CLASSIFICATION

SYRACUSE UNIVERSITY
School of Information Studies

WHAT IS DATA MINING?

"Non-trivial extraction of implicit, previously unknown and potentially useful information from data." – Gregory Piatetsky-Shapiro, founder of kdnuggets.com

In support of decision-making

TYPICAL DATA MINING TASKS

Classification

Clustering

Association rule mining

WHAT IS CLASSIFICATION?

Given some predefined categories, assign objects to one or more categories.

Is this fruit apple, orange, strawberry, or pear?

?

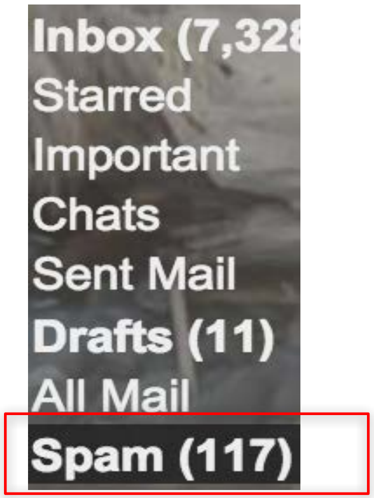
CLASSIFICATION VS. REGRESSION

The prediction output is different.

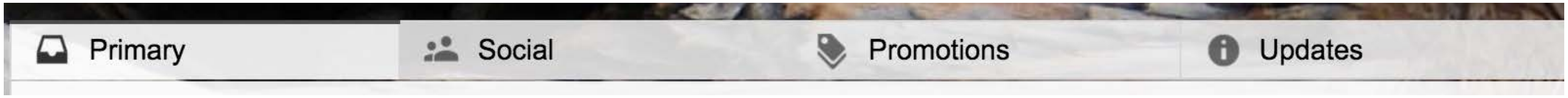
- Classification outputs categorical decisions (e.g., spam or regular e-mails).
  - Uses machine-learning techniques
  - Taught in this class
- Regression outputs numeric values (e.g., stock price, temperature).
  - Uses regression techniques
  - Taught in statistics class

CLASSIFICATION IN EVERYDAY LIFE

Gmail identifies spam e-mails from regular ones.



Gmail categorizes regular e-mails into Primary, Social, Promotion, Updates, etc.



CLASSIFICATION

SYRACUSE UNIVERSITY
School of Information Studies

HOW TO MODEL A CLASSIFICATION PROBLEM

Bank loan approval:

What is the decision to make?

What is the unit of analysis?

What attributes are helpful for classification?

WHAT IS THE DECISION TO MAKE?

What is the decision to make?

“Approve” or “deny” a loan application

The decision is saved in the target attribute

WHAT IS THE UNIT OF ANALYSIS?

The unit of analysis means an example in your data set. A classification decision will be made for each example.

For bank loan classification, an individual application is an example, which will be either approved or denied.

An individual person may not be good unit of analysis, because one person may submit multiple applications over time, and each deserves a decision.

WHAT ATTRIBUTES ARE HELPFUL FOR CLASSIFICATION?

What attributes are useful for classification?

Potentially useful attributes:

E.g., applicant's age, job title, income, credit score, amount requested

Some might be more useful than others.

Classification algorithms can rank the attributes by their contribution to classification.

SAMPLE DATA FOR BANK LOAN CLASSIFICATION

Application	Job Title	Income	Credit Score	Decision
1	teacher	50K	700	approve
2	manager	60K	300	deny

Each row is an example.
Each column is an attribute.
The last attribute is the decision to make, the target attribute.

HOW TO TEACH A COMPUTER TO CLASSIFY?

Step 1: Collect training data.

