

**Before class starts, if you can, please vote on
this poll via your smartphone, tablet, or
laptop (I recommend downloading the Poll
Everywhere app as we will be doing this
again.) The main reason I am interested in IS
733 Data Mining is:**

I am broadly interested in
data science and machine
learning, and it seems like

I want to use data mining
techniques in my future
career in industry.

I plan to use data mining
for my applied research.

I plan to do foundational
research on data mining.

IS 733: Data Mining

Instructor:

Email:

Office Hours:

Dr. Nirmalya Roy

nroy@umbc.edu

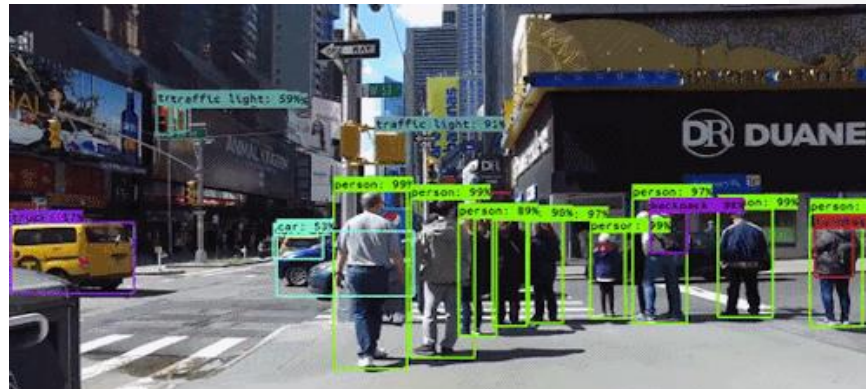
Tuesdays 9:00 – 10:00 am online (or by appointment)

Course website:

<https://mpsc.umbc.edu/courses/is-733-data-mining>

Poll everywhere:

PollEv.com/nirmalyaroy910



Eating



Bathing



Dressing



Transferring



Toileting



Walking or moving around



How is this all going to work?

- This is going to be a rough semester. Let's support each other as best we can.
- Online synchronous instruction (Blackboard Collaborate)
 - In-class exercises (Poll Everywhere)
 - Notes, lectures, homeworks and announcements (Course website)
 - Q&A, Submissions and grades (Blackboard)

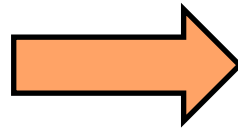
Participating in Class

- If you have a question, please use the “**raise hand**” feature (then speak when called on), and/or write your question in the **chat**.
- If I ask a question, you can **answer via microphone or chat**, whichever you’re comfortable with.
- **Breakout rooms** – you need **microphones on**. **Cameras recommended** but optional. No need for cameras to be on during lecture time, unless you want to.

Data Mining

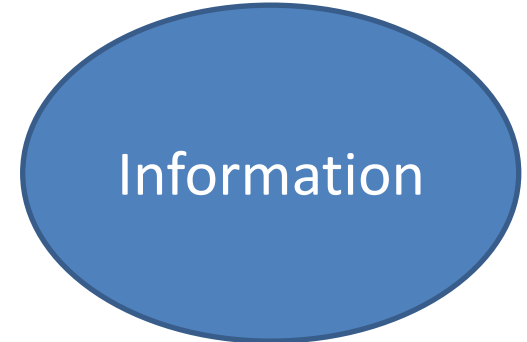


Complicated, noisy,
high-dimensional



Data Mining

Find patterns



Explore, understand, predict

What Is Data Mining?



- Data mining (knowledge discovery from data)
 - Extraction of interesting (non-trivial, implicit, previously unknown and potentially useful) patterns or knowledge from huge amount of data
 - Data mining: a misnomer?
 - “Knowledge mining from data”?
- Alternative names
 - Knowledge discovery (mining) in databases (KDD), knowledge extraction, data/pattern analysis, data archeology, data dredging, information harvesting, business intelligence, etc.



Why Data Mining?

- The **Explosive Growth of Data**: from terabytes to petabytes
 - **Data collection and data availability**
 - Automated data collection tools, database systems, Web, computerized society
 - **Major sources of abundant data**
 - **Business**: Web, e-commerce, transactions, stocks, ...
 - **Science**: Remote sensing, bioinformatics, scientific simulation, ...
 - **Society and everyone**: news, digital cameras, YouTube
- We are drowning in data, but starving for knowledge!
- “Necessity is the mother of invention”—Data mining—Automated analysis of massive data sets

Information is crucial: Life and death!

- Example 1: *in vitro* fertilization
 - Given: embryos described by 60 features
 - Problem: selection of embryos that will survive
 - Data: historical records of embryos and outcome

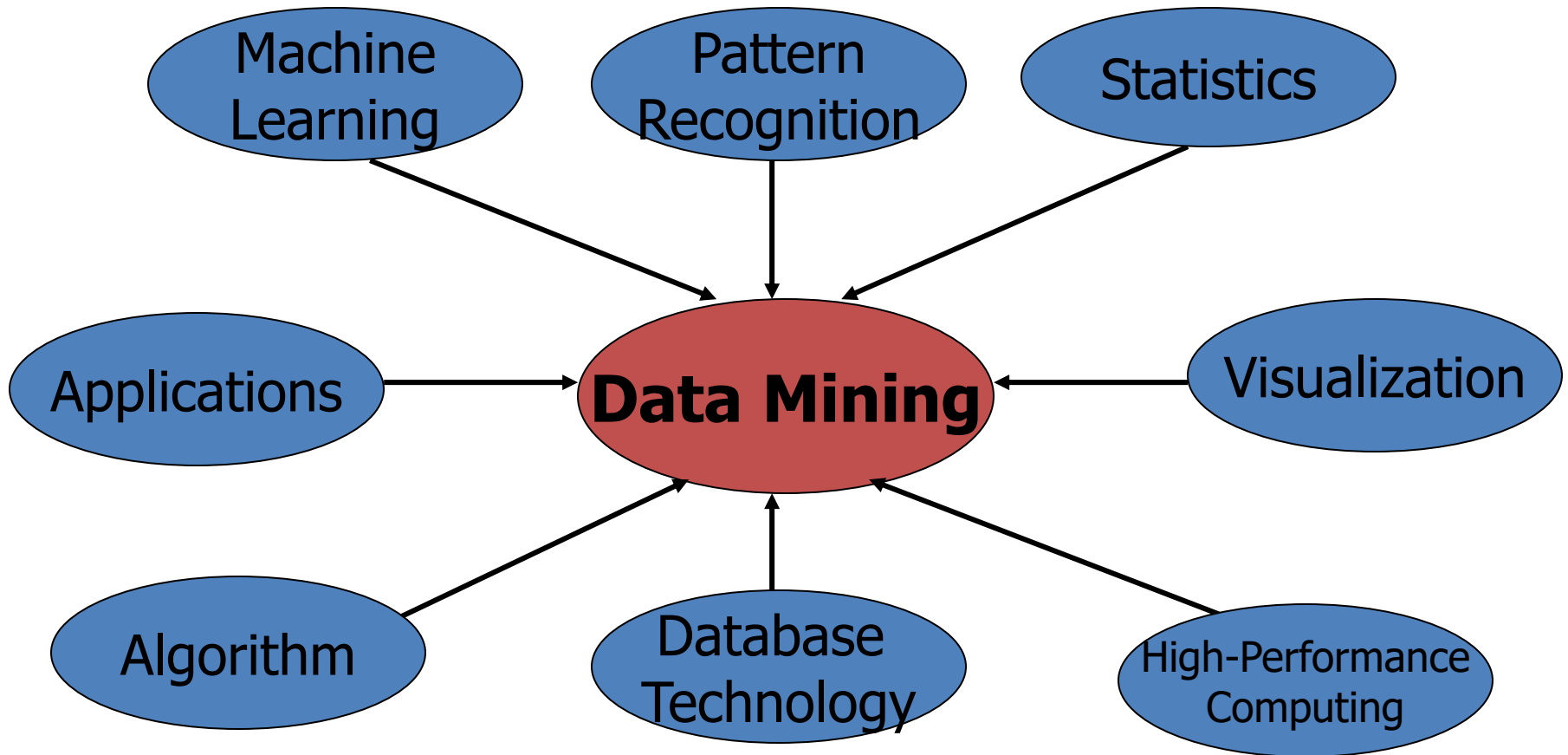


Information is crucial: Life and death!

