Tomáš Horváth

INTRODUCTION TO DATA MINING

Lecture 0

Introduction

Institute of Computer Science, Faculty od Science

Pavol Jozef Šafárik University in Košice

Slovak Republic

Why ML?

Mainly, because

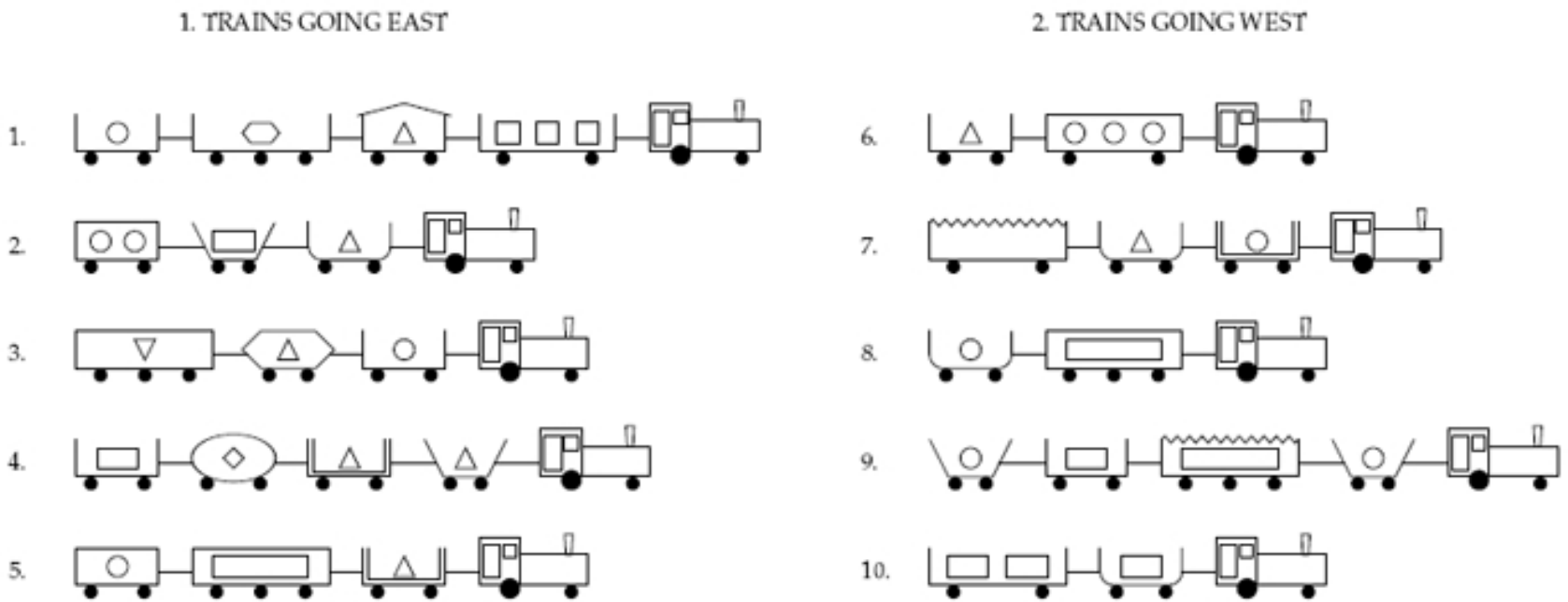
- there is **large data** outside
  - social web, industrial/financial transactions, e-commerce, ...
- **intelligent systems** are demanded
  - face/speech/handwriting recognition, weather forecasting, autonomous robots/agents, ...
- smart **decision support** is needed
  - recommendations, credit risk analysis, electric load control, ...
- the new trend is the **data-intensive scientific discovery**
  - T. Hey, S. Tansley and K. Tolle. The Fourth Paradigm: Data-Intensive Scientific Discovery. Microsoft Research, 2009.

