

CSE 332

INTRODUCTION TO VISUALIZATION

DATA SOURCES AND PREPARATION

KLAUS MUELLER

COMPUTER SCIENCE DEPARTMENT  
STONY BROOK UNIVERSITY

Lecture	Topic	Projects
1	Intro, schedule, and logistics	
2	Applications of visual analytics, basic tasks, data types	
3	Introduction to D3, basic vis techniques for non-spatial data	Project #1 out
4	Data assimilation and preparation	
5	Data reduction and notion of similarity and distance	
6	Visual perception and cognition	
7	Visual design and aesthetics	Project #1 due
8	Statistics foundations	Project #2 out
9	Data mining techniques: clusters, text, patterns, classifiers	
10	Data mining techniques: clusters, text, patterns, classifiers	
11	Computer graphics and volume rendering	
12	Techniques to visualize spatial (3D) data	Project #2 due
13	Scientific and medical visualization	Project #3 out
14	Scientific and medical visualization	
15	Midterm #1	
16	High-dimensional data, dimensionality reduction	Project #3 due
17	Big data: data reduction, summarization	
18	Correlation and causal modeling	
19	Principles of interaction	
20	Visual analytics and the visual sense making process	Final project proposal due
21	Evaluation and user studies	
22	Visualization of time-varying and time-series data	
23	Visualization of streaming data	
24	Visualization of graph data	Final Project preliminary report due
25	Visualization of text data	
26	Midterm #2	
27	Data journalism	
	Final project presentations	Final Project slides and final report due

# 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 features (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 the 5th row highlighted in red.

	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

Find some interesting data on the web

- something that challenges and interests you
- there are many data sources on the web
- use google and some imagination

Criteria for selection

- more than 500 data points (observations)
- more than 10 attributes
- the more the better (you can always reduce it)

Deliverables

- 2-page report that describes the data and justifies your choice
- a URL to the data source

Due date

- Tuesday, September 18, 11:59pm

# 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



# SOME GOOD SOURCES FOR DATA

[Kaggle](#) – lots of data for data science

[NYC Open Data](#) – all kinds of data related to NYC operations

[Kaiser Foundation](#) – numerous data related to public health

[Data.gov](#) – open data site with US government data

[Forbes](#) – site with links to data sites

[Data Quest](#) – another site with links to data sites

[Quandl](#) – mostly financial and economics data

[Open Data Inception](#) – map w/data portals around the world

[World Bank](#) – collection of global development data

[UCI repository](#) – site that has been around for a long time

[Analytics Vidhya](#) – another site with many links to data sites

Wikipedia also has lots of data in tables

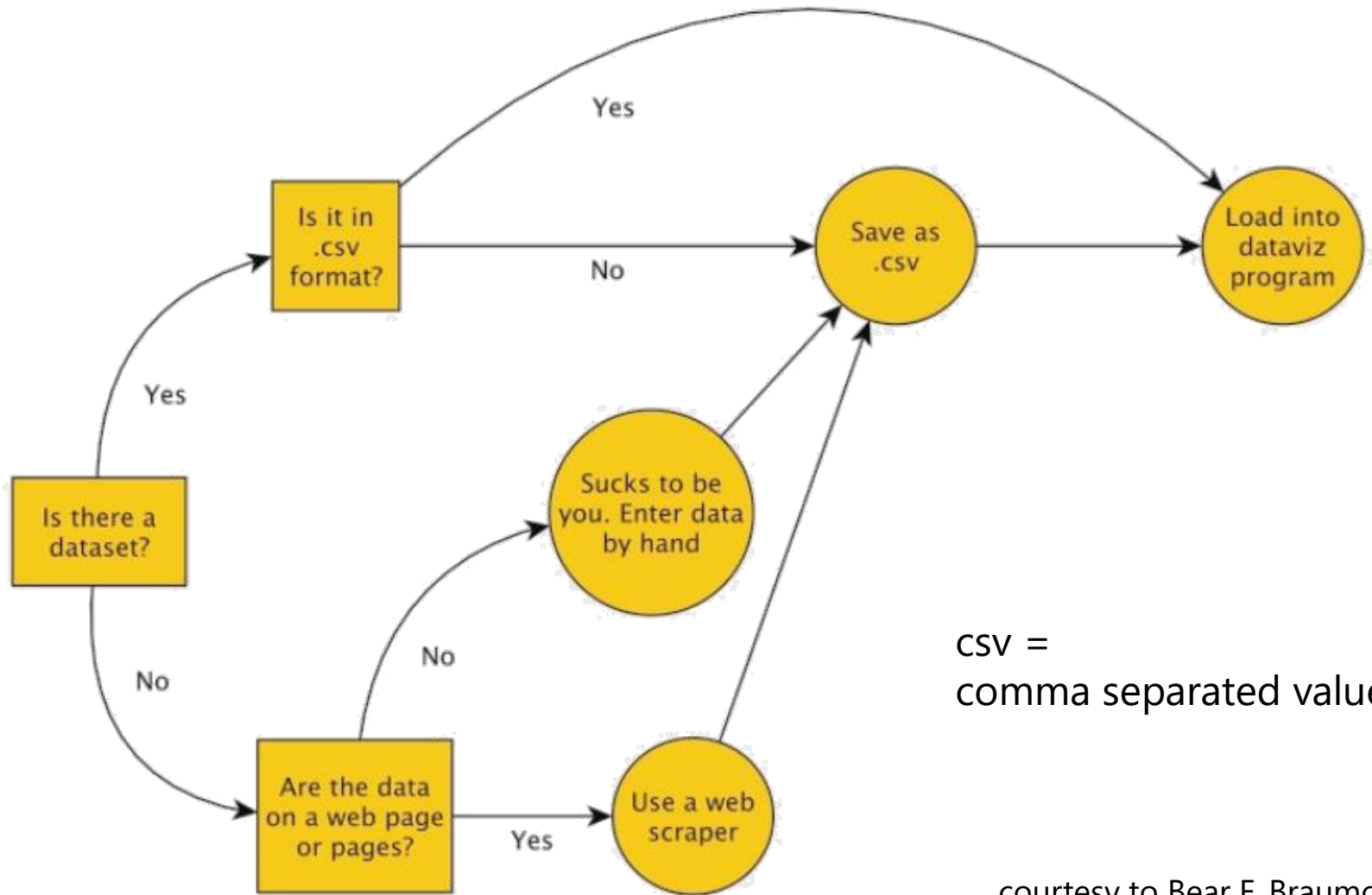
# NOTES ON DATASET

## Some advice

- avoid datasets where the attributes only have a few different values
  - this applies to both categorical and numerical attributes
- convert textual categories into numbers by assigning a numerical ID
- aim for datasets with more than 500 data points and 10 attributes
- if your dataset is larger, pick 500 sample points at random (for now)
- if you have too many attributes keep the ones of interest
  - you could use the ones least correlated
- fuse multiple datasets together to get a more holistic analysis and interesting results (recall the scientific method loop)
  - highly recommended
  - for example: data on crime stats of a city plus data on education and data on demographics (3 files from different sources)
- produce a spreadsheet of rows (data items) and attributes (columns)



# How To Import DATA?



CSV =  
comma separated values file

courtesy to Bear F. Braumoeller

# TABLES ON WEBPAGES

If the data are already in a rectangular table

- try cut and paste into Excel

If the data are on one page but cut/paste is not working

- try a web scraper like [Outwit Hub](#)

If the data are spread across multiple webpages

- try Outwit Hub's automators
- use python
- do it by hand (probably not)

## Table 1. Consumer Price Index for All Urban Consumers (CPI-U): U. S. city average, by expenditure category

**Table 1. Consumer Price Index for All Urban Consumers (CPI-U): U.S. city average, by expenditure category, February 2013**  
[1982-84=100, unless otherwise noted]

Expenditure category	Relative importance Jan. 2013	Unadjusted indexes			Unadjusted percent change		Seasonally adjusted percent change		
		Feb. 2012	Jan. 2013	Feb. 2013	Feb. 2012- Feb. 2013	Jan. 2013- Feb. 2013	Nov. 2012- Dec. 2012	Dec. 2012- Jan. 2013	Jan. 2013- Feb. 2013
All items	100.000	227.663	230.280	232.166	2.0	0.8	0.0	0.0	0.7
Food	14.327	232.486	236.341	236.301	1.6	0.0	0.2	0.0	0.1
Food at home	8.622	231.180	234.240	234.033	1.2	-0.1	0.2	0.0	0.1
Cereals and bakery products	1.232	267.821	269.078	269.304	0.6	0.1	0.2	0.1	-0.2
Meats, poultry, fish, and eggs	1.951	228.610	232.461	233.041	1.9	0.2	0.1	0.0	0.5
Dairy and related products(1)	0.906	219.377	220.319	219.526	0.1	-0.4	0.2	0.4	-0.4
Fruits and vegetables	1.306	281.072	293.714	293.742	4.5	0.0	0.3	0.3	1.4
Nonalcoholic beverages and beverage materials	0.948	169.758	169.593	168.977	-0.5	-0.4	0.2	-0.5	0.0
Other food at home	2.280	204.001	205.387	204.763	0.4	-0.3	0.2	-0.2	-0.6
Food away from home(1)	5.705	235.603	240.713	240.930	2.3	0.1	0.1	0.1	0.1
Energy	9.580	242.663	234.624	248.146	2.3	5.8	-0.8	-1.7	5.4
Energy commodities	5.793	310.685	292.609	320.258	3.1	9.4	-1.5	-3.0	8.6
Fuel oil(1)	0.233	384.747	381.889	393.782	2.3	3.1	0.0	-0.2	3.1
Motor fuel	5.460	206.240	200.100	216.500	2.2	0.0	1.6	2.2	0.0

the Catch (0)

