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

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 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 Century	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 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**

Data types

Quantitative (Numerical)

Categorical (Ordinal)

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

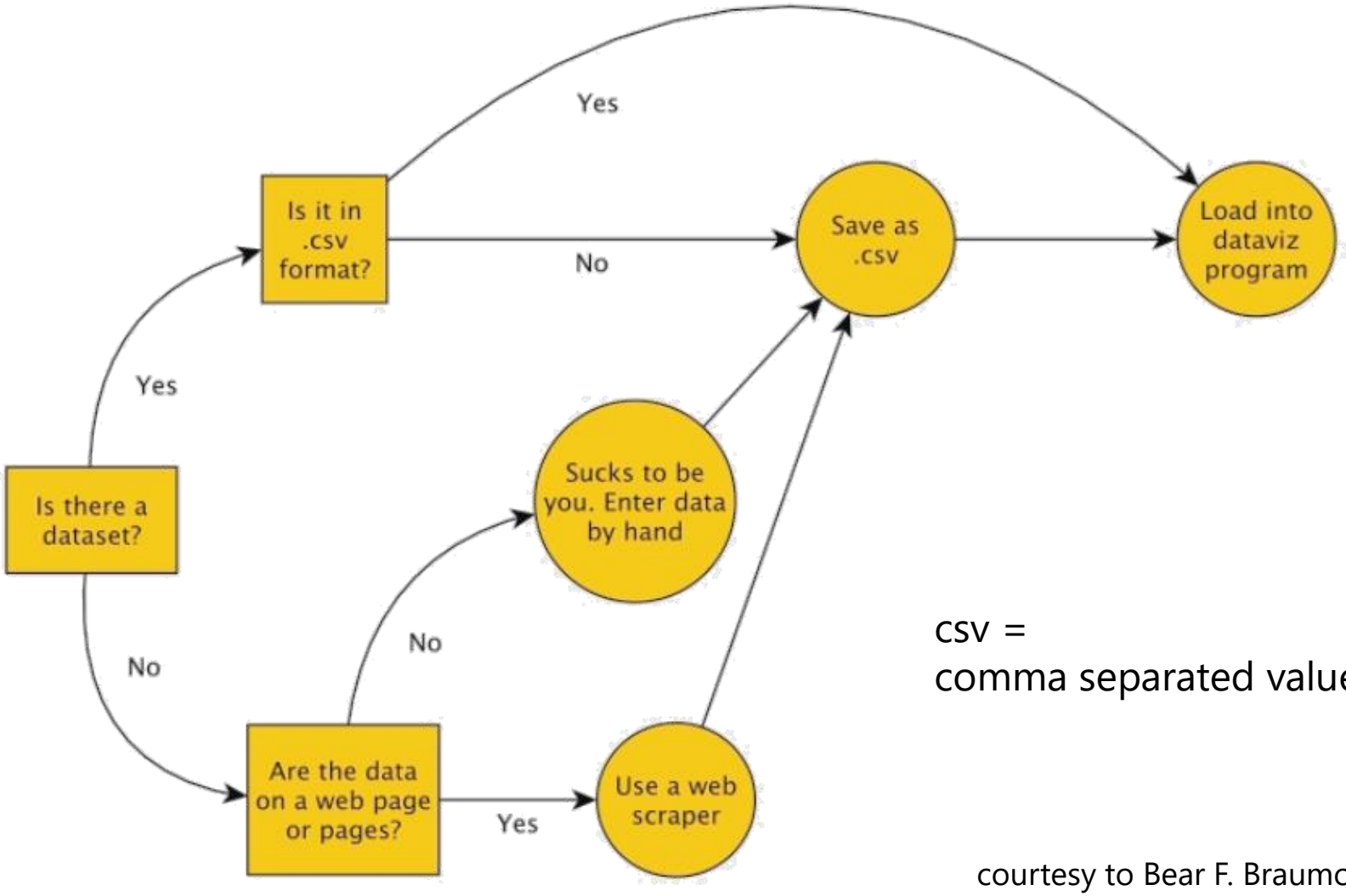
Wikipedia also has lots of data in tables

