Lecture 24. Revision

(the content of this deck is non-examinable)

COMP90051 Statistical Machine Learning

Semester 2, 2016
Lecturers: Trevor Cohn, WHard, brainRover, Andrey Kan

This lecture

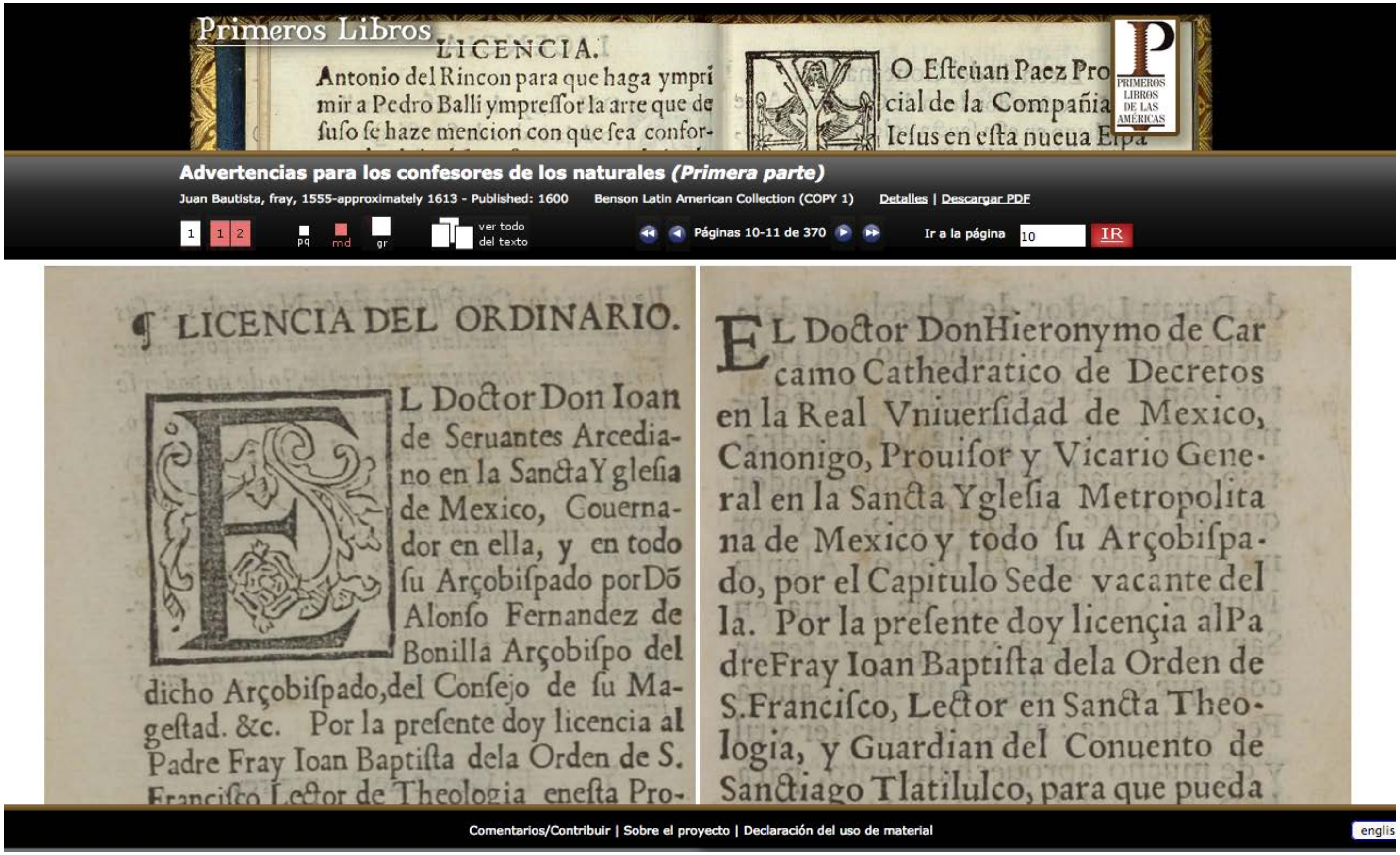
- Project notes and team presentations
- Exam tips
- A deeper insight
- Reflections on the subject
- Q&A session / office hour

Project Notes and Team Presentations

Well done every team!

Project Data: Primeros Libros

Texts printed in Mexico in the 1500s

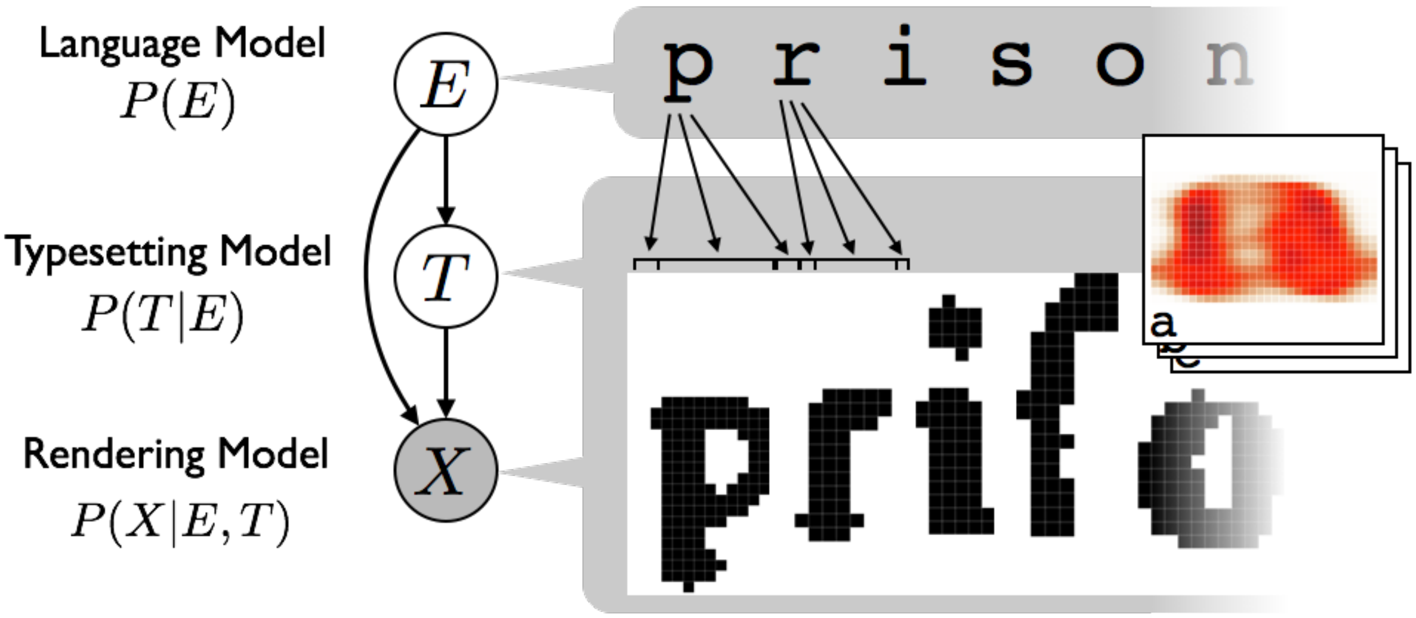


Data & slides adapted from: **Dan Garrette, H. Alpert-Abrams, T. Berg-Kirkpatrick & D. Klein** (2013, 2015, 2016)