Detail

id R P Collection Time

Source Url

Load webpage into Outwit

Rating  Priority  ☐ Save incoming files

Empty

Export

Local IP: 75.60.206.216 Remote IP: 146.142.4.22

Table 1. Consumer Price Index for All Urban Consumers (CPI-U) average, by expenditure category - (41 HTML table rows)

id	Url	Expenditure Category	Relative Importance Jan 2013	Unadjusted Indexes	Unadjusted Indexes 2	Unadjusted Indexes 3
1		All items	100.000	227.663	230.280	232.166
2		Food	14.327	232.486	236.341	236.301
3		Food at home	8.622	231.180	234.240	234.033
4		Cereals and bakery products	1.232	267.821	269.078	269.304
5		Meats, poultry, fish, and eggs	1.951	228.610	232.461	233.041
6		Dairy and related products(1)	0.906	219.377	220.319	219.526
7		Fruits and vegetables	1.306	281.072	293.714	293.742
8		Nonalcoholic beverages and beverage materi...	0.948	169.758	169.593	168.977
9		Other food at home	2.280	204.001	205.387	204.763
10		Food away from home(1)	5.705	235.603	240.713	240.930
11						
12		Energy	9.580	242.663	234.624	248.146
13		Energy commodities	5.793	310.685	292.609	320.258
14		Fuel oil(1)	0.233	384.747	381.889	393.782
15		Motor fuel	5.460	306.348	288.108	316.580
16		Gasoline (all types)	5.273	305.076	286.417	315.243
17		Energy services(2)	3.787	187.962	189.444	189.679
18		Electricity(2)	2.881	193.183	194.525	194.739
19		Utility (piped) gas service(2)	0.906	169.753	171.597	171.888
20						
21		All items less food and energy	76.093	227.865	231.612	232.432
22		Commodities less food and energy commodi...	19.530	146.628	146.492	147.093
23		Apparel	3.526	123.312	124.687	126.303



Select row if Expenditure Category does not contain

Catch

Limit to

999

Options

☒ Clean Text ☐ Deduplicate

On page load

☒ Empty ☐ Catch selection

the Catch (0)

id	R	P	Collection Time	Source Url	Url	Detail
					Expenditure Category	NOTE: Index applies to a month as a whole, not to any specific date.
					Relative Importance Jan 2013	
					Unadjusted Indexes	
					Unadjusted Indexes 2	
					Unadjusted Indexes 3	
					Unadjusted Percent Change	
					Unadjusted Percent Change 2	

Select table

Rating ☐ Priority ☐ ☐ Save incoming files

Empty

Export

Local IP: 75.60.206.216 Table 1. Consumer Price Index for All Urban Consumers (CPI-U) - All items, by expenditure category - (41 HTML table rows) Remote IP: 146.142.4.22

id	Url	Expenditure Category	Relative Importance Jan 2013	Unadjusted Indexes	Unadjusted Indexes 2	Unadjusted Indexes 3
1		All items	100.000	227.663	230.280	232.166
2		Food	14.327	232.486	236.341	236.301
3		Food at home	8.622	231.180	234.240	234.033
4		Cereals and bakery products	1.232	267.821	269.078	269.304
5		Meats, poultry, fish, and eggs	1.951	228.610	232.461	233.041
6		Dairy and related products(1)	0.906	219.377	220.319	219.526
7		Fruits and vegetables	1.306	281.072	293.714	293.742
8		Nonalcoholic beverages and beverage materi...	0.948	169.758	169.593	168.977
9		Other food at home	2.280	204.001	205.387	204.763
10		Food away from home(1)	5.705	235.603	240.713	240.930
11		Energy	9.580	242.663	234.624	248.146
12		Energy commodities	5.793	310.685	292.609	320.258
13		Fuel oil(1)	0.233	384.747	381.889	393.782
14		Motor fuel	5.460	306.348	288.108	316.580
15		Gasoline (all types)	5.273	305.076	286.417	315.243
16		Energy services(2)	3.787	187.962	189.444	189.679
17		Electricity(2)	2.881	193.183	194.525	194.739
18		Utility (piped) gas service(2)	0.906	169.753	171.597	171.888
19						
20						
21		All items less food and energy				232.432
22		Commodities less food and energy				147.093
23		Apparel				126.303

Select row if Expenditure Category does not contain Limit to 999 Options Clean Text Deduplicate On page load Empty Catch selection

Catch data, then cut to CSV

the Catch (41)

id	R	P	Collection Time	Source Url	Url	Expenditure Category
17			03/15/2013 15:32:17	http://www.bls.gov/news.release/cpi.t01.htm		Energy services(2)
18			03/15/2013 15:32:17	http://www.bls.gov/news.release/cpi.t01.htm		Electricity(2)
19			03/15/2013 15:32:17	http://www.bls.gov/news.release/cpi.t01.htm		Utility (piped) gas service(2)
20			03/15/2013 15:32:17	http://www.bls.gov/news.release/cpi.t01.htm		
21			03/15/2013 15:32:17	http://www.bls.gov/news.release/cpi.t01.htm		All items less food and energy
22			03/15/2013 15:32:17	http://www.bls.gov/news.release/cpi.t01.htm		Commodities less food and energy co
23			03/15/2013 15:32:17	http://www.bls.gov/news.release/cpi.t01.htm		Apparel
24			03/15/2013 15:32:17	http://www.bls.gov/news.release/cpi.t01.htm		New vehicles
25			03/15/2013 15:32:17	http://www.bls.gov/news.release/cpi.t01.htm		Used cars and trucks

Excel  
CSV  
TXT  
HTML  
SQL



OutWit catch export - Table\_1\_Consumer\_Price\_Index\_for\_All\_Urban\_Consumers\_CPI\_U\_U\_S\_city\_average\_by\_expenditure\_category.csv

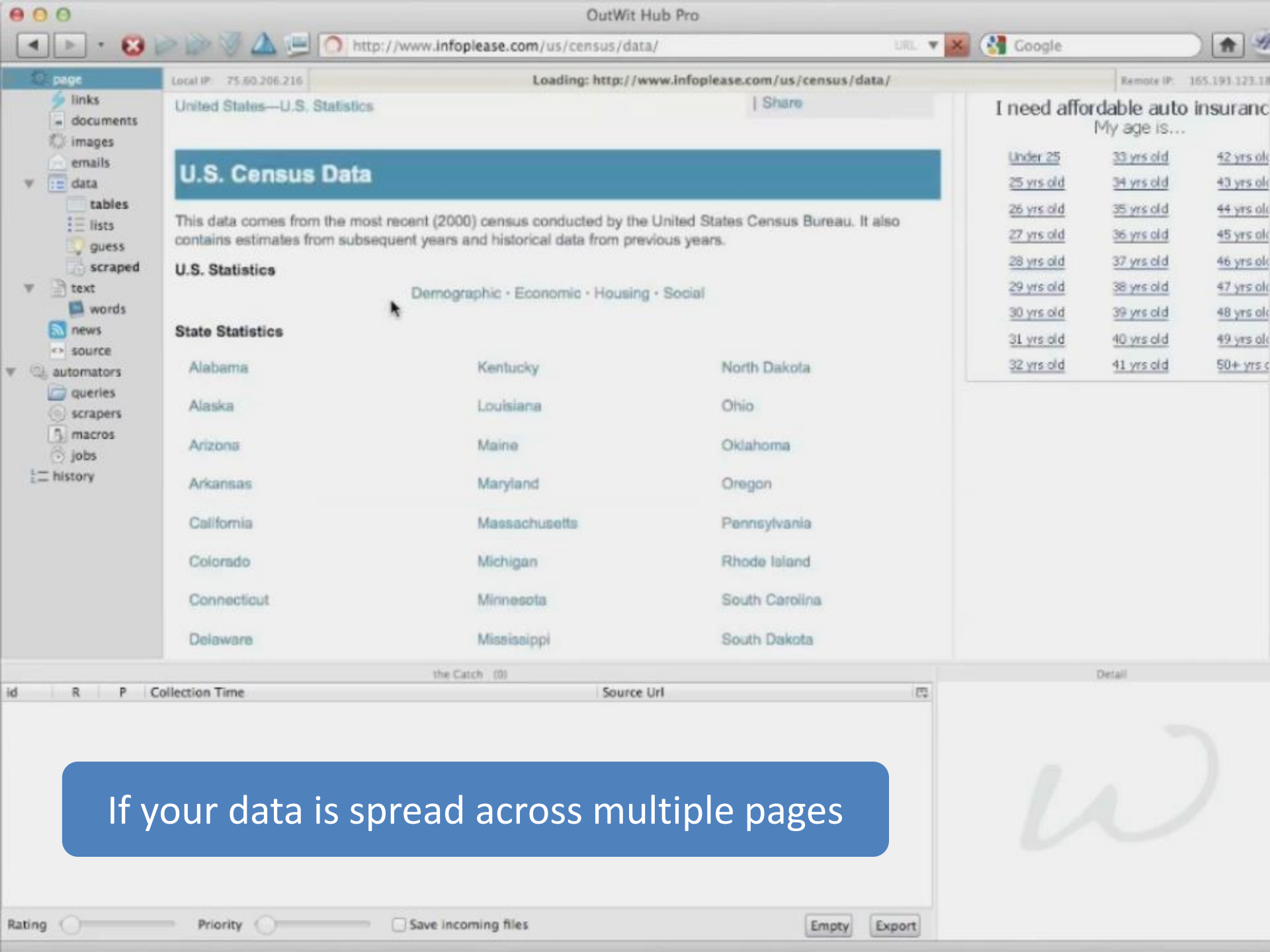
Home Layout Tables Charts SmartArt Formulas Data Review