E.g., a collection of past loan decisions made by financial experts

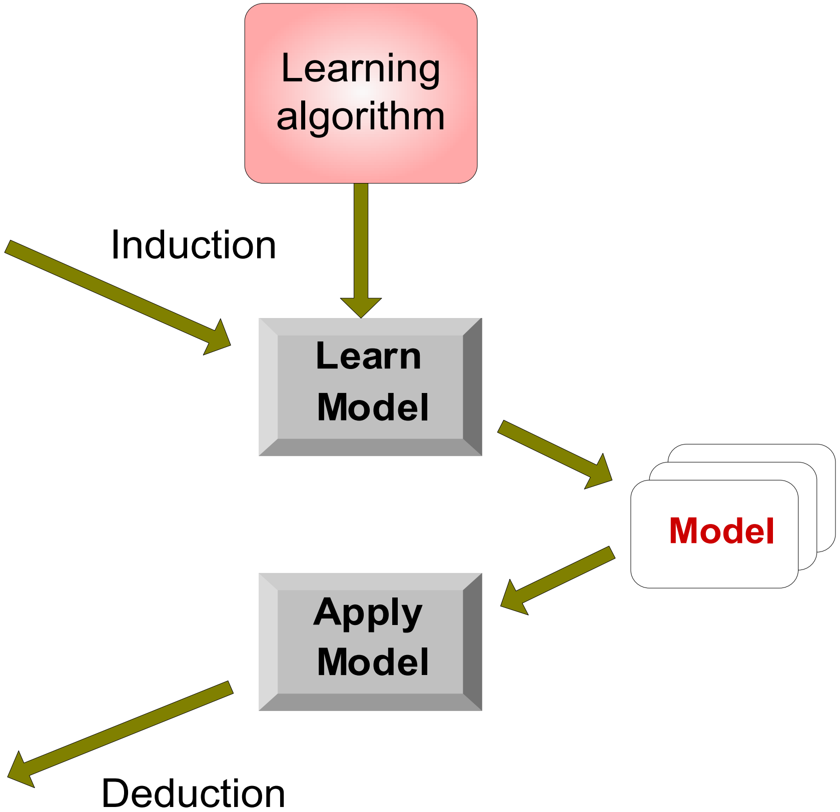
Step 2: Use a machine-learning algorithm to build a classifier based on relevant variables.

Step 3: Apply the classifier to new data.

ILLUSTRATING CLASSIFICATION TASK

Tid	Attrib1	Attrib2	Attrib3	Class
1	Yes	Large	125K	No
2	No	Medium	100K	No
3	No	Small	70K	No
4	Yes	Medium	120K	No
5	No	Large	95K	Yes
6	No	Medium	60K	No
7	Yes	Large	220K	No
8	No	Small	85K	Yes
9	No	Medium	75K	No
10	No	Small	90K	Yes

Training Set



Tid	Attrib1	Attrib2	Attrib3	Class
11	No	Small	55K	?
12	Yes	Medium	80K	?
13	Yes	Large	110K	?
14	No	Small	95K	?
15	No	Large	67K	?

Test Set

Two Steps

ARE WE DONE?

No. Prediction models need maintenance.

What if an approved loan defected?

- Add defective loans to the “deny” pool and retrain the model

What if a denied application was approved by another bank and performed well?

- No good solution without data sharing

ANOMALY DETECTION

Detect significant deviations from normal behavior

Applications:
- Credit card fraud detection
- Network intrusion detection

Can be modeled as classification problem
- Classify each transaction as fraud or not

CLUSTERING

SYRACUSE UNIVERSITY
School of Information Studies

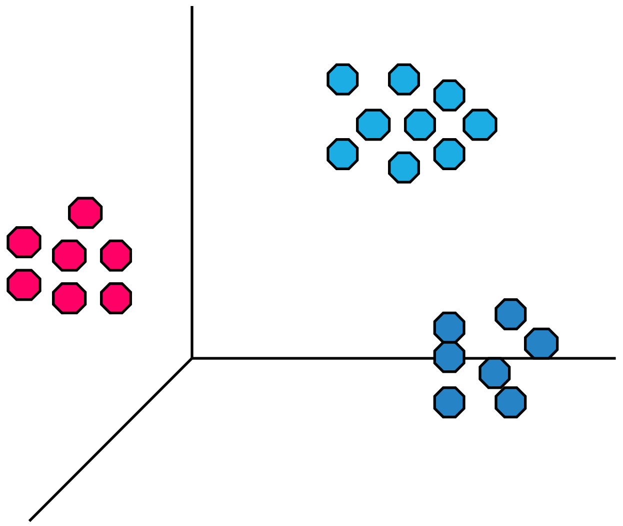
CLUSTERING

Given a set of data points, each having a set of attributes, and a similarity measure among them, find clusters such that:

Data points in one cluster are more similar to one another.

Data points in separate clusters are less similar to one another.

Intracluster distances are minimized.



Intercluster distances are maximized.

SIMILARITY MEASURES FOR CLUSTERING

Similarity measures:

Euclidean distance, if attributes are continuous

Other problem-specific measures

CLUSTERING APPLICATION 1

Market-customer segmentation

Goal: To find the “subgroups” among a large customer base

Approach:

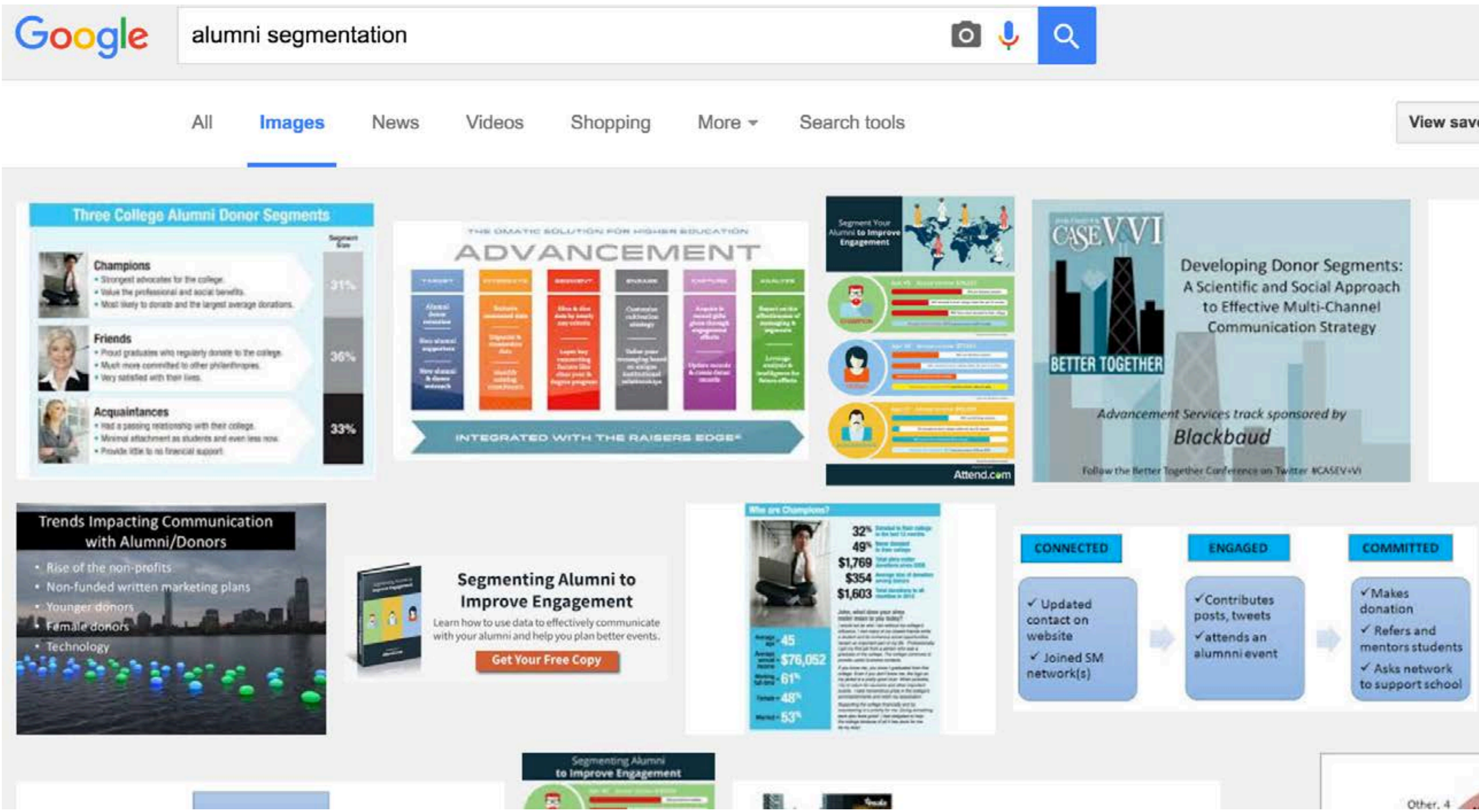
- Collect some “attributes” about the customers – their age, income, favorite brands, etc.
- Calculate the “similarity” between the customers.
- “Cluster” similar customers together.

WHAT DOES A CLUSTER MEAN?

Although a clustering algorithm can group or cluster similar customers together, it does not tell us what each cluster means.

Data analysts need to understand and interpret the meaning of clusters; e.g., one cluster may be interpreted as “tree huggers,” “money savers,” or “luxury fans.”

WHAT DOES A CLUSTER MEAN?



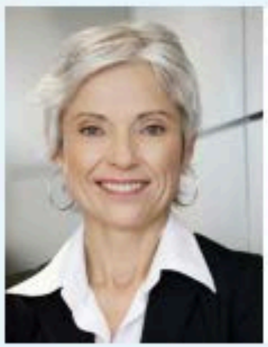
Three College Alumni Donor Segments

Segment Size

Champions

- Strongest advocates for the college.
- Value the professional and social benefits.
- Most likely to donate and the largest average donations.

31%



Friends

- Proud graduates who regularly donate to the college.
- Much more committed to other philanthropies.
- Very satisfied with their lives.

36%

Acquaintances

- Had a passing relationship with their college.
- Minimal attachment as students and even less now.
- Provide little to no financial support.

