Data Preparation for Data Mining

MAS DSE

March 2015

Outline

- **Motivation and Goals**
- **What is data?**
- **Data Preparation:**
  - Organizing data (structural issues)
  - Preprocessing (data value issues)
  - Exploring Variables and Descriptive Statistics
  - Exploring the Data Matrix
  - Outliers, Anomalies, and Visualizations

On the Importance of Data Prep

- **“Garbage in, garbage out”**
- **A crucial step of the DM process**
- **Could take 60-80% of the whole data mining effort**

Working definition

- **Data Preparation:**
  - cleaning, filtering, transforming, and organizing the data
  - preparing data for modeling

Prerequisites

- **Data understanding:**
  - Descriptors, values, ranges, labels
- **Data history**
- **Domain Knowledge**
  - Meaning and data relations
- **Questions to be addressed**

Input/Output

- **Inputs:**
  - raw data
- **Outputs:**
  - two data sets: training and test (if available)
  - Training further broken into training and validation

End Product: Quality Data

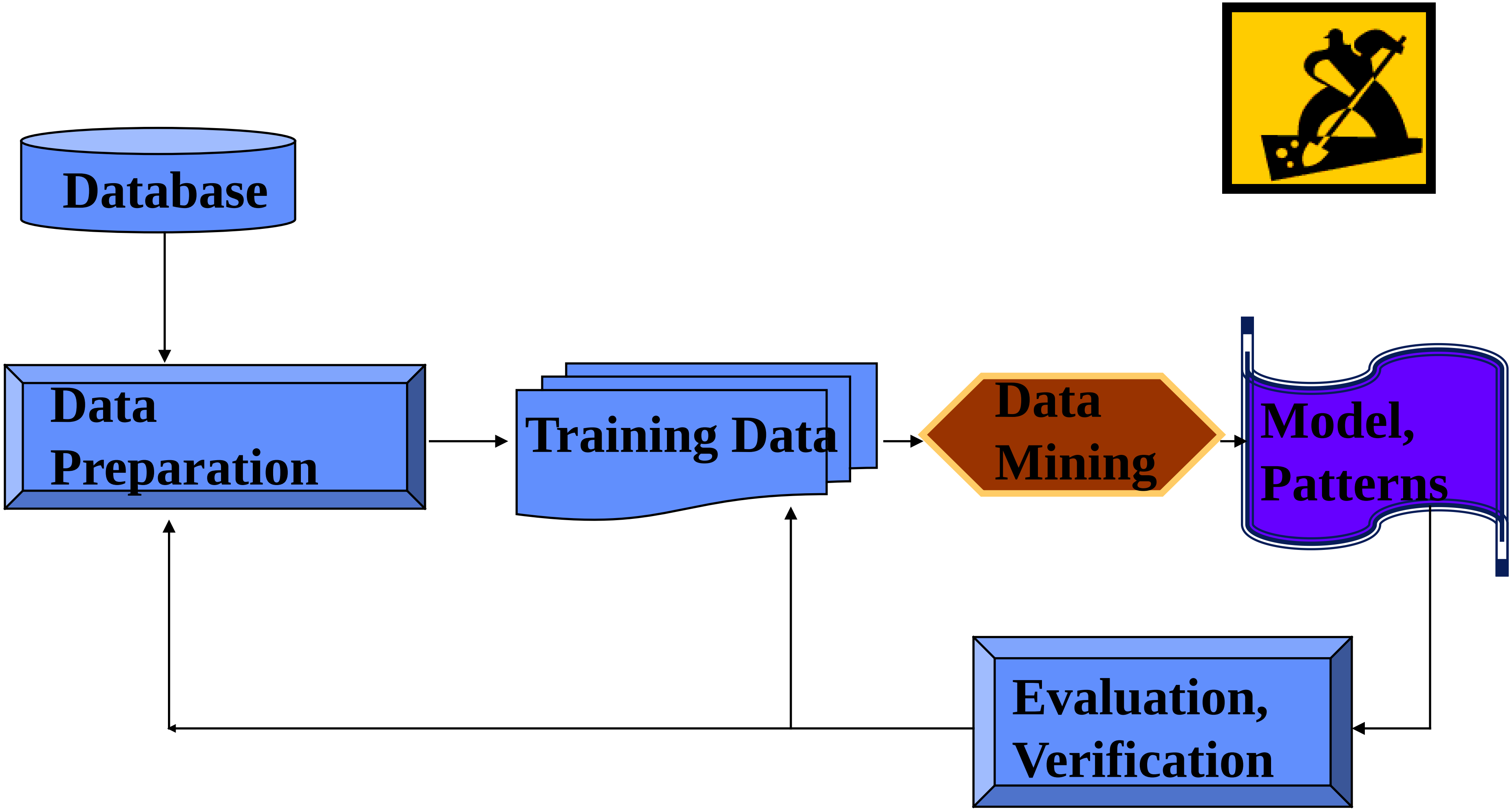
- Accurate
- Complete
- Consistent
- Interpretable

In other words: Good data ⟶ Better results!

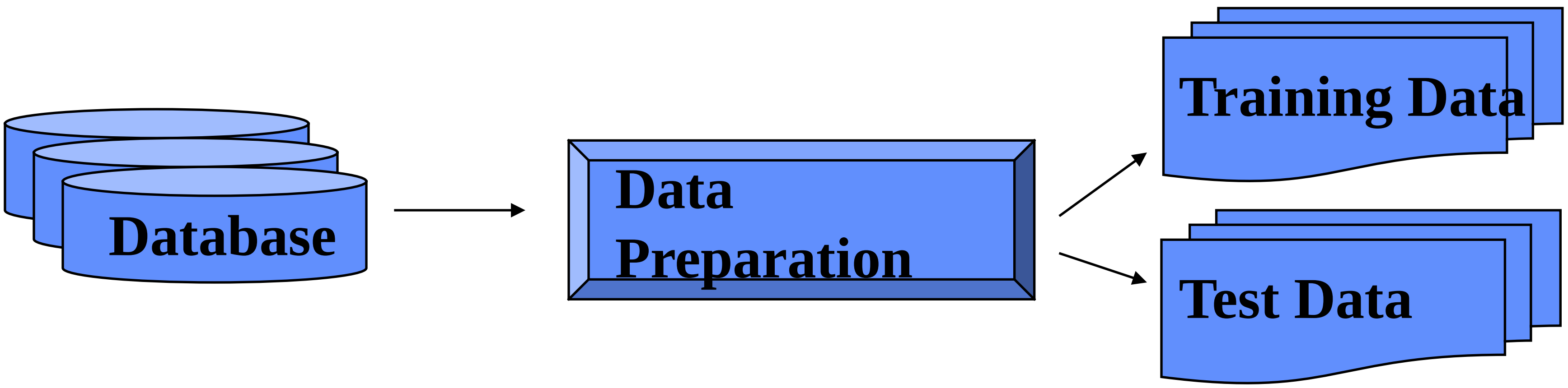
Outline

- **Motivation and Goals**
- **What is data?**
- **Data Preparation:**
  - Organizing data (structural issues)
  - Preprocessing (data value issues)
  - Exploring Variables and Descriptive Statistics
  - Exploring Data Matrix
  - Outliers, Anomalies, and Visualizations

Recall the KDD Process



From Data Source To Algorithm Input



User Decides:

- Selection Criteria –
- Joins => denormalize
- How much data?