Edit Font Alignment Number Format Cells Themes

Calibri (Body) 12 A A abc Wrap Text General Normal Bad Insert Delete Format Themes Aa

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S
			Collection Title	Source URL	Url	Expenditure	Relative Imp	Unadjusted I	Unadjusted I	Unadjusted I	Unadjusted F	Unadjusted F	Seasonally A	Seasonally A	Seasonally Adjusted	Percent Change	3		
2			*****	http://www.bls.gov/news.rele	All Items		100	227.663	230.28	232.166		2	0.8	0	0	0.7			
3			*****	http://www.bls.gov/news.rele	Food		14.327	232.486	236.341	236.301		1.6	0	0.2	0	0.1			
4			*****	http://www.bls.gov/news.rele	Food at hom		8.622	231.18	234.24	234.033		1.2	-0.1	0.2	0	0.1			
5			*****	http://www.bls.gov/news.rele	Cereals and b		1.232	267.821	269.078	269.304		0.6	0.1	0.2	0.1	-0.2			
6			*****	http://www.bls.gov/news.rele	Meats, poultr		1.951	228.61	232.461	233.041		1.9	0.2	0.1	0	0.5			
7			*****	http://www.bls.gov/news.rele	Dairy and rel		0.906	219.377	220.319	219.526		0.1	-0.4	0.2	0.4	-0.4			
8			*****	http://www.bls.gov/news.rele	Fruits and ve		1.306	281.072	293.714	293.742		4.5	0	0.3	0.3	1.4			
9			*****	http://www.bls.gov/news.rele	Nonalcoholic		0.948	169.758	169.593	168.977		-0.5	-0.4	0.2	-0.5	0			
10			*****	http://www.bls.gov/news.rele	Other food a		2.28	204.001	205.387	204.763		0.4	-0.3	0.2	-0.2	-0.6			
11			*****	http://www.bls.gov/news.rele	Food away fr		5.705	235.603	240.713	240.93		2.3	0.1	0.1	0.1	0.1			
12			*****	http://www.bls.gov/news.release/cpi.t01.htm															
13			*****	http://www.bls.gov/news.rele	Energy		9.58	242.663	234.624	248.146		2.3	5.8	-0.8	-1.7	5.4			
14			*****	http://www.bls.gov/news.rele	Energy comm		5.793	310.685	292.609	320.258		3.1	9.4	-1.5	-3	8.6			
15			*****	http://www.bls.gov/news.rele	Fuel oil(1)		0.233	384.747	381.889	393.782		2.3	3.1	0	-0.2	3.1			
16			*****	http://www.bls.gov/news.rele	Motor fuel		5.46	306.348	288.108	316.58		3.3	9.9	-1.6	-3.2	9			
17			*****	http://www.bls.gov/news.rele	Gasoline (all		5.273	305.076	286.417	315.243		3.3	10.1	-1.9	-3	9.1			
18			*****	http://www.bls.gov/news.rele	Electricity serv		3.787	187.962	189.444	189.679		0.9	0.1	0.3	0.4	0.5			
19			*****	http://www.bls.gov/news.rele	Electricity(2)		2.881	193.183	194.525	194.739		0.8	0.1	0.2	1.1	0.3			
20			*****	http://www.bls.gov/news.rele	Utility (piped		0.906	169.753	171.597	171.888		1.3	0.2	0.7	-1.7	1.2			
21			*****	http://www.bls.gov/news.release/cpi.t01.htm															
22			*****	http://www.bls.gov/news.rele	All Items less		76.093	227.865	231.612	232.432		2	0.4	0.1	0.3	0.2			
23			*****	http://www.bls.gov/news.rele	Commodities		19.53	146.628	146.492	147.093		0.3	0.4	-0.1	0.2	0			
24			*****	http://www.bls.gov/news.rele	Apparel		3.526	123.312	124.687	126.303		2.4	1.3	0.1	0.8	-0.1			
25			*****	http://www.bls.gov/news.rele	New vehicles		3.195	144.326	145.871	145.925		1.1	0	0.2	0.1	-0.3			
26			*****	http://www.bls.gov/news.rele	Used cars an		1.839	147.011	145.26	146.718		-0.2	1	-0.3	0.2	0.8			
27			*****	http://www.bls.gov/news.rele	Medical care		1.716	331.867	334.046	334.405		0.8	0.1	-0.3	0.1	-0.4			
28			*****	http://www.bls.gov/news.rele	Alcoholic bev		0.95	230.704	232.558	233.898		1.4	0.6	0.3	-0.1	0.4			
29			*****	http://www.bls.gov/news.rele	Tobacco and		0.807	847.88	867.646	865.607		2.1	-0.2	0.5	0.5	-0.2			
30			*****	http://www.bls.gov/news.rele	Services less		56.563	277.027	283.284	284.231		2.6	0.3	0.2	0.3	0.2			
31			*****	http://www.bls.gov/news.rele	Shelter		31.678	254.931	260.039	260.72		2.3	0.3	0.1	0.2	0.2			
32			*****	http://www.bls.gov/news.rele	Rent of prim		6.54	258.184	264.7	265.256		2.7	0.2	0.2	0.2	0.3			
33			*****	http://www.bls.gov/news.rele	Owners' equ		24.016	262.812	267.995	268.448		2.1	0.2	0.1	0.2	0.2			
34			*****	http://www.bls.gov/news.rele	Medical care		5.46	434.832	448.226	451.625		3.9	0.8	0.3	0.2	0.3			
35			*****	http://www.bls.gov/news.rele	Physicians' se		1.617	343.564	351.25	352.266		2.5	0.3	0	0.1	0			
36			*****	http://www.bls.gov/news.rele	Hospital serv		1.562	250.56	260.035	264.071		5.4	1.6	0.7	0.2	0.8			
37			*****	http://www.bls.gov/news.rele	Transportati		5.84	269.535	277.406	277.96		3.1	0.2	0.4	0.5	0.1			
38			*****	http://www.bls.gov/news.rele	Motor vehicl		1.15	256.968	259.752	260.234		1.3	0.2	0	0.4	0.2			
39			*****	http://www.bls.gov/news.rele	Motor vehicl		2.494	395.516	415.51	416.147		5.2	0.2	0.5	0.5	0.2			
40			*****	http://www.bls.gov/news.rele	Airline fare		0.771	298.477	306.603	309.283		3.6	0.9	0.8	1.1	-0.3			
41			*****	http://www.bls.gov/news.rele	Footnotes:(1) Not seasonally adjusted. (2) This index series was calculated using a Laspeyres estimator. All other item stratum index series were calculated using a geometric means e														
42			*****	http://www.bls.gov/news.rele	NOTE: Index applies to a month as a whole, not to any specific date.														

Paste into CSV



page

Local IP: 75.60.206.216

Loading: http://www.infoplease.com/us/census/data/

Remote IP: 165.193.123.18

United States—U.S. Statistics

Share

## U.S. Census Data

This data comes from the most recent (2000) census conducted by the United States Census Bureau. It also contains estimates from subsequent years and historical data from previous years.

### U.S. Statistics

[Demographic](#) · [Economic](#) · [Housing](#) · [Social](#)

### State Statistics

<a href="#">Alabama</a>	<a href="#">Kentucky</a>	<a href="#">North Dakota</a>
<a href="#">Alaska</a>	<a href="#">Louisiana</a>	<a href="#">Ohio</a>
<a href="#">Arizona</a>	<a href="#">Maine</a>	<a href="#">Oklahoma</a>
<a href="#">Arkansas</a>	<a href="#">Maryland</a>	<a href="#">Oregon</a>
<a href="#">California</a>	<a href="#">Massachusetts</a>	<a href="#">Pennsylvania</a>
<a href="#">Colorado</a>	<a href="#">Michigan</a>	<a href="#">Rhode Island</a>
<a href="#">Connecticut</a>	<a href="#">Minnesota</a>	<a href="#">South Carolina</a>
<a href="#">Delaware</a>	<a href="#">Mississippi</a>	<a href="#">South Dakota</a>

I need affordable auto insurance  
My age is...

