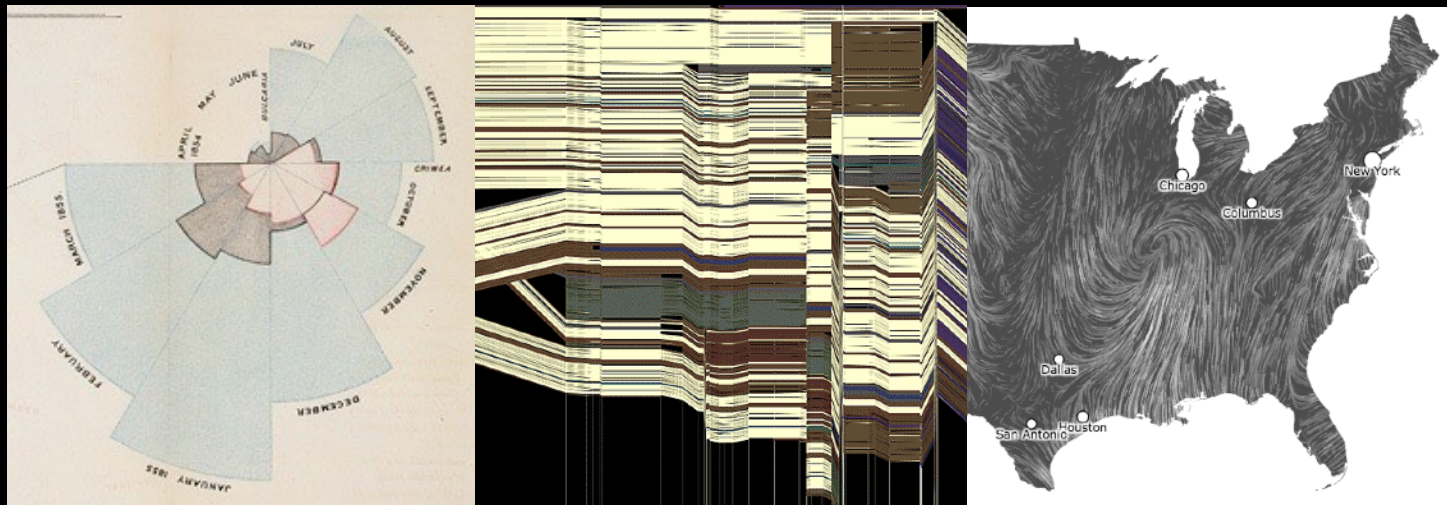


CSE 442 - Data Visualization

Data Transformation



Jeffrey Heer University of Washington

Session Outline

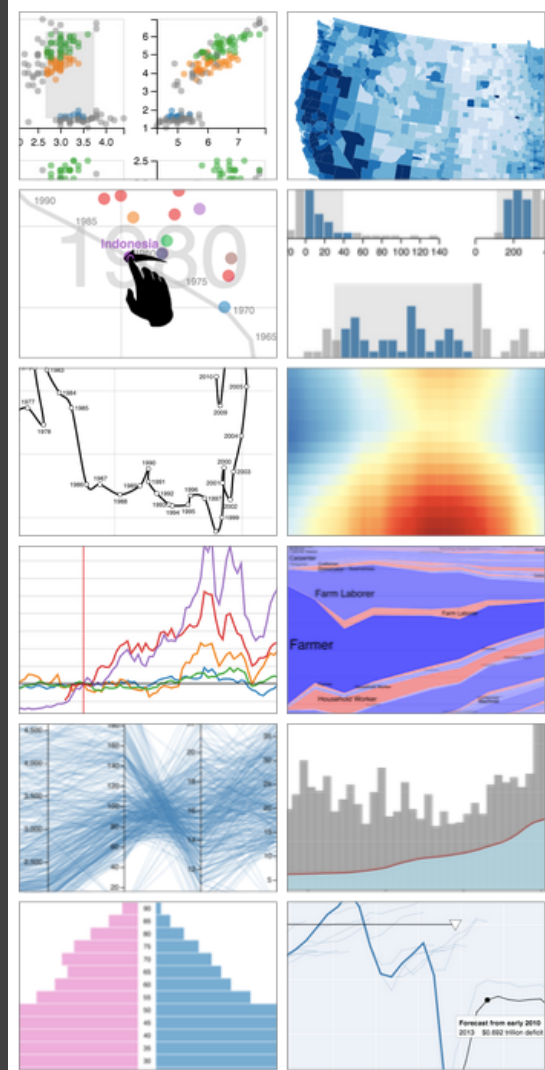
Data Models

Data Tables & Transformations

Data Wrangling & Profiling

Visualizing Distributions

Dimensionality Reduction



Data Models

Data Models / Conceptual Models

Data models are formal descriptions

Math: sets with operations on them

Example: integers with + and x operators

Conceptual models are mental constructions

Include semantics and support reasoning

Examples (data vs. conceptual)

1D floats vs. temperatures

3D vector of floats vs. spatial location

Types of Variables

Physical Types

Characterized by storage format

Characterized by machine operations

Example: bool, int32, float, double, string, ...

Abstract Types

Provide descriptions of the data

May be characterized by methods / attributes

May be organized into a hierarchy

Example: plants, animals, metazoans, ...

Taxonomy of Data Types (?)

1D (sets and sequences)

Temporal

2D (maps)

3D (shapes)

nD (relational)

Trees (hierarchies)

Networks (graphs)

Are there others?

The eyes have it: A task by data type
taxonomy for information visualization
[Shneiderman 96]

Nominal, Ordinal & Quantitative

Nominal, Ordinal & Quantitative

N - Nominal (labels or categories)

- Fruits: apples, oranges, ...

Nominal, Ordinal & Quantitative

N - Nominal (labels or categories)

- Fruits: apples, oranges, ...

O - Ordered

- Quality of meat: Grade A, AA, AAA

Nominal, Ordinal & Quantitative

N - Nominal (labels or categories)

- Fruits: apples, oranges, ...

O - Ordered

- Quality of meat: Grade A, AA, AAA

Q - Interval (location of zero arbitrary)

- Dates: Jan, 19, 2006; Location: (LAT 33.98, LONG -118.45)
- Only differences (i.e., intervals) may be compared

Nominal, Ordinal & Quantitative

N - Nominal (labels or categories)

- Fruits: apples, oranges, ...

O - Ordered

- Quality of meat: Grade A, AA, AAA

Q - Interval (location of zero arbitrary)

- Dates: Jan, 19, 2006; Location: (LAT 33.98, LONG -118.45)
- Only differences (i.e., intervals) may be compared

Q - Ratio (zero fixed)

- Physical measurement: Length, Mass, Time duration, ...
- Counts and amounts

Nominal, Ordinal & Quantitative

N - Nominal (labels or categories)

- Operations: =, $\neq$

O - Ordered

- Operations: =, $\neq$, $<$, $>$

Q - Interval (location of zero arbitrary)

- Operations: =, $\neq$, $<$, $>$, -
- Can measure distances or spans

Q - Ratio (zero fixed)

- Operations: =, $\neq$, $<$, $>$, -, %
- Can measure ratios or proportions

From Data Model to N, O, Q

Data Model

32.5, 54.0, -17.3, ...

Floating point numbers

Conceptual Model

Temperature (°C)

Data Type

Burned vs. Not-Burned (N)

Hot, Warm, Cold (O)

Temperature Value (Q-interval)

Dimensions & Measures

Dimensions (~ independent variables)

Often discrete variables describing data (N, O)

Categories, dates, binned quantities

Measures (~ dependent variables)

Data values that can be aggregated (Q)

Numbers to be analyzed

Aggregate as sum, count, avg, std. dev...

Not a strict distinction. The same variable may be treated either way depending on the task.

Example: U.S. Census Data

Example: U.S. Census Data

People Count: # of people in group

Year: 1850 – 2000 (every decade)

Age: 0 – 90+

Sex: Male, Female

Marital Status: Single, Married, Divorced, ...

Example: U.S. Census

People Count

Year

Age

Sex

Marital Status

2,348 data points

	A	B	C	D	E
1	year	age	marst	sex	people
2	1850	0	0	1	1483789
3	1850	0	0	2	1450376
4	1850	5	0	1	1411067
5	1850	5	0	2	1359668
6	1850	10	0	1	1260099
7	1850	10	0	2	1216114
8	1850	15	0	1	1077133
9	1850	15	0	2	1110619
10	1850	20	0	1	1017281
11	1850	20	0	2	1003841
12	1850	25	0	1	862547
13	1850	25	0	2	799482
14	1850	30	0	1	730638
15	1850	30	0	2	639636
16	1850	35	0	1	588487
17	1850	35	0	2	505012
18	1850	40	0	1	475911
19	1850	40	0	2	428185
20	1850	45	0	1	384211
21	1850	45	0	2	341254
22	1850	50	0	1	321343
23	1850	50	0	2	286580
24	1850	55	0	1	194080
25	1850	55	0	2	187208
26	1850	60	0	1	174976
27	1850	60	0	2	162236
28	1850	65	0	1	106827
29	1850	65	0	2	105534
30	1850	70	0	1	73677
31	1850	70	0	2	71762
32	1850	75	0	1	40834
33	1850	75	0	2	40229
34	1850	80	0	1	23449
35	1850	80	0	2	22949
36	1850	85	0	1	8186
37	1850	85	0	2	10511
38	1850	90	0	1	5259
39	1850	90	0	2	6569
40	1860	0	0	1	2120846
41	1860	0	0	2	2092162

Census: N, O, Q-Interval, Q-Ratio?

People Count

Q-Ratio

Year

Q-Interval (O)

Age

Q-Ratio (O)

Sex

N

Marital Status

N

Census: Dimension or Measure?

People Count

Measure

Year

Dimension

Age

Depends!

Sex

Dimension

Marital Status

Dimension

Census Data Demo

[demo link: us-population-1850-2000](#)

Data Tables & Transformations

Relational Data Model

Represent data as a **table** (or *relation*)

Each **row** (or *tuple*) represents a record

Each record is a fixed-length tuple

Each **column** (or *field*) represents a variable

Each field has a *name* and a *data type*

A table's **schema** is the set of names and types

A **database** is a collection of tables (relations)

Relational Algebra [Codd '70] / SQL

Operations on Data Tables: table(s) in, table out

Relational Algebra [Codd '70] / SQL

Operations on Data Tables: table(s) in, table out

Project (select): select a set of columns

Filter (where): remove unwanted rows

Sort (order by): order records

Aggregate (group by, sum, min, max, ...):

partition rows into groups + summarize

Combine (join, union, ...):

integrate data from multiple tables

Relational Algebra [Codd '70] / SQL

Project (select): select a set of columns
select day, stock

day	stock	price
10/3	AMZN	957.10
10/3	MSFT	74.26
10/4	AMZN	965.45
10/4	MSFT	74.69



day	stock
10/3	AMZN
10/3	MSFT
10/4	AMZN
10/4	MSFT

Relational Algebra [Codd '70] / SQL

Filter (where): remove unwanted rows

```
select * where price > 100
```

day	stock	price
10/3	AMZN	957.10
10/3	MSFT	74.26
10/4	AMZN	965.45
10/4	MSFT	74.69



day	stock	price
10/3	AMZN	957.10
10/4	AMZN	965.45

Relational Algebra [Codd '70] / SQL

Sort (order by): order records

`select * order by stock`

day	stock	price
10/3	AMZN	957.10
10/3	MSFT	74.26
10/4	AMZN	965.45
10/4	MSFT	74.69



day	stock	price
10/3	AMZN	957.10
10/4	AMZN	965.45
10/3	MSFT	74.26
10/4	MSFT	74.69

Relational Algebra [Codd '70] / SQL

Aggregate (group by, sum, min, max, ...):

```
select stock, min(price) group by stock
```

day	stock	price
10/3	AMZN	957.10
10/3	MSFT	74.26
10/4	AMZN	965.45
10/4	MSFT	74.69



stock	min(price)
AMZN	957.10
MSFT	74.26

Relational Algebra [Codd '70] / SQL

Join (join) multiple tables together

day	stock	price
10/3	AMZN	957.10
10/3	MSFT	74.26
10/4	AMZN	965.45
10/4	MSFT	74.69



day	stock	price	min
10/3	AMZN	957.10	957.10
10/3	MSFT	74.26	74.26
10/4	AMZN	965.45	957.10
10/4	MSFT	74.69	74.26

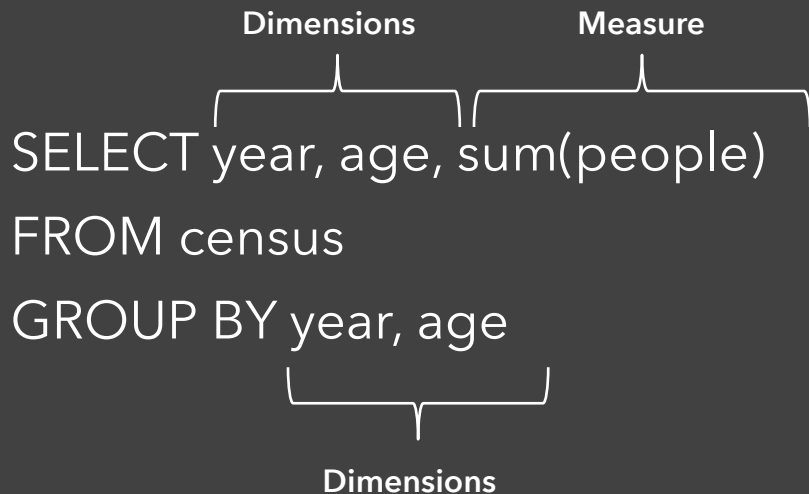
stock	min
AMZN	957.10
MSFT	74.26

```
select t.day, t.stock, t.price, a.min
from table as t, aggregate as a
where t.stock = a.stock
```

Roll-Up and Drill-Down

Want to examine population by year and age?

Roll-up the data along the desired dimensions



The diagram shows a SQL query with two annotations. The first annotation, labeled 'Dimensions', is a bracket above 'year, age' in the 'GROUP BY' clause. The second annotation, labeled 'Measure', is a bracket above 'sum(people)' in the 'SELECT' clause.

```
SELECT year, age, sum(people)
FROM census
GROUP BY year, age
```

Dimensions

Measure

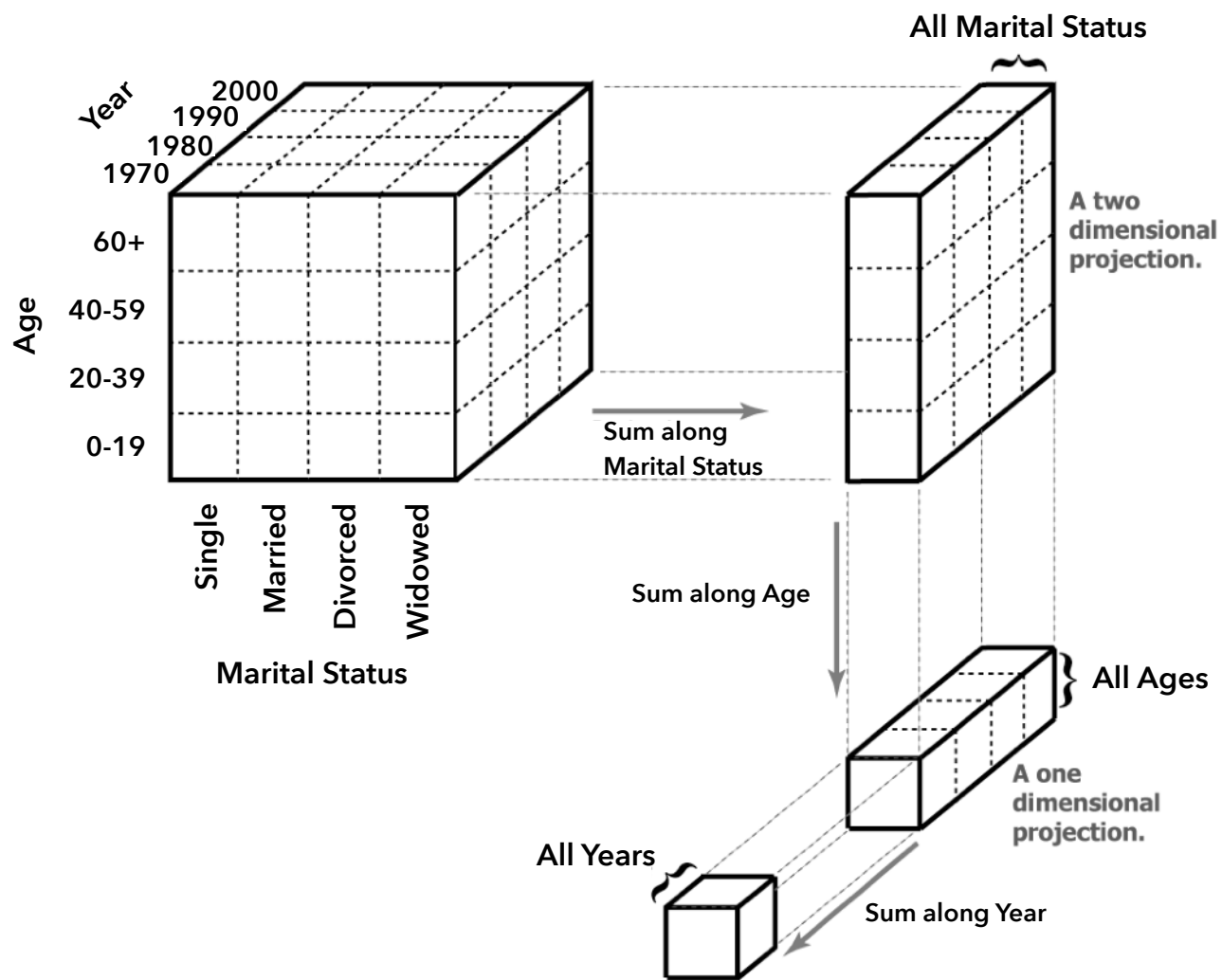
Dimensions

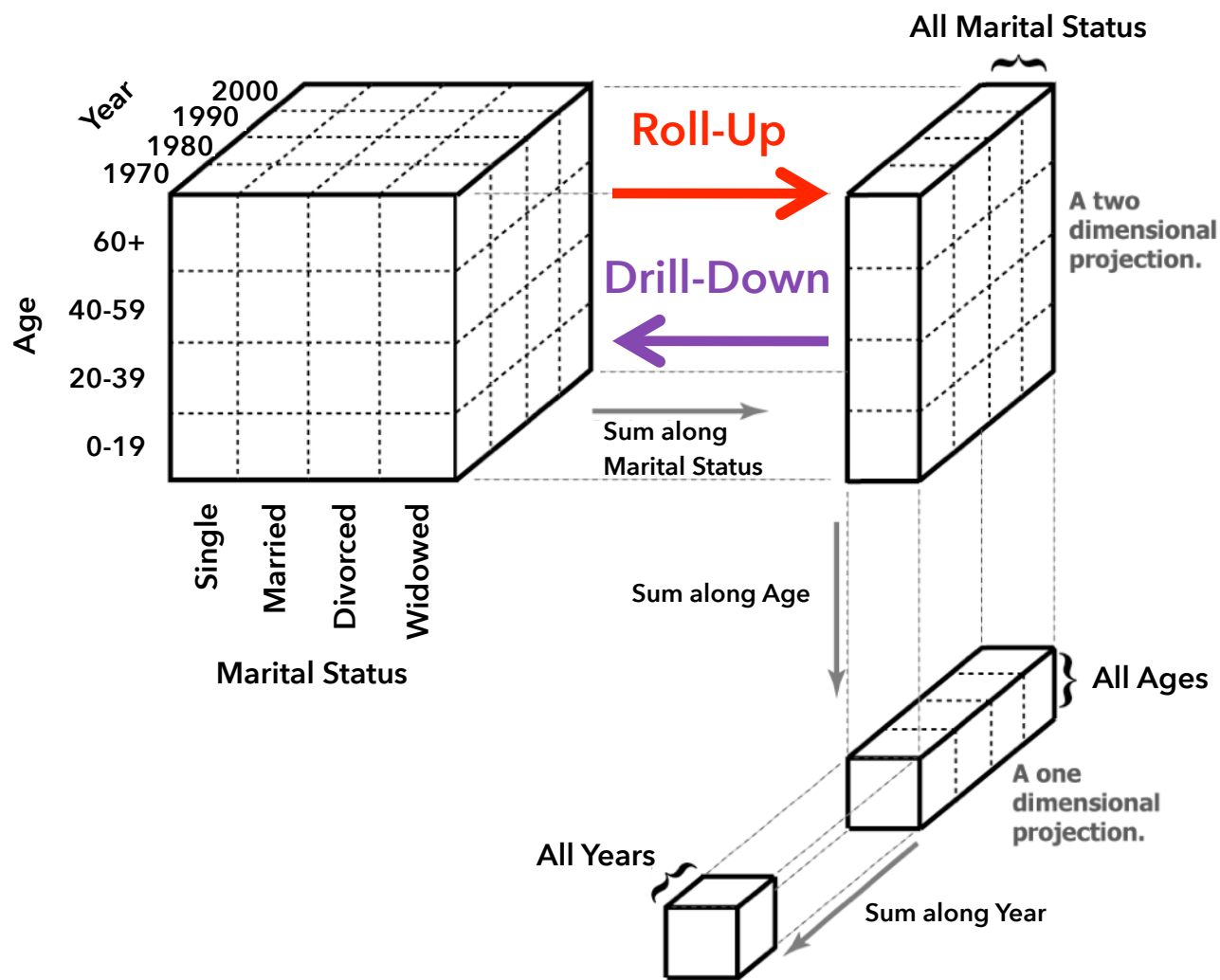
Roll-Up and Drill-Down

Want to see the breakdown by marital status?

Drill-down into additional dimensions

```
SELECT year, age, marst, sum(people)
FROM census
GROUP BY year, age, marst
```





ORIGINAL

YEAR	AGE	MARST	SEX	PEOPLE
1850	0	0	1	1,483,789
1850	5	0	1	1,411,067
1860	0	0	1	2,120,846
1860	5	0	1	1,804,467

...

PIVOTED (or CROSS-TABULATION)

AGE	MARST	SEX	1850	1860	...
0	0	1	1,483,789	2,120,846	...
5	0	1	1,411,067	1,804,467	...

...

Which format might we prefer? Why?

Tidy Data [Wickham 2014]

How do rows, columns, and tables match up with observations, variables, and types? In “tidy” data:

1. Each variable forms a column.
2. Each observation forms a row.
3. Each type of observational unit forms a table.

The advantage is that this provides a flexible starting point for analysis, transformation, and visualization.

Our pivoted table variant was not “tidy”!

(This is a variant of normalized forms in DB theory)

Common Data Formats

CSV: Comma-Separated Values (d3.csv)

```
year,age,marst,sex,people  
1850,0,0,1,1483789  
1850,5,0,1,1411067  
...
```


Common Data Formats

CSV: Comma-Separated Values (d3.csv)

```
year,age,marst,sex,people  
1850,0,0,1,1483789  
1850,5,0,1,1411067  
...
```

JSON: JavaScript Object Notation (d3.json)

```
[  
  {"year":1850,"age":0,"marst":0,"sex":1,"people":1483789},  
  {"year":1850,"age":5,"marst":0,"sex":1,"people":1411067},  
  ...  
]
```

Common Data Formats

CSV: Comma-Separated Values (d3.csv)

```
year,age,marst,sex,people  
1850,0,0,1,1483789  
1850,5,0,1,1411067  
...
```

JSON: JavaScript Object Notation (d3.json)

```
[  
  {"year":1850,"age":0,"marst":0,"sex":1,"people":1483789},  
  {"year":1850,"age":5,"marst":0,"sex":1,"people":1411067},  
  ...  
]
```

Binary Formats: Arrow, Parquet, ...

