CIS 3715
Principles of Data Science

Hongchang Gao

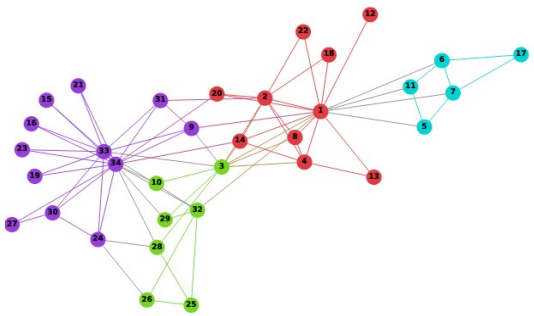
Spring 2024

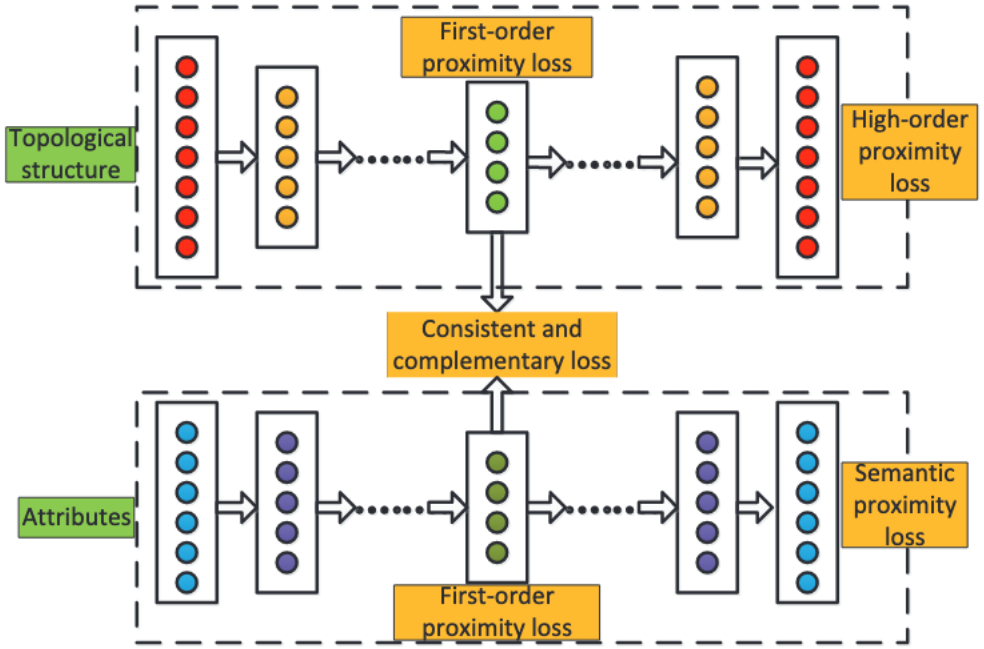
Computer and Information Sciences
Temple University

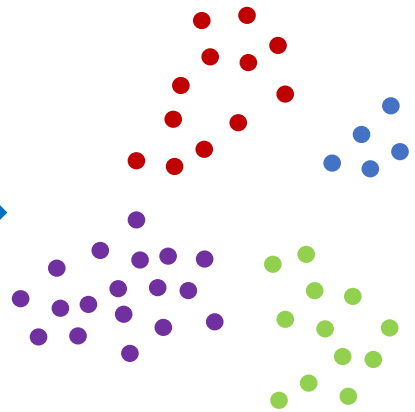
Instructor

- Research areas:
  - Machine/Deep Learning

$$L = -\sum_{E_{ij}>0} \log p_{ij}^{M} - \sum_{E_{ij}>0} \log p_{ij}^{Z} + \sum_{i=1}^{n} \|\hat{M}_{i\cdot} - M_{i\cdot}\|_2^2 + \sum_{i=1}^{n} \|\hat{Z}_{i\cdot} - Z_{i\cdot}\|_2^2 - \sum_{i} \{\log p_{ii} - \sum_{E_{ij}=0} \log(1 - p_{ij})\}$$



