

COMP 527: Data Mining and Visualization

Danushka Bollegala



Introductions

- Lecturer: Danushka Bollegala
- Office: 2.24 Ashton Building (Second Floor)
- Email: danushka@liverpool.ac.uk
- Personal web:
 - <http://danushka.net/>
- Research interests
 - Natural Language Processing (NLP)

Course web site

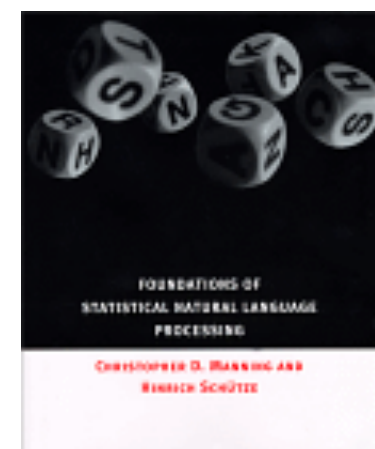
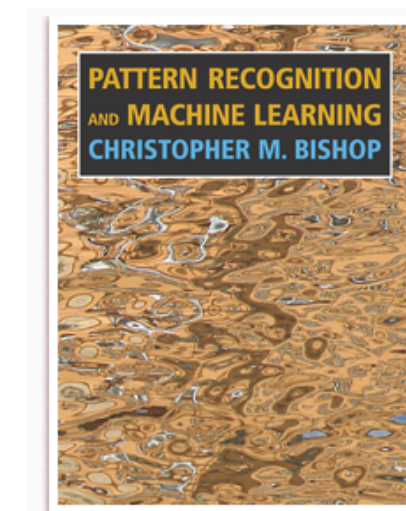
- <http://danushka.net/lect/dm>
- Course notes, lecture schedule, assignments, references are uploaded to the course web site
- Discussion board (QA) on vital available.
- Do not email me your questions. Instead post them on the discussion board so that others can also benefit from your QA.

Evaluation

- 75% End of Year Exam
 - 2.5 hrs
 - short answers and/or essay type questions
 - Select 4 out of 5 questions
 - Past papers are available on the lecture web site
 - Some of the review questions might appear in the exam as well!
- 25% Continuous Assessment
 - Assignment 1: 12%
 - Assignment 2: 13%
- Both assignments are programming oriented (in Python)
- Attend lab sessions for Python+Data Mining (once a week)

References

- Data Mining, Witten
- Pattern recognition and machine learning (PRML), Bishop.
- Fundamentals of Statistical Natural Language Processing (FSNLP), Manning



Course summary

- Data preprocessing (missing values, noisy data, scaling)
- Classification algorithms
 - Decision trees, Naive Bayes, k-NN, logistic regression, SVM
- Clustering algorithms
 - k-Means, k-Medoids, Hierarchical clustering
- Text Mining, Graph Mining, Information Retrieval
- Neural networks and Deep Learning
- Dimensionality reduction
- Visualization theory, t-SNE, embeddings
- Word embedding learning

Data Mining Intro

Danushka Bollegala



What is data mining?

- Various definitions
 - *The nontrivial extraction of implicit, previously unknown, and potentially useful information from data (Piatetsky-Shapiro)*
 - *...the automated or convenient extraction of patterns representing knowledge implicitly stored or captured in large databases, data warehouses, the Web, ... or data streams (Han, page xxi)*
 - *...the process of discovering patterns in data. The process must be automatic or (more usually) semiautomatic. The patterns discovered must be meaningful..." (Witten, page 5)*

