TRIFACTA

Skills of the Agile Data Wrangler

Joe Hellerstein @joe_hellerstein
Jeffrey Heer @jeffrey_heer

YOUR HOSTS

JOE HELLERSTEIN

CEO Trifacta
Professor, Berkeley CS

MADlib, Bloom, Telegraph
Data Wrangler

JEFFREY HEER

CXO Trifacta
Professor, UW CSE

D3.js, Vega, Protovis
Data Wrangler

PLAN FOR THE TUTORIAL

Focus on Goals, Objectives & Strategy (Less Tactical)

OUTLINE

The Wrangling Problem

Secrets of the Agile Data Wrangler

Putting it Together

…and a peek at Trifacta's approach

ADDITIONAL STRATA ACTIVITY

- Trifacta Data Transformation session: Weds 4:50PM, Ballroom F
- Big Data Moonshots and Ground Control: Thurs 8:50AM Keynote

- Jeffrey Heer Office Hours: Weds 1:40PM, Table C
- Joe Hellerstein Office Hours: Thurs 10:10AM, Table A

The Wrangling Problem

WORD ON THE STREET

80%

of the work in any data project is cleaning the data.

DJ PATIL

Retweeted by Joe Hellerstein

Kirk Borne @KirkDBorne · Feb 10

#BigData #quote : "#Analytics is what #DataScientists do for fun after they've done all the tedious work" insideanalysis.com/wp-content/upl... #briefr

Expand

Reply Retweeted Favorite ••• More

Retweeted by Joe Hellerstein

Kirk Borne @KirkDBorne · Feb 10

#BigData #quote : "#Analytics is what #DataScientists do for fun after they've done all the tedious work" insideanalysis.com/wp-content/upl… #briefr

Expand · Reply · Retweeted · Favorite · More

http://smu.gs/1jqH3jO

ds@b

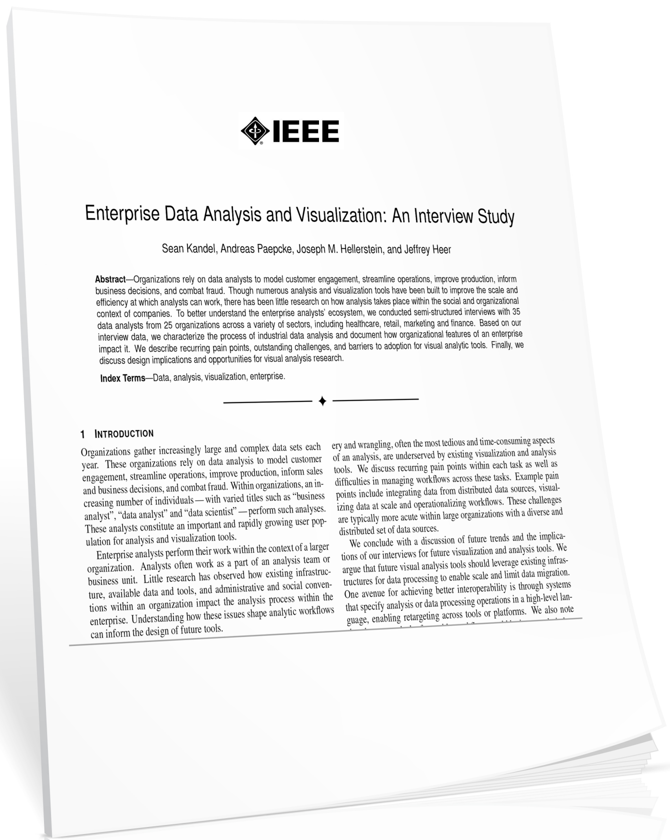
datascience@berkeley @BerkeleyData · Dec 5

Sad truth of cleaning up data: 80% of time spent cleaning up data, and 20% of the time spent COMPLAINING about cleaning up data. #DataBeat

Expand · Reply · Retweet · Favorite · More

"Enterprise Data Analysis and Visualization: An Interview Study"

Kandel, Paepcke, Hellerstein and Heer
IEEE Visual Analytics Science & Technology 2012



IEEE

Enterprise Data Analysis and Visualization: An Interview Study

Sean Kandel, Andreas Paepcke, Joseph M. Hellerstein, and Jeffrey Heer

Abstract—Organizations rely on data analysts to model customer engagement, streamline operations, improve production, inform business decisions, and combat fraud. Though numerous analysis and visualization tools have been built to improve the scale and efficiency at which analysts can work, there has been little research on how analysis takes place within the social and organizational context of companies. To better understand the enterprise analysts' ecosystem, we conducted semi-structured interviews with 35 data analysts from 25 organizations across a variety of sectors, including healthcare, retail, marketing and finance. Based on our interview data, we characterize the process of industrial data analysis and document how organizational features of an enterprise impact it. We describe recurring pain points, outstanding challenges, and barriers to adoption for visual analytic tools. Finally, we discuss design implications and opportunities for visual analysis research.

Index Terms—Data, analysis, visualization, enterprise.

1 INTRODUCTION

Organizations gather increasingly large and complex data sets each year. These organizations rely on data analysis to model customer engagement, streamline operations, improve production, inform sales and business decisions, and combat fraud. Within organizations, an increasing number of individuals — with varied titles such as "business analyst", "data analyst" and "data scientist" — perform such analyses. These analysts constitute an important and rapidly growing user population for analysis and visualization tools.

Enterprise analysts perform their work within the context of a larger organization. Analysts often work as a part of an analysis team or business unit. Little research has observed how existing infrastructure, available data and tools, and administrative and social conventions within an organization impact the analysis process within the enterprise. Understanding how these issues shape analytic workflows can inform the design of future tools.

ery and wrangling, often the most tedious and time-consuming aspects of an analysis, are underserved by existing visualization and analysis tools. We discuss recurring pain points within each task as well as difficulties in managing workflows across these tasks. Example pain points include integrating data from distributed data sources, visualizing data at scale and operationalizing workflows. These challenges are typically more acute within large organizations with a diverse and distributed set of data sources.

We conclude with a discussion of future trends and the implications of our interviews for future visualization and analysis tools. We argue that future visual analysis tools should leverage existing infrastructures for data processing to enable scale and limit data migration. One avenue for achieving better interoperability is through systems that specify analysis or data processing operations in a high-level language, enabling retargeting across tools or platforms. We also note

SEAN KANDEL

CTO Trifacta
PhD, Stanford CS

Citadel Investment Group

Data Wrangler

HIGH FRICTION FOR DATA SCIENTISTS

FRUSTRATION: WRANGLING BOTTLENECK

"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."

LOST OPPORTUNITY FOR BUSINESS ANALYSTS

POTENTIAL: END-USER SELF-SERVICE

> “Most of the time once you transform the data…the insights can be scarily obvious.”

LOST OPPORTUNITY FOR BUSINESS ANALYSTS

REALITY: HEAVYWEIGHT INTERACTION WITH IT

"All data is in a relational database. When I get it, it's out of the database and into an Excel format that I can start pivoting. I ask the IT team to pull it."

THIS IS THE BIG DEAL

The biggest bottleneck in current practice

The biggest roadblock to a data-driven future

A problem that goes outside technical boundaries

DEFINITIONAL ISSUES

DEFINITIONAL ISSUES

What is *"clean"* data? What is *"clean enough"*?

DEFINITIONAL ISSUES

What is *"clean"* data? What is *"clean enough"*?
Better yet, is the data ***"fit for a purpose"***?

DEFINITIONAL ISSUES

What is *"clean"* data? What is *"clean enough"*?
Better yet, is the data ***"fit for a purpose"***?

DEFINITIONAL ISSUES

What is *"clean"* data? What is *"clean enough"*?
Better yet, is the data ***"fit for a purpose"***?

Can I work with the data? (Is it *usable*)

DEFINITIONAL ISSUES

What is *"clean"* data? What is *"clean enough"*?
Better yet, is the data ***"fit for a purpose"***?

Can I work with the data? (Is it *usable*)

DEFINITIONAL ISSUES

What is *"clean"* data? What is *"clean enough"*?
Better yet, is the data ***"fit for a purpose"***?

Can I work with the data? (Is it *usable*)

Do I trust the data? (Is it *credible*)

DEFINITIONAL ISSUES

What is *“clean”* data? What is *“clean enough”*?
Better yet, is the data ***“fit for a purpose”***?

Can I work with the data? (Is it *usable*)

Do I trust the data? (Is it *credible*)

DEFINITIONAL ISSUES

What is *"clean"* data? What is *"clean enough"*?
Better yet, is the data ***"fit for a purpose"***?

Can I work with the data? (Is it *usable*)

Do I trust the data? (Is it *credible*)

Can I learn from it? (Is it *useful*)

USABILITY, CREDIBILITY & USEFULNESS

USABILITY, CREDIBILITY & USEFULNESS

Data is ***usable*** if it can be parsed and manipulated by computational tools. Data usability is thus defined in conjunction with the tools by which it is to be processed.

USABILITY, CREDIBILITY & USEFULNESS

Data is ***usable*** if it can be parsed and manipulated by computational tools. Data usability is thus defined in conjunction with the tools by which it is to be processed.

USABILITY, CREDIBILITY & USEFULNESS

Data is ***usable*** if it can be parsed and manipulated by computational tools. Data usability is thus defined in conjunction with the tools by which it is to be processed.

Data is ***credible*** if, according to one's subjective assessment, it is suitably representative of a phenomenon to enable productive analysis.

USABILITY, CREDIBILITY & USEFULNESS

Data is ***usable*** if it can be parsed and manipulated by computational tools. Data usability is thus defined in conjunction with the tools by which it is to be processed.

Data is ***credible*** if, according to one's subjective assessment, it is suitably representative of a phenomenon to enable productive analysis.

USABILITY, CREDIBILITY & USEFULNESS

Data is ***usable*** if it can be parsed and manipulated by computational tools. Data usability is thus defined in conjunction with the tools by which it is to be processed.

Data is ***credible*** if, according to one's subjective assessment, it is suitably representative of a phenomenon to enable productive analysis.

Data is ***useful*** if it is usable, credible, and *responsive to one's inquiry*.

STANDARD APPROACHES

Custom Code (ideally in Domain-Specific Languages)

SAS Data Step

```
data newlist;
  set newdata.maillist;
/* Extract month, day and year */
/* from the date character vara */
  m = scan(date,1,’ ‘);
  d = scan(date,2,’ ‘);
  y = scan(year,2,’,’);
  dd = compress(d||m||y,’ ,’);
/* Convert mon, day, year into */
/* new date variableb */
  newdate = input(dd,date9.);
run;
```

http://analytics.ncsu.edu/sesug/2001/P-818.pdf

STANDARD APPROACHES

Custom Code (ideally in Domain-Specific Languages)

SAS Data Step

Python Pandas

```
noise_complaint_counts = noise_complaints['Borough'].value_counts()
complaint_counts = complaints['Borough'].value_counts()
noise_complaint_counts / complaint_counts.astype(float)
```

http://pandas.pydata.org/pandas-docs/dev/tutorials.html#pandas-cookbook

STANDARD APPROACHES

Custom Code (ideally in Domain-Specific Languages)

SAS Data Step

Python Pandas

Manual Manipulation in Spreadsheet Interfaces

Spreadsheets

Social Security No.	Last Name	Middle Name	First Name	DOB	Date of Letter
123456789	freehafer	drew	nancy	19780128	7/30/11
123456789	cencini	a	andrew	19850521	8/2/11
123456789	kotas	bollen	jan	19591013	8/5/11
123456789	sergienko		mariya	19710805	8/11/11
123456789	thorpe	j	steven	19720508	8/19/11
123456789	neipper	john	michael	19520520	8/24/11
123456789	zare	b	robert	19750227	8/30/11
123456789	giussani		laura	19561211	8/31/11
123456789	hellung-larsen	marie	anne	19750923	9/1/11

123-45-6789	1/28/78
123-45-6789	5/21/85
123-45-6789	10/13/59
123-45-6789	8/5/71
123-45-6789	5/8/72
123-45-6789	5/20/52
123-45-6789	2/27/75
123-45-6789	12/11/56
123-45-6789	9/23/75

For the Social Security Number values, I applied a built-in format for Social Security Numbers. Click Ctrl+Shift+F to bring up the **Format Cells** dialog box and, on the **Number** tab, under **Category**, click **Special** and then select **Social Security Number**.

Freehafer, Nancy D
Cencini, Andrew A
Kotas, Jan B
Sergienko, Mariya
Thorpe, Steven J
Neipper, Michael J
Zare, Robert B
Giussani, Laura
Hellung-Larsen, Anne M

```
=PROPER(TRIM(TRIM(B2)&", "&D2&" "&LEFT(C2,1)))
```

The **TRIM** function removes unneeded leading and trailing spaces from the last names, first names, and middle names. The **LEFT** function gets the first letter of the middle name. The & characters combine the names, including a comma between the last and the first name.

This is how a name broken out into three cells became a last name, first name, and middle initial, all in one cell.

http://blogs.office.com/2011/09/20/clean-up-imported-or-pasted-data-in-excel/

STANDARD APPROACHES

Custom Code (ideally in Domain-Specific Languages)

SAS Data Step

Python Pandas

Manual Manipulation in Spreadsheet Interfaces

Spreadsheets

Schema Mapping and Workflow in Enterprise Software

Informatica Power Center

http://www.iri.com/blog/data-transformation2/informatica-pushdown-optimization-with-cosort/

STANDARD APPROACHES

Custom Code (ideally in Domain-Specific Languages)

SAS Data Step

Python Pandas

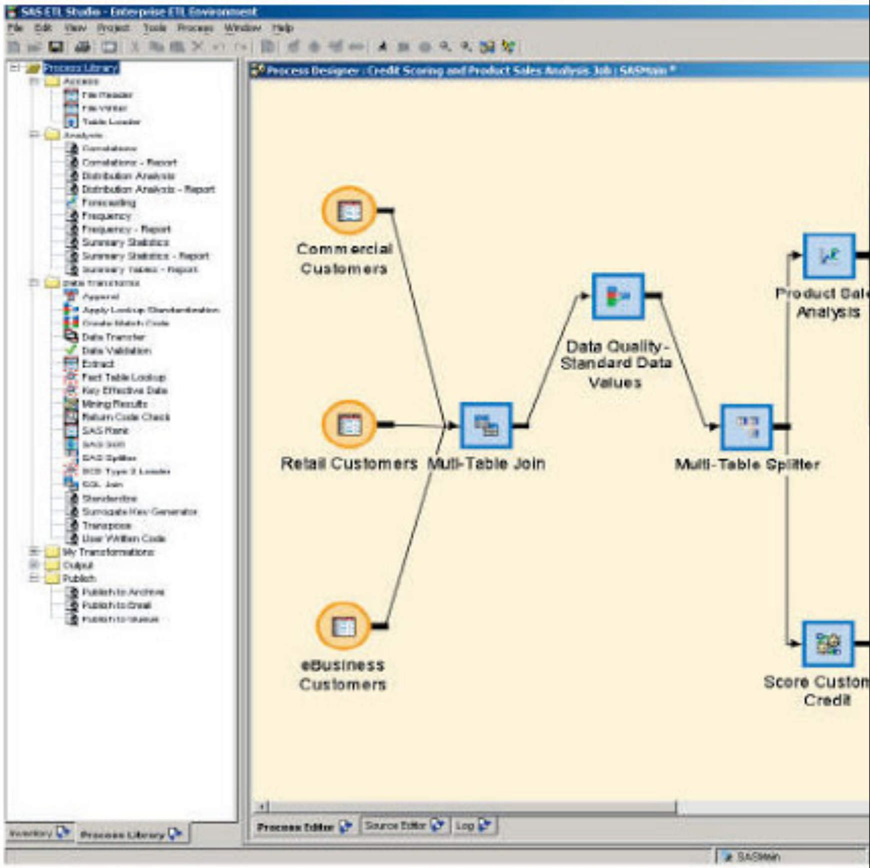
Manual Manipulation in Spreadsheet Interfaces

Spreadsheets

Schema Mapping and Workflow in Enterprise Software

Informatica Power Center

SAS DI Studio



The graphical process designer in SAS Data Integration Studio all
maintain complex processes.

http://saslearn.blogspot.com/2012/05/etl-processing-using-sas-data.html

hy·per·bo·le

/hīˈpərbəlē/

noun

1. exaggerated statements or claims not meant to be taken literally.
   synonyms: exaggeration, overstatement, magnification, embroidery, embellishment, excess, overkill, rhetoric; More

A GOOD DAY FOR A MODERN DATA SCIENTIST

9AM: Hypothesis formed

10AM-12PM: Land and examine various data sets

12-12:45PM: Delicious, healthy food

12:45-3PM: Wrangle chosen data

3-4:30PM: Analyze chosen data

4:30-4:45: Wheatgrass shot

4:45-6PM: Insight, Storytelling

A GOOD MONTH FOR A 2007 DATA ANALYST

02/01: Business use case identified. Consult Warehouse Schema and MDM Master Data for relevant "golden data".

02/02: Land and examine various data sets in staging filer.

02/03: Request private data "sandbox" alongside EDW to house new data

> *Note: "The sandbox phenomenon … carries a significant risk to the IT organization and EDW architecture because it could create isolated and incompatible stovepipes of data"*
>
> http://www.montage.co.nz/assets/Brochures/DataWarehouseBigDataAnalyticsKimball.pdf

02/03: Define schemas and write specifications for ETLing data into sandbox.

02/10: Receive notice from IT that sandbox is loaded. Begin profiling

02/11: Revise specifications for ETLing data and request reload

02/15: Receive notice from IT that sandbox is reloaded. Begin profiling

02/16: Further in-database wrangling and view definition

02/17: Analyze chosen data, engage in storytelling

02/18: Request schema modification in EDW to accommodate data

02/19: Begin writing spec to recode ETL/wrangling

> *"At that point, tracking applications that may have been implemented in the sandbox using a quick and dirty prototyping language, are usually reimplemented by other personnel in the EDW environment using corporate standard tools"*
>
> http://www.montage.co.nz/assets/Brochures/DataWarehouseBigDataAnalyticsKimball.pdf

DIFFERENCES IN GOALS, PROCESSES, DATA

DIFFERENCES IN GOALS, PROCESSES, DATA

Data Warehousing

Single source of truth: Engineered structure.

Waterfall design process.

Garbage-In-Garbage-Out: only golden data stored

DIFFERENCES IN GOALS, PROCESSES, DATA

Data Warehousing

Single source of truth: Engineered structure.

Waterfall design process.

Garbage-In-Garbage-Out: only golden data stored

"There is no point bringing data … into the data warehouse without integrating it".

— Bill Inmon, *Building the Data Warehouse,* 2005

DIFFERENCES IN GOALS, PROCESSES, DATA

Data Warehousing

Single source of truth: Engineered structure.

Waterfall design process.

Garbage-In-Garbage-Out: only golden data stored

"There is no point bringing data … into the data warehouse without integrating it".

— Bill Inmon, *Building the Data Warehouse,* 2005

Data Science

Exploration and provisional truth

Agile design

Signal out of noise: all data stored

DIFFERENCES IN GOALS, PROCESSES, DATA

Data Warehousing

Single source of truth: Engineered structure.

Waterfall design process.

Garbage-In-Garbage-Out: only golden data stored

"There is no point bringing data … into the data warehouse without integrating it".

— Bill Inmon, *Building the Data Warehouse*, 2005

Data Science

Exploration and provisional truth

Agile design

Signal out of noise: all data stored

"Get into the mindset to collect and measure everything you can"

—DJ Patil, *Building Data Science Teams*, 2011

DIFFERENCES IN GOALS, PROCESSES, DATA

Data Warehousing

Single source of truth: Engineered structure.

Waterfall design process.

Garbage-In-Garbage-Out: only golden data stored

"There is no point bringing data … into the data warehouse without integrating it".

— Bill Inmon, *Building the Data Warehouse*, 2005

Data Science

Exploration and provisional truth

Agile design

Signal out of noise: all data stored

"Get into the mindset to collect and measure everything you can"

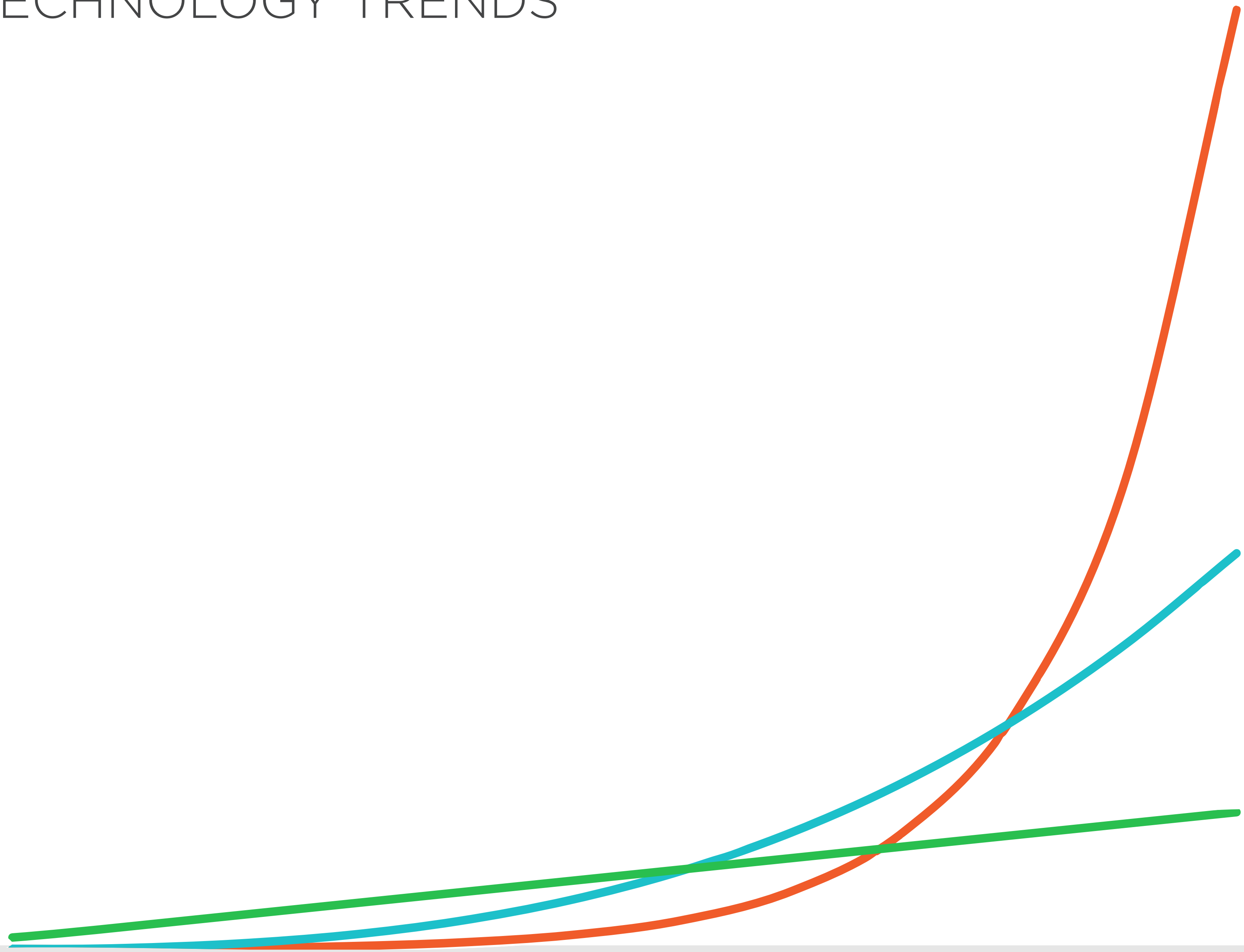
—DJ Patil, *Building Data Science Teams*, 2011

Rational people from both communities know these need to coexist

The former is high value, low variety and volume

The latter is growing value, variety, volume

TECHNOLOGY TRENDS



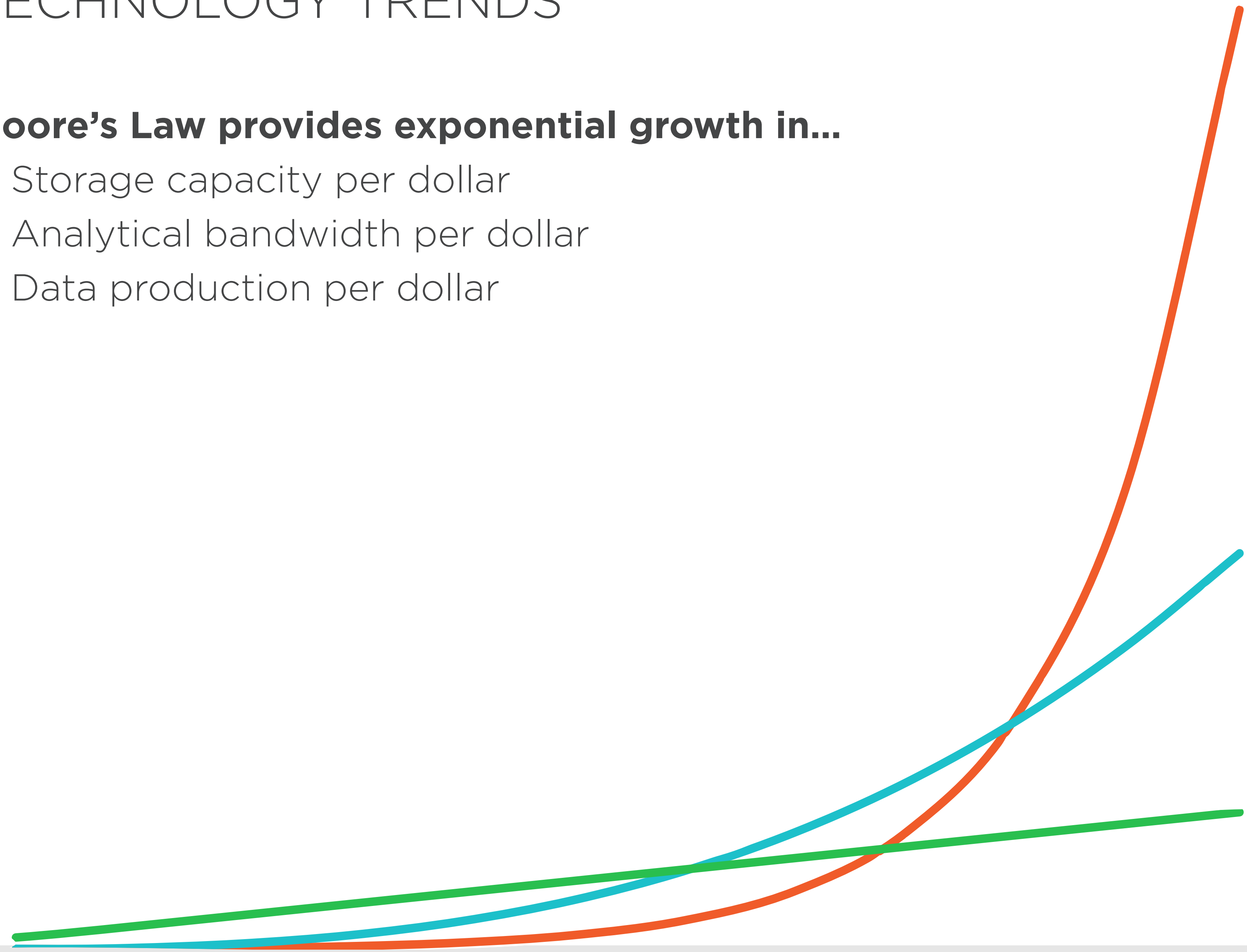
TECHNOLOGY TRENDS

Moore's Law provides exponential growth in...