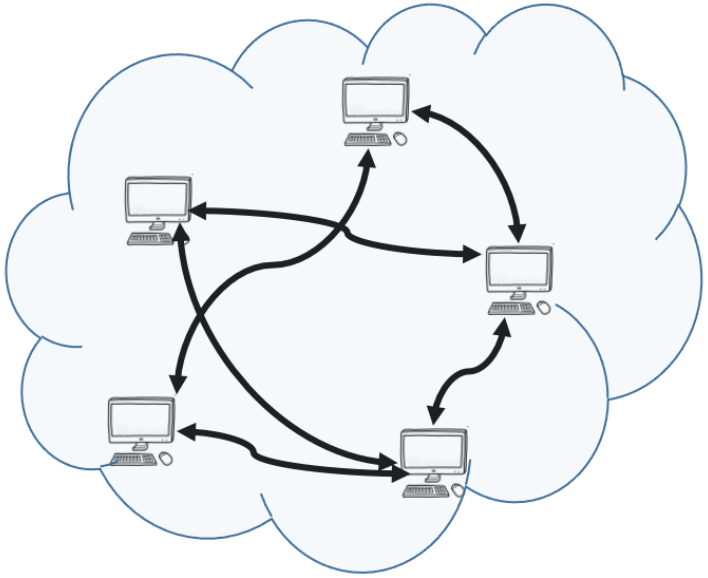


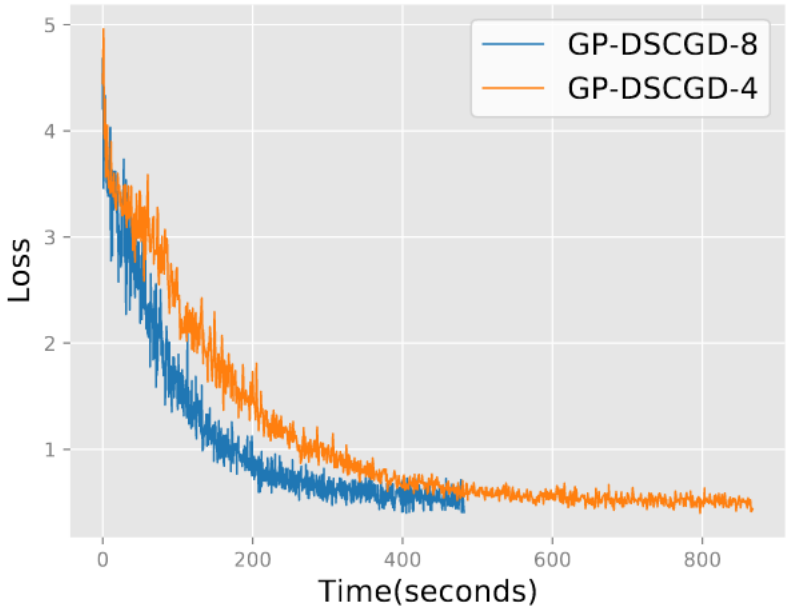


Instructor

- Research areas:
  - Large-scale optimization

$$\min_{x\in\mathbb{R}^d} \frac{1}{K}\sum_{k=1}^{K} \mathbb{E}_\zeta\left[f^{(k)}\left(\mathbb{E}_\xi[g^{(k)}(x;\xi)];\zeta\right)\right]$$





Outline

- Course overview
- Introduction to data science

Course Logistics

- Lecture section:
  - The class meets at 9:30-10:50am on Tue and Thu
  - BEURY, 00162
- Lab section:
  - The class meets at 9:00-10:50am on Mon
  - TA for Section 001:
    - Xinwen (Ellen) Zhang
    - SERC, 00204
  - TA for Section 002:
    - Mathew Kuruvilla
    - SERC, 00206

Course Logistics

- Office hour
  - Instructor: Hongchang Gao, hongchang.gao@temple.edu
    - 11:00am-12:00pm Tuesday,
    - SERC 318
  - Section 001 TA: Xinwen (Ellen) Zhang, ellenz@temple.edu
    - 2:00pm-4:00pm Tuesday,
    - SERC 303
  - Section 002 TA: Mathew Kuruvilla, mathewkuruvilla@temple.edu
    - 11:30am-12:30pm Tuesday and Thursday
    - SERC 357

Course Logistics

- Prerequisites
  - CIS 2166 or linear algebra, CIS 1051 or 1057 or 1068
  - Be familiar with basic mathematical knowledge about algebra and statistics. For example, it is expected that you know vector, matrix, mean, variance.
  - Good programming skills in Python. It is expected that you can use Python to preprocess data and implement state-of-the-art data mining methods to analyze data.

Course Logistics

- Class Materials
  - No required textbooks, but recommend to read
    - Peter Bruce, “Practical Statistics for Data Scientists: 50 Essential Concepts,” 2017.
    - Wes McKinney. “Python for Data Analysis: Data Wrangling with Pandas, NumPy, and IPython.” O'Reilly Media, 2012.
    - Foster Provost, Tom Fawcett. “Data Science for Business: What You Need to Know About Data Mining and Data-Analytic Thinking.” O'Reilly Media, 2013.
    - Cole Nussbaumer Knaflic, “Storytelling with Data: A Data Visualization Guide for Business Professionals,” 2015.

Grading Policy

- Class attendance and participation: 10%
  - Check in for every lab section
- Labs and homework: 25%
  - 10 lab assignments
- Quizzes: 20%
  - Weekly quizzes
- Midterm: 25%
- Final project: 20%

Lab Assignments

- How to submit?
  - Upload via Canvas
  - Not accept email submission

- Policy
  - NOT accept late submission
  - Do NOT copy others' solutions
  - Feel free to use resources from the web, but make sure to acknowledge the sources.

Final Project

- A list of potential topics will be given
- Performed individually or in groups of up to 3 students
- You will need to submit
  - A brief proposal:
    - What’s topic you plan to work on?
    - Why you choose this topic?
    - Which aspects you will focus on?
  - Final report
    - An entire pipeline of this project: data preprocessing, data understanding, data analysis with machine learning models, result analysis.

Course Schedule

Week	Date	Topic(s)	Assignments	Due
1	01/16	Course Introduction		
	01/18	Introduction Python		
2	01/23	Mathematics foundation	Lab 1, 01/22	
	01/25			
3	01/30	Data Preprocessing	Lab 2, 01/29	
	02/01	Exploratory Data Analysis		Lab 1 due, 02/06
4	02/06	Supervised Learning	Lab 3, 02/05	
	02/08			Lab 2 due, 02/13
5	02/13	Supervised Learning	Lab 4, 02/12	
	02/15			Lab 3 due, 02/20
6	02/20	Supervised Learning	Lab 5, 02/19	
	02/22			Lab 4 due, 02/27
7	02/27	Midterm review		
	02/29	**Midterm exam**		
8	03/05	**Spring Break**		
	03/07			

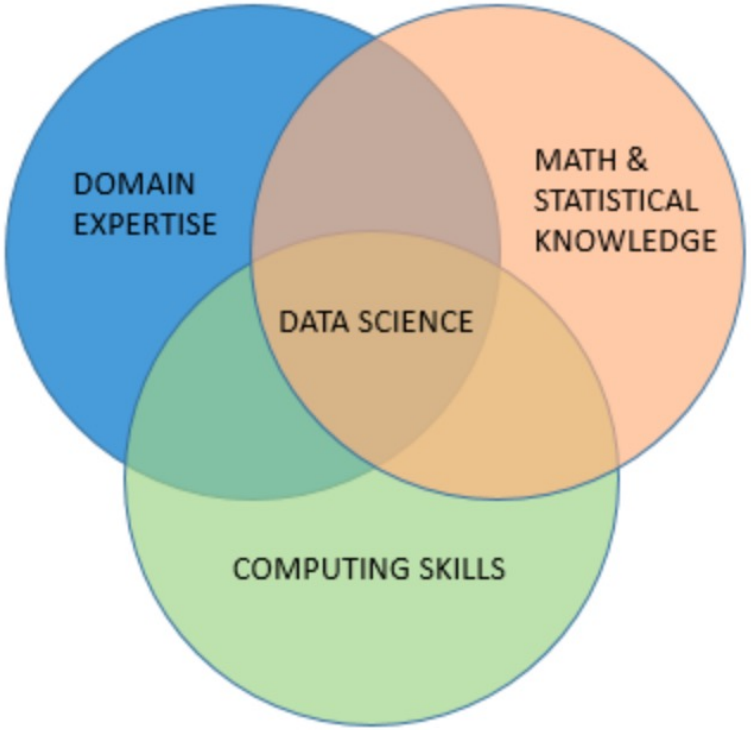
9	03/12	Unsupervised Learning	Lab 6, 03/11	
	03/14			Lab 5 due, 03/19
10	03/19	Unsupervised Learning	Lab 7, 03/18	
	03/21		Final project out	Lab 6 due, 03/26
11	03/26	Document analysis	Lab 8, 03/25	
	03/28			Lab 7 due, 04/02
12	04/02	Recommendation system	Lab 9, 04/01	
	04/04			Lab 8 due, 04/09
13	04/09	Deep neural networks	Lab 10, 04/08	
	04/11			
14	04/16	Deep neural networks		Lab 9 due, 04/16
	04/18			
15	04/23	Final project presentation		Final project due, 04/23
	04/25			
16	04/30	Study days		Lab 10 due, 04/30

