Computational Systems Biology
Deep Learning in the Life Sciences

6.802 6.874 20.390 20.490 HST.506

David Gifford

Lecture 1

February 7, 2017

http://mit6874.github.io

Today's lecture

- Course overview
- How to be scientific
- TensorFlow as a computing paradigm

Welcome to a new approach to life sciences research

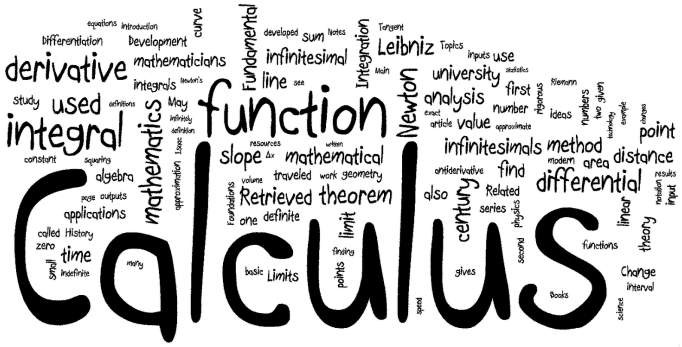
- Enabled by the convergence of three things
  - Inexpensive, high-quality, collection of large data sets (sequencing, imaging, etc.)
  - New machine learning methods (including ensemble methods)
  - High-performance Graphics Processing Unit (GPU) machine learning implementations
- Result is completely transformative

Subject number differences

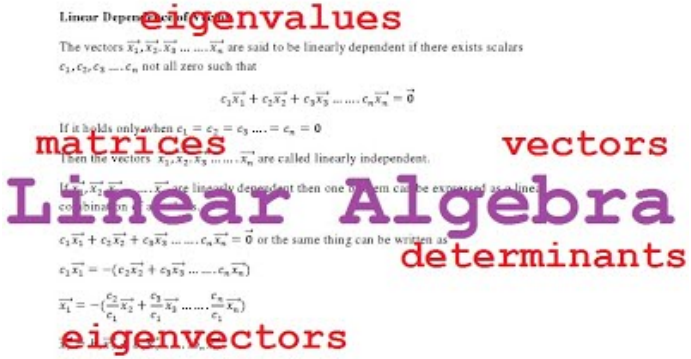
- 6.802 20.390
  - Identical undergrad versions
- 20.490 HST.506
  - Extra part on problem sets
- 6.874
  - EECS Area 2 TQE
  - Extra part on problem sets
  - Solo project (unless permission granted)

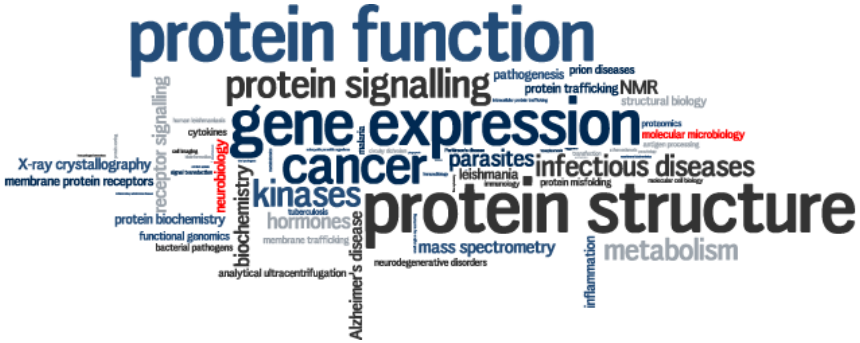
Your background

- Calculus, Linear Algebra
- Probability, Programming
- Introductory Biology




```
def loadstable(ver):
    return _loadversion(ver, prefix="_stable_")

def loadunstable(ver):{...}
def loadexact(ver):{...}
def _loadversion(ver, prefix):
    targetname = prefix + ver.replace('.', '_')
    mainpackage = _original__import__("tools", globals(), locals(),
        [targetname])
    global currentversion
    currentversion = getattr(mainpackage, targetname)

    # Let users change versions after choosing this one
    currentversion.loadstable = loadstable
    currentversion.loadunstable = loadunstable
    currentversion.loadexact = loadexact

    return currentversion

currentversion = None
```





Our goal for you

- This subject is not an encyclopedic summary of contemporary methods in systems biology and genomics.
- We will explore both conventional and deep learning approaches to key problems in the life sciences, comparing and contrasting their power and limits.
- Enable you to execute on new enabling solutions that can have large impact

There are alternative MIT subjects

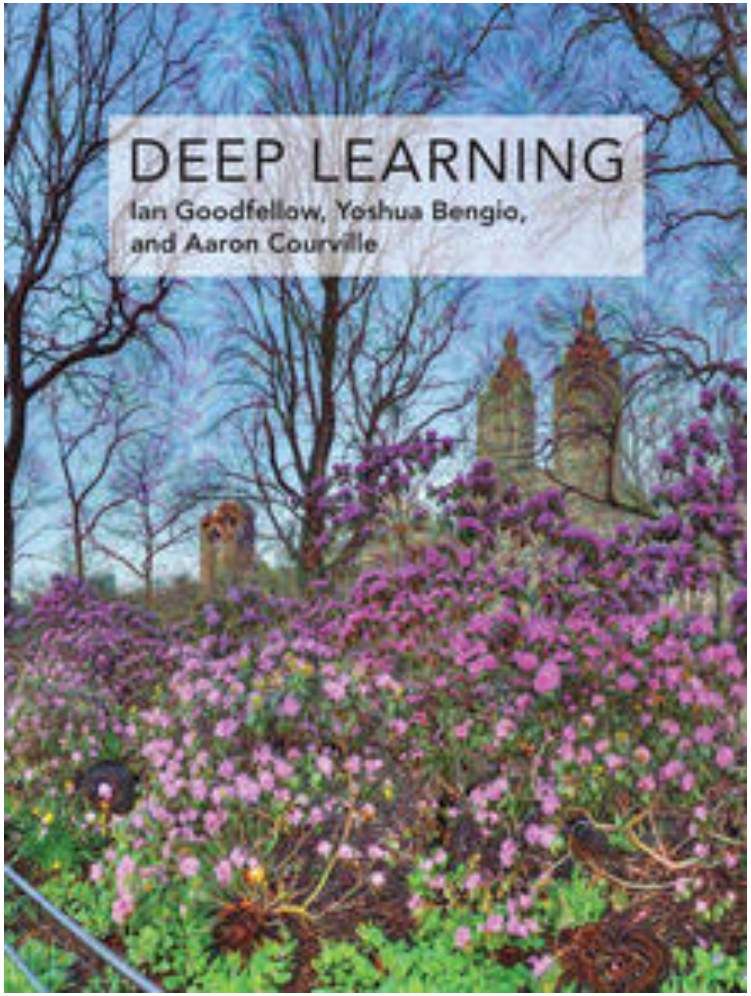
- 6.047 / 6.878 Computational Biology: Genomes, Networks, Evolution
- 8.592 Statistical Physics in Biology
- 7.09 Quantitative and Computational Biology
- 7.33 Evolutionary Biology: Concepts, Models and Computation
- 7.57 Quantitative Biology for Graduate Students
- 18.417 Introduction to Computational Molecular Biology
- 20.482 Foundations of Algorithms and Computational Techniques in Systems Biology


This subject is the red pill

Machine Learning is the ability to improve on the task performance with more training data

