`https://www.cs.ubc.ca/~fwood/CS340/`

Lecture II

Lecture roughly follows: `http://www-users.cs.umn.edu/~kumar/dmbook/dmslides/chap2_data.pdf`
Slides: `https://www.cs.ubc.ca/~fwood/CS340/lectures/L2.pdf`

Data Mining: Some Typical Steps

1. Learn about the application.
2. Identify data mining task.
3. Collect data.
4. Clean and preprocess the data.
5. Transform data or select useful subsets.
6. Choose data mining algorithm.
7. Data mining!
8. Evaluate, visualize, and interpret results.
9. Use results for profit or other goals.
   (often, you'll go through cycles of the above)

What is data?

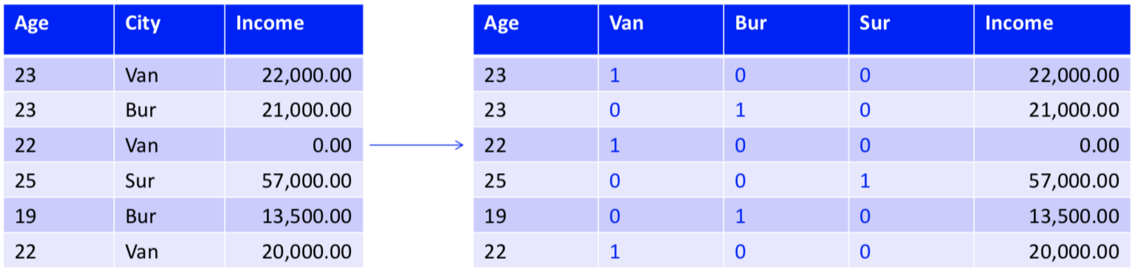
We'll define data as a collection of **examples**, and their **features**

Types of data

- Categorical features come from an unordered set
  - Binary: Job done or not?
  - Nominal: city
- Numerical features come from an ordered sets
  - Discrete counts: age
  - Ordinal: rating
  - Continuous/real-values: height

Converting to numerical features

It is very often more desirable to have real-values example representation.

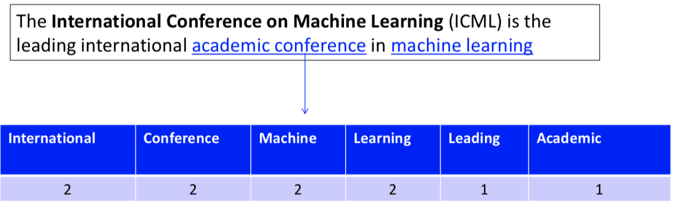


Age	City	Income
23	Van	22,000.00
23	Bur	21,000.00
22	Van	0.00
25	Sur	57,000.00
19	Bur	13,500.00
22	Van	20,000.00

Age	Van	Bur	Sur	Income
23	1	0	0	22,000.00
23	0	1	0	21,000.00
22	1	0	0	0.00
25	0	0	1	57,000.00
19	0	1	0	13,500.00
22	1	0	0	20,000.00

This is called **1 of k encoding**, and we can now interpret examples as points in space (E.g., first example is at (23,1,0,0,22000))

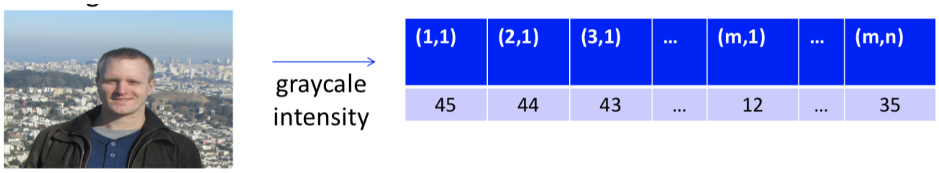
Approximating Text with Numerical Features



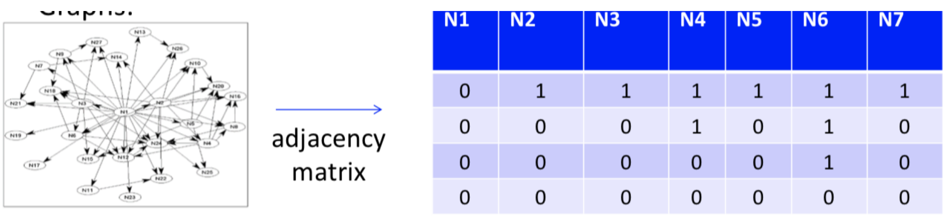
The **International Conference on Machine Learning** (ICML) is the leading international academic conference in machine learning

ICML	International	Conference	Machine	Learning	Leading	Academic
1	2	2	2	2	1	1

Approximating Images and Graphs



(1,1)	(2,1)	(3,1)	...	(m,1)	...	(m,n)
45	44	43	...	12	...	35



N1	N2	N3	N4	N5	N6	N7
0	1	1	1	1	1	1
0	0	0	1	0	1	0
0	0	0	0	0	1	0
0	0	0	0	0	0	0

Data Cleaning

ML+DM typically assume 'clean' data. Ways that data might not be 'clean' :

- Noise (e.g., distortion on phone).
- Outliers (e.g., data entry or instrument error).
- Missing values (no value available or not applicable)
- Duplicated data (repetitions, or different storage formats).

Any of these can lead to problems in analyses

- Want to fix these issues, if possible.
- Some ML methods are robust to these.
- Often, ML is the best way to detect/fix these.