CSE 564
Visualization & Visual Analytics

Data Preparation & Representation

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 preparation and reduction	
5	Data types, 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

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

	A	B	C	D	E	F	G	H	I
1	Name	Country	Miles Per Gallon	Accceleration	Horsepower	weight	cylinders	year	price
2	Volkswagen Rabbit Dl	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

Rectangular Dataset

Also called the *Data Matrix*

Car performance metrics

or Survey question responses

or Patient characteristics

....

One data item

Car models

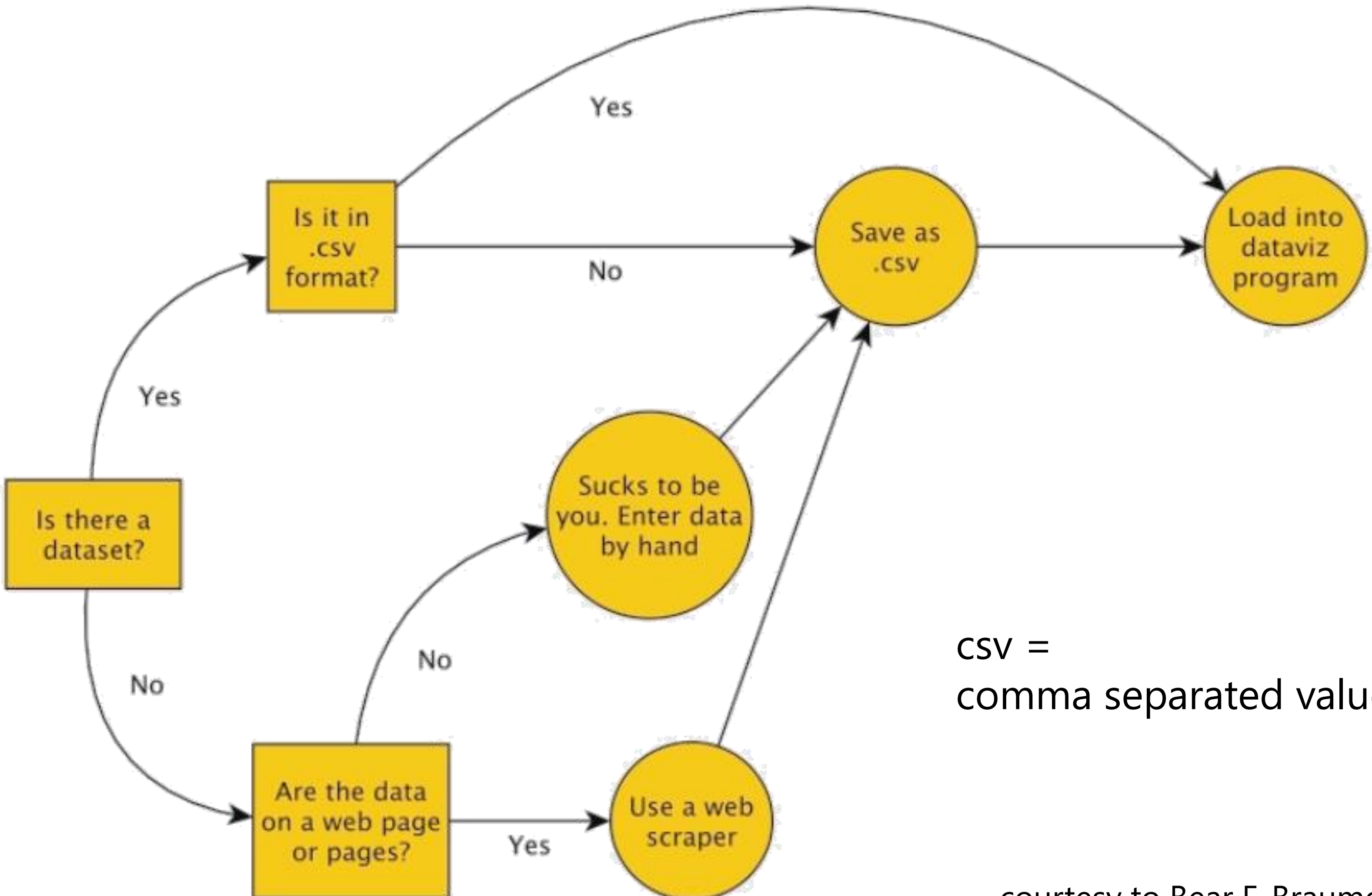
or Survey respondents

or Patients

....

	A	B	C	D	E	F	
1	**Name**	**Country**	**Miles Per Gallon**	**Accceleration**	**Horsepower**	**weight**	**cylir**
2	Volkswagen Rabbit Dl	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 Century	USA	20,6	15,8	105	3380	

How To Import Data?



csv =
comma separated values file

courtesy to Bear F. Braumoeller

How To Get data? (1)

Use Google

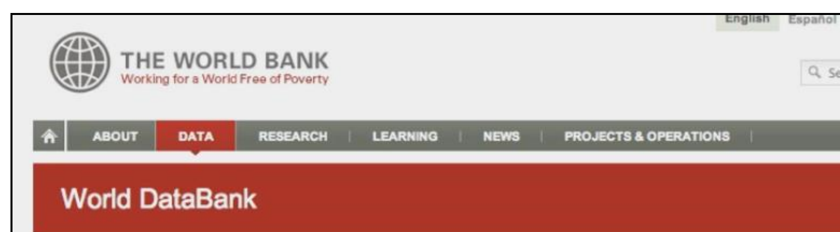
- type the topic you like and perhaps ‘data”, ‘’database’, csv’, etc.

Other sources:

- https://www.data.gov/

- https://fedstats.sites.usa.gov/

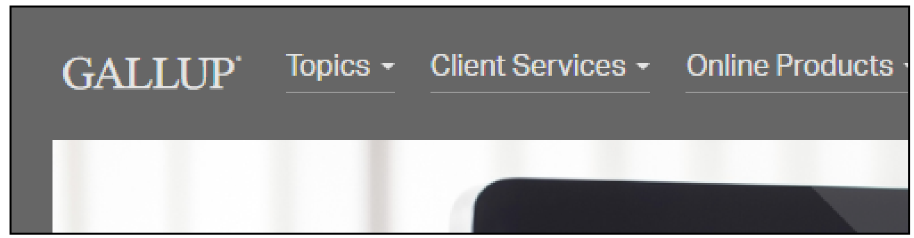
- http://data.worldbank.org/



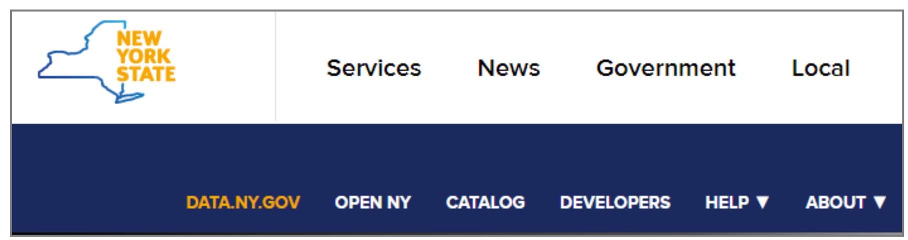
How To Get data? (2)

Other sources:

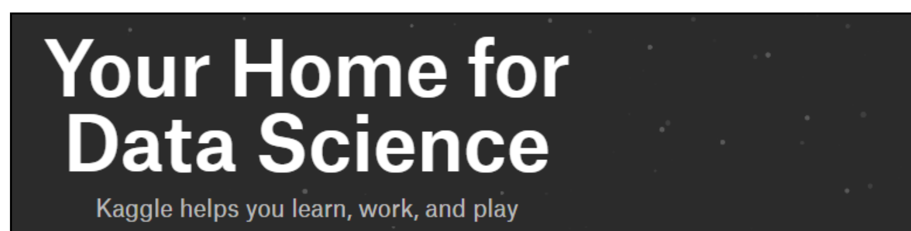
- http://www.gallup.com/products/184157/gallup-analytics-universities-colleges.aspx



- https://data.ny.gov/



- https://www.kaggle.com/



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 Dl	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

Categorical **Quantitative** Categorical (Ordinal) Quantitative

<u>Data types</u>

Quantitative (Numerical)

Categorical (Ordinal)

Notes on Dataset

Some advice

- avoid datasets where the majority of data is categorical (not overly exciting for binning, clustering, and so on)
- convert 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 (prefer quantitative attributes)
- if the data set has text, images, video, logs, etc. convert them to numbers via appropriate mechanism as discussed in class
- produce a spreadsheet of rows (data items) and attributes (columns)

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)