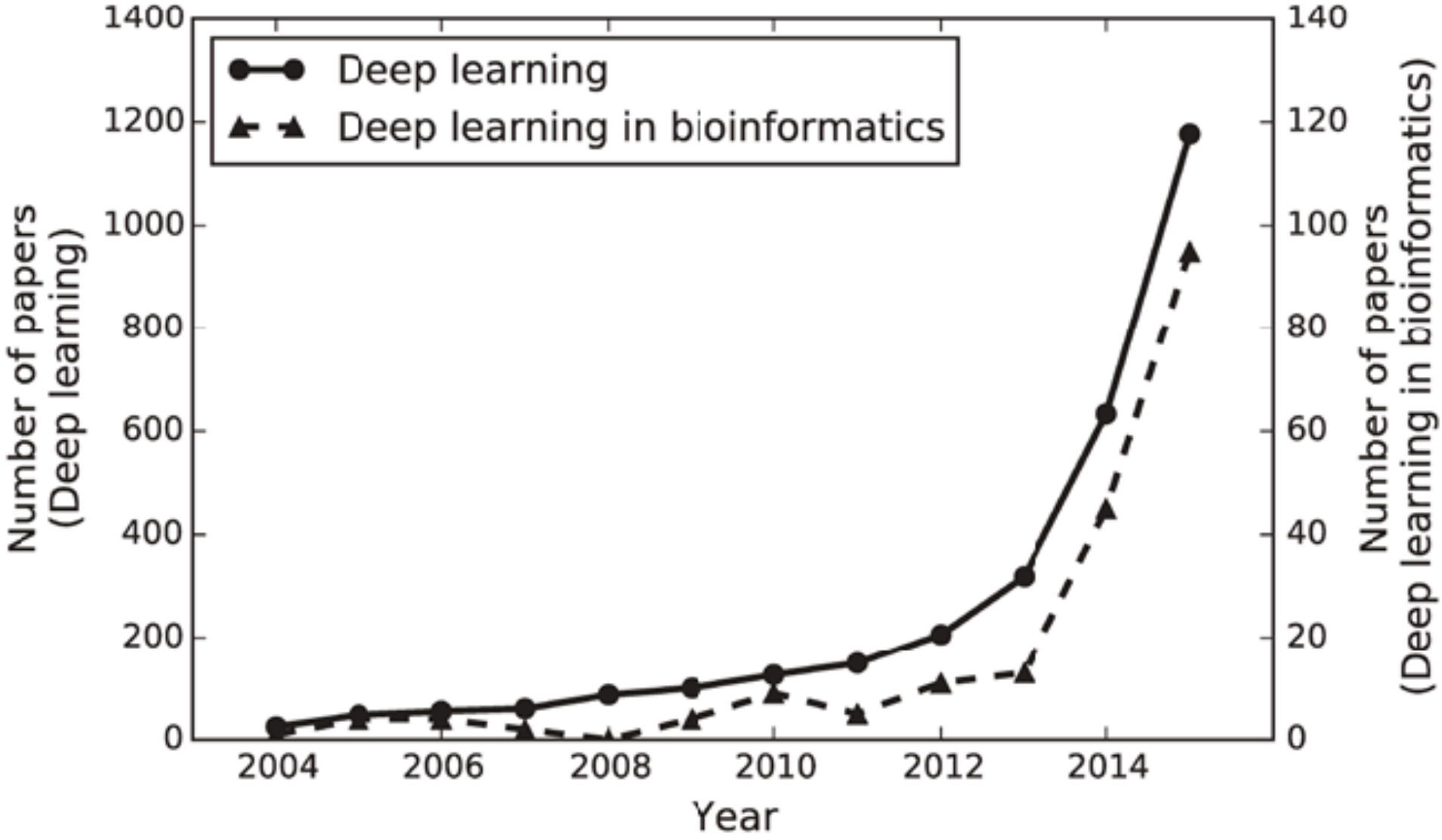
- Task T to be performed
  - Classification, Regression, Transcription, Translation, Structured Output, Anomaly Detection, Synthesis, Imputation, Denoising
- Measured by Performance Measure P
- Trained on Experience E (Training Data)

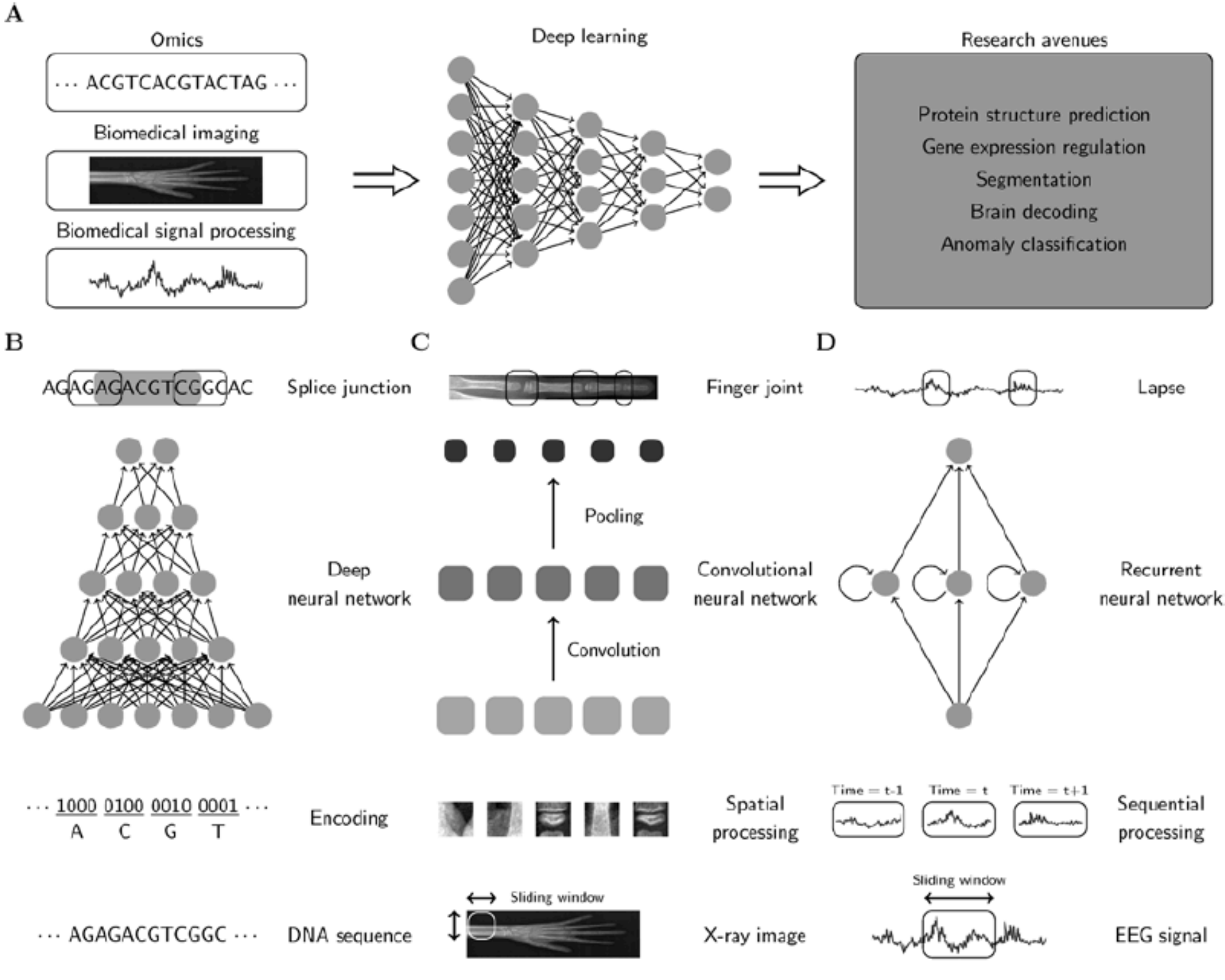
We will use this book deeplearningbook.org



20% Discount for MIT Students at MIT Press Bookstore = $64

Approximately 8% of deep learning publications are in bioinformatics




A
Omics
··· ACGTCACGTACTAG ···
Biomedical imaging
Biomedical signal processing
Deep learning
Research avenues
Protein structure prediction
Gene expression regulation
Segmentation
Brain decoding
Anomaly classification
B
C
D
AGAGAGACGTCGGCAC
Splice junction
Finger joint
Lapse
Pooling
Convolution
Deep
neural network
Convolutional
neural network
Recurrent
neural network
··· 1000 0100 0010 0001 ···
A C G T
Encoding
Spatial
processing
Time = t-1
Time = t
Time = t+1
Sequential
processing
··· AGAGACGTCGGC ···
DNA sequence
Sliding window
X-ray image
Sliding window
EEG signal

Your guides

Jonas Muller
jonasm@mit.edu

Michael Sun
micsun@mit.edu

Haoyang Zeng
zenghaoyang@gmail.com

http://mit6874.github.io


Your computing resource

Your programming environment

Problem 2

In this problem, we wish to use CNN to learn the motif of CTCF from sequences with similar di-nucleotide frequency. The positive samples are 101bp sequences centered at CTCF ChIP-seq peaks from GM12878 cell line. The negative sequences are generated by permuting the nucleotides in the positive sequences while keeping the di-nucleotide frequency.

We will provide functions for loading data, training and testing. You will:

- implement a CNN model with given specifications
- specify the initilization of parameters in the model
- train the model and evaluate on the test set

All the places where you need to fill in begins with "TODO" and ends with "END OF YOUR CODE".

