

From hypothesis testing to hypothesis formation

Data mining in virtual worlds

Antti Ukkonen
Antti.Ukkonen@tkk.fi

About me

- I'm a computer scientist.
 - M. Sc., TKK (2004)
 - coming up: D.Tech., TKK
(defense on June 4th, in two days!)
- I've worked at:
TKK, Nokia, CERN, Yahoo! Research

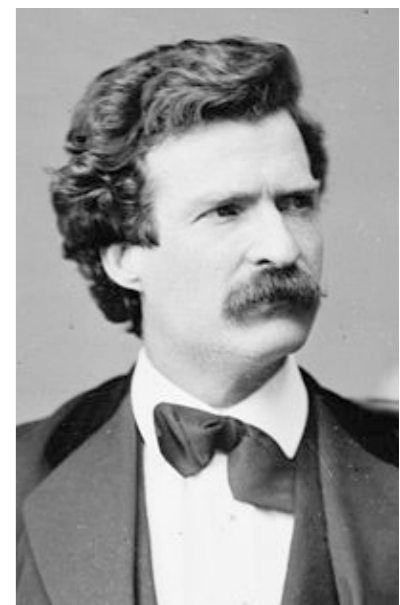
About the talk

- Part I: Data mining vs. statistics
- Part II: Examples of data mining
(with applications to virtual economies)

Part I
Data mining
vs.
statistics

Statistics

- *“There are three kinds of lies: lies, damned lies, and statistics.”*
- Benjamin Disraeli
- *“Facts are stubborn things, but statistics are more pliable.”*
- Mark Twain



Statistics

- Uses data to argue that a proposed theory is unlikely to be false.
- Aims at hypothesis *testing*.

“Diamond salesman”

- A stranger approaches you with a shiny stone and offers to sell it for 1000 euros.
- Null-hypothesis:
The stone is worthless.
- Alternative hypothesis:
The stone is a diamond.
- This is statistics.



“Diamond mine”

- Suppose you arrive at a mountain with a team of miners.
- You tell the miners to find all shiny stones that weigh more than 20g and are larger than 3mm.
- Results may vary.
- This is “data mining”.



Data mining

- *“Data mining is the analysis of (often large) observational data sets to find unsuspected relationships and to summarize the data in novel ways that are both understandable and useful to the data owner.”*

(Hand, Mannila, Smyth. Principles of Data Mining. MIT Press. 2001.)

Data mining

- *“Data mining is the analysis of (often large) observational data sets to find **unsuspected relationships** and to summarize the data in novel ways that are both understandable and useful to the data owner.”*

(Hand, Mannila, Smyth. Principles of Data Mining. MIT Press. 2001.)

Unsuspected relationships?

- E.g. “patterns“ in market basket data:
 - Customers who buy X and Y (beer and sausages) also tend to buy Z (mustard).
 - A customer who bought certain books might also be interested in these other books.

