

CSE/ISE 332

INTRODUCTION TO VISUALIZATION

DATA WRANGLING AND PREPARATION

KLAUS MUELLER

COMPUTER SCIENCE DEPARTMENT
STONY BROOK UNIVERSITY

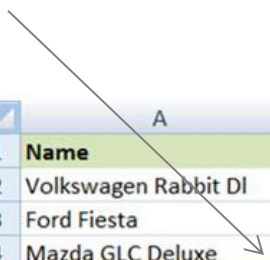
Lecture	Topic	Projects
1	Intro and logistics	
2	Basic visualizations and tasks, data types, examples, ethical considerations	
3	Data preparation (cleaning, imputation, data set integration)	Project #1 out
4	AI-assisted coding for VIS applications (design, debugging, refactoring)	
5	Big data and data reduction (distance/sim metrics, intro to clustering)	
6	High-D data: concept, subspaces, dimension reduction, PCA	
7	Cluster analysis: hierarchical, density, model, embedding, temporal	Project #2(a) out
8	Perception and cognition (human visual system, color, contrast)	
9	Visual design and aesthetics	
10	Visualization of multivariate and high-D data: linear methods, projections	
11	Vis. of multivariate and high-D data: non-linear methods, embeddings	Project #2(b) out
12	Visualization and AI: mutual support and capabilities (VIS4AI, AI4VIS)	
13	Principles of interaction: drive what is visualized, analyzed & how (HCI4VIS)	
14	Midterm recap	
15	Midterm	
16	Visual analytics, human-centered AI, mixed-initiative & collaborative VA	Capstone proposal due
17	Visualization system design & evaluation: nested model, study designs	
18	Visualization of hierarchical data	
19	Visualization of maps and data with geo-reference	
20	Vis. of time-varying, time-series, streaming data, progressive visualization	
21	Critiquing visualization: Tufte's views, integrity, uncertainty, responsibility	Capstone prelim report due
22	Design of effective infographics	
23	Foundations scientific & medical visualization, computer graphics	
24	Intro to volume rendering	
25	Scientific visualization	
26	Capstone project demos (4-5 minutes + 2 minutes Q&A)	All capstone materials due
27	Capstone project demos (4-5 minutes + 2 minutes Q&A)	
28	Capstone project demos (4-5 minutes + 2 minutes Q&A)	
Final	Final exam	

RECTANGULAR DATASET

One data item

The variables

→ the attributes or properties we measured



	A	B	C	D	E	F	G	H	I
1	Name	Country	Miles Per Gallon	Accceleration	Horsepower	weight	cylinders	year	price
2	Volkswagen Rabbit DI	Germany	43,1	21,5	48	1985	4	78	2400
3	Ford Fiesta	Germany	36,1	14,4	66	1800	4	78	1900
4	Mazda GLC Deluxe	Japan	32,8	19,4	52	1985	4	78	2200
5	Datsun B210 GX	Japan	39,4	18,6	70	2070	4	78	2725
6	Honda Civic CVCC	Japan	36,1	16,4	60	1800	4	78	2250
7	Oldsmobile Cutlass	USA	19,9	15,5	110	3365	8	78	3300
8	Dodge Diplomat	USA	19,4	13,2	140	3735	8	78	3125
9	Mercury Monarch	USA	20,2	12,8	139	3570	8	78	2850
10	Pontiac Phoenix	USA	19,2	19,2	105	3535	6	78	2800
11	Chevrolet Malibu	USA	20,5	18,2	95	3155	6	78	3275
12	Ford Fairmont A	USA	20,2	15,8	85	2965	6	78	2375
13	Ford Fairmont M	USA	25,1	15,4	88	2720	4	78	2275
14	Plymouth Volare	USA	20,5	17,2	100	3430	6	78	2700
15	AMC Concord	USA	19,4	17,2	90	3210	6	78	2300
16	Buick Century	USA	20,6	15,8	105	3380	6	78	3300
17	Mercury Zephyr	USA	20,8	16,7	85	3070	6	78	2425
18	Dodge Aspen	USA	18,6	18,7	110	3620	6	78	2700
19	AMC Concord D1	USA	18,1	15,1	120	3410	6	78	2425
20	Chevrolet MonteCarlo	USA	19,2	13,2	145	3425	8	78	3900
21	Buick RegalTurbo	USA	17,7	13,4	165	3445	6	78	4400
22	Ford Futura	Germany	18,1	11,2	139	3205	8	78	2525
23	Dodge Magnum XE	USA	17,5	13,7	140	4080	8	78	3000
24	Chevrolet Chevette	USA	30	16,5	68	2155	4	78	2100

The data items
→ the samples
(observations)
we obtained
from the
population of
all instances

RECTANGULAR DATASET

Also called the *Data Matrix*

Car performance metrics

or Survey question responses

or Patient characteristics

One data item

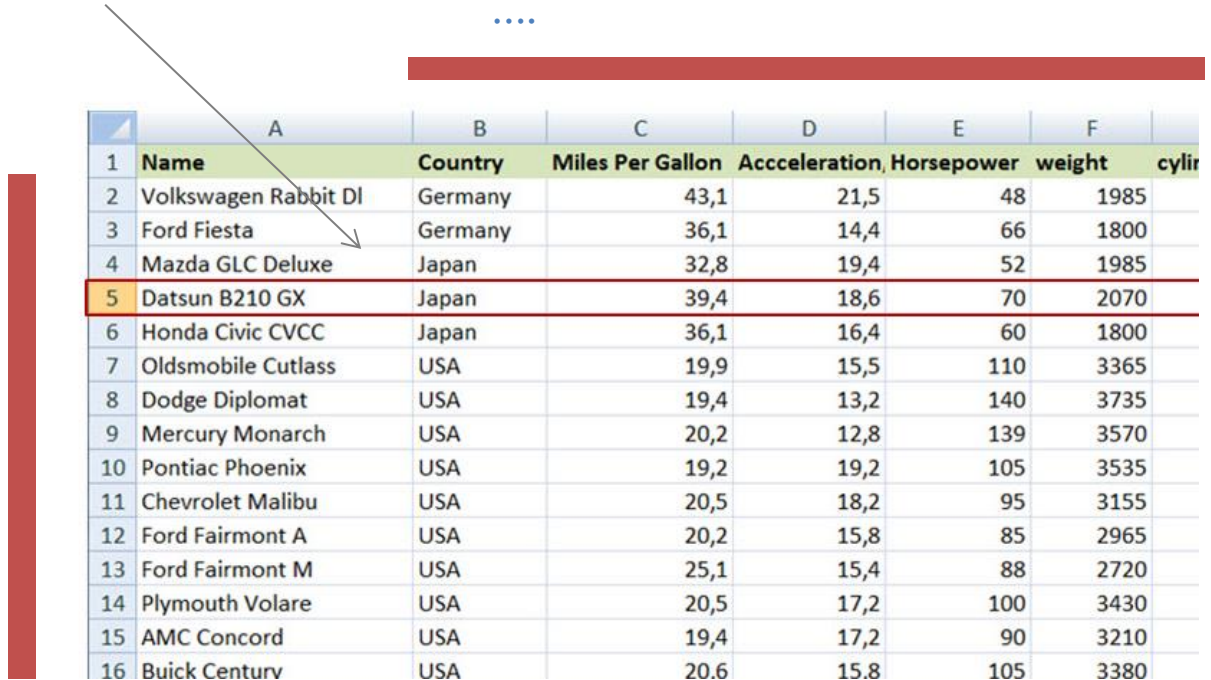
....

