Introduction to unstructured data and deep learning for social scientists

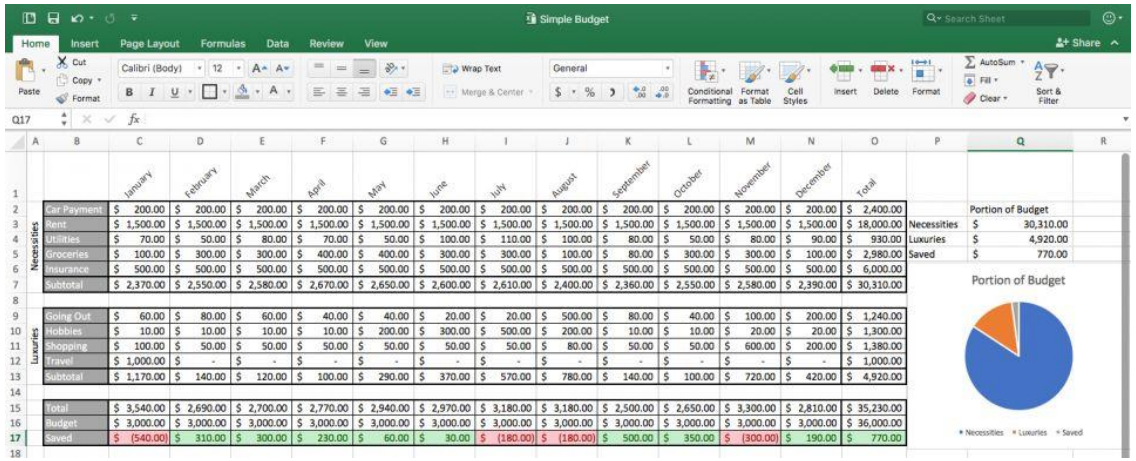
Snorre Ralund, Ph.D Fellow, SoDaS, UCPH

KØBENHAVNS UNIVERSITET

Unstructured data as the most direct data source

Classic quantitative data

National Statistics Offices, Registers, and Surveys

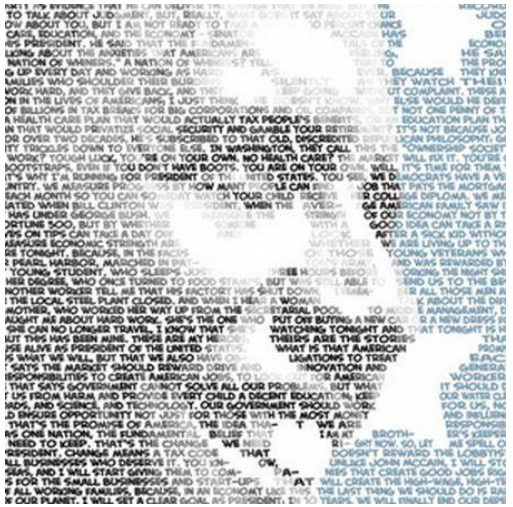


		January	February	March	April	May	June	July	August	September	October	November	December	Total
Necessities	Car Payment	$ 200.00	$ 200.00	$ 200.00	$ 200.00	$ 200.00	$ 200.00	$ 200.00	$ 200.00	$ 200.00	$ 200.00	$ 200.00	$ 200.00	$ 2,400.00
	Rent	$ 1,500.00	$ 1,500.00	$ 1,500.00	$ 1,500.00	$ 1,500.00	$ 1,500.00	$ 1,500.00	$ 1,500.00	$ 1,500.00	$ 1,500.00	$ 1,500.00	$ 1,500.00	$ 18,000.00
	Utilities	$ 70.00	$ 50.00	$ 80.00	$ 70.00	$ 50.00	$ 100.00	$ 110.00	$ 100.00	$ 80.00	$ 50.00	$ 80.00	$ 90.00	$ 930.00
	Groceries	$ 100.00	$ 300.00	$ 300.00	$ 400.00	$ 400.00	$ 300.00	$ 300.00	$ 100.00	$ 80.00	$ 300.00	$ 300.00	$ 100.00	$ 2,980.00
	Insurance	$ 500.00	$ 500.00	$ 500.00	$ 500.00	$ 500.00	$ 500.00	$ 500.00	$ 500.00	$ 500.00	$ 500.00	$ 500.00	$ 500.00	$ 6,000.00
	Subtotal	$ 2,370.00	$ 2,550.00	$ 2,580.00	$ 2,670.00	$ 2,650.00	$ 2,600.00	$ 2,610.00	$ 2,400.00	$ 2,360.00	$ 2,550.00	$ 2,580.00	$ 2,390.00	$ 30,310.00
Luxuries	Going Out	$ 60.00	$ 80.00	$ 60.00	$ 40.00	$ 40.00	$ 20.00	$ 20.00	$ 500.00	$ 80.00	$ 40.00	$ 100.00	$ 200.00	$ 1,240.00
	Hobbies	$ 10.00	$ 10.00	$ 10.00	$ 10.00	$ 200.00	$ 300.00	$ 500.00	$ 200.00	$ 10.00	$ 10.00	$ 20.00	$ 20.00	$ 1,300.00
	Shopping	$ 100.00	$ 50.00	$ 50.00	$ 50.00	$ 50.00	$ 50.00	$ 50.00	$ 80.00	$ 50.00	$ 50.00	$ 600.00	$ 200.00	$ 1,380.00
	Travel	$ 1,000.00	$ -	$ -	$ -	$ -	$ -	$ -	$ -	$ -	$ -	$ -	$ -	$ 1,000.00
	Subtotal	$ 1,170.00	$ 140.00	$ 120.00	$ 100.00	$ 290.00	$ 370.00	$ 570.00	$ 780.00	$ 140.00	$ 100.00	$ 720.00	$ 420.00	$ 4,920.00
	Total	$ 3,540.00	$ 2,690.00	$ 2,700.00	$ 2,770.00	$ 2,940.00	$ 2,970.00	$ 3,180.00	$ 3,180.00	$ 2,500.00	$ 2,650.00	$ 3,300.00	$ 2,810.00	$ 35,230.00
	Budget	$ 3,000.00	$ 3,000.00	$ 3,000.00	$ 3,000.00	$ 3,000.00	$ 3,000.00	$ 3,000.00	$ 3,000.00	$ 3,000.00	$ 3,000.00	$ 3,000.00	$ 3,000.00	$ 36,000.00
	Saved	$ (540.00)	$ 310.00	$ 300.00	$ 230.00	$ 60.00	$ 30.00	$ (180.00)	$ (180.00)	$ 500.00	$ 350.00	$ (300.00)	$ 190.00	$ 770.00