Data Wrangling

I spend more than half of my time integrating, cleansing and transforming data without doing any actual analysis. Most of the time I'm lucky if I get to do any "analysis" at all.

Anonymous Data Scientist
from our 2012 interview study





**Big Data
Borat**

@BigDataBorat



Following

In Data Science, 80% of time spent prepare data, 20% of time spent complain about need for prepare data.



Reported crime in Alabama

Year	Population	Property crime rate	Burglary rate	Larceny-theft rate	Motor vehicle theft rate
2004	4525375 4029.3	987 2732.4 309.9			
2005	4548327 3900	955.8 2656 289			
2006	4599030 3937	968.9 2645.1 322.9			
2007	4627851 3974.9	980.2 2687 307.7			
2008	4661900 4081.9	1080.7 2712.6 288.6			

Reported crime in Alaska

Year	Population	Property crime rate	Burglary rate	Larceny-theft rate	Motor vehicle theft rate
2004	657755 3370.9	573.6 2456.7 340.6			
2005	663253 3615	622.8 2601 391			
2006	670053 3582	615.2 2588.5 378.3			
2007	683478 3373.9	538.9 2480 355.1			
2008	686293 2928.3	470.9 2219.9 237.5			

Reported crime in Arizona

Year	Population	Property crime rate	Burglary rate	Larceny-theft rate	Motor vehicle theft rate
2004	5739879 5073.3	991 3118.7 963.5			
2005	5953007 4827	946.2 2958 922			
2006	6166318 4741.6	953 2874.1 914.4			
2007	6338755 4502.6	935.4 2780.5 786.7			
2008	6500180 4087.3	894.2 2605.3 587.8			

Reported crime in Arkansas

Year	Population	Property crime rate	Burglary rate	Larceny-theft rate	Motor vehicle theft rate
2004	2750000 4033.1	1096.4 2699.7 237			
2005	2775708 4068	1085.1 2720 262			
2006	2810872 4021.6	1154.4 2596.7 270.4			
2007	2834797 3945.5	1124.4 2574.6 246.5			
2008	2855390 3843.7	1182.7 2433.4 227.6			

DataWrangler

The screenshot displays the DataWrangler interface. On the left, a 'Suggestions' panel lists four actions: 'Delete rows 8,10', 'Delete empty rows', 'Delete rows where Property_crime_rate is null', and 'Delete rows where Year is null'. Below this is a 'Script' panel with two suggestions: 'Split data repeatedly on newline into rows' and 'Split data repeatedly on \''. An 'Export' button is located to the right of the script panel. The main area on the right shows a data table with 408 rows. The table has two columns: '# Year' and '# Property_crime_rate'. The data is organized into sections: 'Reported crime in Alabama' (rows 1-7), an empty row (8), 'Reported crime in Alaska' (row 9), and another empty row (10). Rows 11-14 show data for the years 2004, 2005, 2006, and 2007 respectively, with corresponding crime rates.

#	Year	#	Property_crime_rate
1	Reported crime in Alabama		
2			
3	2004		4029.3
4	2005		3900
5	2006		3937
6	2007		3974.9
7	2008		4081.9
8			
9	Reported crime in Alaska		
10			
11	2004		3370.9
12	2005		3615
13	2006		3582
14	2007		3373.9

Wrangler: Interactive Visual Specification of Data Transformation Scripts

Kandel et al. [CHI 2011]

▶	Split	Cut	Extract	Edit	Fill	Translate	Drop	Merge	Wrap	Delete	Promote	Fold	Pivot	Transpose	
Transform Suggestions															
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20		#	split	#	split1	#	split2	#	split3	#	split4				
		Reported crime in Alabama													
		Year	Population			Property crime rate		Burglary rate		Larceny-theft					
		2004	4525375			4029.3		987		2732.4					
		2005	4548327			3900		955.8		2656					
		2006	4599030			3937		968.9		2645.1					
		2007	4627851			3974.9		980.2		2687					
		2008	4661900			4081.9		1080.7		2712.6					
		Reported crime in Alaska													
		Year	Population			Property crime rate		Burglary rate		Larceny-theft					
		2004	657755			3370.9		573.6		2456.7					
		2005	663253			3615		622.8		2601					
		2006	670053			3582		615.2		2588.5					
		2007	683478			3373.9		538.9		2480					
		2008	686293			2928.3		470.9		2219.9					
		Reported crime in Arizona													
ROWS: 458															

Transform Script

Export

▶ Split data repeatedly on **newline** into rows

▶ Split data repeatedly on **'tab'**

2

•

3

7

9

10

Export

16

18

ROWS: 458

Transform Suggestions												
Delete row 2		#	split	#	split1	#	split2	#	split3	#	split4	
	1	Reported crime in Alabama										
Delete empty rows	+	2										
	3	Year			Population		Property crime rate		Burglary rate		Larceny-theft	
Delete rows where split is null	4	2004			4525375		4029.3		987		2732.4	
	5	2005			4548327		3900		955.8		2656	
Delete rows where split1 is null	6	2006			4599030		3937		968.9		2645.1	
	7	2007			4627851		3974.9		980.2		2687	
Delete rows where split2 is null	8	2008			4661900		4081.9		1080.7		2712.6	
	9											
Delete rows where split3 is null	10	Reported crime in Alaska										
	11											
Fold using 2 as a key	12	Year			Population		Property crime rate		Burglary rate		Larceny-theft	
	13	2004			657755		3370.9		573.6		2456.7	
	14	2005			663253		3615		622.8		2601	
Transform Script	Export	15	2006			670053		3582		615.2		2588.5
► Split data repeatedly on newline into rows		16	2007			683478		3373.9		538.9		2480
		17	2008			686293		2928.3		470.9		2219.9
		18										
► Split data repeatedly on 'tab'		19	Reported crime in Arizona									
		20										
ROWS: 458												

Transform Script Export

► Split Cut Extract Edit Fill Translate Drop Merge Wrap Delete Promote Fold Pivot Transpose

Transform Suggestions

	#	split	#	split1	#	split2	#	split3	#	sp
1		Reported crime in Alabama								
2		Year		Population		Property crime rate		Burglary rate		Larceny-theft
3		2004		4525375		4029.3		987		2732.4
4		2005		4548327		3900		955.8		2656
5		2006		4599030		3937		968.9		2645.1
6		2007		4627851		3974.9		980.2		2687
7		2008		4661900		4081.9		1080.7		2712.6
8		Reported crime in Alaska								
9		Year		Population		Property crime rate		Burglary rate		Larceny-theft
10		2004		657755		3370.9		573.6		2456.7
11		2005		663253		3615		622.8		2601
12		2006		670053		3582		615.2		2588.5
13		2007		683478		3373.9		538.9		2480
14		2008		686293		2928.3		470.9		2219.9
15		Reported crime in Arizona								
16		Year		Population		Property crime rate		Burglary rate		Larceny-theft
17		2004		5739879		5073.3		991		3118.7
18		2005		5953007		4827		946.2		2958
19		2006		6166318		4741.6		953		2874.1
20		2007		6338755		4502.6		935.4		2780.5

ROWS: 357

Transform Script

Export

► Split data repeatedly on newline into rows

► Split data repeatedly on 'tab'

► Delete empty rows

Transform Suggestions													
		#	Year	#	Population	#	Property_crime_rate	#	Burglary_rate	#	Larceny-		
	1	Reported crime in Alabama											
	2	2004		4525375		4029.3		987		2732.4			
	3	2005		4548327		3900		955.8		2656			
	4	2006		4599030		3937		968.9		2645.1			
	5	2007		4627851		3974.9		980.2		2687			
	6	2008		4661900		4081.9		1080.7		2712.6			
	7	Reported crime in Alaska											
	8	Year		Population		Property crime rate		Burglary rate		Larceny-theft			
	9	2004		657755		3370.9		573.6		2456.7			
	10	2005		663253		3615		622.8		2601			
	11	2006		670053		3582		615.2		2588.5			
	12	2007		683478		3373.9		538.9		2480			
	13	2008		686293		2928.3		470.9		2219.9			
	14	Reported crime in Arizona											
Transform Script		Export		15	Year		Population		Property crime rate		Burglary rate		Larceny-theft
► Split data repeatedly on newline into rows		16	2004		5739879		5073.3		991		3118.7		
		17	2005		5953007		4827		946.2		2958		
► Split data repeatedly on 'tab'		18	2006		6166318		4741.6		953		2874.1		
		19	2007		6338755		4502.6		935.4		2780.5		
► Delete empty rows		20	2008		6500180		4087.3		894.2		2605.3		
		ROWS: 356											
► Promote row 2 to header													

Transform Suggestions												
Delete row 8			#	Year	#	Population	#	Property_crime_rate	#	Burglary_rate	#	Larceny-
		1	Reported crime in Alabama									
Delete rows where Year = 'Year'		2	2004		4525375		4029.3		987		2732.4	
		3	2005		4548327		3900		955.8		2656	
Delete rows where Population = 'Population'		4	2006		4599030		3937		968.9		2645.1	
		5	2007		4627851		3974.9		980.2		2687	
Delete rows where Property_crime_rate = 'Property crime ...'		6	2008		4661900		4081.9		1080.7		2712.6	
		7	Reported crime in Alaska									
Delete rows where Burglary_rate = 'Burglary rate'		8	Year		Population		Property crime rate		Burglary rate		Larceny-theft	
		9	2004		657755		3370.9		573.6		2456.7	
		10	2005		663253		3615		622.8		2601	
Delete rows where Larceny-theft_rate = 'Larceny-theft ra...		11	2006		670053		3582		615.2		2588.5	
		12	2007		683478		3373.9		538.9		2480	
Fill row 8 with values from the left		13	2008		686293		2928.3		470.9		2219.9	
		14	Reported crime in Arizona									
Transform Script		15	Year		Population		Property crime rate		Burglary rate		Larceny-theft	
▶ Split data repeatedly on newline into rows		16	2004		5739879		5073.3		991		3118.7	
		17	2005		5953007		4827		946.2		2958	
▶ Split data repeatedly on 'tab'		18	2006		6166318		4741.6		953		2874.1	
		19	2007		6338755		4502.6		935.4		2780.5	
▶ Delete empty rows		20	2008		6500180		4087.3		894.2		2605.3	
▶ Promote row 2 to header		ROWS: 356										

Transform Suggestions											
Delete row 8		#	Year	#	Population	#	Property_crime_rate	#	Burglary_rate	#	Larceny-
	1	Reported crime in Alabama									
Delete rows where Year = 'Year'	+	2	2004		4525375		4029.3		987		2732.4
	3	2005			4548327		3900		955.8		2656
Delete rows where Population = 'Population'		4	2006			4599030		3937		968.9	2645.1
	5	2007			4627851		3974.9		980.2		2687
Delete rows where Property_crime_rate = 'Property crime ...'		6	2008			4661900		4081.9		1080.7	2712.6
	7	Reported crime in Alaska									
Delete rows where Burglary_rate = 'Burglary rate'		8	Year		Population		Property crime rate		Burglary rate		Larceny-theft
	9	2004			657755		3370.9		573.6		2456.7
	10	2005			663253		3615		622.8		2601
Delete rows where Larceny-theft_rate = 'Larceny-theft ra...		11	2006			670053		3582		615.2	2588.5
	12	2007			683478		3373.9		538.9		2480
Fill row 8 with values from the left		13	2008			686293		2928.3		470.9	2219.9
	14	Reported crime in Arizona									
Transform Script	Export	15	Year		Population		Property crime rate		Burglary rate		Larceny-theft
► Split data repeatedly on newline into rows		16	2004			5739879		5073.3		991	3118.7
		17	2005			5953007		4827		946.2	2958
► Split data repeatedly on 'tab'		18	2006			6166318		4741.6		953	2874.1
		19	2007			6338755		4502.6		935.4	2780.5
► Delete empty rows		20	2008			6500180		4087.3		894.2	2605.3
► Promote row 2 to header		ROWS: 356									

Transform Suggestions																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																												
-----------------------	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--

ROWS: 306

Transform Suggestions

	#	Year	Abc	extract	#	Population	#	Property_crime_rate	#	Burgla
1		Reported crime in Alabama	Alabama							
2	2004				4525375		4029.3		987	
3	2005				4548327		3900		955.8	
4	2006				4599030		3937		968.9	
5	2007				4627851		3974.9		980.2	
6	2008				4661900		4081.9		1080.7	
7		Reported crime in Alaska	Alaska							
8	2004				657755		3370.9		573.6	
9	2005				663253		3615		622.8	
10	2006				670053		3582		615.2	
11	2007				683478		3373.9		538.9	
12	2008				686293		2928.3		470.9	
13		Reported crime in Arizona	Arizona							
14	2004				5739879		5073.3		991	
15	2005				5953007		4827		946.2	
16	2006				6166318		4741.6		953	
17	2007				6338755		4502.6		935.4	
18	2008				6500180		4087.3		894.2	
19		Reported crime in Arkansas	Arkansas							
20	2004				2750000		4033.1		1096.4	

ROWS: 306

Transform Script

Export

Split data repeatedly on newline into rows

Split data repeatedly on 'tab'

Delete empty rows

Promote row 2 to header

Delete rows where Year = 'Year'

Extract from Year after 'in '

Transform Suggestions

	#	Year	Abc	extract	#	Population	#	Property_crime_rate	#	Burgla
1		Reported crime in Alabama	Alabama							
2	2004				4525375		4029.3		987	
3	2005				4548327		3900		955.8	
4	2006				4599030		3937		968.9	
5	2007				4627851		3974.9		980.2	
6	2008				4661900		4081.9		1080.7	
7		Reported crime in Alaska	Alaska							
8	2004				657755		3370.9		573.6	
9	2005				663253		3615		622.8	
10	2006				670053		3582		615.2	
11	2007				683478		3373.9		538.9	
12	2008				686293		2928.3		470.9	
13		Reported crime in Arizona	Arizona							
14	2004				5739879		5073.3		991	
15	2005				5953007		4827		946.2	
16	2006				6166318		4741.6		953	
17	2007				6338755		4502.6		935.4	
18	2008				6500180		4087.3		894.2	
19		Reported crime in Arkansas	Arkansas							
20	2004				2750000		4033.1		1096.4	

ROWS: 306

Transform Script

Export

Split data repeatedly on newline into rows

Split data repeatedly on 'tab'

Delete empty rows

Promote row 2 to header

Delete rows where Year = 'Year'

Extract from Year after 'in '

Transform Suggestions													
		#	Year	Abc	extract	#	Population	#	Property_crime_rate	#	Burgla		
	1	Reported crime in Alabama		Alabama									
	2	2004				4525375		4029.3		987			
	3	2005				4548327		3900		955.8			
	4	2006				4599030		3937		968.9			
	5	2007				4627851		3974.9		980.2			
	6	2008				4661900		4081.9		1080.7			
	7	Reported crime in Alaska		Alaska									
	8	2004				657755		3370.9		573.6			
	9	2005				663253		3615		622.8			
	10	2006				670053		3582		615.2			
	11	2007				683478		3373.9		538.9			
	12	2008				686293		2928.3		470.9			
	13	Reported crime in Arizona		Arizona									
Transform Script		Export		14	2004				5739879		5073.3		991
► Split data repeatedly on newline into rows		15	2005				5953007		4827		946.2		
		16	2006				6166318		4741.6		953		
		17	2007				6338755		4502.6		935.4		
► Split data repeatedly on 'tab'		18	2008				6500180		4087.3		894.2		
		19	Reported crime in Arkansas		Arkansas								
► Delete empty rows		20	2004				2750000		4033.1		1096.4		
► Promote row 2 to header		ROWS: 306											

Transform Script Export

- Split data repeatedly on newline into rows
- Split data repeatedly on 'tab'
- Delete empty rows
- Promote row 2 to header
- Delete rows where Year = 'Year'
- Extract from Year after 'in '

Split Cut Extract Edit Fill

Translate Drop Merge Wrap Delete Promote Fold Pivot Transpose

Transform Suggestions

Fill **extract** with values from **above**

Fill **extract** with values from **below**

Drop **extract**

Fold **extract** using **header** as a key

Fold **extract** using **1** as a key

Fold **extract** using **1, 2** as keys

Fold **extract** using **1, 2, 3** as keys

Transform Script

Export

- Split **data repeatedly** on **newline** into **rows**

- Split **data repeatedly** on 'tab'

- ▶ Delete **empty rows**

- ▶ Promote row 2 to header

- ▶ Delete rows where Year = 'Year'

► Extract from **Year** after 'in '

[illegible]

Transform Suggestions

	#	Year	Abc	extract	#	Population	#	Property_crime_rate	#	Burgla
1		Reported crime in Alabama	Alabama							
2	2004		Alabama		4525375		4029.3		987	
3	2005		Alabama		4548327		3900		955.8	
4	2006		Alabama		4599030		3937		968.9	
5	2007		Alabama		4627851		3974.9		980.2	
6	2008		Alabama		4661900		4081.9		1080.7	
7		Reported crime in Alaska	Alaska							
8	2004		Alaska		657755		3370.9		573.6	
9	2005		Alaska		663253		3615		622.8	
10	2006		Alaska		670053		3582		615.2	
11	2007		Alaska		683478		3373.9		538.9	
12	2008		Alaska		686293		2928.3		470.9	
13		Reported crime in Arizona	Arizona							
14	2004		Arizona		5739879		5073.3		991	
15	2005		Arizona		5953007		4827		946.2	
16	2006		Arizona		6166318		4741.6		953	
17	2007		Arizona		6338755		4502.6		935.4	
18	2008		Arizona		6500180		4087.3		894.2	
19		Reported crime in Arkansas	Arkansas							
20	2004		Arkansas		2750000		4033.1		1096.4	

ROWS: 306

Transform Script

Export

► Split data repeatedly on **newline** into **rows**

► Split data repeatedly on **'tab'**

► Delete **empty rows**

► Promote row **2** to header

► Delete **rows where Year = 'Year'**

► Extract from **Year** after **'in '**

► Fill **extract** with values from **above**

Transform Suggestions

Extract from **Year** between positions 0, 17



Extract from **Year** on 'Reported crime in'

Extract from **Year** on 'Reported crime any word '

Extract from **Year** on 'Reported crime any lowercase word '

Extract from **Year** on 'Reported any word in'

Extract from **Year** on 'Reported any word any word '

Cut from **Year** between positions 0, 17

Transform Script

Export

▶ Split **data repeatedly** on **newline** into **rows**

▶ Split **data repeatedly** on 'tab'

▶ Delete **empty rows**

▶ Promote row 2 to header

▶ Delete **rows where Year = 'Year'**

▶ Extract from **Year** after 'in '

▶ Fill **extract** with values from **above**

#	Year	Abc extract1	Abc extract	# Population	# Property_
1	Reported crime in Alabama	Reported crime in	Alabama		
2	2004		Alabama	4525375	4029.3
3	2005		Alabama	4548327	3900
4	2006		Alabama	4599030	3937
5	2007		Alabama	4627851	3974.9
6	2008		Alabama	4661900	4081.9
7	Reported crime in Alaska	Reported crime in	Alaska		
8	2004		Alaska	657755	3370.9
9	2005		Alaska	663253	3615
10	2006		Alaska	670053	3582
11	2007		Alaska	683478	3373.9
12	2008		Alaska	686293	2928.3
13	Reported crime in Arizona	Reported crime in	Arizona		
14	2004		Arizona	5739879	5073.3
15	2005		Arizona	5953007	4827
16	2006		Arizona	6166318	4741.6
17	2007		Arizona	6338755	4502.6
18	2008		Arizona	6500180	4087.3
19	Reported crime in Arkansas	Reported crime in	Arkansas		
20	2004		Arkansas	2750000	4033.1
ROWS: 306					

Export

	#	Year	Abc	extract	#	Population	#	Property_crime_rate	#	Burgla
1		Reported crime in Alabama	Alabama							
2		2004	Alabama		4525375		4029.3		987	
3		2005	Alabama		4548327		3900		955.8	
4		2006	Alabama		4599030		3937		968.9	
5		2007	Alabama		4627851		3974.9		980.2	
6		2008	Alabama		4661900		4081.9		1080.7	
7		Reported crime in Alaska	Alaska							
8		2004	Alaska		657755		3370.9		573.6	
9		2005	Alaska		663253		3615		622.8	
10		2006	Alaska		670053		3582		615.2	
11		2007	Alaska		683478		3373.9		538.9	
12		2008	Alaska		686293		2928.3		470.9	
13		Reported crime in Arizona	Arizona							
14		2004	Arizona		5739879		5073.3		991	
15		2005	Arizona		5953007		4827		946.2	
16		2006	Arizona		6166318		4741.6		953	
17		2007	Arizona		6338755		4502.6		935.4	
18		2008	Arizona		6500180		4087.3		894.2	
19		Reported crime in Arkansas	Arkansas							
20		2004	Arkansas		2750000		4033.1		1096.4	

ROWS: 306

Transform Suggestions

	#	Year	Abc	extract	#	Population	#	Property_crime_rate	#	Burgla
1	2004		Alabama		4525375		4029.3		987	
2	2005		Alabama		4548327		3900		955.8	
3	2006		Alabama		4599030		3937		968.9	
4	2007		Alabama		4627851		3974.9		980.2	
5	2008		Alabama		4661900		4081.9		1080.7	
6	2004		Alaska		657755		3370.9		573.6	
7	2005		Alaska		663253		3615		622.8	
8	2006		Alaska		670053		3582		615.2	
9	2007		Alaska		683478		3373.9		538.9	
10	2008		Alaska		686293		2928.3		470.9	
11	2004		Arizona		5739879		5073.3		991	
12	2005		Arizona		5953007		4827		946.2	
13	2006		Arizona		6166318		4741.6		953	
14	2007		Arizona		6338755		4502.6		935.4	
15	2008		Arizona		6500180		4087.3		894.2	
16	2004		Arkansas		2750000		4033.1		1096.4	
17	2005		Arkansas		2775708		4068		1085.1	
18	2006		Arkansas		2810872		4021.6		1154.4	
19	2007		Arkansas		2834797		3945.5		1124.4	
20	2008		Arkansas		2855390		3843.7		1182.7	

ROWS: 255

Transform Script

Export

► Split data repeatedly on **newline** into **rows**

► Split data repeatedly on **'tab'**

► Delete **empty rows**

► Promote row **2** to header

► Delete **rows where Year = 'Year'**

► Extract from **Year** after **'in '**

► Fill **extract** with values from **above**

► Delete **rows where Year starts with 'Reported crime in'**

Transform Suggestions

- ☒ Data
- ☐ Script

Comma-Separated Values (CSV)

Back to Wrangling

```
Year,extract,Population,Property_crime_rate,Burglary_rate,Larceny-
theft_rate,Motor_vehicle_theft_rate
2004,Alabama,4525375,4029.3,987,2732.4,309.9
2005,Alabama,4548327,3900,955.8,2656,289
2006,Alabama,4599030,3937,968.9,2645.1,322.9
2007,Alabama,4627851,3974.9,980.2,2687,307.7
2008,Alabama,4661900,4081.9,1080.7,2712.6,288.6
2004,Alaska,657755,3370.9,573.6,2456.7,340.6
2005,Alaska,663253,3615,622.8,2601,391
2006,Alaska,670053,3582,615.2,2588.5,378.3
2007,Alaska,683478,3373.9,538.9,2480,355.1
2008,Alaska,686293,2928.3,470.9,2219.9,237.5
2004,Arizona,5739879,5073.3,991,3118.7,963.5
2005,Arizona,5953007,4827,946.2,2958,922
2006,Arizona,6166318,4741.6,953,2874.1,914.4
2007,Arizona,6338755,4502.6,935.4,2780.5,786.7
2008,Arizona,6500180,4087.3,894.2,2605.3,587.8
2004,Arkansas,2750000,4033.1,1096.4,2699.7,237
2005,Arkansas,2775708,4068,1085.1,2720,262
2006,Arkansas,2810872,4021.6,1154.4,2596.7,270.4
2007,Arkansas,2834797,3945.5,1124.4,2574.6,246.5
2008,Arkansas,2855390,3843.7,1182.7,2433.4,227.6
2004,California,35842038,3423.9,686.1,2033.1,704.8
2005,California,36154147,3321,692.9,1915,712
2006,California,36457549,3175.2,676.9,1831.5,666.8
2007,California,36553215,3032.6,648.4,1784.1,600.2
2008,California,36756666,2940.3,646.8,1769.8,523.8
2004,Colorado,4601821,3918.5,717.3,2679.5,521.6
2005,Colorado,4663295,4041,745.1,2736,560
2006,Colorado,4753377,3441.8,682,2325.1,434.8
2007,Colorado,4861515,2991.3,588.5,2061.1,341.7
2008,Colorado,4939456,2856.7,571.4,2013.7,271.6
2004,Connecticut,3498966,2684.9,456.1,1908.3,320.5
2005,Connecticut,3500701,2579,435.5,1840,303
2006,Connecticut,3504809,2575.442,6,1839,8,292.6
```

Transform Script

Export

- ▶ Split **data repeatedly** on **newline** into **rows**
- ▶ Split **data repeatedly** on **'tab'**
- ▶ Delete **empty rows**
- ▶ Promote row **2** to header
- ▶ Delete **rows where Year = 'Year'**
- ▶ Extract from **Year** after **'in '**
- ▶ Fill **extract** with values from **above**
- ▶ Delete **rows where Year starts with 'Reported crime in'**

DataWrangler

The screenshot displays the DataWrangler interface. On the left, a 'Suggestions' panel lists four actions: 'Delete rows 8,10', 'Delete empty rows', 'Delete rows where Property_crime_rate is null', and 'Delete rows where Year is null'. Below this is a 'Script' panel with two suggestions: 'Split data repeatedly on newline into rows' and 'Split data repeatedly on \''. An 'Export' button is located to the right of the script panel. The main area on the right shows a data table with 408 rows. The table has two columns: 'Year' and 'Property_crime_rate'. The data is grouped by state, with 'Reported crime in Alabama' and 'Reported crime in Alaska' as section headers. The table includes row numbers and data for years 2004 through 2007.