In [1]:
```
import tensorflow as tf, sys, numpy as np, h5py
from os.path import join,dirname,basename,exists,realpath
from os import makedirs
from tensorflow.examples.tutorials.mnist import input_data
from sklearn.metrics import roc_auc_score
```

In [2]:
```
data_folder = '../data/motif_disc'
batch_size = 128
valid_size = 2000
epochs = 20
best_model_file = join('../output',basename(data_folder),'best_model.ckpt')
if not exists(dirname(best_model_file)):
    makedirs(dirname(best_model_file))
```

In [3]:
```
# Function to load the data embedded in the previous problem and their labels
def load_data(mydir):
    train = h5py.File(join(mydir,'train.h5'),'r')
```

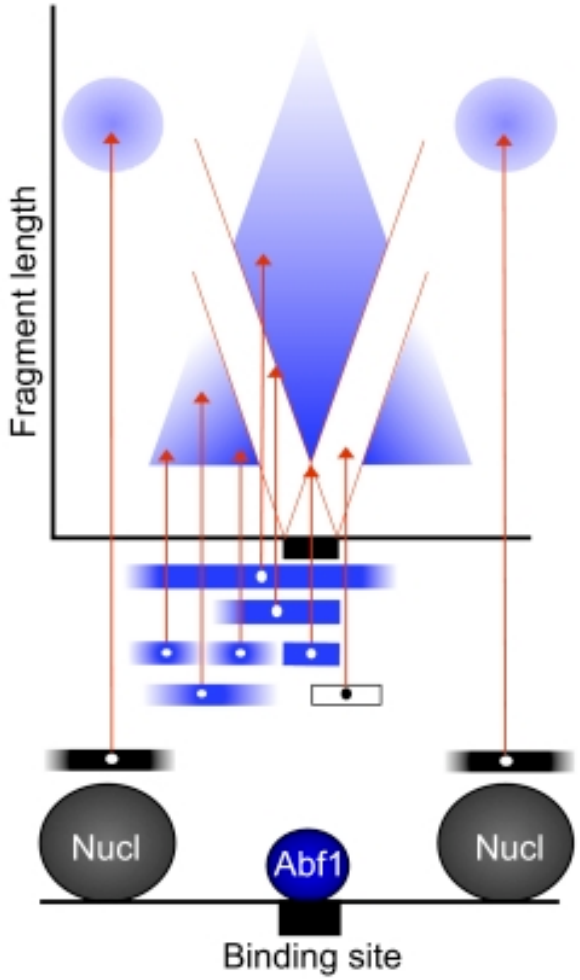
Your grade is based on 3 problem sets, an exam, and a final project

- Three Problem Sets (40%)
  - Individual contribution
  - Done using Google Cloud, Jupyter Notebook
  - One paper presentation can substitute for a PS
- In class exam (1.5 hours), one sheet of notes (30%)
- Final Project (30%)
  - Done in as individually or in teams (6.874 by permission)
  - Substantial question

We will have three modules

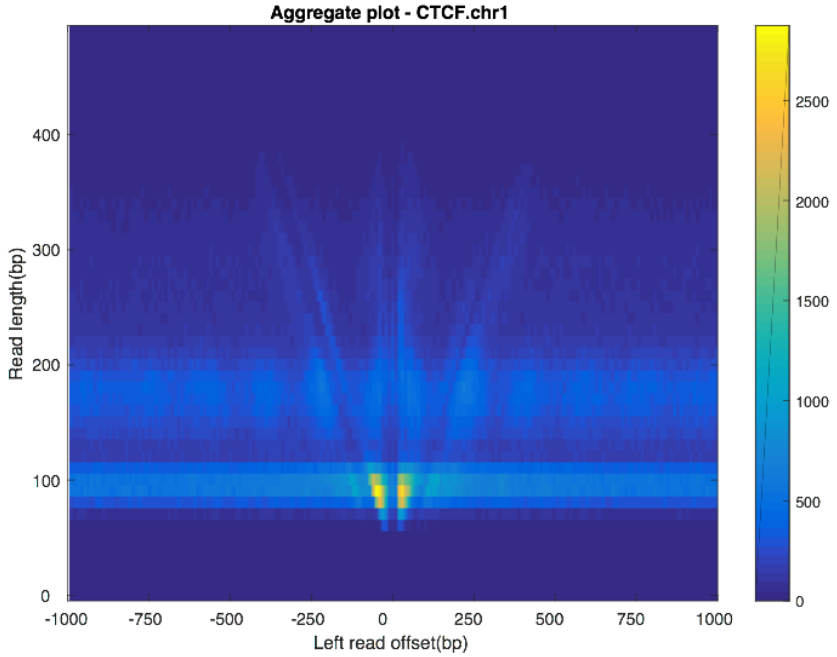
- Module 1: Machine learning principles
  - Introduction, Neural Networks, Tensor Flow, CNNs, RNNs, PCA, t-SNE, Autoencoders
- Module 2: Problems in life sciences
  - Regulatory Genome, Image Understanding, Text Understanding, Variant Prioritization
- Module 3: Project
  - Meetings with mentors

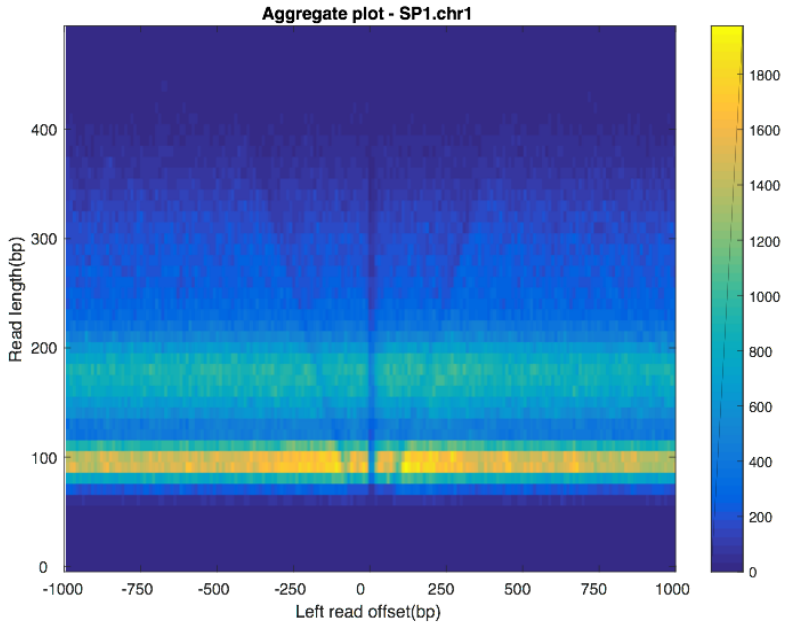
PS 1: Interpreting ATAC-seq V-Plots

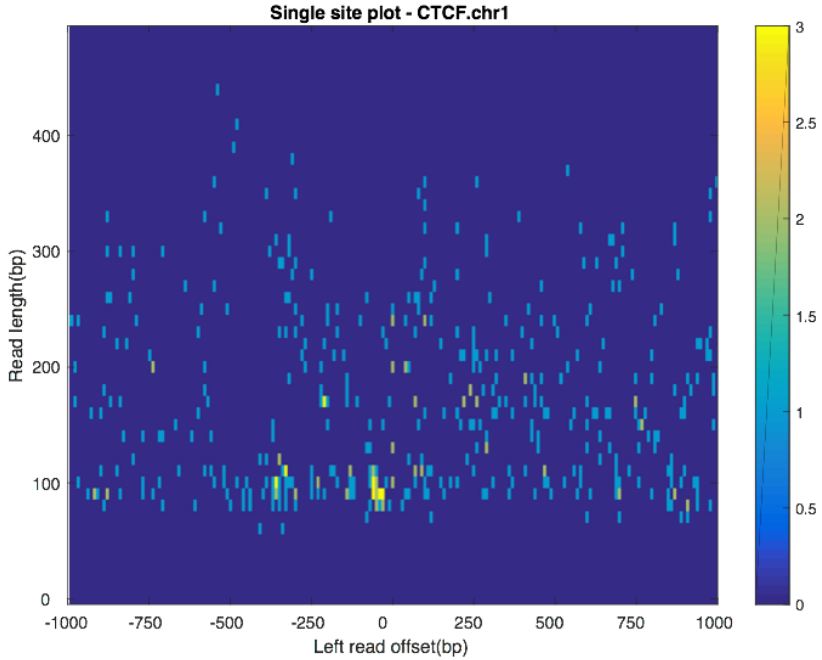


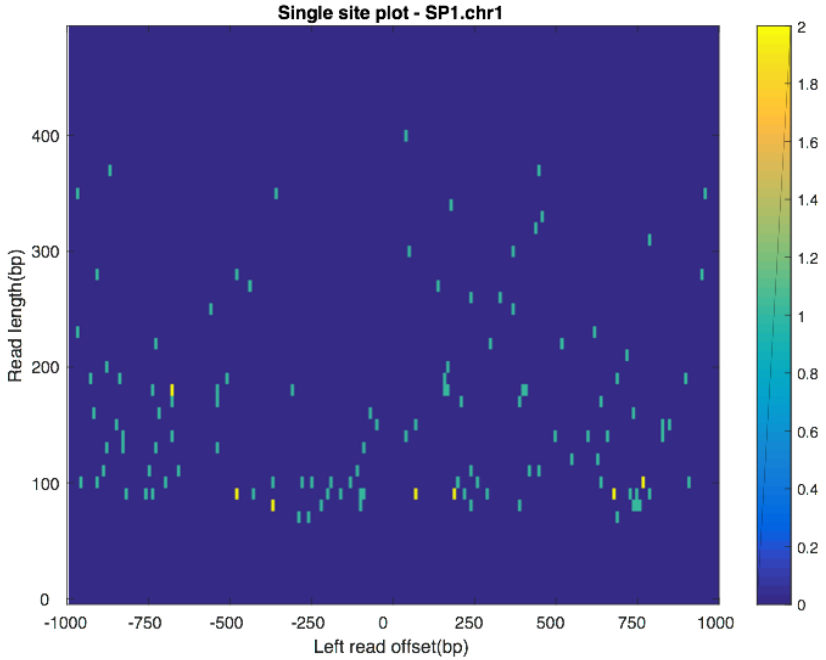
https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3215028/

