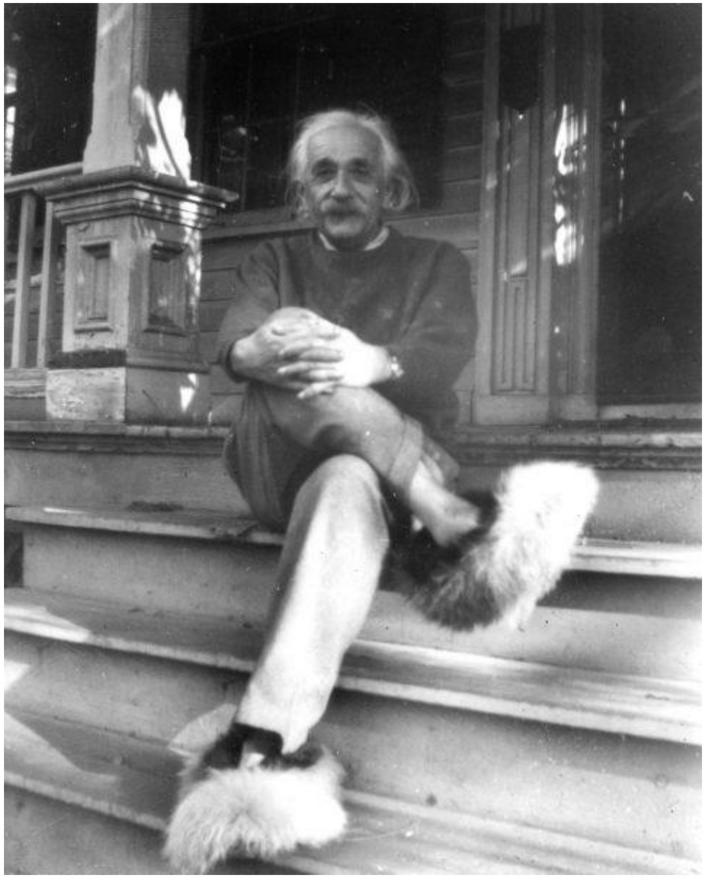
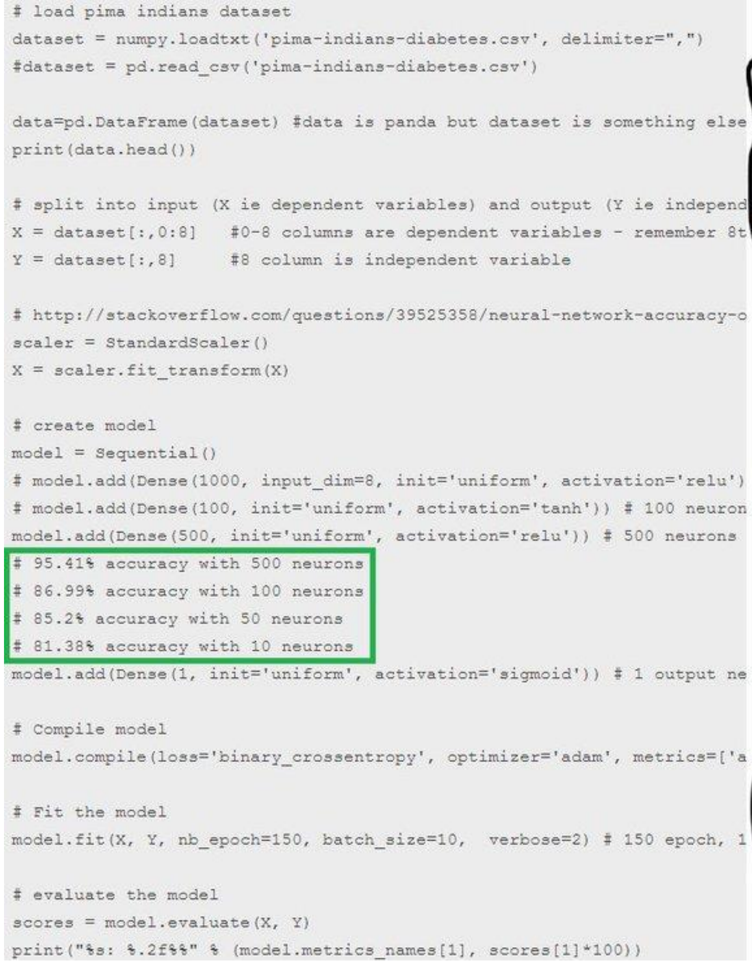
Research in ML and CSC2515 projects