<a href="#">Under 25</a>	<a href="#">33 yrs old</a>	<a href="#">42 yrs old</a>
<a href="#">25 yrs old</a>	<a href="#">34 yrs old</a>	<a href="#">43 yrs old</a>
<a href="#">26 yrs old</a>	<a href="#">35 yrs old</a>	<a href="#">44 yrs old</a>
<a href="#">27 yrs old</a>	<a href="#">36 yrs old</a>	<a href="#">45 yrs old</a>
<a href="#">28 yrs old</a>	<a href="#">37 yrs old</a>	<a href="#">46 yrs old</a>
<a href="#">29 yrs old</a>	<a href="#">38 yrs old</a>	<a href="#">47 yrs old</a>
<a href="#">30 yrs old</a>	<a href="#">39 yrs old</a>	<a href="#">48 yrs old</a>
<a href="#">31 yrs old</a>	<a href="#">40 yrs old</a>	<a href="#">49 yrs old</a>
<a href="#">32 yrs old</a>	<a href="#">41 yrs old</a>	<a href="#">50+ yrs old</a>

the Catch (0)

Detail

id	R	P	Collection Time	Source Url
----	---	---	-----------------	------------

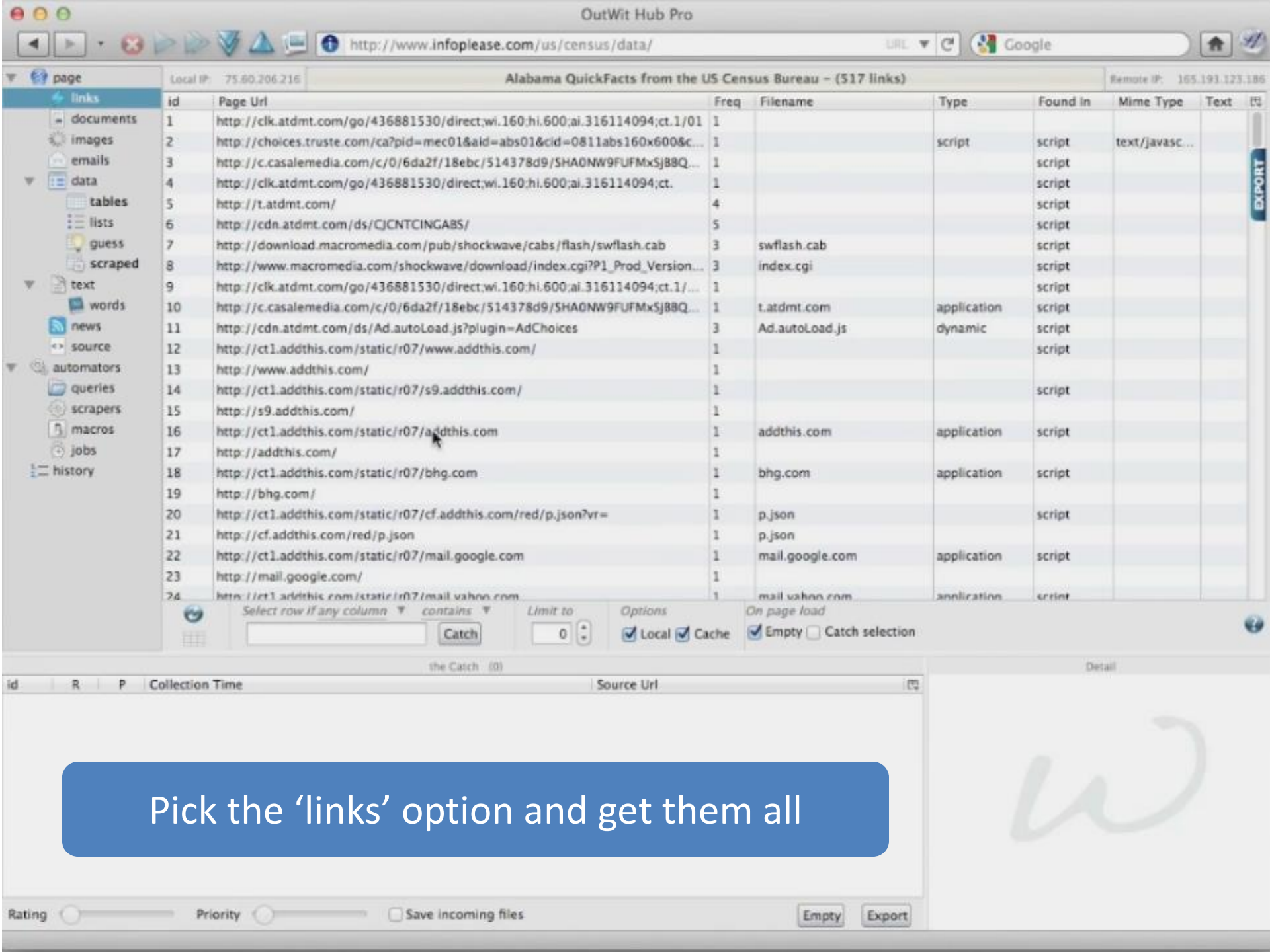
If your data is spread across multiple pages

Rating ☐ Priority ☐ ☐ Save incoming files

Empty

Export





Pick the 'links' option and get them all

Reset All Views  
Clear History  
Downloads  
Add-ons  
Preferences  
Error Console  
Apply Scraper ⌘ S  
Apply Macro ⌘ M

Local IP: 75.60.206.216

id	Page Url
19	http://www.infoplease.com/
20	http://www.infoplease.com/
21	http://www.infoplease.com/
22	http://www.infoplease.com/
23	http://www.infoplease.com/
24	http://www.infoplease.com/us/census/data/south-dakota/
25	http://www.infoplease.com/us/census/data/district-of-columbia/
26	http://www.infoplease.com/us/census/data/missouri/
27	http://www.infoplease.com/us/census/data/tennessee/
28	http://www.infoplease.com/us/census/data/florida/
29	http://www.infoplease.com/us/census/data/montana/
30	http://www.infoplease.com/us/census/data/texas/
31	http://www.infoplease.com/us/census/data/georgia/
32	http://www.infoplease.com/us/census/data/nebraska/
33	http://www.infoplease.com/us/census/data/utah/
34	http://www.infoplease.com/us/census/data/hawaii/
35	http://www.infoplease.com/us/census/data/nevada/
36	http://www.infoplease.com/us/census/data/vermont/
37	http://www.infoplease.com/us/census/data/idaho/
38	http://www.infoplease.com/us/census/data/new-hampshire/
39	http://www.infoplease.com/us/census/data/virginia/
40	http://www.infoplease.com/us/census/data/illinois/
41	http://www.infoplease.com/us/census/data/new-jersey/
42	http://www.infoplease.com/us/census/data/washington/

U.S. Census Data - (51 links)

Freq	Filename	Type	Found in	Mime Type	Text
1			a		Connecticut
1			a		Minnesota
1			a		South Carolina
1			a		Delaware
1			a		Mississippi
1			a		South Dakota
1			a		District of Columbia
1			a		Missouri
1			a		Tennessee
1			a		Florida
1			a		Montana
1			a		Texas
1			a		Georgia
1			a		Nebraska
1			a		Utah
1			a		Hawaii
1			a		Nevada
1			a		Vermont
1			a		Idaho
1			a		New Hampshire
1			a		Virginia
1			a		Illinois
1			a		New Jersey
1			a		Washington

Select row if any column contains Limit to Options On page load

Catch  0 ☒ Local ☒ Cache ☒ Empty ☐ Catch selection

the Catch: (0)

id	R	P	Collection Time	Source Url
----	---	---	-----------------	------------

Detail

Page Url	http://www.infoplease.com/us/census/data/wyoming/
Freq	1
Filename	
Type	
Found in	a
Mime Type	
Text	Wyoming

Pick the URLs you want to scrape  
and apply scraper

Scraper(s) applied successfully - (51 scraped rows) (ended at 15:44:20)

id	Name	Pct White	Pct Black	Pct More Than1	Pct Latino
29	Montana	91.1	0.4	1.5	2.4
30	Texas	83.2	11.7	1.1	35.1
31	Georgia	66.1	29.8	1.0	7.1
32	Nebraska	92.0	4.3	1.1	7.1
33	Utah	93.8	1.0	1.3	10.9
34	Hawaii	26.8	2.3	20.1	8.0
35	Nevada	82.0	7.7	2.6	23.5
36	Vermont	96.9	0.6	1.1	1.1
37	Idaho	95.5	0.6	1.3	9.1
38	New Hampshire	96.1	1.0	1.0	2.2
39	Virginia	73.6	19.9	1.6	6.0
40	Illinois	79.4	15.1	1.1	14.3
41	New Jersey	76.6	14.5	1.3	15.2
42	Washington	85.0	3.5	3.0	8.8
43	Indiana	88.6	8.8	1.1	4.5
44					43.4
45					0.9
46					3.7
47					16.1
48					4.5
49	Kansas	89.4	5.9	1.6	8.3
50	North Carolina	74.1	21.8	1.0	6.4
51	Wyoming	94.8	0.9	1.2	6.7

