Cheaper Beds, **Better** Breakfasts

Aid users in their discovery of relevant listings

For now, we are interested not in listings themselves but the context around the listing

Which neighborhoods are the most explorable?

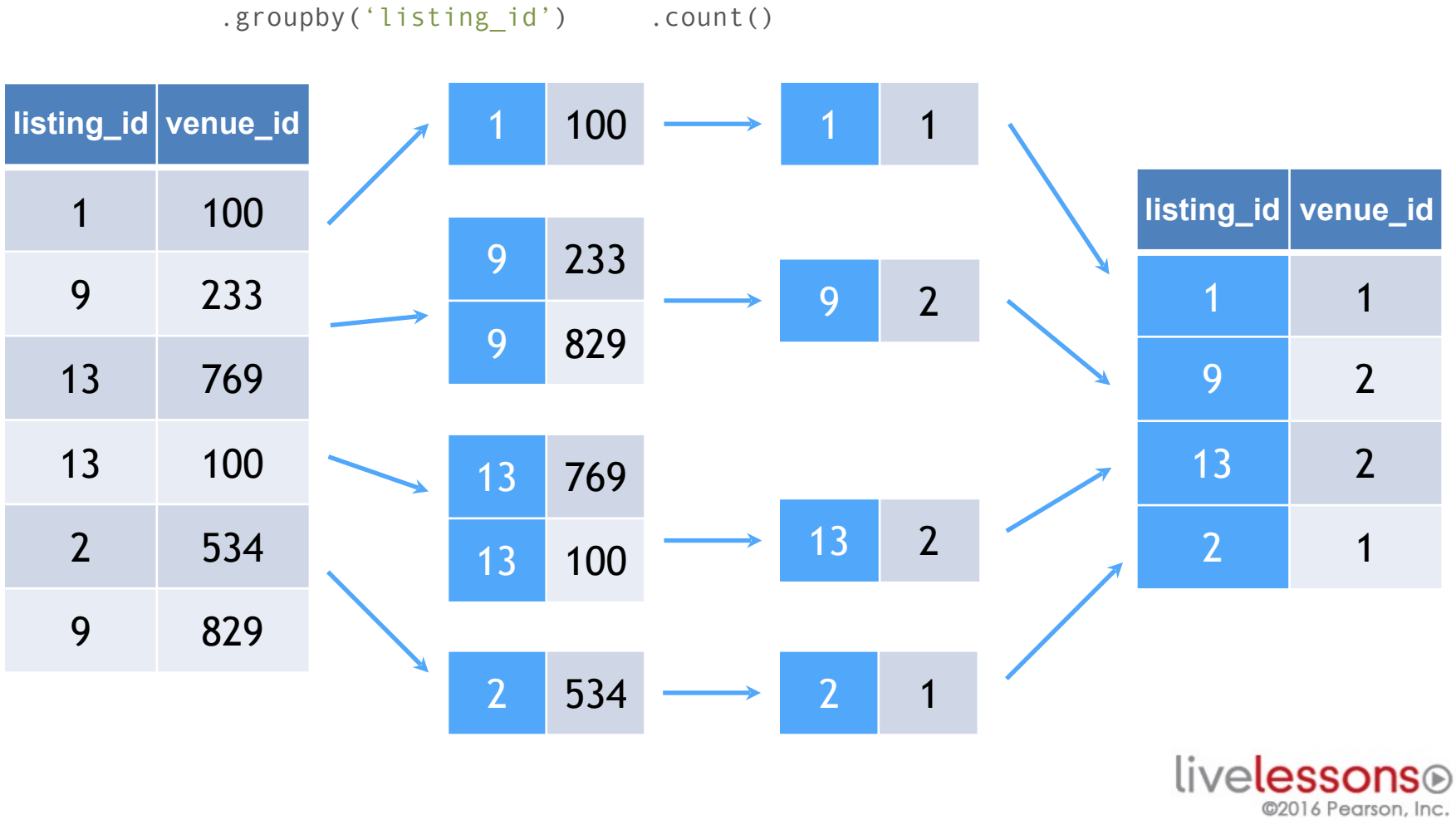
Which listings have the most lively and active POI around them?

