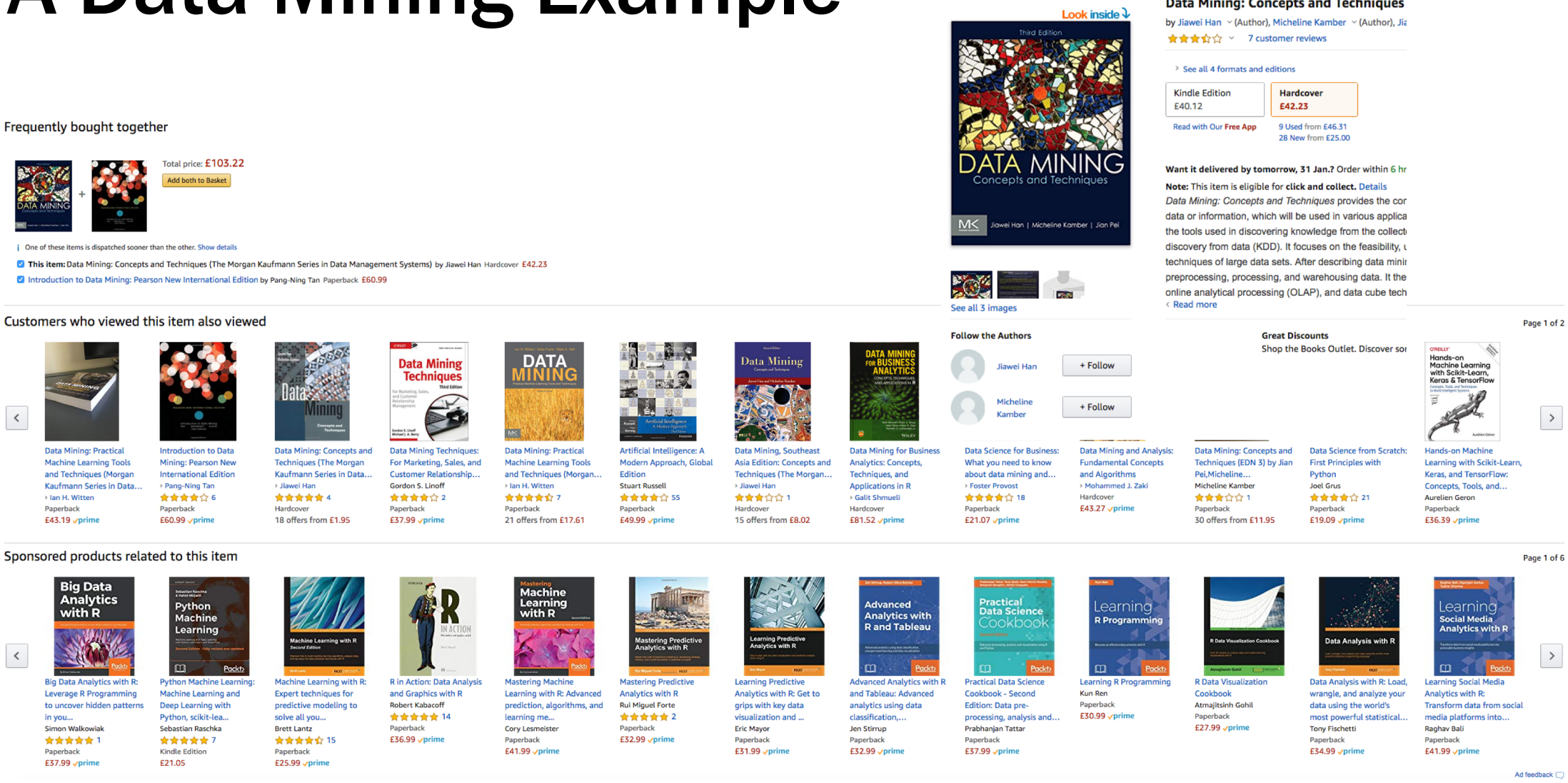
Big Data and Data Mining

01. Introduction to Data Mining

A Data Mining Example

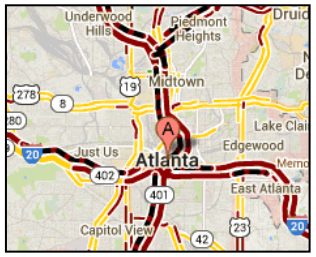


a) Image Source from www.amazon.co.uk

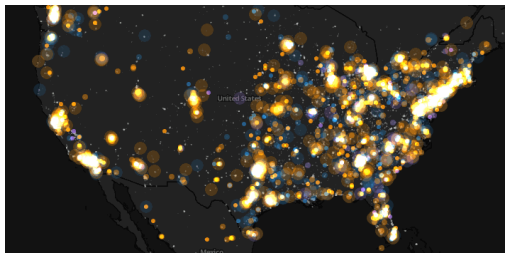
Big Data Era – How Data Is Generated

- Global Population
  - 3.54 Billion (1968)
  - 6.74 Billion (2008)
  - 7.70 Billion (2018)
- Mobile Users
  - 4.57 Billion (2018)
  - 3G ➔ 4G ➔ 5G
- Acquisition ⇧
  - Search on Website
- Demand ⇧
  - Item Recommendation

E-Commerce

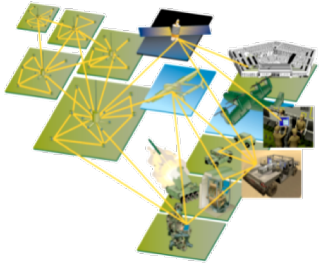


Traffic Flows

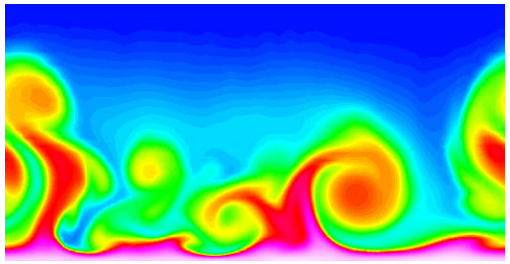


Social Networking: Twitter

Mobile Network



Sensor Networks



Computational Simulations

More.....

a) Introduction to Data Mining, 2nd Edition, by Tan, Steinbach, Karpatne, Kumar
b) Image Source from Google Search

Big Data Era – How Data Is Stored

- Floppy Disk ➔ USB Sticker ➔ Solid State Disk
- The Size of Storage
  - 10^6 (Mega Bytes) ➔ 10^9 (Giga Bytes) ➔ 10^{12} (Tera Bytes)
- The Speed of Data Transfer
  - FDD (720KB) ➔ USB2.0 (60MB) ➔ USB3.0 (600MB) ➔ TB3 (5GB)
  - IDE (133MB) ➔ SATA3.0 (600MB) ➔ SAS4.0 (3GB)
  - * 1 Byte = 8 bits

a) Image Source from Google Search

Big Data Era – How Data Is Processed

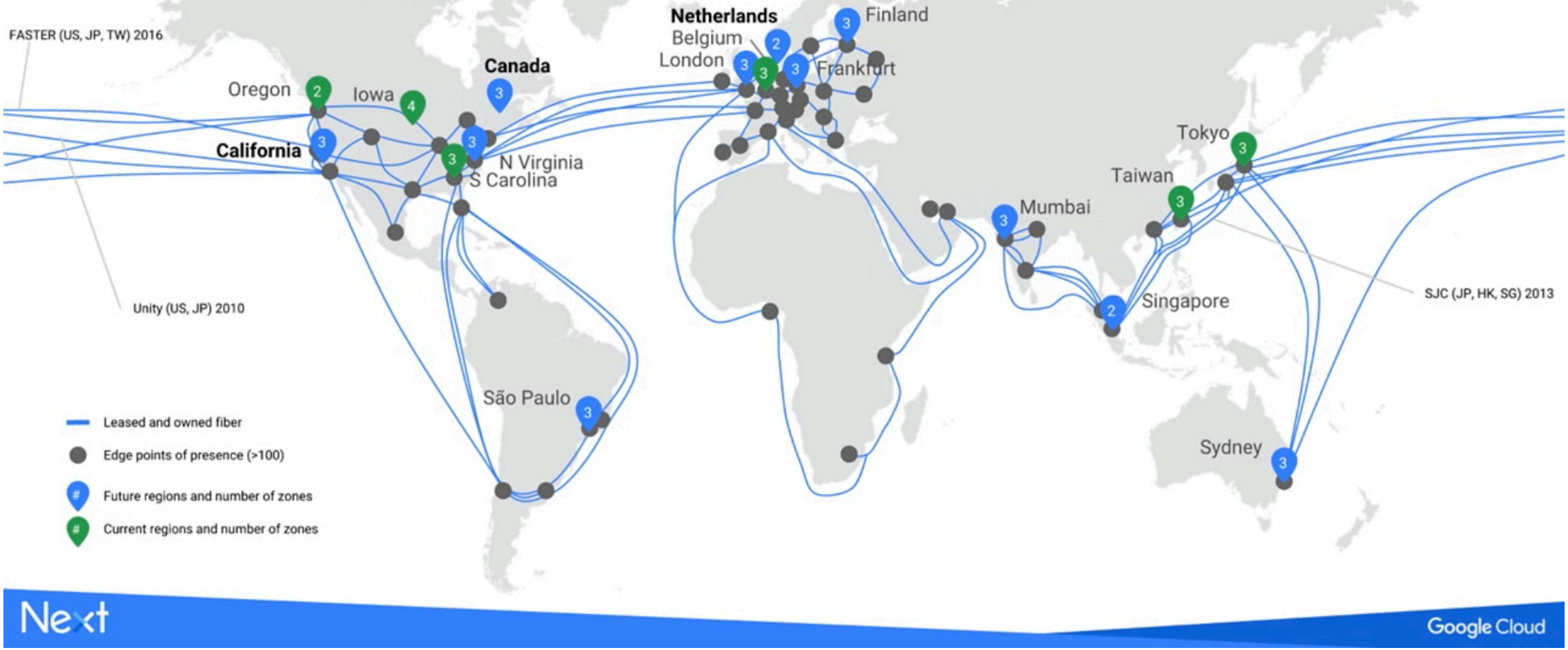
Server & Workstation

Cloud Data Centre

a) https://www.datacenterknowledge.com/google-data-center-faq-part-2/
b) https://www.youtube.com/watch?time_continue=5&v=zDAYZU4A3w0

GCP Infrastructure

6 regions, 18 zones, over 100 points of presence, and a well-provisioned global network comprised of hundreds of thousands of miles of fiber optic cable.



a) https://www.datacenterknowledge.com/google-data-center-faq-part-2/

b) Google VP of data centers Joe Kava's presentation at Google Cloud Next 2017 in San Francisco

Why Data Mining? – Commercial View

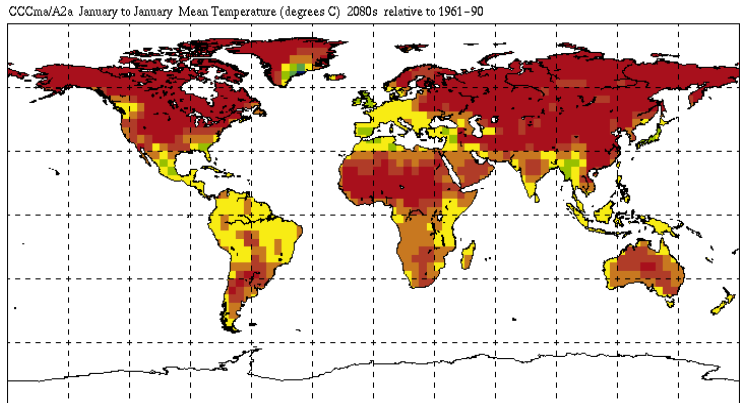
- Google searches per day:
  - 2000: 32.8 million searches
  - 2010: 1+ billion searches
  - 2012: 3.2 billion searches
  - 2018: 5.5 billion searches (63,000/s)
- Search markets 2018
  - The average cost per click in Google Ads is between $1 and $2 on the search network.
  - Better Search ➔ User Click ➔ Ads Income
  - Data Mining Revenue
  - Same Principal Applies to
    - Grocery Stores: Tesco …
    - E-Commerce: Amazon …

