Data Mining

IT University of Copenhagen, Spring 2016

Sebastian Risi

Today's menu

- Introduction to data mining (the field)
- Details about data mining (the course)
  - Structure
  - Teachers
  - Projects
  - Exam

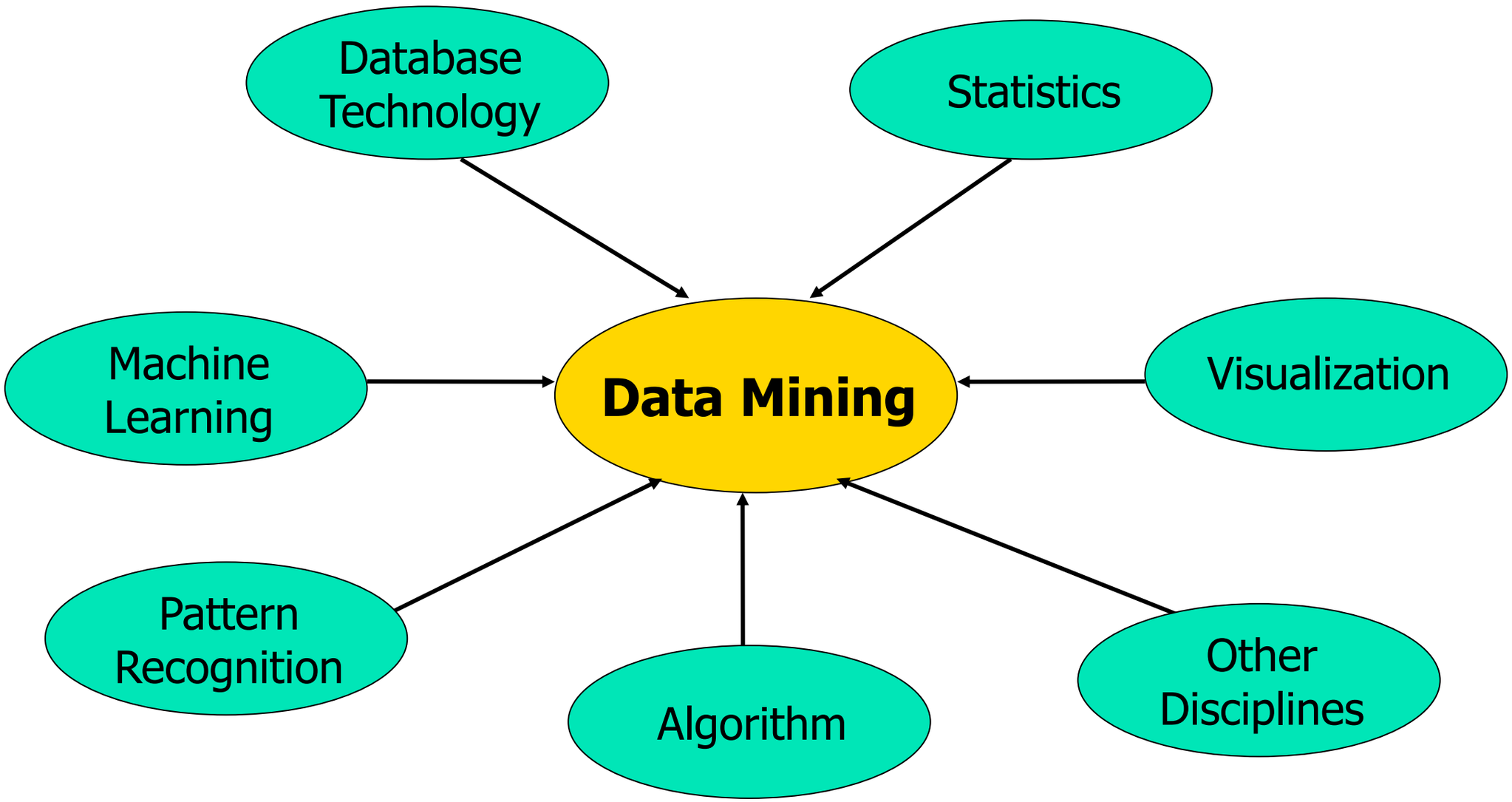
What is data mining?