Portion of Budget	
Necessities	$ 30,310.00
Luxuries	$ 4,920.00
Saved	$ 770.00

Raw unstructured data

Collections of Text, Images and Sound

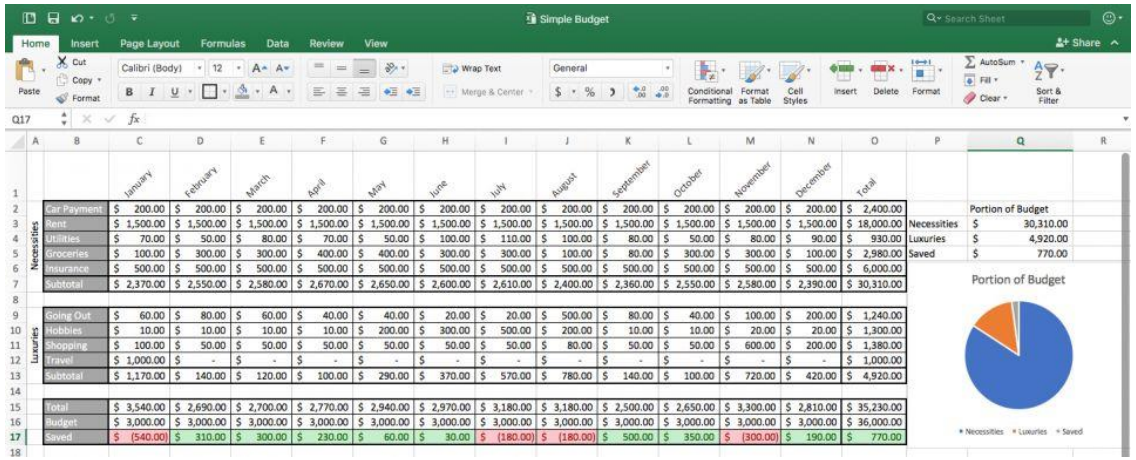


Unstructured data as the most direct data source

Classic quantitative data

National Statistics Offices, Registers, and Surveys

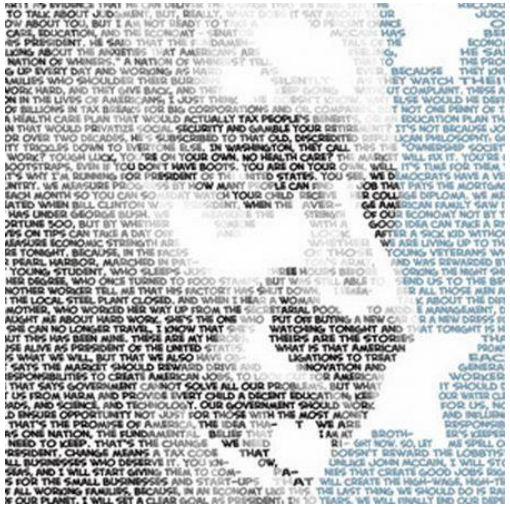
- Distant: From the lived life
- Predetermined by other party
- Reduced to a number