Google	89.1%
bing	6.76%
Yahoo!	2.25%
MSN	1.08%

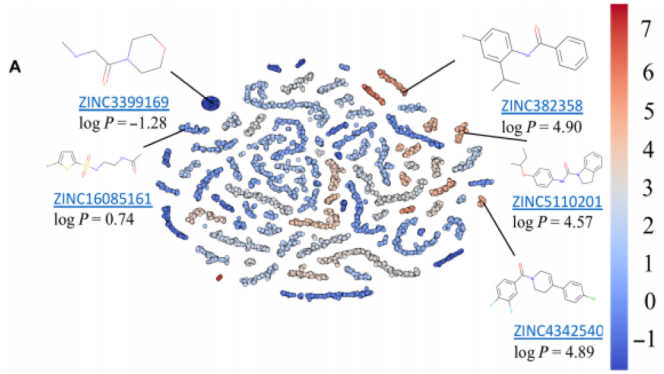
a) https://ardorseo.com/blog/how-many-google-searches-per-day-2018/
b) https://www.wordstream.com/blog/ws/2015/05/21/how-much-does-adwords-cost

Why Data Mining? – Scientific View

Improving health care and reducing costs



Predicting the impact of climate change



Drug design

Finding alternative/ green energy sources

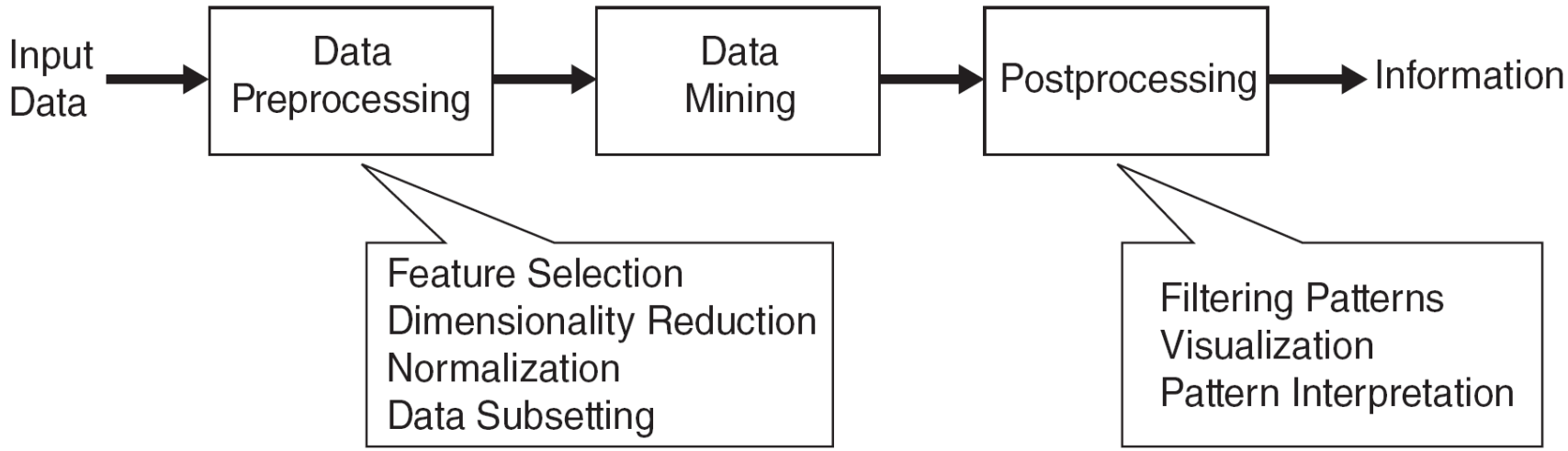
Reducing hunger and poverty by increasing agriculture production

Global economic analysis

a) Introduction to Data Mining, 2nd Edition, by Tan, Steinbach, Karpatne, Kumar
b) Deep reinforcement learning for de novo drug design, Science, 2018
c) https://finance.yahoo.com/news/happened-stock-market-today-221404831.html

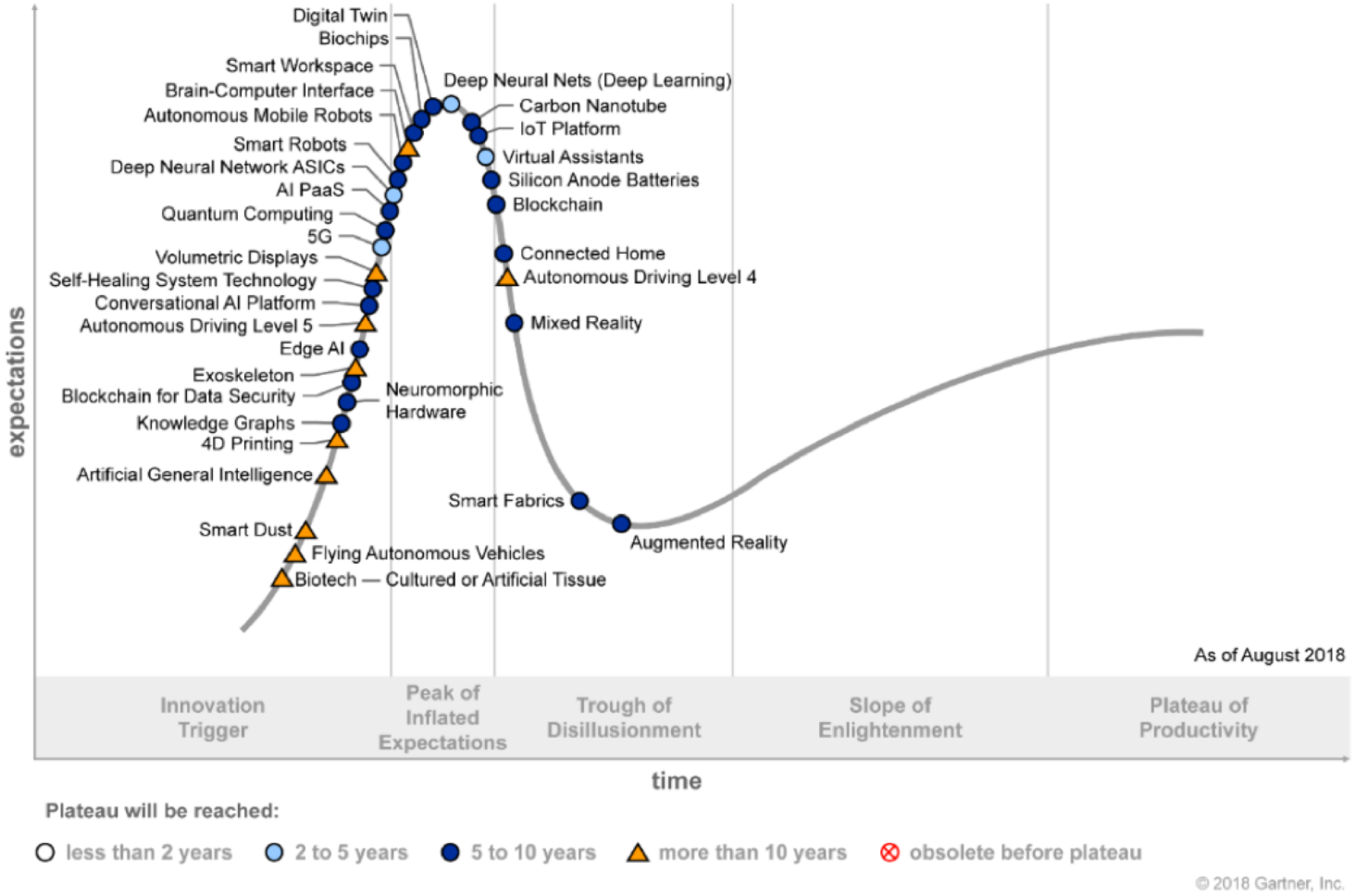
What is Data Mining?

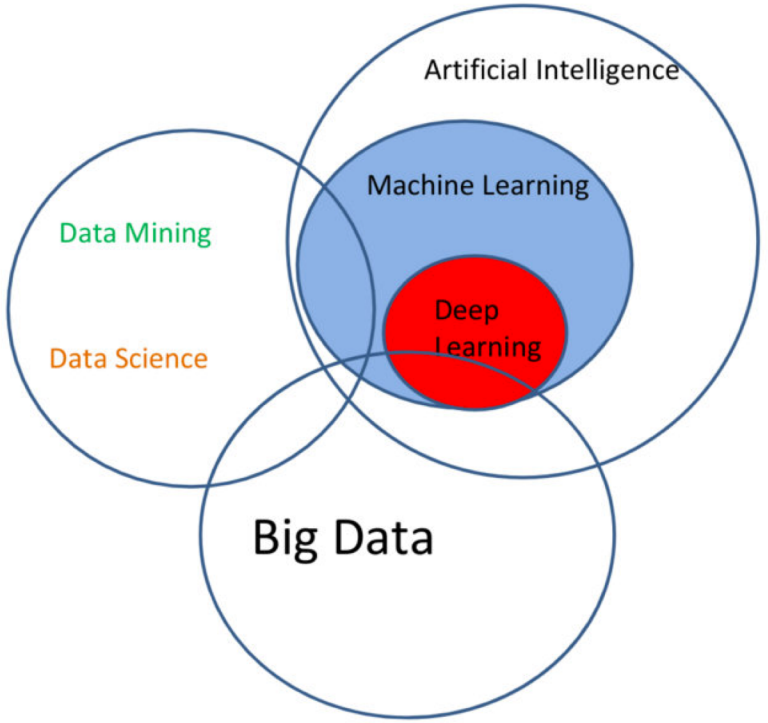
- Different Definitions
  - Non-trivial extraction of **implicit**, previously **unknown** and potentially **useful** information from data
  - Exploration & analysis, by automatic or semi-automatic means, of large quantities of data in order to discover meaningful patterns



a) Introduction to Data Mining, 2nd Edition, by Tan, Steinbach, Karpatne, Kumar

Data Mining and Emerging Techs

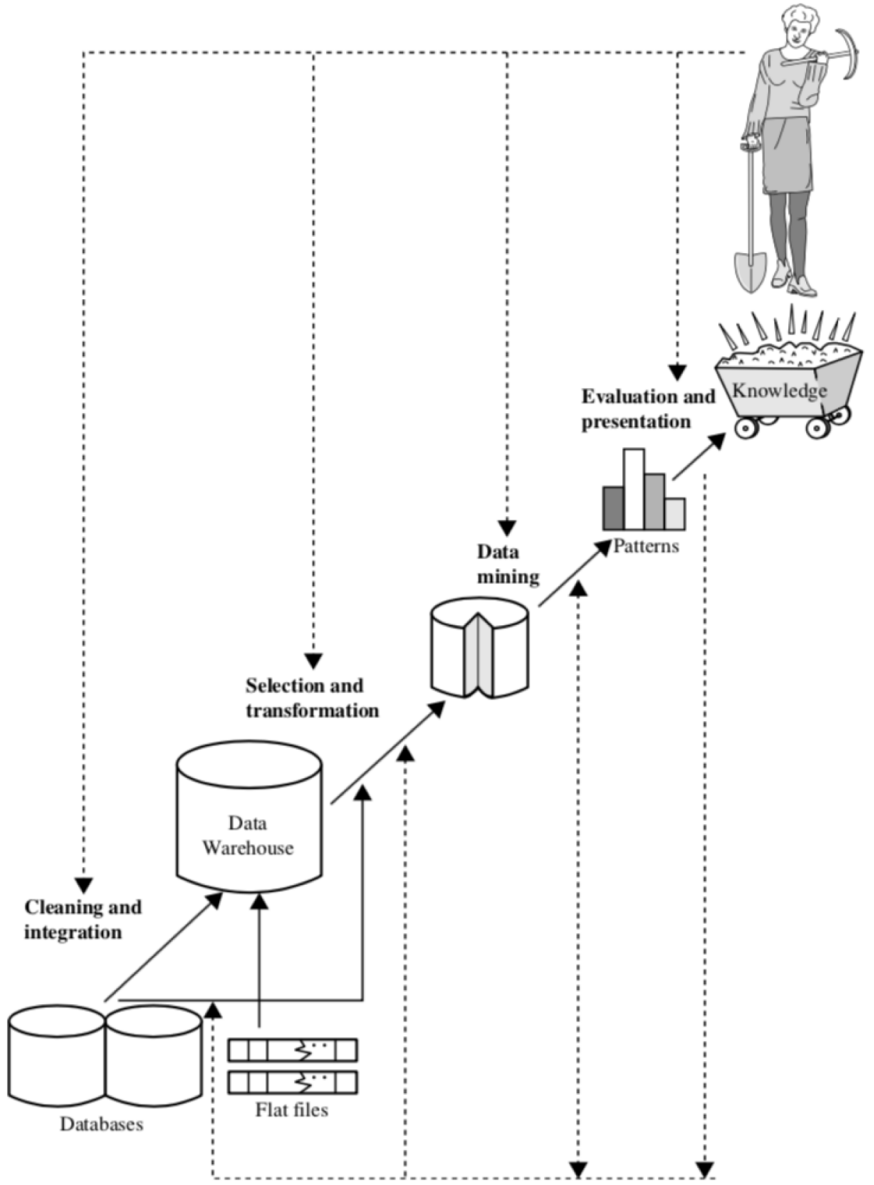




a) https://www.gartner.com/en/newsroom/press-releases/2018-08-20-gartner-identifies-five-emerging-technology-trends-that-will-blur-the-lines-between-human-and-machine
b) https://www.kdnuggets.com/2016/03/data-science-puzzle-explained.html

