Data Mining
資料探勘

Syllabus

Hung-Yu Kao, Fall 2019

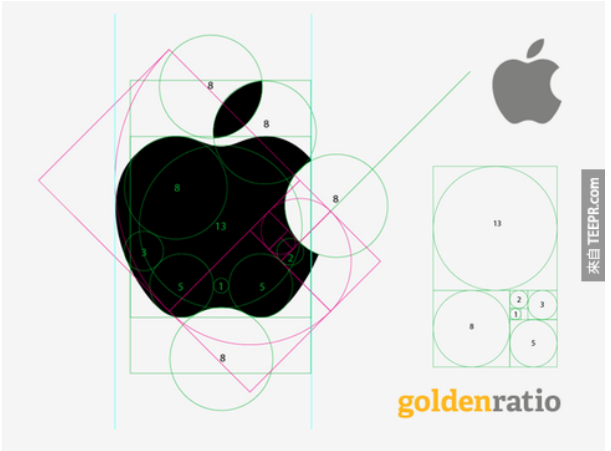
What do you see/recognize ?

FedEx Express

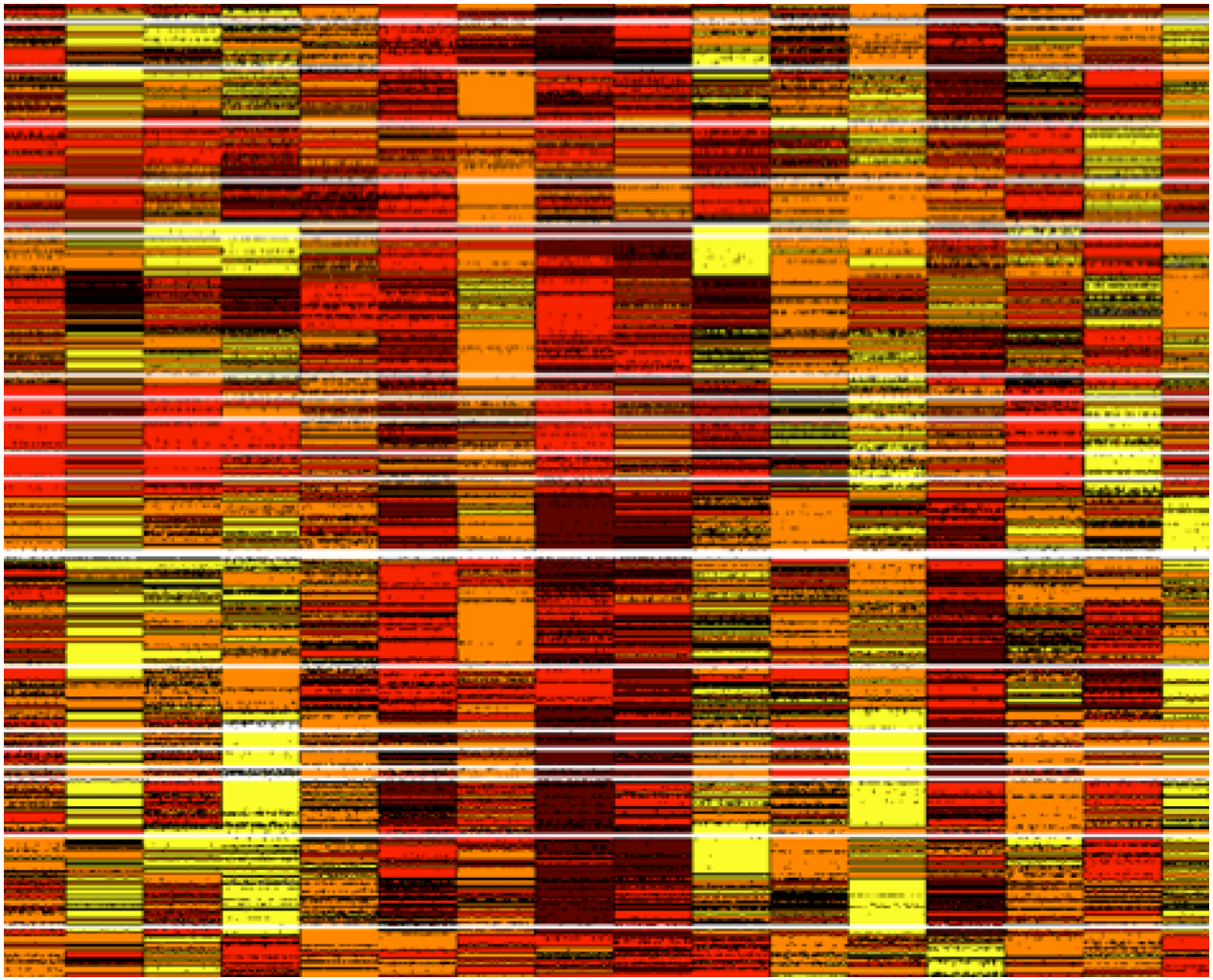
VAIO

amazon.com

Mining: Get Hidden/non-trivial Information!



What is this? What do you get from this figure?



Powerful !