Outline

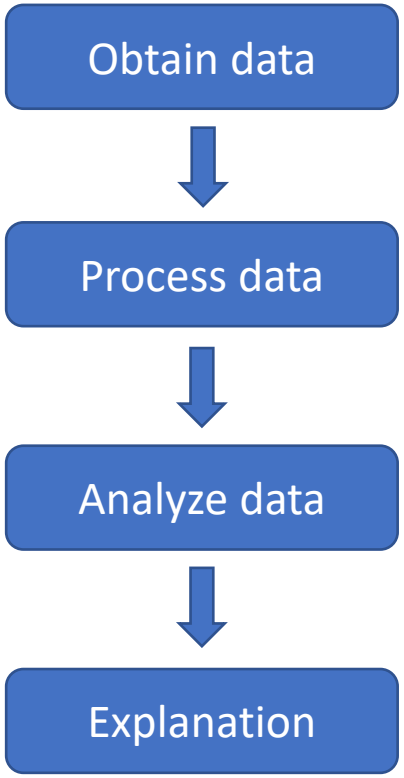
- Course overview
- Introduction to data science

Introduction to Data Science

- What is data science?
  - Extract knowledge from data



https://library.osu.edu/site/it/where-does-data-science-fit-in/



Example: Spam detection

- Step 1: load data

```
message_data = pd.read_csv("spam.csv",encoding = "latin")
message_data.head()
```

	v1	v2	Unnamed: 2	Unnamed: 3	Unnamed: 4
0	ham	Go until jurong point, crazy.. Available only ...	NaN	NaN	NaN
1	ham	Ok lar... Joking wif u oni...	NaN	NaN	NaN
2	spam	Free entry in 2 a wkly comp to win FA Cup fina...	NaN	NaN	NaN
3	ham	U dun say so early hor... U c already then say...	NaN	NaN	NaN
4	ham	Nah I don't think he goes to usf, he lives aro...	NaN	NaN	NaN

Example: Spam detection

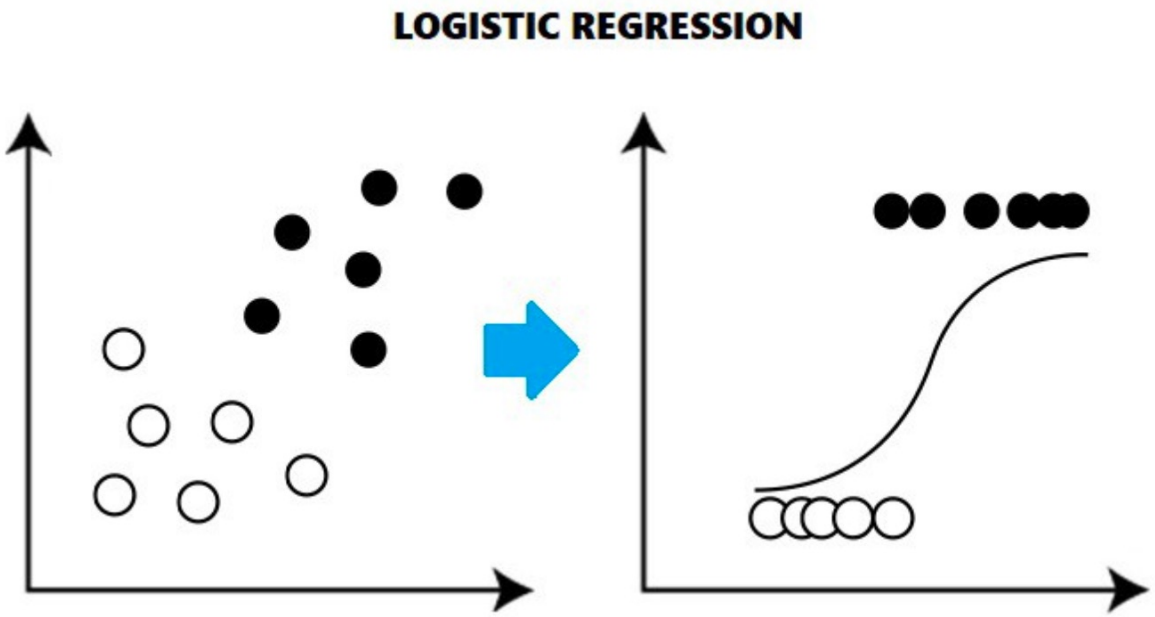
- Step 2: data preprocessing
  - Some features are missing or not important
  - Transform words to numerical values

Build a dictionary → [world, Hello, computer, math, PA, Temple, campus]

Hello Temple → [0, 1, 0, 0, 0, 1, 0]