Questions from you are most welcome during lectures

What is Data Mining?

- Extraction of interesting (non-trivial, implicit, previously unknown and potentially useful) patterns or knowledge from (huge amounts of) data
- Alternative name: knowledge discovery in databases (KDD)

What is Data Mining?



Main data mining topics

- Descriptive statistics: how do we efficiently summarize large amounts of data?
- Databases / data cubes: how do we store large amounts of data so that it can be promptly accessed by data mining algorithms?

Main data mining topics

- Prediction: how can we efficiently learn to predict an attribute from other attributes on unseen instances, based on a large data set?
- Classification: how can we efficiently create a model that classifies unseen instances into one of several categories, based on a large data set?

Main data mining topics

- Association mining: how can we efficiently find attributes that frequently co-occur?
- Clustering: how can we efficiently find clusters of instances
- Evaluation: how reliable and interesting are these patterns?

What is <u>not</u> data mining?

- Simple search
- Query processing
- Expert systems / deductive logics
- Reinforcement learning
- (machine learning on small data sets?)

The new “gold fever”

- Most companies want to profit from their data
  - Sell advertisements
  - Personalise services
  - Recommend apps/games/items/shows/dates
  - Debugging
  - And many more
- National security
  - Find threat to national security
- Science!
  - Find patterns in nature

> “Do you seek to engage in terrorist activities while in the United States or have you ever engaged in terrorist activities?”
>
> —Visa waiver questionnaire