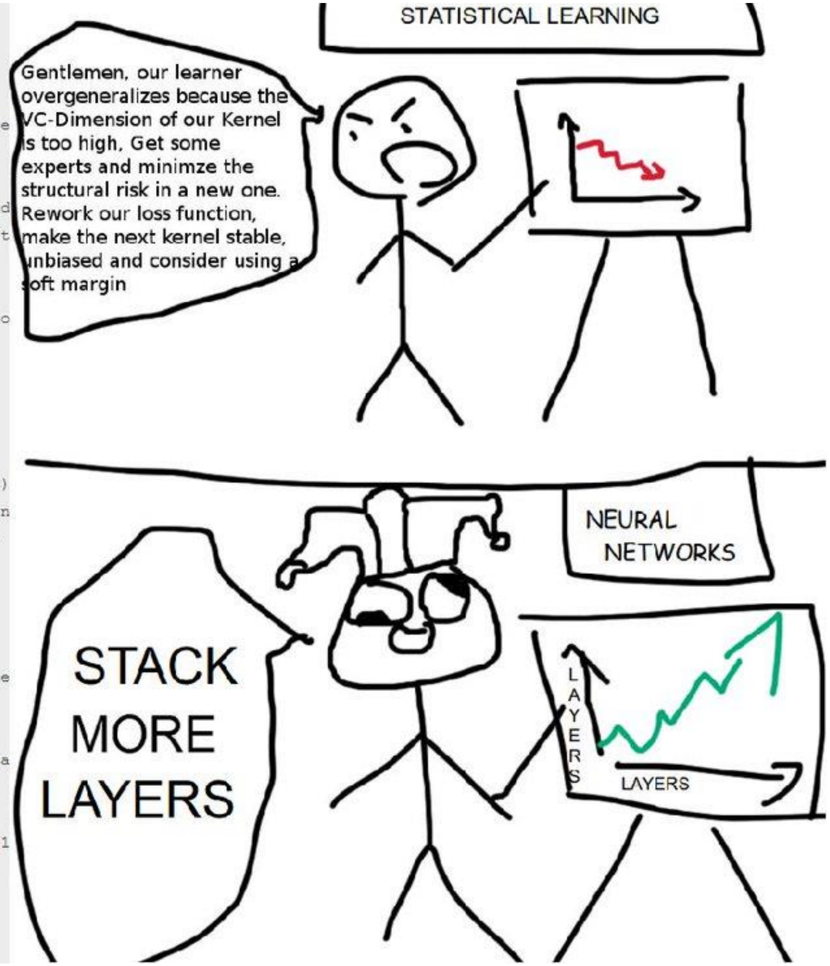


CSC2515: Machine Learning and Data Mining, Winter 2018

Michael Guerzhoy and Lisa Zhang

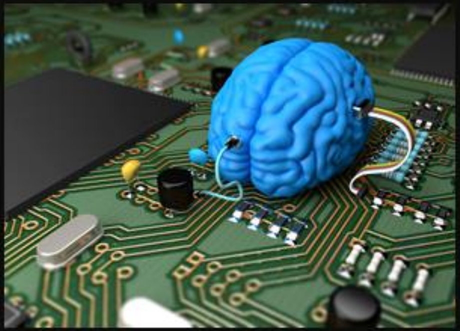


```
# load pima indians dataset
dataset = numpy.loadtxt('pima-indians-diabetes.csv', delimiter=",")
#dataset = pd.read_csv('pima-indians-diabetes.csv')

data=pd.DataFrame(dataset) #data is panda but dataset is something else
print(data.head())

# split into input (X ie dependent variables) and output (Y ie independ
X = dataset[:,0:8]   #0-8 columns are dependent variables - remember 8t
Y = dataset[:,8]     #8 column is independent variable

# http://stackoverflow.com/questions/39525358/neural-network-accuracy-o
scaler = StandardScaler()
X = scaler.fit_transform(X)

# create model
model = Sequential()
# model.add(Dense(1000, input_dim=8, init='uniform', activation='relu')
# model.add(Dense(100, init='uniform', activation='tanh')) # 100 neuron
model.add(Dense(500, init='uniform', activation='relu')) # 500 neurons
# 95.41% accuracy with 500 neurons
# 86.99% accuracy with 100 neurons
# 85.2% accuracy with 50 neurons
# 81.38% accuracy with 10 neurons
model.add(Dense(1, init='uniform', activation='sigmoid')) # 1 output ne

# Compile model
model.compile(loss='binary_crossentropy', optimizer='adam', metrics=['a

# Fit the model
model.fit(X, Y, nb_epoch=150, batch_size=10,  verbose=2) # 150 epoch, 1

# evaluate the model
scores = model.evaluate(X, Y)
print("%s: %.2f%%" % (model.metrics_names[1], scores[1]*100))
```