#	Year	#	Property_crime_rate
1	Reported crime in Alabama		
2			
3	2004		4029.3
4	2005		3900
5	2006		3937
6	2007		3974.9
7	2008		4081.9
8			
9	Reported crime in Alaska		
10			
11	2004		3370.9
12	2005		3615
13	2006		3582
14	2007		3373.9

Wrangler: Interactive Visual Specification of Data Transformation Scripts

Kandel et al. [CHI 2011]

The first sign that a visualization is good is that it shows you a problem in your data.

Every successful visualization that I've been involved with has had this stage where you realize, "Oh my God, this data is not what I thought it would be!" So already, you've discovered something.

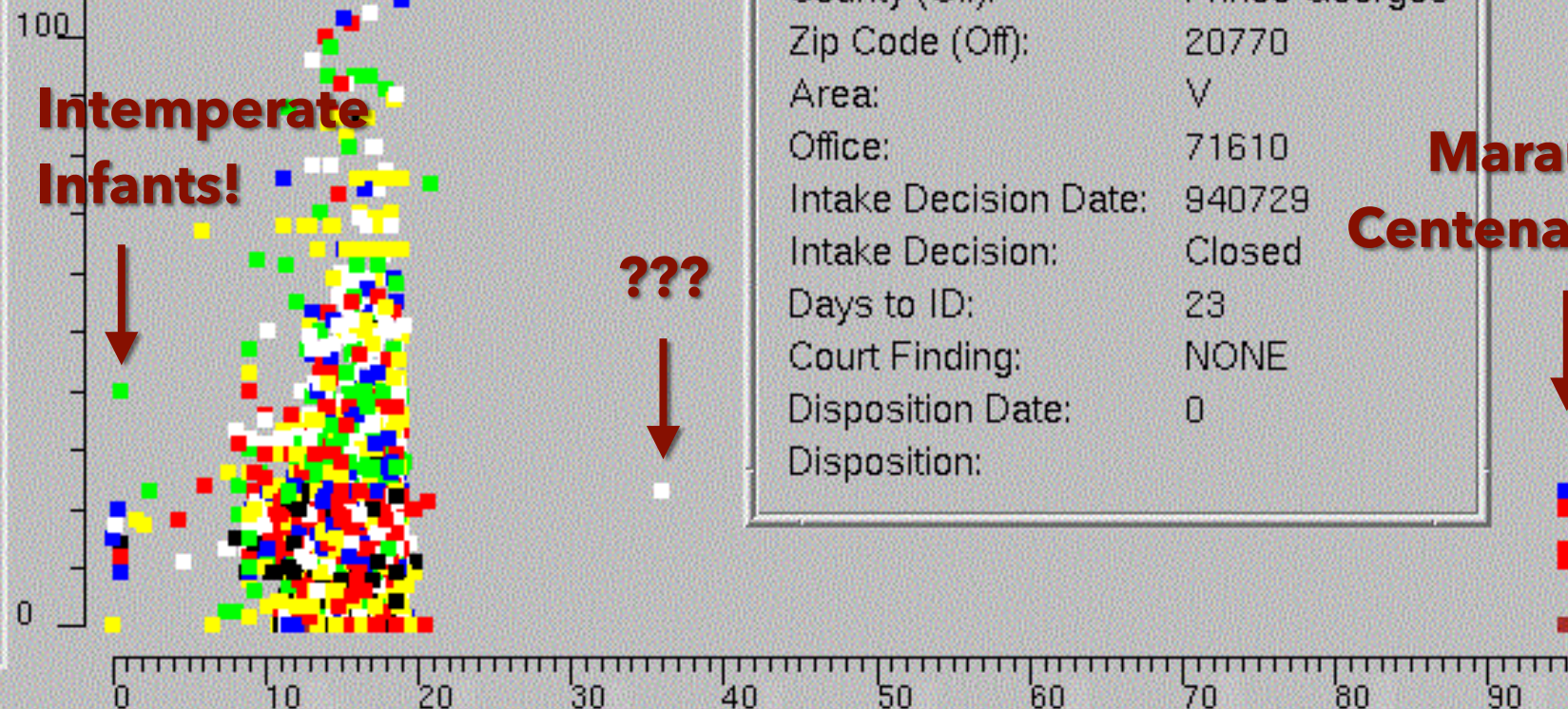
Martin Wattenberg [ACM Queue '09]

Alleged Offense:	HARAS
Offense Level:	2 - Misdemeanor
County (Off):	Prince Georges
Zip Code (Off):	20770
Area:	V
Office:	71610
Intake Decision Date:	940729
Intake Decision:	Closed
Days to ID:	23
Court Finding:	NONE
Disposition Date:	0
Disposition:	

**Intemperate
Infants!**

???

**Marauding
Centenarians!**

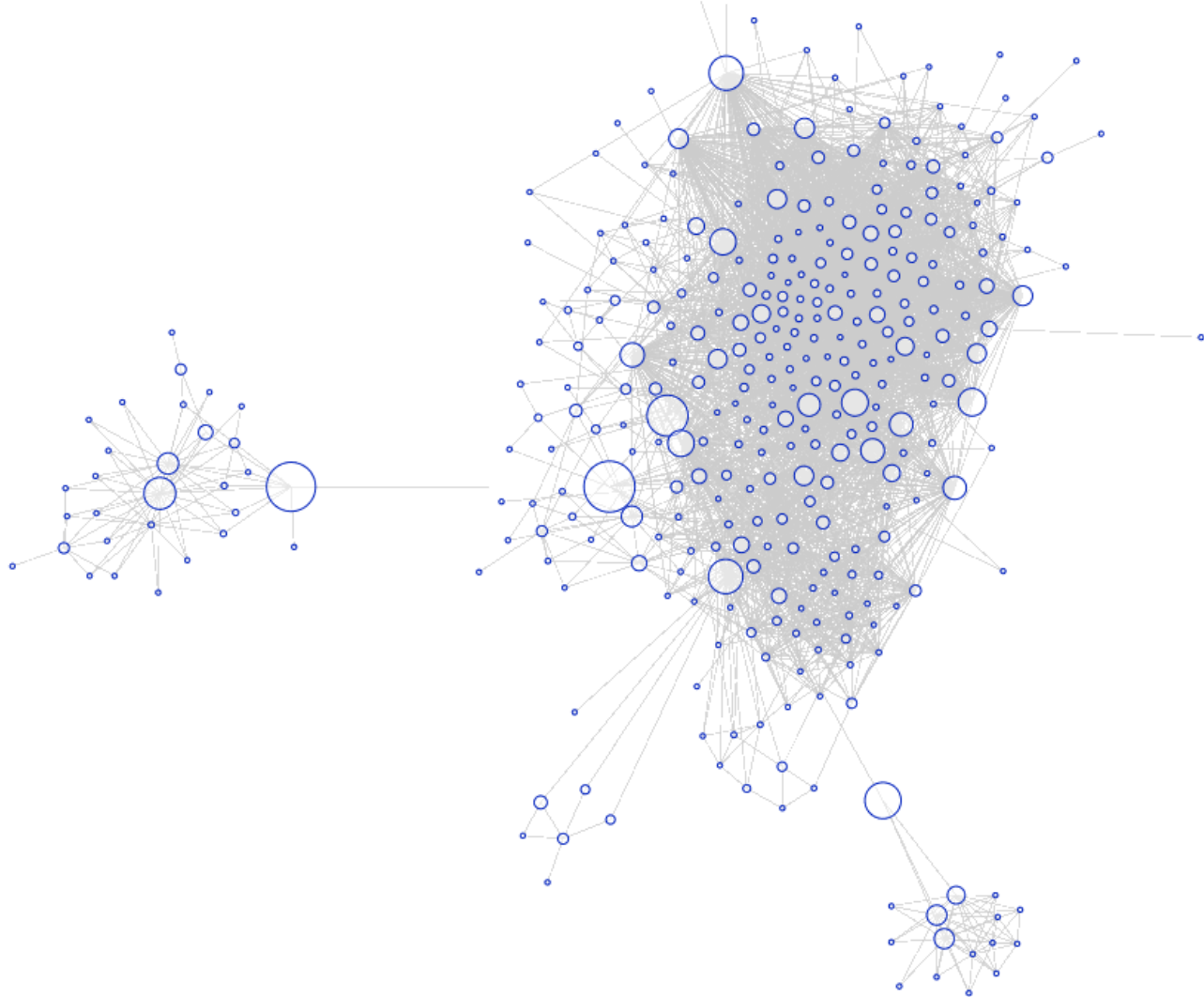
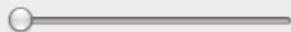
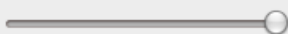


Age

Query Result: 4792 out of 4792 (100%)

None

Edge centrality filters:



Graph Viewer

Roll-up by:

All

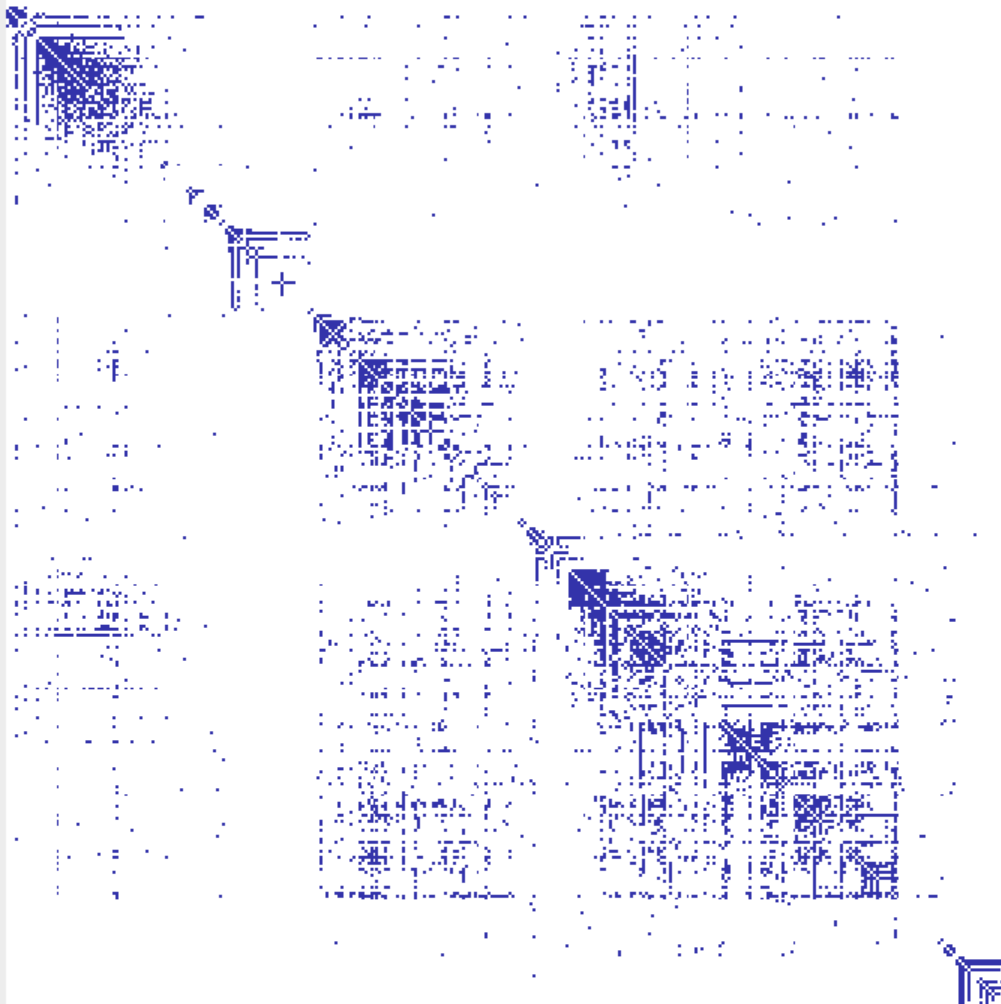
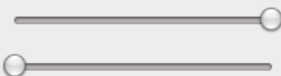
Visualization:

Matrix

Sort by:

Linkage

Edge centrality filters:





Graph Viewer

Roll-up by:

All

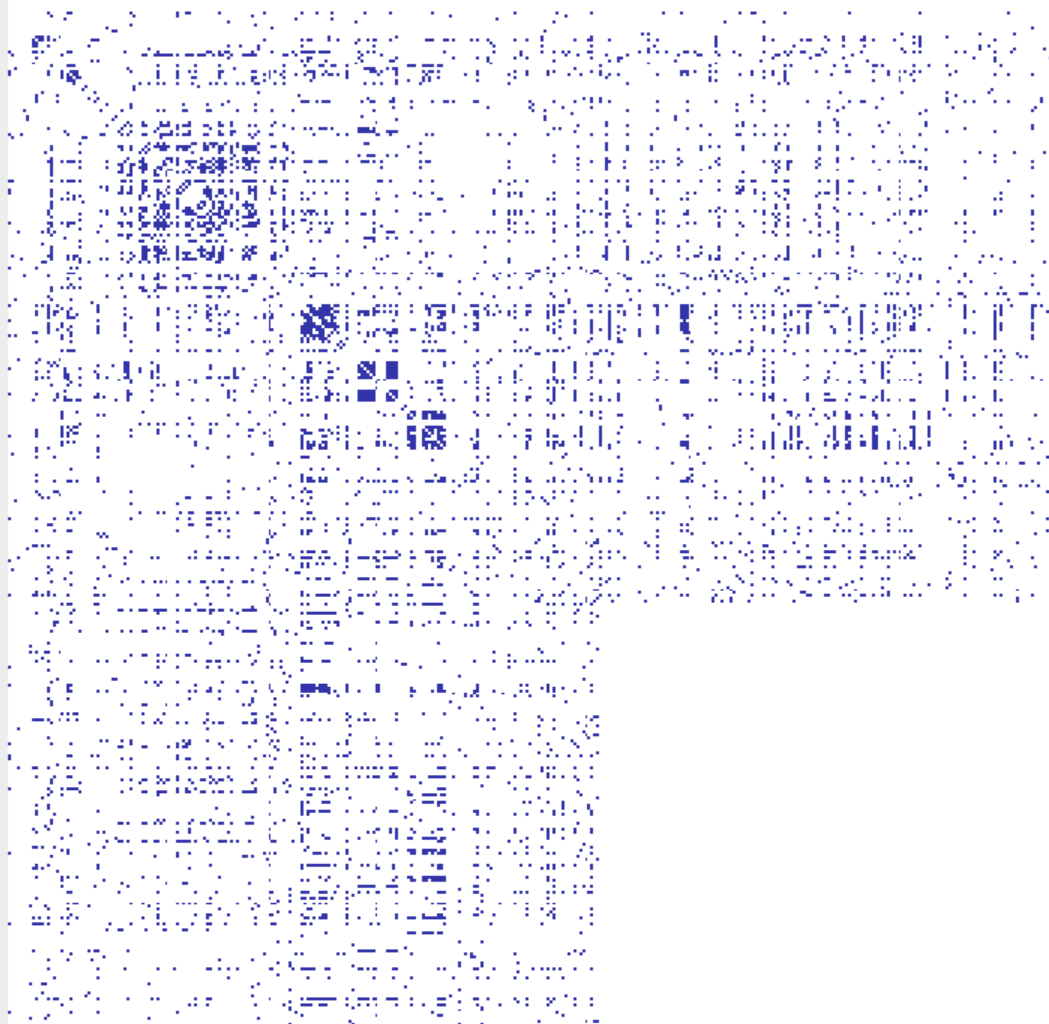
Visualization:

Matrix

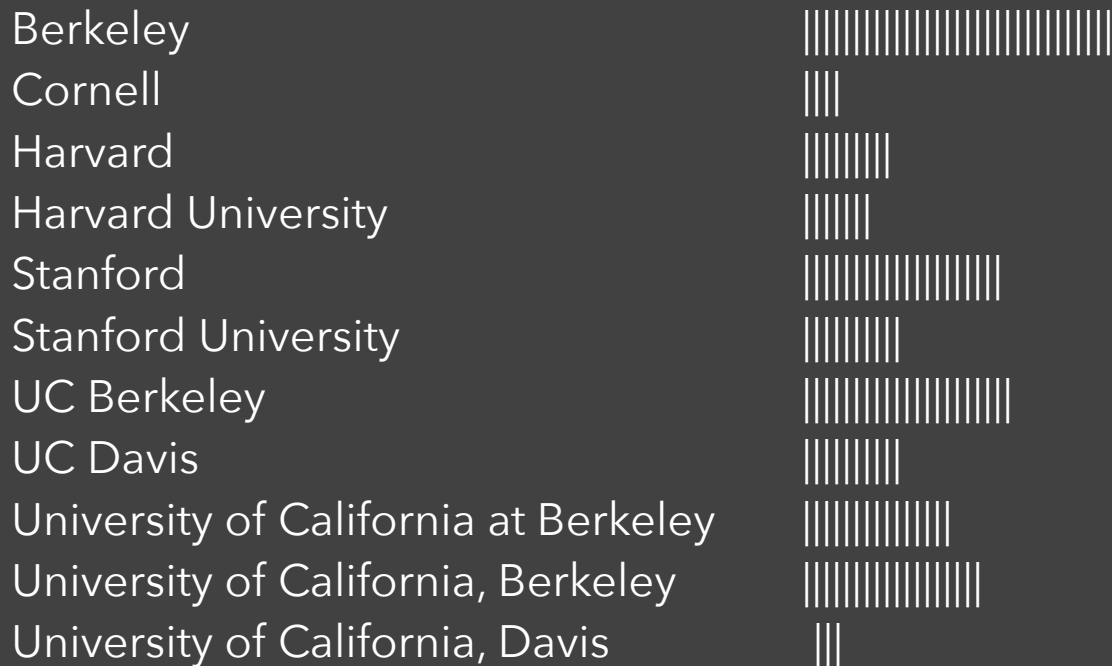
Sort by:

None

Edge centrality filters:



Visualize Friends by School?



Data Quality Hurdles

Missing Data	no measurements, redacted, ...?
Erroneous Values	misspelling, outliers, ...?
Type Conversion	e.g., zip code to lat-lon
Entity Resolution	diff. values for the same thing?
Data Integration	effort/errors when combining data

Anticipate problems with your data!

Data Wrangling Tools

Libraries

JavaScript: Arquero

Python: Pandas, Polars

R: dplyr

Databases

DuckDB + SQL queries

Graphical Tools

We'll look at some of these next!