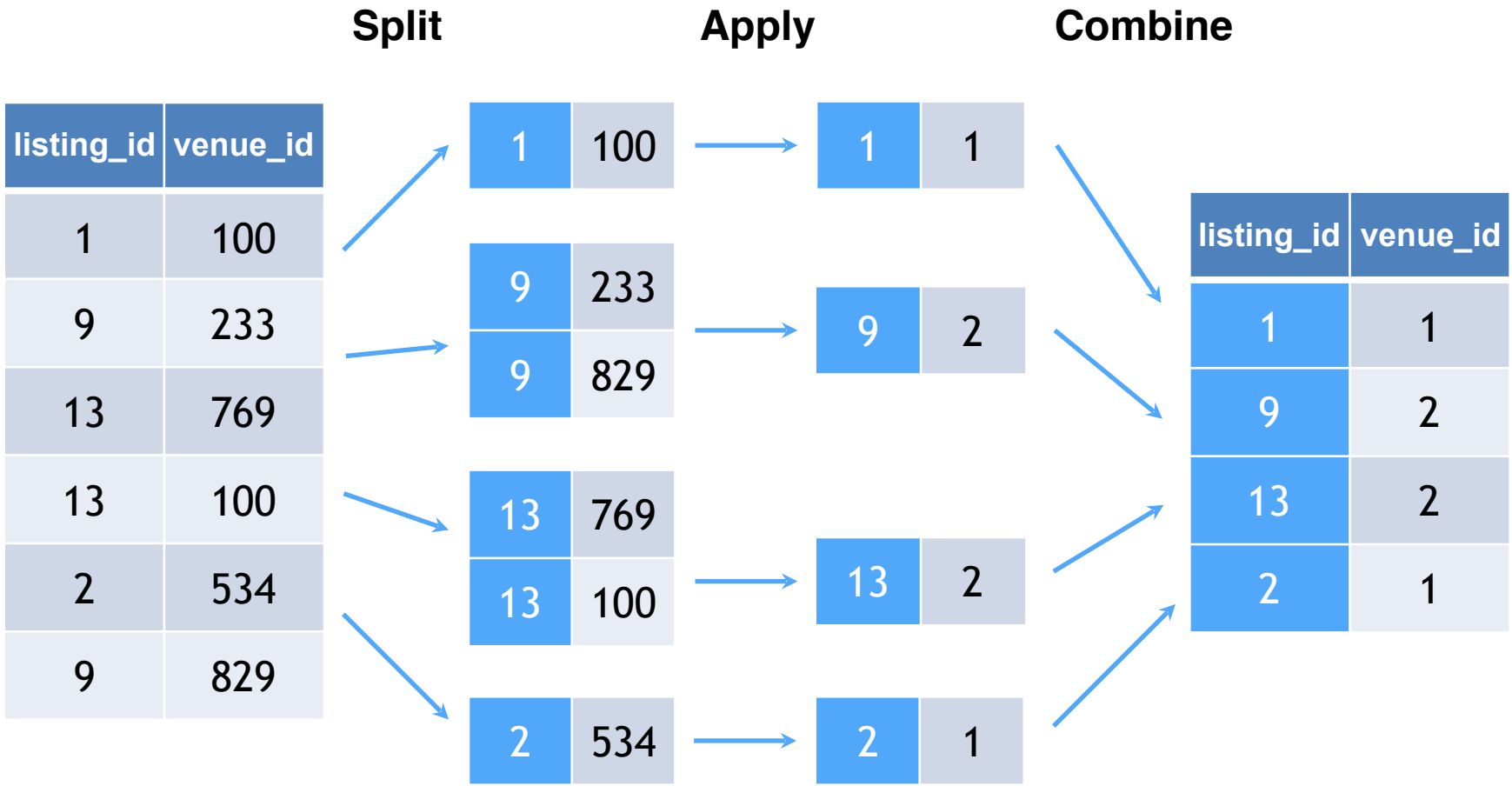
ETL Pipeline

ETL Pipeline

ETL(Q) Pipeline

In pandas**!**





In pandas**!**

Extract	Transform	Load	Query
read_csv()	transform()	to_csv()	query()
read_excel()	dropna()	to_excel()	sort_values()
read_json()	fillna()	to_json()	unique()
read_sql()	groupby()	to_sql()	value_counts()
read_html()	apply()	to_html()	describe()
read_clipboard()	filter()	to_clipboard()	corr()
…	…	…	…

Your One Stop Data Shop

source: http://animezingartist.deviantart.com/art/Rudy-King-of-the-Pandas-Royal-312614172

pandas vs peewee (and SQL)

Databases

- Combines Data and Computation
- Optimized for data modeling and representation/persistence
- Can be more performant (and expressive) for simple queries
- Shared/collaborative

pandas
(and client libraries)

- Purely client based, offloads persistence to other technologies
- Bounded my memory (unless extra measures are taken)
- Can express more complex queries and computation (Turing complete)
- Optimized for single user experience

Systems Dichotomy

Databases	pandas (and client libraries)
Storage	**Computation**

Systems Dichotomy

Databases	pandas (and client libraries)
(mainly) Storage	**Computation**

Systems Dichotomy

Databases

Disk

pandas
(and client libraries)