Catch, cut, and paste to CSV as usual



Select row if any column contains

Catch

Limit to

999

Options

☒ Clean Text☐ Keep Order☐ Deduplicate

On page load

☒ Empty☐ Catch selection

the Catch (51)

id	R	P	Collection Time	Source Url	Name	Pct White	Pct Black	Detail
1			03/15/2013 15:44:33	<a href="http://www.infoplease.com/us/census/data...">http://www.infoplease.com/us/census/data...</a>	Alabama	71.4	26.4	Name
2			03/15/2013 15:44:33	<a href="http://www.infoplease.com/us/census/data...">http://www.infoplease.com/us/census/data...</a>	Kentucky	90.4	7.5	Pct White
3			03/15/2013 15:44:33	<a href="http://www.infoplease.com/us/census/data...">http://www.infoplease.com/us/census/data...</a>	North Dakota	92.3	0.8	Pct Black
4			03/15/2013 15:44:33	<a href="http://www.infoplease.com/us/census/data...">http://www.infoplease.com/us/census/data...</a>	Alaska	70.5	3.7	Pct More Than1
5			03/15/2013 15:44:33	<a href="http://www.infoplease.com/us/census/data...">http://www.infoplease.com/us/census/data...</a>	Louisiana	64.1	33.1	Pct Latino
6			03/15/2013 15:44:33	<a href="http://www.infoplease.com/us/census/data...">http://www.infoplease.com/us/census/data...</a>	Ohio	85.1	11.9	
7			03/15/2013 15:44:33	<a href="http://www.infoplease.com/us/census/data...">http://www.infoplease.com/us/census/data...</a>	Arizona	87.4	3.6	
8			03/15/2013 15:44:33	<a href="http://www.infoplease.com/us/census/data...">http://www.infoplease.com/us/census/data...</a>	Maine	96.9	0.8	
9			03/15/2013 15:44:33	<a href="http://www.infoplease.com/us/census/data...">http://www.infoplease.com/us/census/data...</a>	Oklahoma	78.5	7.7	

# 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

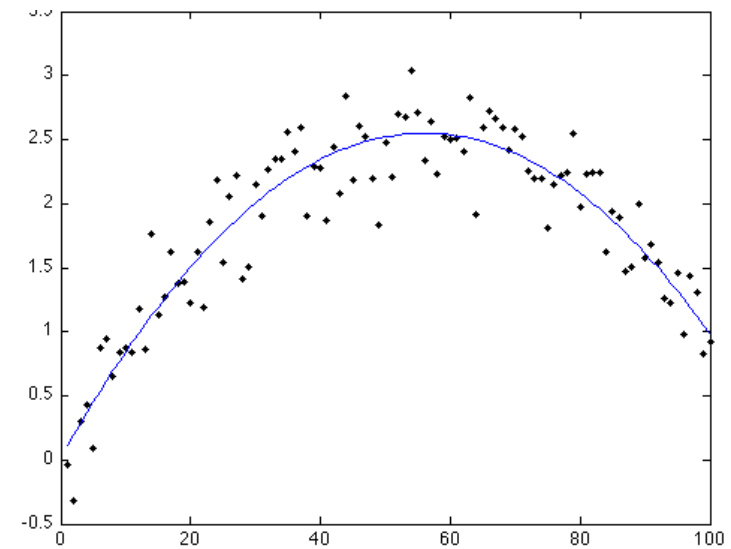
Age	Income	Team	Gender
23	24,200	Mets	M
39	50,245	Yankees	F
45	45,390	Yankees	F
22	32,300	Mets	M
52		Yankees	F
27	28,300	Mets	F
48	53,100	Yankees	M

- How would you estimate the missing value for income?
  - ignore or put in a default value (will decimate the usable data)
  - manually fill in (can be tedious or infeasible for large data)
  - 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)

# 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



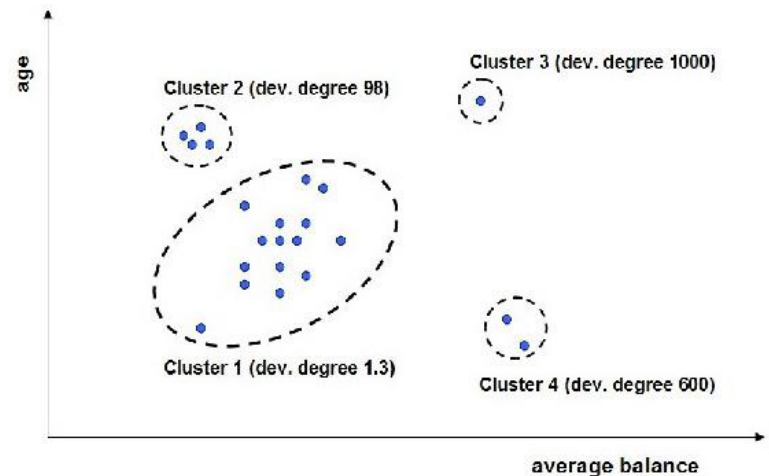
# NOISY DATA – WHAT TO DO

## Binning method

- discussed last lecture

## Clustering

- detect and remove outliers

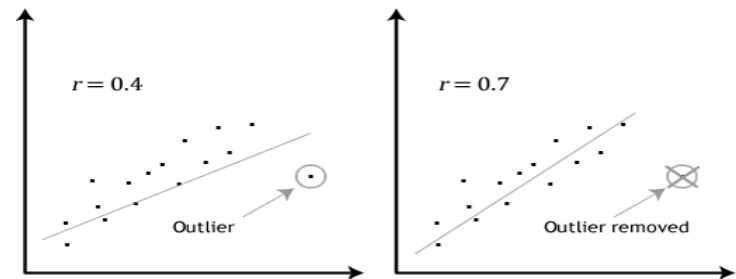


## Semi-automated method

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

## Regression

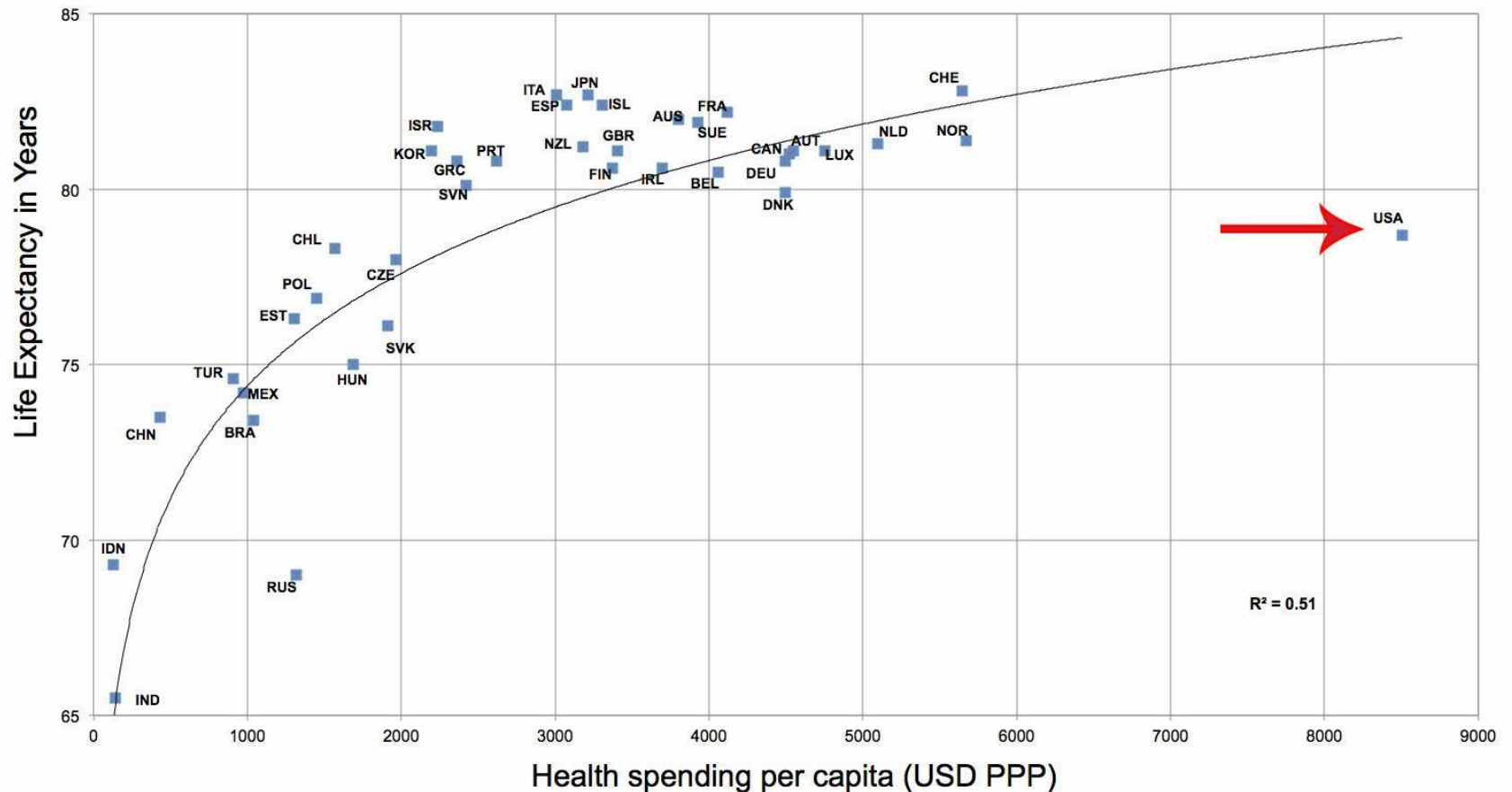
- smooth by fitting the data to a regression function