Raw unstructured data

Collections of Text, Images and Sound

- Close to the lived life
- Undetermined
- Rich in information



Raw traces of society

Example: Street view raw data

Example: Street view raw data

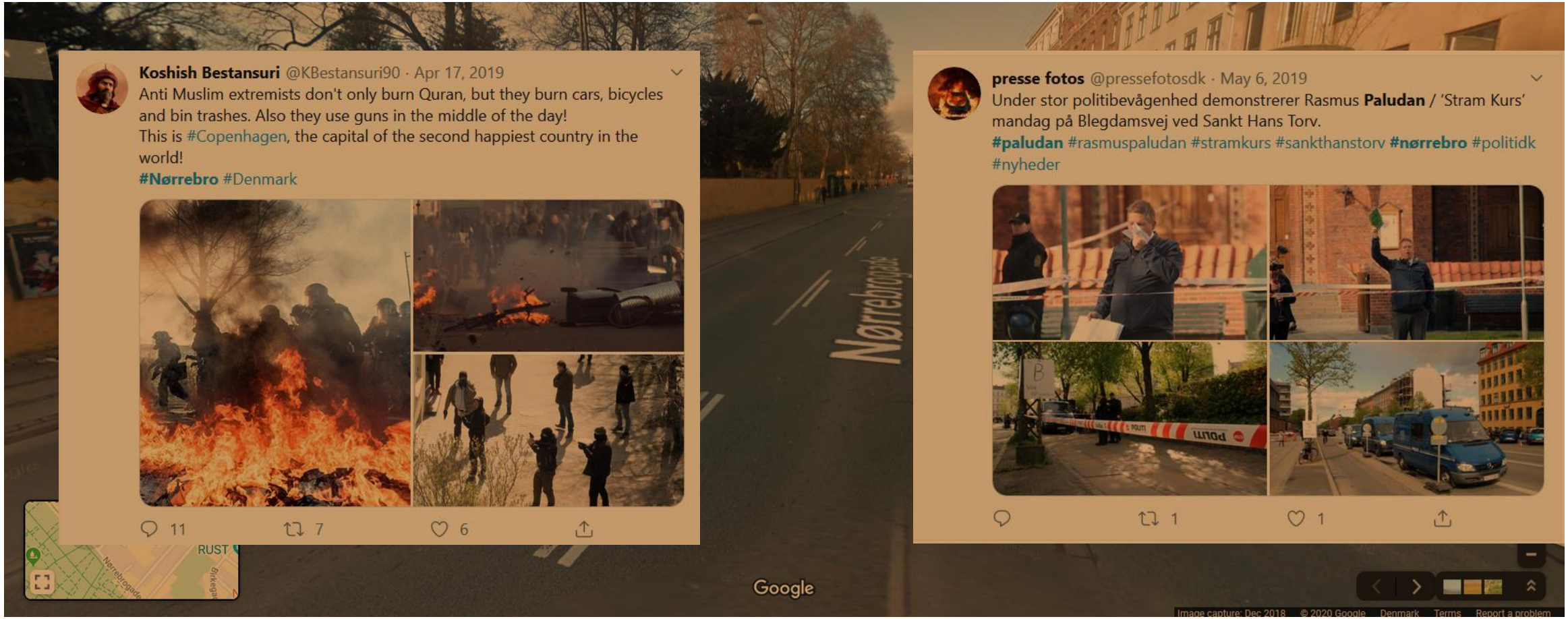


Image data in social science research

- Gebru, Timnit, et al. 2017: "Using deep learning and Google Street View to estimate the demographic makeup of neighborhoods across the United States."
  - High granularity measurement of income, race, education, and voting patterns.
- Gender in the presentation of jobs: https://www.pewsocialtrends.org/2018/12/17/gender-and-jobs-in-online-image-searches/
  - Relates to the paper on gendered discourses around jobs:
    - Bolukbasi, Tolga, et al. "Man is to computer programmer as woman is to homemaker? debiasing word embeddings." *Advances in neural information processing systems*. 2016.

Unstructured data processed by complex heuristics

Text, images and sound are complex data forms.

- Needs new methods

Unstructured data processed by complex heuristics

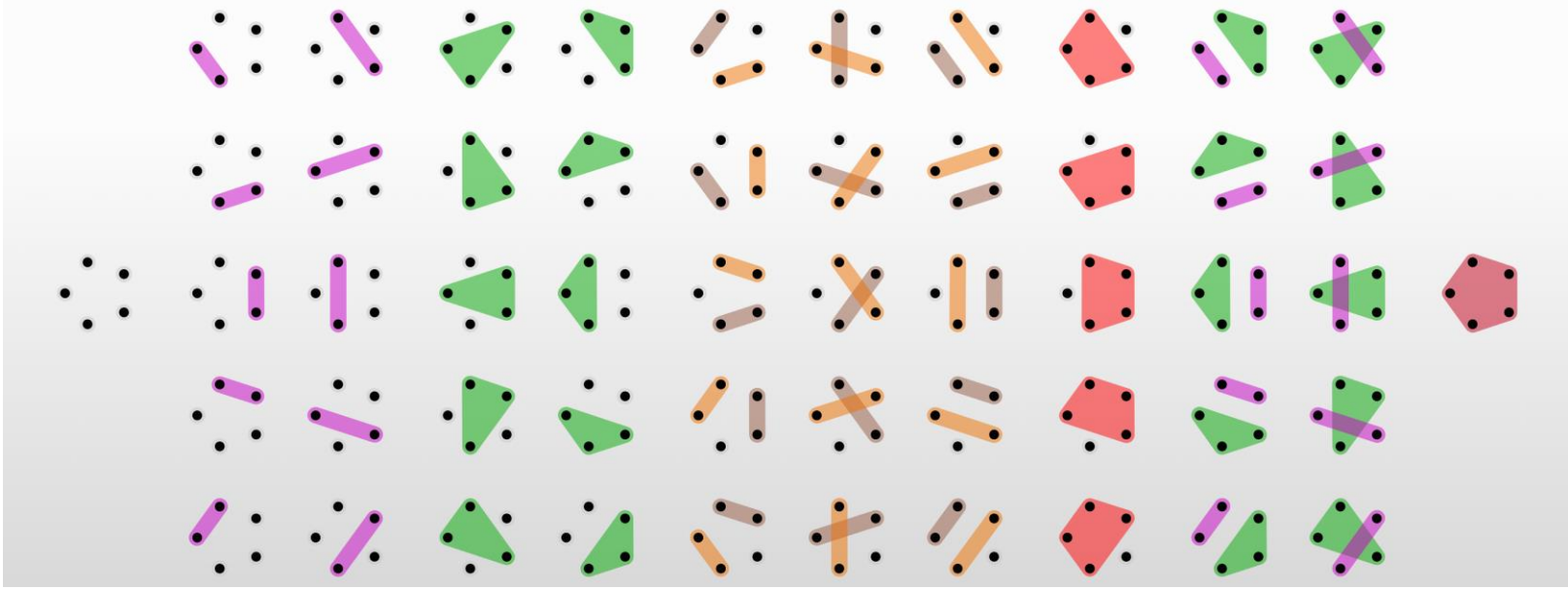
Text, images and sound are complex data forms.

- Rules are hard to explicate and combinatorial space is extreme.
  - Epistatic, sequential and compositional features.
    - E.g “He really do like me.”, “He Really do not like me, does he?”

Unstructured data processed by complex heuristics

Text, images and sound are complex data forms.

