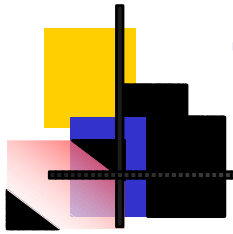
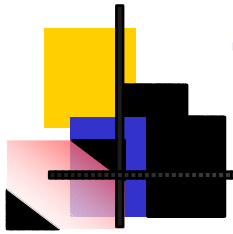


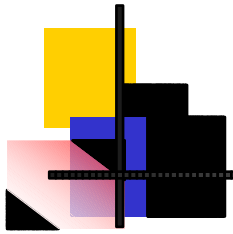
Today's Program

- Logistics, advice and course overview
- Background on data mining
- Data mining challenges
- Data mining tasks
- Data mining vs. machine learning
- The data mining process



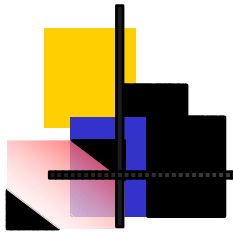
Today's Program

- Logistics, advice and course overview
- Background on data mining
- Data mining challenges
- Data mining tasks
- Data mining vs. machine learning
- The data mining process



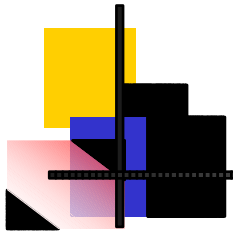
Course Style

- Provide a broad survey of several important and well-know subfields
- "Hands on" experience, interactive lectures/discussions
- Develop an overall sense of how to extract information from data in a systematic way
  - The How: Gain insight into the working of specific algorithms
  - The Why: Understand the “big picture” of data mining



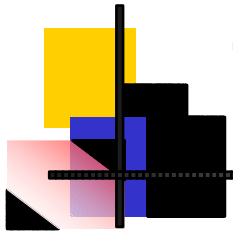
Course Goals

- Understand the challenges faced in data min
- Understand what a data mining algorithm should do
- Understand how current systems work
  - Algorithmically
  - Empirically
  - Their shortcomings
- Think about how we could improve algorithm



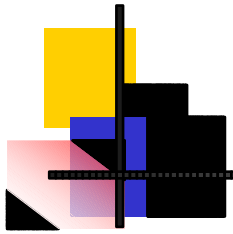
How To Do Well In This Class

- Attend all lectures and exercise sessions
- Actively participate in class activities
  - Ask questions
  - Think critically about course material
- Make sure you can apply learned concepts in different settings
- If you have a question, ask!



Today's Program

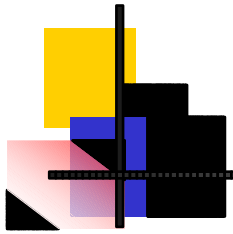
- Logistics, advice and course overview
- Background on data mining
- Data mining challenges
- Data mining tasks
- Data mining vs. machine learning
- The data mining process



What Is Data Mining?

Many definitions are possible:

- Phrase to put on CV to get hired
- Non-trivial extraction of implicit, previously unknown and useful information from data
- Buzzword used to get money from funding agencies and venture capital firms
- (Semi-)automated exploration and analysis o large dataset to discover meaningful patterns

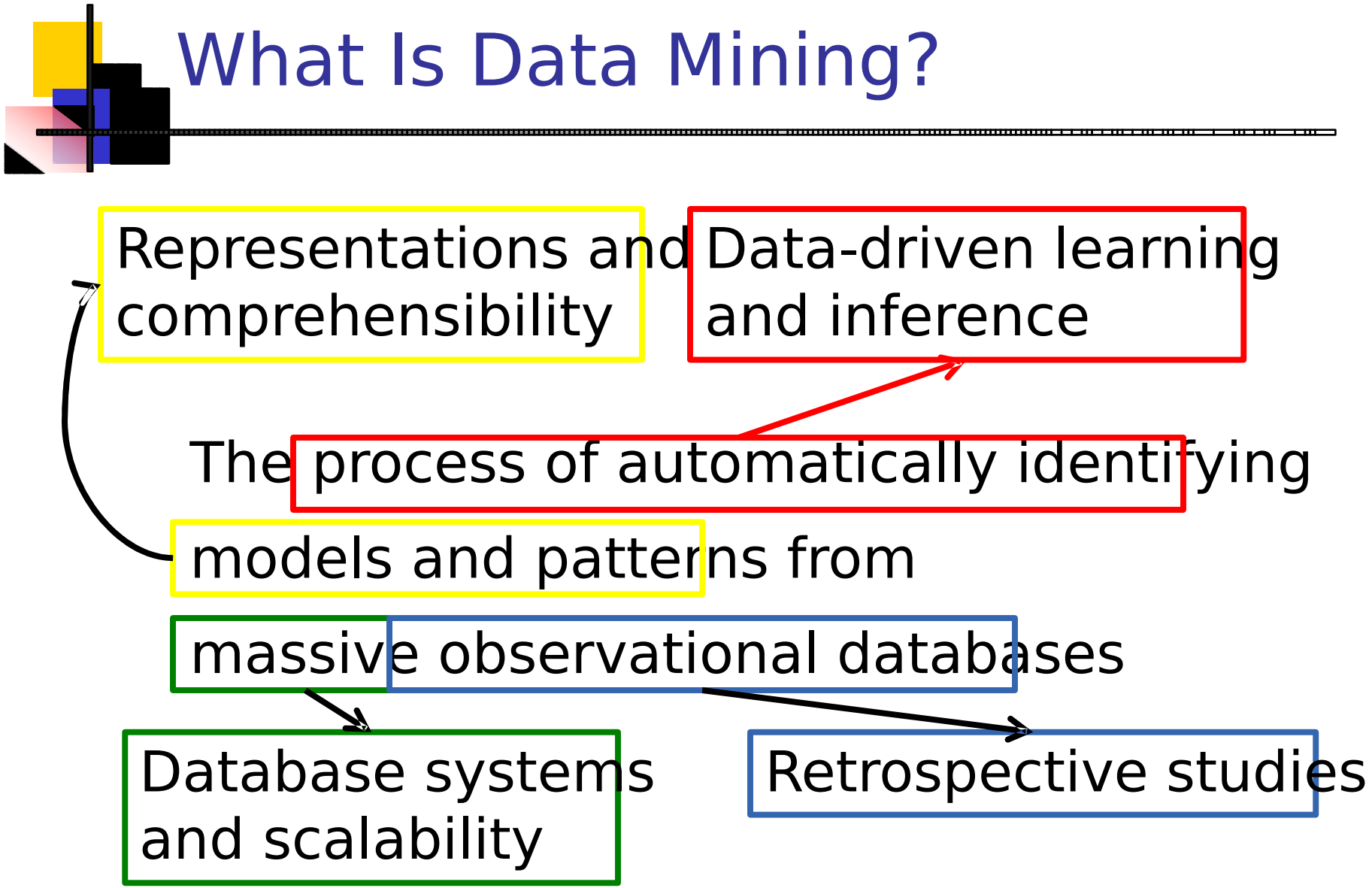


What Is Data Mining?