Books used

Gante	(1553)	motlacatilia:ynica ſacramento Baptiſ-
Sahagún	(1583)	Yoan oquilhui in Em̃perador, in tlaça
Rincón	(1595)	ction.v.g.tetlaçotlaliztli.amatio, vel,
Bautista	(1600)	ſtimo, hæc ſupra dictus doctor Medina. Mas

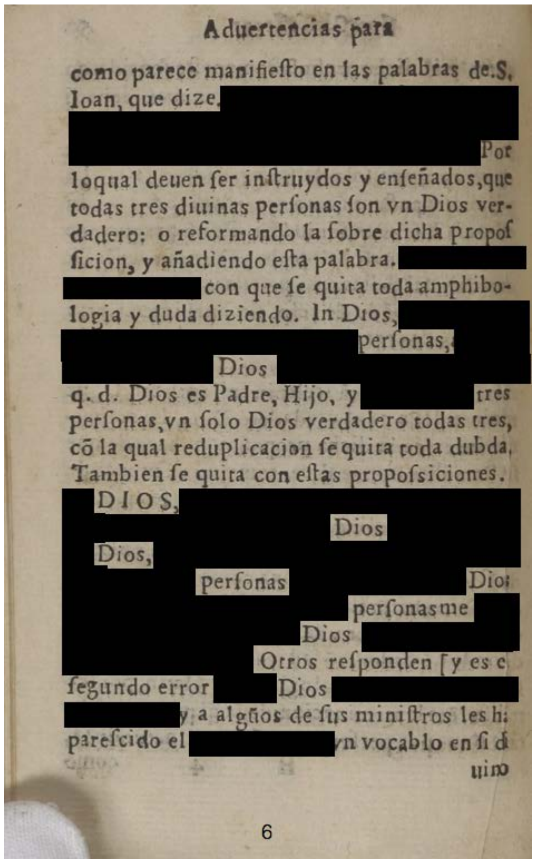
Overview: Ocular OCR system (Berg-Kirkpatrick et al, 2013)

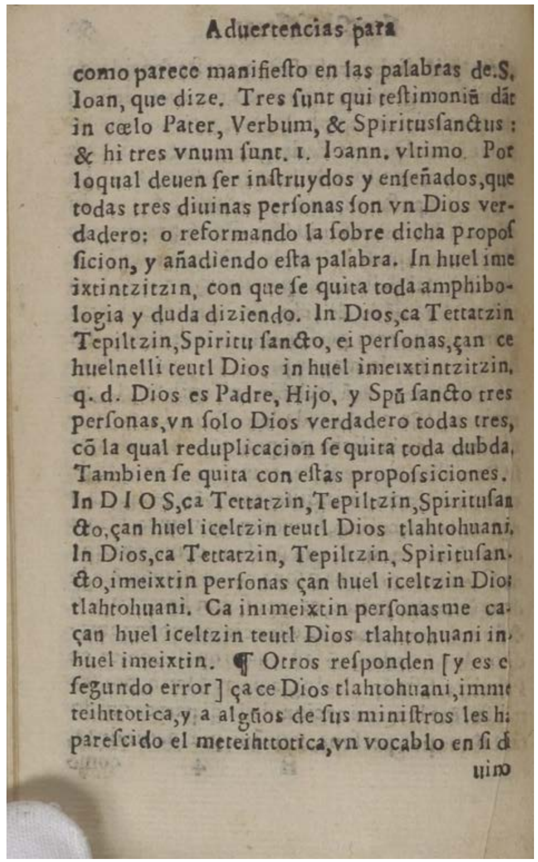


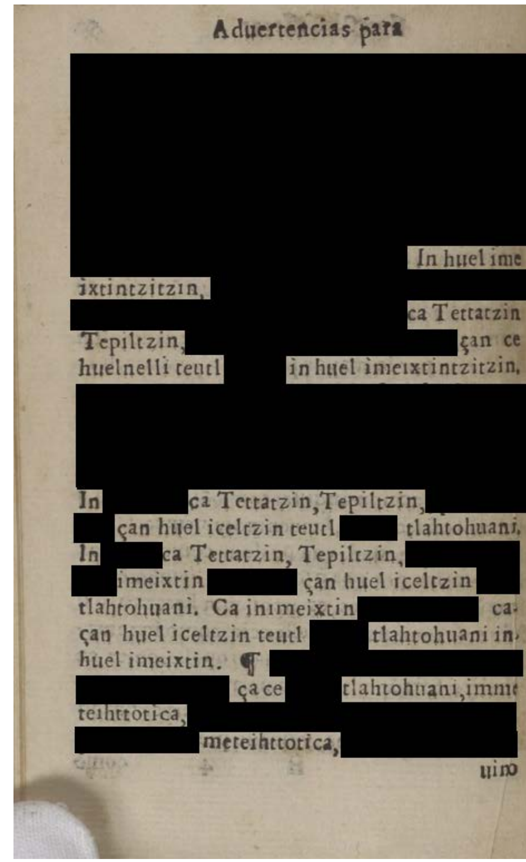
Observe ONLY the image; infer the font and rendering components using EM algorithm.

(Also know likely character sequences in the language[s] from modern texts.)

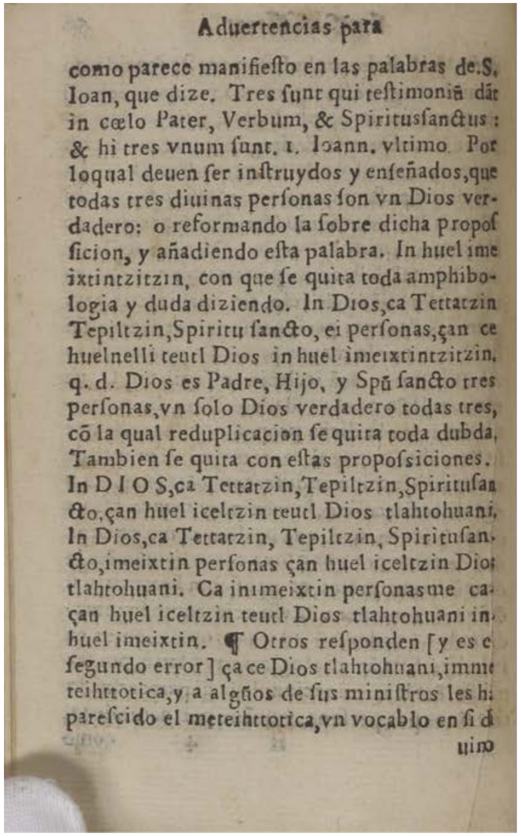
Multilingual texts: Spanish, Latin & Nahuatl