- Rules are hard to explicate and combinatorial space is extreme.
  - Epistatic, sequential and compositional features.
    - E.g “He really do like me.”, “He Really do not like me, does he?”



Unstructured data processed by complex heuristics

Text, images and sound are complex data forms.

Classic approach

- Feature Extraction (unsupervised learning)
- Hand-coded rules complex feature extraction and very domain specific models + expertise

Deep learning approach

- Feature Learning (supervised)
- Generic framework
- Model Capacity: Network structure (Size, Layers, and connection structure), and neuron type.

Unstructured data processed by complex heuristics

Text, images and sound are complex data forms.

TABLE I
DIFFERENT FEATURE LEARNING APPROACHES

Approaches	**Learning steps**				
Rule based	Input	Hand-design features	Output		
Traditional Machine Learning	Input	Hand-design features	Mapping from features	Output	
Representation Learning	Input	Features	Mapping from features	Output	
Deep Learning	Input	Simple features	Complex features	Mapping from features	Output

Alom et. al 2018

Unstructured data processed by complex heuristics

Text, images and sound are complex data forms.

- **The rise of Deep Learning:**

AlexNet (LSVRC 2012) ImageNet Competition

Team name	Filename	Error (5 guesses)	Description
SuperVision	test-preds-141-146.2009-131-137-145-146.2011-145f.	0.15315	Using extra training data from ImageNet Fall 2011 release
SuperVision	test-preds-131-137-145-135-145f.txt	0.16422	Using only supplied training data
ISI	pred_FVs_wLACs_weighted.txt	0.26172	Weighted sum of scores from each classifier with SIFT+FV, LBP+FV, GIST+FV, and CSIFT+FV, respectively.

← SOTA on Imagenet 2012

← Computer vision Classic

Unstructured data processed by complex heuristics

Text, images and sound are complex data forms.

- **The rise of Deep Learning:**

AlexNet (LSVRC 2012) ImageNet Competition

We use multi-class online learning and late fusion techniques with multiple image features.

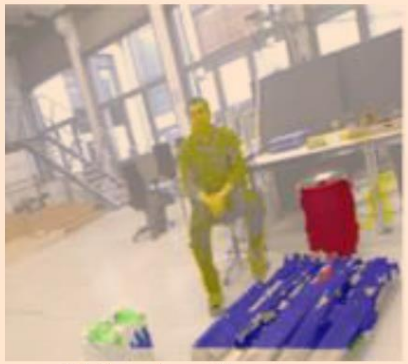
We extract conventional Fisher Vectors (FV) [Sanchez et al., CVPR 2011] and streamlined version of Graphical Gaussian Vectors (GGV) [Harada, NIPS 2012]. For extraction, we use not only common SIFT and CSIFT, but also LBP and GIST in a dense-sampling manner.

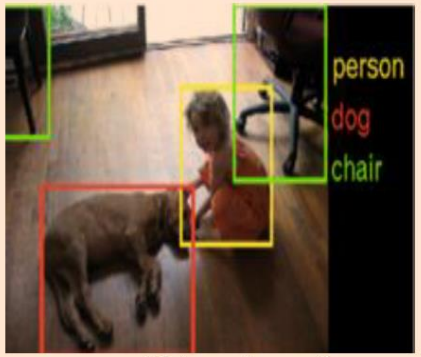
We train linear classifiers using Passive-Aggressive (PA) algorithm [Crammer et al., JMLR 2006].

Our model is a large, deep convolutional neural network trained on raw RGB pixel values. The neural network, which has 60 million parameters and 650,000 neurons, consists of five convolutional layers, some of which are followed by max-pooling layers, and three globally-connected layers with a final 1000-way softmax. It was trained on two NVIDIA GPUs for about a week. To make training faster, we used non-saturating neurons and a very efficient GPU implementation of convolutional nets. To reduce overfitting in the globally-connected layers we employed hidden-unit "dropout", a recently-developed regularization method that proved to be very effective.

Deep learning dominate unstructured data

- Sound to text, Computer Vision, NLP Annotation of video and images.


Object localization



Object detection


Image or video Segmentation


Autonomous Car

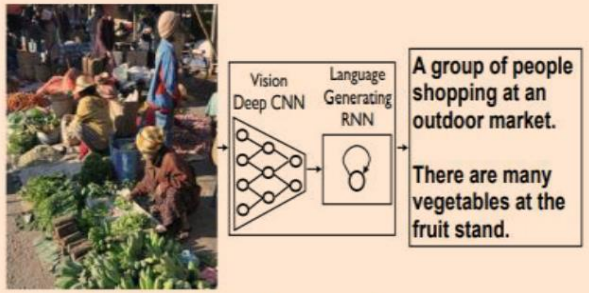
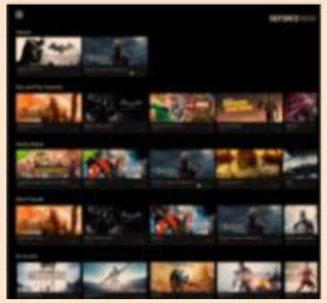