HOW GOOGLE IS USING TARGETED ADVERTISING TO STOP ISIS RECRUITS

BY UTILIZING SEARCH TO FUNNEL BETTER INFORMATION

By **Kate Baggaley** September 8, 2016

http://www.popsci.com/how-google-is-using-targeted-advertising-to-stop-isis-recruits

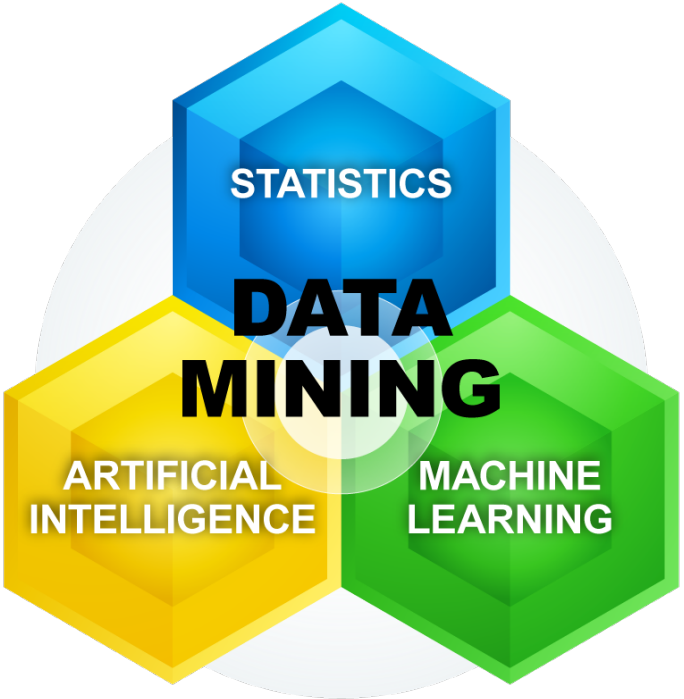
But, you will do like that…

Alien Statistics !