Depends on needs and domain knowledge about what's relevant

User Performs:

- Cleaning data and Transformations

Depends on domain knowledge, data itself and possibly on algorithms

Terminology from data source…

- **Data consists of:**

  Examples, observations, measurements, events, transactions, records..

- **Data can be:**

  Structured (e.g. database rows) or unstructured (e.g. text)

… To Algorithm Input

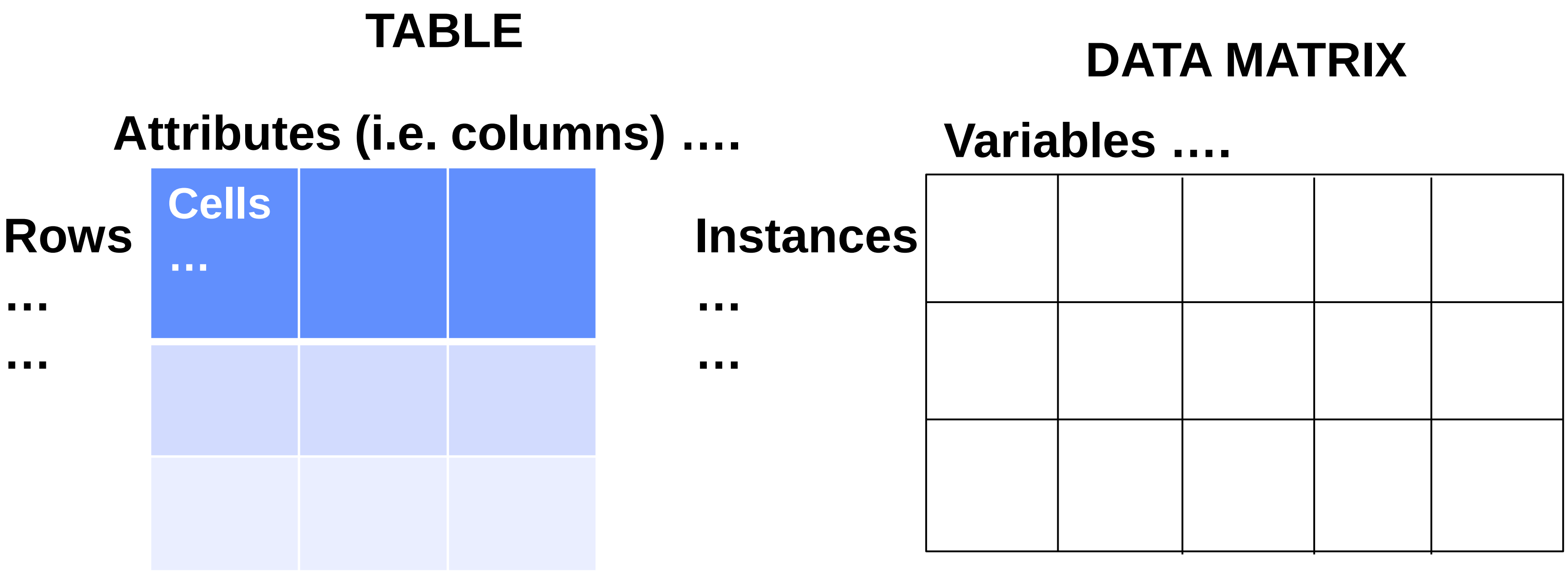
- **Instance = specific example**
  - thing to be classified, associated, or clustered
  - instances may be labeled as a class, or as an outcome
  - If no labels available you can either do unsupervised learning or try to get labels
- **Set of instances comprise the input dataset**
  - Often represented as a single flat file or *data matrix*

Algorithm Input Detail

- **Each instance described by a predefined set of “attributes” or “variables”**

- **Attributes’ values, or it’s existence, may or may not be dependent on each other**
  - e.g. height and weight may be correlated
  - e.g. spouse name depends on marital status

Terms from database to math

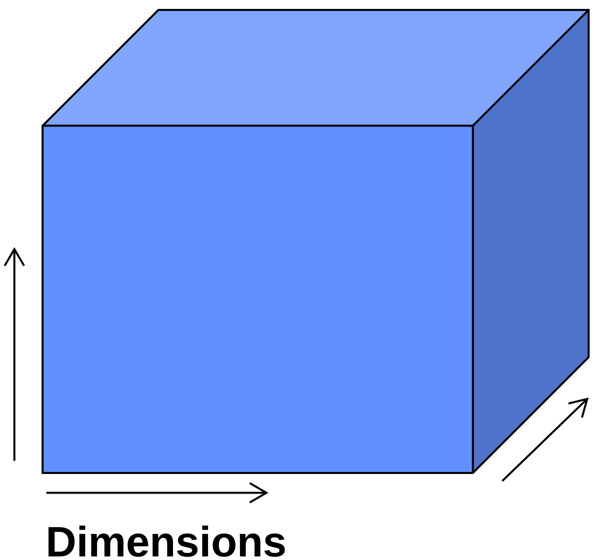


attributes in the database relate to variables in the data matrix

Terms database to math

TABLES can 2 or more dimensions (multi-way) given by discrete attributes called Factors

In DATA MATRIX each variable is a dimension in some coordinate space



Row Vector is a Coordinate Pt. ….

x1	x2	x3	….	

- **Matrix Variables can also be Factors**
- **Factor Tables can also be treated mathematically**

Variables and Features terms

- **Variables and their transformations are features**
- **Instance labels are outcomes or dependent variables (as in supervised learning)**
- **No instance labels available then use unsupervised learning**

Outline

- **Motivation and Goals**
- **What is data?**
- **Data Preparation:**
  - Organizing data (structural issues)
  - Preprocessing   (data value issues)
  - Exploring Variables and Descriptive Statistics
  - Exploring Data Matrix
  - Outliers, Anomalies, and Visualizations

Database to Data Matrix

- **Goal: gather all relevant information into each instance in one data matrix**
  - Typical models are: *instance outcomes = F(row values)*

- **Key: the functions you model and questions you pose determine what variables are bought together and how they are presented**

Organizing data example

Customer	Item	Price	Date
John	Acme Mower	100	Jan 2000
John	Acme Wrench	10	Sept 2000
Jane	Ace Mower	120	Mar 2003
Jane	Ace Rake	20	Mar 2003
Fred	Ace Hammer	15	July 2002

Customer	Zip
John	99000
Jane	11000
Fred	99000

2 tables, keyed on customer id

Simple descriptive queries

Customer	Total Spent
John	110
Jane	140
Fred	15

A data matrix using Aggregation Levels

Relevant Questions involve customers and totals

Database to Data Matrix

Customer	Zip
John	99000
Jane	11000
Fred	99000

Customer	Item	Price	Date
John	Acme Mower	100	Jan 2000
John	Acme Wrench	10	Sept 2000
Jane	Ace Mower	120	Mar 2003
Jane	Ace Rake	20	Mar 2003
Fred	Ace Hammer	15	July 2002

- **What would the data matrix be for a relationship question:**

How similar are zip codes?