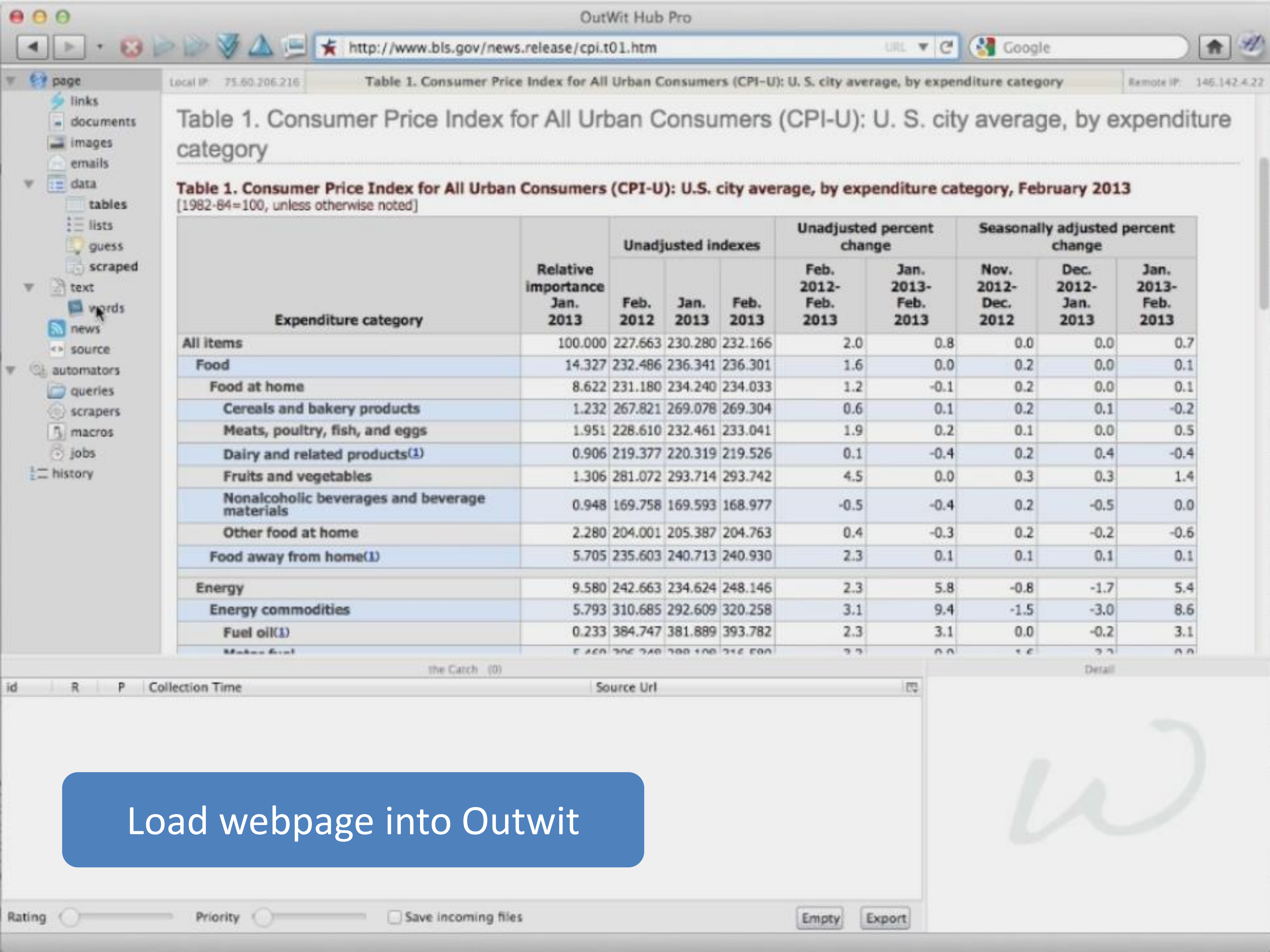
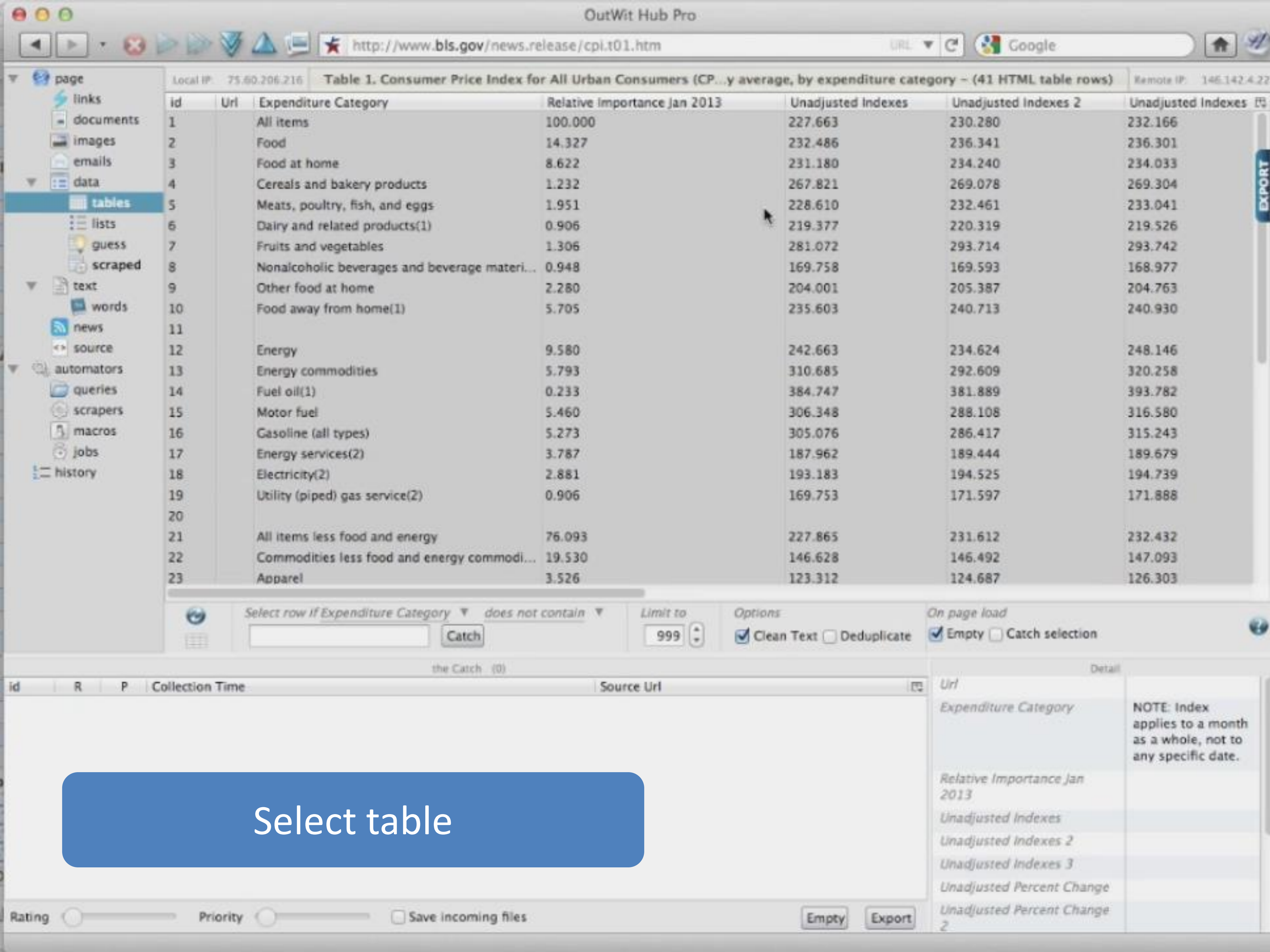


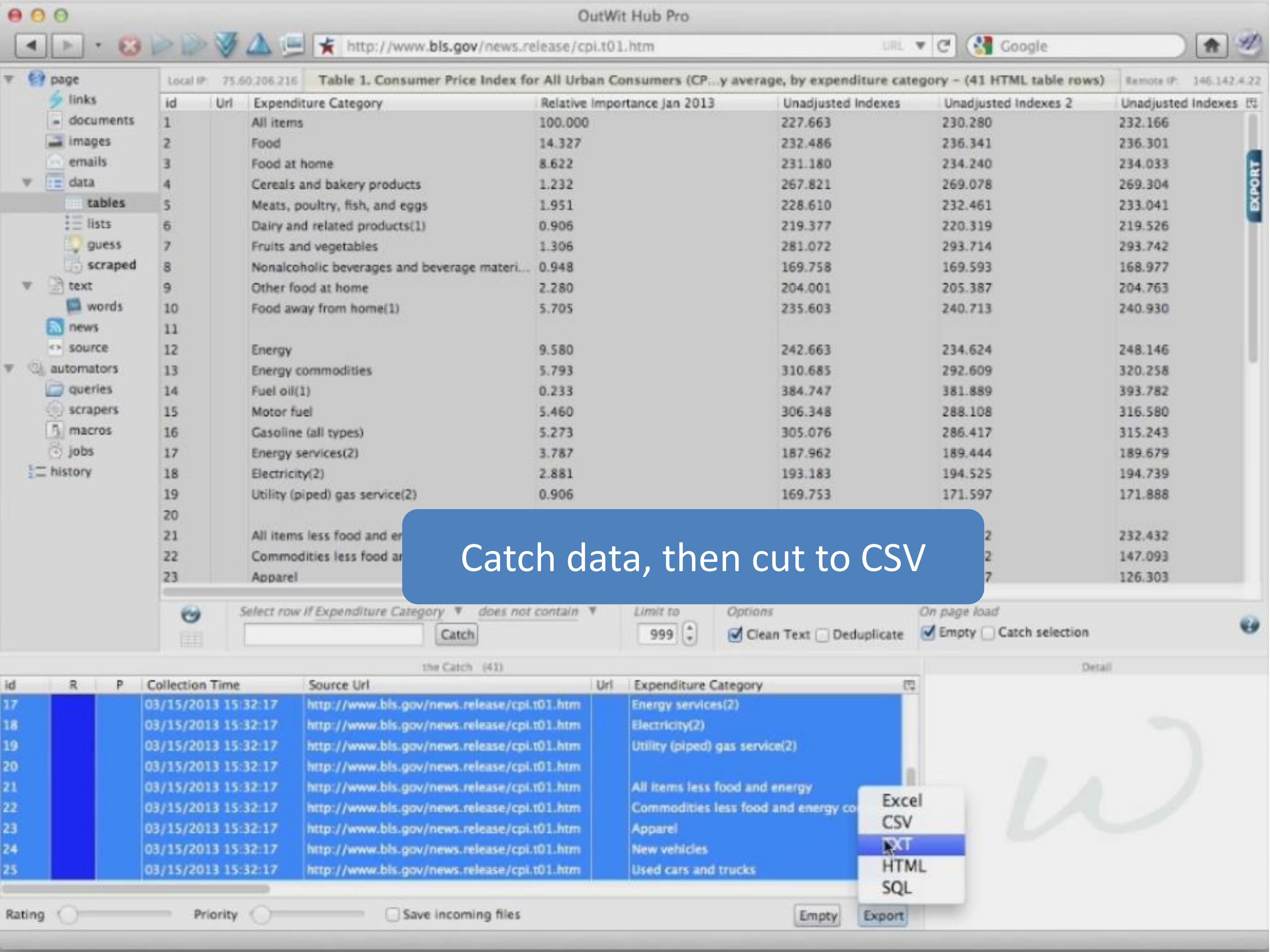
Table 1. Consumer Price Index for All Urban Consumers (CPI-U): U.S. city average, by expenditure category, February 2013
[1982-84=100, unless otherwise noted]