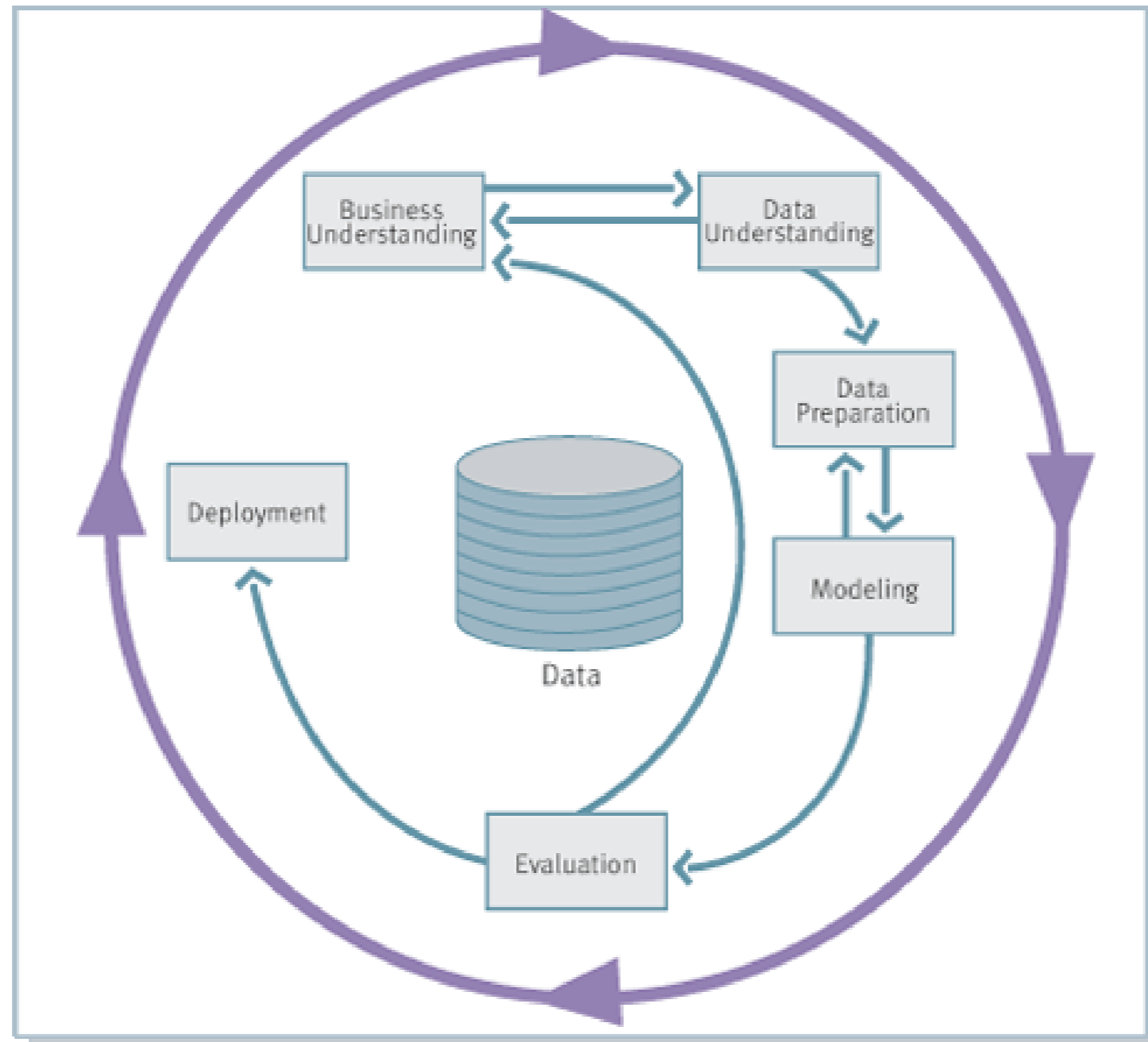
Patterns

- Depends on your data.
- Examples:
 - Sets of products that tend to occur together
 - Pairs of commodities whose prices behave together in a certain way
 - Sequences of avatar actions that are usually associated with a certain type of use

Patterns vs. queries

- Patterns are similar to database queries:
 - DB Query: *“Get the records of all Finnish customers who have been logged in for longer than 1000 hours in total.”*
 - Pattern query: *“Get all combinations of commodities that rapidly become very common among the users.”*

Data mining process



<http://www.crisp-dm.org/Process/index.htm>

Some differences

Statistics

tools available

*data gathered for
a particular purpose*

hypothesis verification

Data mining

*tends to require
custom-made software*

*data gathered
“just for fun”*

hypothesis formation

Part II: Data mining in virtual worlds

Virtual worlds/economies

- Interesting from a data miners perspective:
 - There's lots(!) of data.
 - Possibility to combine different types of data.
 - The application area is novel.

“Mispriced” commodities

- Why do people buy certain products?
- Query: “Get all pairs of commodities that are substitutes but have a price difference larger than d .”
 - (Called a statistical arbitrage in the securities business.)
- What factors lead to the price difference?

Finding substitutes

- Pepsi vs. Coke
- 2 shopping baskets are similar if they contain similar products.
- 2 products are similar if they are contained in similar shopping baskets!
- Query: “Get all pairs of products that behave similarly with respect to products in the set S .”

Innovation mining

- How fast do innovations spread in a virtual economy?
- What is an innovation?
- How to measure rate of adoption?

Innovation?

- In Habbo Hotel the players come up with novel uses for some items.
- In EVE online it may be advantageous to fit the spaceship with an appropriate combination of equipment.
- We define:
An innovation is a combination of items.

Frequent itemsets

- Classical data mining problem:
 - Given the contents of all shopping baskets purchased this week, find all *frequent itemsets*, i.e., combinations of items that appear in more than $x\%$ of the baskets.
- Replace shopping baskets with inventories of players to find innovations.

Emerging innovations

- An innovation is *emerging*, if the combination of items is about to become frequent.
- How fast does the innovation propagate among the players?
- How do innovations die out?

Macro miners

- “Thanks” to RMT it is possible to make real money with MMORPGs.
- Serious players tend to (more than!) frown upon such activities.
- How to identify players that use client-side scripting to automate game-play tasks?

Conclusion

- Data mining
 - Discovering patterns
 - (What to do with the patterns...?)
 - Probably requires collaborating with a data mining specialist
 - Keep an open mind!

Thank you!