Database to Data Matrix

Customer	Zip
John	99000
Jane	11000
Fred	99000

Customer	Item	Price	Date
John	Acme Mower	100	Jan 2000
John	Acme Wrench	10	Sept 2000
Jane	Ace Mower	120	Mar 2003
Jane	Ace Rake	20	Mar 2003
Fred	Ace Hammer	15	July 2002

- **Coding Issues among variables**
  - implicit domain knowledge: customers buy items
  - large number of categorical values: number of items bought
  - spurious regularities, e.g. "item" predicts "supplier"
  - usual data issues, e.g. date/time, composite fields, entity resolution, etc..

Database to Data Matrix

Customer	Zip
John	99000
Jane	11000
Fred	99000

Customer	Item	Price	Date
John	Acme Mower	100	Jan 2000
John	Acme Wrench	10	Sept 2000
Jane	Ace Mower	120	Mar 2003
Jane	Ace Rake	20	Mar 2003
Fred	Ace Hammer	15	July 2002

How similar are zip codes?

‘similar’ wrt to what entities?

‘similar’ implies a comparison?

An approach: instances are transpose of items, cell values are counts

Customer Zip	Acme Mower	Ace Mower	Acme Wrench	Ace Wrench	…	(last item)
99000	1	0	1	0		
11000	0	1	0	0		
…						

Get related measurements down row into separate columns of the same instance

How do zip codes compare?
What items go together?
How do they impact purchases?

Instance are counts, but aggregated across item types

Custom er Zip	Mower	Wrench	Rake	Hammer	…	(last item)
99000	1	1	1	1		
11000	1	0	0	0		
…						

What questions can we ask now?
Should we include customer name and zip code?

Customer	Zip
John	99000
Jane	11000
Fred	99000

Customer	Item	Price	Date
John	Acme Mower	100	Jan 2000
John	Acme Wrench	10	Sept 2000
Jane	Ace Mower	120	Mar 2003
Jane	Ace Rake	20	Mar 2003
Fred	Ace Hammer	15	July 2002

Can also compare customer-item pairs

	Mower	Wrench	Rake	Hammer	…	(last item)
John	1	1	0	0		
Jane	1	0	1	0		
Fred	0	0	0	1		

Would John buy a Rake too?

Should 0 indicate 'not yet bought'?

We can compare customers, or products.
Can we use customer-item pairs collaboratively?

Data Wrangling Cautions

- **Beware of data integration:**
  different names for same data
  different data for same names

Outline

- **Motivation and Goals**
- **What is data?**
- **Data Preparation:**
  - Organizing data (structural issues)
  - Preprocessing (data value issues)
  - Exploring Variables and Descriptive Statistics
  - Exploring Data Matrix
  - Outliers, Anomalies, and Visualizations

4 Preprocessing data values and QA

- **Preprocessing involves:**
  - Cleansing data
  - Missing data
  - Exploring variable characteristics
  - Re-representing variables (normalizing, discretizing, transforming)

Because real data is incomplete, inconsistent, noisy, etc…

Data Preparation is Variable Prep

- **Know the meanings (domain knowledge!)**
- **Know types of variables**
- **Know statistical properties**
- **Do QA (clean, fill-in, fix errors)**
- **Do enhance or re-represent**
  - add more data as needed
  - apply domain knowledge to ease the work of the tool

Types of Measurements

- **Nominal (names)**
- **Categorical (zip codes)**

Qualitative (unordered, non-scalar)

- **Ordinal (H,M,L)**
- **Real Numbers**
  - May or may not have a Natural Zero Point?
    - If not comparisons are OK but not multiplication (e.g. dates)

Quantitative (ordered, scalar)

Know variable properties

- **Explore characteristics of each variable:**
  - typical values, min, max, range etc.
  - entirely empty or constant variables can be discarded
  - explore variable dependencies
- **Sparsity**
  - missing, N/A, or 0?
- **Monotonicity**
  - increasing without bound, e.g. dates, invoice numbers
  - new values not in the training set
- **Visualize the distribution**
  - Check skews, outliers

Noise in Data

- **Noise is unknown error source**
  - sometimes assumed to be independent and random
- **Approaches to Address Noise**
  - Detect suspicious values and remove outliers
  - Smooth by averaging with neighbors
    - but then how many neighbors?
  - Smooth by fitting the data with other variables

Data Errors are also Noise

- **Incorrect attribute values**
  - data collection errors
  - data entry errors
  - duplicate records
  - Etc..
- **Approaches to Address Problems**
  - apply domain knowledge to replace values
  - model error process to reverse engineer correct value
    - e.g. common misspellings and typos

Missing Data

- **Data values not present**

  e.g. customer income in sales data not easy to get

  e.g. sensor malfunction

- **Or data available but missing due to**
  - deletions
  - not entered

Missing Data

- **Important: review statistics of a missing variable**
  - Are missing cases random?
  - Are missing cases random but dependent on other variable(s)?
  - Are other variables missing data in same instances?
  - Is there a relation between missing cases and outcome variable?
  - What is frequency of missing cases?

Quick Approaches to Handle Missing Data

- If there's enough data and missing seems random
  - Delete instances with missing attribute values
  - Delete attributes with high "missingness"
- Use the attribute mean to fill in (impute) the missing value
- Use the attribute mean for all samples belonging to the same class

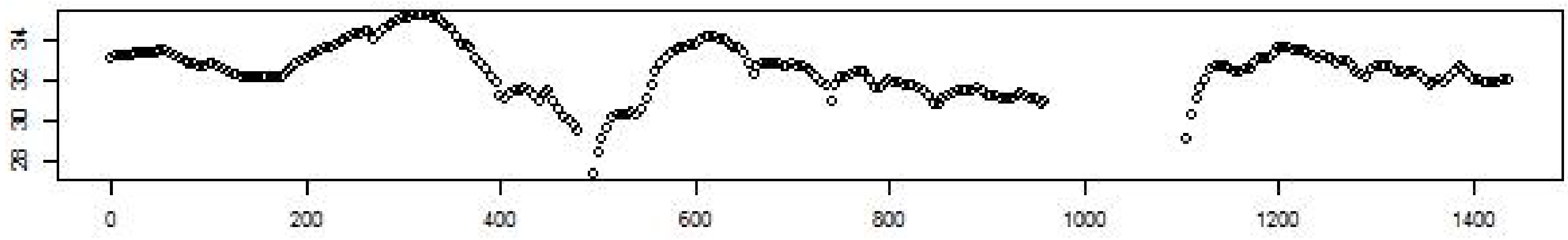
Additional Approaches to Handle Missing Data

- Use a model (based on other attributes) to infer missing value
- Use a global constant to fill in the missing value, e.g. “unknown”, and let algorithms figure it out (e.g. Decision Trees)
- Add a new indicator variable (1 or 0) to indicate missing and let algorithms figure it out (e.g Linear Models)

Missing Data Example

Time series of glucose measurements over 24hours.

raw data of glucose level



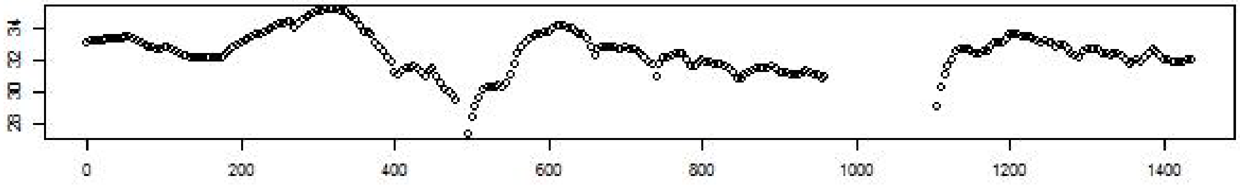
Time (minutes)

Can we ignore missing values?
Should we fill it in with a constant (eg last value)? Or with a mean? Or a model?