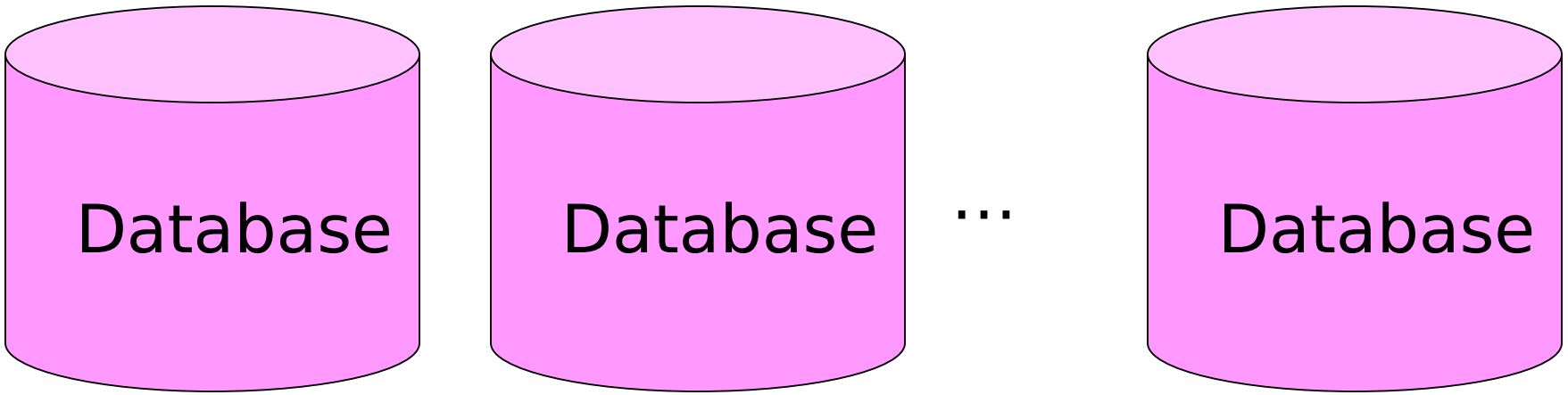
The process of automatically identifying models and patterns from massive observatio databases that are

- Valid: hold on new data with some certainty
- Novel: non-obvious to the system
- Useful: should be possible to act on the item
- Understandable: humans should be able to interpret the pattern

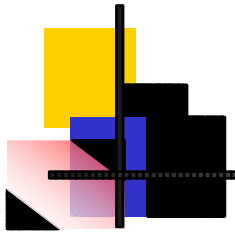
What Is Data Mining?



Simply Stated:
Three Goals of Data Mining

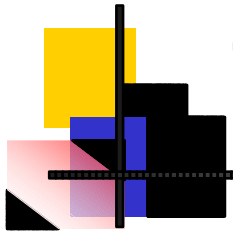


1) Understand the data
2) Extract knowledge from the data
3) Make predictions about the future



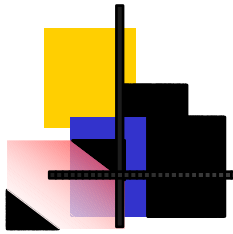
Why Is Data Popular Now?

- 25 years ago basically no data mining
- Now, it is hugely popular and successful
  - Frequently in popular press
  - Used in companies
  - Taught in academia
- Two main reasons
  - **Possible:** Technology has greatly improve
  - **Needed:** Databases and the Web means everyone has data



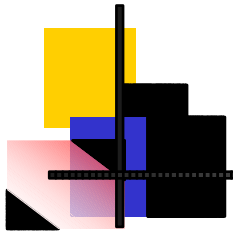
Technological Advances

- Storage is larger and cheaper
  - Moore’s law for magnetic disk density: “capacity doubles every 18 months”
  - Storage cost per byte falling rapidly
- Improvements in computing power
  - Super computer of 15 years ago is equivalently powerful as modern desktops
  - Cloud computing
- Improvements in machine learning algorithm



Many Large Datasets

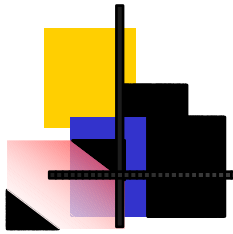
- Online text sources
  - MEDLINE has 19 million published articles
  - Wikipedia has huge number of articles
- Web search engines
  - Multiple billion Web pages indexed
  - 100’s of millions of site visitors per day
- Retail transaction data
  - Ebay, Amazon, Walmart: >100 million transactions per day
  - Visa, Mastercard: similar or larger numbers



Motivation For Data Mining

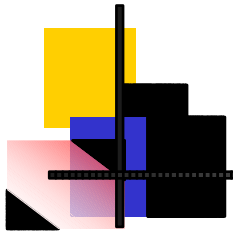
- We have lots and lots of data
- There is often information “hidden” in the dat that is not readily evident
  - Human analysts take weeks to discover useful information
  - Much of the data is never analyzed at all

- “We’re drowning in information, but starving knowledge.” (John Naisbett)



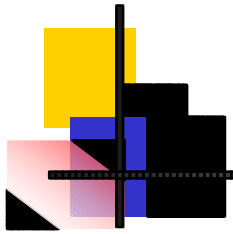
Scientifically Useful

- Data collected and stored at GB/hour
  - Remote sensors on a satellite
  - Microarrays generating gene expression da
  - Scientific simulations
- Traditional techniques infeasible for raw data
- Data mining helps scientists to
  - Classifying and segmenting data
  - Form hypotheses
  - Find hidden patterns and correlations



