CSI-777
Principles of Knowledge Mining

Class 2

William G. Kennedy,
PhD, CAPT, USN (Ret.)
Center for Social Complexity
Computational and Data Sciences Dept.
College of Science

Outline

- Review last lecture
- Discuss homework
- New material

Review of Previous Class

How will class meetings work?

- Class schedule: Thursdays 4:30-7:10pm
- In class, basic outline:
  - Discussion of previous readings & any exercise
  - Lecture on new material
  - Break(s)?
  - At end of semester, project presentations
- Outside of class (between classes)
  - Some readings & some written reviews
  - Some exercises
  - Contact me with questions, feedback, time of day, etc.

Other topics or questions?

Science Facts

Attribute	Description	Examples	Operations
Nominal	Names (only)	Color(names),sex, IDs, zip codes	Entropy, correlation, shuffle
Ordinal	Ordered objects	Street numbers, mineral hardness, qualitative terms (good, better, best), grades	Sort, percentiles, rank correlation
Interval	Difference meaningful (within range) units	Dates, temperatures (in ^{o}C&F)	Mean, standard deviation, correlation, t & F tests, fixed linear transformations
Ratio	Diff. and ratios meaningful	Temperature (K), Money, counts, age, mass, length, …	Geometric mean, %variation, linear scaling