# 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 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

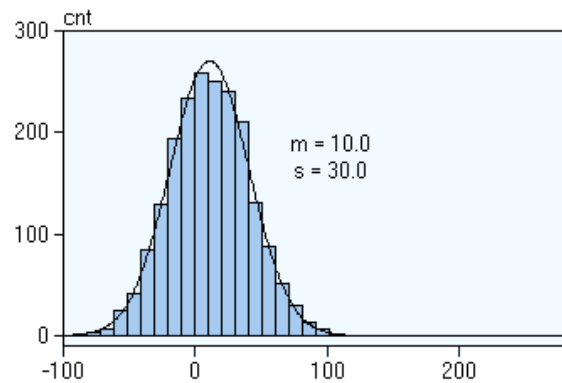
- min-max normalization

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

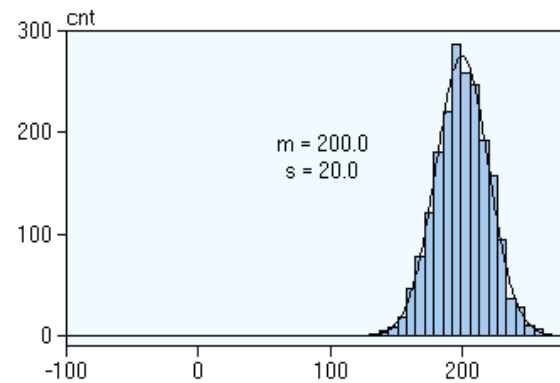
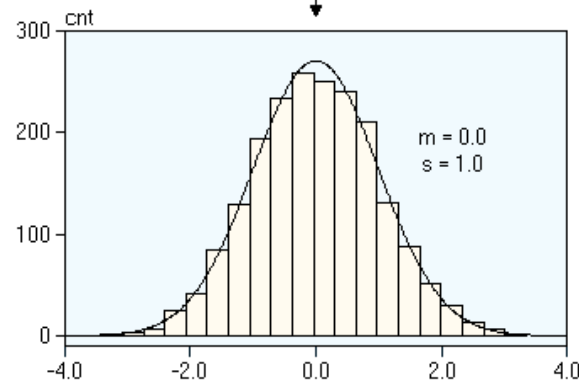
- standardization / z-score normalization

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

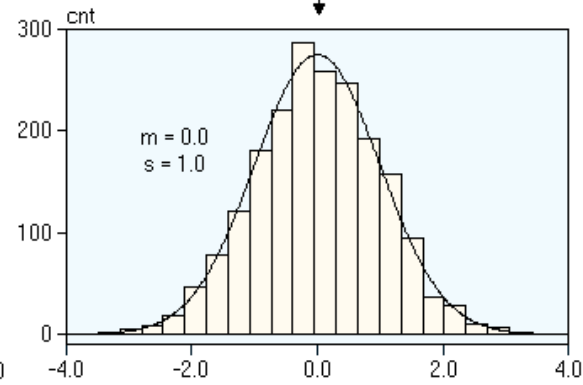
# STANDARDIZATION



Standardisation

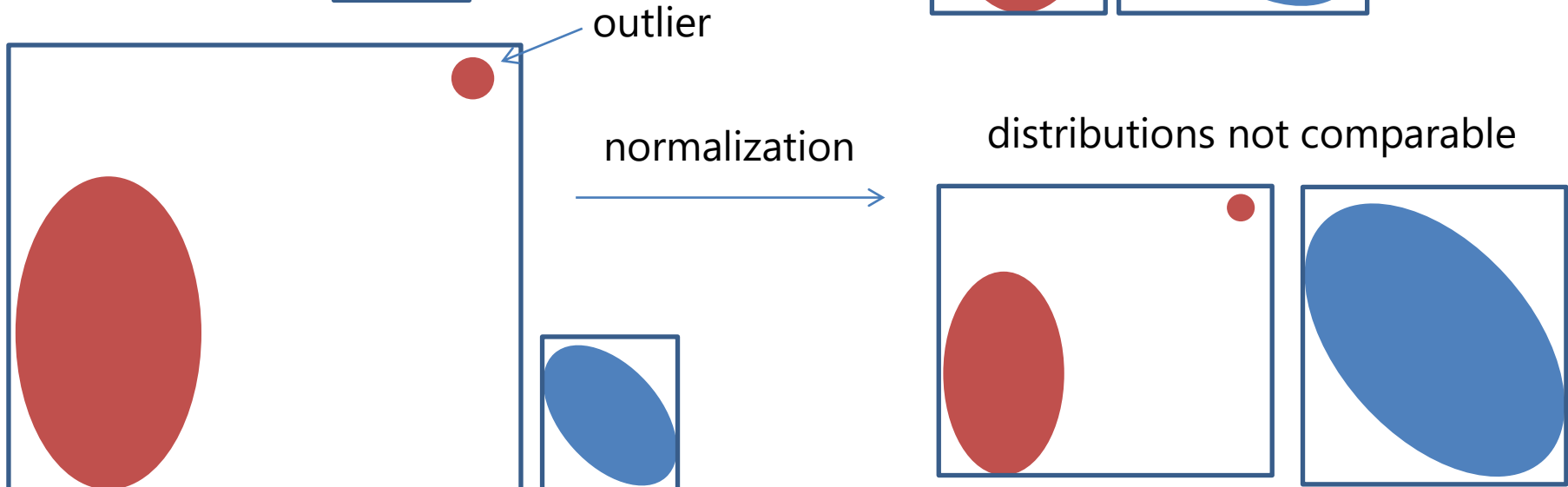
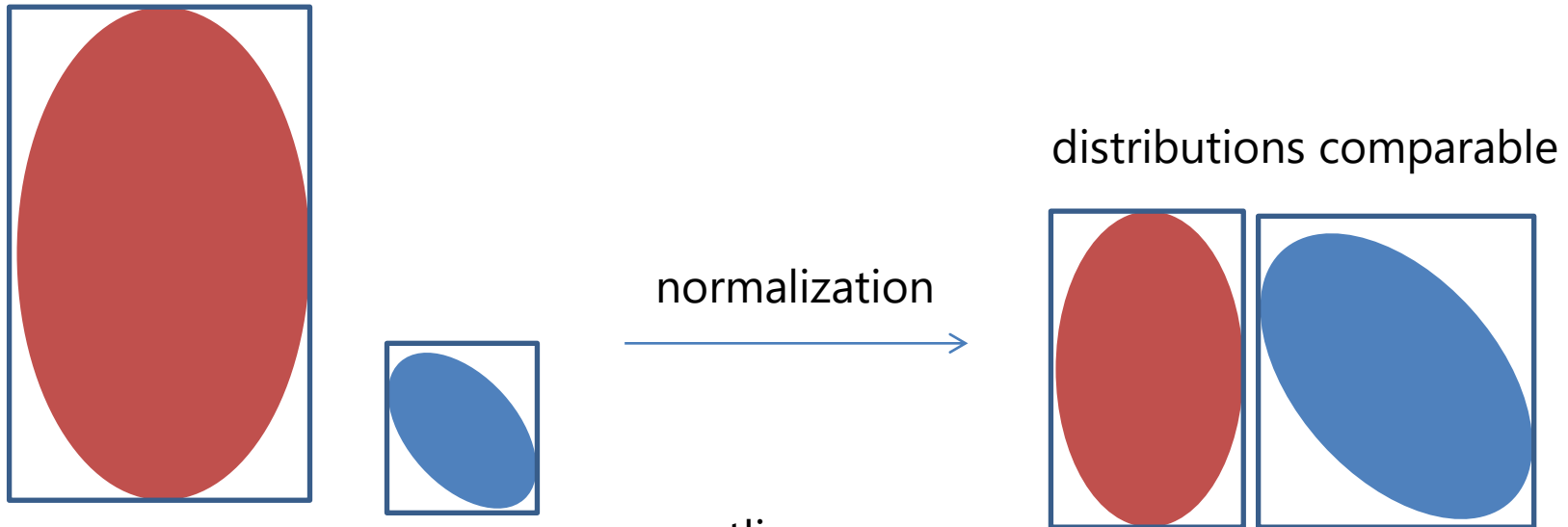


Standardisation



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

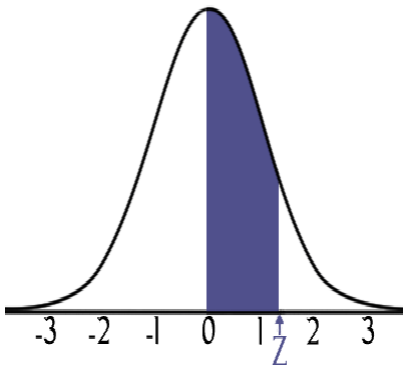
# NORMALIZATION



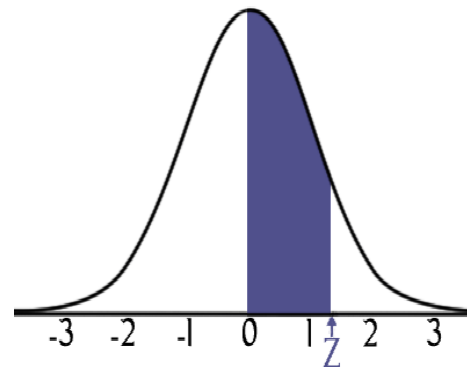


# STANDARDIZATION

Is standardization less or more sensitive to outliers?



without outlier



with outlier (just slightly extended)

# 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

But data integration can also bring ethical problems – see next

# PRIVACY

Can you identify a person from these medical records?