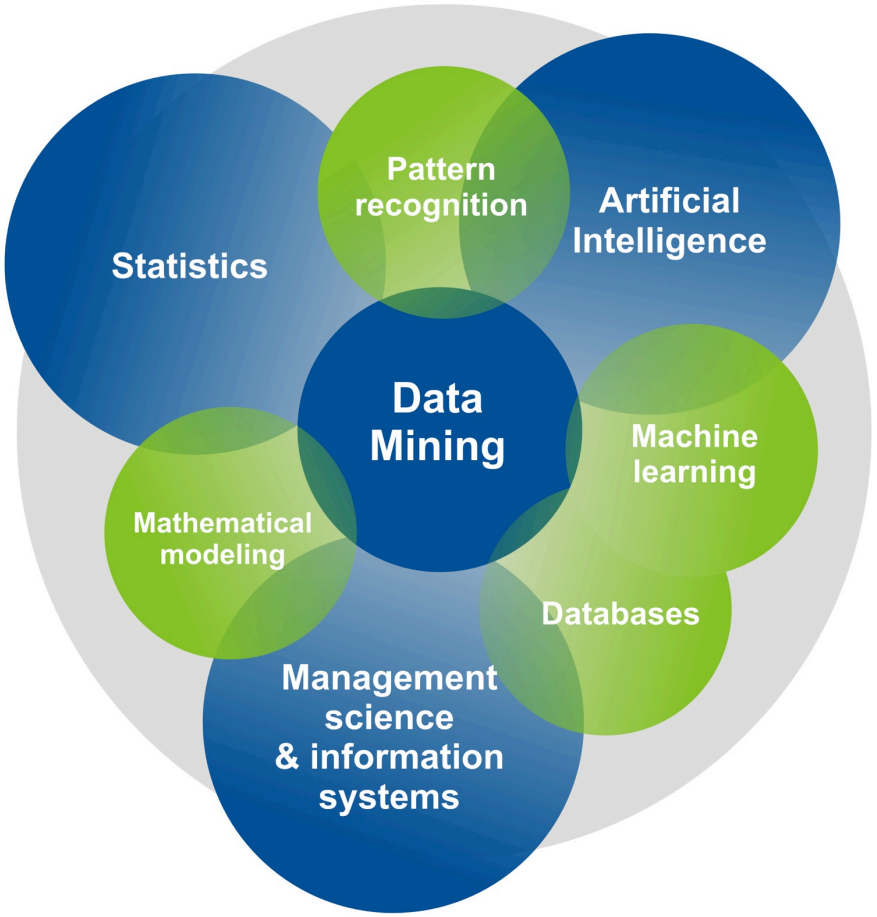
In this course

- Data Mining
  - Association Rule Mining
  - Frequent Pattern Mining / Time Series Mining
  - Classification / Prediction
    - Machine Learning (partial)
  - Cluster Analysis
  - Text Mining
    - Natural Language Processing (partial)
  - Graph / Link Mining

What's their difference?