Example: Spam detection

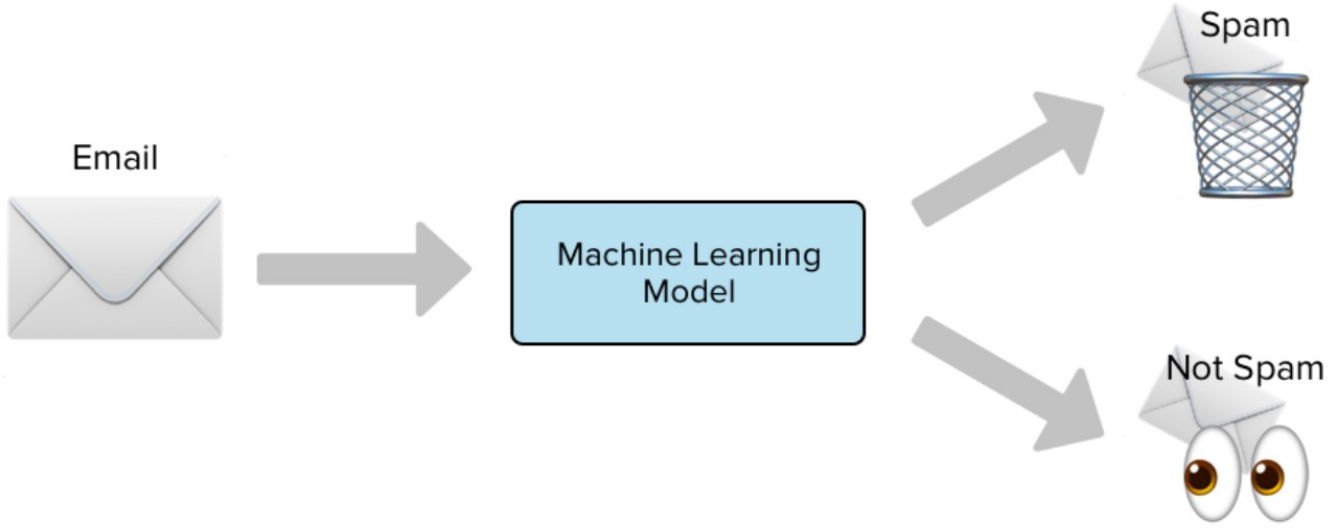
- Step 3: build the classifier



https://medium.com/analytics-vidhya/understanding-logistic-regression-b3c672deac04

Example: Spam detection

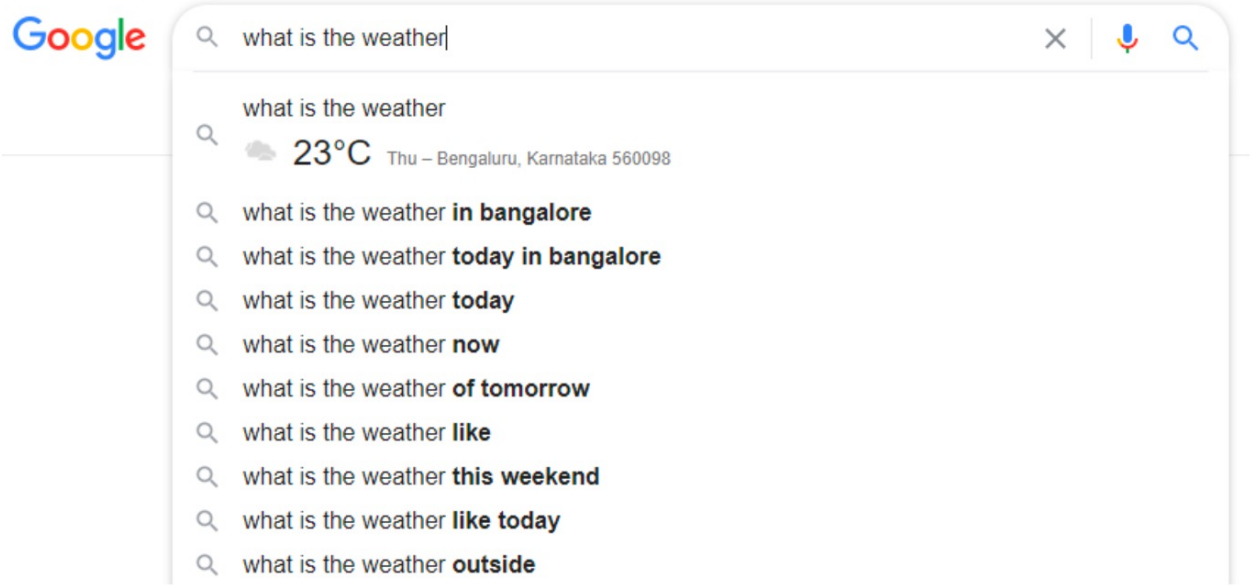
- Step 4: evaluation

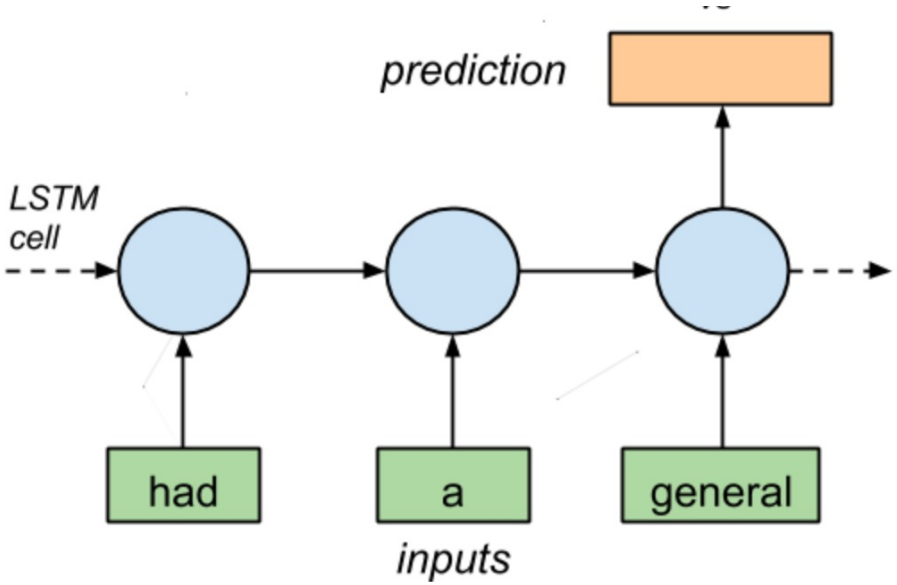


https://towardsdatascience.com/spam-detection-with-logistic-regression-23e3709e522

