SOSC 4300/5500
Image Data

The Hong Kong University of Science and Technology
Division of Social Science

Han Zhang

Contents

Motivation

- Social scientists have been familiar with survey data for a long time
- Big data era offers opportunities to use other types of data
  - Networks
  - Texts: study cultural changes, meanings, etc.

A picture is worth a thousand words

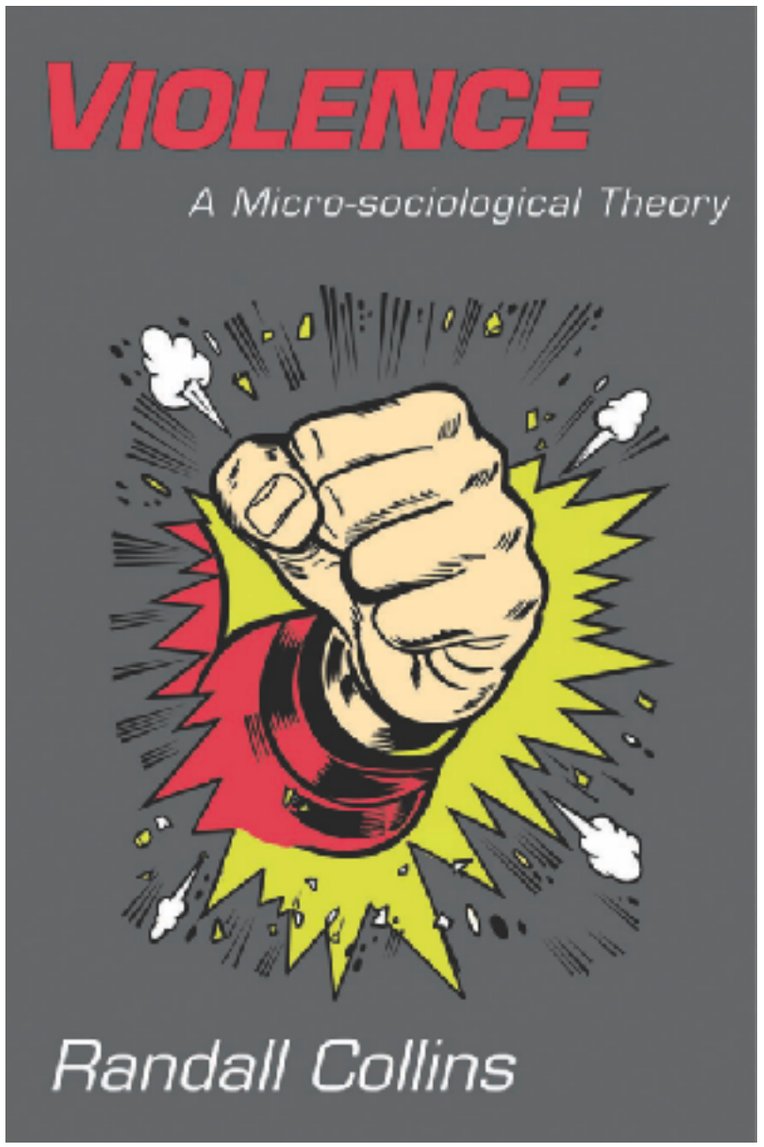
Power of Visual Information

- Psychological roots
  - Dual-coding theory: human beings process visual information quicker than they process text information (Allan Paivio, 1971)
  - Visual information are more effective in triggering attention, improve credibility, increase audience engagement (Pieters & Wedel, 2004; Cancer & Poole, 2017)
- Historical perspective:
  - Earliest painting goes back to 32,000 years ago
  - Written text emerge only around 3000BC
  - But before digital age, visual data are harder to store and produce, compared with texts

Power of Visual Information

- Visual cues are transferrable across culture
- Vs. surveys: more intrusive; sometimes able to observe peoples performance
- Complements other types of big data
  - Photos with captions
  - Video: composites of images, sounds, and texts (transcripts)

Visual Content Analysis in Sociology



- micro-situational analysis: using video recordings to analyze interaction styles and emotions

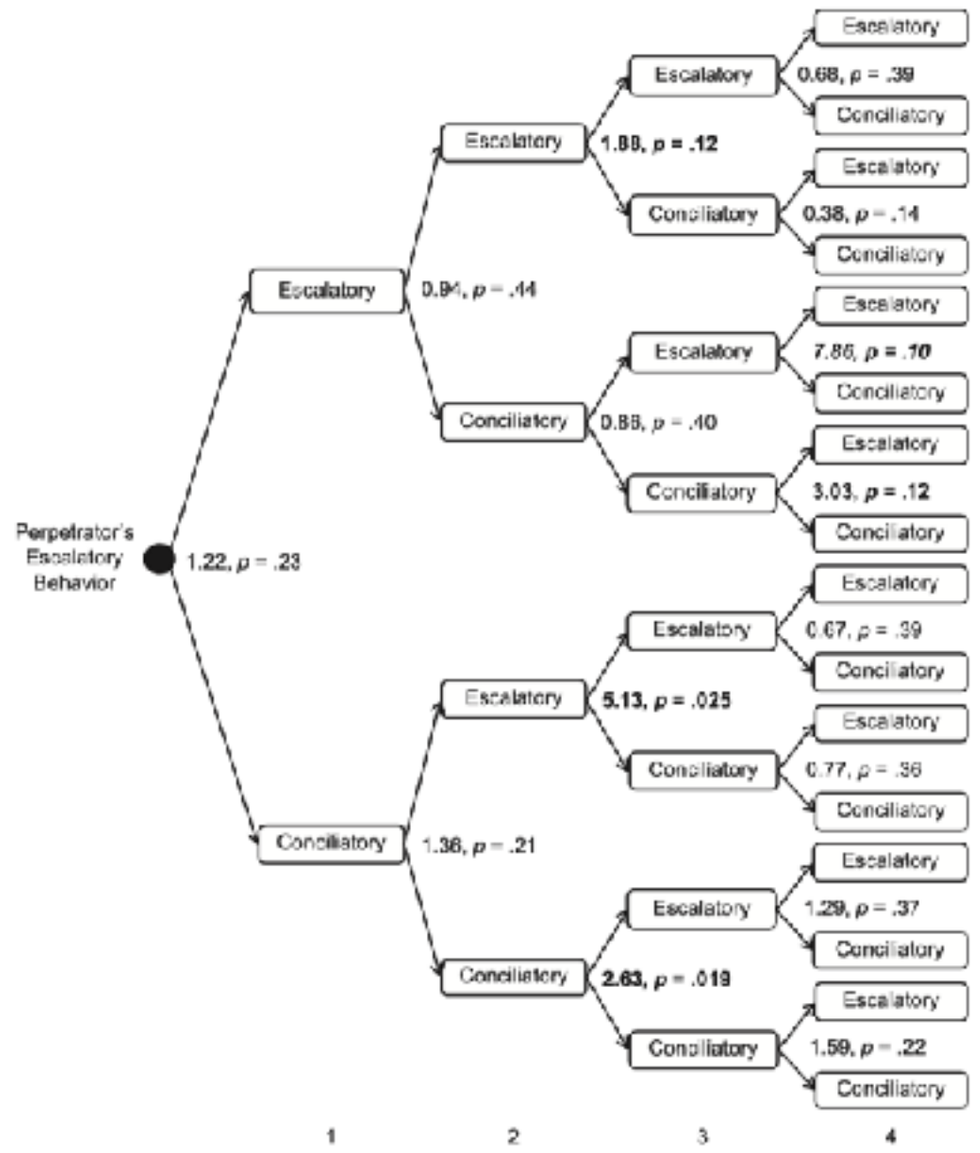
- E.g., Randall Collins (2008) uses visual data to support his theory that violence emerge from situational emotions, instead of predetermined strategies and motivations

- We are “entering the Golden Age of visual sociology” (Collins 2016)

- Levine, Mark, Paul J. Taylor, and Rachel Best. “Third Parties, Violence, and Conflict Resolution: The Role of Group Size and Collective Action in the Microregulation of Violence.” *Psychological Science* 22, no. 3 (March 1, 2011): 406–12. https://doi.org/10.1177/0956797611398495.

- Watch 42 CCTV video tapes involving 312 people, some ends up with violent fighting and others not

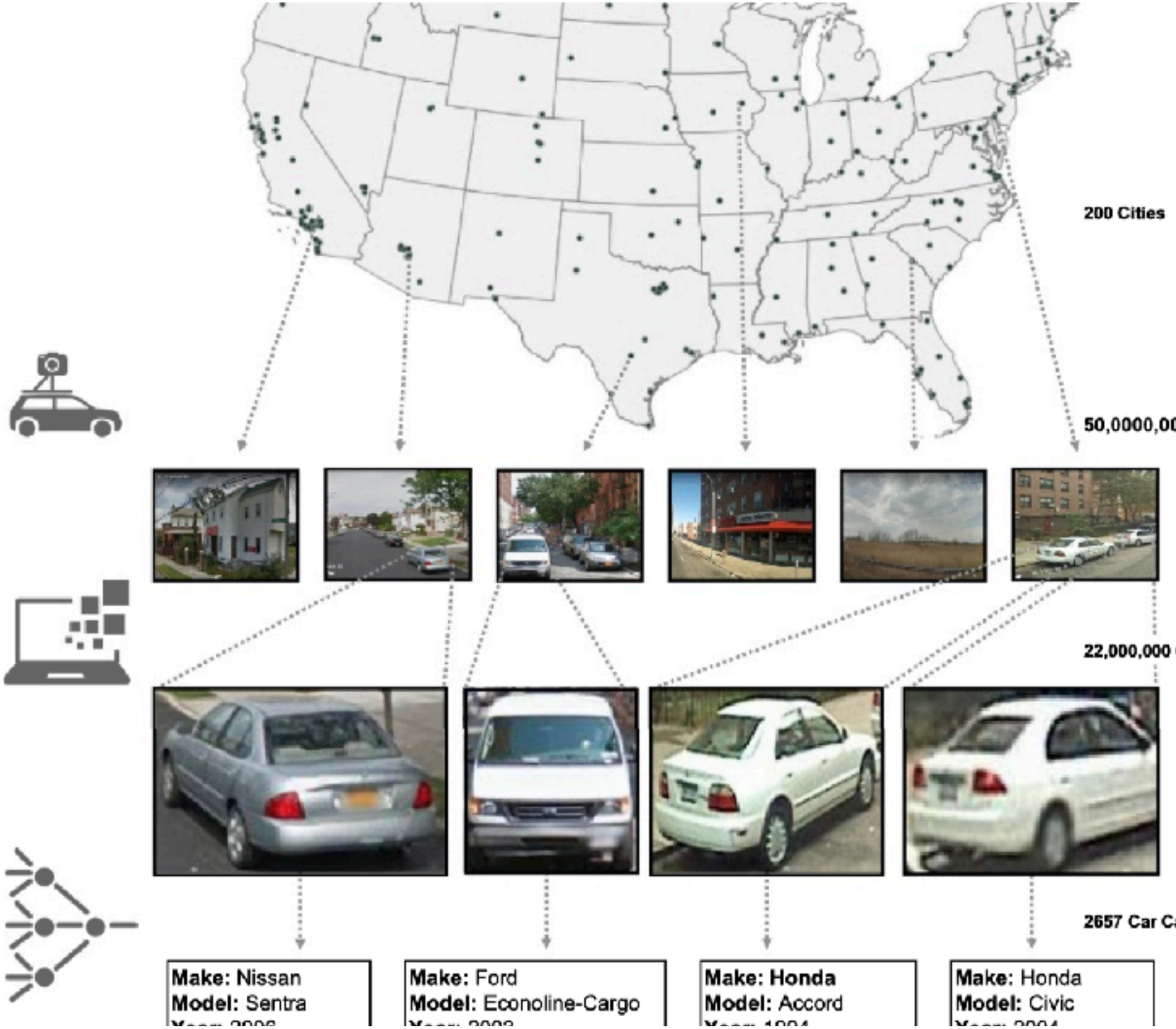
Fig. 1. Examples of escalatory and conciliatory behavior: (a) a perpetrator directing a punch toward the victim and (b) a third party inserting himself between the perpetrator and the victim. Faces in these still frames have been blurred to protect the identities of the people involved.

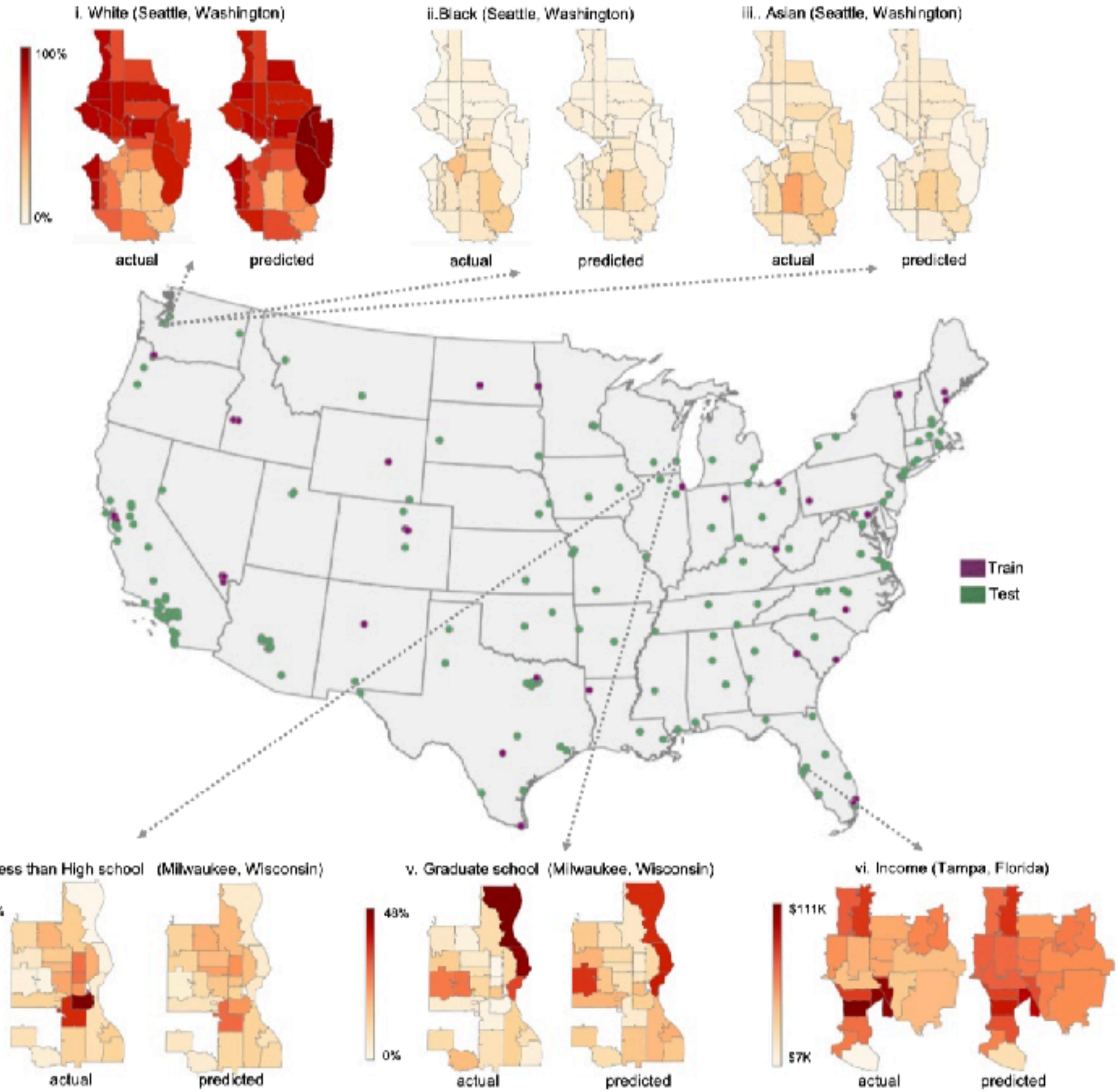


Images reveal socio-demographics conditions

- Gebru, Timnit, et al. "Using deep learning and Google Street View to estimate the demographic makeup of neighborhoods across the United States." Proceedings of the National Academy of Sciences 114.50 (2017): 13108-13113.

- Intuition:
  - Fine-grained survey data are hard to obtain;
    - E.g., American Community Survey, $250 million per year; 2.5 year lag
  - Use Google Street View to obtain reasonably good measures of demographic makeup


200 Cities
50,0000,000 Images
22,000,000 Cars Analyzed
2657 Car Categories
Make: Nissan
Model: Sentra
Make: Ford
Model: Econoline-Cargo
Make: Honda
Model: Accord
Make: Honda
Model: Civic


i. White (Seattle, Washington)
ii.Black (Seattle, Washington)
iii.. Asian (Seattle, Washington)
100%
0%
actual
predicted
actual
predicted
actual
predicted
Train
Test
iv. Less than High school (Milwaukee, Wisconsin)
34%
0%
actual
predicted
v. Graduate school (Milwaukee, Wisconsin)
48%
0%
actual
predicted
vi. Income (Tampa, Florida)
$111K
$7K
actual
predicted

Large Visual Datasets

- Surge in social media platforms
  - From pure text-based Internet content such as blog or static websites
  - To text/image hybrid: Facebook, Twitter, Weibo.
  - To image/video dominant: Instagram, Youtube, TikTok

Modern Visual Data (cont'd)

- And many other digitized platforms
- Google/Baidu maps
- Satellite images
- Digitalized, historical archives
  - Historical maps
- Others?