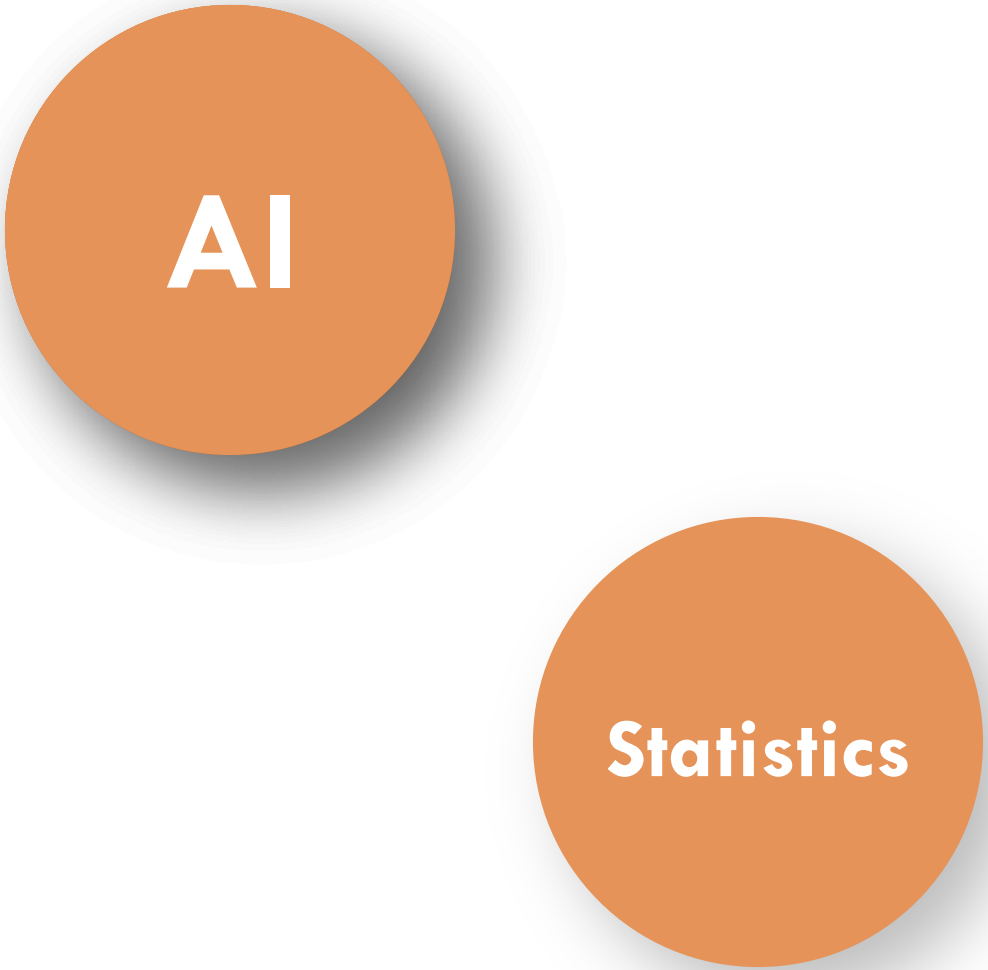


from sas.com

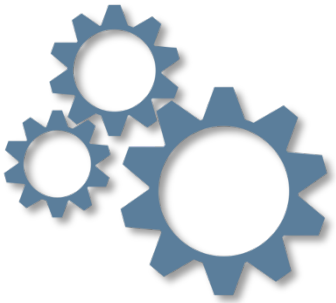


from http://frontender.com/blog/enablers/data-mining/

What's their difference?



Artificial Intelligence



Reasoning

Artificial Intelligence

Knowledge

Learning

Communication

Planning

Machine Learning

- *"Gives computers the ability to learn without being explicitly programmed" -- Arthur Samuel, 1959.*

Learning from experience

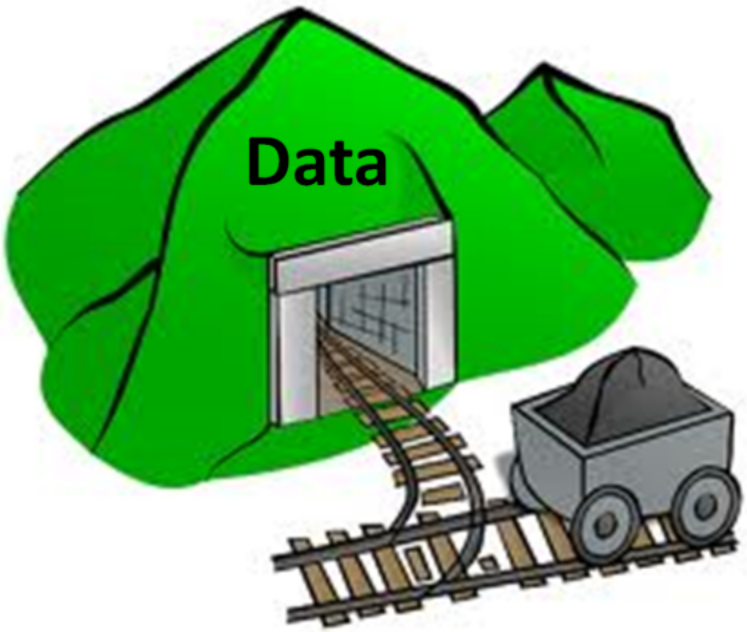
Teaching, Reading/ Observation, Memorizing/ Understanding

