DATA CLEANING & DATA MANIPULATION

PETRA ISENBERG

VISUAL ANALYTICS

With slides by Wesley Willett

WHAT IS "DIRTY DATA"?

BEFORE WE CAN TALK ABOUT CLEANING, WE NEED TO KNOW ABOUT TYPES OF ERROR AND WHERE THEY COME FROM

SOURCES OF ERROR

DATA ENTRY ERRORS

MEASUREMENT ERRORS

DISTILLATION ERRORS

DATA INTEGRATION ERRORS

[HELLERSTEIN 2008]

DATA ENTRY ERROR

LOTS OF DATA IS ENTERED BY HAND

TYPOGRAPHIC ERRORS

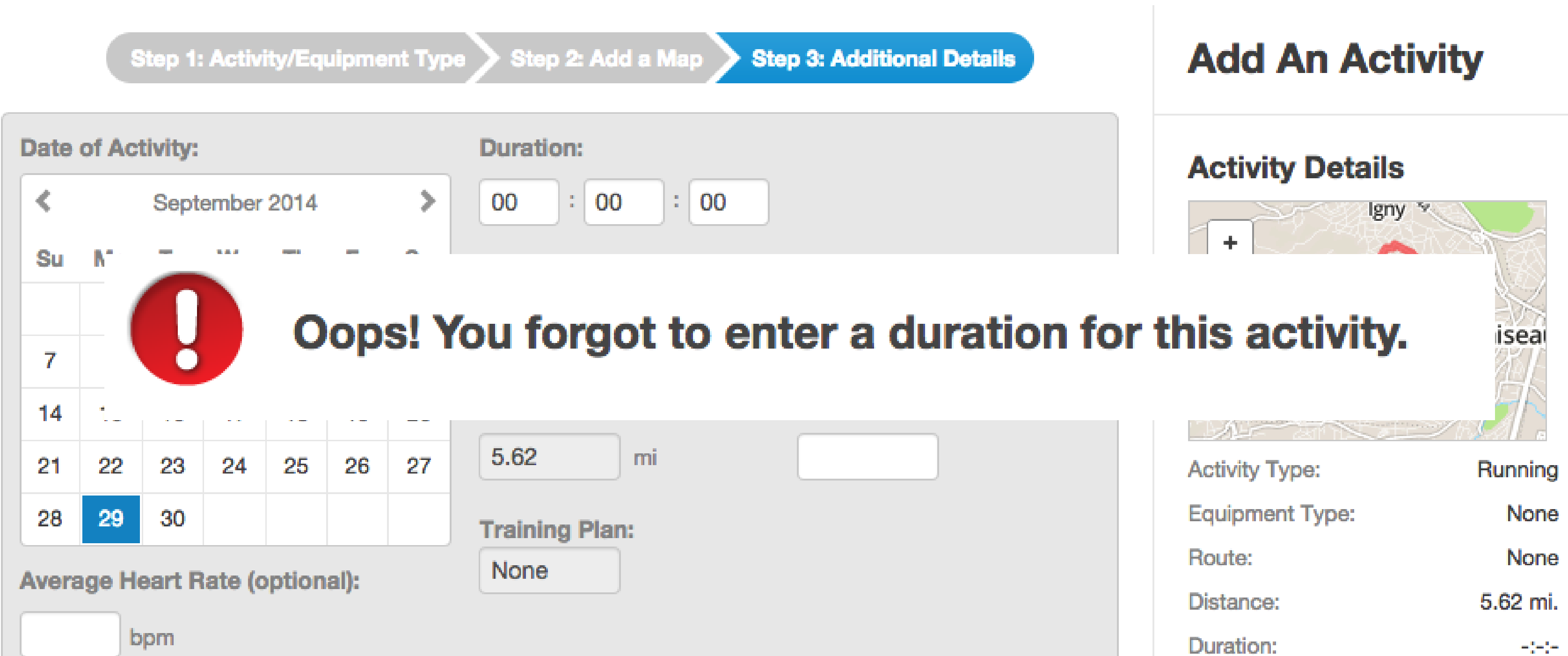
MISUNDERSTANDING DATA OR CONVENTIONS

“SPURIOUS INTEGRITY”

“SPURIOUS INTEGRITY”

ENTERING BAD DATA IN RESPONSE TO (OFTEN WELL-INTENTIONED) INTERFACE CONSTRAINTS

“SPURIOUS INTEGRITY”



MEASUREMENT ERRORS

- SENSOR ISSUES
- MALFUNCTIONS
- PLACEMENT
- INTERFERENCE
- MISCALIBRATION

DISTILLATION ERRORS

SOME DATA MAY BE LOST OR COMPRESSED BEFORE IT ENTERS THE DATABASE

0.345413 → 0.35

National Price Index → NPI

1985, $2, Apples
1985, $2, Oranges → 1985, $2, "Apples,Oranges,Cucumbers"
1985, $2, Cucumbers