Trifacta Wrangler (now part of Alteryx)

Campaign Finance 2016 > cn16

1 → 0 Generate Results

Columns: All Transformed: 3 Columns Rows: All Transformed: 4,859 Rows

Filter in grid

Grid Columns Full Dataset - 461.78kB 17 Columns 4,864 Rows 3 Data Types

Source to be dropped Preview

CAND_ID	CAND_NAME	CAND_NAME1	CAND_NAME2	CAND_PARTY_AFFILIATION	CAND_ELECTION_YEAR
4,864 Categories	4,760 Categories	3,416 Categories	3,677 Categories	76 Categories	1986 - 2052
H0AK00097	COX, JOHN R.	COX	JOHN R.	REP	2014
H0AL02087	ROBY, MARTHA	ROBY	MARTHA	REP	2016
H0AL02095	JOHN, ROBERT E JR	JOHN	ROBERT E JR	IND	2016
H0AL05049	CRAMER, ROBERT E "BUD" JR	CRAMER	ROBERT E "BUD" JR	DEM	2008
H0AL05163	BROOKS, MO	BROOKS	MO	REP	2016
H0AL06088	COOKE, STANLEY KYLE	COOKE	STANLEY KYLE	REP	2010
H0AL07086	SEWELL, TERRI A.	SEWELL	TERRI A.	DEM	2016
H0AL07094	HILLIARD, EARL FREDERICK JR	HILLIARD	EARL FREDERICK JR	DEM	2010
H0AL07177	CHAMBERLAIN, DON	CHAMBERLAIN	DON	REP	2012
H0AR01083	CRAWFORD, ERIC ALAN RICK	CRAWFORD	ERIC ALAN RICK	REP	2016
H0AR01091	GREGORY, JAMES CHRISTOPHER	GREGORY	JAMES CHRISTOPHER	DEM	2010
H0AR01109	CAUSEY, CHAD	CAUSEY	CHAD	DEM	2010
H0AR01125	SMITH, PRINCELLA D	SMITH	PRINCELLA D	REP	2010
H0AR02107	GRIFFIN, JOHN TIMOTHY	GRIFFIN	JOHN TIMOTHY	REP	2014
H0AR02131	ELLIOTT, JOYCE ANN	ELLIOTT	JOYCE ANN	DEM	2010
H0AR03022	SKOCH, BERNARD KURT 'BERNIE'	SKOCH	BERNARD KURT 'BERNIE'	REP	2010
H0AR03030	WHITAKER, DAVID JEFFREY	WHITAKER	DAVID JEFFREY	DEM	2010
H0AR03055	WOMACK, STEVE	WOMACK	STEVE	REP	2016
H0AS00018	FALEOMAVAEGA, ENI	FALEOMAVAEGA	ENI	DEM	2014
H0AZ01184	FLAKE, JEFF MR.	FLAKE	JEFF MR.	REP	2012
H0AZ01259	GOSAR, PAUL ANTHONY	GOSAR	PAUL ANTHONY	REP	2016

SUGGESTIONS

Split CAND_NAME into 2 columns on '{delim-ws}'

CAND_NAME	CAND_NAME1	CAND_NAME2
COX, JOHN R.	COX	JOHN R.
ROBY, MARTHA	ROBY	MARTHA
JOHN, ROBERT E JR	JOHN	ROBERT E JR

Affects 1 column, 4859 rows Creates 2 columns

Extract '{delim-ws}' from CAND_NAME

CAND_NAME	CAND_NAME1
COX, JOHN R.	,
ROBY, MARTHA	,
JOHN, ROBERT E JR	,

Affects 1 column, 4859 rows Creates 1 column

Count occurrences of '{delim-ws}'

CAND_NAME
COX, JOHN R.
ROBY, MARTHA
JOHN, ROBERT E JR

Affects 1 column, 4859 rows

AWS Glue DataBrew

☰

nyccitibikes

Dataset: citibike

Sample: First n sample (500 rows)

No job runs, no job runs scheduled

Run job

Job Details

Lineage

Actions

Datasets

Undo

Redo

Filter

Column

Format

Clean

Extract

Missing

Invalid

Duplicates

Split

Merge

Create

Functions

Unpivot

Pivot

Group

Join

Union

Text

Scale

Mapping

Encode

Recipes

Projects

Viewing 21 columns 500 rows

View highlighted

Grid

Schema

Profile

Recipes

Jobs

Community

SOURCE

start station latitude

will be deleted

SOURCE

start station longitude

will be deleted

PREVIEW

Alt latlong

end station id

Total 500

Unique 334

6 1.2%

5 1%

5 1%

84 96.8%

Min 40.46

Median 40.74

Mean 40.74

Mode 40.72

Max 40.85

Total 500

Unique 334

6 1.2%

5 1%

5 1%

84 96.8%

Min -74.02

Median -73.98

Mean -73.98

Mode -74

Max -73.9

Total 500

Unique 334

6 1.2%

5 1%

5 1%

84 96.8%

Min 40.72210379

Median 40.741177603

Mean 40.741177603

Mode 40.741177603

Max 40.741177603

Total 500

Unique 330

6 1.2%

5 1%

5 1%

84 96.8%

Min 3966

Median 3668

Mean 3668

Mode 3668

Max 3668

40.819241

-73.941057

40.819241, -73.941057

3966

40.68691865

-73.976682

40.68691865, -73.976682

3668

40.7689738

-73.95482273

40.7689738, -73.95482273

3164

40.7619557

-73.968087

40.7619557, -73.968087

3906

40.71638032

-73.94821286

40.71638032, -73.94821286

128

40.704508

-73.9351

40.704508, -73.9351

3774

40.741177603

-74.00149746

40.741177603, -74.00149746

462

40.72110063

-73.9919254

40.72110063, -73.9919254

470

40.75038009

-73.98338988

40.75038009, -73.98338988

312

40.7668

-73.9347774

40.7668, -73.9347774

372

40.72362738

-73.99949601

40.72362738, -73.99949601

400

40.733765

-73.96222088

40.733765, -73.96222088

405

40.825125

-73.941616

40.825125, -73.941616

3629

40.70870368

-73.9448625

40.70870368, -73.9448625

3070

40.73314259

-73.97573881

40.73314259, -73.97573881

487

40.71882

-73.93948

40.71882, -73.93948

3585

40.65539977

-74.01062787

40.65539977, -74.01062787

3041

40.73124

-73.95161

40.73124, -73.95161

3119

40.72210379

-73.99724901

40.72210379, -73.99724901

325

40.78414472

-73.98562492

40.78414472, -73.98562492

3160

40.7652654

-73.98192338

40.7652654, -73.98192338

468

40.72706363

-73.99662137

40.72706363, -73.99662137

3812

40.7627204

-73.971888

40.7627204, -73.971888

500

Zoom

100%

Merge columns

Merge columns Info

Merge columns and create a new column

Source column

Select two or more columns in the order to merge

start station latitude

start station longitude

Add a column

Separator - Optional

Concatenated values are separated by this

,

New column name

Name of the target column to merge into

latlong

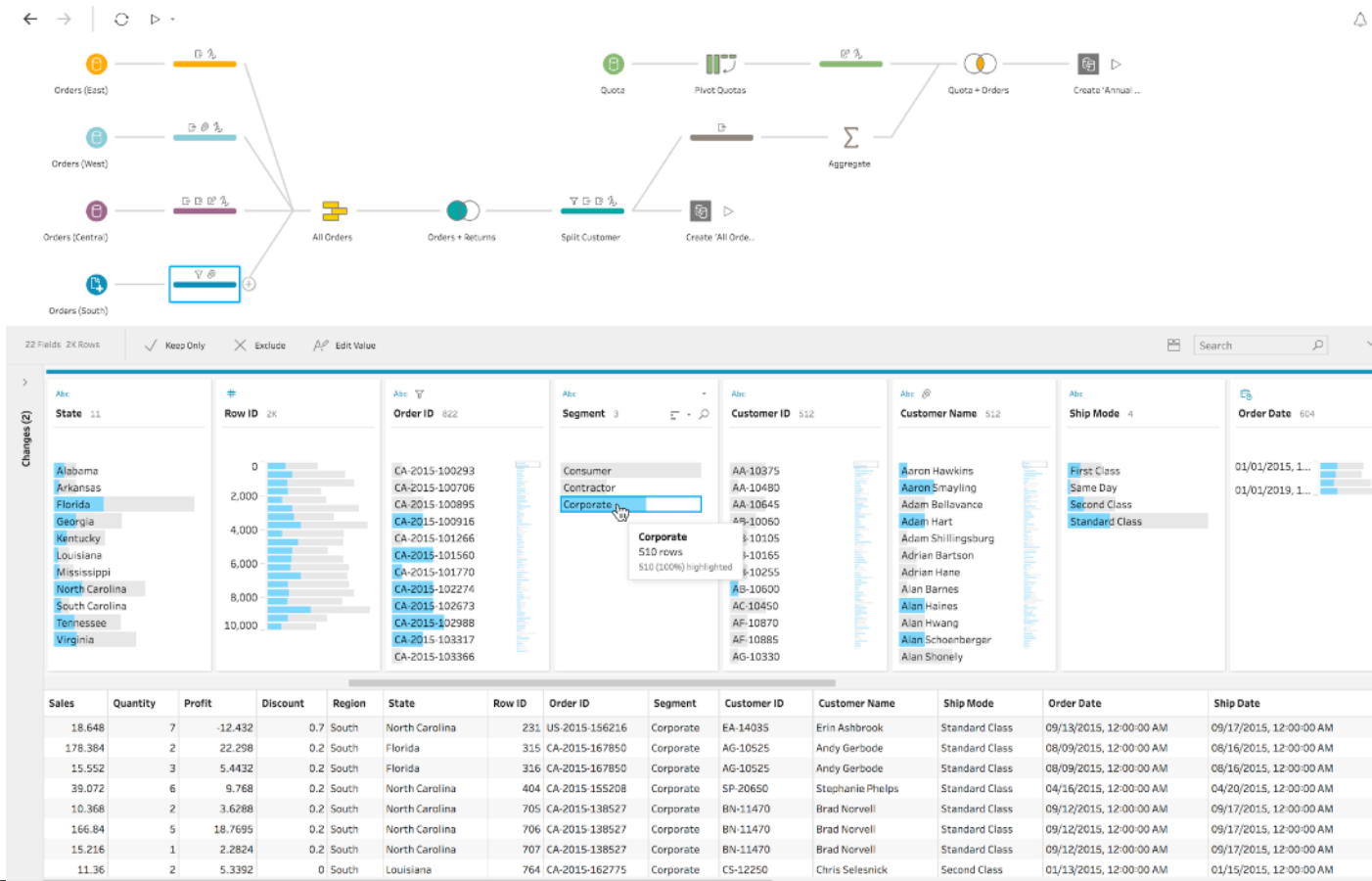
Valid characters are alphanumeric, underscore, and space

Preview shown

Cancel

Apply

Tableau Prep



Deepnote

df.head(50)

[13]

Visualize

DEPENDENTS bool

False 82%
True 18%

TECHSUPPORT ob...

No 62%
Yes 26%
No internet service 12%

CONTRACT object

Month-to-month 66%
Two year 18%
One year 16%

PAPERLESSBILL...

true 58%
false 42%

MONTHLYCHAR...

19.35 - 110.15

TOTALCHARGES f...

44.7 - 7998.8

CHURNVALUE flo...

0.0 - 1.0

TENUREMONTHS 1...

1 - 72

false	Yes	Month-to-month	true	83.4	83.4	0	1	Y
false	No	One year	true	100.05	6254.2	1	64	Y
false	No	Month-to-month	true	69.1	69.1	1	1	Y
false	No	Month-to-month	true	85.35	1375.15	1	16	Y
true	No	Month-to-month	true	79.25	1111.65	1	13	Y
false	No	Month-to-month	false	74.4	434.1	1	6	Y
false	No internet service	Two year	false	19.35	1099.6	1	59	Y

Observable Data Table Cells

cell 1072

= chinook

customers

59 rows

Run

Filter

Columns 13

Sort

Slice [0, 100]

SQL

▼	CustomerId number	FirstName string	LastName string	Company string	Address string	City string	St
	060	93% unique	59 unique values	83% NULL/EMPTY	59 unique values	80%	
	0	60	57 categories	11 categories		53 categories	26
0	1	Luís	Gonçalves	Embraer - Empresa Brasileira de Aeronáutica S.A.	Av. Brigadeiro Faria Lima, 2170	São José dos Campos	SF
1	2	Leonie	Köhler	NULL	Theodor-Heuss-Straße 34	Stuttgart	NI
2	3	François	Tremblay	NULL	1498 rue Bélanger	Montréal	QC
3	4	Bjørn	Hansen	NULL	Ullevålsveien 14	Oslo	NI
4	5	František	Wichterlová	JetBrains s.r.o.	Klanova 9/506	Prague	NI
5	6	Helena	Holý	NULL	Rilská 3174/6	Prague	NI
6	7	Astrid	Gruber	NULL	Rotenturmstraße 4, 1010 Innere Stadt	Vienne	NI
7	8	Daan	Peeters	NULL	Grétrystraat 63	Brussels	NI
8	9	Kara	Nielsen	NULL	Sønder Boulevard 51	Copenhagen	NI
9	10	Eduardo	Martins	Woodstock Discos	Rua Dr. Falcão Filho, 155	São Paulo	SF

10 per page ▼

page 1 of 6

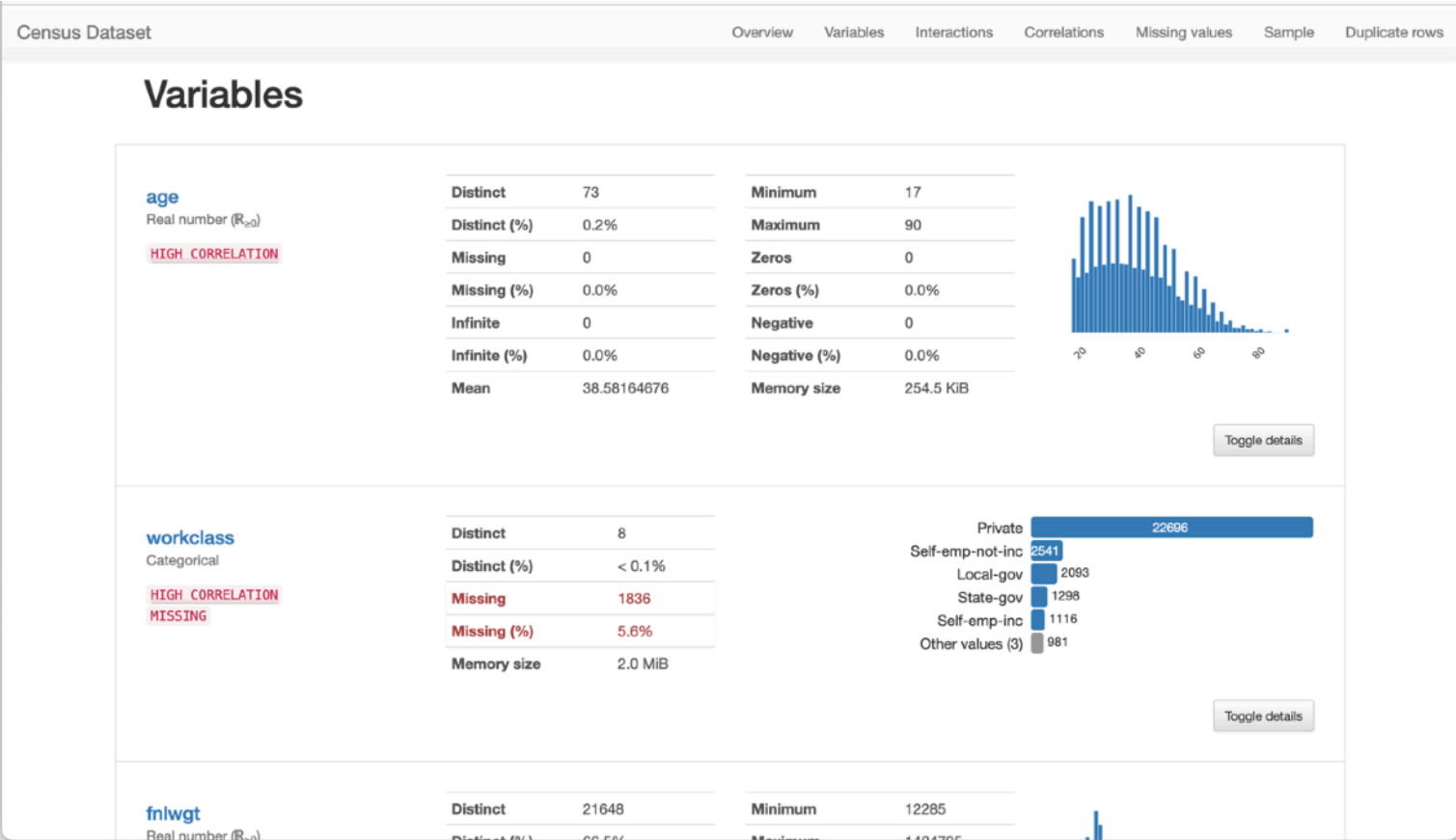
◀ ▶

Quak widget usable in Jupyter Notebooks

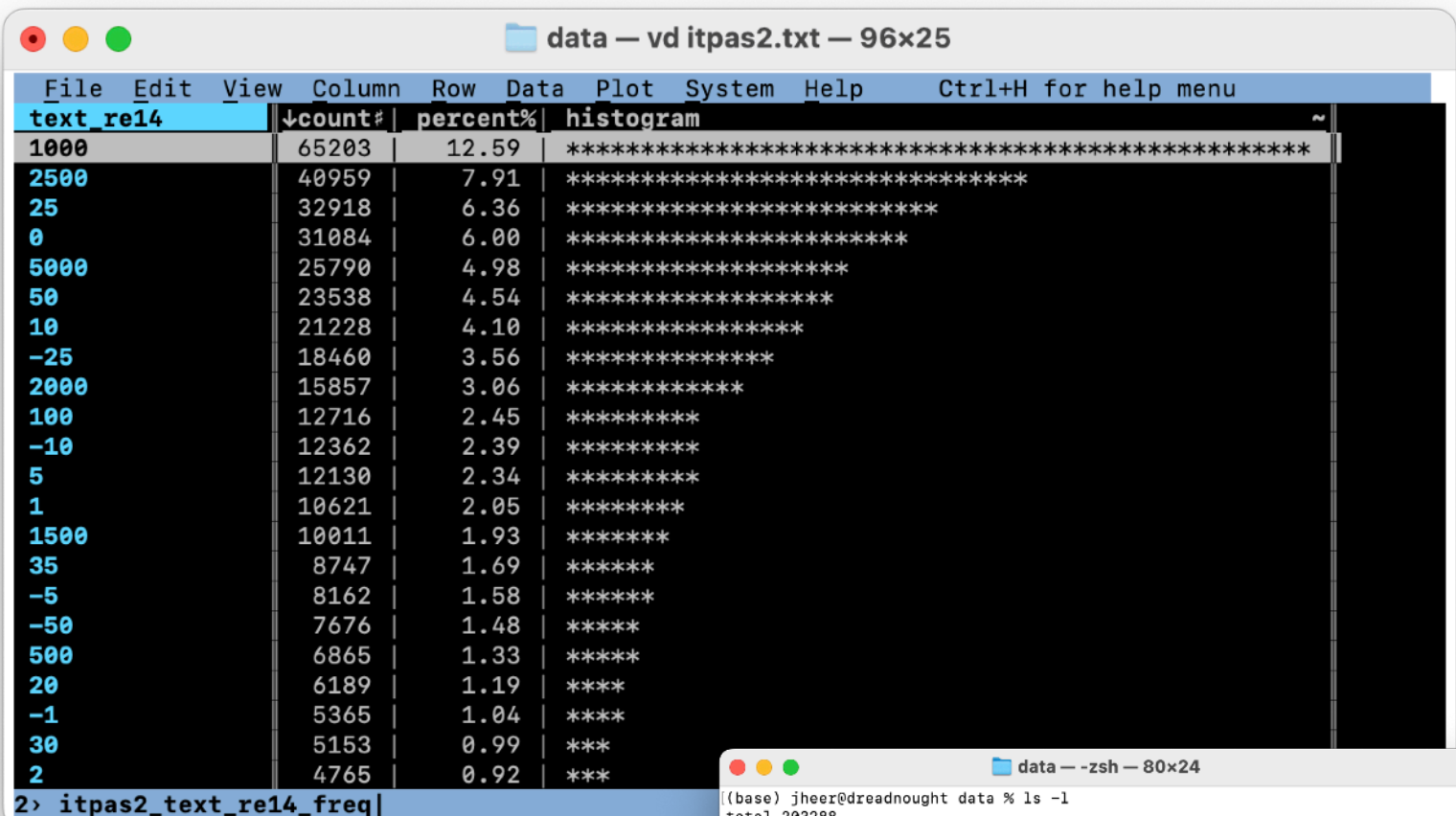
	name utf8 unique 22 categories nationality utf8 207 categories sex utf8 male female female height float64 1.22.3 weight int64 20180 sport utf8 ath 28 categories gold int64 05.5 silver int64 0-								
0	A Lam Shin	KOR	female	1.68	56	fencing	0		
1	Auri Lorena Bokesa	ESP	female	1.8	62	athletics	0		
2	Abbey Weitzeil	USA	female	1.78	68	aquatics	1		
3	Abbie Brown	GBR	female	1.76	71	rugby sevens	0		
4	Abby Erceg	NZL	female	1.75	68	football	0		
5	Abdoulkarim Fawzi...	CMR	female	1.8	67	volleyball	0		
6	Abigel Joo	HUN	female	1.83	76	judo	0		

Reset 3,420 of 11,538 rows

Pandas Profiling



VisiData



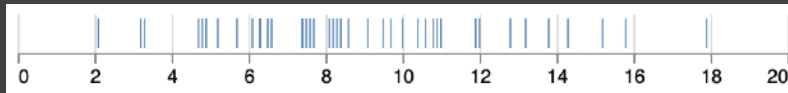
data — -zsh — 80x24

```
(base) jheer@dreadnought data % ls -l
total 203288
-rw-r--r-- 1 jheer staff 2336013 Oct 12 15:24 cm.txt
-rw-r--r-- 1 jheer staff 814306 Oct 12 15:24 cn.txt
-rw-r--r--@ 1 jheer staff 90781486 Oct 12 15:25 itpas2.txt
(base) jheer@dreadnought data % vd cn.txt
```