Learning from ~~experience~~ Data

Designed instructions

Data Mining

- Looking for Gold nuggets in the Data



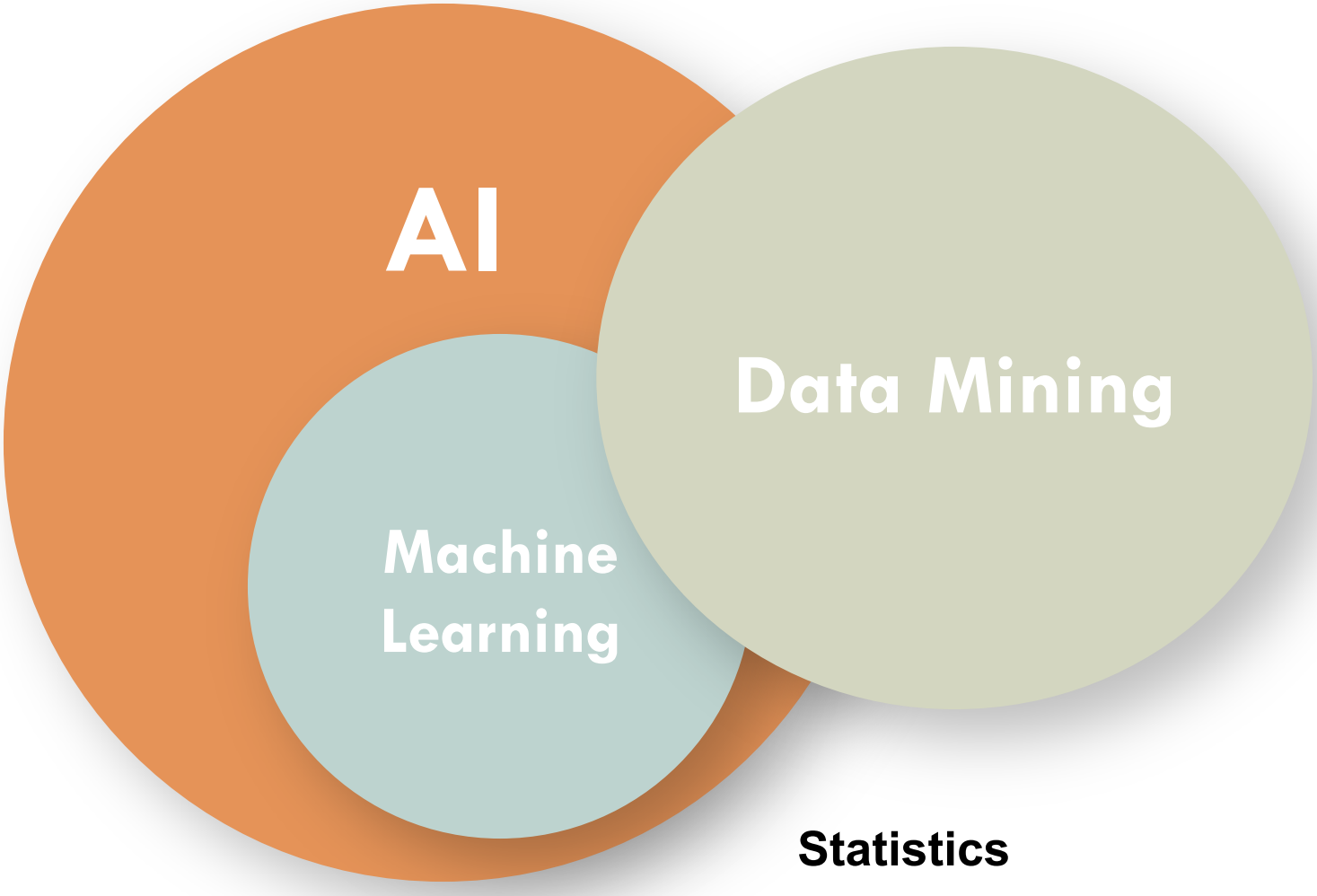
Useful knowledge (patterns)

How ?

Data Mining Techniques

- Classification
- Clustering
- Regression
- Outlier Detection
- Sequential Pattern
- Prediction
- Association Rules

What's their difference?