Image or Video Captioning


Media and entertainment

Machine translation

Speech recognition

Alom et. al 2018

Deep Development

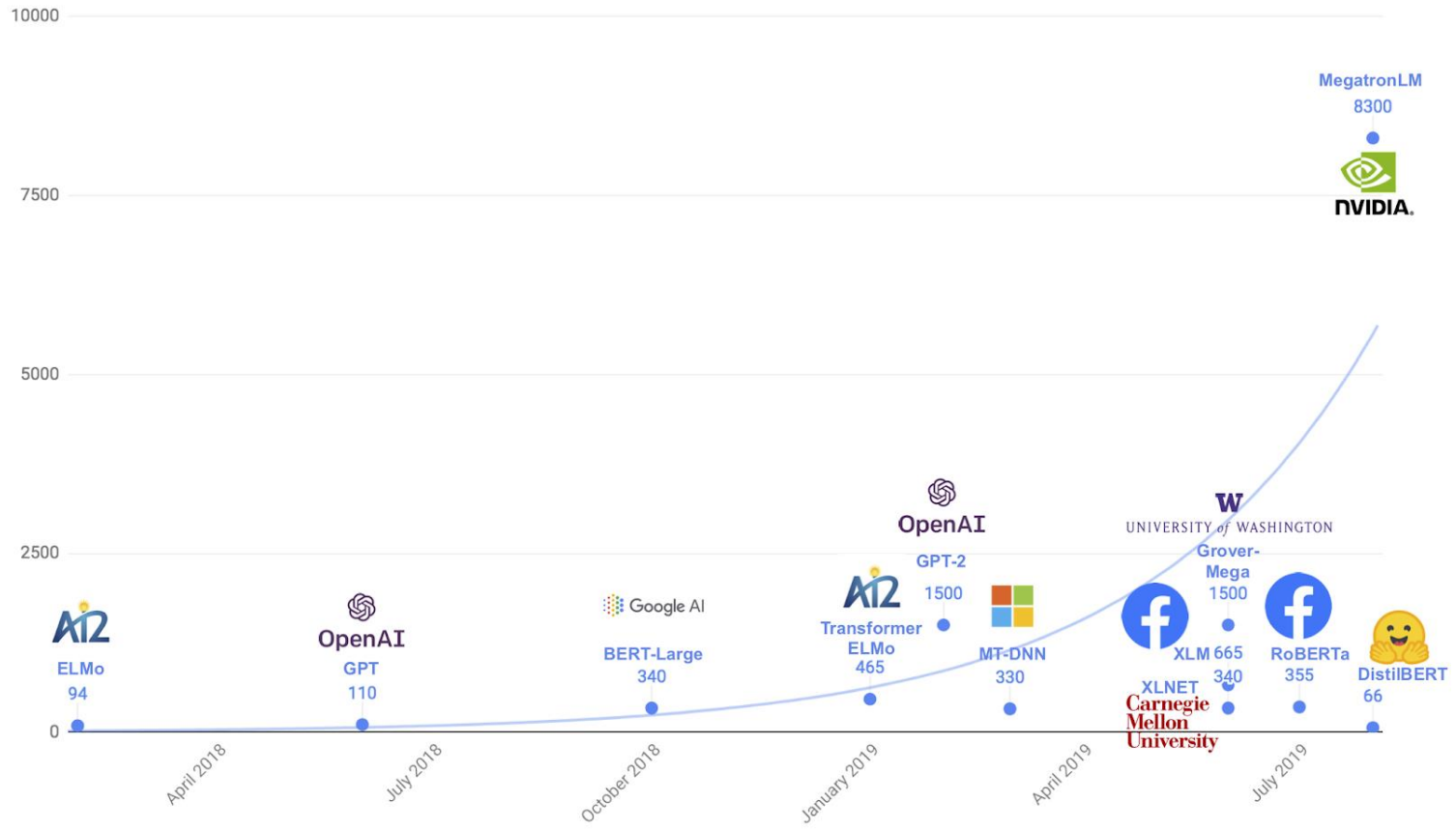
- Objectives: Efficient training, reliable optimization, model capacity.
- Architectures:
  - Convolutional Neural Network (CNN), Recurrent Neural Network (RNN) including Long Short Term Memory (LSTM), Auto-Encoder (AE), Generative Adversarial Network (GAN), Transformers.
- Optimizers:
  - SGD, Momentum, Adam, Adagrad
- Activation functions:
  - Sigmoid, TanH, ReLu, swish, mish,
- Regularization:
  - Dropout, Skipconnections, BatchNorm, Limited or mixed precision training.

Deep Development

Transfer Learning

Big models needs big data.

Models can be “recycled”!



Deep Development

Transfer Learning

Big models needs big data.

Models can be “recycled”

- Transfer learning CV: Features learned from imagenet generalize (Donahue et al., 2014)
- Transfer learning NLP: Word2Vec(Mikolov et al 2013), ELMO (Peters et al 2018), BERT (Devlin et al. 2018) , ULMFit (Ruder and Howard 2018)

Computer scientists and Social Scientist

How to appropriately appropriate methods from computer science?

- They have really fancy names and abbreviations.
- They compete and win competitions.

kaggle

177 Grandmasters | 1,381 Masters | 5,329 | 53,283 Contributors | 72,120 Novices

Rank	Tier	User		Medals	Points
1		bestfitting	joined 3 years ago	26 4 0	257,414
2		Guanshuo Xu	joined 4 years ago	12 15 2	227,712
3		Giba	joined 7 years ago	48 38 26	151,224

stackoverflow

Jon Skeet 739 8346 8690 member for: 11 years, 4 months	#11 week rank	+3 change	1,160,760 total reputation
VonC 364 3121 3675 member for: 11 years, 4 months	#9 week rank	+3 change	926,255 total reputation
BalusC 321 3330 3336 member for: 10 years, 5 months	#29 week rank	+2 change	911,327 total reputation

Methodological Differences: I

Computer Scientist

Evaluation of a Heuristic / Software Tool

- Serving a User
- Query -- Usefullness

Social Scientist

Evaluation of a Measurement device.