Exercise

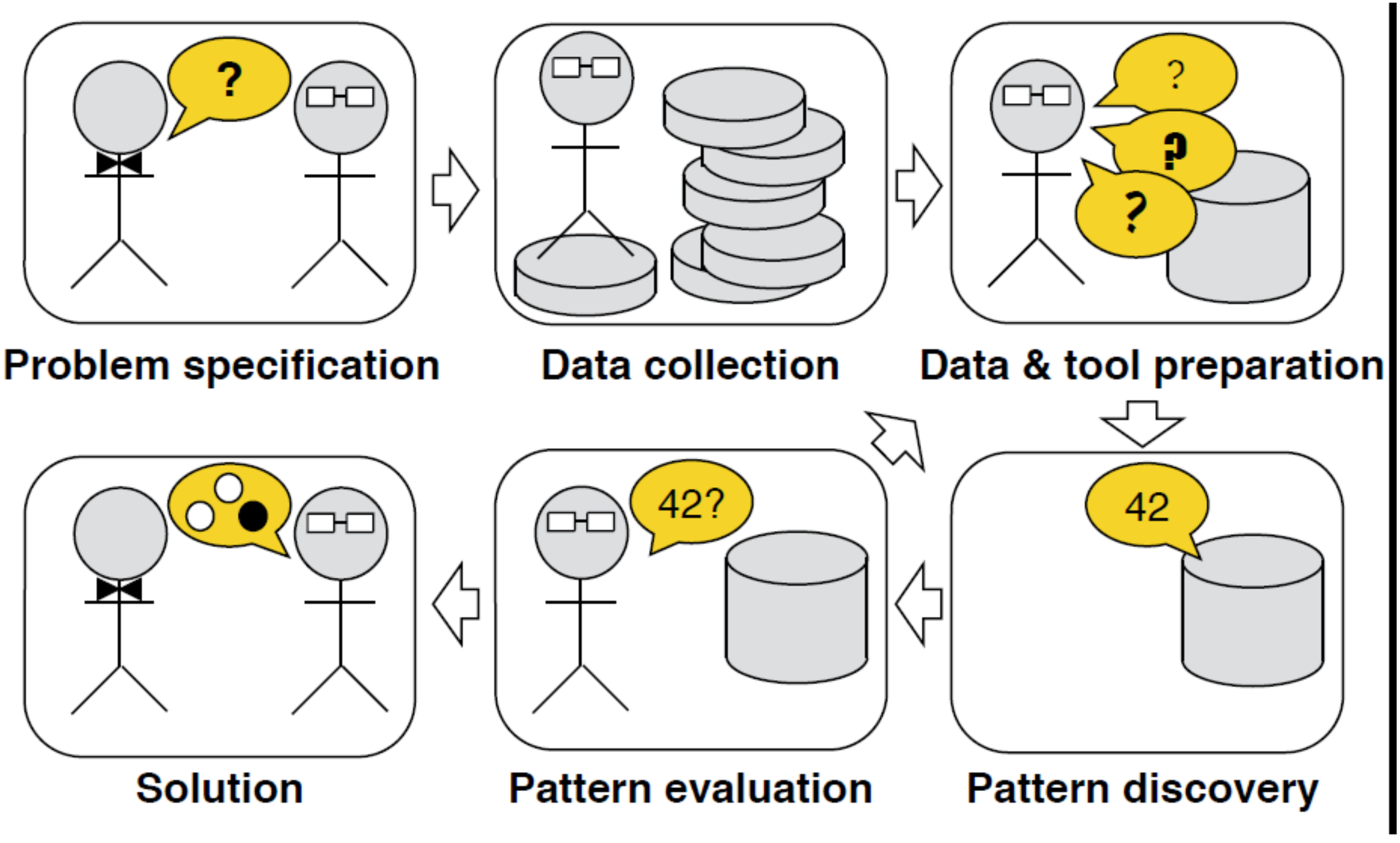
- 10 minutes to fill in questionnaire

LearnIT

or

http://goo.gl/forms/dGXTvxHbFl

Workflow


?
Problem specification
Data collection
?
?
Data & tool preparation