- Human codings are slow and may not be consistent

- To analyze large-scale image dataset, we need help from computer science tools

  - Computer vision (subfield of CS that design algorithms to analyze visual data)

  - Machine learning (subfield of CS that learns pattern from large-scale data)

  - After 2010s, the dominant approach is deep learning

2. Automated image analysis

Two approaches

- Explorative: not a specific variable to measure; explore themes in large-scale image datasets
    - unsupervised ML algorithm; cluster analysis
- Explanatory: have a clear independent or dependent variable that you want to measure from images
    - **Rule-based** method
        - Very similar to dictionary-based methods.
    - supervised ML algorithm
- Nowadays, both unsupervised and supervised algorithms rely on deep learning

Unsupervised approaches

- Zhang, Han, and Yilang Peng. “Image Clustering: An Unsupervised Approach to Categorize Visual Data in Social Science Research.” *Sociological Methods and Research, 2022.*

- Example: images from CASM-China dataset in first half of 2016 (n = 14K)

- What themes do you expect from Chinese protesters’ images on Weibo?

1 A B C D E F G H I J










2










3










4










5

6

Explanatory study

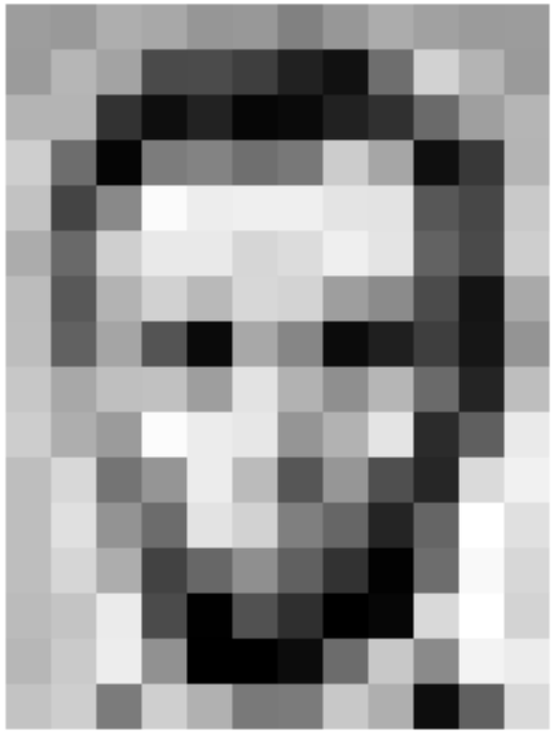
- You have a clear concept/variable you want to measure with image data

    - Basic image features: brightness (dim or bright); color tones.. (**no machine learning needed**)
    - Objects: crowds; fire; symbols; crowds, new buildings…
        - Texts (OCR): a specific type of object.
    - Scenes: e.g., outdoor/indoor; type of places…
    - Variables related to face (e.g. is there a face? Facial expressions, sentiments.)

2. 1

Image representation and rule-based methods

Raw Data



157	153	174	168	150	152	129	151	172	161	155	156
155	182	163	74	75	62	33	17	110	210	180	154
180	180	50	14	34	6	10	33	48	106	159	181
206	109	5	124	131	111	120	204	166	15	56	180
194	68	137	251	237	239	239	228	227	87	71	201
172	105	207	233	233	214	220	239	228	98	74	206
188	88	179	209	185	215	211	158	139	75	20	169
189	97	165	84	10	168	134	11	31	62	22	148
199	168	191	193	158	227	178	143	182	106	36	190
205	174	155	252	236	231	149	178	228	43	95	234
190	216	116	149	236	187	86	150	79	38	218	241
190	224	147	108	227	210	127	102	36	101	255	224
190	214	173	66	103	143	96	50	2	109	249	215
187	196	235	75	1	81	47	0	6	217	255	211
183	202	237	145	0	0	12	108	200	138	243	236
195	206	123	207	177	121	123	200	175	13	96	218

157	153	174	168	150	152	129	151	172	161	155	156
155	182	163	74	75	62	33	17	110	210	180	154
180	180	50	14	34	6	10	33	48	106	159	181
206	109	5	124	131	111	120	204	166	15	56	180
194	68	137	251	237	239	239	228	227	87	71	201
172	105	207	233	233	214	220	239	228	98	74	206
188	88	179	209	185	215	211	158	139	75	20	169
189	97	165	84	10	168	134	11	31	62	22	148
199	168	191	193	158	227	178	143	182	106	36	190
205	174	155	252	236	231	149	178	228	43	95	234
190	216	116	149	236	187	86	150	79	38	218	241
190	224	147	108	227	210	127	102	36	101	255	224
190	214	173	66	103	143	96	50	2	109	249	215
187	196	235	75	1	81	47	0	6	217	255	211
183	202	237	145	0	0	12	108	200	138	243	236
195	206	123	207	177	121	123	200	175	13	96	218

- **Grayscale Representation of an 16 * 12 image**
- **Each cell is from 0 to 255 (0 is black and 255 is white)**
- **Matrix can further be simplified as 16*12 = 192 dimension vector**

- **RGB Representation of an 6*5 image**
- **This matrix can then be simplified as a 6*5*3= 90 dimensional vector**

165	187	209	58	7
14	125	233	201	98
253	144	120	251	41
67	100	32	241	23
209	118	124	27	58
210	236	105	169	19
35	178	199	197	4
115	104	34	111	19
32	69	231	203	74

159	147	204	30	79	156
165	201	14	218	196	218

Challenges

- What if we simply regress Y on X_{raw} using linear regression or GLM?
  - $Y = \beta X_{raw}$
- It is not going work, because:
  - Algorithm side: linear regression and generalized linear models usually do not work well with high-dimensional data
    - E.g., for a typical 600*600 colored image, the raw vector length is **1,080,000**
    - You cannot fit a linear regression if there are more variables than observations
  - Data side:
    - The raw representation contains too much noises

Noise

Stalin in Image?

It would be much easier if first clean the data first

Noises

If our goal is to perform facial recognition in a picture the following things are noises:

Lightening

https://alitarhini.wordpress.com/2010/12/05/face-recognition-an-introduction/

Noises

Angle


Figure 7: − 90°
Figure 8: − 80°
Figure 9: − 70°
Figure 10: − 60 °
Figure 11: − 50°


Figure 12: − 40°
Figure 13: − 30°
Figure 14: − 20°
Figure 15: − 10°
Figure 16: 00°

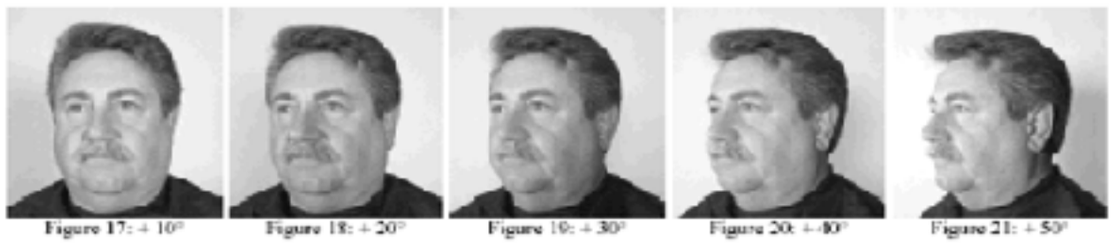


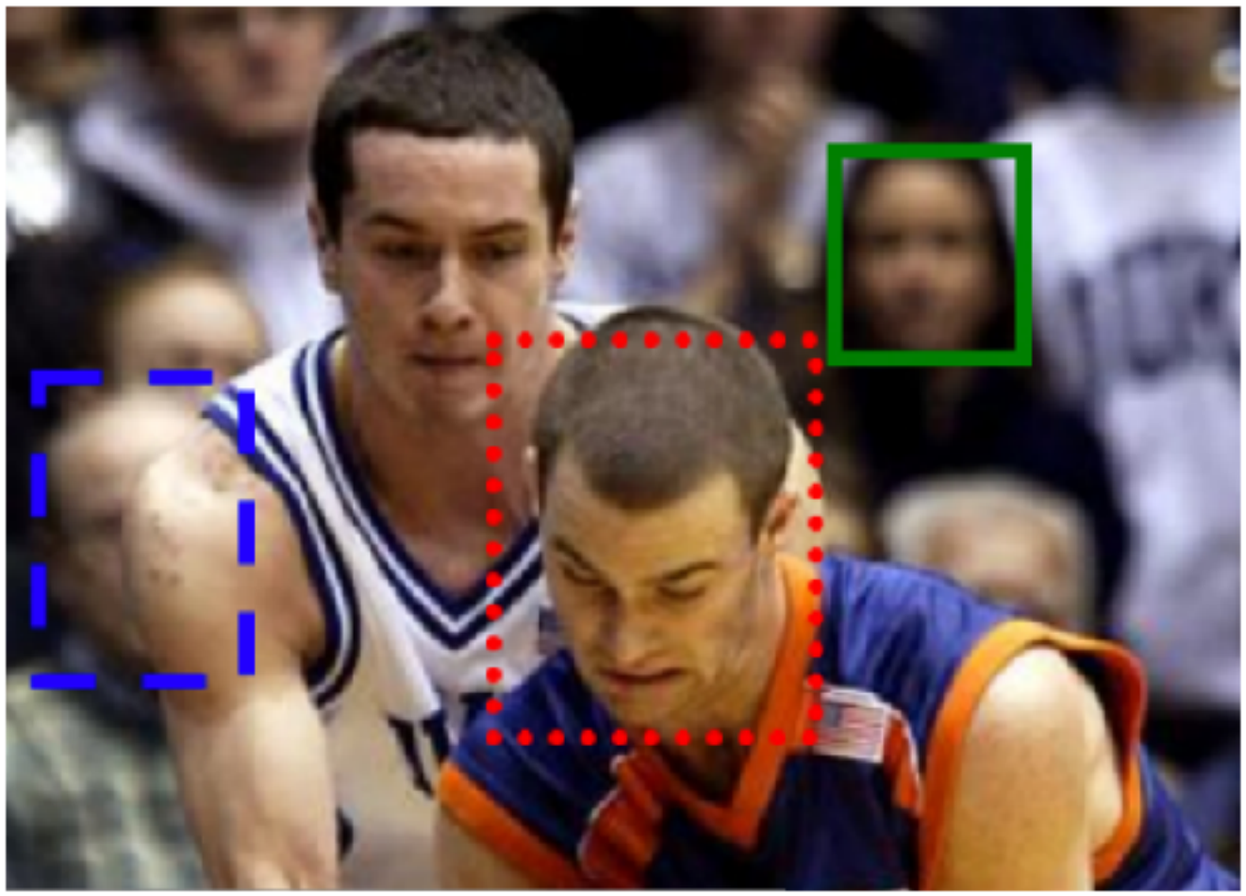
Figure 17: + 10°
Figure 18: + 20°
Figure 19: + 30°
Figure 20: + 40°
Figure 21: + 50°


Figure 22: + 60°
Figure 23: + 70°
Figure 24: + 80°
Figure 25: + 90°

https://alitarhini.wordpress.com/2010/12/05/face-recognition-an-introduction/

Noises

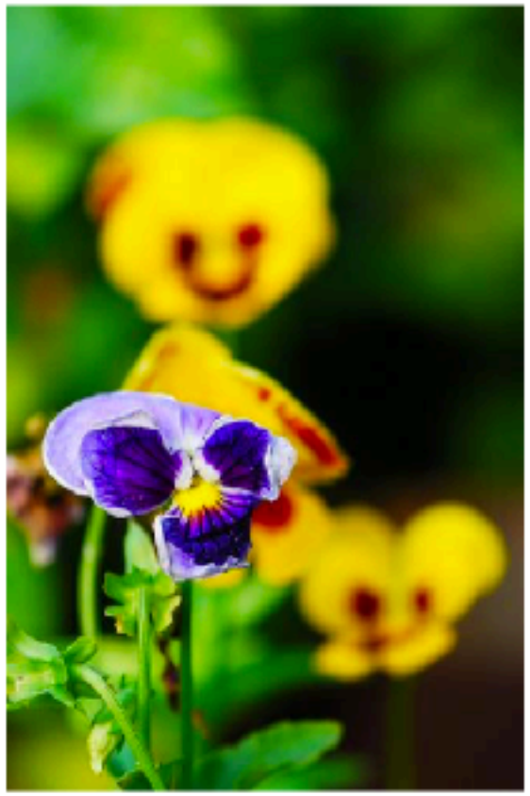
Background



Jain, Vidit, and Erik Learned-Miller. *Fddb: A benchmark for face detection in unconstrained settings*. Vol. 88. Technical Report UM-CS-2010-009, University of Massachus
Amherst, 2010.

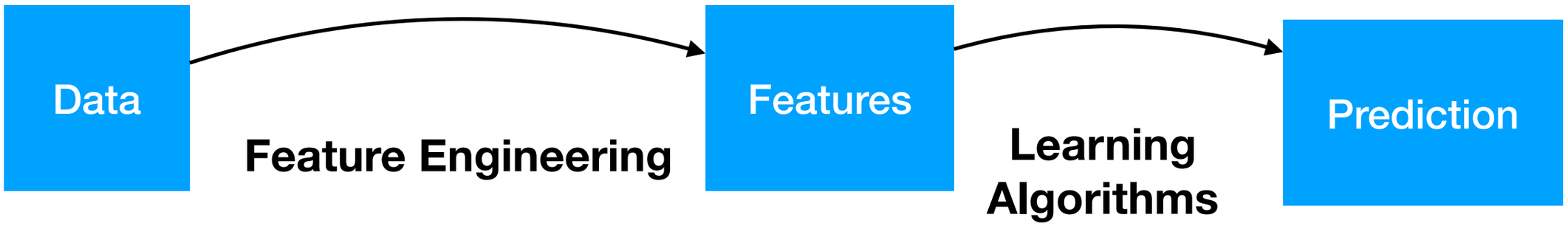
Noises

multiple faces/ no faces



https://alitarhini.wordpress.com/2010/12/05/face-recognition-an-introduction/

Feature extraction

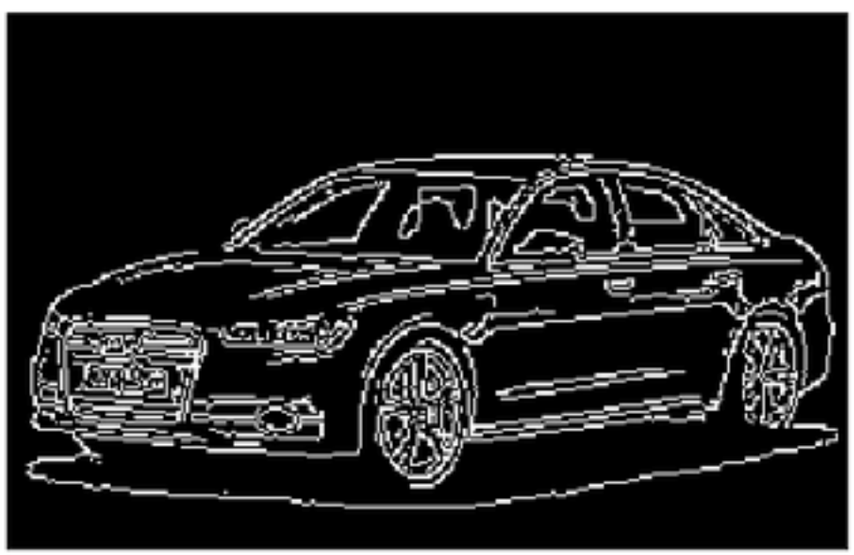


X_{raw} X_{cleaned} Y

Using a text analogy, extract features from images
is similar to
transforming texts into document-term matrix / embeddings

Feature Extraction Example

Edge detection is one of the simplest method to remove noises and extract shape of an image

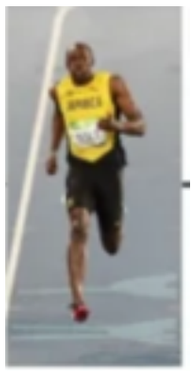


(https://www.analyticsvidhya.com/blog/2019/09/feature-engineering-images-introduction-hog-feature-descriptor/

How edge detection is performed?