Commercially Useful

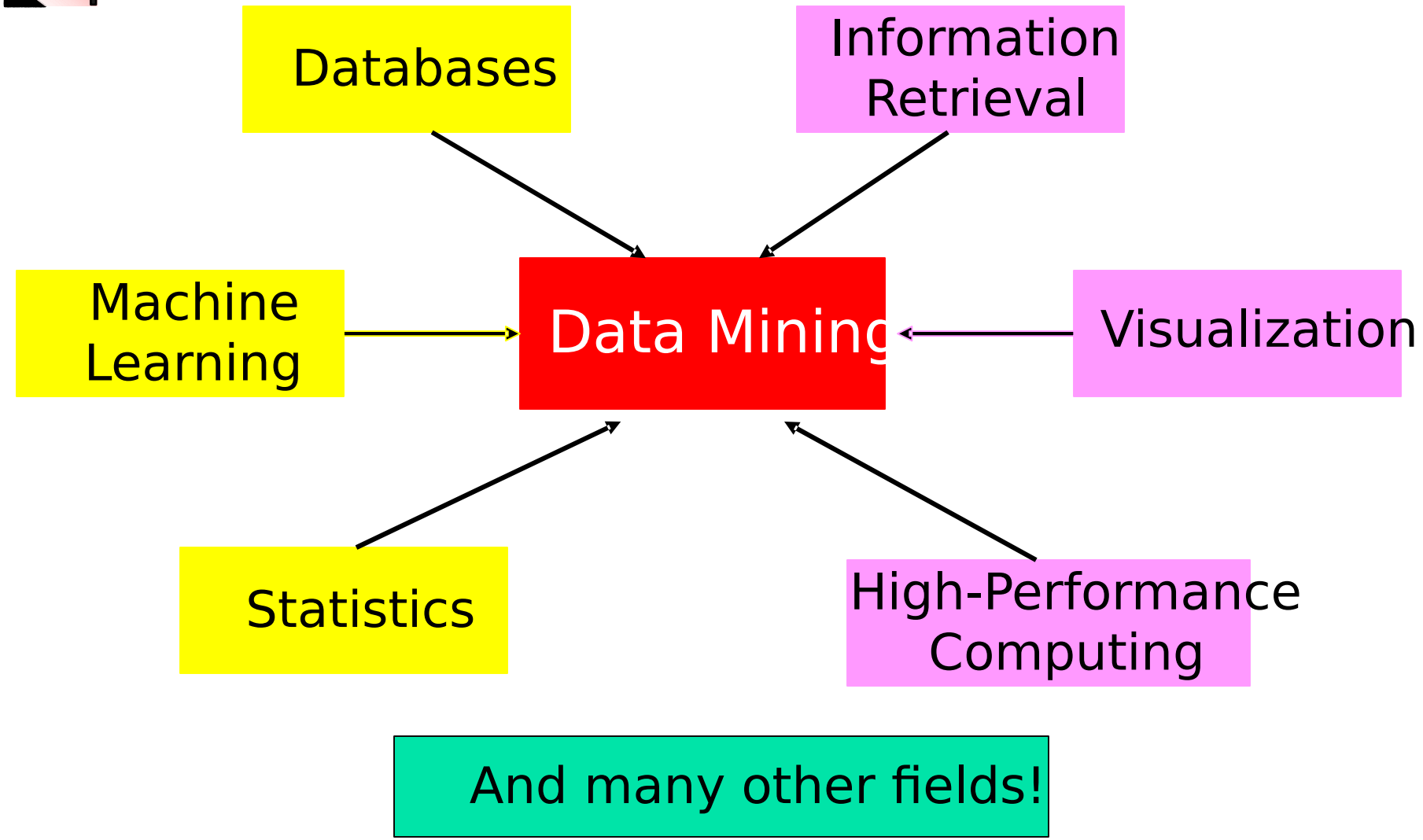
- Many companies collect and store data
  - **Search-engines:** click data
  - **Stores:** purchases records
  - **Banks:** credit card transactions
  - Many many more
- Computers are cheap and powerful
- Strong competition

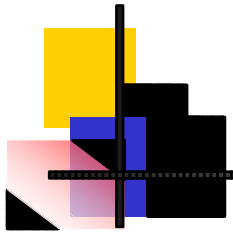


Commercially Useful

- Data mining can help provide better, customized services
  - Better search results
  - Target advertising
  - Viral marketing
  - Manage inventory
  - Many more

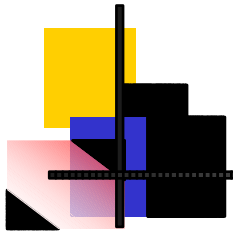
Data Mining Draws From Many Disciplines





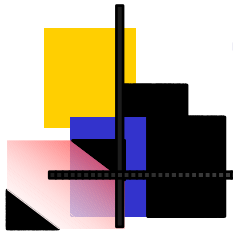
Data Mining vs. Statistics

- Traditional statistics
  - Hypothesize, then collect data, then analyz
  - Often model-oriented
- Data mining is different
  - Usually no hypothesis
  - Focus on data driven analysis of existing data
  - Algorithms vs. models
- Ideas from statistical are very useful in data mining, particularly in evaluation



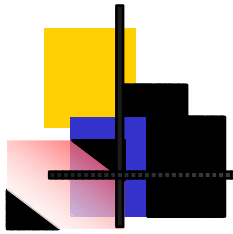
Data Mining vs. Machine Learning

- High-level view: fields are very similar
- Data mining focuses more on
  - Scalability, i.e., data resides in relational D
  - Applications
  - Term used more in industrial setting
- Machine learning
  - More theoretical emphasis
  - Term more used in research/academia



Today's Program

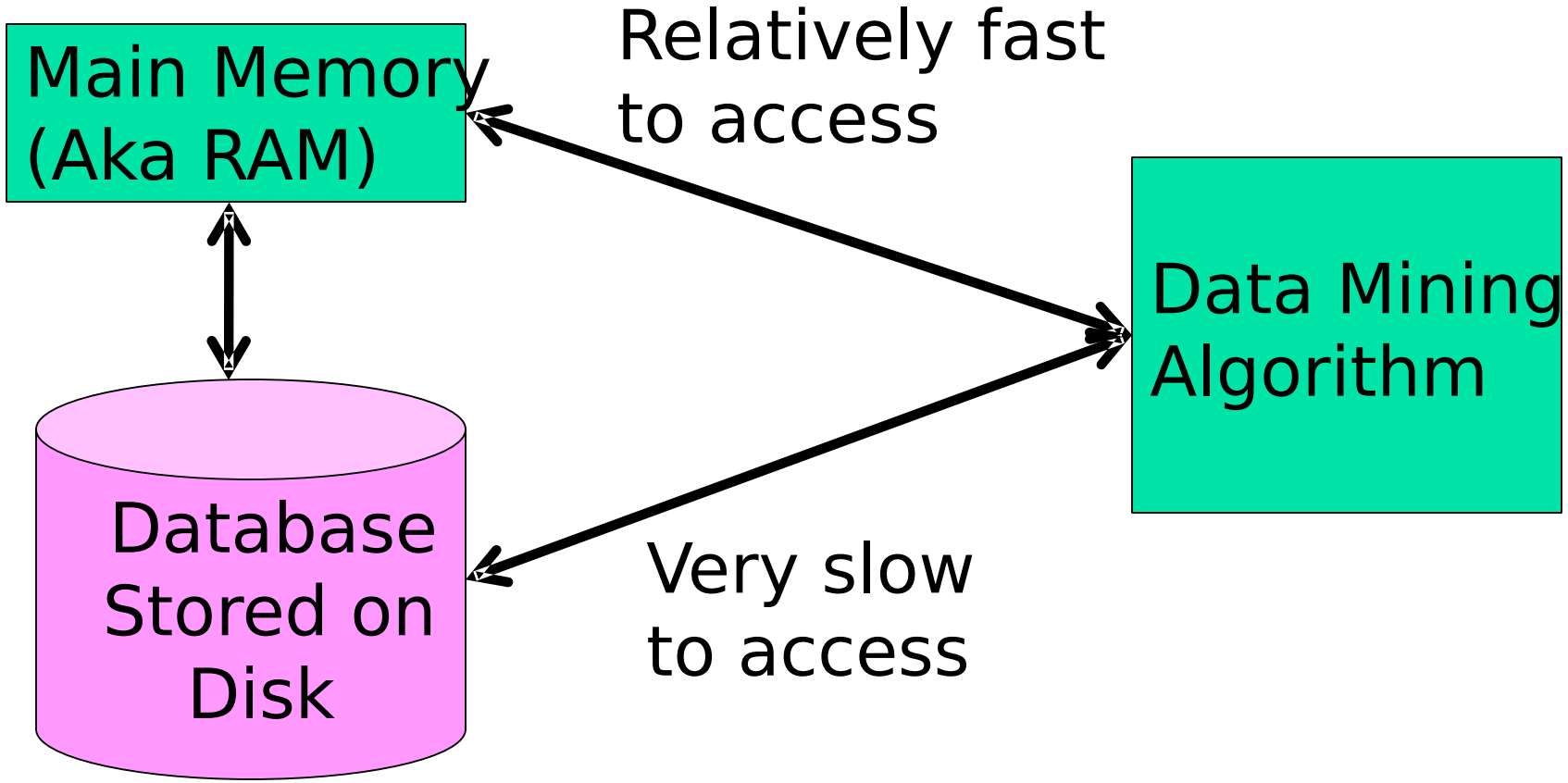
- Logistics, advice and course overview
- Background on data mining
- Data mining challenges
- Data mining tasks
- Data mining vs. machine learning
- The data mining process