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

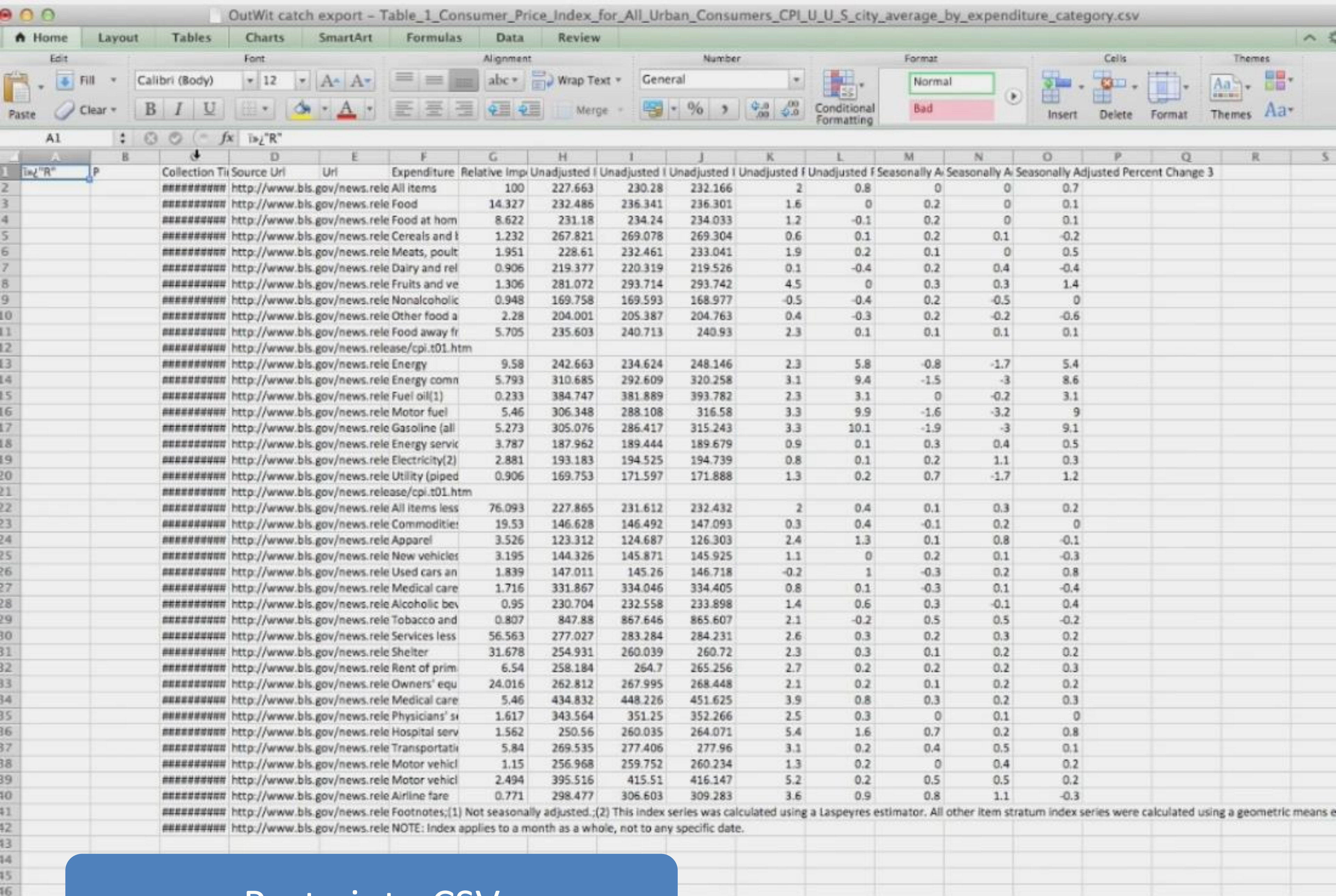
Load webpage into Outwit



Select table

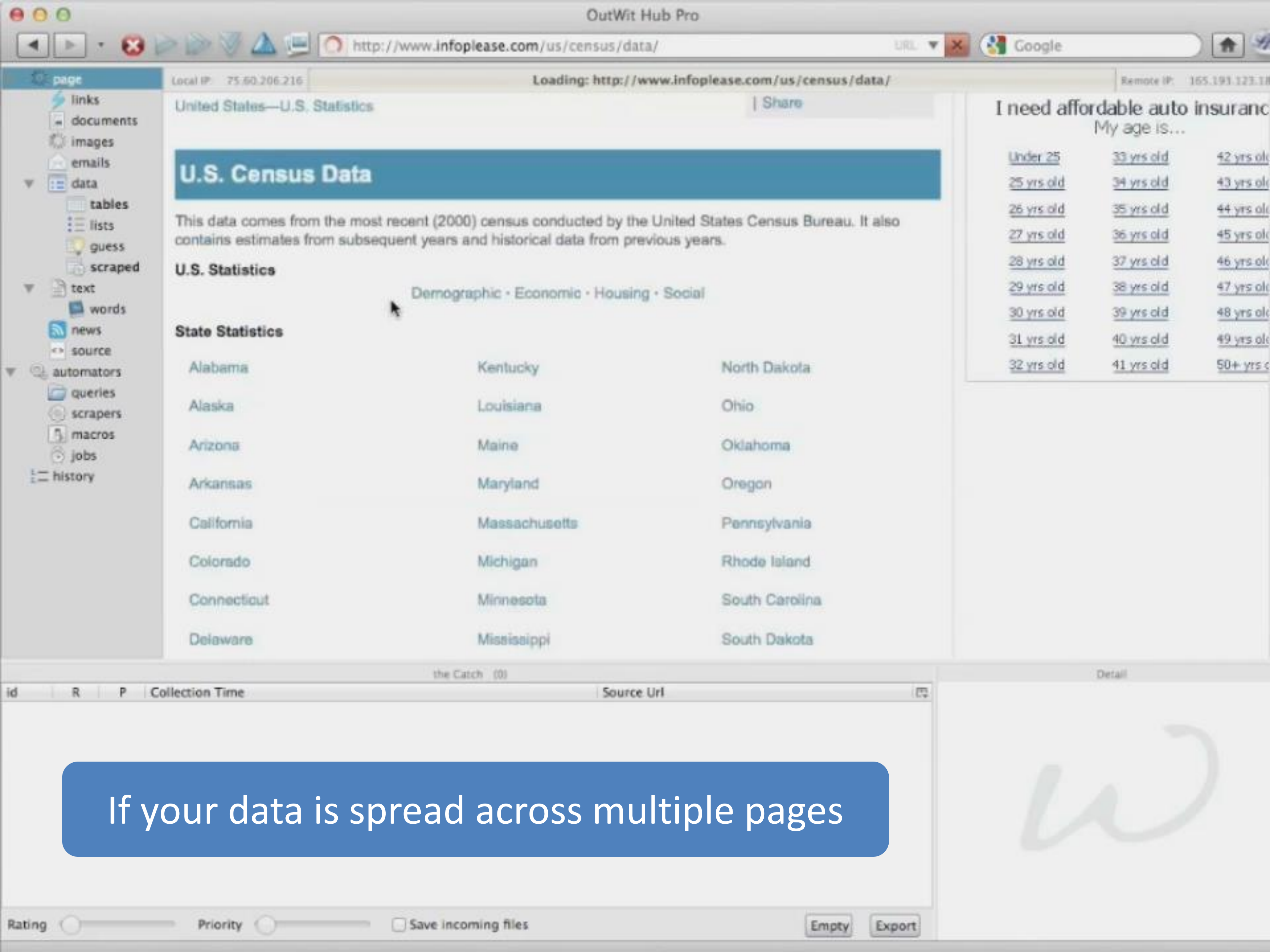


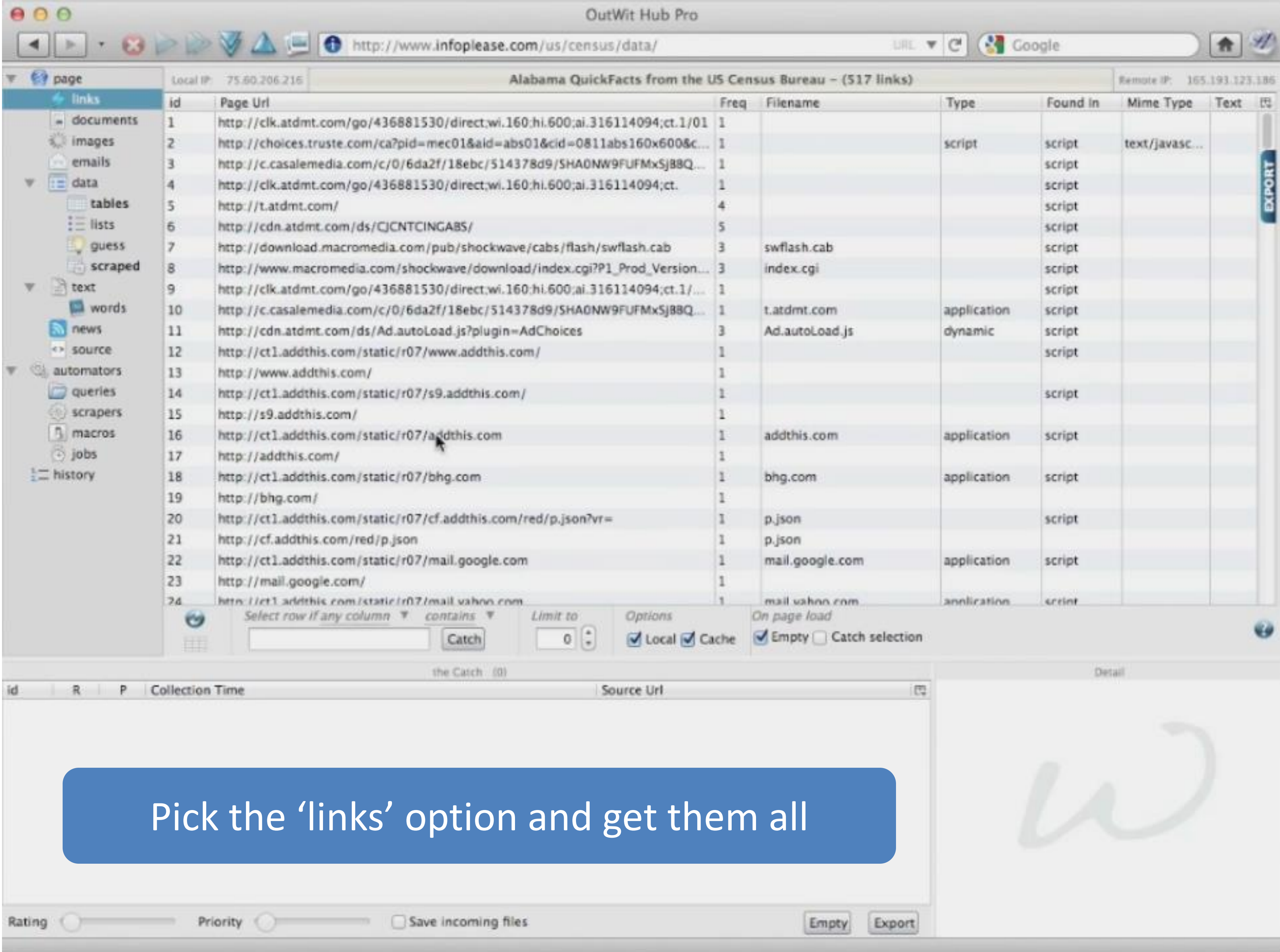
OutWit Hub Pro
http://www.bls.gov/news.release/cpi.t01.htm
Table 1. Consumer Price Index for All Urban Consumers (CP...y average, by expenditure category - (41 HTML table rows)
Catch data, then cut to CSV
Excel
CSV
TXT
HTML
SQL
Empty
Export



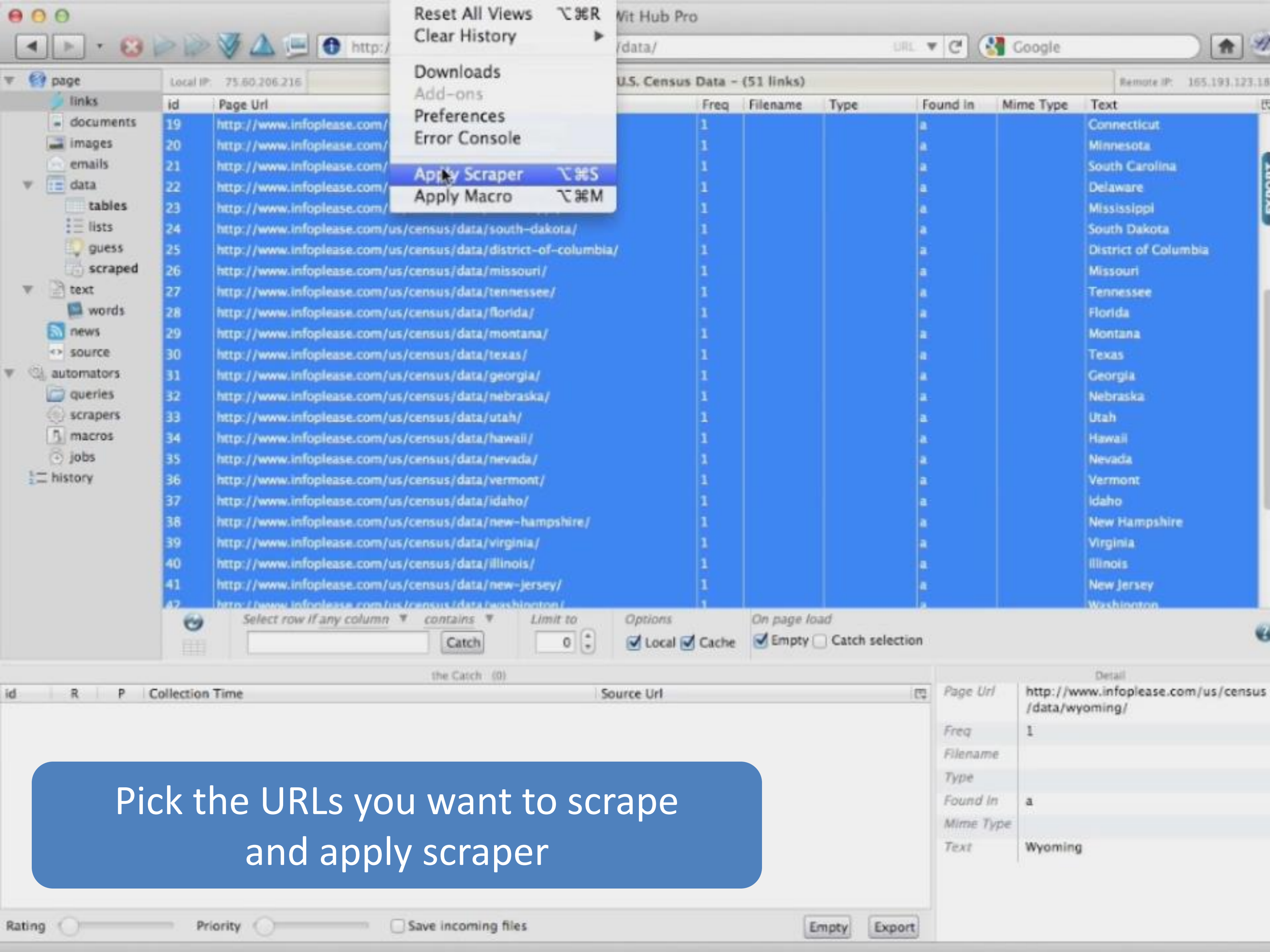
	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O
1	ï»¿"R"	P	Collection Ti	Source Url	Url	Expenditure	Relative Imp	Unadjusted I	Unadjusted I	Unadjusted I	Unadjusted F	Unadjusted F	Seasonally A	Seasonally A	Seasonally Adjusted Percent Change 3
2			##########	http://www.bls.gov/news.rele		All items	100	227.663	230.28	232.166	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, poult	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 comn	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		Energy servic	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

