

# Data Mining

## Introduction

IT University of Copenhagen, Spring 2018

Sebastian Risi

# Today's menu

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

# Who am I?

- Associate Professor at ITU
- From Germany
- Doing research in evolutionary algorithms, neural networks, neuroevolution, deep learning, etc.

# Who are the other guys?

- Paolo Burelli (lectures)
- Carolina Bermejo, Daniel Fritsdal Sørensen, Stefania Santagati (TAs)

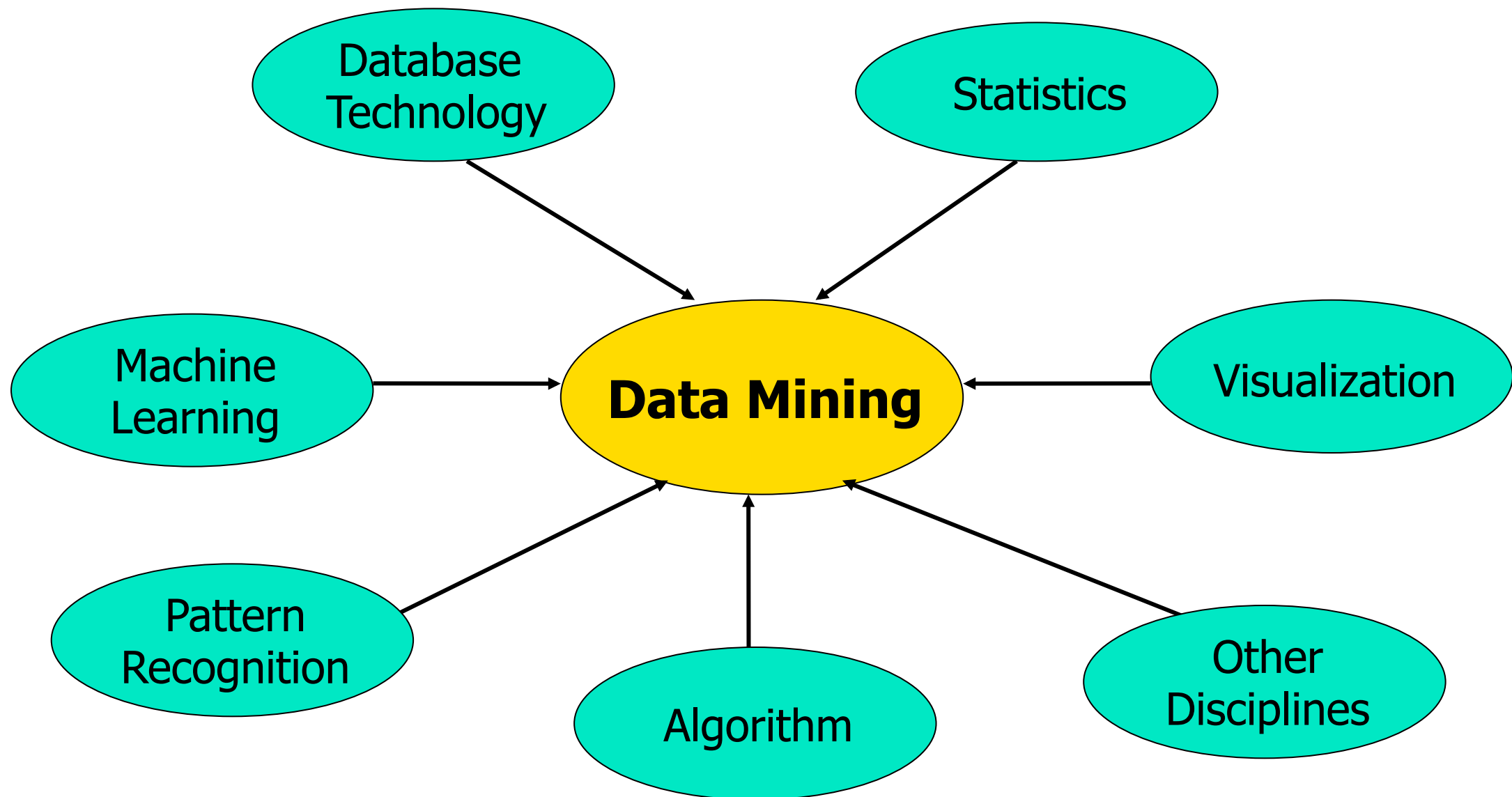
# What is data mining?