DATA INTEGRATION ERRORS

DATA OFTEN COMES FROM MULTIPLE SOURCES

SCHEMAS CHANGE OVER TIME

DATA IS OFTEN COERCED FROM
ONE TYPE TO ANOTHER

CAN LEAD TO DATA LOSS,
DUPLICATION, AND OTHER

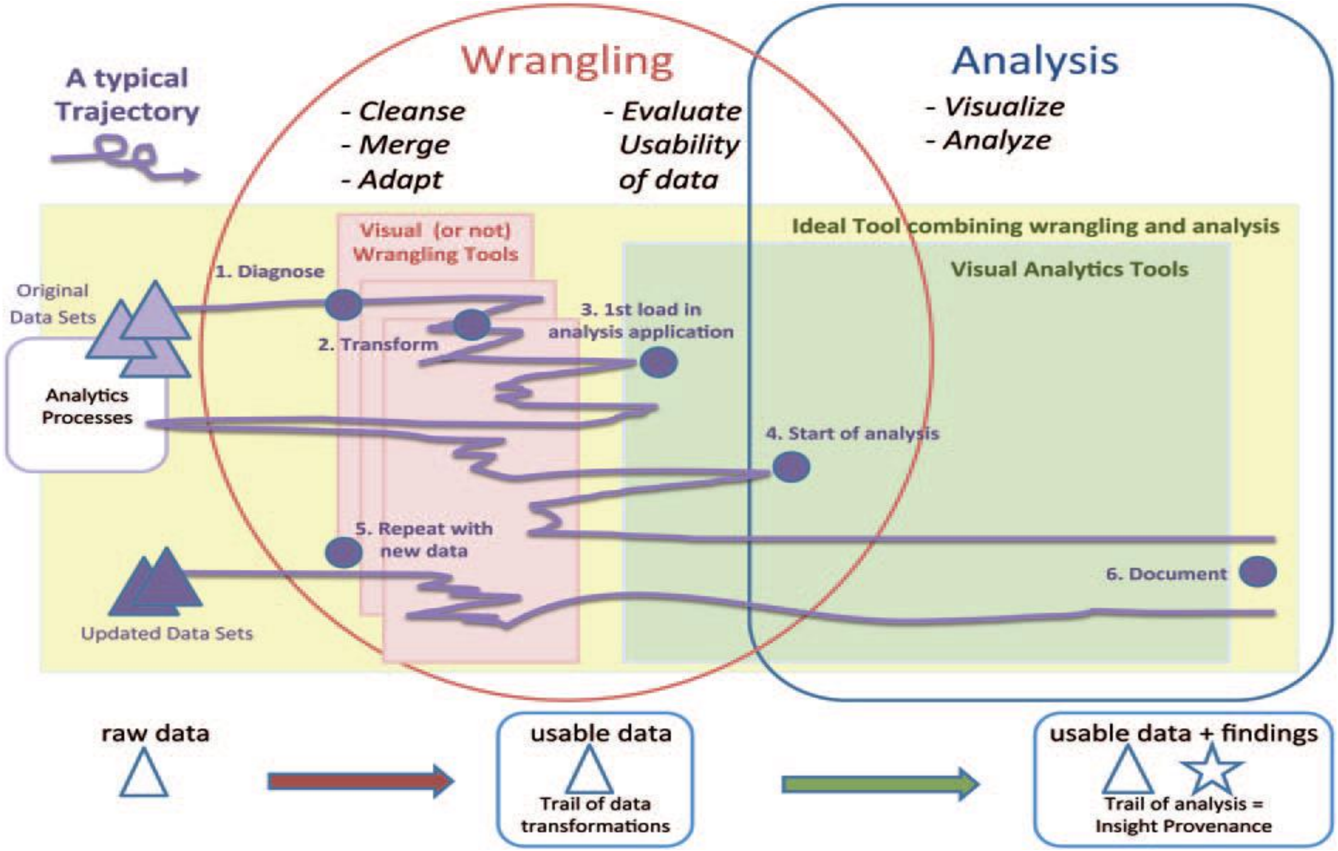
WHY IS THIS IMPORTANT?

<u>MOST</u> OF THE TIME IN THE DATA ANALYSIS PROCESS IS ACTUALLY SPENT HERE!

“I spend more than half 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.”

