Visual Tools & Methods for Data Cleaning

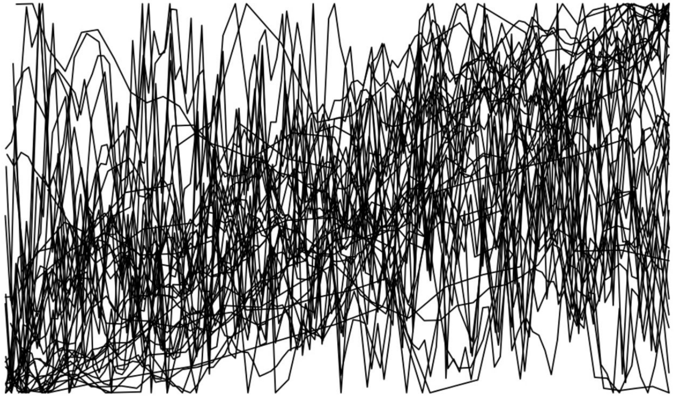
Lyon1 M2 Dataviz - 2018/2019- Cours #4

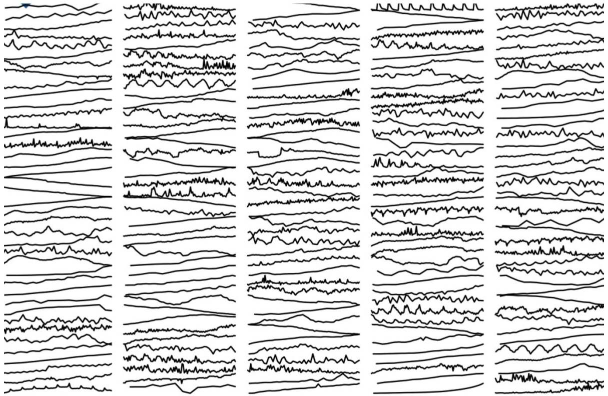
http://romain.vuillemot.net/

@romsson

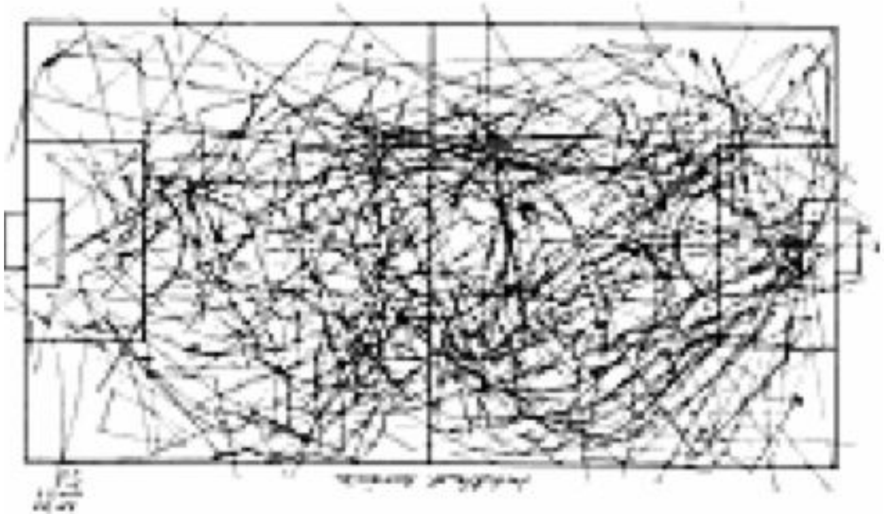
Reality

* Time series

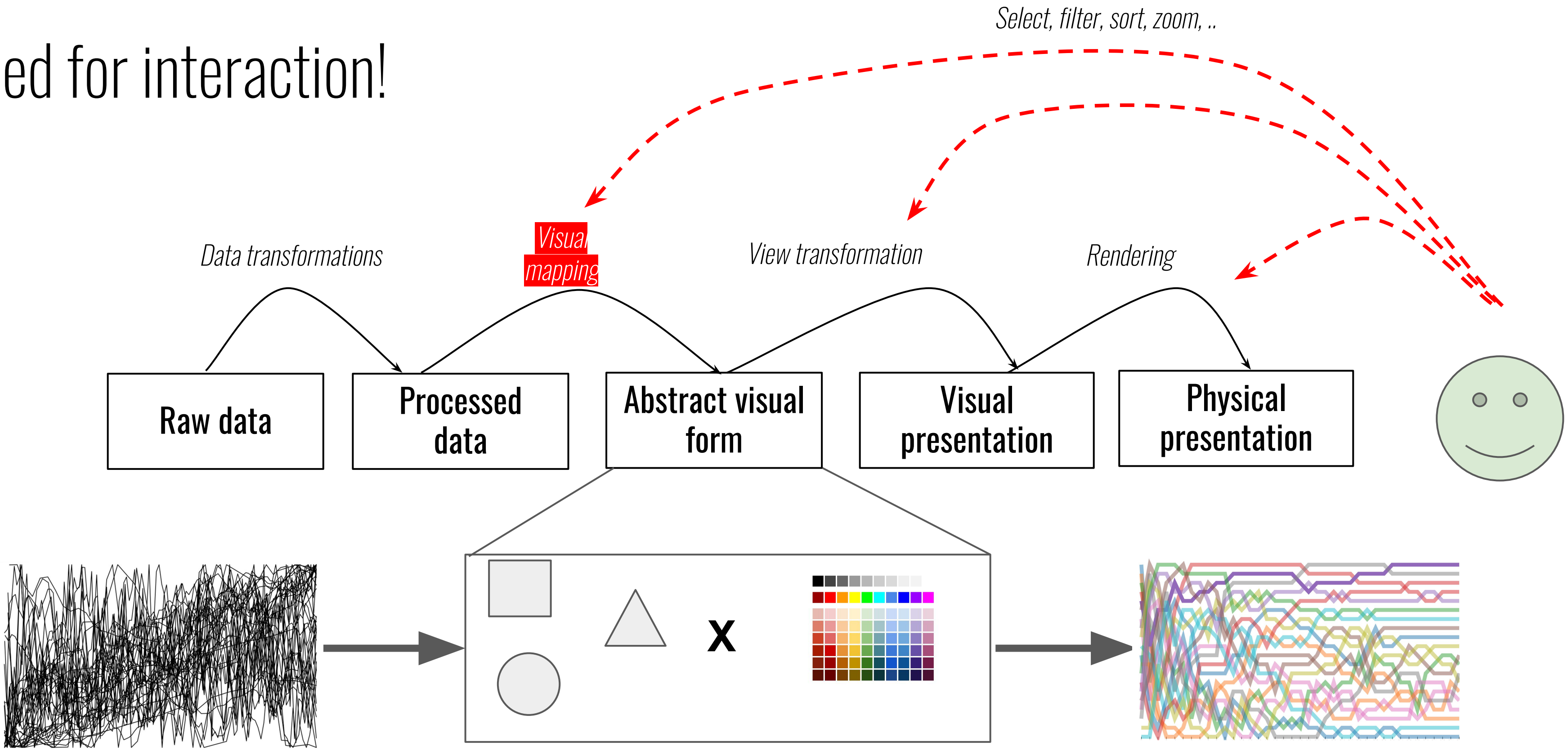




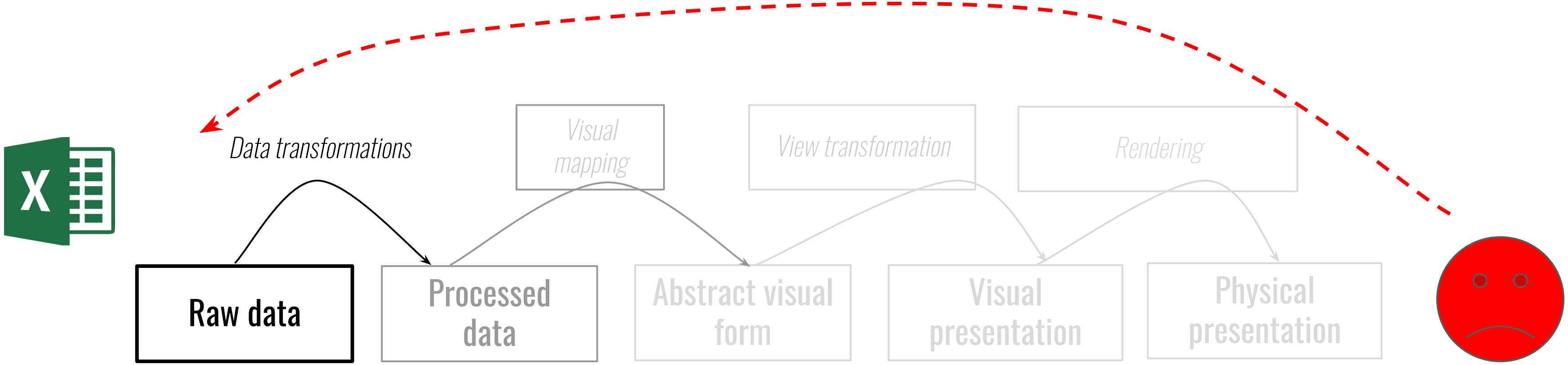
* Geo-spatial data



Need for interaction!