and, also

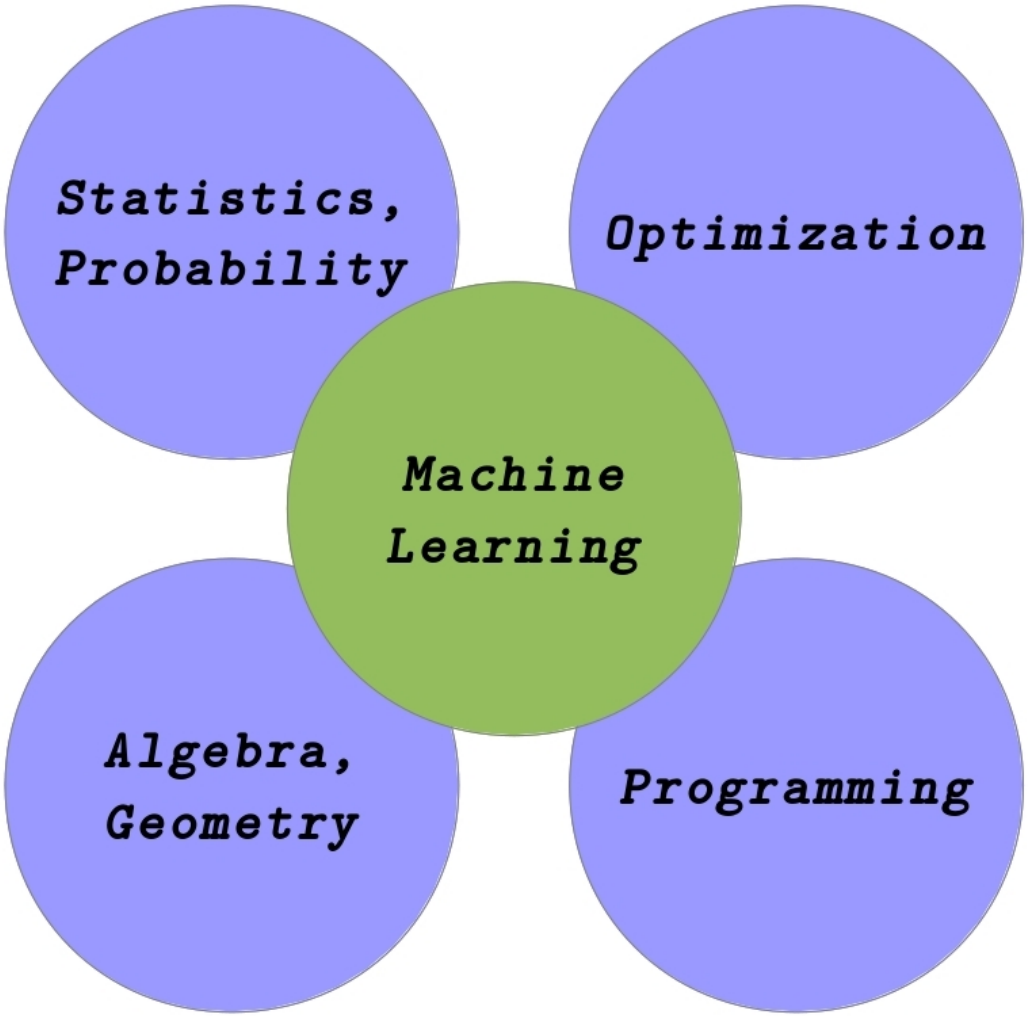
- after 100 years of the theory of relativity, the world needed some new topic for snobby discussions ;)

Need ML 'cause even in small data is hard to decide...

...for example, how trains going East differ from those going West?



What is ML?


Statistics,
Probability
Optimization
Machine
Learning
Algebra,
Geometry
Programming

Some Applications

Handwritten digit recognition

Image source: Subhransu Maji and Jitendra Malik: Fast and Accurate Digit Classification. Technical Report No. UCB/EECS-2009-159, Berkeley, 2009.

Some Applications

Spam filtering

Image source: Royce's spam collection, `http://xrl.us/rspam`

Some Applications

Robotics

Image source: `http://asimo.honda.com/`

Some Applications

fMRI (functional magnetic resonance imaging)