- Example 2: cow culling
 - Given: cows described by 700 features
 - Problem: selection of cows that should be culled
 - Data: historical records and farmers' decisions

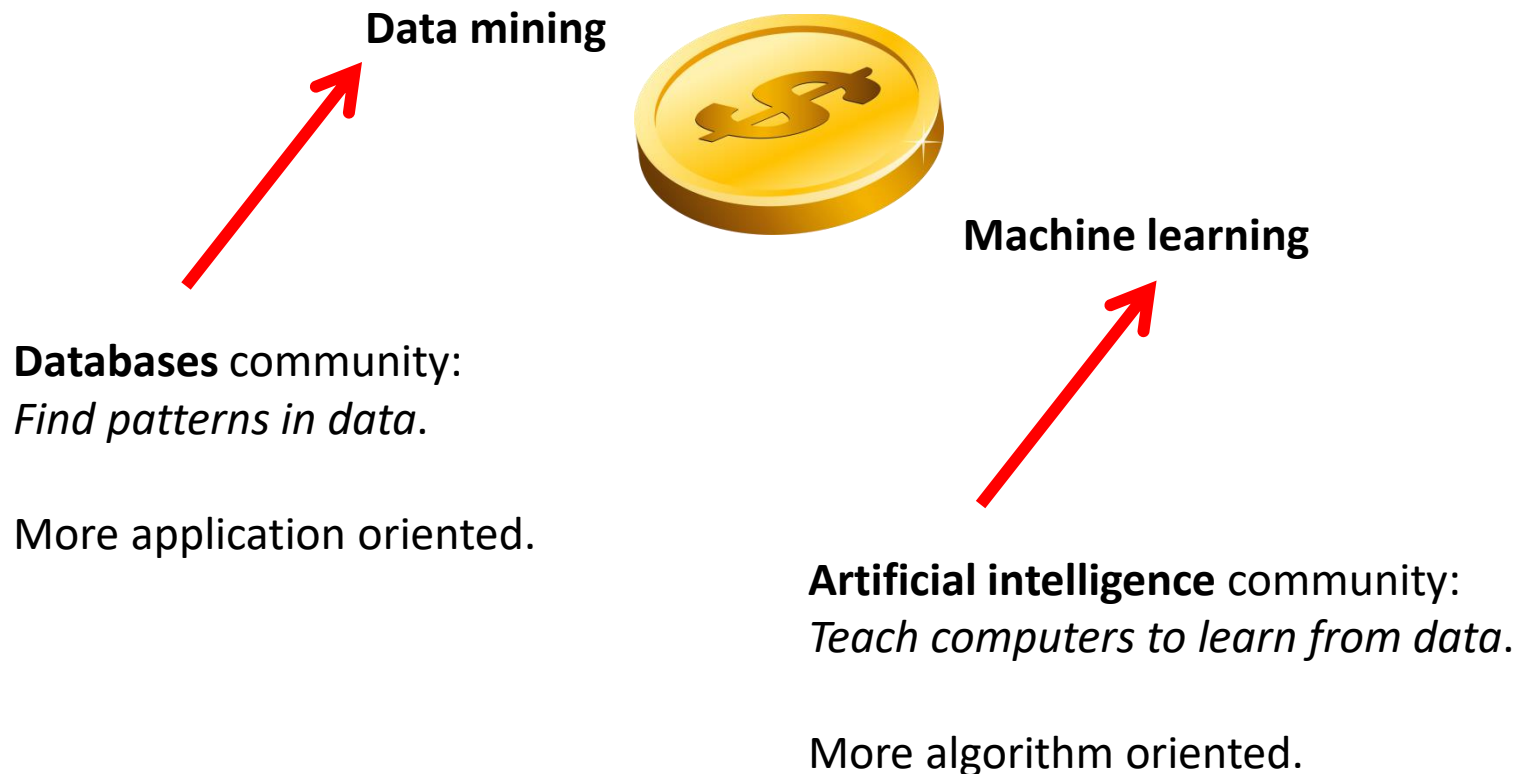


Data Mining: Confluence of Multiple Disciplines



Data Mining vs Machine Learning

- Two sides of the same coin



Learning Goals

- By the end of the course you will be able to:
 - **Apply** a variety of **data mining techniques** to real-world situations,
 - **select appropriate strategies** for each step in the data mining process, and
 - discuss the **underlying theoretical principles** behind data mining methods, and the **practical implications** of these.

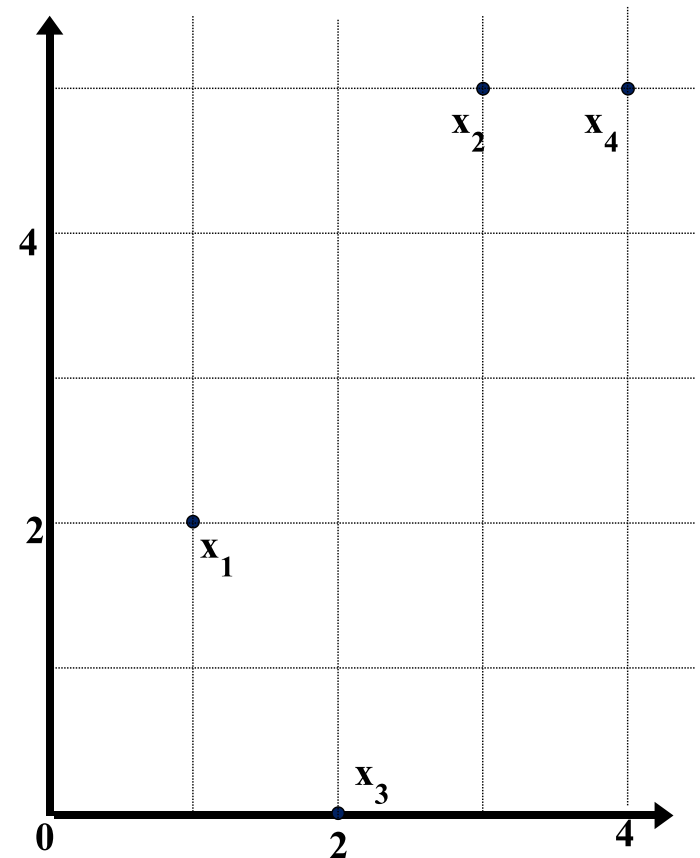
Week 1

- Course overview
- Introduction to data mining
 - Applications
 - The data mining process
 - Data mining ethics



Week 2

- Know your data
 - Instances and attributes
 - plotting and visualization



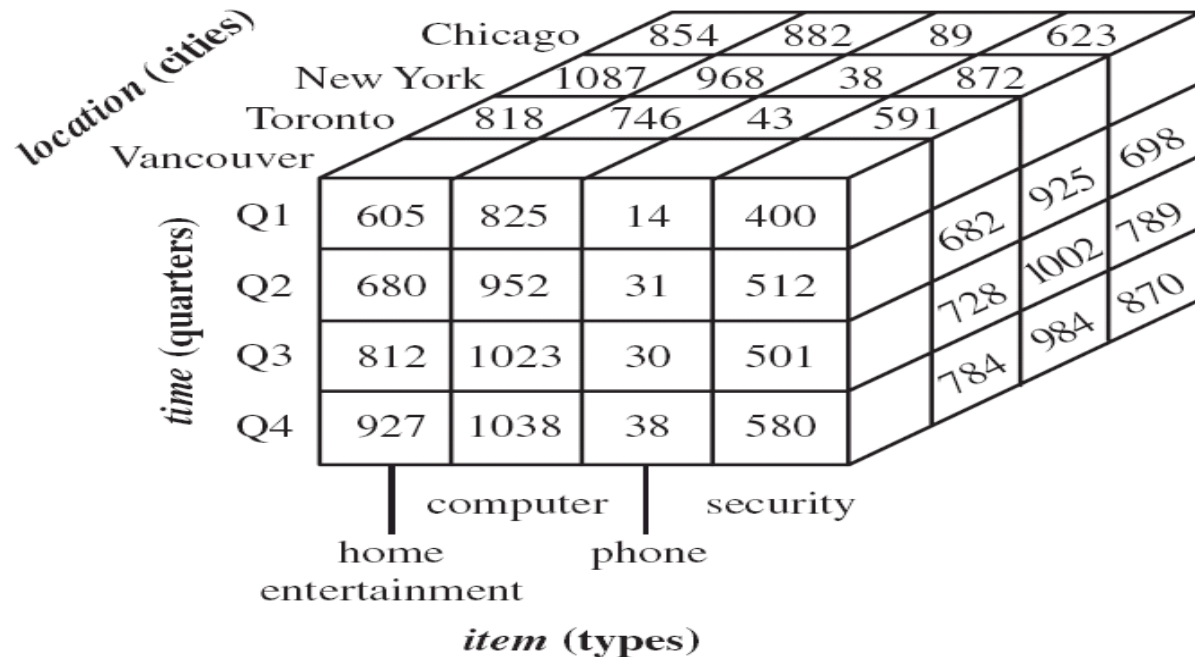
Week 3

- Data preprocessing
 - Data cleaning, integration, transformation, reduction, discretization.