33%

Who are Champions?

Average age = 45

Average annual income = $76,052

Working full-time = 61%

Female = 48%

Married = 53%

32% Donated to their college in the last 12 months

49% Never donated to their college

$1,769 Total alma mater donations since 2006

$354 Average size of donation among donors

$1,603 Total donations to all charities in 2010

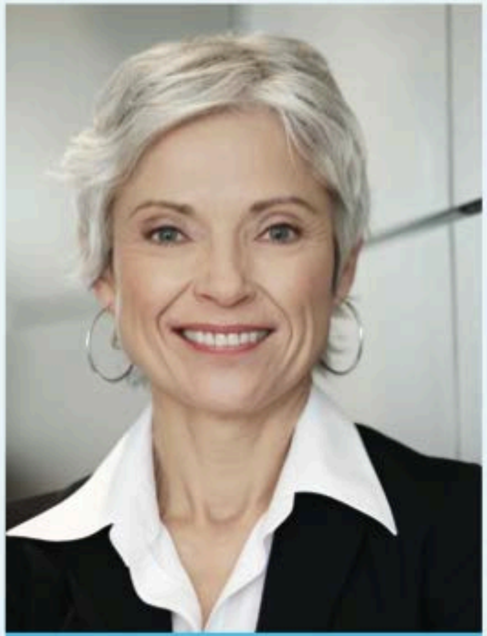
John, what does your alma mater mean to you today?

I would not be who I am without my college's influence. I met many of my closest friends while a student and its numerous social opportunities remain an important part of my life. Professionally, I got my first job from a person who was a graduate of the college. The college continues to provide useful business contacts.

If you know me, you know I graduated from this college. Even if you don't know me, the logo on my jacket is a pretty good clue! When possible, I try to return for reunions and other important events. I take tremendous pride in the college's accomplishments and relish my association.

Supporting the college financially and by volunteering is a priority for me. Giving something back also feels good! I feel obligated to help the college because of all it has done for me. Its my duty!

Who are Friends?



Average age = 56

Average annual income = $77,601

Working full-time = 40%

Female = 61%

Married = 68%

24% Donated to their college in the last 12 months

56% Never donated to their college

$985 Total alma mater donations since 2006

$197 Average size of donation among donors

$2,750 Total donations to all charities in 2010

Susan, what does your alma mater mean to you today?

I am very proud of my college! Academically, it has always been a great school and I am fortunate to have attended. I am not the type of alumnus who wears college sweatshirts or puts decals on my car, but I certainly enjoy talking about the college when somebody asks. I rarely get back to campus, so the alumni magazine is a nice way to keep in touch.

I am very happy with my life and grateful to the college. I have no regrets for having attended my college. It is a great school. Nonetheless, I have not been involved with the college since my graduation. I am really not sure why, except my other interests take up all my time.

Yes, I regularly make modest donations to the college. It just seems like the right thing to do -- more of habit than a passion. Organizations providing food and health services are in greater need of my money and time.

Who are Acquaintances?

Average age = 51

Average annual income = $69,935

Working full-time = 49%

Female = 59%

Married = 54%

5% Donated to their college in the last 12 months

86% Never donated to their college

$226 Total alma mater donations since 2006

$45 Average size of donation among donors

$1,300 Total donations to all charities in 2010

Kate, what does your alma mater mean to you today?

Gosh, I really haven't thought much about my college since graduating 30 years ago. I wasn't a flag waving student and certainly have not become one as an alumnus! I didn't even attend commencement for my graduation. Honestly, I don't understand why people have strong feelings toward colleges. For me, college was a place where I earned my degree – no more, no less. I paid dearly for that degree, so why am I supposed to be grateful to them?

Yes, the college contacts me each year requesting a donation. I just say no and wait for their call next year when I say no again. I don't even read the alumni magazine they send. My annual refusal to give them money is my only contact with the college. Why do they keep calling? They should know by now that I am not going to give them anything. Calling me is a waste of their money and my time. I don't get it.

CLUSTERING: APPLICATION 2

Document clustering

Goal: To find groups of documents that are similar to each other based on the important terms appearing in them

Approach: To identify frequently occurring terms in each document, form a similarity measure based on the frequencies of different terms; use it to cluster.

Gain: Search engines can organize search results by document clusters.