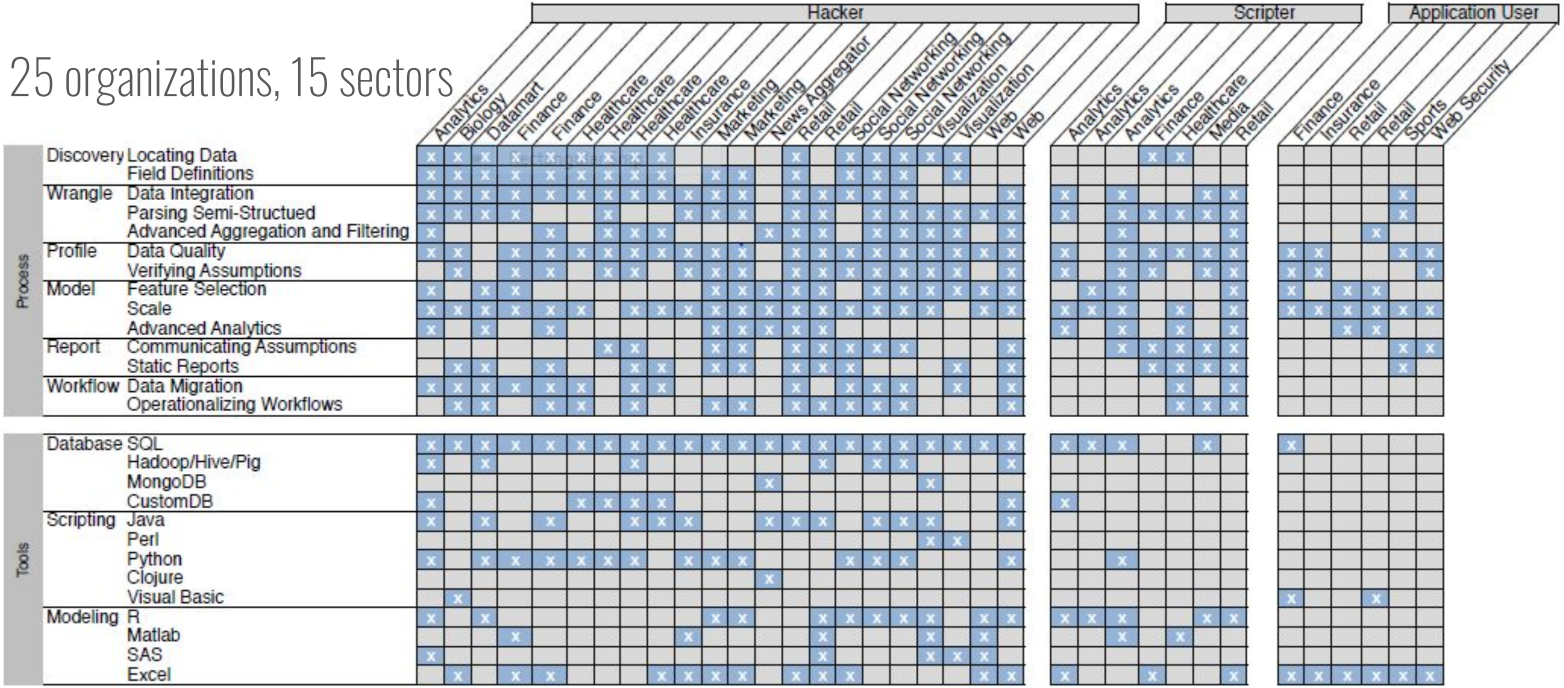


Need for interaction **with Raw Data**



Empirical study

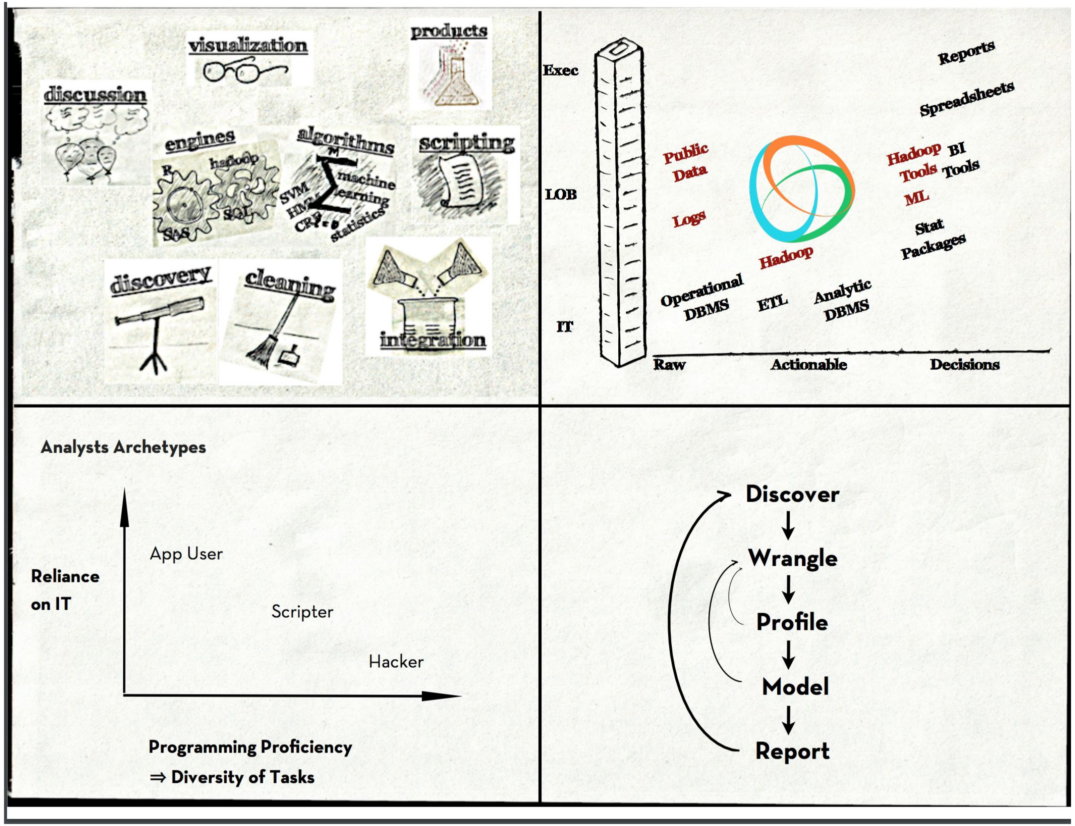
35 data analysts, 25 organizations, 15 sectors



			Hacker																						Scripter							Application User					
			Analytics	Biology	Datamart	Finance	Finance	Healthcare	Healthcare	Healthcare	Healthcare	Insurance	Marketing	Marketing	News Aggregator	Retail	Retail	Social Networking	Social Networking	Social Networking	Visualization	Visualization	Web	Web	Analytics	Analytics	Analytics	Finance	Healthcare	Media	Retail	Finance	Insurance	Retail	Retail	Sports	Web Security
Process	Discovery	Locating Data	x	x	x	x	x	x	x	x	x					x		x	x	x	x	x						x	x								
		Field Definitions	x	x	x	x	x	x	x	x	x		x	x		x		x	x	x		x															
	Wrangle	Data Integration	x	x	x	x	x	x	x	x	x	x	x	x		x	x	x	x	x				x	x		x			x	x					x	
		Parsing Semi-Structued	x	x	x	x			x			x	x	x		x	x		x	x	x	x	x	x	x		x	x	x	x	x					x	
		Advanced Aggregation and Filtering	x				x		x	x	x				x	x	x		x	x	x	x		x			x				x				x		
	Profile	Data Quality	x	x		x	x	x	x	x	x	x	x	x		x	x	x	x	x	x	x	x	x	x		x	x	x	x	x	x	x			x	x
		Verifying Assumptions		x		x	x		x	x		x	x	x		x	x	x	x	x	x	x		x	x		x	x		x	x	x	x				x
	Model	Feature Selection	x		x	x							x	x	x	x	x		x	x	x	x	x	x		x	x				x	x		x	x		
		Scale	x	x	x	x	x	x		x	x	x	x	x	x	x	x	x	x	x	x		x	x	x	x	x		x		x	x	x	x	x	x	x
		Advanced Analytics	x		x		x						x	x	x	x	x								x		x		x		x			x	x		
	Report	Communicating Assumptions							x	x			x	x		x	x	x	x	x				x			x	x	x	x	x					x	x
		Static Reports		x	x		x			x	x		x	x		x	x	x				x		x				x	x	x	x					x	
	Workflow	Data Migration	x	x	x	x	x	x		x	x					x		x	x	x		x		x					x		x						
		Operationalizing Workflows		x	x		x	x		x			x	x		x	x	x	x	x				x					x	x	x						
Tools	Database	SQL	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x			x		x					
		Hadoop/Hive/Pig	x		x					x							x		x	x				x													
		MongoDB													x						x																
		CustomDB	x					x	x	x	x													x	x												
	Scripting	Java	x		x		x			x	x	x			x	x	x		x	x	x			x													
		Perl																			x	x															
		Python	x		x	x	x	x	x	x		x	x	x				x	x	x				x			x										
		Clojure													x																						
		Visual Basic		x																												x			x		
	Modeling	R	x		x								x	x			x	x	x	x	x		x	x	x	x	x			x	x						
		Matlab				x						x					x				x		x				x		x								
		SAS	x														x				x	x	x														
		Excel		x		x	x				x	x	x	x		x	x	x					x	x	x			x			x	x	x	x	x	x	x

Kandel, Sean, et al. "Enterprise data analysis and visualization: An interview study." IEEE Transactions on Visualization and Computer Graphics 18.12 (2012): 2917-2926. (pdf)

Empirical study

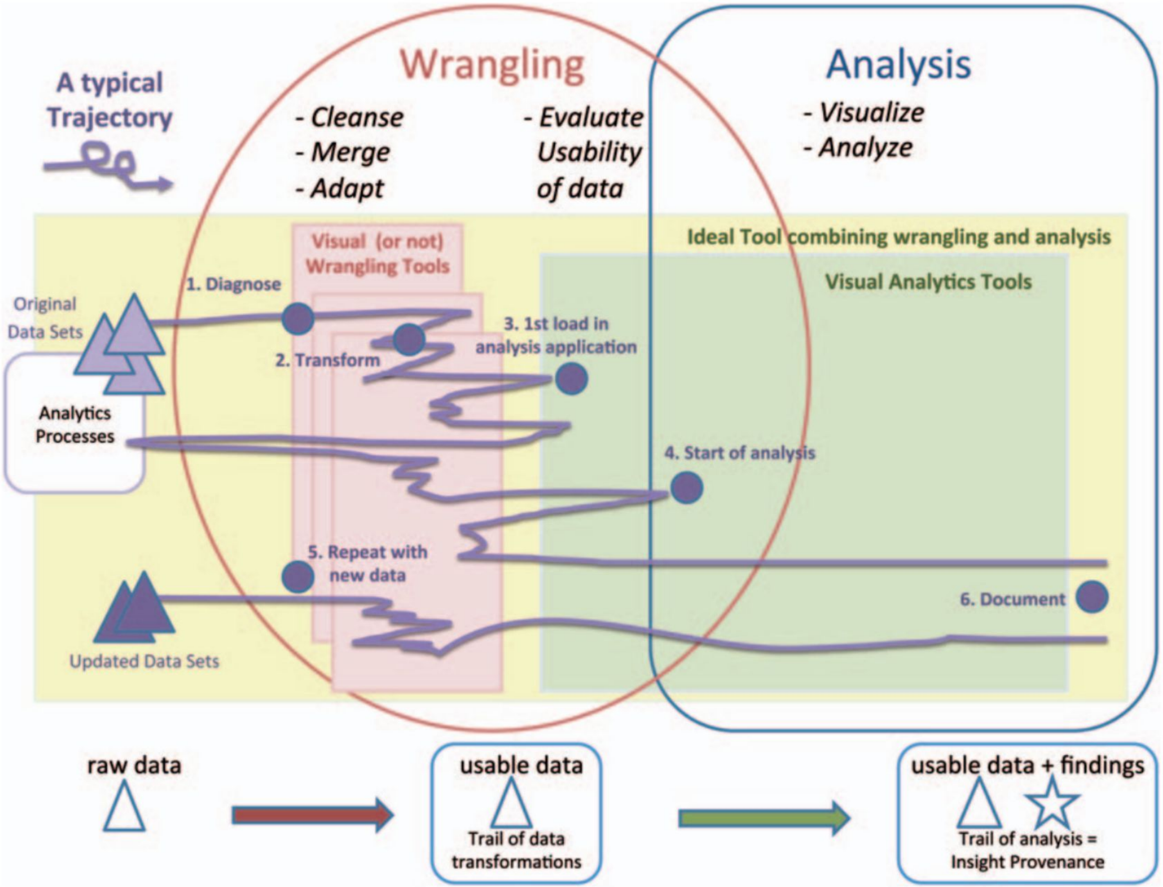


Joe Hellerstein “Data wrangling” BERKELEY & Trifacta (pdf)

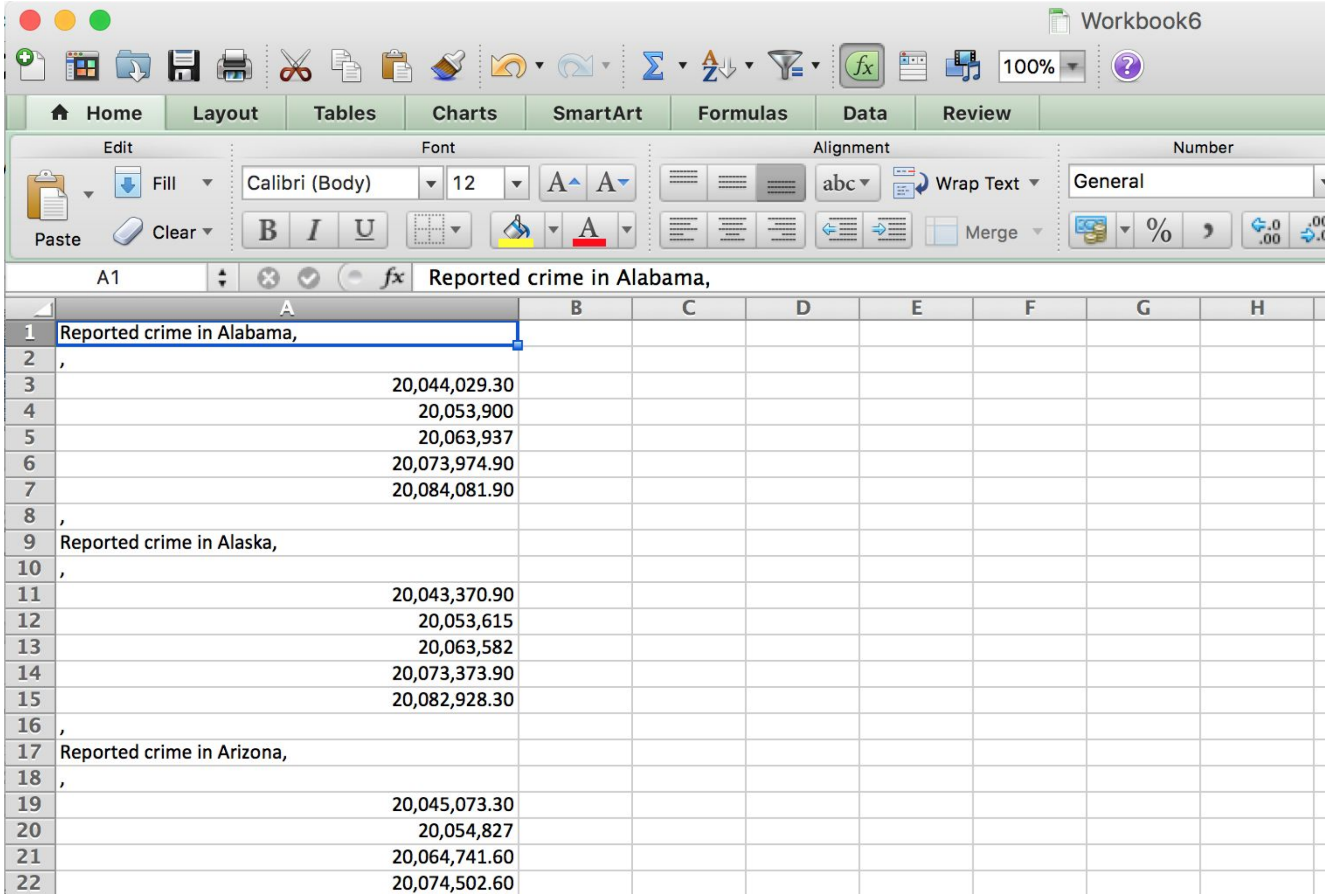
Wrangling and analysis process

* Iterative, non-linear process

Kandel, S., Heer, J., Plaisant, C., Kennedy, J., van Ham, F., Riche, N. H., & Buono, P. (2011). “Research directions in data wrangling: Visualizations and transformations for usable and credible data. Information Visualization”