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

# 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 not data mining?

- Simple search
- Query processing
- Expert systems / deductive logics
- Reinforcement learning

# 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

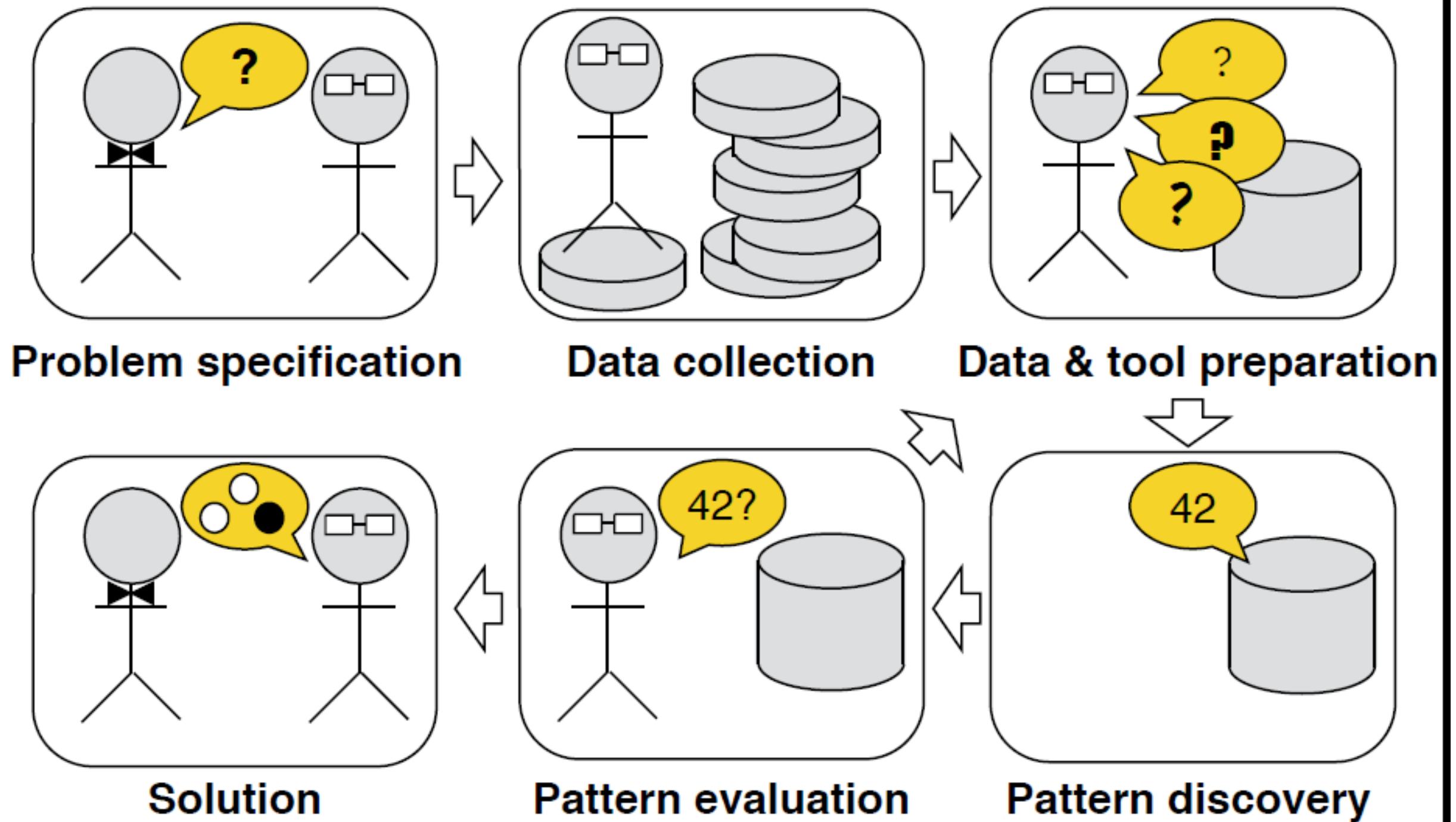
10 minutes to fill in questionnaire

Link on LearnIT

Or

<https://goo.gl/7DitTh>

# Workflow



# Databases

- Data “in the wild” not always well-suited to data mining...
- Incomplete, inaccurate, inconsistent, noisy
- Organised in strange ways making retrieval inefficient
- Relevant data needs to be selected