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

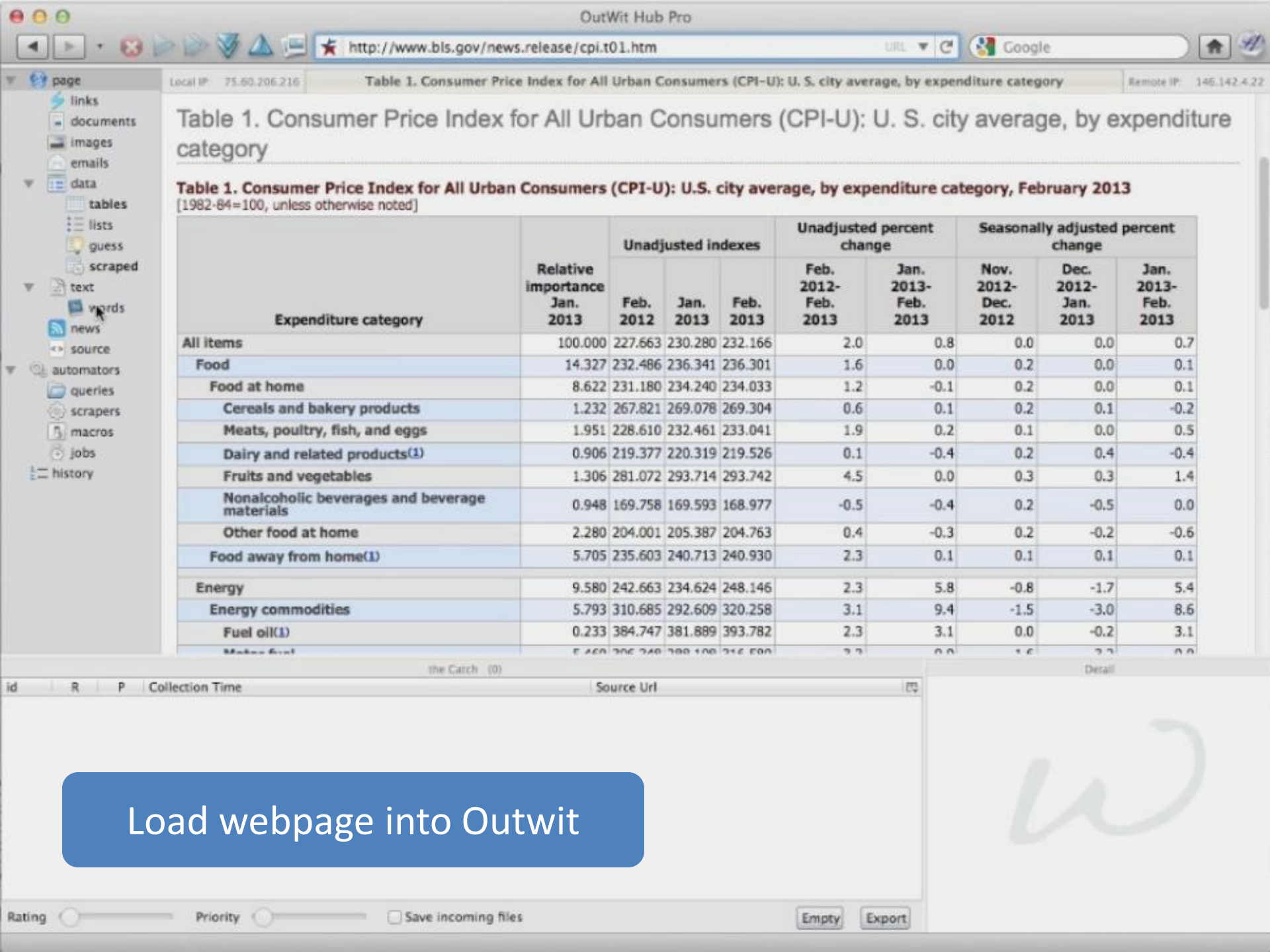
- try cut and paste into Excel

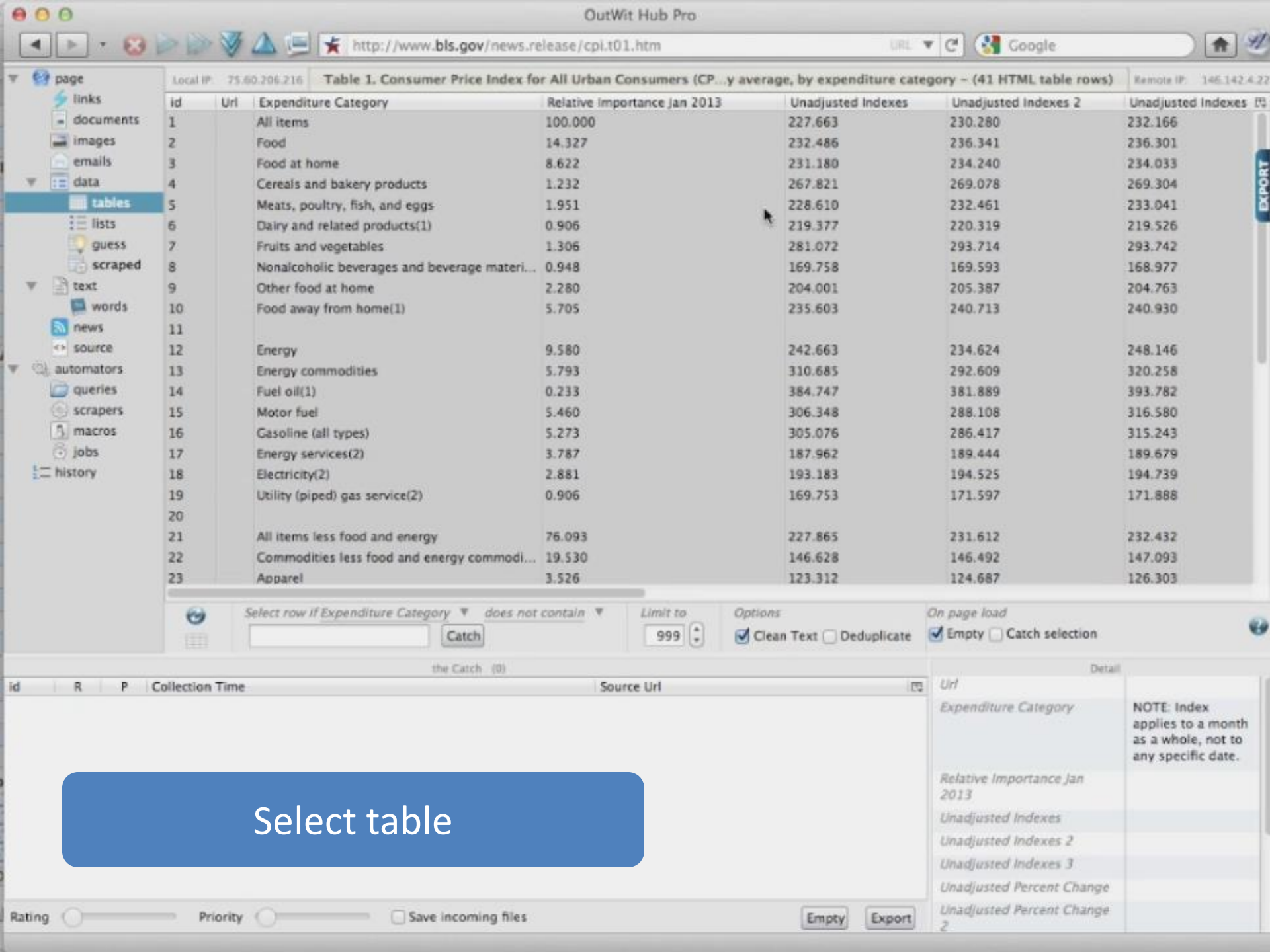
If the data are on one page but cut/paste is not working