Week 4

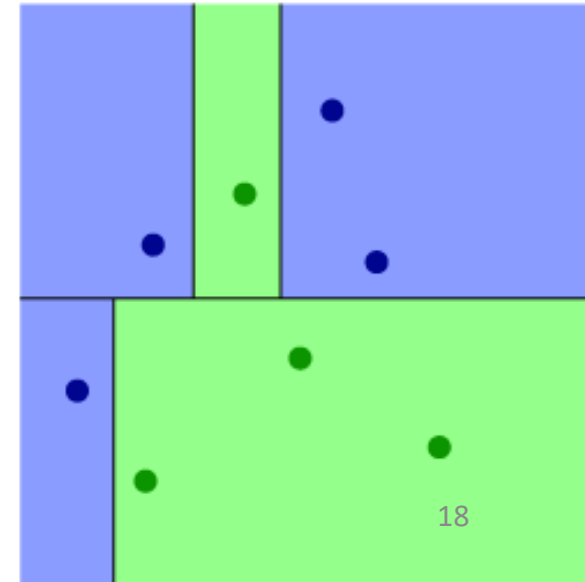
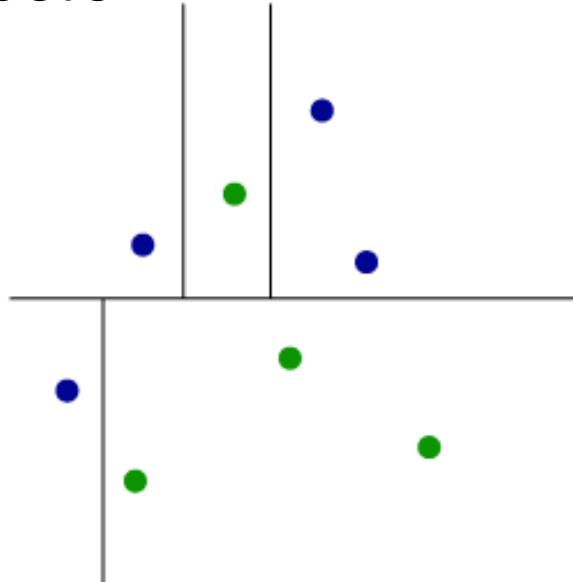
- Data warehousing
 - OLAP vs OLTP
 - Data cubes
- Project brainstorming.



Week 5

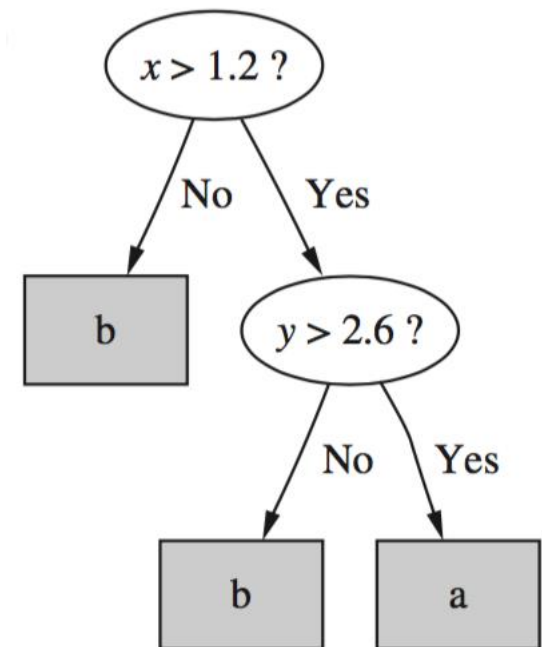
- Knowledge representation
 - Linear models
 - Trees
 - Rules
 - Nearest neighbors

- Sharing project ideas.



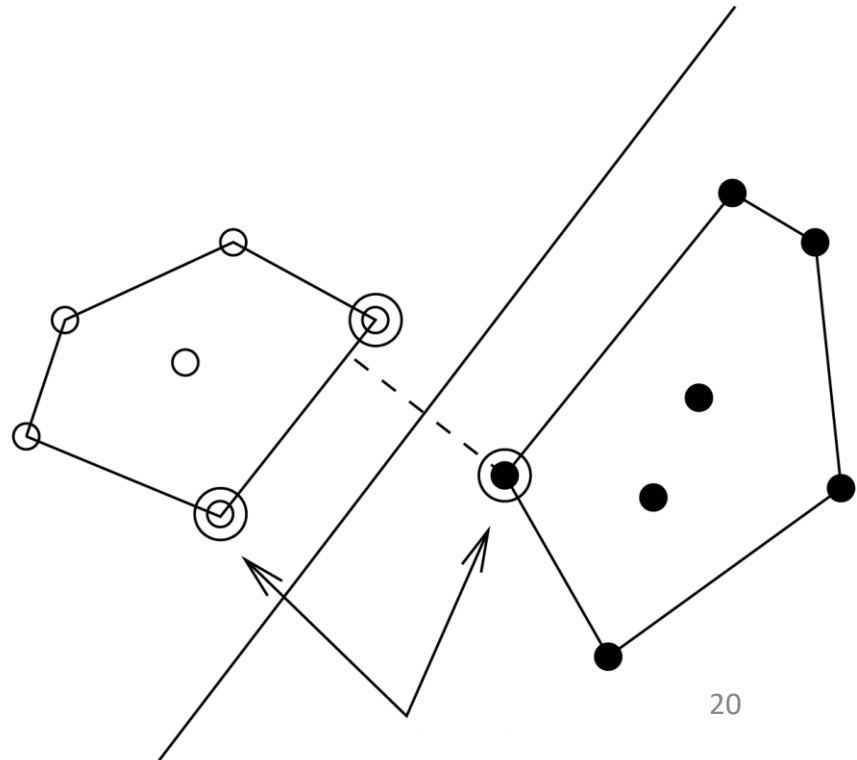
Week 6

- Supervised learning
 - Decision trees
 - Decision rules
 - Ethical thinking:
fairness in classification