Storage capacity per dollar

Analytical bandwidth per dollar

Data production per dollar



TECHNOLOGY TRENDS

Moore's Law provides exponential growth in...

Storage capacity per dollar

Analytical bandwidth per dollar

Data production per dollar

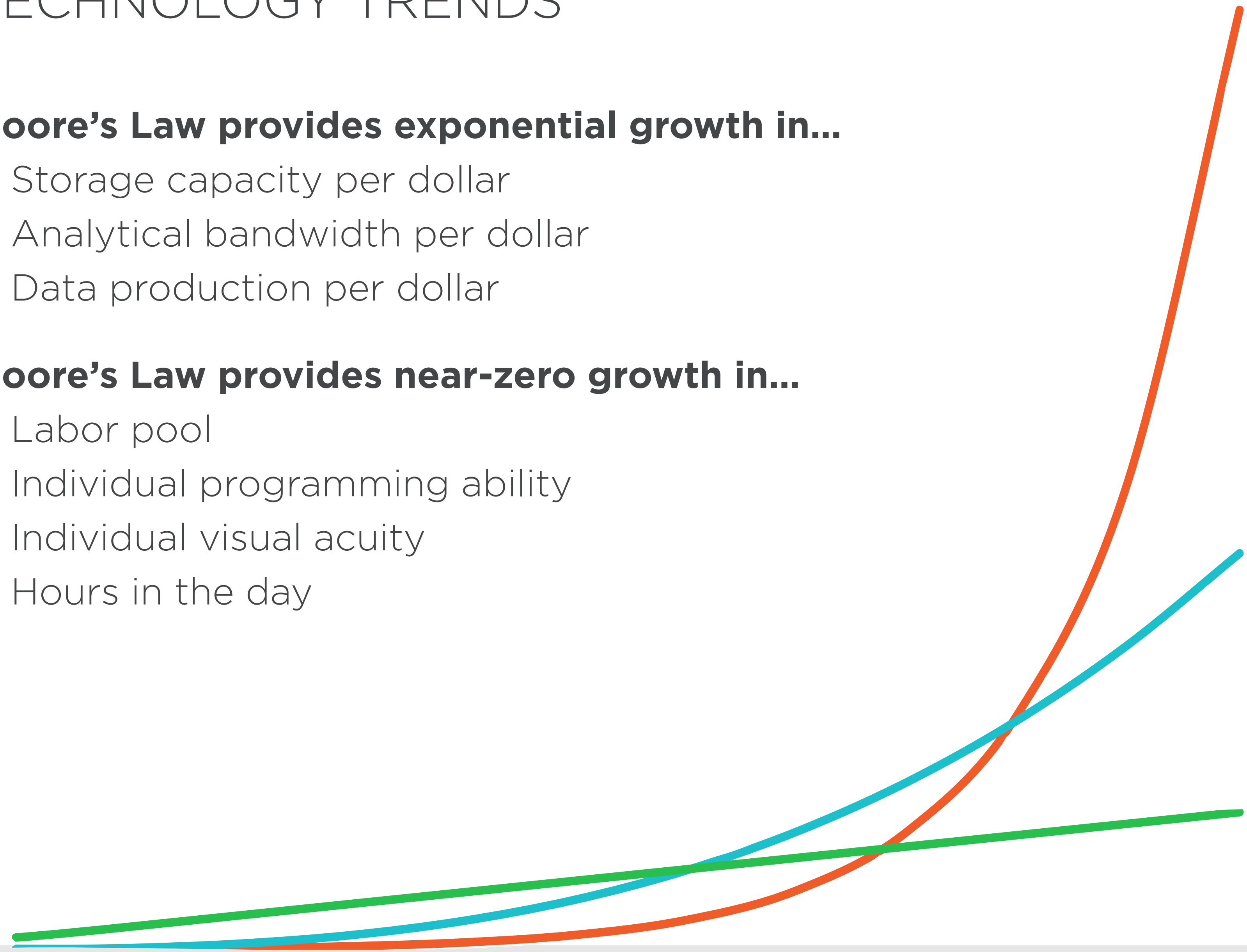
Moore's Law provides near-zero growth in...

Labor pool

Individual programming ability

Individual visual acuity

Hours in the day



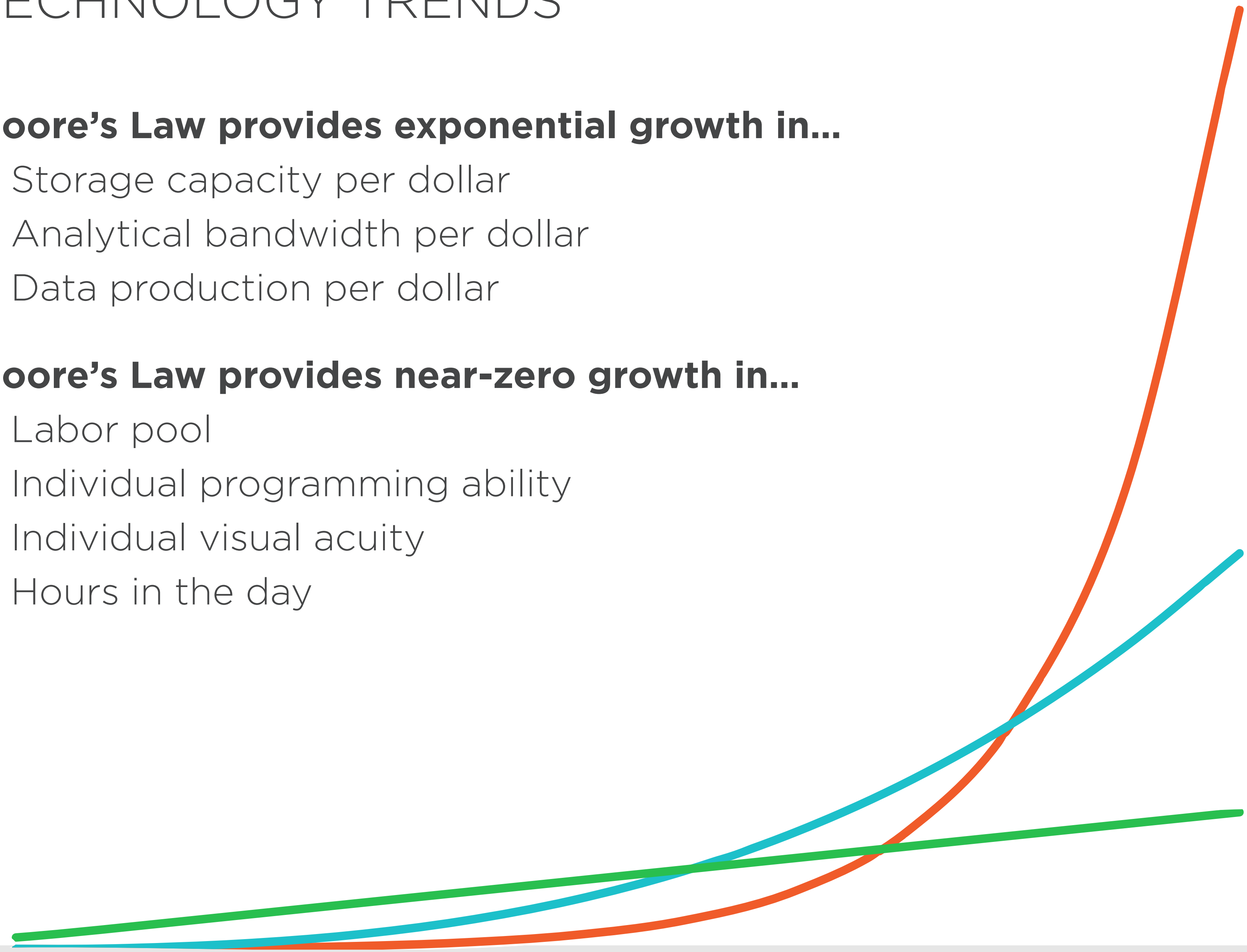
TECHNOLOGY TRENDS

Moore's Law provides exponential growth in...

- Storage capacity per dollar
- Analytical bandwidth per dollar
- Data production per dollar

Moore's Law provides near-zero growth in...

- Labor pool
- Individual programming ability
- Individual visual acuity
- Hours in the day



TECHNOLOGY TRENDS

Moore's Law provides exponential growth in…

- Storage capacity per dollar
- Analytical bandwidth per dollar
- Data production per dollar

Moore's Law provides near-zero growth in…

- Labor pool
- Individual programming ability
- Individual visual acuity
- Hours in the day

Possible Solutions?

- Better automation
- Higher-level abstractions for humans
- *Combinations of the two*

IS AUTOMATION THE ANSWER?

Example: SF Restaurant Violation Data.

http://blog.zipfianacademy.com/post/57158627293/how-to-data-science-mapping-sf-restaurant-inspection

https://data.sfgov.org/Public-Health/Restaurant-Scores/stya-26eb

```
"business_id","date","description"
10,"20121114","Unclean or degraded floors walls or ceilings  [ date violation corrected:  ]"
10,"20120403","Unclean or degraded floors walls or ceilings  [ date violation corrected: 9/20/2012 ]"
10,"20110428","Inadequate and inaccessible handwashing facilities  [ date violation corrected: 6/1/2011 ]"
12,"20120420","Food safety certificate or food handler card not available  [ date violation corrected: 11/20/2012 ]"
17,"20120823","Inadequately cleaned or sanitized food contact surfaces  [ date violation corrected: 9/6/2012 ]"
17,"20120823","High risk food holding temperature   [ date violation corrected: 9/6/2012 ]"
17,"20120823","Unclean nonfood contact surfaces  [ date violation corrected: 9/6/2012 ]"
```

```
"business_id","name","address","city","state","postal_code","latitude","longitude","phone_number"
10,"TIRAMISU KITCHEN","033 BELDEN PL","San Francisco","CA","94104","37.791116","-122.403816",""
12,"KIKKA","250 EMBARCADERO  7/F","San Francisco","CA","94105","37.788613","-122.393894",""
17,"GEORGE'S COFFEE SHOP","2200 OAKDALE AVE ","San Francisco","CA","94124","37.741086","-122.401737","+14155531470"
19,"NRGIZE LIFESTYLE CAFE","1200 VAN NESS AVE, 3RD FLOOR","San Francisco","CA","94109","37.786848","-122.421547",""
24,"OMNI S.F. HOTEL - 2ND FLOOR PANTRY","500 CALIFORNIA ST, 2ND  FLOOR","San Francisco","CA","94104","37.792888","-122.403135",""
29,"CHICO'S PIZZA","131 06TH ST ","San Francisco","CA","94103","37.774722","-122.406761","+14155251111"
31,"NORMAN'S ICE CREAM AND FREEZES","2801 LEAVENWORTH ST ","San Francisco","CA","94133","37.807155","-122.419004",
```

```
"business_id","Score","date","type"
10,"98","20121114","routine"
10,"98","20120403","routine"
10,"100","20110928","routine"
10,"96","20110428","routine"
10,"100","20101210","routine"
12,"100","20121120","routine"
```

IS AUTOMATION THE ANSWER?

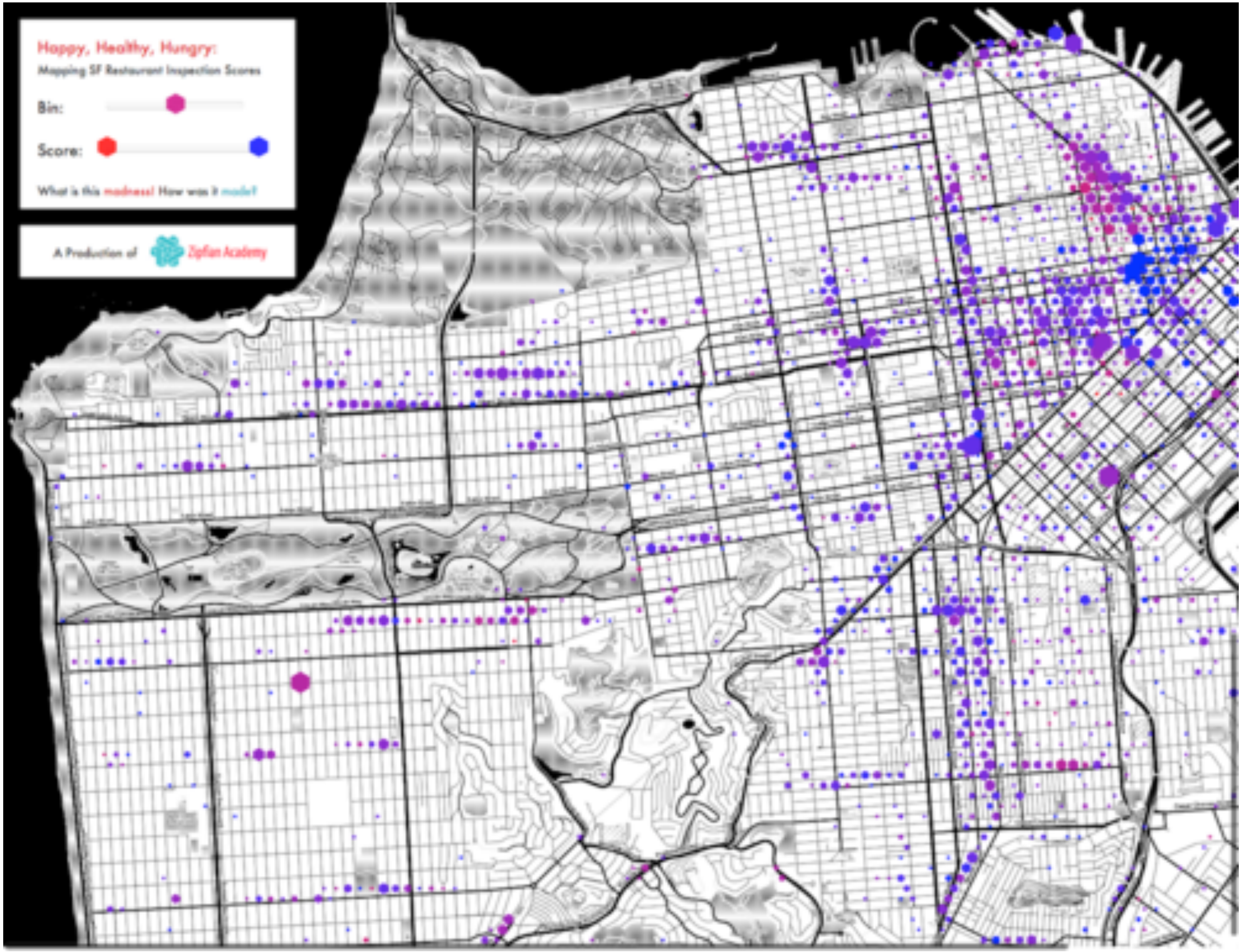
Example: SF Restaurant Violation Data.

http://blog.zipfianacademy.com/post/57158627293/how-to-data-science-mapping-sf-restaurant-inspection
https://data.sfgov.org/Public-Health/Restaurant-Scores/stya-26eb

```
"business_id","date","description"
10,"20121114","Unclean or degraded floors walls or ceilings  [ date violation corrected:  ]"
10,"20120403","Unclean or degraded floors walls or ceilings  [ date violation corrected: 9/20/2012 ]"
10,"20110428","Inadequate and inaccessible handwashing facilities  [ date violation corrected: 6/1/2011 ]"
12,"20120420","Food safety certificate or food handler card not available  [ date violation corrected: 11/20/2012 ]"
17,"20120823","Inadequately cleaned or sanitized food contact surfaces  [ date violation corrected: 9/6/2012 ]"
17,"20120823","High risk food holding temperature   [ date violation corrected: 9/6/2012 ]"
17,"20120823","Unclean nonfood contact surfaces  [ date violation corrected: 9/6/2012 ]"
```

```
"business_id","name","address","city","state","postal_code","latitude","longitude","phone_number"
10,"TIRAMISU KITCHEN","033 BELDEN PL","San Francisco","CA","94104","37.791116","-122.403816",""
12,"KIKKA","250 EMBARCADERO  7/F","San Francisco","CA","94105","37.788613","-122.393894",""
17,"GEORGE'S COFFEE SHOP","2200 OAKDALE AVE ","San Francisco","CA","94124","37.741086","-122.401737","+14155531470"
19,"NRGIZE LIFESTYLE CAFE","1200 VAN NESS AVE, 3RD FLOOR","San Francisco","CA","94109","37.786848","-122.421547",""
24,"OMNI S.F. HOTEL - 2ND FLOOR PANTRY","500 CALIFORNIA ST, 2ND  FLOOR","San Francisco","CA","94104","37.792888","-122.403135",""
29,"CHICO'S PIZZA","131 06TH ST ","San Francisco","CA","94103","37.774722","-122.406761","+14155251111"
31,"NORMAN'S ICE CREAM AND FREEZES","2801 LEAVENWORTH ST ","San Francisco","CA","94133","37.807155","-122.419004",
```

```
"business_id","Score","date","type"
10,"98","20121114","routine"
10,"98","20120403","routine"
10,"100","20110928","routine"
10,"96","20110428","routine"
10,"100","20101210","routine"
12,"100","20121120","routine"
```



http://zipfianacademy.com/maps/h3/

IS AUTOMATION THE ANSWER?

Example: SF Restaurant Violation Data.

http://blog.zipfianacademy.com/post/57158627293/how-to-data-science-mapping-sf-restaurant-inspection

https://data.sfgov.org/Public-Health/Restaurant-Scores/stya-26eb

```
"business_id","date","description"
10,"20121114","Unclean or degraded floors walls or ceilings  [ date violation corrected:  ]"
10,"20120403","Unclean or degraded floors walls or ceilings  [ date violation corrected: 9/20/2012 ]"
10,"20110428","Inadequate and inaccessible handwashing facilities  [ date violation corrected: 6/1/2011 ]"
12,"20120420","Food safety certificate or food handler card not available  [ date violation corrected: 11/20/2012 ]"
17,"20120823","Inadequately cleaned or sanitized food contact surfaces  [ date violation corrected: 9/6/2012 ]"
17,"20120823","High risk food holding temperature   [ date violation corrected: 9/6/2012 ]"
17,"20120823","Unclean nonfood contact surfaces  [ date violation corrected: 9/6/2012 ]"
```

```
"business_id","name","address","city","state","postal_code","latitude","longitude","phone_number"
10,"TIRAMISU KITCHEN","033 BELDEN PL","San Francisco","CA","94104","37.791116","-122.403816",""
12,"KIKKA","250 EMBARCADERO  7/F","San Francisco","CA","94105","37.788613","-122.393894",""
17,"GEORGE'S COFFEE SHOP","2200 OAKDALE AVE ","San Francisco","CA","94124","37.741086","-122.401737","+14155531470"
19,"NRGIZE LIFESTYLE CAFE","1200 VAN NESS AVE, 3RD FLOOR","San Francisco","CA","94109","37.786848","-122.421547",""
24,"OMNI S.F. HOTEL - 2ND FLOOR PANTRY","500 CALIFORNIA ST, 2ND  FLOOR","San Francisco","CA","94104","37.792888","-122.403135",""
29,"CHICO'S PIZZA","131 06TH ST ","San Francisco","CA","94103","37.774722","-122.406761","+14155251111"
31,"NORMAN'S ICE CREAM AND FREEZES","2801 LEAVENWORTH ST ","San Francisco","CA","94133","37.807155","-122.419004",
```

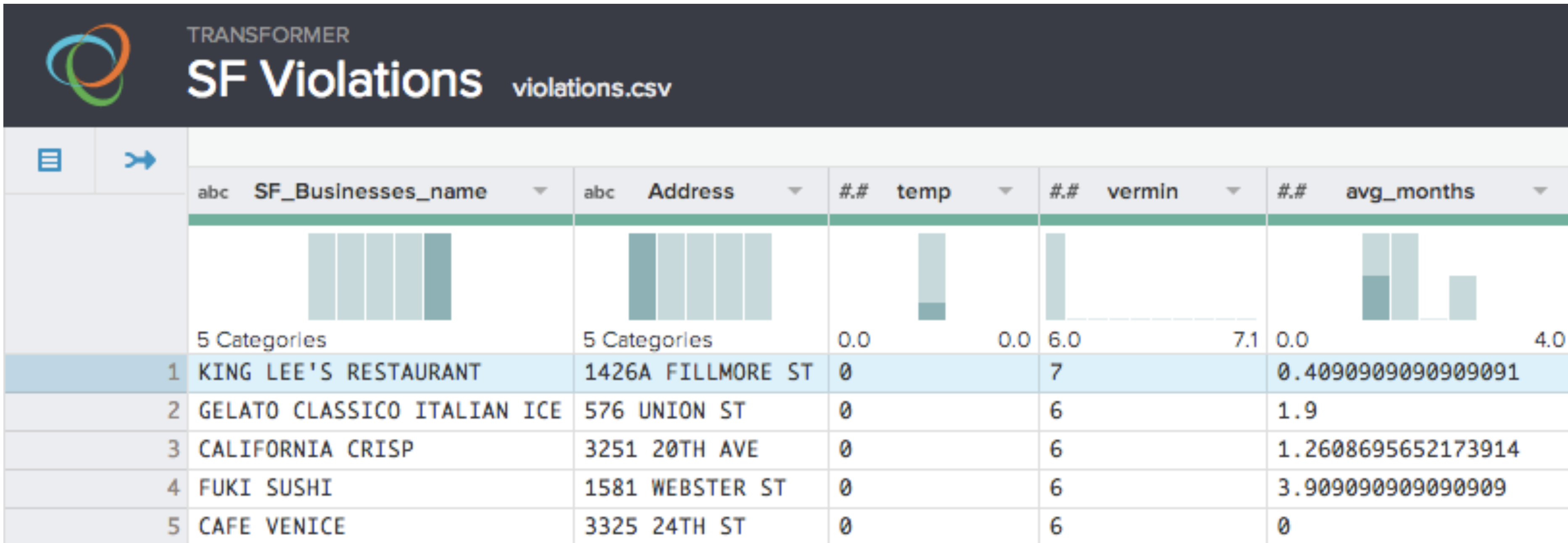
```
"business_id","Score","date","type"
10,"98","20121114","routine"
10,"98","20120403","routine"
10,"100","20110928","routine"
10,"96","20110428","routine"
10,"100","20101210","routine"
12,"100","20121120","routine"
```

```
// BASIC STRUCTURE
splitrows col: column1 on: '\r\n' quote: '\"'
split col: column1 on: ',' limit: 2 quote: '\"'
replace col: * on: `"` with: '' global: true
header

// EXTRACT KEYWORDS
countpattern col: description on: `vermin`
rename col: countpattern_description to: 'vermin'
countpattern col: description on: `temp|hot|therm|cold|cool`
rename col: countpattern_description to: 'temp'

// DATE WHACKING
split col: date at: 4,4
split col: date3 at: 2,2
merge col: date32,'\/',date33,'\/',date2
drop col: date2
drop col: date32
drop col: date33
rename col: column1 to: 'date'
extract col: description after: `: ` before: ` `
rename col: description2 to: 'date_corrected'
derive value: ((year(date_corrected) - year(date)) * 12) + (month(date_corrected) - month(date))
rename col: column1 to: 'delay_months'

// SUMMARIZE, CLEAN, LOOKUP
aggregate value: sum(temp),sum(vermin),mean(delay_months) group: business_id
set col: mean_delay_months value: valid(mean_delay_months, ['Float']) ? mean_delay_months : 0
lookup with: SF Businesses col: {SF Restaurant Violations}.business_id
        key: {SF Businesses}.business_id
```