- Take first-order differences between each pixels (vertical or horizontal directions)
  - It's called gradient in the computer vision literature
- Unchanged part will be 0 (thus black in the processed image)
- Then only boundary (edge) or an object will be left
- I am showing its simplest intuition below
  - The commonly used one (Canny edge detector 1985) advanced edge detector algorithms

Magnitude of edge



100 x 200

Horizontal difference

g_x

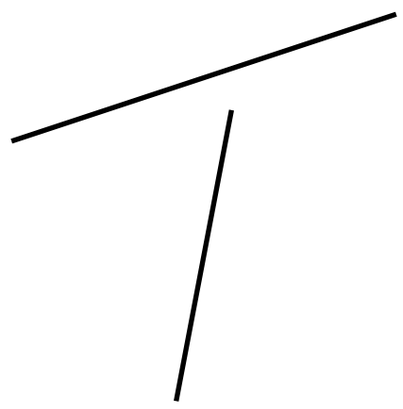
vertical difference

g_y

Magnitude: $\sqrt{g_x^2 + g_y^2}$

Direction of edge

- We can also calculate direction (orientation) of edge
- g_x and g_y are first-order differences for horizontal and vertical axes
  - g_x / g_y
    - if > 1 :
    - if < 1

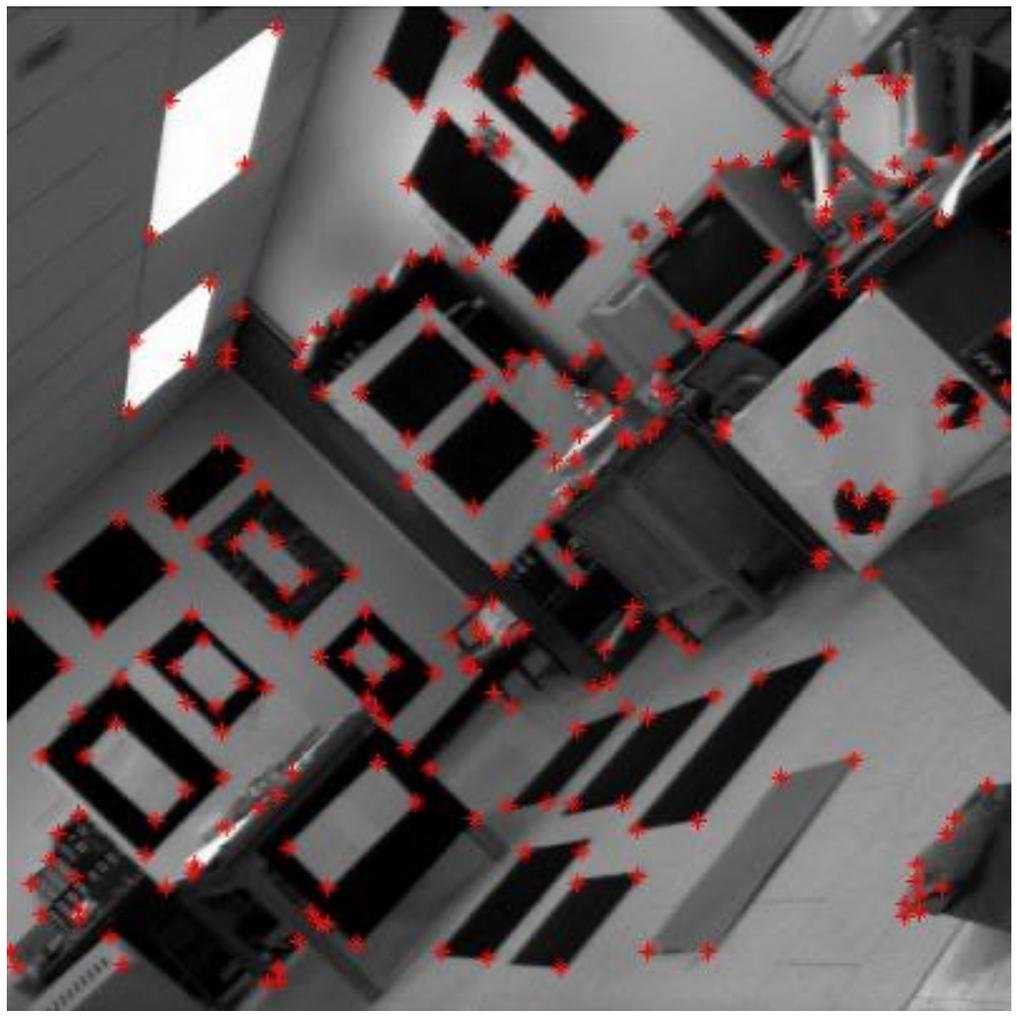


- Technically, direction is measured as the angle $arctan(\frac{g_x}{g_y})$

Many other features

- Corner detection

  - Edge detection may not work



https://www.researchgate.net/publication/327828684_Sparse_Least-Squares_Support_Vector_Machines_via_Accelerated_Segmented_Test_a_Dual_Approach/figures?lo=1&utm_source=google&utm_medium=organic

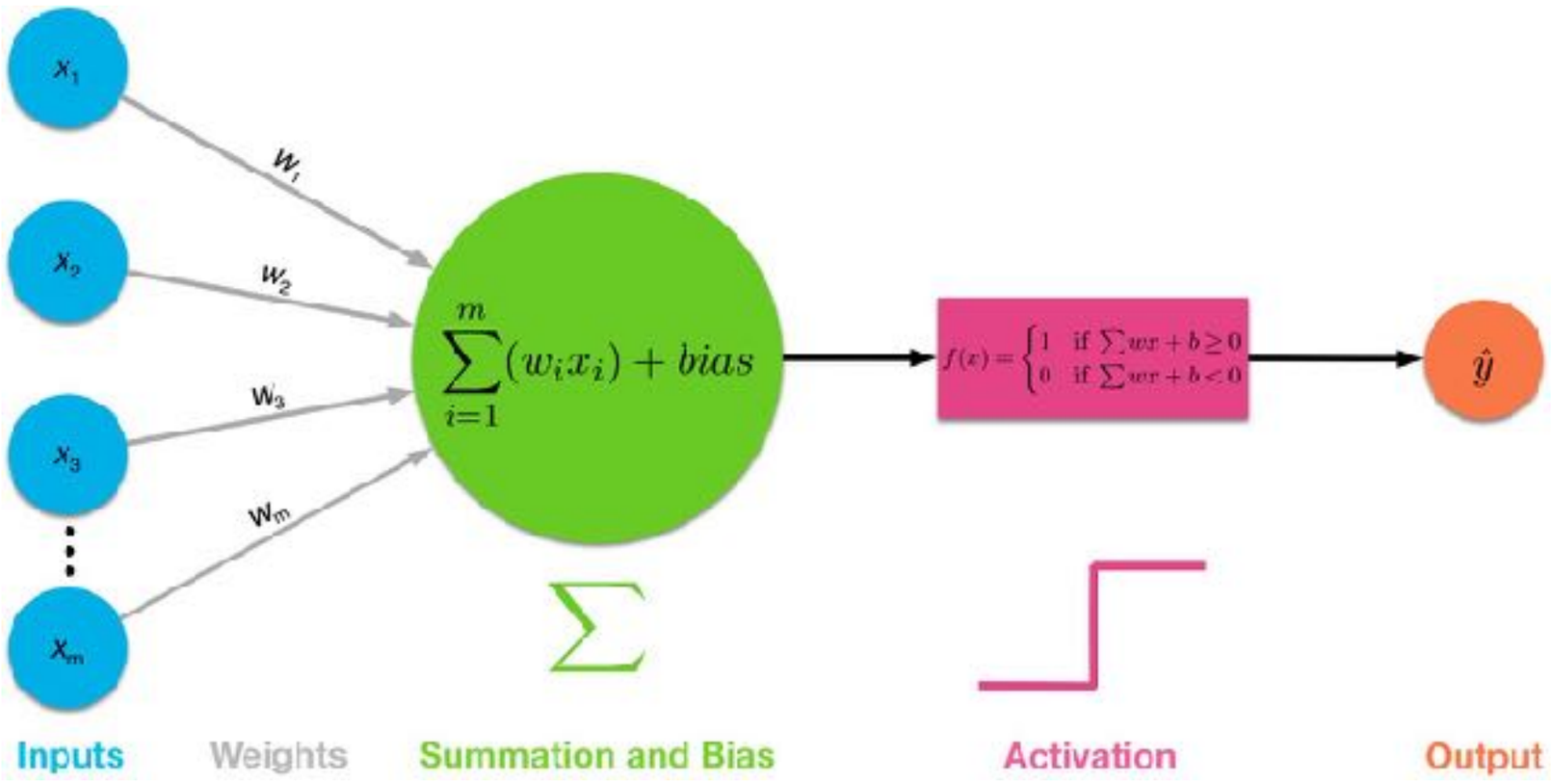
Many problems remain

- There are many arbitrary choices during feature extraction steps
  - General feature extractors can only obtain **low-level** representations
  - Domain-specific extractors lacks generalizability

2.2

Modern Approaches: Deep Learning

Single-layer neural network



- f() is called **activation** function (nonlinear)
- w: **weights (regression coefficients)**
- **bias = intercept**

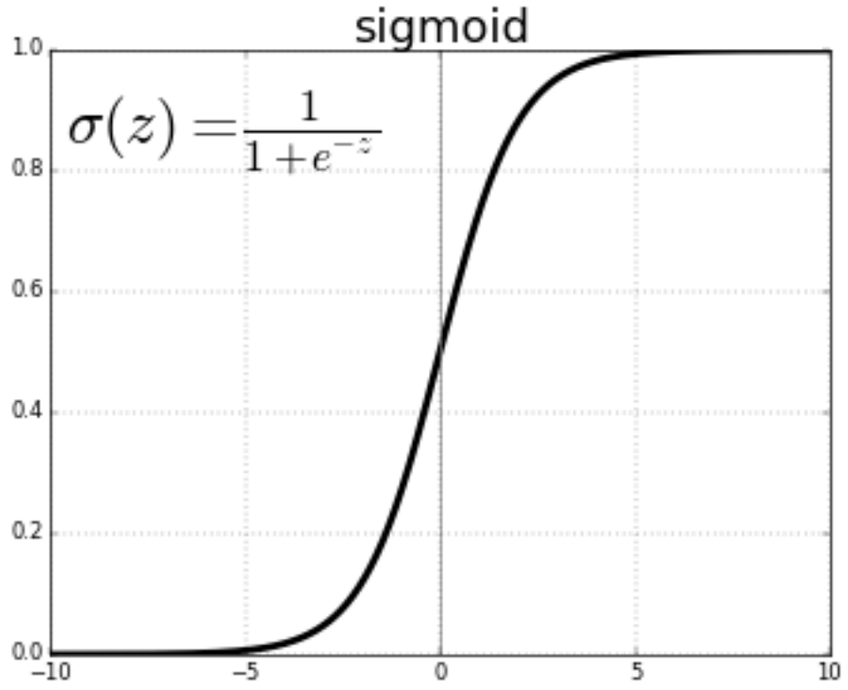
Activation function

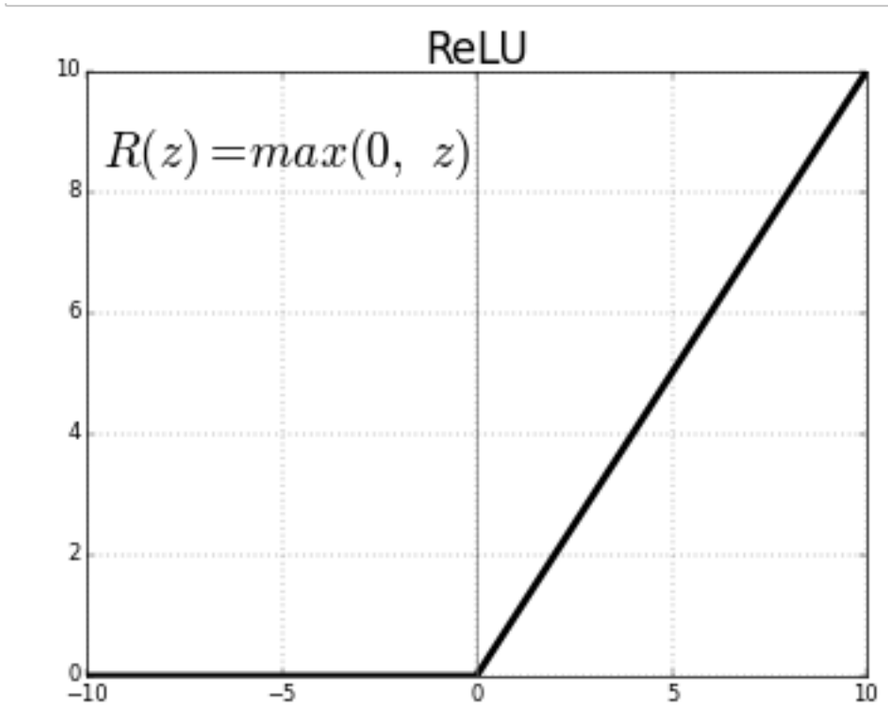
- Activation function adds non-linearity to the linear weighted sum of hidden units.

- If f() is logistic function (sigmoid), then **single-layer neural network becomes logistic regression**

  - Not used in reality; due to some numerical issues

- Common choices: step function, or ReLu (Rectified Linear Unit)

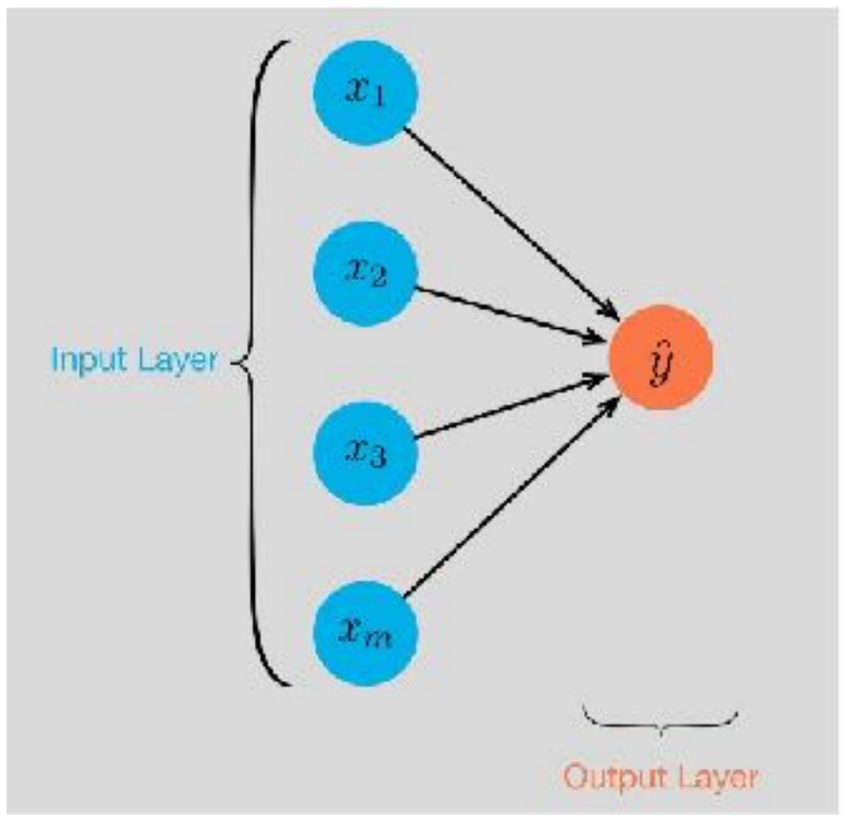


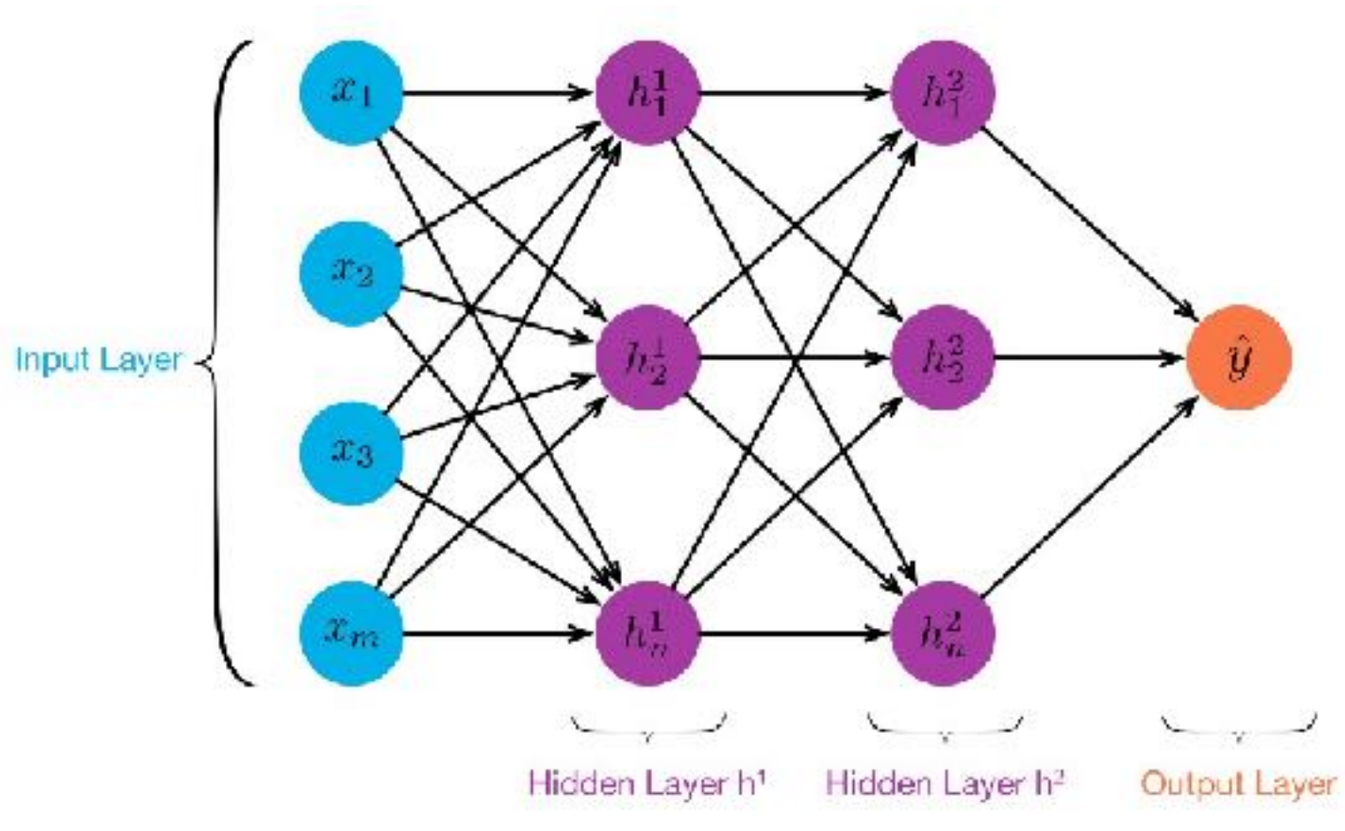


Multi-layer Neural Networks

- **There are hidden units; the algorithms will learn their values**

Rumelhart, David E., Geoffrey E. Hinton, and Ronald J. Williams. "Learning representations by back-propagating errors." *Nature* 323, no. 6088 (1986): 533.



