Data

Data: Outline

- Types of data
  - data objects and attributes
  - unordered vs ordered data
- Quality of data
  - measurement errors
  - missing values, inconsistent values, duplicates
- Preprocessing
  - aggregation, sampling, feature extraction
  - discretization and variable transformations
- Similarity and dissimilarity
  - desirable properties
  - common measures
- Summary statistics and visualization

Standard data model

A data set can often be viewed as a *collection of data objects*:

- A *data object* (also called a record, data point, data vector, case, sample point, observation, entity) is described by a set of attributes
- An *attribute* (also called a variable, characteristic, field, feature, or dimension) describes one aspect of a data object

Example:

Student ID	Name	Date of Birth	Credits	Average	...
2328193	Matti	23.09.1985	123	3.6	...
9819234	Tuuli	12.03.1986	98	3.8	...
...	...	...	...	...	...

Here: each row is one data object, each column is an attribute.

Attributes: Number of values

- How many distinct values can an attribute take
  1. Binary attributes: only 2 possible states
     (e.g. on/off, pass/fail)
  2. Discrete attributes with $N < \infty$ states
     (e.g. course grade: 0, 1, 2, 3, 4, or 5)
  3. Discrete attributes with (countably) infinite states
     (e.g. counts: 0,1,2,3, . . .)
  4. Continuous attributes (uncountably infinite)
     (e.g. length or weight: $\in \mathbb{R}$)

- Asymmetric attributes
  1. Binary sparse attributes ('on' infrequent and significant, 'off' common and insignificant)
  2. Other attributes may also have 'special' states, need to take into account

Attributes: Measurement scales

- What operations are valid?
  1. Distinctness: $=$ vs $\neq$
  2. Order: $<$, $\leq$, $>$, $\geq$
  3. Addition and subtraction: $+$, $-$
  4. Multiplication and division: $*$ and $/$

- These define four types of attribute measurement scales:
  1. Nominal (categorical)
  2. Ordinal
  3. Interval
  4. Ratio

  Each measurement scale allows all the operations with number *smaller than or equal to* the number of the measurement scale. Example: 'Interval' attributes allow all operations except multiplication and division.

Attributes: Measurement scales – examples

Attribute Type	Description	Examples	Operations
Nominal	The values of a nominal attribute are just different names, i.e., nominal attributes provide only enough information to distinguish one object from another. (=, ≠)	zip codes, employee ID numbers, eye color, sex: {*male, female*}	mode, entropy, contingency correlation, χ^2 test
Ordinal	The values of an ordinal attribute provide enough information to order objects. (<, >)	hardness of minerals, {*good, better, best*}, grades, street numbers	median, percentiles, rank correlation, run tests, sign tests
Interval	For interval attributes, the differences between values are meaningful, i.e., a unit of measurement exists. (+, -)	calendar dates, temperature in Celsius or Fahrenheit	mean, standard deviation, Pearson's correlation, *t* and *F* tests
Ratio	For ratio variables, both differences and ratios are meaningful. (*, /)	temperature in Kelvin, monetary quantities, counts, age, mass, length, electrical current	geometric mean, harmonic mean, percent variation

General characteristics of data

- Number of data points vs dimensionality (number of attributes)
  1. Most traditional data analysis methods assume (many) more data points than dimensions
  2. Many interesting datasets have extremely high dimensions and few data objects
- Sparsity (relatively few non-zero values)
  1. Efficient storage and computation, for some methods
  2. May be important for modeling
- Resolution (spatial, temporal, . . .)
  1. How small details can the sensors reliably detect
  2. How large datasets can the methods handle

Examples – Record data

- Data ‘matrix’
- Sparse matrix representation?