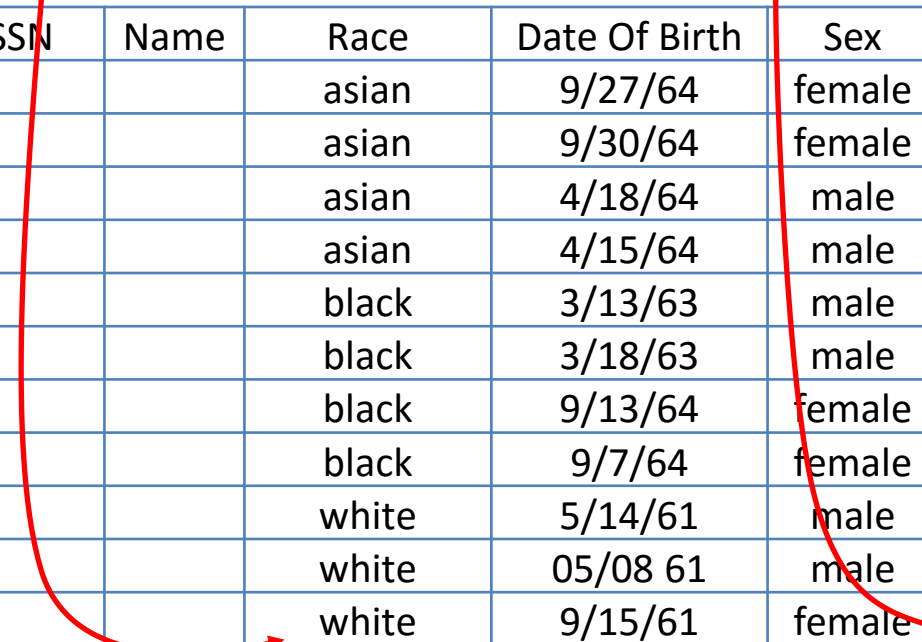
SSN	Name	Race	Date Of Birth	Sex	ZIP	Marital Status	Health Problem
		asian	9/27/64	female	94139	divorced	hypertension
		asian	9/30/64	female	94139	divorced	obesity
		asian	4/18/64	male	94139	married	chest pain
		asian	4/15/64	male	94139	married	obesity
		black	3/13/63	male	94138	married	hypertension
		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



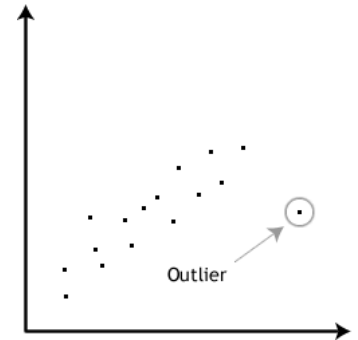
# DATA FUSION VS. DATA PRIVACY

Data fusion can bring insight

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

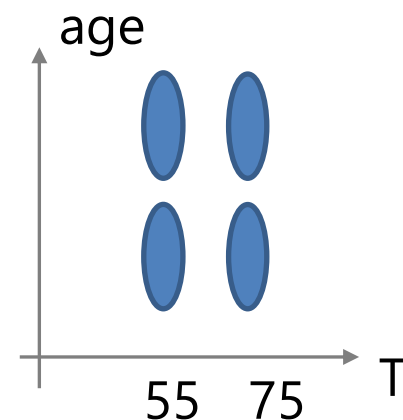
Visualization can bring insight

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



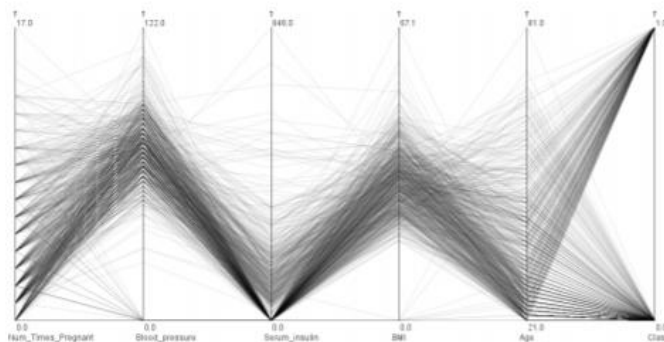
How to obfuscate for protection?

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

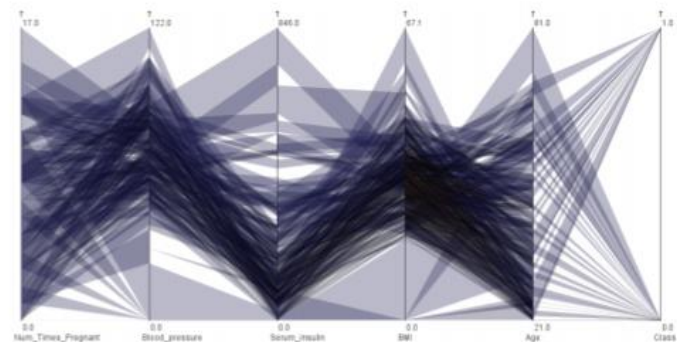


# DATA PRIVACY WITH PARALLEL COORDINATES USING K-ANONYMITY

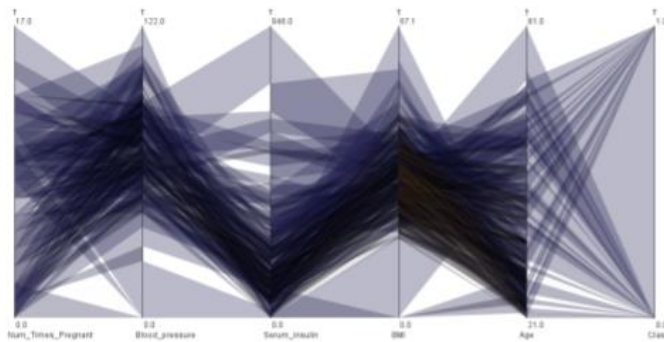
- Cluster records into  $k$ -sized bins for each variable/dimension
- Dasgupta and Kosara show this for parallel coordinates [\[TVCG, 2011\]](#)



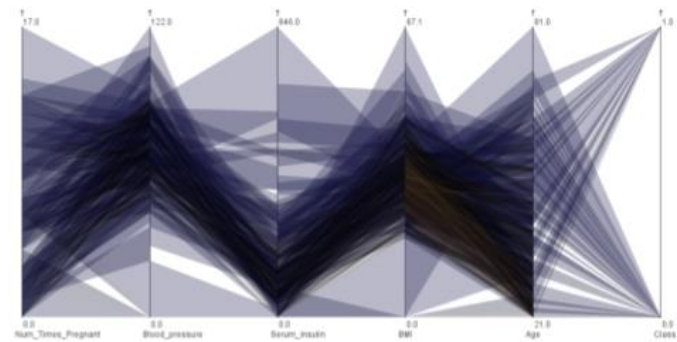
(a) Original View of the raw dataset



(b) Anonymization with  $k=2$



(c) Anonymization with  $k=3$



(d) Anonymization with  $k=4$

# DATA REDUCTION

## Sampling

- random
- stratified



## Data summarization

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



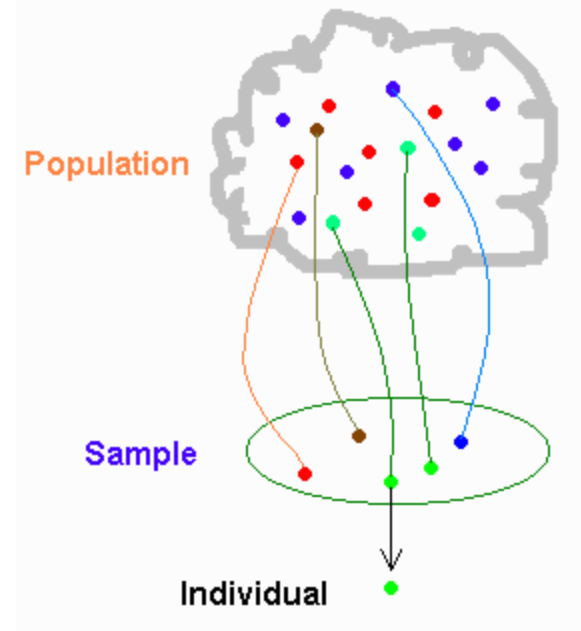
# SAMPLING

## The goal

- pick a representative subset of the data

## Random sampling

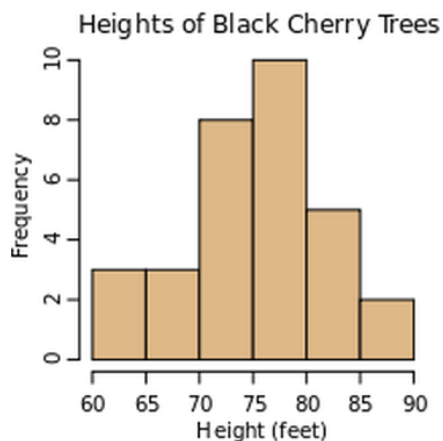
- pick sample points at random
- will work if the points are distributed uniformly
- this is usually not the case
- outliers will likely be missed
- so the sample will not be representative