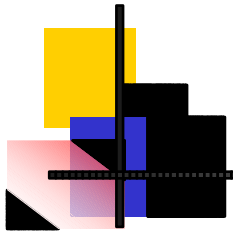


Data Mining Challenges

- Scalability
- Dimensionality
- Retrospective data
- Complex andheterogeneous data
- Data quality
- Data ownership anddistribution
- Privacy preservation
- Streaming Data

Scalability



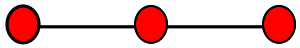


Curse of Dimensionality

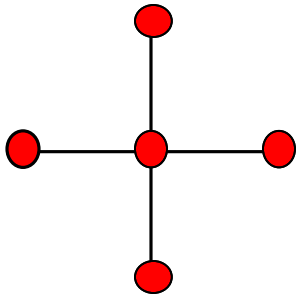
- Imagine instances are described by 1000 attributes, but only two are relevant to the concept
- Curse of dimensionality
  - With lots of features, can end up with spurious correlations
  - Nearest neighbors are easily mislead in hig dim
  - Easy problems in low-dim are hard in high-
  - Low-dim intuition doesn’t apply in high-dim

Example: Points on Hypergrid

- In 1-D space: 2 NN are equidistant



- In 2-D space: 4 NN are equidistant



Spurious Correlations in Data

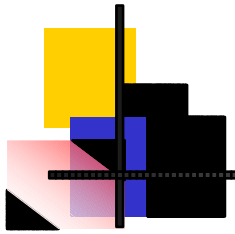
- A big data mining risk is that you will “discove patterns that are meaningless
- Bonferroni’s principle: (roughly) if you look in more places for interesting patterns than you amount of data will support, you are bound t find crap
- Another way: if more variables than example some variables will be correlated by chance

Retrospective Data

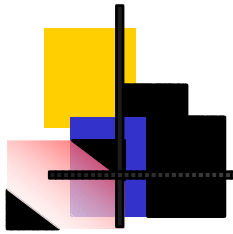
- Generally speaking, two types of data
  - Experimental data
  - Observational data

- What type of data you have influences the conclusions you can draw from it

Traditional Scientific Experimental Design

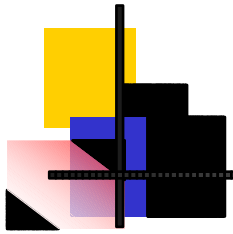


- Traditional approach:
  - Develop a hypothesis H
  - Design experiment, with controls, to test H
  - Collect data
  - Analyze results see if they confirm H
- Examples: clinical trials, gene knockout experiments, etc.
- Also called prospective studies
- Very expensive and time consuming



Observational Data

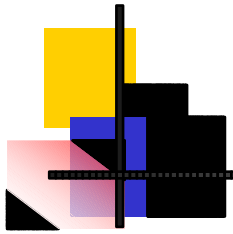
- Now we have huge observational data sets
- Examples: Web logs, customer transactions a retail stores, human genome, etc.
- Makes sense to leverage available data
  - May contain useful information
  - Very cheap to collect
- Assumptions of experimental design violated
  - How can we use such data to do science?
  - Can we do model exploration, hypothesis testing?
- Also called retrospective studies



Complex and Heterogeneous Data

- Data are not simple
  - Comes in different forms
  - From different sources
  - Collected under different conditions
  - Collected with different equipment

- All of these factors cause problems for analys



Complex Data: Structured

Patient

PID	Gender	Birthday
P1	M	3/22/63

Drugs

PID	Date	Medication	Dose	Duration
P1	5/17/98	zoloft	10mg	3 months

Diseases

PID	Date	Symptoms	Diagnosis
P1	1/1/01	palpitations	hypoglycemic
P1	2/1/03	fever, aches	influenza

Lab Tests

PID	Date	Lab Test	Result
P1	1/1/01	blood glucose	42
P1	1/9/02	blood glucose	45

- Dependencies between tables
- Dependencies between rows in table

Complex Data: Semi-Structured

Structured information

Free text

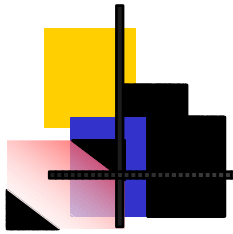
Full name [edit]

Albert's full name is *Albert Félix Humbert Théodore Christian Eugène Marie* in French (pronounced: [albɛʁ feliks œbɛʁ teodɔʁ kʁistjɑ̃ øʒɛn maʁi]), *Albert Felix Humbert Theodoor Christiaan Eugène Marie* in Dutch (pronounced ['ʔɑlbərt 'felɪks 'ɦʏmbərt te'jodoːr 'krɪstiːjaːn ʔøː'ʒɛːn mə'riː]), and *Albert Felix Humbert Theodor Christian Eugen Maria* in German (pronounced ['ʔalbɛʁt 'feːlɪks 'hʊmbɛʁt 'teːodoːɐ̯ 'kʁɪstian 'ʔɔʏgən ma'ʁiːa]).[1]

King of the Belgians	
Reign	9 August 1993 – present
Predecessor	Baudouin
Heir apparent	Philippe, Duke of Brabant
Prime Ministers	*See list* [show]
Spouse	Princess Paola Ruffo di Calabria (1959–present)
Issue	*Detail*
	Prince Philippe, Duke of Brabant Princess Astrid, Archduchess of Austria-Este Prince Laurent
House	House of Belgium (Saxe-Coburg-Gotha)
Father	Leopold III of Belgium
Mother	Astrid of Sweden
Born	6 June 1934 (age 77) Stuyvenberg Castle, Belgium
Signature	
Religion	Roman Catholicism

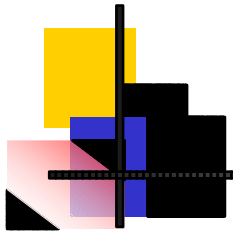
Complex Data: Unstructured

This project addresses the problem of real-world abductiveinference finding the best explanation for evidence when the latter is incomp noisy, possibly contradictory, and in multiple modalities (e.g., senso networks, video, audio, text, etc.). This capability is crucial for supp situation assessment and decision-making by military commanders today's urban theaters of operation. Traditionally, approaches to abductivereasoning have either been based on first-order logic, by determining assumptions sufficient to deduce the observations to b explained, or based on Bayesian networks, by using probabilistic in to compute the posterior probability of alternative explanations giv set of observations. Both of these approaches have significant limit The logical approach is unable to reason under uncertainty and esti the likelihood of alternative explanations. The Bayes-net approach i unable to handle structured representations, and therefore is incap effectively reasoning about situations involving multiple entities wit various relations between them.