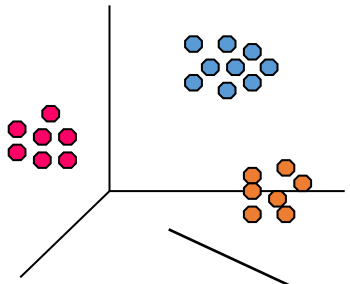
DM *v.s.* ML

- Data Mining
  - Data **Knowledge**
  - Use huge data to find property that is interesting to us.
  - Database system (Historical reason)
  - We found that …
- Machine Learning
  - **Intelligent** Machinery
  - To find a machine that is able to find interesting property.
  - Algorithm (Logic, Biology)
  - Machine can do …



a) Data Mining Concepts and Techniques, 3rd Edition, by Jiawei Han *et al.*

Data Mining Task

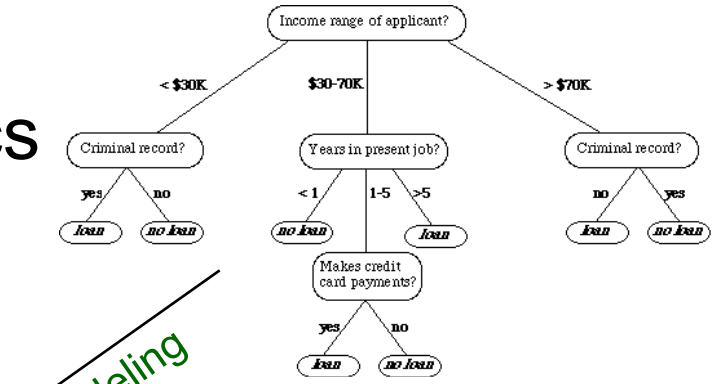


Probability & Statistics

Clustering

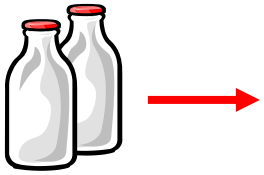
Data

Tid	Refund	Marital Status	Taxable Income	Cheat
1	Yes	Single	125K	No
2	No	Married	100K	No
3	No	Single	70K	No
4	Yes	Married	120K	No
5	No	Divorced	95K	Yes
6	No	Married	60K	No
7	Yes	Divorced	220K	No
8	No	Single	85K	Yes
9	No	Married	75K	No
10	No	Single	90K	Yes
11	No	Married	60K	No
12	Yes	Divorced	220K	No
13	No	Single	85K	Yes
14	No	Married	75K	No
15	No	Single	90K	Yes



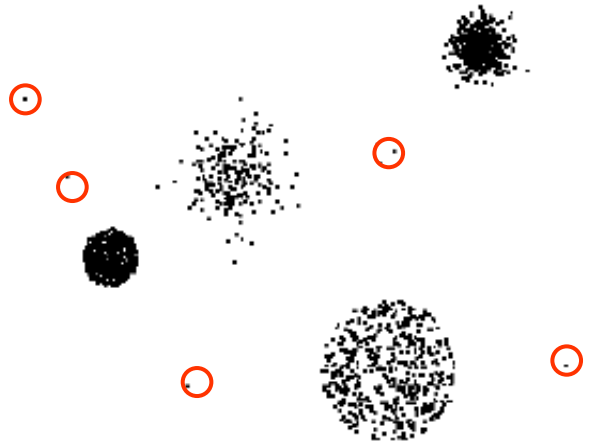
Predictive Modeling

Association Rules



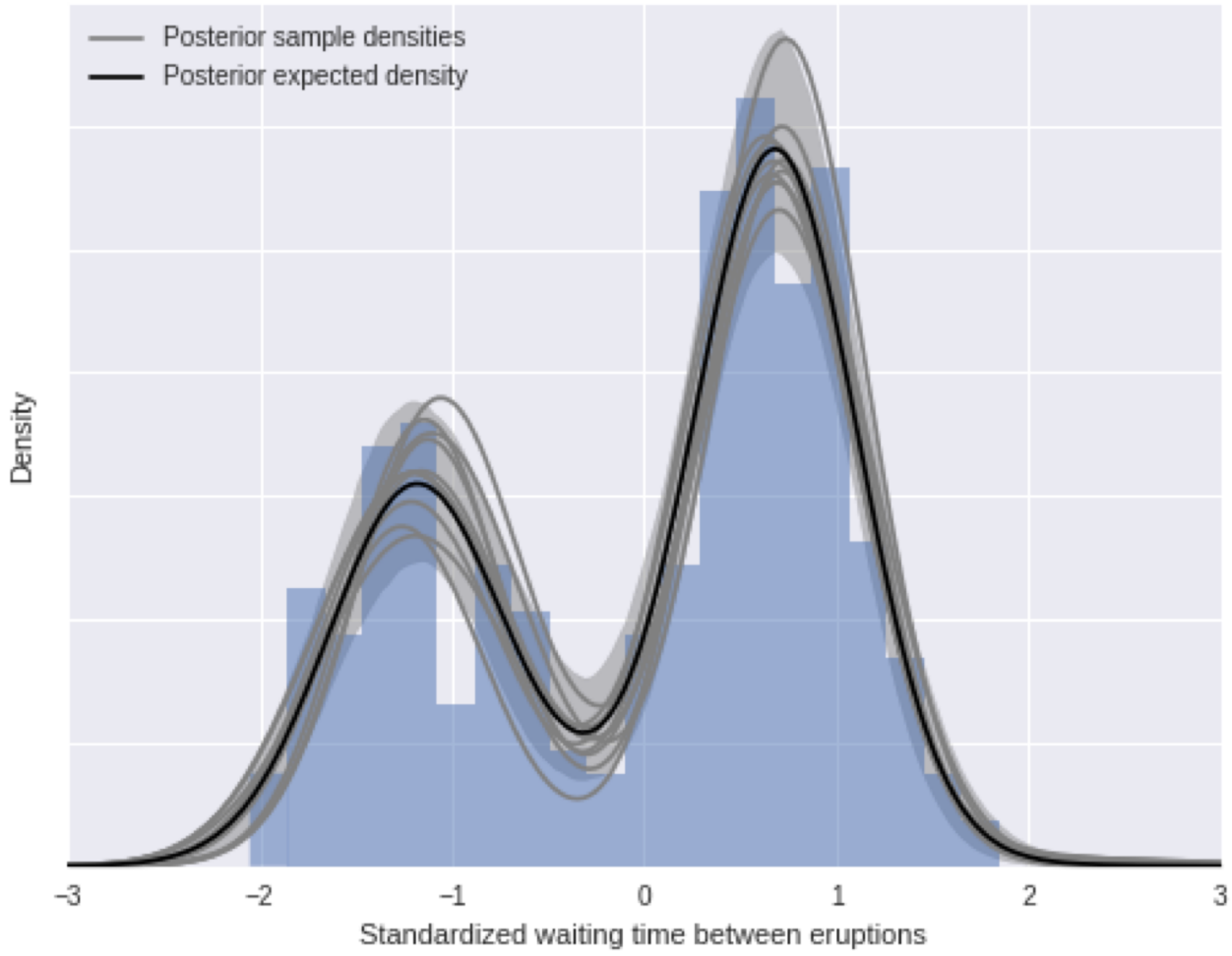
Milk

Anomaly Detection



a) Introduction to Data Mining, 2nd Edition, by Tan, Steinbach, Karpatne, Kumar

Probabilistic & Statistic Methods



BAYES' THEOREM

$$P(\text{Model} \mid \text{Data}) = \frac{P(\text{Model}) \times P(\text{Data} \mid \text{Model})}{P(\text{Data})}$$

Prior Distribution

Likelihood

Whatever this is, it is **constant**, so it affects all models the same

a) https://www.slideshare.net/andreslopezsepulcre/foundations-of-statistics-in-ecology-and-evolution-8-bayesian-statistics
b) https://en.wikipedia.org/wiki/Thomas_Bayes

Associate Rule

- Given a set of records each of which contain some number of items from a given collection
  - Produce dependency rules which will predict occurrence of an item based on occurrences of other items.

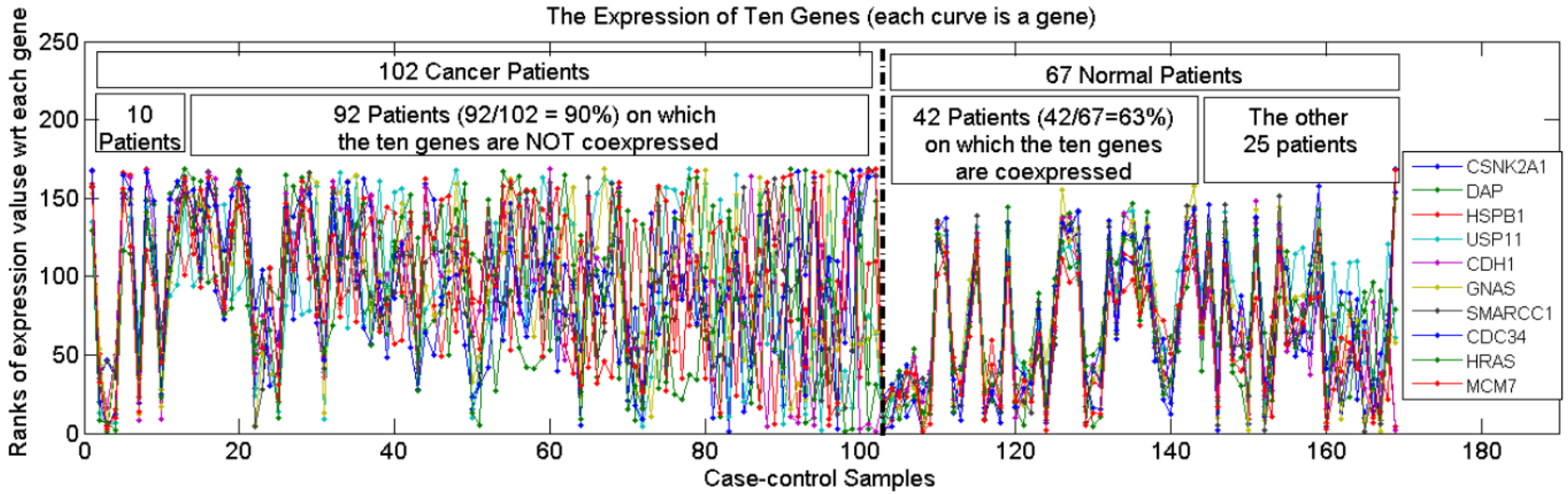
TID	*Items*
1	Bread, Coke, Milk
2	Beer, Bread
3	Beer, Coke, Diaper, Milk
4	Beer, Bread, Diaper, Milk
5	Coke, Diaper, Milk