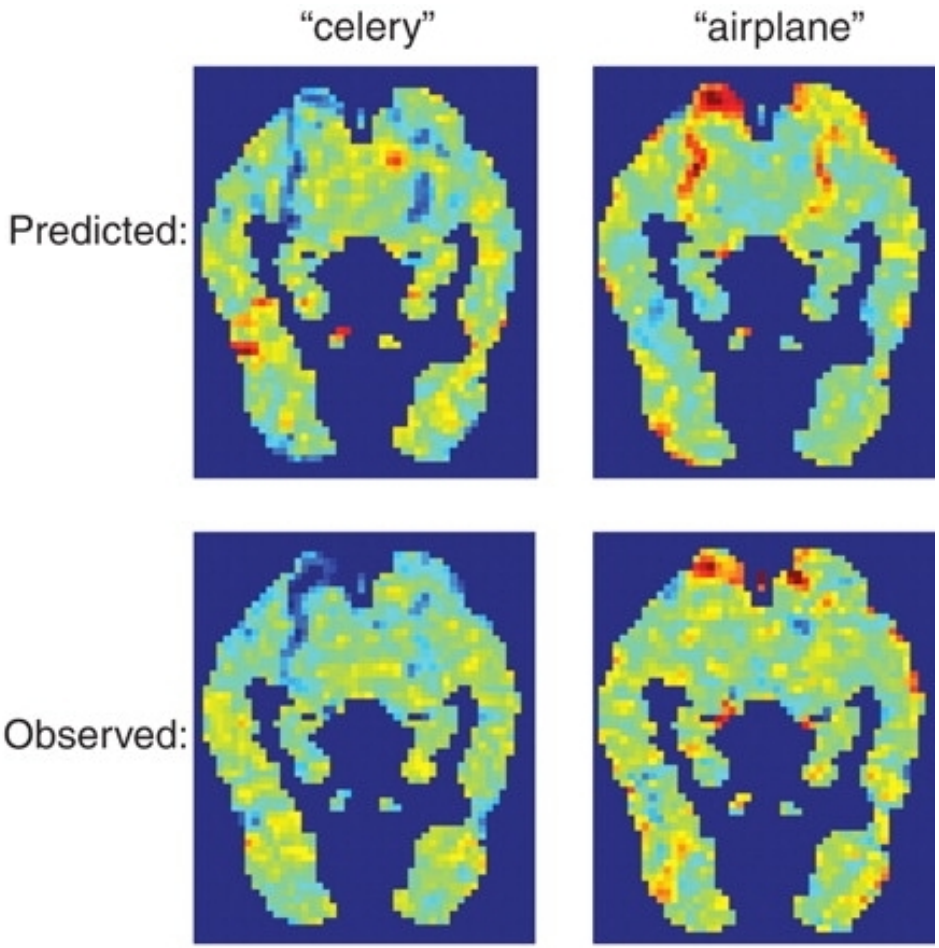
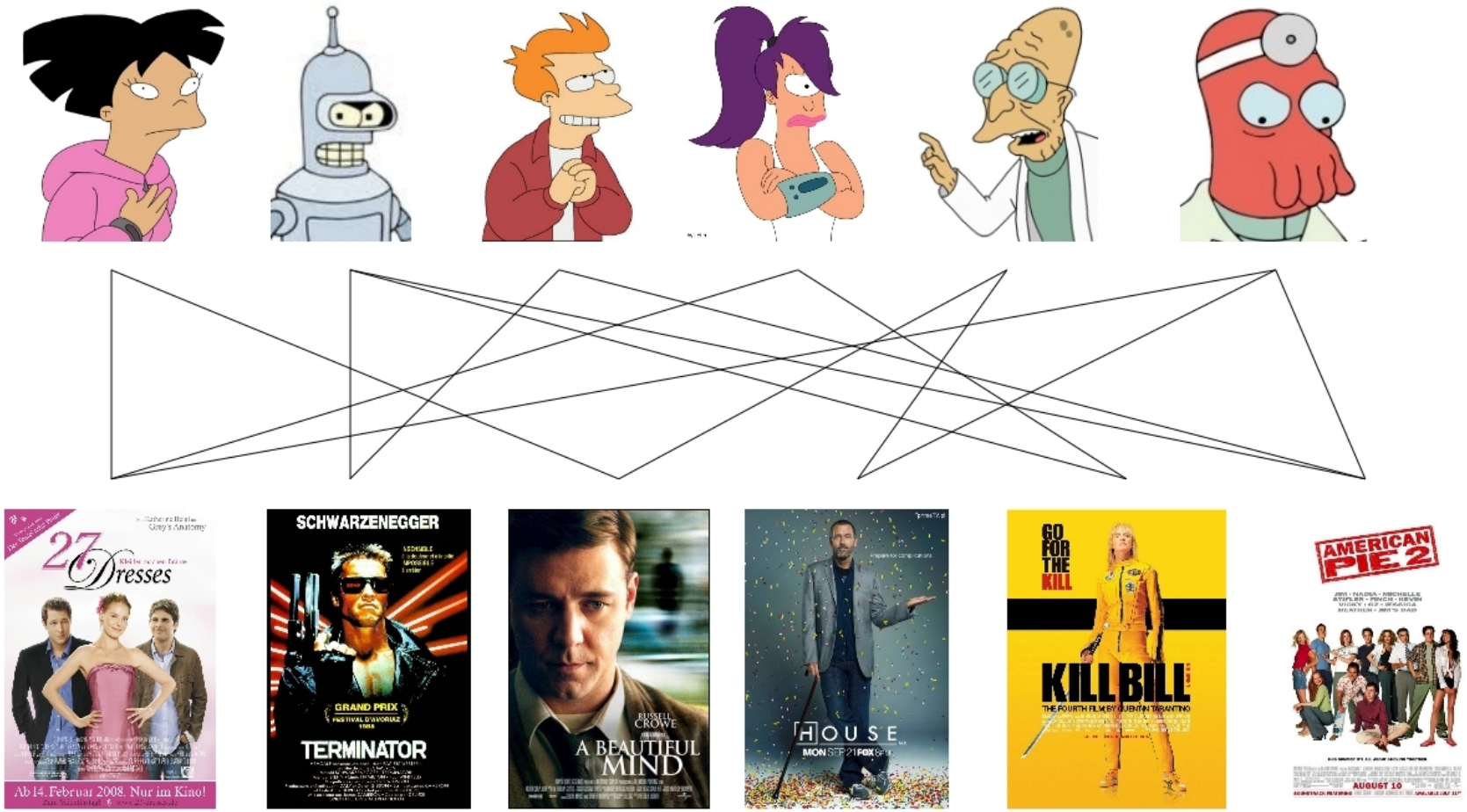