GOLD HILL
EST. — 1859
ELEV. — 8463
POP. — 118
TOTAL 10440


ASHTON
POPULATION 1219
FOUNDED 1906
ELEVATION 5262
TOTAL 8387

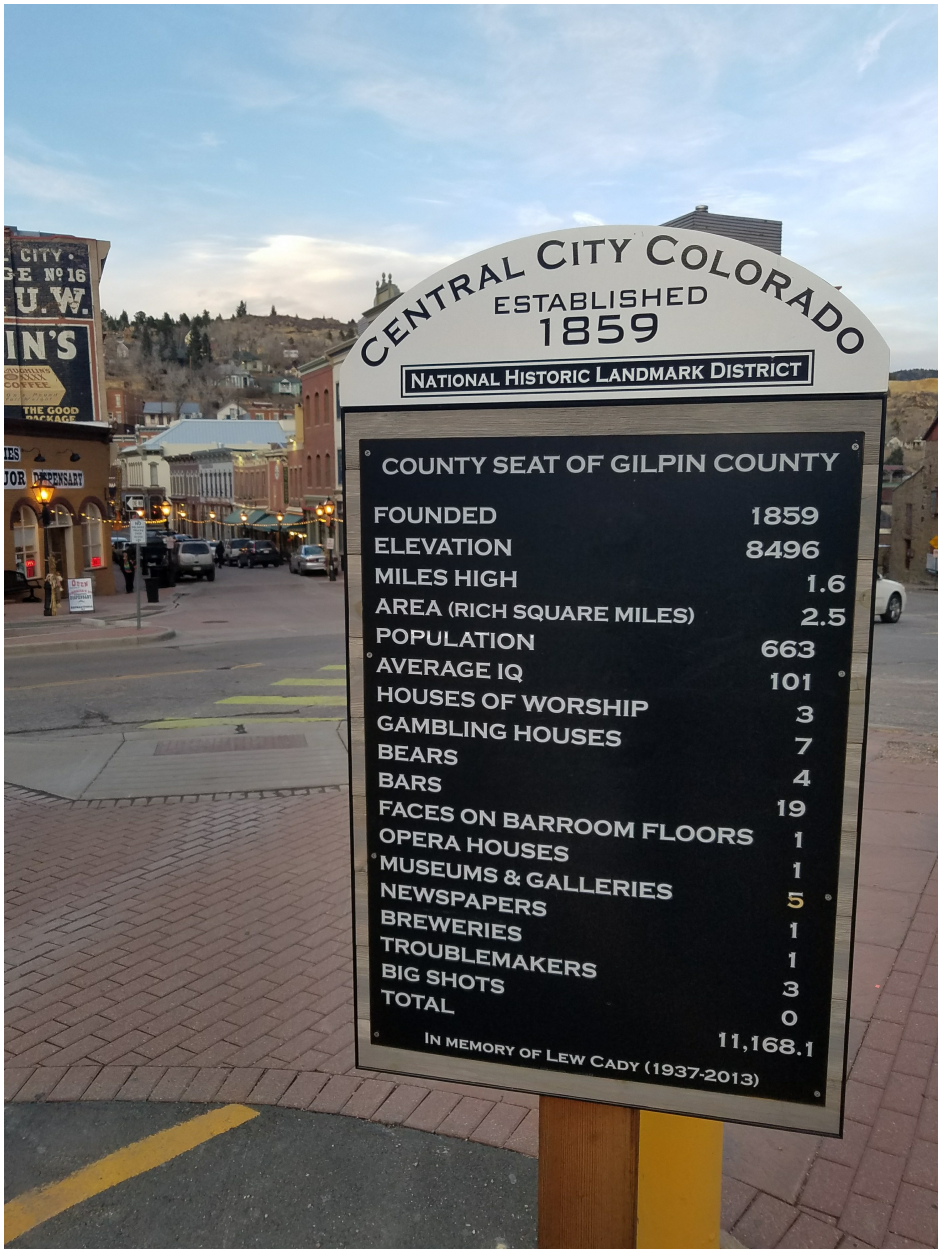

PAGE
ARIZONA
FOUNDED - 1957
ELEVATION - 4,300
POPULATION - 9,743
TOTAL - 16,000


NEW CUYAMA
Population 562
Ft above sea level 2150
Established 1951
TOTAL 4663

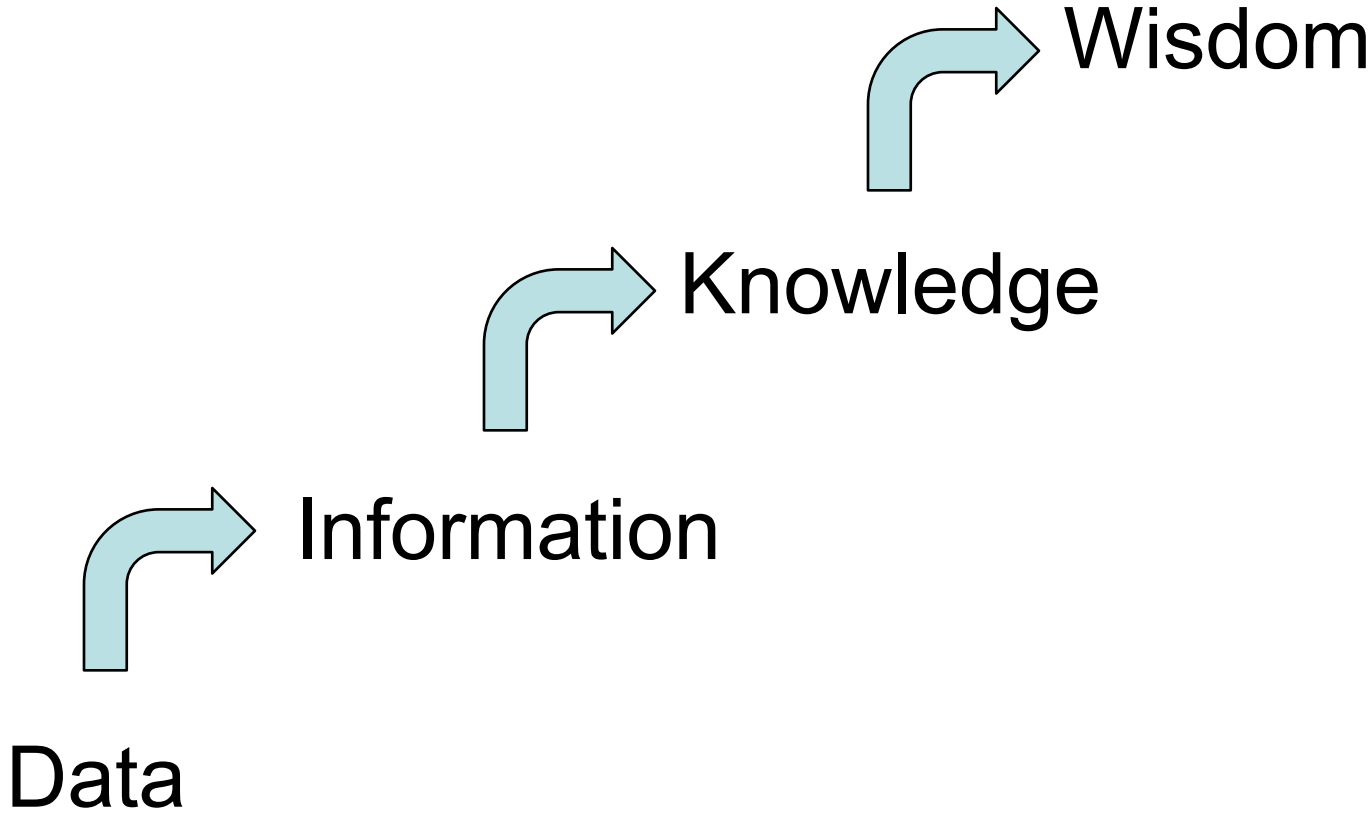

TOWN OF
SNOWMASS VILLAGE
ESTABLISHED 1967
ELEVATION 8388
POPULATION 2824
TOTAL 13,181

Math is hard


CENTRAL CITY.
COUNTY SEAT OF GILPIN COUNTY.
FOUNDED 1859
ELEVATION 8496
MILES HIGH 1.6
AREA (RICH SQUARE MILES) 2.5
POPULATION 663
AVERAGE IQ 101
HOUSES OF WORSHIP 3
GAMBLING HOUSES 7
BEARS 4
BARS 19
FACES ON BARROOM FLOORS 1
OPERA HOUSES 1
MUSEUMS & GALLERIES 5
NEWSPAPERS 1
BREWERIES 1
TROUBLEMAKERS 3
BIG SHOTS 0
TOTAL 11,168.1