Rules Discovered:
{Milk} --> {Coke}
{Diaper, Milk} --> {Beer}

a) Introduction to Data Mining, 2nd Edition, by Tan, Steinbach, Karpatne, Kumar

Associate Rule Application

- Market-basket analysis
  - Rules are used for sales promotion, shelf management, and inventory management
- Telecommunication alarm diagnosis
  - Rules are used to find combination of alarms that occur together frequently in the same time period
- Medical Informatics
  - Rules are used to find combination of patient symptoms and test results associated with certain diseases



a) Introduction to Data Mining, 2nd Edition, by Tan, Steinbach, Karpatne, Kumar

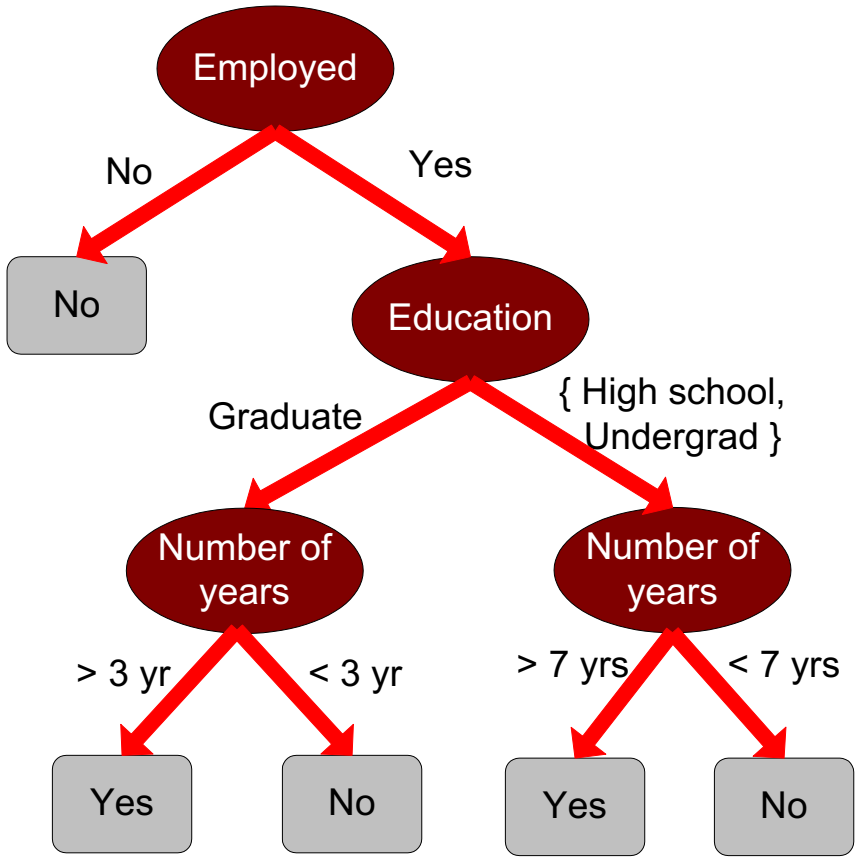
Prediction – Classification

Class

Tid	Employed	Level of Education	# years at present address	Credit Worthy
1	Yes	Graduate	5	Yes
2	Yes	High School	2	No
3	No	Undergrad	1	No
4	Yes	High School	10	Yes
…	…	…	…	…

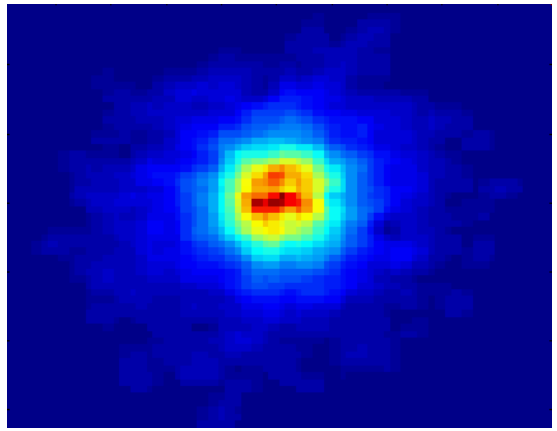
- Predict a value of a given discrete/categorical valued variable based on the values of other variables, assuming a linear or nonlinear model of dependency.

Model for predicting credit worthiness

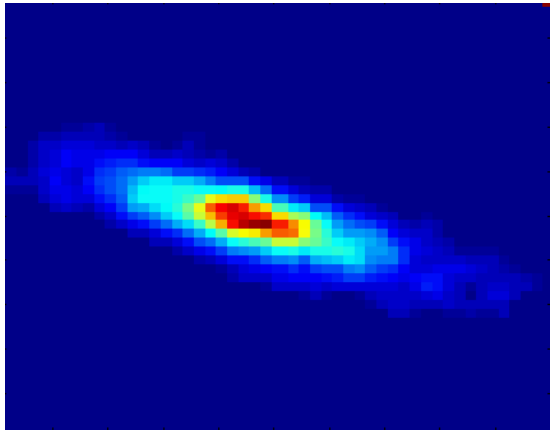


Classification – Example

Early



Intermediate

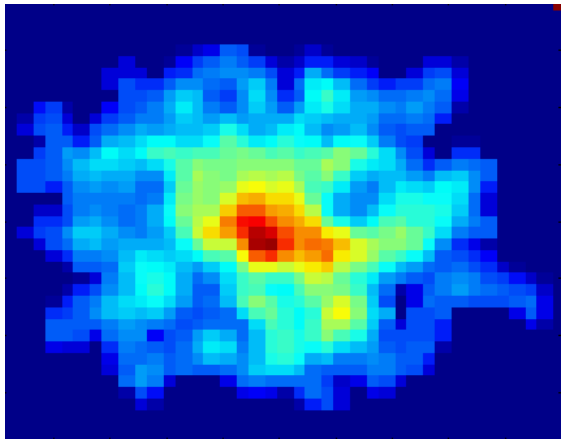


Class: Stages of Formation

Attributes:

- Image features,
- Characteristics of light waves received, etc.

Late



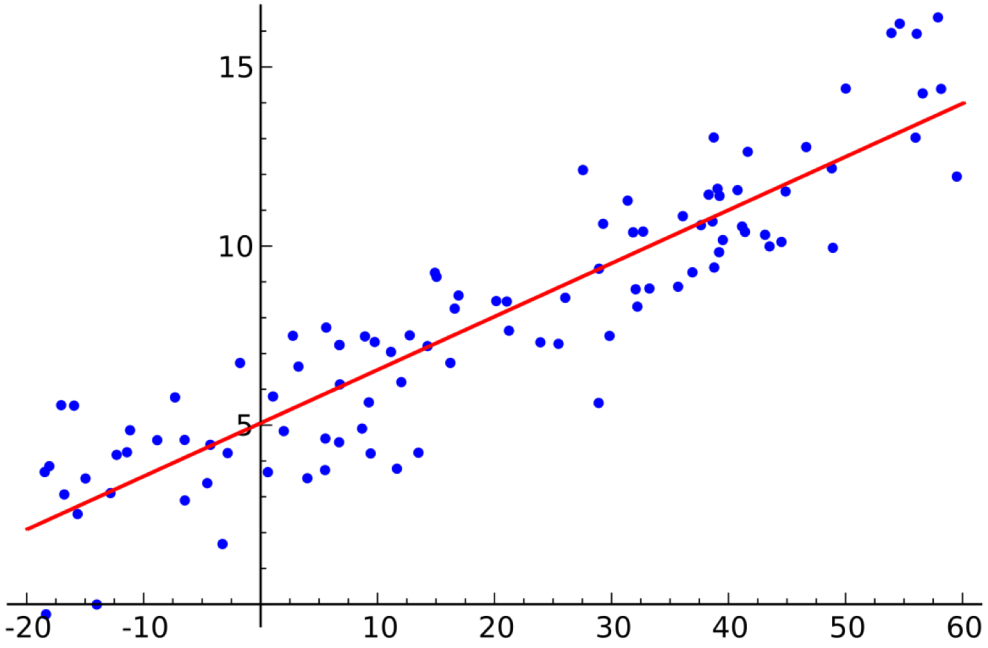
Data Size:

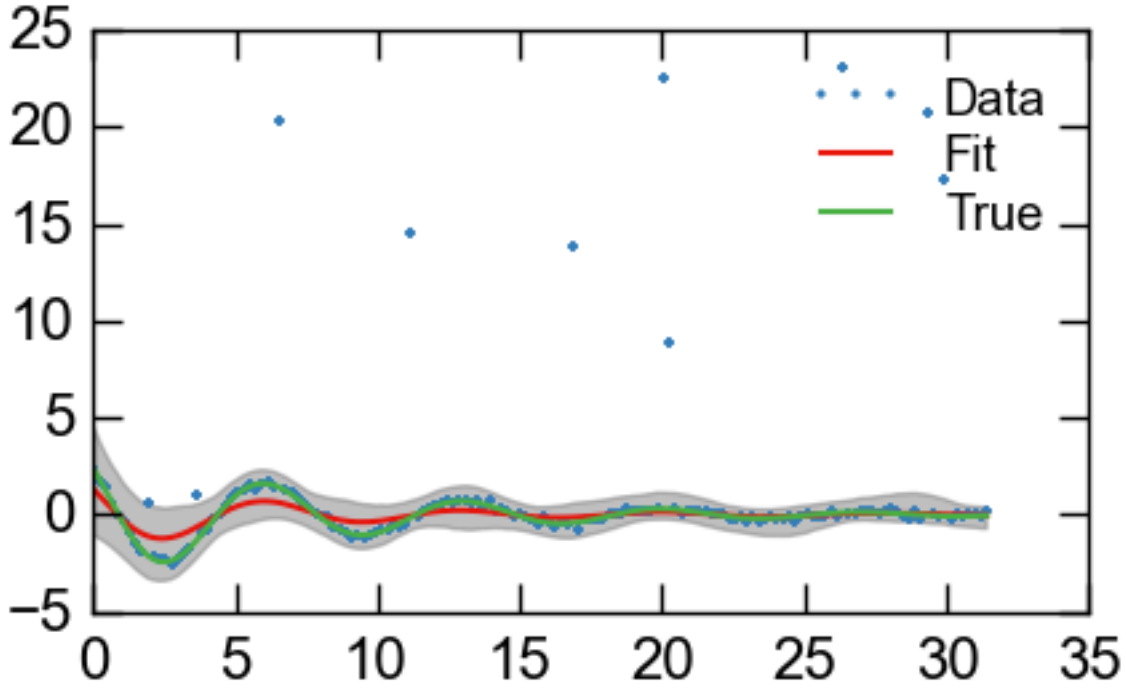
- 72 million stars, 20 million galaxies
- Object Catalog: 9 GB
- Image Database: 150 GB