Heterogeneous Data

- Data are different
  - Companies use different databases schema
  - Different terminology for the same concept
  - Dates stored differently
  - Full name vs. nick names
- Gives raise to complicated problems
  - Schema matching
  - Ontology matching
  - Entity resolution



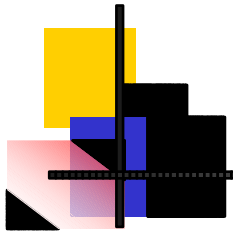
Data Quality

- Data often missing or incomplete
  - Forms have optional fields
  - People are intentionally misleading
- Many measurements are inexact
  - How can Google measure a user's satisfaction with the search results?
  - Did you display the right ads? Would someone have clicked a different ad?
- Known biases in data

Data Ownership

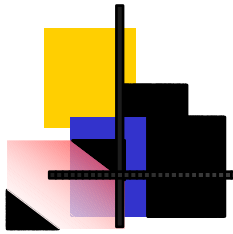
- Data is valuable and people and companies often not interested in sharing

- Can get into trouble for using existing data
  - E.g., most Websites don’t want you to crawl them
  - Statistics from sports matches, etc.
  - Images: rights often retained by photographer (or agency)



Privacy Issues

- Often underestimated by technology people
- Privacy breaches get much attention
  - Massachusetts health records
  - AOL search logs
- Unclear what measures need to be taken to ensure that data is not identifiable
- Can data mining be performed such that resu guarantee the privacy of individuals?
- Correlations/predictors may be discriminatory

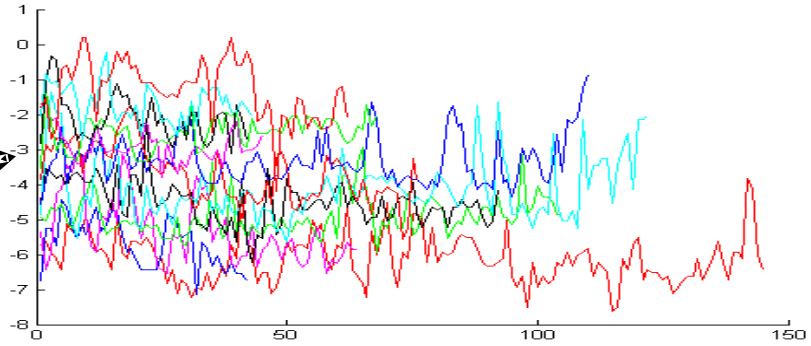


Streaming Data

- Data is not static
  - Stock prices
  - News tickers
  - Cameras
  - Sensor networks
- Actually have so much data that it isn’t possi to permanently store all of it
  - What should we store?
  - How can we make use of the data we see?

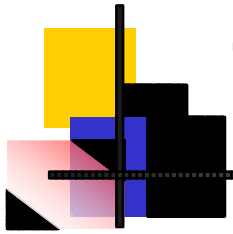
Example: Data Streams

100s of numeric sensors



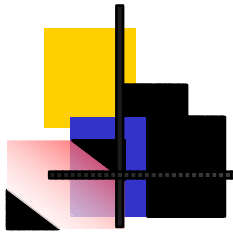
Patient info

Age, weight, gender, etc.



Today's Program

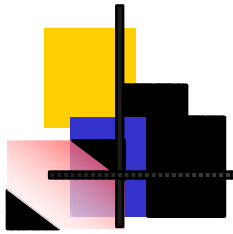
- Logistics, advice and course overview
- Background on data mining
- Data mining challenges
- Data mining tasks
- Data mining vs. machine learning
- The data mining process



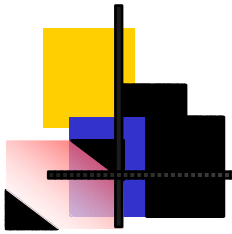
Data mining tasks

- Exploratory data analysis
- Descriptivemodeling
  - Clustering
  - Probability estimation
- Predictivemodeling
  - Classification
  - Regression
- Discoveringpatterns
  - Association detection
  - Trend and deviation detection
- Many others

Exploratory Analysis: Know Your Data

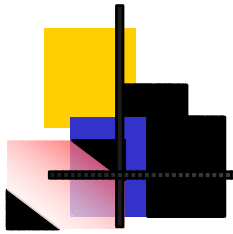


- First step in any data mining problem
- Inspect the data and try to get a feel for what is going on, that is, debug the data
- Getting a sense for
  - What challenges exist
  - What is possible/realistic



What Should You Look For?

- Good to look at simple statistics of
  - Number of variables
  - Size of data
  - Missing values
  - Skew
- For each attribute, look at
  - Discrete: number of possible values, are th ordered, frequency of each value, etc.
  - Numeric: mean, min, max, etc.