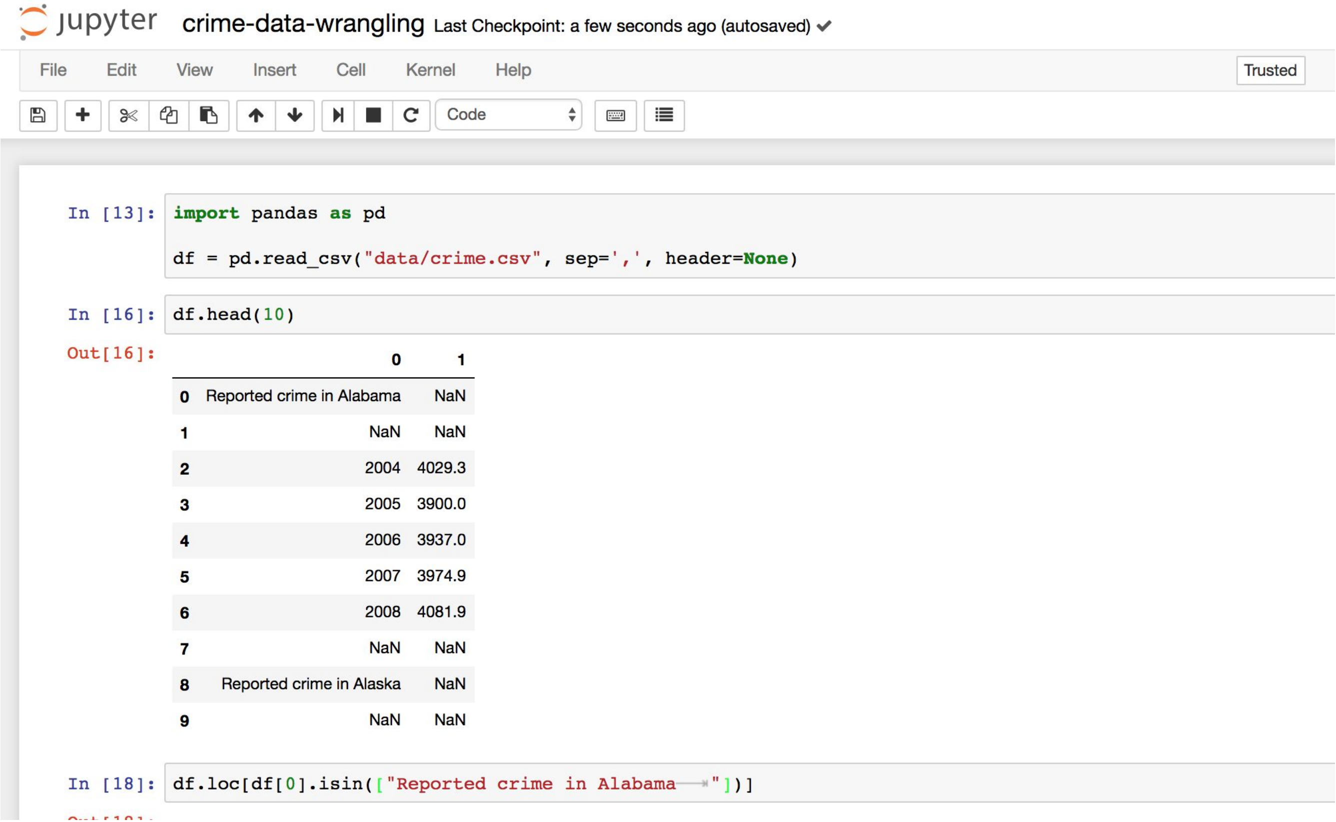
Microsoft Excel



A1 Reported crime in Alabama,

	A	B	C	D	E	F	G	H
1	Reported crime in Alabama,							
2	,							
3	20,044,029.30							
4	20,053,900							
5	20,063,937							
6	20,073,974.90							
7	20,084,081.90							
8	,							
9	Reported crime in Alaska,							
10	,							
11	20,043,370.90							
12	20,053,615							
13	20,063,582							
14	20,073,373.90							
15	20,082,928.30							
16	,							
17	Reported crime in Arizona,							
18	,							
19	20,045,073.30							
20	20,054,827							
21	20,064,741.60							
22	20,074,502.60							

Python Notebook


jupyter
crime-data-wrangling Last Checkpoint: a few seconds ago (autosaved) ✓
File
Edit
View
Insert
Cell
Kernel
Help
Trusted
Code

```
In [13]: import pandas as pd

         df = pd.read_csv("data/crime.csv", sep=',', header=None)
```

```
In [16]: df.head(10)
```

Out[16]:

	0	1
0	Reported crime in Alabama	NaN
1	NaN	NaN
2	2004	4029.3
3	2005	3900.0
4	2006	3937.0
5	2007	3974.9
6	2008	4081.9
7	NaN	NaN
8	Reported crime in Alaska	NaN
9	NaN	NaN

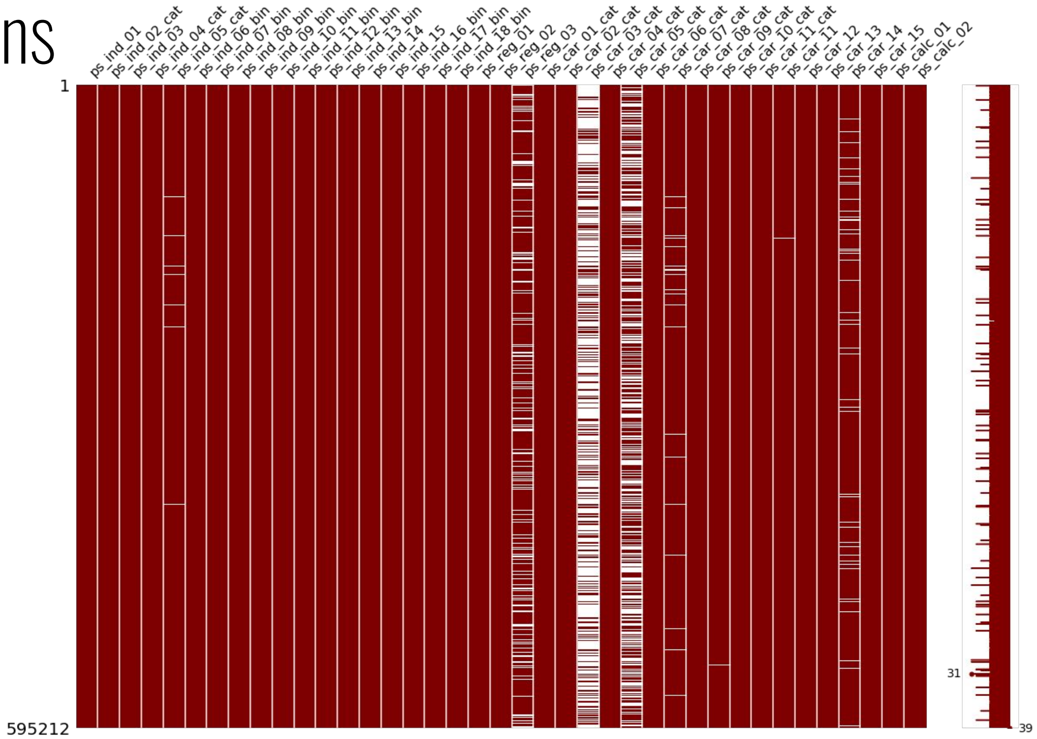
```
In [18]: df.loc[df[0].isin(["Reported crime in Alabama→"])]
```

Low-level scripts & visualizations

- Python / Perl / ..

- Pipeline / Batch process

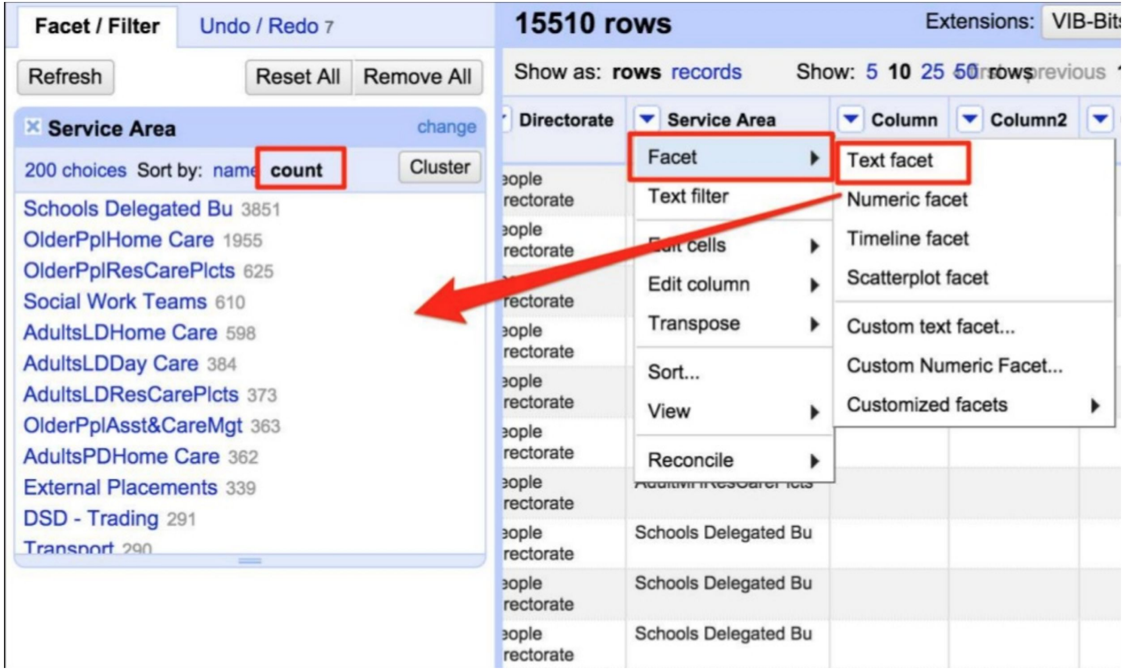
- …



Example:
SafeDriver - data cleaning & visualization (webpage)

Google / Open Refine (2010 - ...)