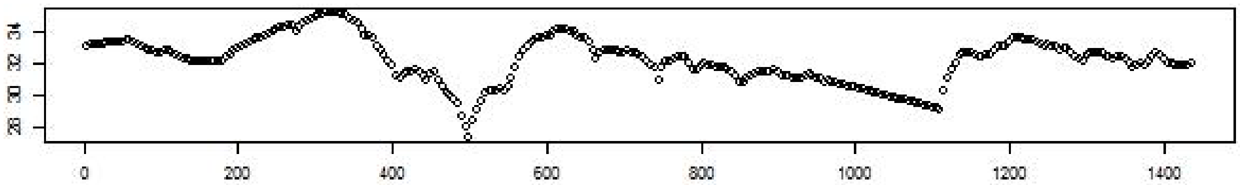
Missing Data Example

Time series of glucose measurements

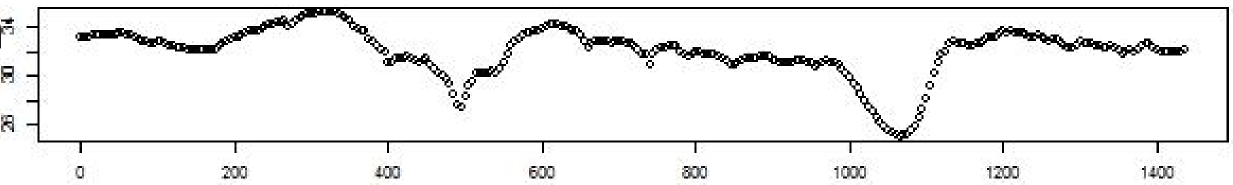
raw data



linear interpolation
(too linear)



polynomial interpolation
(too nonlinear)

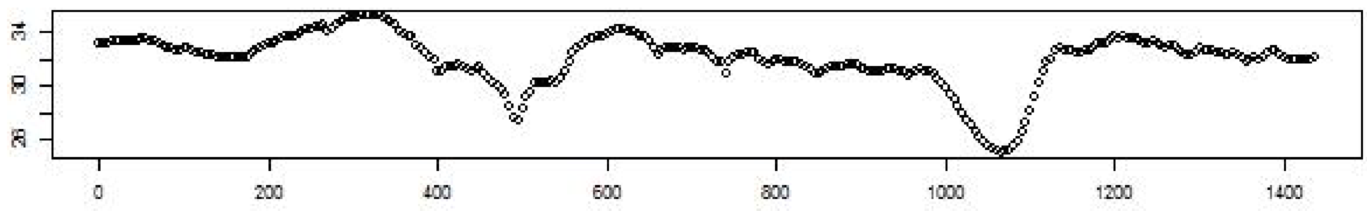


Time (minutes)

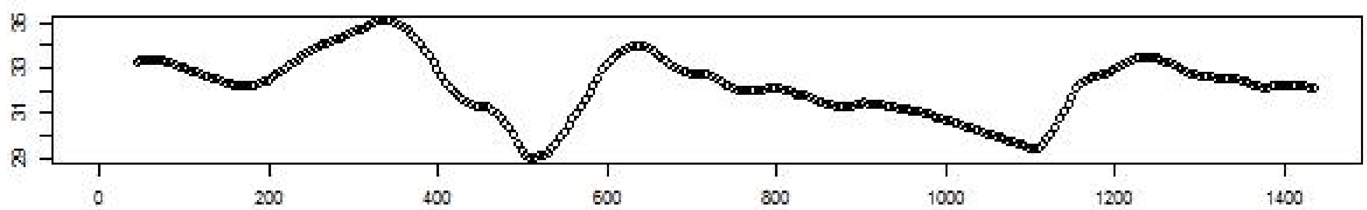
Missing Data Example

Time series of glucose measurements

polynomial interpolation (too nonlinear)



polynomial interpolation then smoothed by averaging over windows (better, but trade offs?)



Time (minutes)

Variable Transformations

- **Why transform data?**
  - **Combine attributes**

    ratios can be more useful
  - **Normalizing data**

    to same scale
  - **Simplifying data**

    discrete data is often more intuitive for user and algorithm and helps the algorithms

Feature Engineering is Variable Enhancement

- **Use Domain and world knowledge to help model**
- **Example: variables exist that represent date and location of doctor visits**
  - deduce a new variable for Number-of-1st-time-visits
  - deduce a new variable for Number-of-visits-over-25-miles
  - deduce a new variable for Amount-of-time-between-visits

Adding Information As Variable Enhancement

- **Example: zip codes**
  - Change ZIP to latitude and longitude
  - Change ZIP to miles to a reference point
  - Change ZIP to known category (H,M,L income)
  - Change ZIP to set of indicator variables (1 per ZIP)

Discretization/Binning May Enhance Data

- **Discretization**
  - A continuous attribute divided into intervals and replaced by Interval labels
  - E.g. replace age by functional concepts (such as young, middle-aged, or senior) which may have better predictive value

Discretization/Binning Options

- **E.g. Equal-width (distance) partitioning:**
  - *N* intervals of equal size, but outliers skew range

64 65 68 69	70 71 72	75	80 81 83	85
Yes No Yes Yes	Yes No No Yes	Yes Yes	No Yes Yes	No

- **E.g. Equal-depth (frequency) partitioning:**
  - *N* intervals, of equal sample frequency, can help scale data

64 65	68 69 70	71 72	75 80	81 83 85
Yes No	Yes Yes Yes	No No Yes	Yes Yes No	Yes Yes No

Is 85 special?