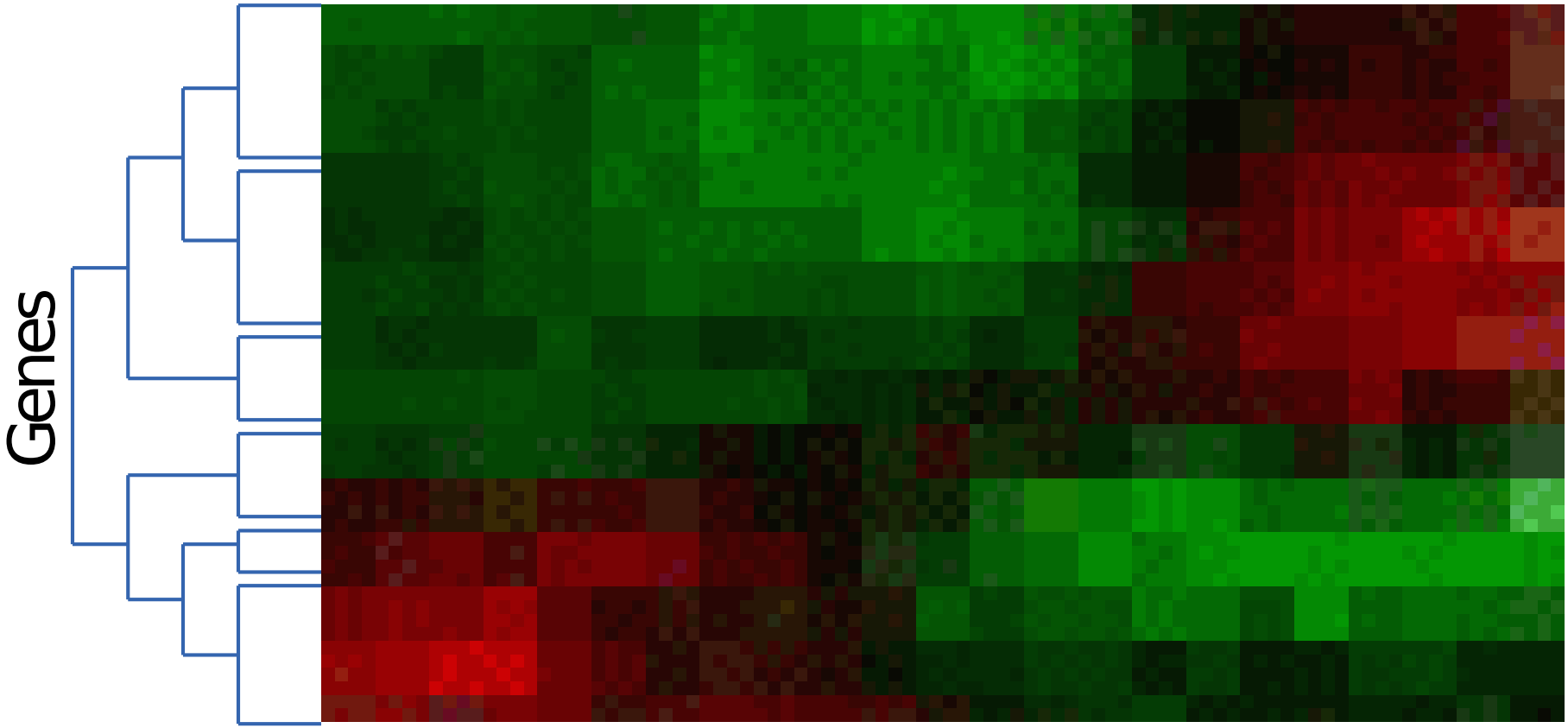


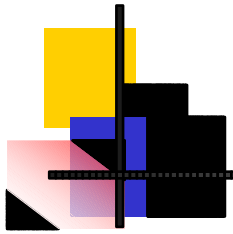
Descriptive Modeling

- Build model that can be
  - Describe or summarize the data
  - Simulate the data
  - Model the process that generated the data
- Techniques
  - Clustering
  - Density estimation/probabilistic models

Example: Gene Expression

(Green= up-regulated, Red= down-regulated)





Example: Document Clustering

- Web search is not great
  - System perspective: covers small coverage Web (<16%), dead links, out of date pages
  - IR perspective: very short queries, huge database, novice users
- One solution: document clustering
  - User receives many (200 -5000) document from Web search engine
  - Group documents in clusters by topic
  - Present clusters as interface

web news images wikipedia blogs jobs more »

ncaa basketball tournament Search advanced preferences

clusters sources sites

remix

All Results (216)

- Brackets (40)
- Tickets (38)
- March Madness (30)
- NCAA Men's Basketball Tournament (18)
- Women's (17)
- Pool (9)
- Photos (11)
- Hoops (5)
- Memphis (7)
- Programs (5)

more | all clusters

find in clusters: Find

Font size: A A A A

Top **212** results of at least **2,237,000** retrieved for the query **ncaa basketball tournament** (details)

Sponsored Results

College Hoops Contest - Know college **basketball** ? Prove it. Go for the $100,000 grand prize! - www.wagerline.com

Sports Contest Promotions - Run a sports contest promotion for your business or website. - www.poolhost.com

Search Results

1. **NCAA** Men's Division I **Basketball** Championship - Wikipedia, the free ...
   The **NCAA** Men's Division I **Basketball** Championship is a single elimination **tournament** held each spring featuring 65 [1] college **basketball** teams in the United States. This **tournament**, organized by the National Collegiate Athletic Association (**NCAA**), was first developed by the National Association of **Basketball** Coaches in 1939. [2]**Tournament** format · Format history · March Madness and ...
   en.wikipedia.org/wiki/NCAA_Men's_Division_I_Basketball_Championship - [cache] - Live, Ask, Gigablast

2. Welcome To Your Official **NCAA** Web Sites
   Enter **NCAA**.com For complete March Madness coverage, brackets and other championship **tournament** information for all **NCAA** sports. Enter **NCAA**.org For information about the **NCAA**,
   www.ncaa.org - [cache] - Live, Gigablast

3. 2009 **NCAA Basketball Tournament** | CollegeHoops.net
   2009 **NCAA Tournament** preview, schedule, bracket, and bracketology.
   www.collegehoopsnet.com/ncaatournament - [cache] - Live, Ask

4. **NCAA Tournament** Tickets, 2009 **NCAA Basketball Tournament** Info, Final
   March Madness is here and GoTickets.com has your 2009 **NCAA**® Men's **Basketball Tournament** tickets and Final Four® tickets.
   www.gotickets.com/sports/college_basketball/ncaa_tournament.php - [cache] - Ask, Gigablast

5. **NCAA**.com - The Official Web Site of the **NCAA**
   Selection Sunday Challenge. Think you deserve a seat on the **NCAA** Men's **Basketball** Selection Committee? See if you can pick the correct field of 65.
   www.ncaa.com - [cache] - Live, Gigablast

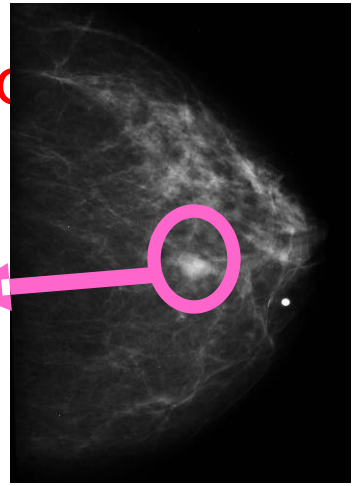
6. **NCAA Tournament** Tickets, 2009 **NCAA Basketball Tournament** Ticket ...
   NCAA Tournament Tickets from TickCo Premium Seating; rapid delivery on **NCAA Tournament**/March Madness tickets order and save today!
   www.tickco.com/sports_basketball_ncaa_tournament_tickets.htm - [cache] - Live, Ask

7. Working Class Software
   NCAA basketball tournament program.
   www.wcsoftware.com - [cache] - Open Directory, Ask, Gigablast

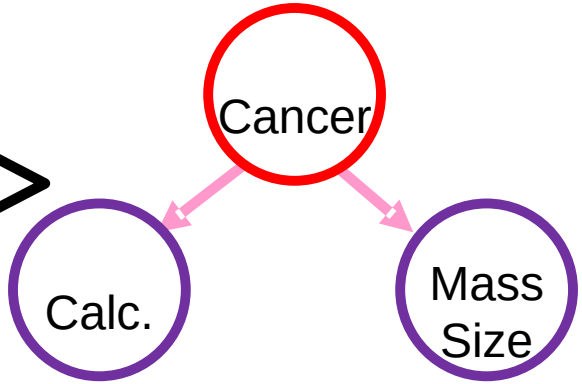
8. **NCAA Basketball Tournament** Most Outstanding Player - Wikipedia, the ...
   At the conclusion of the **NCAA** men's and women's Division I **basketball** championships (the "Final Four" **tournaments**), the Associated Press selects a Most Outstanding Player. The MOP need not be, but almost always is a member of the Championship team. The last man to win the award despite not being on the Championship team was Hakeem Olajuwon in 1983; the last woman to do so was Dawn Staley in 1991.
   en.wikipedia.org/wiki/NCAA_Basketball_Tournament_Most_Outstanding_Player - [cache] - Live, Ask

Example: Medical Diagnoses

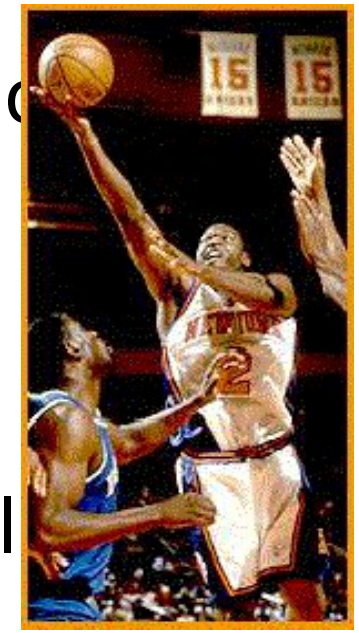
Patient	Date	Calcification Fine/Linear		...	Mass Size	Loc	Canc
P1	5/02	Present			3mm	RU4	No