Car models

or Survey respondents

or Patients

....



The diagram illustrates a rectangular dataset matrix. A red vertical bar on the left represents the set of data items (rows), and a red horizontal bar at the top represents the set of attributes (columns). An arrow points from the text 'One data item' to the 5th row of the table. The table itself contains 16 rows of car data with 7 columns: Name, Country, Miles Per Gallon, Accceleration, Horsepower, weight, and cylir.

	A	B	C	D	E	F	
1	Name	Country	Miles Per Gallon	Accceleration	Horsepower	weight	cylir
2	Volkswagen Rabbit DI	Germany	43,1	21,5	48	1985	
3	Ford Fiesta	Germany	36,1	14,4	66	1800	
4	Mazda GLC Deluxe	Japan	32,8	19,4	52	1985	
5	Datsun B210 GX	Japan	39,4	18,6	70	2070	
6	Honda Civic CVCC	Japan	36,1	16,4	60	1800	
7	Oldsmobile Cutlass	USA	19,9	15,5	110	3365	
8	Dodge Diplomat	USA	19,4	13,2	140	3735	
9	Mercury Monarch	USA	20,2	12,8	139	3570	
10	Pontiac Phoenix	USA	19,2	19,2	105	3535	
11	Chevrolet Malibu	USA	20,5	18,2	95	3155	
12	Ford Fairmont A	USA	20,2	15,8	85	2965	
13	Ford Fairmont M	USA	25,1	15,4	88	2720	
14	Plymouth Volare	USA	20,5	17,2	100	3430	
15	AMC Concord	USA	19,4	17,2	90	3210	
16	Buick Centurv	USA	20.6	15.8	105	3380	

PROJECT #1: DATASET EXAMPLE

Multivariate - Quantitative data and Categorical data

Data Items

	A	B	C	D	E	F	G	H	I
1	Name	Country	Miles Per Gallon	Accceleration	Horsepower	weight	cylinders	year	price
2	Volkswagen Rabbit DI	Germany	43,1	21,5	48	1985	4	78	2400
3	Ford Fiesta	Germany	36,1	14,4	66	1800	4	78	1900
4	Mazda GLC Deluxe	Japan	32,8	19,4	52	1985	4	78	2200
5	Datsun B210 GX	Japan	39,4	18,6	70	2070	4	78	2725
6	Honda Civic CVCC	Japan	36,1	16,4	60	1800	4	78	2250
7	Oldsmobile Cutlass	USA	19,9	15,5	110	3365	8	78	3300
8	Dodge Diplomat	USA	19,4	13,2	140	3735	8	78	3125
9	Mercury Monarch	USA	20,2	12,8	139	3570	8	78	2850
10	Pontiac Phoenix	USA	19,2	19,2	105	3535	6	78	2800
11	Chevrolet Malibu	USA	20,5	18,2	95	3155	6	78	3275
12	Ford Fairmont A	USA	20,2	15,8	85	2965	6	78	2375
13	Ford Fairmont M	USA	25,1	15,4	88	2720	4	78	2275
14	Plymouth Volare	USA	20,5	17,2	100	3430	6	78	2700
15	AMC Concord	USA	19,4	17,2	90	3210	6	78	2300
16	Buick Century	USA	20,6	15,8	105	3380	6	78	3300
17	Mercury Zephyr	USA	20,8	16,7	85	3070	6	78	2425
18	Dodge Aspen	USA	18,6	18,7	110	3620	6	78	2700
19	AMC Concord D1	USA	18,1	15,1	120	3410	6	78	2425
20	Chevrolet MonteCarlo	USA	19,2	13,2	145	3425	8	78	3900
21	Buick RegalTurbo	USA	17,7	13,4	165	3445	6	78	4400
22	Ford Futura	Germany	18,1	11,2	139	3205	8	78	2525
23	Dodge Magnum XE	USA	17,5	13,7	140	4080	8	78	3000
24	Chevrolet Chevette	USA	30	16,5	68	2155	4	78	2100
25	Toyota Corona	Japan	27,5	14,2	95	2560	4	78	2975

Data types

Quantitative (Numerical)

Categorical (Ordinal)

↑
Categorical

↑
Quantitative

↑
Categorical (Ordinal)
Quantitative

SCRAPE DATA

Simple case

- Download excel sheets from a website
- Data are in a webpage table – cut & paste into a local excel sheet

Harder case

- Data are structured or semi-structured but cut & paste does not work
- Can use LLMs to get these data

LLMs are useful when data is

- Semi-structured
- Inconsistently formatted
- Human-readable but machine-hostile

PROMPT PATTERN 1

SCHEMA-FIRST EXTRACTION

When to use:

- You know what columns you want, but the page is messy
- LMM fills the table, finds values that match the columns

perl

 Copy code

You are **given** the text content of a web page.

Extract the following fields:

- **<field 1>**
- **<field 2>**
- **<field 3>**

Return the result as a JSON array.

If a **field** is missing **or** unclear, **use** null.

Do **not** infer **values** that are **not** explicitly stated.

Text:

<<<PASTE PAGE TEXT HERE>>>

PROMPT PATTERN 2

TABLE RECONSTRUCTION

When to use:

- The page looks tabular but isn't HTML-table clean
- LLM rebuilds a table, reconstruct rows and alignment

vbnet

 Copy code

The following **text** describes a table, but formatting **is** inconsistent.

Reconstruct the table **with** the following columns:

<col1>, <col2>, <col3>

Return the table **as** CSV.

Preserve original wording.

Do not normalize units **or** values.

Text:

<<<PASTE PAGE **TEXT** HERE>>>

PROMPT PATTERN 3

REPEATED ENTITY EXTRACTION

When to use:

- Lists of items (people, organizations, events, products)
- LLM discovers how many entities exist, groups attributes by entity

vbnet

 Copy code

Identify all entities of type <ENTITY TYPE> mentioned in the text.

For each entity, extract:

- name
- associated attributes explicitly stated

Return a JSON list.

Ignore commentary and opinions.

Text:

<<<PASTE PAGE TEXT HERE>>>

OTHER PROMPTS

THE POSSIBILITIES ARE ENDLESS

There is a lot more information that might be extracted

- Provenance-aware extraction
- Uncertainty
-

vbnet

 Copy code

Extract the following fields.

For each field, assign:

- confidence: high / medium / low
based only on clarity of the text.

Do not guess missing values.

Text:

<<<PASTE PAGE TEXT HERE>>>

AFTER DOWNLOADING THE DATA ...

Do you think data are always clean and perfect?