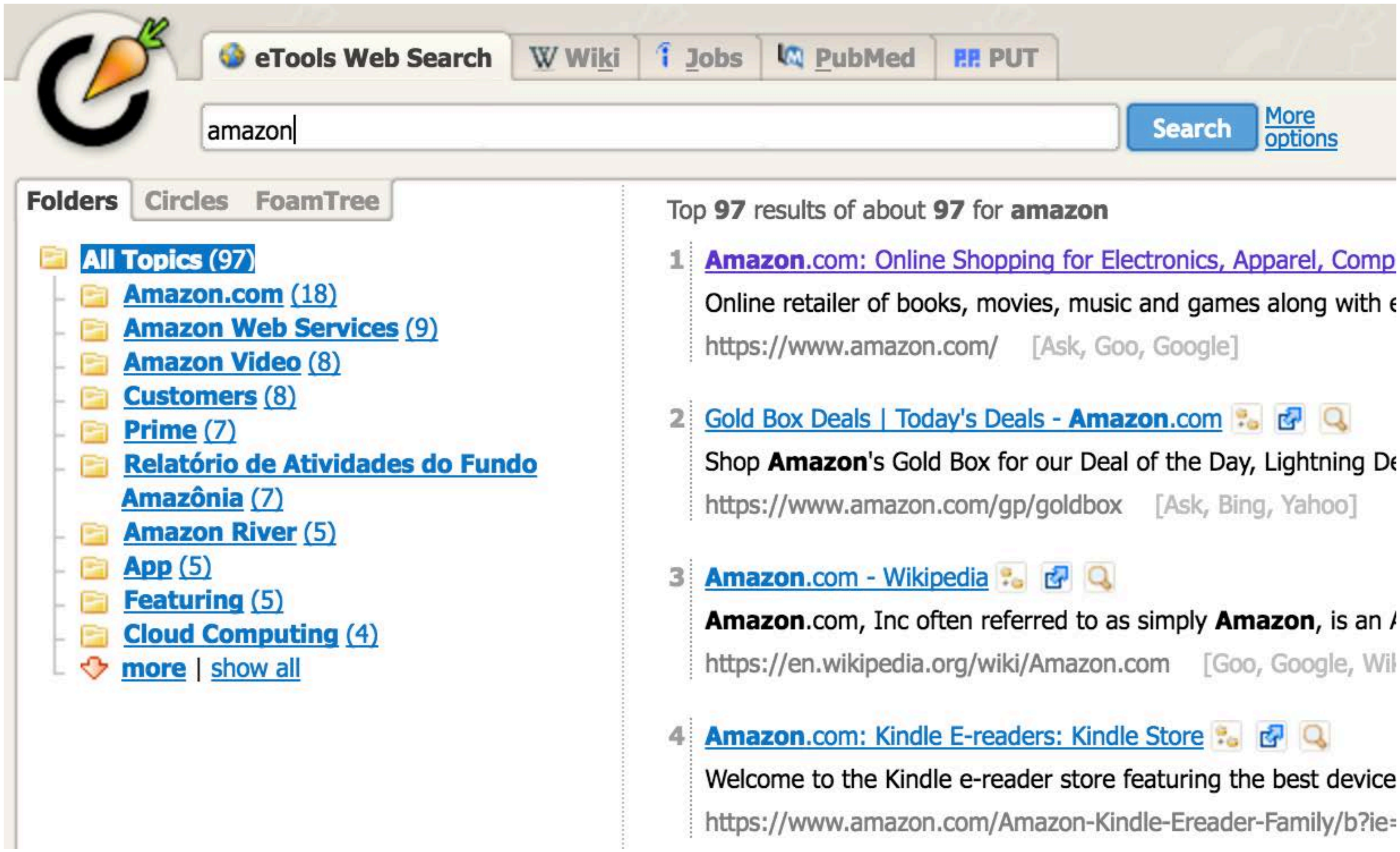
SEARCH ENGINE BASED ON DOCUMENT CLUSTERING

http://search.carrot2.org/stable/search

Search "Amazon" and see the returned results organized into clusters with labels.

The pioneer clustering-based search engine Vivisimo was acquired by IBM in 2012.

CARROT2 SEARCH ENGINE BASED ON DOCUMENT CLUSTERING



CLASSIFICATION VS. CLUSTERING

Classification: Supervised learning

Clustering: Unsupervised learning

- No training data
- No predefined target variable
- More suitable for exploratory analysis for data sets that we don't know much about

CAN A CLUSTERING MODEL DO CLASSIFICATION?

Yes, sometimes serves as the first step, to define the categories

E.g., after a customer base is clustered into three clusters, we can examine these clusters and “label” them in “tree huggers,” “savers,” and “luxury fans” categories.

Given a new customer, we can predict which cluster this customer belongs to, based on his or her similarity with the members in each cluster.