Applications of Text Mining

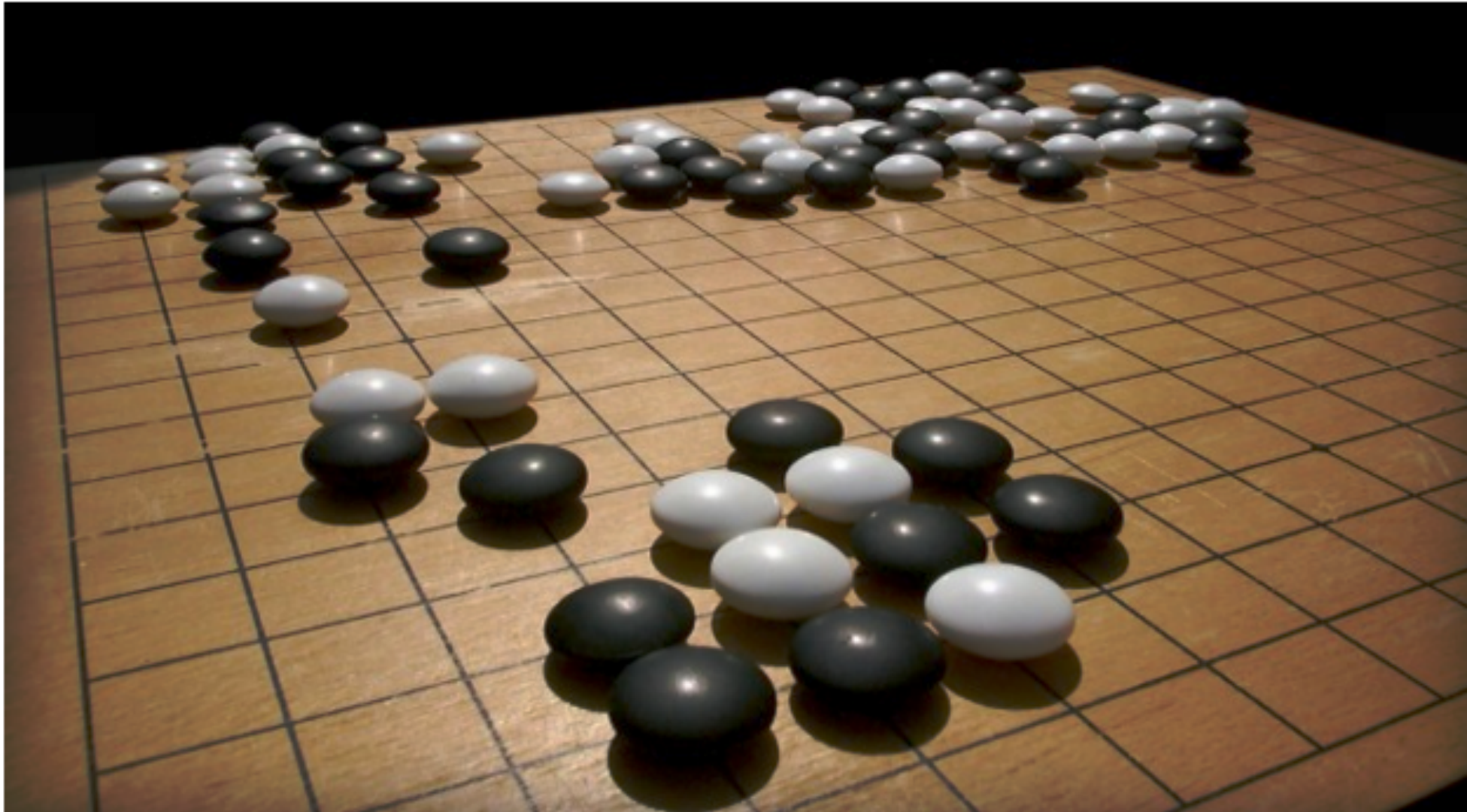


Computer program wins Jeopardy contest in 2011! 9