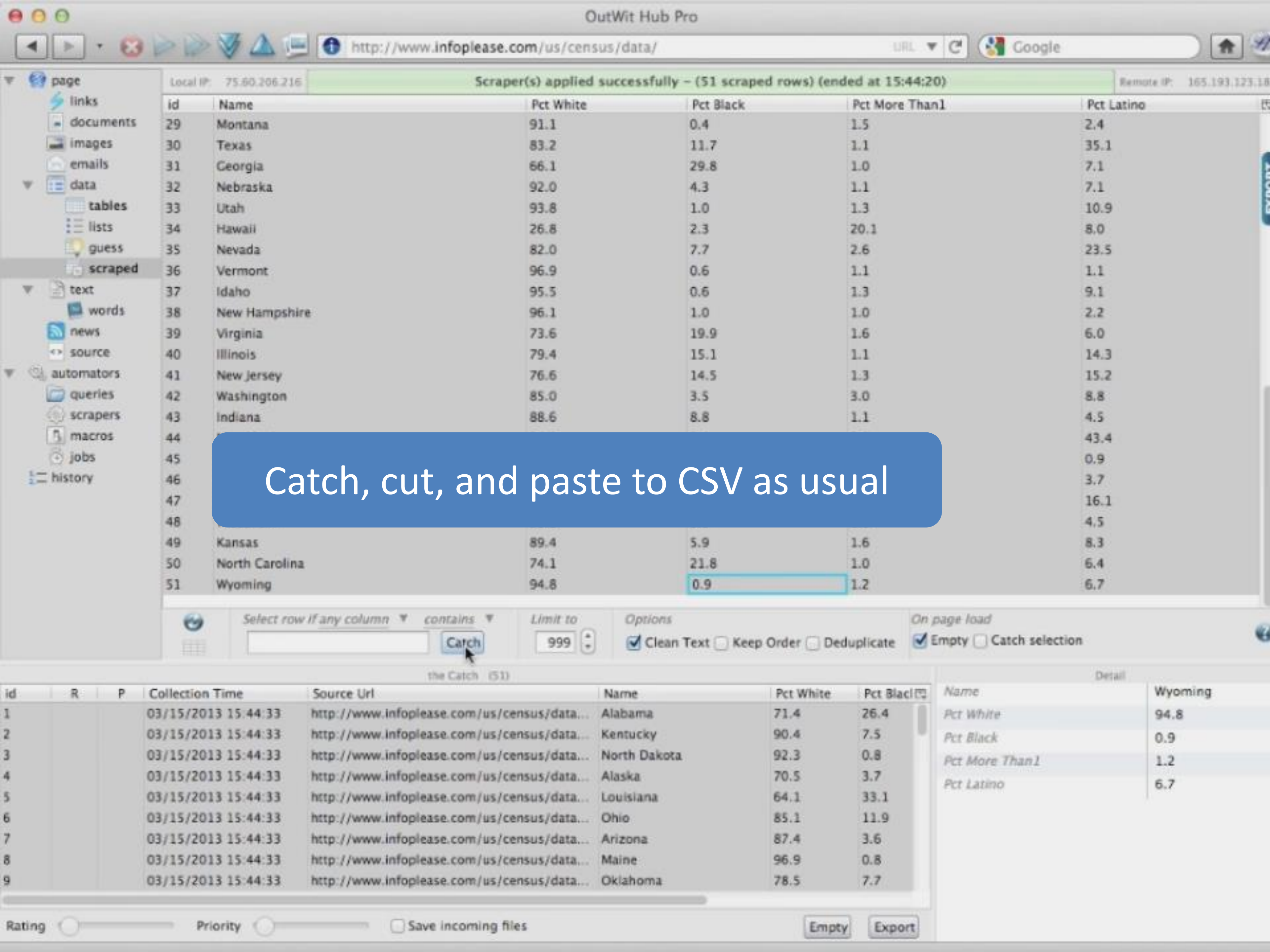

OutWit Hub Pro
http://www.infoplease.com/us/census/data/
Google
Loading: http://www.infoplease.com/us/census/data/
page
links
documents
images
emails
data
tables
lists
guess
scraped
text
words
news
source
automators
queries
scrapers
macros
jobs
history
United States—U.S. Statistics
| Share
U.S. Census 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
Alabama
Alaska
Arizona
Arkansas
California
Colorado
Connecticut
Delaware
Kentucky
Louisiana
Maine
Maryland
Massachusetts
Michigan
Minnesota
Mississippi
North Dakota
Ohio
Oklahoma
Oregon
Pennsylvania
Rhode Island
South Carolina
South Dakota
My age is...
Under 25
25 yrs old
26 yrs old
27 yrs old
28 yrs old
29 yrs old
30 yrs old
31 yrs old
32 yrs old
33 yrs old
34 yrs old
35 yrs old
36 yrs old
37 yrs old
38 yrs old
39 yrs old
40 yrs old
41 yrs old
the Catch (0)
Detail
id
R
P
Collection Time
Source Url
Rating
Priority
Save incoming files
Empty
Export
If your data is spread across multiple pages



Pick the 'links' option and get them all



Pick the URLs you want to scrape and apply scraper



Catch, cut, and paste to CSV as usual
OutWit Hub Pro
http://www.infoplease.com/us/census/data/
Google
page
links
documents
images
emails
data
tables
lists
guess
scraped
text
words
news
source
automators
queries
scrapers
macros
jobs
history
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
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
1 03/15/2013 15:44:33 http://www.infoplease.com/us/census/data... Alabama 71.4 26.4
2 03/15/2013 15:44:33 http://www.infoplease.com/us/census/data... Kentucky 90.4 7.5
3 03/15/2013 15:44:33 http://www.infoplease.com/us/census/data... North Dakota 92.3 0.8
4 03/15/2013 15:44:33 http://www.infoplease.com/us/census/data... Alaska 70.5 3.7
5 03/15/2013 15:44:33 http://www.infoplease.com/us/census/data... Louisiana 64.1 33.1
6 03/15/2013 15:44:33 http://www.infoplease.com/us/census/data... Ohio 85.1 11.9
7 03/15/2013 15:44:33 http://www.infoplease.com/us/census/data... Arizona 87.4 3.6
8 03/15/2013 15:44:33 http://www.infoplease.com/us/census/data... Maine 96.9 0.8
9 03/15/2013 15:44:33 http://www.infoplease.com/us/census/data... Oklahoma 78.5 7.7
Detail
Name Wyoming
Pct White 94.8
Pct Black 0.9
Pct More Than1 1.2
Pct Latino 6.7
Rating
Priority
Save incoming files
Empty
Export