STATISTICAL LEARNING
Gentlemen, our learner overgeneralizes because the VC-Dimension of our Kernel is too high, Get some experts and minimze the structural risk in a new one. Rework our loss function, make the next kernel stable, unbiased and consider using a soft margin
NEURAL NETWORKS
STACK MORE LAYERS
LAYERS
LAYERS

Deep Learning

What society thinks I do

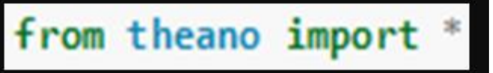


What my friends think I do

What other computer scientists think I do

What mathematicians think I do

What I think I do



```
from theano import *
```

What I actually do

CSC2515 projects

- Ideally: you make a project that turns into an academic paper or a start-up, and get an A+
    - There is not expectation that you produce anything publishable
    - If you are close, we'll help you try to get there in the summer
- Our hope: you make a project that follows the spec, maybe shows some small amount of original insight, and get an A+

Research in Computer Science

- Most research gets published in conference papers
    - Papers that are accepted are published in conferences proceedings, and the authors go to the conferences to present the papers
- Each area of CS has 2-4 “major” conferences where the top papers get published, and dozens of minor conferences
    - Major ML conferences: NIPS, ICML, ICLR, KDD
    - Major Computer Vision conferences: CVPR, ICCV, ECCV
    - Major Computational Linguistics conferences: ACL, NAACL, Interspeech, COLING
    - AI conference: AAAI

Other conferences

- National conferences
    - Canadian AI, Canadian Computer and Robot Vision
    - European […]
    - …
    - Generally respectable, but not nearly as important as the major conferences. Much easier to publish in
- Niche Conferences
    - International Conference on Medical Image Computing and Computer Assisted Intervention
    - …
    - Perfectly OK, generally not as competitive

ArXiV

- A recent trend is for ML researchers to post their papers on arxiv.org
- Andrej Karpathy’s Arxiv Sanity Preserver:
    - http://www.arxiv-sanity.com/top

Reading papers

- Google the conference name + year
- Enter the name of the papers your interested in into Google Scholar
- Get the pdf
- If the paper is paywalled and you are off campus, transform

https://link.springer.com/chapter/10.1007/978-3-319-06483-3_12 ->

https://link.springer.com.myaccess.library.utoronto.ca/chapter/10.1007/978-3-319-06483-3_12

Reading papers

- To get papers on your topic, you need to know the standard terminology for what you're trying to do
    - Ask us if you don't know!
- Find a "survey paper" or a well-cited paper on Google Scholar, and see which papers it cites
- Look which papers cite the paper that you are reading as well (you can check that on Google Scholar)