TRANSFORMER

SF Violations violations.csv

	abc SF_Businesses_name	abc Address	#.# temp	#.# vermin	#.# avg_months
	5 Categories	5 Categories	0.0 0.0	6.0 7.1	0.0 4.0
1	KING LEE'S RESTAURANT	1426A FILLMORE ST	0	7	0.4090909090909091
2	GELATO CLASSICO ITALIAN ICE	576 UNION ST	0	6	1.9
3	CALIFORNIA CRISP	3251 20TH AVE	0	6	1.2608695652173914
4	FUKI SUSHI	1581 WEBSTER ST	0	6	3.909090909090909
5	CAFE VENICE	3325 24TH ST	0	6	0

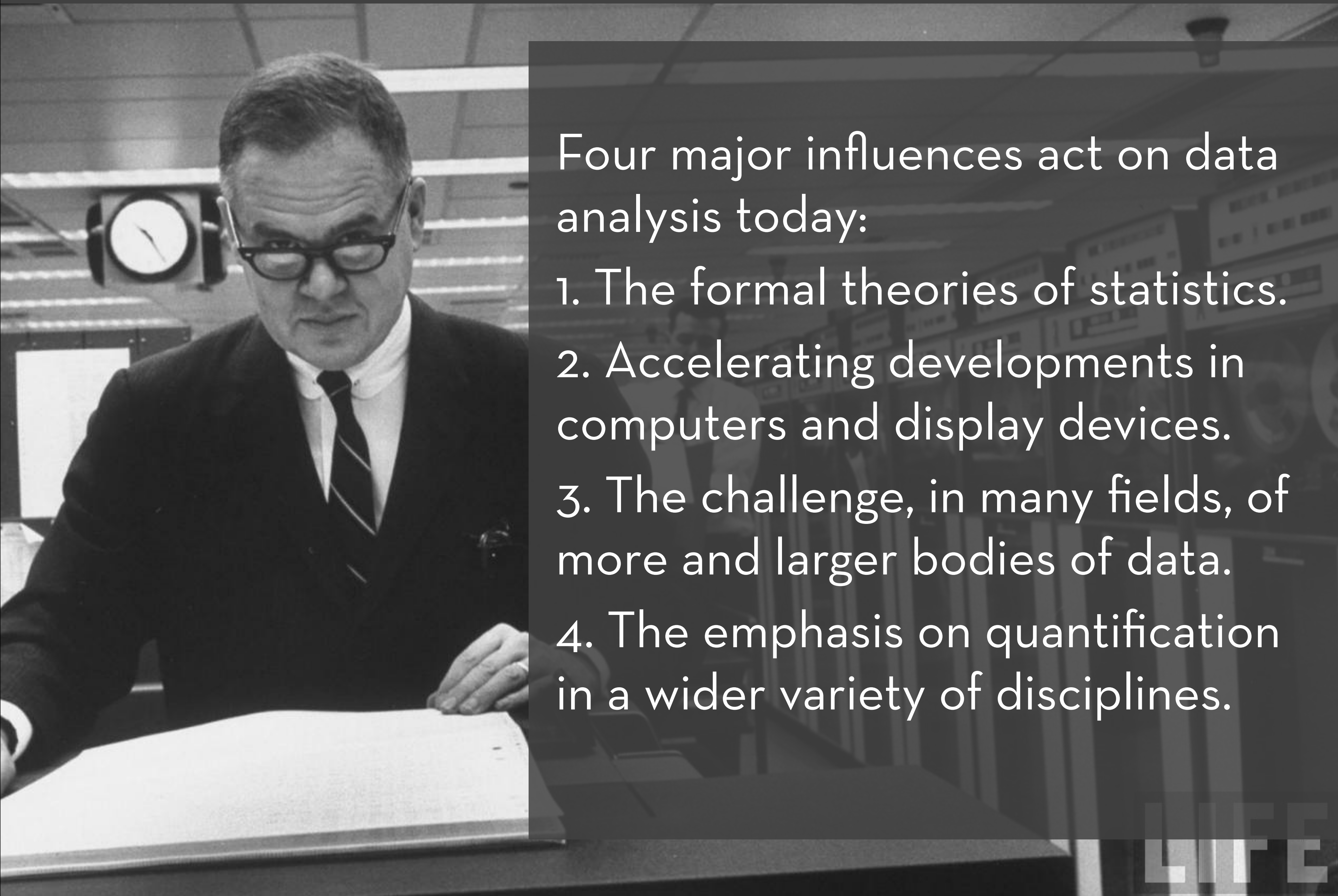
Secrets of the Agile Data Wrangler

Secrets of the Agile Data Wrangler

Secrets of the Agile Data Wrangler

1. **Data is never clean**

Data Analysis & Statistics, Tukey 1965



Four major influences act on data analysis today:

1. The formal theories of statistics.
2. Accelerating developments in computers and display devices.
3. The challenge, in many fields, of more and larger bodies of data.
4. The emphasis on quantification in a wider variety of disciplines.

Nothing - not the careful logic of mathematics, not statistical models and theories, not the awesome arithmetic power of modern computers - nothing can substitute here for the **flexibility of the informed human mind**.

Accordingly, both approaches and techniques need to be structured so as to **facilitate human involvement and intervention**.

Some implications for effective data analysis are: (1) that it is essential to have convenience of **interaction of people and intermediate results** and (2) that at all stages of data analysis, the nature and detail of output, both actual and potential, need to be **matched to the capabilities of the people who use it and want it.**

Visualization

Acquisition

Cleaning

Integration

Visualization

Modeling

Presentation

Dissemination

Acquisition

↓

Cleaning

↓

Integration

↓

Visualization

↓

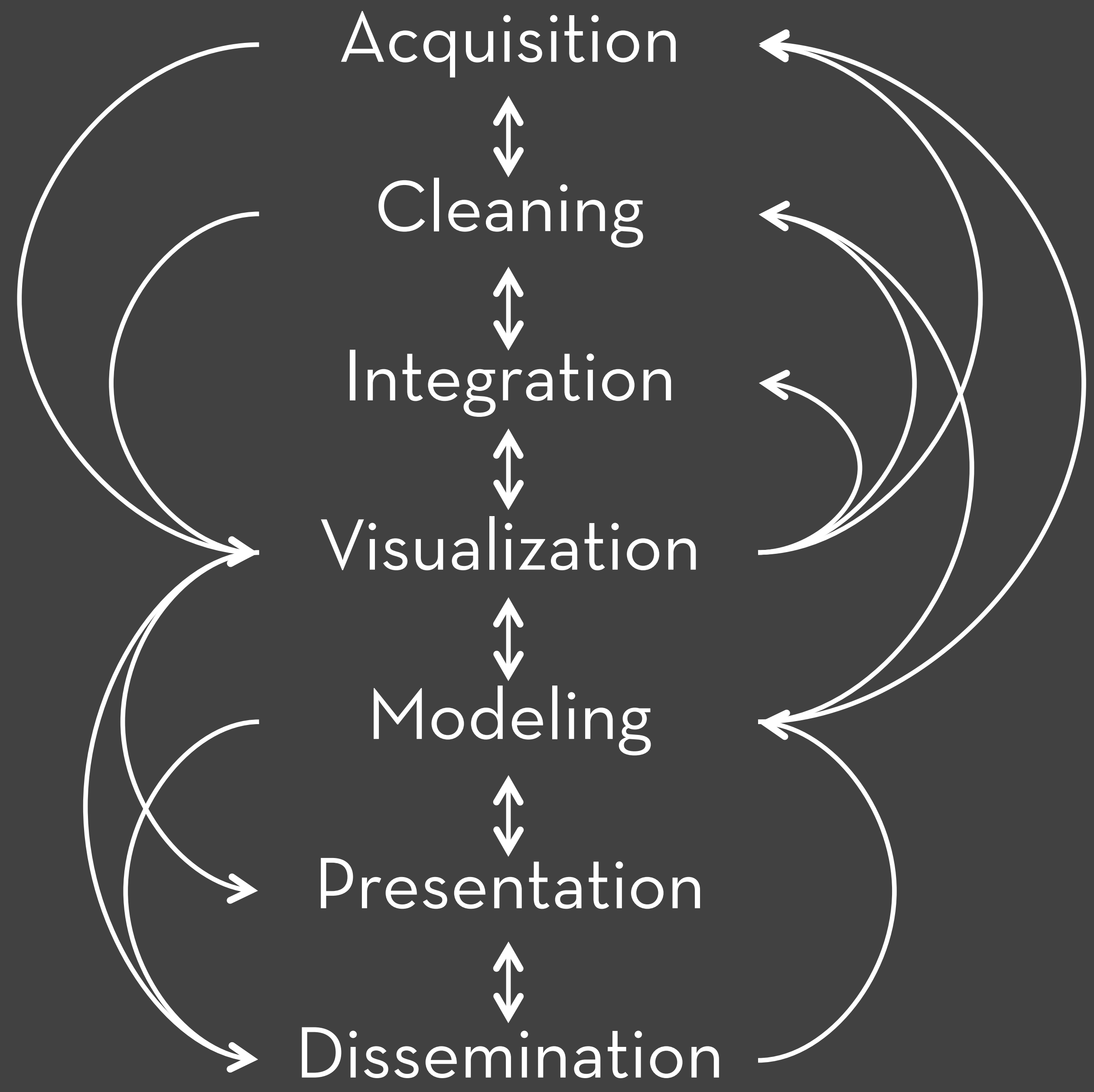
Modeling

↓

Presentation

↓

Dissemination


Acquisition
Cleaning
Integration
Visualization
Modeling
Presentation
Dissemination

Secrets of the Agile Data Wrangler

1. **Data is never clean**

Secrets of the Agile Data Wrangler

1. Data is never clean
2. **Function follows form**

Raw Data: Government Contacts

Bureau of I.A.	
Regional Director	Numbers
Niles C.	Tel: (800)645-8397
	Fax: (907)586-7252
Jean H.	Tel: (918)781-4600
	Fax: (918)781-4604
Frank K.	Tel: (615)564-6500
	Fax: (615)564-6701

Bureau of I.A.	
Regional Director	Numbers
Niles C.	Tel: (800)645-8397
	Fax: (907)586-7252
Jean H.	Tel: (918)781-4600
	Fax: (918)781-4604
Frank K.	Tel: (615)564-6500
	Fax: (615)564-6701

Filter First Two Rows

Bureau of I.A.	
Regional Director	Numbers
Niles C.	Tel: (800)645-8397
	Fax: (907)586-7252
Jean H.	Tel: (918)781-4600
	Fax: (918)781-4604
Frank K.	Tel: (615)564-6500
	Fax: (615)564-6701

→

Niles C.	Tel: (800)645-8397
	Fax: (907)586-7252
Jean H.	Tel: (918)781-4600
	Fax: (918)781-4604
Frank K.	Tel: (615)564-6500
	Fax: (615)564-6701

Niles C.	Tel: (800)645-8397
	Fax: (907)586-7252
Jean H.	Tel: (918)781-4600
	Fax: (918)781-4604
Frank K.	Tel: (615)564-6500
	Fax: (615)564-6701

Split on ":" Delimiter

Niles C.	Tel: (800)645-8397
	Fax: (907)586-7252
Jean H.	Tel: (918)781-4600
	Fax: (918)781-4604
Frank K.	Tel: (615)564-6500
	Fax: (615)564-6701

→

Niles C.	Tel	(800)645-8397
	Fax	(907)586-7252
Jean H.	Tel	(918)781-4600
	Fax	(918)781-4604
Frank K.	Tel	(615)564-6500
	Fax	(615)564-6701

Niles C.	Tel	(800)645-8397
	Fax	(907)586-7252
Jean H.	Tel	(918)781-4600
	Fax	(918)781-4604
Frank K.	Tel	(615)564-6500
	Fax	(615)564-6701

Delete Empty Rows

Niles C.	Tel	(800)645-8397
	Fax	(907)586-7252
Jean H.	Tel	(918)781-4600
	Fax	(918)781-4604
Frank K.	Tel	(615)564-6500
	Fax	(615)564-6701

→

Niles C.	Tel	(800)645-8397
	Fax	(907)586-7252
Jean H.	Tel	(918)781-4600
	Fax	(918)781-4604
Frank K.	Tel	(615)564-6500
	Fax	(615)564-6701

Niles C.	Tel	(800)645-8397
	Fax	(907)586-7252
Jean H.	Tel	(918)781-4600
	Fax	(918)781-4604
Frank K.	Tel	(615)564-6500
	Fax	(615)564-6701

Fill Values Down

Niles C.	Tel	(800)645-8397
	Fax	(907)586-7252
Jean H.	Tel	(918)781-4600
	Fax	(918)781-4604
Frank K.	Tel	(615)564-6500
	Fax	(615)564-6701

Niles C.	Tel	(800)645-8397
Niles C.	Fax	(907)586-7252
Jean H.	Tel	(918)781-4600
Jean H.	Fax	(918)781-4604
Frank K.	Tel	(615)564-6500
Frank K.	Fax	(615)564-6701

Niles C.	Tel	(800)645-8397
Niles C.	Fax	(907)586-7252
Jean H.	Tel	(918)781-4600
Jean H.	Fax	(918)781-4604
Frank K.	Tel	(615)564-6500
Frank K.	Fax	(615)564-6701

Pivot Number on Type

Niles C.	Tel	(800)645-8397
Niles C.	Fax	(907)586-7252
Jean H.	Tel	(918)781-4600
Jean H.	Fax	(918)781-4604
Frank K.	Tel	(615)564-6500
Frank K.	Fax	(615)564-6701

→

	Tel	Fax
Niles C.	(800)645-8397	(907)586-7252
Jean H.	(918)781-4600	(918)781-4604
Frank K.	(615)564-6500	(615)564-6701

Reformatted Data

	Tel	Fax
Niles C.	(800)645-8397	(907)586-7252
Jean H.	(918)781-4600	(918)781-4604
Frank K.	(615)564-6500	(615)564-6701

Map Transforms: Per-Tuple Actions

Map Transforms: Per-Tuple Actions

Rows	Fill Values Left, Right
	Filter

Map Transforms: Per-Tuple Actions

Rows	Fill Values Left, Right
	Filter
Cells	Extract
	Replace
	Edit

Map Transforms: Per-Tuple Actions

Rows	Fill Values Left, Right
	Filter
Cells	Extract
	Replace
	Edit
Columns	Drop
	Split
	Merge
	Shift

Table Transforms

Table	Promote, Demote Header
	Fill Values Down, Up
	Transpose
	Pivot
	Fold

Table Transforms: Reshaping

Table Transforms: Reshaping

	Boys	Girls
Australia	1	2
Austria	3	4
Belgium	5	6
China	7	8

Table Transforms: Reshaping

Fold

	Boys	Girls
Australia	1	2
Austria	3	4
Belgium	5	6
China	7	8

Australia	Boys	1
Australia	Girls	2
Austria	Boys	3
Austria	Girls	4
Belgium	Boys	5
Belgium	Girls	6
China	Boys	7
China	Girls	8

Table Transforms: Reshaping

Fold

	Boys	Girls
Australia	1	2
Austria	3	4
Belgium	5	6
China	7	8

Australia	Boys	1
Australia	Girls	2
Austria	Boys	3
Austria	Girls	4
Belgium	Boys	5
Belgium	Girls	6
China	Boys	7
China	Girls	8

Pivot

Bureau of Justice Statistics - Data Online
http://bjs.ojp.usdoj.gov/

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

Reported crime in California

Year	Population	Property crime rate	Burglary rate	Larceny-theft rate	Motor vehicle theft rate
2004	35842038	3423.9	686.1	2033.1	704.8
2005	36154147	3321	692.9	1915	712
2006	36457549	3175.2	676.9	1831.5	666.8
2007	36553215	3032.6	648.4	1784.1	600.2
2008	36756666	2940.3	646.8	1769.8	523.8

Reported crime in Colorado

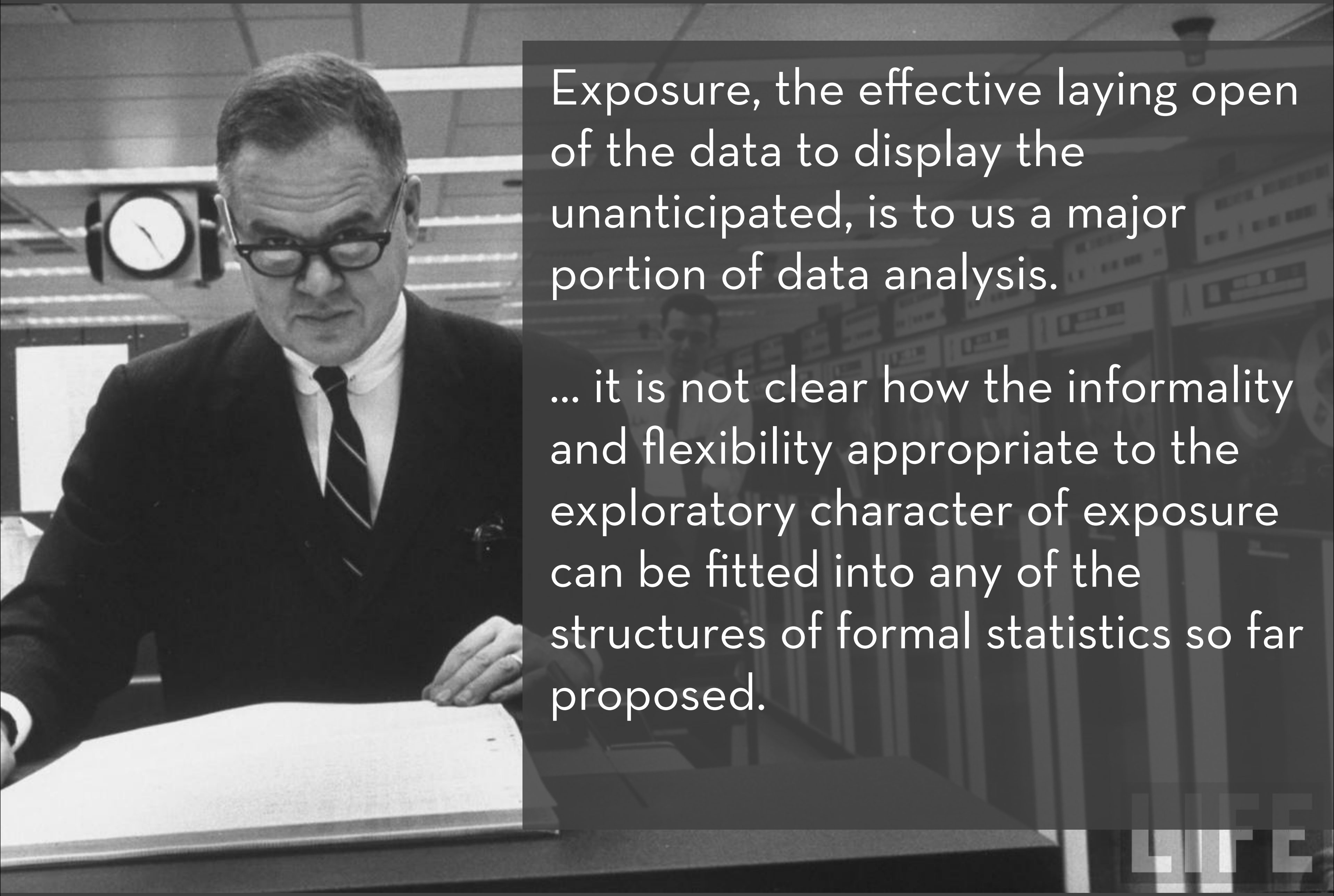
Year	Population	Property crime rate	Burglary rate	Larceny-theft rate	Motor vehicle theft rate
2004	4601821	3918.5	717.3	2679.5	521.6

Secrets of the Agile Data Wrangler

1. Data is never clean
2. **Function follows form**

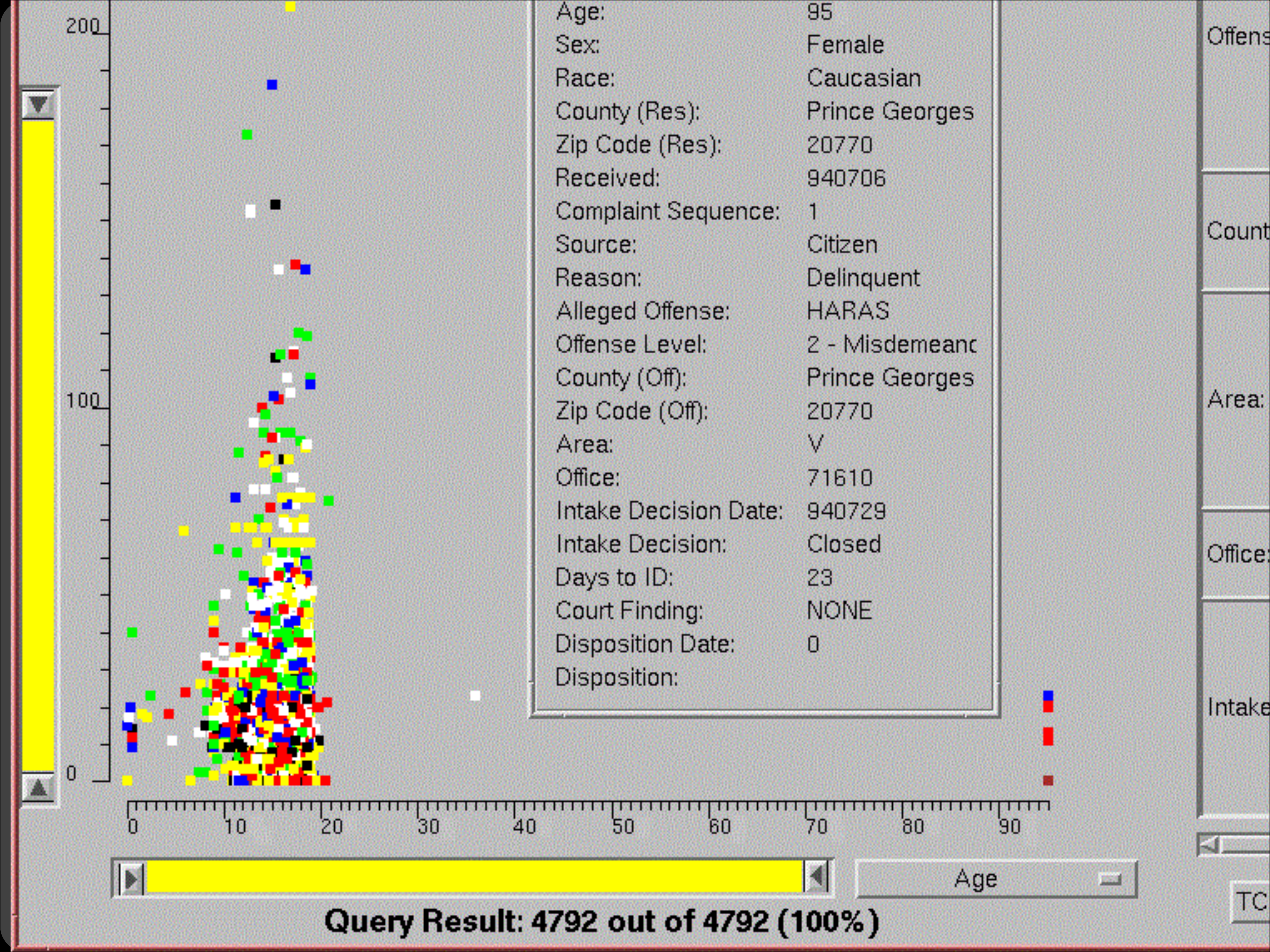
Secrets of the Agile Data Wrangler

1. Data is never clean
2. Function follows form
3. **Expose your data**



Exposure, the effective laying open of the data to display the unanticipated, is to us a major portion of data analysis.