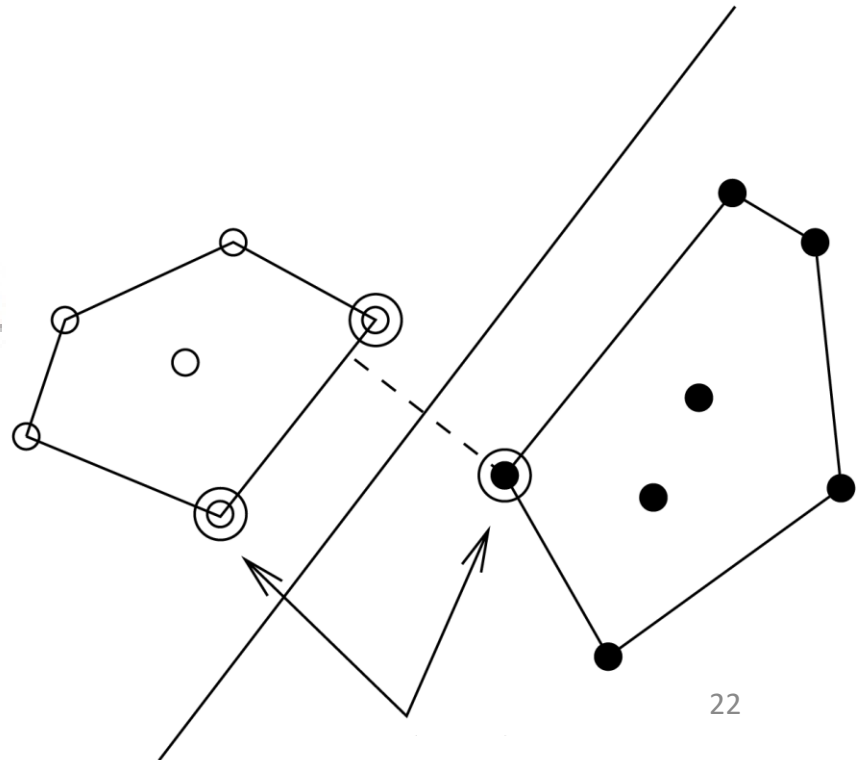
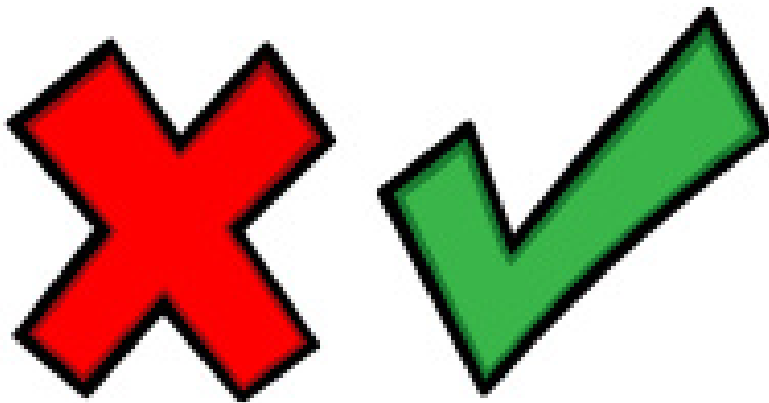
Week 7

- Supervised learning (continued)
 - Naive Bayes
 - logistic regression
 - support vector machines.



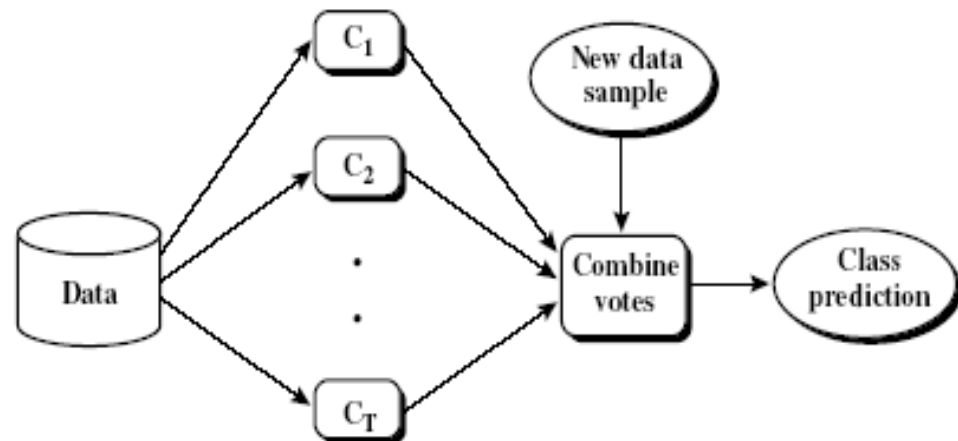
Week 9

- Evaluation of supervised learning
 - Hold-out method
 - Cross validation
 - ROC curves



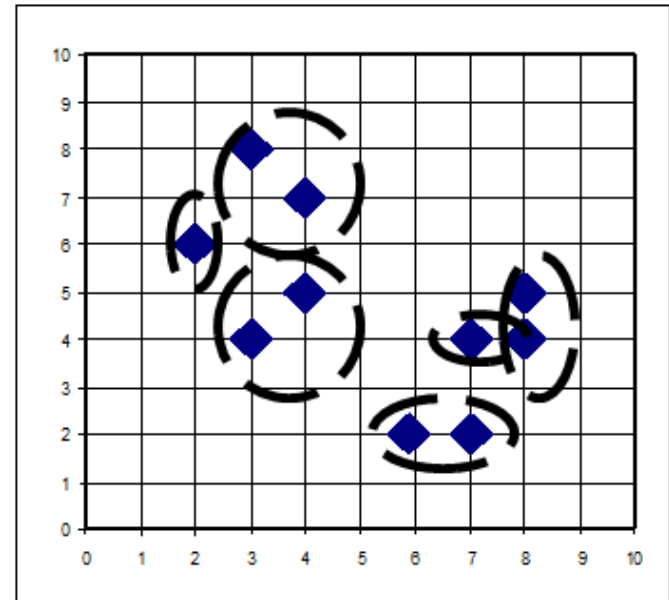
Week 10

- Ensemble methods
 - Bagging
 - Boosting
 - Random forests
 - Stacking



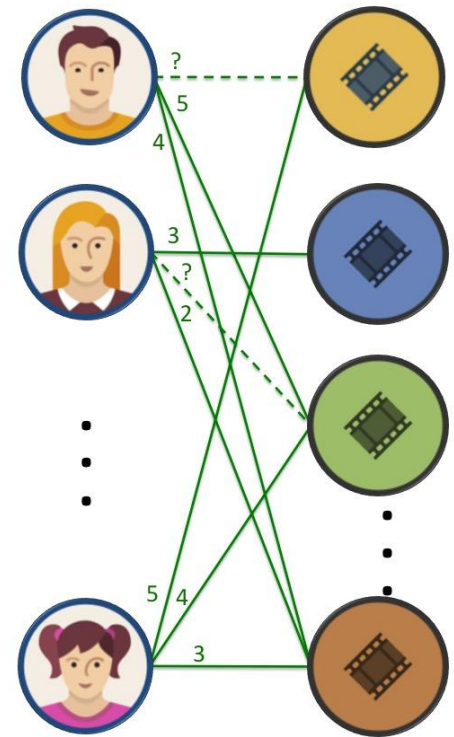
Week 11

- Unsupervised learning
 - Association rule learning
 - K-means
 - Hierarchical clustering



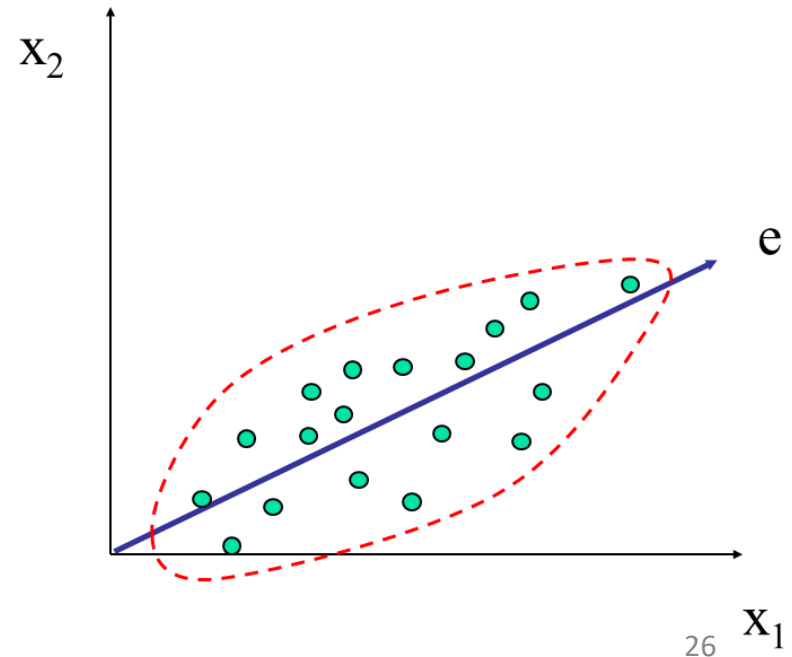
Week 12

- Recommender systems
 - Content filtering
 - Collaborative filtering



Week 13

- Dimensionality reduction
 - Principal components analysis (PCA)
 - Independent components analysis (ICA)
 - t-SNE
 - Random projections