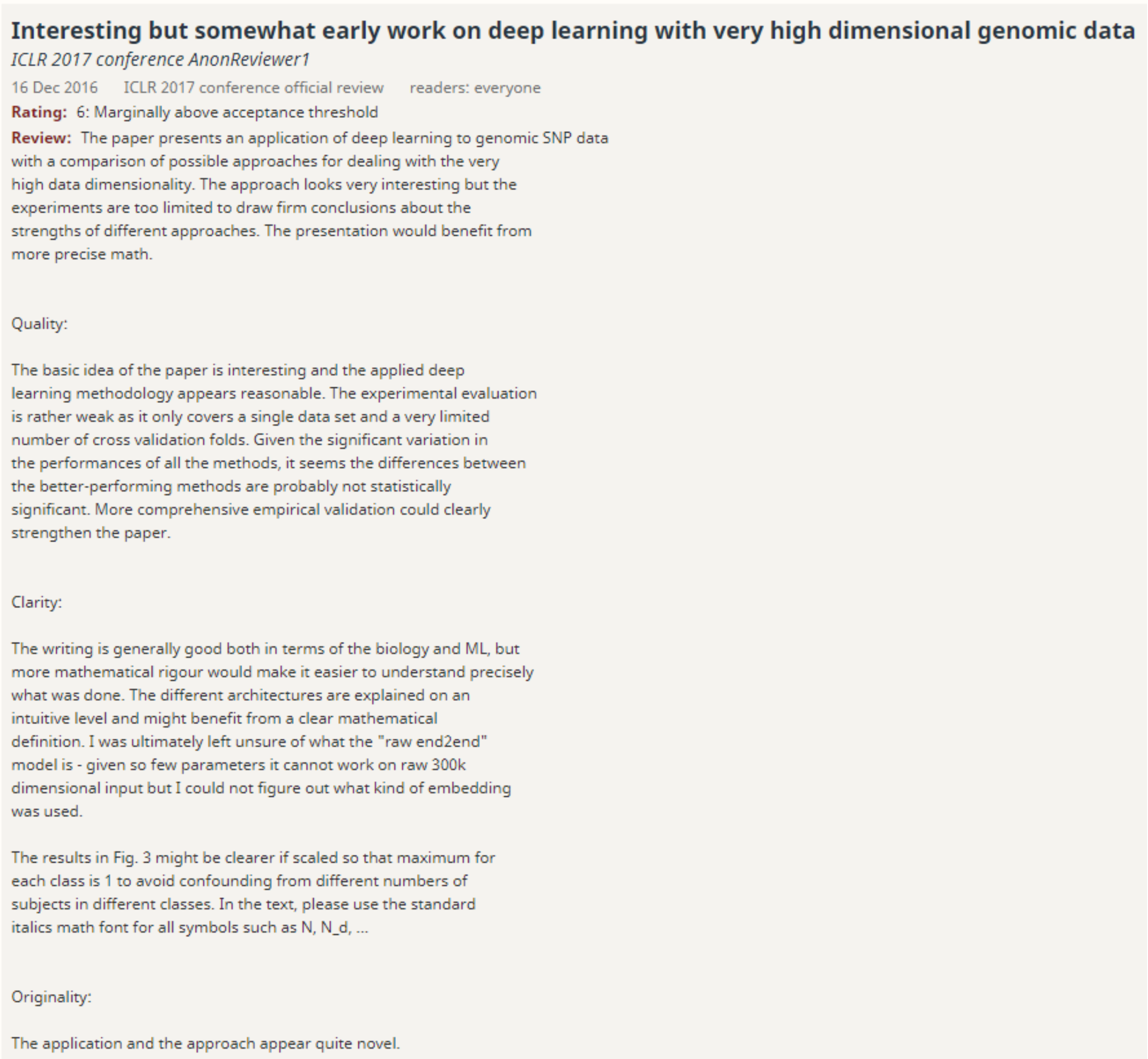
Suggested reading

- Top papers:
  - Start with NIPS 2017 or CVPR 2017
- OK papers (mostly)
  - Start with Canadian AI 2017, look at the "Applications of AI" papers

Conference review process

- Usually ~3 peer-reviewers get assigned to review each paper
- The conference's program committee/area chairs are responsible for selecting the papers to accept, based on the feedback of the peer reviewers
- ICLR's process is (unusually) open – you can see all the reviews and get an inside look into the process
- https://openreview.net/group?id=ICLR.cc/2017/conference

Conference reviews



Interesting but somewhat early work on deep learning with very high dimensional genomic data

ICLR 2017 conference AnonReviewer1

16 Dec 2016 ICLR 2017 conference official review readers: everyone

Rating: 6: Marginally above acceptance threshold

Review: The paper presents an application of deep learning to genomic SNP data with a comparison of possible approaches for dealing with the very high data dimensionality. The approach looks very interesting but the experiments are too limited to draw firm conclusions about the strengths of different approaches. The presentation would benefit from more precise math.

Quality:

The basic idea of the paper is interesting and the applied deep learning methodology appears reasonable. The experimental evaluation is rather weak as it only covers a single data set and a very limited number of cross validation folds. Given the significant variation in the performances of all the methods, it seems the differences between the better-performing methods are probably not statistically significant. More comprehensive empirical validation could clearly strengthen the paper.