… it is not clear how the informality and flexibility appropriate to the exploratory character of exposure can be fitted into any of the structures of formal statistics so far proposed.



Age: 95
Sex: Female
Race: Caucasian
County (Res): Prince Georges
Zip Code (Res): 20770
Received: 940706
Complaint Sequence: 1
Source: Citizen
Reason: Delinquent
Alleged Offense: HARAS
Offense Level: 2 - Misdemeanc
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:
200
100
0
0
10
20
30
40
50
60
70
80
90
Age
Offens
Count
Area:
Office:
Intake
TC
Query Result: 4792 out of 4792 (100%)

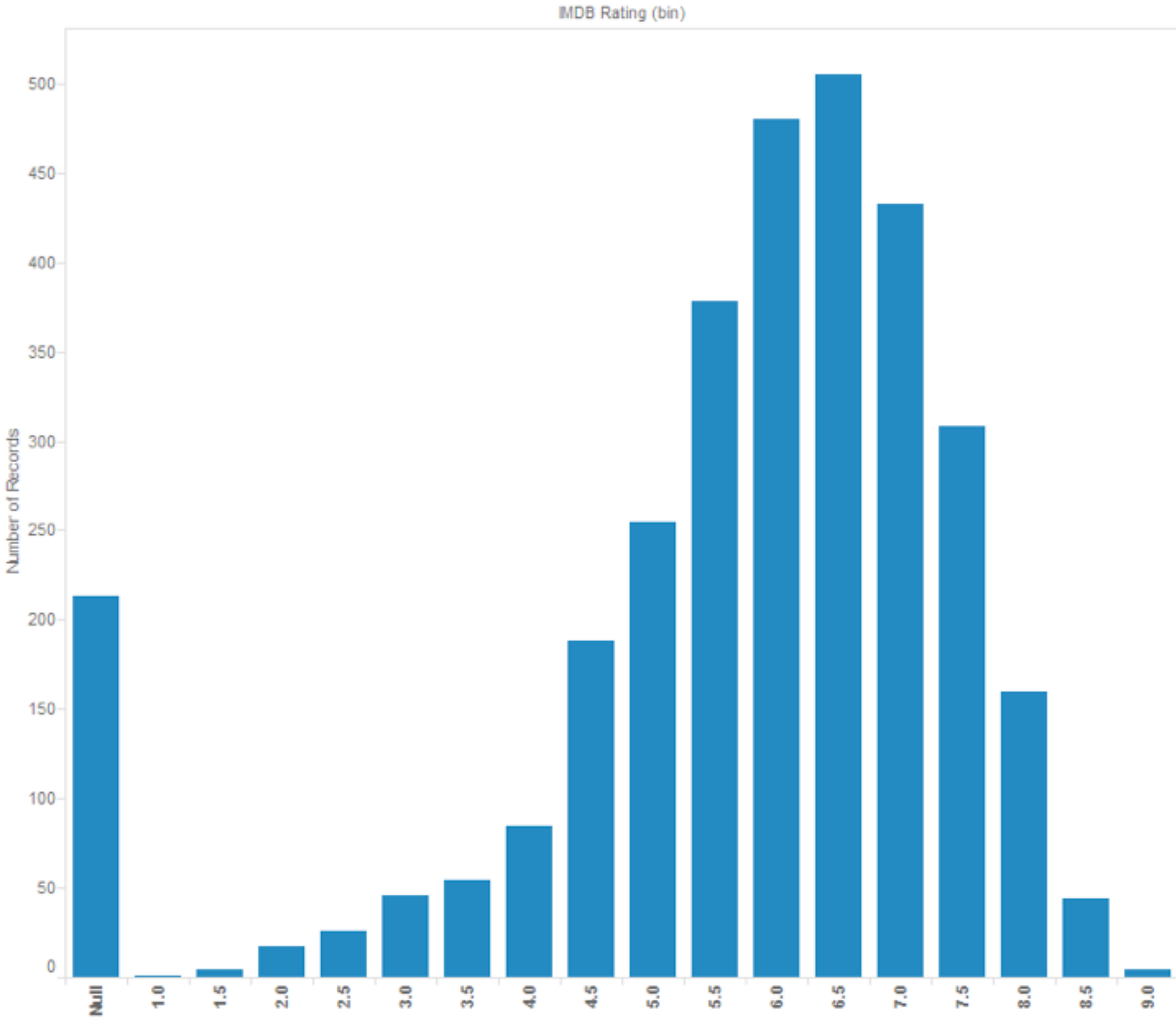
Example:

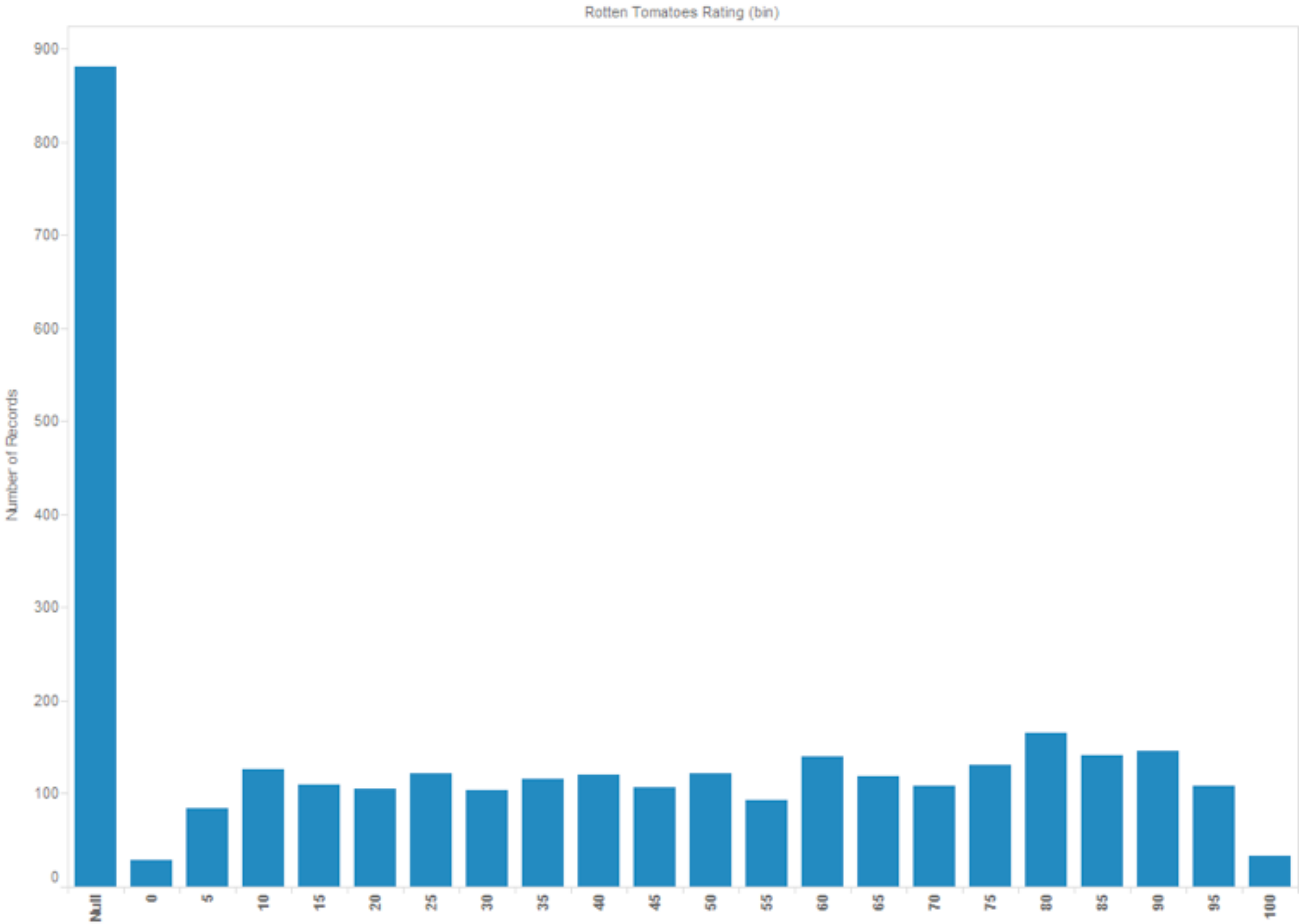
Motion Pictures Data

Motion Pictures Data

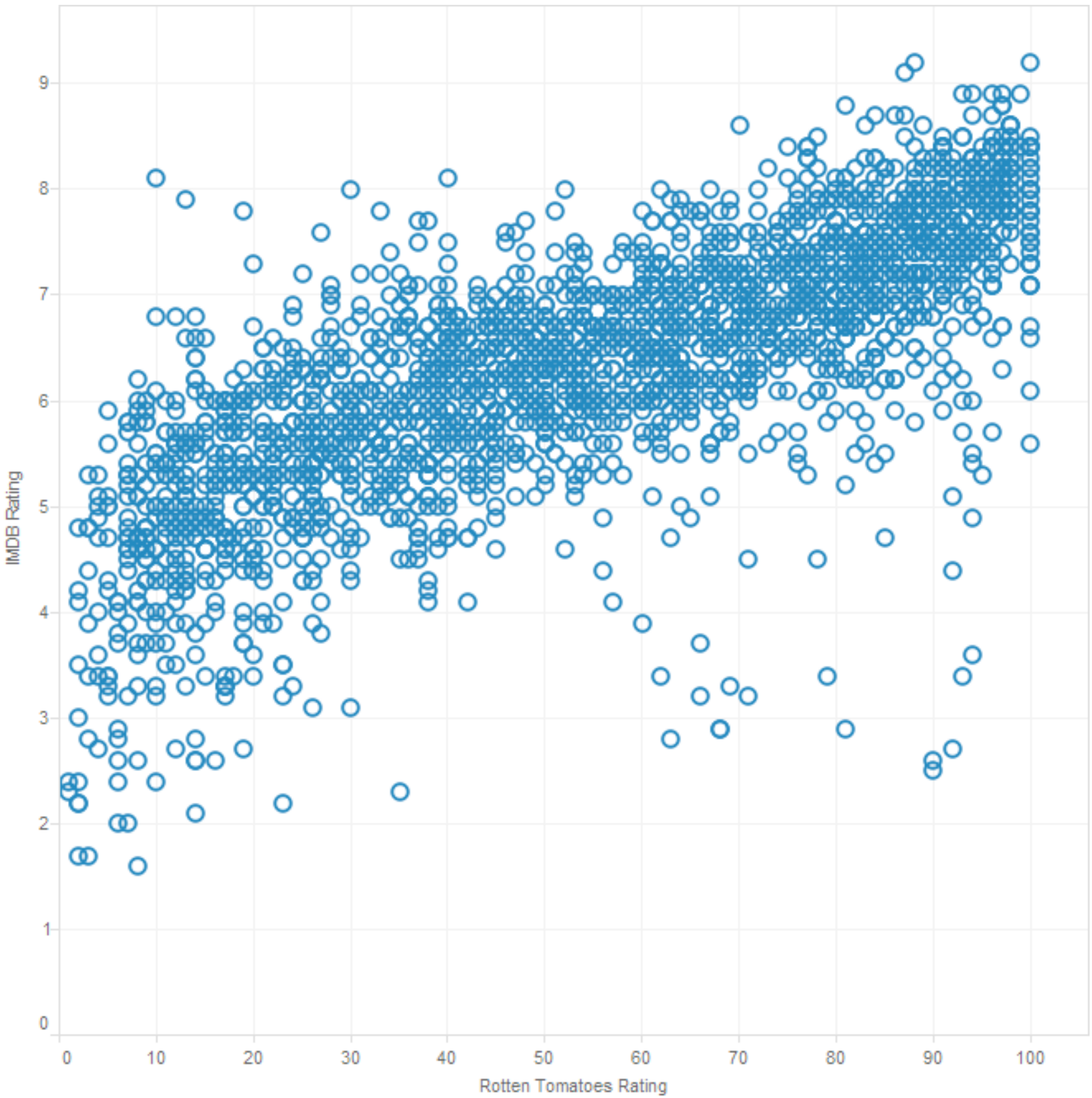
Title	String
IMDB Rating	Number
Rotten Tomatoes Rating	Number
MPAA Rating	String
Release Date	Date
Worldwide Gross	Number

Integrated data from IMDB, Rotten Tomatoes and The Numbers, joined on film title.

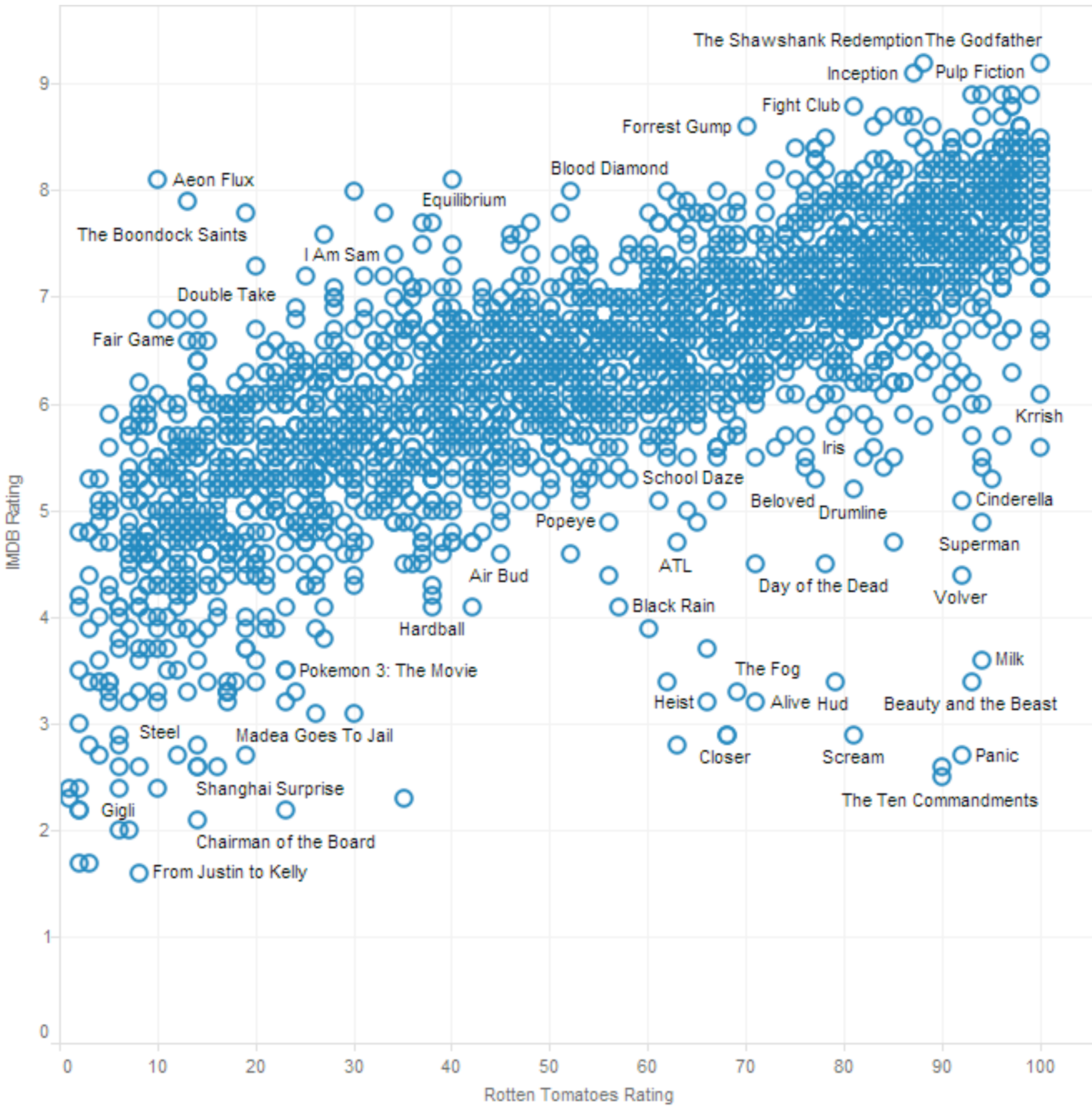

IMDB Rating (bin)
Number of Records
500
450
400
350
300
250
200
150
100
50
0
Null
1.0
1.5
2.0
2.5
3.0
3.5
4.0
4.5
5.0
5.5
6.0
6.5
7.0
7.5
8.0
8.5
9.0



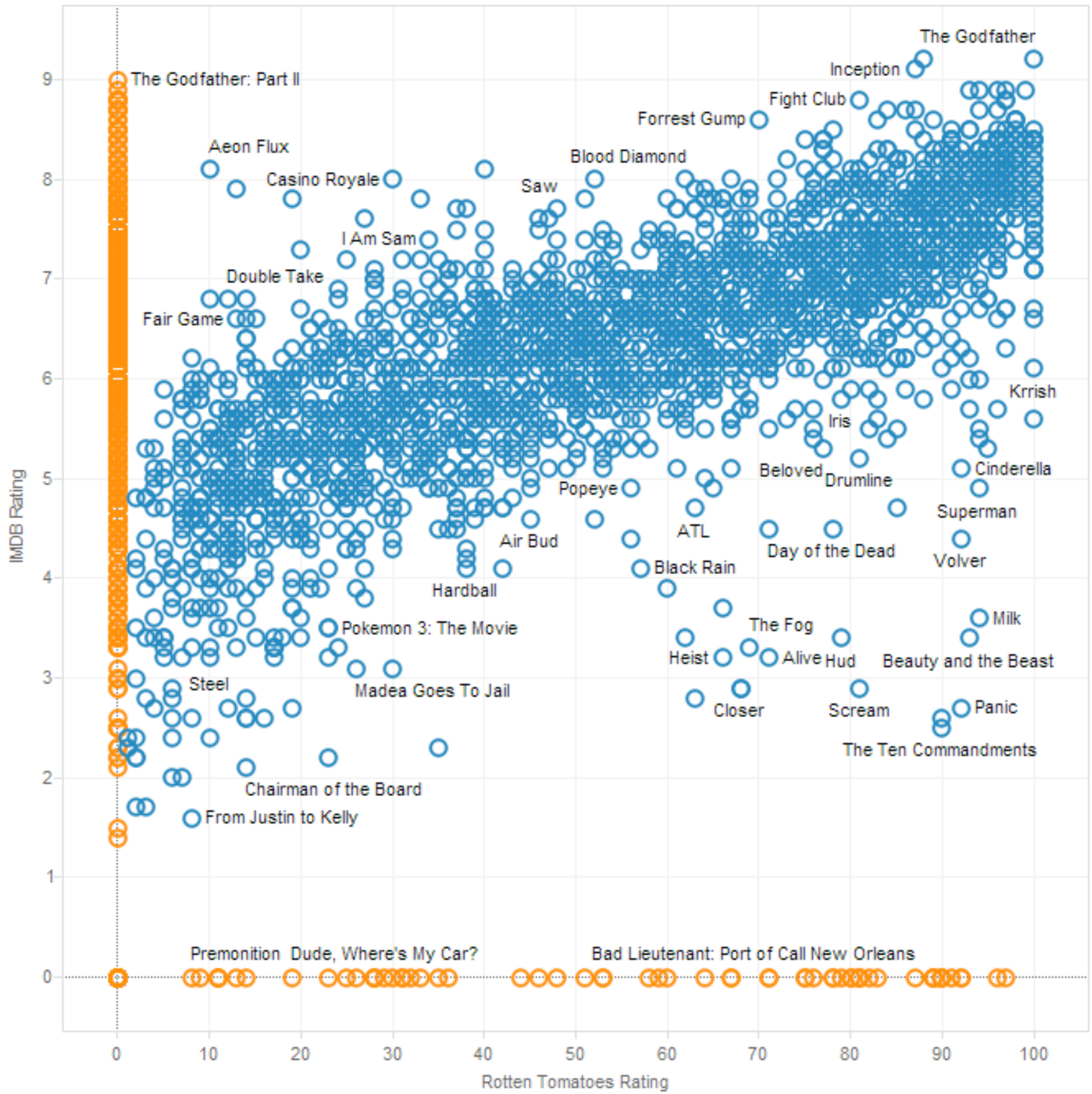
Rotten Tomatoes Rating (bin)
Number of Records
900
800
700
600
500
400
300
200
100
0
Null
0
5
10
15
20
25
30
35
40
45
50
55
60
65
70
75
80
85
90
95
100



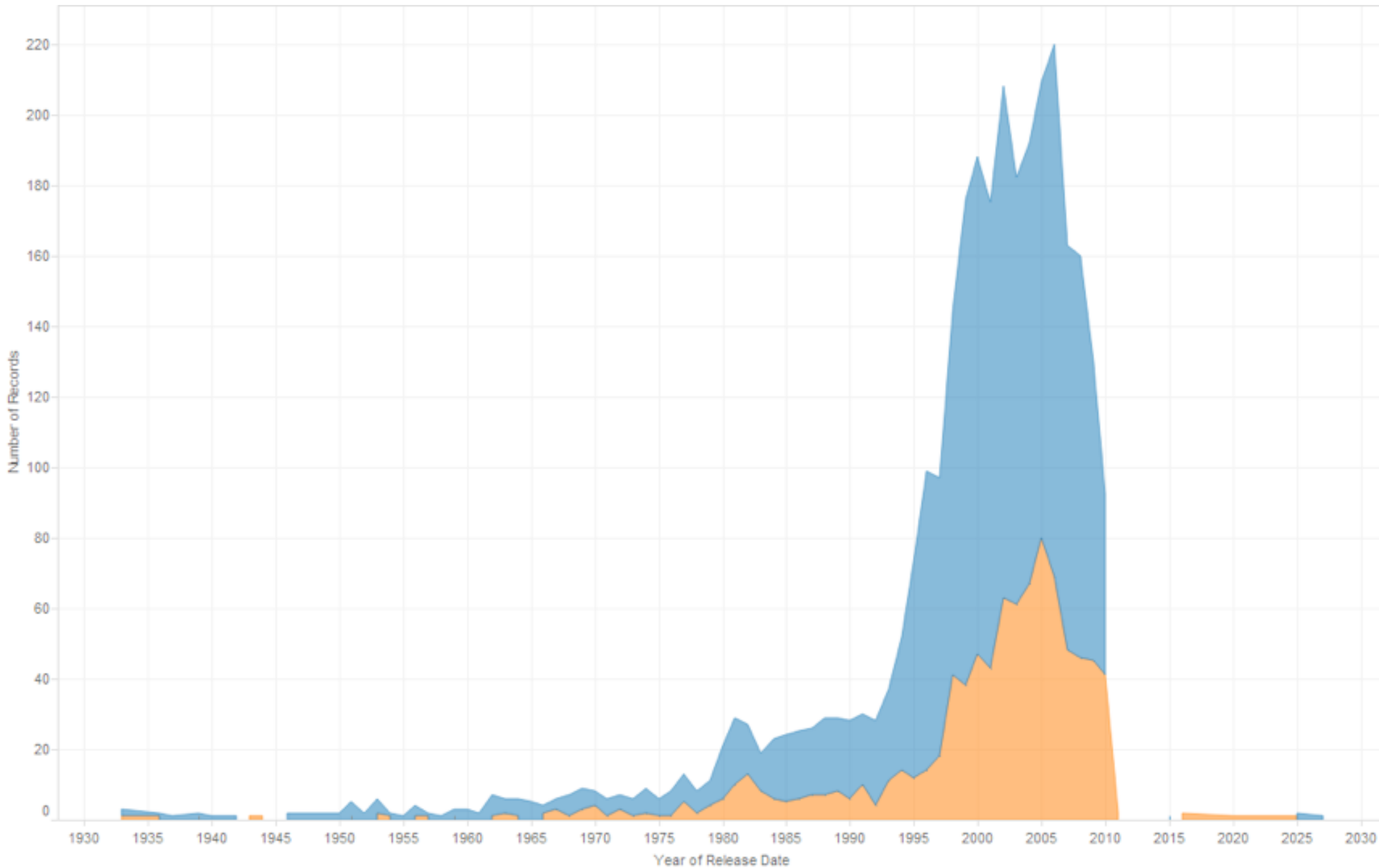
IMDB Rating
Rotten Tomatoes Rating
0
1
2
3
4
5
6
7
8
9
0
10
20
30
40
50
60
70
80
90
100