Applications of Deep Learning

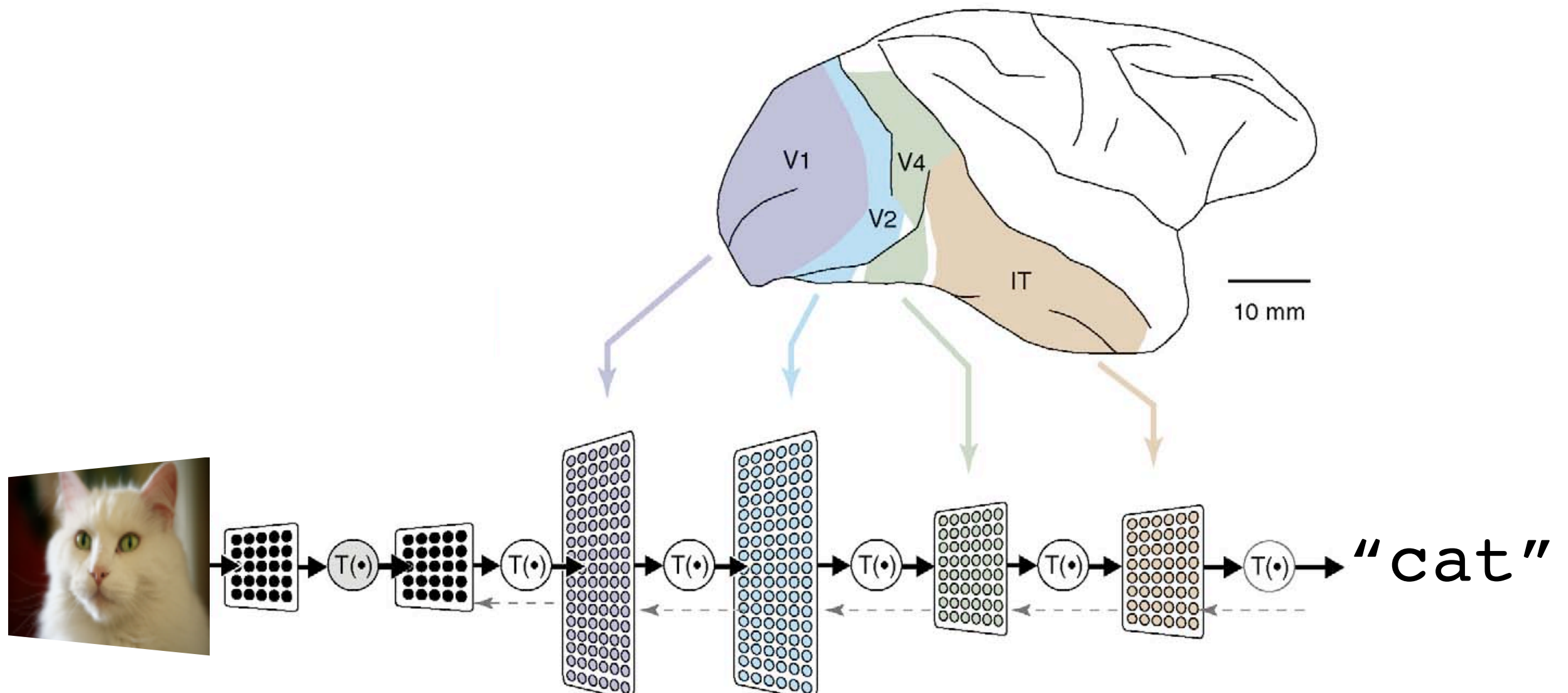
Google's DeepMind AI beats humans at the massively complex game Go