CPU

Systems Dichotomy

Databases

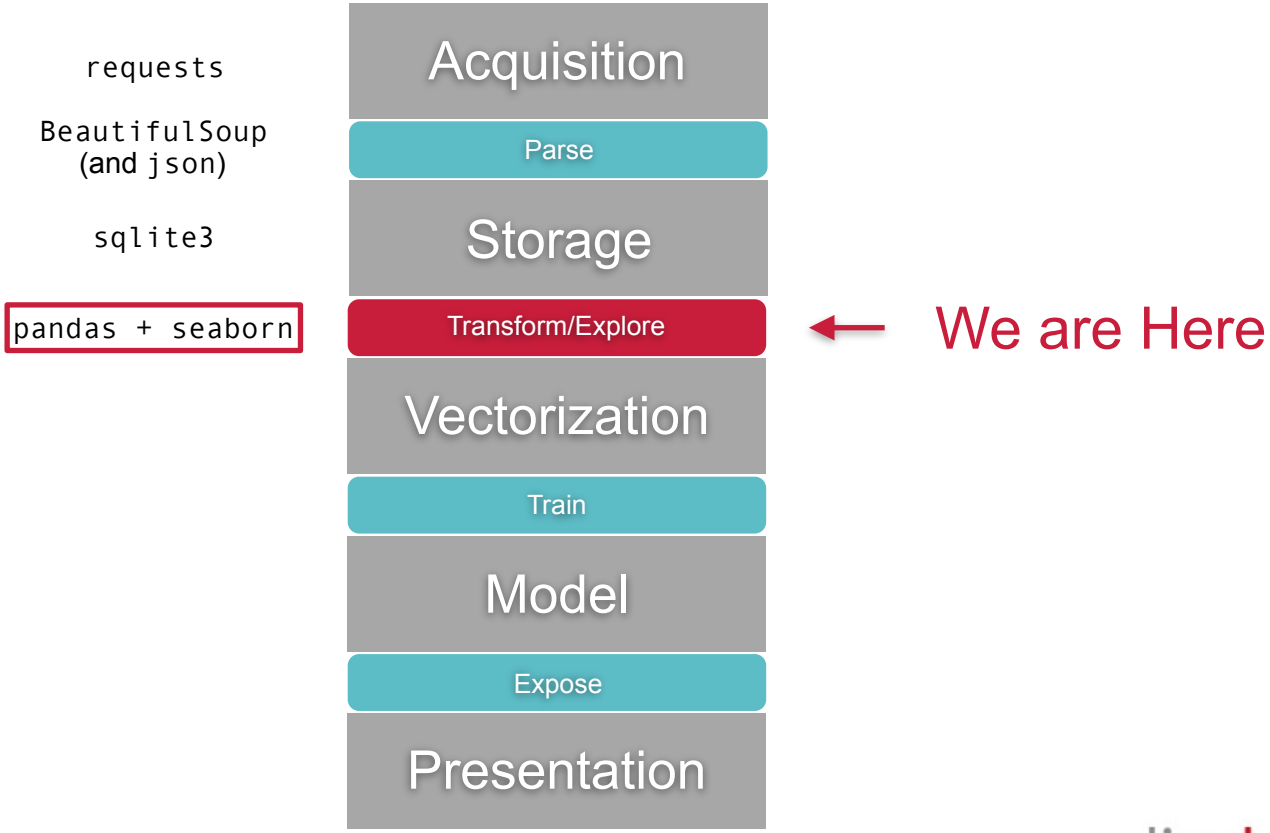
Space

pandas
(and client libraries)