Think again

Real world data are dirty

Data cleaning (wrangling)

- fill in **missing values**
- smooth **noisy data**
- identify or remove **outliers**
- resolve **inconsistencies**
- **standardize/normalize** data
- **fuse/merge** disjoint data



MISSING VALUES

Data is not always available

- e. g, many tuples have no recorded value for several attributes, such as customer income in sales data

Missing data may be due to

- equipment malfunction
- inconsistent with other recorded data and thus deleted
- data not entered due to misunderstanding
- certain data may not be considered important at the time of entry
- many more reasons

MISSING DATA – EXAMPLE

Assume you get these baseball fan data

Age	Income	Team	Gender
23	24,200	Mets	M
39	50,245	Yankees	F
45	45,390	Yankees	F
22	32,300	Mets	M
52		Yankees	F
27	28,300	Mets	F
48	53,100	Yankees	M

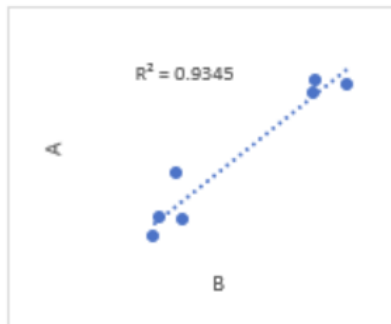
- How would you estimate the missing value for income (imputation)?
 - ignore or put in a default value (will decimate the usable data)
 - manually fill in (can be tedious or infeasible for large data)
 - use the available value of the nearest neighbors
 - average over all incomes
 - average over incomes of Yankee fans
 - average over incomes of female Yankees fans
 - use a probabilistic method (regression, Bayesian, decision tree)
 - use a neural network trained on complete data

MORE SOPHISTICATED IMPUTATION SCHEMES

MULTIPLE IMPUTATION BY CHAINED EQUATIONS (MICE)

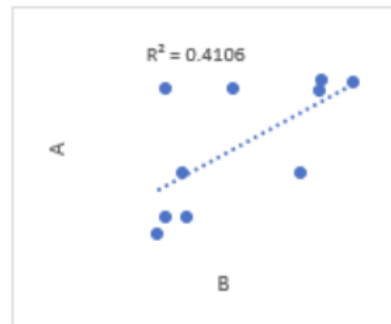
Missing data is in red. There is a strong correlation between A and B, so let's try to impute A using B and C.

A	B	C
0.93	1.40	1.53
0.24	0.46	0.76
	0.80	
0.95	1.24	1.46
0.23	0.57	
0.90		1.28
0.15	0.42	
0.47	0.54	0.63
	1.14	
0.89	1.23	1.45



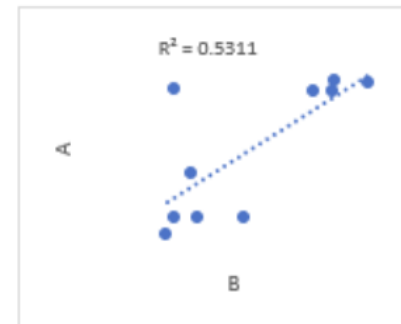
Missing data is filled in randomly. This dilutes the correlations, but allows us to impute using all available data.

A	B	C
0.93	1.40	1.53
0.24	0.46	0.76
	0.80	
0.95	1.24	1.46
0.23	0.57	
0.90		1.28
0.15	0.42	
0.47	0.54	0.63
	1.14	
0.89	1.23	1.45



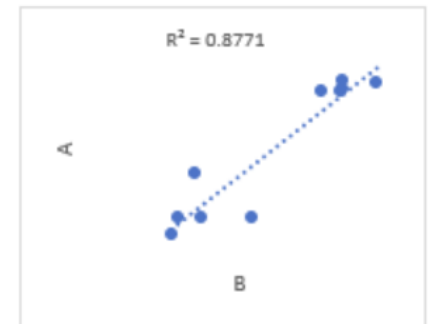
A random forest is used to predict A with B and C. Notice the correlation between A and B improved.

A	B	C
0.93	1.40	1.53
0.24	0.46	0.76
	0.80	
0.95	1.24	1.46
0.23	0.57	
0.90		1.28
0.15	0.42	
0.47	0.54	0.63
	1.14	
0.89	1.23	1.45



After Imputing B using A and C, we have achieved a correlation between A and B much closer to the original data.

A	B	C
0.93	1.40	1.53
0.24	0.46	0.76
	0.80	
0.95	1.24	1.46
0.23	0.57	
0.90		1.28
0.15	0.42	
0.47	0.54	0.63
	1.14	
0.89	1.23	1.45

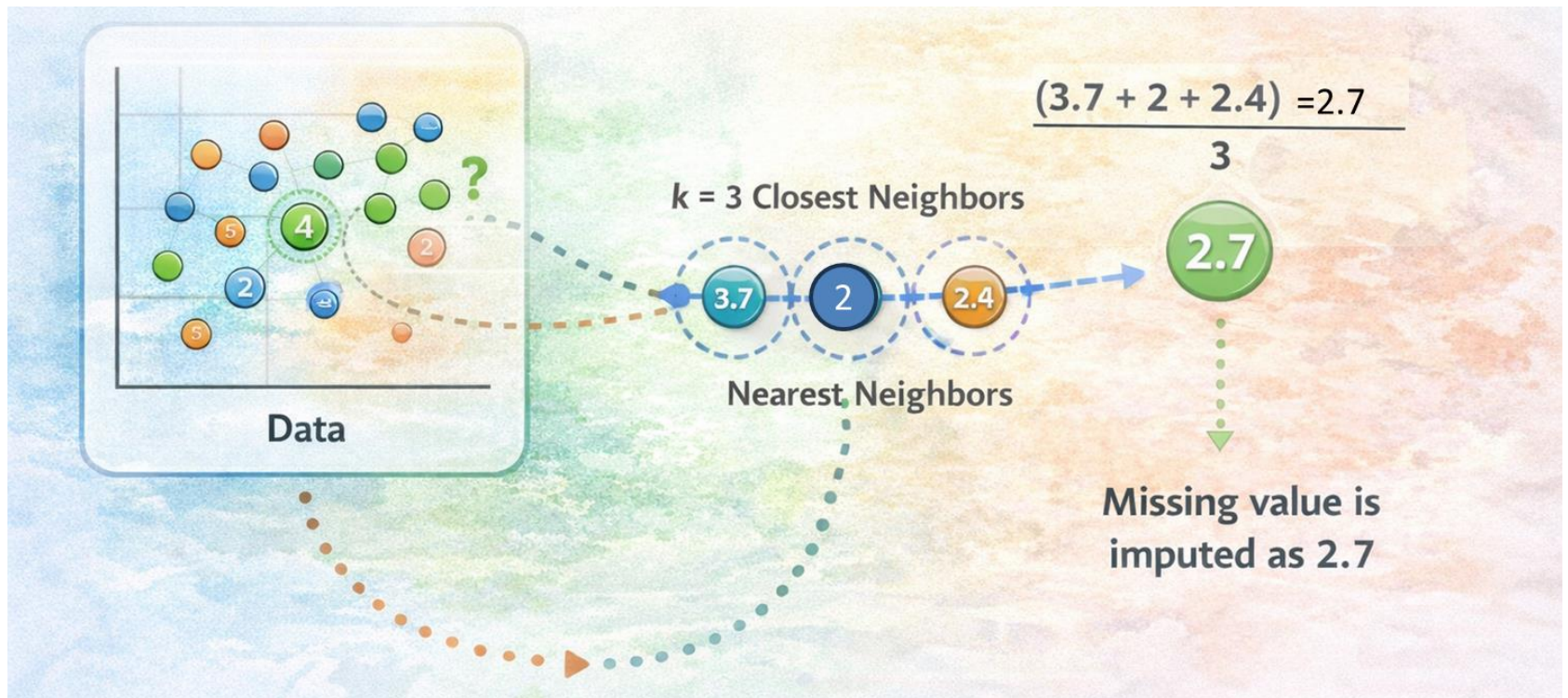


continue till all specified variables have been imputed, may need to do more iterations (<5)