PS 1 Task: Label single sites









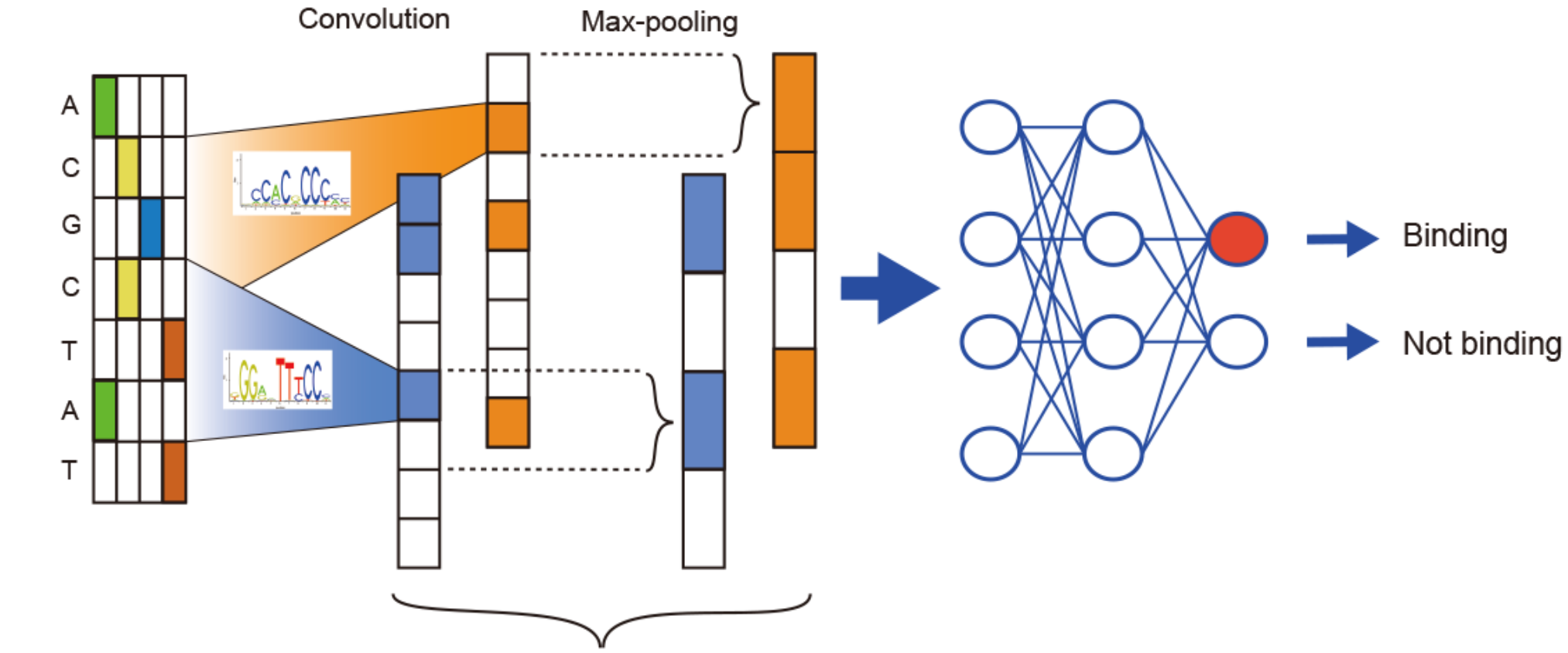
PS 2: Understanding medical records

discharge instructions: if bleeding recurs or symptoms worsen, including lightheadedness, palor skin, or weakness, please call or go to the emergency room. otherwise pleae follow up with dr. [*omitted*] (below) in 1-2 weeks. continue taking your plavix.$ discharge instructions: please call your primary care physician or return to a local emergency room if [*omitted*] experience any worsening headaches, visual changes, speech or language disturbances, focal numbness, weakness, incoordination. . we have started [*omitted*] on a medication called zocor for elevated cholesterol. we have also started [*omitted*] on a medication called lopressor for blood pressure control. . please take all of your medications as directed. . please keep all of your follow up appointments; [*omitted*] will need repeat neuroimaging and/or another angiogram in 4-6 weeks. [*omitted*] will also need an mrv of the brain and tte of the heart.$

PS 2 Task: Synthetic discharges

Sample 1: discharge instructions: please contact your primary care physician or return to the emergency room if [*omitted*] develop any constipation. [*omitted*] should be had stop transferred to [*omitted*] with dr. [*omitted*] or started on a limit your medications. * [*omitted*] see fult dr. [*omitted*] office and stop in a 1 mg tablet to tro fever great to your pain in postions, storale. [*omitted*] will be taking a cardiac catheterization and take any anti-inflammatory medicines diagness or any other concerning symptoms.

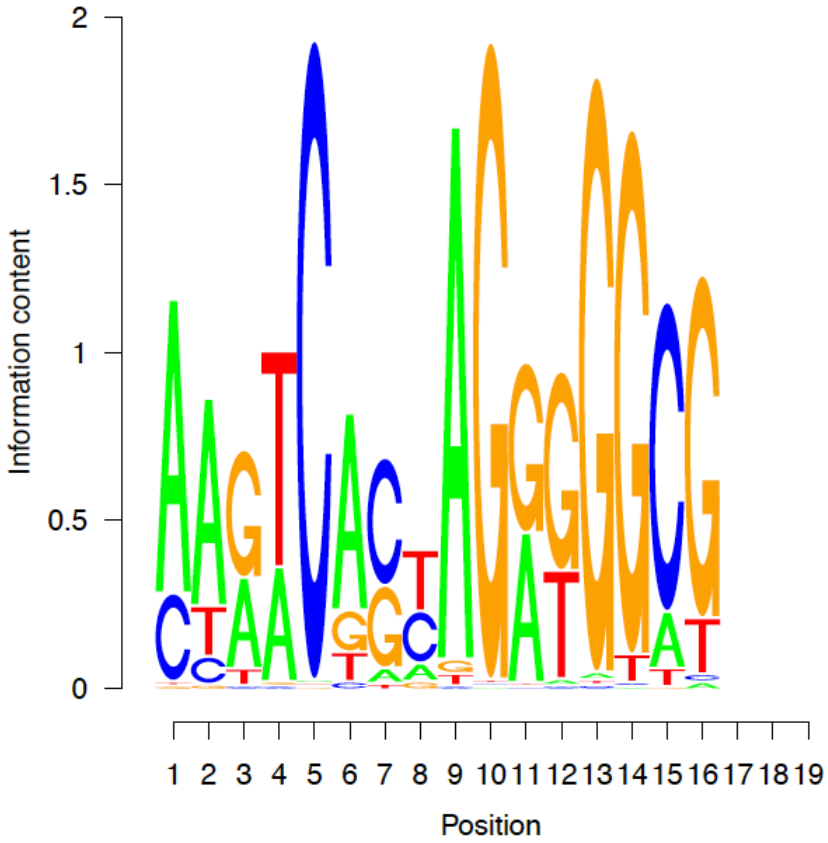
PS 3: Identifying regulatory codes



One-hot encoded sequence context → Stacks of convolutional and max-pooling layers → Fully-connected layers → TF binding prediction