Data Science Applications

- Internet search
  - Autocomplete feature



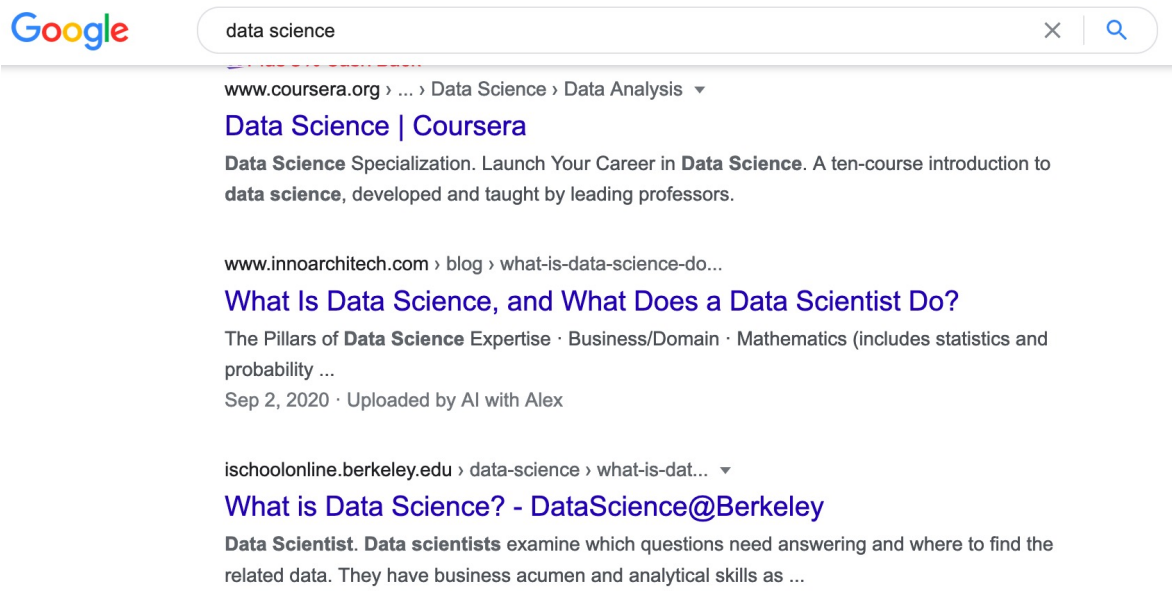


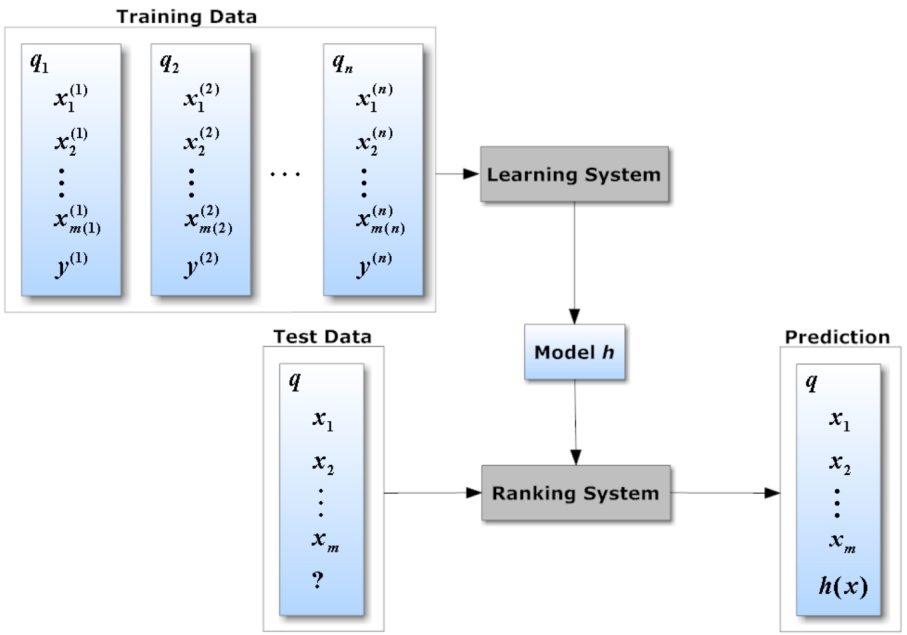
https://towardsdatascience.com/next-word-prediction-with-nlp-and-deep-learning-48b9fe0a17bf

https://tomaxent.com/2017/04/26/LSTM-by-Example-using-Tensorflow-Text-Generate/

Data Science Applications

- Internet search
  - Autocomplete feature
  - Ranking results



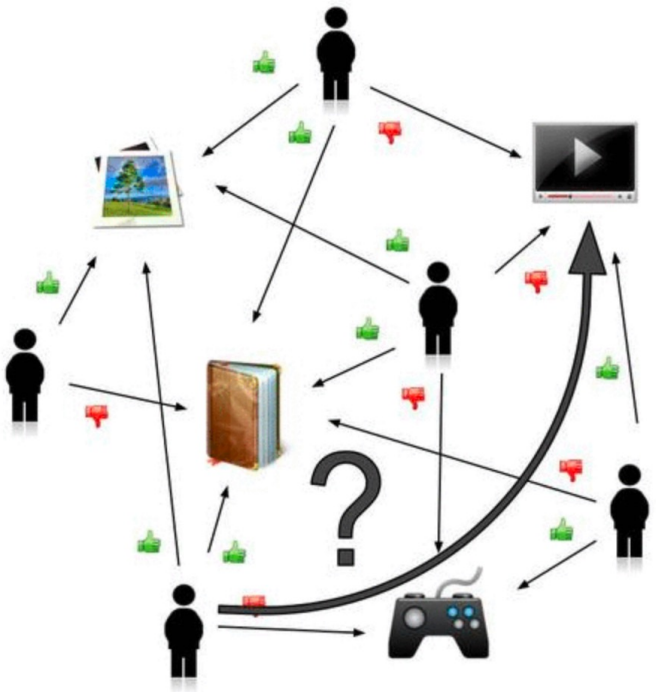


(Photo coutesy: Catarina Moreira)

https://medium.com/@nikhilbd/intuitive-explanation-of-learning-to-rank-and-ranknet-lambdarank-and-lambdamart-fe1e17fac418