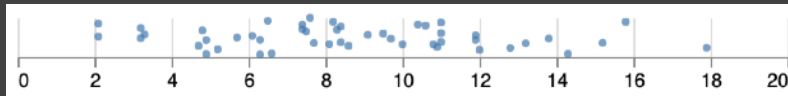
Visualizing Distributions

Distribution Visualizations

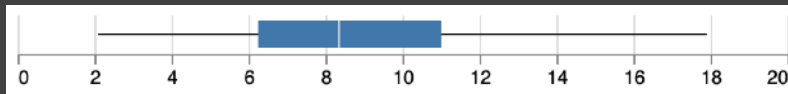
Strip Plot



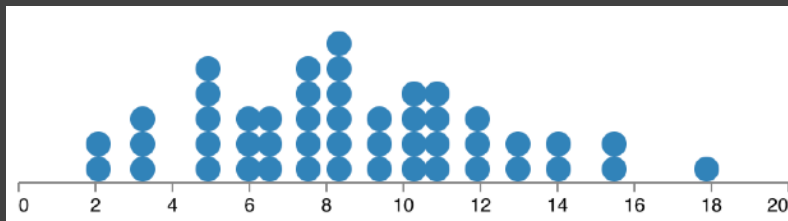
Jittered Plot



Box Plot



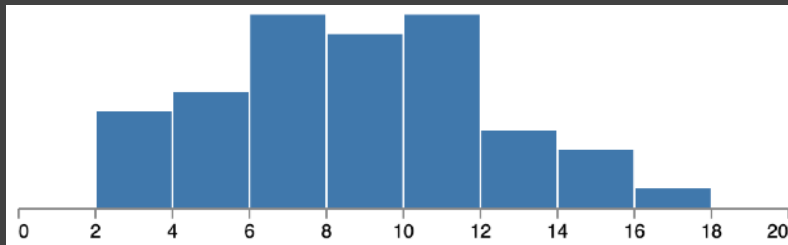
Dot Plot



Distribution Visualizations

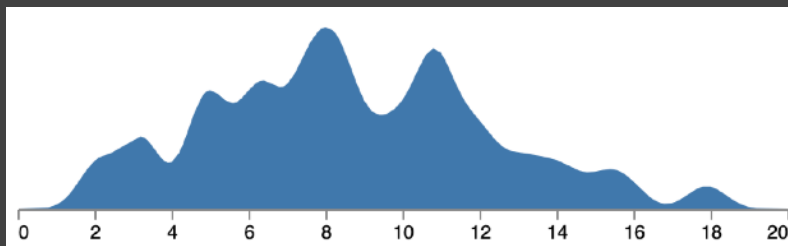
Histogram

bin size = 2



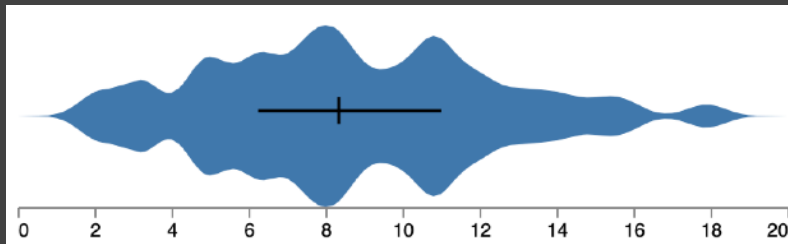
Density Plot

kde, $\sigma = 0.5$



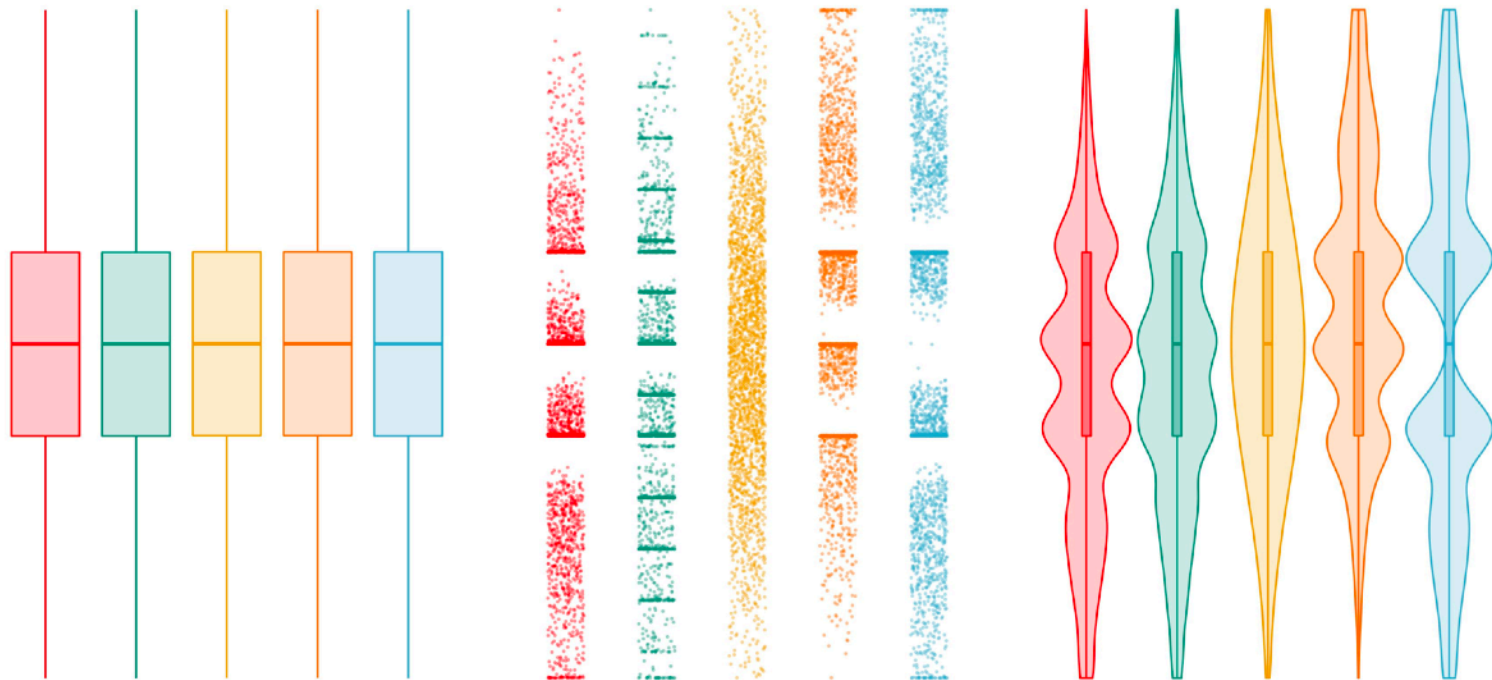
Violin Plot

kde, $\sigma = 0.5$

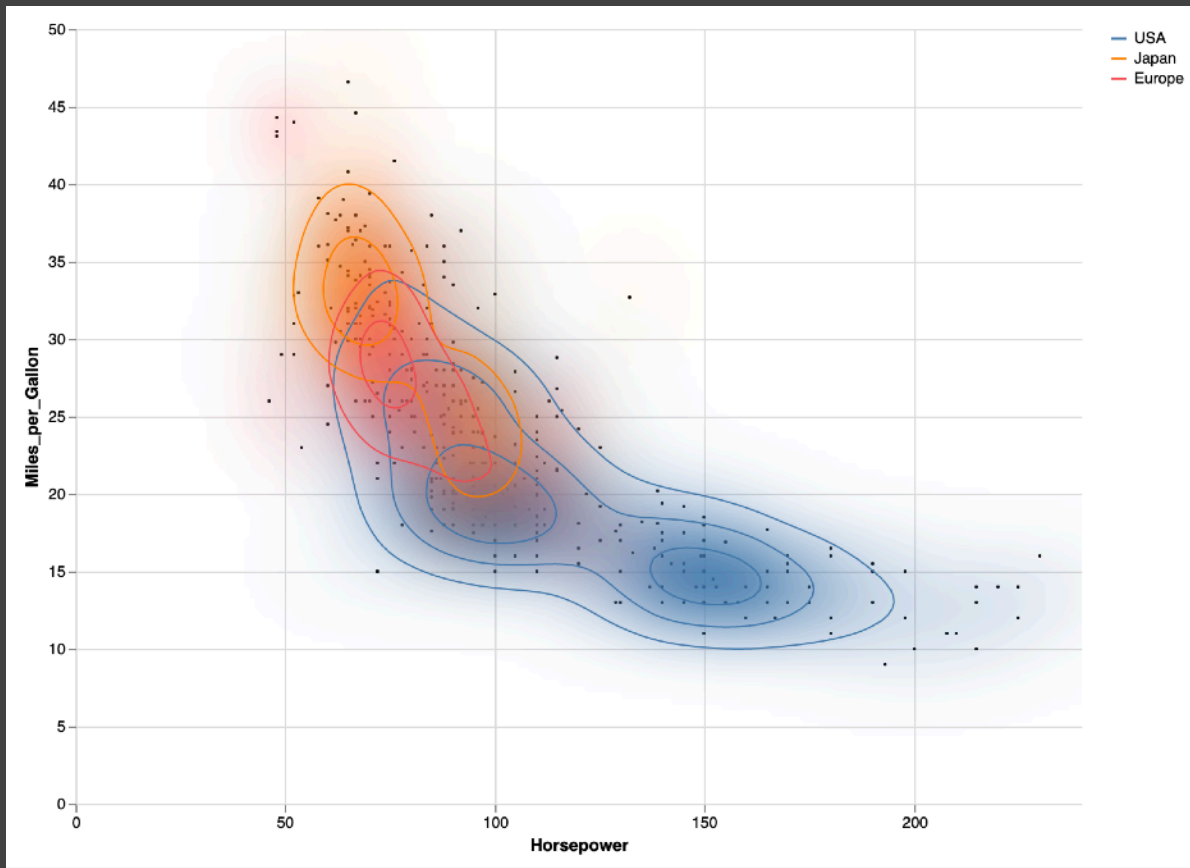


Identical boxplots, different distributions

Boxplots are great. They show medians and ranges and enable comparison of different groups. However, boxplots can be misleading. Different datasets can have the same descriptive statistics (left), but quite different underlying distributions (middle). Therefore, it is crucial to visualize the distribution in addition to descriptive statistics. Violin plots with integrated boxplots are great for this.

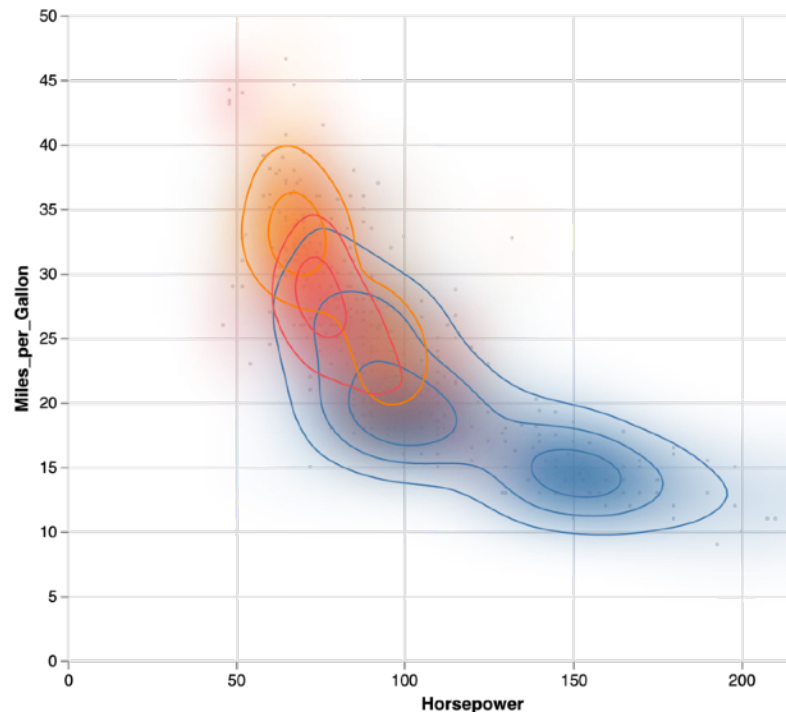
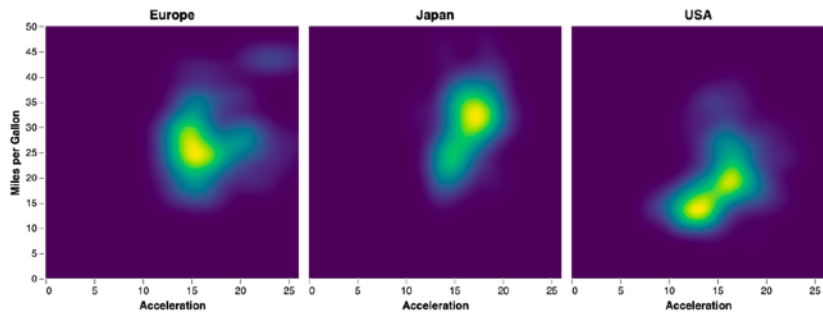
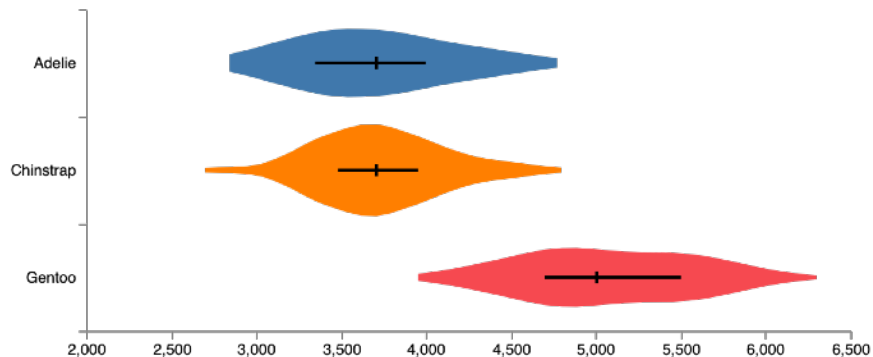


Now in 2D! Heatmaps, Contours



Kernel Density Estimation (KDE)

Enables violin plots, heat maps, contour plots...



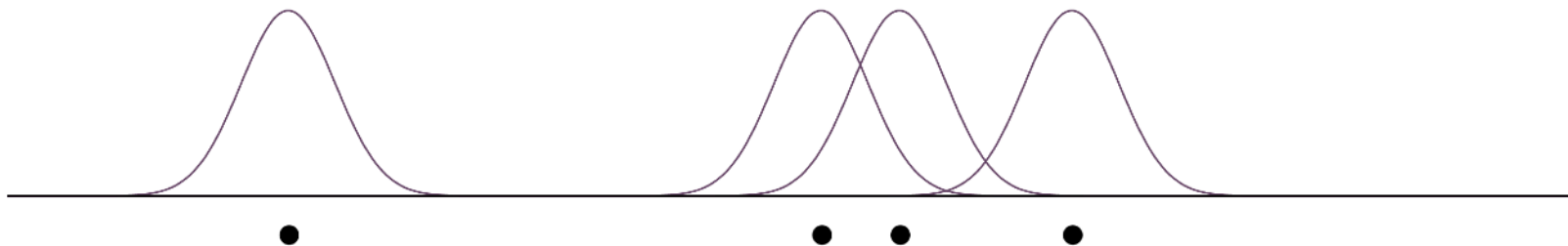
Kernel Density Estimation

For a set of input data points...



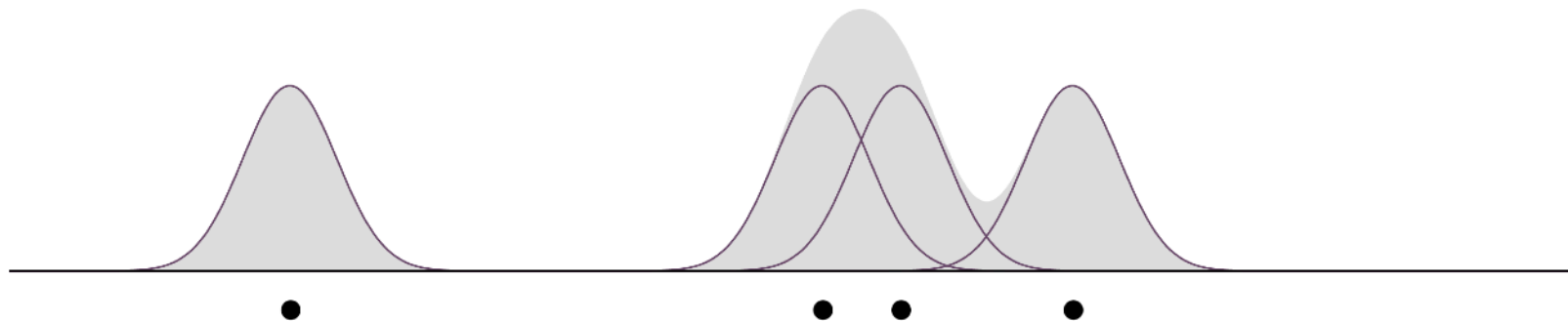
Kernel Density Estimation

Represent each point with a “kernel” distribution



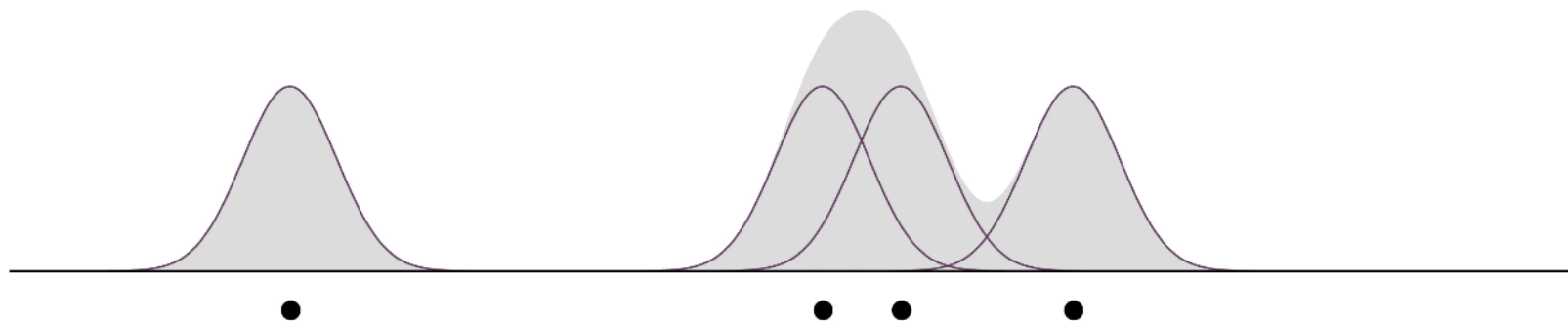
Kernel Density Estimation

Sum the kernels to form a density estimate



Kernel Density Estimation

Sized by bandwidth (standard deviation)



Dimensionality Reduction

Dimensionality Reduction (DR)

Project n D data to 2D or 3D for viewing. Often used to interpret and sanity check high-dimensional representations fit by machine learning methods.

Different DR methods make different trade-offs: for example to **preserve global structure** (e.g., PCA) or **emphasize local structure** (e.g., nearest-neighbor approaches, including t-SNE and UMAP).

In contrast, multidimensional scaling (MDS) attempts to **preserve pairwise distances**.

Reduction Techniques

LINEAR - PRESERVE GLOBAL STRUCTURE

Principal Components Analysis (PCA)

Linear transformation of basis vectors, ordered by amount of data variance they explain.

NON-LINEAR - PRESERVE LOCAL TOPOLOGY

t-Dist. Stochastic Neighbor Embedding (t-SNE)

Probabilistically model distance, optimize positions.

Uniform Manifold Approx. & Projection (UMAP)

Identify local manifolds, then stitch them together.

Mapping Emoji Images



Q Latent Dimensions: 32 ▾ Projection: t-SNE ▾ Perplexity: 30 ▾

t-SNE



Q Latent Dimensions: 32 ▾ Projection: UMAP ▾ Neighbors: 15 ▾ Distance: 0.1 ▾

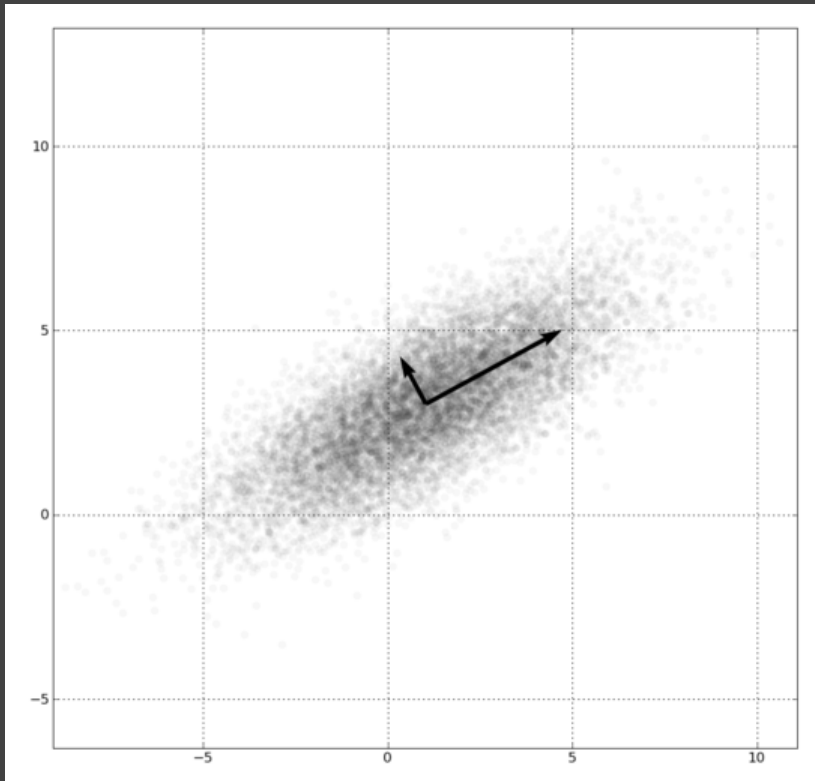
UMAP



Q Latent Dimensions: 32 ▾ Projection: PCA ▾ X-Axis: PC1 ▾ Y-Axis: PC2 ▾

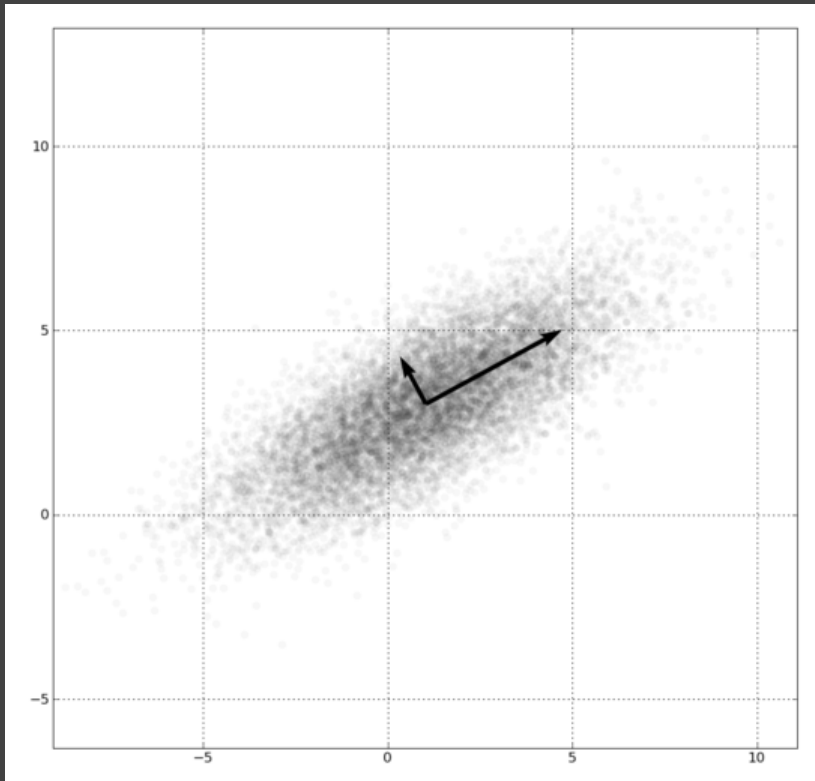
PCA

Principal Components Analysis



1. Mean-center the data.
2. Find $\perp$ basis vectors that maximize the data variance.
3. Plot the data using the top vectors.