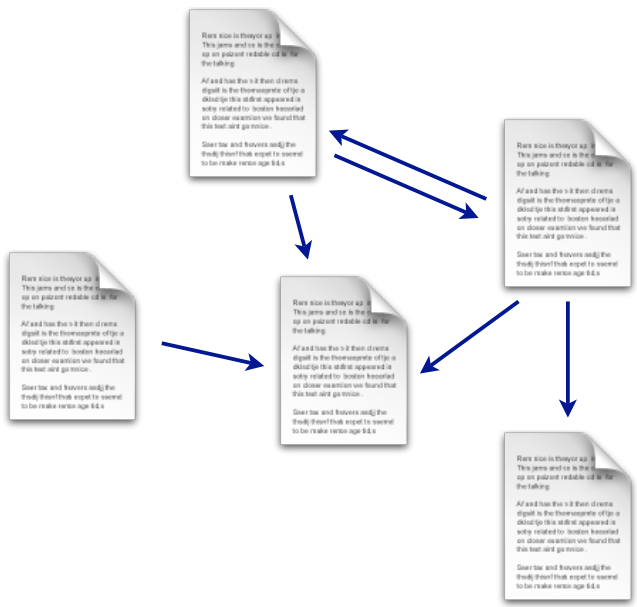
Tid	Refund	Marital Status	Taxable Income	Cheat
1	Yes	Single	125K	No
2	No	Married	100K	No
3	No	Single	70K	No
4	Yes	Married	120K	No
5	No	Divorced	95K	Yes
6	No	Married	60K	No
7	Yes	Divorced	220K	No
8	No	Single	85K	Yes
9	No	Married	75K	No
10	No	Single	90K	Yes

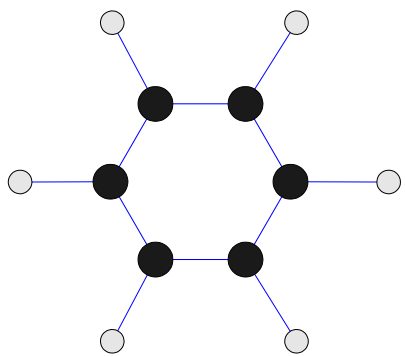
	team	coach	play	ball	score	game	win	lost	timeout	season
Document 1	3	0	5	0	2	6	0	2	0	2
Document 2	0	7	0	2	1	0	0	3	0	0
Document 3	0	1	0	0	1	2	2	0	3	0

TID	*Items*
1	Bread, Coke, Milk
2	Beer, Bread
3	Beer, Coke, Diaper, Milk
4	Beer, Bread, Diaper, Milk
5	Coke, Diaper, Milk

Examples – Graph data

- Relationships between objects
  - World wide web
  - Facebook users
  - Scientific papers
- Objects are graphs

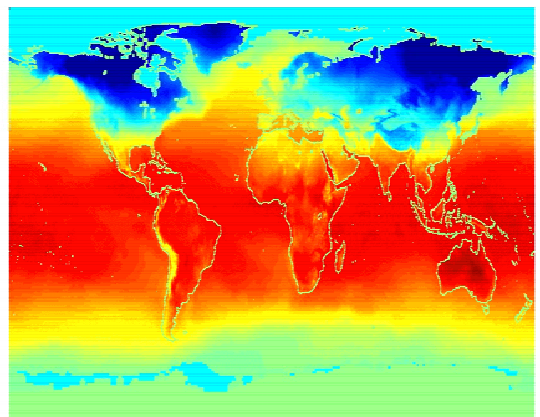




Examples – Ordered data

- Data objects with natural ordering
  - Sequential data (e.g. logins, credit card purchases, . . .)
  - Sequences (e.g. genome)
  - Time-series (e.g. climate science, financial data, . . .)
  - Spatial data (e.g. journey planner data, surface/sea temperatures, multi-spectral images)