- Serving a research project.
- Sample -- Population

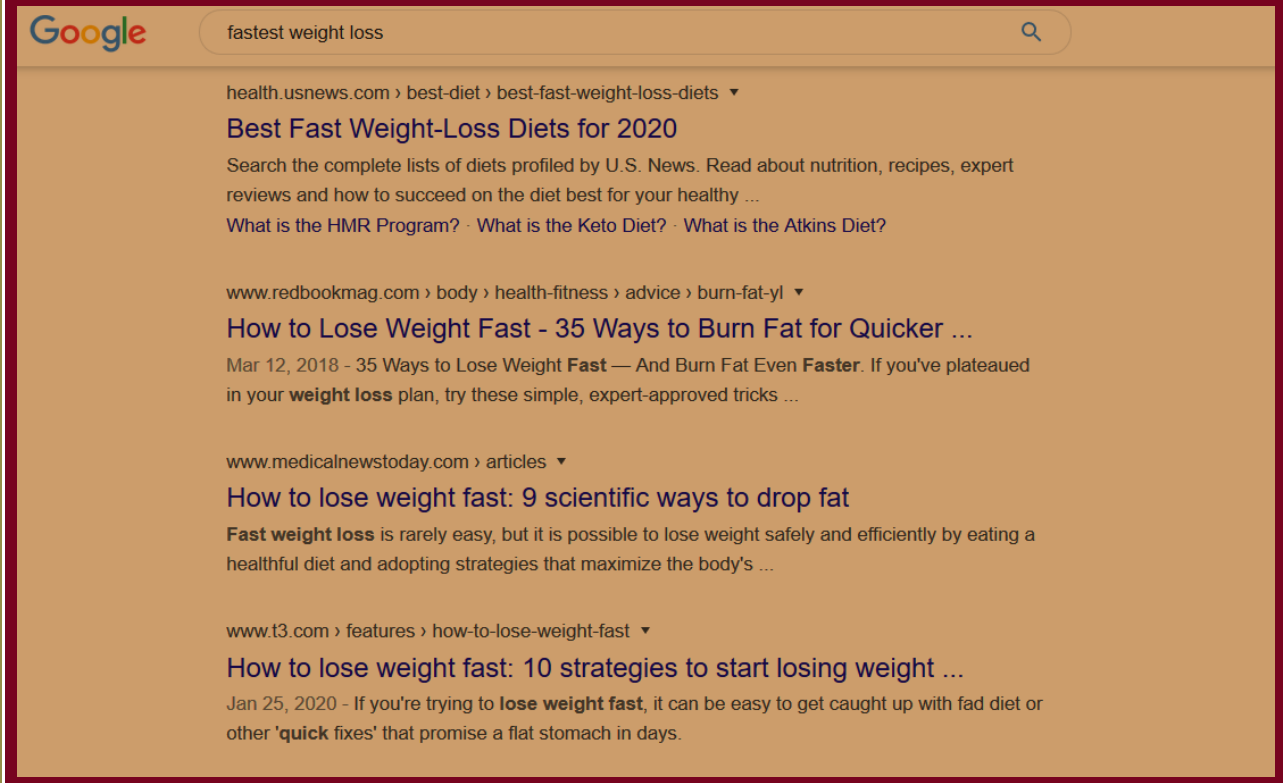
Methodological Differences: I

Computer Scientist

Evaluation of a Heuristic / Software Tool

Example the Search Engine.

- 1. **Huge** database of unstructured text and html. **User** needs to find a needle in a haystack, a site around weight loss.



Methodological Differences: I

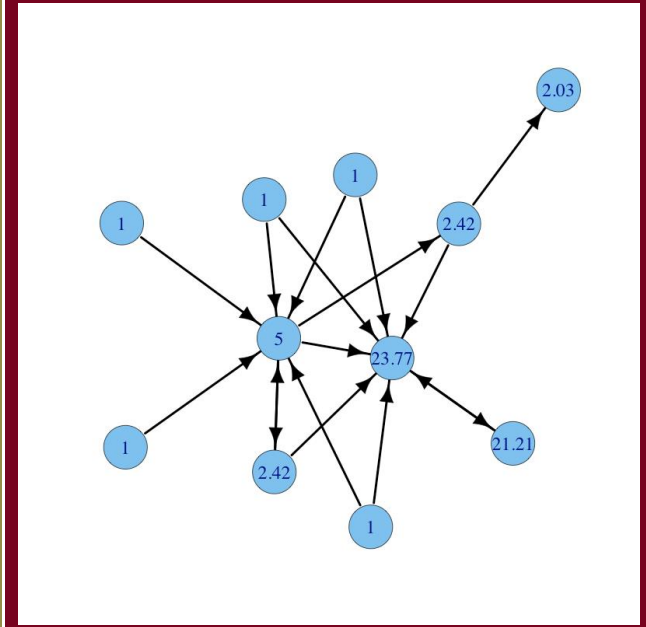
Computer Scientist

Evaluation of a Heuristic / Software Tool

Example the Search Engine.

- 1. **Huge** database of unstructured text and html. **User** needs to find a needle in a haystack, a site around weight loss.

- 2. Use Network analysis and Pagerank to locate most popular pages based on “inlinks”.



Methodological Differences: I

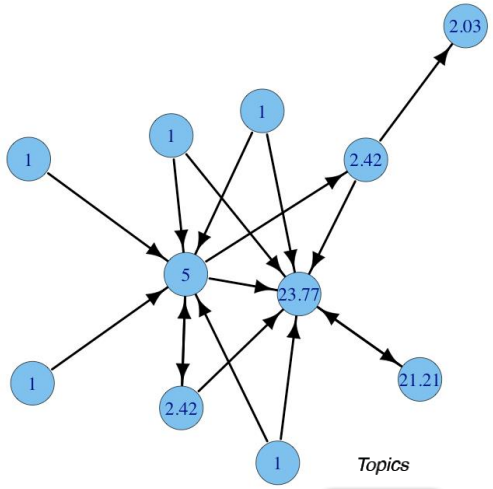
Computer Scientist

Evaluation of a Heuristic / Software Tool

Example the Search Engine.

- 1. **Huge** database of unstructured text and html. **User** needs to find a needle in a haystack, a site around weight loss.