* Loading
* Checking
* Exploring
* Cleaning
* Reshaping
* Annotating
* Saving



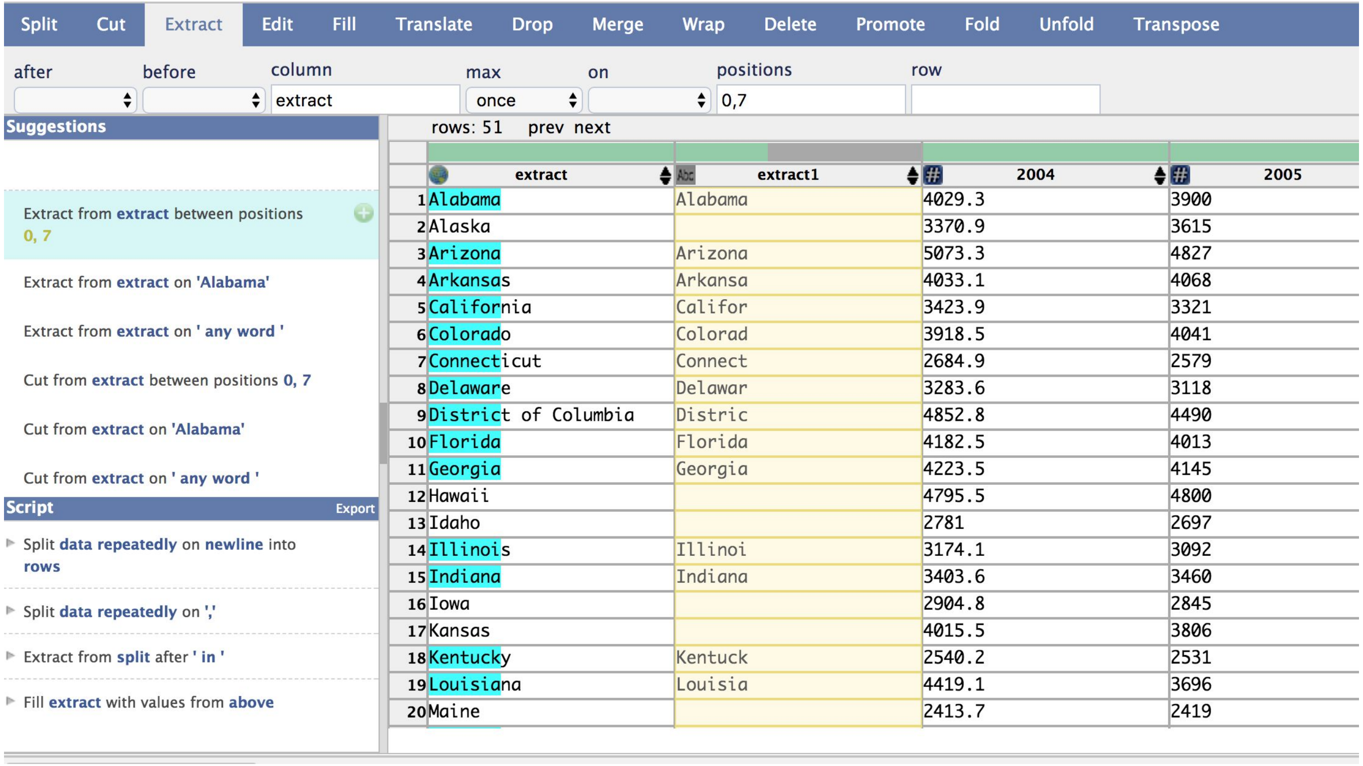
Schools Delegated Bu	3851
OlderPplHome Care	1955
OlderPplResCarePlcts	625
Social Work Teams	610
AdultsLDHome Care	598
AdultsLDDay Care	384
AdultsLDResCarePlcts	373
OlderPplAsst&CareMgt	363
AdultsPDHome Care	362
External Placements	339
DSD - Trading	291
Transport	290

Directorate	Service Area
People Directorate	
People Directorate	
People Directorate	
People Directorate	
People Directorate	
People Directorate	
People Directorate	AdultMHResCarePlcts
People Directorate	Schools Delegated Bu
People Directorate	Schools Delegated Bu
People Directorate	Schools Delegated Bu

https://github.com/OpenRefine/OpenRefine

A Quick Tour of OpenRefine (slides)

Wrangler



after:
before:
column: extract
max: once
on:
positions: 0,7
row:

Suggestions

Extract from **extract** between positions **0, 7**

Extract from **extract** on **'Alabama'**

Extract from **extract** on **' any word '**

Cut from **extract** between positions **0, 7**

Cut from **extract** on **'Alabama'**

Cut from **extract** on **' any word '**

Script Export

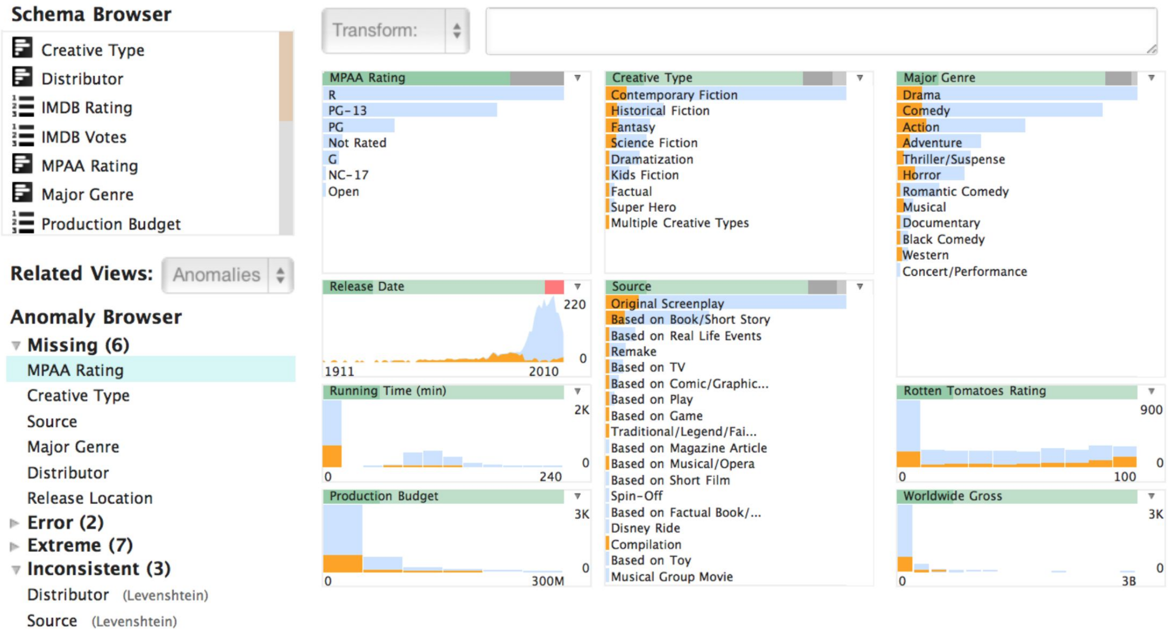
- Split **data repeatedly** on **newline** into **rows**
- Split **data repeatedly** on **','**
- Extract from **split** after **' in '**
- Fill **extract** with values from **above**

rows: 51 prev next

	extract	extract1	2004	2005
1	Alabama	Alabama	4029.3	3900
2	Alaska		3370.9	3615
3	Arizona	Arizona	5073.3	4827
4	Arkansas	Arkansa	4033.1	4068
5	California	Califor	3423.9	3321
6	Colorado	Colorad	3918.5	4041
7	Connecticut	Connect	2684.9	2579
8	Delaware	Delawar	3283.6	3118
9	District of Columbia	Distric	4852.8	4490
10	Florida	Florida	4182.5	4013
11	Georgia	Georgia	4223.5	4145
12	Hawaii		4795.5	4800
13	Idaho		2781	2697
14	Illinois	Illinoi	3174.1	3092
15	Indiana	Indiana	3403.6	3460
16	Iowa		2904.8	2845
17	Kansas		4015.5	3806
18	Kentucky	Kentuck	2540.2	2531
19	Louisiana	Louisia	4419.1	3696
20	Maine		2413.7	2419

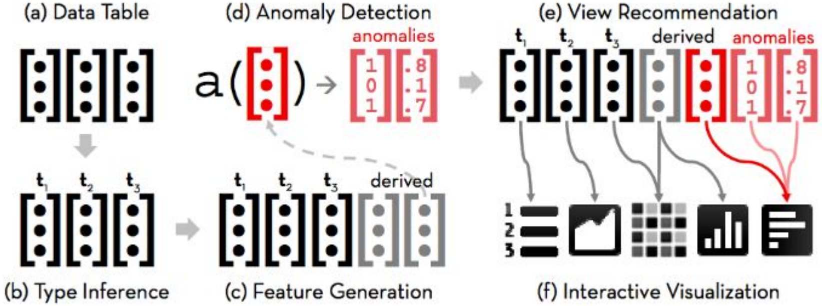
Kandel, S., Paepcke, A., Hellerstein, J., & Heer, J. (2011, May). Wrangler: Interactive visual specification of data transformation scripts. In Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (pp. 3363-3372). ACM. (demo)

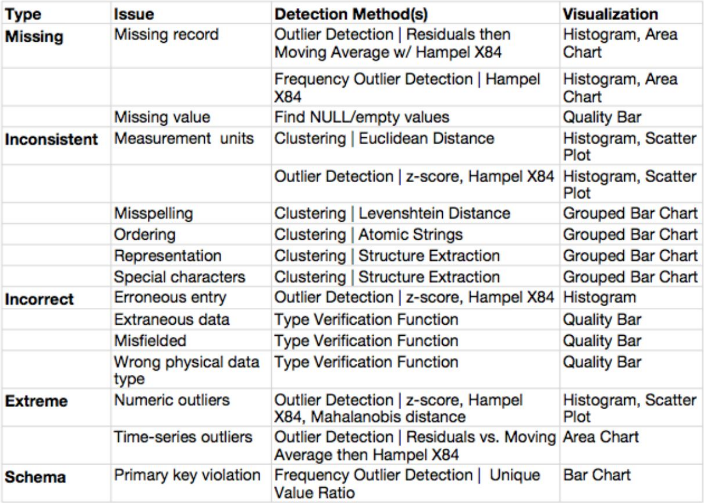
Profiler



Kandel, S., Parikh, R., Paepcke, A., Hellerstein, J. M., & Heer, J. (2012, May). Profiler: Integrated statistical analysis and visualization for data quality assessment. In Proceedings of the International Working Conference on Advanced Visual Interfaces (pp. 547-554). ACM. (pdf)