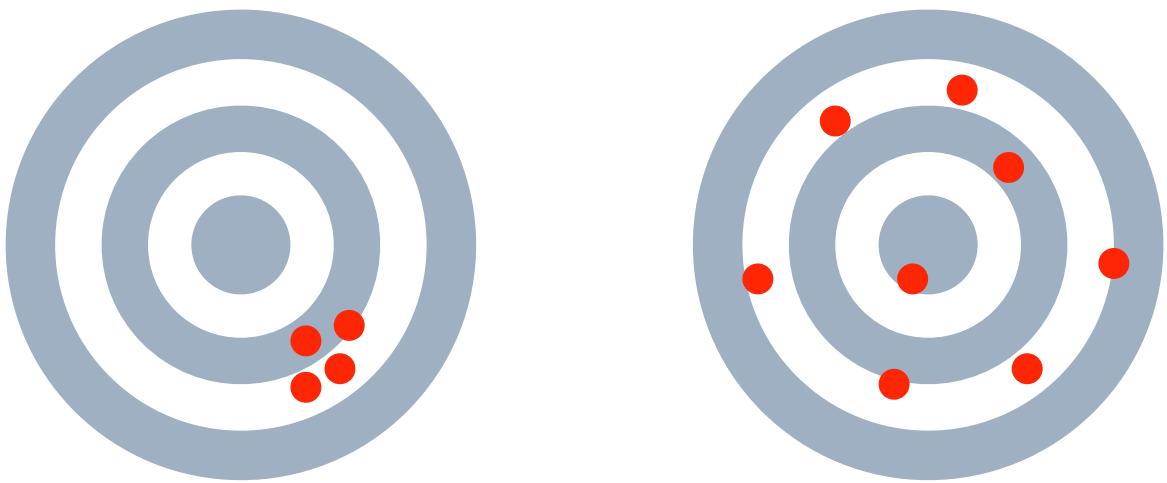
```
GGTTCCGCCTTCAGCCCCGCGCC
CGCAGGGCCCGCCCCGCGCCGTC
GAGAAGGGCCCGCCTGGCGGGCG
GGGGGAGGCGGGGCCGCCCGAGC
CCAACCGAGTCCGACCAGGTGCC
CCCTCTGCTCGGCCTAGACCTGA
GCTCATTAGGCGGCAGCGGACAG
GCCAAGTAGAACACGCGAAGCGC
TGGGCTGCCTGCTGCGACCAGGG
```



Data quality

- Measurement error ('noise')
  - Continuous data: Gaussian, Student's t distribution, ...
  - Binary data: Bernoulli, ...
- Precision and bias
  - Precision: closeness of repeated measurements to each other
  - Bias: systematic deviation from the true underlying value

Left: high precision, high bias. Right: low precision, low bias.



- Outliers i.e. 'anomalous' objects:

  (Objects that are very different from the others)

  - Noise?
  - Data collection error?
  - Legitimate, interesting objects?

- Missing values:

  (Some attribute values are missing for some data objects)

  - Missing at random? Need to model the process?
  - Just eliminating such data objects or attributes?
  - Estimating and imputing missing values?
  - Ignoring or explicitly taking them into account?

- Duplicate data, inconsistent data

- Timeliness, relevance, application-specific prior knowledge

Data preprocessing

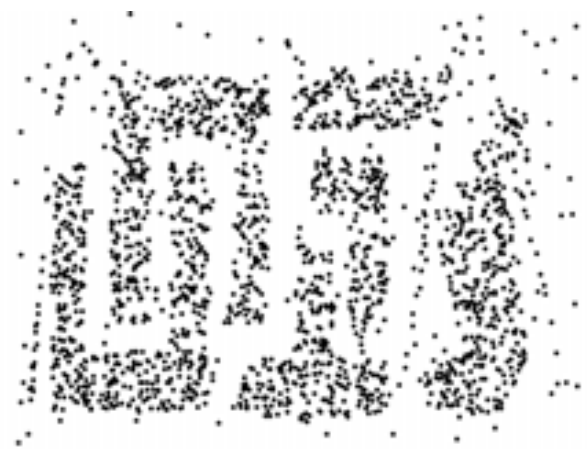
- Aggregation
  - $\Rightarrow$ fewer, less noisy, data points
    - Example: monthly analysis of daily (or hourly) data
    - Loss of resolution

- Sampling (with/without replacement, stratified or not)
  - $\Rightarrow$ fewer data points
    - Visualization, applying computationally demanding data analysis procedures
    - Loss of precision of data statistics