Image source: Tom M. Mitchell, et al. Predicting Human Brain Activity Associated with the Meanings of Nouns. Science 320, 1191 (2008).

Some Applications

Recommender systems



Some Applications

Fraud detection

Image source: http://bdemarest.wordpress.com/

Some (fuzzy) relations to other buzzwords

ML vs. Data Mining

- ML is a part of DM which covers also other steps as data pre-processing, data transformation, etc.
- still not unique opinions

DM vs. Knowledge Discovery in Databases

- used interchangeably, can be considered to be the same

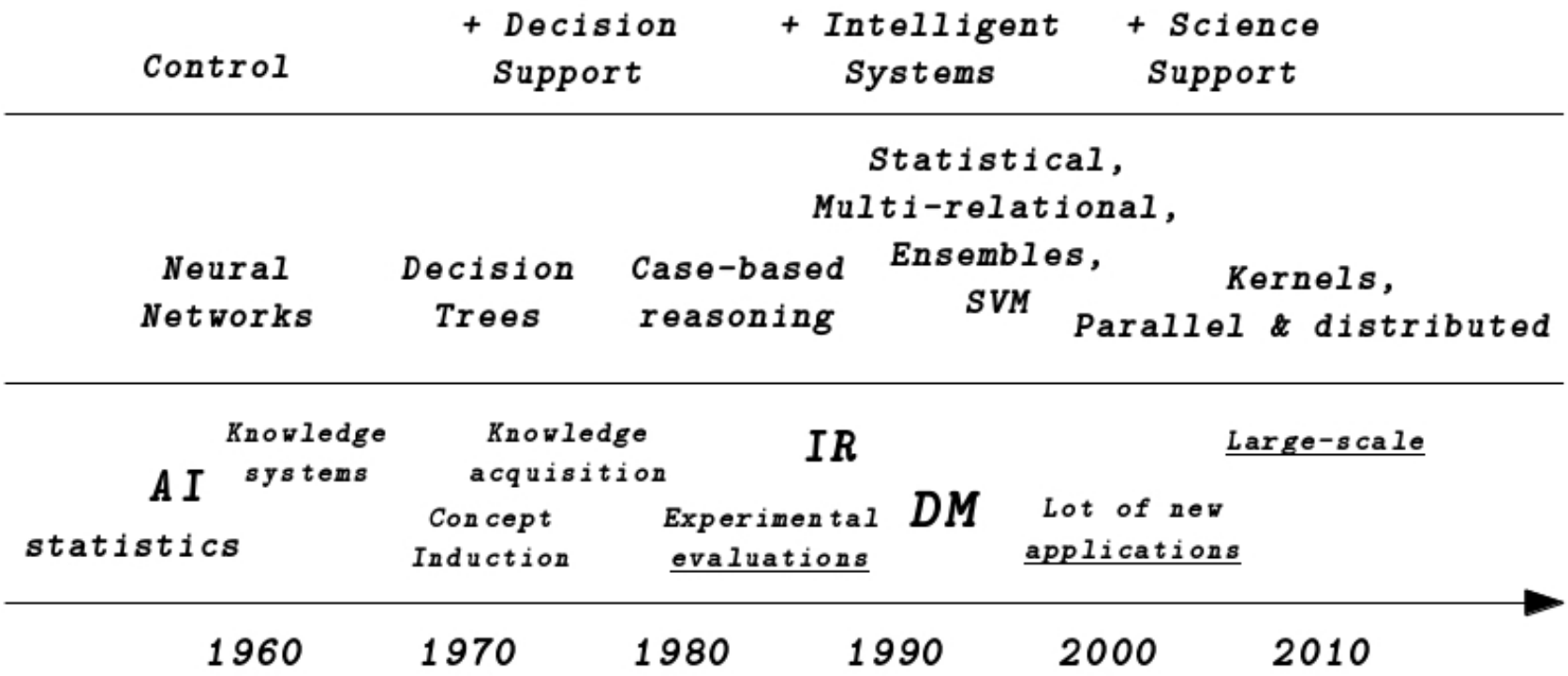
KDD vs. Business Analytics

- BA is concerned with decision support based on extensive use of data analysis, thus KDD is a methodology used in BA

BA vs. Business Intelligence