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

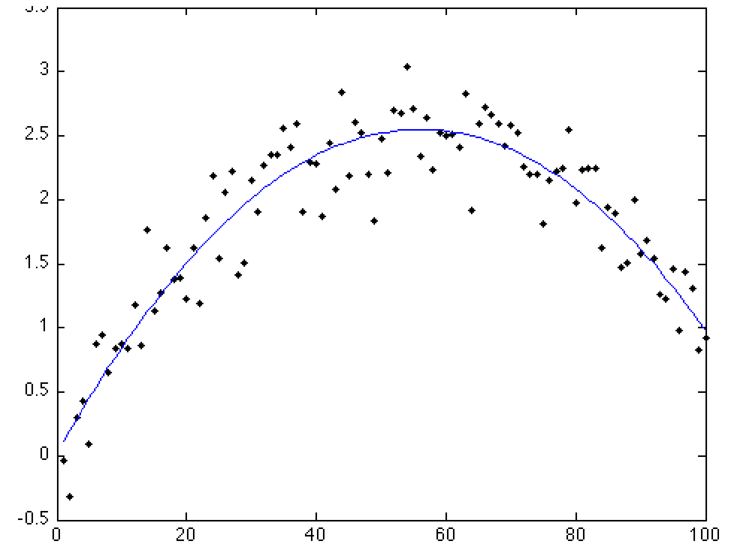
Missing Data – Example

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

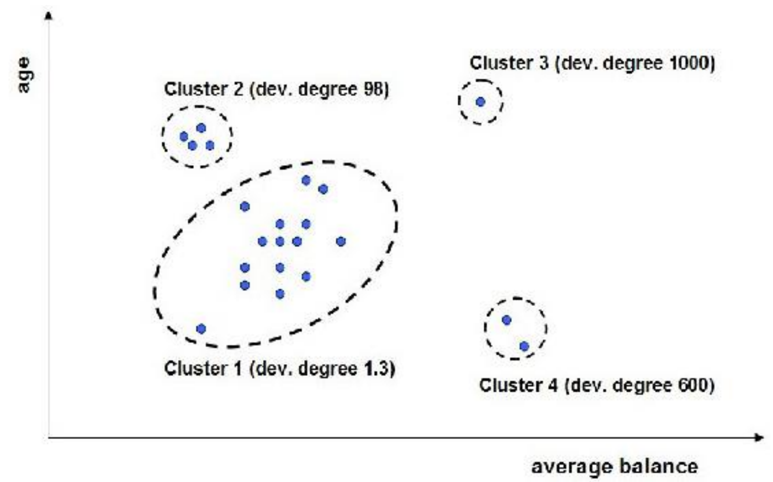
Noisy Data – What To Do

Binning method

- discussed last lecture

Clustering

- detect and remove outliers

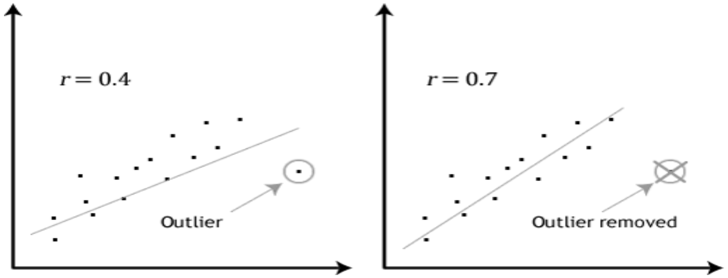


Semi-automated method

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

Regression

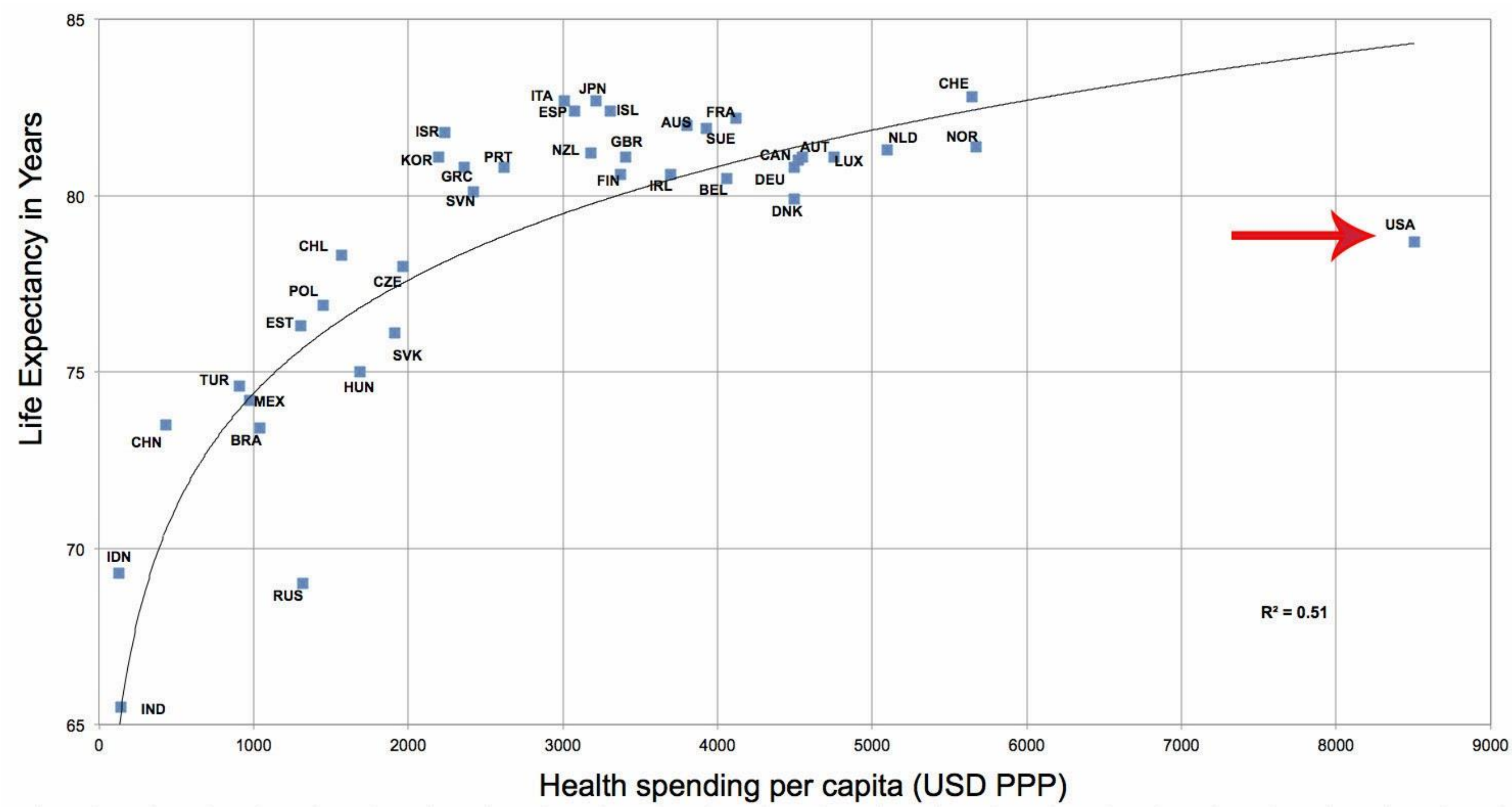
- smooth by fitting the data to a regression function



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

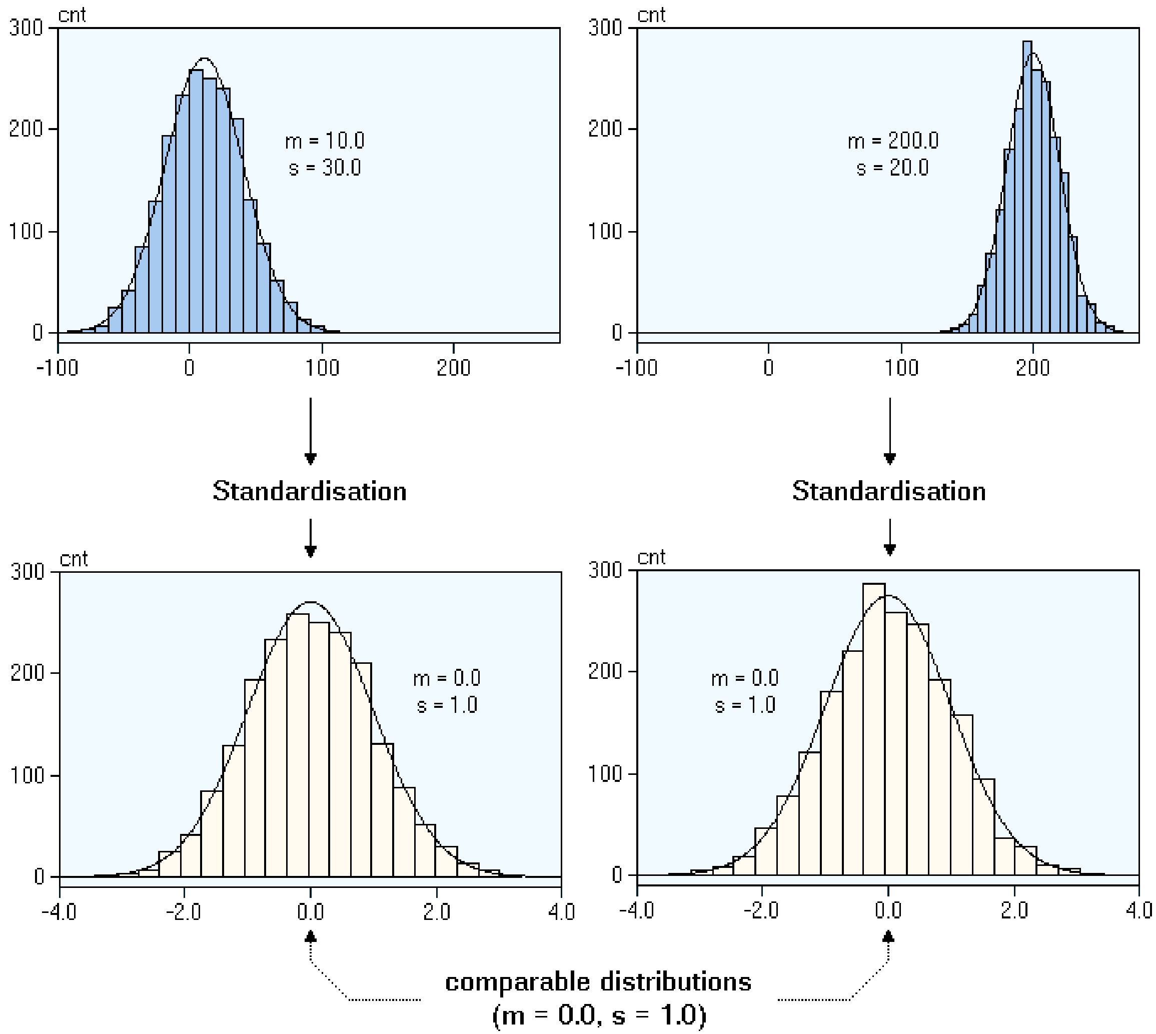
- min-max normalization

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

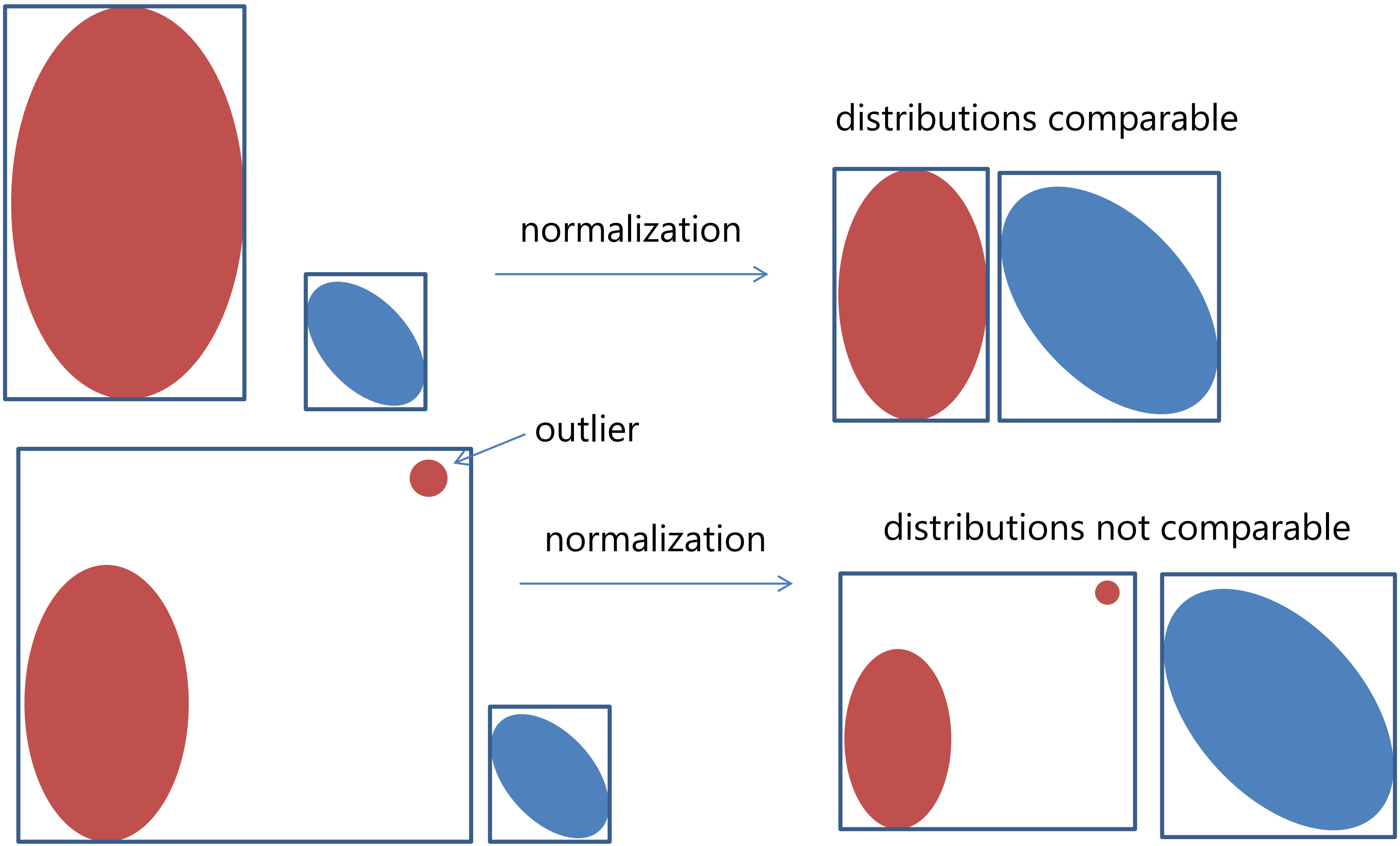
- standardization / z-score normalization

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

Standardization

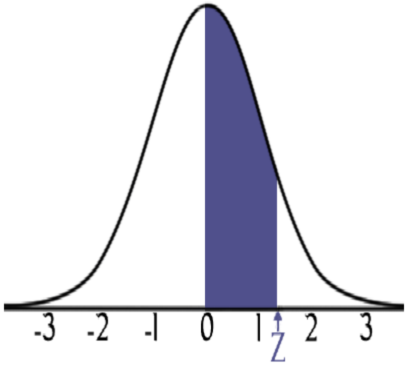


Normalization

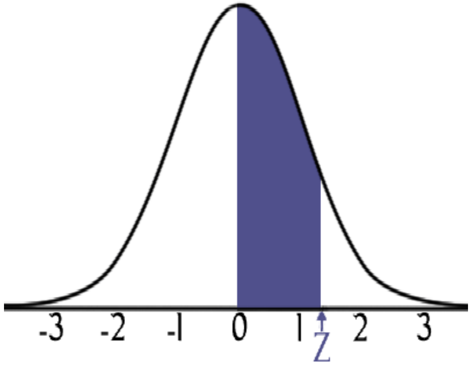


Standardization

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

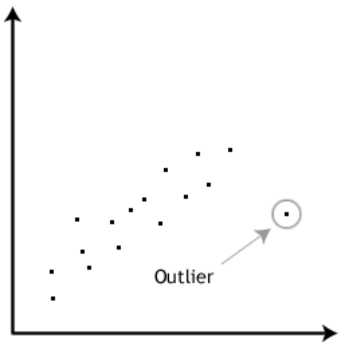
Data Fusion vs. Data Privacy

Data fusion can bring insight

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

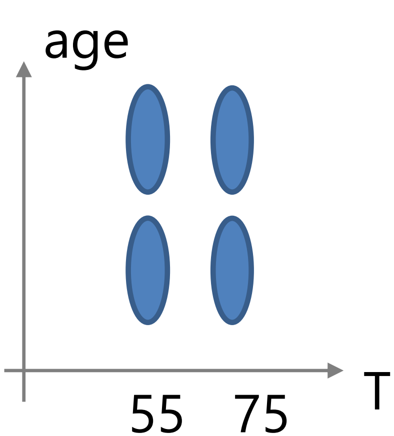
Visualization can bring insight

- 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?