Data preprocessing

- Aggregation
  - $\Rightarrow$ fewer, less noisy, data points
    - Example: monthly analysis of daily (or hourly) data
    - Loss of resolution

- Sampling (with/without replacement, stratified or not)
  - $\Rightarrow$ fewer data points
    - Visualization, applying computationally demanding data analysis procedures
    - Loss of precision of data statistics

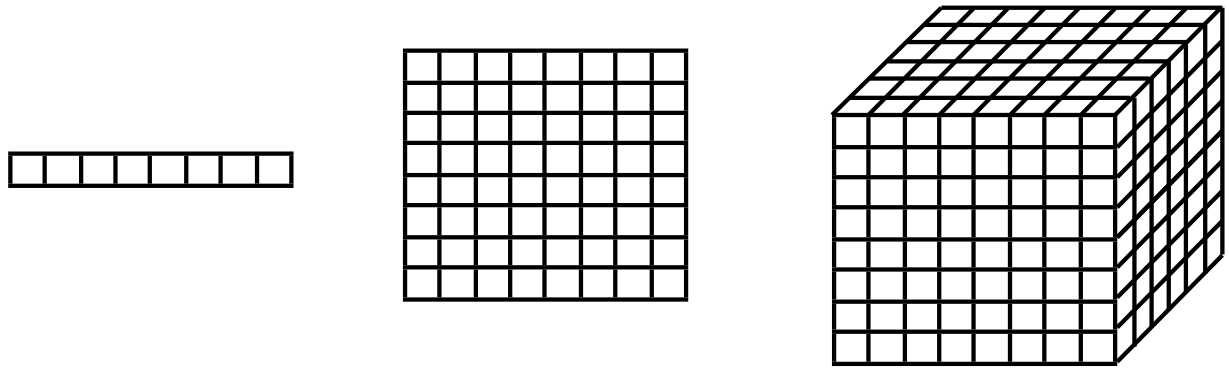


Data preprocessing

- Aggregation
  - $\Rightarrow$ fewer, less noisy, data points
    - Example: monthly analysis of daily (or hourly) data
    - Loss of resolution

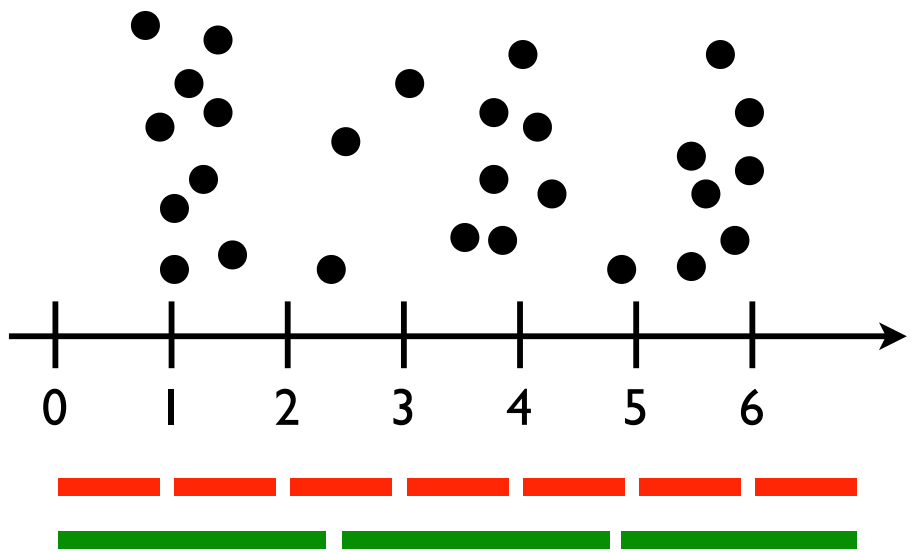
- Sampling (with/without replacement, stratified or not)
  - $\Rightarrow$ fewer data points
    - Visualization, applying computationally demanding data analysis procedures
    - Loss of precision of data statistics

The ‘curse of dimensionality’: As the dimensionality grows, the data becomes increasingly sparse in the space (e.g. n dimensions of binary variables has 2^n joint states)