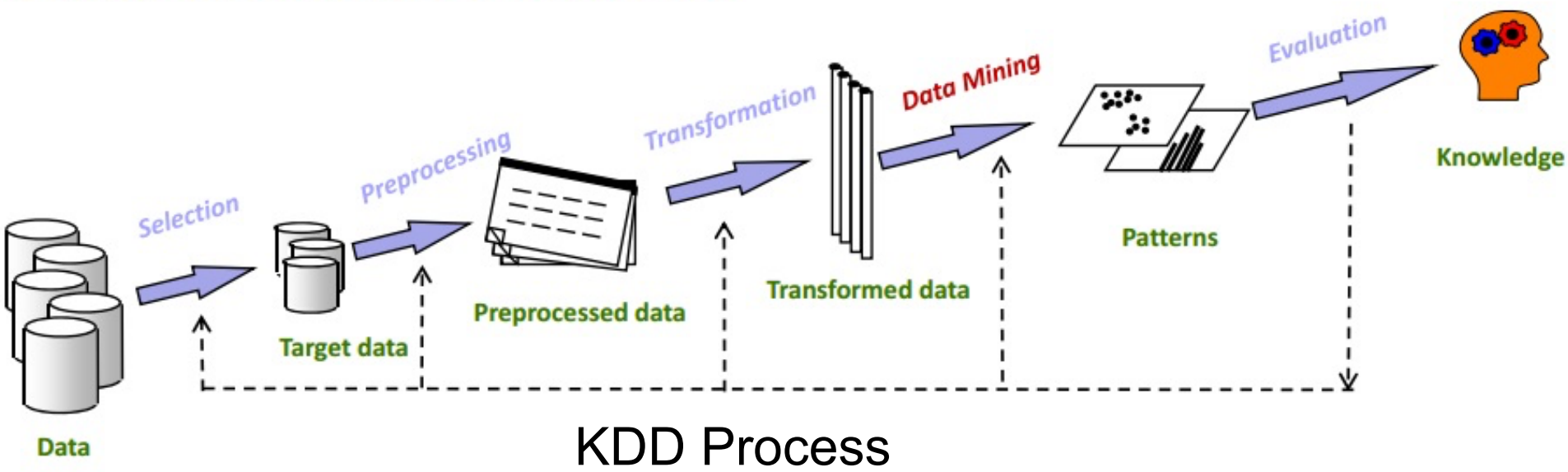

CENTRAL CITY COLORADO
ESTABLISHED
1859
NATIONAL HISTORIC LANDMARK DISTRICT
COUNTY SEAT OF GILPIN COUNTY
FOUNDED 1859
ELEVATION 8496
MILES HIGH 1.6
AREA (RICH SQUARE MILES) 2.5
POPULATION 663
AVERAGE IQ 101
HOUSES OF WORSHIP 3
GAMBLING HOUSES 7
BEARS 4
BARS 19
FACES ON BARROOM FLOORS 1
OPERA HOUSES 1
MUSEUMS & GALLERIES 5
NEWSPAPERS 1
BREWERIES 1
TROUBLEMAKERS 3
BIG SHOTS 0
TOTAL 11,168.1
IN MEMORY OF LEW CADY (1937-2013)

First, Data…



Knowledge vs. Data Mining

[Fayyad, Piatetsky-Shapiro & Smyth, 1996]



KDD Process

(Knowledge Discovery in Databases)

“Big Data”

Examples of “Big Data”

Google’s first setup at Stanford circa 1998

Algorithmic Complexity

scale	n	n log(n)	n^2
Tiny	10^2	$2x10^2$	10^4
Small	10^4	$4x10^4$	10^8
Medium	10^6	$6x10^6$	10^{12}
Large	10^8	$8x10^8$	10^{16}
Huge	10^{10}	10^{11}	10^{20}
Massive	10^{12}	$1.2x10^{13}$	10^{24}
Super massive	10^{15}	$1.5x10^{16}$	10^{30}

Number of operations for algorithms of various computational complexity and various data set sizes

Algorithmic Complexity

scale	n	n log(n)	n^2
Tiny	<1sec.	<1sec	<1sec.
Small	<1sec.	<1sec.	<1sec.
Medium	<1sec.	<1sec.	6.67 sec.[2]
Large	<1sec.	<1sec.	18.5 hrs
Huge	<1sec.	0.67 sec.	21.2 yrs
Massive	6.67 sec.	1.33 min	211 millennia
Super massive	1.85 hrs	1.16 days	2.11×10^{12} yrs >>age of universe

Computational Feasibility on a Intel Core i7 Processor (150 Gigaflops).
All times in seconds unless otherwise stated.

#1 (June 2017) PRC's 93 petaflops: 10^6 x faster… still 10^6 years
#1 (June 2018) US's DOE "Summit" by IBM 122.3 pentaflopw (not since 2012)

Intro to ”Data Mining”

- Why Data Mining?
- What is Knowledge Discovery in Databases?
- Potential Applications
  - Fraud Detection
  - Manufacturing Processes
  - Targeting Markets
  - Scientific Data Analysis
  - Risk Management
  - Web Intelligence

Intro to ”Data Mining”

Data Mining: On what kind of data?

- Relational Databases
- Data Warehouses
- Transactional Databases
- Advanced
  - Object-relational
  - Spatial, Temporal, Spatiotemporal
  - Text, www
  - Heterogeneous, Legacy, Distributed

Intro to ”Data Mining”

Data Mining: Why now?

– Confluence of multiple disciplines
  - Database systems, data warehouses
  - Machine learning
  - Statistical and data analysis methods
  - Visualization
  - Mathematical programming
  - High performance computing

Intro to ”Data Mining”

Why do we need data mining?

- Large number of records (cases) (10^8-10^{12} bytes)
- High dimensional data (variables) (10-10^4 attributes)

How do you explore millions of records, tens or hundreds of fields, and find patterns?

Intro to ”Data Mining”

Why do we need data mining?

- Only a small portion, typically 5% to 10%, of the collected data is ever analyzed.
- Data that may never be explored continues to be collected out of fear that something that may prove important in the future may be missing.
- Magnitude of data precludes most traditional analysis (more on complexity later).

Data Preparation

KDD and data mining have roots in traditional database technology

- As database grow, the ability of the decision support process to exploit traditional (I.e. Boolean) query languages is limited.
  - Many queries of interest are difficult/impossible to state in traditional query languages
  - "Find all cases of fraud in IRS tax returns."
  - "Find all individuals likely to ignore Census questionnaires."
  - "Find all documents relating to this customer's problem."