By Ryan Whitwam on January 27, 2016 at 4:00 pm | [11 Comments](#)



Google acquired the British

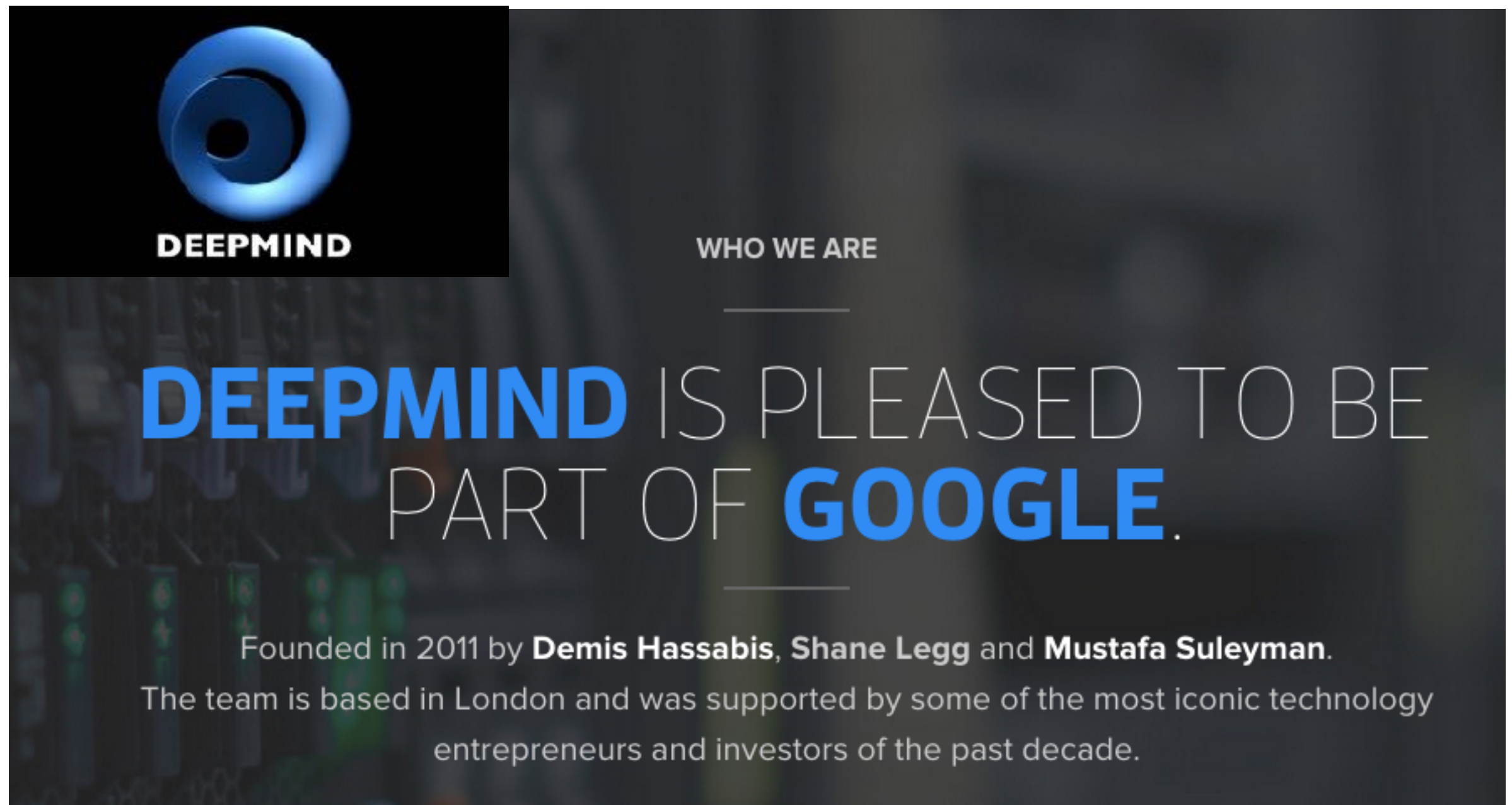
Deep Learning



An unsupervised neural network learns to recognize cats when trained using millions of you tube videos! (2012)



Deep Learning



Google acquires London-based AI (gaming) startup for USD 400M!

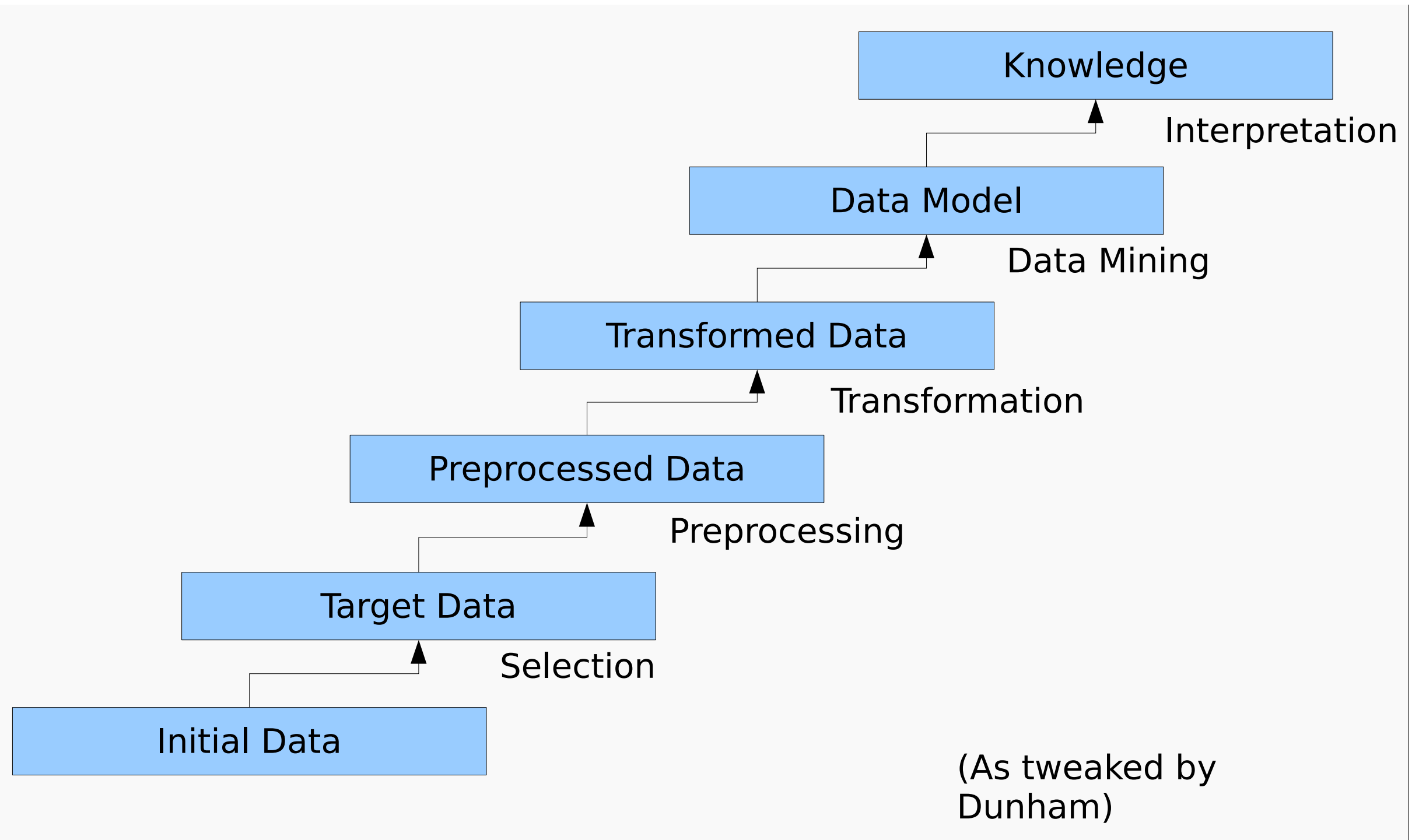
Industrial Interests

- Data Mining (DM)/ Machine Learning (ML)/ Natural Language Processing (NLP) experts are sought after by the CS industry
- Google research (Geoff Hinton/NN)
- Facebook AI research (Yann LeCun/Deep ML)
- Baidu (Andrew Ng)
- The ability to apply the algorithms we learn in this lecture (and their complex combinations) will greatly improve your employability in CS industries