Learn model that maps features to cancer



Example: NBA Data

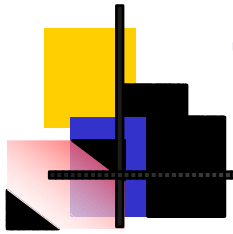


- NBA logs all play by play informatic
  - Which players are in the game
  - Shots attempts
  - Etc.
- Questions: Which lineups work wel
  - Offensive efficiency
  - Defensive efficiency
  - Etc.
- See: http://www.synergysportstech.com/

Example: Frequent Itemsets

Items
Bread, Cheese,Wine
Chips, Salsa,Wine
Bread, Cheese,Wine
Buns,Hamburger Meat, Ketch
Cheese,Wine
Chips, Coke,Salsa
Hamburger Meat, Ketchup
Beer,Chips, Salsa
Bread,Cheese, Wine

- Items co-purchased
  - Cheese and Wine
  - Chips and Salsa
  - Hamburger Meat and Ketchup
- Associations: {Cheese,W}→ {Br}, w/confidence = 0.75

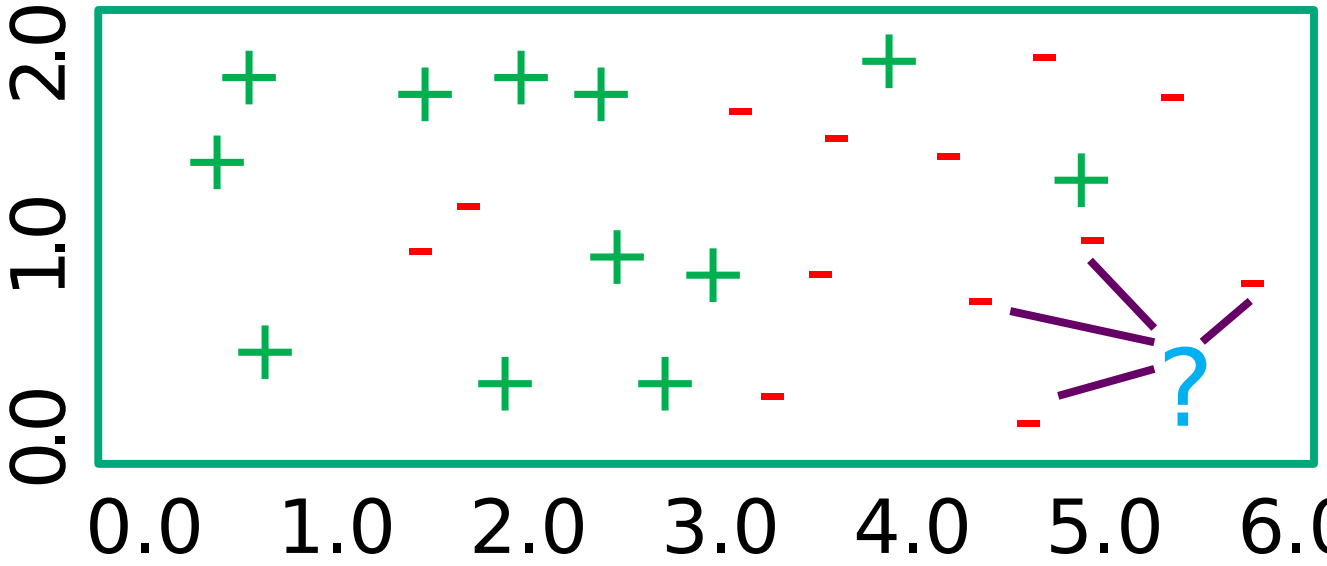


Today's Program

- Logistics, advice and course overview
- Background on data mining
- Data mining challenges
- Data mining tasks
- Data mining vs. machine learning
- The data mining process

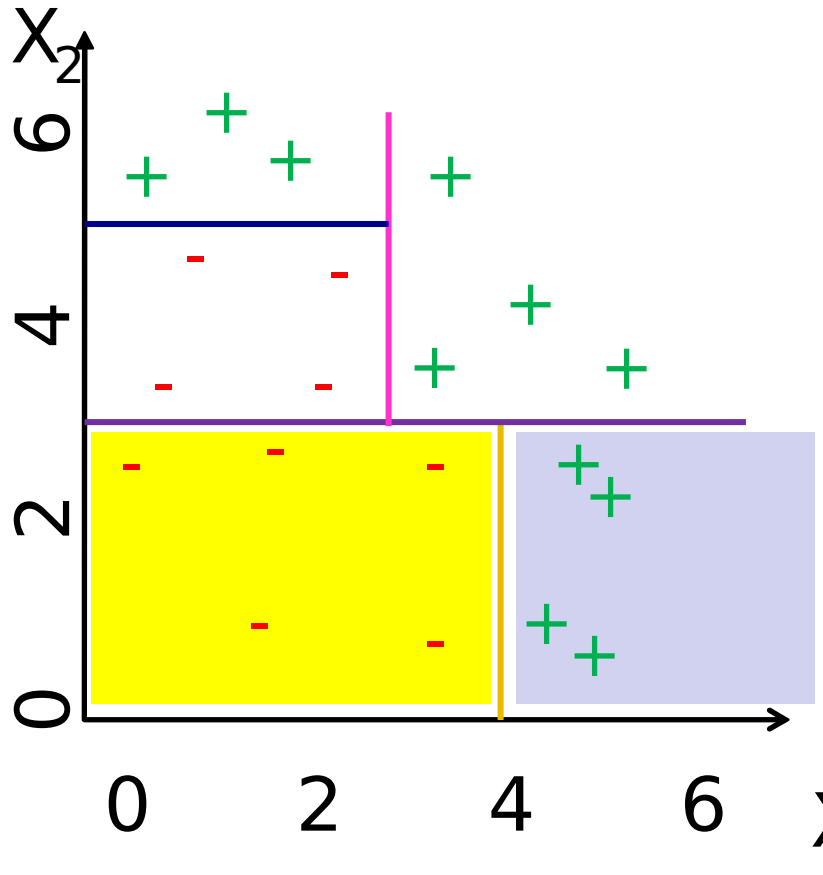
Instance-Based Learning

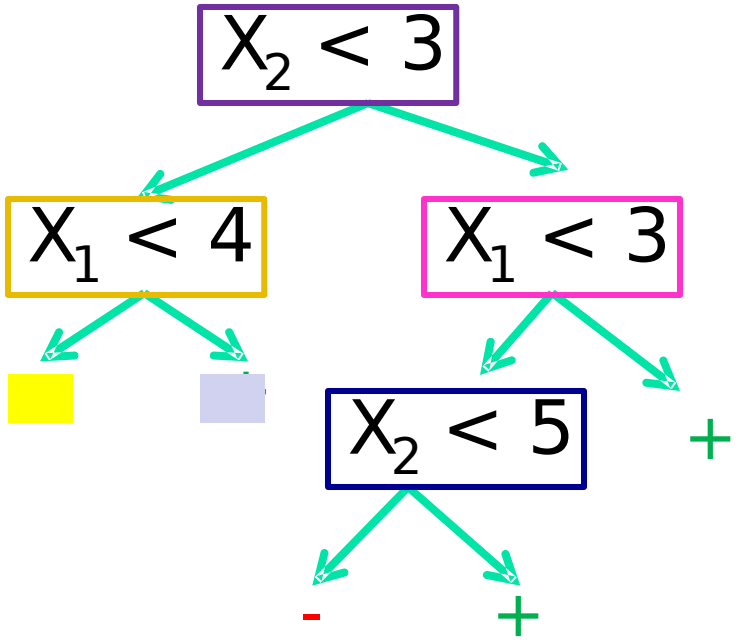
- Learning ≈ memorize training examples
- Find most similar example
  - Classification: output its category
  - Regression: output its value