a) Courtesy: http://aps.umn.edu

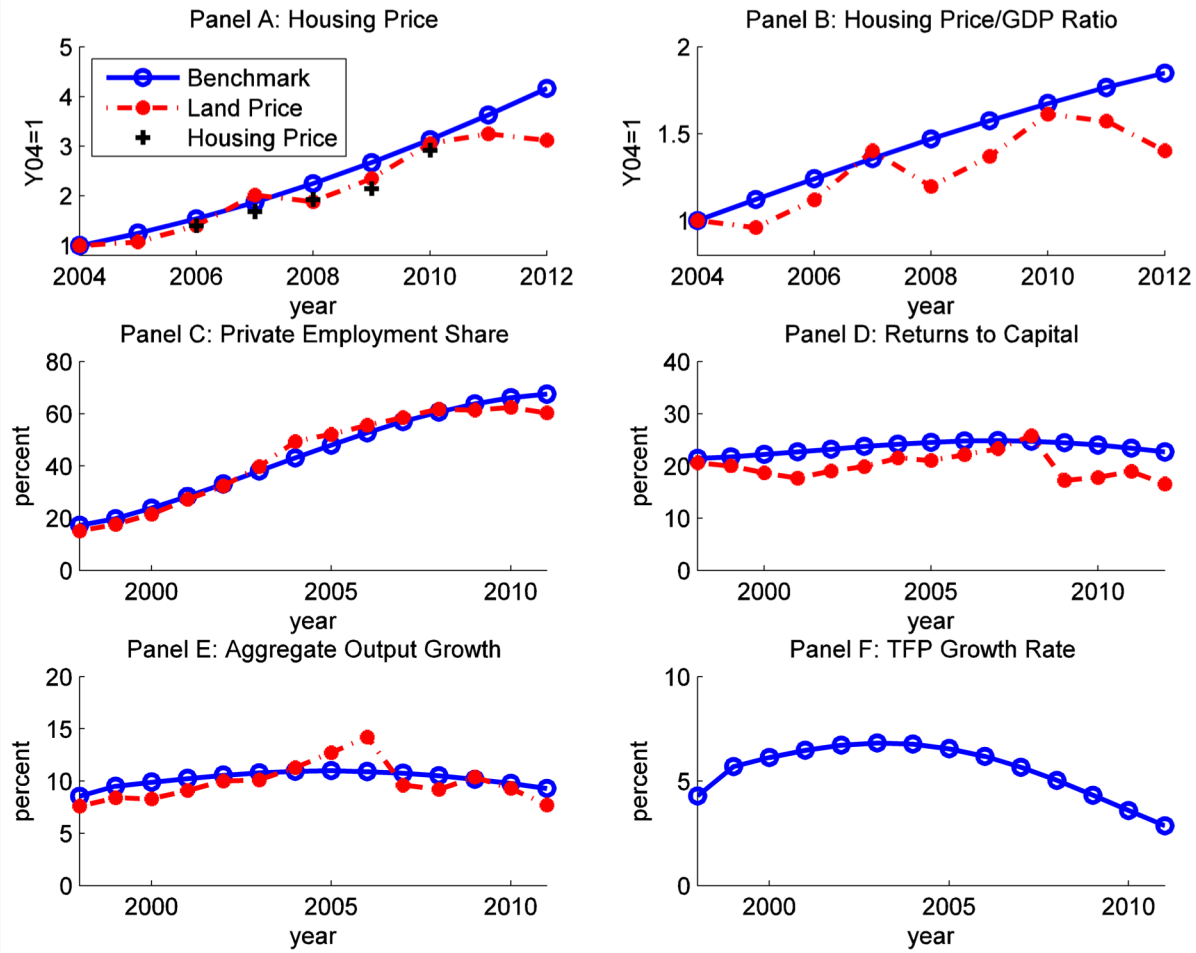
Prediction – Regression

- Predict a value of a given continuous valued variable based on the values of other variables, assuming a linear or nonlinear model of dependency.



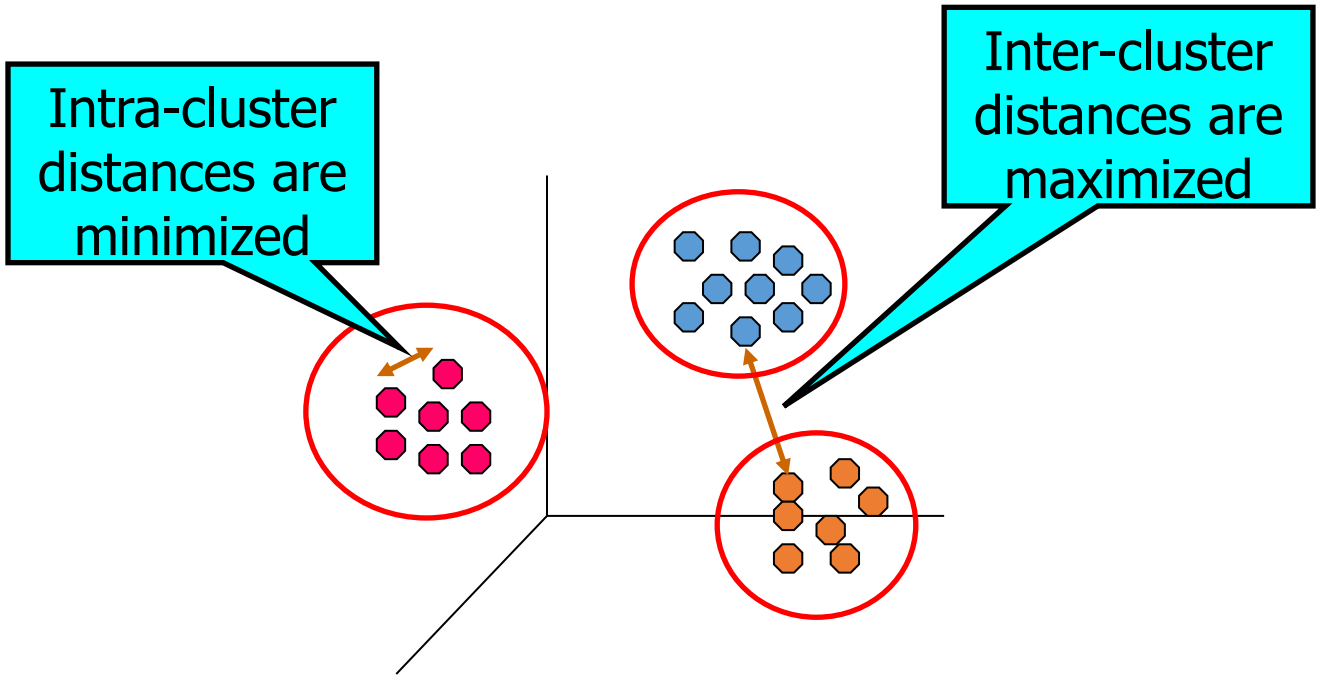


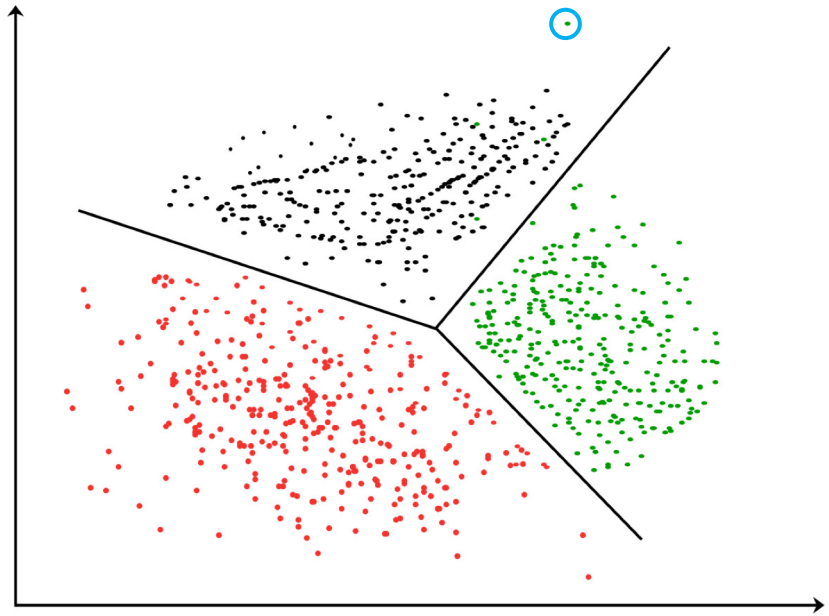
Regression – Example



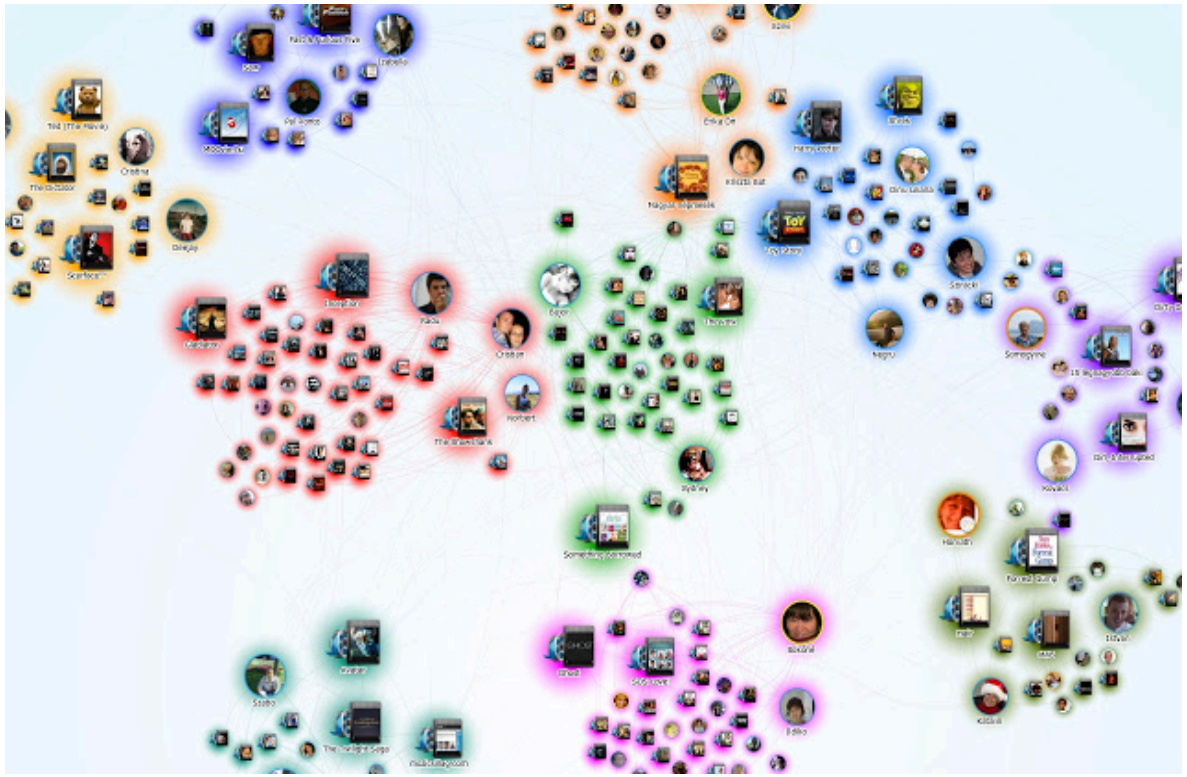
Clustering

- Finding groups of objects such that the objects in a group will be similar (or related) to one another and different from (or unrelated to) the objects in other groups

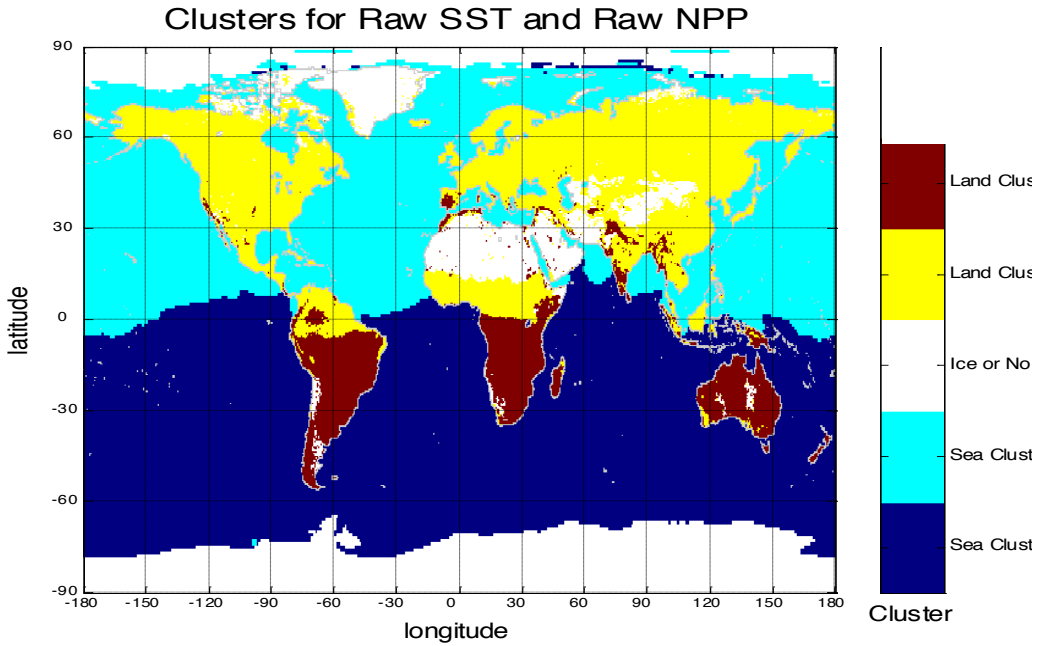




Clustering – Example



Clustering Social Activity Group via Facebook



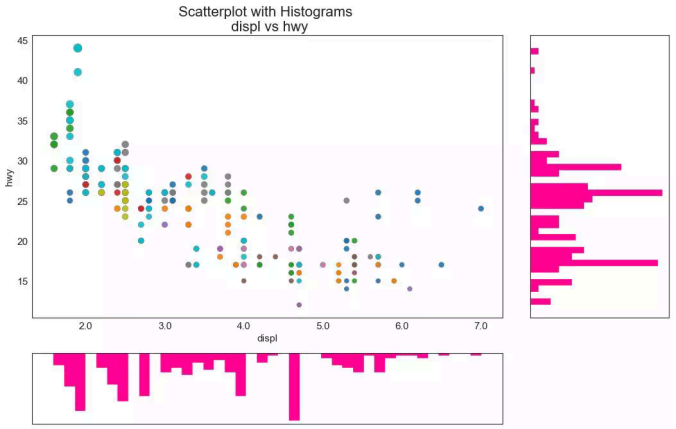
Use of K-means to partition Sea Surface Temperature (SST) and Net Primary Production (NPP) into clusters that reflect the Northern and Southern Hemispheres.