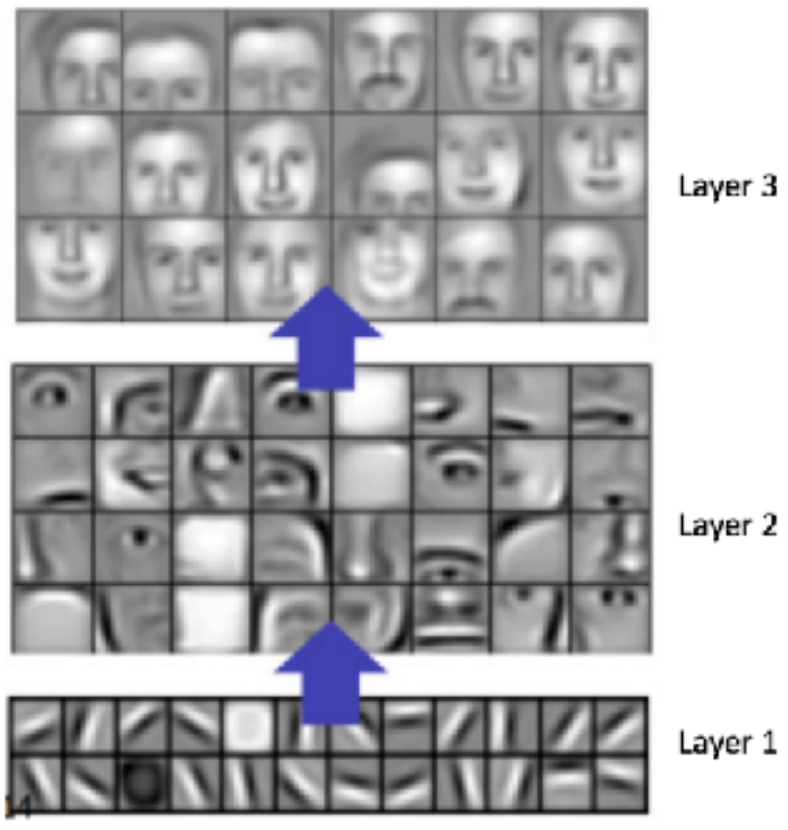


Why Multiple Layers

- Multiple levels of representation

  - lower layers capture lower-level features, such as edges and corners

  - Upper layers capture combinations of motifs.



Honglak Lee, Roger Grosse, Rajesh Ranganath, and Andrew Y. Ng. 2009. Convolutional deep belief networks for scalable unsupervised learning of hierarchical representations. (ICML '09).

Visualization

- https://adamharley.com/nn_vis/mlp/3d.html
- MNIST hand-written digit recognition test
- 28 * 28 images
- 784 (28 * 28) nodes as input
- 300 nodes in the first hidden layer
- 100 nodes in the second hidden layer
- 10 nodes in the output layer (corresponding to the 10 digits).

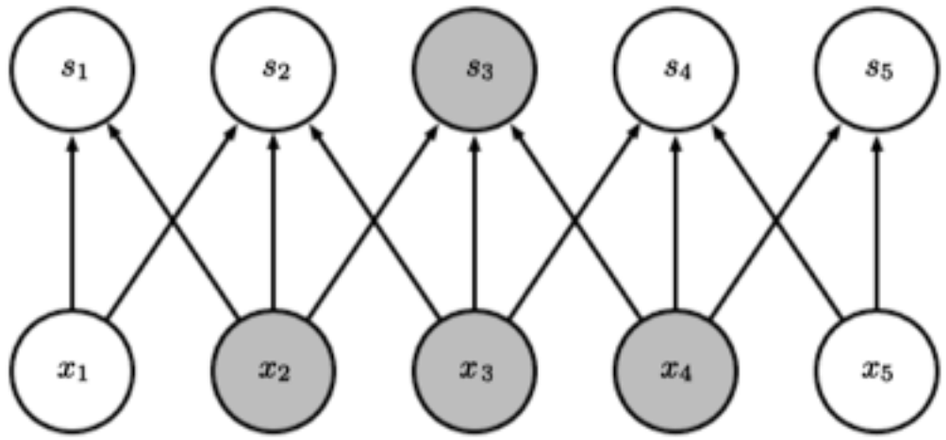
Problems of multi-layer neural networks

- Multi-layer neural networks are **fully connected**
  - Every unit contributes to the next layer
  - It leads to very wide network

- But as we have seem, often only local regions in images matter; not the entire picture
  - This locality assumption is not uncommon beyond images
    - phrases and sentences in texts
    - video/sound clip from a longer video/song

Convolutional Neural Networks (LeCun et al., 1989)

- Local connectivity: each units depends on only on local regions of the previous layer.



- It’s not based on all units of the previous layer

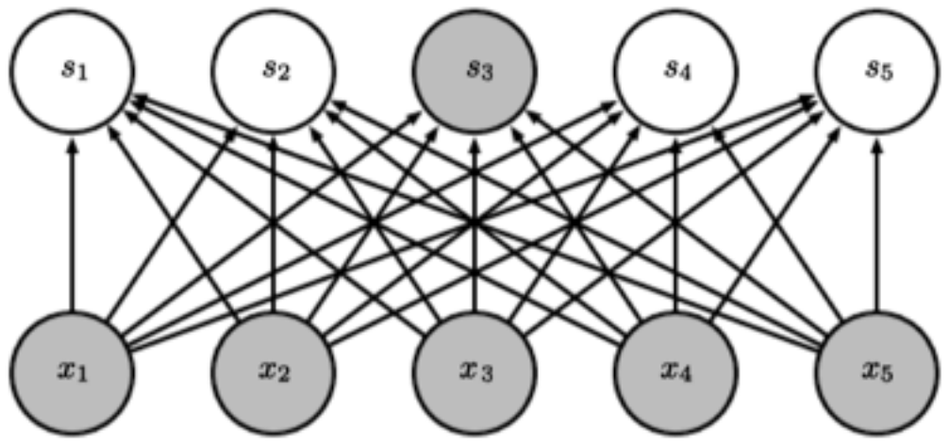


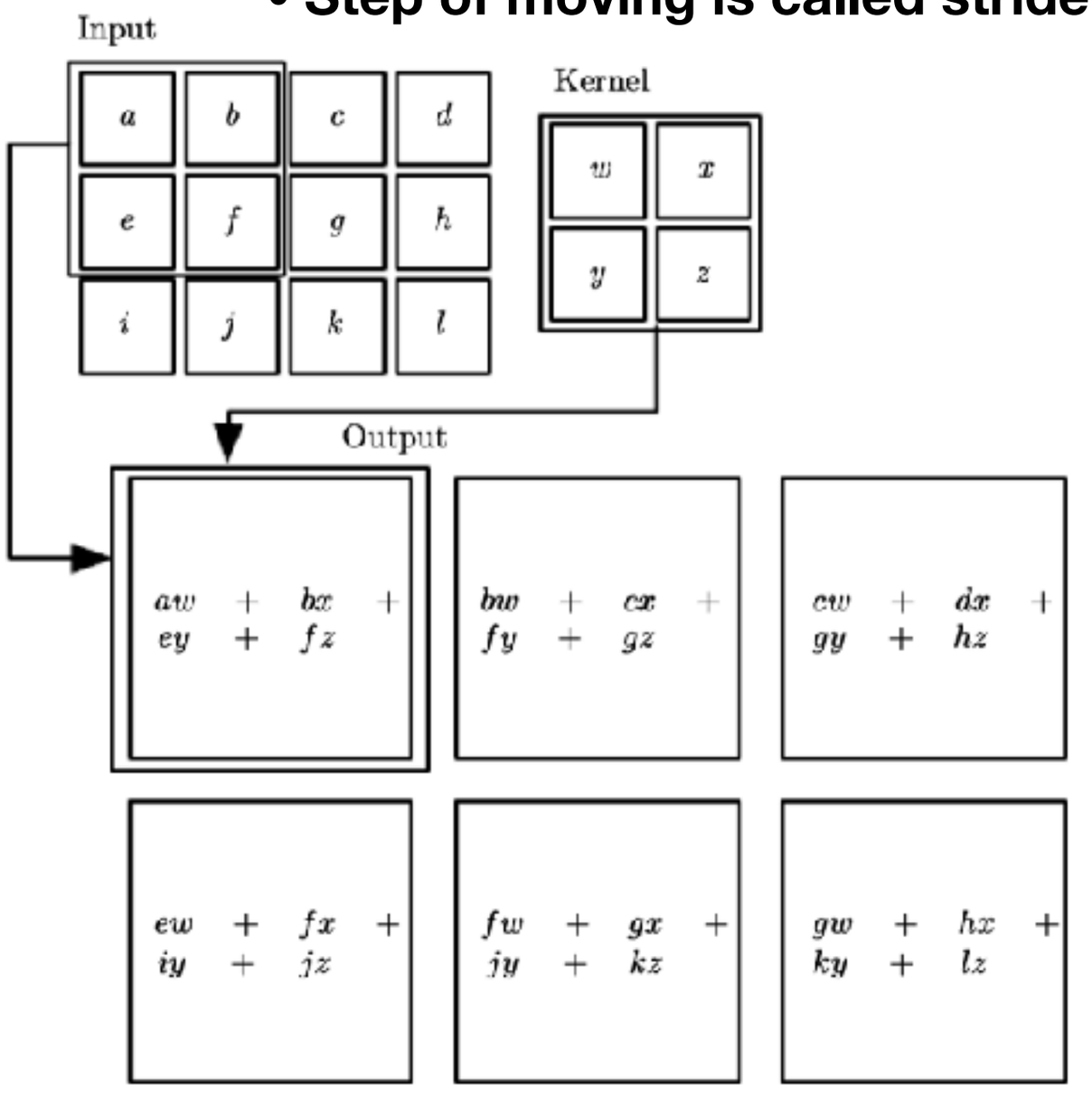
Figure 9.2, Goodfellow et al., 2015

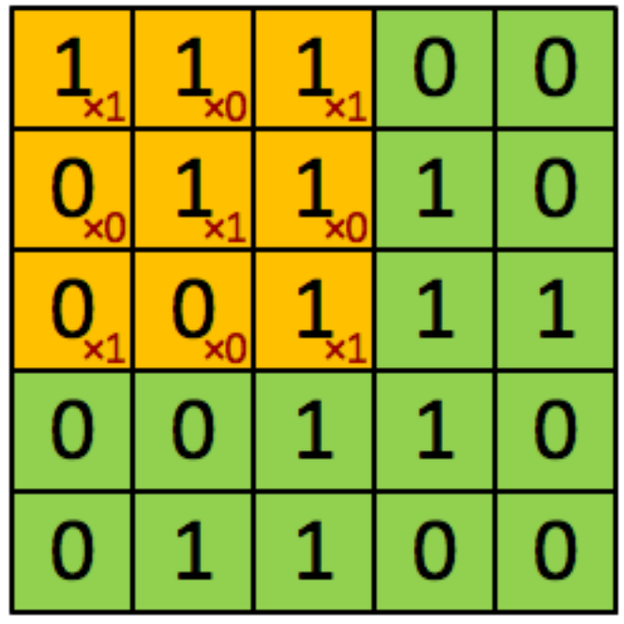

Yoshua Bengio
Geoffrey Hinton
Yann LeCun

Fei-Fei Li, ImageNet

Convolutional Kernels

- **Convolutional Kernel: just a matrix of weights for a subregion**
- **Step of moving is called stride (the below figure has stride 1)**





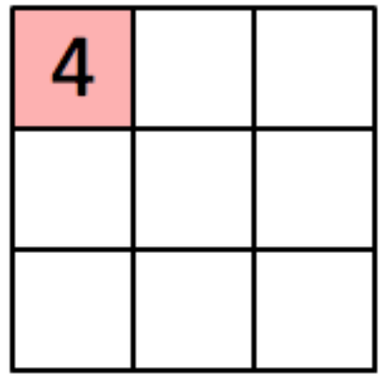


Figure 9.1, Goodfellow et al., 2015

https://hackernoon.com/visualizing-parts-of-convolutional-neural-networks-using-keras-and-cats-5cc01b214e59

Feature Maps

- Convolutional kernel is also called a feature map, or a filter
- It’s called feature map, because it extract one type of feature;
- We usually use multiple feature maps
  - Each capture a different feature (e.g., edge, corners).

One Convolutional Layer

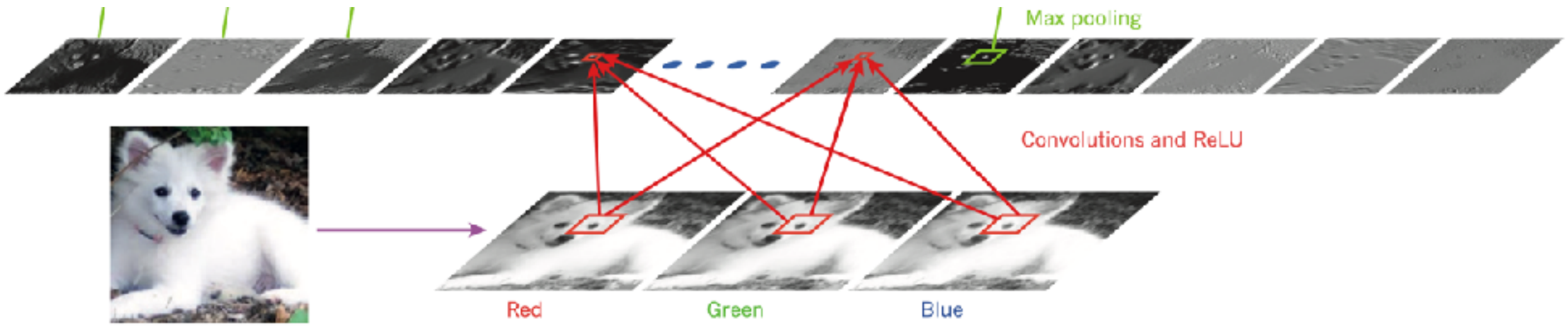


Figure 2,

LeCun, Yann, Yoshua Bengio, and Geoffrey Hinton. "Deep Learning." *Nature* 521.7553 (2015): 436-444.

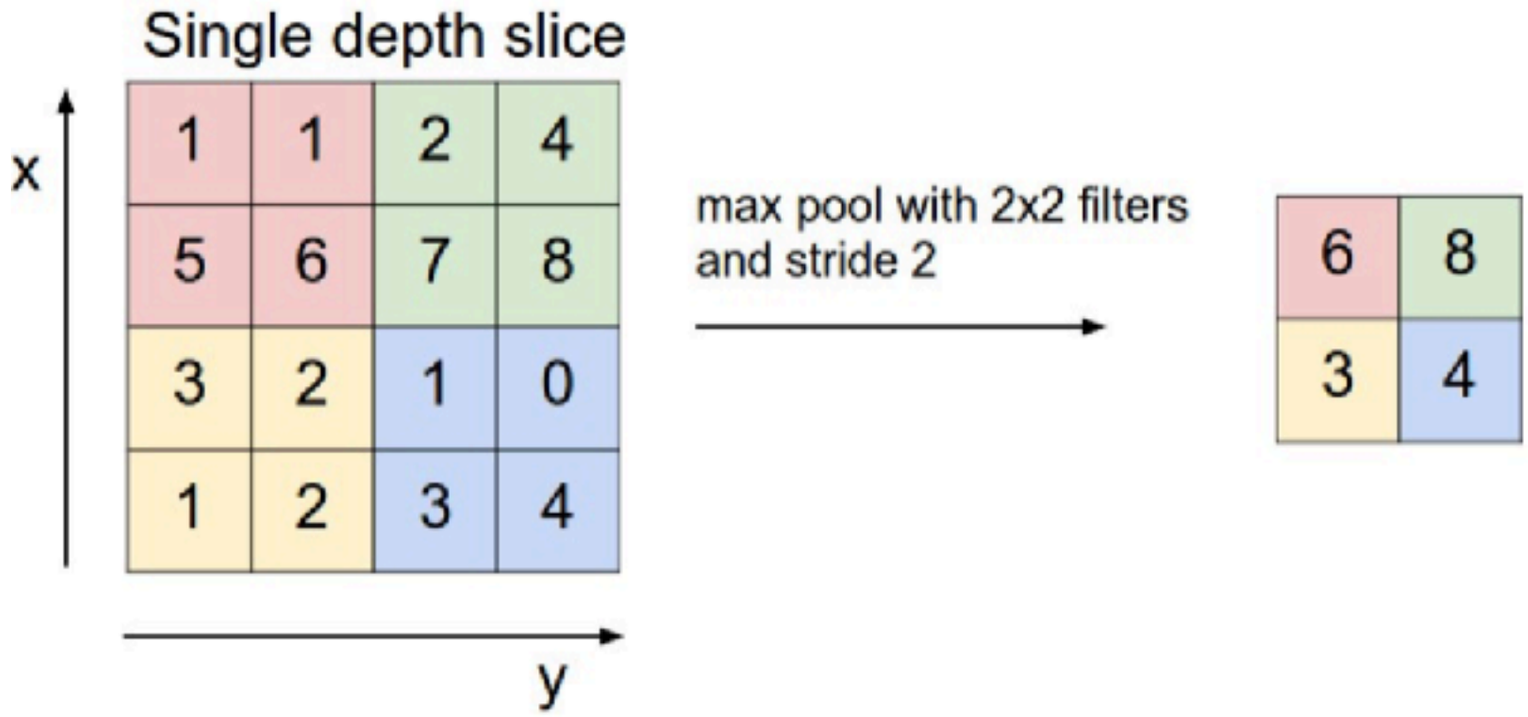
Visualization

https://poloclub.github.io/cnn-explainer/

Pooling

- Max-Pooling further reduces data dimensions, by focusing on only **large** values in data