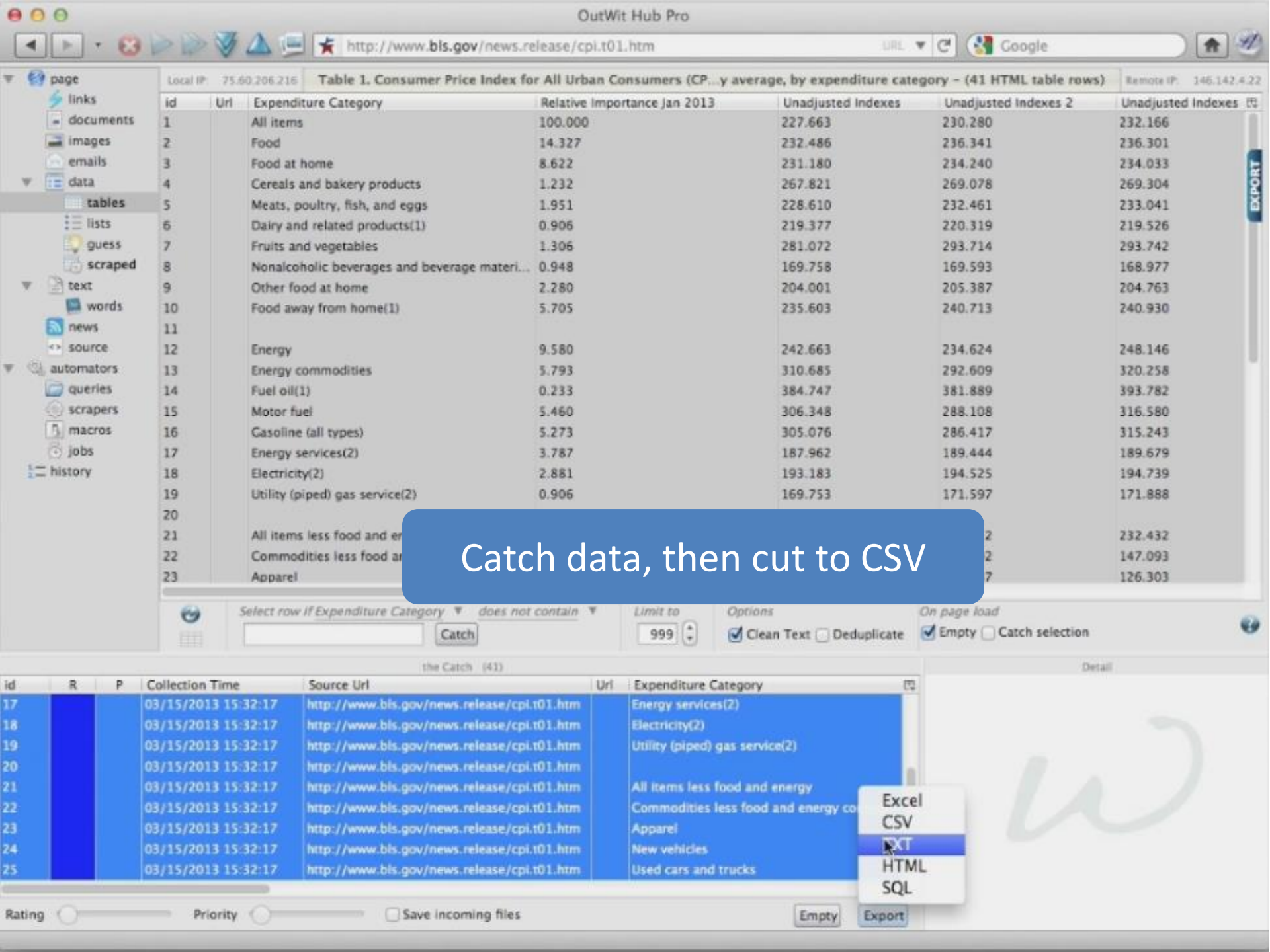
- try a web scraper like Outwit Hub

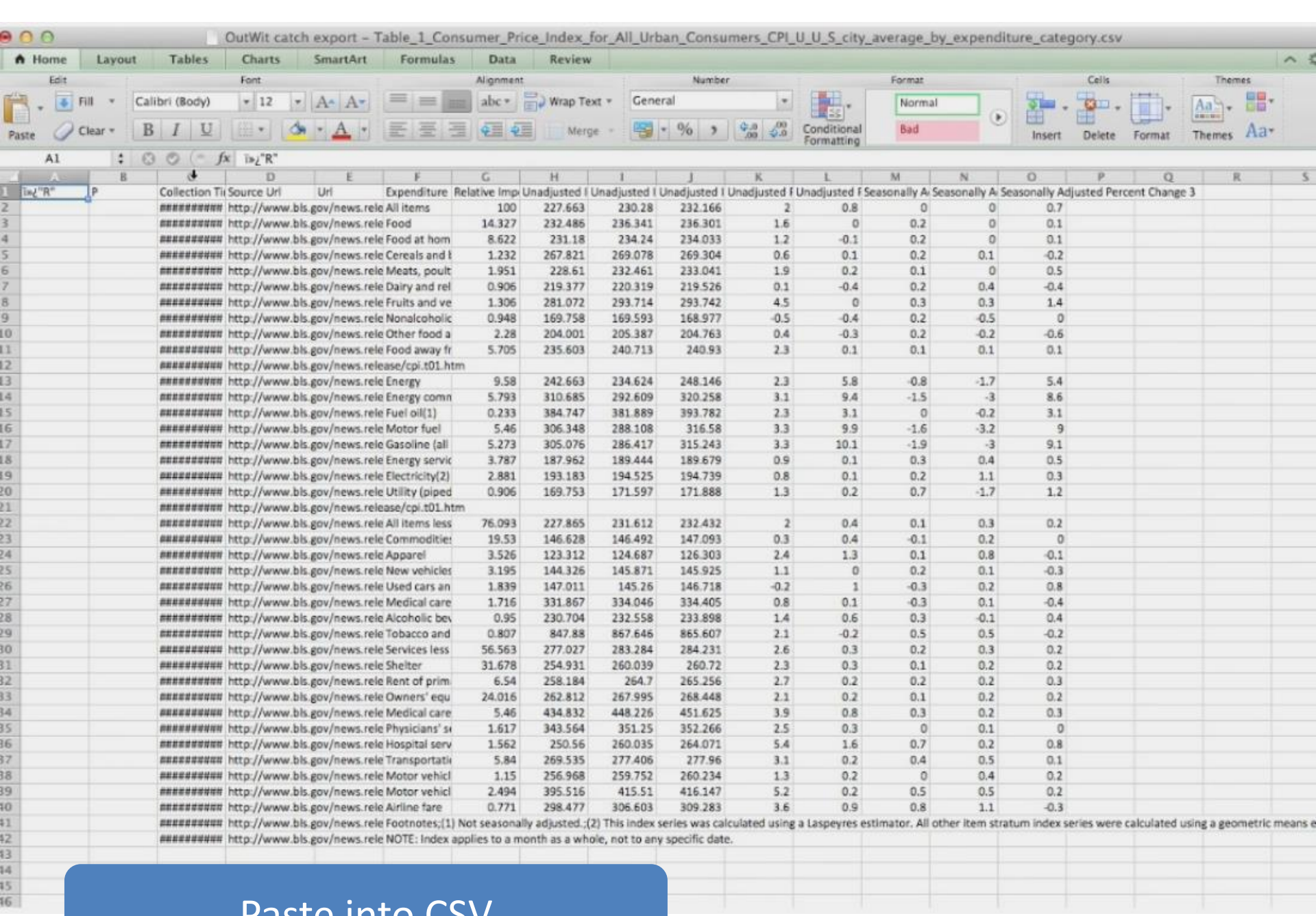
If the data are spread across multiple webpages

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

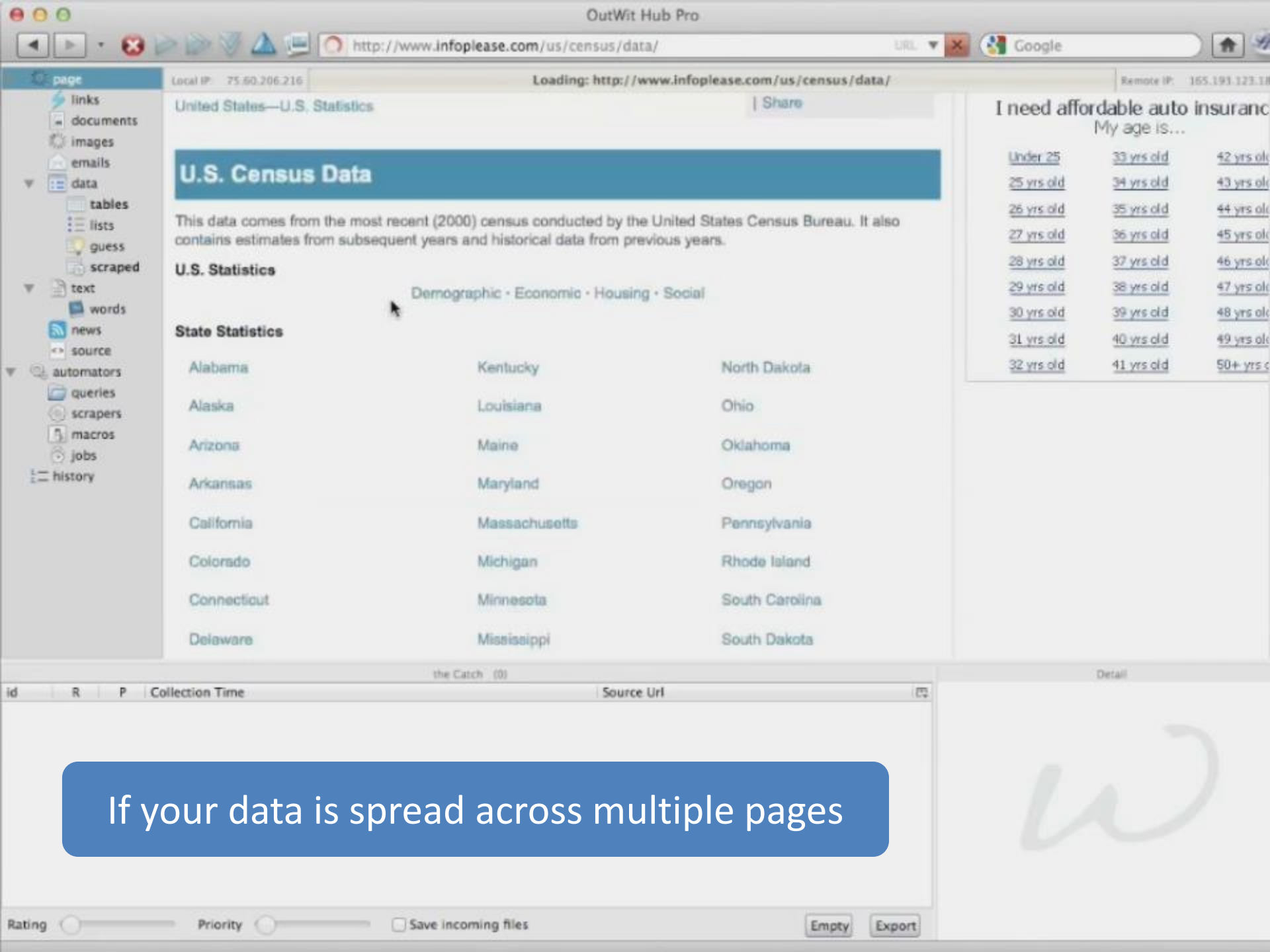

OutWit Hub Pro
http://www.bls.gov/news.release/cpi.t01.htm
Google
page
links
documents
images
emails
data
tables
lists
guess
scraped
text
words
news
source
automators
queries
scrapers
macros
jobs
history
Table 1. Consumer Price Index for All Urban Consumers (CPI-U): U. S. city average, by expenditure category
Table 1. Consumer Price Index for All Urban Consumers (CPI-U): U. S. city average, by expenditure category
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
Feb. 2012
Jan. 2013
Feb. 2013
Unadjusted percent change
Feb. 2012-Feb. 2013
Jan. 2013-Feb. 2013
Seasonally adjusted percent change
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
the Catch (0)
Detail
id
R
P
Collection Time
Source Url
Load webpage into Outwit
Rating
Priority
Save incoming files
Empty
Export


OutWit Hub Pro
http://www.bls.gov/news.release/cpi.t01.htm
Google
page
links
documents
images
emails
data
tables
lists
guess
scraped
text
words
news
source
automators
queries
scrapers
macros
jobs
history
Table 1. Consumer Price Index for All Urban Consumers (CP...y average, by expenditure category - (41 HTML table rows)
id | Url | Expenditure Category | Relative Importance Jan 2013 | Unadjusted Indexes | Unadjusted Indexes 2 | Unadjusted Indexes
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
EXPORT
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
Detail
Url
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