<https://cran.r-project.org/web/packages/miceRanger/vignettes/miceAlgorithm.html>

KNN IMPUTATION

Replaces missing values with a distance-weighted average of the k most similar observations (using known features)



MISSING DATA CONSEQUENCES

MCAR (Missing Completely At Random)

- No relationship between missing data and any values
- Imputation meaningless

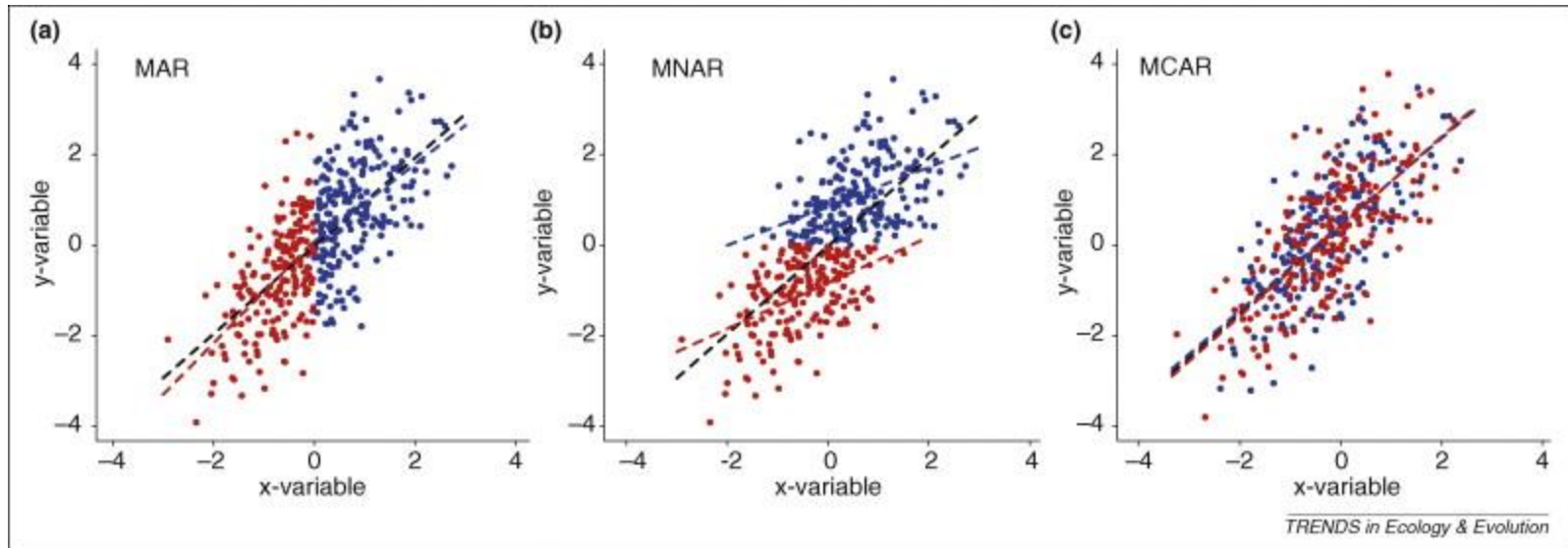
MAR (Missing At Random)

- Missingness is related to observed data
- Sweet spot for imputation as it depends on observed data

MNAR (Missing Not At Random)

- Missingness is related to the missing values themselves
- Examples
 - Health: Individuals with severe symptoms are less likely to complete a follow-up survey.
 - Income: High-income individuals are more likely to skip questions about their income.
- Imputation is unreliable as it depends on unobserved values

MAR, MNAR, MCA COMPARISON



Missingness depends on x

Missingness depends
on x and y.

Missingness does not
depend on x or y.

DATA TRANSFORMATION

Can help reduce influence of extreme values

See our discussion last lecture

DATA NORMALIZATION

Sometimes we like to have all variables on the same scale

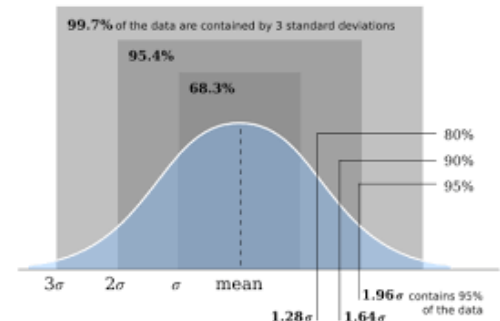
- min-max normalization

$$v' = \frac{v - \min}{\max - \min}$$

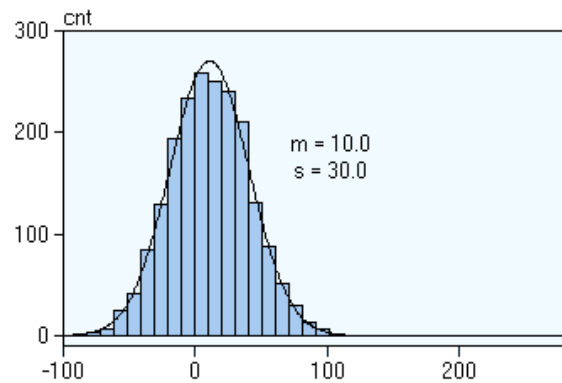
- standardization / z-score normalization

$$v' = \frac{v - \bar{v}}{\sigma_v}$$

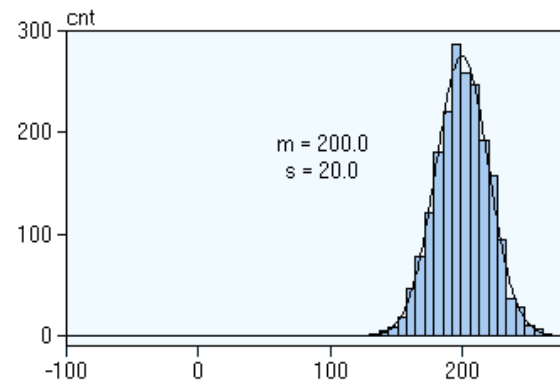
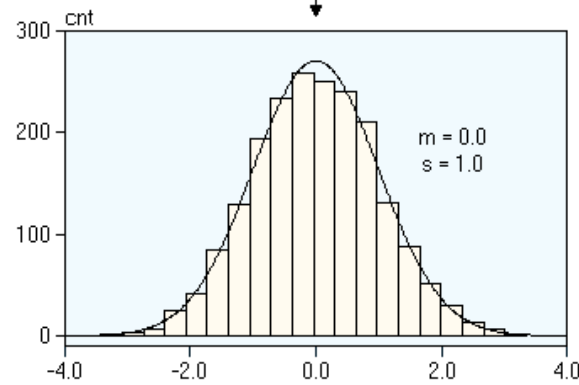
- clipping tails and outliers
 - set all values beyond $\pm 3\sigma$ to value at 3σ
 - set values $<5\%$ ($>95\%$) to value at 5% (95%)