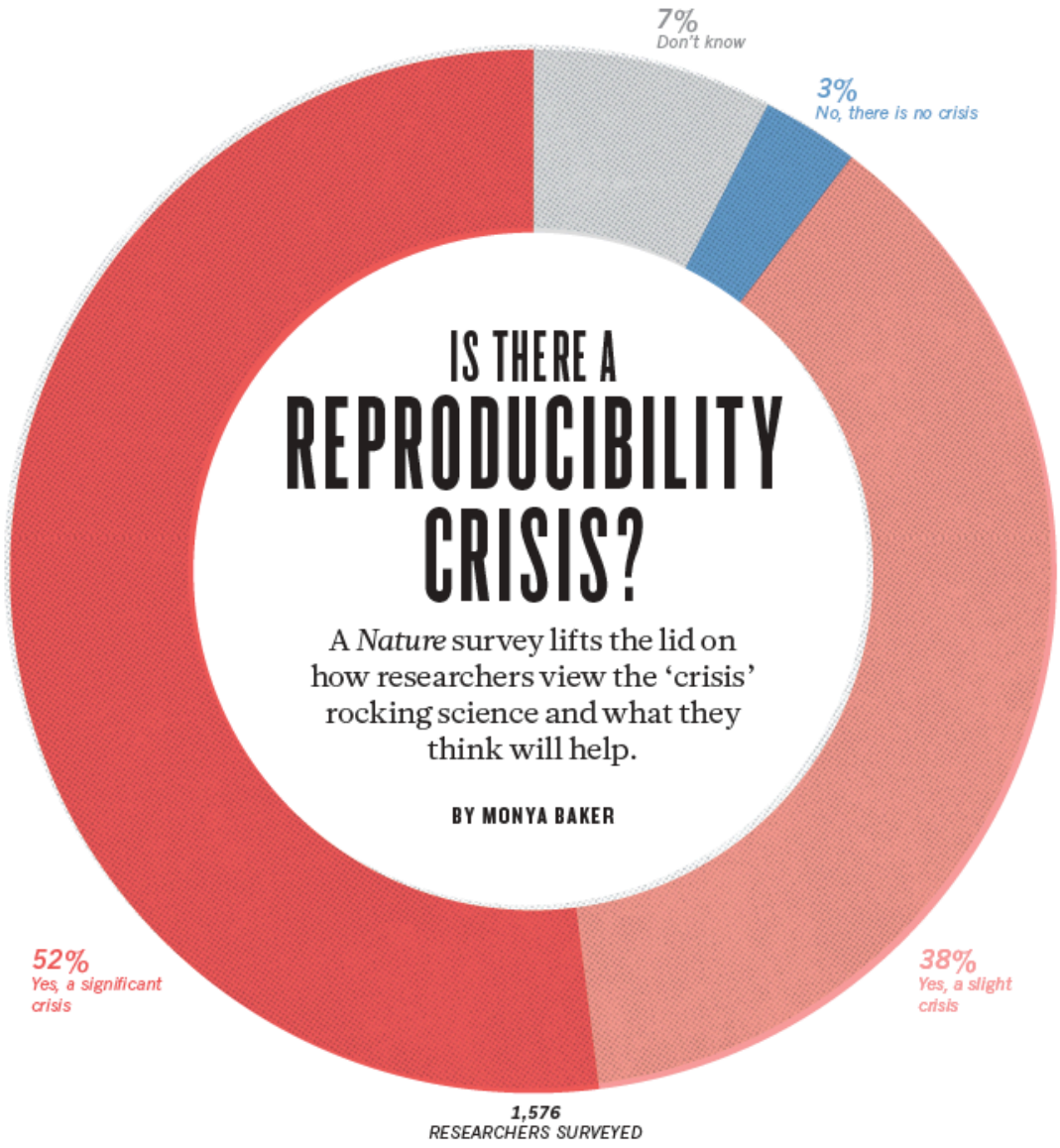
PS 3: Identifying regulatory codes

```
TGAGGCAGACCACCAGAGGGCGTCGCAGGGGTGTCCCTCG
TTCACTCGACCGCTAGGGGAAACTAAGGGGCTTTCCGTAT
CCAAGCTTGCCAGCAGGGGGCGCTATGGGGATTCCCGGTA
GCATGCACACCACTAGAGGTCGCCCCGGGGATTCCCTCGC
TGGTTGGGGCCACCAGGAGGCGGCGAAGGGGATCCCCTCC
TACCACGTGCCCCCAGGTGACGTTGAGGGGATCCCCGCGG
GTTCCGCTGCCTCTAGAGGGCGCCATACGGGGATCCCGGC
CCACCTTGACCACCAGGTGGAACTGAAGGGGATCCCCCGA
CAGGCACTGCCAGCAGGGGGAGCCGCGGGGGATTCCCACG
TACCAATGGCCACCAGAGGGCAGTGTCGGGGATTCCCACG
```



Today's lecture

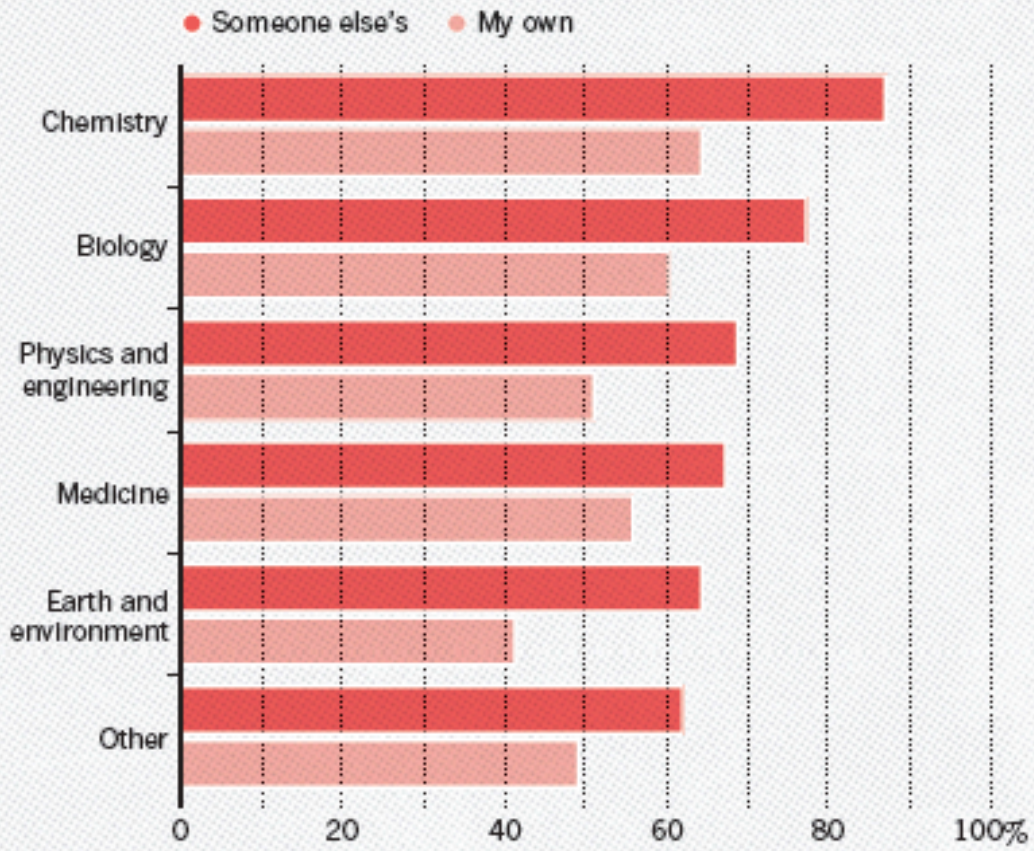
- Course overview
- How to be scientific
- TensorFlow as a computing paradigm



http://www.nature.com/polopoly_fs/1.19970!/menu/main/topColumns/topLeftColumn/pdf/533452a.pdf

