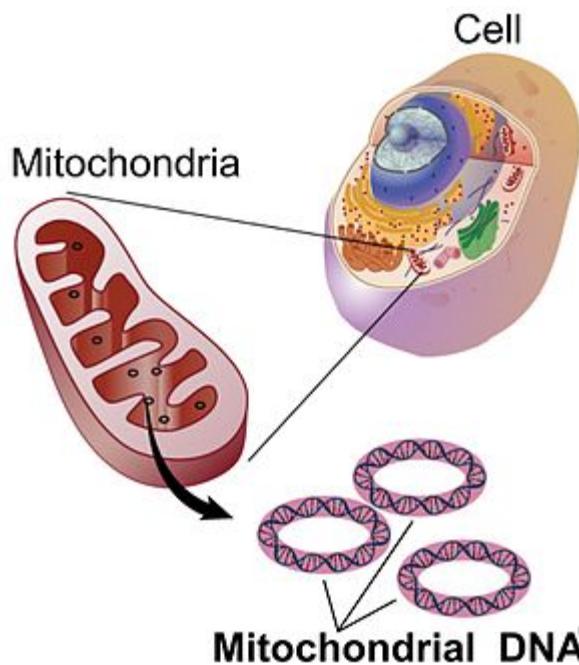


Use of MANN: Matching nets



#REF: Lanchantin, Jack, Ritambhara Singh, and Yanjun Qi. "Memory Matching Networks for Genomic Sequence Classification." *arXiv preprint arXiv:1702.06760* (2017).

Mitochondrial DNA



MtDNA versus aging

(on going work @PraDA/Deakin)

On average, 500 copies per cell

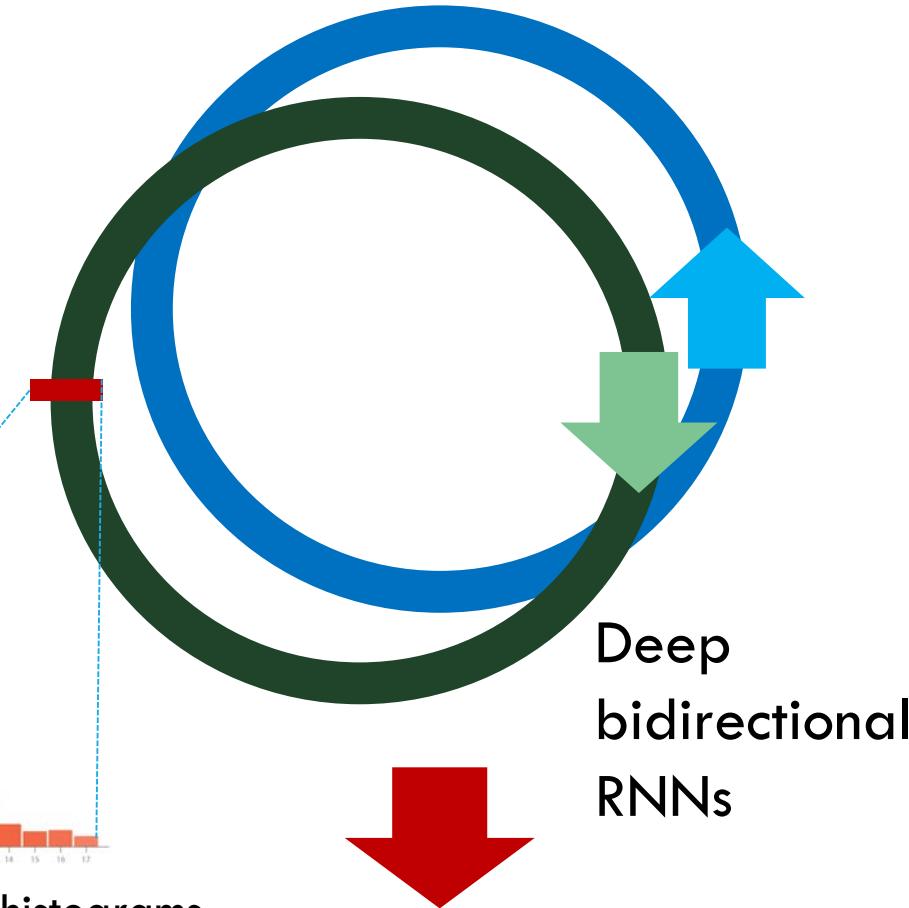
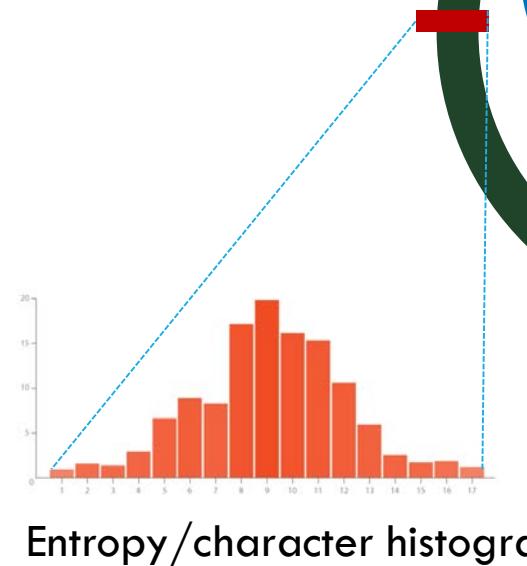
Reading is per shot segments

- Suggests multiple instance learning

The entire MtDNA is a ring

- Suggests Bidirectional RNNs

For aging, the entropy at each location seems to matter



More models/frameworks

DragoNN

DeepChrome

DeepSEA

Basset

DeepBound

...

The screenshot displays the GitHub repository for DragoNN. It includes:

- Code snippets for `motif_density_localization_simulation_parameters` and `one_filter_dragonn_parameters`.
- A learning curve plot titled `SequenceDNN_learning_curve(one_filter_dragonn)`.
- A sequence motif visualization titled `interpret_SequenceCNN_filters(multi_layer_dragonn, simulation_data)`, showing motifs CCA and CTG.
- A multi-filter interpretation plot titled `interpret_data_with_SequenceCNN(multi_filter_dragonn, simulation_data)`, showing three filters (Filter 5, Filter 9, Filter 13) with corresponding sequence contexts.
- A central box labeled `DragoNN` containing:
 - `IPython Notebook Tutorials` and `Command Line Interface` buttons.
 - `SimDNA`, `Keras`, and `DeepLIFT` buttons.
 - `TensorFlow` and `Theano` buttons.
 - `CPU` and `GPU` buttons.
- Textual usage instructions:

```
usage: dragonn [-h] {train,test,predict,interpret}
```
- Positional arguments:

```
{train,test,predict,interpret}
```
- Help options:
 - `dragonn command help`
 - `model training help`
 - `model testing help`
 - `model prediction help`
 - `model interpretation help`

<http://kundajelab.github.io/dragonn>

The outlook

Read an extremely long book and answer any queries about it

- Memory-augmented neural networks (MANN), and
- Multiple hierarchical attentions and grammars

Instead of read, write (DNA/viruses/RNA/proteins)

Super-rich genome SNP annotation

The society of things (DNA/RNA/protein)

Transfer learning between cell types, tissues and diseases

Biology-driven deep nets (e.g., knowledge as memory)

Handling rare events (e.g., the role of memory)



and break

We're hiring

PhD & Postdocs

truyen.tran@deakin.edu.au