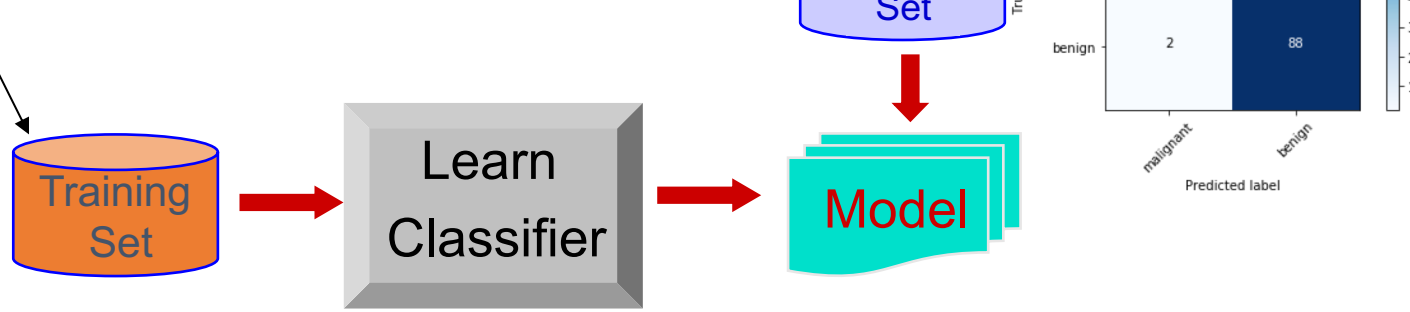
Typical Pipeline

categorical categorical quantitative class

Tid	Employed	Level of Education	# years at present address	Credit Worthy
1	Yes	Graduate	5	Yes
2	Yes	High School	2	No
3	No	Undergrad	1	No
4	Yes	High School	10	Yes
…	…	…	…	…

Tid	Employed	Level of Education	# years at present address	Credit Worthy
1	Yes	Undergrad	7	?
2	No	Graduate	3	?
3	Yes	High School	2	?
…	…	…	…	…

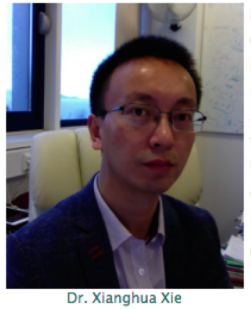




Advanced Topic

- Computer Vision: Image Data
  - CSCM77: Computer Vision and Deep Learning
  - Biomedical Image Analysis
- Quantitative Finance: Time-Series Pricing Data
- Intelligent Networking: Graph and Network Data
- Healthcare Informatics: Medical Record Data
- Scientific Simulation: Fluid Dynamic for Cardiovascular System
- *More …*
  - **SwanseaVision Group @ Swansea University**
  - **Research Focus: Computer Vision And Machine Learning**

Principal Investigator



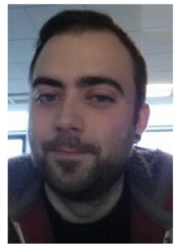
Dr. Xianghua Xie is an Associate Professor in the Visual Computing Group at the Department of Computer Science, Swansea University. He held an RCUK Academic Fellowship between September 2009 and March 2012, and he was a Senior Lecturer between October 2012 and March 2013. Prior to his position at Swansea, he was a Research Associate in the Computer Vision Group, Department of Computer Science, University of Bristol, where he obtained his PhD in Computer Science and MSc in Advanced Computing (with commendation) in 2006 and 2002, respectively.

CLICK HERE TO FIND OUT MORE...

Dr. Xianghua Xie

Academic Member

Dr. Jingjing Deng, Lecturer



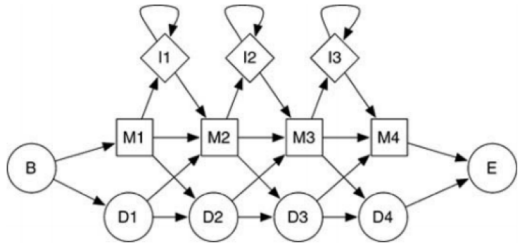
Dr. Michael Edwards, Assistant Lecturer

a) http://csvision.swansea.ac.uk/

Application – Protein Analysis

```
VSPAGMASGYD
:  |  |  ||  ||
I-P-GKAS-YD
```

Matched amino acids sequence



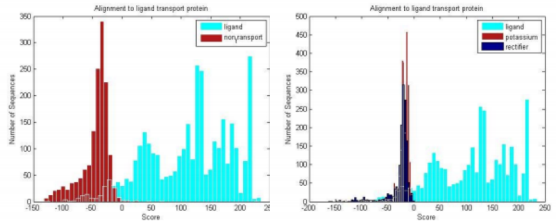


Fig. 5. *Left:* Non-transport and ligand proteins aligned to the ligand HMM model. *Right:* the three different sub-types of transmembrane proteins aligned to the ligand HMM model.

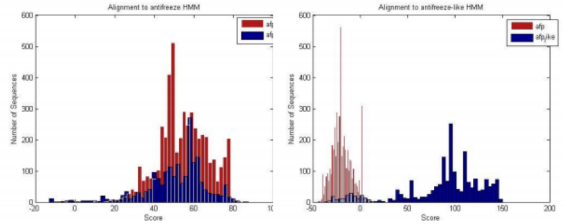
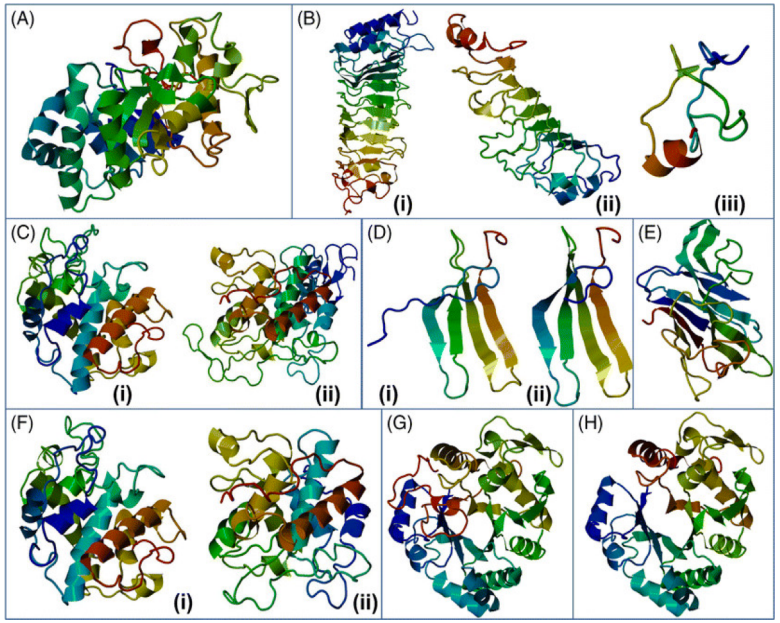


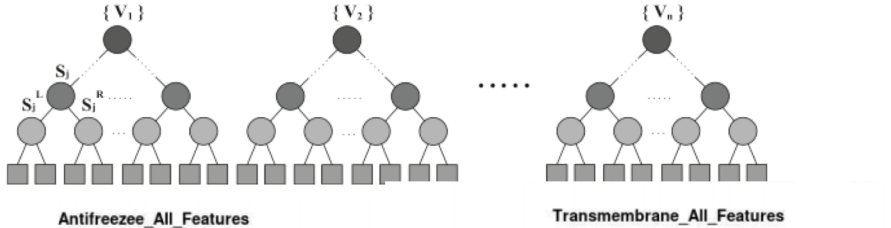
Fig. 6. *Left:* Non-transport and ligand proteins aligned to the ligand HMM model. *Right:* the three different sub-types of antifreeze proteins aligned to the ligand HMM model.

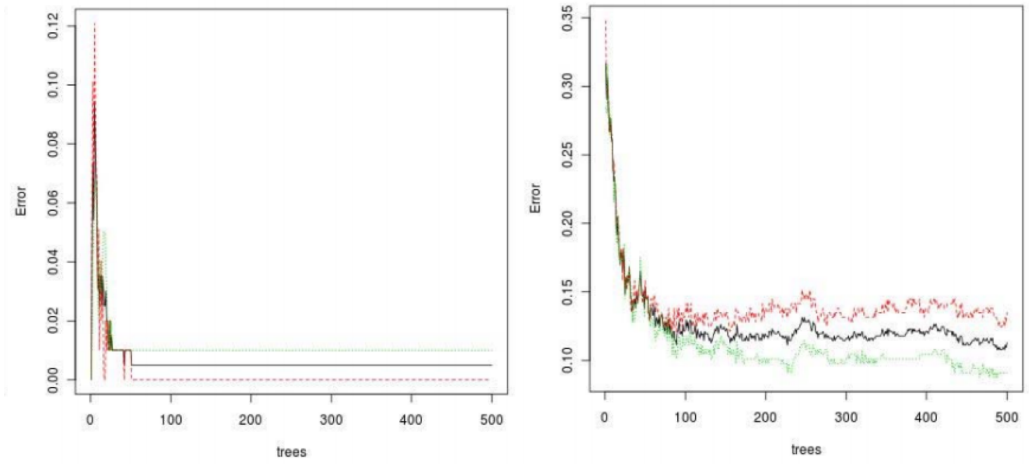
Protein features

#	Features	No. of features
1	Amino acids	20
2	Functional groups	17
3	Chemical properties	6
4	Secondary Structure	60
	Total	**93**

1) *Frequency of amino acids:* The frequency of each of the 20 naturally occuring acids was calculated.
2) *Frequency of functional groups:* The amino acids found within each protein sequence were categorized into 17 functional groups such as phenylene, valine, leucine, proline and hydroxyl, where the frequency of each functional group was calculated.
3) *Secondary Structure* Frequency of helix, beta sheet and coil structures within each protein were predicted using PSIPRED. The frequency of each amino acid found within each secondary structure element was calculated. PSIPRED [11].
4) *Physio-chemical properties:* The frequency of physio-chemical properties of each protein was derived from the amino acid index (AAINDEX) [12] database, and their methodologies were used to calculate the isoelectric point, aromaticity, grand average of hydropathicity index, instability index as well as molecular mass of all protein sequences [13].