The Shawshank Redemption
The Godfather
Inception
Pulp Fiction
Fight Club
Forrest Gump
Blood Diamond
Aeon Flux
Equilibrium
The Boondock Saints
I Am Sam
Double Take
Fair Game
Krrish
Iris
School Daze
Beloved
Drumline
Cinderella
Popeye
Superman
ATL
Air Bud
Day of the Dead
Volver
Black Rain
Hardball
Pokemon 3: The Movie
The Fog
Milk
Heist
Alive
Hud
Beauty and the Beast
Steel
Madea Goes To Jail
Closer
Scream
Panic
Shanghai Surprise
The Ten Commandments
Gigli
Chairman of the Board
From Justin to Kelly
IMDB Rating
Rotten Tomatoes Rating
0
1
2
3
4
5
6
7
8
9
10
20
30
40
50
60
70
80
90
100



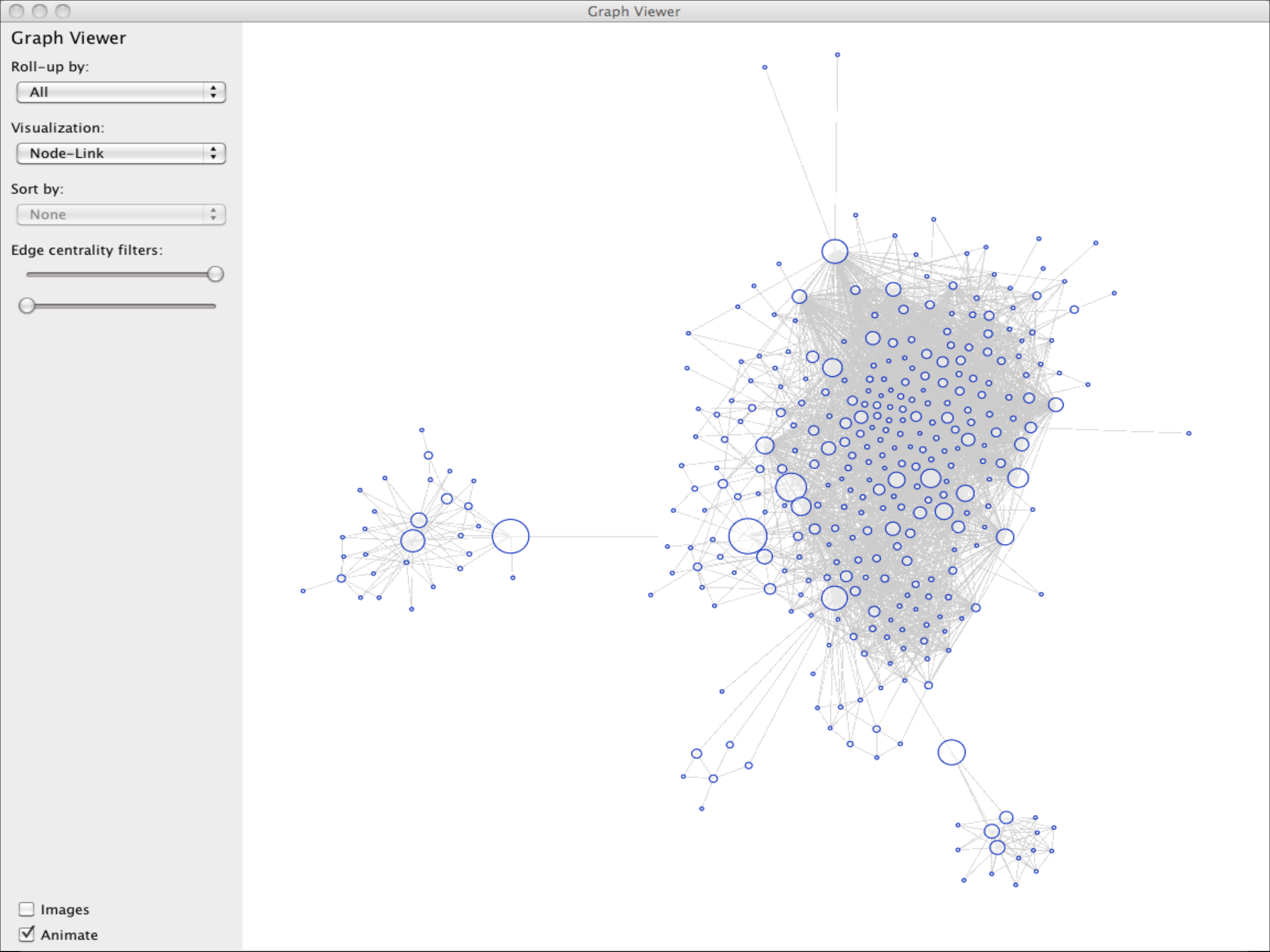
The Godfather
Inception
The Godfather: Part II
Fight Club
Forrest Gump
Aeon Flux
Blood Diamond
Casino Royale
Saw
I Am Sam
Double Take
Fair Game
Krrish
Iris
Beloved
Drumline
Cinderella
Popeye
Superman
ATL
Air Bud
Day of the Dead
Volver
Black Rain
Hardball
Milk
Pokemon 3: The Movie
The Fog
Heist
Alive
Hud
Beauty and the Beast
Steel
Madea Goes To Jail
Closer
Scream
Panic
The Ten Commandments
Chairman of the Board
From Justin to Kelly
Premonition
Dude, Where's My Car?
Bad Lieutenant: Port of Call New Orleans
IMDB Rating
Rotten Tomatoes Rating

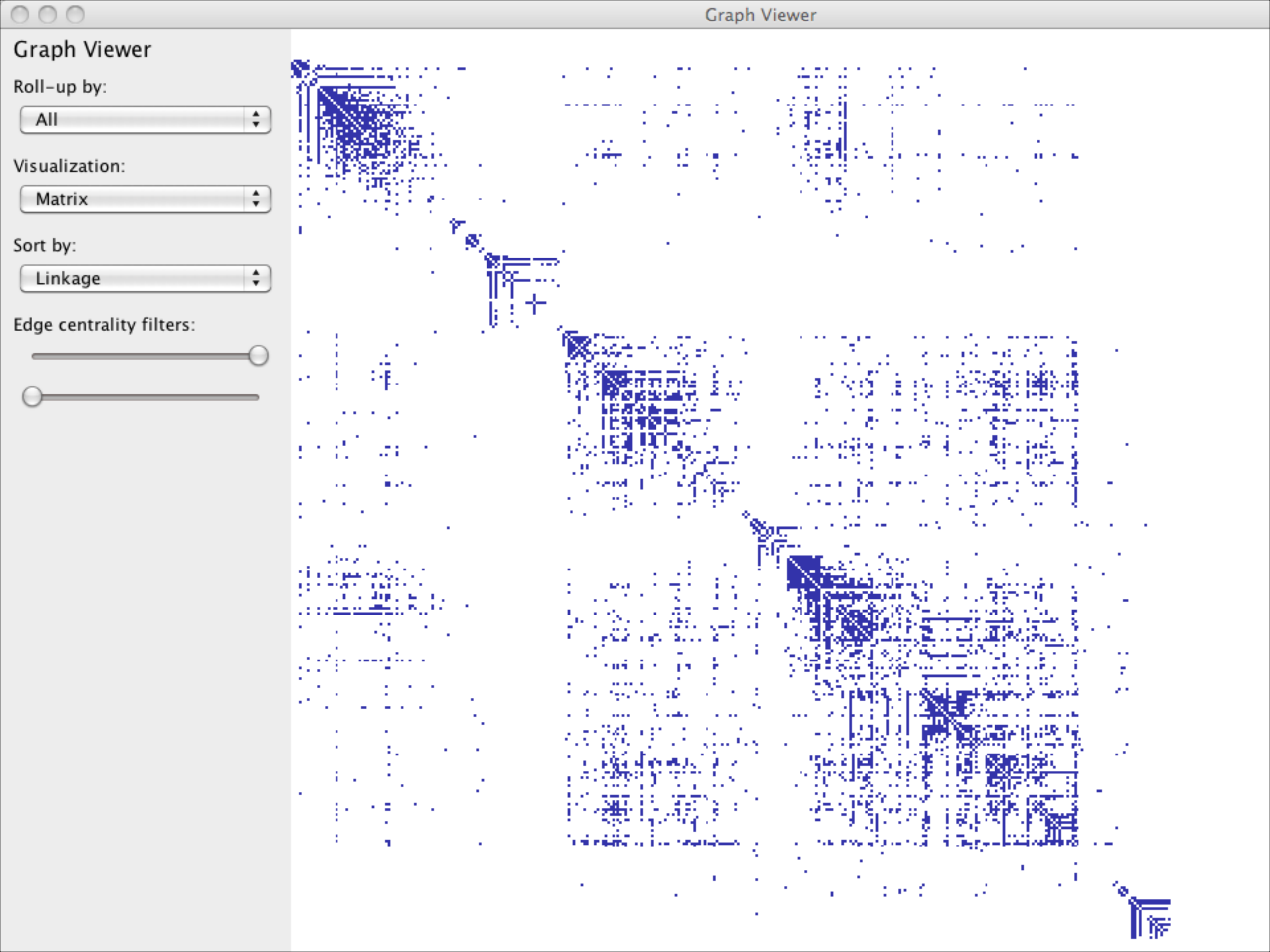


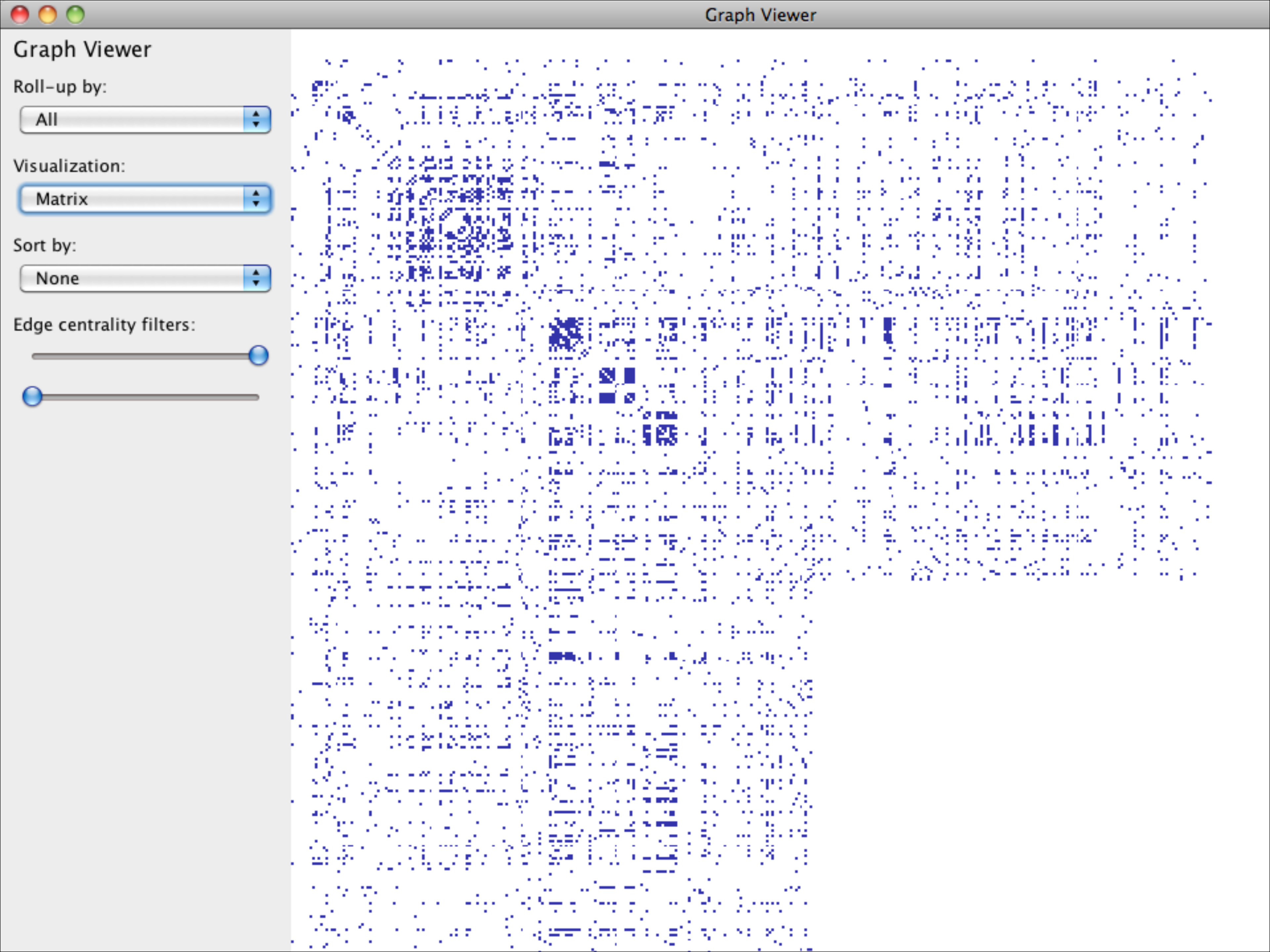
220
200
180
160
140
120
100
80
60
40
20
0
Number of Records
1930
1935
1940
1945
1950
1955
1960
1965
1970
1975
1980
1985
1990
1995
2000
2005
2010
2015
2020
2025
2030
Year of Release Date

Example:

Facebook Social Graph


Graph Viewer
Graph Viewer
Roll-up by:
All
Visualization:
Node-Link
Sort by:
None
Edge centrality filters:
Images
Animate


Graph Viewer
Graph Viewer
Roll-up by:
All
Visualization:
Matrix
Sort by:
Linkage
Edge centrality filters:

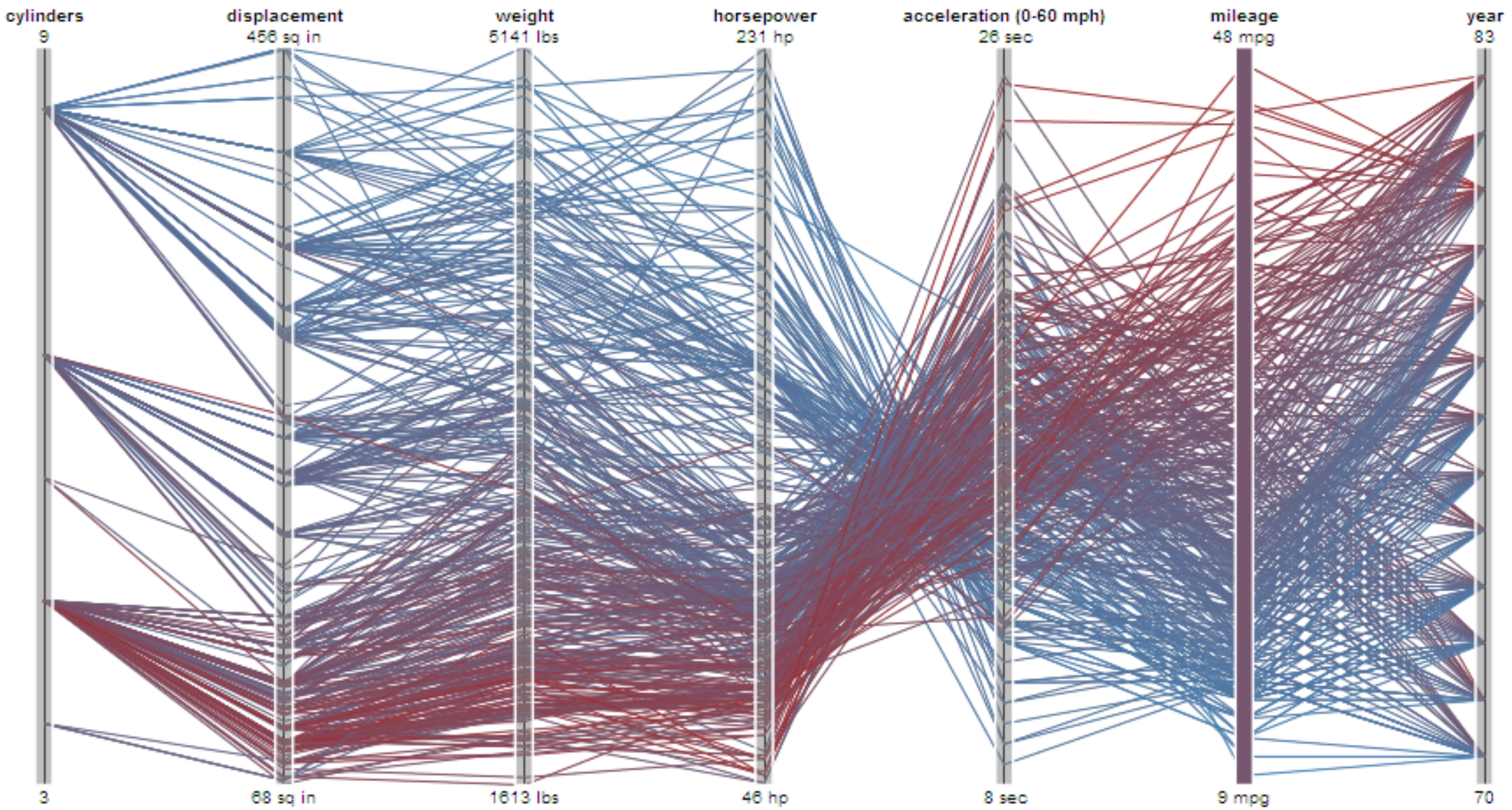

Graph Viewer
Graph Viewer
Roll-up by:
All
Visualization:
Matrix
Sort by:
None
Edge centrality filters:

Count Friends by School

<table>
<tr><td>Berkeley</td><td>|||||||||||||||||||||||||||||||</td></tr>
<tr><td>Cornell</td><td>||||||||||||||</td></tr>
<tr><td>Cornell College</td><td>|||</td></tr>
<tr><td>Cornell University</td><td>||||||</td></tr>
<tr><td>Harvard</td><td>||||||||||||</td></tr>
<tr><td>Harvard University</td><td>|||||||</td></tr>
<tr><td>Stanford</td><td>||||||||||||||||||||</td></tr>
<tr><td>Stanford University</td><td>||||||||||</td></tr>
<tr><td>UC Berkeley</td><td>|||||||||||||||||||||</td></tr>
<tr><td>University of California at Berkeley</td><td>|||||||||||||||</td></tr>
<tr><td>University of California, Berkeley</td><td>||||||||||||||||||</td></tr>
</table>

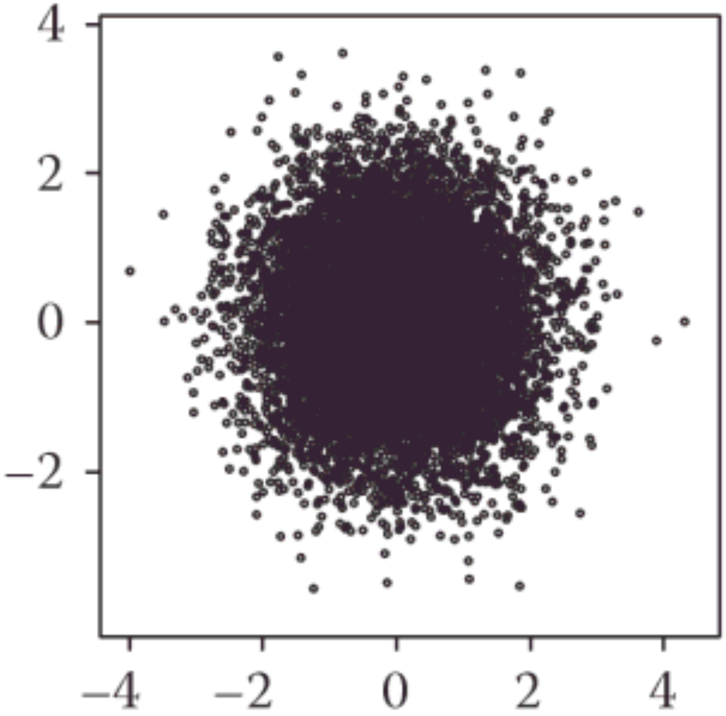
Challenges

High-Dimensional Data

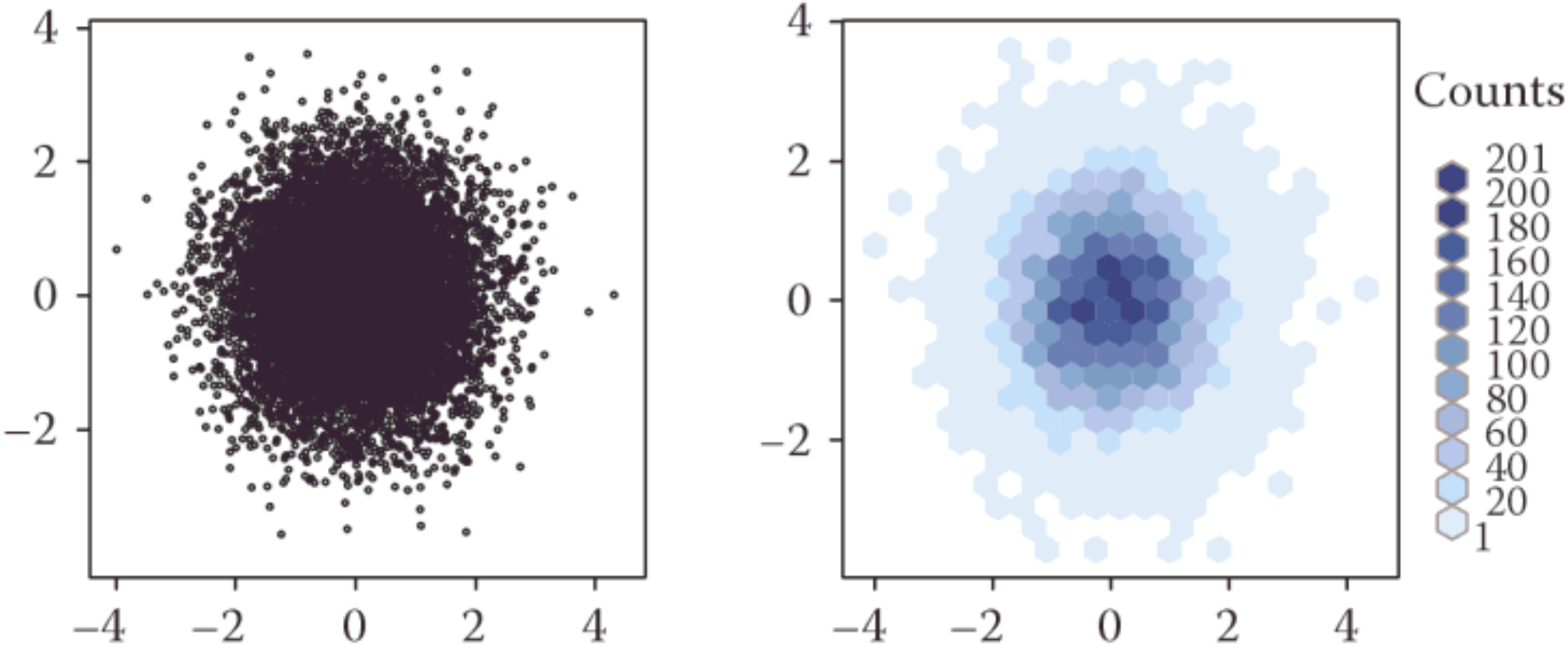


Parallel Coordinates [Inselberg]

Scalable Representations



Scalable Representations



Binned Scatterplot, adapted from Carr, 1987

Secrets of the Agile Data Wrangler

1. Data is never clean
2. Function follows form
3. **Expose your data**

Secrets of the Agile Data Wrangler

1. Data is never clean
2. Function follows form
3. Expose your data
4. **Statistics & graphics: better together**

It is too much to ask for close and effective guidance for data analysis from any highly formalized structure, either now or in the near future.

Data analysis can gain much from formal statistics, but only if the connection is kept adequately loose.