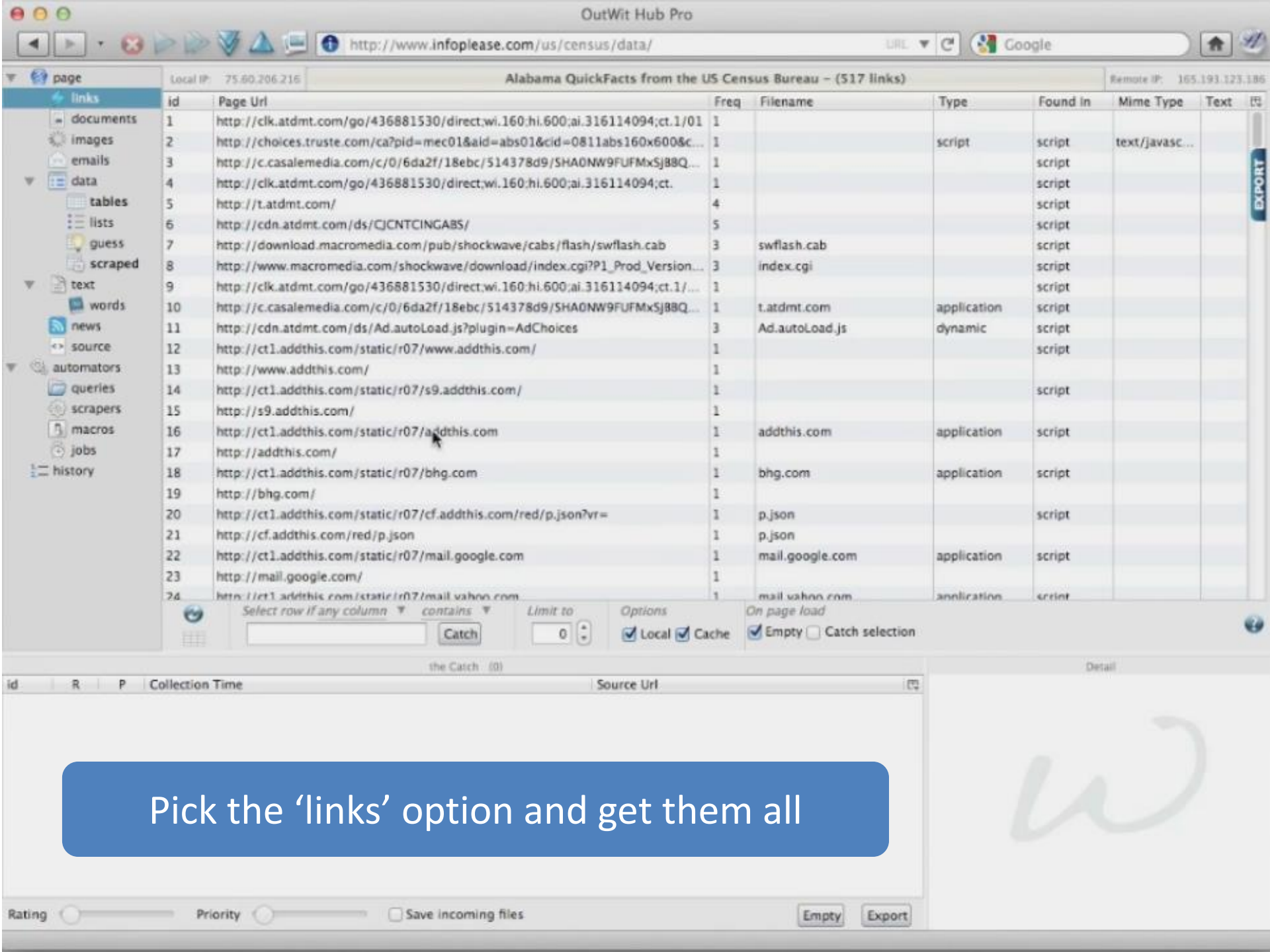

OutWit Hub Pro
http://www.bls.gov/news.release/cpi.t01.htm
Google
page
links
documents
images
emails
data
tables
lists
guess
scraped
text
words
news
source
automators
queries
scrapers
macros
jobs
history
Table 1. Consumer Price Index for All Urban Consumers (CP...y average, by expenditure category - (41 HTML table rows)
id | Url | Expenditure Category | Relative Importance Jan 2013 | Unadjusted Indexes | Unadjusted Indexes 2 | Unadjusted Indexes
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 en | 232.432
22 | Commodities less food ar | 147.093
23 | Apparel | 126.303
EXPORT
Catch data, then cut to CSV
Select row if Expenditure Category does not contain
Catch
Limit to 999
Options
Clean Text Deduplicate
On page load
Empty Catch selection
the Catch (41)
Detail
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
Rating
Priority
Save incoming files
Empty
Export