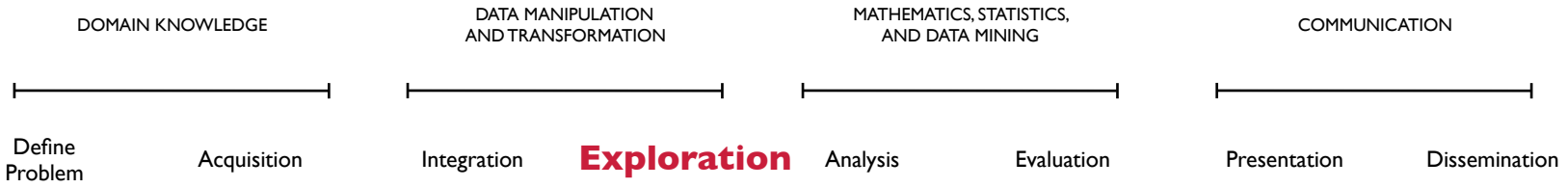
Time

http://pandas.pydata.org/pandas-docs/stable/comparison_with_sql.html

Process + Tools



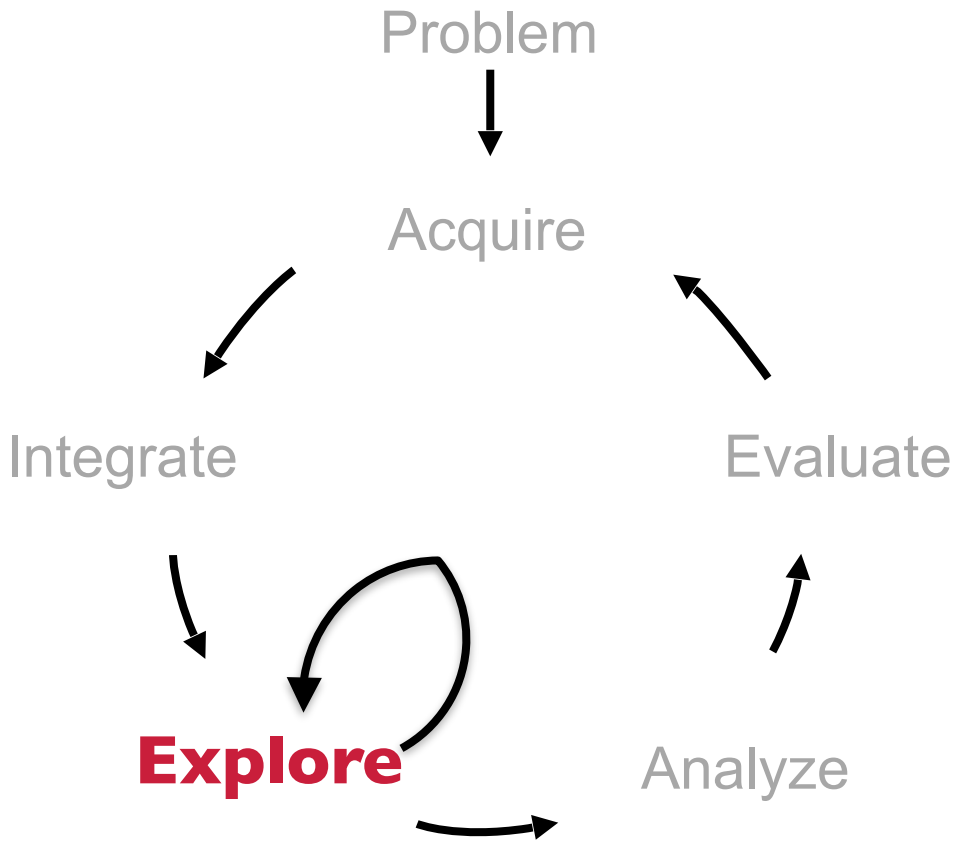
Process



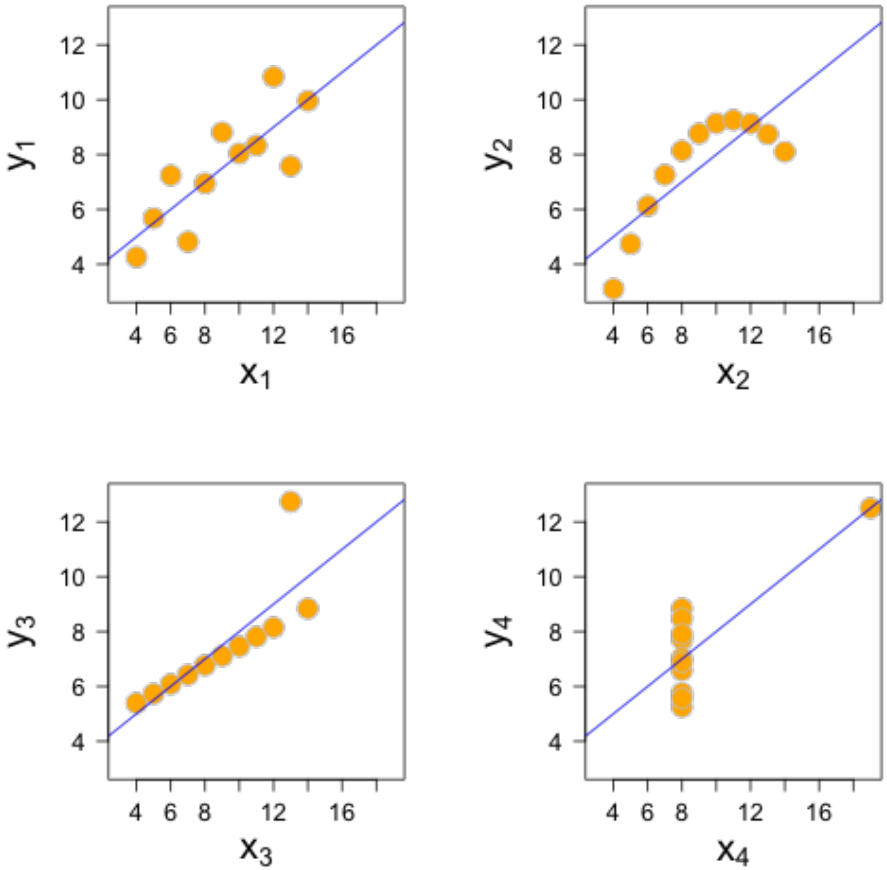
What was missing?

- Visual Inspection (visualization)
- Advanced Statistics/Computation
- (also iteration/recursion)

Process + Iteration



Anscombe's Quartet



Mean (x)	9
Sample Variance (x)	11
Mean (y)	7.50
Sample Variance (y)	4.127
Correlation	0.816
Linear Regression	y = 3.00 + 0.500x

A picture tells a 1000 words

A visualization tells a 1000 statistics

Common Questions

- How many records in total are there?
- How many missing (or null) values are there?
- How many unique values does each column contain?
- What are the most common values for each column?
- For numeric columns, what are the summary statistics?
- How are the values of each column distributed?