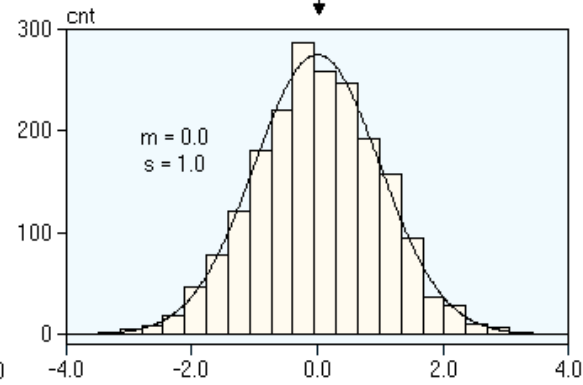
STANDARDIZATION



Standardisation

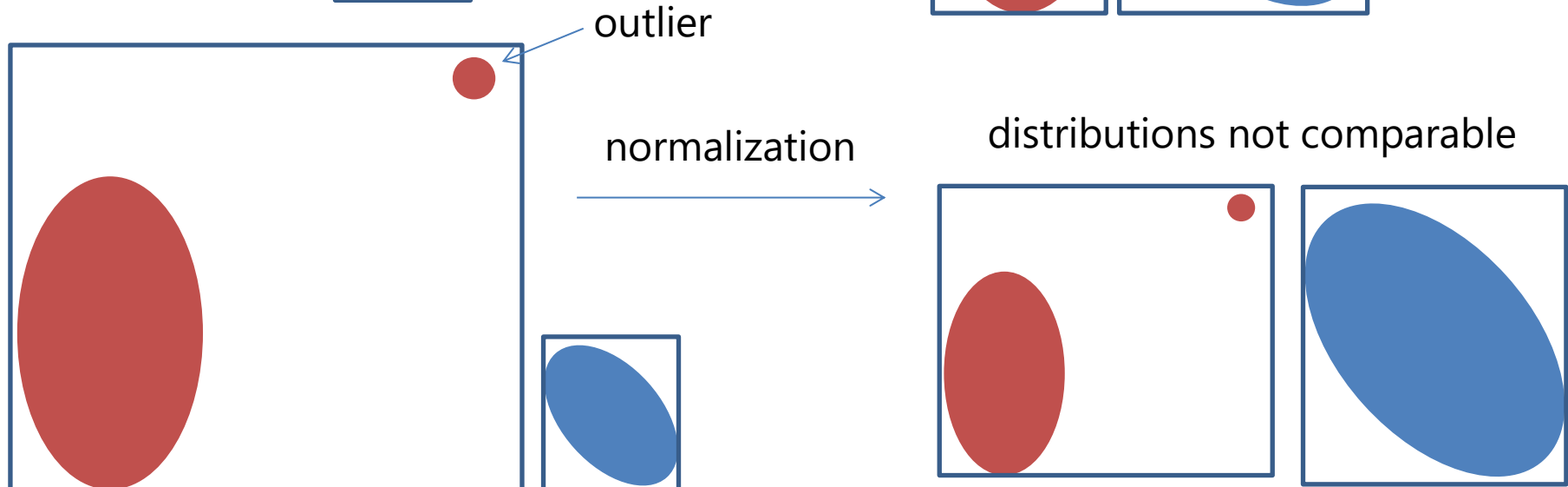
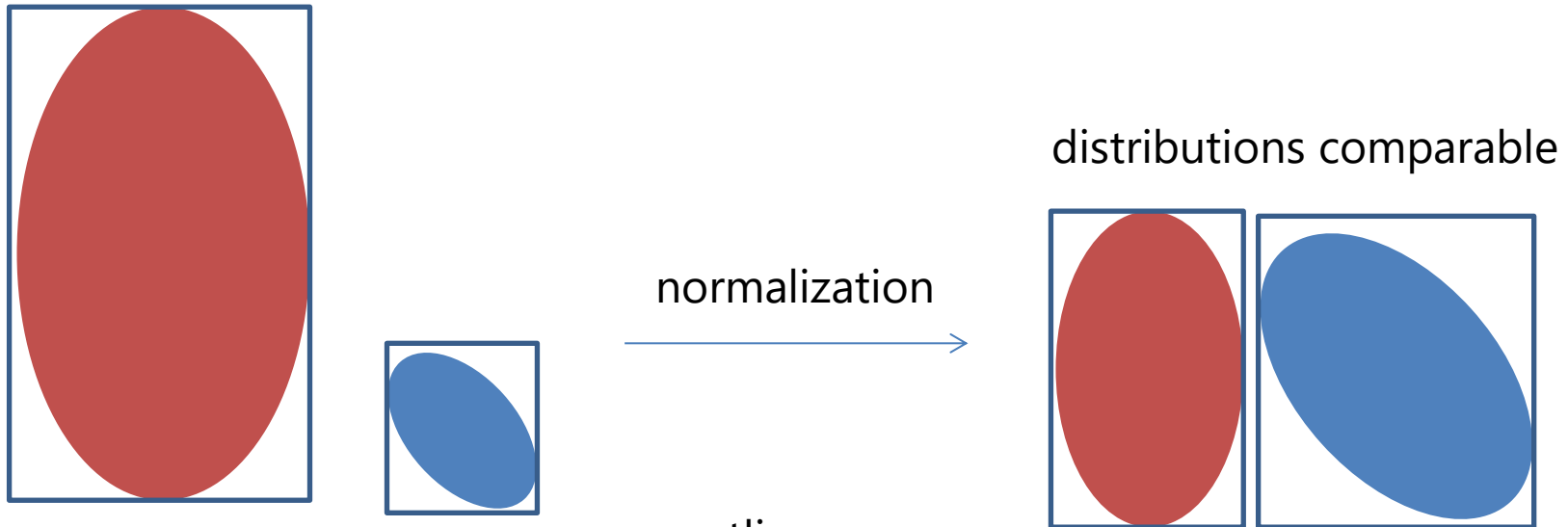


Standardisation



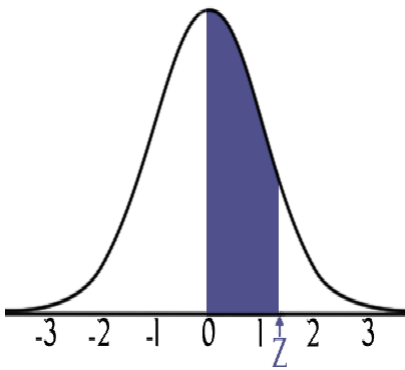
comparable distributions
($m = 0.0$, $s = 1.0$)

NORMALIZATION

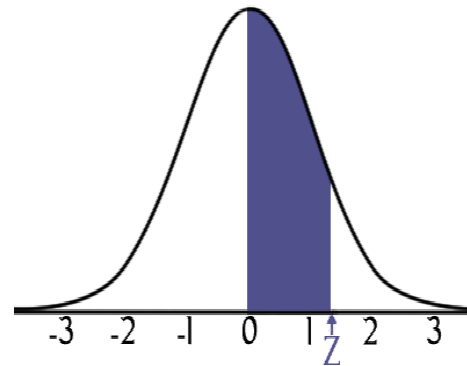


STANDARDIZATION

Is standardization less or more sensitive to outliers?



without outlier



with outlier (just slightly extended)