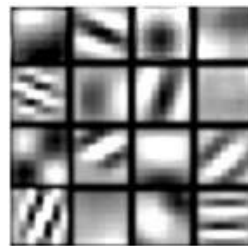
Week 14

- Text mining
 - Bag of words representation
 - n-grams
 - topic models
 - word embeddings



Week 15

- Deep learning
 - Deep feedforward networks
 - Backpropagation
 - Convolutional neural nets



First Layer



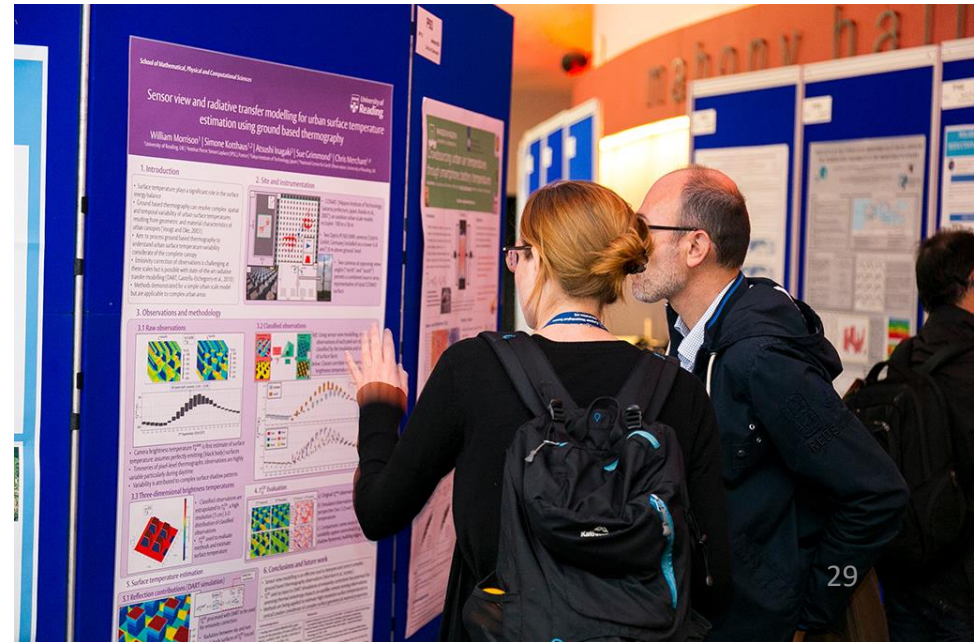
Second Layer



Third Layer

Week 16

- Group project presentations
 - (virtual) Poster session! Online due to COVID-19
 - Present posters to peers and instructor in Blackboard Collaborate breakout rooms

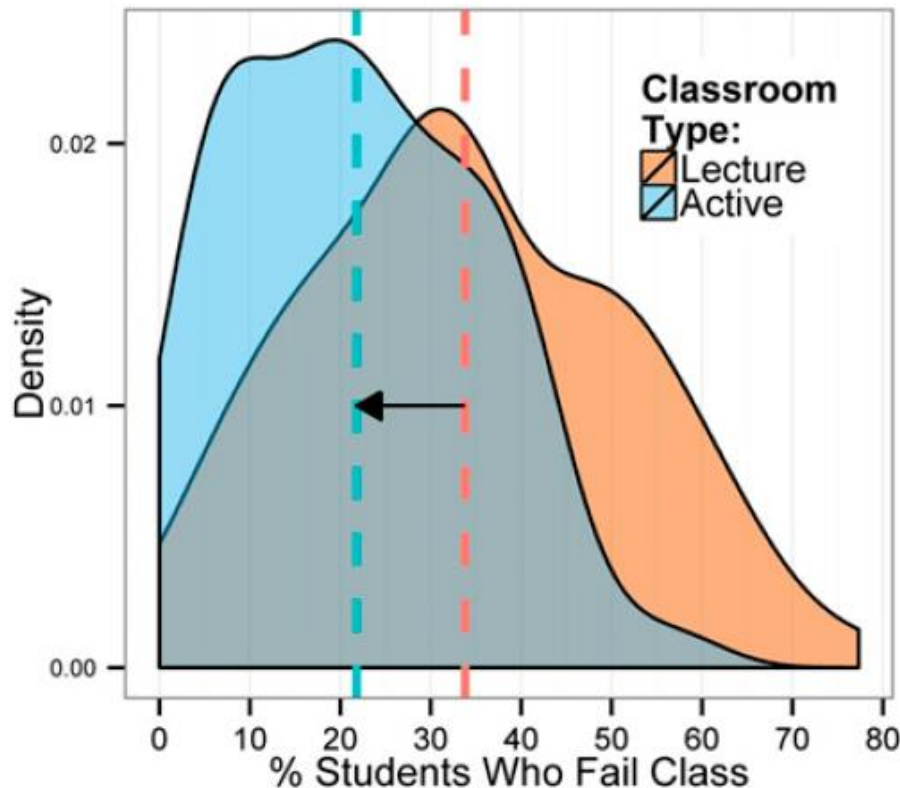


Pedagogy: Active learning

- Student-centered instruction
- Actively engage with the material in class
- In-class quizzes and polls, peer instruction, discussion with peers

Active learning increases student performance in science, engineering, and mathematics

Scott Freeman^{a,1}, Sarah L. Eddy^a, Miles McDonough^a, Michelle K. Smith^b, Nnadozie Okoroafor^a, Hannah Jordt^a, and Mary Pat Wenderoth^a



“If the experiments [were] medical interventions, ...

the control condition might be discontinued

because the treatment being tested was clearly more beneficial.”

Pedagogy: Active learning

- **Everyone gets to participate**, not just students in the front row
- With traditional lectures only, course content is “transmitted” in class, and you have to do the hard yards of **learning on your own**
- With active learning to augment lectures, **learning, synthesis, and integration** with prior knowledge **occur in class**, with support from instructor and peers

Peer instruction

- No need to buy a clicker: *polleverywhere.com*

The screenshot displays the Poll Everywhere website interface. At the top, the navigation bar includes the Poll Everywhere logo, links for Plans & Pricing, Take a tour, Help & FAQ, My polls, and Log out. The main heading is "Live Audience Participation" with the subtext "Poll Everywhere lets you engage your audience or class in real time". Below this, there is a red button labeled "Create your first poll" and a link to "Watch our 2 min video". A note states "Takes 30 seconds. No signup required". On the right, a callout says "Use your phone to text a vote now!" with an arrow pointing to a smartphone. The smartphone screen shows a "New Message" interface with a contact named "Lion" and a phone number "22333". The laptop screen displays a poll titled "What's your favorite animal?" with the instruction "Text a KEYWORD to 37607". The poll results are shown as a bar chart with three categories: LION (17%), TURTLE (50%), and GRANDPA (33%). A small note at the bottom of the laptop screen says "Message and data rates may apply".

What's your favorite animal?

Text a **KEYWORD** to 37607

Animal	Percentage
LION	17%
TURTLE	50%
GRANDPA	33%

Message and data rates may apply

Peer instruction

- You can respond to Poll Everywhere polls with your laptop, tablet, or smartphone (bring it to class!)

[PollEv.com/nirmalyaroy910](https://pollev.com/nirmalyaroy910)

- I recommend using the [Poll Everywhere app](#), for Android and iPhone. Find it in the app store.
- If you do not have a smartphone or laptop, please let me know and we can work something out.

Peer instruction: Participation grades

- I'll start recording participation using Poll Everywhere starting next week, so please sign up for an account (and participate in polls!)

Your **account needs to be linked to the course**, which is a separate step! The link to register is

<https://pollev.com/nirmalyaroy910/register>

- Blackboard will also be factored into participation grades.

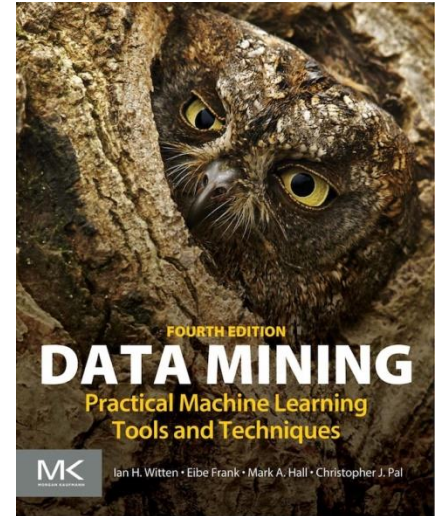
Course Readings

- Required textbook:

Data Mining: Practical Machine Learning Tools and Techniques, Fourth Edition (Witten et al.)