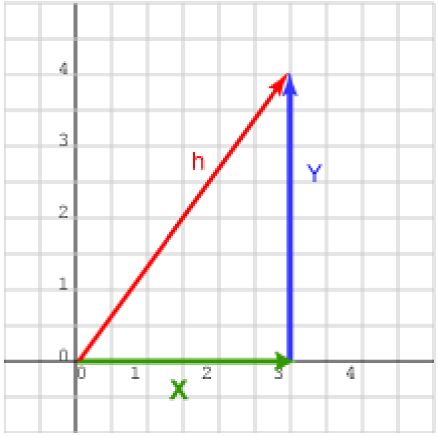
- k-anonymity (generalize)
- make data less specific → binning
- age *groups*, zip code *groups*, etc…
- make blobs instead of points

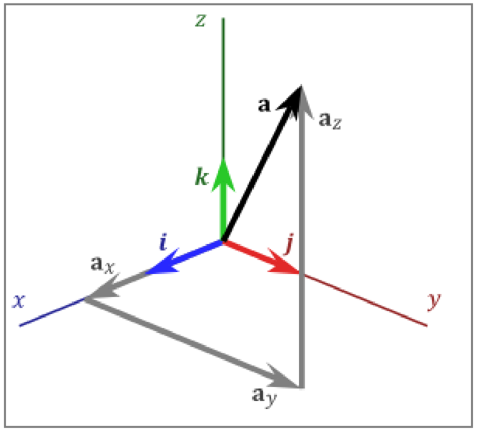


REPRESENTATION

Each data item is an N-dimensional vector (N variables)

- recall 2D and 3D vectors in 2D and 3D space, respectively

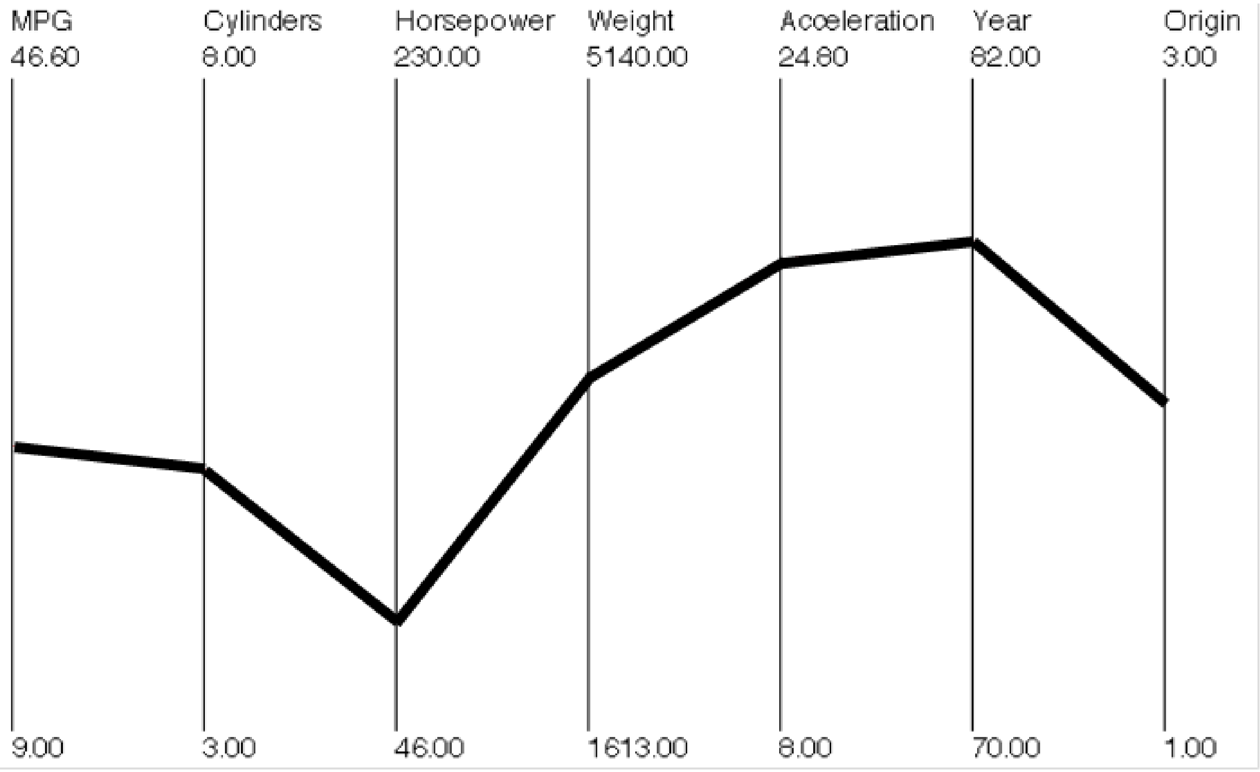




Now we have N-D attribute space

- the data axes extend into more than 3 orthogonal directions
- hard to imagine?
- that's why need good visualization methods

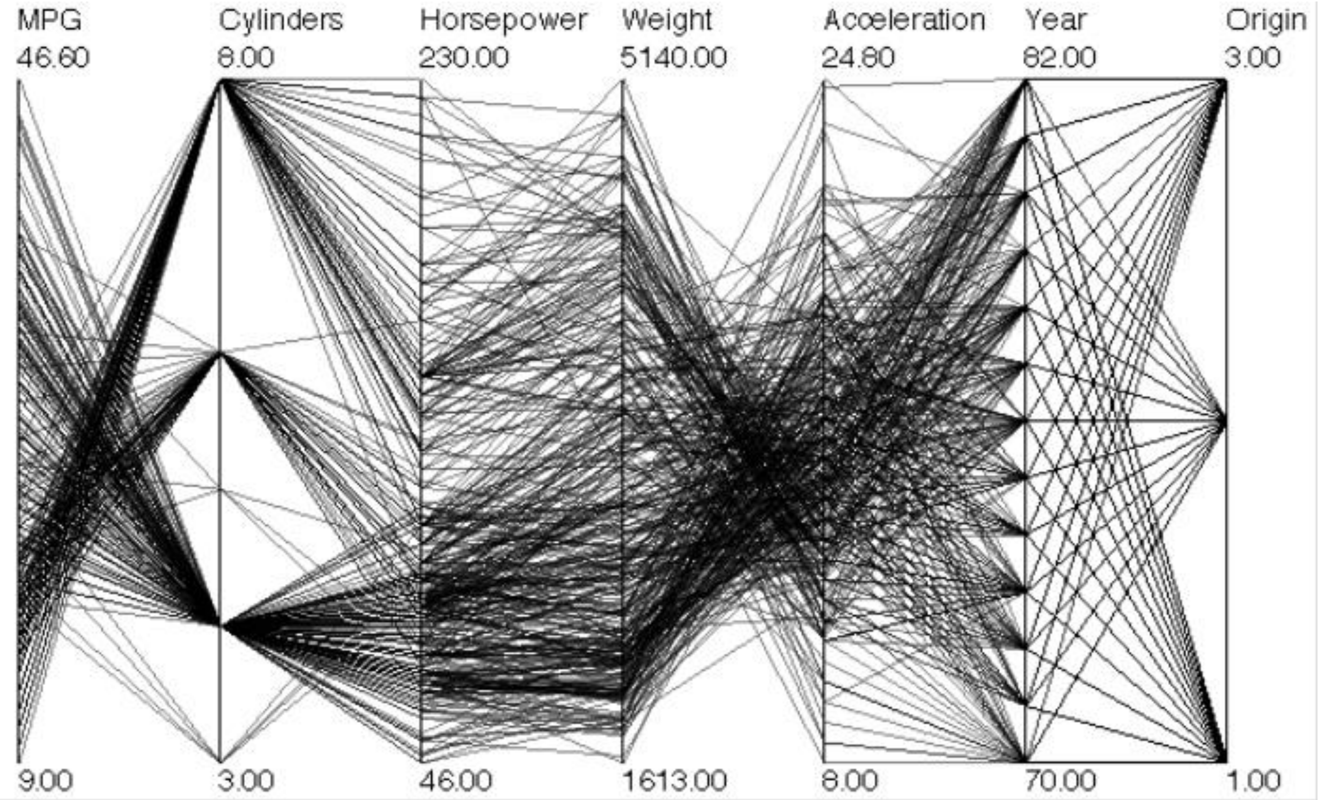
Parallel Coordinates – 1 Car



The N=7 data axes are arranged side by side

- in parallel

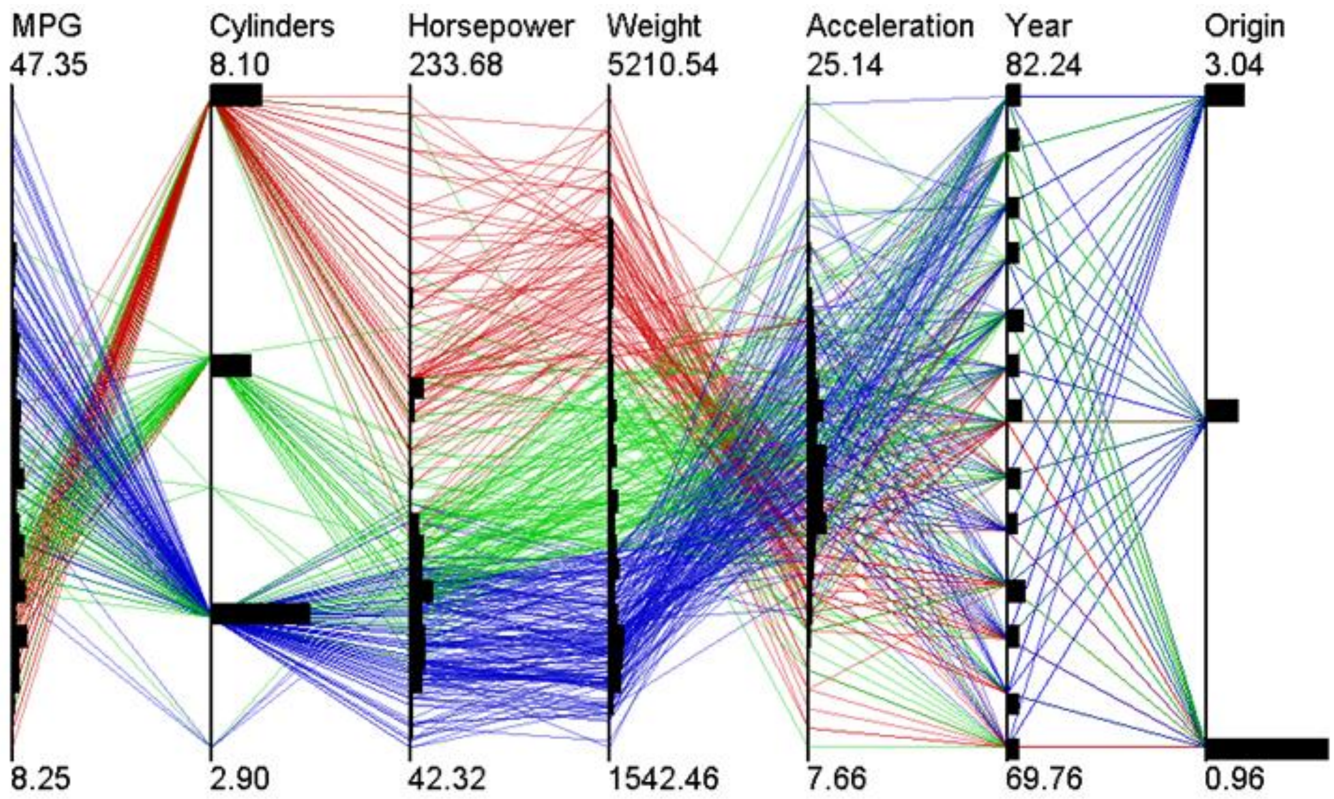
PARALLEL COORDINATES – 100 CARS



Hard to see the individual cars?