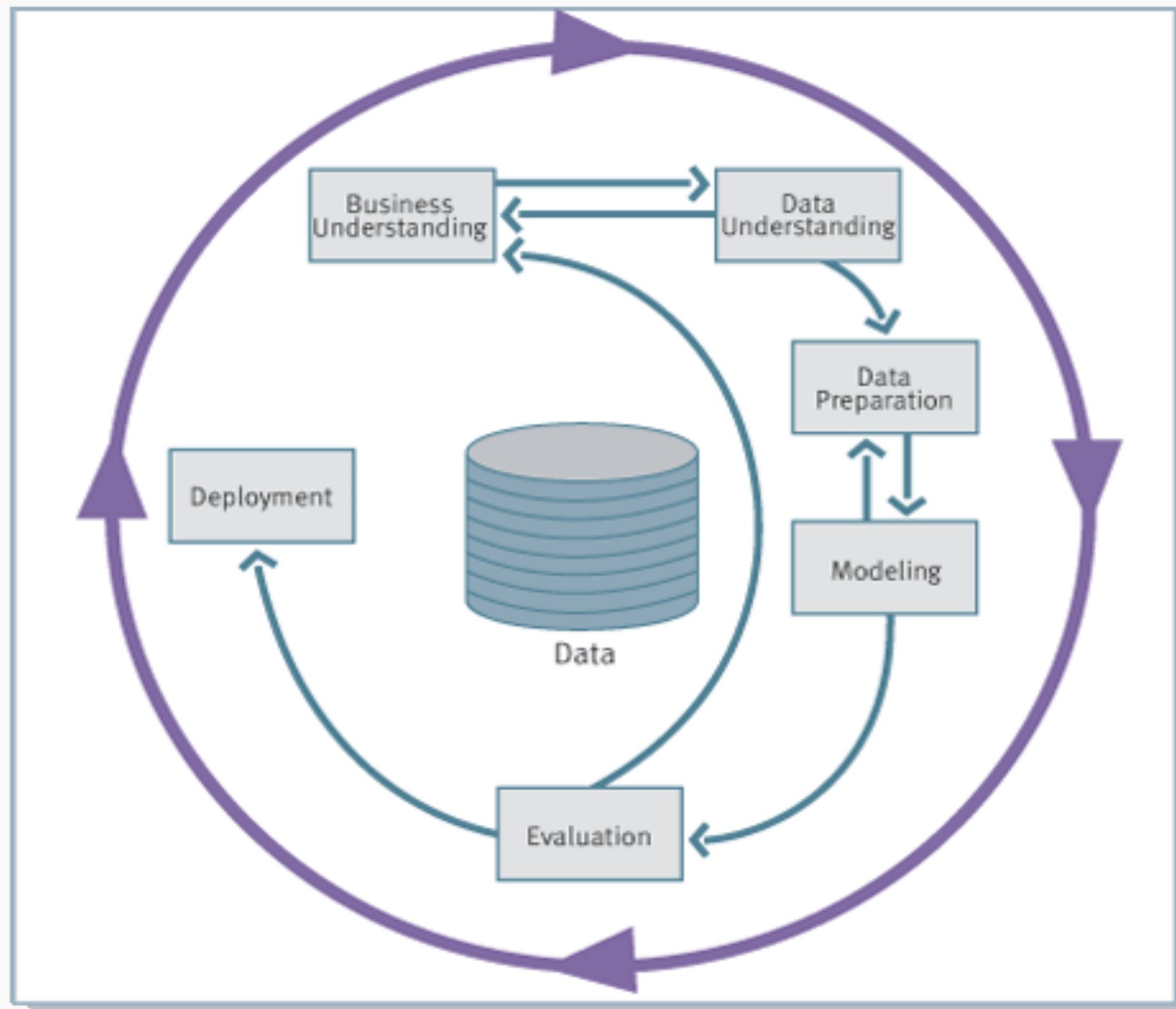
Academic Interests

- DM is an active research field.
- Top conferences
 - Knowledge Discovery and Data Mining (KDD) [<http://www.kdd.org/kdd2018/>]
 - Annual Conference of the Association for Computational Linguistics (ACL) [<http://acl2018.org/>]
 - International Word Wide Web Conference (WWW) [www2018.thewebconf.org]
 - International Conference on Machine Learning (ICML)
 - Neural and Information Processing (NIPS)
 - International Conference on Learning Representations (ICLR)