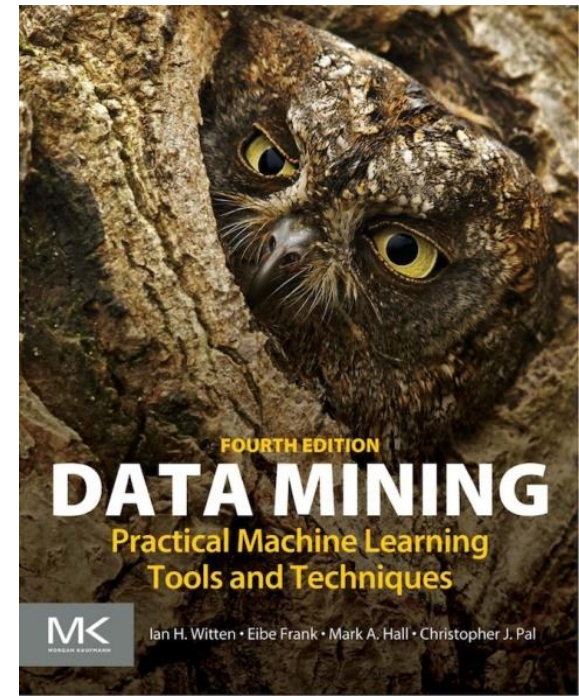
See: <http://www.cs.waikato.ac.nz/ml/weka/book.html>

- Readings need to be completed **before each class**. We will do reading quizzes at the start of each class.
- It is **very important that you do the readings** so that we can make effective use of our limited lecture time together (a “flipped classroom” approach.)



Reading for Today's Lecture

- [Usama Fayyad, Gregory Piatetsky-Shapiro, and Padhraic Smyth, From Data Mining to Knowledge Discovery in Databases, AI Magazine 17\(3\) 1996, up to page 7.](#)
- Or if you have the textbook:
- Witten et al., Chapter 1, up to page 15
- This is a good overview of what IS 733 is all about, if you're still deciding whether this course is for you.



Blackboard

- Please use blackboard to ask questions about the course, instead of emailing me directly, so that everyone in the class can benefit from the answer
- Blackboard will also be used for announcements, including information on the readings

Assessment

- **Homeworks** 25% (5 of them, 5% each)
- **Group Project** 35%
- **Final** 35%
- **Participation** 5%
 - Poll questions 4%
 - At least two blackboard posts 1%
(can be either questions or answers)

Group Project

- Groups of 2
- An open-ended project, to give you an opportunity to explore the techniques and principles covered in the course
- May overlap with your other research, but not any other class project

Group Project

- Milestones / deliverables
 - *Groups formed by 2/16/2021*
 - *Proposal 5% (due 2/23/2021)*
 - *Mid-term report 5% (due 4/6/2021)*
 - *Group project poster 10%*
(presented in class 5/11/2021, digital copy due at the same time)
 - *Final report 15% (due **Friday** 05/14/2021)*
- Note that you may have to read ahead to start your project. All readings are listed on the syllabus, on the course webpage.

How to Succeed in IS 733

- While the course will be challenging in the sense that we have a lot of material to cover in 15-16 weeks, the course is designed so that **everyone has the opportunity to succeed**. I do not grade on a curve.
- **Learning goals** for each lesson will be clearly stated.
 - **if you achieve these you will be well prepared for the exam.**
- **Homeworks** are designed to give you practice and feedback on the learning goals.
- The 5% **participation** marks are there for the taking
 - Participate in peer instruction/class discussions/etc (4%)
 - Blackboard (2 posts for 1% of course grade, can be questions, answers, comments)
- Come to **office hours**!!! Tuesdays 9 -10 am

Required Knowledge



- Requires IS 620, or consent of the instructor.
- Required knowledge:
 - Basic **programming ability** in a high-level language such as Java, Python, R, or Matlab
 - No previous background in data mining is required.
 - Although relatively non-technical, a basic understanding of elementary concepts in continuous and discrete mathematics will be needed (**linear algebra**, **Boolean logic**, **graphs** and **trees**, ...).

Academic Integrity

- UMBC's policies on academic integrity will be strictly enforced
- **All of your work must be your own.**
- **Acknowledge** and **cite** source material in your papers or assignments.
- While you may verbally discuss assignments with your peers, **you may not copy or look at anyone else's written assignment work or code, or share your own solutions.**
- **Any exceptions will result in a zero on the assessment in question, and may lead to further disciplinary action.**

Academic Integrity

- “Cheating, fabrication, plagiarism, and helping others to commit these acts are all forms of academic dishonesty, and they are wrong.”
-(UMBC's academic integrity overview)
- "Students shall not submit as their own work **any work which has been prepared by others.**"
-(USM policy document)





**Which is worth the least percentage points
of your grade for the course, assuming 100%
of the corresponding points are earned?**

The final exam

Five homeworks

Four homeworks plus
participation

The group project





**Which is worth the least percentage points
of your grade for the course, assuming 100%
of the corresponding points are earned?**

The final exam

Five homeworks

Four homeworks plus
participation



The group project



**If I create a Poll Everywhere account, my
responses are automatically linked with the
IS 733 course.**

True

False



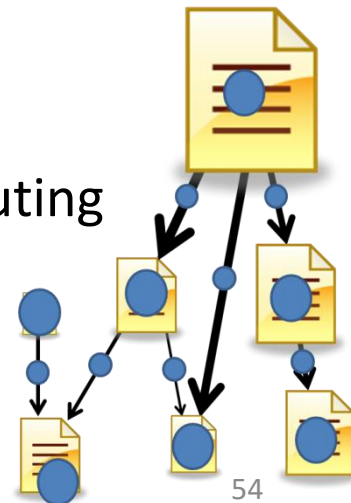
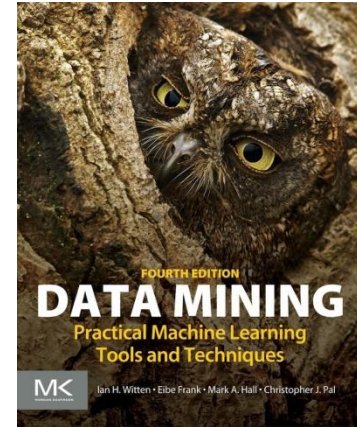
**The group project proposal, mid-term report
and poster are cumulatively worth more
towards the total grade than the final
report.**

True

False

Interlude: About your instructor

- Undergrad in CSE at Jadavpur University, India and Master's in CSE at the University of Texas at Arlington, where I worked on Grid Computing
- Ph.D. at University of Texas at Arlington, working on reinforcement learning & game theoretic RL models for user activity and actions in smart home environments
- Postdocs at University of Texas at Austin, working on quality-of-service aware optimization in pervasive computing applications
- Started faculty position at UMBC Fall 2013