- what can we do?

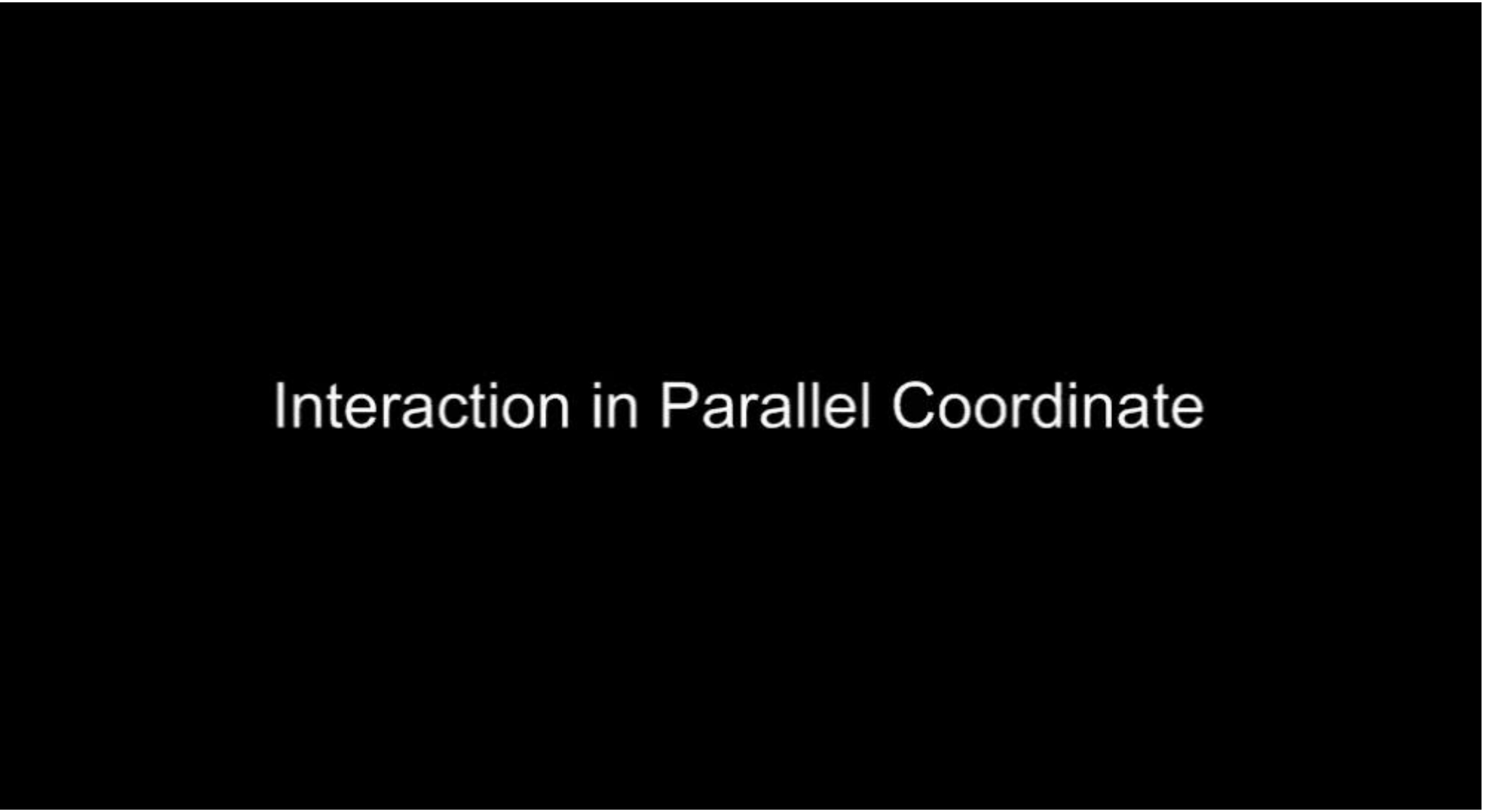
Parallel Coordinates – 100 Cars



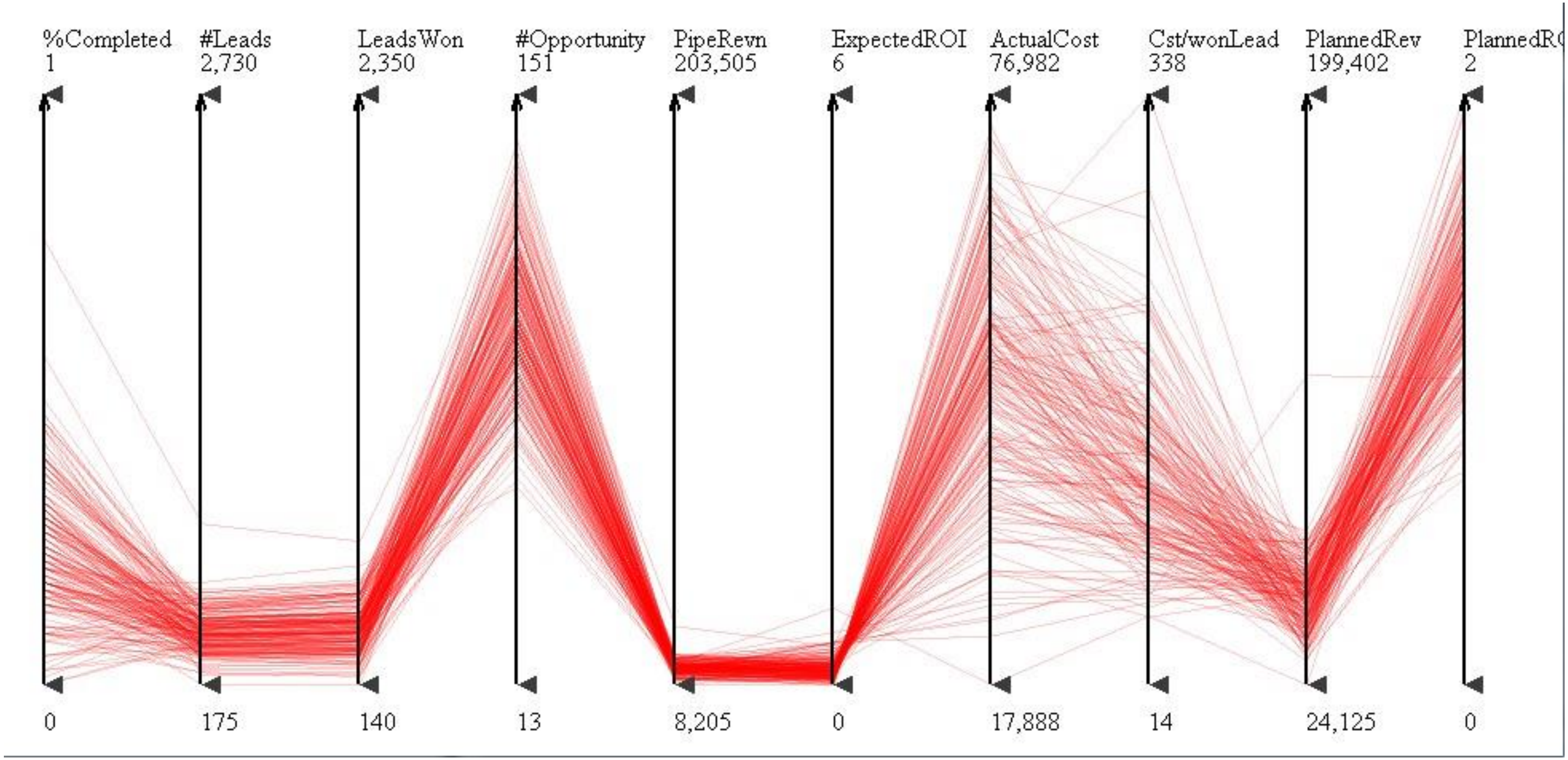
Grouping the cars into sub-populations

- this is called *clustering*
- can be automated or interactive (put the user in charge)