42
Pattern discovery
42?
Pattern evaluation
Solution

Is data mining good or evil?

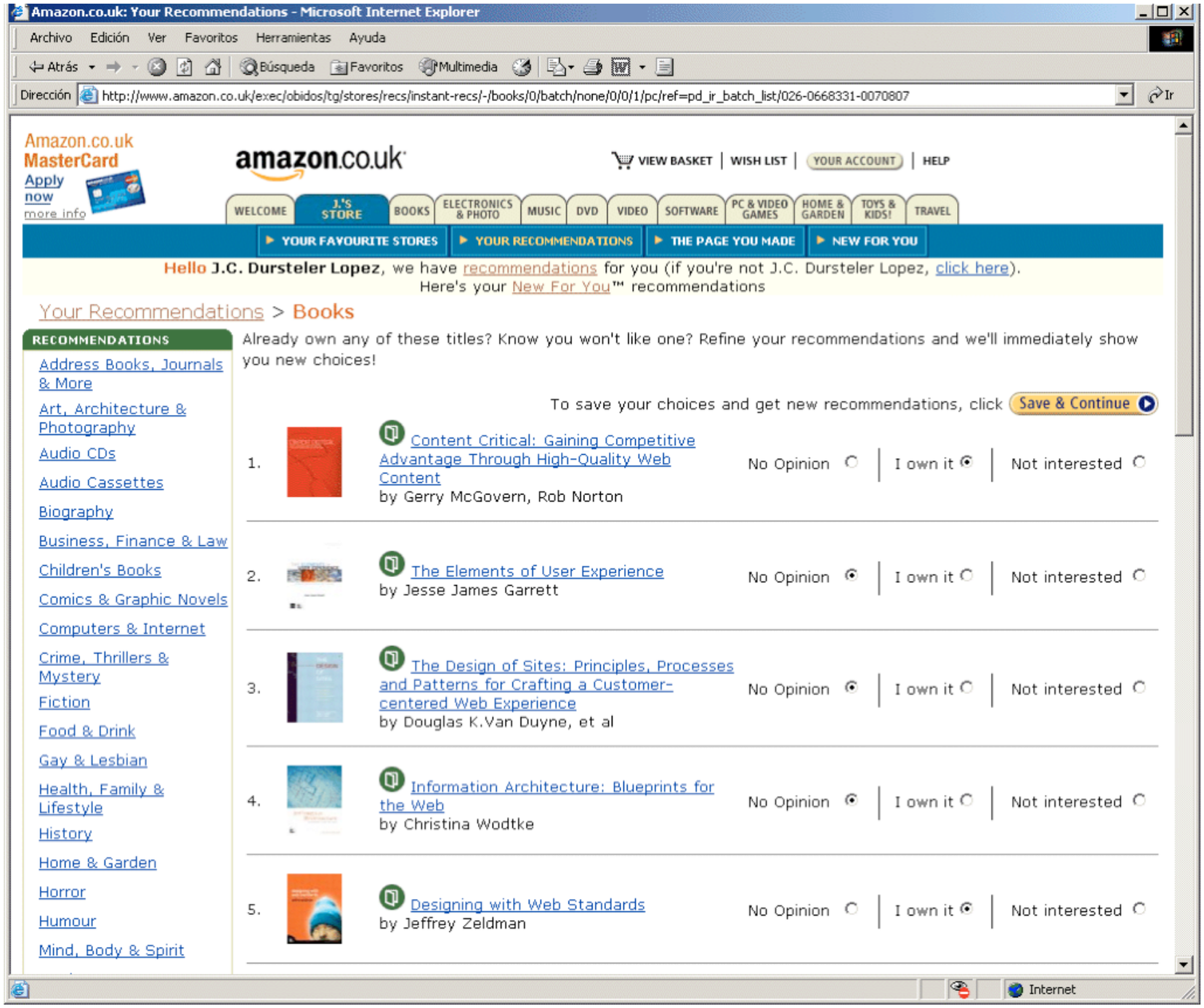
- Privacy
- Objectivity / sampling
- Is prediction understanding?