Set A		Set B		Set C		Set D	
X	Y	X	Y	X	Y	X	Y
10	8.04	10	9.14	10	7.46	8	6.58
8	6.95	8	8.14	8	6.77	8	5.76
13	7.58	13	8.74	13	12.74	8	7.71
9	8.81	9	8.77	9	7.11	8	8.84
11	8.33	11	9.26	11	7.81	8	8.47
14	9.96	14	8.1	14	8.84	8	7.04
6	7.24	6	6.13	6	6.08	8	5.25
4	4.26	4	3.1	4	5.39	19	12.5
12	10.84	12	9.11	12	8.15	8	5.56
7	4.82	7	7.26	7	6.42	8	7.91
5	5.68	5	4.74	5	5.73	8	6.89

Set A X	Set A Y	Set B X	Set B Y	Set C X	Set C Y	Set D X	Set D Y
10	8.04	10	9.14	10	7.46	8	6.58
8	6.95	8	8.14	8	6.77	8	5.76
13	7.58	13	8.74	13	12.74	8	7.71
9	8.81	9	8.77	9	7.11	8	8.84
11	8.33	11	9.26	11	7.81	8	8.47
14	9.96	14	8.1	14	8.84	8	7.04
6	7.24	6	6.13	6	6.08	8	5.25
4	4.26	4	3.1	4	5.39	19	12.5
12	10.84	12	9.11	12	8.15	8	5.56
7	4.82	7	7.26	7	6.42	8	7.91
5	5.68	5	4.74	5	5.73	8	6.89

Summary Statistics

$u_X = 9.0$ $\sigma_X = 3.317$

$u_Y = 7.5$ $\sigma_Y = 2.03$

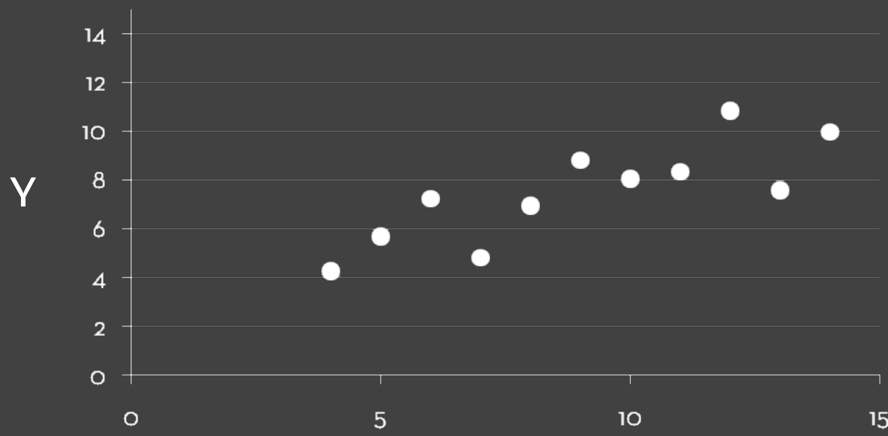
Linear Regression

$Y = 3 + 0.5X$

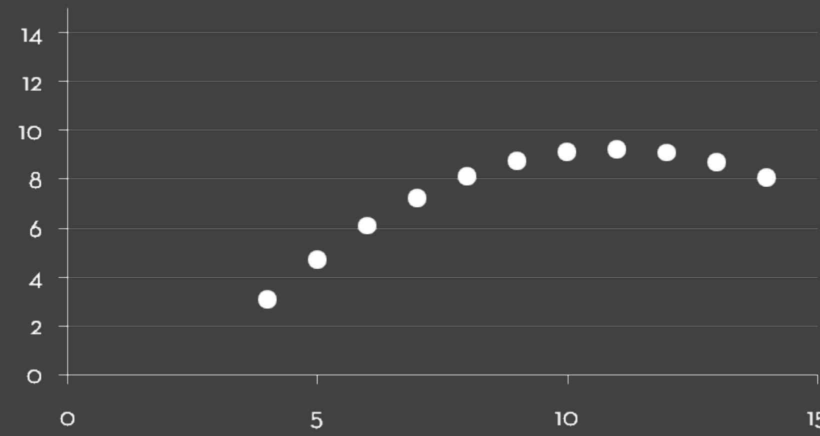
$R^2 = 0.67$

Anscombe 1973

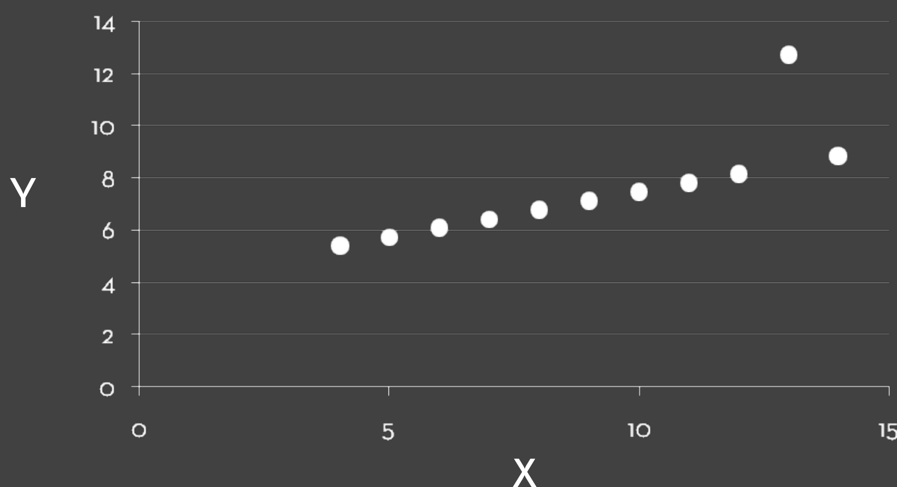
Set A



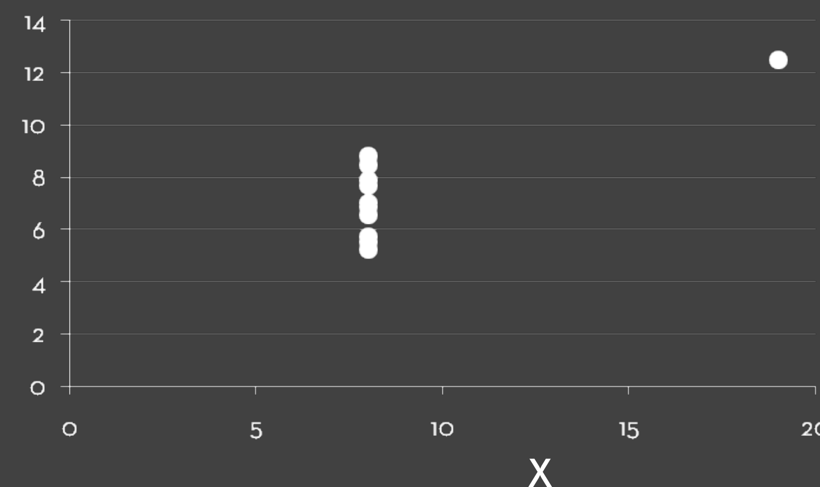
Set B

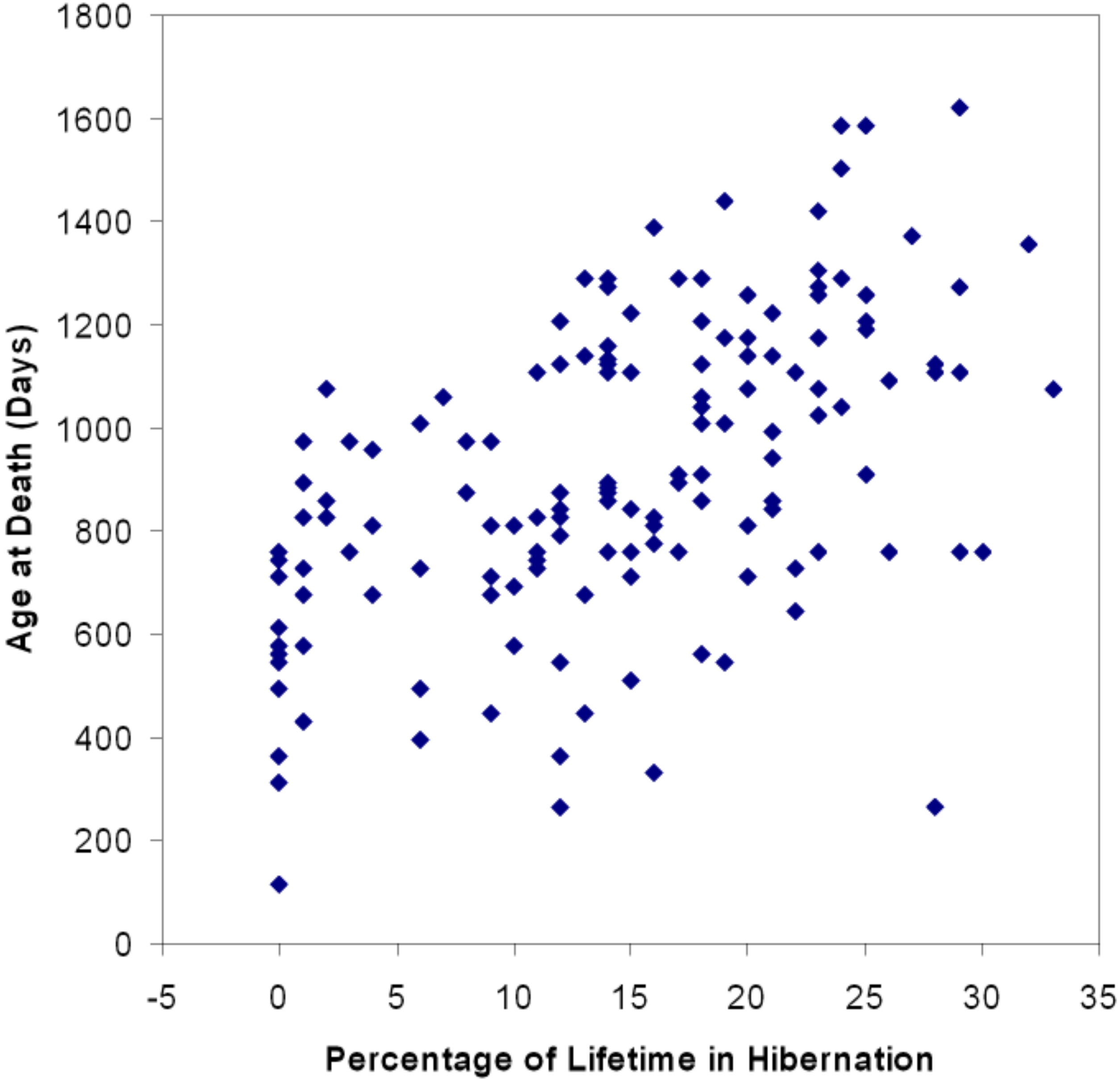


Set C

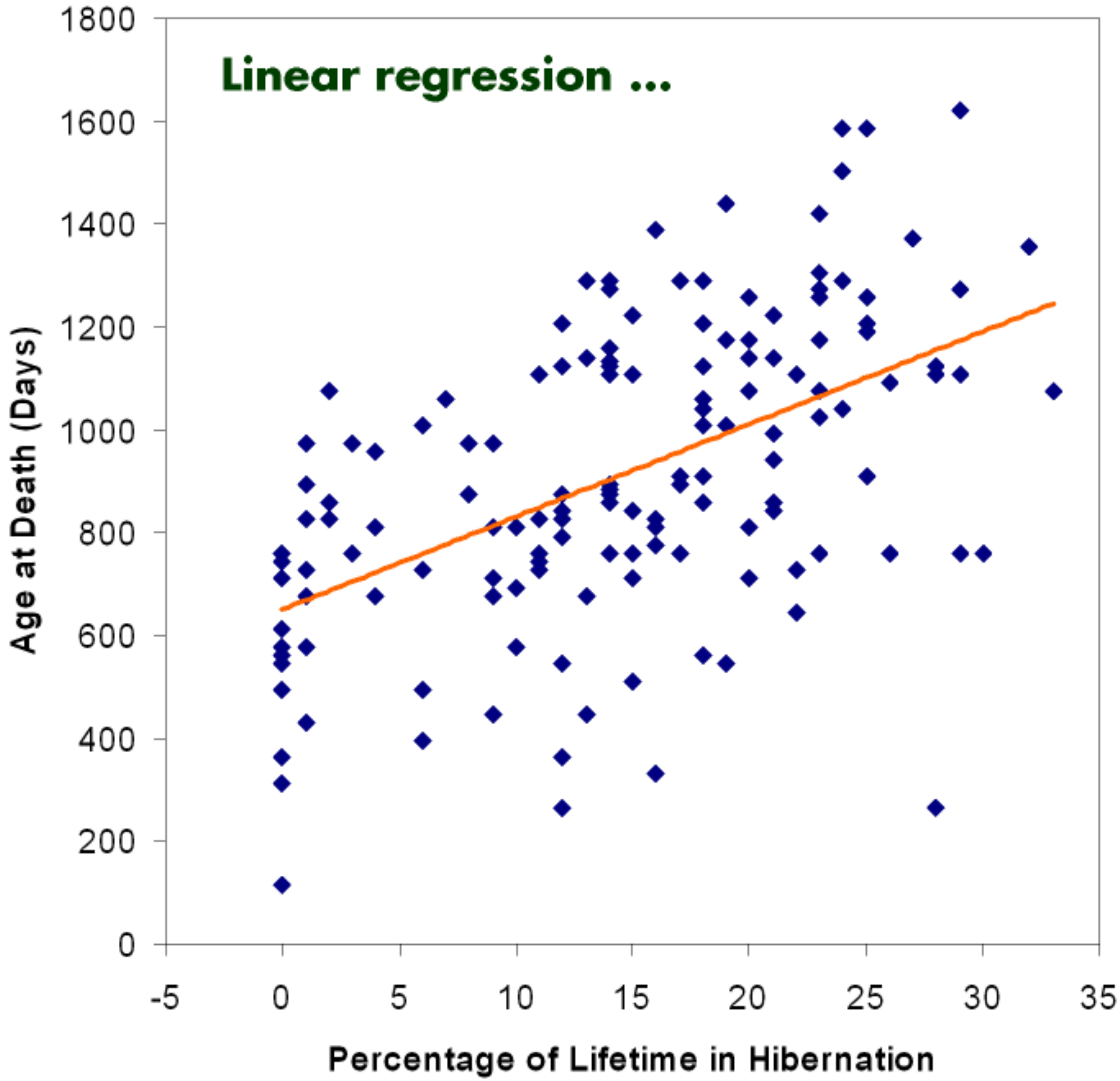


Set D

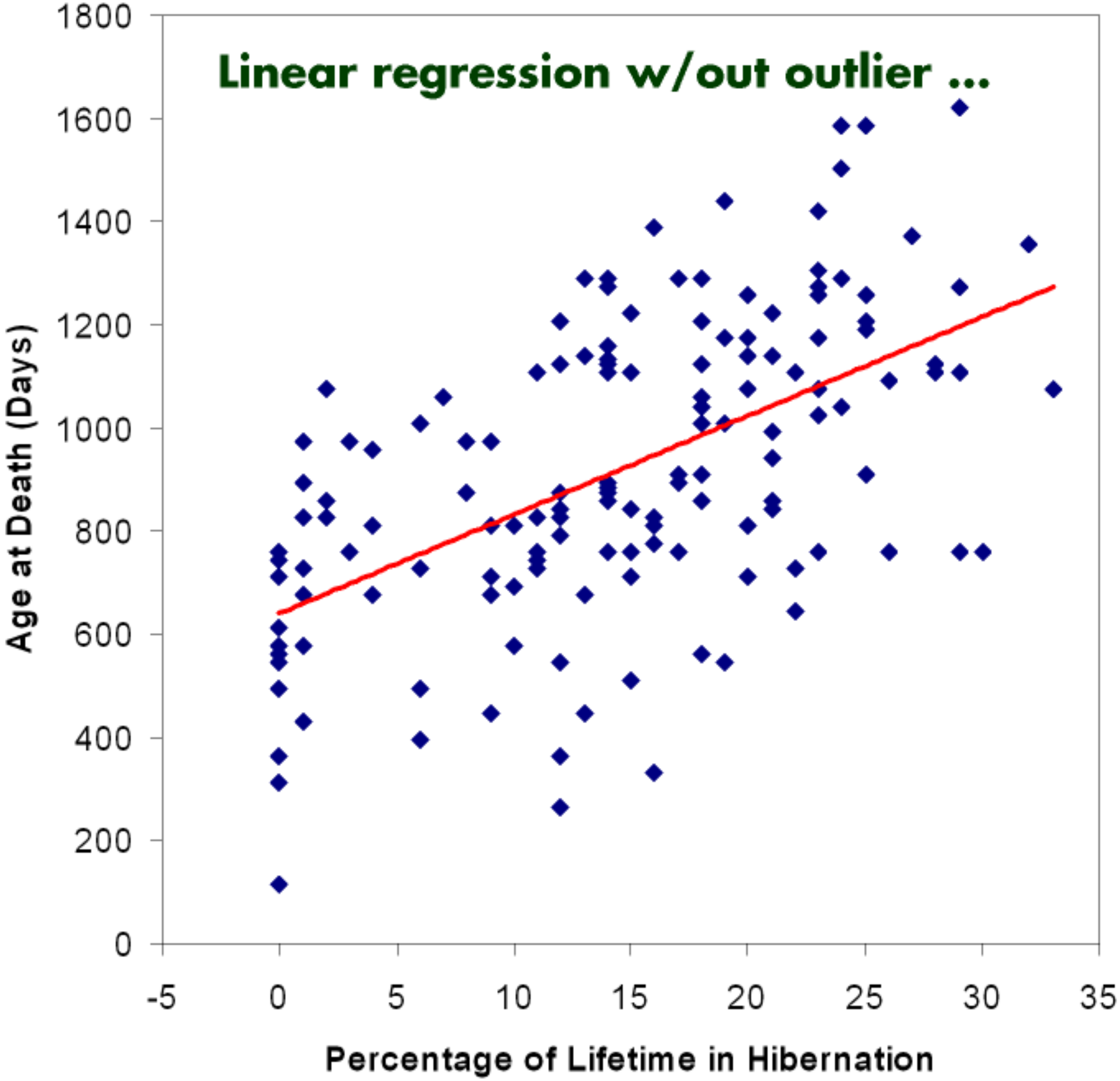




[The Elements of Graphing Data. Cleveland 94]



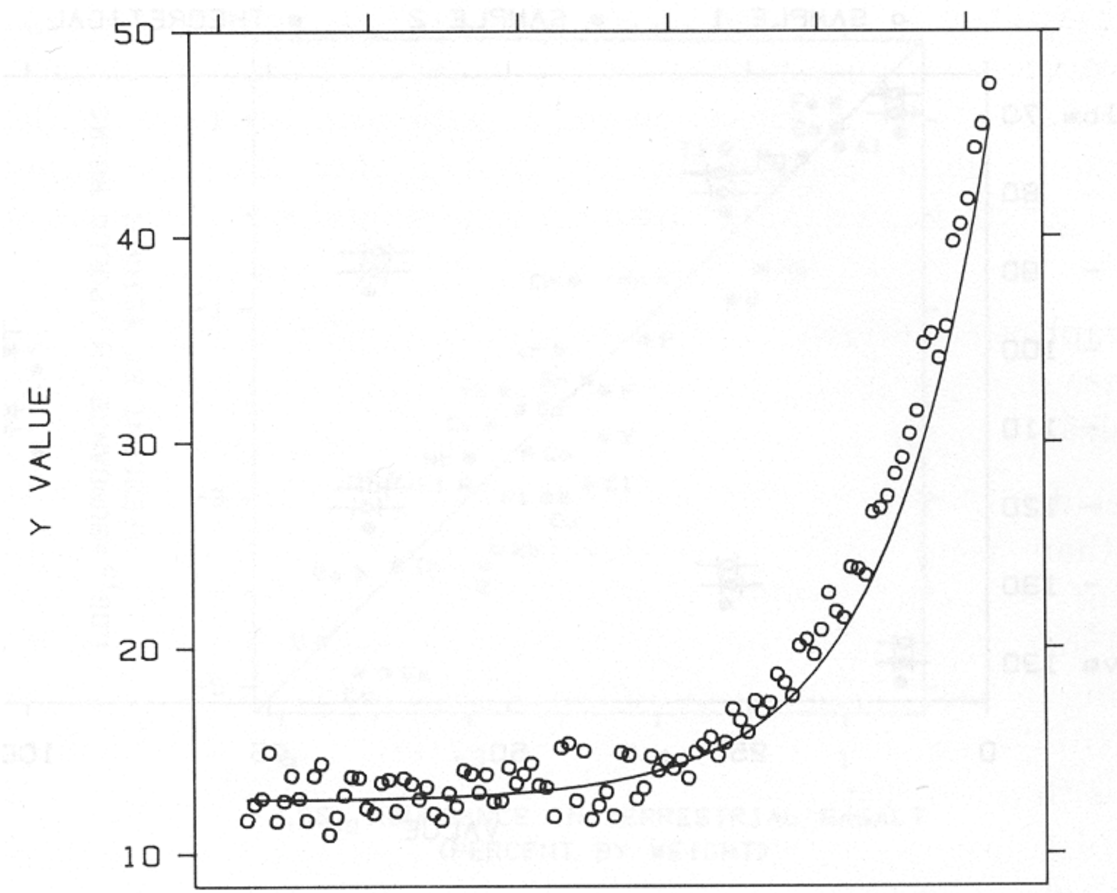
[The Elements of Graphing Data. Cleveland 94]



[The Elements of Graphing Data. Cleveland 94]

Transforming Data

How well does the curve fit data?