- 2. Use Network analysis and Pagerank to locate most popular pages based on “inlinks”.
- 3. Cluster texts based on similarity or Bayesian probabilistic models: Tf-Idf, LSI, LDA.

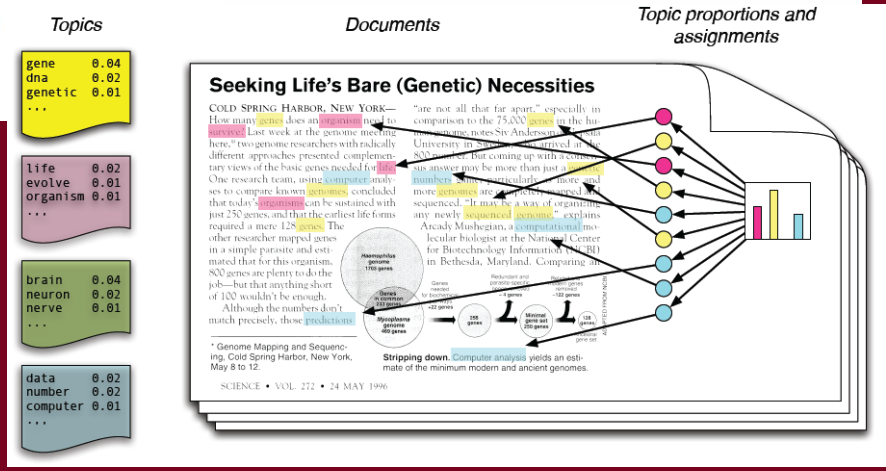


$$w_{i,j} = tf_{i,j} \times \log\left(\frac{N}{df_i}\right)$$

tf_{ij} = number of occurrences of i in j

df_i = number of documents containing i

N = total number of documents



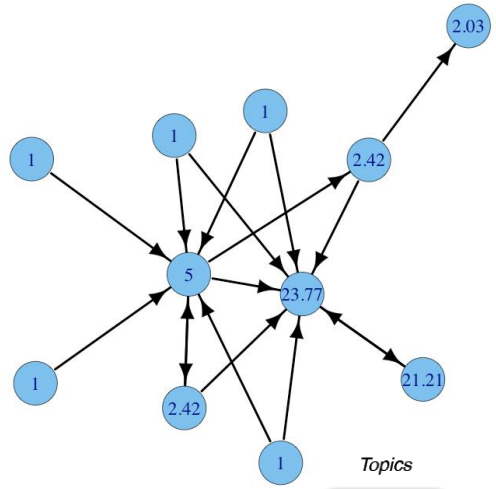
Methodological Differences: I

Computer Scientist

Evaluation of a Heuristic / Software Tool

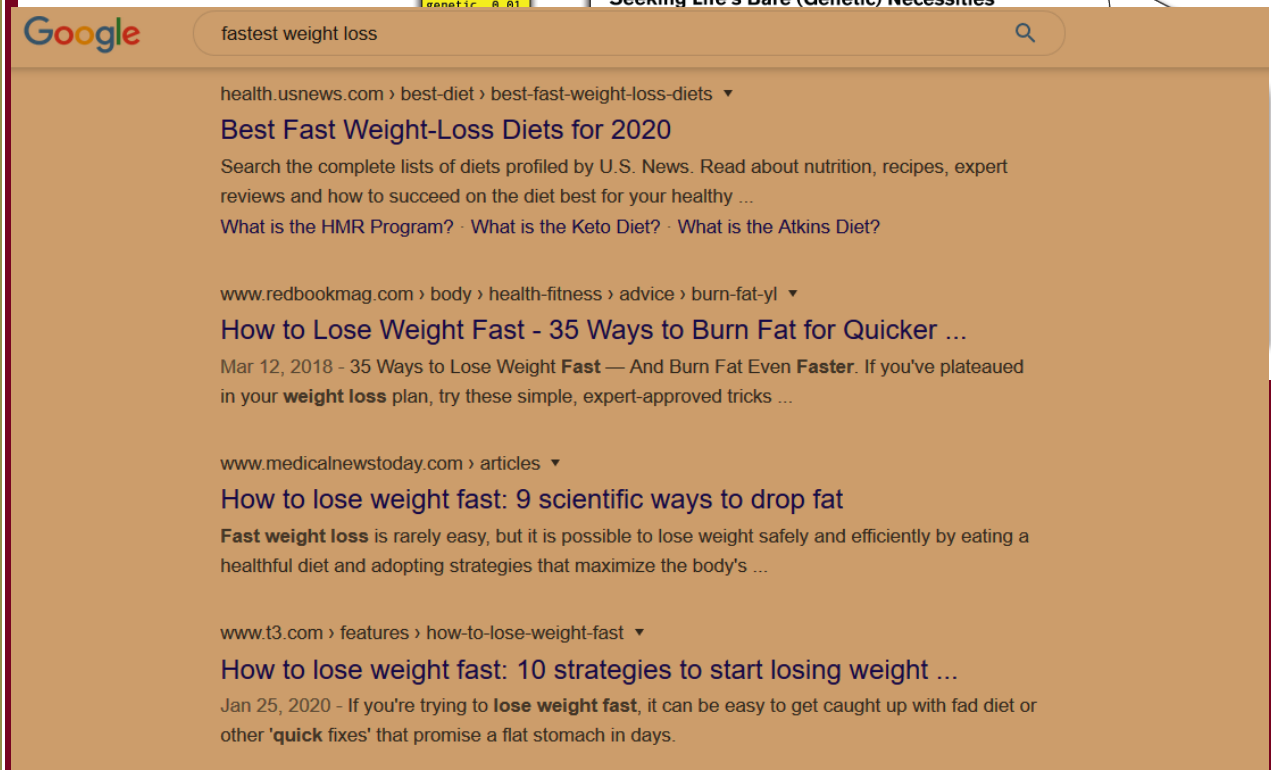
Example the Search Engine.