- BI is reporting what was happened, where the problem is, etc. while BA is trying to answer why is this happening, what will happen next, etc.

Brief History



The aim of this lecture

Understand the basic concepts of ML, DM to be able to deal with the following questions

- How to to prepare the data?
- What to be aware of during the DM process?
- How to choose a model and assess its qualities?
- ...

Free-diving instead of scuba-diving

- we won't focus on exact, algorithmic explanation of specific ML techniques nor deep theoretical description of different models

Admin

Lectures

- Monday, 8:55 – 10:25, P14

Tutorials (100 points in total)

- Monday, 10:30 – 12:00, P14

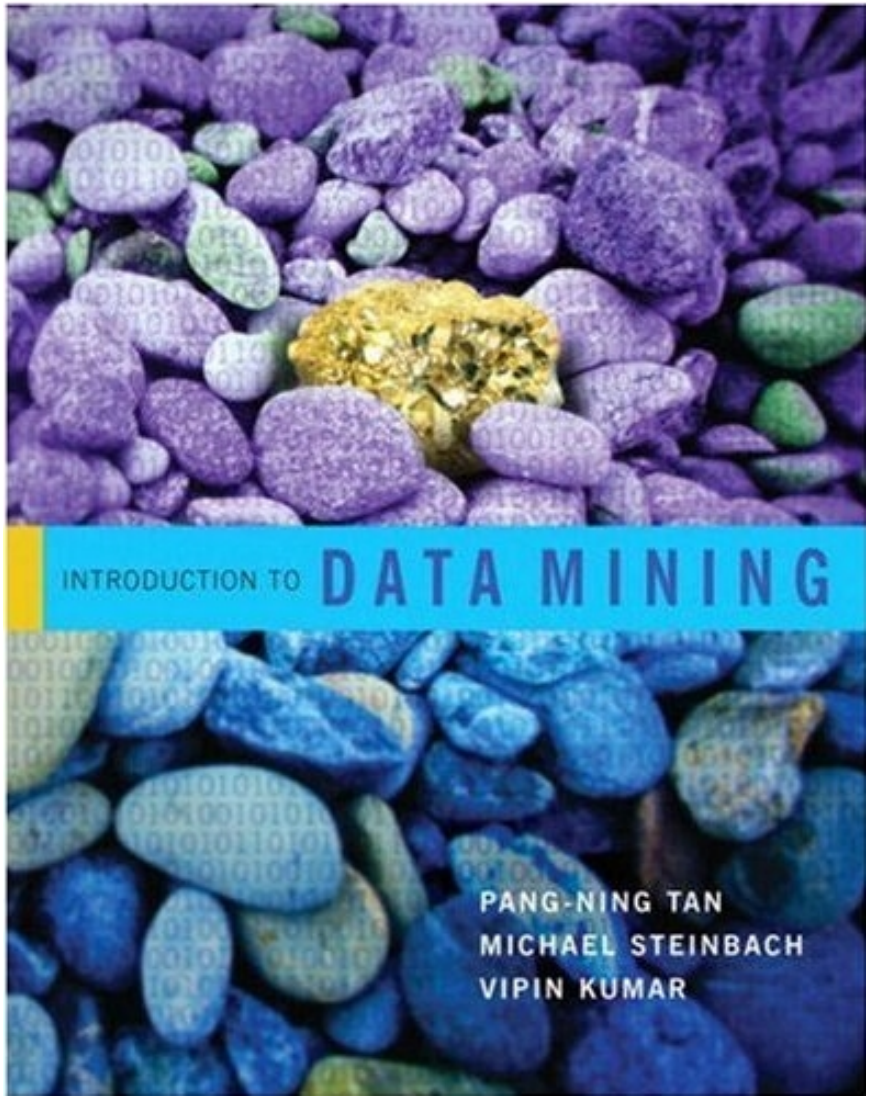
Examination (100 points in total)