Massive Data Sets

One Terabyte Dataset

vs

One Million Megabyte Data Sets

Both difficult to analyze

but for different reasons

Data Mining of Massive Data Sets

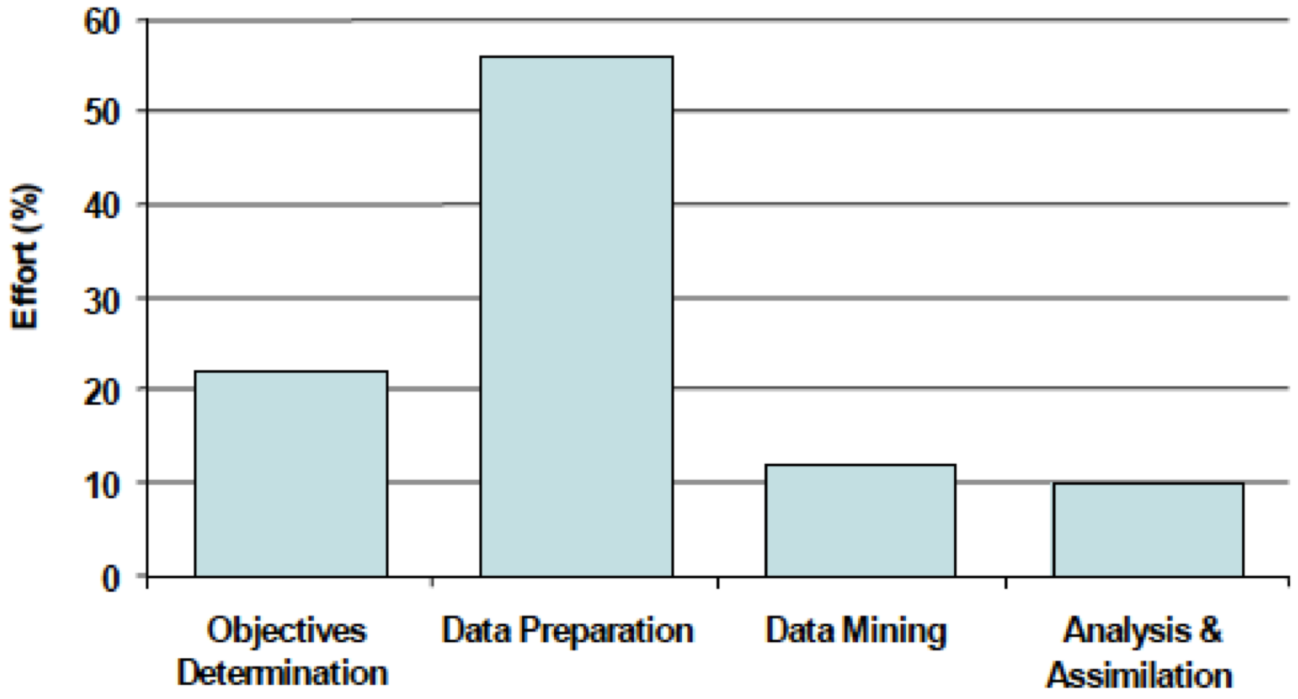
Data Mining is a kind of **Exploratory Data Analysis** with **Little or No Human Interaction** using **Computationally Feasible Techniques**, i.e., the Attempt to find Interesting Structure unknown a priori

Massive Data Sets

- Major issues:
  - Complexity
  - Non-homogeneity
- Examples
  - Air Traffic Control: megabyte per min.
  - Highway maintenance: maintenance records over decades, uneven quality & missing data…

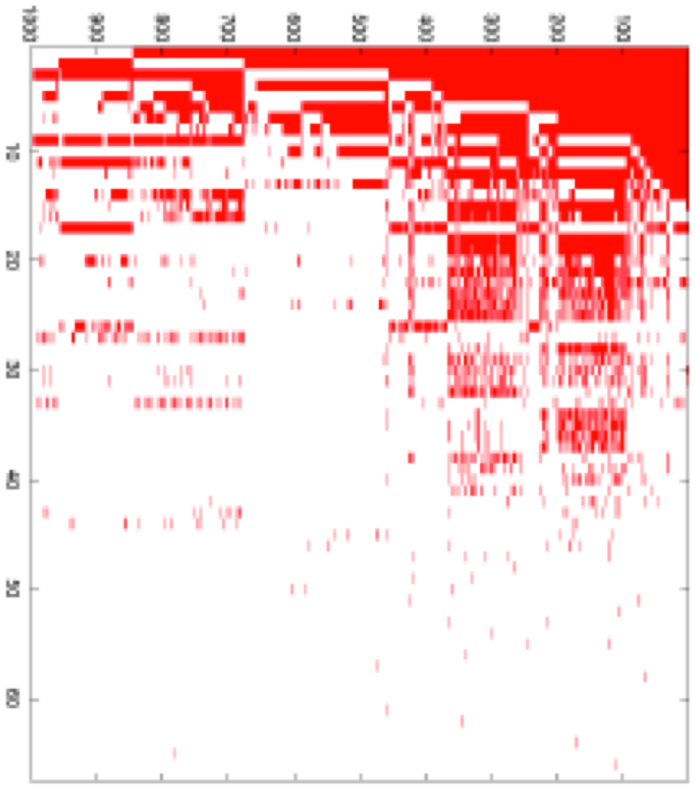
Data Preparation

Data Preparation



Missing Data

- Missing Value Plot
  - A plot of variables by cases
  - Missing values colored red
  - Special case of "color histogram" with binary data
  - "Color histogram" also known as "data image"
  - This example is 67 dimensions by 1000 cases



Data Compression

- Sampling
- Quantization

Sampling

- May be practical rather than exhaustive processing
- Tools can help select nec'y data
- To work, data must satisfy certain conditions (avoid biases)
- Sampling a DBMS can be more expensive than sequential full processing

Data Quantization

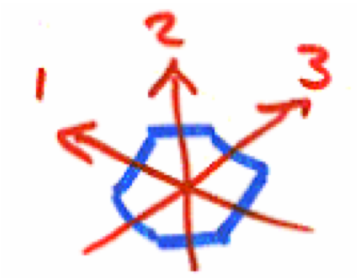
- 3 approaches

Geometry-based Quantization

- Need space-filling tessellations
- Need congruent tiles
- Need as spherical as possible

Geometry-based Quantization

- (polytope = shape with flat sides)
- 1D data: only possible polytope is line segment
- 2D data: only polytopes: equilateral triangle, square, and hexagons



Quantization Strategies

- Geometry-based Summary:
  - Geometric approach good for 4-5 D
  - Adaptive tilings may improve growth rate but may increase distortion
  - Good for large n, but weaker for large d

Quantization Strategies

- **Another approach**: distance based clustering (in EE, ”vector” quantization)
- Form bins via clustering $O(n^2)$
  - Poor for >>n
  - Not too sensitive to dimension, d
  - Clusters may not be round...
- Bottom line: good for large d, not large n, not particularly useful for “massive” data

Quantization Strategies

- **Third approach**: Density-based clustering
- Density estimation O(n)
- Not distance-based, not $O(n^2)$
- Roundness may be a problem

What about syllabus?