Piatetsky-Shapiro View



CRISP-DM View



Two main goals in DM

- Prediction
 - Build models that can predict future/unknown values of variables/patterns based on known data
 - Machine learning, Pattern recognition
- Description
 - Analyse given datasets to identify novel/interesting/useful patterns/rules/trends that can describe the dataset
 - clustering, pattern mining, associative rule mining

Broad classification of Algorithms



Classification Algorithms
(k-NN, Naive Bayes, logistic regression, SVM, Neural Networks, Decision Trees)

Clustering Algorithms
(k-means, hierarchical clustering)
visualization algorithms
(t-SNE, PCA)
Dimensionality reduction
(SVD, PCA)
Pattern/sequence mining

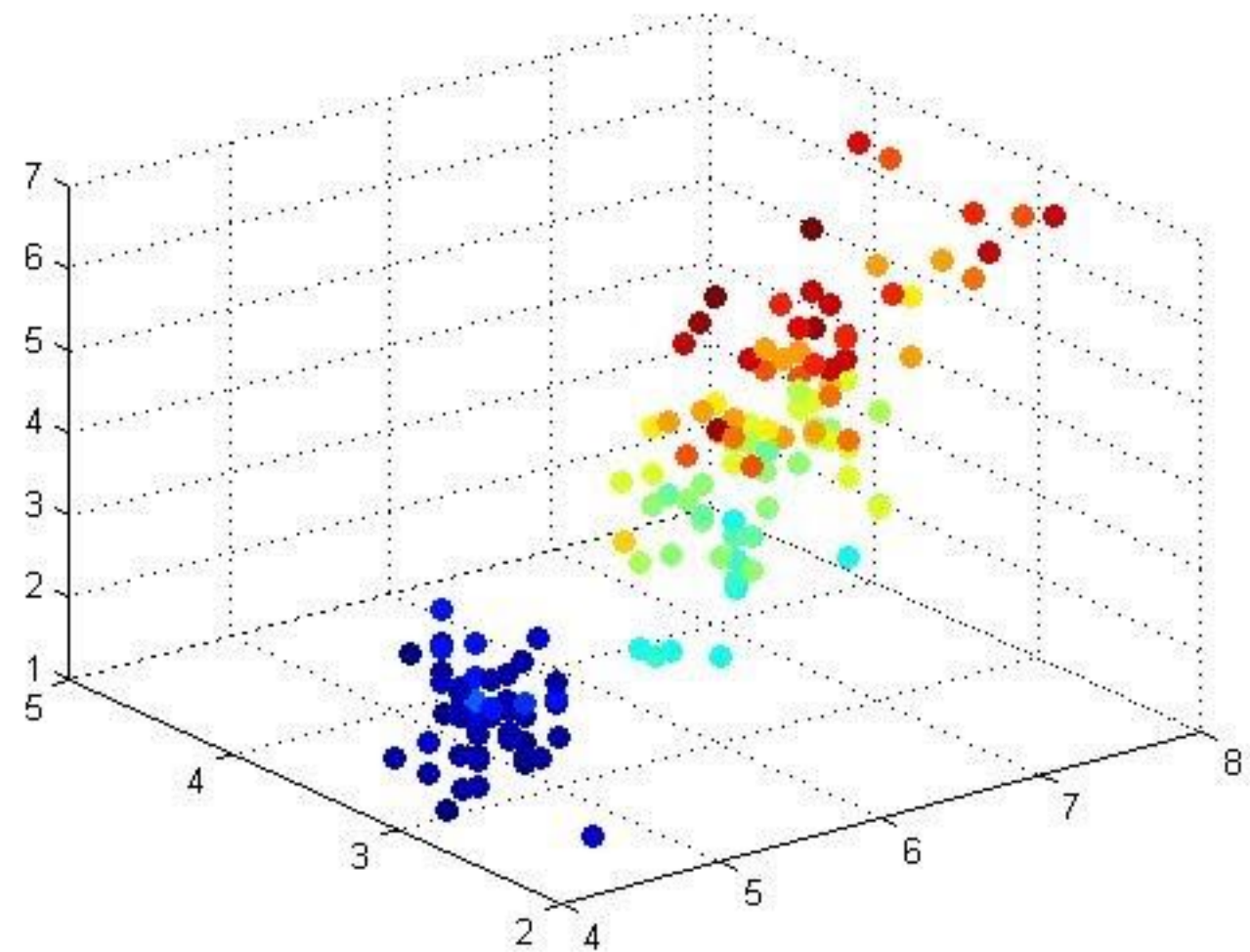
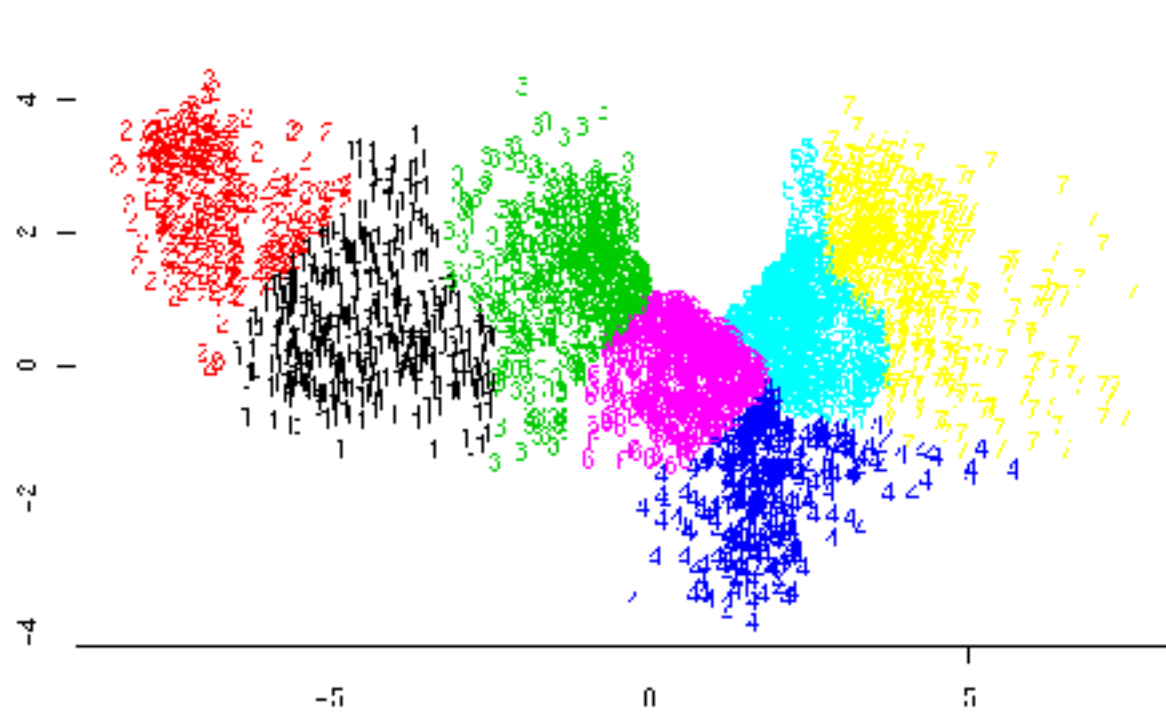
Classification