CLUSTERING FOR CLASSIFICATION

Clustering points: 3,204 Articles of *Los Angeles Times*

Similarity measure: How many words are common in these documents (after some word filtering)

Category	Total Articles	Correctly Placed
Financial	555	364
Foreign	341	260
National	273	36
Metro	943	746
Sports	738	573
Entertainment	354	278

ASSOCIATION RULE MINING

SYRACUSE UNIVERSITY
School of Information Studies

ASSOCIATION RULE (AR) MINING

Given a set of transactions, find:

- Items that co-occur frequently
- Rules such as "if a customer bought *x*, he or she would buy *y*, too"

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

Strong rules

- {Milk} --> {Coke}
- {Diaper, Milk} --> {Beer}

FREQUENT ITEMSETS

Itemset:

A collection of one or more items

k-itemset contains k items

1-itemset:

{A}:3, {B}:3, {C}:2, {D}:4, {E}:3, {F}:2

2-itemset:

{A,B}:1, {A,D}:3

3-itemset:

{A,B,C}:0, {B,E,F}:2

Transaction ID	Items Bought
10	A, B, D
20	A, C, D
30	A, D, E
40	B, E, F
50	B, C, D, E, F

ASSOCIATION RULES

Association rule:

An implication of the form X → Y, where X and Y are itemsets

E.g., {E, F} → {B}

LHS: Left-Hand Side →

Example Rules:

{B, E} → {F}

{E, F} → {B}

{B, F} → {E}

{B} → {E, F}

{E} → {B, F}

{F} → {B, E}

← RHS: Right-Hand Side

Transaction ID	Items Bought
10	A, B, D
20	A, C, D
30	A, D, E
40	B, E, F
50	B, C, D, E, F

AR MINING APPLICATION 1: MARKETING AND SALES PROMOTION

SYRACUSE UNIVERSITY
School of Information Studies

AR MINING APPLICATION 2: SHELF MANAGEMENT

Supermarket shelf management

Goal: To identify items that are bought together by sufficiently many customers

Approach: Process the point-of-sale data collected with barcode scanners to find dependencies among items.

A classic rule:

If a customer buys diapers and milk, then he is very likely to buy beer.

So don’t be surprised if you find six-packs stacked next to diapers!

AR MINING APPLICATION 3: INVENTORY MANAGEMENT

Inventory management

Goal: A consumer-appliance repair company wants to anticipate the nature of repairs on its consumer products and keep the service vehicles equipped with right parts to reduce the number of visits to consumer households.

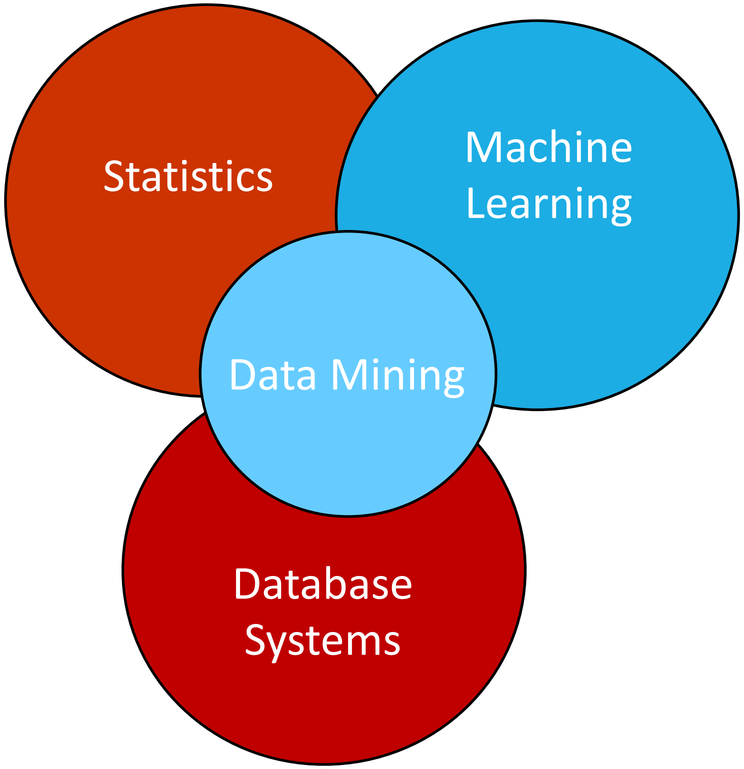
Approach: Process the data on tools and parts required in previous repairs at different consumer locations and discover the co-occurrence patterns.

RELATIONSHIP WITH OTHER FIELDS

SYRACUSE UNIVERSITY
School of Information Studies

ORIGINS OF DATA MINING

Draws ideas from machine learning and artificial intelligence (AI), statistics, and database systems



Terminology problem

- Synonyms:
  - Variable (statistics)
  - Column, attribute, field (database)
  - Feature, attribute (machine learning)

WHAT IS NOT DATA MINING?