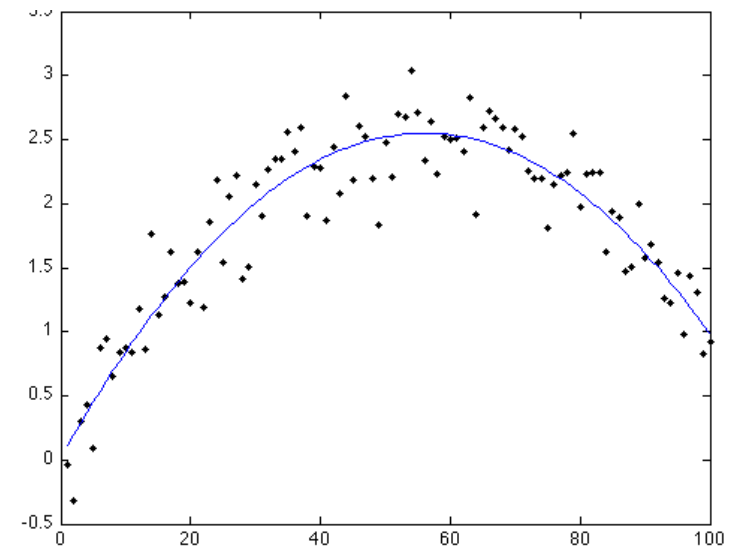
But you need to set a reasonable cut-off point on each side

- normal distributions are infinite

NOISY DATA

Noise = Random error in a measured variable

- faulty data collection instruments
- data entry problems
- data transmission problems
- technology limitation
- inconsistency in naming convention



Other data problems which require data cleaning

- duplicate records
- incomplete data
- inconsistent data

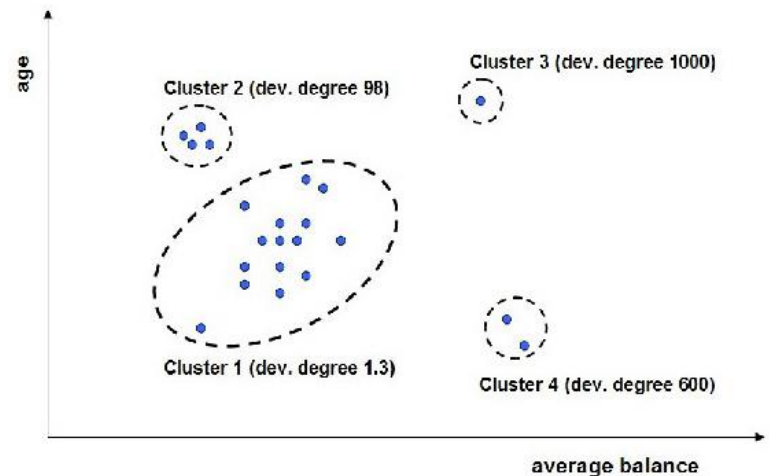
NOISY DATA – WHAT TO DO

Binning method

- discussed last lecture

Clustering

- detect and remove outliers

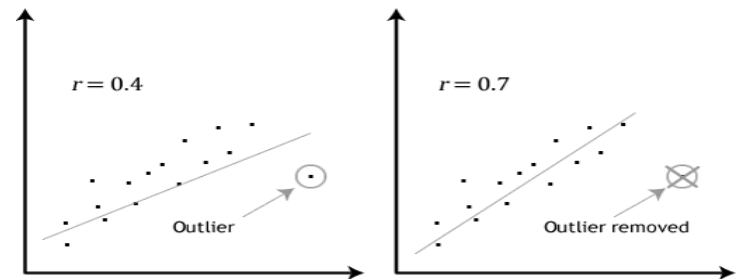


Semi-automated method

- combined computer and human inspection
- detect suspicious values and check manually (need visualization)

Regression

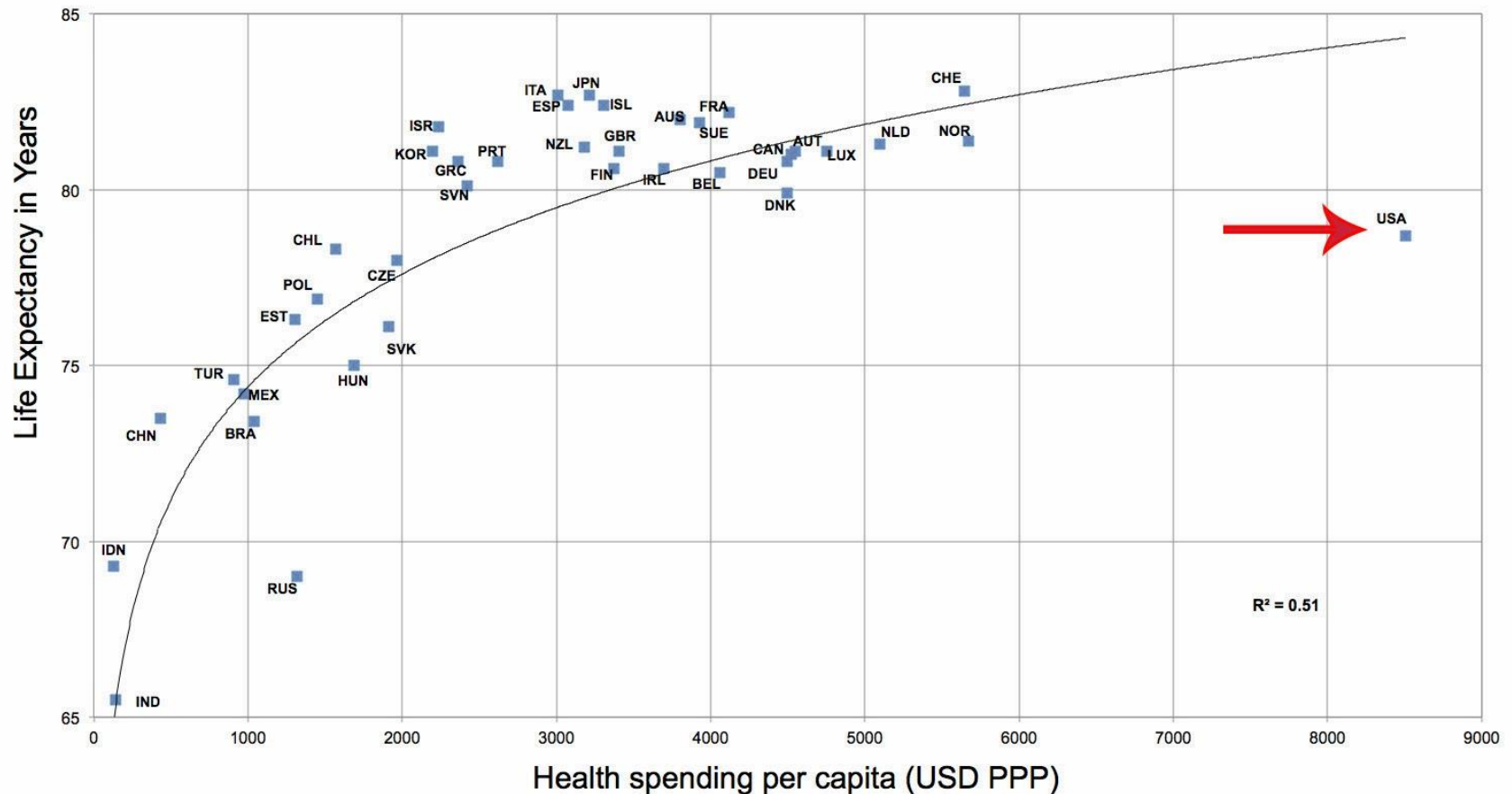
- smooth by fitting the data to a regression function