Example: customer credit rating

Example: customer credit rating

- Data: historical data about who defaulted and who did not
- Attributes might include monthly income, marital status, length of credit history and hundreds others
- Predict which customers will default if given a card
- Cluster into different groups

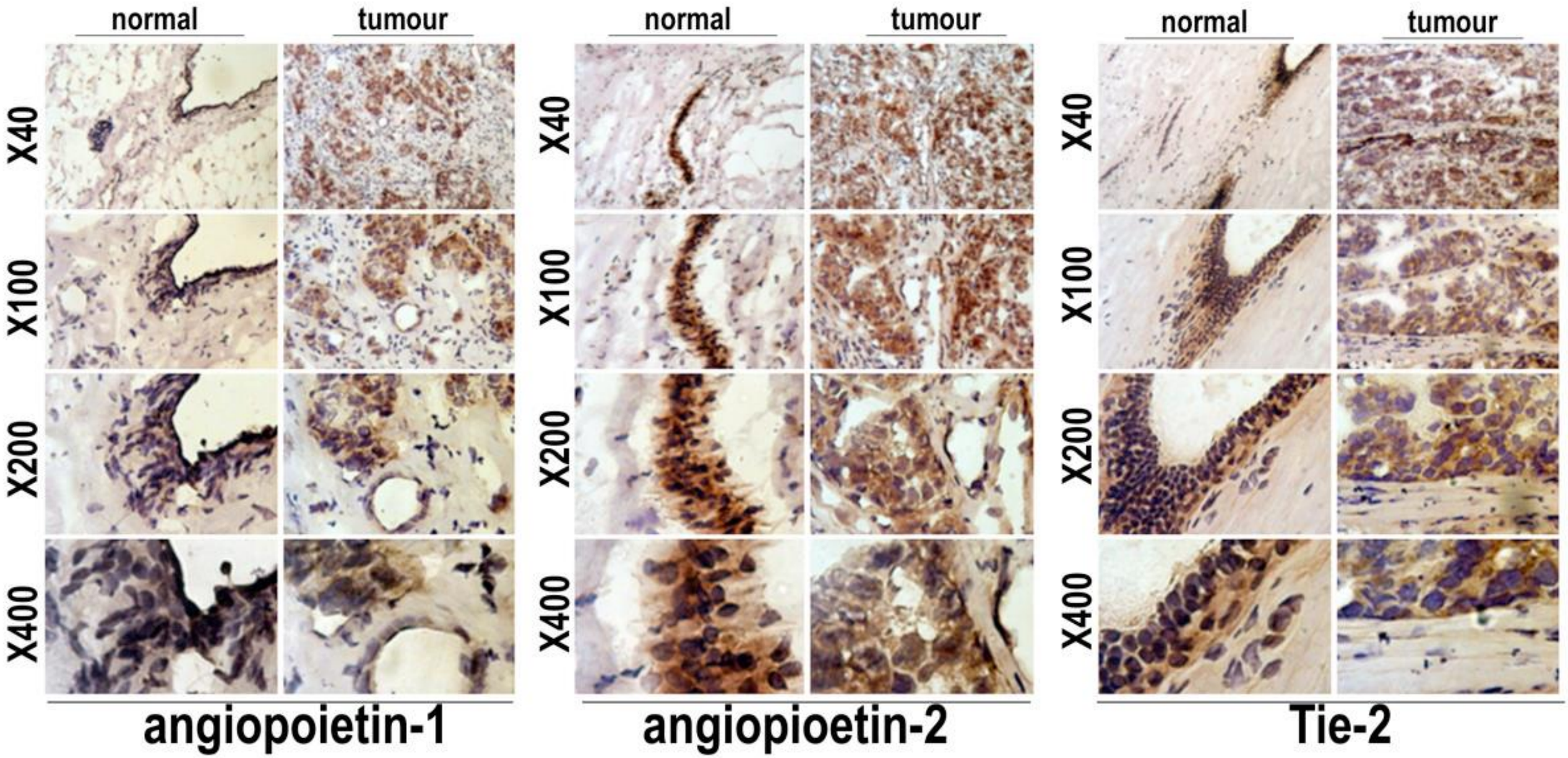
Example: product recommendation



Example: product recommendation

- Find products that are often bought together (associated)
- Cluster customers into relevant groups
- Predict which product recommendation will lead to the customer spending more money

Example: biological data analysis



Example: biological data analysis

- Data: lots of x-rays, or microscope images, or DNA samples labeled with type of cancer
- Classify new samples into the correct cancer type
- Including the reliability of the classification!

Example: computer game adaptation

Optimization of platform game levels for player experience

Chris Pedersen, Julian Togelius, Georgios Yannakakis
IT University of Copenhagen
Rued Langgaards Vej 7, DK-2300 Copenhagen S
Denmark
gammabyte@gmail.com, {juto, yannakakis}@itu.dk

Abstract

We demonstrate an approach to modelling the effects of certain parameters of platform game levels on the players' experience of the game. A version of Super Mario Bros has been adapted for generation of parameterized levels, and experiments are conducted over the web to collect data on the relationship between level design parameters and aspects of player experience. These relationships have been learned using preference learning of neural networks. The acquired models will form the basis for artificial evolution of game levels that elicit desired player emotions.

Introduction

Numerous theories exist regarding what makes computer games fun, as well as which aspects contribute to other types of player experience (Csikszentmihalyi 1990; Koster 2005). Recently, research in player satisfaction modelling has focused on empirically measuring the effects on player experience of changing various aspects of computer games, such as NPC playing styles (Yannakakis and Hallam 2007). Such studies have been conducted using both in-game data collection, questionnaires and physiological measurements (Yannakakis and Hallam 2008a).

Figure 1: Test-bed game screenshot.

of computational models of player experience derived from gameplay interaction which can be used as fitness functions for game content generation.

Test-bed Platform game

Example: computer game adaptation

- Create models of players (playing styles and emotions)
- Cluster player types
- Find commonly co-occurring player traits
- Predict which game modification will lead to higher enjoyment / frustration / retention etc.

Any more examples?