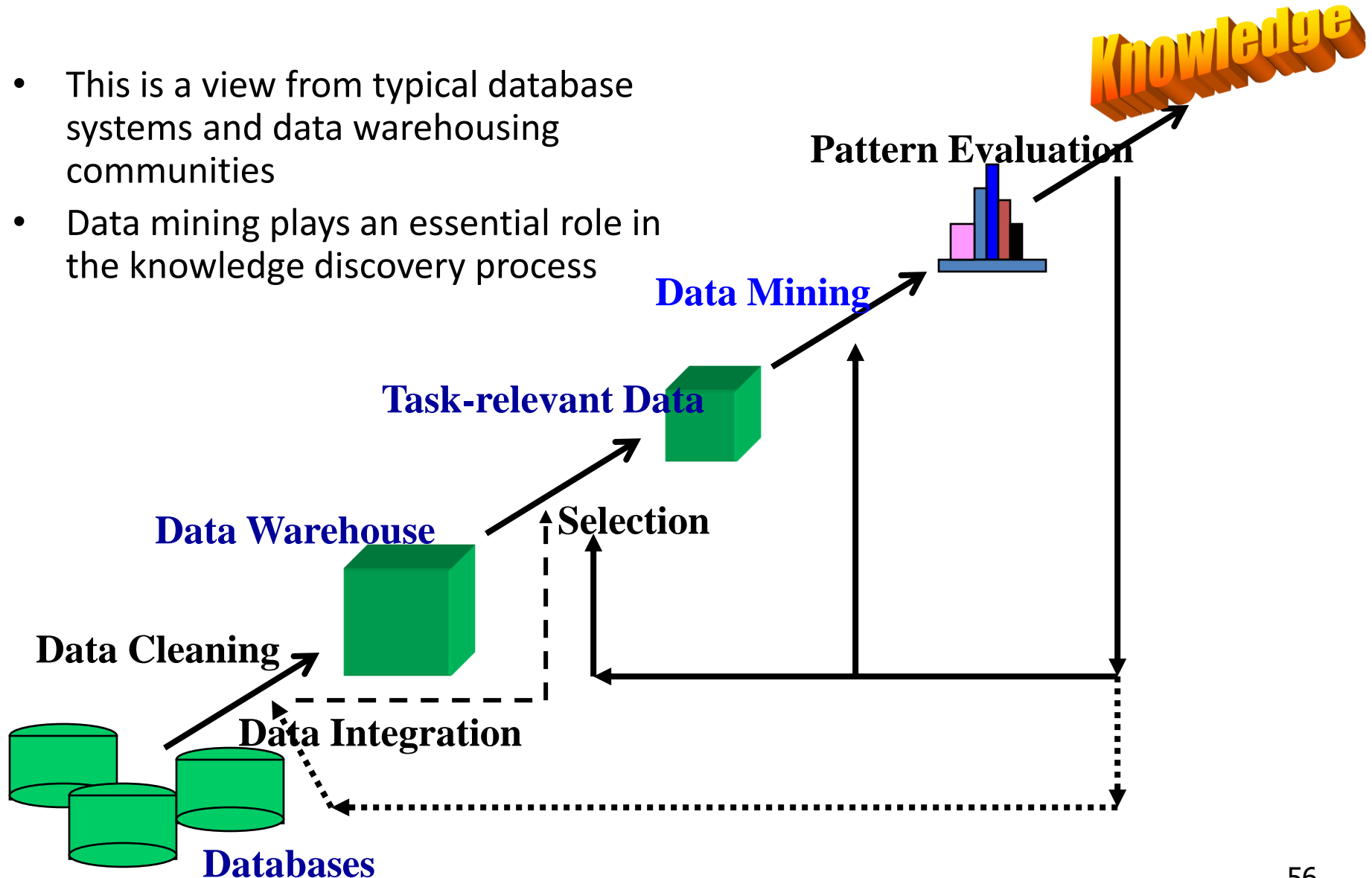


Interlude: About you!

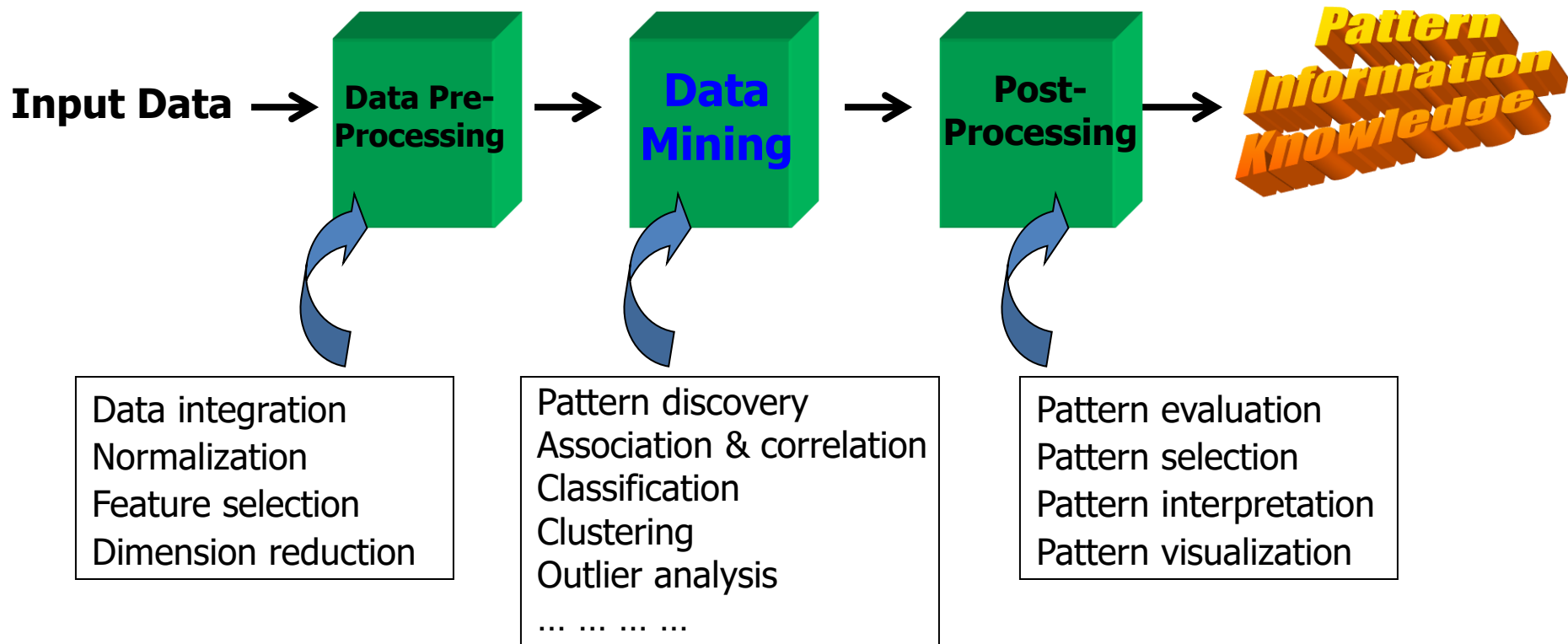
- Please introduce yourself to your ~~neighbor~~ breakout group buddy: your **name**, what you **hope to get out of this course**, and your **favorite thing to do in your spare time**
- You'll then introduce your ~~neighbor~~ breakout group buddy to the class!

Knowledge Discovery (KDD) Process

- This is a view from typical database systems and data warehousing communities
- Data mining plays an essential role in the knowledge discovery process