NOISE REMOVAL – A WORD OF CAUTION

An outlier may not be noise

- it may be an anomaly that is very valuable (e.g., the Higgs particle)



RESOLVE INCONSISTENCIES

Inconsistencies in naming conventions or data codes

- e.g., 2/5/2002 could be 2 May 2002 or 5 Feb 2002

Redundant data

- duplicate tuples, which were received twice should be removed

DATA INTEGRATION

Data integration/fusion

- multiple databases
- data cubes
- files
- notes

Produces new opportunities

- can gain more comprehensive insight (value > sum of parts)
- but watch out for *synonymy and polysemy*
- attributes with different labels may have the same meaning
 - “comical” and “hilarious”
- attributes with the same label may have different meaning
 - “jaguar” can be a cat or a car

DATA FUSION – How To (1)

Goal is to add new thematic aspects

- enable deeper and more far-fetching insights
- can open valuable opportunities for research and \$\$\$

How to do it

- start off with a first dataset, such as a set of listings of houses
- ask, what would house buyers be interested in?
 - education for children (school quality, pre-K, ...)
 - quality of life (entertainment, socializing, fitness, clean air, ...)
 - infrastructure (shopping, airport, roads, ...)
 - what else?
- make Google your best friend
 - determine a good attribute to link to other datasets (e.g. zip code)
 - then ask “primary education by zip code” or “livability by zip code”

EXAMPLE: FUSING DIFFERENT THEMATIC DATASETS

House listing data					
Address	Size	Bedrooms	Baths	Price	Zip Code
5 Nut Str.	2,345 sqft	3	1	\$564k	11794
Education by zip code	Zip Code	School Name	Avg. SAT	Class Size	Cost
	11794	Tree Top	1060	34	Public
Quality of life by zip code	Zip Code	Livability Score	Distance to Airport	Air Quality Score	Electricity Cost
	11794	63	45 miles	89	\$0.34/KW

Make sure that all data are from the same/similar year (when time matters)

Might need different keys for linking different thematic datasets

- for example zip code, state, county, and so on
- find associations for each in all tables and fuse
- duplicate information for coarse grained tables in finer-grained tables

BUT DATA INTEGRATION CAN ALSO BRING
ETHICAL PROBLEMS

PRIVACY

Can you identify a person from these medical records?

SSN	Name	Race	Date Of Birth	Sex	ZIP	Marital Status	Health Problem
		asian	9/27/64	female	94139	divorced	hypertension
		asian	9/30/64	female	94139	divorced	obesity
		asian	4/18/64	male	94139	married	chest pain
		asian	4/15/64	male	94139	married	obesity
		black	3/13/63	male	94138	married	hypertension
		black	3/18/63	male	94138	married	shortness of breath
		black	9/13/64	female	94141	married	shortness of breath
		black	9/7/64	female	94141	married	obesity
		white	5/14/61	male	94138	single	chest pain
		white	05/08 61	male	94138	single	obesity
		white	9/15/61	female	94142	widow	shortness of breath

PRIVACY

What if you had a voter list

Name	Address	City	ZIP	DOB	Sex	Party	
Sue J. Carlson	900 Market St.	San Francisco	94142	9/15/61	female		

SSN	Name	Race	Date Of Birth	Sex	ZIP	Marital Status	Health Problem
		asian	9/27/64	female	94139	divorced	hypertension
		asian	9/30/64	female	94139	divorced	obesity
		asian	4/18/64	male	94139	married	chest pain
		asian	4/15/64	male	94139	married	obesity
		black	3/13/63	male	94138	married	hypertension
		black	3/18/63	male	94138	married	shortness of breath
		black	9/13/64	female	94141	married	shortness of breath
		black	9/7/64	female	94141	married	obesity
		white	5/14/61	male	94138	single	chest pain
		white	05/08/61	male	94138	single	obesity
		white	9/15/61	female	94142	widow	shortness of breath

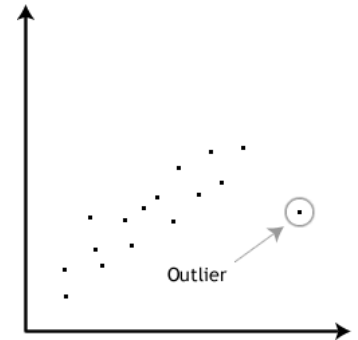
DATA FUSION VS. DATA PRIVACY

Data fusion can bring insight

- the purpose is not always good
- but often it is (criminal justice, market analysis,)

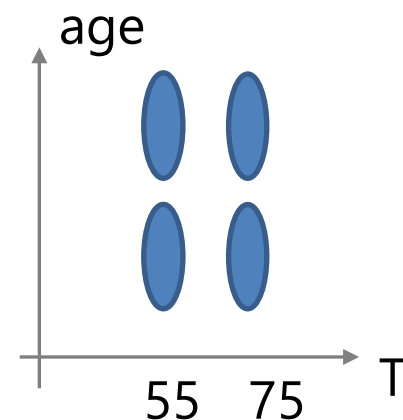
Visualization can bring insight

- the 94142 zip code would have been an outlier
- your visualization would have shown that nicely
- then you could have dug for complementary data



How to obfuscate for protection?

- k-anonymity (generalize)
- make data less specific → use binning
- age *groups*, zip code *groups*, etc...
- make blobs instead of points



THE NEED FOR DATA REDUCTION

Purpose

- reduce the data to a size that can be feasibly stored
- reduce the data so a mining algorithm can be feasibly run

Alternatives

- buy more storage
- buy more computers or faster ones
- develop more efficient algorithms (look beyond O-notation)

In practice, all of this is happening at the same time

- but the growth of data and complexities is faster
- and so data reduction is important

DATA REDUCTION

Sampling

- random
- stratified



Data summarization

- binning (already discussed)
- clustering (see a future lecture)
- dimension reduction (see next lecture)