Profiler

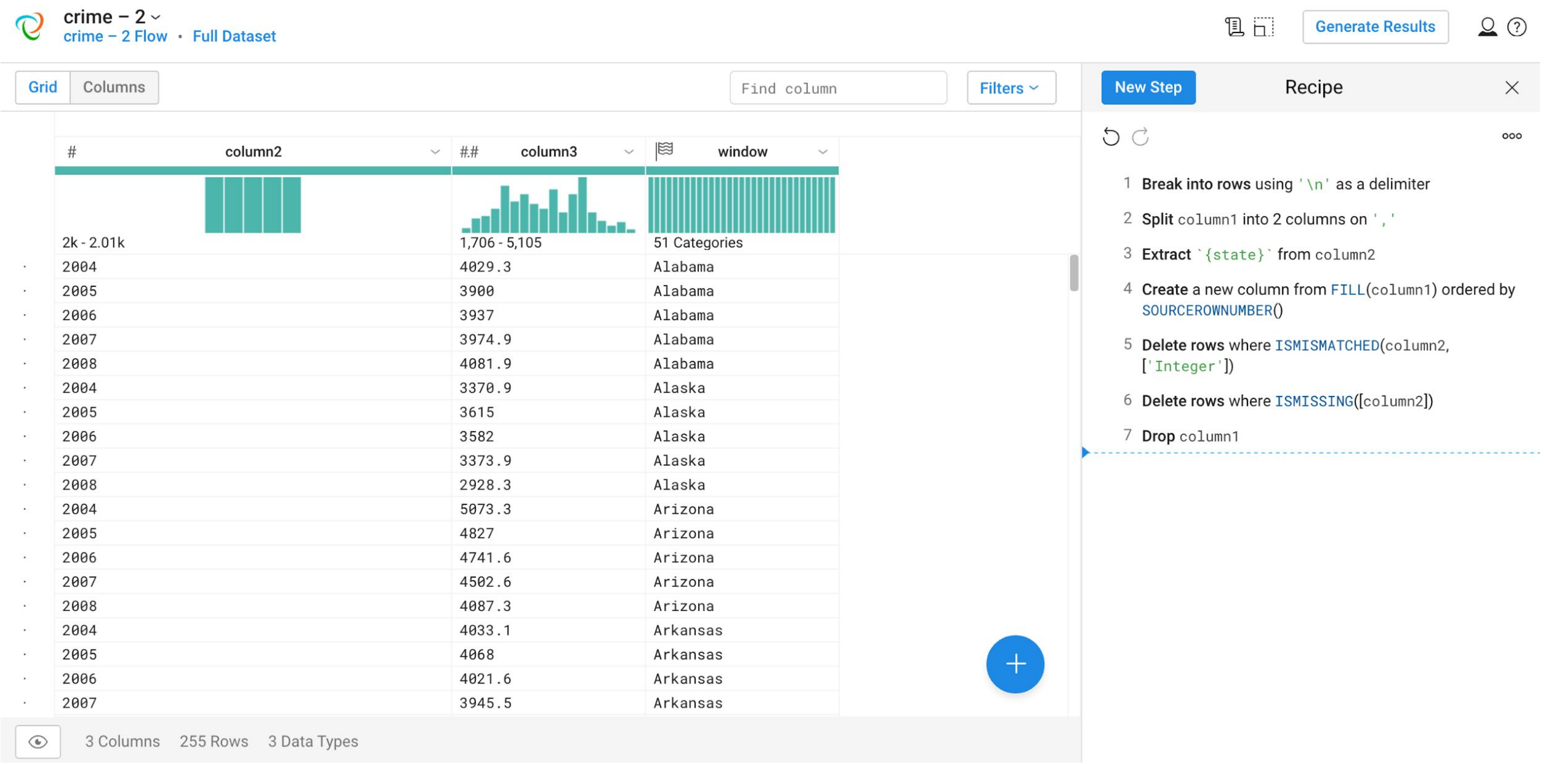




Type	**Issue**	**Detection Method(s)**	**Visualization**
Missing	Missing record	Outlier Detection | Residuals then Moving Average w/ Hampel X84	Histogram, Area Chart
		Frequency Outlier Detection | Hampel X84	Histogram, Area Chart
	Missing value	Find NULL/empty values	Quality Bar
Inconsistent	Measurement units	Clustering | Euclidean Distance	Histogram, Scatter Plot
		Outlier Detection | z-score, Hampel X84	Histogram, Scatter Plot
	Misspelling	Clustering | Levenshtein Distance	Grouped Bar Chart
	Ordering	Clustering | Atomic Strings	Grouped Bar Chart
	Representation	Clustering | Structure Extraction	Grouped Bar Chart
	Special characters	Clustering | Structure Extraction	Grouped Bar Chart
Incorrect	Erroneous entry	Outlier Detection | z-score, Hampel X84	Histogram
	Extraneous data	Type Verification Function	Quality Bar
	Misfielded	Type Verification Function	Quality Bar
	Wrong physical data type	Type Verification Function	Quality Bar
Extreme	Numeric outliers	Outlier Detection | z-score, Hampel X84, Mahalanobis distance	Histogram, Scatter Plot
	Time-series outliers	Outlier Detection | Residuals vs. Moving Average then Hampel X84	Area Chart
Schema	Primary key violation	Frequency Outlier Detection | Unique Value Ratio	Bar Chart

Kandel, S., Parikh, R., Paepcke, A., Hellerstein, J. M., & Heer, J. (2012, May). Profiler: Integrated statistical analysis and visualization for data quality assessment. In Proceedings of the International Working Conference on Advanced Visual Interfaces (pp. 547-554). ACM. (pdf)

Trifacta



# column2	## column3	window
2k - 2.01k	1,706 - 5,105	51 Categories
2004	4029.3	Alabama
2005	3900	Alabama
2006	3937	Alabama
2007	3974.9	Alabama
2008	4081.9	Alabama
2004	3370.9	Alaska
2005	3615	Alaska
2006	3582	Alaska
2007	3373.9	Alaska
2008	2928.3	Alaska
2004	5073.3	Arizona
2005	4827	Arizona
2006	4741.6	Arizona
2007	4502.6	Arizona
2008	4087.3	Arizona
2004	4033.1	Arkansas
2005	4068	Arkansas
2006	4021.6	Arkansas
2007	3945.5	Arkansas

1 **Break into rows** using `'\n'` as a delimiter

2 **Split** `column1` into 2 columns on `','`

3 **Extract** `` `{state}` `` from `column2`

4 **Create** a new column from `FILL(column1)` ordered by `SOURCEROWNUMBER()`

5 **Delete rows** where `ISMISMATCHED(column2, ['Integer'])`

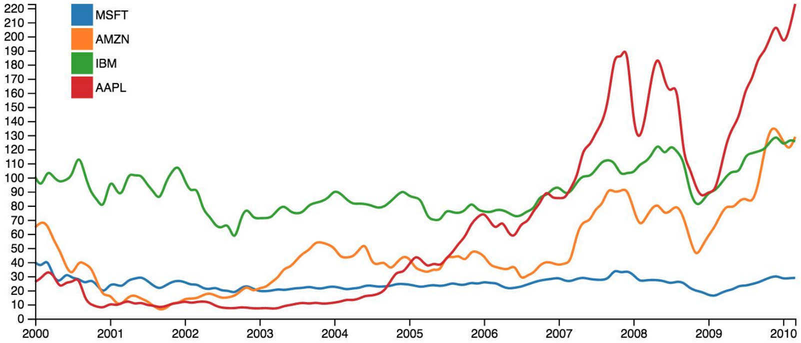
6 **Delete rows** where `ISMISSING([column2])`

7 **Drop** `column1`

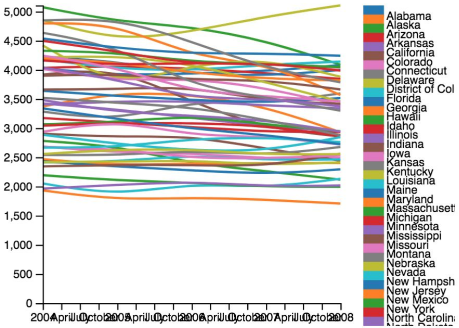
Trifacta https://www.trifacta.com/

Visualization!

```
"year","value","state"
"2004","4029.3","Alabama"
"2005","3900","Alabama"
"2006","3937","Alabama"
"2007","3974.9","Alabama"
"2008","4081.9","Alabama"
"2004","3370.9","Alaska"
"2005","3615","Alaska"
"2006","3582","Alaska"
"2007","3373.9","Alaska"
"2008","2928.3","Alaska"
"2004","5073.3","Arizona"
"2005","4827","Arizona"
"2006","4741.6","Arizona"
"2007","4502.6","Arizona"
"2008","4087.3","Arizona"
```



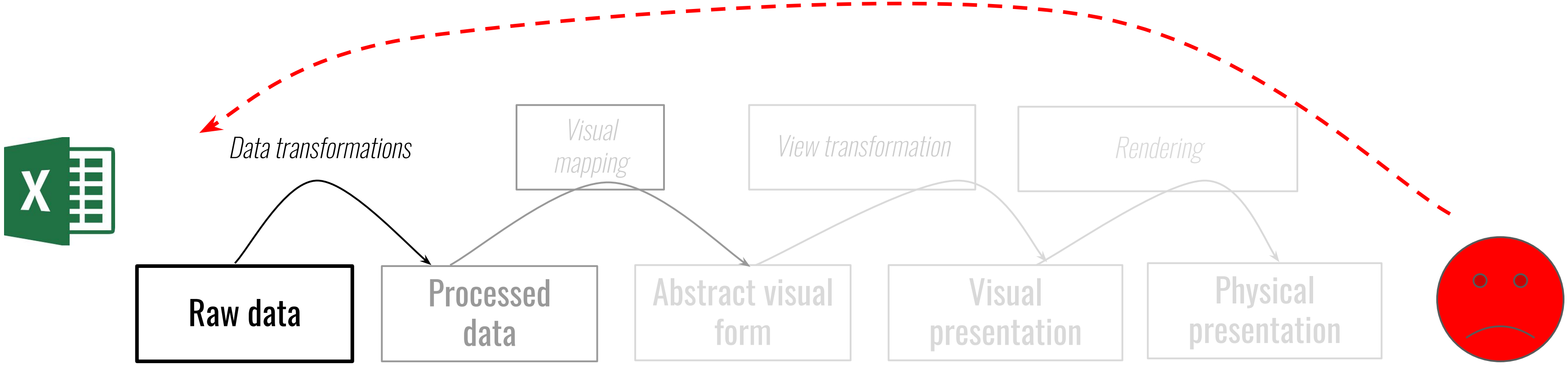
Expected visualization (demo)



Reality (demo)

D3.js https://d3js.org/

Need for interaction **with Raw Data**



Visualization!

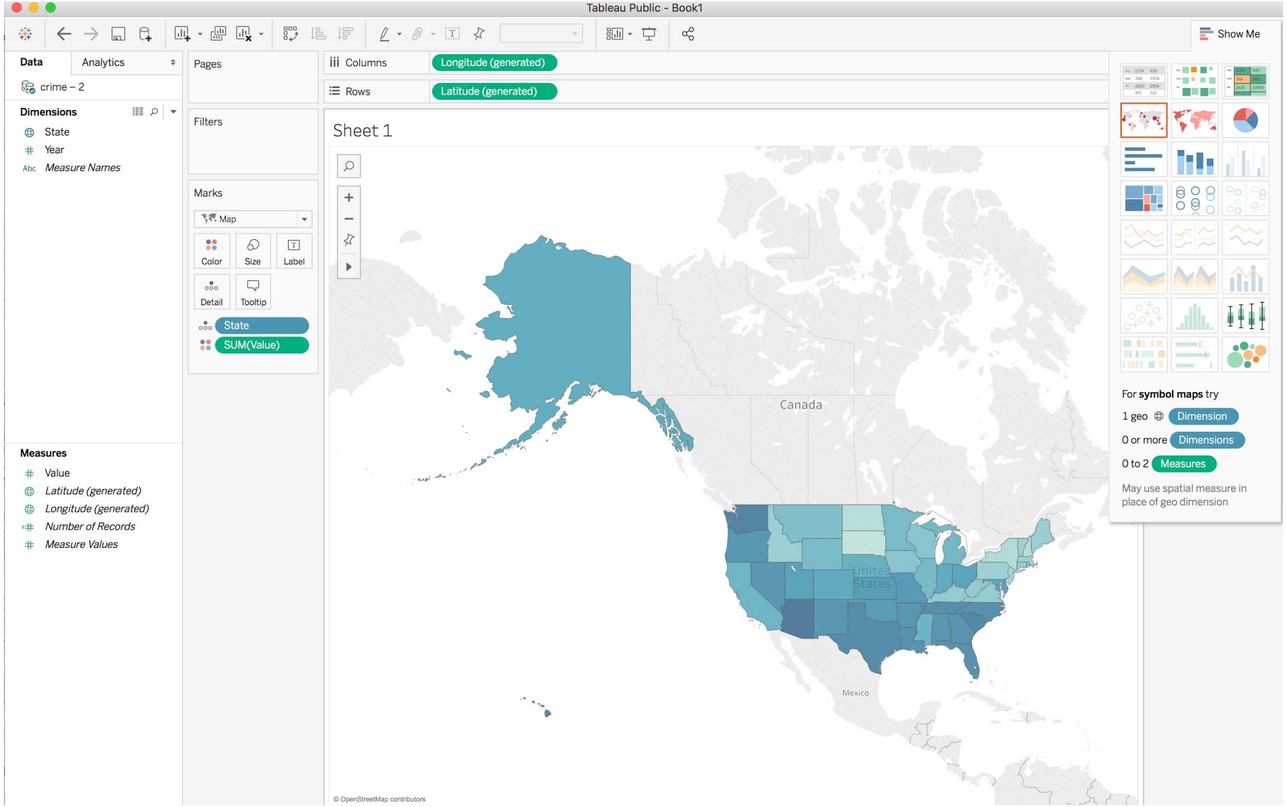


Tableau Public - Book1

Data source: crime – 2

Dimensions: State, Year, *Measure Names*

Measures: Value, *Latitude (generated)*, *Longitude (generated)*, *Number of Records*, *Measure Values*

Columns: Longitude (generated)

Rows: Latitude (generated)

Marks: Map

State

SUM(Value)

Sheet 1

For **symbol maps** try

1 geo Dimension

0 or more Dimensions

0 to 2 Measures

May use spatial measure in place of geo dimension

Tableau Software

Summary

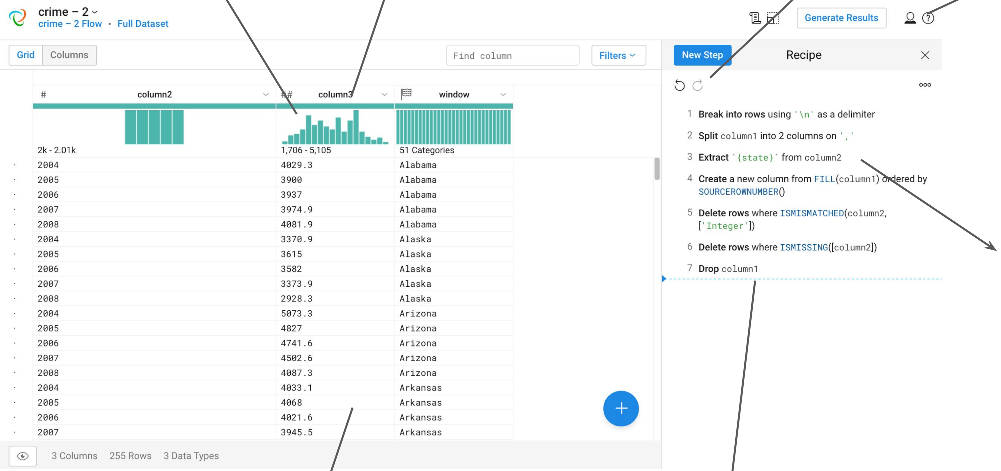
Data distribution

Data quality progress bar

Undo!