What is *not* data mining?

- Search phone number in phone directory
  - Trivial task
  - The answer is not new knowledge
- Query a Web search engine for information about “Amazon”
  - An information retrieval problem
  - Could use data mining techniques to help

DESCRIPTIVE VS. PREDICTIVE ANALYSIS

SYRACUSE UNIVERSITY
School of Information Studies

DESCRIPTION VS. PREDICTION

Predictive analysis

- Uses some variables to predict unknown or future values of other variables: classification, regression

Descriptive analysis

- Derives patterns (average, correlations, trends, clusters, and anomalies) that summarize the underlying relationships in data

Sometimes the difference between descriptive and predictive analysis is not black and white

- Trends, clusters, anomalies

SAMPLE DATA PROBLEM

The marketing department of a financial firm keeps records on customers, including demographic information and number of type of accounts. When launching a new product, such as a Personal Equity Plan (PEP), a direct mail piece advertising the product is sent to 500 existing customers, a sample of its 1 million customers, and a record kept as to whether each customer responded and bought the product. Based on this store of prior experience, the managers decide to use data mining techniques to build customer profile models, which will be used to decide which of the 1 million customers are likely to buy a PEP and thus should receive the advertisement.

http://facweb.cs.depaul.edu/mobasher/classes/csc478/Assignments/bank-data.html

DATA DESCRIPTION

ID	A UNIQUE IDENTIFICATION NUMBER
age	age of customer in years
sex	MALE/FEMALE
region	Inner city/rural/suburban/town
income	income of customer
married	Is the customer married (YES/NO)
children	number of children
car	Does the customer own a car (YES/NO)
save_acct	Does the customer have a saving account (YES/NO)
current_acct	Does the customer have a current account (YES/NO)
mortgage	Does the customer have a mortgage (YES/NO)
pep	Did the customer buy a PEP after the last mailing (YES/NO)

SAMPLE DATA

	A	B	C	D	E	F	G	H	I	J	K	L
1	id	age	sex	region	income	married	children	car	save_act	current_a	mortgage	pep
2	ID12201	54	MALE	INNER_CI	26707.9	YES	1	NO	YES	YES	YES	YES
3	ID12202	27	FEMALE	INNER_CI	11604.4	YES	2	YES	YES	YES	NO	NO
4	ID12203	42	MALE	INNER_CI	15499.9	YES	0	YES	NO	YES	YES	YES
5	ID12204	43	MALE	TOWN	33088.5	NO	0	NO	YES	YES	YES	NO
6	ID12205	64	FEMALE	INNER_CI	34513.6	YES	1	NO	YES	YES	NO	YES
7	ID12206	43	MALE	TOWN	32395.5	YES	3	YES	YES	YES	NO	NO
8	ID12207	49	MALE	RURAL	46633	YES	0	YES	YES	NO	NO	NO
9	ID12208	23	MALE	INNER_CI	13039.9	YES	0	NO	NO	YES	NO	NO
10	ID12209	23	MALE	INNER_CI	12681.9	NO	0	NO	YES	YES	NO	YES
11	ID12210	30	FEMALE	INNER_CI	24031.5	YES	2	YES	YES	YES	YES	NO
12	ID12211	36	MALE	TOWN	37330.5	NO	2	NO	YES	YES	NO	YES
13	ID12212	34	MALE	INNER_CI	25333.2	YES	3	YES	NO	NO	YES	NO
14	ID12213	51	FEMALE	INNER_CI	37094.2	YES	0	YES	NO	YES	NO	NO
15	ID12214	36	MALE	TOWN	33630.6	NO	2	YES	YES	YES	NO	YES
16	ID12215	56	MALE	INNER_CI	43228.2	YES	1	YES	YES	YES	NO	YES
17	ID12216	54	FEMALE	INNER_CI	47796.8	YES	0	NO	YES	YES	NO	NO
18	ID12217	56	FEMALE	TOWN	21730.3	YES	2	NO	YES	NO	NO	NO
19	ID12218	26	MALE	INNER_CI	10044.1	YES	3	NO	YES	YES	YES	NO
20	ID12219	39	MALE	TOWN	17270.1	NO	0	YES	NO	NO	NO	YES
21	ID12220	64	FEMALE	RURAL	45765	YES	3	YES	YES	YES	NO	YES
22	ID12221	46	MALE	RURAL	29525.5	NO	2	NO	YES	NO	YES	NO
23	ID12222	62	FEMALE	RURAL	54863.8	YES	1	YES	YES	YES	NO	YES

DESCRIPTIVE ANALYSIS QUESTIONS

What are the average age and income of the customers?