Aduertencias para
como parece manifieſto en las palabras de S.
Ioan, que dize. [redacted] Por
loqual deuen ſer inſtruydos y enſeñados,que
todas tres diuinas perſonas ſon vn Dios ver-
dadero: o reformando la ſobre dicha propoſ
ficion, y añadiendo eſta palabra. [redacted] con que ſe quita toda amphibo-
logia y duda diziendo. In Dios, [redacted] perſonas, [redacted] Dios [redacted]
q. d. Dios es Padre, Hijo, y [redacted] tres
perſonas,vn ſolo Dios verdadero todas tres,
cõ la qual reduplicacion ſe quita toda dubda,
Tambien ſe quita con eſtas propoſſiciones.
[redacted] D I O S, [redacted] Dios [redacted] Dios, [redacted] perſonas [redacted] Dios [redacted] perſonasme [redacted] Dios [redacted] Otros reſponden [y es e
ſegundo error [redacted] Dios [redacted] y a algũos de ſus miniſtros les ha
pareſcido el [redacted] vn vocablo en ſi d
uino


Aduertencias para
como parece manifieſto en las palabras de S.
Ioan, que dize. Tres ſunt qui teſtimoniũ dãt
in cœlo Pater, Verbum, & Spiritusſanctus:
& hi tres vnum ſunt. 1. Ioann. vltimo. Por
loqual deuen ſer inſtruydos y enſeñados,que
todas tres diuinas perſonas ſon vn Dios ver-
dadero: o reformando la ſobre dicha propoſ
ficion, y añadiendo eſta palabra. In huel ime
ixtintzitzin, con que ſe quita toda amphibo-
logia y duda diziendo. In Dios,ca Tettatzin
Tepiltzin,Spiritu ſancto, ei perſonas,çan ce
huelnelli teutl Dios in huel imeixtintzitzin,
q. d. Dios es Padre, Hijo, y Spũ ſancto tres
perſonas,vn ſolo Dios verdadero todas tres,
cõ la qual reduplicacion ſe quita toda dubda,
Tambien ſe quita con eſtas propoſſiciones.
In D I O S,ca Tettatzin,Tepiltzin,Spirituſan
cto,çan huel iceltzin teutl Dios tlahtohuani,
In Dios,ca Tettatzin, Tepiltzin, Spirituſan.
cto,imeixtin perſonas çan huel iceltzin Dios
tlahtohuani. Ca inimeixtin perſonasme ca-
çan huel iceltzin teutl Dios tlahtohuani in-
huel imeixtin. ¶ Otros reſponden [y es e
ſegundo error] ça ce Dios tlahtohuani,imme
teihttotica,y a algũos de ſus miniſtros les ha
pareſcido el meteihttotica,vn vocablo en ſi d
uino


Aduertencias para
[redacted] In huel ime
ixtintzitzin, [redacted] ca Tettatzin
Tepiltzin, [redacted] çan ce
huelnelli teutl [redacted] in huel imeixtintzitzin,
[redacted] In [redacted] ca Tettatzin, Tepiltzin, [redacted]
[redacted] çan huel iceltzin teutl [redacted] tlahtohuani,
In [redacted] ca Tettatzin, Tepiltzin, [redacted]
[redacted] imeixtin [redacted] çan huel iceltzin [redacted]
tlahtohuani. Ca inimeixtin [redacted] ca-
çan huel iceltzin teutl [redacted] tlahtohuani in-
huel imeixtin. ¶ [redacted] ça ce [redacted] tlahtohuani,imme
teihttotica, [redacted] meteihttotica, [redacted]
uino

Transcriptions: their evaluation vs. our training


Aduertencias para

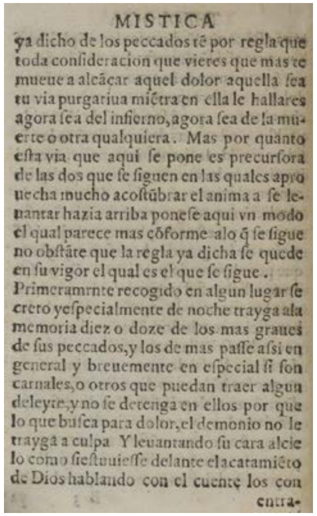
como parece manifiefto en las palabras de S.
Ioan, que dize. Tres funt qui teftimoniũ dãt
in cœlo Pater, Verbum, & Spiritusfanctus:
& hi tres vnum funt. 1. Ioann. vltimo. Por
loqual deuen fer inftruydos y enfeñados, que
todas tres diuinas perfonas fon vn Dios ver-
dadero: o reformando la fobre dicha propof
ficion, y añadiendo efta palabra. In huel ime
ixtintzitzin, con que fe quita toda amphibo-
logia y duda diziendo. In Dios, ca Tettatzin
Tepiltzin, Spiritu fancto, ei perfonas, çan ce
huelnelli teutl Dios in huel imeixtintzitzin,
q. d. Dios es Padre, Hijo, y Spũ fancto tres
perfonas, vn folo Dios verdadero todas tres,
cõ la qual reduplicacion fe quita toda dubda.
Tambien fe quita con eftas propofsiciones.
In DIOS, ca Tettatzin, Tepiltzin, Spiritufan
cto, çan huel iceltzin teutl Dios tlahtohuani.
In Dios, ca Tettatzin, Tepiltzin, Spiritufan.
cto, imeixtin perfonas çan huel iceltzin Dios
tlahtohuani. Ca inimeixtin perfonasme ca-
çan huel iceltzin teutl Dios tlahtohuani in-
huel imeixtin. ¶ Otros refponden [y es e
fegundo error] ça ce Dios tlahtohuani, imme
teihttotica, y a algũos de fus miniftros les h
parefcido el meteihttotica, vn vocablo en fi d
uino