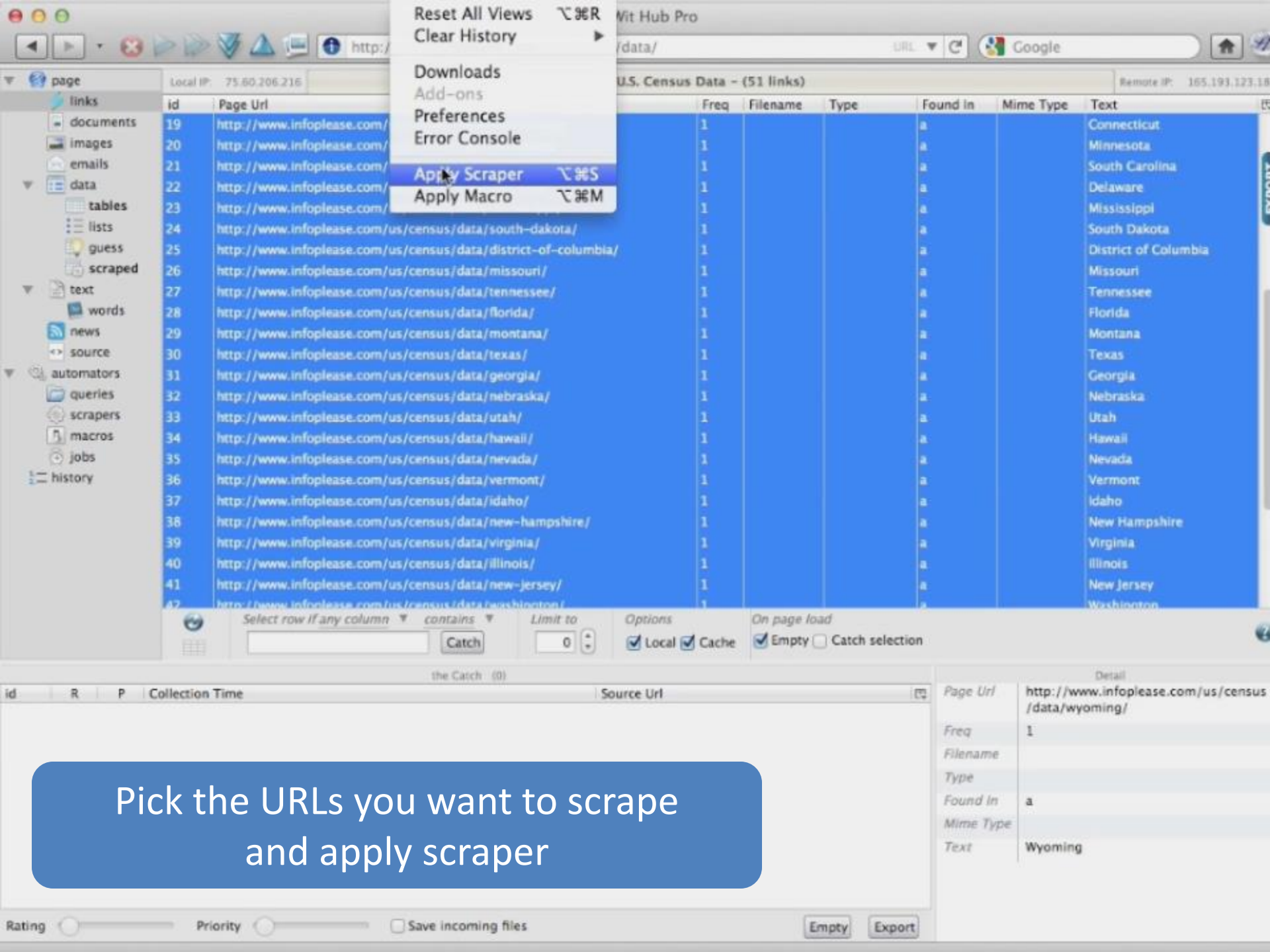
Paste into CSV



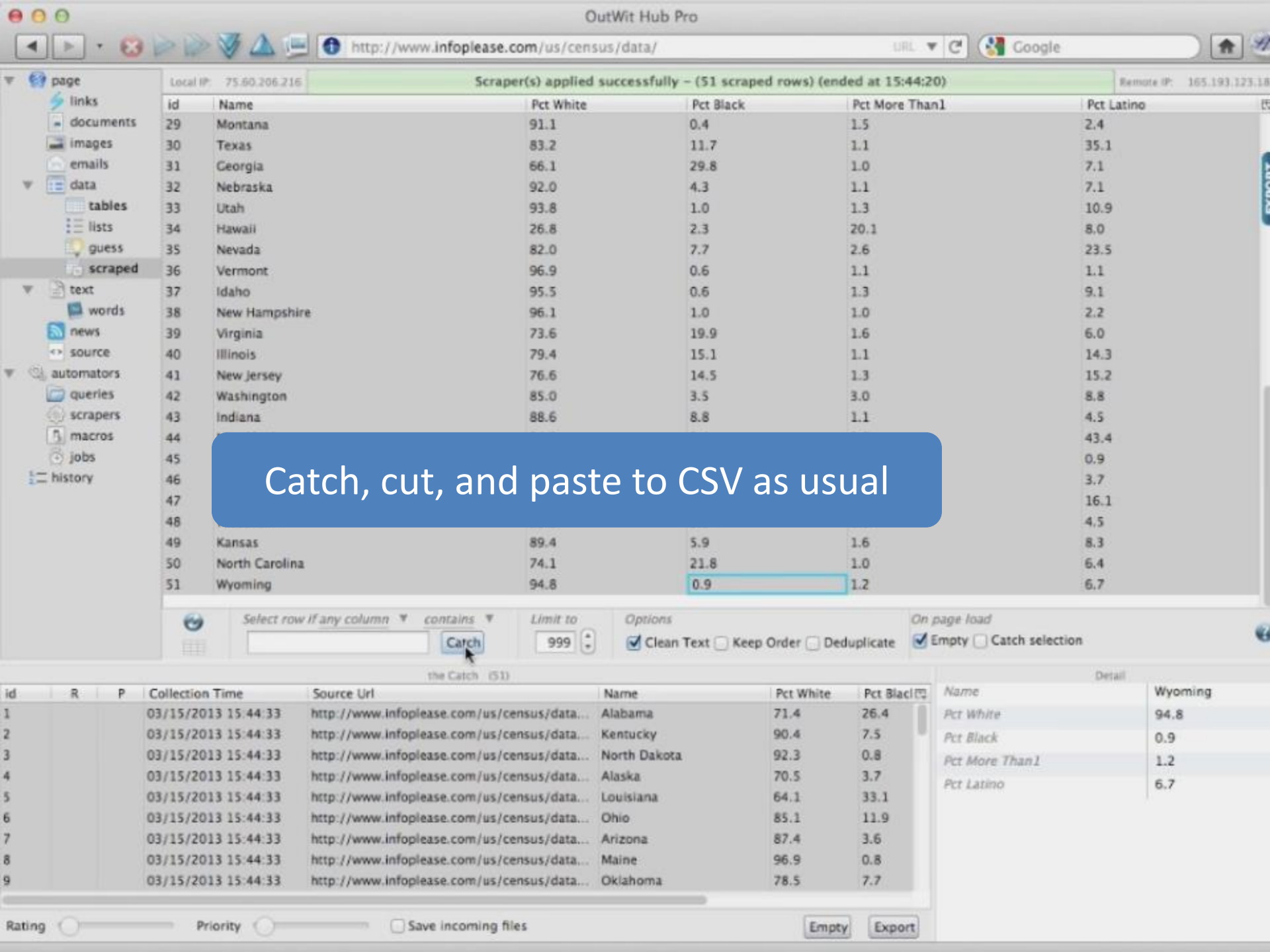
If your data is spread across multiple pages



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

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

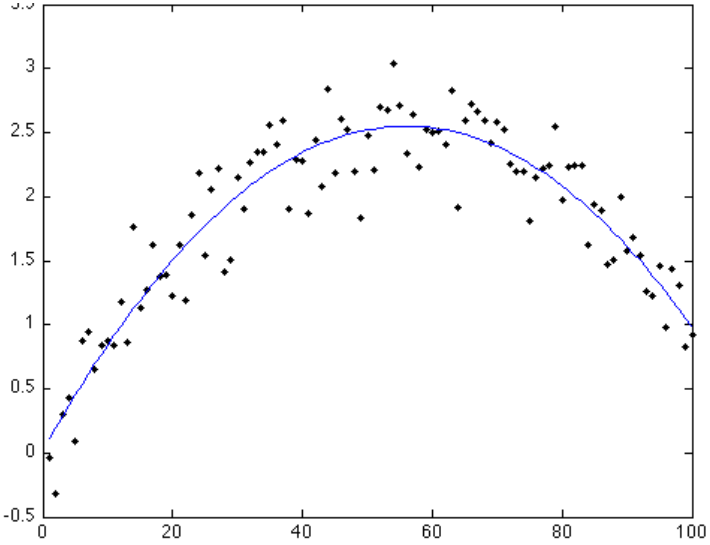
Missing Data – Example

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

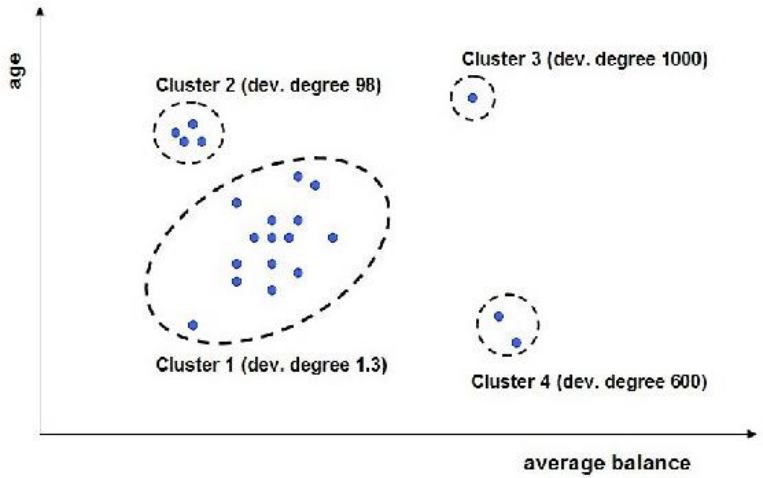
Other data problems which require data cleaning

- duplicate records
- incomplete data
- inconsistent data

Noisy Data – What To Do

Binning method

- discussed last lecture



Clustering

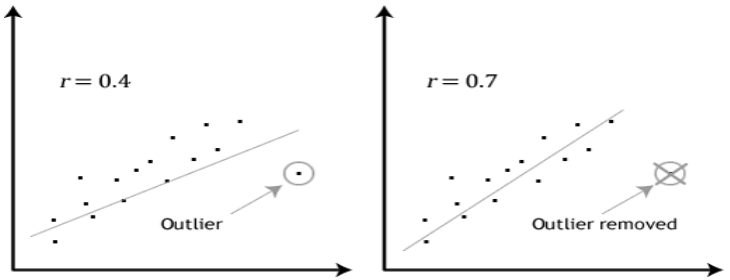
- detect and remove outliers

Semi-automated method

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

Regression

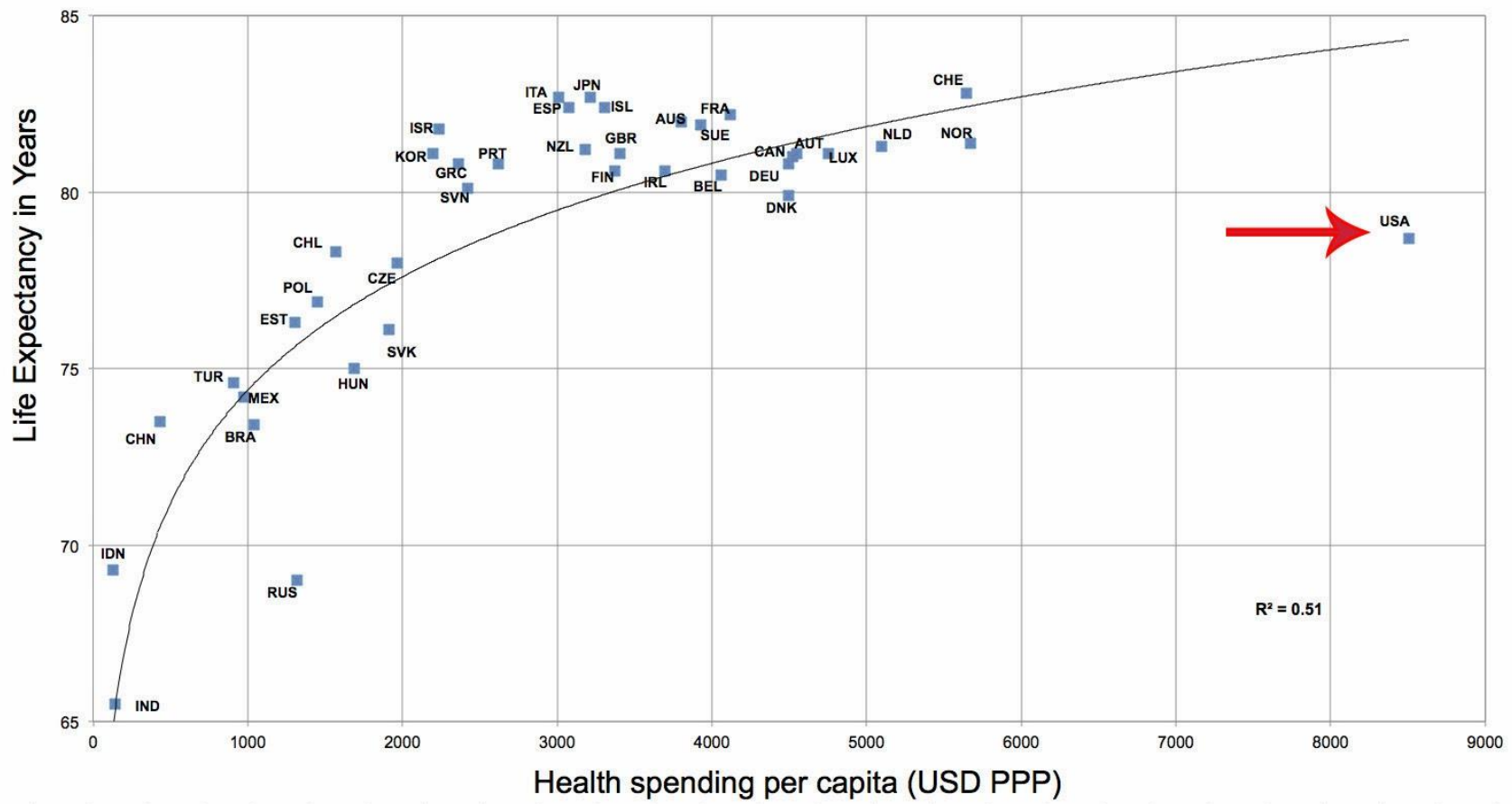
- smooth by fitting the data to a regression function