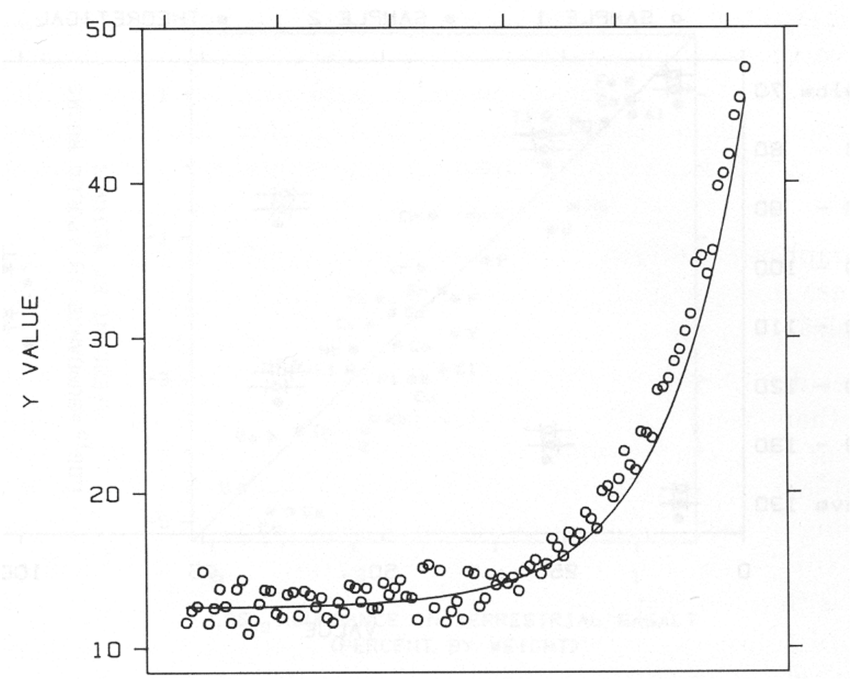


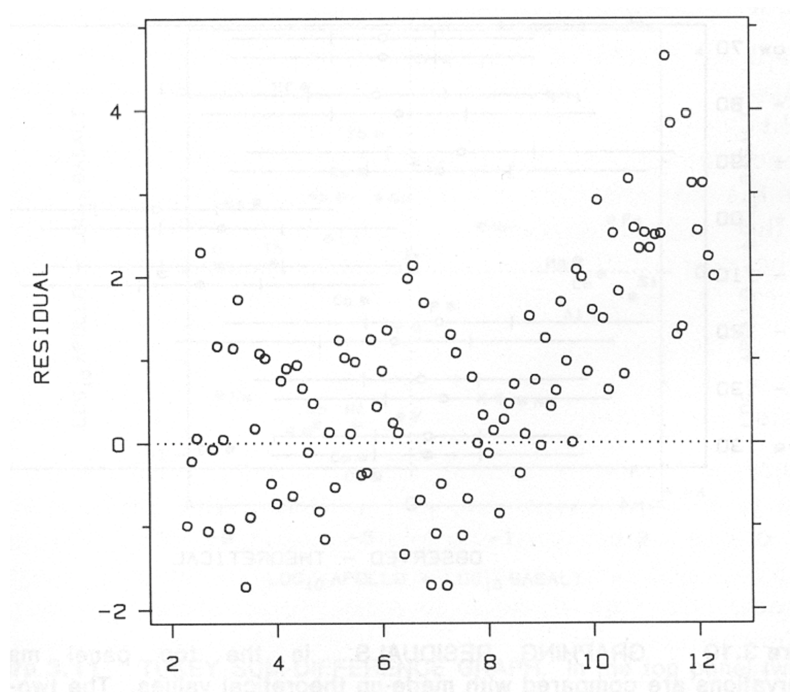
[Cleveland 85]

Plot the Residuals

Plot vertical distance from best fit curve

Residual graph shows accuracy of fit

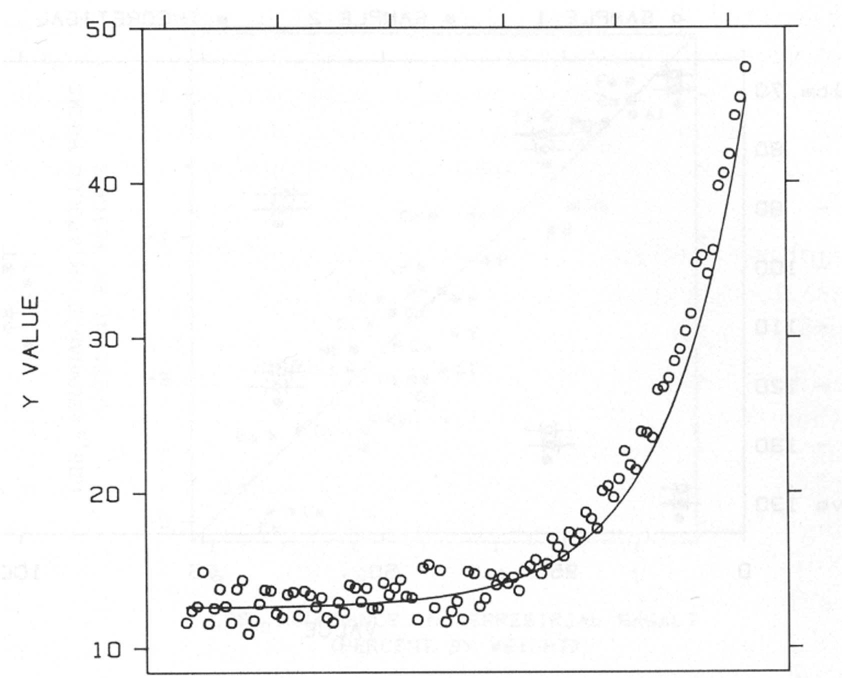




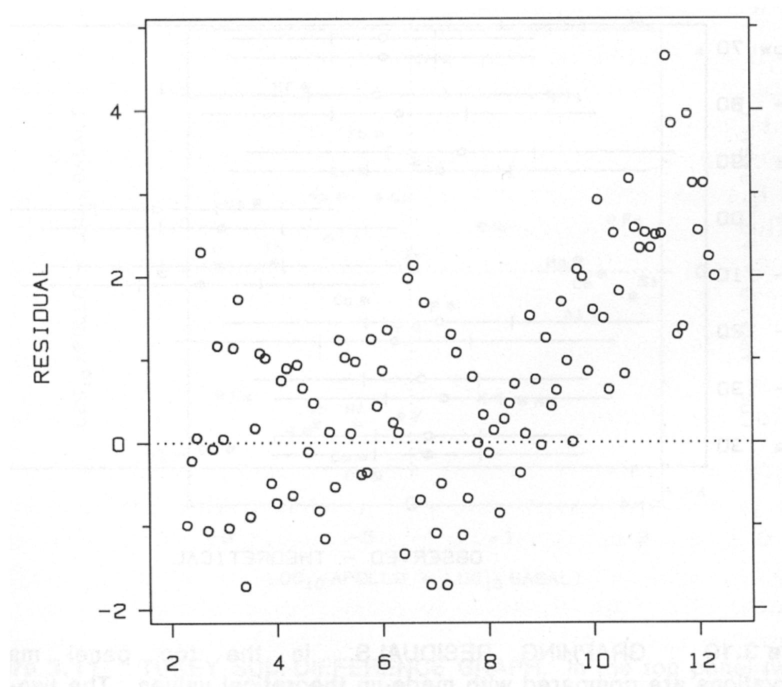
[Cleveland 85]

Multiple Plotting Options

Plot model in data space



Plot data in model space



[Cleveland 85]

What’s an outlier?

Far From the Center

Center

Dispersion

Far From the Center

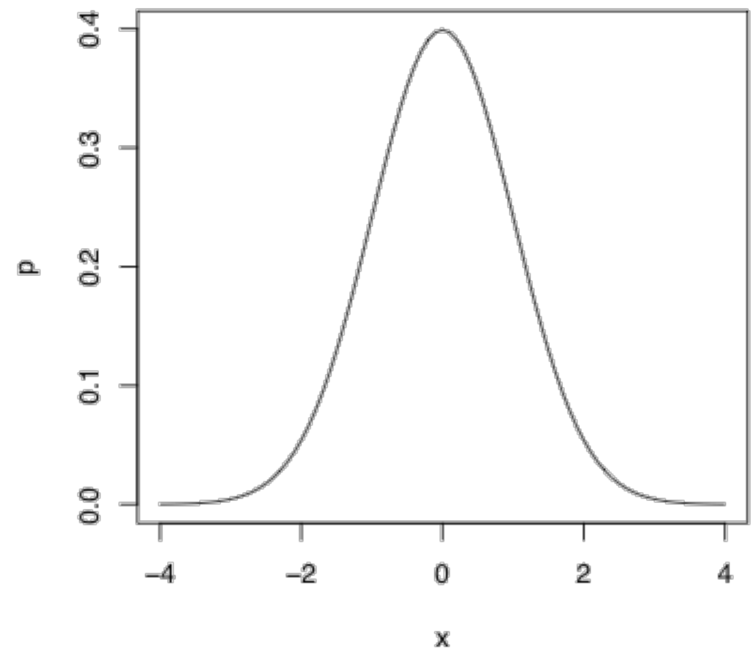
Center

Dispersion

Normal Distribution

Gaussian, bell curve

Mean, Variance



Center & Dispersion (Normal)

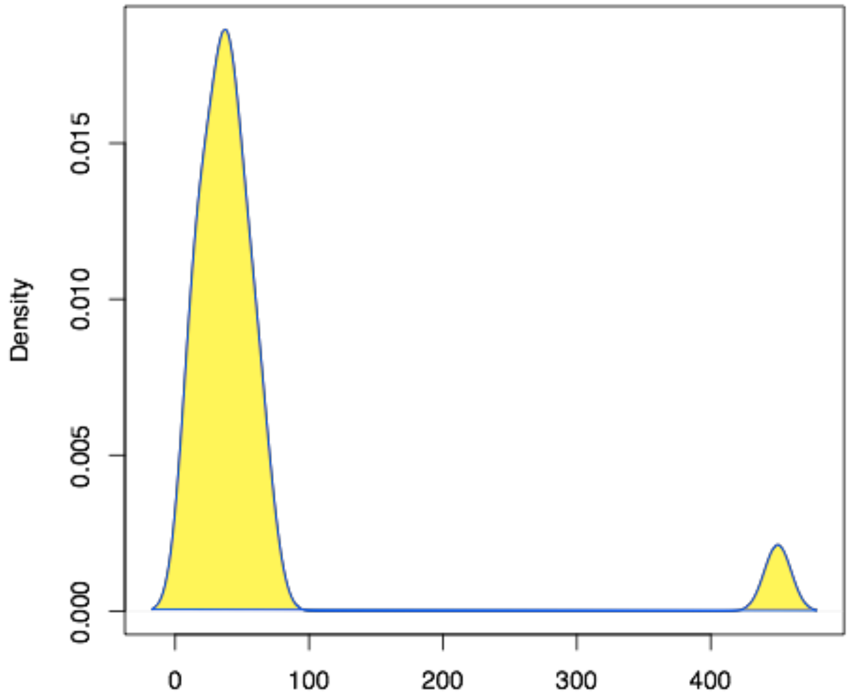
Ages of Employees

12 13 14 21 22 26 33 35 36 37 39 42 45 47 54 57 61 68 450

Center & Dispersion (Normal)

Ages of Employees

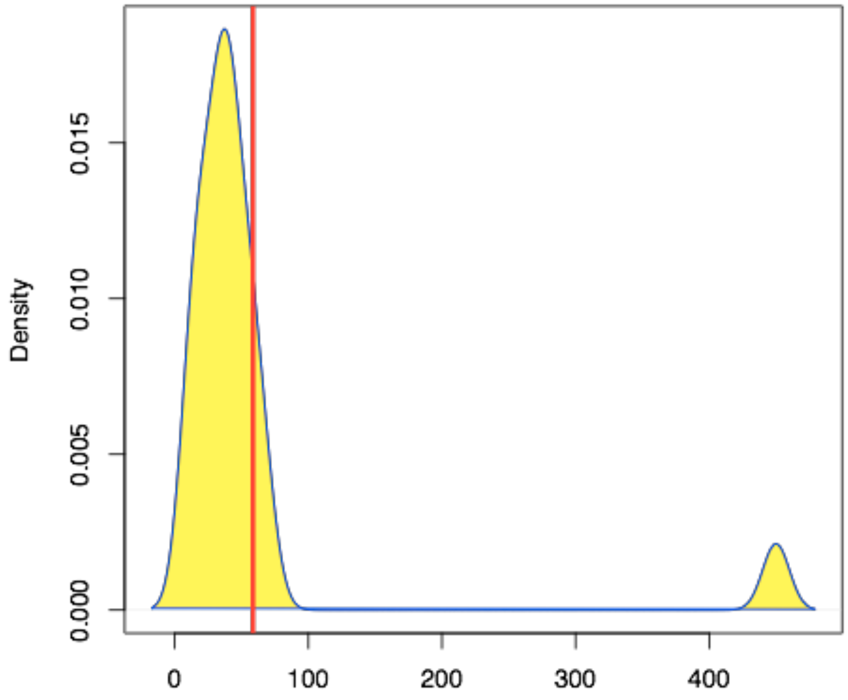
12 13 14 21 22 26 33 35 36 37 39 42 45 47 54 57 61 68 450



Center & Dispersion (Normal)

Ages of Employees

12 13 14 21 22 26 33 35 36 37 39 42 45 47 54 57 61 68 450

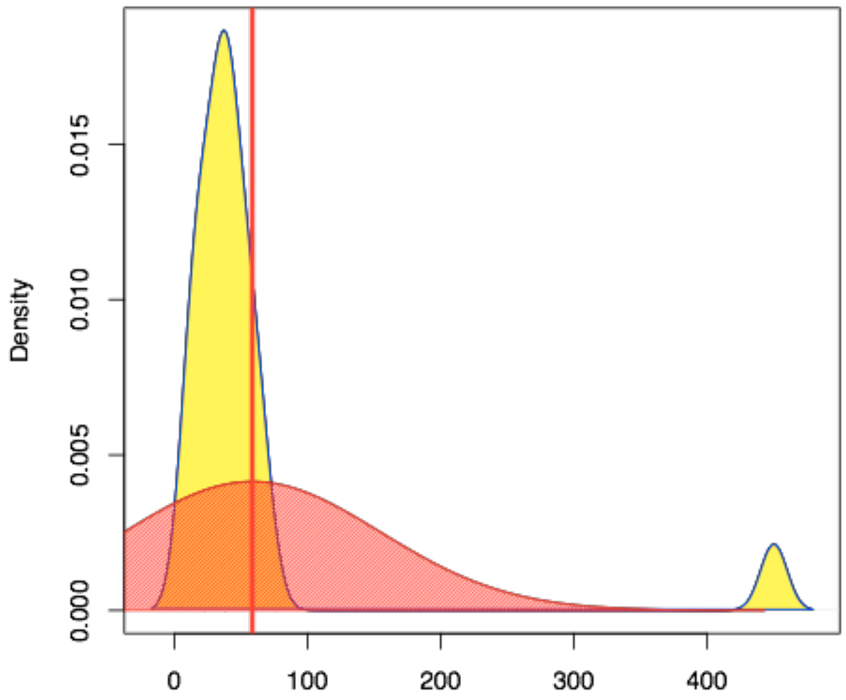


Mean: 58.52632

Center & Dispersion (Normal)

Ages of Employees

12 13 14 21 22 26 33 35 36 37 39 42 45 47 54 57 61 68 450

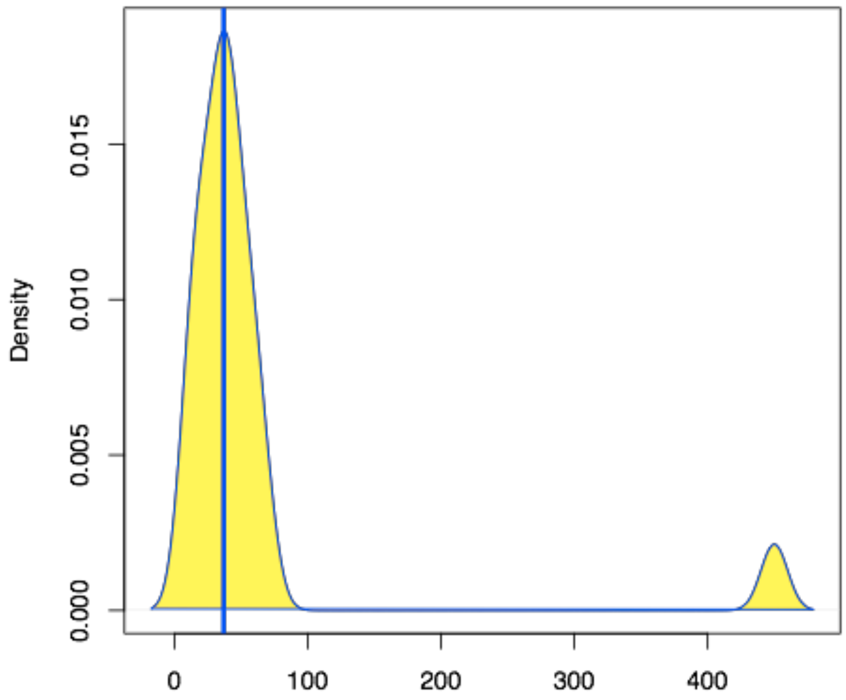


Mean: 58.52632
Variance: 9252.041

Center & Dispersion (Robust)

Ages of Employees

12 13 14 21 22 26 33 35 36 37 39 42 45 47 54 57 61 68 450

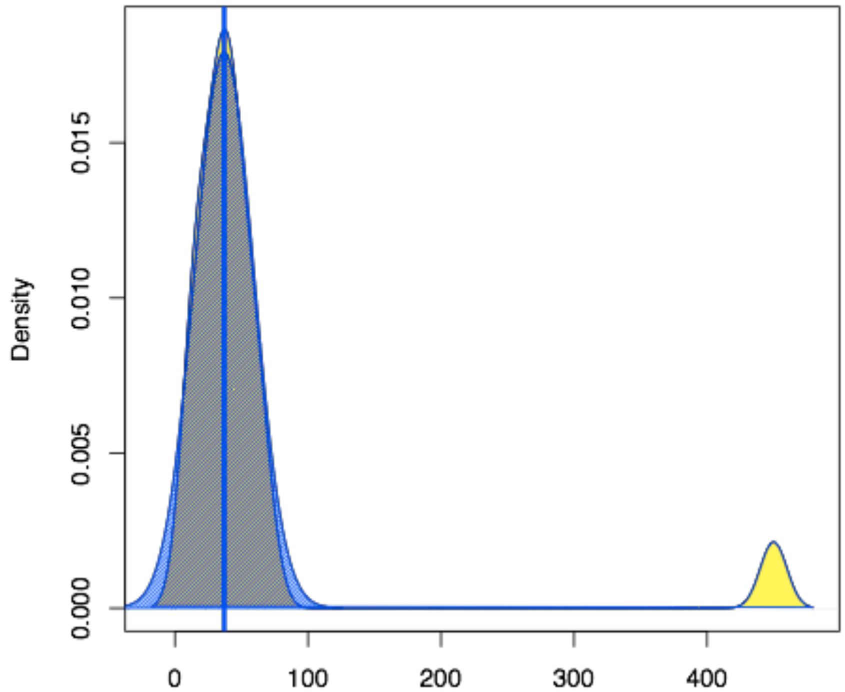


Median: 37

Center & Dispersion (Robust)

Ages of Employees

12 13 14 21 22 26 33 35 36 37 39 42 45 47 54 57 61 68 450



Median: 37
MAD: 22.239
(Median Absolute Deviation)

Subtler Problems

Ages of Employees

12 13 14 21 22 26 33 35 36 37 39 42 45 47 54 57 61 68 450

Subtler Problems

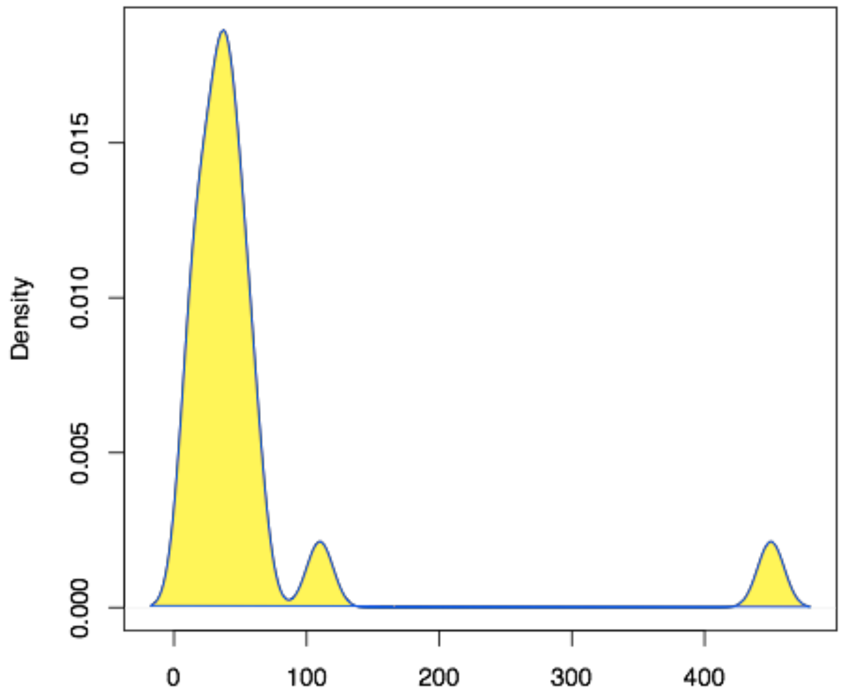
Ages of Employees

12 13 14 21 22 26 33 35 36 37 39 42 45 47 54 57 61 110 450

Subtler Problems

Ages of Employees

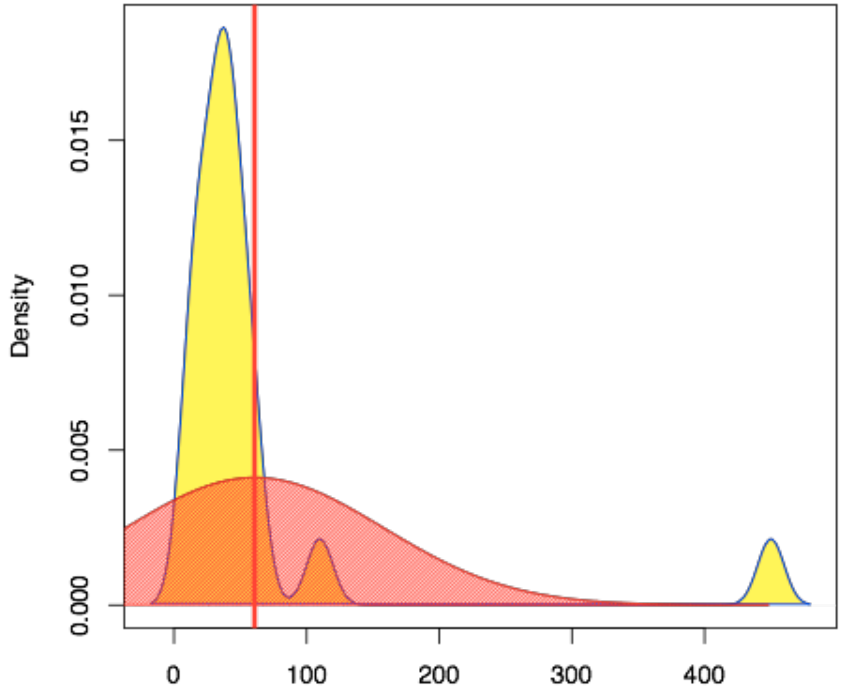
12 13 14 21 22 26 33 35 36 37 39 42 45 47 54 57 61 110 450



Subtler Problems

Ages of Employees

12 13 14 21 22 26 33 35 36 37 39 42 45 47 54 57 61 110 450



Masking

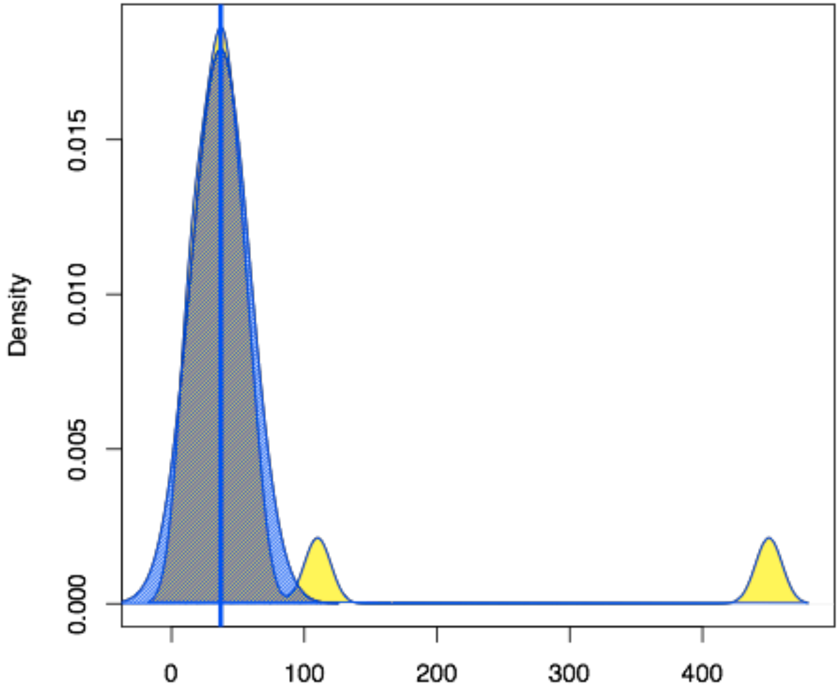
Magnitude of one outlier masks smaller outliers

Makes manual removal of outliers tricky

Subtler Problems

Ages of Employees

12 13 14 21 22 26 33 35 36 37 39 42 45 47 54 57 61 110 450



Robust Statistics

Handle multiple outliers

Robust w.r.t. magnitude of the outliers

Some Robust Centers

Ages of Employees

12 13 14 21 22 26 33 35 36 37 39 42 45 47 54 57 61 110 450

Median

k% Trimmed Mean

k% Winsorized Mean

Some Robust Centers

Ages of Employees

12 13 14 21 22 26 33 35 36 37 39 42 45 47 54 57 61 110 450

Median (37)

Value that evenly splits set into higher & lower halves

k% Trimmed Mean

k% Winsorized Mean

Some Robust Centers

Ages of Employees

12 13 14 21 22 26 33 35 36 37 39 42 45 47 54 57 61 110 450

Median (37)

k% Trimmed Mean (37.933, k=10%)

Remove lowest & highest k% values

Compute mean on remainder

k% Winsorized Mean

Some Robust Centers

Ages of Employees

14 14 14 21 22 26 33 35 36 37 39 42 45 47 54 57 61 *61 61*

Median (37)

k% Trimmed Mean (37.933, k=10%)

k% Winsorized Mean (37.842, k=10%)

Remove lowest & highest k% values

Replace low removed with lowest remaining value

Replace high removed with highest remaining value

Compute mean of resulting set

Model-Driven Validation

A Detective Story

You have accounting records for two firms that are in dispute. One is lying. *How to tell?*

Firm A

283.08	25.23
153.86	385.62
1448.97	12371.32
18595.91	1280.76
21.33	257.64

Amt. Paid: $34823.72

Firm B

283.08	75.23
353.86	185.25
5322.79	9971.42
8795.64	4802.43
61.33	57.64

Amt. Rec'd: $29908.67

A Detective Story

You have accounting records for two firms that are in dispute. One is lying. *How to tell?*

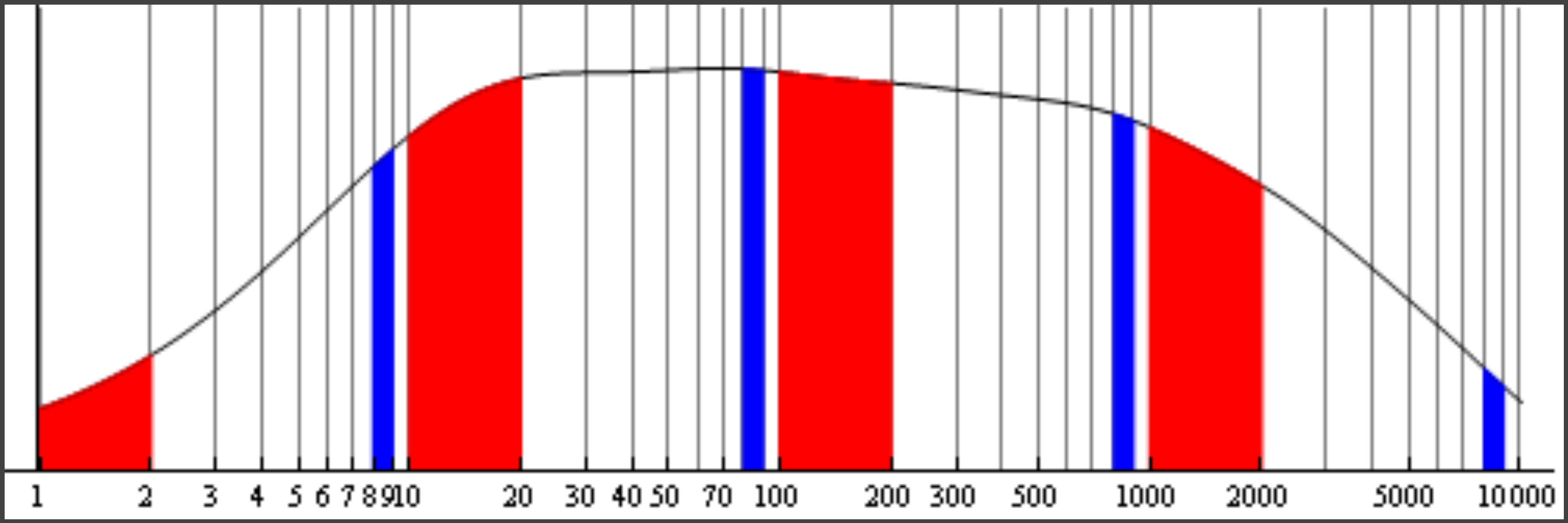
Firm A		*Firm B*	LIARS!
283.08	25.23	283.08	75.23
153.86	385.62	353.86	185.25
1448.97	12371.32	5322.79	9971.42
18595.91	1280.76	8795.64	4802.43
21.33	257.64	61.33	57.64

Amt. Paid: $34823.72

Amt. Rec'd: $29908.67

Benford's Law (Benford 1938, Newcomb 1881)

The *logarithms* of the values (not the values themselves) are uniformly randomly distributed.