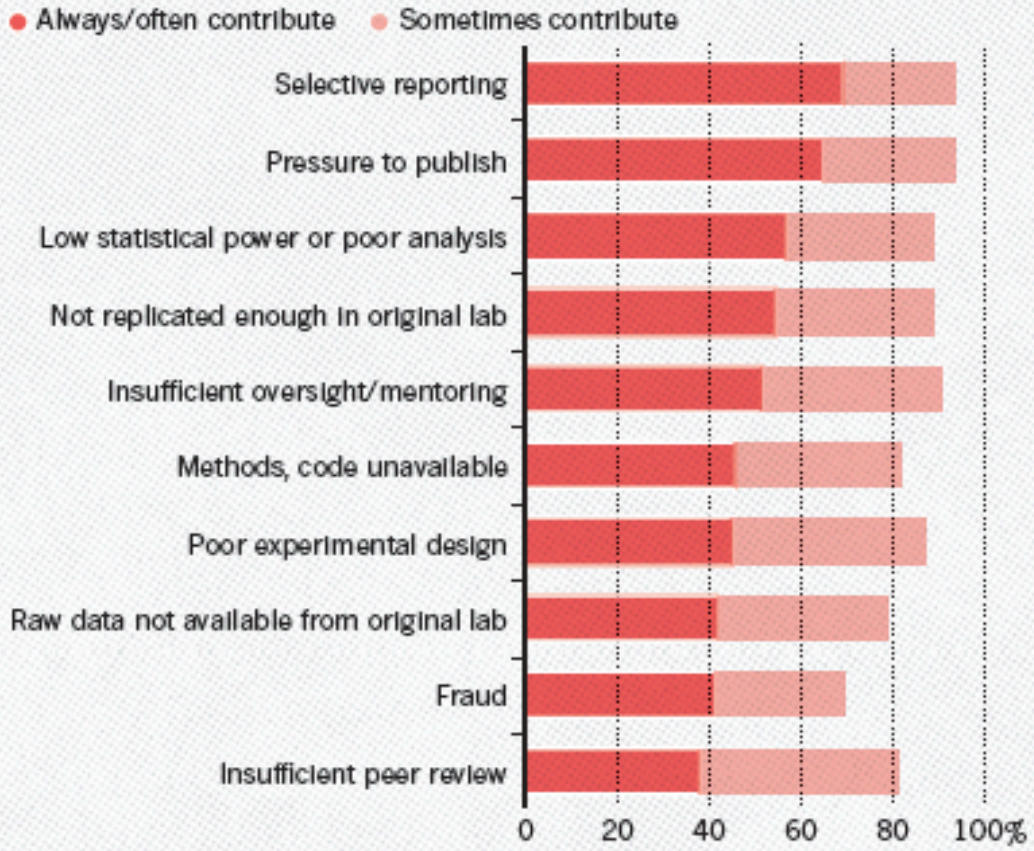
HAVE YOU FAILED TO REPRODUCE AN EXPERIMENT?

Most scientists have experienced failure to reproduce results.



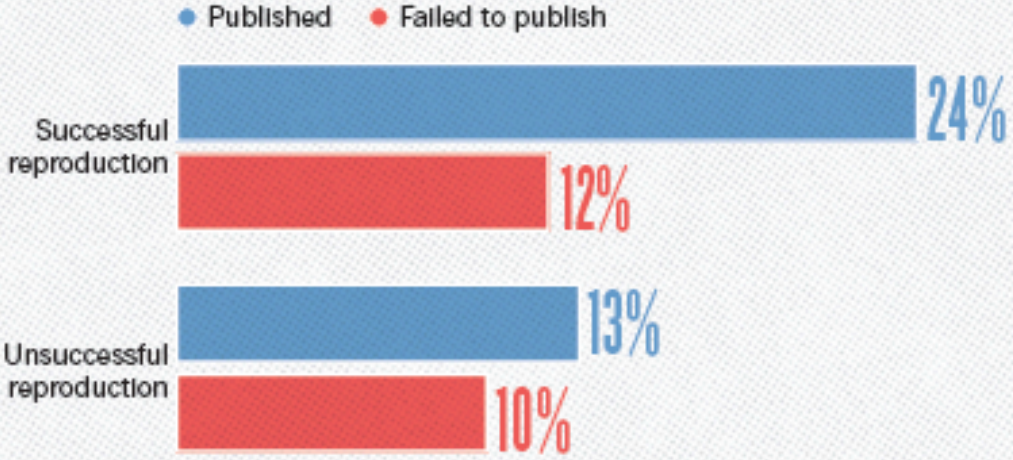
WHAT FACTORS CONTRIBUTE TO IRREPRODUCIBLE RESEARCH?

Many top-rated factors relate to intense competition and time pressure.



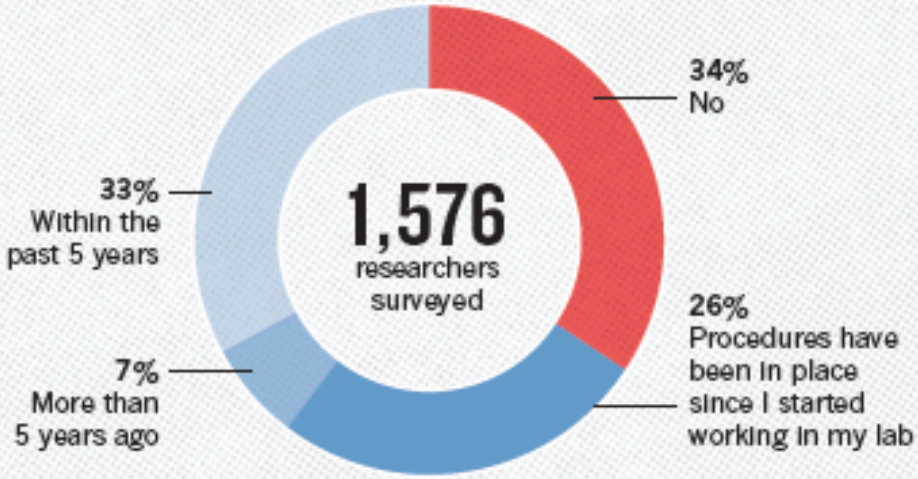
HAVE YOU EVER TRIED TO PUBLISH A REPRODUCTION ATTEMPT?

Although only a small proportion of respondents tried to publish replication attempts, many had their papers accepted.



HAVE YOU ESTABLISHED PROCEDURES FOR REPRODUCIBILITY?

Among the most popular strategies was having different lab members redo experiments.



Number of respondents from each discipline:
Biology **703**, Chemistry **106**, Earth and environmental **95**, Medicine **203**, Physics and engineering **236**, Other **233**

Amgen could not reproduce the findings of 47/53 (89%) landmark preclinical cancer papers

REPRODUCIBILITY OF RESEARCH FINDINGS

Preclinical research generates many secondary publications, even when results cannot be reproduced.

Journal impact factor	Number of articles	Mean number of citations of non-reproduced articles*	Mean number of citations of reproduced articles
>20	21	248 (range 3–800)	231 (range 82–519)
5–19	32	169 (range 6–1,909)	13 (range 3–24)

Results from ten-year retrospective analysis of experiments performed prospectively. The term 'non-reproduced' was assigned on the basis of findings not being sufficiently robust to drive a drug-development programme.
*Source of citations: Google Scholar, May 2011.

http://www.nature.com/nature/journal/v483/n7391/pdf/483531a.pdf

Direct and conceptual replication is important

- **Direct replication** is defined as attempting to reproduce a previously observed result with a procedure that provides no a priori reason to expect a different outcome
- **Conceptual replication** uses a different methodology (such as a different experimental technique or a different model of a disease) to test the same hypothesis; tries to avoid confounders

https://elifesciences.org/content/6/e23383

Reproducibility Project: Cancer Biology Registered Report/Replication Study Structure

- A **Registered Report** details the experimental designs and protocols that will be used for the replications, and experiments cannot begin until this report has been peer reviewed and accepted for publication.
- The results of the experiments are then published as a **Replication Study**, irrespective of outcome but subject to peer review to check that the experimental designs and protocols were followed.

https://elifesciences.org/content/6/e23383

Muddy waters

Results of the first five replication studies run by the Reproducibility Project: Cancer Biology.

Paper	Conclusion	Focus of key experiment	Replication results	Example problem
Sirota, M. *et al. Sci. Transl. Med.* **3,** 96ra77 (2011)	Public gene expression data can identify unintuitive uses for old drugs	Growth of tumours treated with an anti-ulcer drug	Substantially reproduced	Disagreements over appropriateness of statistical analysis
Sugahara, K. N. *et al. Science* **328,** 1031–1035 (2010)	A tumour-penetrating peptide enhances the effects of cancer drugs	Growth of peptide-treated tumours	Not reproduced	Potential differences in peptide synthesis or solutions
Willingham, S. B. *et al. Proc. Natl Acad. Sci. USA* **109,** 6662–6667 (2012)	Blocking contact between CD47 and another protein inhibits tumour	Growth and metastasis of treated tumours	Uninterpretable (Treated tumours were larger, but not significantly so)	Some tumours spontaneously regressed
Delmore, J. E. *et al. Cell* **146,** 904–917 (2011)	Blocking a protein sequence damps down pro-cancer genes	Gene expression in treated cells; growth of treated tumours	Substantially reproduced	Bioluminescence/survival for the control groups differed markedly
Berger, M. F. *et al. Nature* **485,** 502–506 (2012)	Sequencing reveals gene that is frequently mutated in melanoma and accelerates growth	Tumour formation in cells carrying mutations	Uninterpretable	Tumours without mutations grew too fast for any accelerated growth to be detected