Data Science Applications

- Recommendation system
  - Recommend products to users

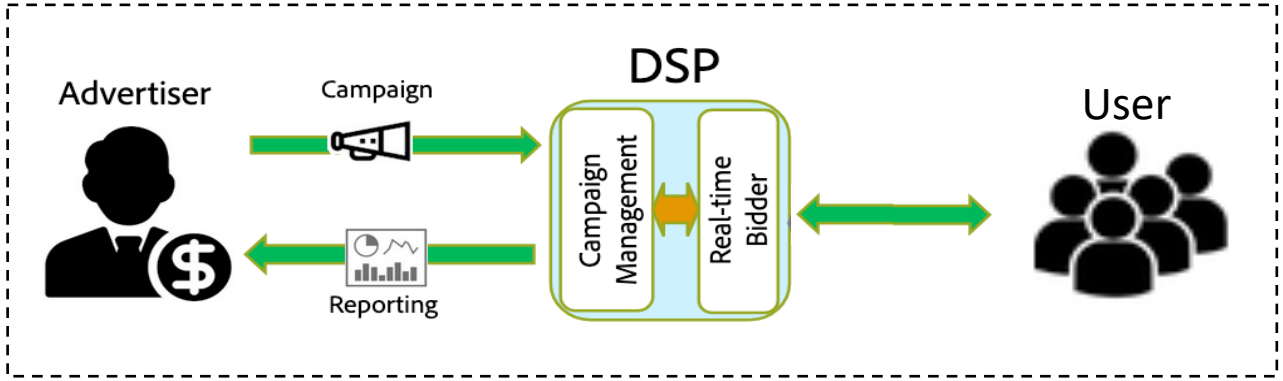


https://upload.wikimedia.org/wikipedia/commons/5/52/Collaborative_filtering.gif

https://heartbeat.fritz.ai/recommender-systems-with-python-part-iii-collaborative-filtering-singular-value-decomposition-5b5dcb3f242b

Data Science Applications

- Target advertising
  - Deciding which ads to show
  - How to show them



https://medium.com/@vinitvssingh5/explaining-the-field-aware-factorization-machines-ffms-for-ctr-prediction-2e98b2949dd6

Data Science Applications

- Social network analysis

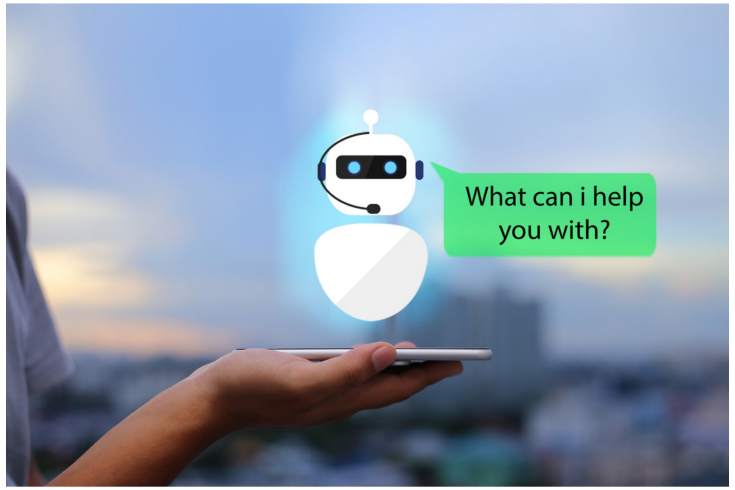
https://learningnets.github.io/KDD19_Tutorial/1_Motivations.pdf

Node Classification

https://slideplayer.com/slide/3131845/

Data Science Applications

- Natural language processing



Chat Robot

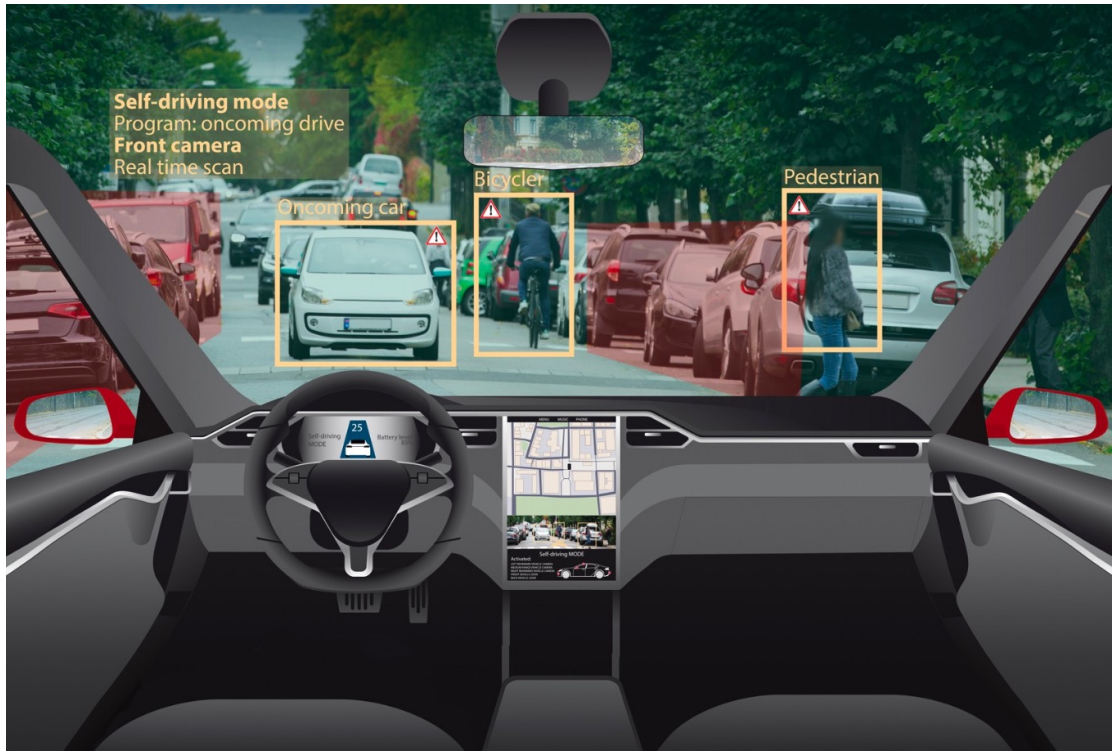
https://www.analyticsinsight.net/nlp-augments-the-power-of-chatbots-and-voice-in-2019/