Variable Transformation Summary

- **Smoothing: remove noise from data**
- **Aggregation: summarization, data cube construction**
- **Introduce/re-label/categorize variable values**
- **Normalization: scaled to fall within a small, specified range**
- **Attribute/feature construction**

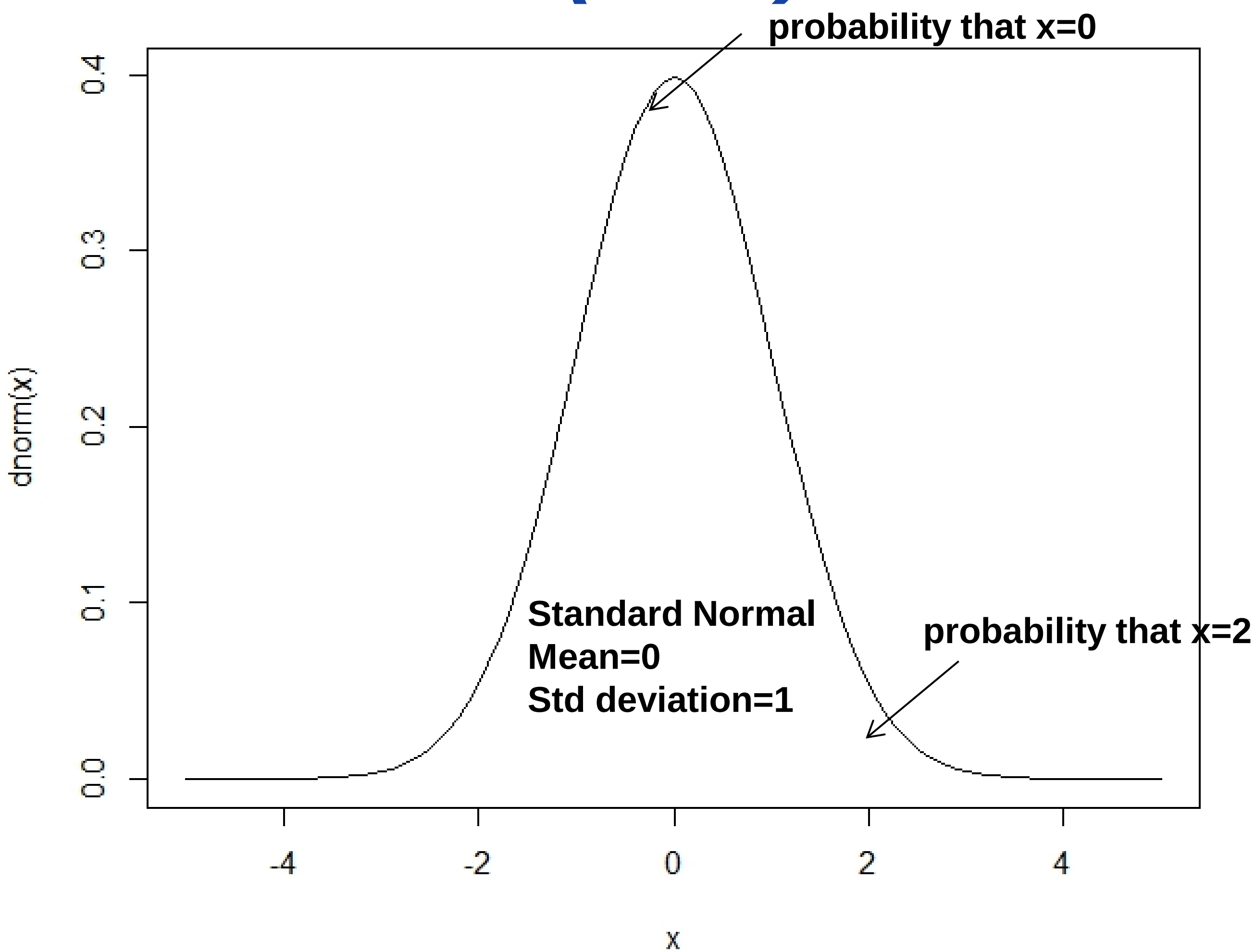
Outline

- **Motivation and Goals**
- **What is data?**
- **Data Preparation:**
  - Organizing data (structural issues)
  - Preprocessing (data value issues)
  - Exploring Variables and Descriptive Statistics
  - Exploring Data Matrix
  - Outliers, Anomalies, and Visualizations

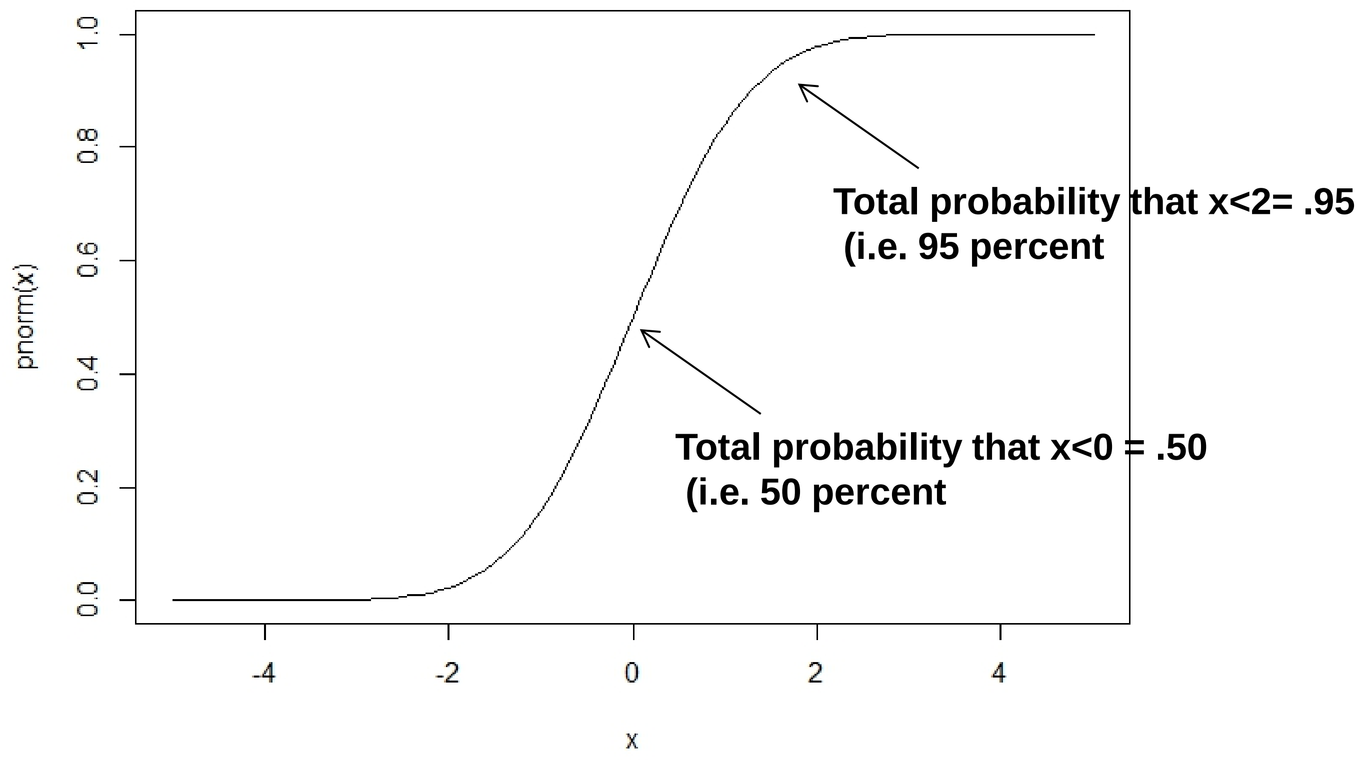
Stats for Data Preprocessing

- **Distributions and histograms**
  - Continuous variables (functions and graphs)
  - Discrete variables (sets and counting)
- **Normalizations**
- **Correlations**