Clarity:

The writing is generally good both in terms of the biology and ML, but more mathematical rigour would make it easier to understand precisely what was done. The different architectures are explained on an intuitive level and might benefit from a clear mathematical definition. I was ultimately left unsure of what the "raw end2end" model is - given so few parameters it cannot work on raw 300k dimensional input but I could not figure out what kind of embedding was used.

The results in Fig. 3 might be clearer if scaled so that maximum for each class is 1 to avoid confounding from different numbers of subjects in different classes. In the text, please use the standard italics math font for all symbols such as N, N_d, ...

Originality:

The application and the approach appear quite novel.

Conference Reviews

Dear Mr. Turing,

We regret to inform you that your submission

"On Computable Numbers, With an Application to the Entscheidungsproblem"

was not accepted to appear in FOCS 1936. The Program Committee received a record 4 submissions this year, many of them of high quality, and scheduling constraints unfortunately made it impossible to accept all of them.

Below please find some reviews on your submission. The reviews are *not* intended as an explanation for why your paper was rejected. This decision depended on many factors, including discussions at the PC meeting and competition from other papers.

Best wishes,

FOCS 1936 Program Committee

https://www.scottaaronson.com/blog/?p=253

Conference Reviews

```
---------------------------------------- review 1 ---------------------------------------
-
```

seems like a trivial modification of godel's result from STOC'31

```
---------------------------------------- review 2 ---------------------------------------
-
```

The author shows that Hilbert's Entscheidungsproblem (given a mathematical statement, decide whether it admits a formal proof) is unsolvable by any finite means. While this seems like an important result, I have several concerns/criticisms:

1. The author defines a new "Turing machine" model for the specific purpose of proving his result. This model was not defined in any previous papers; thus, the motivation is unclear.

2. I doubt Hilbert's goal of "automating mathematical thought" was ever really taken seriously by anyone (including Hilbert himself). Given this, the negative result comes as no surprise -- a positive result would have been much more interesting.

3. It's hard to find any technical "meat" in this paper. Once the author sets up the problem, the main result follows immediately by a standard diagonalization argument.

4. The whole philosophical discussion in Section 9, about what it means to compute something, is out of place (even slightly embarrassing) and should be deleted entirely.

Summary: While this paper deserves to be published somewhere -- SODA? ICALP? FSTTCS? -- it certainly isn't FOCS caliber.

What makes for a really good paper?

- Any one of
  - A really new idea for an algorithm
  - A new idea for an algorithm that works really well
  - A very carefully-engineered and rigorously evaluated system that uses a combination of recent ideas to obtain state-of-the-art results on datasets that everyone is working on
  - A new application area no one has thought of
- Clear introduction and conclusion that relate the idea to work by other researchers

What makes for an OK paper?

- A partially-failed attempt at a really good paper
- An idea that's very similar to what a good researcher in the field could tell you would probably work, applied to a dataset that's not really novel, with good experiments comparing the new idea to reasonable baselines (algorithms that have previously been used for similar datasets), with a reasonable overview of has been done in the field before, with reasonably good experimental results
- An interesting application of a variation of a well-known method
  - http://www.cs.toronto.edu/~guerzhoy/humantravel/
- An interesting analysis of a well-known method
  - http://www.cs.toronto.edu/~guerzhoy/oriviz/crv17.pdf

What makes for a good “side-project”?

- Something that works and does something interesting that looks difficult
- Good clean code

CS231n projects

- Line Drawing Colorization
  - http://cs231n.stanford.edu/reports/2017/pdfs/425.pdf

- Classifying U.S. Houses by Architectural Style Using Convolutional Neural Networks
  - http://cs231n.stanford.edu/reports/2017/pdfs/126.pdf

Canadian AI paper

- Investigating Citation Linkage with Machine Learning
  - https://link-springer-com.myaccess.library.utoronto.ca/chapter/10.1007/978-3-319-57351-9_10

Tips

- If you want to implement a twist on a relatively new method, look for code that does similar things, modify it, and acknowledge (and acknowledge it)
  - Don’t reinvent the wheel
    - Unless you want your project title is “Deep Convolutional Wheels”
- It is fine to work on non-cutting-edge topics
- If you have an idea for a new application/new dataset, work on that
- Talk to us/Google about standard datasets to work with
  - Images: ImageNet, PASCAL, …
  - Medical text: i2b2
  - …

Logistics

- Submit proposal earlier => get feedback earlier (hopefully)
- TA and instructor office hours