Aduertencias para

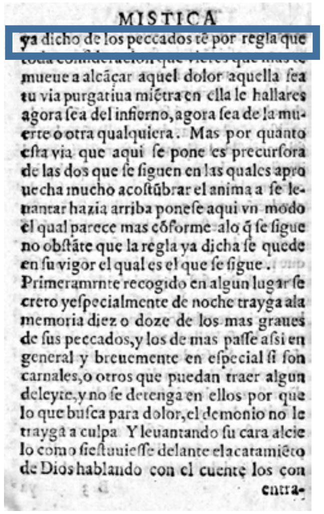
como parece manifiefto en las palabras de S.
Ioan, que dize. Tres funt qui teftimoniu dãt
in cœlo Pater, Verbum, & Spiritus fanctus:
& hi tres vnum funt. 1. Ioann. vltimo. Por
lo qual deuen fer inftruydos y enfeñados, que
todas tres diuinas perfonas fon vn Dios ver-
dadero: o reformando la fobre dicha propof
ficion, y añadiendo efta palabra. In huel ime
ixtintzitzin, con que fe quita toda amphibo-
logia y duda diziendo. In Dios, ca Tettatzin
Tepiltzin, Spiritu fancto, ei perfonas, çan ce
huelnelli tuetl Dios, in huel imeixtintzitzin,
q. d. Dios es Padre, Hijo, y Spu fancto tres
perfonas, vn folo Dios verdadero todas tres,
cõ la qual reduplicacion fe quita toda dubda.
Tambien fe quita con eftas propofsiciones.
In DIOS, ca Tettatzin, Tepiltzin, Spiritu fan
cto, çan huel iceltzin teutl Dios tlahtohuani.
In Dios, ca Tettatzin, Tepiltzin, Spiritu fan
cto, imeixtin perfonas çan huel iceltzin Dios
tlahtohuanì. Ca inimeixtin perfonas me ca-
çan huel iceltzin teutl Dios tlahtohuani in-
huel imeixtin. ¶Otros refponden [y es e
fegundo error] ça ce Dios tlahtohuani, imme
teihttotica, y a alguos de fus miniftros les ha
parefcido el meteihttotica, vn vocablo en fi di
uino

Processing pipeline

1. Normalise colour
2. Extract lines
3. Apply tesseract OCR
4. Edit distance alignment
   to manual
   transcription
5. Extract example for
   (char, region from
   image)



MISTICA
ya dicho de los peccados tẽ por regla que
toda consideracion que vieres que mas te
mueue a alcãçar aquel dolor aquella sea
tu via purgatiua mie̅tra en ella le hallares
agora sea del infierno, agora sea de la mu-
erte o otra qualquiera. Mas por quanto
esta via que aqui se pone es precursora
de las dos que se siguen en las quales apro
uecha mucho acostũbrar el anima a se le-
uantar hazia arriba ponese aqui vn modo
el qual parece mas cõforme alo q̃ se sigue
no obstãte que la regla ya dicha se quede
en su vigor el qual es el que se sigue.
Primeramrnte recogido en algun lugar se
creto y especialmente de noche trayga ala
memoria diez o doze de los mas graues
de sus peccados, y los de mas passe asi en
general y breuemente en especial si son
carnales, o otros que puedan traer algun
deleyte, y no se detenga en ellos por que
lo que busca para dolor, el demonio no le
traygã a culpa Y leuantando su cara alcie
lo como si estuuiesse delante ela catamie̅to
de Dios hablando con el cuente los con
entra-




```
Ya dícHo de los Pcccados té por regla'Qé
```

```
Y 2 2 23 42 0
a 24 15 36 33 0
d 44 15 59 41 0
í 62 15 67 40 0
c 70 15 81 31 0
H 82 16 100 42 0
o 101 16 115 32 0
```

```
ya dicho de los peccados t\~e por regla que
```

Team presentations

- WHard
  - Ranked 3rd with the score of 0.83986
- brainRover
  - Ranked 2nd with the score of 0.84715

OUR SOLUTION

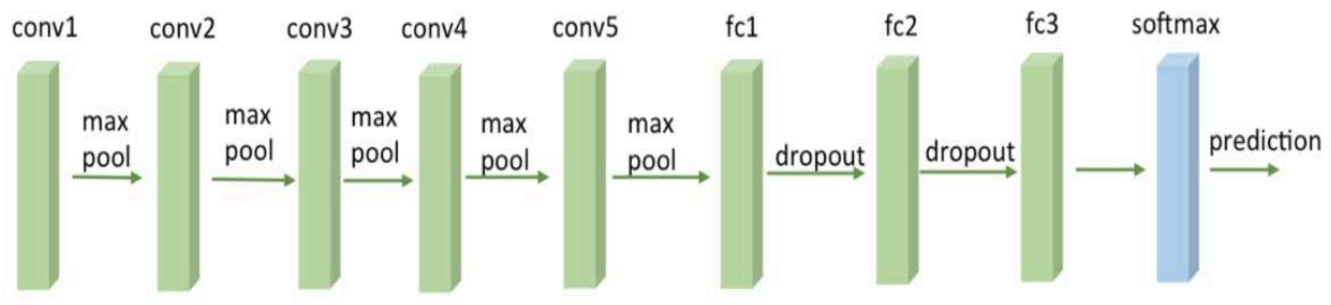
Team : Whard

Team members : Han Yu, Yiheng Wang, Xing Han

PREPROCESSING

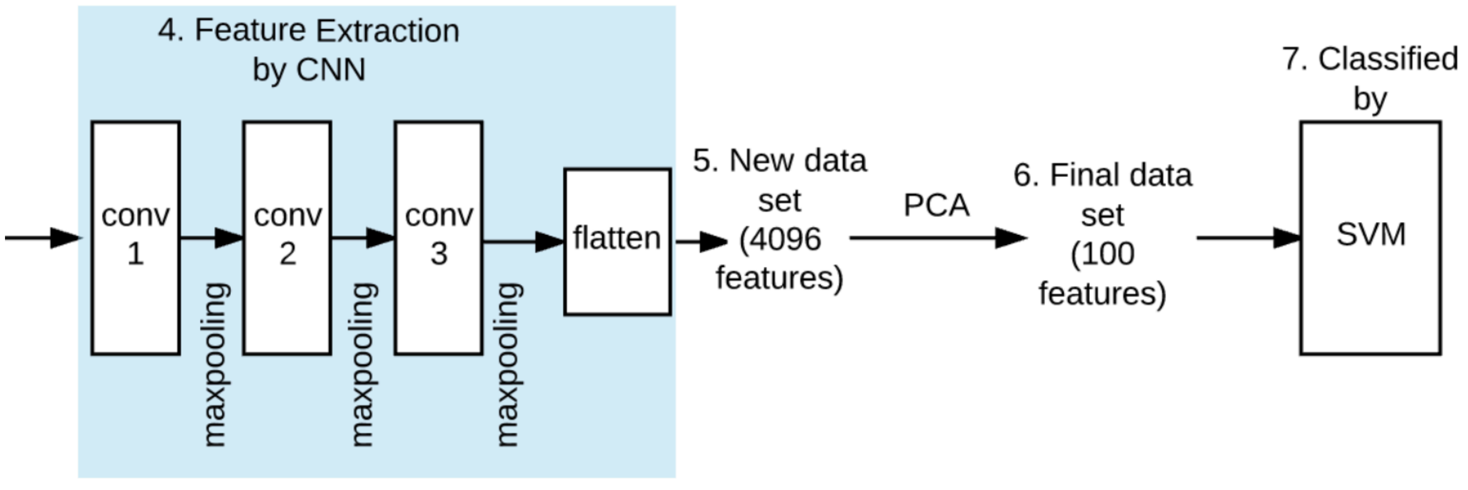
- padding image to 33*33
- move character to the center of image
- resize it to 32*32
- transform it to RGB