Normal probability density function (PDF)

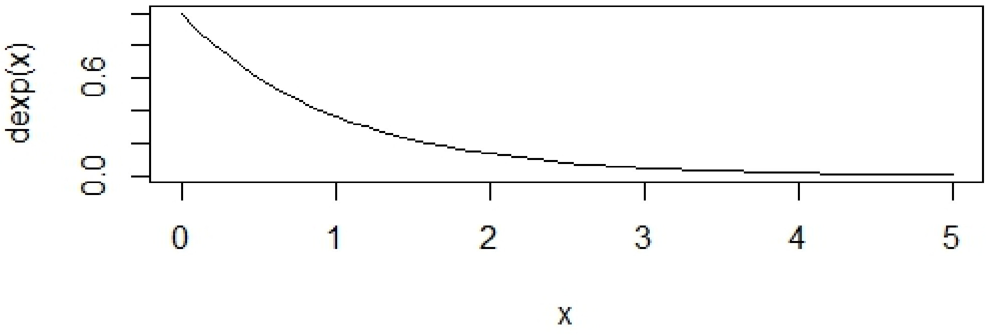


Normal cumulative distribution

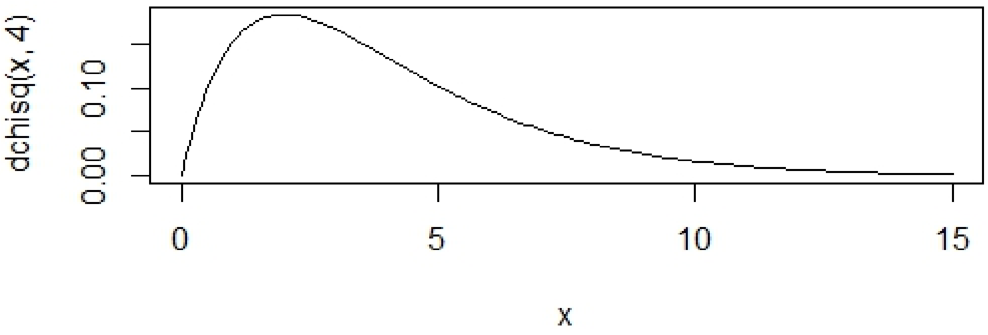


Exponential and Chi-squared density functions

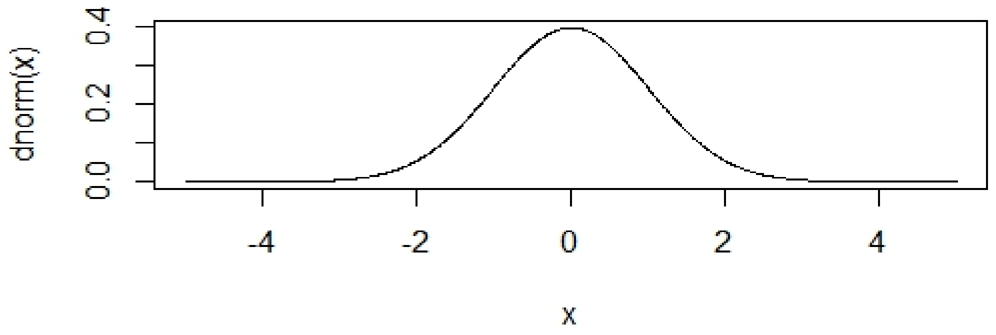
Exponential is good for 'counts', 'events', etc…,
ie, items that are >0, usually near 0, and higher values more rare



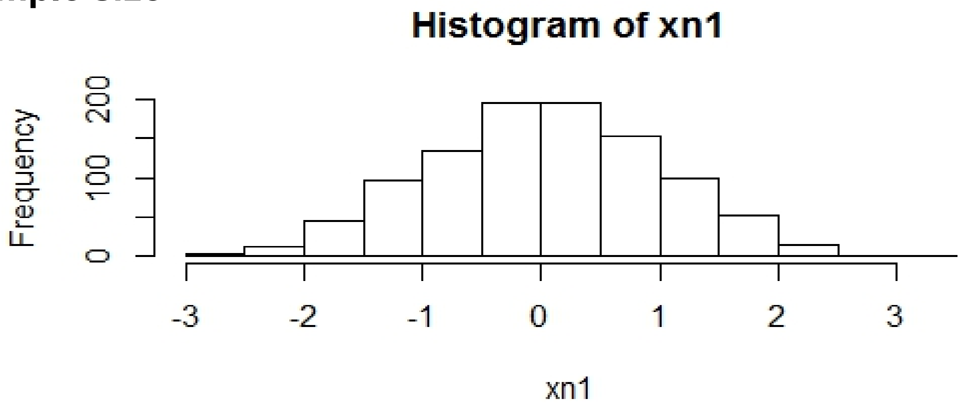
Chi Square is good for 'costs', 'rates', 'salaries', etc….,
ie, items that are > 0, usually not near 0, and higher values more rare



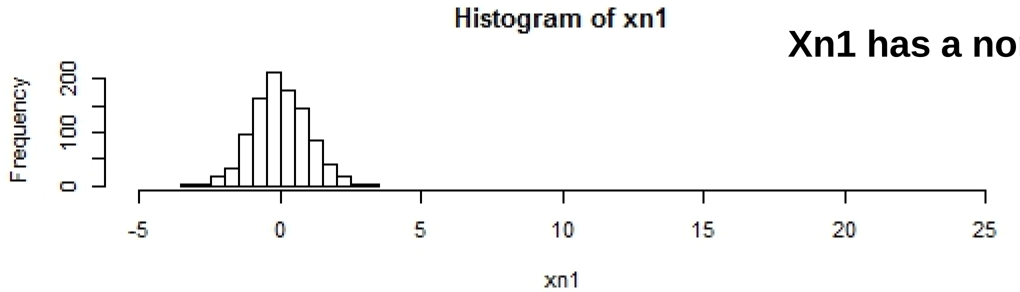
Histogram is a sample PDF



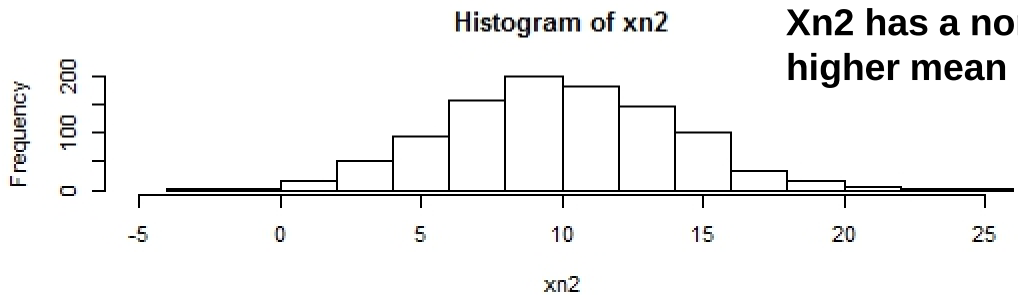
Frequency count ~
probability times sample size