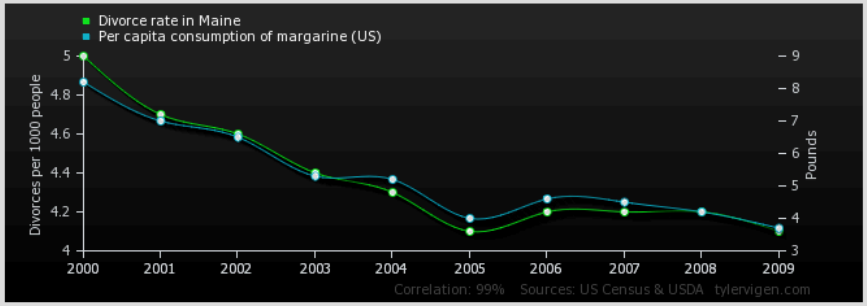
Claim precision is key to science

- “We have discovered the regulatory elements”
- “We have predicted the regulatory elements”

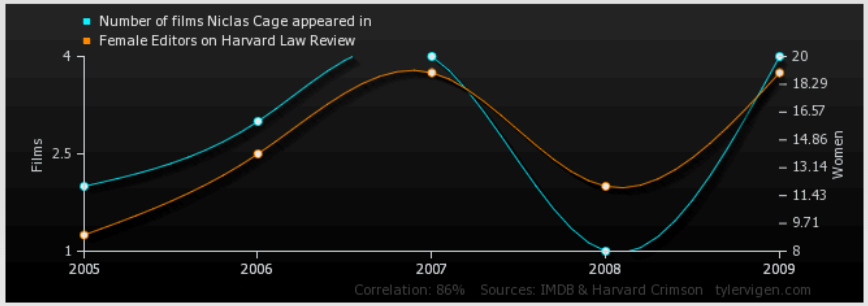
- “The variant causes a difference in gene expression”
- “The variant is associated with a difference in gene expression”

Correlation ≠ Causation

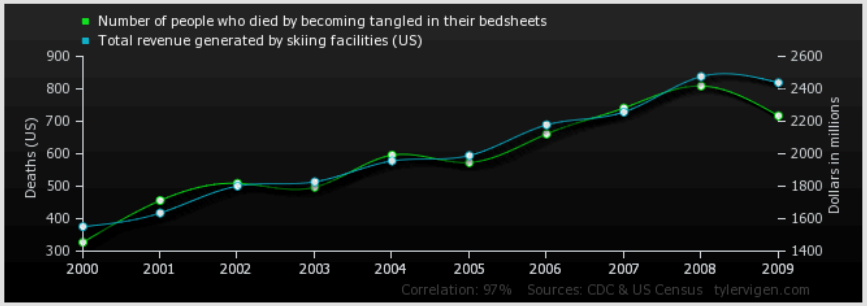
Divorce rate in Maine
correlates with
Per capita consumption of margarine (US)



Number of films Niclas Cage appeared in
correlates with
Female Editors on Harvard Law Review

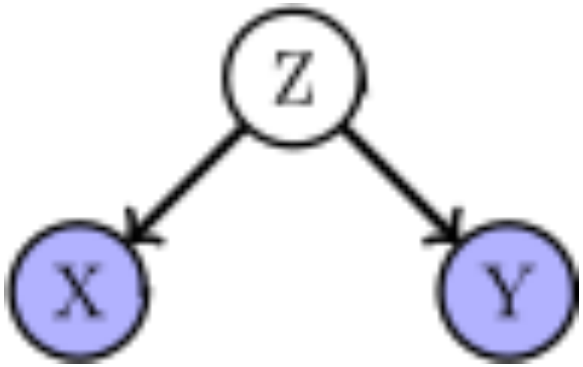


Number of people who died by becoming tangled in their bedsheets
correlates with
Total revenue generated by skiing facilities (US)



Interventions enable causal statements

- Observation only data can be influenced by confounders
- A confounder is an unobserved variable that explains an observed effect
- Interventions on a variable allow for the detection of its direct and indirect effects



Today's lecture

- Course overview
- How to be scientific
- TensorFlow as a computing paradigm

What is TensorFlow?

- TensorFlow takes computations described using a dataflow-like model and maps them onto a wide variety of different hardware platforms
- A tensor is a typed, multi-dimensional array (signed and unsigned integers, IEEE float, complex numbers, byte arrays).

Who is TensorFlow?

TensorFlow:
Large-Scale Machine Learning on Heterogeneous Distributed Systems

Martín Abadi, Ashish Agarwal, Paul Barham, Eugene Brevdo, Zhifeng Chen, Craig Citro,
Greg S Corrado, Andy Davis, Jeffrey Dean, Matthieu Devin, Sanjay Ghemawat, Ian Goodfellow,
Andrew Harp, Geoffrey Irving, Michael Isard, Yangqing Jia, Lukasz Kaiser, Manjunath Kudlur,
Josh Levenberg, Dan Mané, Rajat Monga, Sherry Moore, Derek Murray, Chris Olah,
Jonathon Shlens, Benoit Steiner, Ilya Sutskever, Kunal Talwar, Paul Tucker,
Vincent Vanhoucke, Vijay Vasudevan, Fernanda Viégas, Oriol Vinyals, Pete Warden,
Martin Wattenberg, Martin Wicke, Yuan Yu, and Xiaoqiang Zheng
Google Research*

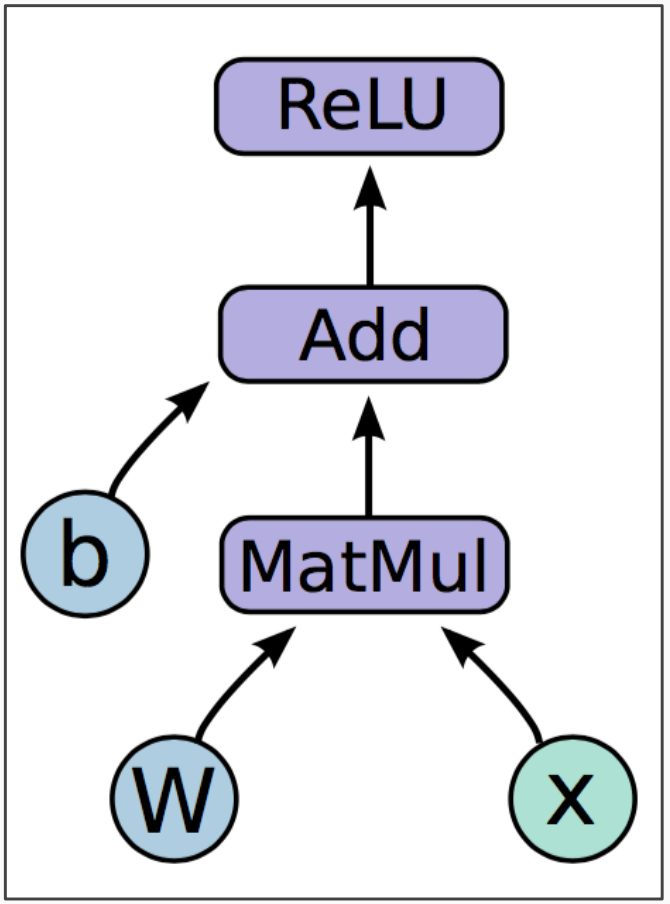
Programming model

Big idea: Express a numeric computation as a **graph**.

- Graph nodes are **operations** which have any number of inputs and outputs
- Graph edges are **tensors** which flow between nodes

Programming model: NN feedforward

$$h_i = \mathrm{ReLU}(Wx + b)$$



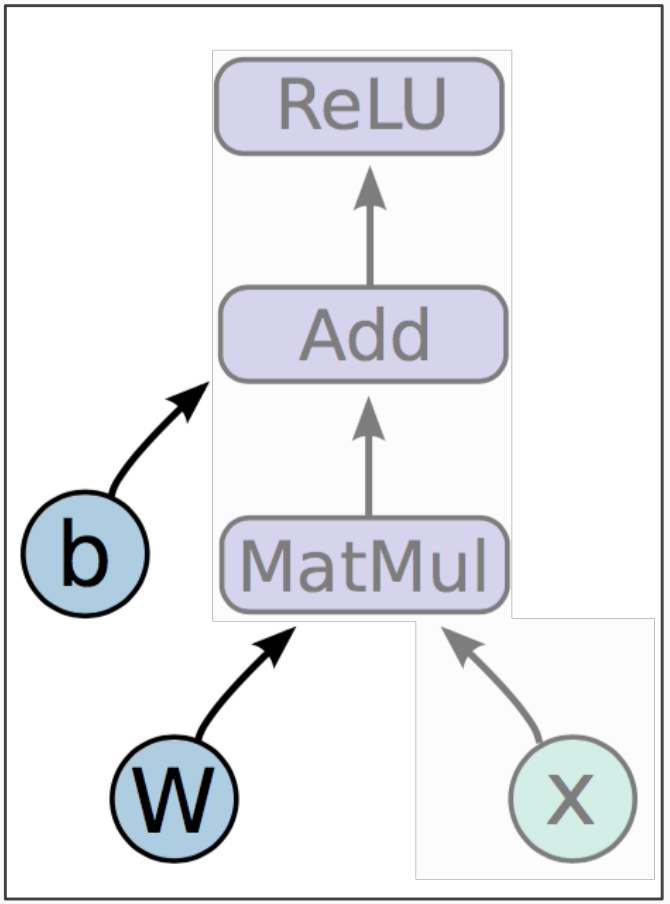
Programming model: NN feedforward

$$h_i = \mathrm{ReLU}(Wx + b)$$

Variables are 0-ary stateful nodes which output their current value.