Syllabus

- Points, readings, exercises, project
- Readings (35%): evidence you read and considered the material
- Exercises (40%): exercise breadth of methods of knowledge mining
- Project (25%): whole cycle for your topic

Review of Homework

Comments

- Read and write a review of Rowley (2006)
- Install software

- Timing
- Observations

New Material

This week:
Data Mining Concepts

This week:
Data Mining Concepts

Types of data/knowledge mining

- Classification learning - from classified examples, learn how to classify others
- Association learning – finding associations, not just defining ones
- Clustering – group together
- Numeric prediction – interested in a predicted value, not just class ID

"Concept" – thing to be learned, based on approach

Data Mining concepts

“Instance” – set of instances are input to learning process

“Example” –used for general discussion, instances are input data

“Attributes” –values of an instance

Data set: matrix of instances vs. attributes

Data Mining Concepts

Table 2.1

Iris Data as a Clustering Problem

	Sepal Length	Sepal Width	Petal Length	Petal Width
1	5.1	3.5	1.4	0.2
2	4.9	3.0	1.4	0.2
3	4.7	3.2	1.3	0.2
4	4.6	3.1	1.5	0.2
5	5.0	3.6	1.4	0.2
…				
51	7.0	3.2	4.7	1.4
52	6.4	3.2	4.5	1.5
53	6.9	3.1	4.9	1.5
54	5.5	2.3	4.0	1.3
55	6.5	2.8	4.6	1.5
…				
101	6.3	3.3	6.0	2.5
102	5.8	2.7	5.1	1.9
103	7.1	3.0	5.9	2.1
104	6.3	2.9	5.6	1.8
105	6.5	3.0	5.8	2.2
…				

Attributes

Instances

Values

Data Mining Concept: Relations

Instances

First person	Second person	Sister of?
Peter	Peggy	No
Peter	Steven	No
…	…	…
Steven	Peter	No
Steven	Graham	No
Steven	Pam	Yes
Steven	Grace	No
…	…	…
Ian	Pippa	Yes
…	…	…
Anna	Nikki	Yes
…	…	…
Nikki	Anna	Yes

Attributes

Values

Data Mining Concept: Relations

First person	Second person	Sister of?
Steven	Pam	Yes
Graham	Pam	Yes
Ian	Pippa	Yes
Brian	Pippa	Yes
Anna	Nikki	Yes
Nikki	Anna	Yes
All the rest		No

Data Mining Concept: Relations

Table 2.3

Family Tree

Name	Gender	Parent1	Parent2
Peter	Male	?	?
Peggy	Female	?	?
Steven	Male	Peter	Peggy
Graham	Male	Peter	Peggy
Pam	Female	Peter	Peggy
Ian	Male	Grace	Ray

Using concept of family tree to organize sister data

Data Mining Concept: Relations

Table 2.4

The Sister-of Relation

First Person				Second Person				
Name	Gender	Parent1	Parent2	Name	Gender	Parent1	Parent2	Sister-of?
Steven	Male	Peter	Peggy	Pam	Female	Peter	Peggy	Yes
Graham	Male	Peter	Peggy	Pam	Female	Peter	Peggy	Yes
Ian	Male	Grace	Ray	Pippa	Female	Grace	Ray	Yes
Brian	Male	Grace	Ray	Pippa	Female	Grace	Ray	Yes
Anna	Female	Pam	Ian	Nikki	Female	Pam	Ian	Yes
Nikki	Female	Pam	Ian	Anna	Female	Pam	Ian	Yes
All the rest								No

Using concept of family tree to organize sister data

Data Mining Concept: Relations

However, relations don’t deal well with noisy data…

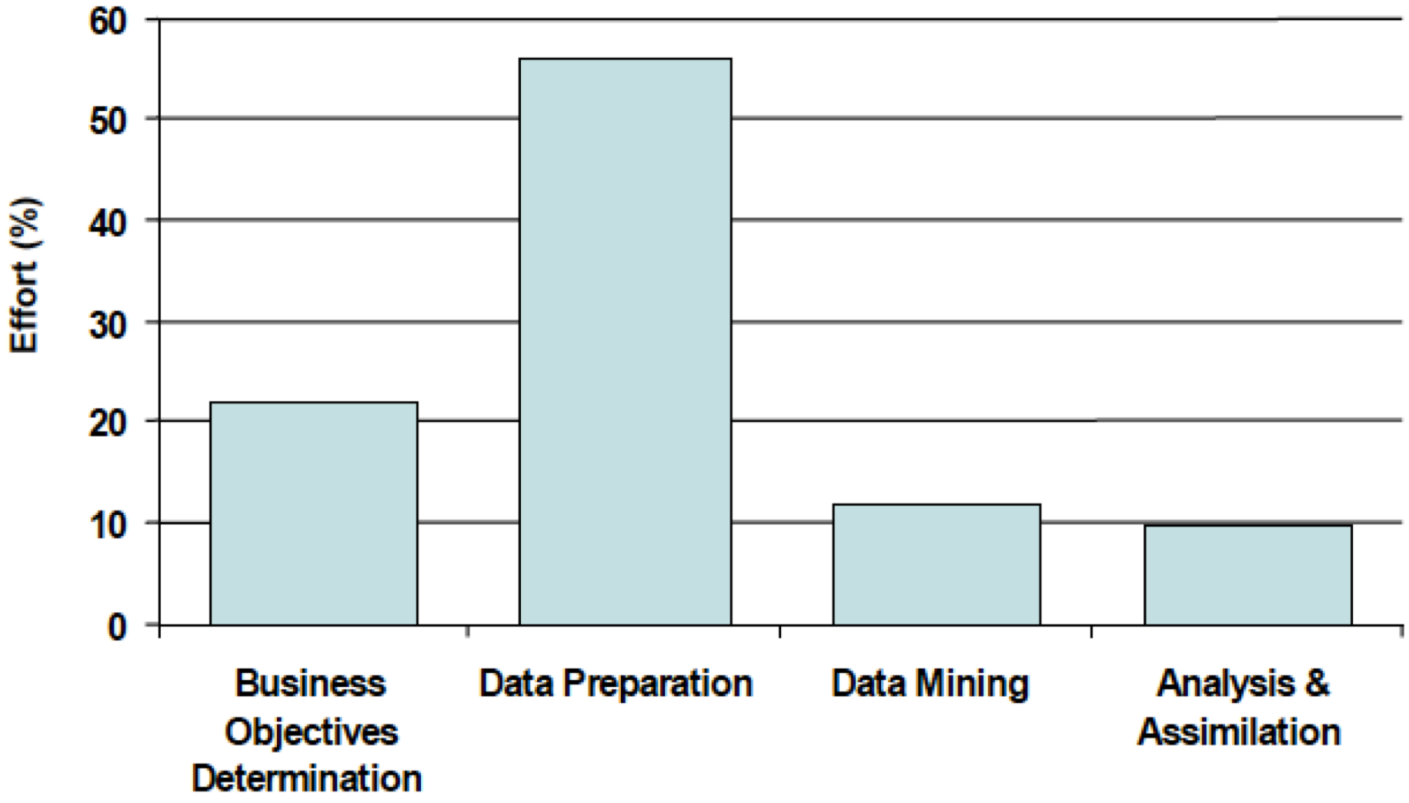
Data Mining Concept: Attributes 2

- Instances have attributes w/ fixed values
- Some instances may have different attributes
- (NIH data example)
- Nominal, Ordinal, Interval, Ratio values for attributes