One histogram as mixture

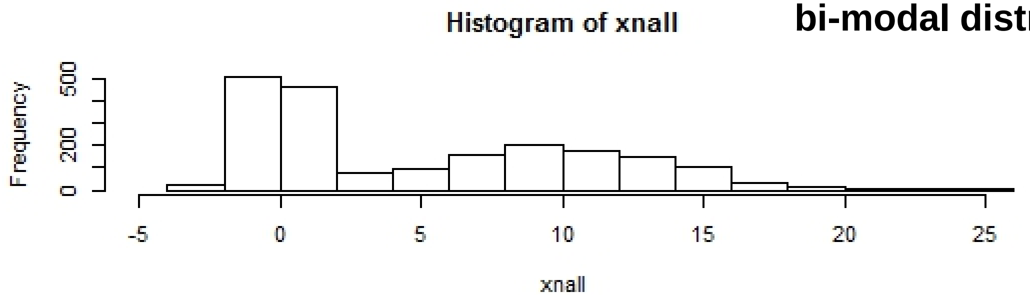


Xn1 has a normal distribution



Xn2 has a normal distribution with higher mean and higher variance

Xn1 + Xn2 has a bi-modal distribution



Descriptive Statistics

- **Mean and Std Dev summarize variables**

$$\text{std}(x, y) = \sqrt{\text{mean}((x - \text{mean}(x))^2)}$$

- **Transformations and Functions also summarize**
  - E.g. take the highest amount charged for customers in a zip code, take that for each zip code and get a new distribution
  - E.g. take the difference of 75th to 25th percentile of all customers in a zip code, take that for each zip code and get a new distribution

Data Transformation: Normalizations (to help with scaling)

- **Mean center** $$x_{new} = x - \text{mean}(x)$$
- **z-score** $$z - score = \frac{x - \text{mean}(x)}{\text{std}(x)}$$
- **Scale to [0…1]** $$x_{new} = \frac{x - \min(x)}{\max(x) - \min(x)}$$
- **log scaling** $$x_{new} = \log(x)$$

More Descriptive Statistics

- **Covariance between 2 variables**

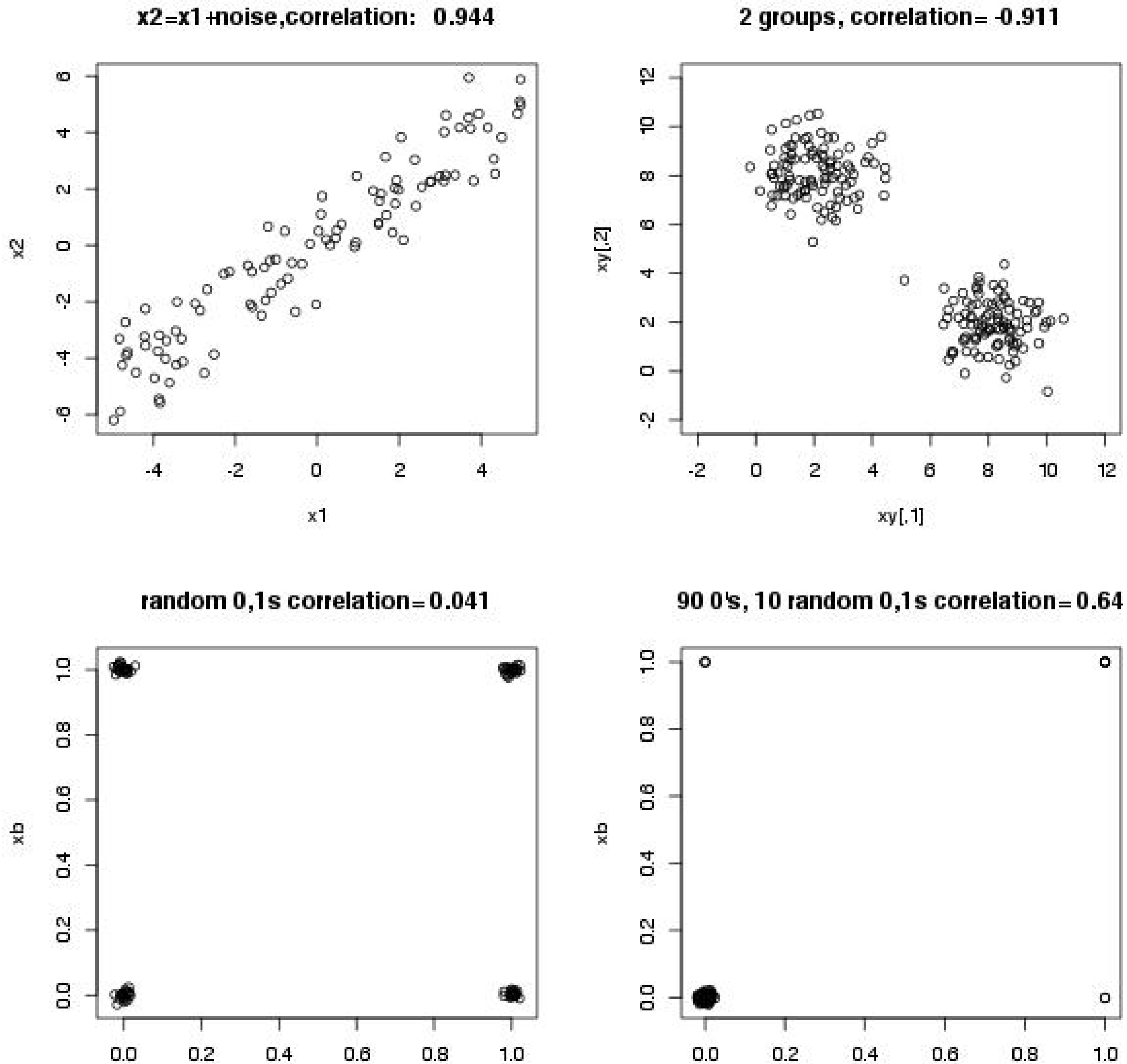
$$\mathrm{cov}(x, y) \sim \mathrm{mean}((x - \mathrm{mean}(x))(y - \mathrm{mean}(y)))$$

- **Correlation between 2 variables**

$$\mathrm{corr}(x, y) \sim \frac{\mathrm{cov}(x, y)}{\mathrm{std}(x)\,\mathrm{std}(y)}$$