Export

Programming by demonstration



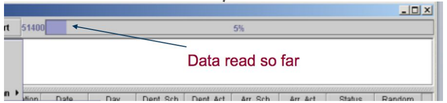
column2	column3	window
2k - 2.01k	1,706 - 5,105	51 Categories
2004	4029.3	Alabama
2005	3900	Alabama
2006	3937	Alabama
2007	3974.9	Alabama
2008	4081.9	Alabama
2004	3370.9	Alaska
2005	3615	Alaska
2006	3582	Alaska
2007	3373.9	Alaska
2008	2928.3	Alaska
2004	5073.3	Arizona
2005	4827	Arizona
2006	4741.6	Arizona
2007	4502.6	Arizona
2008	4087.3	Arizona
2004	4033.1	Arkansas
2005	4068	Arkansas
2006	4021.6	Arkansas
2007	3945.5	Arkansas

3 Columns 255 Rows 3 Data Types

1 Break into rows using `'\n'` as a delimiter
2 Split `column1` into 2 columns on `','`
3 Extract `` `{state}` `` from `column2`
4 Create a new column from `FILL(column1)` ordered by `SOURCEROWNUMBER()`
5 Delete rows where `ISMISMATCHED(column2, ['Integer'])`
6 Delete rows where `ISMISSING([column2])`
7 Drop `column1`

Preview transformation application

Data sampling progress



Data samples as table

Suggested transformations using a declarative language

Research directions

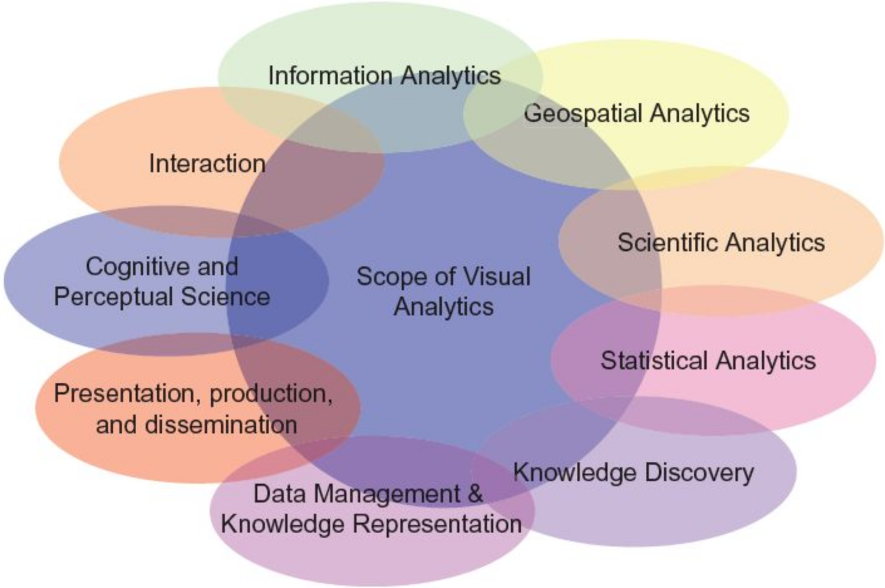
Abedjan, Z., Chu, X., Deng, D., Fernandez, R. C., Ilyas, I. F., Ouzzani, M., ... & Tang, N. (2016). Detecting Data Errors: Where are we and what needs to be done?. Proceedings of the VLDB Endowment, 9(12), 993-1004.

Kandel, S., Heer, J., Plaisant, C., Kennedy, J., van Ham, F., Riche, N. H., ... & Buono, P. (2011). Research directions in data wrangling: Visualizations and transformations for usable and credible data. Information Visualization, 10(4), 271-288.

Chu, X., Ilyas, I. F., Krishnan, S., & Wang, J. (2016, June). Data cleaning: Overview and emerging challenges. In Proceedings of the 2016 International Conference on Management of Data (pp. 2201-2206). ACM.

“Combine with visual analytics” [Kandel, 2011]

“Data wrangling also constitutes a promising direction for visual analytics research, as it requires combining automated techniques (e.g. discrepancy detection, entity resolution, semantic data type inference) with interactive visual interfaces”


Information Analytics
Geospatial Analytics
Interaction
Scientific Analytics
Cognitive and Perceptual Science
Scope of Visual Analytics
Statistical Analytics
Presentation, production, and dissemination
Knowledge Discovery
Data Management & Knowledge Representation

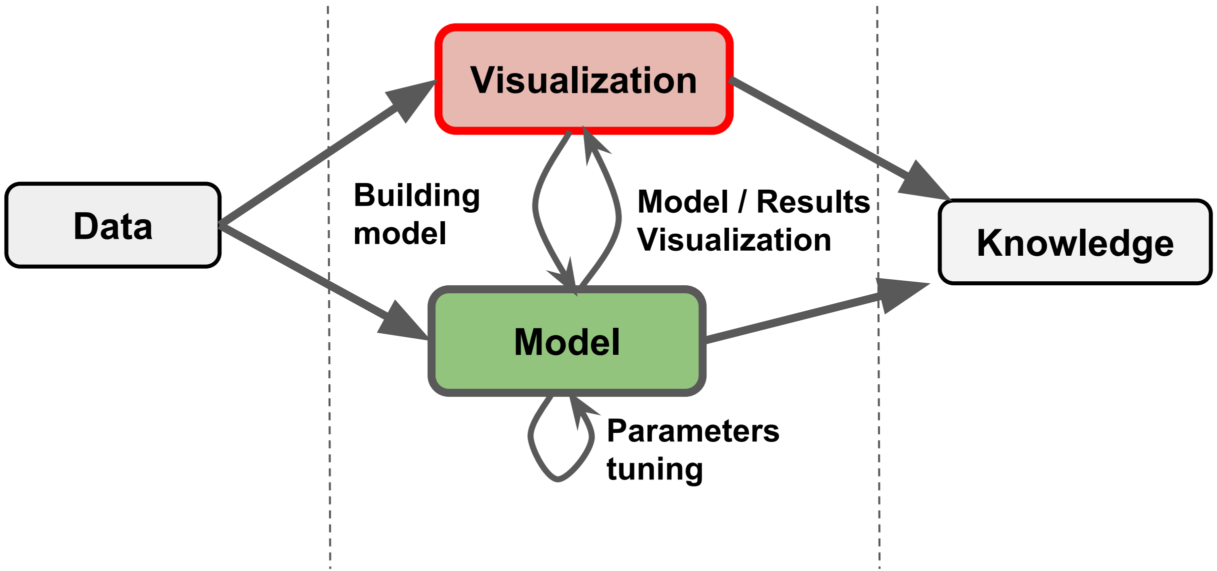
http://www.infovis-wiki.net/index.php?title=File:Keim06visual-analytics-disciplines.png

Visual Analytics

“

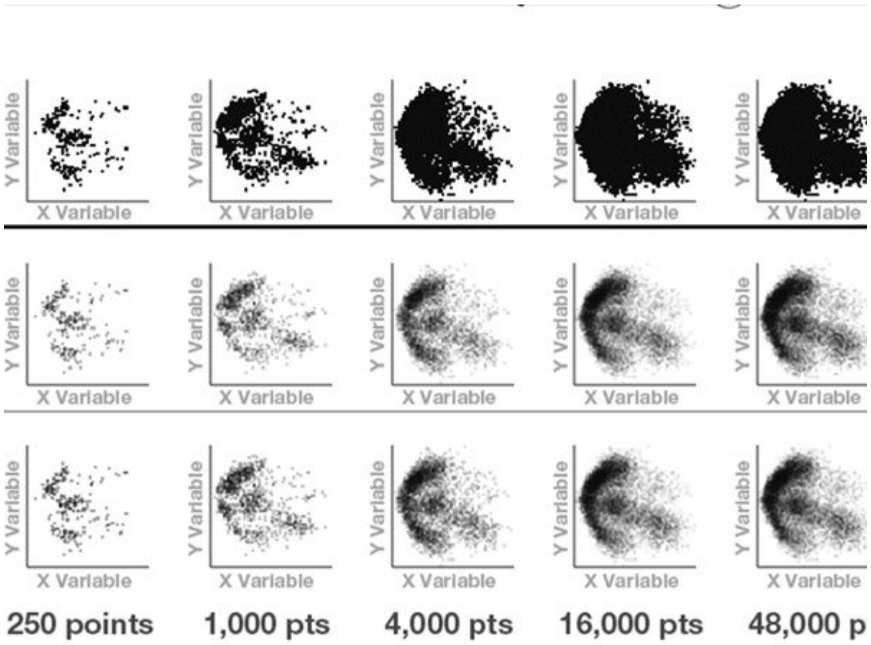
The science of analytical reasoning facilitated by interactive visual interfaces.