- A **test** trying to uncover your hidden potential ;)

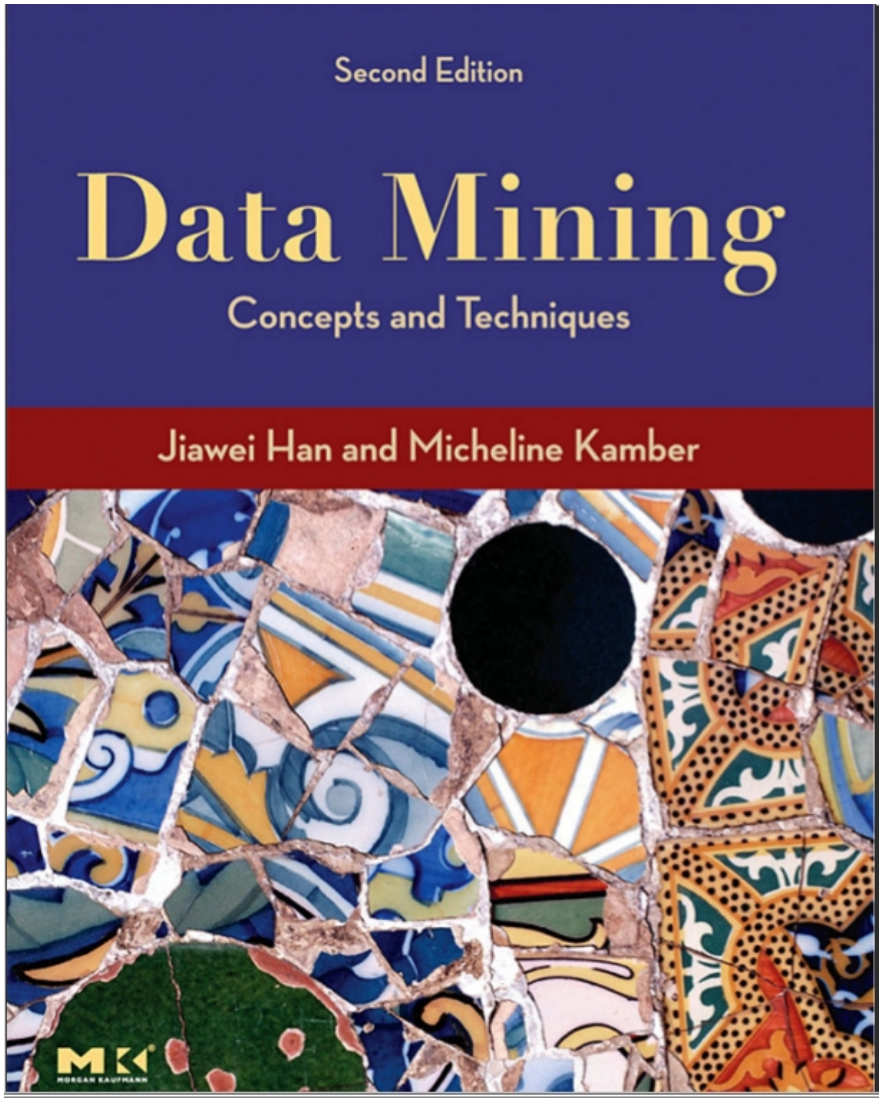
final points = 0.2 . points for tutorials + 0.8 . points for the exam

final points	*final grade*
91 – 100	A
81 – 90	B
71 – 80	C
61 – 70	D
51 – 60	E
0 – 50	F