- This is a kind of regularization

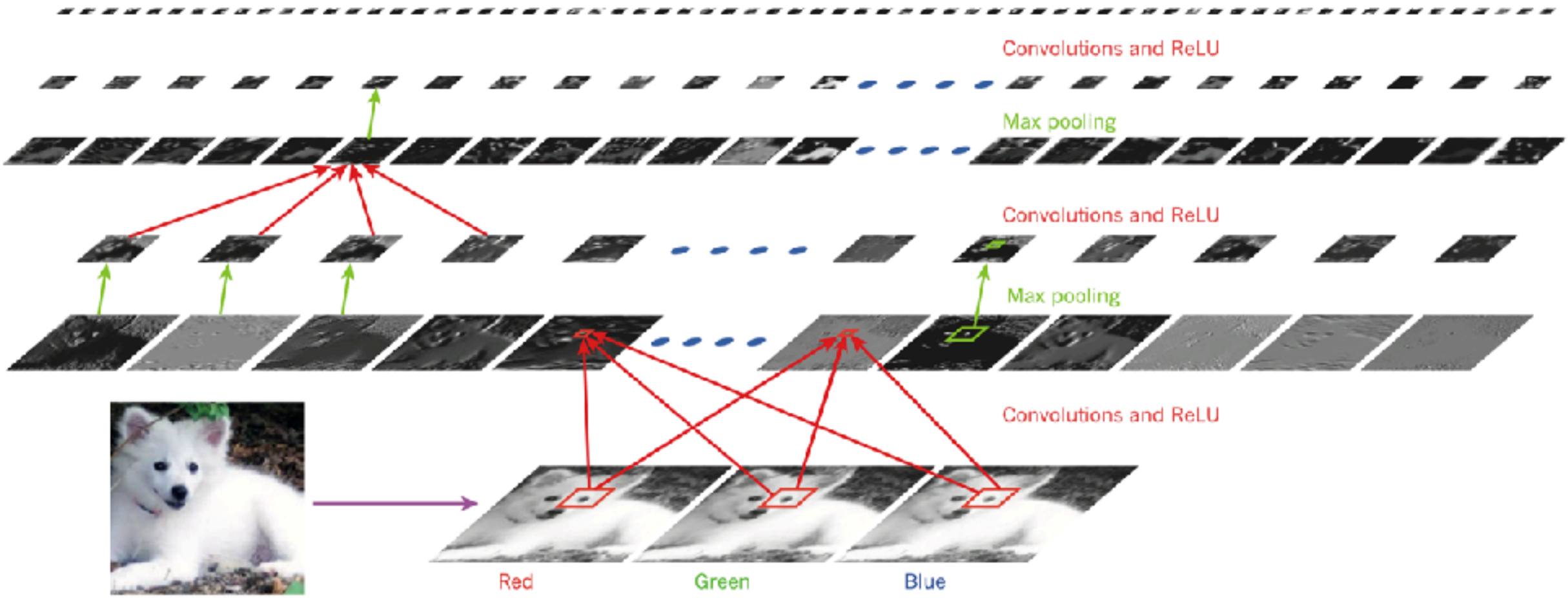


http://cs231n.github.io/convolutional-networks/

Deep Learning

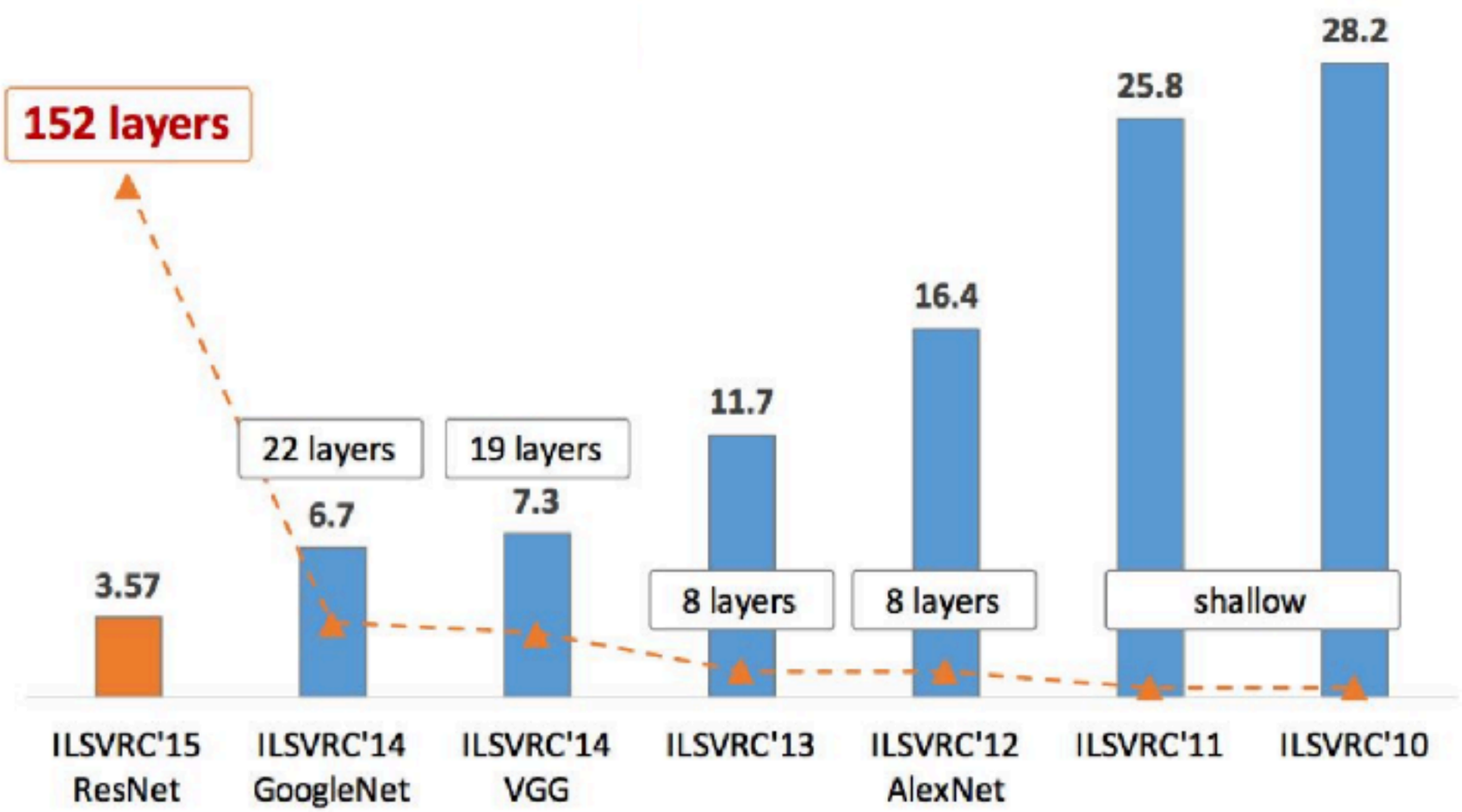
- Combines convolutional network
- And multi-layers architecture
- So it’s both **deep** and **sparse** (not too wide)

Convolutional Layer + Pooling Layer



ImageNet Large Scale Visual Recognition Challenge (ILSVRC) winners

Deep learning models are getting deeper



Visualization

- https://adamharley.com/nn_vis/cnn/3d.html
- 1024 nodes on the bottom layer (32 * 32 image)
- 6 convolutional kernels of 5x5 (stride 1) size
- followed by 16 convolutional filters of 5x5 (stride 1) size
- Then three fully-connected layers,
    - with 120 nodes in the first,
    - 100 nodes in the second,
    - 10 nodes in the third.
- Each convolutional layers followed by downsampling layer that does 2x2 max pooling (with stride 2).

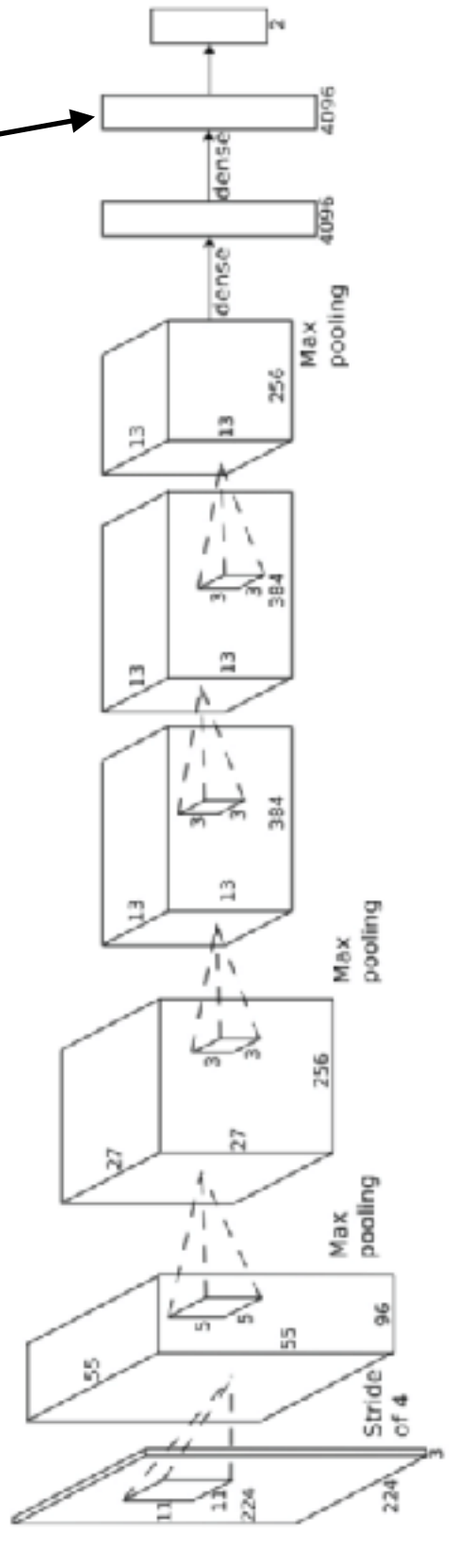
Last step: fully connected layers

- The last (or last several) layers has transformed raw images into a vector

- This vector can be considered as a kind of **extracted image representation**
  - **Similar to word embedding**

- How do we know our learned representations are good?
  - If two images look similar, their learned representations should also have high similarity
  - Similar to word embedding's idea

Test image L2 Nearest neighbors in feature space



http://cs231n.stanford.edu/slides/2017/cs231n_2017_lecture12.pdf

3.

Automated Image Analysis: practice

- Easiest case: you only care about basic features (color, presence of edges)
    - Just apply rules
- Second easiest case:
    - You want to measure something more complex from images
    - There are mature ML algorithms that do similar tasks
        - Often trained in deep learning
        - Often referred as **pre-trained models**
    - So you don’t need to collect training data; just apply existing models

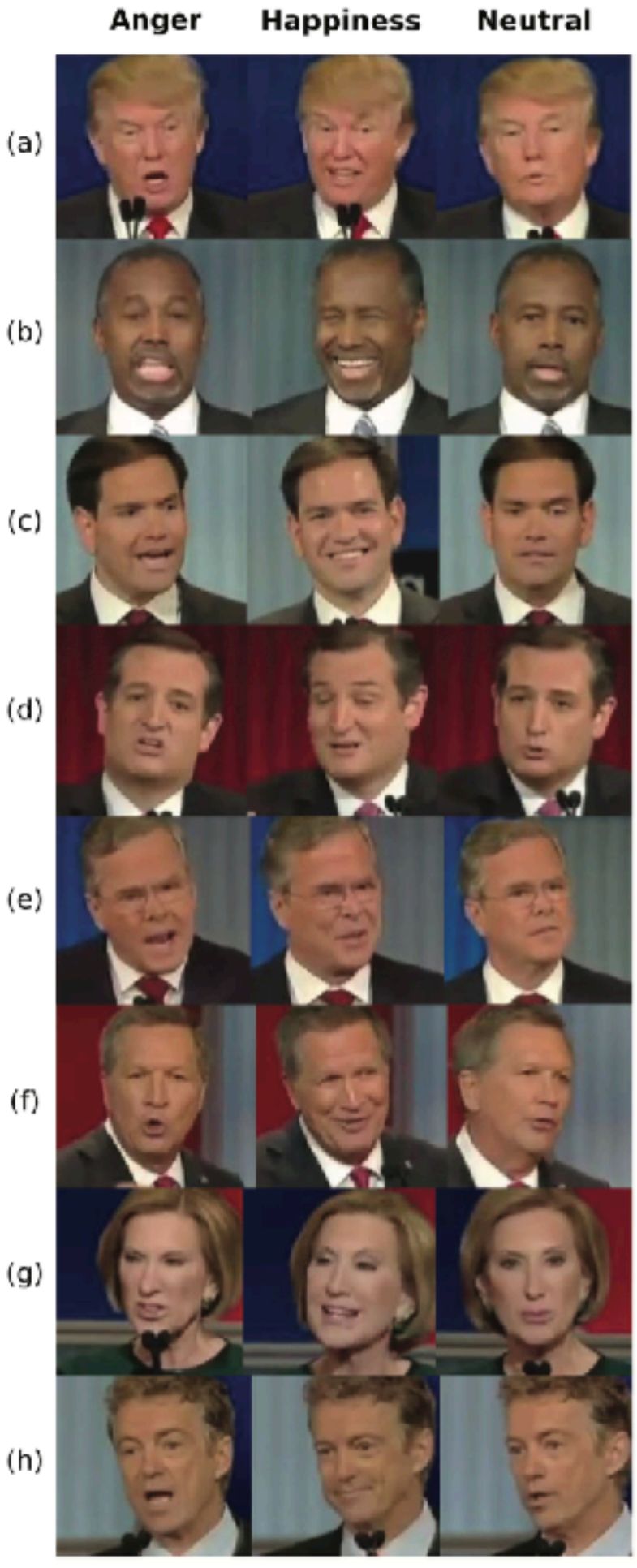
When existing tool is available

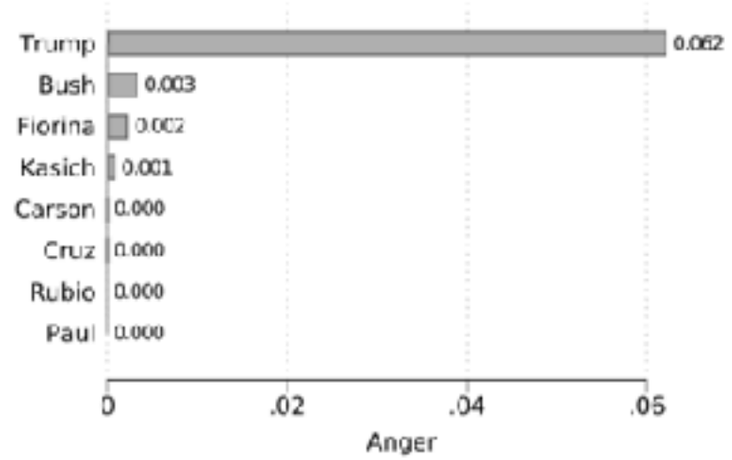
- Boussalis, Constantine, and Travis G. Coan. “Facing the Electorate: Computational Approaches to the Study of Nonverbal Communication and Voter Impression Formation.” *Political Communication* 38, no. 1–2 (March 15, 2021): 75–97. https://doi.org/10.1080/10584609.2020.1784327.

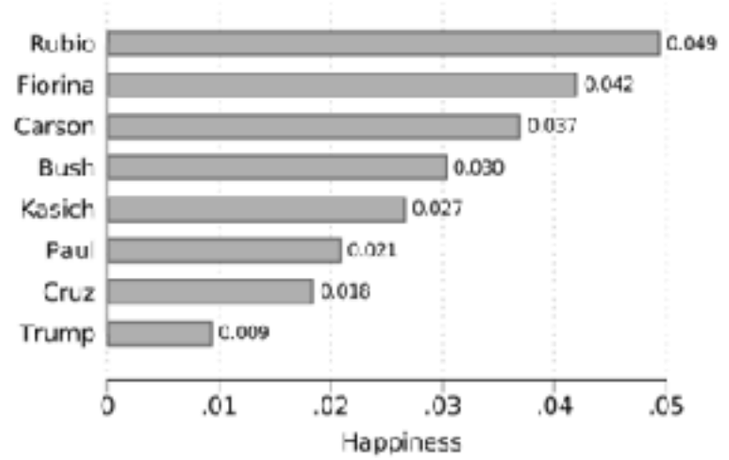
- How does politicians’ emotions (through facial expressions) impacts voters’ support for them?