Hence the leading digit **1** has a ~30% likelihood. Larger digits are increasingly less likely.

Benford’s Law (Benford 1938, Newcomb 1881)

The *logarithms* of the values (not the values themselves) are uniformly randomly distributed.

Benford's Law (Benford 1938, Newcomb 1881)

The *logarithms* of the values (not the values themselves) are uniformly randomly distributed.

Holds for many (but certainly not all) real-life data sets: Addresses, Bank accounts, Building heights, ...

Benford's Law (Benford 1938, Newcomb 1881)

The *logarithms* of the values (not the values themselves) are uniformly randomly distributed.

Holds for many (but certainly not all) real-life data sets: Addresses, Bank accounts, Building heights, ...

Benford's Law (Benford 1938, Newcomb 1881)

The *logarithms* of the values (not the values themselves) are uniformly randomly distributed.

Holds for many (but certainly not all) real-life data sets: Addresses, Bank accounts, Building heights, ...

Data spanning multiple orders of magnitude.

Benford's Law (Benford 1938, Newcomb 1881)

The *logarithms* of the values (not the values themselves) are uniformly randomly distributed.

Holds for many (but certainly not all) real-life data sets: Addresses, Bank accounts, Building heights, ...

Data spanning multiple orders of magnitude.

Benford's Law (Benford 1938, Newcomb 1881)

The *logarithms* of the values (not the values themselves) are uniformly randomly distributed.

Holds for many (but certainly not all) real-life data sets: Addresses, Bank accounts, Building heights, ...

Data spanning multiple orders of magnitude.

Evidence that records do not follow Benford's Law is admissible in a U.S. court of law.

Secrets of the Agile Data Wrangler

1. Data is never clean
2. Function follows form
3. Expose your data
4. **Statistics & graphics: better together**

Secrets of the Agile Data Wrangler

1. Data is never clean
2. Function follows form
3. Expose your data
4. Statistics & graphics: better together
5. **Wrangling is not distinct from analysis**

Bringing it All Together

TECHNOLOGY TRENDS

Moore's Law provides exponential growth in…

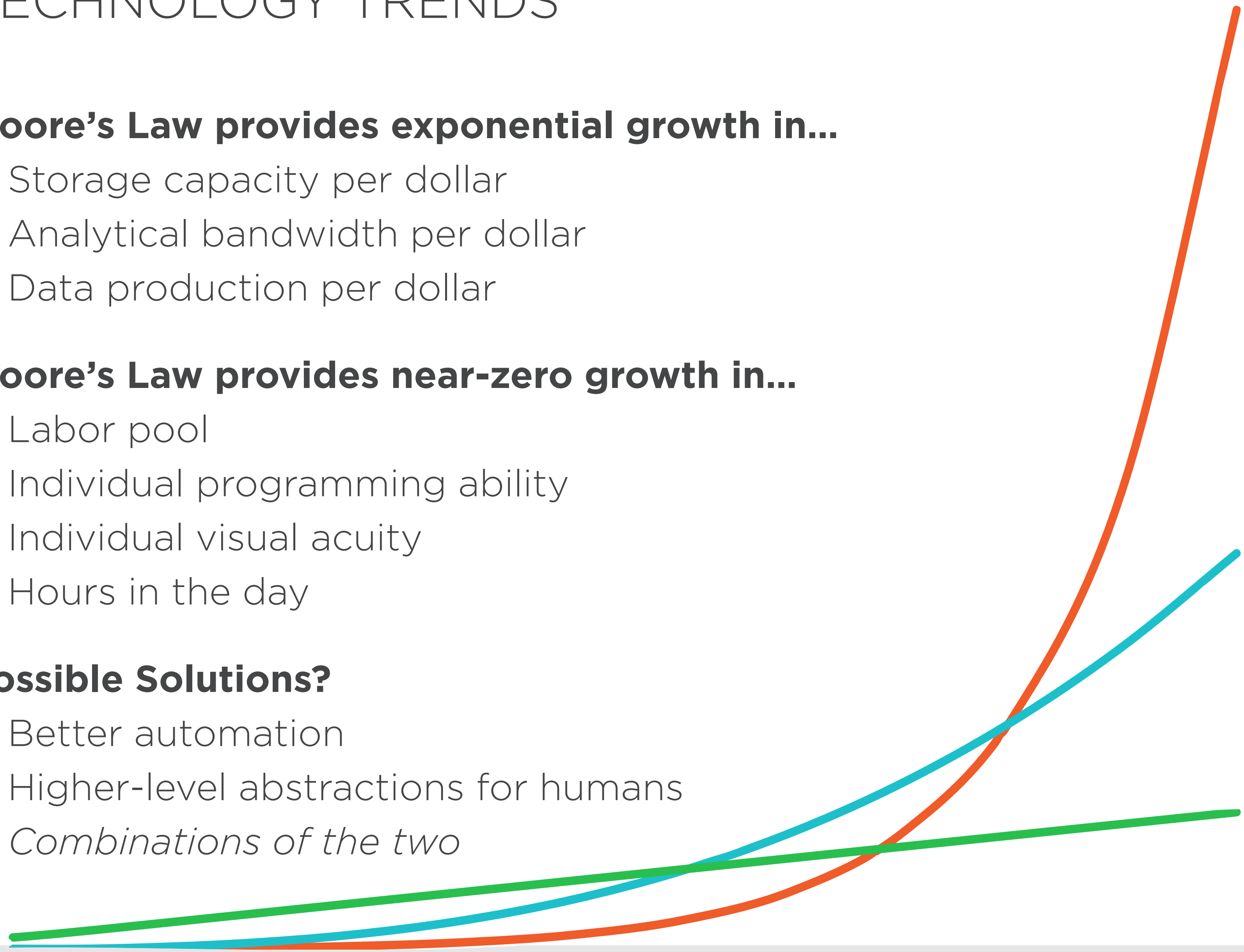
- Storage capacity per dollar
- Analytical bandwidth per dollar
- Data production per dollar

Moore's Law provides near-zero growth in…

- Labor pool
- Individual programming ability
- Individual visual acuity
- Hours in the day

Possible Solutions?

- Better automation
- Higher-level abstractions for humans
- *Combinations of the two*

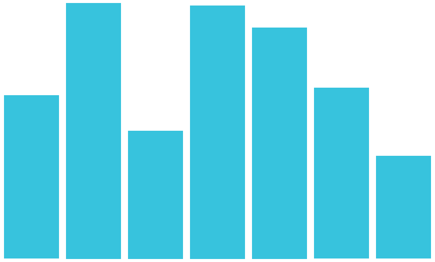


Predictive Interaction

VISUALIZATION

Predictive Interaction

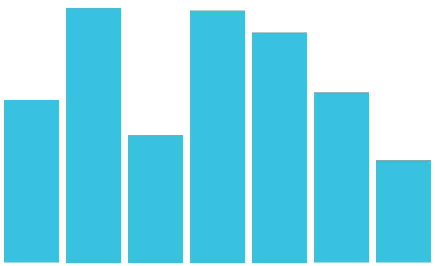
VISUALIZATION



User interacts with
data visualizations

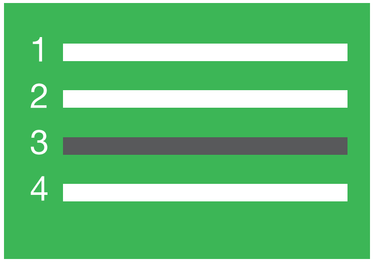
Predictive Interaction

VISUALIZATION



User interacts with
data visualizations

PREDICTION



Algorithms predict desired
action based on data +
context

Predictive Interaction

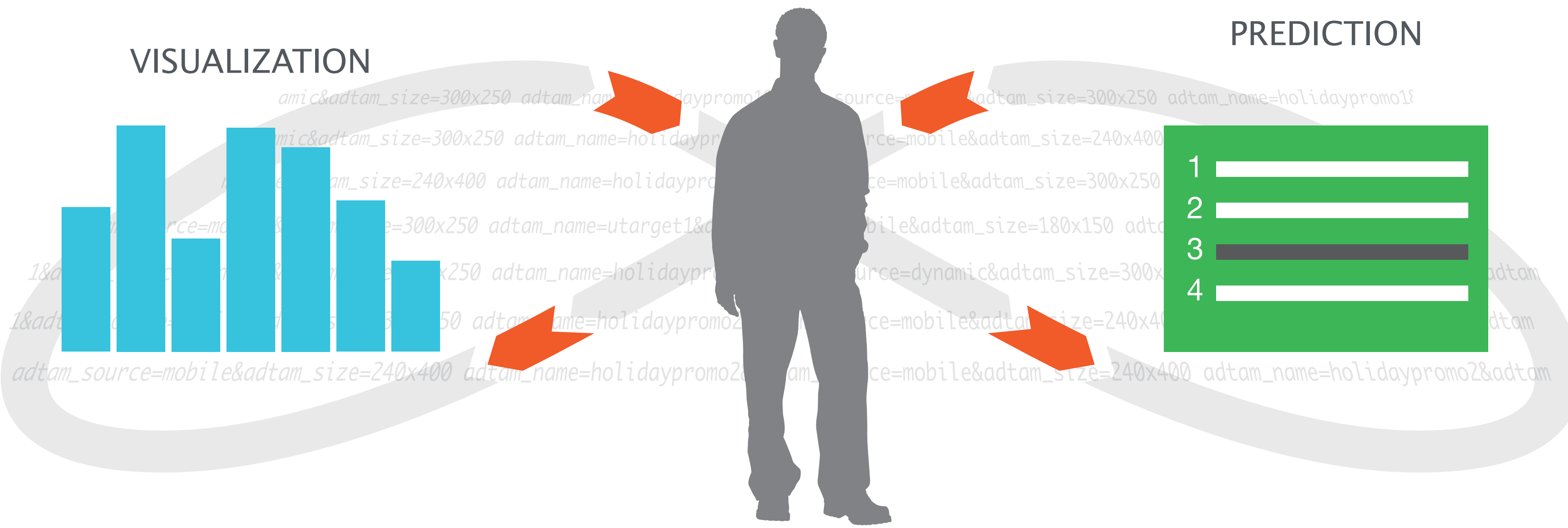
VISUALIZATION

PREDICTION

Data previews allow user to choose, adjust and ratify

Algorithms predict desired action based on data + context

Predictive Interaction



PREDICTIVE INTERACTION ™

Think of transformation happening on two planes:

Data Visualization

Code

Good wranglers "trampoline"

Work at the coding level, periodically validate at the visual level

Predictive interaction flips the paradigm

Work at the high level: visual feature identification

Software predicts the low level: auto-generated code

Choose (low) and preview (high)

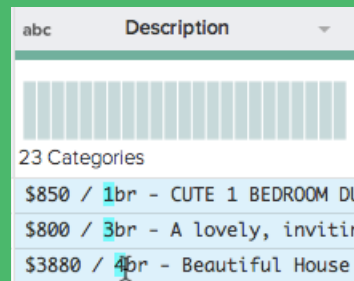
PREDICTIVE INTERACTION ™

Visualization and Interaction

Grounded Syntax

PREDICTIVE INTERACTION ™

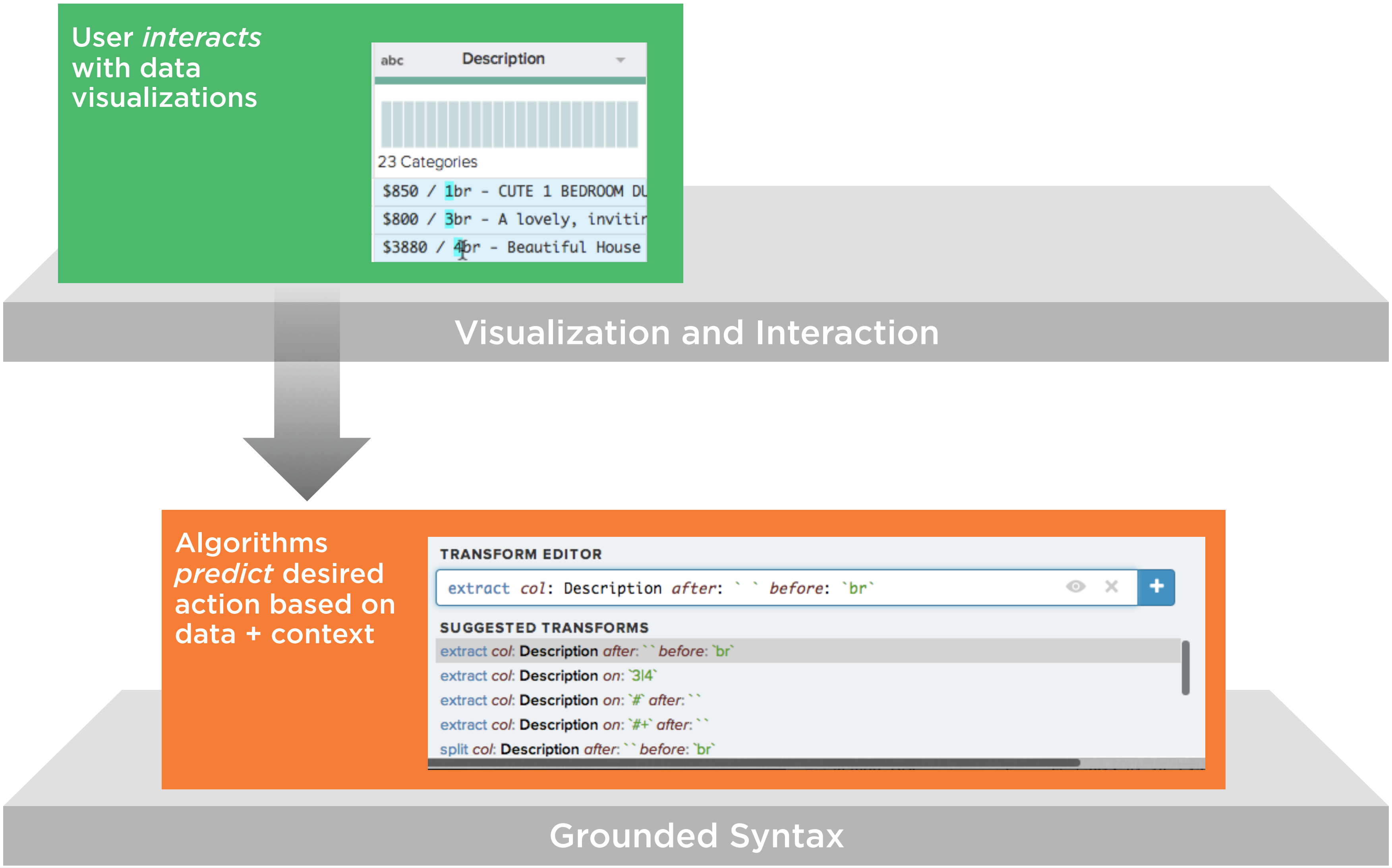
User *interacts* with data visualizations



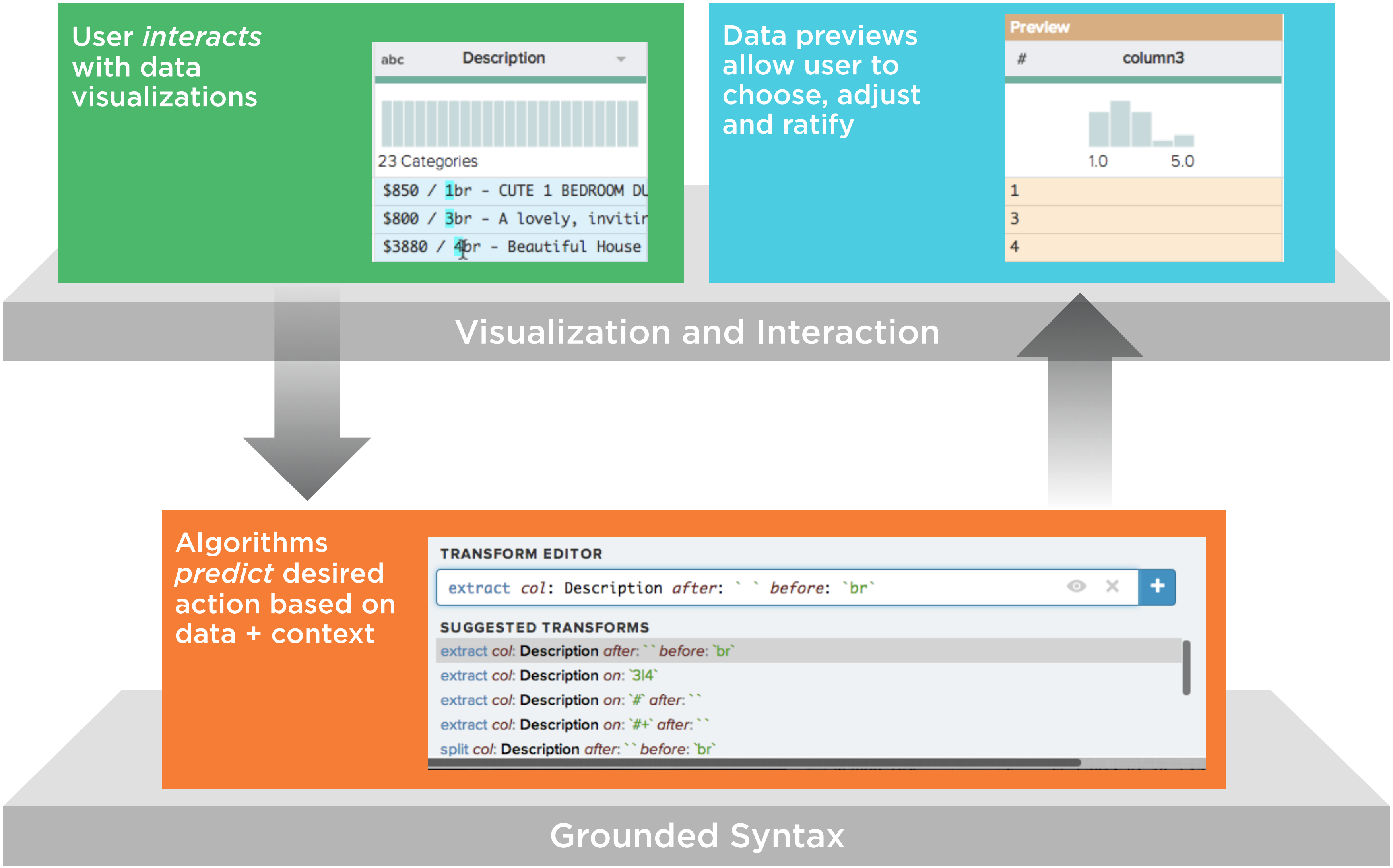
Visualization and Interaction

Grounded Syntax

PREDICTIVE INTERACTION ™



PREDICTIVE INTERACTION ™



IS AUTOMATION THE ANSWER?

Example: SF Restaurant Violation Data.

http://blog.zipfianacademy.com/post/57158627293/how-to-data-science-mapping-sf-restaurant-inspection

https://data.sfgov.org/Public-Health/Restaurant-Scores/stya-26eb

```
"business_id","date","description"
10,"20121114","Unclean or degraded floors walls or ceilings  [ date violation corrected:  ]"
10,"20120403","Unclean or degraded floors walls or ceilings  [ date violation corrected: 9/20/2012 ]"
10,"20110428","Inadequate and inaccessible handwashing facilities  [ date violation corrected: 6/1/2011 ]"
12,"20120420","Food safety certificate or food handler card not available  [ date violation corrected: 11/20/2012 ]"
17,"20120823","Inadequately cleaned or sanitized food contact surfaces  [ date violation corrected: 9/6/2012 ]"
17,"20120823","High risk food holding temperature   [ date violation corrected: 9/6/2012 ]"
17,"20120823","Unclean nonfood contact surfaces  [ date violation corrected: 9/6/2012 ]"
```

```
"business_id","name","address","city","state","postal_code","latitude","longitude","phone_number"
10,"TIRAMISU KITCHEN","033 BELDEN PL","San Francisco","CA","94104","37.791116","-122.403816",""
12,"KIKKA","250 EMBARCADERO  7/F","San Francisco","CA","94105","37.788613","-122.393894",""
17,"GEORGE'S COFFEE SHOP","2200 OAKDALE AVE ","San Francisco","CA","94124","37.741086","-122.401737","+14155531470"
19,"NRGIZE LIFESTYLE CAFE","1200 VAN NESS AVE, 3RD FLOOR","San Francisco","CA","94109","37.786848","-122.421547",""
24,"OMNI S.F. HOTEL - 2ND FLOOR PANTRY","500 CALIFORNIA ST, 2ND  FLOOR","San Francisco","CA","94104","37.792888","-122.403135",""
29,"CHICO'S PIZZA","131 06TH ST ","San Francisco","CA","94103","37.774722","-122.406761","+14155251111"
31,"NORMAN'S ICE CREAM AND FREEZES","2801 LEAVENWORTH ST ","San Francisco","CA","94133","37.807155","-122.419004",
```

```
"business_id","Score","date","type"
10,"98","20121114","routine"
10,"98","20120403","routine"
10,"100","20110928","routine"
10,"96","20110428","routine"
10,"100","20101210","routine"
12,"100","20121120","routine"
```

Resources

W. Cleveland, The Elements of Graphing Data

http://www.amazon.com/dp/0963488414

W. Cleveland, Visualizing Data.

http://www.amazon.com/dp/0963488406

Kandel, et al. Enterprise Data Analysis and Visualization: An Interview Study.

http://vis.stanford.edu/files/2012-EnterpriseAnalysisInterviews-VAST.pdf

Dasu & Johnson, Exploratory Data Mining and Data Cleaning

http://www.amazon.com/Exploratory-Data-Mining-Cleaning/dp/0471268518

Hellerstein, Quantitative Data Cleaning for Large Databases

http://db.cs.berkeley.edu/jmh/papers/cleaning-unece.pdf

Getoor & Machanavajjhala: Entity Resolution: Theory, Practice & Open Challenges http://www.cs.umd.edu/~getoor/Tutorials/ER_VLDB2012.pdf

TRIFACTA