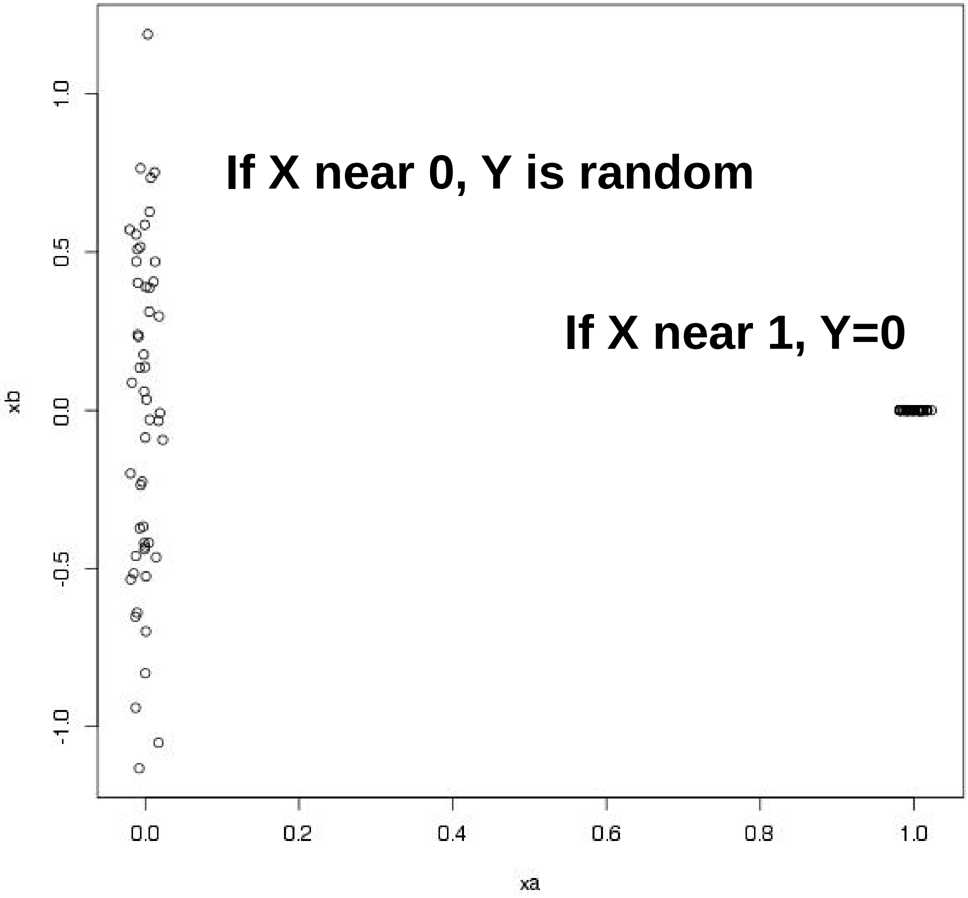
  - Ranges -1 to 1
  - Represents linear relationship

Correlation demos



Correlation vs. Independence

- **No Correlation $\not\Rightarrow$ Independence**



Correlation = .021
But Y depends on X

More Descriptive Statistics

- **(Spearman) Rank correlation between 2 variables**
  - Rank the instances of each variable
    (now there are 2 ordinal rank variables)
  - Take correlation coefficient of ranks
  - Represents monotonic relationship
- **Confidence interval wrt mean or percentiles**

$$\text{mean}(x) - \text{std}(x), \text{mean}(x) + \text{std}(x)$$

$$15th \text{ percentile}, 85th \text{ percentile}$$

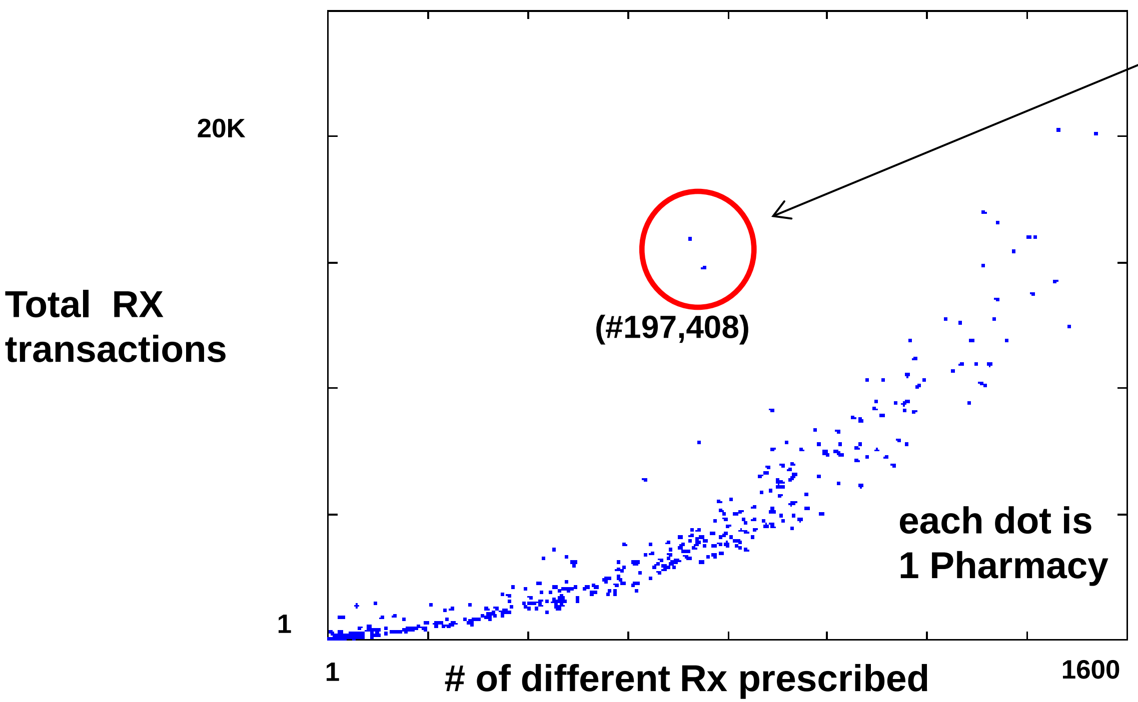
Outline

- **Motivation and Goals**
- **What is data?**
- **Data Preparation:**
  - Organizing data (structural issues)
  - Preprocessing (data value issues)
  - Exploring Variables and Descriptive Statistics
  - Exploring Data Matrix
  - Outliers, Anomalies, and Visualizations

Anomalies

- **3 working definitions of an anomaly**
  - statistical outlier (far from mean)
  - distance based (farthest point to its neighbors)
  - deviance based (model quantity, take biggest error to model)
- **Making decisions and cutoffs**
  - anomalies can be ranked
  - but decisions depend on some cutoff

The importance of normalization and varieties of deviance



Not an outlier in
Rx or in total Rx

Outliers for #Rx/total Rx

Far from others

Deviant wrt main trend

Visualizations

- **For communication and exploration**
- **MultiDimensional Scaling (MDS)**
  - Find points in 2D that preserve relative distances in P-dimensions of full data matrix
  - In some cases similar to PC1 and PC2
- **Plotting relations between variables**
- **Heat Maps over vectors**
  - Discretize into bins and labeled by a few colors

Summary

- **Data preparation is a key issue for mining**
- **Lots of techniques**
- **Partly an art that depends on data and algorithm knowledge**
- **Partly a science that depends on statistical principles**

Reading Material

- **Data Preparation for Data Mining by Dorian Pyle**
  - http://www.ebook3000.com/Data-Preparation-for-Data-Mining_88909.html
- **Data mining – Practical Machine learning tools and techniques by Witten & Frank**
  - http://books.google.com