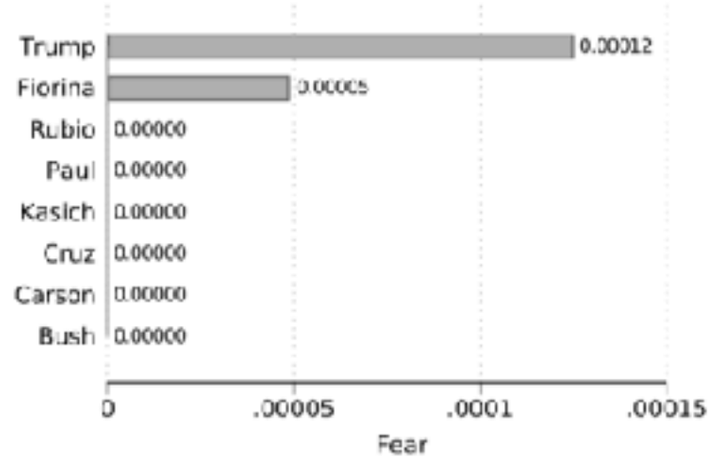
  - Anger; Happiness; Fear /evasion

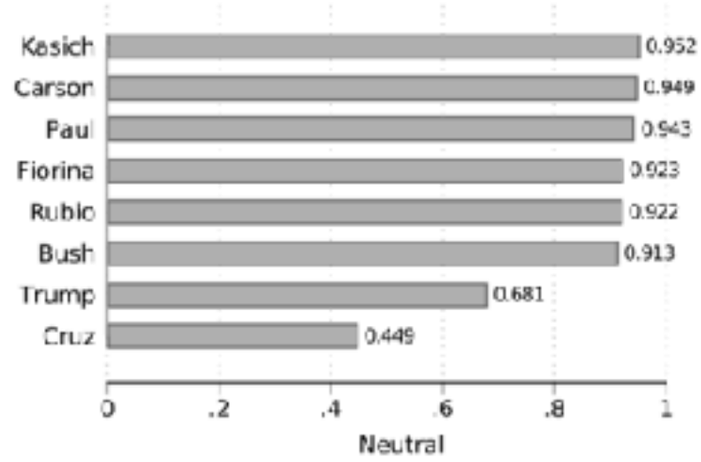
- Republican primary debate on November 10th, 2015
- Debate video is 1:58:41 at 29.97 frames per second
  - 213,442 frames
- Using Microsoft Face API to turn each frame into emotion indexes
  - Many other alternatives available
- Outcome: a focus groups' real time responses to videos











Result: anger predicts more support from viewers

Validation, Validation, Validation

- Always validation ML algorithm's prediction performance on your dataset

- Don't trust softwares's claim that they are the best

- Create a carefully annotated dataset and compared ML's predictions with human labels

- If machine predictions are not accurate enough, you have to do more work to create your model

When existing tool is not available

- Supervised machine learning
  - Need humans to create training data; can be the most resource-demanding part
  - Give training to machine and let them to learn from your training data

Train from scratch

- Follows standard supervised ML procedures: create training data -> tune model through cross-validation -> apply on larger datasets

- Pros: customized to your own problems

- Cons: much more demanding:
    - create training data
    - train deep learning models
    - programming skills

Transfer Learning

- Transfer learning:
  - Intuition: someone learned to fit a model Y = f(g(…(x))) using much bigger data than yours
    - But their outcome Y is different from yours
  - Then you take the model, only train one of the last layer (here, f(x))
    - This is called fine-tuning

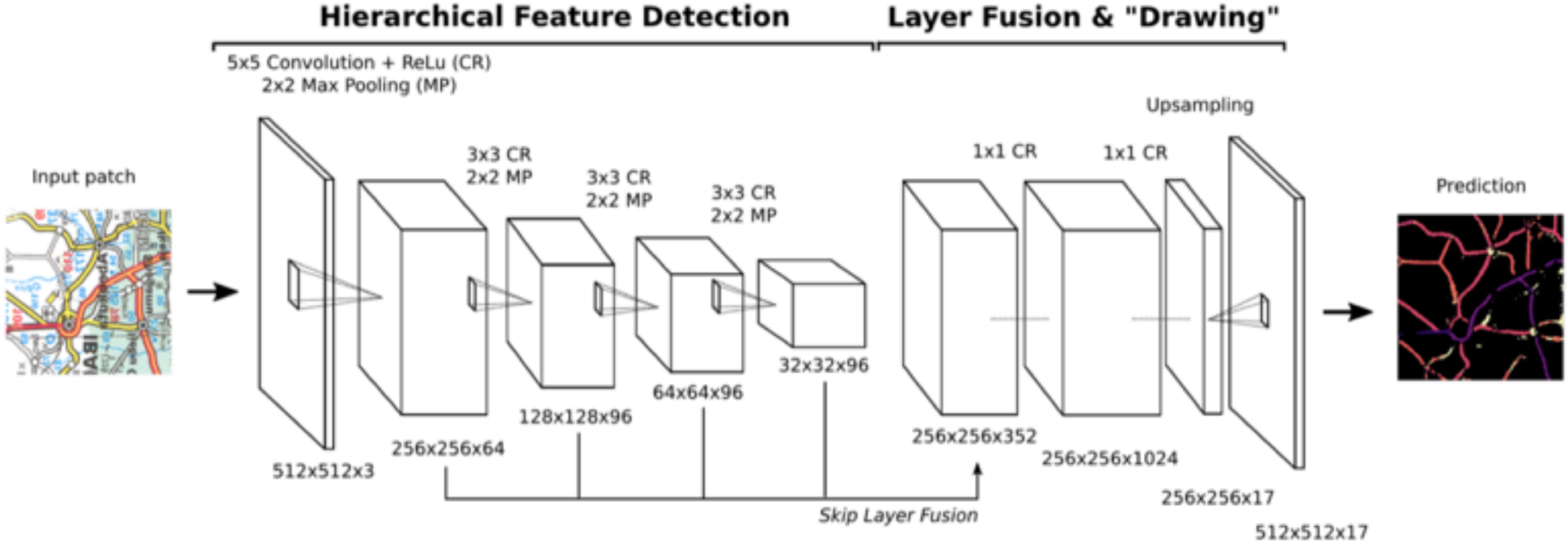
Fine-tuning

- AlexNet, VGG and ResNet is recommended
- Why fine-tuning works: lower layers are basic features which are useful for every task;
    - If they are trained on millions of images, they are much better than what yours
    - You just need to customize it to your own purpose

4. Applications

Historical Maps

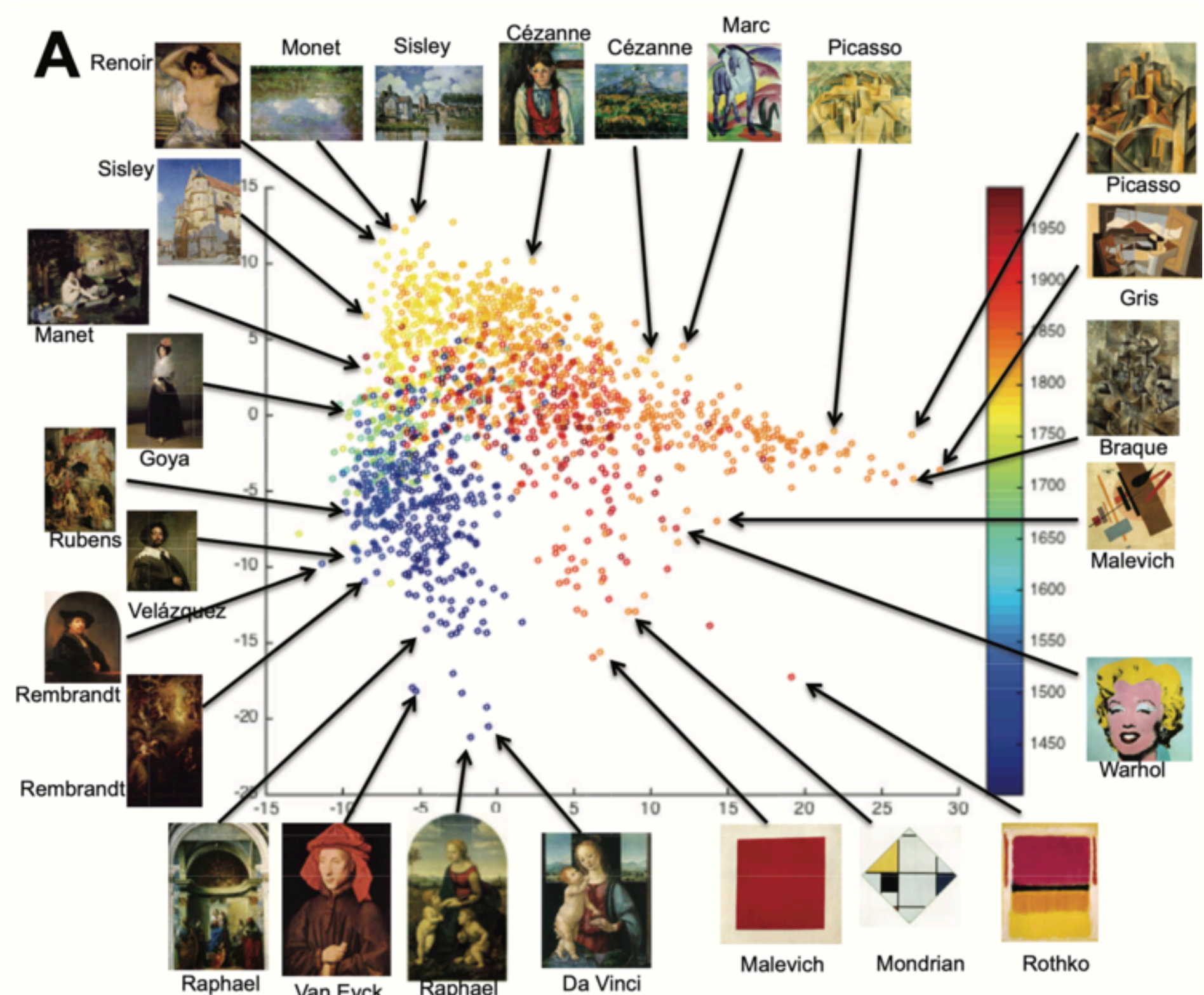
- Theory: regions with low state capacity is more likely to have civil conflicts
- How to measure state capacity?
- Muller-Crepon, Carl, Philipp Hunziker, and Lars-Erik Cederman. “Roads to Rule, Roads to Rebel: Relational State Capacity and Conflict in Africa,” 2020. (R&R at *Journal of Conflict Resolution*)’
- Idea: use road networks
- Problem: only modern measure of road capacity
- Raw data:
    - collection of historical maps in Africa from 1966 to 1990



Findings: civil wars are more likely in areas that lack road to political centers but have interconnected road

Paintings

- Unsupervised clustering



Politician's Instagram Style

- Yilang Peng (2020), What makes politicians' Instagram posts popular? Analyzing social media strategies with computer vision. *The International Journal of Press/Politics*

- US congressmen's instagram photos (~59,000).

- Just feed the images into VGG, and extract the last year as feature representations

- Then use K-means to cluster images and observe patterns

- Last, regress instagram audience engagement on clusters

https://osf.io/9pcf5/

Protesters' visual framing

- Zhang, Han, and Yilang Peng. "Image Clustering: An Unsupervised Approach to Categorize Visual Data in Social Science Research." *Sociological Methods and Research, 2022.*

- Example: images from CASM-China dataset in first half of 2016 (n = 14K)

- Use the last-layer from a pre-trained model (VGG)

    - Then perform K-means or hierarchical clustering algorithms

1 A B C D E F G H I J










2










3










4










5

6

Predict protests from social media

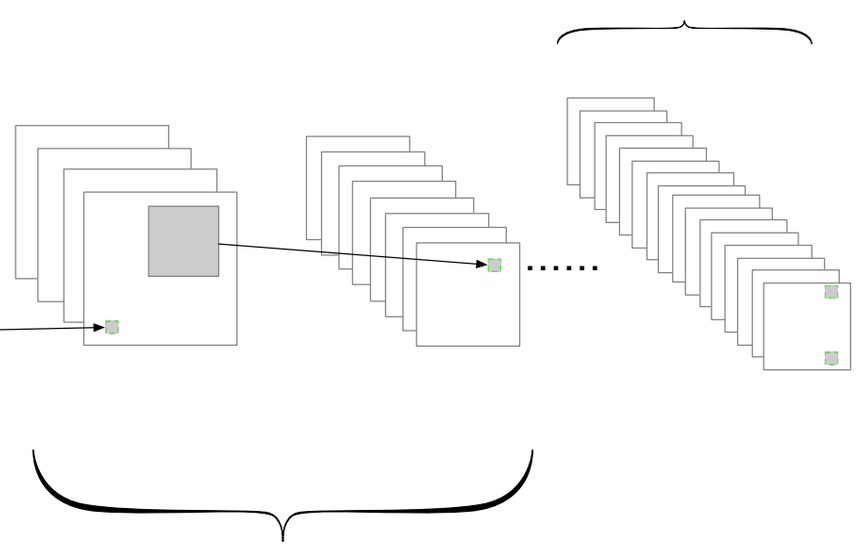
- Zhang, Han, and Jennifer Pan. "CASM: A Deep-Learning Approach for Identifying Collective Action Events With Text and Image Data From Social Media." *Sociological Methodology* 49, no. 1 (2019): 1-57.

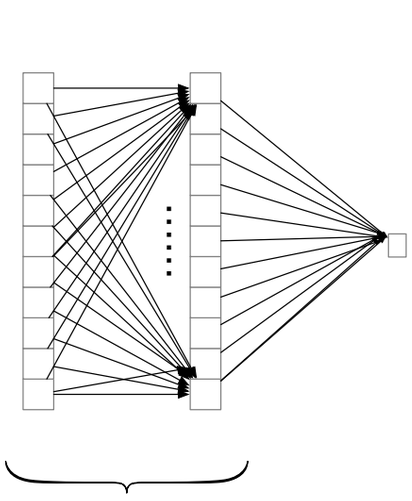
- Goal: give social media post (containing both text and iamges)
  - Predict whether it's about a protest
  - Image are from Weibo

Solution

- Turn each image into a 4096-dimension vector representation
- Start from VGG
- And only train the last 4 layers
- Based on over 200,000 images

Fine Tune
last 4
VGG Layers





Prob (is a protest)

Input image

VGG
12 Convolutional Layers

Fully Connected
Layers

Results

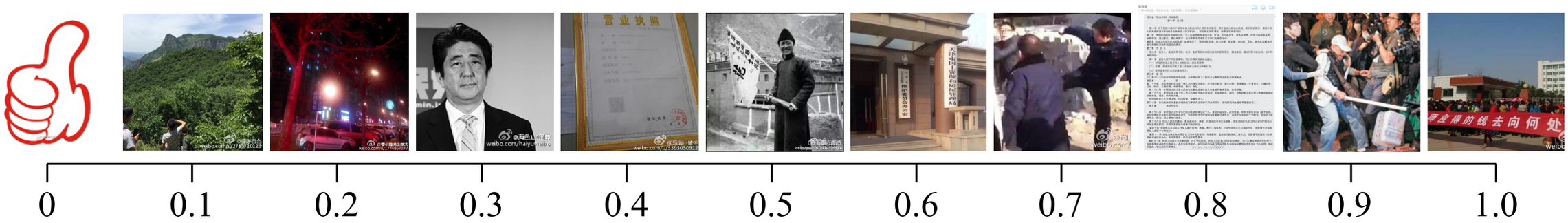


Image alone is
not enough;
texts and image
are comp

Both text and

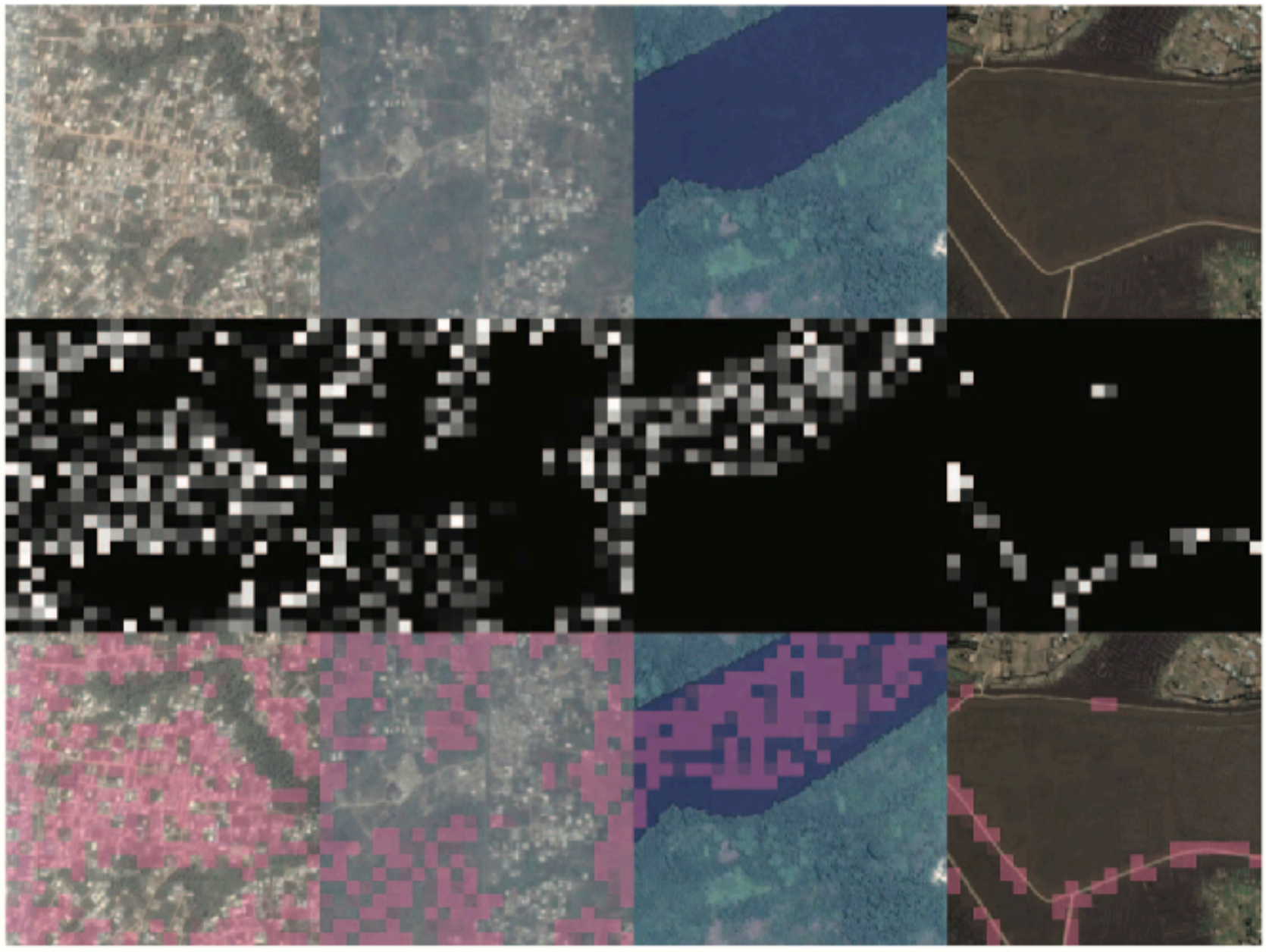
Image alone is not enough: we need both text and images.