Principal Components Analysis

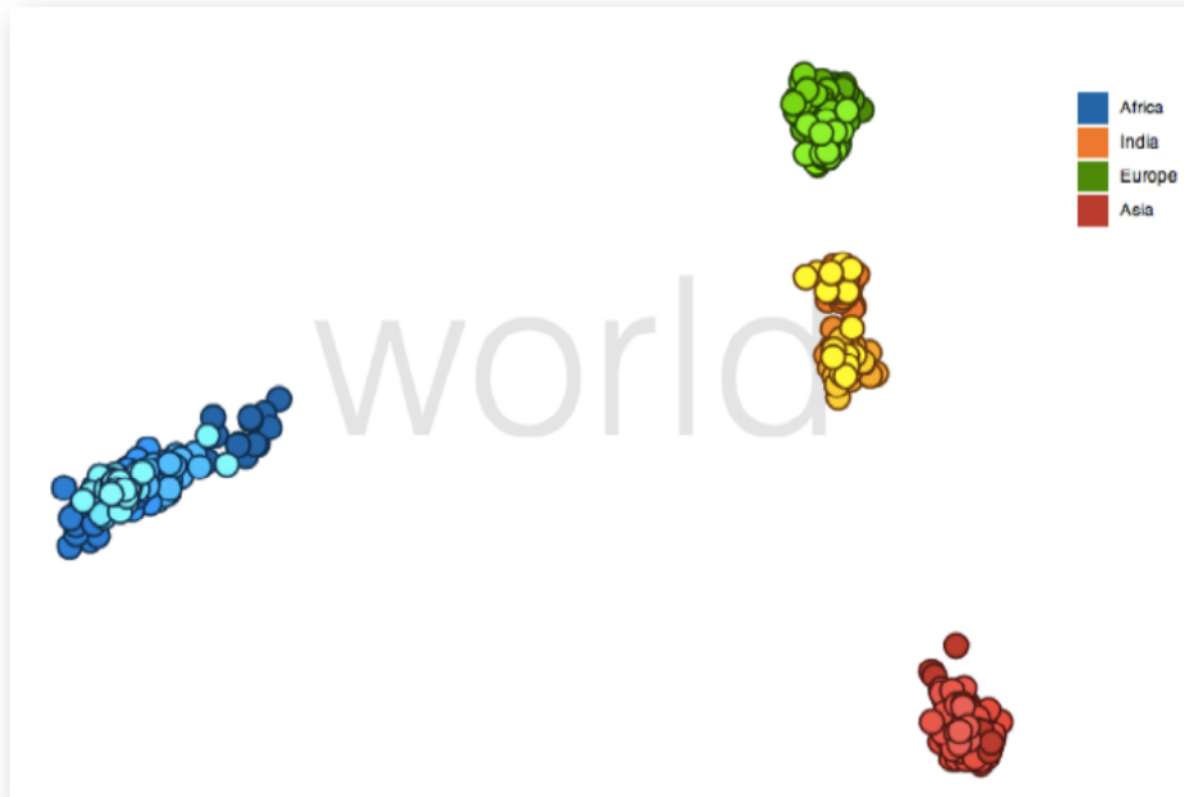


Linear transform:
scale and rotate
original space.

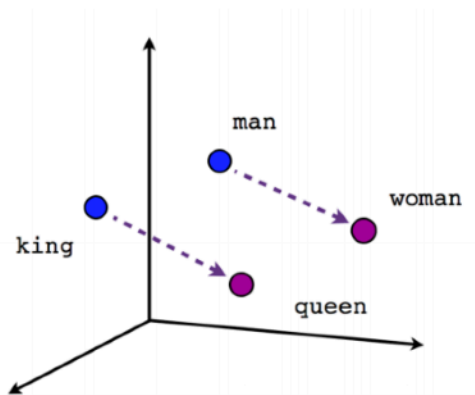
Lines (vectors)
project to lines.

Preserves global
distances.

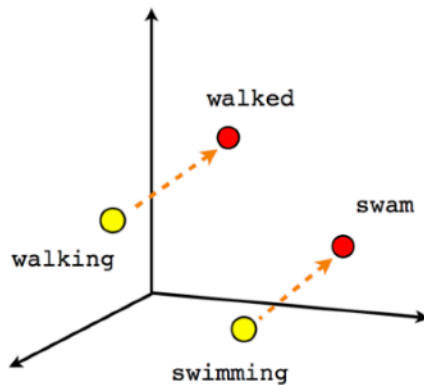
PCA of Genomes [Demiralp et al. '13]



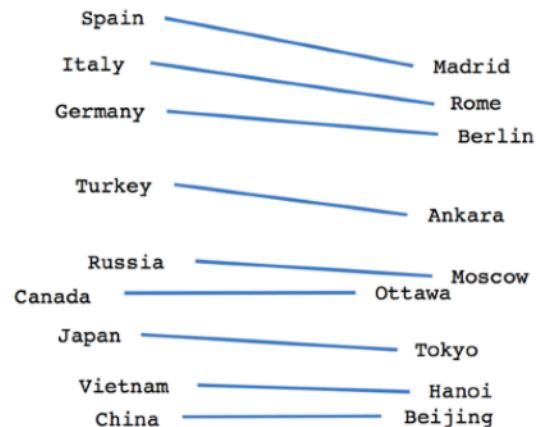
Vector Space Word Embeddings (word2vec, GloVe)



Male-Female



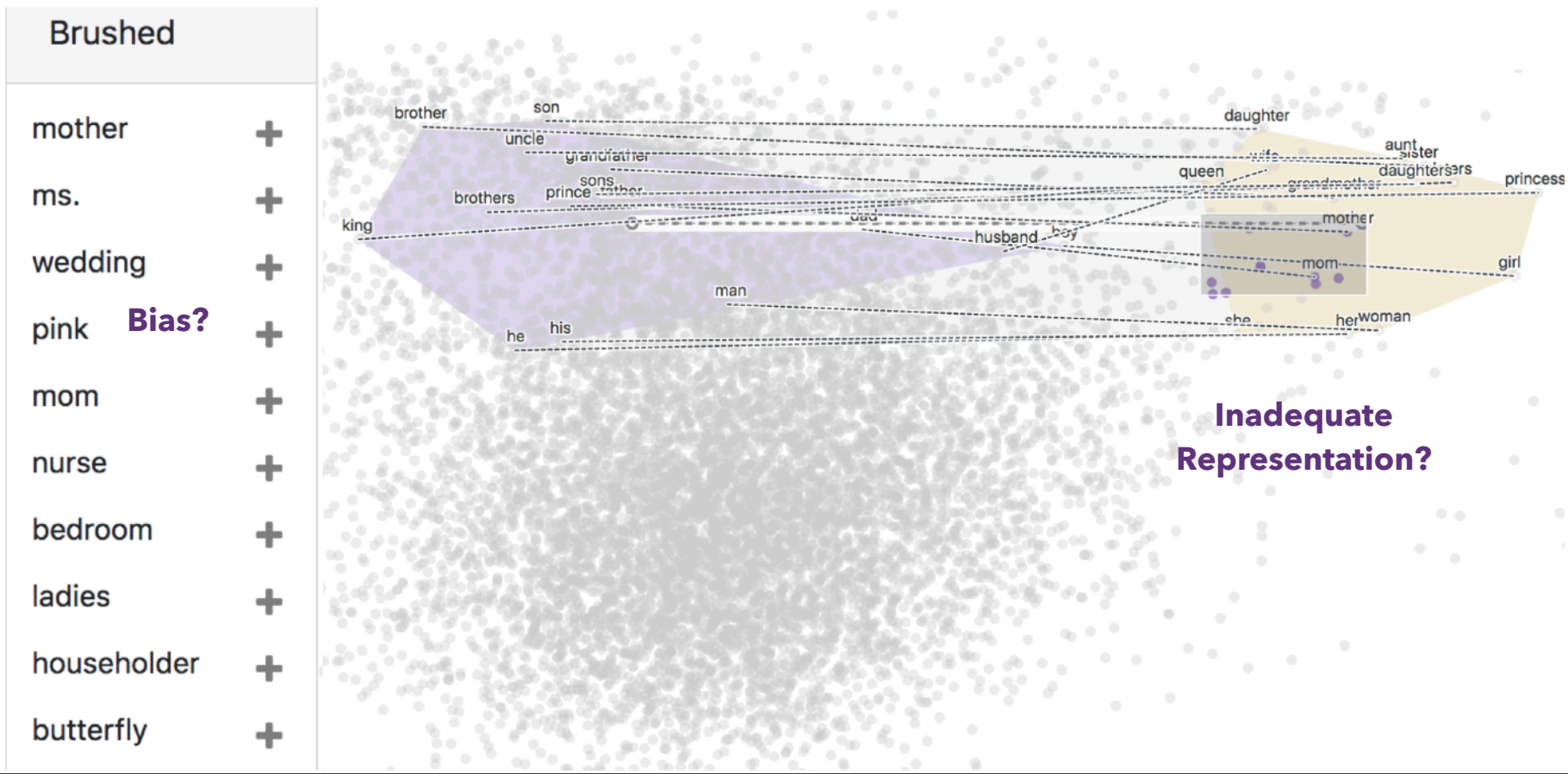
Verb tense



Country-Capital

Mapping Machine-Learned Latent Spaces

[Liu et al. 2019]



Non-Linear Techniques

Distort the space, trade-off preservation of global structure to emphasize local neighborhoods. Use topological (nearest neighbor) analysis.

Two popular contemporary methods:

t-SNE - probabilistic interpretation of distance

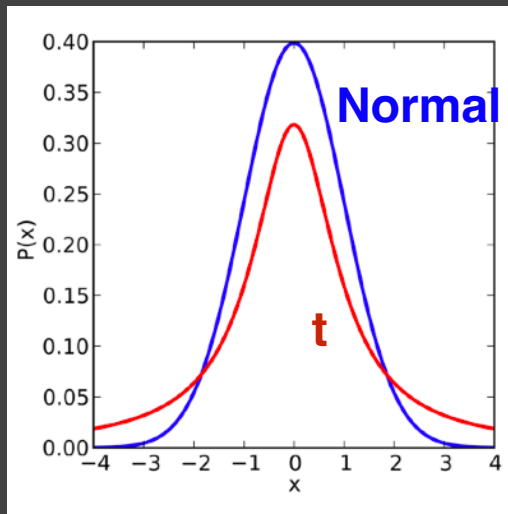
UMAP - tries to balance local/global trade-off

t-SNE [Maaten & Hinton 2008]

1. Model probability $\mathbf{P}$ of one point “choosing” another as its neighbor in the original space, using a Gaussian distribution defined using the distance between points. Nearer points have higher probability than distant ones.

t-SNE [Maaten & Hinton 2008]

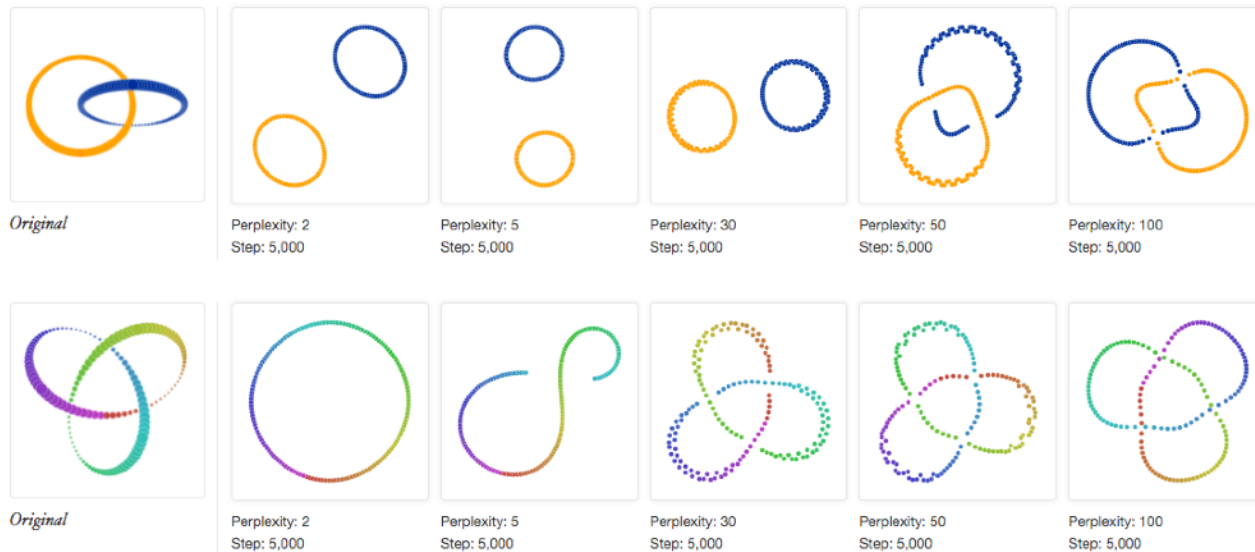
2. Define a similar probability $\mathbf{Q}$ in the low-dimensional (2D or 3D) embedding space, using a Student's t distribution (*hence the "t-" in "t-SNE"!*). The t -distribution is heavy-tailed, allowing distant points to be even further apart.



t-SNE [Maaten & Hinton 2008]

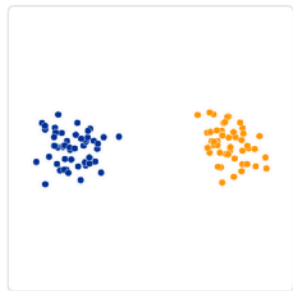
1. Model probability $\mathbf{P}$ of one point “choosing” another as its neighbor in the original space, using a Gaussian distribution defined using the distance between points. Nearer points have higher probability than distant ones.
2. Define a similar probability $\mathbf{Q}$ in the low-dimensional (2D or 3D) embedding space, using a Student’s t distribution (*hence the “t-” in “t-SNE”!*). The t -distribution is heavy-tailed, allowing distant points to be even further apart.
3. Optimize to find the positions in the embedding space that minimize the Kullback-Leibler divergence between the $\mathbf{P}$ and $\mathbf{Q}$ distributions: $KL(P \parallel Q)$

Visualizing t-SNE [Wattenberg et al. '16]

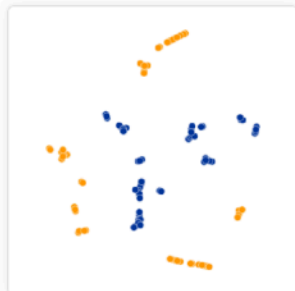


Results can be highly sensitive to the algorithm parameters!
Are you seeing real structures, or algorithmic hallucinations?

Hyperparameters matter!



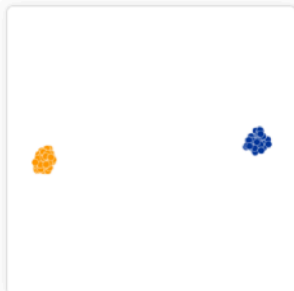
Original



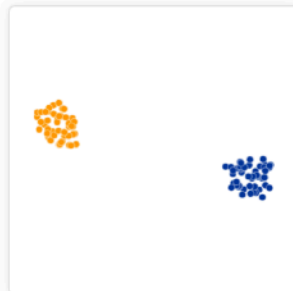
Perplexity: 2
Step: 5,000



Perplexity: 5
Step: 5,000



Perplexity: 30
Step: 5,000



Perplexity: 50
Step: 5,000



Perplexity: 100
Step: 5,000



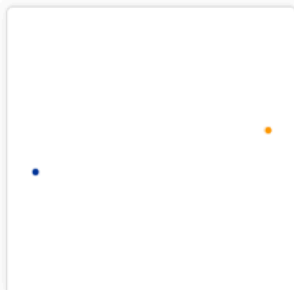
Original



Perplexity: 30
Step: 10



Perplexity: 30
Step: 20



Perplexity: 30
Step: 60

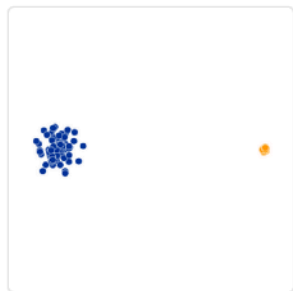


Perplexity: 30
Step: 120

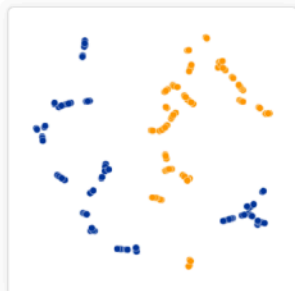


Perplexity: 30
Step: 1,000

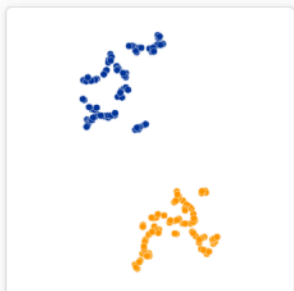
Cluster sizes mean nothing...



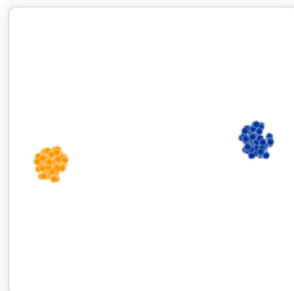
Original



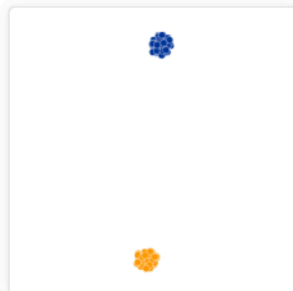
Perplexity: 2
Step: 5,000



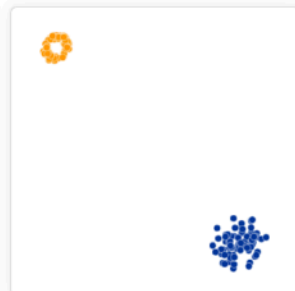
Perplexity: 5
Step: 5,000



Perplexity: 30
Step: 5,000

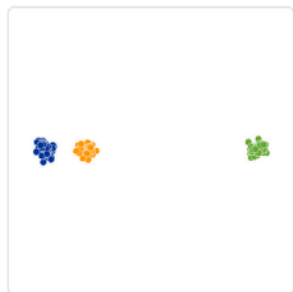


Perplexity: 50
Step: 5,000

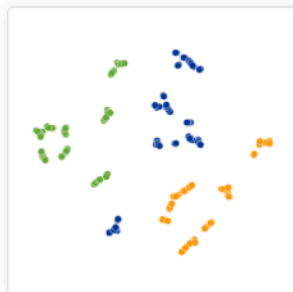


Perplexity: 100
Step: 5,000

Cluster distances are illusive



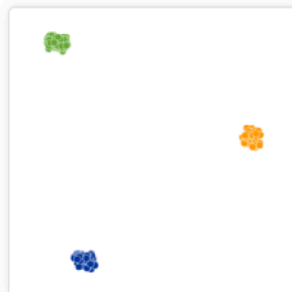
Original



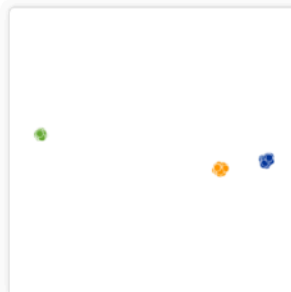
Perplexity: 2
Step: 5,000



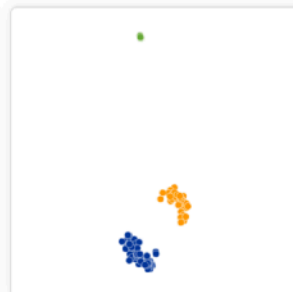
Perplexity: 5
Step: 5,000



Perplexity: 30
Step: 5,000



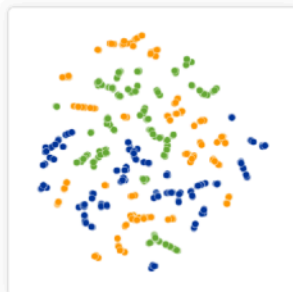
Perplexity: 50
Step: 5,000



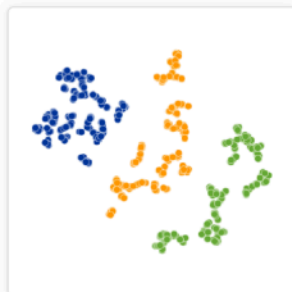
Perplexity: 100
Step: 5,000



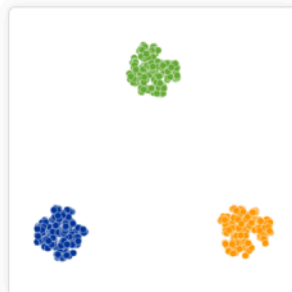
Original



Perplexity: 2
Step: 5,000



Perplexity: 5
Step: 5,000



Perplexity: 30
Step: 5,000



Perplexity: 50
Step: 5,000



Perplexity: 100
Step: 5,000

Random noise may not look like it



Original



Perplexity: 2
Step: 5,000



Perplexity: 5
Step: 5,000



Perplexity: 30
Step: 5,000



Perplexity: 50
Step: 5,000



Perplexity: 100
Step: 5,000

You can see shapes, sometimes



Original



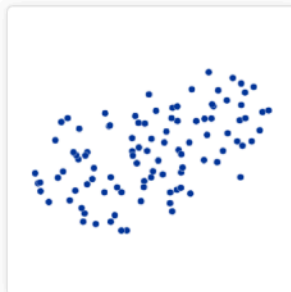
Perplexity: 2
Step: 5,000



Perplexity: 5
Step: 5,000



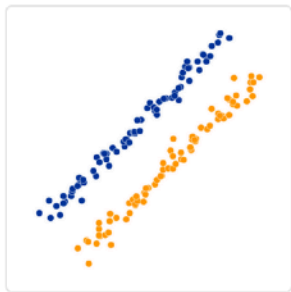
Perplexity: 30
Step: 5,000



Perplexity: 50
Step: 5,000



Perplexity: 100
Step: 5,000



Original



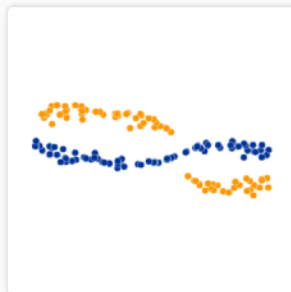
Perplexity: 2
Step: 5,000



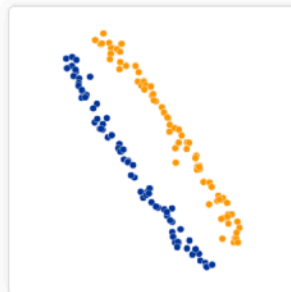
Perplexity: 5
Step: 5,000



Perplexity: 30
Step: 5,000

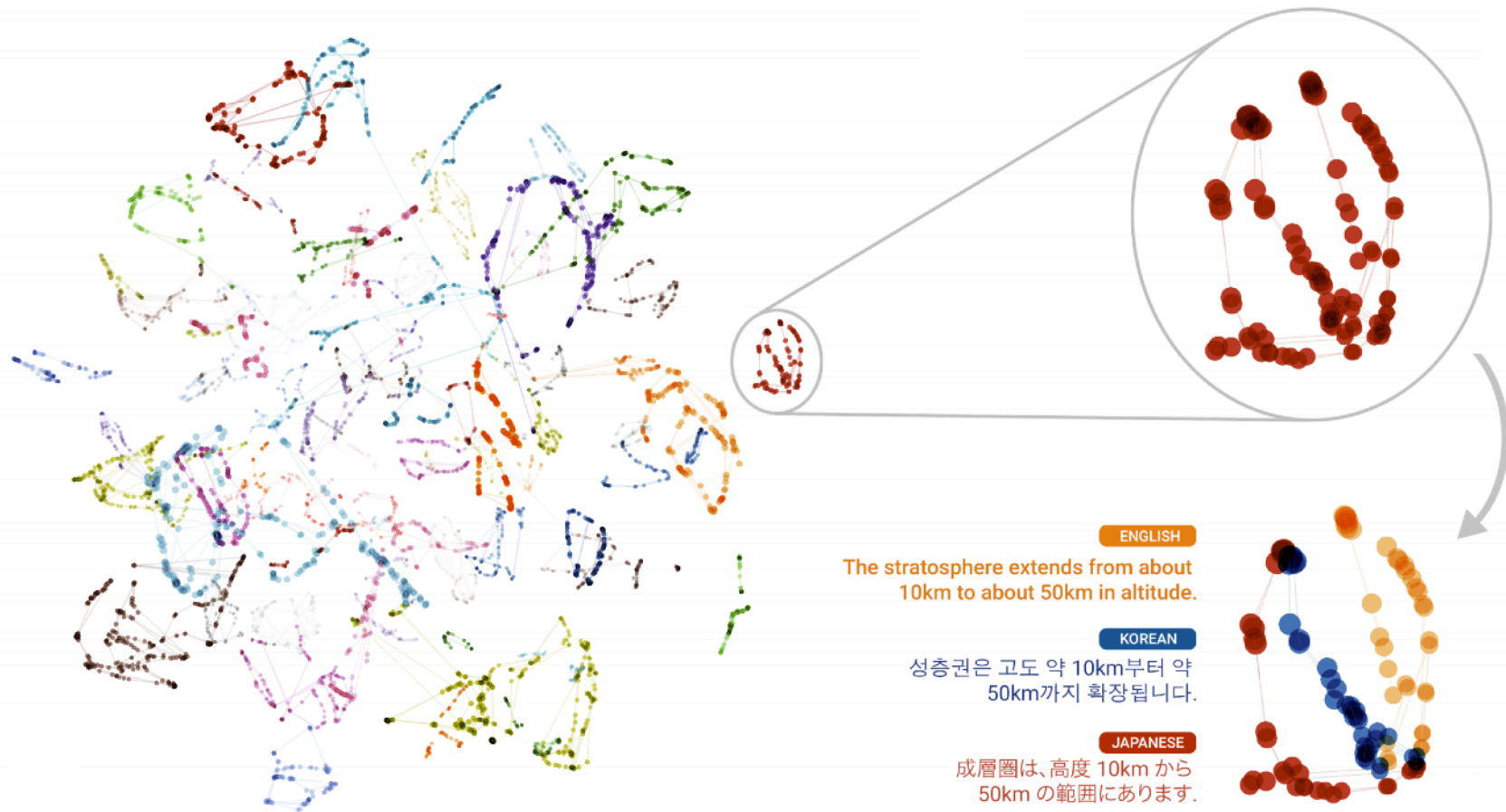


Perplexity: 50
Step: 5,000



Perplexity: 100
Step: 5,000

Multi-Lingual Word Embedding [Google 2016]



UMAP

[McInnes et al. 2018]

Form weighted nearest neighbor graph, then layout the graph in a manner that balances embedding of local and global structure.