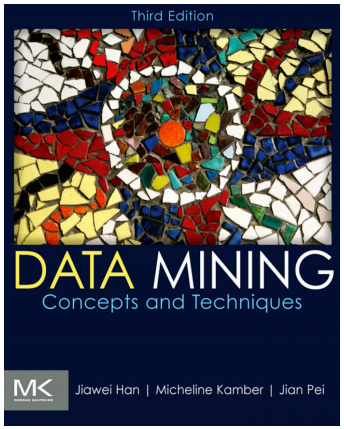


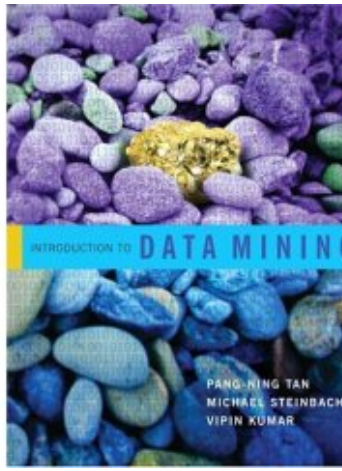
In this course

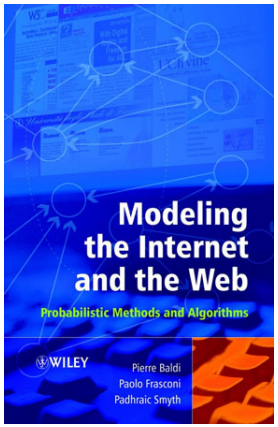
- Things not to expect …
  - Fundamental Database
    - Data warehouse
  - Fundamental Machine Learning
- Prerequisites
  - Basic programming skill
  - Background in database

Text book & References

- "Data Mining: Concepts and Techniques, "Jiawei Han, Micheline Kamber and Jian Pei, http://hanj.cs.illinois.edu/bk3/
- "Introduction to Data Mining," **Pang-Ning Tan**, Michael Steinbach, **Vipin Kumar, http://www-users.cs.umn.edu/~kumar/dmbook/index.php**
- "Modeling the Internet and the Web -- Probabilistic Methods and Algorithms," P. Baldi, P. Frasconi, P. Smyth. WILEY, 2003
- Web data Mining - Exploring Hyperlinks, Contents and Usage Data", By Bing Liu, Second Edition, Springer, July 2011, ISBN 978-3-642-19459-7







Grading

- **4** Projects: 20, 15, 15, 20% (group)
  - The first 3 projects are technical projects, and the last one trends to be a research/competition one.
  - The topic of the last project may related to a competition.
- **1** midterm: 15%
- **1** competition: 15% (group) (from kaggle)
  - Some selected datasets from Kaggle
  - Kaggle Kernel construction

Projects

Project ID	Title / description	Estimated working hours	Coding Difficulty	Demo	Discussion
1	**Association rule / sequential pattern**	20	High	Selected	Yes
2	**Classification / clustering**	20	Medium	Selected	Yes
3	**Social Text Mining**: Crawl, Classify, and Detect event on Social Text	20	Medium	Selected	Yes
4	**Final project**	(depend)	High	Yes	Very High
5	**Competition**	(depend)	High	Yes	High

Schedule

	Topics	*#Slides*
1 (9/10)	Syllabus	
2	Data Mining Fundamental	
3	Data Mining Fundamental, Project 1	
4	Classification / Clustering	
5	Classification / Clustering	
6	Classification / Clustering / Evaluation	
7	Classification / Clustering, Project 2	
8	Text Mining	
9	Text Mining	
10	Text Mining / Competition Introduction	
11	Graph / Link Mining, Project 3	
12	Exam	
13	Competition Report (1/3) (Problem)	
14	Competition Report (2/3) (solution)	
15	Final project report (1/2)	
16	Competition Report (3/3) (kaggle kernel)	
17 (12/31)	Homework Presentation	
18 (1/7)	Final project report (2/2)	

Instructors and TAs

- *Instructors:*
  - *Hung-Yu Kao 資訊系館 12F, Room 65C11 (hykao@mail.ncku.edu.tw)*
  - *http://myweb.ncku.edu.tw/~hykao*
- *TAs:*
  - *資訊系館 9F, Room 65903 IKM lab.*
  - *nckudm@gmail.com*
- *Course website NCKU Moodle*

Related top conferences and journals

- Conference
  - Database: ICDE, SIGMOD, VLDB
  - Data Mining: SIGKDD, ICDM, SDM (SIAM), ASONAM
  - IR: SIGIR, CIKM, TREC
  - Web: WWW, WSDM, WI (Web Intelligence)
  - Machine Learning: AAAI, ICML
- Journal
  - IEEE TKDE (Impact factor = 2.236)
  - ACM TOIS, TOIT, TOW
  - VLDB Journal (Impact factor = 6.800)
  - Information Systems (Elsevier)
  - Knowledge and Information Systems (Springer)
  - Machine Learning, Journal of American Society of Information Sciences

Question ?