Harder Questions

- How are the records of one table related to another?
- Deduplication and Entity Resolution
- Are there outliers?
- What real world process does each column represent?
- How was the data collected? And what bias has this introduced?

Point Statistics versus Distributions

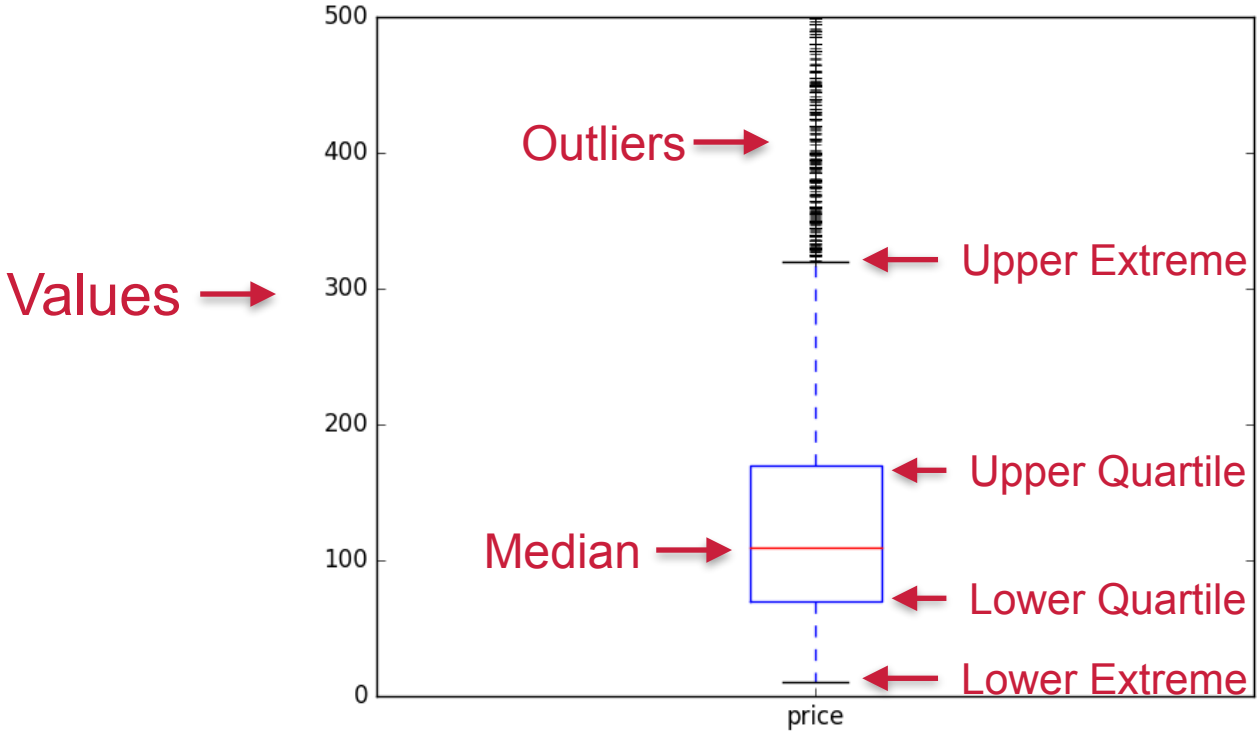
Point Statistics

- Easy to compare many at once
- Possible to automate comparison
- Can miss outliers