VGG16



- pretrained VGG16 cnn which trained on Image net dataset
- use as feature extractor
- remove layer after conv3

SVM + VGG16


4. Feature Extraction
by CNN
conv
1
maxpooling
conv
2
maxpooling
conv
3
maxpooling
flatten
5. New data
set
(4096
features)
PCA
6. Final data
set
(100
features)
7. Classified
by
SVM

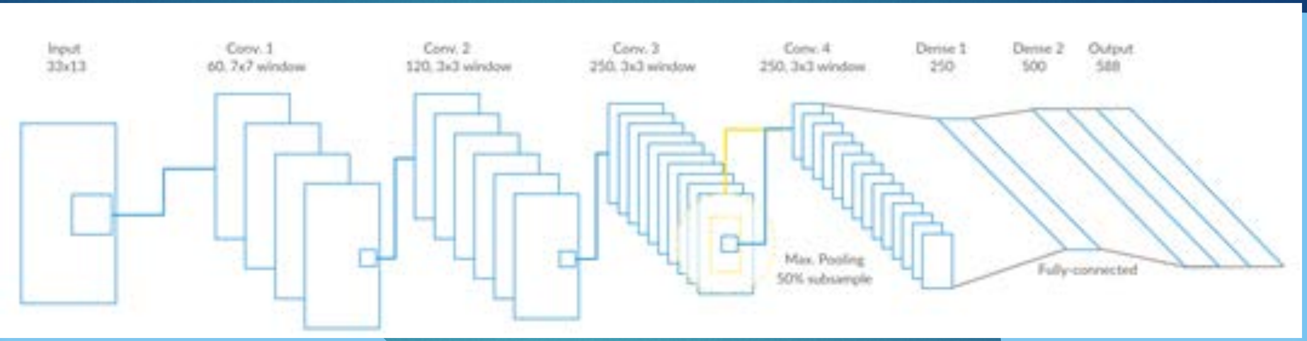
COMP90051: Statistical Machine Learning

***Project:** OCR for Historical Documents*
Notes on Team BrainRover's Approach

Preprocessing

1. Quality control and culling obvious errors
   - Particularly important for rare classes
2. Oversampling rare classes
3. Affine transformations using ImageDataGenerator in Keras
   - Create believable transformations to emulate unseen data
   - Random shear, zoom, rotation, width/height shift
   - Reduce overfitting by enlarging the dataset

Topology



1. Input
2. C1: 60, 7x7
3. C2: 120, 3x3
4. C3: 250, 3x3

   - Max. pooling

5. C4: 250, 3x3
6. D1: 250
7. D2: 500
8. Output (softmax): 588 (98x6)

- Rectified linear units (ReLUs)
- Dropout imposed on D1 and D2

Training

1. Loss function: Categorical cross-entropy
2. Optimisation: Adadelta
3. Samples/epoch: ~50,000, in batches of 800
4. Epochs: 60
5. GPU: NVIDIA GTX1070

Exam Tips

Don’t panic ☺

Exam tips

- Don’t panic
- Attempt all questions
  - Do your best guess whenever you don’t know the answer
- Finish easy questions first
- Quizzes and practice exam questions are representative of what you might get at the exam
- Make sure you understand solutions for each quiz and practice exam

What’s non-examinable?

- Green slides
- All of this deck
- Deck 7: Slides on Gaussian blur and Sobel kernel
- Deck 8: Slides on CNNs & RNNs (beyond motivation)
- Deck 16: Slides on LDA, HMMs, vision MRFs (beyond motivation)
- Something that was in workshops but not in lectures
- Note: material covered in the reading is fair-game

A Deeper Insight

A selection of additional topics with the aim to provide a deeper insight into main lectures content

Kernelised perceptron (1/3)

When classified correctly, weights are unchanged

When misclassified: $\boldsymbol{w}^{(k+1)} = -\eta(\pm\boldsymbol{x})$
($\eta > 0$ is called *learning rate)*

<u>If $y = 1$, but $s < 0$</u>
$w_i \leftarrow w_i + \eta x_i$
$w_0 \leftarrow w_0 + \eta$

<u>If $y = -1$, but $s \geq 0$</u>
$w_i \leftarrow w_i - \eta x_i$
$w_0 \leftarrow w_0 - \eta$

Suppose weights are initially set to 0

First update: $\boldsymbol{w} = \eta y_{i_1} \boldsymbol{x}_{i_1}$
Second update: $\boldsymbol{w} = \eta y_{i_1} \boldsymbol{x}_{i_1} + \eta y_{i_2} \boldsymbol{x}_{i_2}$
Third update $\boldsymbol{w} = \eta y_{i_1} \boldsymbol{x}_{i_1} + \eta y_{i_2} \boldsymbol{x}_{i_2} + \eta y_{i_3} \boldsymbol{x}_{i_3}$
etc.

Kernelised perceptron (2/3)

- Weights always take the form $\boldsymbol{w} = \sum_{i=1}^{n} \alpha_i y_i \boldsymbol{x}_i$, where $\boldsymbol{\alpha}$ some coefficients
- Perceptron weights are always a linear combination of data!
- Recall that prediction for a new point $\boldsymbol{x}$ is based on sign of $w_0 + \boldsymbol{w}'\boldsymbol{x}$
- Substituting $\boldsymbol{w}$ we get $w_0 + \sum_{i=1}^{n} \alpha_i y_i \boldsymbol{x}_i'\boldsymbol{x}$
- The dot product $\boldsymbol{x}_i'\boldsymbol{x}$ can be replaced with a kernel

Kernelised perceptron (3/3)

1. Initialisation: set $\boldsymbol{\alpha} = \mathbf{0}$
2. For each training example $\{\boldsymbol{x}_j, y_j\}$
   a) Make prediction based on $w_0 + \sum_{i=1}^{n} \alpha_i y_i \boldsymbol{x}_i' \boldsymbol{x}_j$
   b) If the example is misclassified, update $\alpha_j \leftarrow \alpha_j + 1$
3. Repeat Step 2 predefined number of times (number of epochs)