"Our algorithm is competitive with t-SNE for visualization quality and arguably preserves more of the global structure with superior run time performance." - McInnes et al. 2018

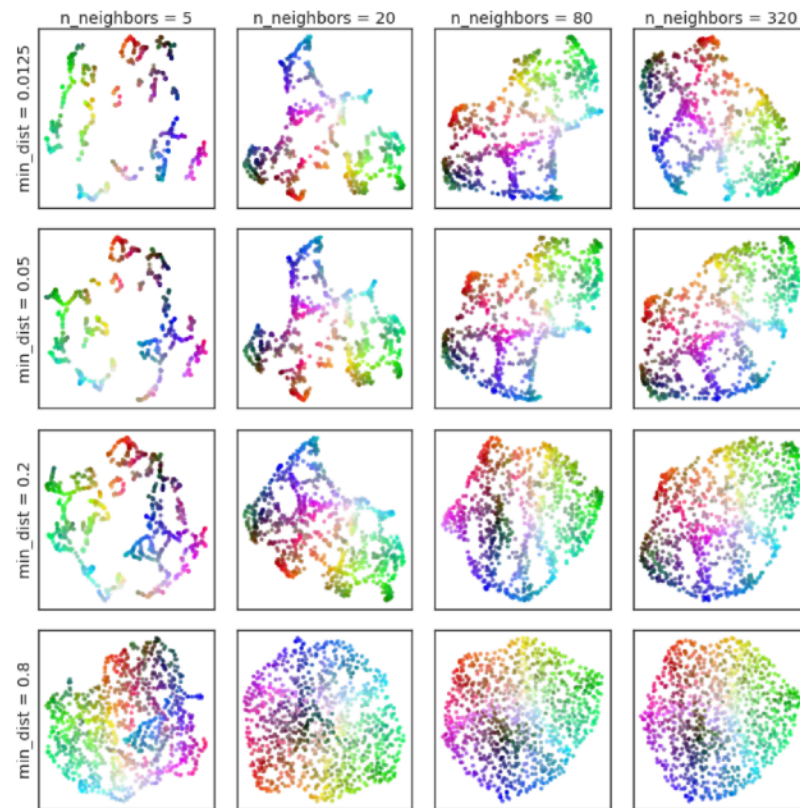


Figure 1: Variation of UMAP hyperparameters n and min-dist result in different embeddings. The data is uniform random samples from a 3-dimensional color-cube, allowing for easy visualization of the original 3-dimensional coordinates in the embedding space by using the corresponding RGB colour. Low values of n spuriously interpret structure from the random sampling noise – see Section 6 for further discussion of this phenomena.

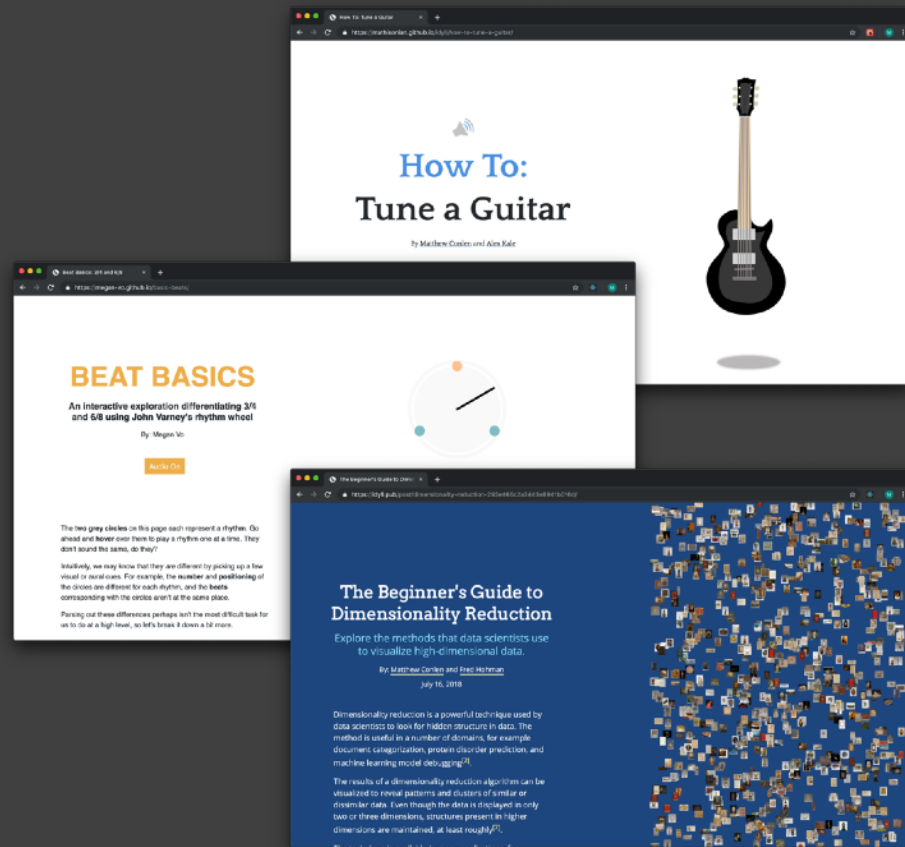
User Engagement with Interactive Articles

Provide an overview of usage patterns of interactive features.

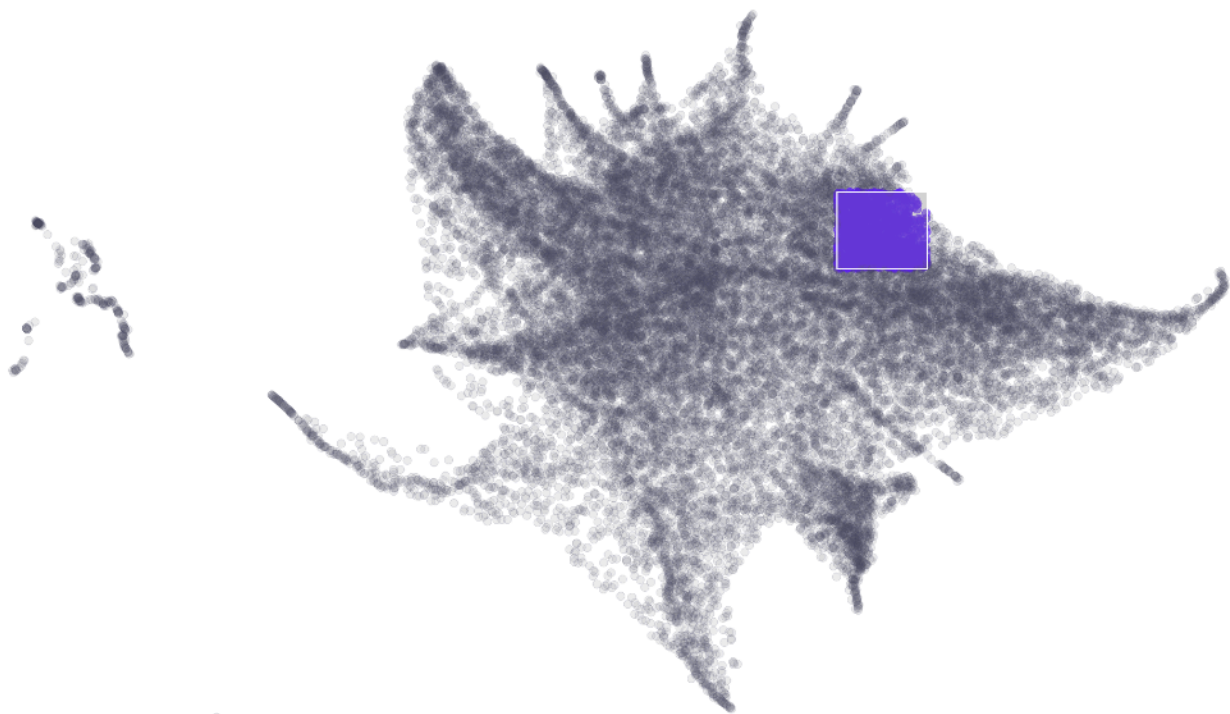
Identify variations in usage

Represent reader sessions as a feature vector with:

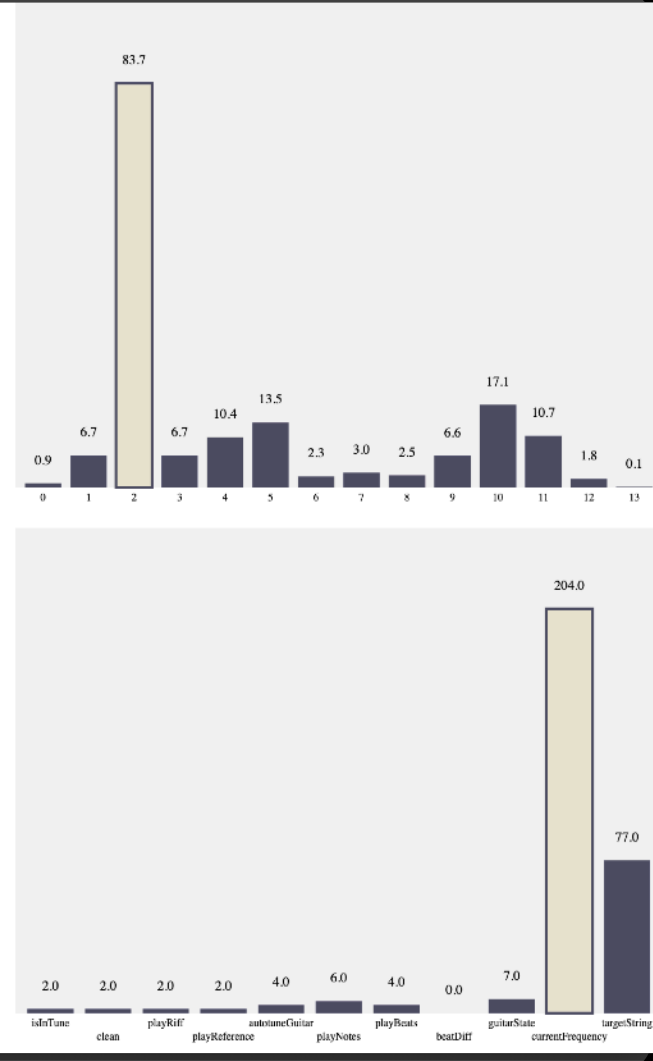
- time spent in each section
- count of variable changes



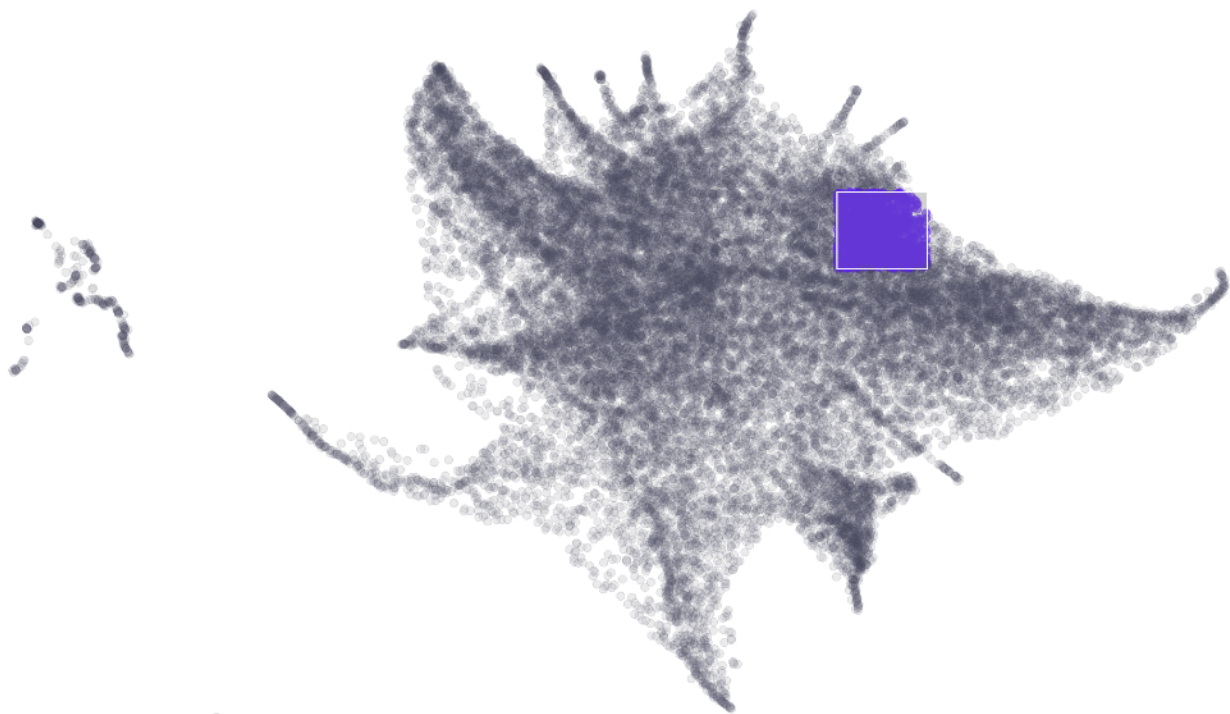
Showing 1233 users.



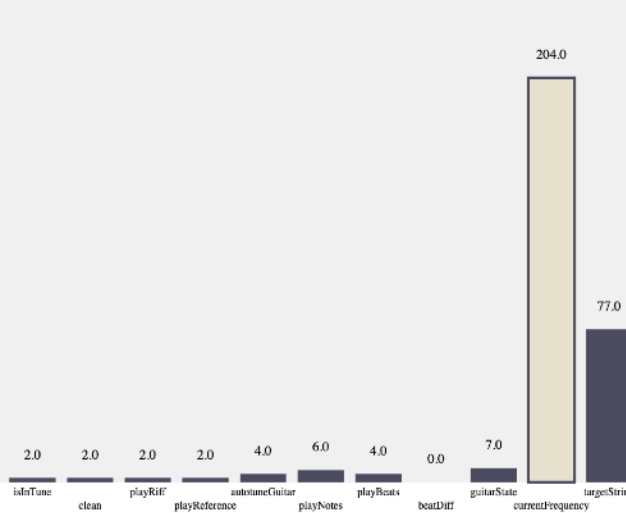
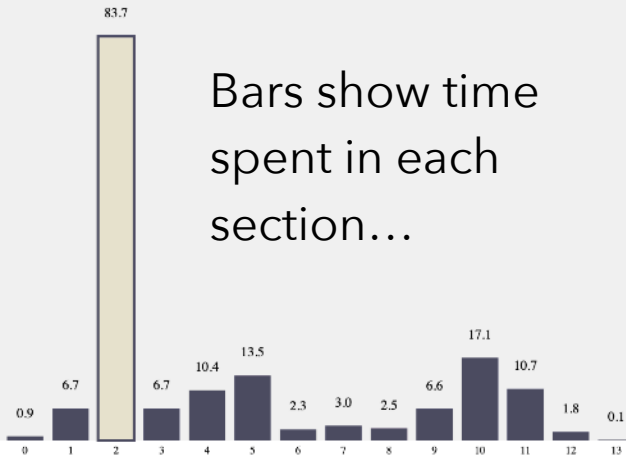
Each point represents a readers session, projected via UMAP.



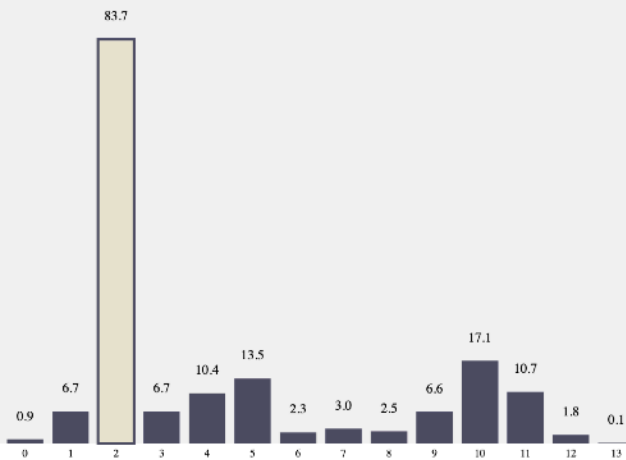
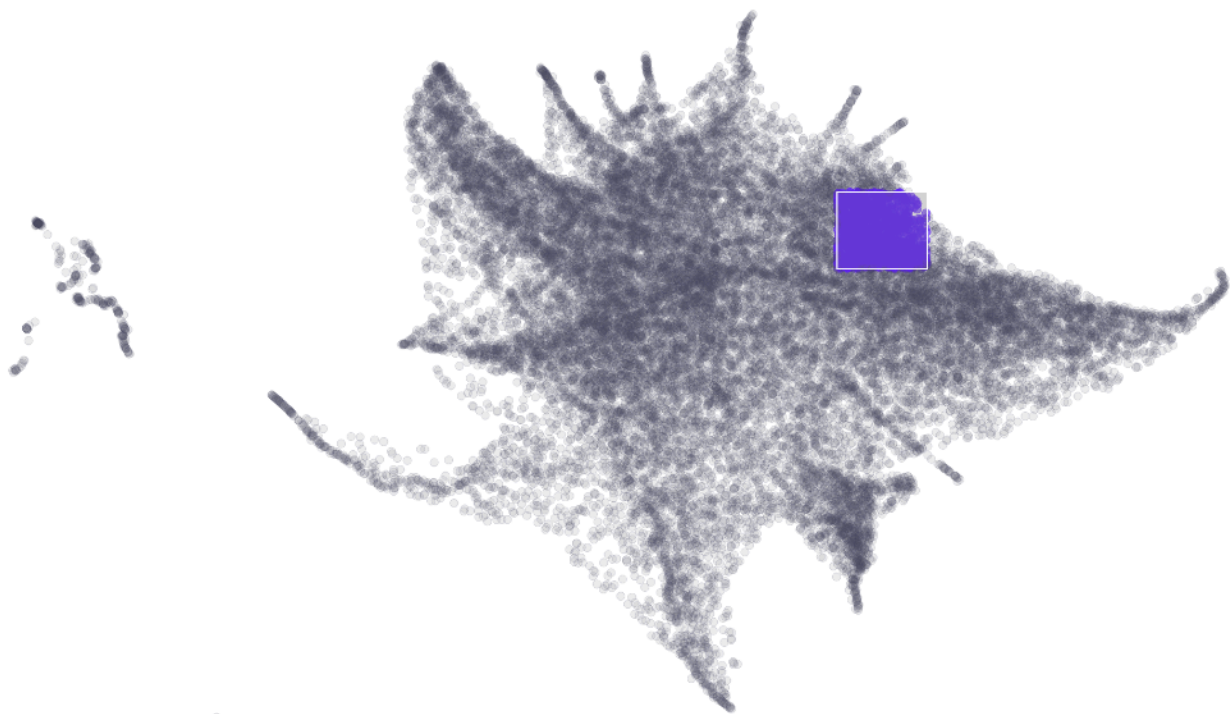
Showing 1233 users.



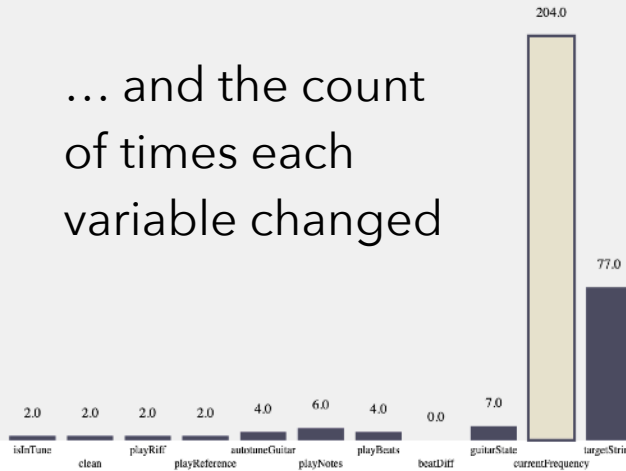
Bars show time spent in each section...