Is there correlation between age and income?

How many people have 0, 1, 2, 3, or more children?

Is there correlation between the number of children and the decision to buy a PEP?

PREDICTIVE ANALYSIS QUESTIONS

Given a customer's demographic profile, what is the chance that he or she would buy the bank product PEP?

CHALLENGES OF DATA MINING

SYRACUSE UNIVERSITY
School of Information Studies

CHALLENGES OF DATA MINING

Scalability

Dimensionality

Complex and heterogeneous data

Data quality, security, and ownership

Privacy preservation

DATA COMMUNICATION SKILLS

SYRACUSE UNIVERSITY
School of Information Studies

DATA COMMUNICATION SKILLS

Hone your data communication skills while learning data mining techniques.

- E.g., information presentation and visualization

Combine them to become a great “data storyteller,” a critical characteristic of the new data scientist.

- http://www.seas.harvard.edu/news/2014/01/big-data-heralds-new-kind-of-analyst

QUOTES FROM THE NEWS REPORT

http://www.seas.harvard.edu/news/2014/01/big-data-heralds-new-kind-of-analyst

"As computational tools become more sophisticated, the field of data science risks alienating non-experts. Investigative journalists, for instance, have much to gain from accessible research tools."

"It is important for practitioners of computational science and engineering to be able to accurately and engagingly communicate the results of an investigation to others outside their field."

"There's an element we can learn from journalists—hearing how they tell stories and investigate and ask questions, and how they find what's actually interesting to other people," explained Schutt. "It's important in communicating about data [to know] exactly what's objective and what's subjective ... and [to make] sure you're transparent about the data collection process and your modeling process."

AN EXAMPLE OF INFORMATION PRESENTATION

How would you compare the income distribution of men and women shown in the following table?

INCOME (IN THOUSAND DOLLARS)						
	<$50K	[$50K,$60K)	[$60K,$70K)	[$70K,$80K)	>$80K	Total
MEN	10	20	100	30	40	200
WOMEN	10	20	50	20	10	100

ARE THESE DESCRIPTIONS CORRECT?

1. There are equal numbers of men and women in the lower income group (up to $60K).
2. More men than women earned higher incomes (above $70K)

INCOME (IN THOUSAND DOLLARS)						
GENDER	<$50K	[$50K,$60K)	[$60K,$70K)	[$70K,$80K)	>$80K	Total
MALE	10	20	100	30	40	200
FEMALE	10	20	50	20	10	100

PERCENTAGE VS. RAW COUNT

Let's convert the original data to the percentages. Actually, the percentage of women in lower income group (up to $60K) is 30%, and the percentage is 15% for men.

INCOME (IN THOUSAND DOLLARS)						
GENDER	<$50K	[$50K,$60K)	[$60K,$70K)	[$70K,$80K)	>$80K	Total
MALE	.05	.10	.50	.15	.20	1.00
FEMALE	.10	.20	.50	.20	.10	1.00

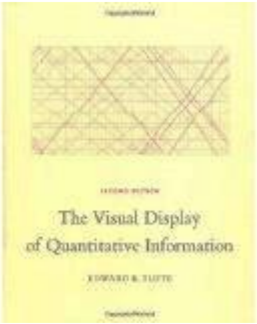
BUT PERCENTAGE IS NOT ALWAYS BETTER

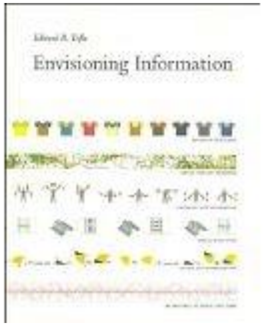
A statement: “The number of students from Mars doubled this year.”

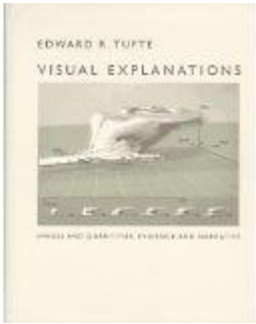
Additional fact: “There was one student from Mars last year.”

HOW TO IMPROVE DATA STORYTELLING SKILLS?

Some general data science books may be fun to read and can improve your data storytelling skills, such as Edward Tufte's books on data presentation and visualization.







BILINGUALISM IN MATH

Steven Strogatz, *The Joy of x: A guided Tour of Math, from One to Infinity*

Check out Professor Strogatz's 15-part series on math in the *New York Times*

http://opinionator.blogs.nytimes.com/category/steven-strogatz/?module=BlogCategory&version=Blog%20Post&action=Click&contentCollection=Opinion&pgtype=Blogs&region=Header