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

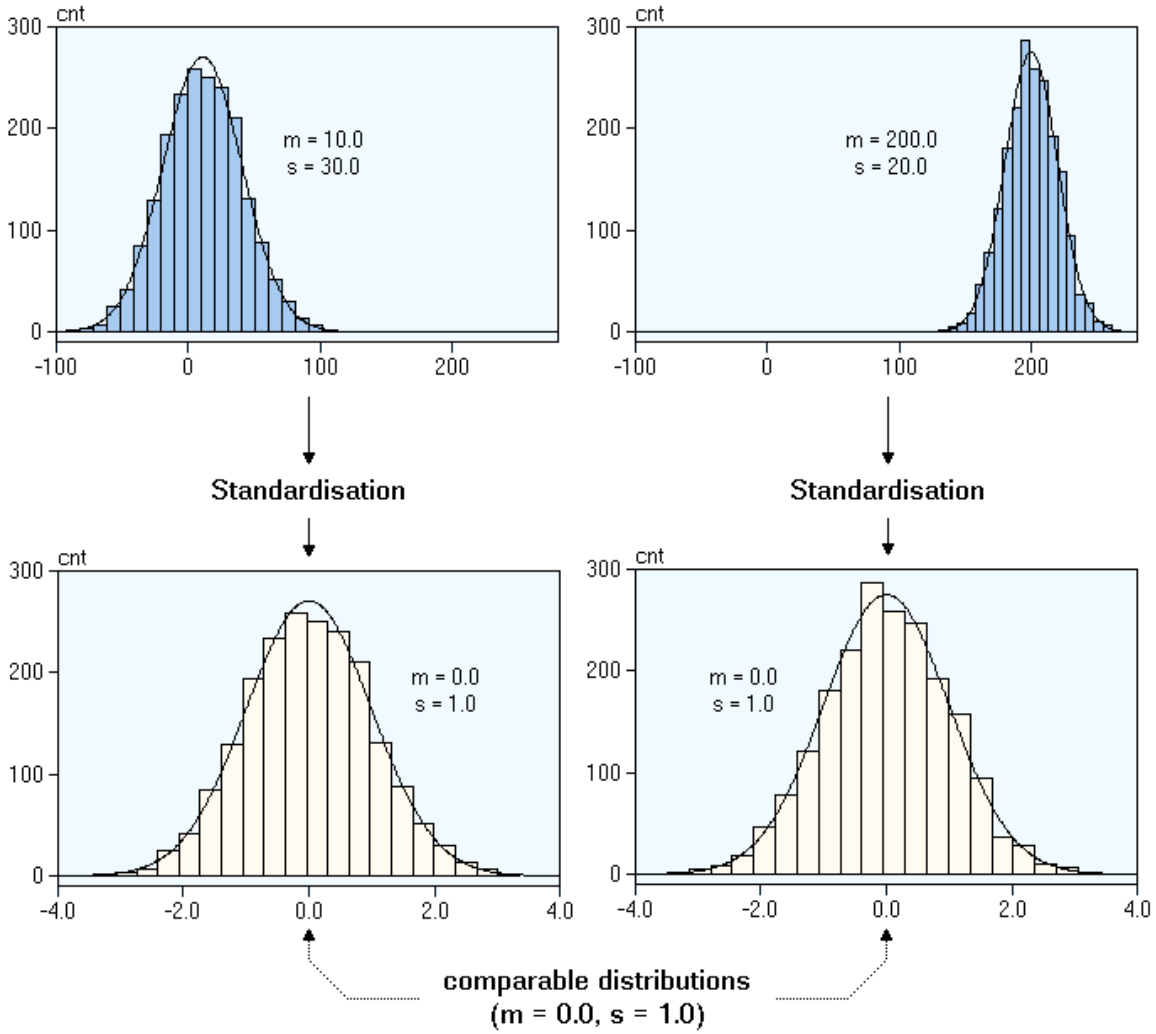
- min-max normalization

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

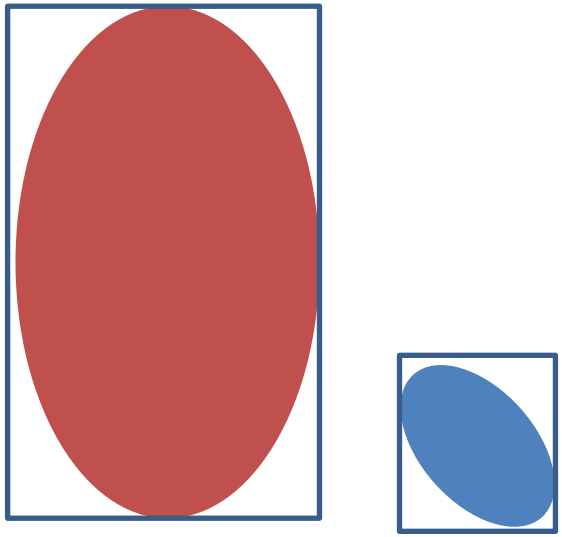
- standardization / z-score normalization

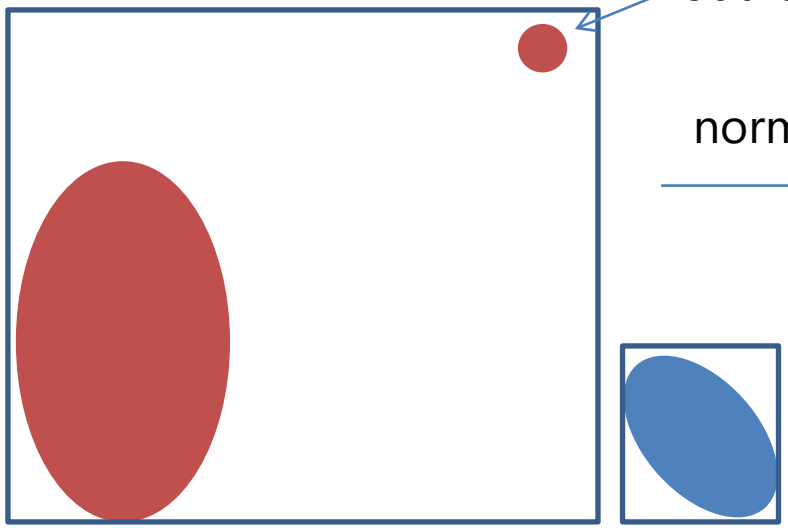
$$v' = \frac{v - \overline{v}}{\sigma_v}$$