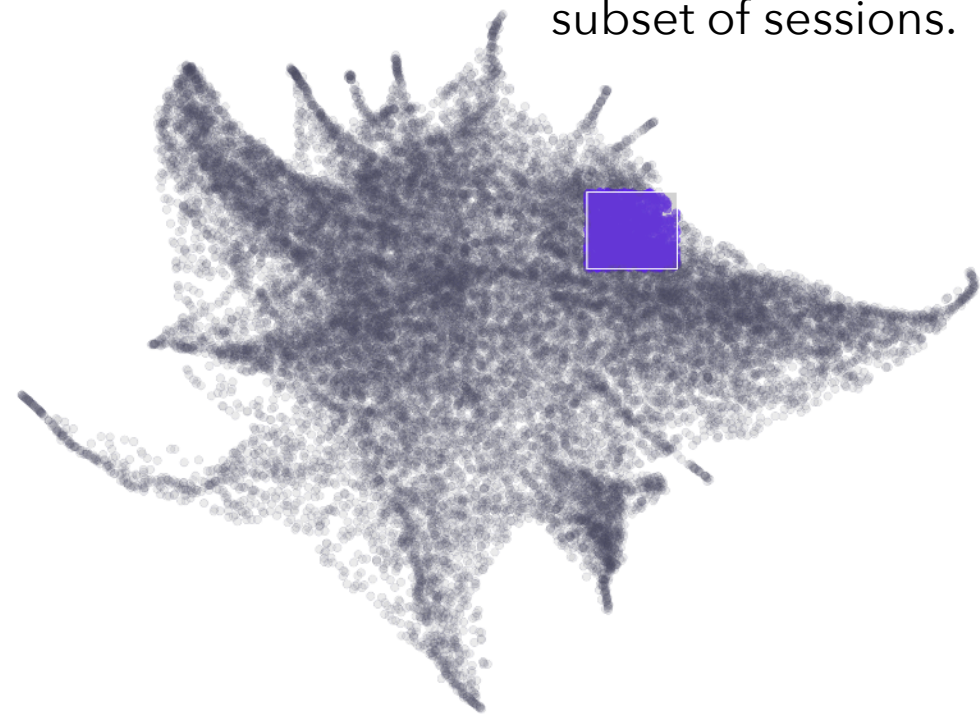
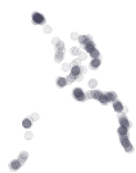
Showing 1233 users.



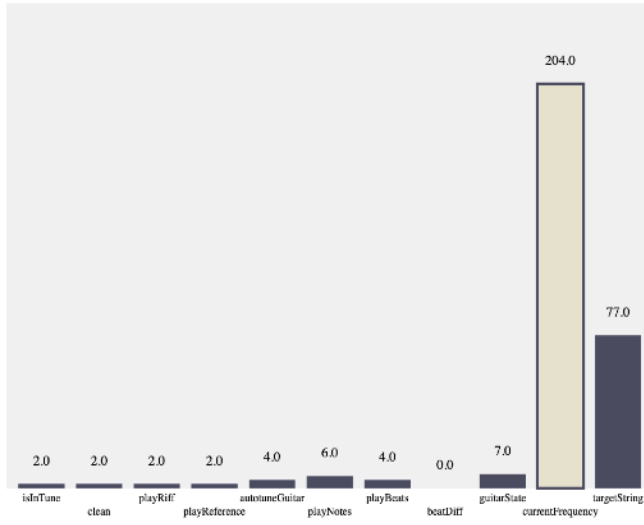
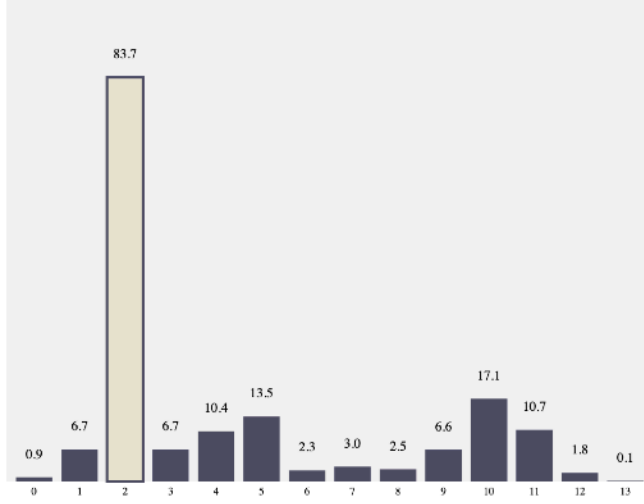
... and the count of times each variable changed

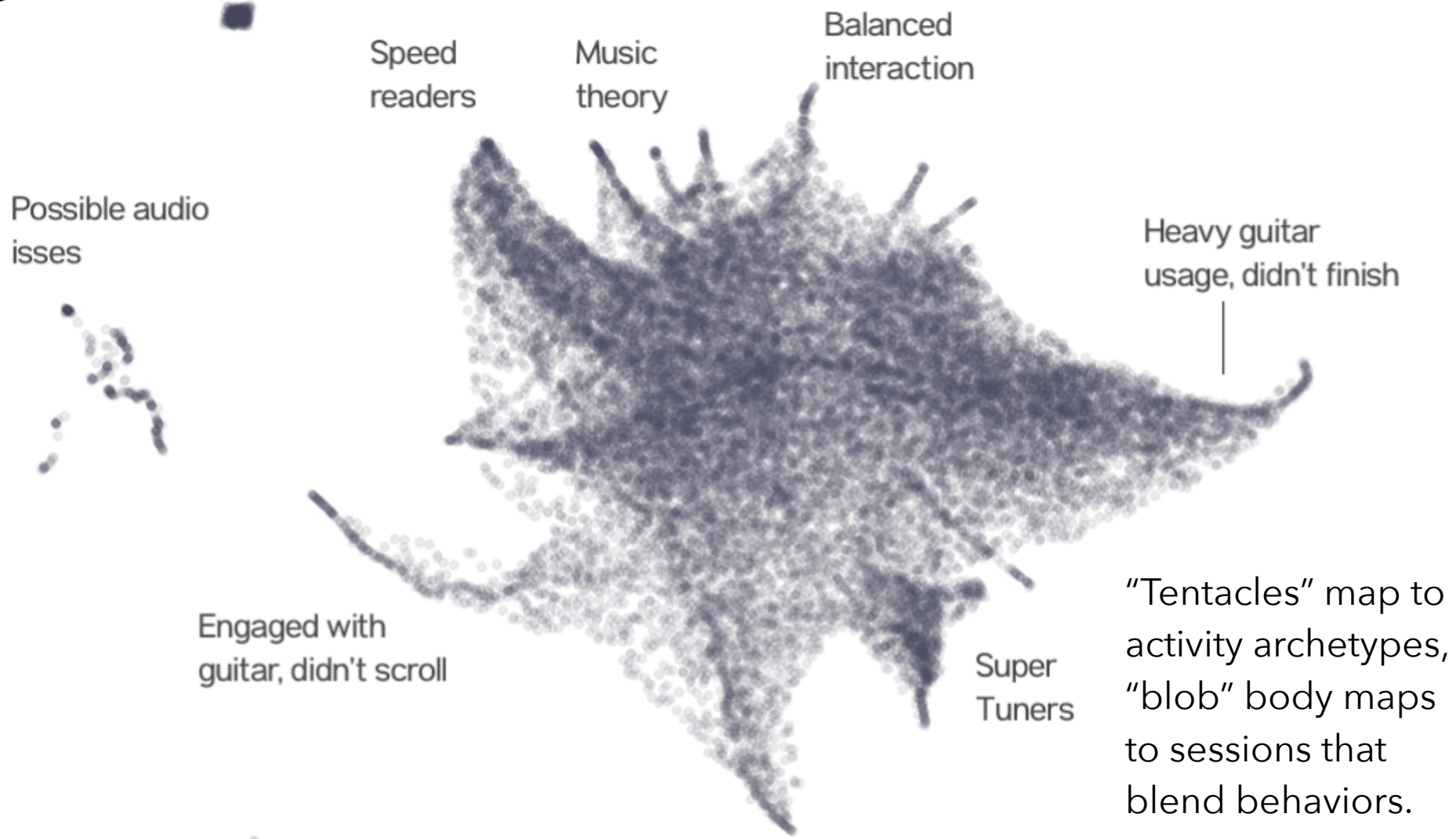


Showing 1233 users.



Brushing to select a subset of sessions.





Mapping Emoji Images



Q Latent Dimensions: 32 ▾ Projection: t-SNE ▾ Perplexity: 30 ▾

t-SNE



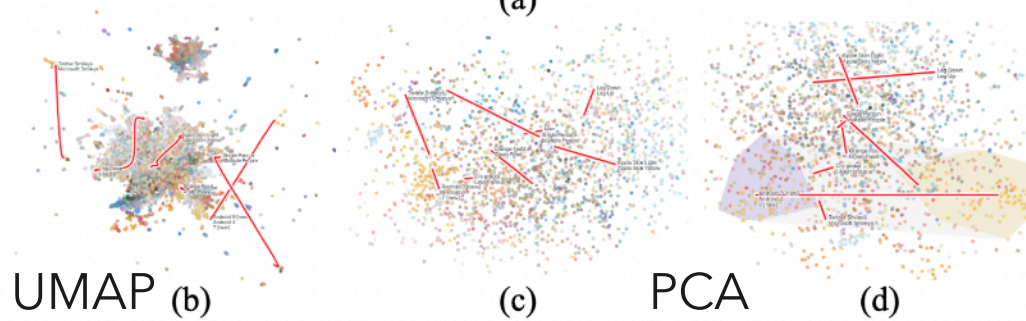
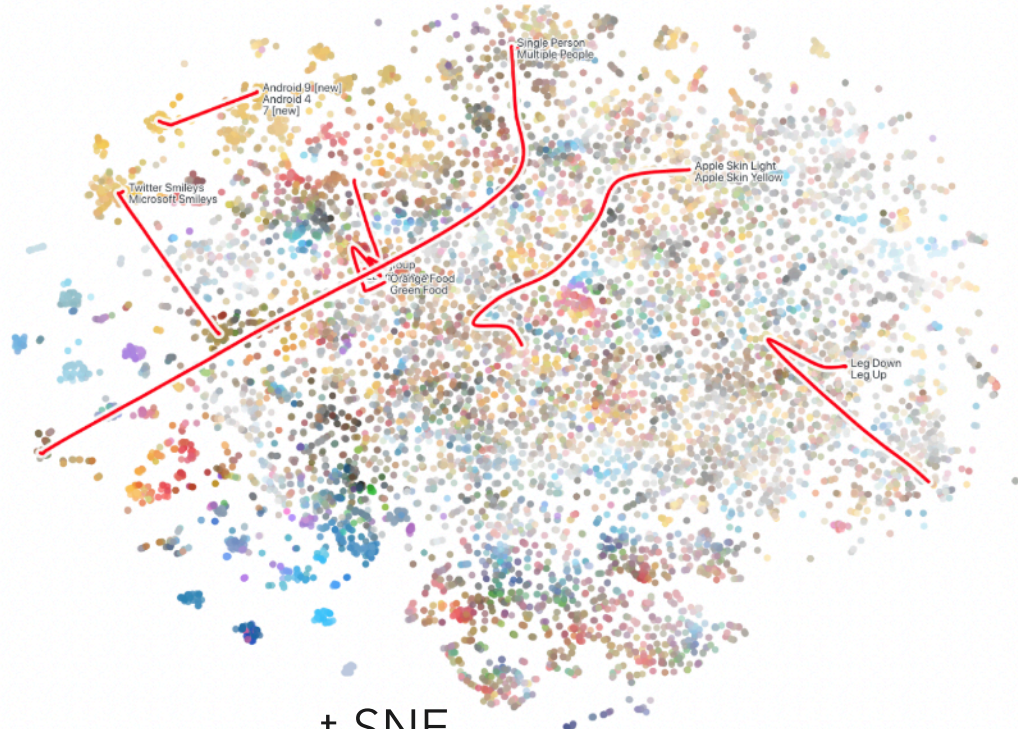
Q Latent Dimensions: 32 ▾ Projection: UMAP ▾ Neighbors: 15 ▾ Distance: 0.1 ▾

UMAP



Q Latent Dimensions: 32 ▾ Projection: PCA ▾ X-Axis: PC1 ▾ Y-Axis: PC2 ▾

PCA



Dimensionality Reduction Issues

Reproducible?

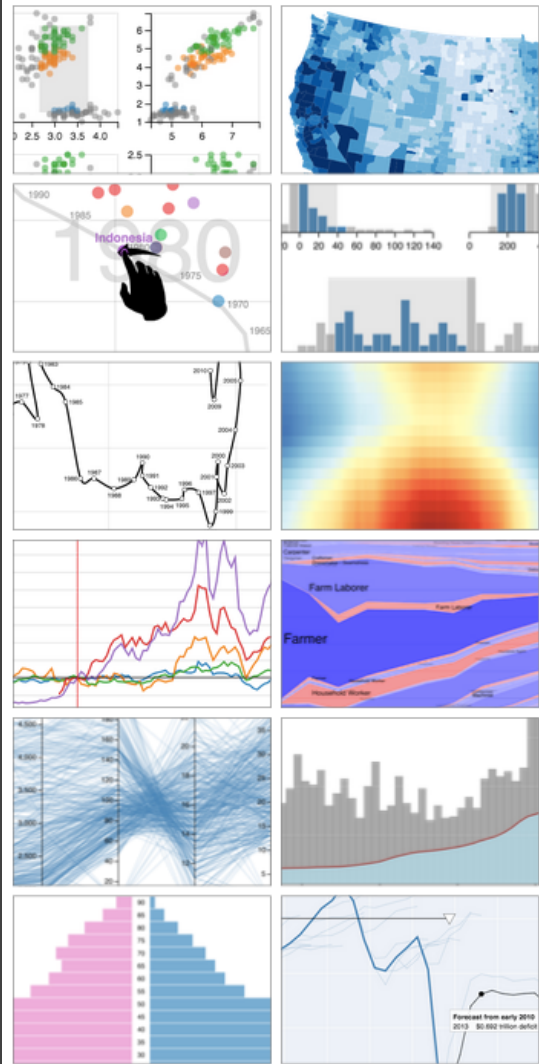
Projections are *data-dependent*. Fitting a new projection with different data can give rise to different results.

Reusable?

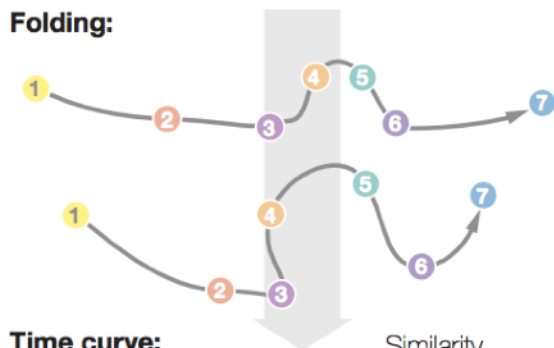
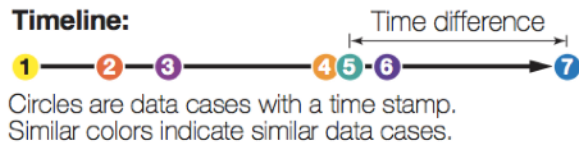
PCA and UMAP provide reusable projection functions that can map new points from high-D to low-D. t-SNE (and others, like MDS) do not provide this.

Interpretable?

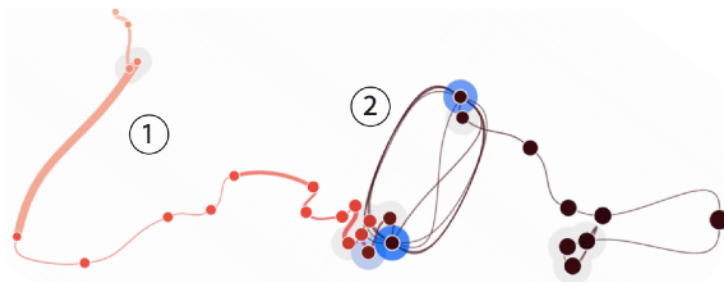
DR plots are hard to interpret! Try multiple methods and hyperparameter settings. Inspect via interaction!



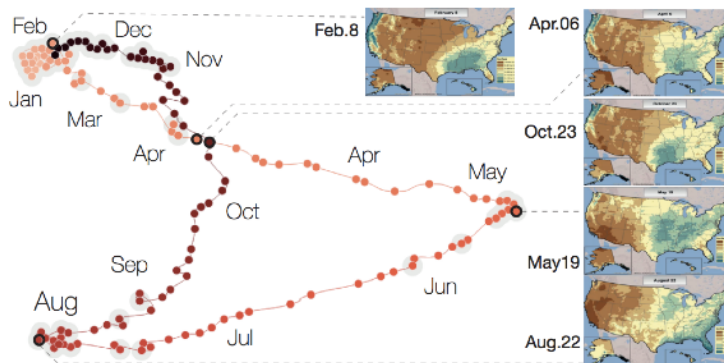
Time Curves [Bach et al. '16]



(a) Folding time



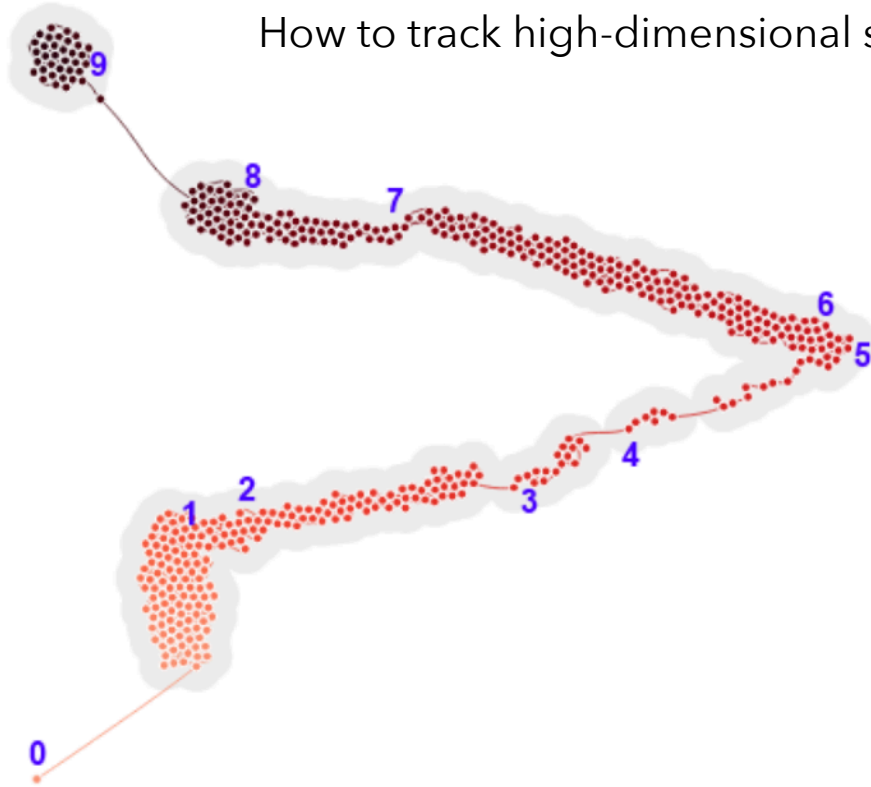
Wikipedia "Chocolate" Article



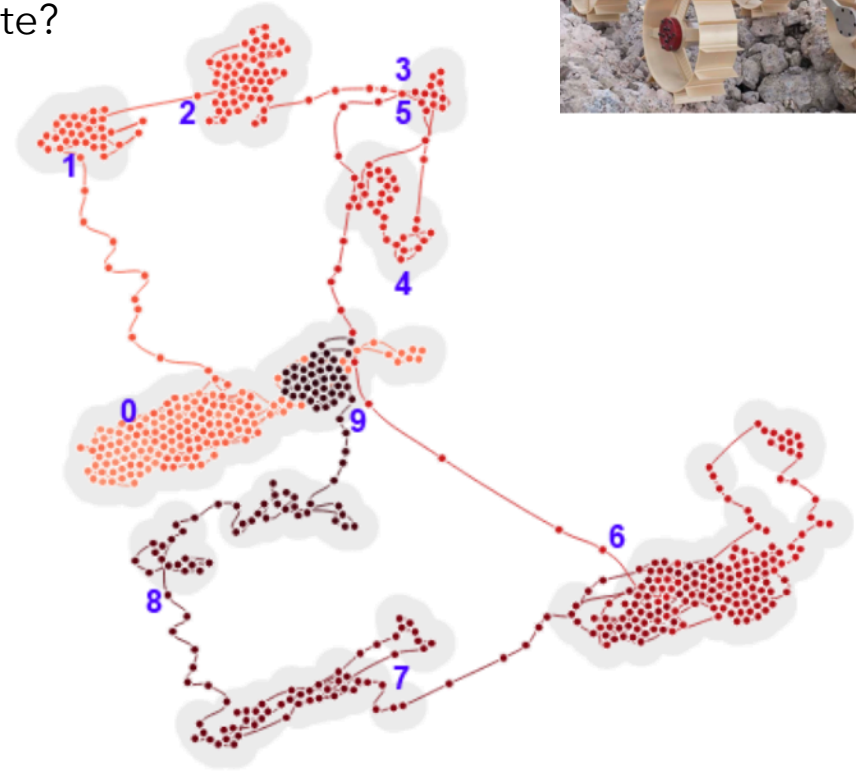
U.S. Precipitation over 1 Year

Rover Telemetry [Guy '16]

How to track high-dimensional state?



Using Raw Multi-D Data

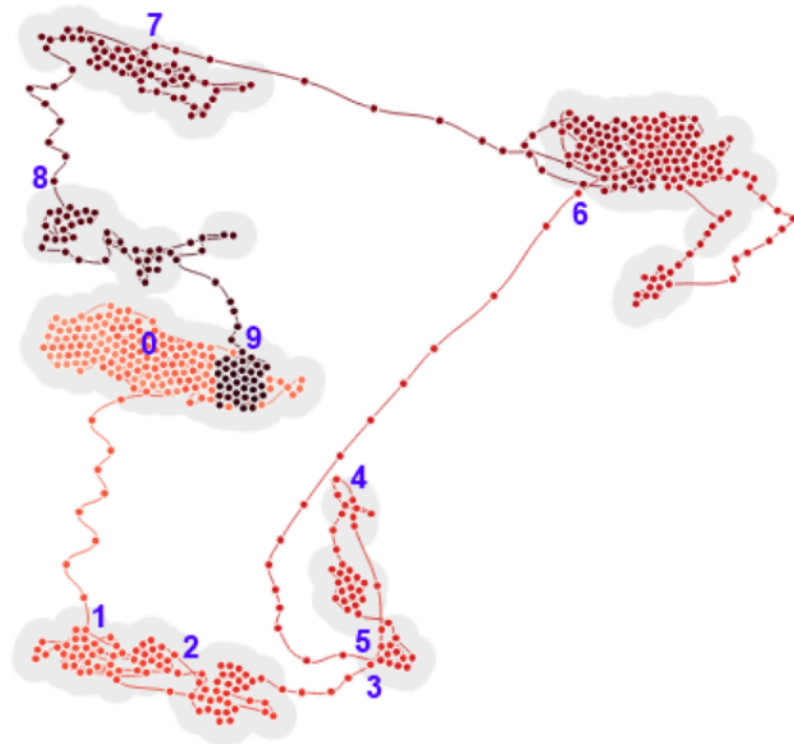


Using Pearson Correlation Matrix

Rover Telemetry [Guy '16]

How to track high-dimensional state?

Event Number	Timestamp	Event Description
0	0s	Rover begins traveling forward along smooth terrain.
1	188s	Rover begins descending into crater.
2	223s	Rover loses line of sight with lander and packet drops begin.
3	287s	Rover enters shade, causing temp, comms, and power drops.
4	300s	Rover begins traversing smooth bottom of crater.
5	330s	Rover begins climbing out of crater.
6	343s	Rover exits shade; continues uphill.
7	534s	Rover emerges from crater and enters smooth terrain.
8	594s	Rover enters choppy terrain.
9	643s	Rover wheel has fault; rover stops moving.



Using Spearman Correlation Matrix