- Feature subset selection
  (selecting only some of the attributes, manually or automatically)
- Feature extraction
  (computing new ‘attributes’ to replace existing ones, e.g. image color/structure features)
- Dimensionality reduction
  (principal component analysis, other unsupervised methods)

- Discretization of continuous attributes (required for some machine learning algorithms)
  - e.g. $\mathbb{R}$ is divided into $(-\infty, x_1], (x_1, x_2], (x_2, x_3], (x_3, \infty)$
  - Unsupervised or supervised
  - Ideally: use application-specific knowledge



- Binarization
  - Some algorithms *require* binary attributes
  - How NOT to represent nominal variables:

Categorical value	Integer value	x_1	x_2	x_3
Toyota	0	0	0	0
Volkswagen	1	0	0	1
Chevrolet	2	0	1	0
Saab	3	0	1	1
Volvo	4	1	0	0

- Binarization
  - Some algorithms *require* binary attributes
  - How NOT to represent nominal variables:

Categorical value	Integer value	x_1	x_2	x_3
Toyota	0	0	0	0
Volkswagen	1	0	0	1
Chevrolet	2	0	1	0
Saab	3	0	1	1
Volvo	4	1	0	0

  - How better to represent nominal variables:

Categorical value	Integer value	x_1	x_2	x_3	x_4	x_5
Toyota	0	1	0	0	0	0
Volkswagen	1	0	1	0	0	0
Chevrolet	2	0	0	1	0	0
Saab	3	0	0	0	1	0
Volvo	4	0	0	0	0	1

  - Useful even when binary variables are not strictly required!

- Variable transformations
  - Simple functions (e.g. $\log(x)$ often used for strictly positive variables)
  - Ideally: use application-specific knowledge
  - Normalization / standardization:

$$x' = \frac{x - \bar{x}}{s_x}, \tag{1}$$

where $\bar{x}$ is the mean of the attribute values and s_x is the standard deviation. This yields x' with zero mean and a standard deviation of 1.

Useful to ensure that the scale (units) of attribute values do not affect the results.

Similarity and dissimilarity

- Many machine learning algorithms use measures of similarity or dissimilarity between data objects
- Examples:
  - Handwritten letters
  - Segments of DNA
  - Text documents

TCGATTGC ATCCTGTG ACCTGTCG

"Parliament overwhelmingly approved amendments to the Firearms Act on Wednesday. The new law requires physicians to inform authorities of individuals they consider unfit to own guns. It also increases the age for handgun ownership from 18 to 20."

"Parliament's Committee for Constitutional Law says that persons applying for handgun licences should not be required to join a gun club in the future. Government is proposing that handgun users be part of a gun association for at least two years."

"The cabinet on Wednesday will be looking at a controversial package of proposed changes in the curriculum in the nation's comprehensive schools. The most divisive issue is expected to be the question of expanded language teaching."

- Similarity: s
  - Numerical measure of the degree to which two objects are *alike*
  - Higher for objects that are alike
  - Typically between 0 (no similarity) and 1 (completely similar)
- Dissimilarity: d
  - Numerical measure of the degree to which two objects are *different*
  - Higher for objects that are different
  - Typically between 0 (no difference) and ∞ (completely different)
- Transformations
  - Converting from one to the other
- Use similarity or dissimilarity measures?
  - Method-specific

- Proximity between *attribute values*

Attribute Type	Dissimilarity	Similarity
Nominal	$d = \begin{cases} 0 & \text{if } p = q \\ 1 & \text{if } p \neq q \end{cases}$	$s = \begin{cases} 1 & \text{if } p = q \\ 0 & \text{if } p \neq q \end{cases}$
Ordinal	$d = \frac{\lvert p-q \rvert}{n-1}$ (values mapped to integers 0 to $n-1$, where n is the number of values)	$s = 1 - \frac{\lvert p-q \rvert}{n-1}$
Interval or Ratio	$d = \lvert p - q \rvert$	$s = -d$, $s = \frac{1}{1+d}$ or $s = 1 - \frac{d - min\_d}{max\_d - min\_d}$