Preparing Input

- First and biggest step in the process
- Topics:
  - Gathering the data (databases+)
  - Selecting a format
    - ARFF
  - Sparse data
  - Attribute value types: nominal & numeric
  - Missing
  - Noisy
  - Unbalanced
  - … getting to know your data

Relative Effort



Gathering the Data

- KDD & Data Mining roots in database technology
- Simplest is a single Excel spreadsheet
- Relational Databases & Structured Query Language (SQL) have long history
- Relational Databases (RD) include “tables”

Sample Data

- Sample table called "empinfo"

first	last	id	age	city	state
John	Jones	99980	45	Payson	Arizona
Mary	Jones	99982	25	Payson	Arizona
Eric	Edwards.	88232	32	San Diego	California
Mary Ann	Edwards.	88233	32	Phoenix	Arizona
Ginger	Howell	98002	42	Cottonwood	Arizona
Sebastian	Smith	92001	23	Gila Bend	Arizona
Gus	Gray	22322	35	Bagdad	Arizona
Mary Ann	May	32326	52	Tucson	Arizona
Erica	Williams	32327	60	Show Low	Arizona

…

SQL

- `select` <column1[,col2, …]>
  `from` <table>
  [`where` <condition>];
- Where clause: =,>,<,>=,<=,<>, & “like”
- `select first, last, city from empinfo where first LIKE ‘Er%’;`
- `select * from empinfo where first LIKE ’Eric’;`
- Commands for create, insert, update, and delete

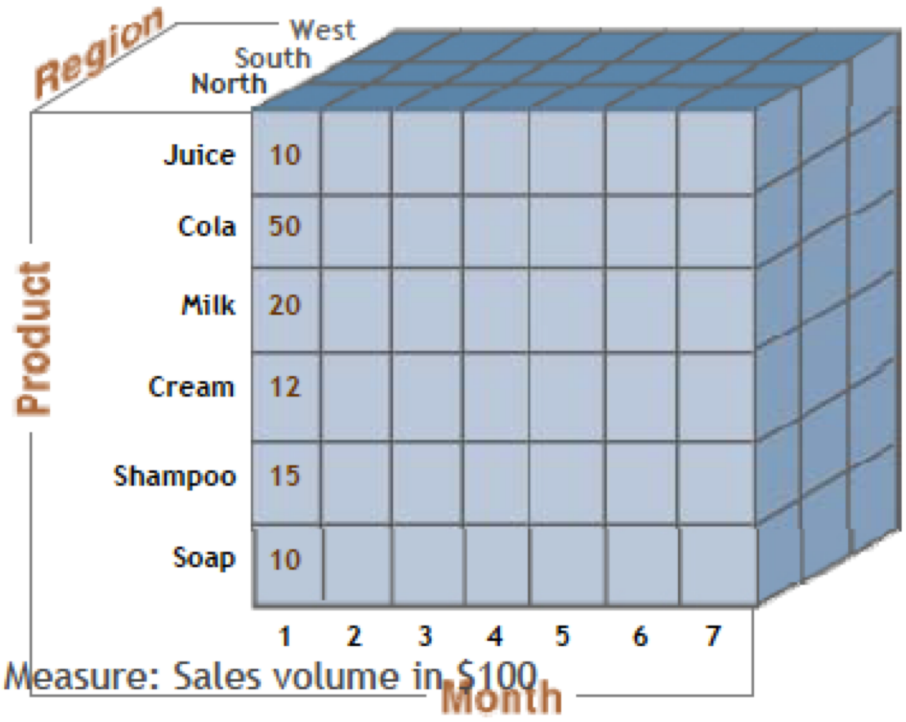
Databases

- Computer scientists prefer RD & SQL
- Statisticians prefer with flat files e.g., text/numbers w/ space, tab, comma separators
- RD can have more structure, more flexibility, hence more overhead…
- Back to flat files for massive data analysis

Databases

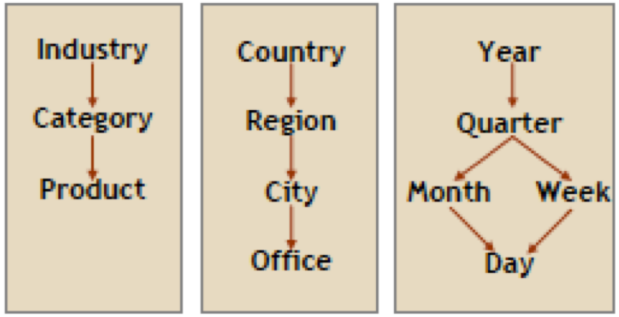
- Data Cubes, OLAP (online analysis processing)
- For business management
- Local databases assembled into central facility, a data warehouse

Data Cube



Dimensions:
Product
Region
Week

Hierarchical Summarization Paths:



Data Cube

- Multidimensional array of data
- Each dimension: set of sets representing domain content such as time or geography
- Dimensions scaled categorically: e.g., region of country, state, quarter of year, week of quarter.
- Cells: aggregated measures (usually counts) of variables
- Exploration: **drill down, drill up, drill though**.