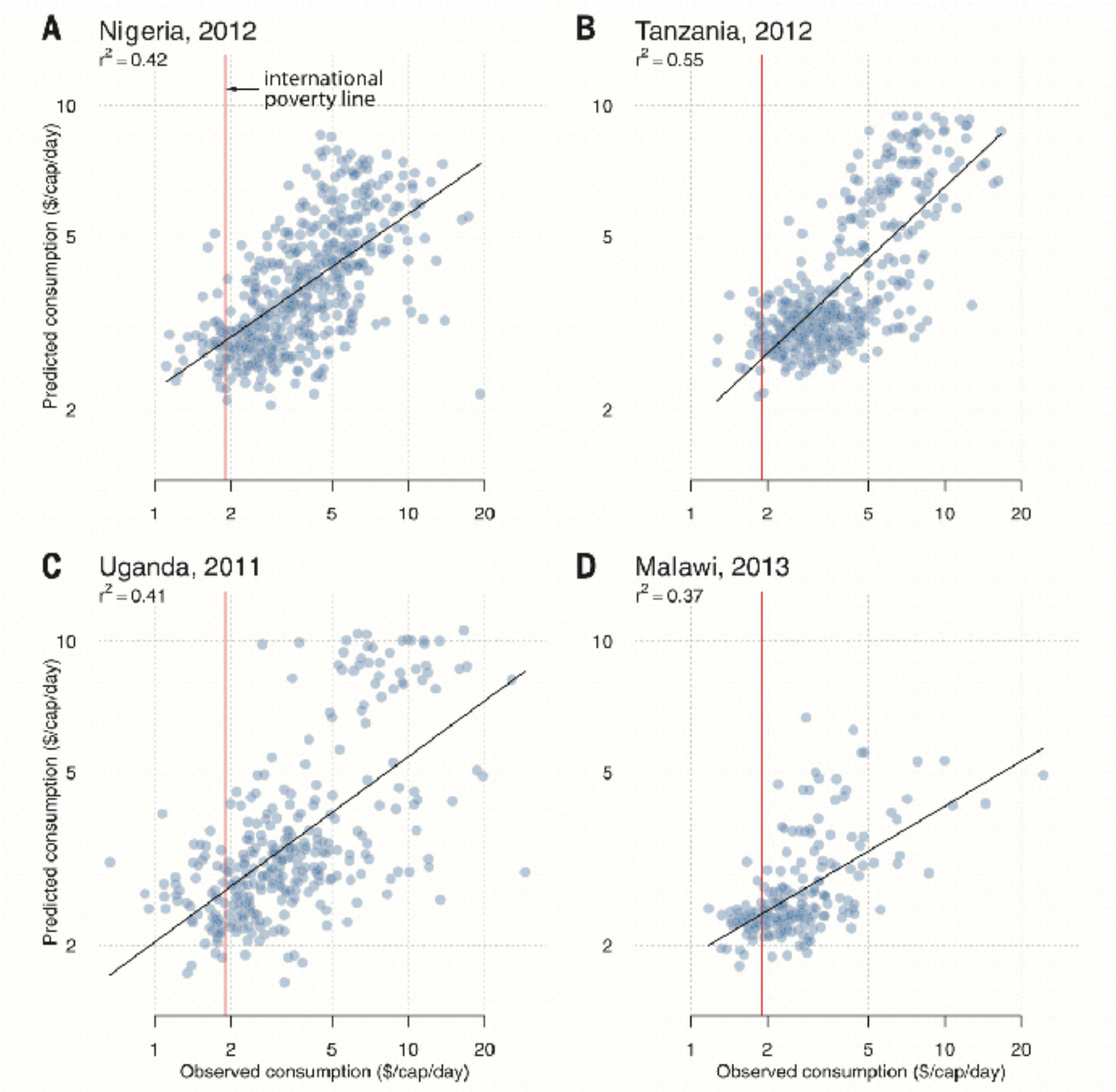
Today, high school teachers in Yanling county petition at the government offices. School masters call the police and hope to prevent teachers from going, but still over a thousand go. Hundreds of female teachers raised slogans and yell out "right protection" to attract bystanders. How shameful are those who do not join protests!

Thousands are hurt in the 2014 Shanghai stampede during the New Year's Eve. Police are coming to keep order.

Jean, Neal, Marshall Burke, Michael Xie, W. Matthew Davis, David B. Lobell, and Stefano Ermon. “Combining Satellite Imagery and Machine Learning to Predict Poverty.” Science 353, no. 6301 (August 19, 2016): 790–94. https://doi.org/10.1126/science.aaf7894.

Tanzania, Nigeria, Uganda and Malawi




A
Nigeria, 2012
$r^2 = 0.42$
international poverty line
B
Tanzania, 2012
$r^2 = 0.55$
C
Uganda, 2011
$r^2 = 0.41$
D
Malawi, 2013
$r^2 = 0.37$
Predicted consumption ($/cap/day)
Observed consumption ($/cap/day)
Observed consumption (S/cap/day)
1
2
5
10
20

Visual Data in Sociology

- Urban sociology: understanding landscapes, such as gentrification and segregation patterns

- Hwang, Jackelyn, and Robert J. Sampson. “Divergent Pathways of Gentrification: Racial Inequality and the Social Order of Renewal in Chicago Neighborhoods.” *American Sociological Review* 79, no. 4 (August 1, 2014): 726–51. https://doi.org/10.1177/0003122414535774.

- Using Google Street View to quantify different signs of gentrification over time:

    - Mix of old and new buildings
    - visible beautification efforts
    - lack of disorder and decay

Fig. 1. Modern bus stop in Chicago
Address: 1809 West Polk Street

Fig. 2. Modern public trash can in Chicago
Address: 2986 North Sheridan Road

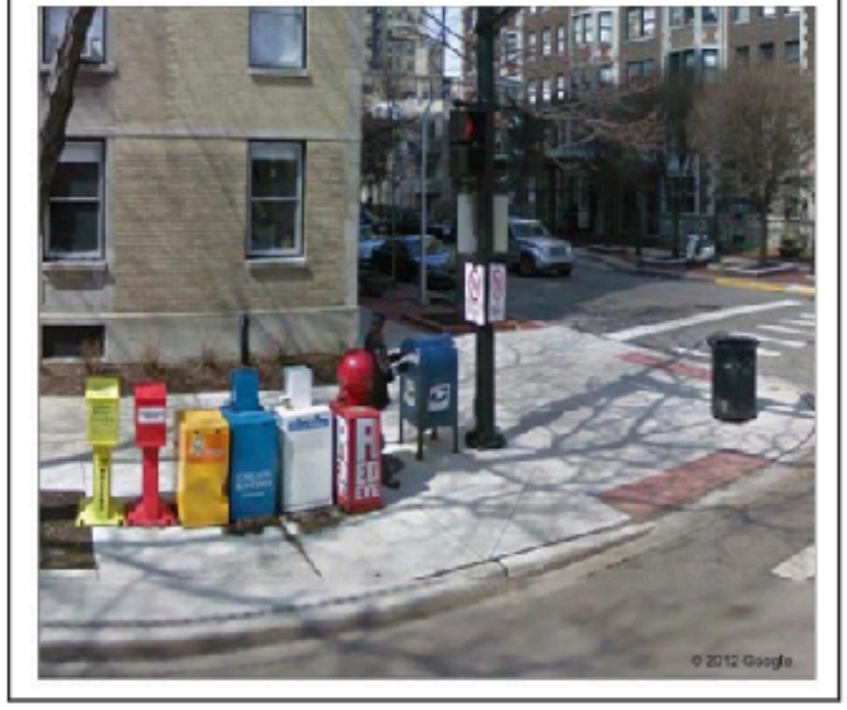


Fig. 3. Converted industrial use
Address: 1962 South Halsted Street

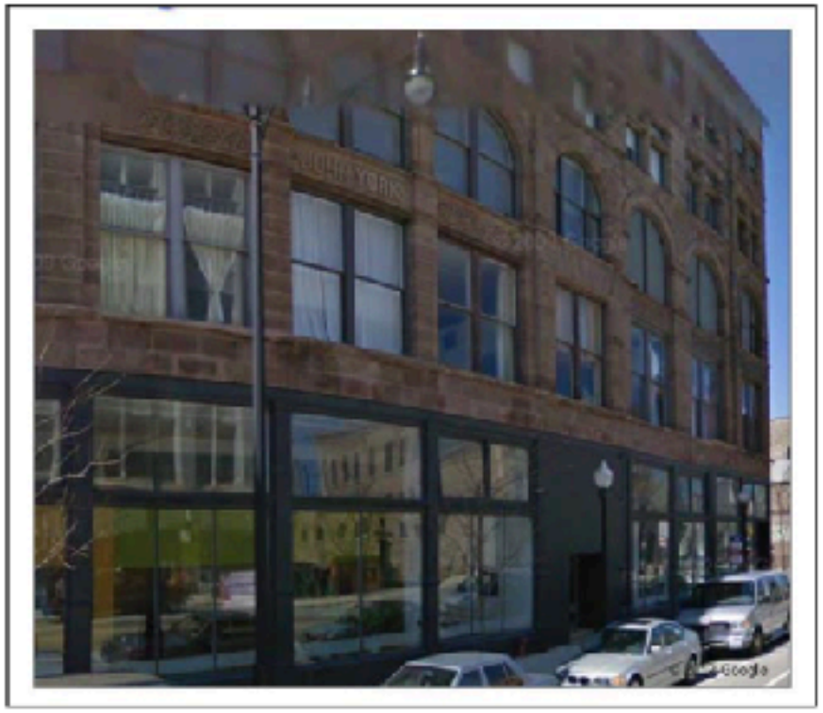
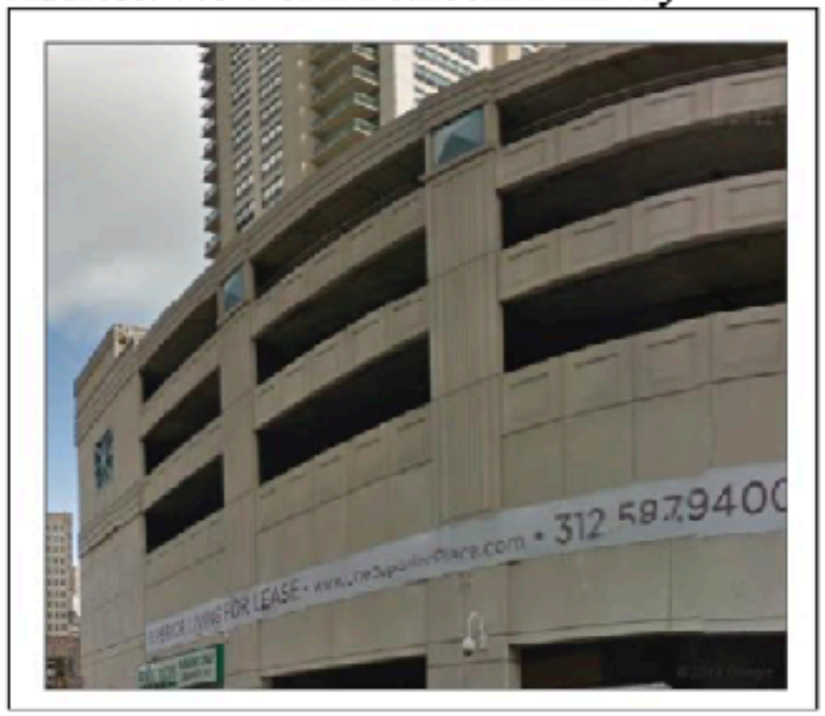


Fig. 4. Luxury high-rise condominiums
Address: 705 North Dearborn Parkway



Summary

- Promising future of visual data
- Basic steps of automated image analysis:
  - Goal: supervised vs. unsupervised
  - Tools: use existing ones vs develop your own
- Use modern tools to answer classical questions

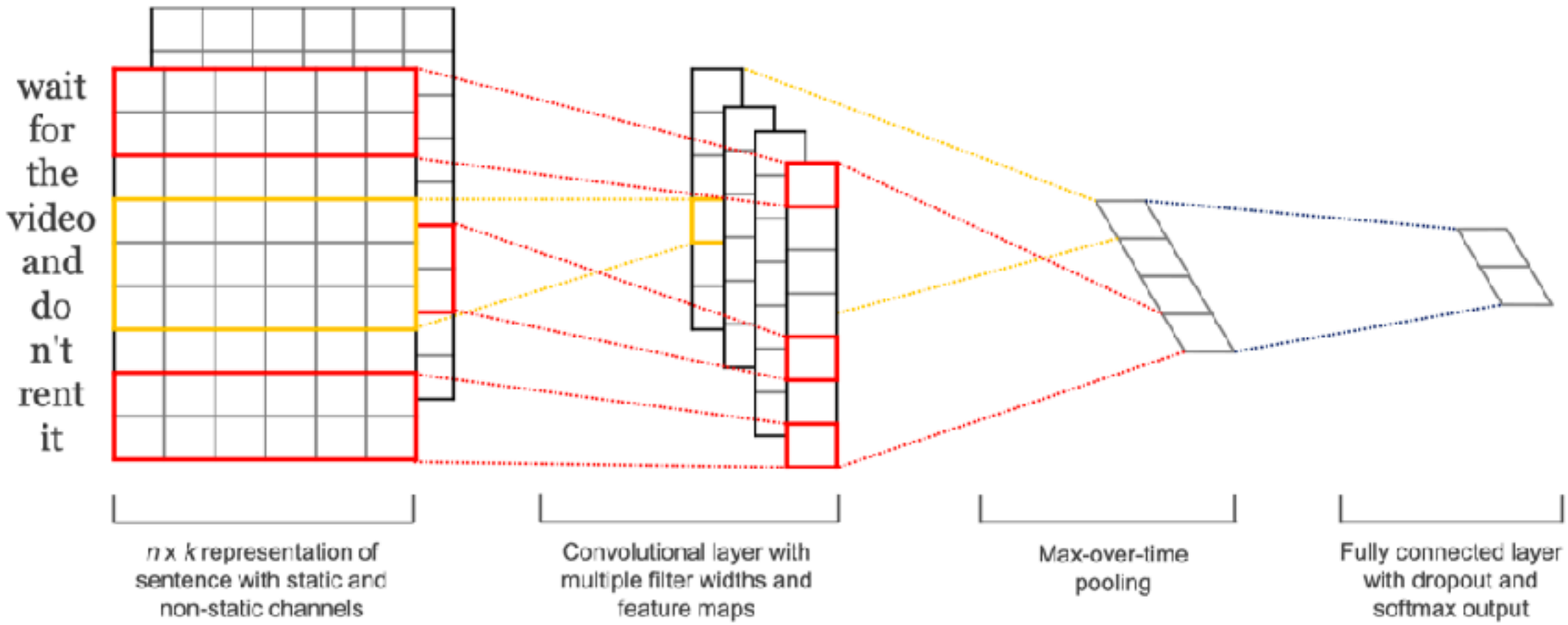
Appendix

Lots of parameter choices

- Number of layers:
  - More layers means better representation ability
  - may overfit; increase computational complexity.
- Number of feature maps: usually double by each convolutional layers.
- Filter size, stride size.
- It’s been common to follow award-winners in ImageNet challenge
  - Such AlexNet or VGG16 (especially the latter)

CNN for texts

- Text Classification



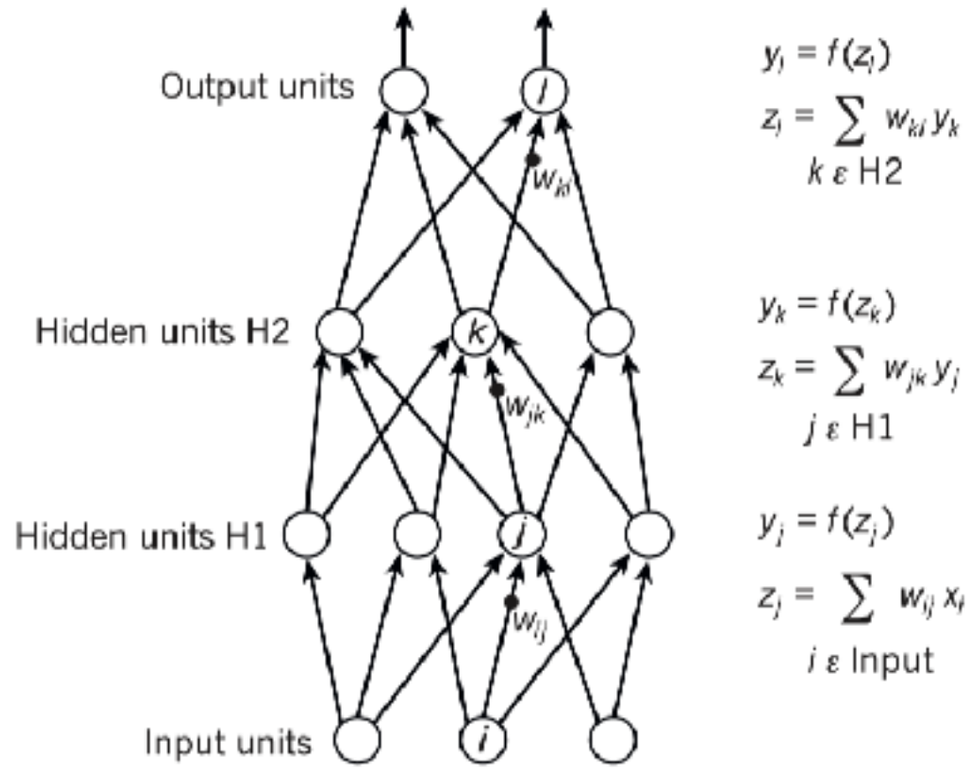
Kim,. 2014. “Convolutional Neural Networks for Sentence Classification.” ArXiv:1408.5882, August. http://arxiv.org/abs/1408.5882.