Dissimilarities between *objects*:

- Minkowski distance for vectors in R^n

$$d(\mathbf{x}, \mathbf{y}) = \left(\sum_{k=1}^{n} |x_k - y_k|^r \right)^{1/r} \tag{2}$$

  - $r = 1$: City-block, hamming
  - $r = 2$: Euclidean
  - $r = \infty$: Supremum

- Distance metric $d(\mathbf{x}, \mathbf{y})$ properties:
  - Positivity: $d(\mathbf{x}, \mathbf{y}) \geq 0$ for all $\mathbf{x}$ and $\mathbf{y}$, with $d(\mathbf{x}, \mathbf{y}) = 0$ if and only if $\mathbf{x} = \mathbf{y}$
  - Symmetry: $d(\mathbf{x}, \mathbf{y}) = d(\mathbf{y}, \mathbf{x})$ for all $\mathbf{x}$ and $\mathbf{y}$
  - Triangle inequality: $d(\mathbf{x}, \mathbf{z}) \leq d(\mathbf{x}, \mathbf{y}) + d(\mathbf{y}, \mathbf{z})$ for all $\mathbf{x}$, $\mathbf{y}$, $\mathbf{z}$

- Most dissimilarity measures satisfy positivity. Depending on the application, it may or may not be important to satisfy symmetry and the triangle inequality.

Similarities between objects:

- Typical properties of $s(\mathbf{x}, \mathbf{y})$:
  - $0 \leq s(\mathbf{x}, \mathbf{y}) \leq 1$, with $s(\mathbf{x}, \mathbf{y}) = 1$ if only if $\mathbf{x} = \mathbf{y}$
  - $s(\mathbf{x}, \mathbf{y}) = s(\mathbf{y}, \mathbf{x})$ for all $\mathbf{x}$ and $\mathbf{y}$ (symmetry)

- Examples (binary vectors):
  (f_{ab} is the number of attributes for which $\mathbf{x} = a$ and $\mathbf{y} = b$)
  - Simple matching coefficient

$$\text{SMC} = \frac{f_{11} + f_{00}}{f_{11} + f_{00} + f_{01} + f_{10}} \tag{3}$$

  - Jaccard coefficient

$$J = \frac{f_{11}}{f_{11} + f_{01} + f_{10}} \tag{4}$$

- Examples (arbitrary vectors):
  - Cosine similarity (angle between vectors)

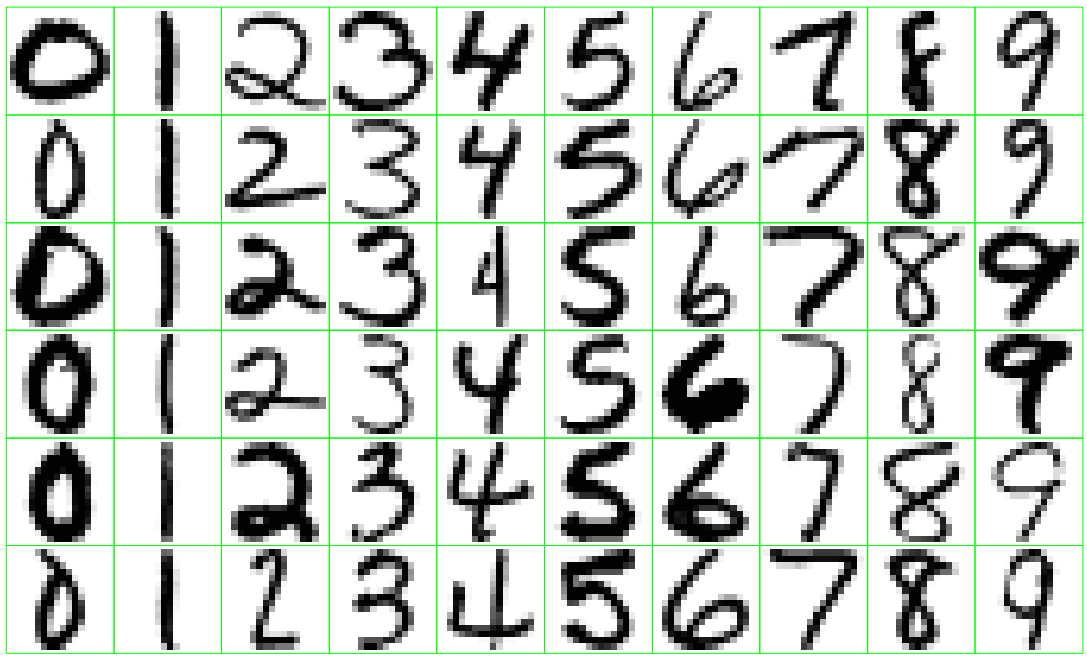
$$\cos(\mathbf{x}, \mathbf{y}) = \frac{\mathbf{x}^T \mathbf{y}}{||\mathbf{x}|| \, ||\mathbf{y}||} \tag{5}$$