# Discovered patterns

- Not all are interesting
- Many could be redundant / correlated / overlapping
- Might be “artefacts” (generalise badly)



# 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 (i.e. wasn't able to pay back the credit) 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

Amazon.co.uk: Your Recommendations - Microsoft Internet Explorer

Archivo Edición Ver Favoritos Herramientas Ayuda

Dirección [http://www.amazon.co.uk/exec/obidos/tg/stores/recs/instant-recs/-/books/0/batch/none/0/0/1/pc/ref=pd\\_ir\\_batch\\_list/026-0668331-0070807](http://www.amazon.co.uk/exec/obidos/tg/stores/recs/instant-recs/-/books/0/batch/none/0/0/1/pc/ref=pd_ir_batch_list/026-0668331-0070807)

Amazon.co.uk MasterCard Apply now more info

amazon.co.uk

VIEW BASKET | WISH LIST | YOUR ACCOUNT | HELP

WELCOME J.'S STORE BOOKS ELECTRONICS & PHOTO MUSIC DVD VIDEO SOFTWARE PC & VIDEO GAMES HOME & GARDEN TOYS & KIDS! TRAVEL

YOUR FAVOURITE STORES YOUR RECOMMENDATIONS THE PAGE YOU MADE NEW FOR YOU

Hello J.C. Dursteler Lopez, we have [recommendations](#) for you (if you're not J.C. Dursteler Lopez, [click here](#)). Here's your [New For You](#)™ recommendations






[Your Recommendations](#) > [Books](#)

**RECOMMENDATIONS**

[Address Books, Journals & More](#)  
[Art, Architecture & Photography](#)  
[Audio CDs](#)  
[Audio Cassettes](#)  
[Biography](#)  
[Business, Finance & Law](#)  
[Children's Books](#)  
[Comics & Graphic Novels](#)  
[Computers & Internet](#)  
[Crime, Thrillers & Mystery](#)  
[Fiction](#)  
[Food & Drink](#)  
[Gay & Lesbian](#)  
[Health, Family & Lifestyle](#)  
[History](#)  
[Home & Garden](#)  
[Horror](#)  
[Humour](#)  
[Mind, Body & Spirit](#)

Already own any of these titles? Know you won't like one? Refine your recommendations and we'll immediately show you new choices!

To save your choices and get new recommendations, click [Save & Continue](#)

-  [Content Critical: Gaining Competitive Advantage Through High-Quality Web Content](#)  
by Gerry McGovern, Rob Norton  
No Opinion ☐ | I own it ☒ | Not interested ☐
-  [The Elements of User Experience](#)  
by Jesse James Garrett  
No Opinion ☒ | I own it ☐ | Not interested ☐
-  [The Design of Sites: Principles, Processes and Patterns for Crafting a Customer-centered Web Experience](#)  
by Douglas K. Van Duyne, et al  
No Opinion ☒ | I own it ☐ | Not interested ☐
-  [Information Architecture: Blueprints for the Web](#)  
by Christina Wodtke  
No Opinion ☒ | I own it ☐ | Not interested ☐
-  [Designing with Web Standards](#)  
by Jeffrey Zeldman  
No Opinion ☐ | I own it ☒ | Not interested ☐