Standardization



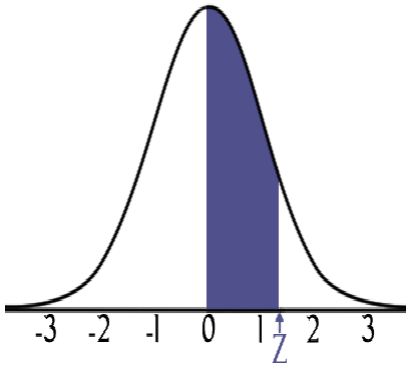
Normalization



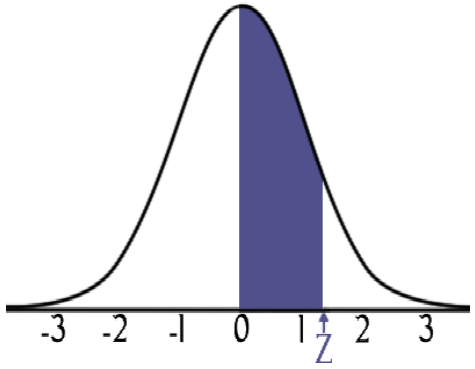


Standardization

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

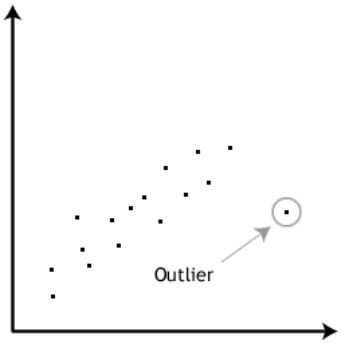
Data Fusion vs. Data Privacy

Data fusion can bring insight

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

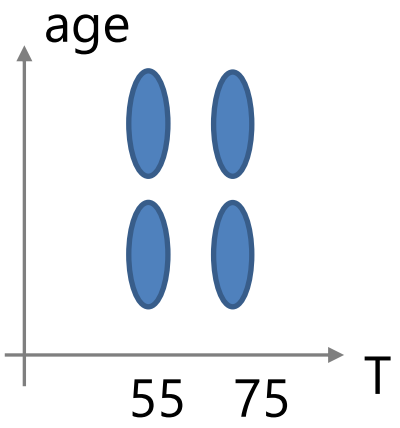
Visualization can bring insight

- 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?

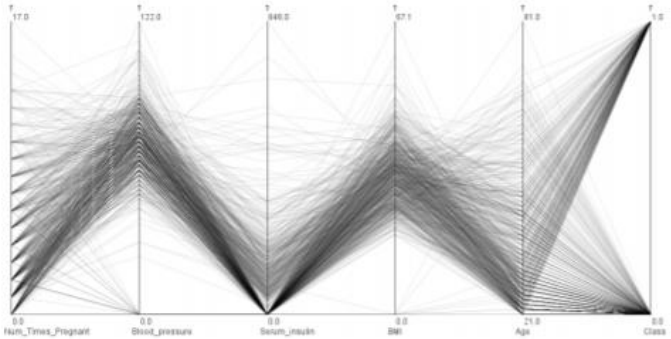
- k-anonymity (generalize)
- make data less specific → binning
- age *groups*, zip code *groups,* etc…
- make blobs instead of points



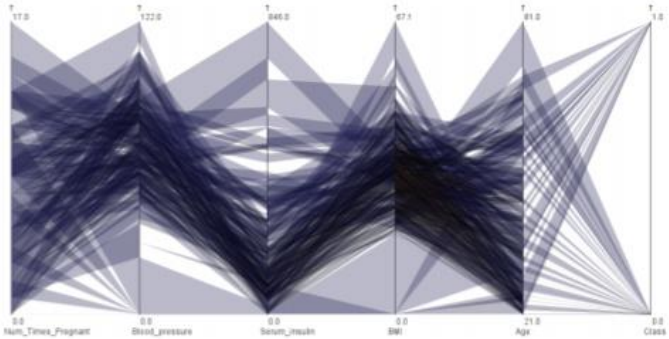
Data Privacy With Parallel Coordinates Using k-Anonymity

Cluster records intro k-sized bins for each variable/dimension

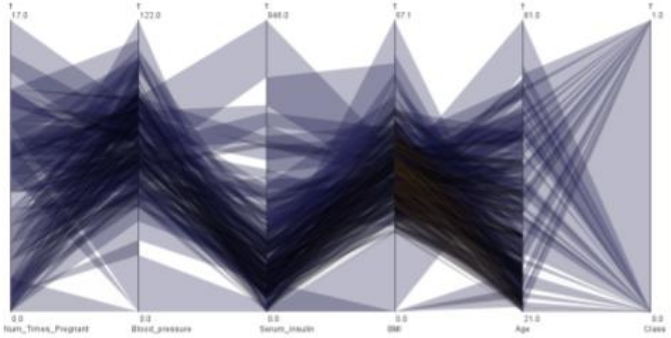
- Dasgupta and Kosara show this for parallel coordinates [TVCG, 2011]



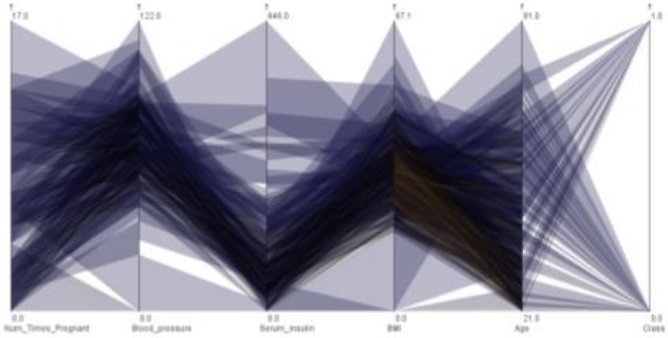
(a) Original View of the raw dataset



(b) Anonymization with k=2



(c) Anonymization with k=3



(d) Anonymization with k=4

Data Reduction

Sampling

- random
- stratified

Data summarization

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

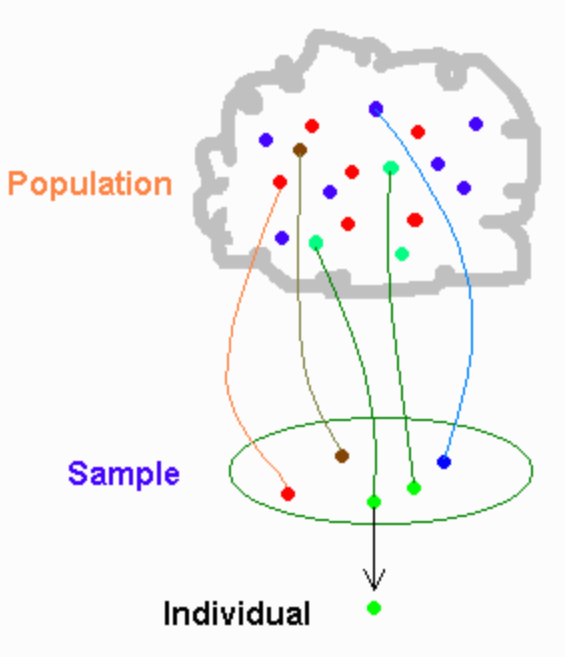
Sampling

The goal

- pick a representative subset of the data

Random sampling

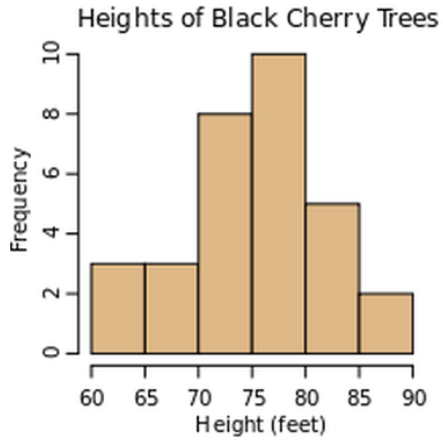
- pick sample points at random
- will work if the points are distributed uniformly
- this is usually not the case
- outliers will likely be missed
- so the sample will not be representative

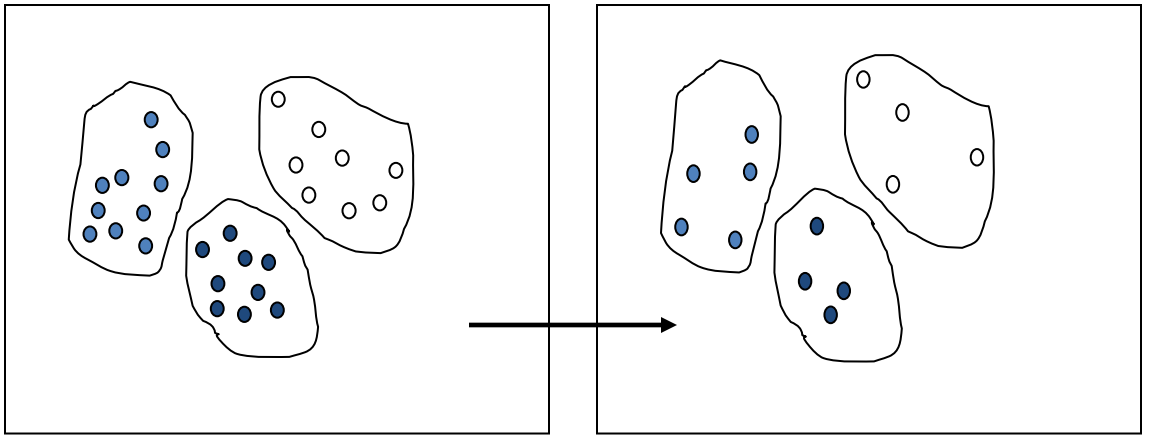


Adaptive Sampling

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 ~ bin height

sampling rate ~ cluster size

What's When Your data Is Too Small

Can you "hallucinate" or "invent" realistic data?