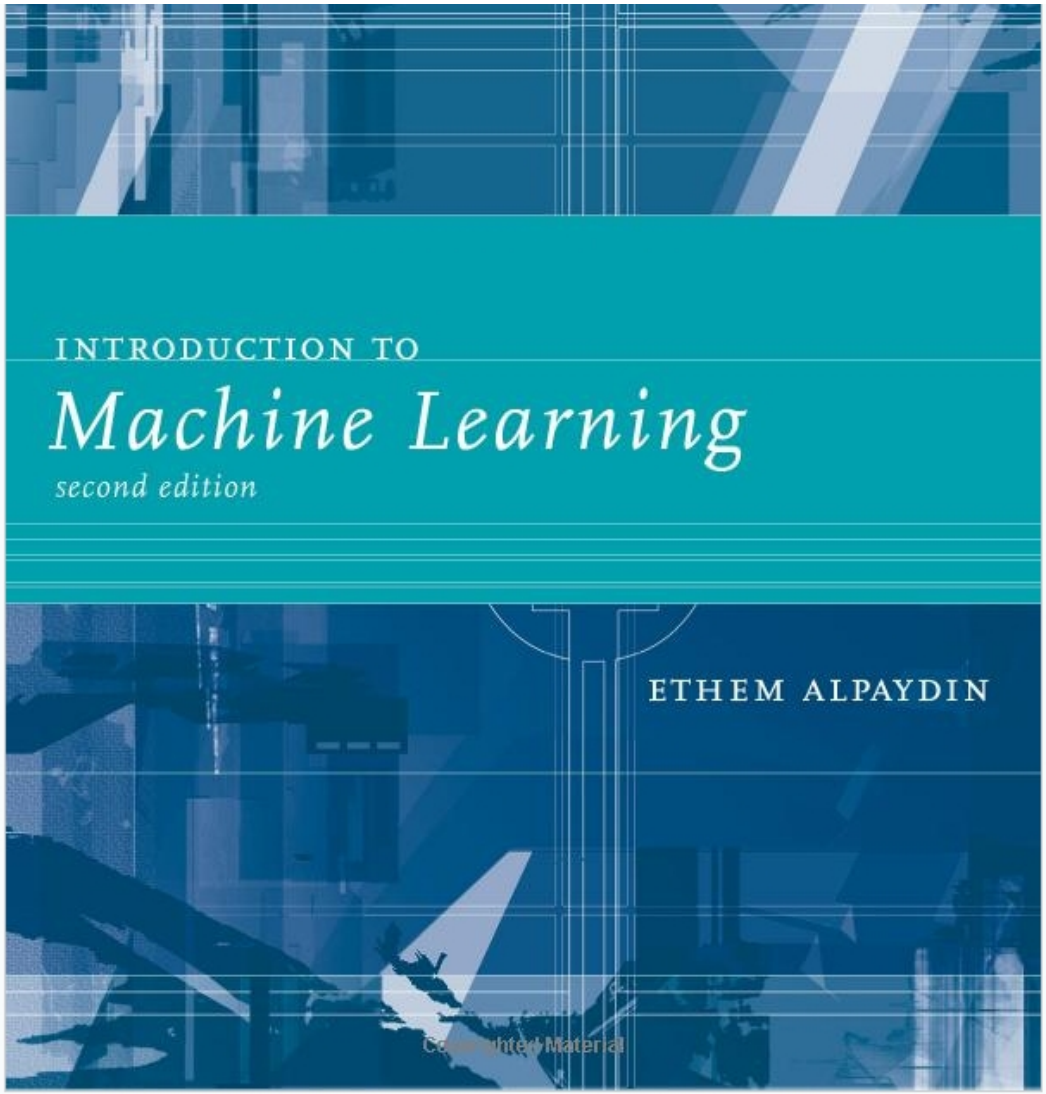
Textbook (1)


INTRODUCTION TO DATA MINING
PANG-NING TAN
MICHAEL STEINBACH
VIPIN KUMAR

Textbook (2)


Second Edition
Data Mining
Concepts and Techniques
Jiawei Han and Micheline Kamber
MK
MORGAN KAUFMANN

Textbook (3)


INTRODUCTION TO
Machine Learning
second edition
ETHEM ALPAYDIN
Copyrighted Material


That's all Folks!
Thanks for your attention

Questions?

Tomas.Horvath@upjs.sk

http://www.ics.upjs.sk/~horvath