Networks in real life: the Internet

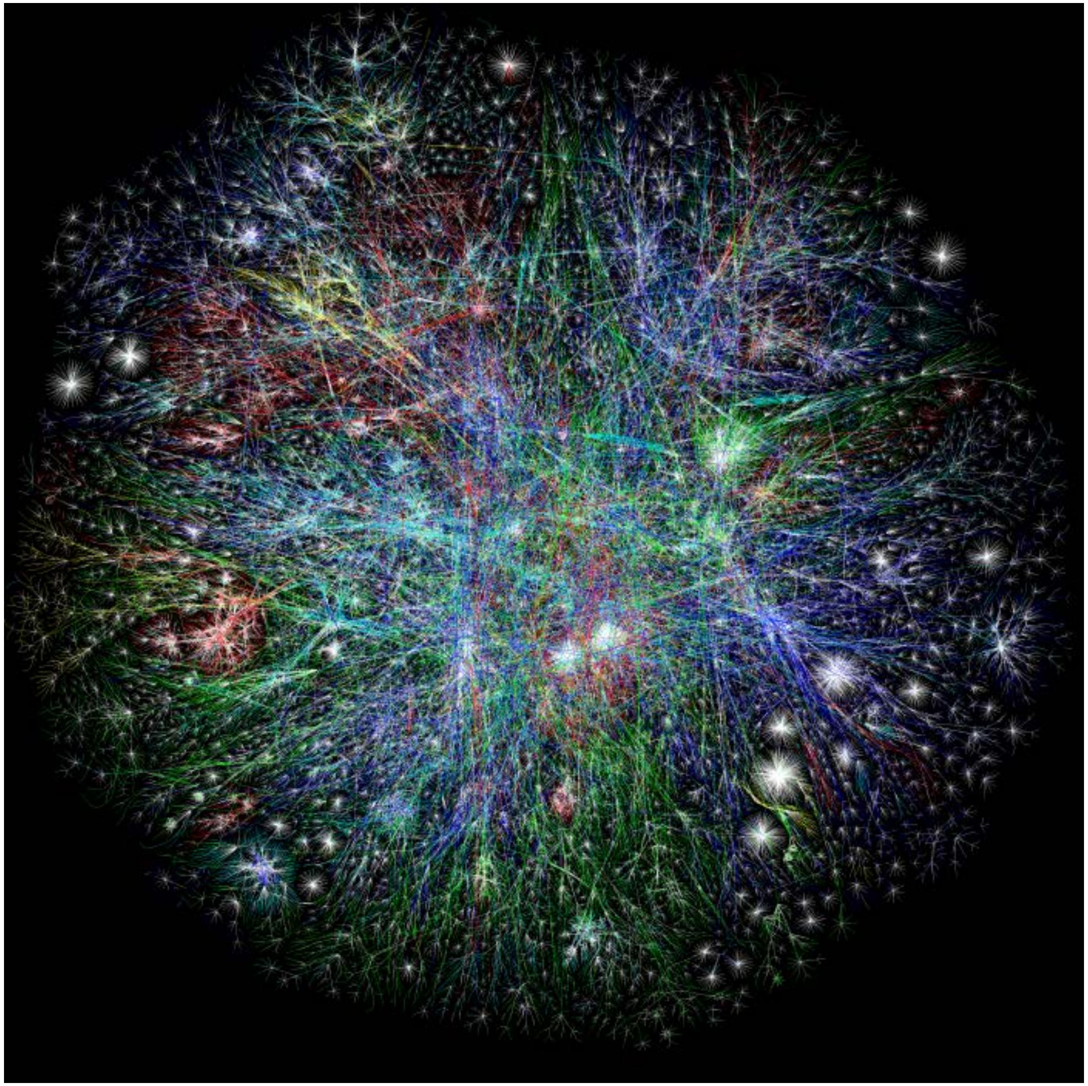
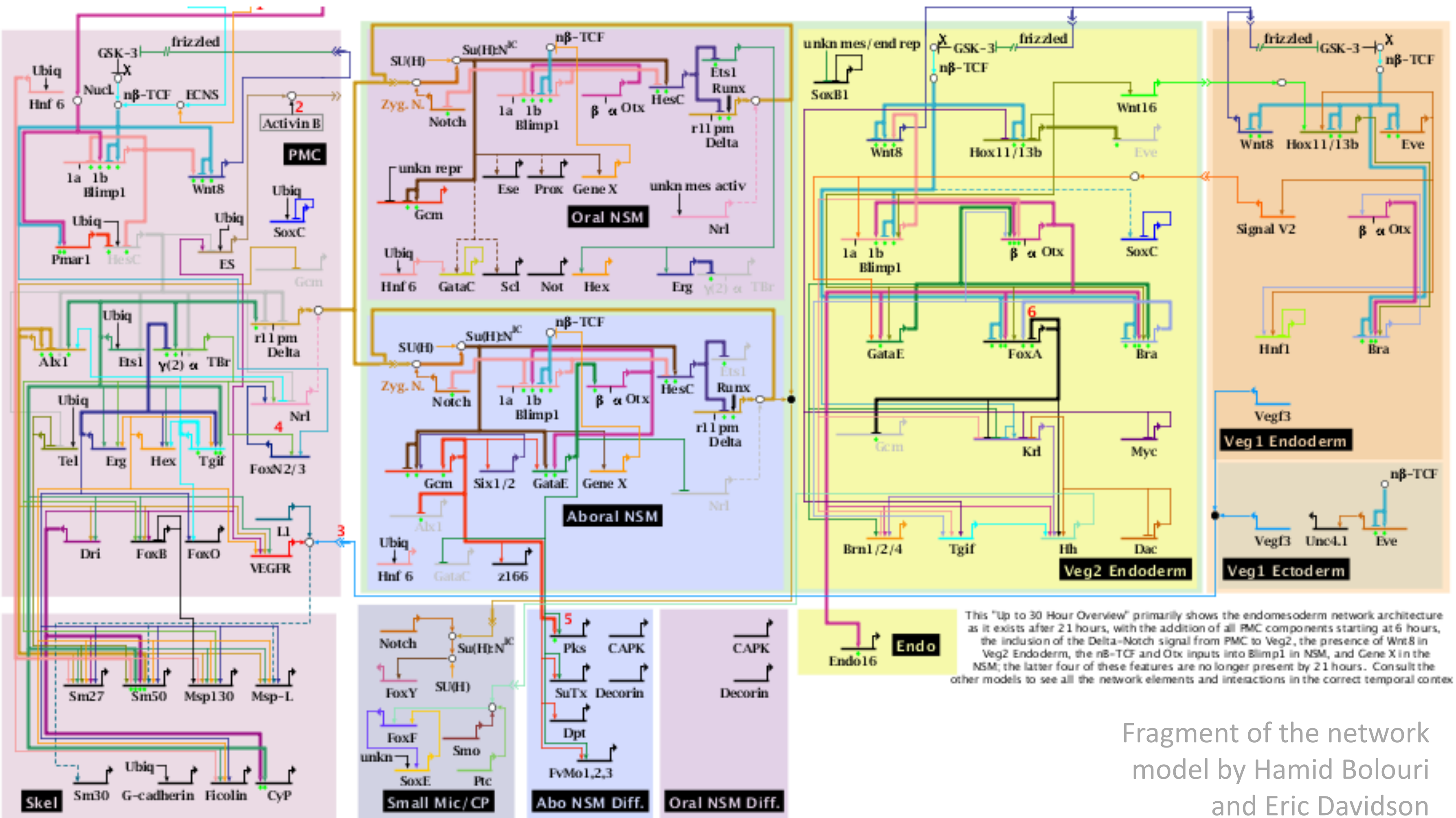