“



Thomas, J., Cook, K.: Illuminating the Path: Research and Development Agenda for Visual Analytics. IEEE-Press (2005)"

“Better Data *Exploration* tools (rather than *communication* tools)”



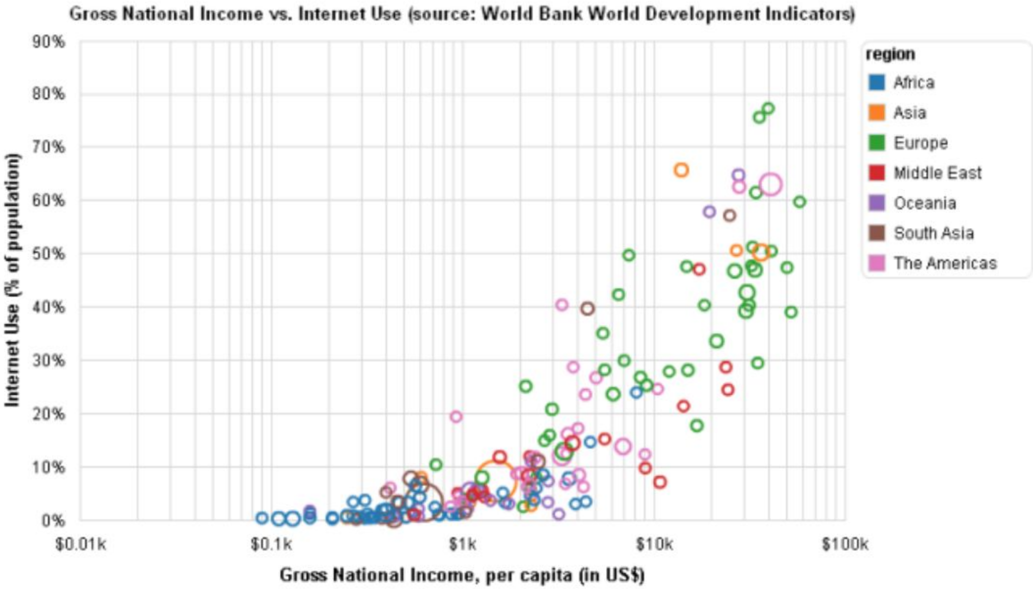
Matejka, Justin, Fraser Anderson, and George Fitzmaurice. "Dynamic opacity optimization for scatter plots." Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems. ACM, 2015. (pdf)

“Combine with query relaxation”

* We interact with **pixels**
Ex: brushing/selection

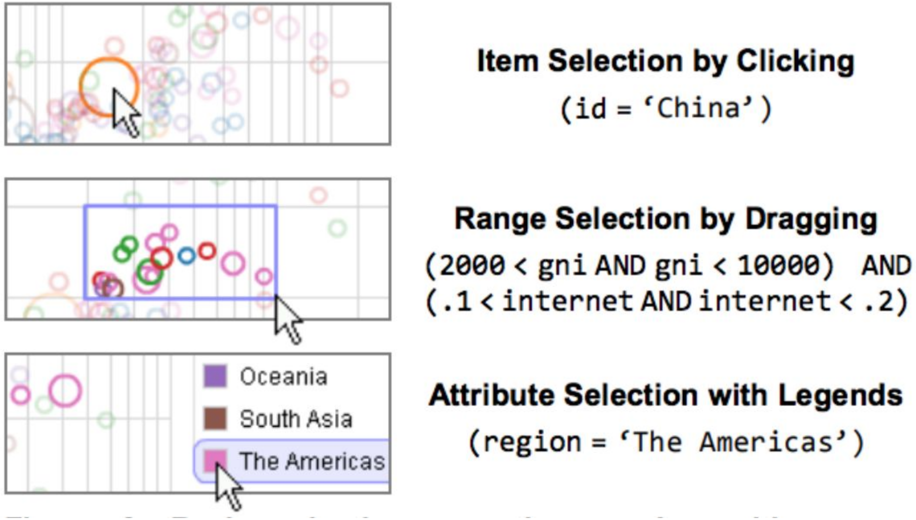
X < 300px && X > 600px
&& Y > 400px && Y < 700px

* Turn pixels into semantic



Heer, Jeffrey, Maneesh Agrawala, and Wesley Willett. *"Generalized selection via interactive query relaxation."* Proceedings of the SIGCHI Conference on Human Factors in Computing Systems. ACM, 2008. (pdf)

“Combine with query relaxation”



Item Selection by Clicking

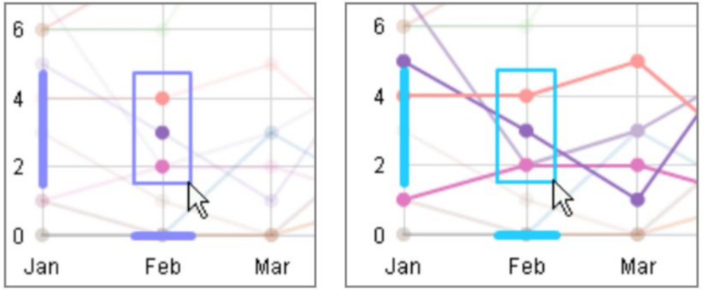
(id = ‘China’)

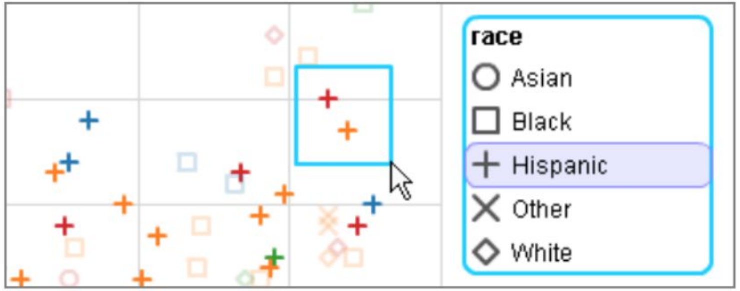
Range Selection by Dragging

(2000 < gni AND gni < 10000) AND (.1 < internet AND internet < .2)

Attribute Selection with Legends

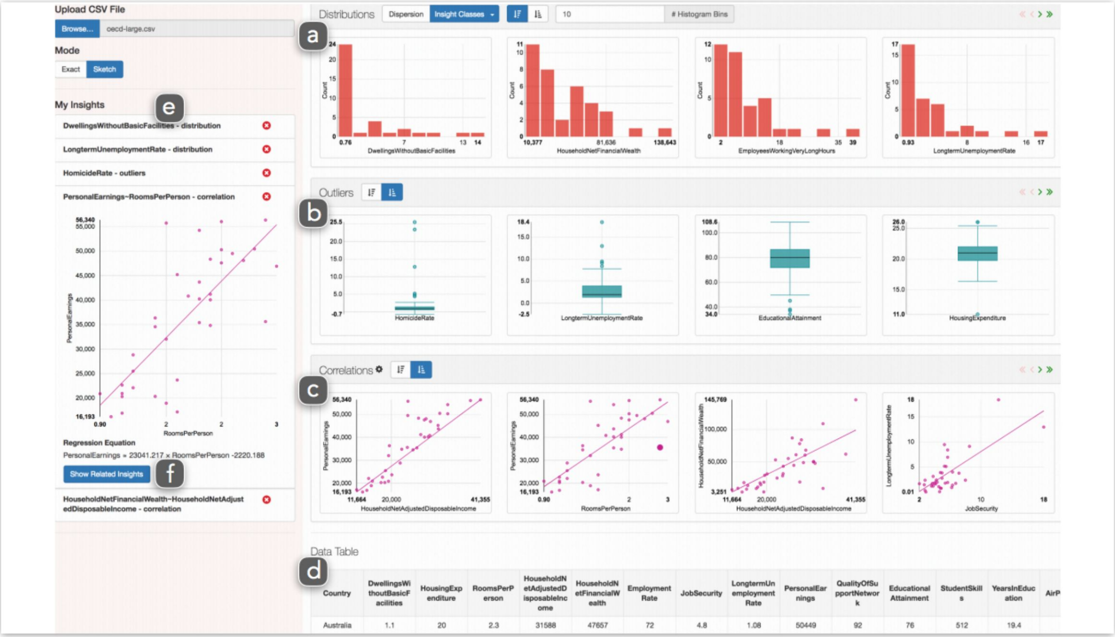
(region = ‘The Americas’)





Heer, Jeffrey, Maneesh Agrawala, and Wesley Willett. *"Generalized selection via interactive query relaxation."* Proceedings of the SIGCHI Conference on Human Factors in Computing Systems. ACM, 2008. (pdf)

“Guide users exploratory process”



Upload CSV File

oecd-large.csv

Mode

Exact | Sketch

My Insights

- DwellingsWithoutBasicFacilities - distribution
- LongtermUnemploymentRate - distribution
- HomicideRate - outliers
- PersonalEarnings~RoomsPerPerson - correlation
- HouseholdNetFinancialWealth~HouseholdNetAdjustedDisposableIncome - correlation

Regression Equation

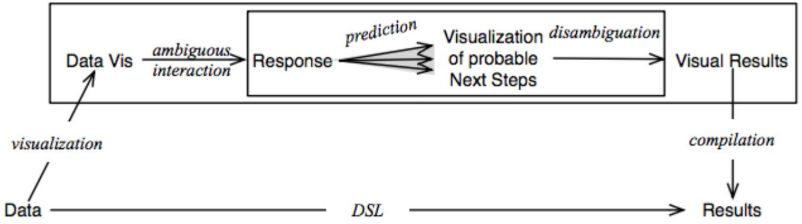
PersonalEarnings = 23041.217 × RoomsPerPerson -2220.188

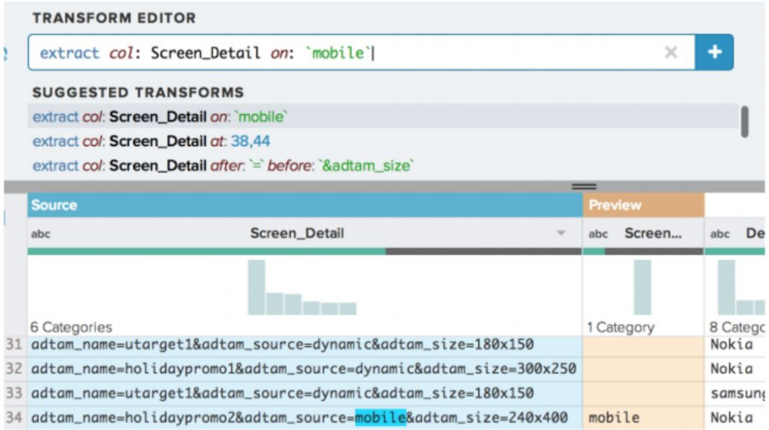
Data Table

Country	DwellingsWithoutBasicFacilities	HousingExpenditure	RoomsPerPerson	HouseholdNetAdjustedDisposableIncome	HouseholdNetFinancialWealth	EmploymentRate	JobSecurity	LongtermUnemploymentRate	PersonalEarnings	QualityOfSupportNetwork	EducationalAttainment	StudentSkills	YearsInEducation	AirP
Australia	1.1	20	2.3	31588	47657	72	4.8	1.08	50449	92	76	512	19.4	

Demiralp, Ç., Haas, P. J., Parthasarathy, S., & Pedapati, T. (2017). Foresight: Rapid Data Exploration Through Guideposts.

“Predict next interaction”





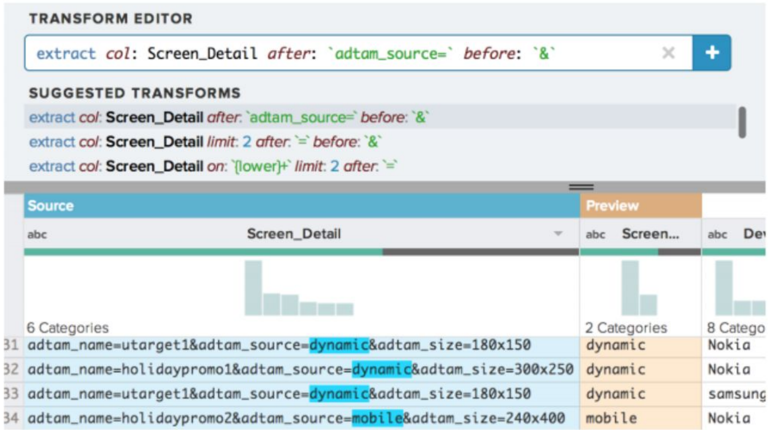
TRANSFORM EDITOR

```
extract col: Screen_Detail on: `mobile`|
```

SUGGESTED TRANSFORMS

- extract col: Screen_Detail on: `mobile`
- extract col: Screen_Detail at: 38,44
- extract col: Screen_Detail after: `=` before: `&adtam_size`

	Source	Preview	
	abc Screen_Detail	abc Screen...	abc De
	6 Categories	1 Category	8 Catego
31	adtam_name=utarget1&adtam_source=dynamic&adtam_size=180x150		Nokia
32	adtam_name=holidaypromo1&adtam_source=dynamic&adtam_size=300x250		Nokia
33	adtam_name=utarget1&adtam_source=dynamic&adtam_size=180x150		samsung
34	adtam_name=holidaypromo2&adtam_source=mobile&adtam_size=240x400	mobile	Nokia



TRANSFORM EDITOR

```
extract col: Screen_Detail after: `adtam_source=` before: `&`
```