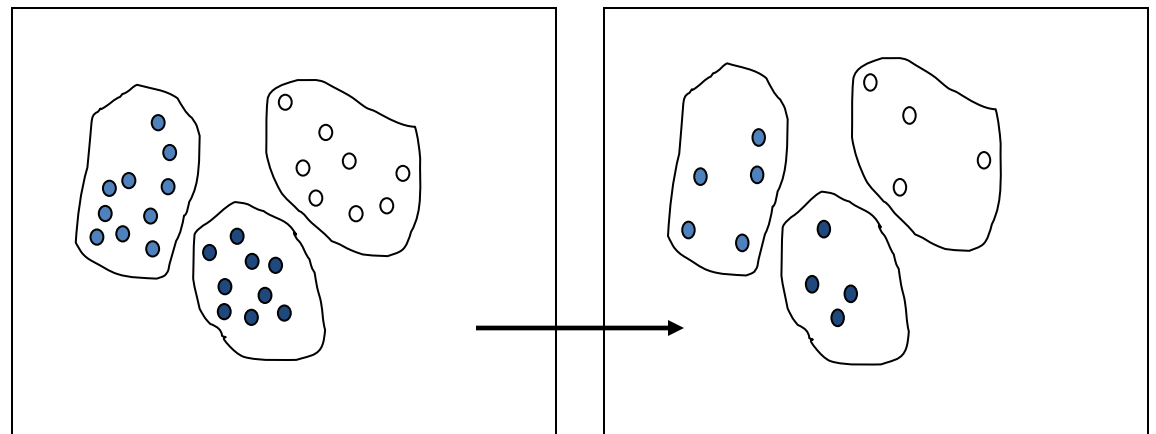
# ADAPTIVE SAMPLING

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  $\sim$  bin height



sampling rate  $\sim$  cluster size

# WHAT'S WHEN YOUR DATA IS TOO SMALL

Can you “hallucinate” or “invent” realistic data?

And if so, how would you go about this?

# HOW TO HALLUCINATE MORE DATA...



# DATA AUGMENTATION

Strategy to artificially synthesize new data from existing data

- go from small data to big data





# DATA AUGMENTATION IN MACHINE LEARNING

Important topic in deep learning

Common techniques are (for images)

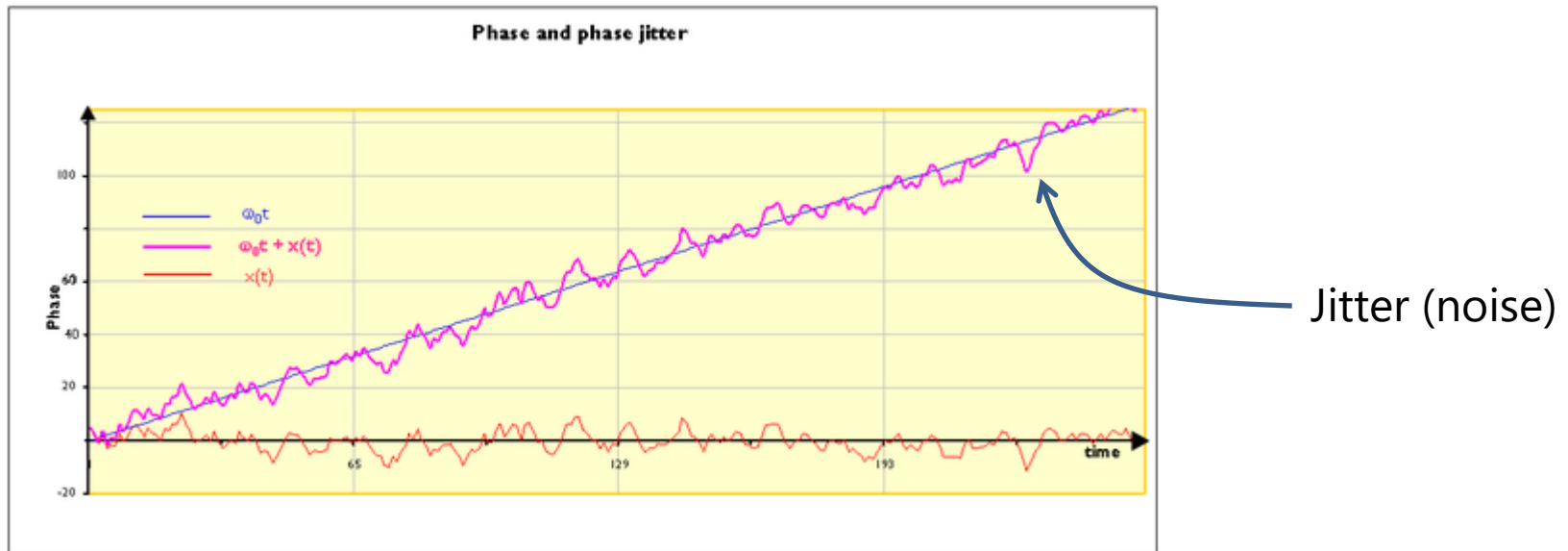
- rotations
- translations
- zooms
- flips
- color perturbations
- crops
- add noise by *jittering*



# WHAT'S JITTERING?

## Definition from dictionary

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



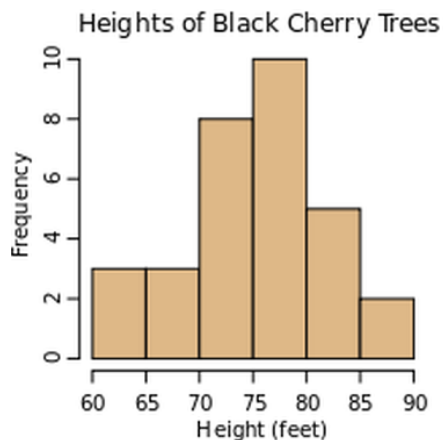
- small random noise about a steady signal



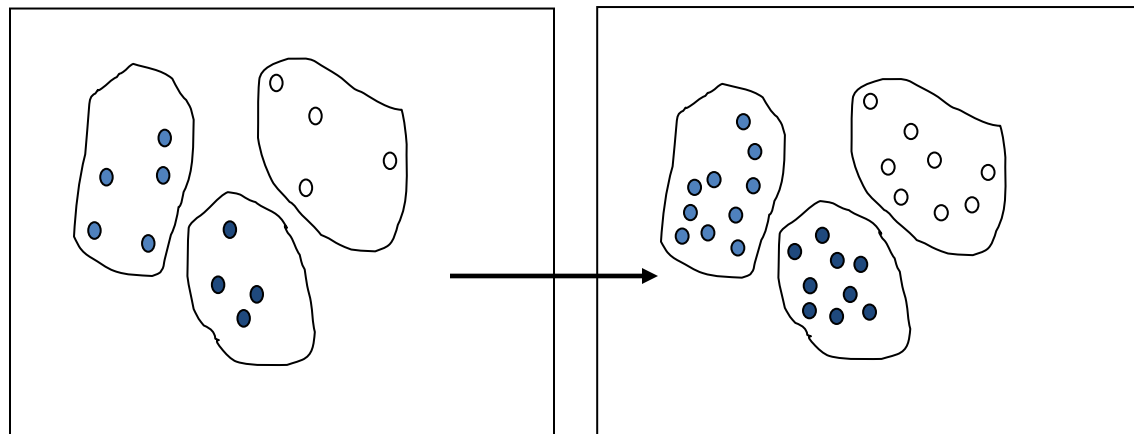
# DATA AUGMENTATION FOR VISUALIZATION & VISUAL ANALYTICS

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



sampling rate  $\sim$  bin height



augmentation rate  $\sim$  cluster size

# QUESTIONS?

The course has been set up with Piazza

- <http://piazza.com/stonybrook/fall2018/cse332/home>
- please let me know if you cannot access it

Make use of this handy discussion forum

- ask questions of general interest
- give advice to peers (those who ask questions)
- give general feedback (observe etiquette)
- but obviously, don't provide actual solutions and aid in cheating