- 1. **Huge** database of unstructured text and html. **User** needs to find a needle in a haystack, a site around weight loss.
- 2. Use Network analysis and Pagerank to locate most popular pages based on “inlinks”.
- 3. Cluster texts based on similarity or Bayesian probabilistic models: Tf-Idf, LSI, LDA.
- 4. Match search with cluster and present most popular as top results.



$$w_{i,j} = tf_{i,j} \times \log\left(\frac{N}{df_i}\right)$$

tf_{ij} = number of occurrences of i in j
df_i = number of documents containing i
N = total number of documents



Methodological Differences: I

Computer Scientist

Evaluation of a Heuristic / Software Tool

- Serving a User
- Query -- Usefullness
- Example the Search Engine.

Social Scientist

Evaluation of a Measurement device.

- Serving a research project.
- Sample -- Population
- Example: Population of all “weight loss” sites. (Top Sites versus All results).
  - How many are there?
  - Which are growing?

Methodological Differences: II

Computer Scientist

Performance of my Algorithm

- Accuracy and State-of-the-art

TREND	DATASET	BEST METHOD	PAPER TITLE	PA
	SST-2 Binary classification	T5-3B	Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer	
	IMDb	NB-weighted-BON + dv-cosine	Sentiment Classification Using Document Embeddings Trained with Cosine Similarity	
	SST-5 Fine-grained classification	BERT large	Fine-grained Sentiment Classification using BERT	
	Yelp Binary classification	BERT large	Unsupervised Data Augmentation	
	Yelp Fine-grained classification	BERT large	Unsupervised Data Augmentation	

Methodological Differences: II

Computer Scientist

Performance of my Algorithm

- Accuracy and State-of-the-art

TREND	DATASET	BEST METHOD	PAPER TITLE	PA
	SST-2 Binary classification	T5-3B	Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer	
	IMDb	NB-weighted-BON + dv-cosine	Sentiment Classification Using Document Embeddings Trained with Cosine Similarity	
	SST-5 Fine-grained classification	BERT large	Fine-grained Sentiment Classification using BERT	
	Yelp Binary classification	BERT large	Unsupervised Data Augmentation	
	Yelp Fine-grained classification	BERT large	Unsupervised Data Augmentation	

Social Scientist

Performance of my Measurement device.

- Calibration and Error Correction (Hopkins and King 2010, Wiedemann 2018, Jerzak et. al Forthcoming)
- Differential Bias of the measurement device variables of interest (social groups, neighborhoods, countries etc).

Methodological Differences: III

Computer Scientist

Theoretical Problem

- Given a (set of) datasets to test.
- Optimize Efficiency

Conditions

- Often Ideal.

Ressources

- Extensive model search and hyperparameter optimization for proving a theoretical point.

Social Scientist

Practical Problem

- Construction of Category Scheme and Training Data. Lessons from Krippendorf 2018.

Conditions

- Unbalanced classes and (extremely) rare cases.
  - E.g. Batchsize for performance rather than efficiency, to minimize “Catastrophic Forgetting”.

Ressources

- Sparse ressources to get a specific model working. Focus on Calibration rather than SOTA.

Methodological Differences: III

Computer Scientist

Theoretical Problem

Ressources

- Extensive model search and hyperparameter optimization for proving a theoretical point.

Consumption	CO_2e (lbs)
Air travel, 1 passenger, NY↔SF	1984
Human life, avg, 1 year	11,023
American life, avg, 1 year	36,156
Car, avg incl. fuel, 1 lifetime	126,000
Training one model (GPU)	
NLP pipeline (parsing, SRL)	39
w/ tuning & experimentation	78,468
Transformer (big)	192
w/ neural architecture search	626,155

(https://arxiv.org/pdf/1906.02243.pdf)

Social Scientist

Practical Problem

Ressources

- Sparse ressources to get a specific model working. Focus on Calibration rather than SOTA.
- Use pre-tested architectures and /or Pretrained models.

Exam and Practical Information

Course info mainly in github:
https://github.com/ulfaslak/sds_tddl_2020/wiki/Syllabus

Exam

Type of assessment	Written assignment, 24 hours individuel take-home assignment. The students are allowed to communicate about the given problem-set but must work on, write and upload the assignment answer individually. Be aware that the plagiarism rules must be complied. The exam assignment is given in English and must be answered in English. ___
Exam registration requirements	During the semester mandatory assignments must be handed in to the teachingassistants not later than the given deadlines. Two mandatory assignments must be approved to be able to sit the exam. ___

Exercise I

- Setting up a server for deep learning: Google Cloud Compute
  - Log in to google. Sign up for 300$ Free Credit.
  - Follow the instructions on https://course.fast.ai/start_gcp.html.

Exercise II : Creating an image dataset

- Downloading an image dataset: formatting them for deep learning.
  - Visit the website: https://images.google.com

  Two Options:
  - Hardest (maybe smartest):
    1. Design a script to input search terms using the selenium package.
    2. Search for the six basic emotions: happiness, sadness, fear, anger, surprise and disgust.
    3. Scroll down *n* times (not to many google ain't stupid), and save the html. – *name files by the emotion.*
  - Easy (and safest):
    - Input the six search terms and download the html manually.
  - Open the html files and use regular expression to extract links to each image.
  - Download all images into dedicated folders.
  - Pick an image from each folder and visualize it in the notebook using matplotlib.

Exercise III : Creating a sentiment dataset

- See exercise 8.2 in the SDS summer school course:
  - https://github.com/abjer/sds/blob/master/material/session_8/exercise_8.ipynb