- Given a data point x , classify it into a set of **discrete classes**
- Example
 - Sentiment classification
 - *The movie was great* +1
 - *The food was cold and tasted bad* -1
 - Spam vs. non-spam email classification
- We want to learn a classifier $f(x)$ that predicts either -1 or +1. We must learn function f that optimises some objective (e.g. number of misclassifications)
- A train dataset $\{x,y\}$ where $y \in \{-1,1\}$ is provided to learn the function f .
 - supervised learning

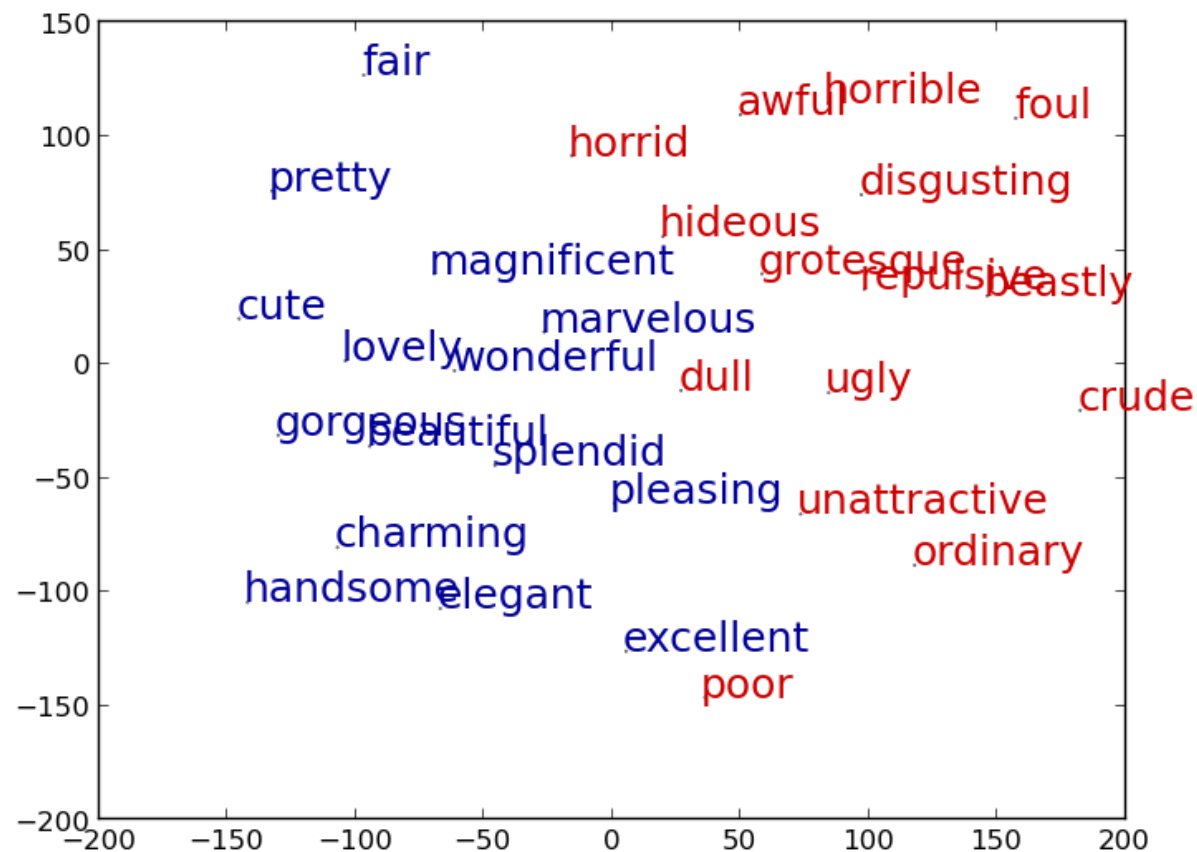
Clustering

- Given a dataset $\{x_1, x_2, \dots, x_n\}$ group the data points into k groups such that data points within the same group have some common attributes/similarities.
- Why we need clusters (groups)
 - If the dataset is large, we can select some representative samples from each cluster
 - Summarise the data, visualise the data

Cluster visualization



Word clusters



Yogatama+14

words that express
similar sentiments
are grouped into
the same cluster

COMP527
Data Mining and Visualisation
Problem Set 0

Danushka Bollegala

Question 1 Consider two vectors $\mathbf{x}, \mathbf{y} \in \mathbb{R}^3$ defined as $\mathbf{x} = (1, 2, -1)^\top$ and $\mathbf{y} = (-1, 0, 1)^\top$. Answer the following questions about these two vectors.

- A. Compute the length (ℓ_2 norm) of \mathbf{x} and \mathbf{y} . **(4 marks)**
- B. Compute the inner product between \mathbf{x} and \mathbf{y} . **(2 marks)**
- C. Compute the cosine of the angle between the two vectors \mathbf{x} and \mathbf{y} . **(4 marks)**
- D. Compute the Euclidean distance between the end points corresponding to the two vectors \mathbf{x} and \mathbf{y} . **(4 marks)**
- E. For any two vectors $\mathbf{x}, \mathbf{y} \in \mathbb{R}^d$ such that $\|\mathbf{x}\|_2 = \|\mathbf{y}\|_2 = 1$ show that the following relationship holds between their cosine similarity $\cos(\mathbf{x}, \mathbf{y})$ and their Euclidean distance $\text{Euc}(\mathbf{x}, \mathbf{y})$. **(6 marks)**

$$\text{Euc}(\mathbf{x}, \mathbf{y})^2 = 2(1 - \cos(\mathbf{x}, \mathbf{y}))$$

Question 2 Consider a matrix $\mathbf{A} \in \mathbb{R}^{2 \times 2}$ defined as follows:

$$\mathbf{A} = \begin{pmatrix} 2 & 1 \\ 1 & 2 \end{pmatrix}$$

Answer the following questions related to \mathbf{A} .

- A. Compute the transpose \mathbf{A}^\top . **(2 marks)**
- B. Compute the determinant $\det(\mathbf{A})$. **(2 marks)**
- C. Compute the inverse \mathbf{A}^{-1} . **(4 marks)**
- D. Compute the eigenvalues and eigenvectors of \mathbf{A} . **(6 marks)**

Question 3

- A. Given $\sigma(x) = \frac{1}{1+\exp(ax+b)}$, compute $\sigma'(x)$, the differential of $\sigma(x)$ with respect to x .
- B. Given $H(p) = -p \log(p) - (1-p) \log(1-p)$, find the value of p that maximises $H(p)$.
- C. Find the maximum value of $g(x, y) = x^2 + y^2$ such that $y \leq -x + 1$.