Interactive Clustering With parallel Coordinates

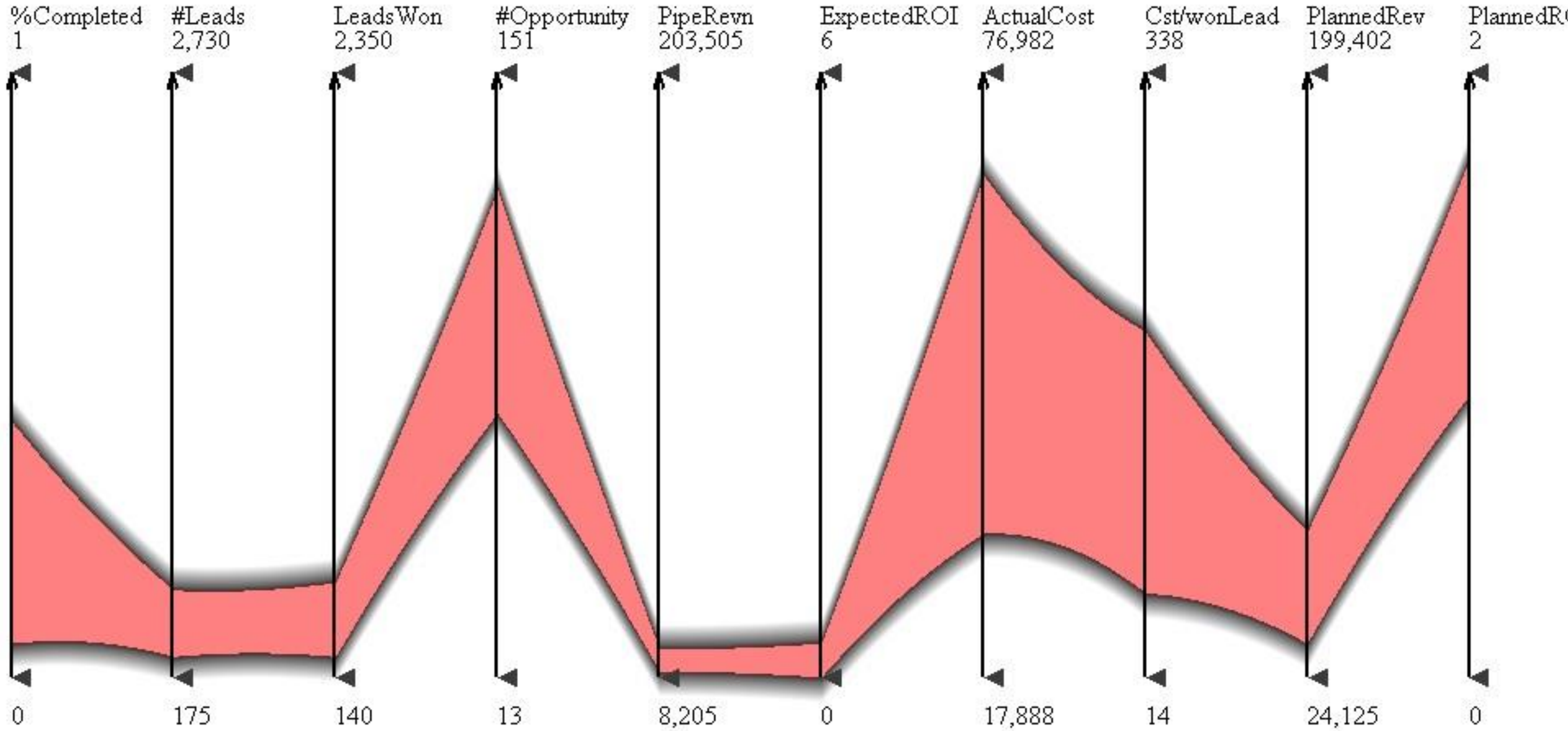


Illustrative Abstraction



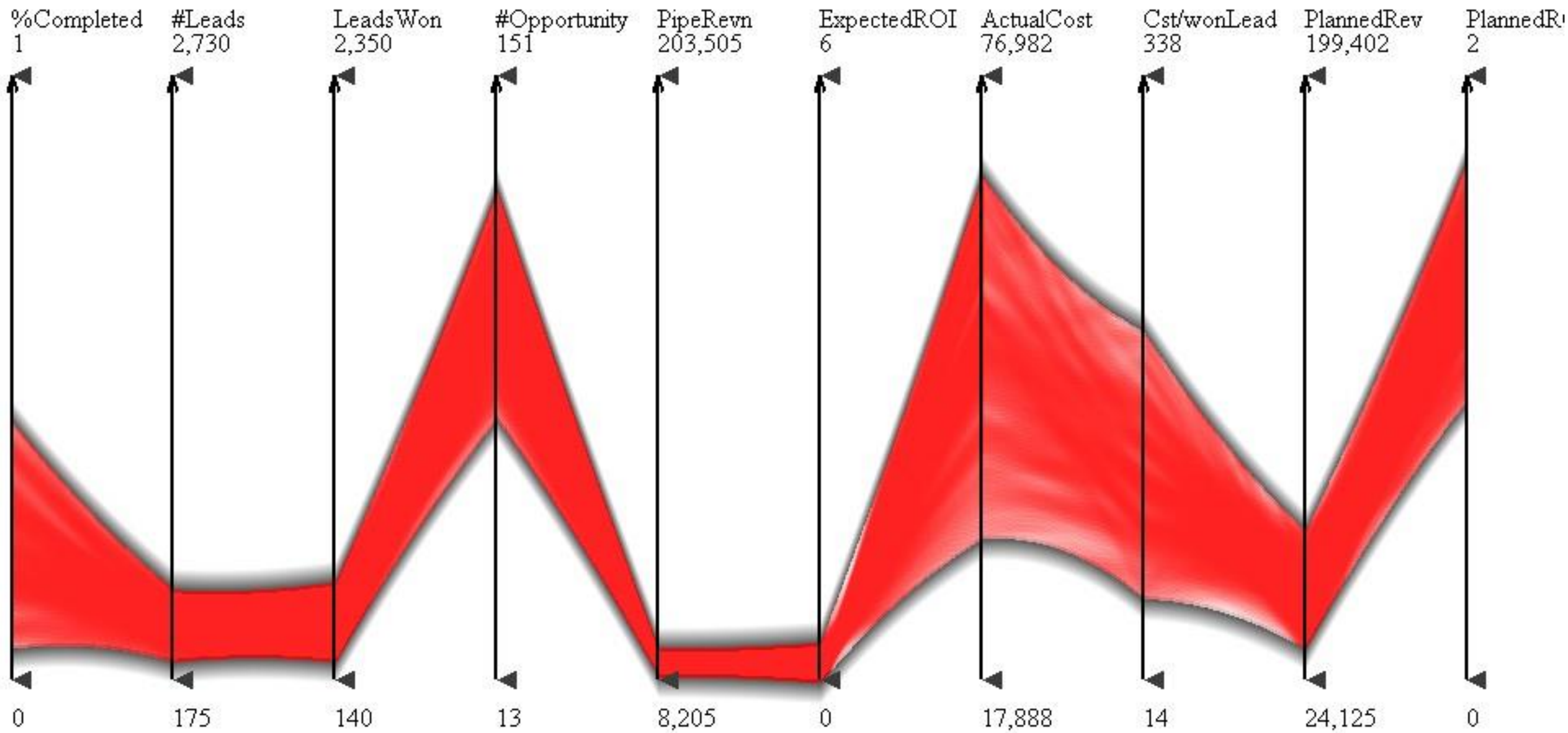
individual polylines

PC With Illustrative Abstraction



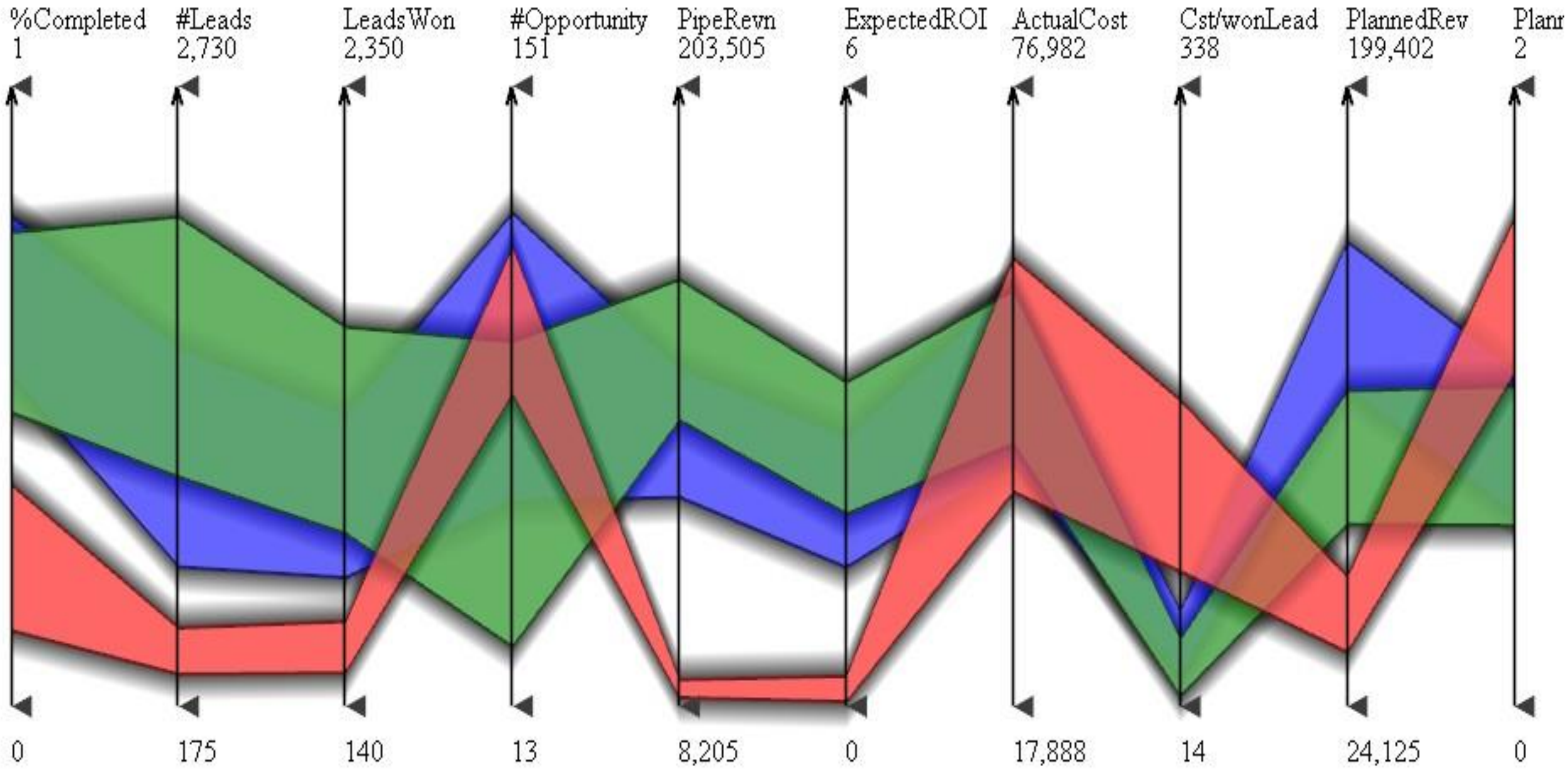
completely abstracted away

PC With Illustrative Abstraction



blended partially

PC With Illustrative Abstraction

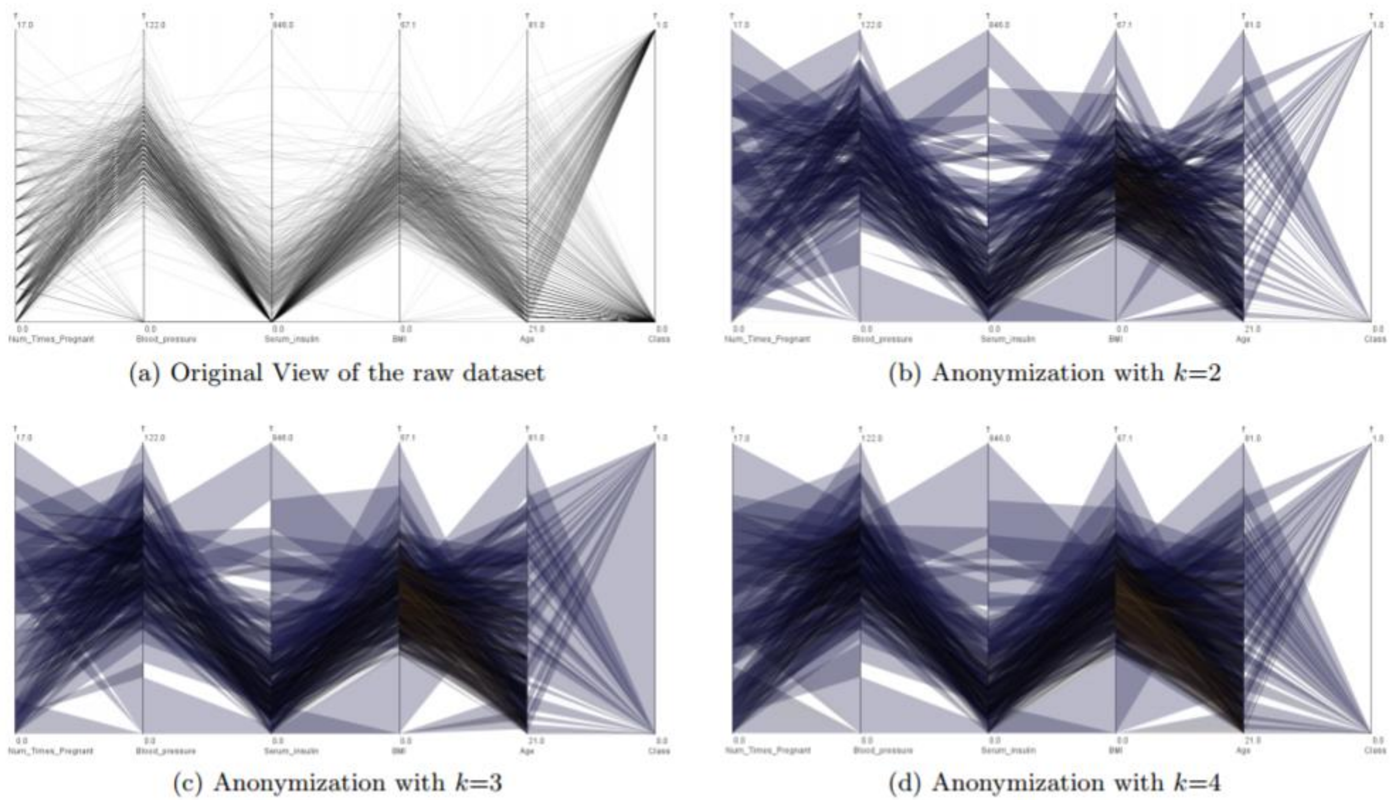


all put together – three clusters

Data Privacy With Parallel Coordinates Using k-Anonymity

Cluster records intro 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