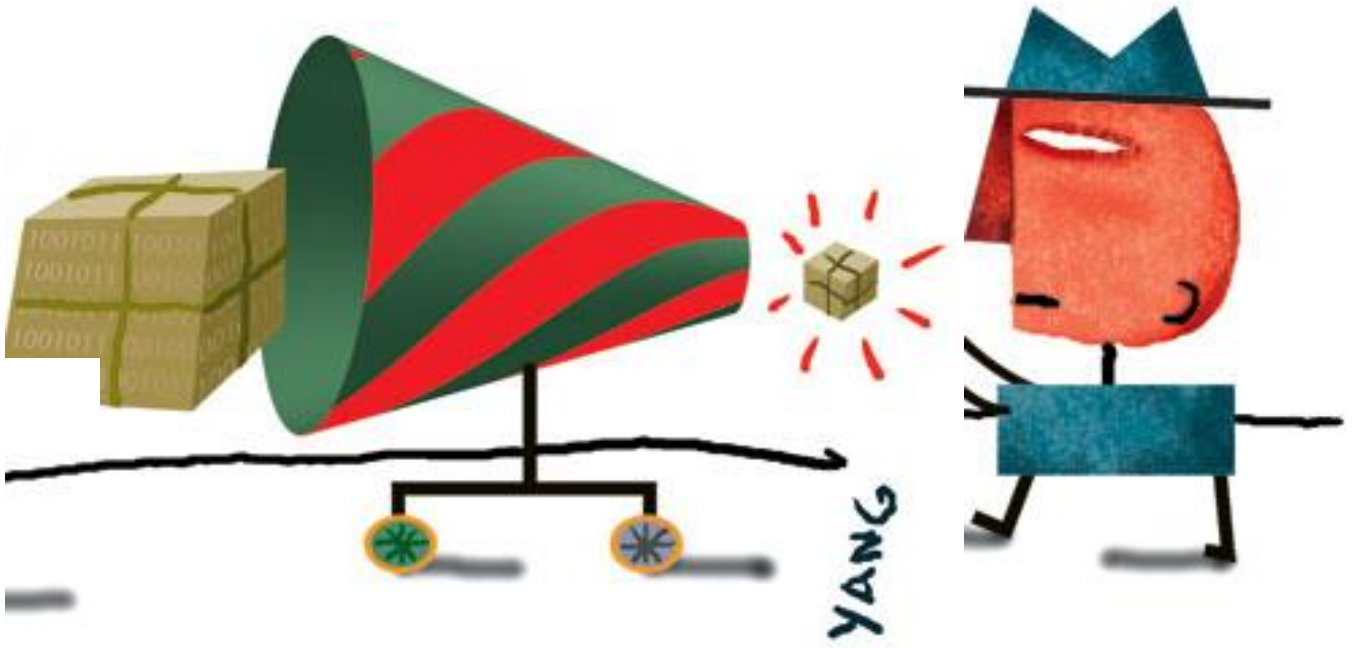
And if so, how would you go about this?

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

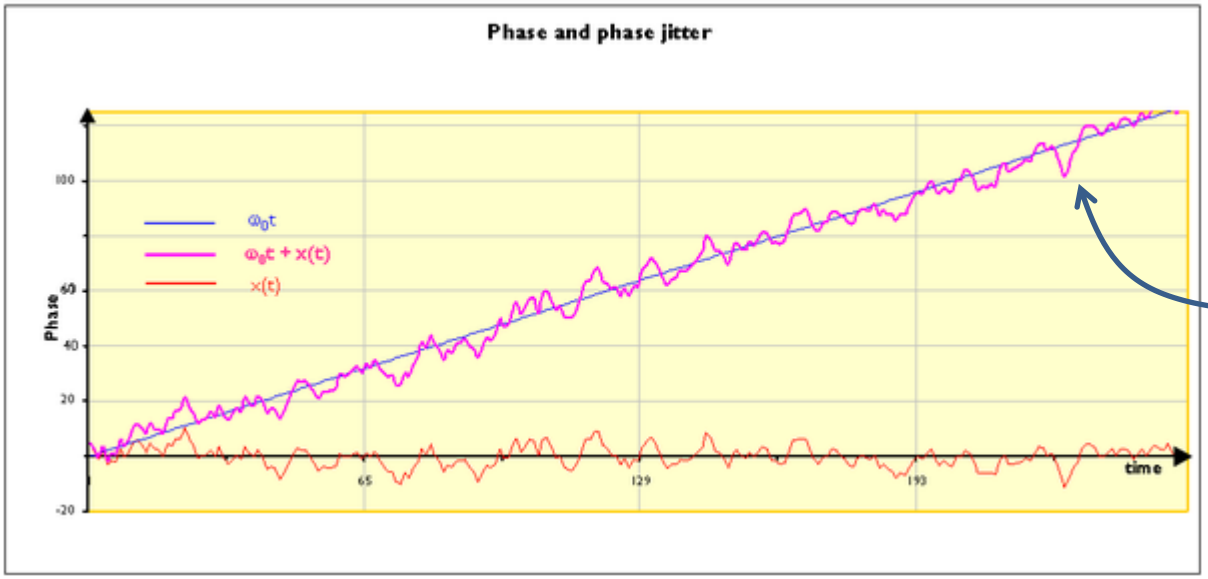
Common techniques are (for images)

- 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"



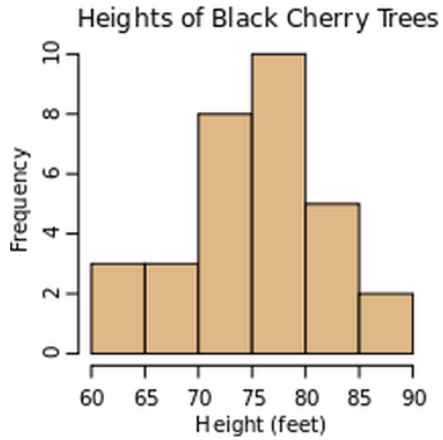
Jitter (noise)

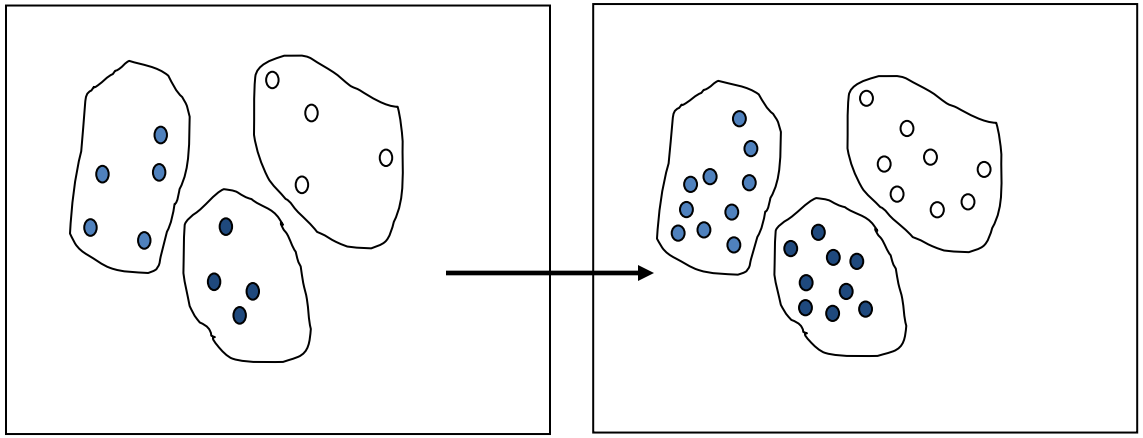
- small random noise about a steady signal

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 ~ bin height

augmentation rate ~ 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