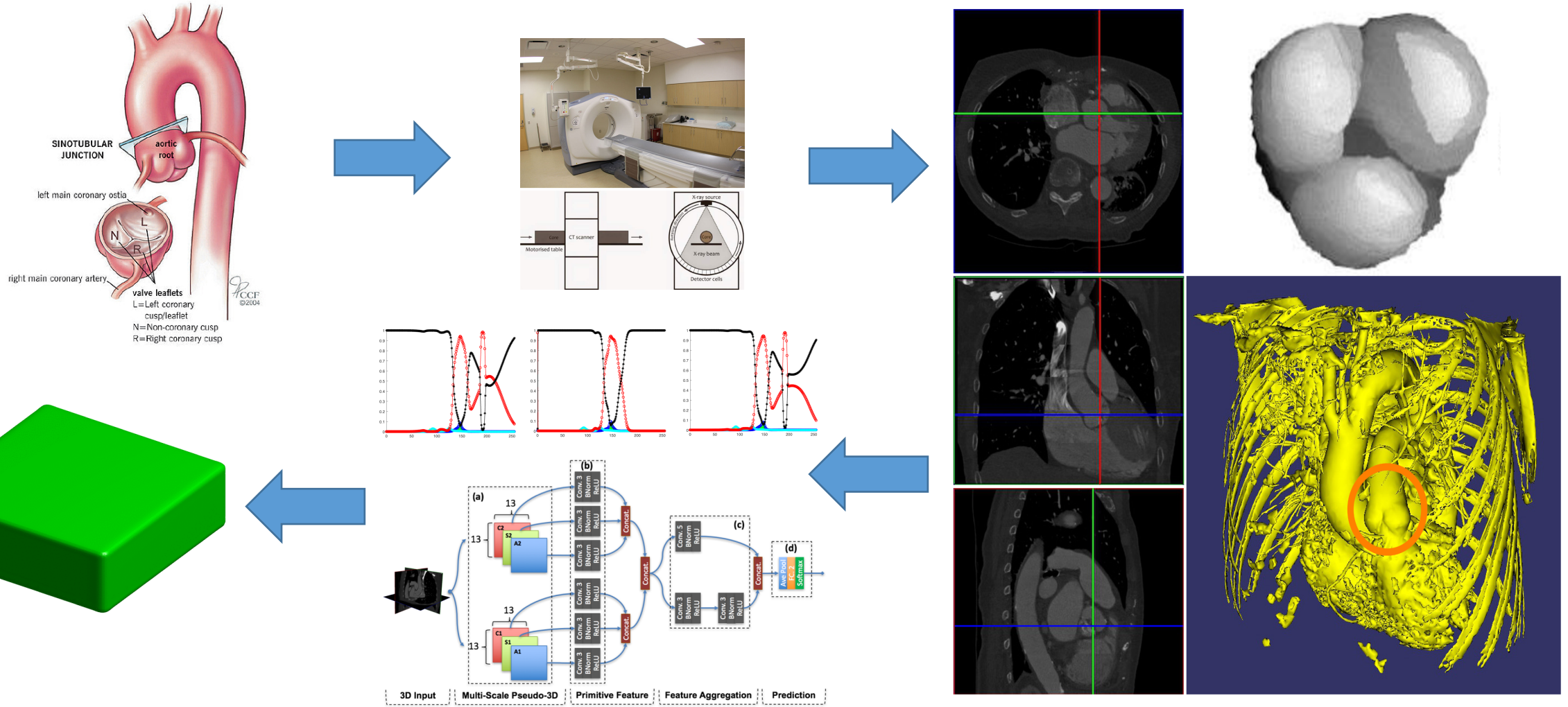






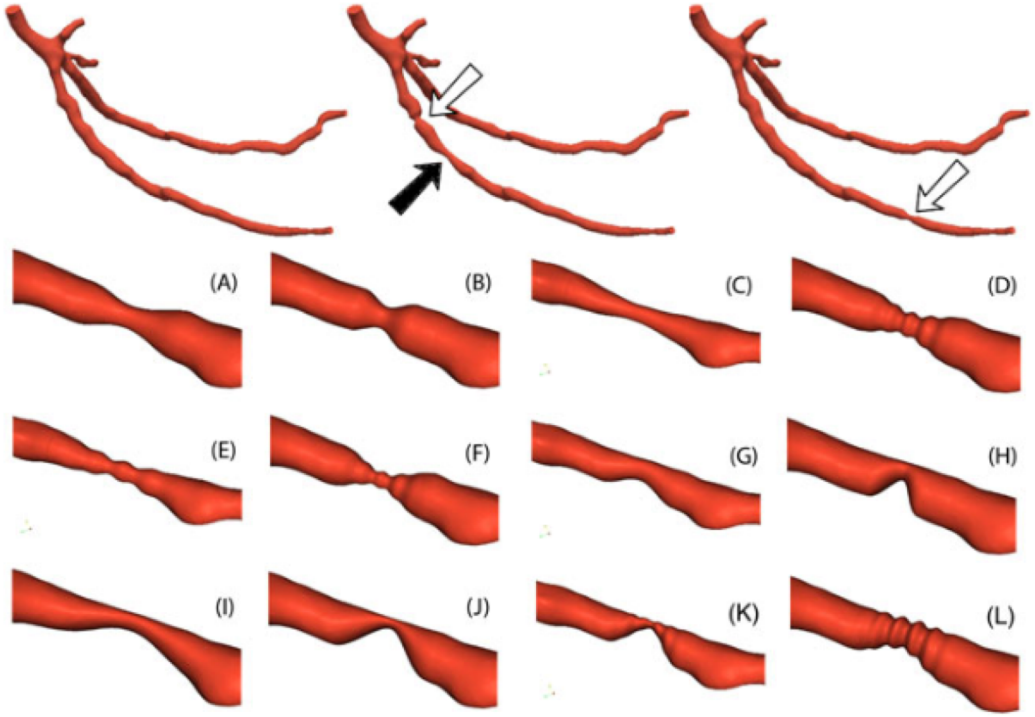
a) A. Lacey, J. Deng, and X. Xie, Protein Classification using Hidden Markov Models and Randomised Decision Trees, In Proc. Int'l Conf. BioMedical Engineering and Informatics, pp. 659-664, October 2014.

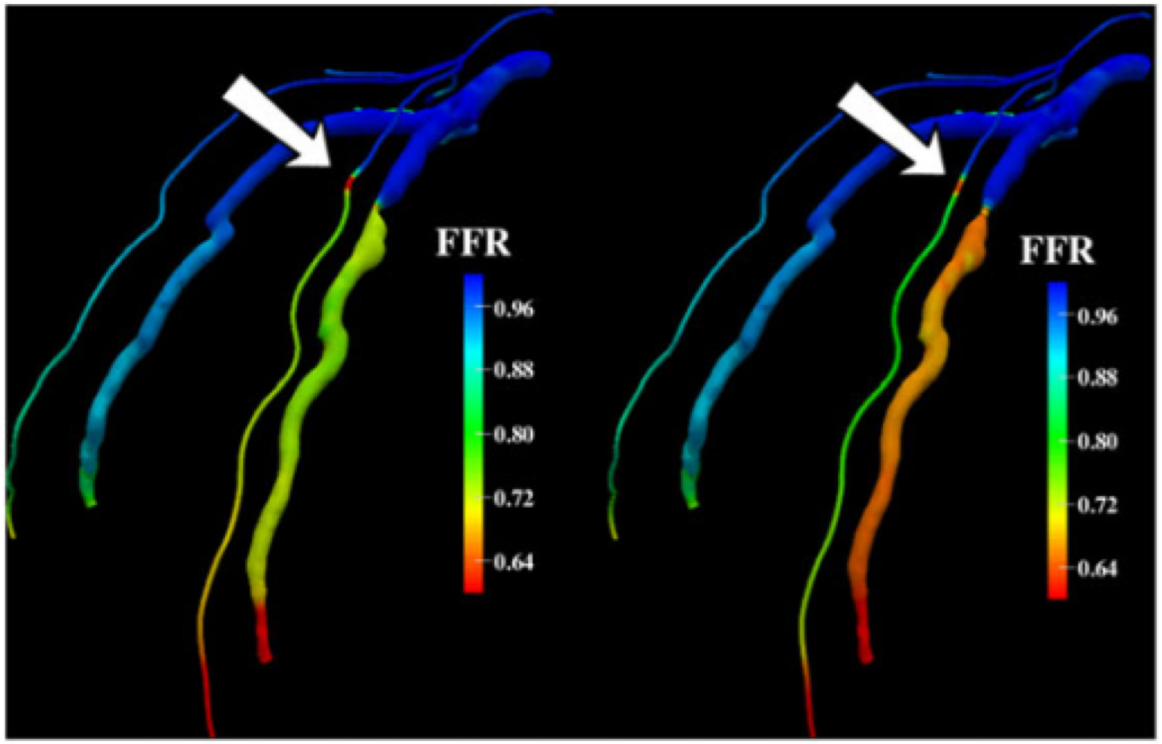
Application – Medical Imaging Analysis



a) Jingjing Deng, Xianghua Xie, 3D Interactive Segmentation with Semi-Implicit Representation and Active Learning, 2018, Under Review

Application – Fluid Dynamic Simulation





a) E. Boileau, S. Pant, C. Roobottom, I. Sazonov, J. Deng, X. Xie, and P. Nithiarasu, Estimating the Accuracy of a Reduced-Order Model for the Calculation of Fractional Flow Reserve (FFR), International Journal for Numerical Methods in Biomedical Engineering, 2017.

Application – Healthcare Informatics

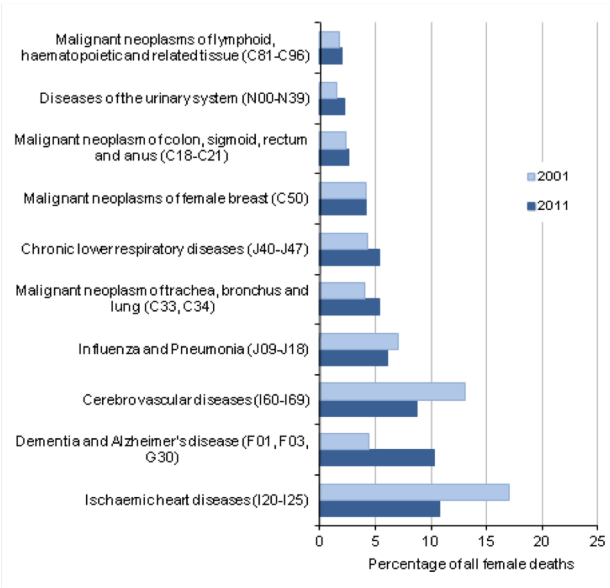
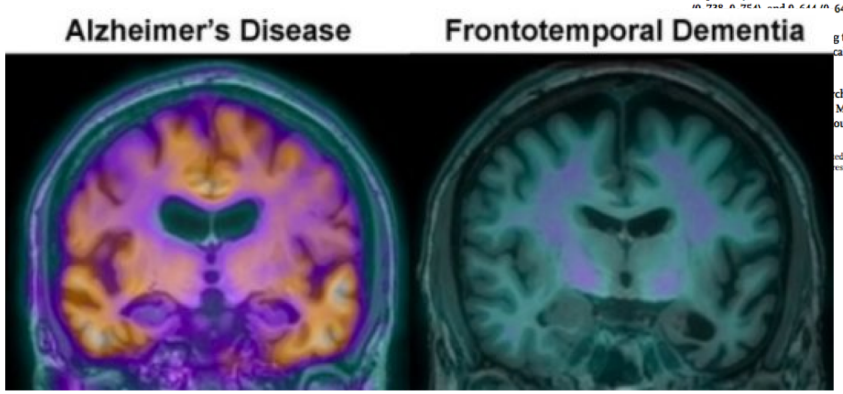


TABLE IV
TOP 10 EVENT CODES RANKED IN ASC. ORDER OF IMPORTANCE AS DETERMINED BY THE PROPOSED METHODOLOGY.

Importance	Event CD	Definition
0.227	9N32.	Third Party Encounter
0.231	E2749	Nightmares
0.234	dh12.	Serc-16 Tablet
0.234	ja1I.	Ibugel gel 100g
0.247	1323.	Social group 3 - skilled
0.254	bxd5.	Simvastatin 40mg tablet
0.259	44Uz.	Blood glucose raised NOS
0.300	n473.	Influvac sub-unit prefilled syringe 0.5mL
0.318	ip3j.	Adcal-D3 1.5g/10ug chewable tablet
0.481	G20..	Essential Hypertension



THE LANCET

Meeting Abstracts

Mining electronic health records to identify influential predictors associated with hospital admission of patients with dementia: an artificial intelligence approach

Shang-Ming Zhou, Gavin Tsang, Xianghua Xie, Lin Huo, Sinead Brophy, Ronan A Lyons

Abstract

Background Early prediction of the outcomes of dementia is important and challenging. This study aimed to identify influential predictors from primary care electronic health records that can robustly predict whether patients with dementia will be admitted to hospital or remain under GP care.

Methods Health records of patients with dementia were collected from general practice (GP) and hospital data in Wales between 1980 and 2015. These records were linked at individual patient level via the Secure Anonymised Information Linkage databank. The GP records of each patient were selected 1 year before diagnosis up to hospital admission. An artificial intelligence technique, neural network with entropy regularisation (a multilayer feedforward neural network whose weights between input layer and the first hidden layer were regularised by an entropy metric into the fitness function during training process) was used to automatically identify the most influential predictors from initial GP read codes, sex, and age. 10-fold cross validation was used to assess the predictive performance of the identified signals.