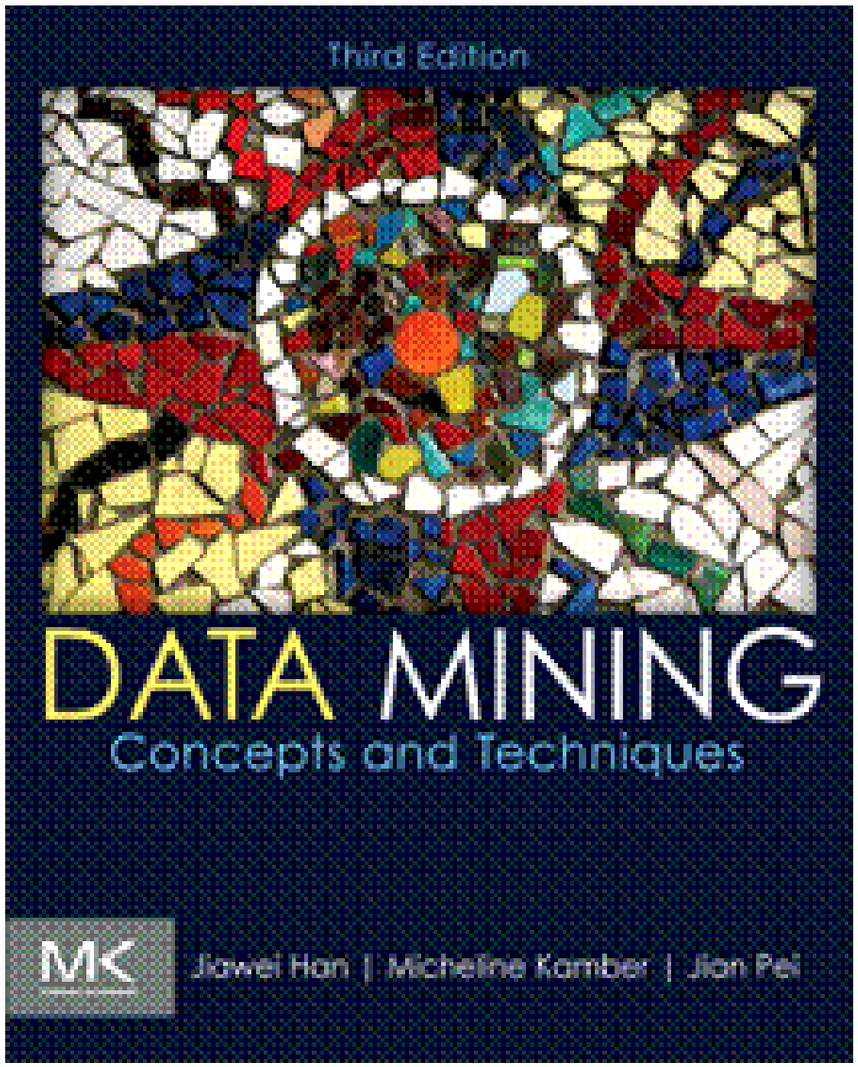
Who am I?

- Associate Professor at ITU
- From Germany
- Worked 2.5 years at ITU
- Do research in evolutionary algorithms, neural networks, neuroevolution, etc.

Who are the other guys?

- Anders Hartzen (lectures, TA)
- Kasra Tahmasebi Shahrebabak (TA)

Course book



- Han, Kamber and Pei: Data Mining: Concepts and Techniques, third edition
- http://www.cs.uiuc.edu/~hanj/bk3/
- Recommendation: get it!

Course book vs. course

- Much more comprehensive (includes more topics) than my lectures
- Lecture slides will be adapted from book web site
- The course will focus less on databases, business methods and buzz words, more on algorithms

Course philosophy

- Algorithm-focused
  - not database-focused
- You learn by doing:
  - exercises based on implementing key algorithms
  - group project during second half of term

Lectures: da rulez

- Interrupt me at any time
- Discussions are welcome
- You don’t have to be here
  - but I appreciate if you turn up, and show attention
  - people with extensive surfing habits please sit in the back
  - be on time

Course plan

- February: basic concepts and core algorithms
- March: advanced concepts and algorithms
- End of March: compulsory assignment hand-in (pass/fail)
- April and May: supervised group projects
- Oral examination in June. Based on group project, lectures and labs

The group project

- Organise yourselves into groups of 3 persons (not 2 or 1)
- Define and conduct your own data mining project using data and tools of your choice
- April 6th: Group Project Proposal Feedback
- Write a good report!

Examples of Past Projects

- Predict world cup winner
- Music genre classification
- Characterise behaviours of players in games
- Analyse groups of users on Steam
- **Anaylse ITUs Wifi Problems!**

The oral exam

- Determines your grade, together with the project report
- Based on both the group project and the lectures and the labs

Intended learning outcome

- After the course the students should be able to:
- - Analyse data mining problems and reason about the most appropriate methods to apply to a given dataset and knowledge extraction need.
- - Implement basic pre-processing, association mining, classification and clustering algorithms.
- - Apply and reflect on advanced pre-processing, association mining, classification and clustering algorithms.
- - Work efficiently in groups and evaluate the algorithms on real-world problems.

That’s it!

- Or is there anything else you want to know about?

- Readings: Chapter 2