Machine Translation

https://finance.yahoo.com/news/google-ceo-sundar-pichai-revealed-004138550.html

Data Science Applications

- Self-driving cars



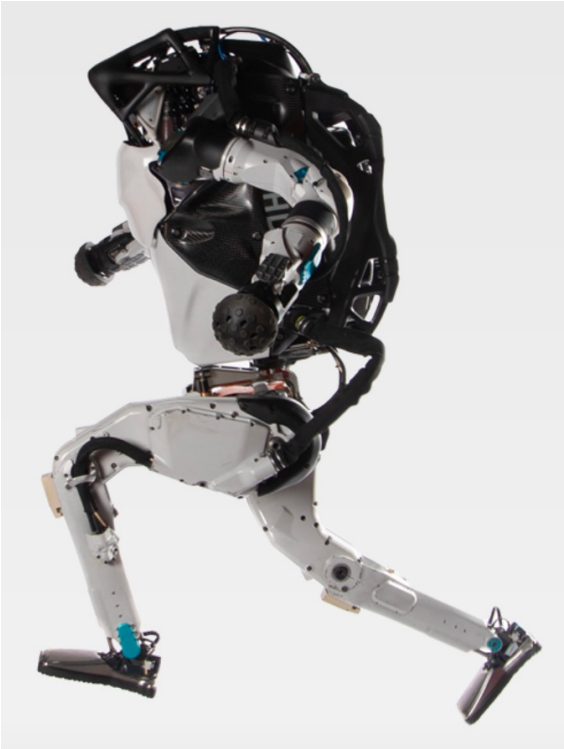
Autonomous Driving

https://medium.com/@webanalytics_31234/

Data Science Applications

- Sequential decision: robotics

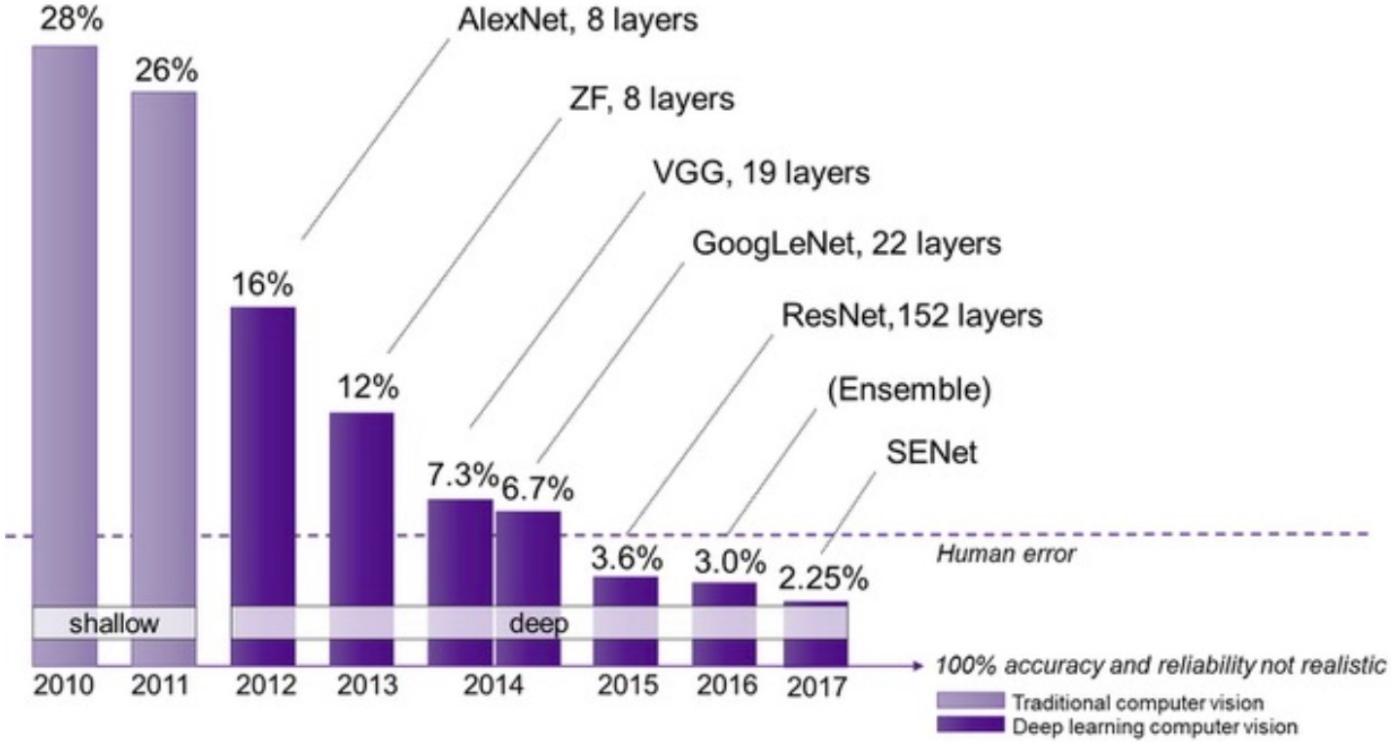
https://arxiv.org/pdf/1504.00702.pdf



https://www.bostondynamics.com/atlas

Exciting Success 1

- Image classification
  - ImageNet competition



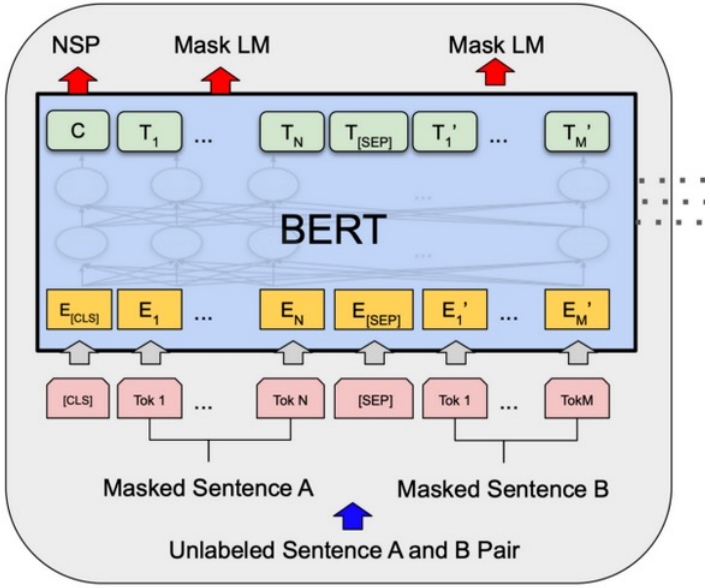
https://semiengineering.com/new-vision-technologies-for-real-world-applications/

Exciting Success 2

- **S**tanford **Qu**estion **A**nswering **D**ataset (SQuAD):
  - a reading comprehension dataset,
  - questions posed by crowd workers on a set of Wikipedia articles