(State is retained across multiple executions of a graph.)

(parameters, gradient stores, eligibility traces, …)

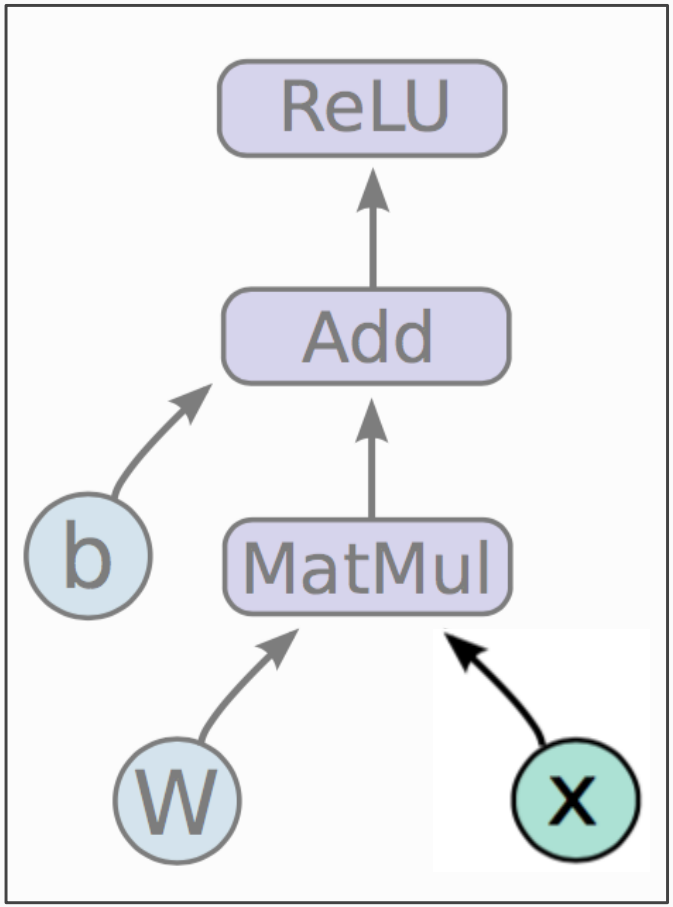


Programming model: NN feedforward

$$h_i = \mathrm{ReLU}(Wx + b)$$

Placeholders are 0-ary nodes whose value is fed in at execution time.

(inputs, variable learning rates, …)



Programming model: NN feedforward

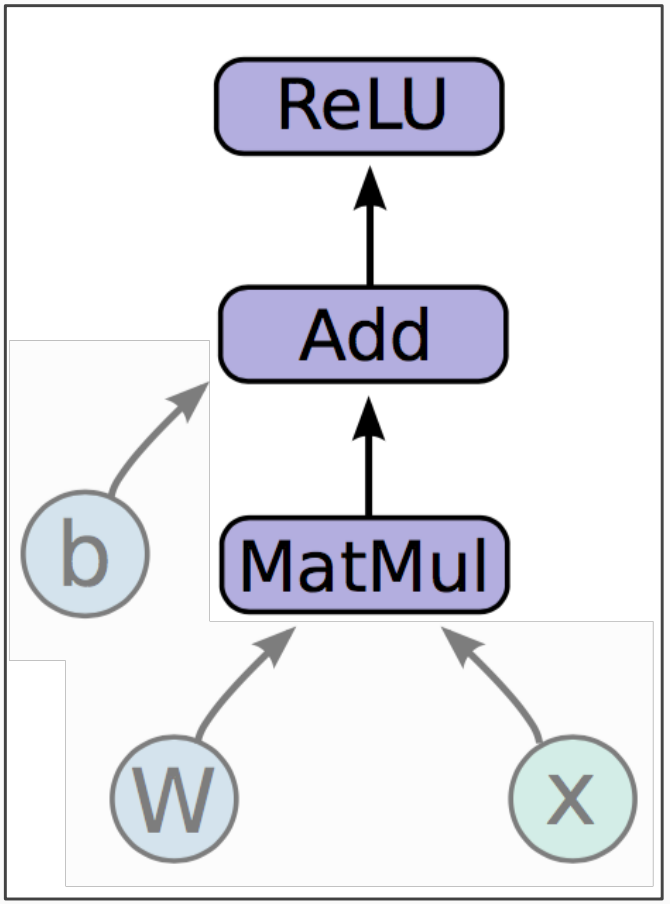
$$h_i = \mathrm{ReLU}(Wx + b)$$

Mathematical operations:

MatMul: Multiply two matrix values.

Add: Add elementwise (with broadcasting).

ReLU: Activate with elementwise rectified linear function.

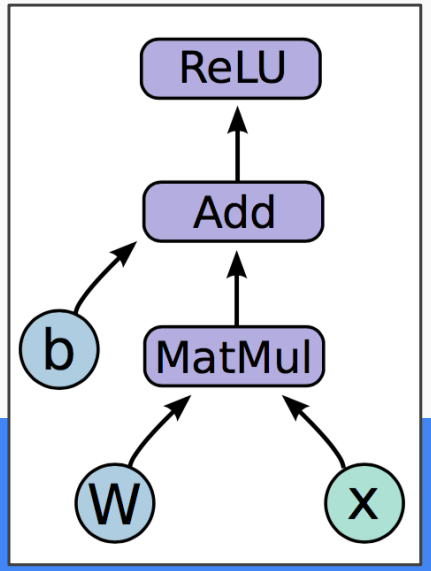


In code, please!

1. Create model weights, including initialization
    a. `W` ~ *Uniform*(-1, 1); `b` = **0**
2. Create input placeholder `x`
    a. *m* * 784 input matrix
3. Create computation graph

```
import tensorflow as tf

1 b = tf.Variable(tf.zeros((100,)))
  W = tf.Variable(tf.random_uniform((784,
  100), -1, 1))
2 x = tf.placeholder(tf.float32, (None,
  784))
3 h_i = tf.nn.relu(tf.matmul(x, W) + b)
```

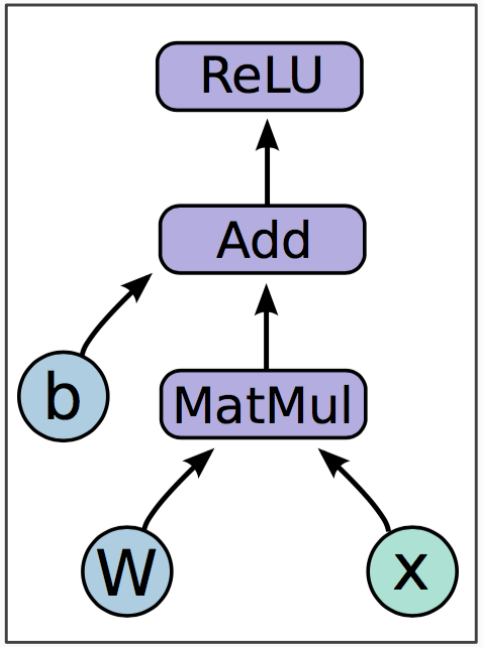
$$h_i = \mathrm{ReLU}(Wx + b)$$



How do we run it?

So far we have defined a **graph**.

We can deploy this graph with a **session**: a binding to a particular execution context (e.g. CPU, GPU)



Getting output

`sess.run(fetches, feeds)`

Fetches: List of graph nodes. Return the outputs of these nodes.

Feeds: Dictionary mapping from graph nodes to concrete values. Specifies the value of each graph node given in the dictionary.

```
import numpy as np
import tensorflow as tf

1 b = tf.Variable(tf.zeros((100,)))
  W = tf.Variable(tf.random_uniform((784,
  100) -1, 1))
2 x = tf.placeholder(tf.float32, (None, 784))
3 h_i = tf.nn.relu(tf.matmul(x, W) + b)


sess = tf.Session()
sess.run(tf.initialize_all_variables())
sess.run(h_i, {x: np.random.random(64,
784)})
```

Basic flow

1. Build a graph
    a. Graph contains parameter specifications, model architecture, optimization process, …
    b. Somewhere between 5 and 5000 lines
2. Initialize a session
3. Fetch and feed data with `Session.run`
    a. Compilation, optimization, etc. happens at this step — you probably won't notice

FIN - Thank You