  - (Pearson's) linear correlation coefficient (goes from -1 to 1)
  - (Spearman's) rank correlation coefficient (also -1 to 1)

- Issues to consider:
  - Proximity measures ideally application-specific
  - Use well-known, established measures for the particular type of objects
  - Standardization to avoid arbitrary units affecting results
  - Try a number of different possibilities and evaluate the results

Course dataset 1

- Handwritten digits:
  http://yann.lecun.com/exdb/mnist/index.html



- each image 28×28 pixels, each pixel a gray value between 0 and 255.

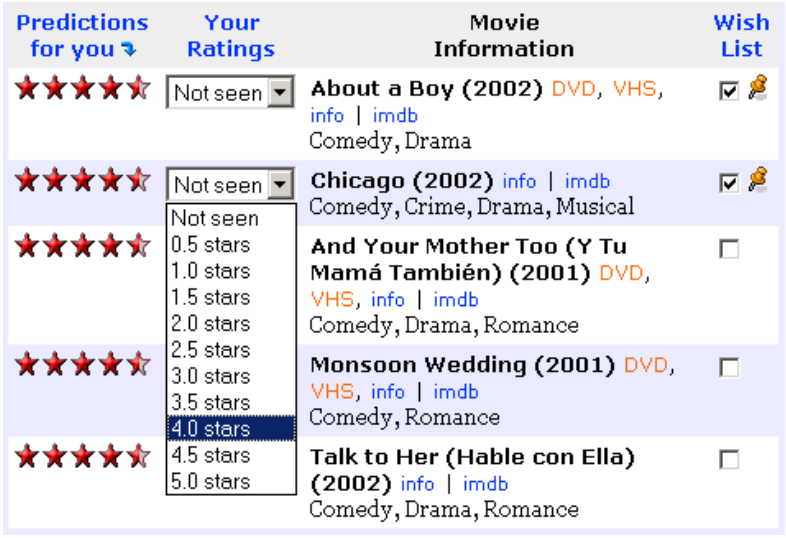
Course dataset 2

- Collection of texts:
  http://people.csail.mit.edu/jrennie/20Newsgroups/
  - Messages from 20 different usenet newsgroups
  - Preprocessed into bag-of-words representation

comp.graphics comp.os.ms-windows.misc comp.sys.ibm.pc.hardware comp.sys.mac.hardware comp.windows.x	rec.autos rec.motorcycles rec.sport.baseball rec.sport.hockey	sci.crypt sci.electronics sci.med sci.space
misc.forsale	talk.politics.misc talk.politics.guns talk.politics.mideast	talk.religion.misc alt.atheism soc.religion.christian

Course dataset 3

- Movielens:
  http://movielens.umn.edu/
  http://www.grouplens.org/node/73



	Seven	Fargo	Aliens	Leon	Avatar
Linda		4		5	5	1		2
			3		4	3		
Jack	1			4		1	5	1
Bill					4	1		
Lucy			2	1	1		5	
John	1				1	4		5
		4				5		5
	2		3				3	