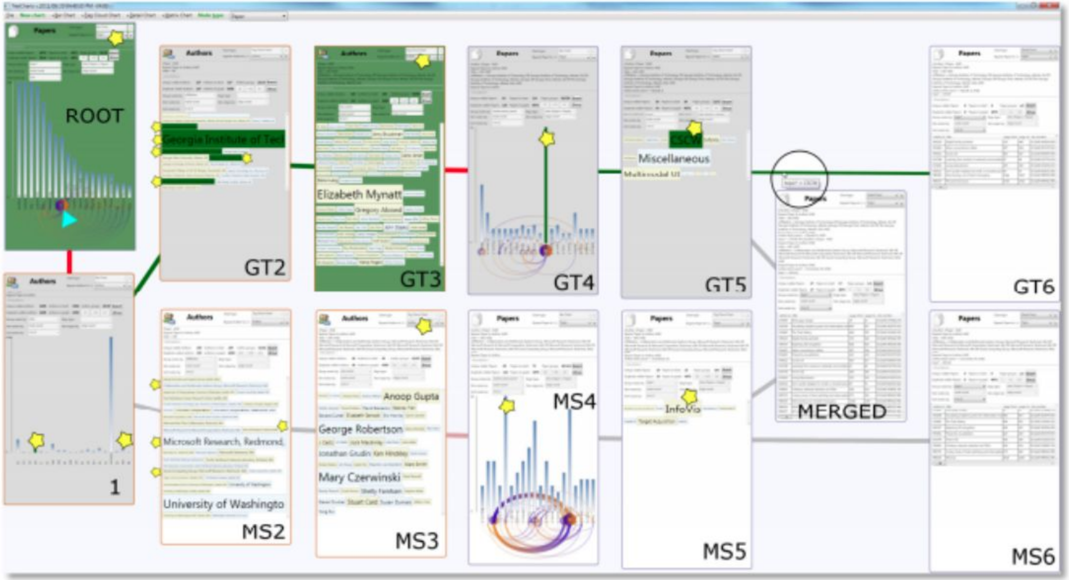
SUGGESTED TRANSFORMS

- extract col: Screen_Detail after: `adtam_source=` before: `&`
- extract col: Screen_Detail limit: 2 after: `=` before: `&`
- extract col: Screen_Detail on: `{lower}+` limit: 2 after: `=`

	Source	Preview	
	abc Screen_Detail	abc Screen...	abc De
	6 Categories	2 Categories	8 Catego
31	adtam_name=utarget1&adtam_source=dynamic&adtam_size=180x150	dynamic	Nokia
32	adtam_name=holidaypromo1&adtam_source=dynamic&adtam_size=300x250	dynamic	Nokia
33	adtam_name=utarget1&adtam_source=dynamic&adtam_size=180x150	dynamic	samsung
34	adtam_name=holidaypromo2&adtam_source=mobile&adtam_size=240x400	mobile	Nokia

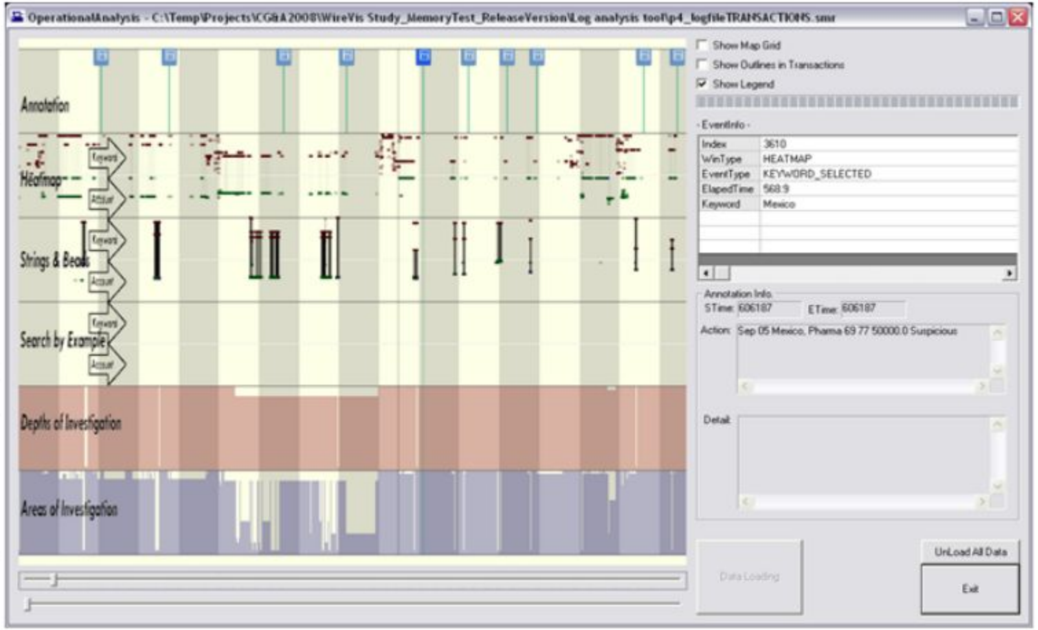
Heer, Jeffrey, Joseph M. Hellerstein, and Sean Kandel. "Predictive Interaction for Data Transformation." CIDR. 2015.

“Support history exploration”



Dunne, C., Henry Riche, N., Lee, B., Metoyer, R., & Robertson, G. (2012, May). GraphTrail: Analyzing large multivariate, heterogeneous networks while supporting exploration history. In Proceedings of the SIGCHI conference on human factors in computing systems (pp. 1663-1672). ACM.

“Help users recall their reasoning process”



OperationalAnalysis - C:\Temp\Projects\CG&A2008\WireVis Study_MemoryTest_ReleaseVersion\Log analysis tool\p4_logfileTRANSACTIONS.smr

□ Show Map Grid
□ Show Outlines in Transactions
☑ Show Legend

EventInfo

Index	3610
WinType	HEATMAP
EventType	KEYWORD_SELECTED
ElapedTime	568.9
Keyword	Mexico

Annotation Info

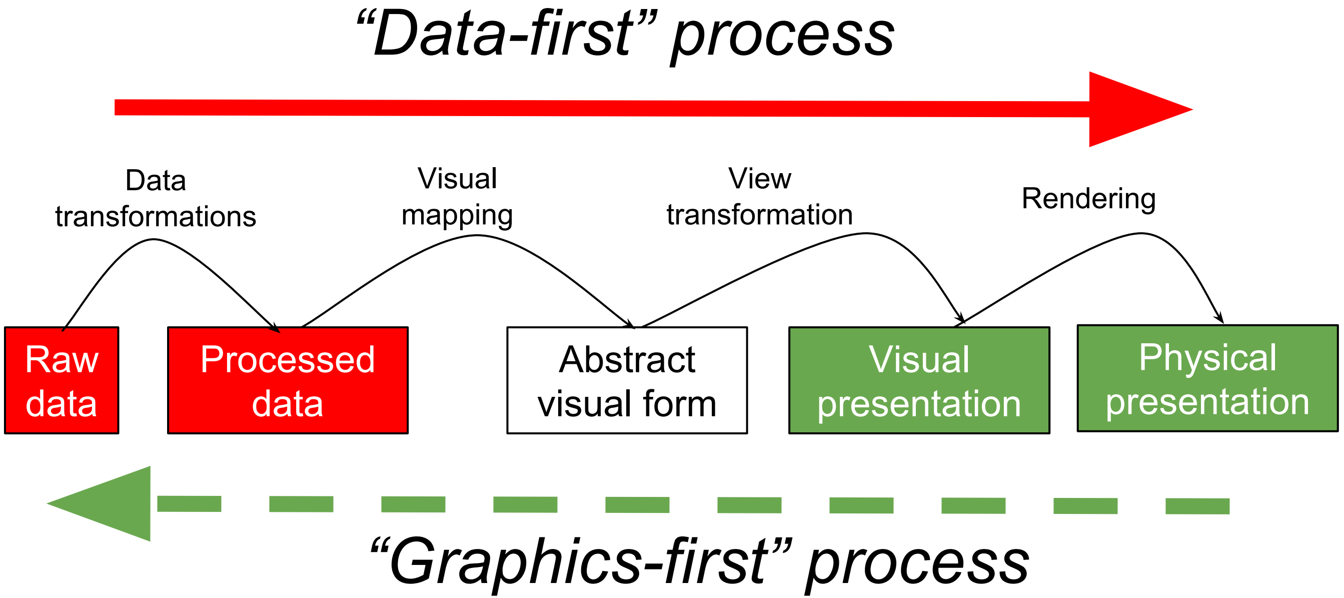
STime: 606187 ETime: 606187

Action: Sep 05 Mexico, Phama 69 77 50000.0 Suspicious

Detail:

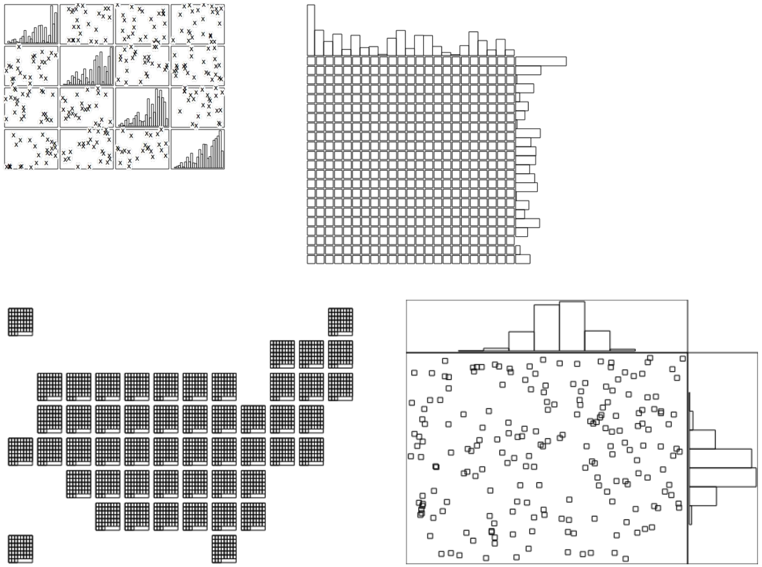
Lipford, H. R., Stukes, F., Dou, W., Hawkins, M. E., & Chang, R. (2010, October). Helping users recall their reasoning process. In Visual Analytics Science and Technology (VAST), 2010 IEEE Symposium on (pp. 187-194). (pdf).

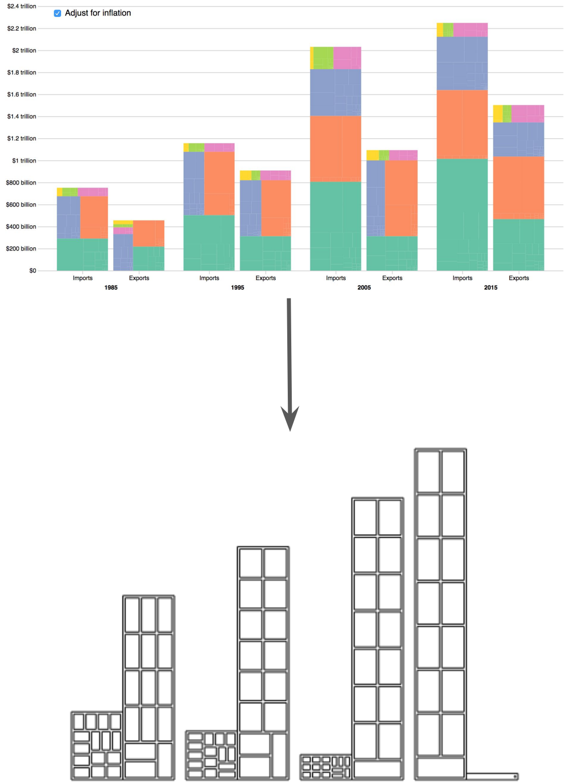
“Start working.. without data! (yet)”



Vuillemot, Romain, and Jeremy Boy. "Structuring Visualization Mock-ups at the Graphical Level by Dividing the Display Space." IEEE transactions on visualization and computer graphics (2017).

“Start working.. without data! (yet)”





Vuillemot, Romain, and Jeremy Boy. "Structuring Visualization Mock-ups at the Graphical Level by Dividing the Display Space." IEEE transactions on visualization and computer graphics (2017).

Future directions

[Abedjan et al., VLDB 2016]

A holistic combination of tools

A data enrichment system

A novel interactive dashboard.

Reasoning on real-world data

[Chu et al. ICMD 2016]

Scalability

User Engagement

Semi-structured and unstructured data

New Applications for Streaming Data

Growing Privacy and Security Concerns

[Kandel et al. IV 2011] (Among many!)

Living with dirty data

Visualize missing and uncertain data

Adapting systems to tolerate error

Sharing data transformations

Feedback from downstream analysts

Thanks!