Findings 52·5 million individual records of 59298 patients (20674 men, 38624 women) with dementia were used. 30178 were admitted to hospital and 29120 remained with GP care. More men were admitted to hospital than stayed with GP care (11233 *vs* 9441), whereas more women stayed with GP care than were admitted (19679 *vs* 18945). From the 54649 initial event codes, the ten most important signals identified for admission for dementia were two diagnostic events (nightmares, essential hypertension), five medication events (betahistine dihydrochloride, ibuprofen gel, simvastatin, influenza vaccine, calcium carbonate and colecalciferol chewable tablets), and three procedural events (third party encounter, social group 3—skilled, blood glucose raised). They performed significantly above chance to predict admission to hospital with sensitivity of 0·758 (95% CI 0·731–0·785), specificity 0·759 (0·71–0·808), precision 0·766 (0·735–0·797), and negative predictive value 0·751 (0·741–0·761). Linear regression with all raw features yielded values of 0·286 (0·26–0·313), 0·804 (0·792–0·816), 0·487 (0·463–0·511), and 0·633 (0·615–0·651), respectively, and with ten identified features yielded values of 0·684 (0·679–0·691), 0·712 (0·705–0·718), 0·747 (0·738–0·754) and 0·644 (0·64–0·651).

Published Online November 22, 2018

Health Data Research UK Wales and Northern Ireland Site, Swansea University Medical School, Swansea, UK (S-M Zhou PhD, S Brophy PhD, R A Lyons MD); and China-ASEAN Research Institute, Guangxi University, Nanning, China (L Huo PhD); and Department of Computer Science Swansea University, Swansea, UK (G Tsang MSc, X Xie PhD)

Correspondence to: Dr Shang-Ming Zhou, Health Data Research UK Wales and Northern Ireland Site, Swansea University Medical School, Swansea SA2 8PP, UK s.zhou@swansea.ac.uk

or

Dr Lin Huo, China-ASEAN Research Institute, Guangxi University, 530004, Nanning, China Lhuo@gxu.edu.cn

g traditional methods, the artificial intelligence technique provides an effective means of
al signals to predict hospital admission of patients with dementia significantly

ch UK Wales and Northern Ireland Site, National Centre for Population Health and
Major Project of National Social Science Foundation of China (16ZDA0092), Guangxi
oud Big Data Security and Mining Technology Innovation Team.

ed data in the cohort. GT, S-MZ, and XX conceived the methods. GT and S-MZ completed the experiments.
esults. S-MZ and LH wrote the abstract. RL, SB, and S-MZ and LH interpreted the results. All authors reviewed

9

a) Shang-Ming Zhou, Gavin Tsang, Xianghua Xie, Lin Huo, Sinead Brophy, and Ronan A Lyons, Mining Electronic Health Records to Identify Influential Predictors Associated with Hospitalisation of Dementia Patients: An Artificial Intelligence Approach, **The Lancet**, 2018. accepted.
b) https://www.ons.gov.uk/peoplepopulationandcommunity/birthsdeathsandmarriages/deaths/bulletins/deathsregisteredinenglandandwalesseriesdr/2012-11-06
c) https://www.sciencedaily.com/releases/2016/02/160224133634.htm

Application – Financial Prediction

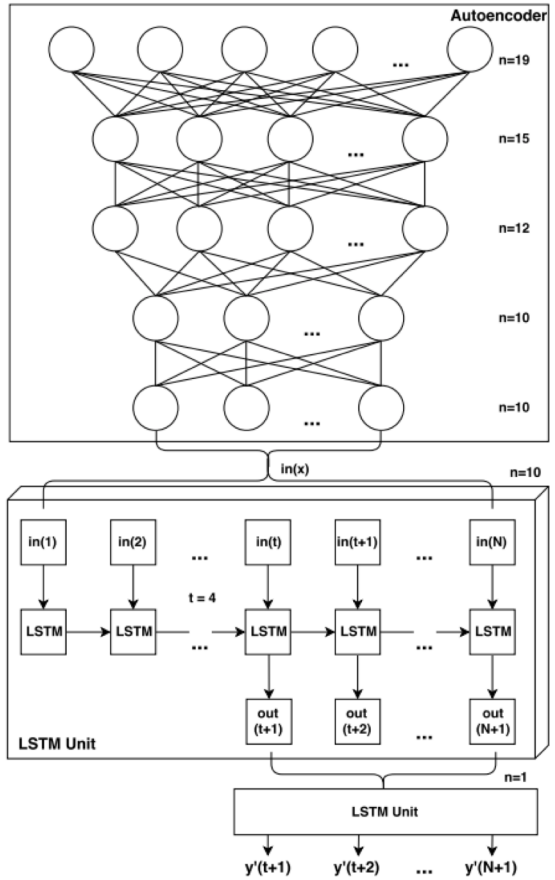
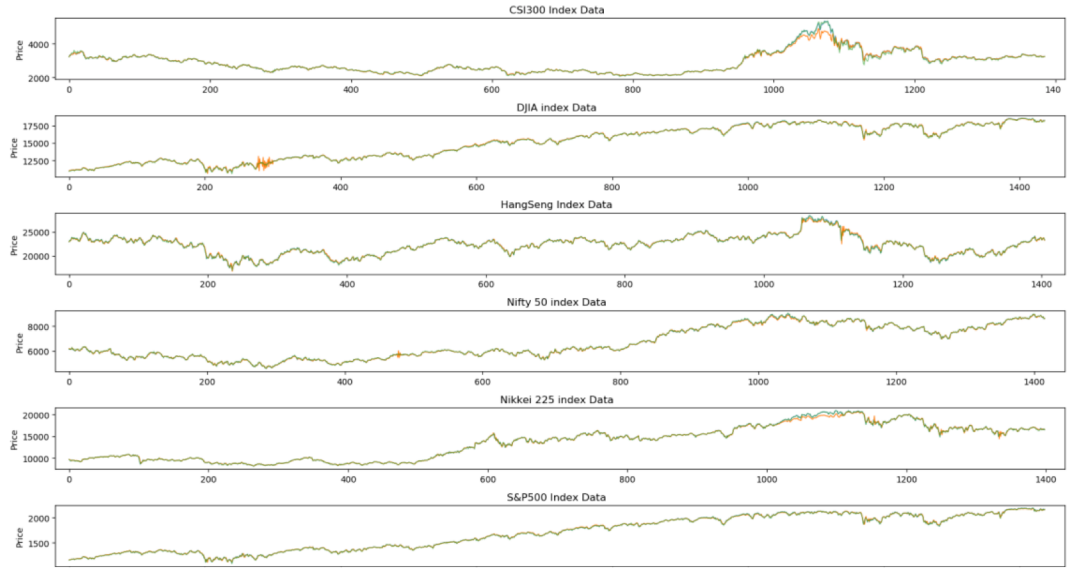


Fig. 3. Diagram of the proposed neural network architecture.

Daily Trading Data	
Open/Close Price	Nominal daily open/close price
High/Low Price	Nominal daily highest/lowest price
Trading Volume	Daily trading volume
Technical Indicators	
MACD	Moving average convergence divergence
CCI	Commodity channel index
ATR	Average true range
BOLL	Bollinger band
EMA20	20 day exponential moving average
MA5/MA10	5/10 day moving average
MTM6/MTM12	6/12 month momentum
ROC	Price rate of change
SMI	Stochastic Momentum Index
WVAD	Williams' variable accumulation/distribution
Economic Factors	
Exchange Rate	US dollar index
Interest Rate	Interbank offered interest rate

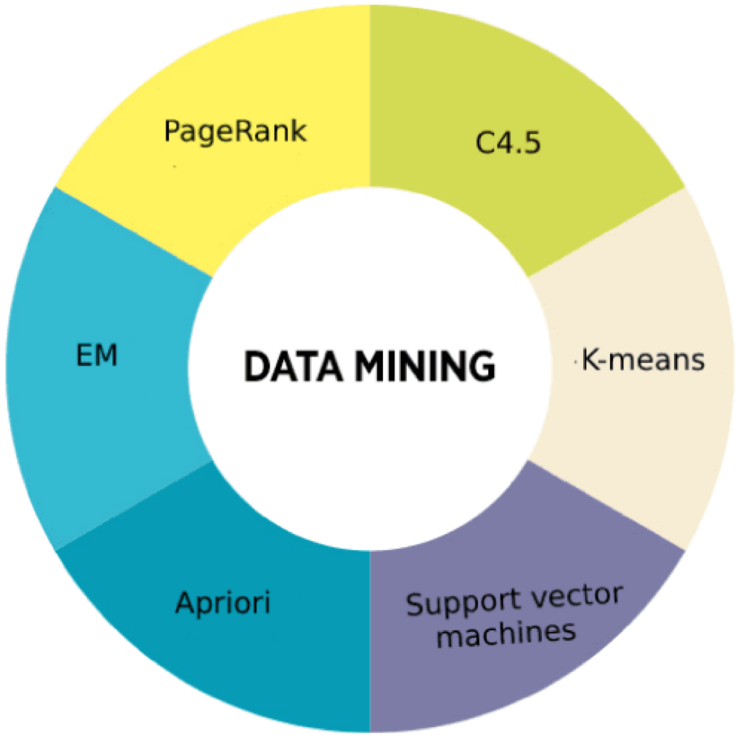


a) Gavin Tsang, Jing Jing Deng, Xianghua Xie, Recurrent Neural Networks for Financial Time-Series Modelling, International Conference on Pattern Recognition, 2018.
b) Alex Momotov, Xianghua Xie, News Sentiment Mining Using LDA-Based Thematic Modelling and Lexicon Approaches, Technical Report, 2018

Challenge

- Scalability
  - Data Scaling: Google BigTable,
  - Computation Scaling: Cloud, HPC, GPU, ASIC, FPGA Chips
- High Dimensionality
  - Large Sense Network
- Heterogeneous and Complex Data
  - Novel Database Techniques: Graph DB
- Data Ownership and Distribution
  - Data Privacy Protection – GDPR
- Non-Traditional Analysis
  - Deep Neural Networks, Reinforcement Learning

Top 10 Data Mining Algorithm



Top Data Mining Algorithms

Knowl Inf Syst (2008) 14:1–37
DOI 10.1007/s10115-007-0114-2

SURVEY PAPER

Top 10 algorithms in data mining

Xindong Wu · Vipin Kumar · J. Ross Quinlan · Joydeep Ghosh · Qiang Yang · Hiroshi Motoda · Geoffrey J. McLachlan · Angus Ng · Bing Liu · Philip S. Yu · Zhi-Hua Zhou · Michael Steinbach · David J. Hand · Dan Steinberg

Received: 9 July 2007 / Revised: 28 September 2007 / Accepted: 8 October 2007
Published online: 4 December 2007

Abstract This paper presents the top 10 data mining algorithms identified by the IEEE International Conference on Data Mining (ICDM) in December 2006: C4.5, *k*-Means, SVM, Apriori, EM, PageRank, AdaBoost, *k*NN, Naive Bayes, and CART. These top 10 algorithms are among the most influential data mining algorithms in the research community. With each algorithm, we provide a description of the algorithm, discuss the impact of the algorithm, and review current and further research on the algorithm. These 10 algorithms cover classification,

X. Wu (✉)
Department of Computer Science, University of Vermont, Burlington, VT, USA
e-mail: xwu@cs.uvm.edu

V. Kumar
Department of Computer Science and Engineering,
University of Minnesota, Minneapolis, MN, USA
e-mail: kumar@cs.umn.edu

J. Ross Quinlan
Rulequest Research Pty Ltd,
St Ives, NSW, Australia
e-mail: quinlan@rulequest.com

J. Ghosh
Department of Electrical and Computer Engineering,
University of Texas at Austin, Austin, TX 78712, USA
e-mail: ghosh@ece.utexas.edu

Q. Yang
Department of Computer Science,
Hong Kong University of Science and Technology,
Honkong, China
e-mail: qyang@cs.ust.hk

H. Motoda
AFOSR/AOARD and Osaka University,
7-23-17 Roppongi, Minato-ku, Tokyo 106-0032, Japan
e-mail: motoda@ar.sanken.osaka-u.ac.jp

Springer

a) Wu, Xindong, Vipin Kumar, J. Ross Quinlan, Joydeep Ghosh, Qiang Yang, Hiroshi Motoda, Geoffrey J. McLachlan et al. "Top 10 algorithms in data mining." *Knowledge and information systems*14, no. 1 (2008): 1-37.