Image: OPTE Project Map (CC2)

Networks in real life: gene regulatory network



Fragment of the network model by Hamid Bolouri and Eric Davidson

Networks in real life: transport map

Network analysis (1/4)

- Analysis of large scale real world networks has recently attracted considerable attention from research and engineering communities
- Networks/graphs is a list of pairwise relations (edges) between a set of objects (vertices)
- Example problems / types of analysis
  - Link prediction
  - Identifying frequent subgraphs
  - Identifying influential vertices
  - Community finding

Network analysis (2/4)

- *Community is a group of vertices* that interact more frequently within its own group than to those outside the group
  - Families
  - Friend circles
  - Websites (communities of webpages)
  - Groups of proteins that maintain a specific function in a cell
- This is essentially a definition of a *cluster* in unsupervised learning

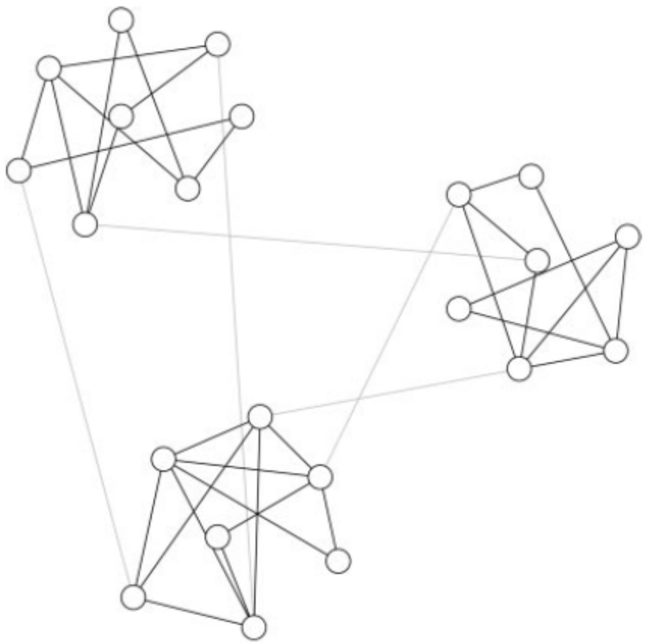


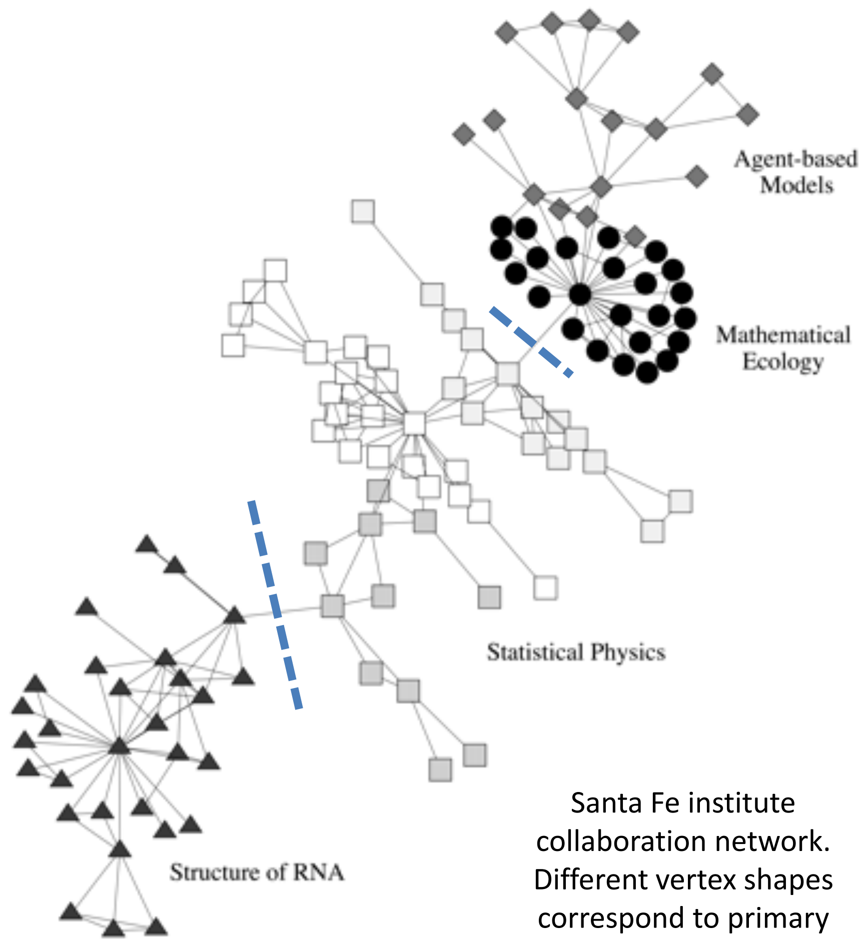
Image: Girvan and Newman, Community structure in social and biological networks, PNAS, 2002

Network analysis (3/4)

- Why community detection?
  - Understanding the system behind the network (e.g., structure of society)
  - Identifying roles of vertices (e.g., hubs, mediators)
  - Summary graphs (vertices – communities, edges – connections between communities)
  - Facilitate distributed computing (e.g., place data from the same community to the same server or core)
- There are many community detection algorithms, let's have a look at only one of the ideas

Network analysis (4/4)

- Communities are connected by a few connections, which tends to form *bridges*
- Cut the bridges to obtain communities
- One of the algorithms is called normalised cuts which is equivalent to spectral clustering



Santa Fe institute collaboration network. Different vertex shapes correspond to primary divisions of the institute