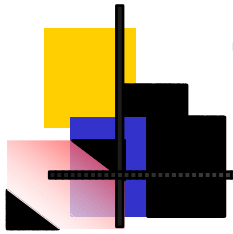
Label based on neighbors

How Algorithms Partition Feature Space





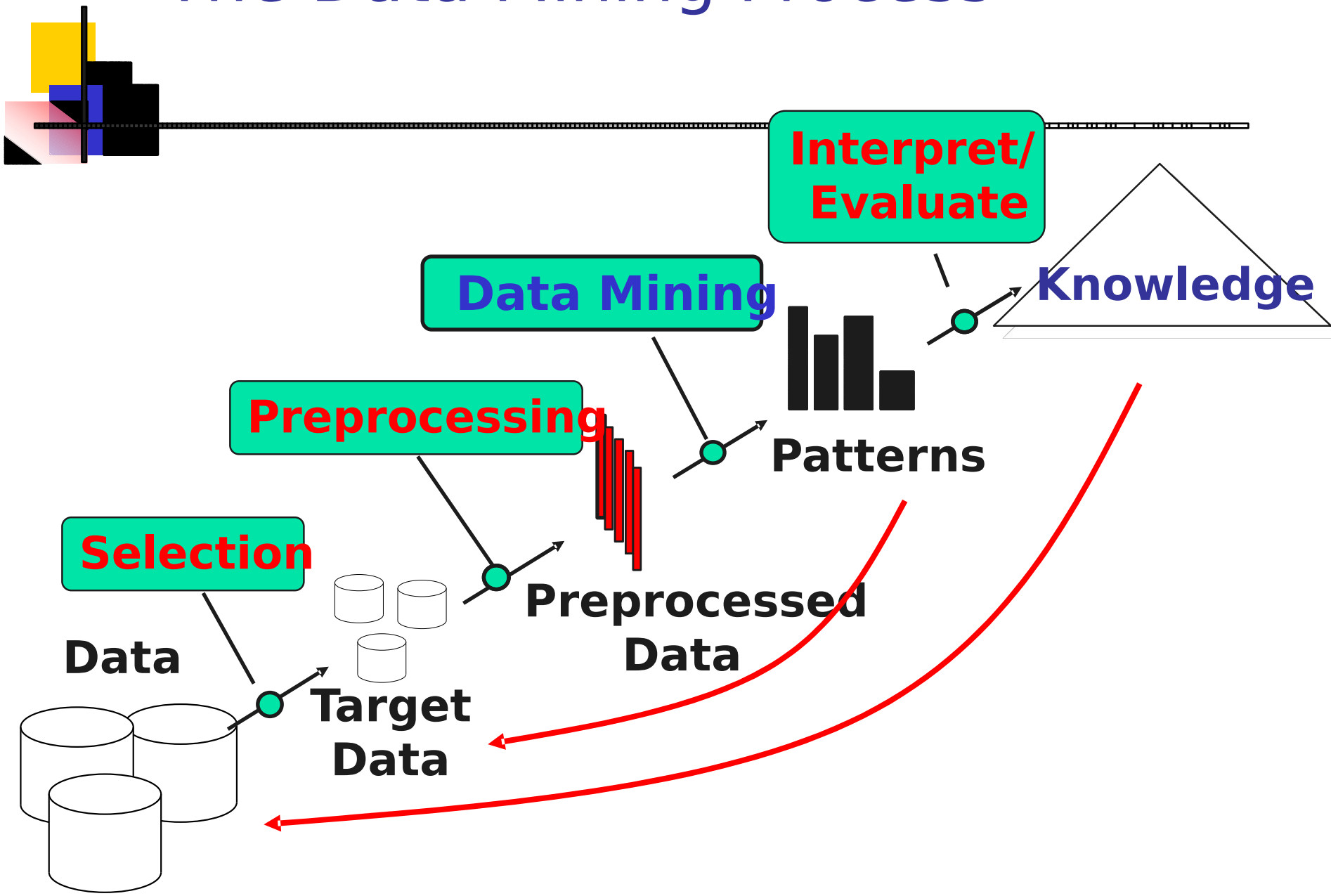
Decision Trees: Divide feature space into axis parallel rectangles and labels each one with one of the K classes



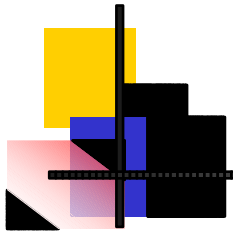
Today's Program

- Logistics, advice and course overview
- Background on data mining
- Data mining challenges
- Data mining tasks
- Data mining vs. machine learning
- The data mining process

The Data Mining Process



Requirements for a Data Mining System



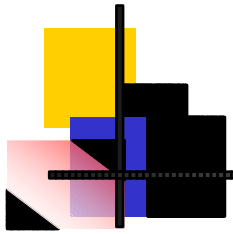
Data mining systems should be

- Computationally sound
  - Scalability time and space complexity
  - Parallelizable, e.g., MAP-Reduce and Hadoc
- Statistically sound
  - Are patterns meaningful?
  - Do our results generalize to new data?
- Ergonomically sound
  - Presents results in a comprehensible mann
  - Does it need 6 PhDs to run it?

Components of a Data Mining Syst

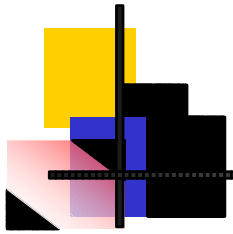
- Representation
- Evaluation
- Search
- Data management
- User interface

} Focus of this course



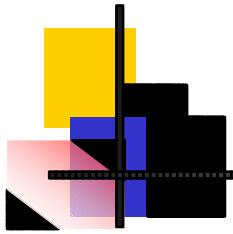
Representation: Data

- Feature vectors
- Relational database
- Free text
- Images
- Graphs
- Etc.



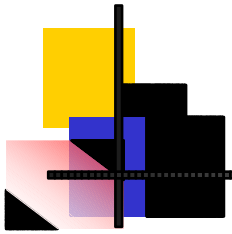
Representation: Model

- Decision trees
- Graphical models
- Rule set
- Association rules
- Graph patterns
- Sequential patterns
- Etc.



Evaluation

- Objective
  - Accuracy
  - Precision and recall
  - Cost / Utility
  - Fast
  - Etc.
- Subjective
  - Interesting
  - Novel
  - Actionable
  - Etc.



Summary

- We live an age where large amounts of data are commonplace
- Data mining is hugely popular and hugely successful because it extracts useful information from this data
- This information comes in many forms
  - Models
  - Patterns
  - Etc.
- Data mining is challenging for many reasons

References

- Datamining Slides, KULeuven, Jesse Davis