Data Cube Extended

- OLAP = On-line Analytical Processing
- MOLAP = Multidimensional OLAP
- Fundamental data object for MOLAP is the Data Cube
- Operations limited to simple measures like counts, means, proportions, standard deviations, but do not work well for non-linear techniques
- Aggregate of the statistic is not the statistic of the aggregate
- ROLAP = Relational OLAP using extended SQL

Database Summary

- Use of database technology is fairly compute intensive
- Touching an observation means using it
- Commercial database technology is challenged by analysis of full data sets above about 10^8
- This limitation applies to many of the algorithms developed by computer scientists for data mining

Preparing Input: Management issues

1. Standards
2. Storage
3. Who's data is it

Ok, ARFF

Preparing Input: Management issues

```
% ARFF file for the weather data with some numeric features
%
@relation weather

@attribute outlook { sunny, overcast, rainy }
@attribute temperature numeric
@attribute humidity numeric
@attribute windy { true, false }
@attribute play? { yes, no }

@data
%
% 14 instances
%
sunny, 85, 85, false, no
sunny, 80, 90, true, no
overcast, 83, 86, false, yes
rainy, 70, 96, false, yes
rainy, 68, 80, false, yes
rainy, 65, 70, true, no
```

Sparse Data

- Representation efficiency
- Sparse vs. missing
- Storage vs. handling efficiency

Numeric Attribute Values

- ARFF only supports nominal and “numeric”
- Techniques treats numeric as ratio data?
- ”Standardizing” to mean=0, std dev = 1

  (subtract mean & divide by std dev)
- If ordinal data, units (distance definition)
- Some (instance-based and regression) deal only with ratio scales due to differences between instances handled as “distances”

Missing Data

- Real world…
- Mark… with out of range value (-1, -9999, …)
- Significance, if any?

Inaccurate Values

- Types: typos, duplicates, measurement error, collection biases, …
- How handle?
- Common sense fixes (spelling names)
- Effects of bad data…

Anscombe Datasets

- Number of observations (n) = 11
- Mean of x's ($\bar{x}$)= 9.0
- Mean of y's ($\bar{y}$)= 7.5
- Regression coefficient (b_1) of y on x = 0.5
- Equation of regression line: $y = 3 + 0.5x$
- Sum of squares of $x - \bar{x} = 110.0$
- Regression sum of squares = 27.50 (1 d.f.)
- Residual sum of squares of y = 13.75 (9 d.f.)
- Estimated standard error for ($b1$) = 0.118
- Multiple $R^2 = 0.667$

Anscombe Datasets

Dataset 1		Dataset 2		Dataset 3		Dataset 4	
x	y	x	y	x	y	x	y
10.0	8.04	10.0	9.14	10.0	7.45	8.0	6.58
8.0	6.95	8.0	8.14	8.0	6.77	8.0	5.76
13.0	7.58	13.0	8.74	13.0	12.74	8.0	7.71
9.0	8.81	9.0	8.77	9.0	7.11	8.0	8.84
11.0	8.33	11.0	9.26	11.0	7.81	8.0	8.47
14.0	9.96	14.0	8.10	14.0	8.84	8.0	7.04
6.0	7.24	6.0	6.13	6.0	6.08	8.0	5.25
4.0	4.26	4.0	3.10	4.0	5.39	19.0	12.50
12.0.	10.84	12.0	9.13	12.0	8.15	8.0	5.56
7.0	4.82	6.0	7.26	7.0	6.42	8.0	7.91
5.0	5.68	5.0	4.74	5.0	5.73	8.0	6.89