Multi-layer Neural Networks

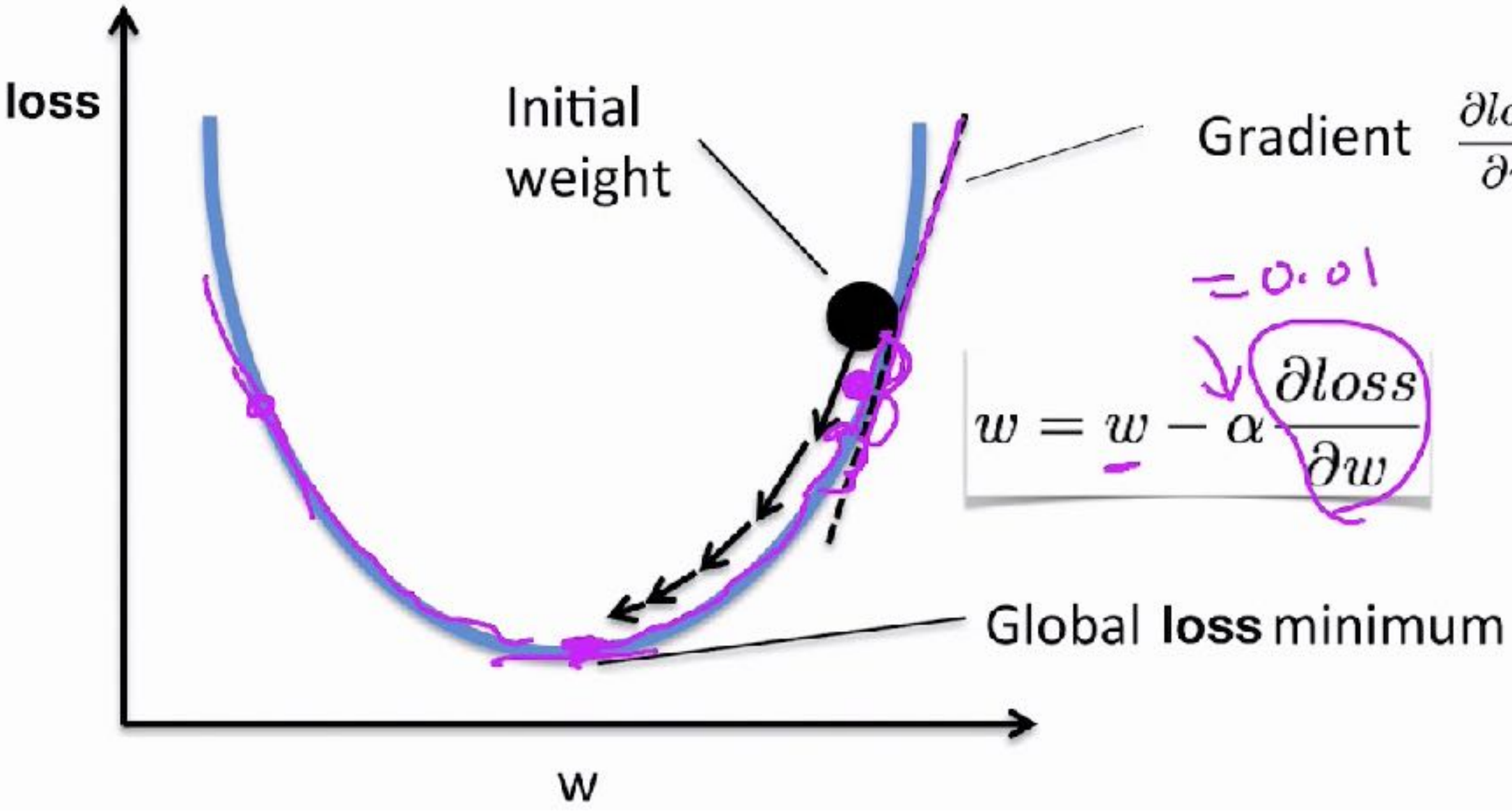
Hinton, Rumelhart, Ronald Williams, 1986

- $f()$:activation function
  - Activation function is nonlinear
- w_{ij} :weights (regression coefficients)
- z_j, z_k : activations, weighted sums of previous layer's units
- y_j, y_k : hidden units

c



Gradient descent algorithm

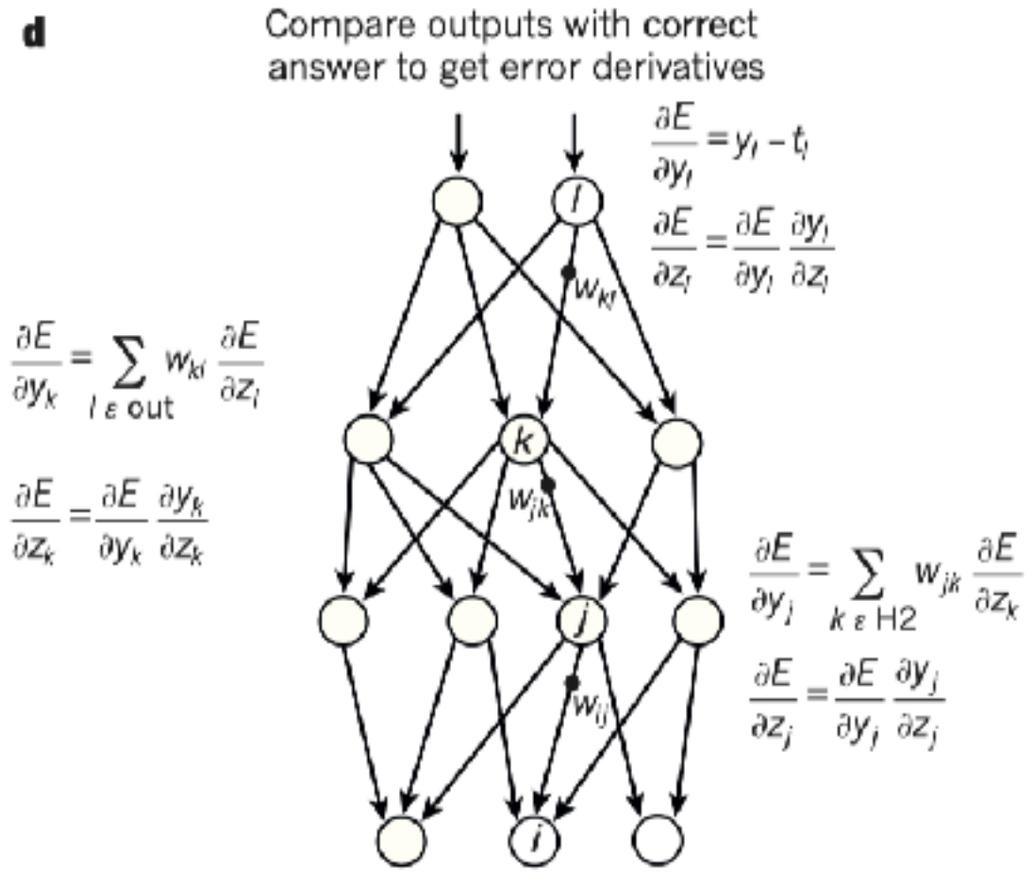


https://www.youtube.com/watch?v=b4Vyma9wPHo

Error Back Propagation

- To use gradient descent, we need to know the value of $\frac{\partial E_l}{\partial w_{kl}}$

- $\frac{\partial E_l}{\partial w_{kl}} = \frac{\partial E_l}{\partial z_l} \frac{\partial z_l}{\partial w_{kl}}$
- $\frac{\partial E_l}{\partial z_l} = \frac{\partial E_l}{\partial y_l} \frac{\partial y_l}{\partial z_l} = \frac{\partial E_l}{\partial y_l} \frac{\partial f(z_l)}{\partial z_l}$
- $\frac{\partial E_l}{\partial y_l} = \frac{\partial \frac{1}{2}(y_l - t_l)^2}{\partial z_l} = y_l - t_l$ **error**
- $\frac{\partial z_l}{\partial w_{kl}} = \frac{\partial \sum_k w_{kl} y_k}{\partial w_{kl}} = y_k$



Rumelhart, David E., Geoffrey E. Hinton, and Ronald J. Williams. "Learning representations by back-propagating errors." *Nature* 323, no. 6088 (1986): 533.

Neural Network: learning

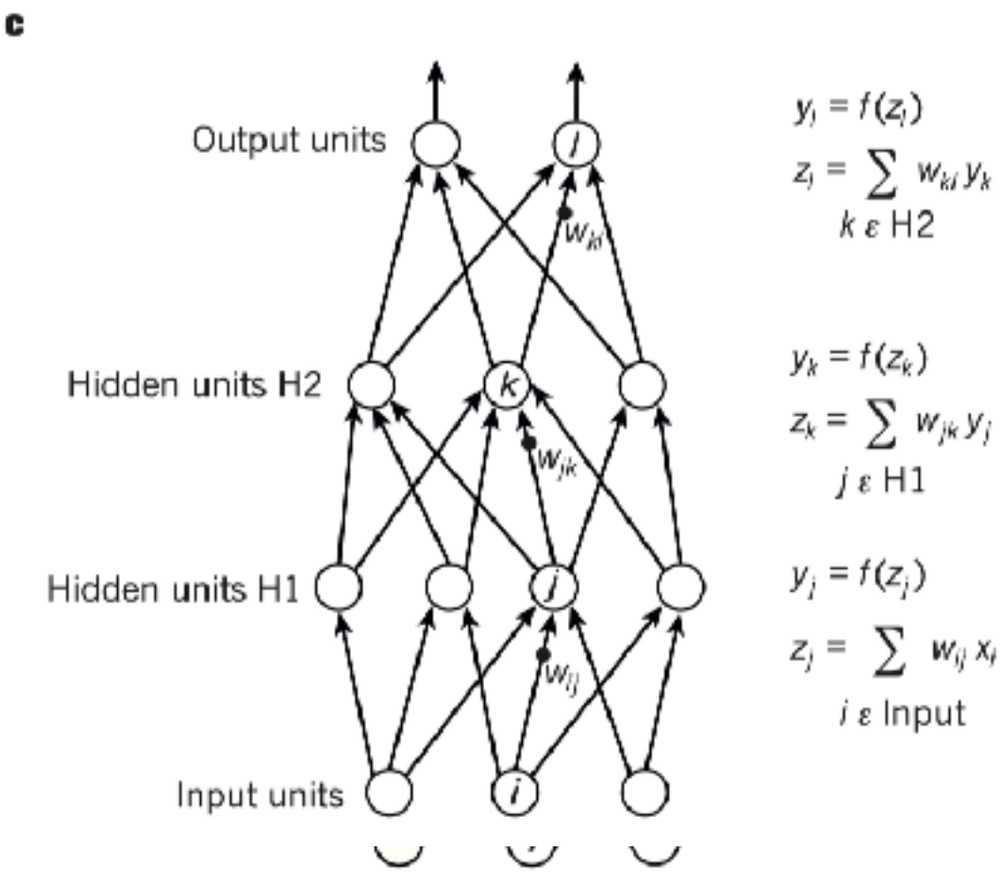
- Initialize: randomly assign some weights (often small random values around 0).

1. Forward pass: take some input units X and calculate activations of all layers

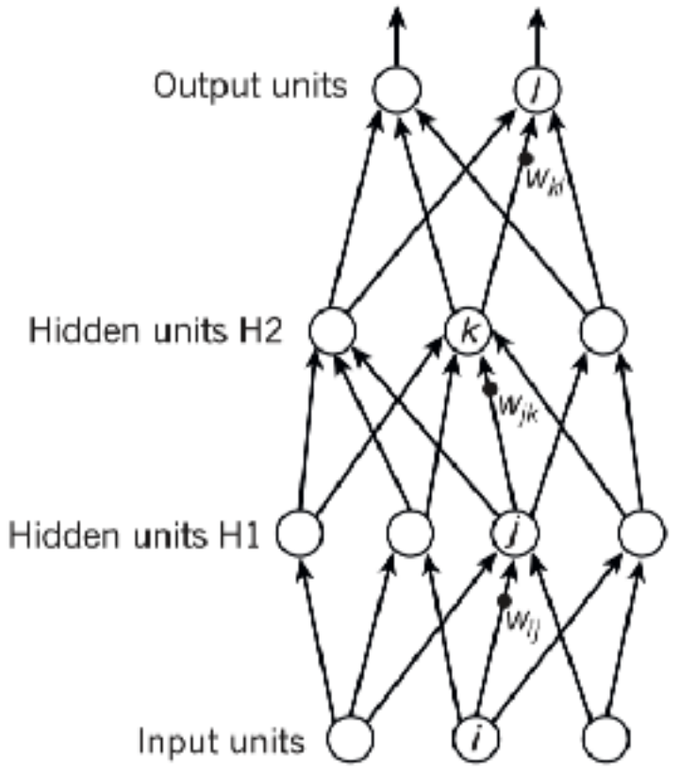
2. Back propagation: obtain partial derivatives of weights using back prop

3. Update weights using gradient descent

4. Repeat 1 - 3 until convergence.



c



$$y_l = f(z_l)$$
$$z_l = \sum_{k \,\varepsilon\, \text{H2}} w_{kl}\, y_k$$

$$y_k = f(z_k)$$
$$z_k = \sum_{j \,\varepsilon\, \text{H1}} w_{jk}\, y_j$$

$$y_j = f(z_j)$$
$$z_j = \sum_{i \,\varepsilon\, \text{Input}} w_{ij}\, x_i$$

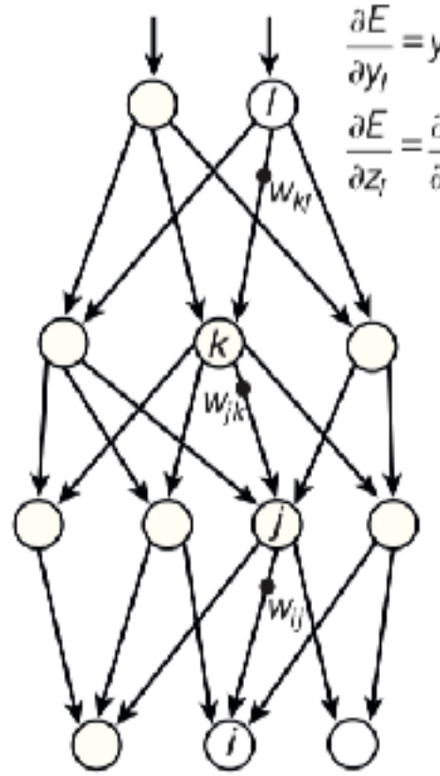
d

Compare outputs with correct answer to get error derivatives

$$\frac{\partial E}{\partial y_l} = y_l - t_l$$
$$\frac{\partial E}{\partial z_l} = \frac{\partial E}{\partial y_l}\,\frac{\partial y_l}{\partial z_l}$$

$$\frac{\partial E}{\partial y_k} = \sum_{l \,\varepsilon\, \text{out}} w_{kl}\,\frac{\partial E}{\partial z_l}$$

$$\frac{\partial E}{\partial z_k} = \frac{\partial E}{\partial y_k}\,\frac{\partial y_k}{\partial z_k}$$



$$\frac{\partial E}{\partial y_j} = \sum_{k \,\varepsilon\, \text{H2}} w_{jk}\,\frac{\partial E}{\partial z_k}$$

$$\frac{\partial E}{\partial z_j} = \frac{\partial E}{\partial y_j}\,\frac{\partial y_j}{\partial z_j}$$

- Give some observed X (here, images) and Y (outcomes, such as whether the image contains cats)

- How do we know values of

  - weights?

  - hidden units (this is our learned representation)?

Learning Weights

- Initialize: randomly assign some weights (often small random values around 0).

1. Forward pass: take some input units X and calculate values of hidden units based on initial guesses of weights
2. Back-propagation to update weights
    - If **predicted y != actual y (error in prediction)**
        - for the k-th layer: $w_{new,k} \leftarrow w_{old,k} - \alpha \cdot h_k \cdot h'_{k+1}$
    - If no prediction error, do not change weights
    - α is called learning rate; it's something scholars has to choose
3. Repeat 1 and 2 multiple times, until values of weights do not change

Learning weights

- Back-propagation: Rumelhart, David E., Geoffrey E. Hinton, and Ronald J. Williams. "Learning representations by back-propagating errors." ***Nature*** 323, no. 6088 (1986): 533.

- I am skipping the derivation of back propagation algorithms;

    - if you are interested, we can go over these at the end of the talk