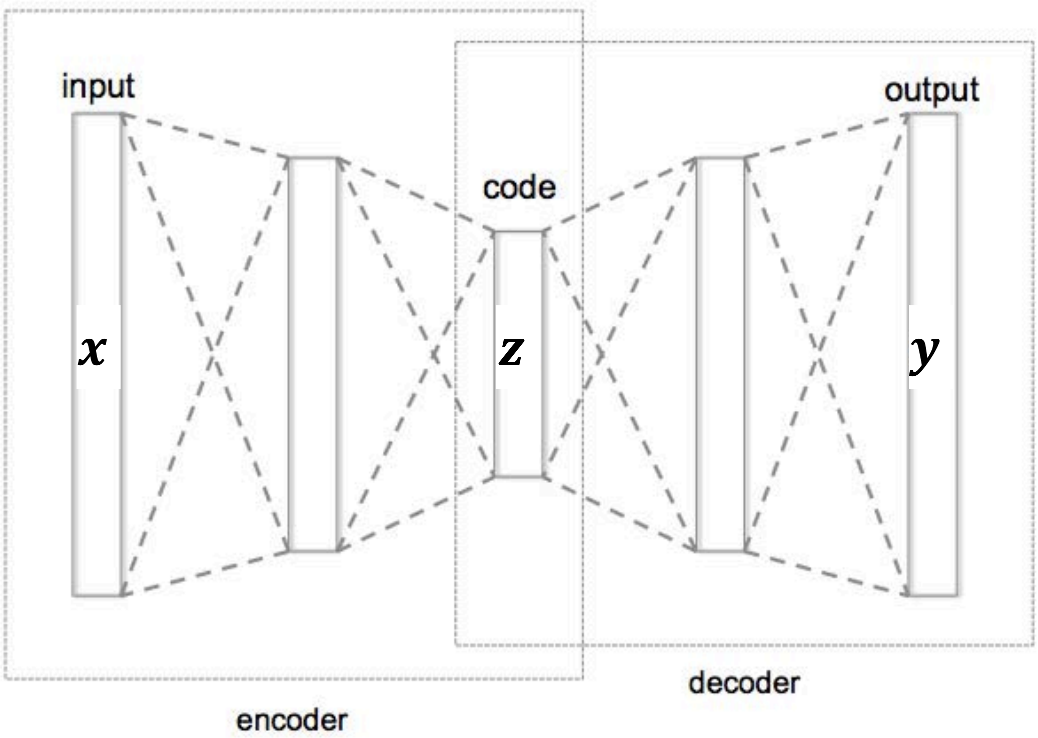
Image: Girvan and Newman, Community structure in social and biological networks, PNAS, 2002

Autoencoder: dimensionality reduction (1/4)

- Supervised learning:
  - Univariate regression: predict y from $\boldsymbol{x}$
  - Multivariate regression: predict $\boldsymbol{y}$ from $\boldsymbol{x}$
- Unsupervised learning: explore $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n$
- For each $\boldsymbol{x}_i$ set $\boldsymbol{y}_i \equiv \boldsymbol{x}_i$
- Train a feed forward ANN to predict $\boldsymbol{y}_i$ from $\boldsymbol{x}_i$
- Pointless?

Autoencoder: dimensionality reduction (2/4)

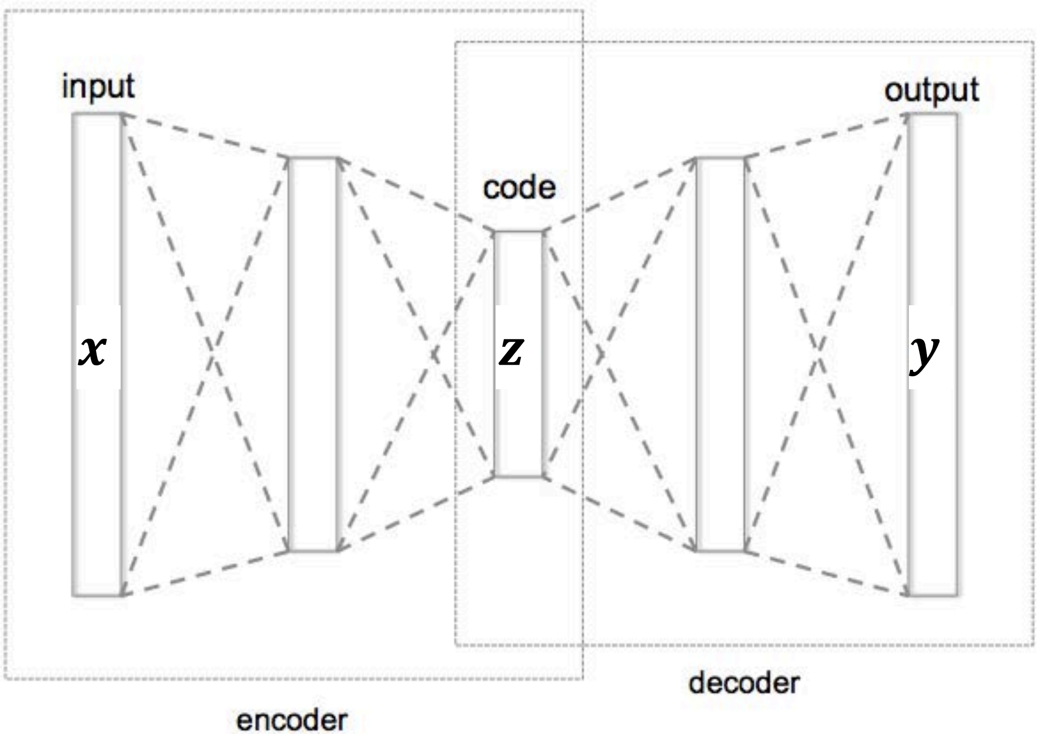
- For each $\boldsymbol{x}_i$ set $\boldsymbol{y}_i \equiv \boldsymbol{x}_i$ and train a feed forward ANN to predict $\boldsymbol{y}_i$ from $\boldsymbol{x}_i$
- Set the hidden layer $\boldsymbol{z}_j$ in the middle "thinner" than the input



adapted from: Chervinskii at Wikimedia Commons (CC4)

Autoencoder: dimensionality reduction (3/4)

- Suppose you managed to train a network that gives a good restoration of the original signal $\boldsymbol{y} \approx \boldsymbol{x}$
- This means that the data structure can be effectively described by a lower dimensional representation $\boldsymbol{z}$



adapted from: Chervinskii at Wikimedia Commons (CC4)

Autoencoder: dimensionality reduction (4/4)

- In general, autoencoders learn a non-linear transformation
- If you use linear activation functions and only one hidden layer, then the setup becomes almost that of PCA
- The difference is that ANN might find a different solution, it doesn't use eigenvalues

Reflections on the Subject

Frequentist supervised learning

- Essentially a task of function approximation
- A function can be defined
  * Theoretically, by listing the mapping
  * Algorithmically
  * Analytically
- Every equation is an algorithm, but not every algorithm is an equation

Frequentist supervised learning

- Simple and more interpretable methods (e.g., linear regression) vs more complicated “black box” models (e.g., random forest)
- Apparent dichotomy: prediction quality vs interpretability
- However, some complex models are interpretable
  - Convolutional Neural Networks
  - In any “black box” model, one can study effects of removing features to get insights what is a useful feature

What is Machine Learning?

- Machine learning
  - *"a set of methods that can automatically detect patterns in data, and then use the uncovered patterns to predict future data, or to perform other kinds of decision making under uncertainty (such as planning how to collect more data!)"* (Murphy)
- Data mining
- Pattern recognition
- Statistics
- Data science
- Artificial intelligence

We'll first stay here, then
move to the office hour room

Thank you and good luck!