  - the answer to every question is a segment of text, or *span*

SQuAD1.1 Leaderboard

Rank	Model	EM	F1
	Human Performance *Stanford University* (Rajpurkar et al. '16)	82.304	91.221
1 Oct 05, 2018	BERT (ensemble) *Google AI Language* https://arxiv.org/abs/1810.04805	**87.433**	**93.160**
2 Sep 09, 2018	nlnet (ensemble) *Microsoft Research Asia*	85.356	91.202
3 Jul 11, 2018	QANet (ensemble) *Google Brain & CMU*	84.454	90.490

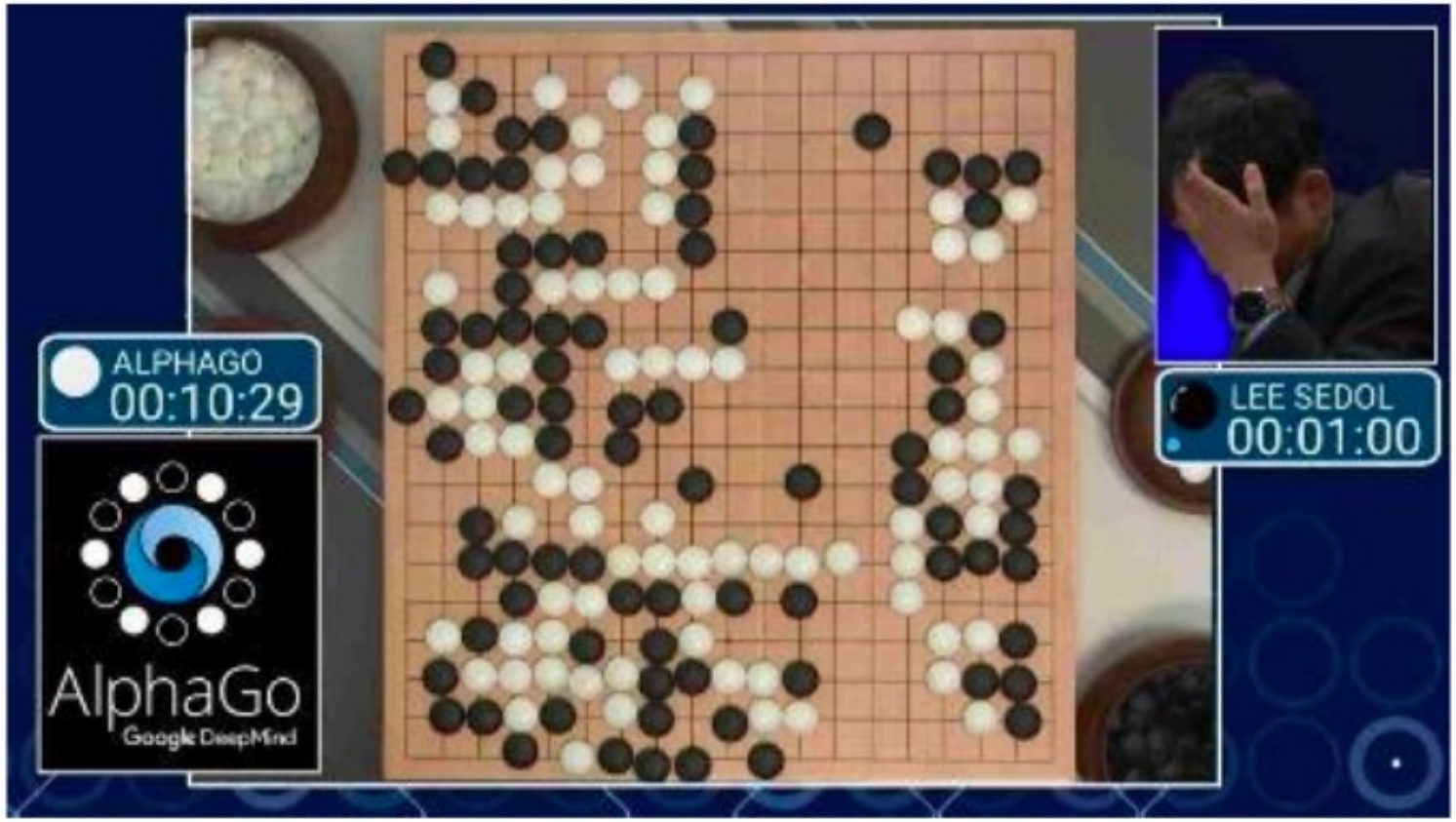


https://ai.googleblog.com/2018/11/open-sourcing-bert-state-of-art-pre.html

https://towardsdatascience.com/bert-why-its-been-revolutionizing-nlp-5d1bcae76a13

Exciting Success 3

- Alpha Go



Conclusion

- Data Science is everywhere
- Data Science is interesting
- Data Science is powerful