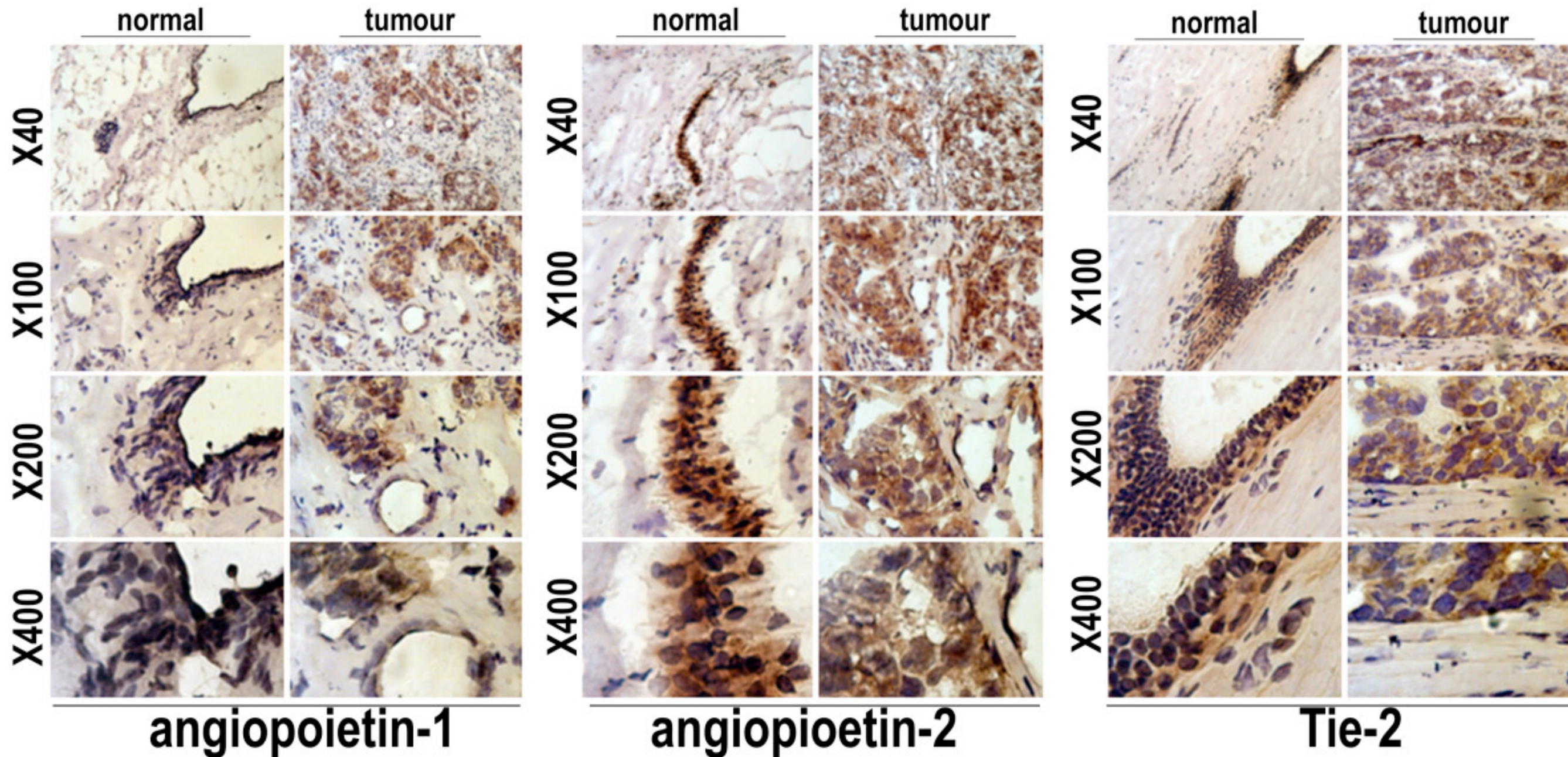
Internet

# 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**

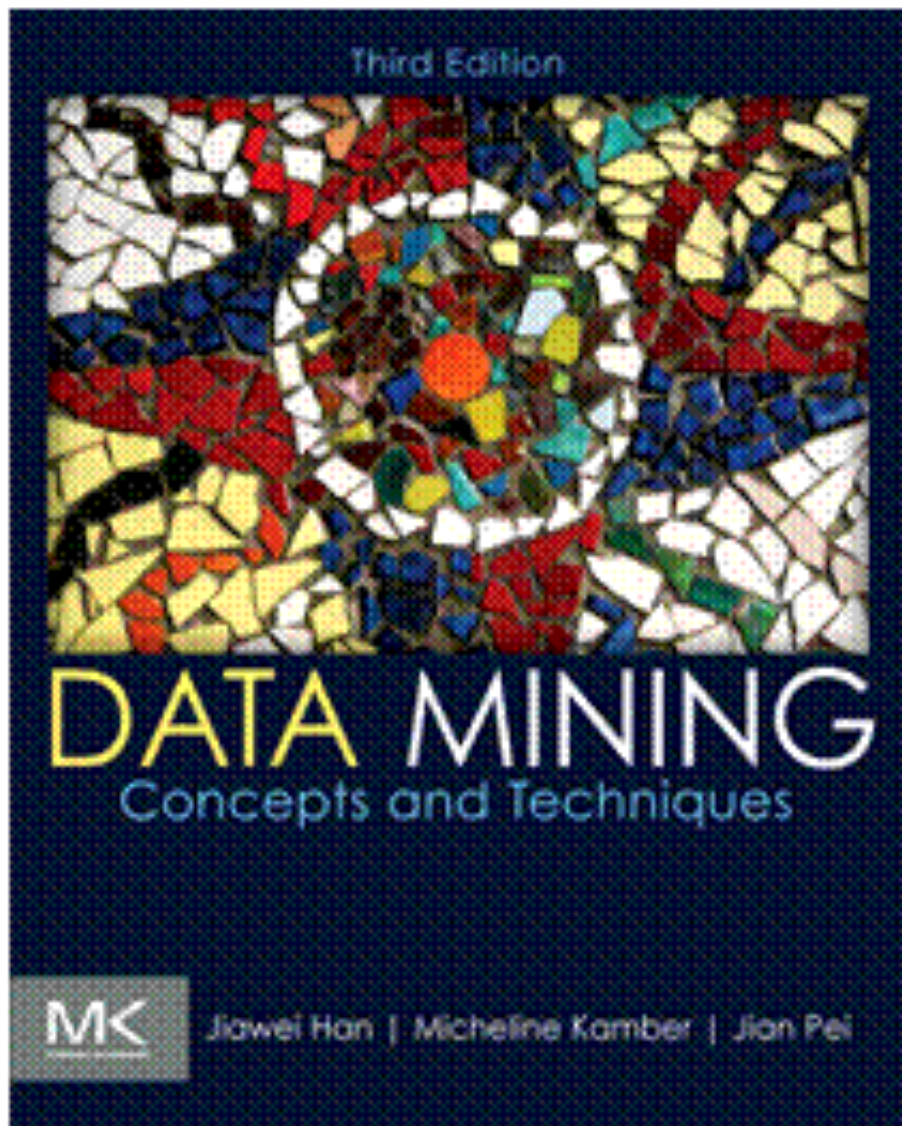


# Example: computer game adaptation

- 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?**

# 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

# “Inspirational reading”

- Ayres: Super Crunchers
- Levitt and Dubner: Freakonomics and Superfreakonomics

# 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
- April 2nd: compulsory assignment hand-in (pass/fail)
- April and May: supervised group projects
- Group Project Hand-in: May 19th
- Oral examination June 11-14 . Based on group project, lectures, readings, and labs



# Lecture plan (tentative)

- Lecture 1 - Introduction
- Lecture 2 - Preprocessing
- Lecture 3 - Pattern and association mining 1
- Lecture 4 - Classification and prediction 1
- Lecture 5 - Clustering 1
- Lecture 6 - Pattern and association mining 2
- Lecture 7 - Classification and prediction 2
- Lecture 8 - Clustering 2
- Lecture 9 – Deep Learning

# The individual assignment

- Based on the lab exercises
- Hand-in: April 2nd
- Mandatory

# Individual Assignment

## (see LearnIT for details)

- You must apply at least two different preprocessing methods, one sequential/frequent pattern mining method, one clustering method and one supervised learning method. All code for these algorithms must be written by yourself. Source code will only be accepted if it's reasonably well commented.
- What to hand in: A two-page report, in which you describe what questions you tried to answer, which methods you used, the results you reached, any problems you encountered along the way, etc.. Please include numerical results and preferably also graphs.
- The assignment is a natural continuation of the work done during the labs. You will be able to reuse the code you wrote during the labs, and even some data set analysis you performed. If you completed all the labs, you will have most of the work on the assignment. Therefore, use the lab sessions effectively, and don't be afraid of asking the TAs and teacher questions!

# The group project

- Organise yourselves into groups of 3-4 persons (not 2 or 1)
- Define and conduct your own data mining project using data and tools of your choice
- May 26th: Group Project Proposal Feedback
- Write a good report!

# Examples Projects

- Predict world cup winner
  - Music genre classification
  - Characterise behaviours of players in games
  - Analyse groups of users on Steam
  - Analyse ITUs Wifi Problems!
- ➔ Can lead to a published paper! Be creative!

# 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 and 3