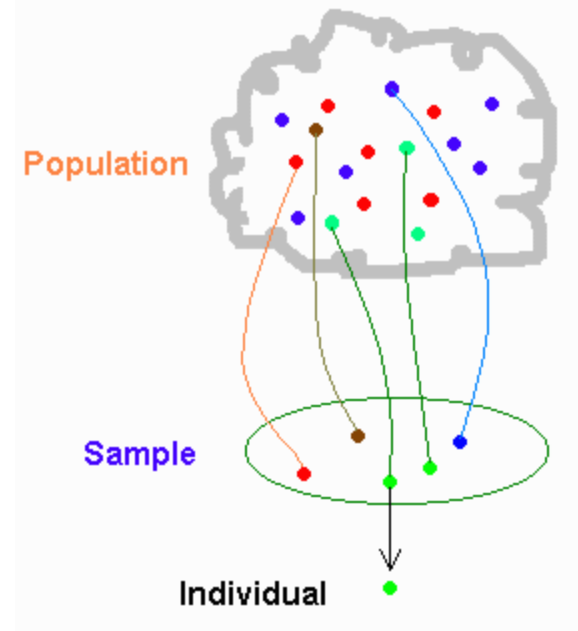
SAMPLING

The goal

- pick a representative subset of the data

Random sampling

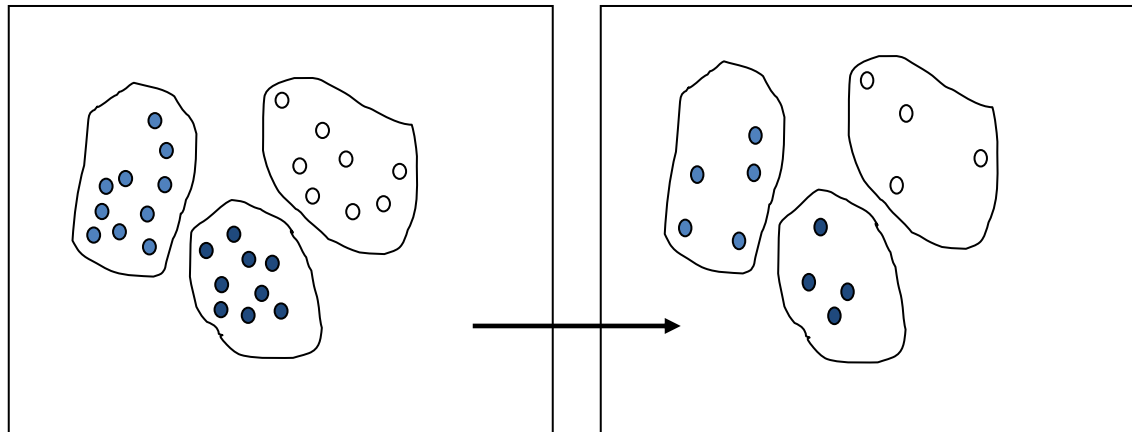
- pick sample points at random
- will work if the points are distributed uniformly
- this is usually not the case
- outliers will likely be missed
- so the sample will not be representative



ADAPTIVE SAMPLING

Pick the samples according to some knowledge of the data distribution

- create a binning of some sort (outliers will form bins as well)
- also called *strata* (stratified sampling)
- the size of each bin represents its percentage in the population
- it guides the number of samples – bigger bins get more samples

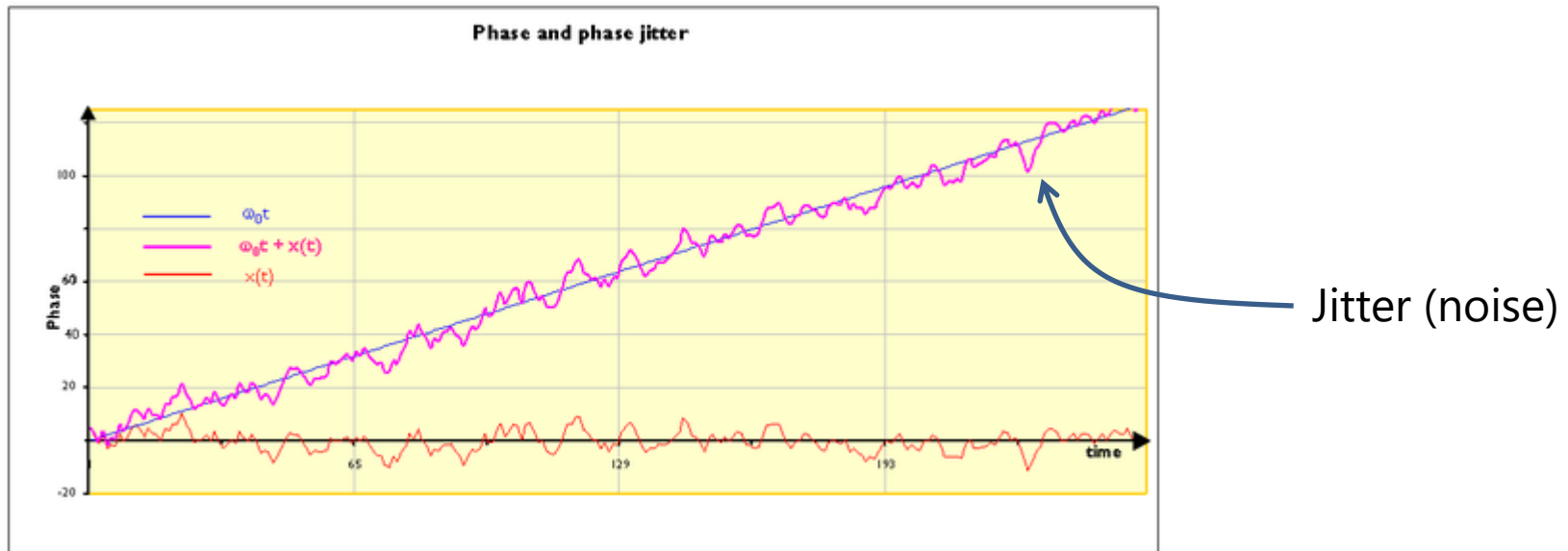


sampling rate $\sim$ cluster size

WHAT'S JITTERING?

Definition from dictionary

- act nervously
- "an anxious student who jittered at any provocation"

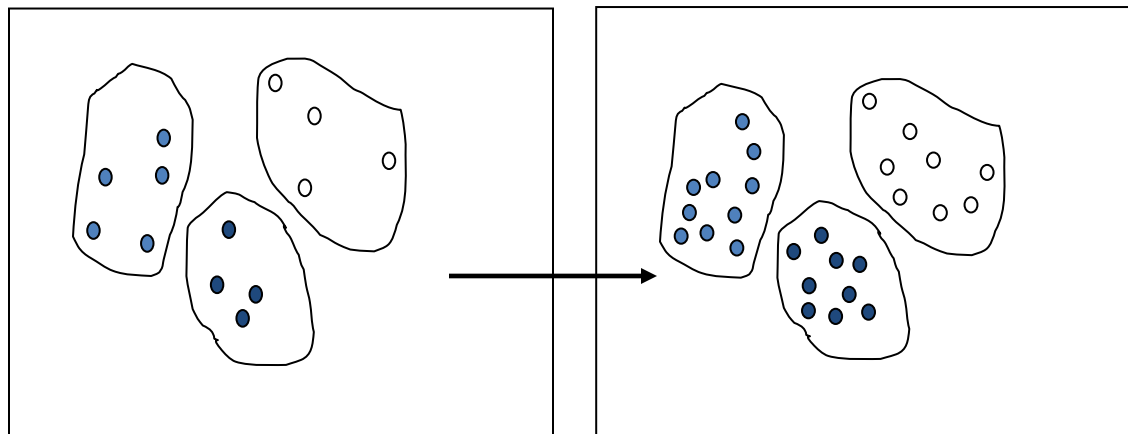


- small random noise about a steady signal

DATA AUGMENTATION FOR VISUALIZATION & VISUAL ANALYTICS

Generate new samples according to the data distributions

- cluster the data (outliers will form clusters as well)
- the size of each cluster represents its percentage in the population
- randomize new samples – bigger clusters get more samples
- add a small randomized value to either the mean or an existing sample
- do this for every dimension of the chosen mean or sample



augmentation rate $\sim$ cluster size

TODAY'S TAKE AWAYS

How to deal with

- missing data
- noisy data and outliers
- uneven and diverse data ranges

Various strategies for segmenting data to

- visualize overall trends and groups
- reduce data
- augment/enrich data
- enable techniques to ensure privacy

Enrich datasets by adding other thematic aspects

- obtained by additional attributes from other sources
- determine proper key attributes helpful for linking