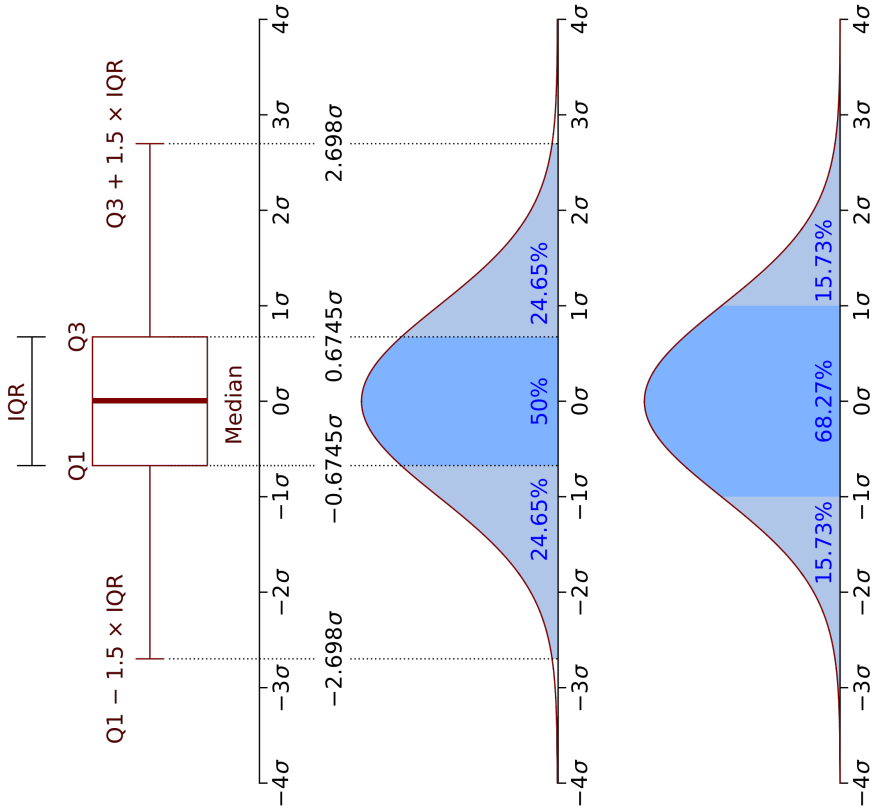
Distributions

- Requires visual/human inspection
- Can reveal complex patterns in data
- Gives a sense of uncertainty/ confidence
- Outliers become visible

The Box Plot



The Box Plot



Box Plot versus Violin Plot

Box Plot

- Easier to interpret (a visual representation of the five-number summary) and more familiar.
- Shows individual outlier data points
- Easier to compare a large number of box plots
- Easier to compare the spread (upper - lower quartile) of the data.

Violin Plot

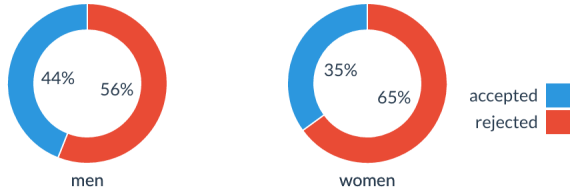
- Shows more information (distribution rather than point statistics).
- Will show bi-modality (or more complex distribution of values).
- Shows density of values/data.

Simpson's Paradox

Girls gone average.
Averages gone wild.

In 1973, the University of California-Berkeley was sued for sex discrimination. The numbers looked pretty incriminating: the graduate schools had just accepted 44% of male applicants but only 35% of female applicants. When researchers looked at the evidence, though, they uncovered something surprising:

> If the data are properly pooled...there is a small but statistically significant bias in favor of women.
>
> — (p. 403)



It was a textbook case of **Simpson's paradox**.

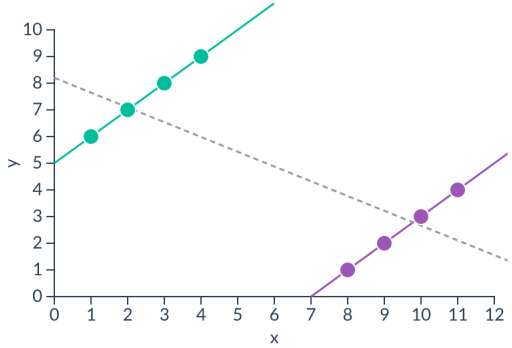
What is Simpson's paradox?