Anscombe Datasets

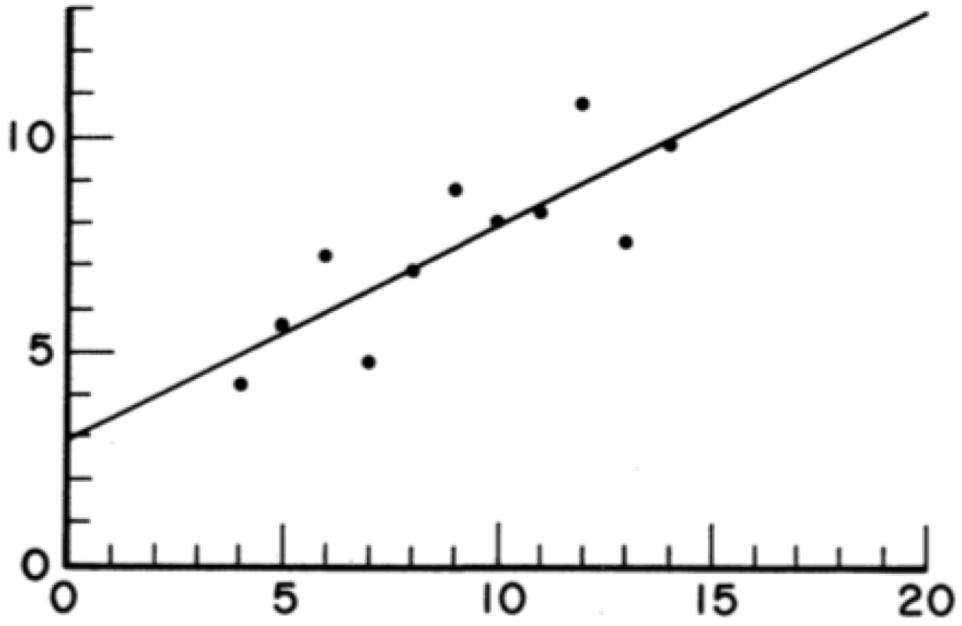


Figure I

Anscombe Datasets

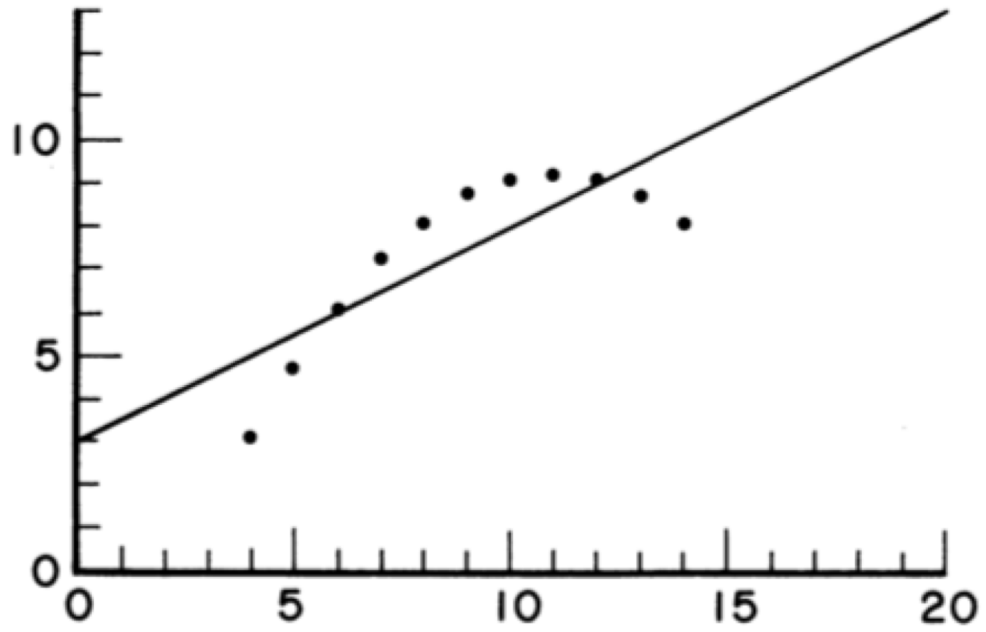


Figure 2

Anscombe Datasets

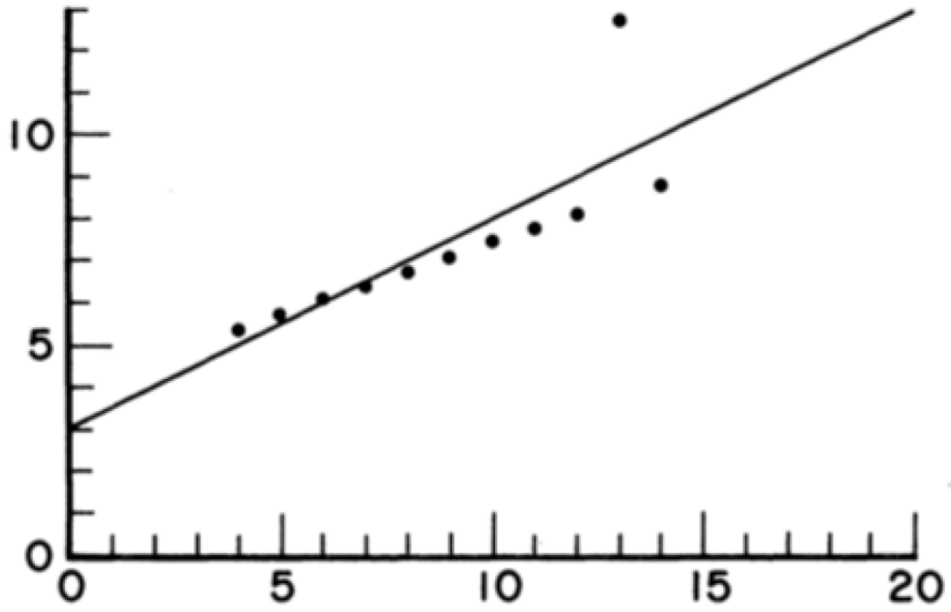


Figure 3

Anscombe Datasets

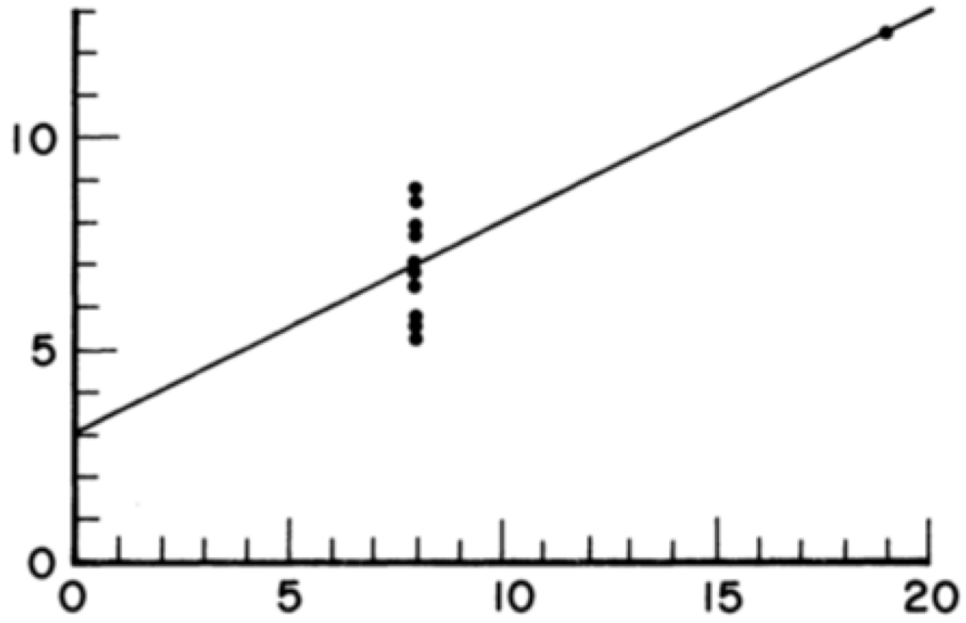


Figure 4

Unbalanced Data

- If 99% of instances have same attribute-value, useful or delete?
- Risk/impact trade off of data

Getting to Know Your Data

- An internal representation of the data
- Ed Tufte's "Show them the data"

Summary

- Data/Knowledge mining concepts (examples, instances, attributes, values)
- Types of learning (classification, association, clustering, numeric prediction)
- And …

Preparing Input

- First and biggest step in the process
- Topics:
  - Gathering the data (databases+)
  - Selecting a format (ARFF)
  - Sparse data
  - Attribute value types: nominal & numeric
  - Missing
  - Noisy
  - Unbalanced
  - … getting to know your data

Homework #1

Ex1: Load the specified weather dataset and remove all instances with high humidity. Submit resulting dataset.