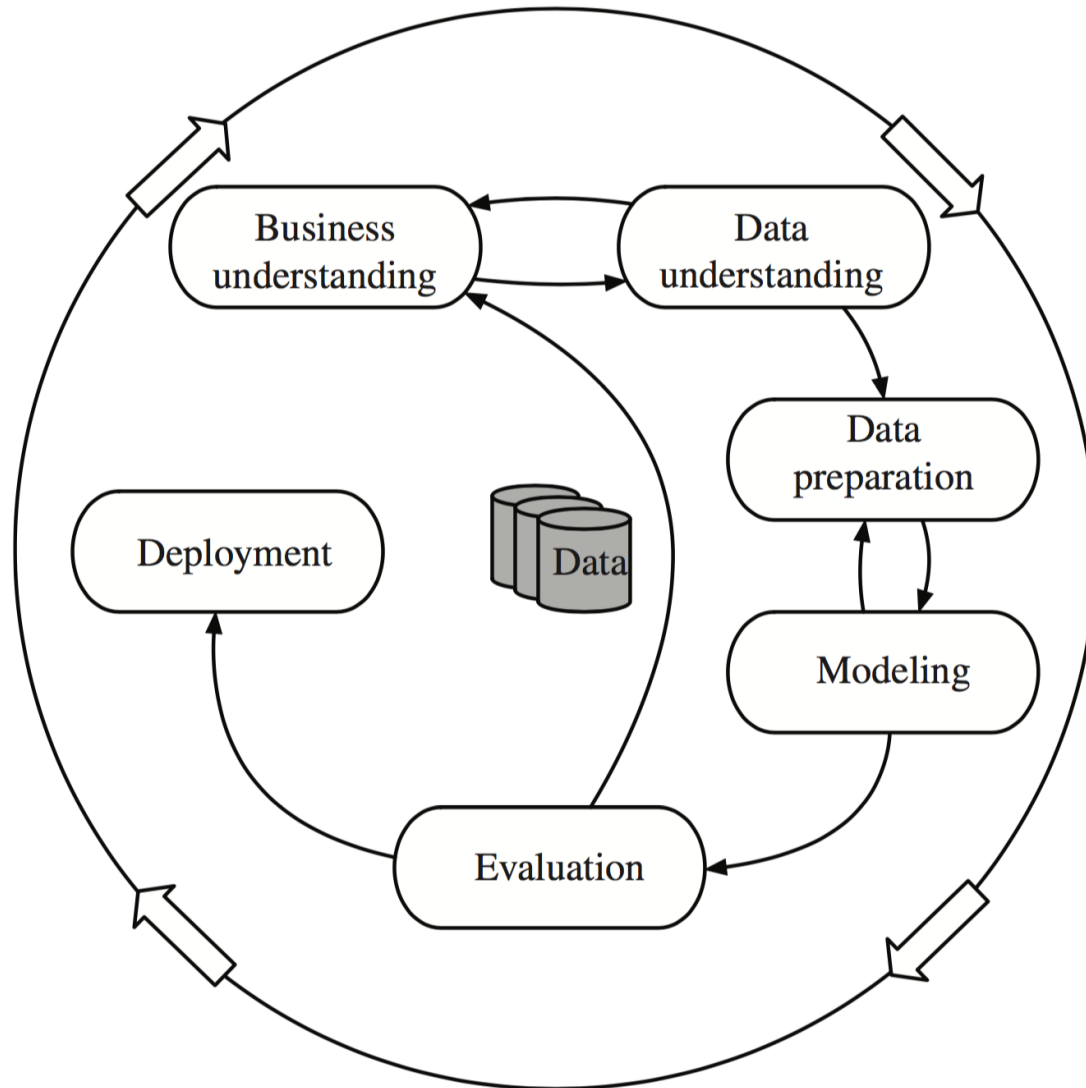


KDD Process: A Typical View from ML and Statistics



- This is a view from typical machine learning and statistics communities

In reality, data mining is an iterative process



Classification

- **Classification and label prediction**
 - Construct models (functions) based on some training examples
 - Describe and distinguish classes or concepts for future prediction
 - E.g., classify countries based on (climate), or classify cars based on (gas mileage)
 - Predict some unknown class labels
- **Typical methods**
 - Decision trees, naïve Bayesian classification, support vector machines, neural networks, rule-based classification, pattern-based classification, logistic regression, ...
- **Typical applications**
 - Credit card fraud detection, direct marketing, classifying stars, diseases, web-pages, ...

The weather problem

- Conditions for playing a certain game

Outlook	Temperature	Humidity	Windy	Play
Sunny	Hot	High	False	No
Sunny	Hot	High	True	No
Overcast	Hot	High	False	Yes
Rainy	Mild	Normal	False	Yes
...

Examples

Features / attributes

Class label
(what we want to predict)

The weather problem

- Conditions for playing a certain game

Outlook	Temperature	Humidity	Windy	Play
Sunny	Hot	High	False	No
Sunny	Hot	High	True	No
Overcast	Hot	High	False	Yes
Rainy	Mild	Normal	False	Yes
...

If outlook = sunny and humidity = high then play = no

If outlook = rainy and windy = true then play = no

If outlook = overcast then play = yes

If humidity = normal then play = yes

If none of the above then play = yes

Classifying iris flowers

	Sepal length	Sepal width	Petal length	Petal width	Type
1	5.1	3.5	1.4	0.2	Iris setosa
2	4.9	3.0	1.4	0.2	Iris setosa
...					
51	7.0	3.2	4.7	1.4	Iris versicolor
52	6.4	3.2	4.5	1.5	Iris versicolor
...					
101	6.3	3.3	6.0	2.5	Iris virginica
102	5.8	2.7	5.1	1.9	Iris virginica
...					

```
If petal length < 2.45 then Iris setosa
If petal width < 2.10 then Iris versicolor
...
```

Fielded applications

- The result of learning—or the learning method itself—is deployed in practical applications
 - Processing loan applications
 - Screening images for oil slicks
 - Marketing and sales...
 - Electricity supply forecasting
 - Diagnosis of machine faults
 - Separating crude oil and natural gas
 - Reducing banding in rotogravure printing
 - Finding appropriate technicians for telephone faults
 - Scientific applications: biology, astronomy, chemistry
 - Automatic selection of TV programs
 - Monitoring intensive care patients

Processing loan applications (American Express)

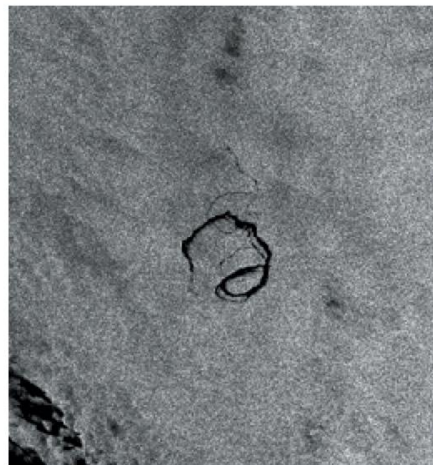
- Given: questionnaire with financial and personal information
- Question: should money be lent?
- Simple statistical method covers 90% of cases
- Borderline cases referred to loan officers
- But: 50% of accepted borderline cases defaulted!
- Solution: reject all borderline cases?
 - No! Borderline cases are most active customers

Enter machine learning

- 1000 training examples of borderline cases
- 20 attributes:
 - age
 - years with current employer
 - years at current address
 - years with the bank
 - other credit cards possessed,...
- Learned rules: correct on 70% of cases
 - human experts only 50%
- Rules could be used to explain decisions to customers

Screening images

- Given: radar satellite images of coastal waters
- Problem: detect oil slicks in those images
- Oil slicks appear as dark regions with changing size and shape
- Not easy: lookalike dark regions can be caused by weather conditions (e.g. high wind)
- Expensive process requiring highly trained personnel



Enter machine learning

- Extract dark regions from normalized image
- Attributes:
 - size of region
 - shape, area
 - intensity
 - sharpness and jaggedness of boundaries
 - proximity of other regions
 - info about background
- Constraints:
 - Few training examples—oil slicks are rare!
 - Unbalanced data: most dark regions aren't slicks
 - Regions from same image form a batch
 - Requirement: adjustable false-alarm rate

Marketing and sales

- Market basket analysis
 - Association techniques find groups of items that tend to occur together in a transaction (used to analyze checkout data)



Diaper



Beer

IS 733: Data Mining

Instructor:

Email:

Office Hours:

Dr. Nirmalya Roy

nroy@umbc.edu

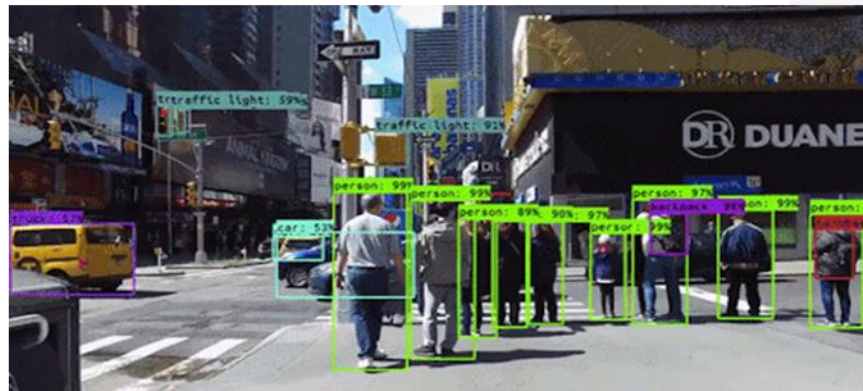
Tuesdays 9:00 – 10:00 am online (or by appointment)

Course website:

<https://mpsc.umbc.edu/courses/is-733-data-mining>

Poll everywhere:

PollEv.com/nirmalyaroy910



Eating



Bathing



Dressing



Transferring



Toileting



Walking or
moving around