Every Simpson's paradox involves at least three variables:

1. the explained
2. the *observed* explanatory
3. the *lurking* explanatory

If the effect of the observed explanatory variable on the explained variable changes directions when you account for the lurking explanatory variable, you've got a Simpson's Paradox.

For example, to the right, *x* appears to have a negative effect on *y*, but the opposite is true when you account for color. *y* is the explained variable, *x* the observed explanatory variable, and color the lurking explanatory variable.



Proper Pooling

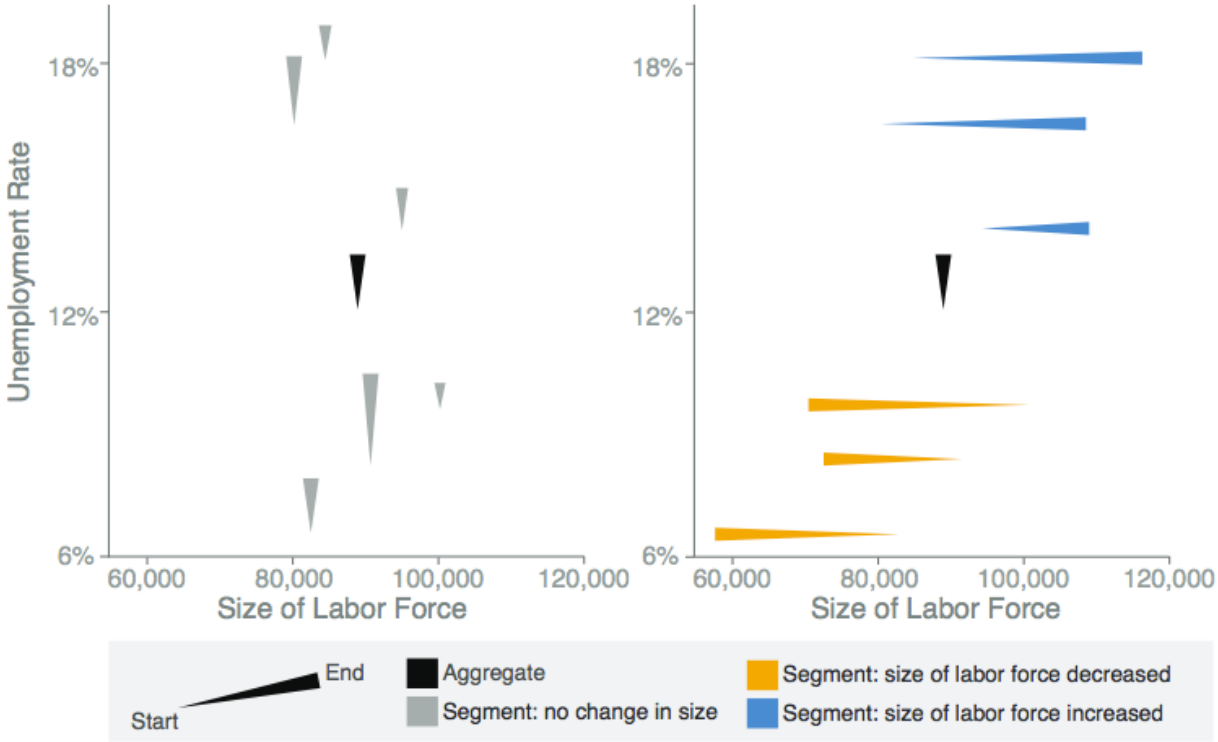
By "properly pooled," the investigators at Berkeley meant "broken down by department." Men more often applied to science departments, while women inclined towards humanities. Science departments require special technical skills but accept a large

Departments

A B C D E F combined

Visualizing Statistical Mix Effects and Simpson's Paradox

Zan Armstrong and Martin Wattenberg



source: Zan Armstrong, 2014 (http://research.google.com/pubs/pub42901.html)

Minimize TTU

Minimize Time To Understanding

Use the right tool for the job