[Kandel 2012]

ANALYSIS TRAJECTORIES



KANDEL ET AL. 2011

SOME DATA QUALITY ISSUES

MISSING DATA	MISSED MEASUREMENTS, REDACTED ITEMS, INCOMPLETE FORMS, ETC.
ERRONEOUS VALUES	MISSPELLINGS, OUTLIERS, “SPURIOUS INTEGRITY”, ETC.
ENTITY RESOLUTION	DIFFERENT VALUES, ABBREVS., 2+ ENTRIES FOR THE SAME THING?
TYPE CONVERSION	E.G., ZIP CODE OR PLACE NAME TO LAT-LON
DATA INTEGRATION	MISMATCHES AND INCONSISTENCIES WHEN COMBINING DATA

SOME APPROACHES FOR IMPROVING DATA QUALITY

TOOLS FOR MANIPULATING AND CLEANING DATA

SOME APPROACHES FOR IMPROVING DATA QUALITY

TOOLS FOR MANIPULATING AND CLEANING DATA

PREVENTING ERROR

CATCHING DIRTY DATA AT THE SOURCE

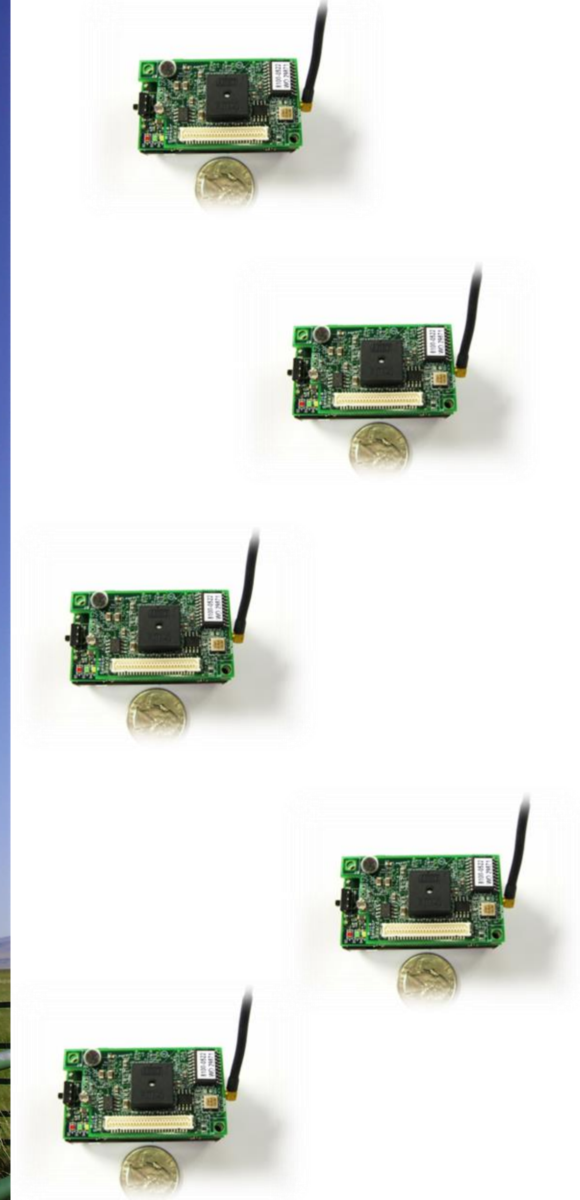
MINIMIZING SENSOR ERROR

CALIBRATE AND VERIFY SENSORS

CHECK SENSORS BEFORE DEPLOYMENT (AND PERIODICALLY REVALIDATE THEM)

USE REDUNDANT SENSORS

CHECK DATA AGAINST HISTORICAL LOGS OR COMPUTED MODELS



TRADE-OFFS BETWEEN **(RE)CALIBRATION** AND **REDUNDANCY**

REDUCING ERROR DURING DATA ENTRY

DOUBLE DATA ENTRY

PERFORM ALL DATA ENTRY TWICE

(IDEALLY BY SEPARATE PEOPLE)

IDENTIFY MISMATCHES AND DISCARD OR REPAIR

(VIA VOTING OR RE-ENTRY)

INTEGRITY CONSTRAINTS

This field is required.

TEMPERATURE xx ºC

INTEGRITY CONSTRAINTS

Temperatures must be between -50ºC and 50ºC.

TEMPERATURE -60 ºC

INTEGRITY CONSTRAINTS

TEMPERATURE ____ ºC

INTEGRITY CONSTRAINTS DO NOT PREVENT BAD DATA

ENFORCING CONSTRAINTS LEADS TO FRUSTRATION

FRICTION AND PREDICTION

USE DATA QUALITY MEASURES TO **PREDICT** HOW LIKELY A VALUE IS TO BE CORRECT.

ADJUST THE INTERFACE TO **ADD FRICTION** WHEN ENTERING UNLIKELY RESPONSES.

[HELLERSTEIN 2008]

FRICTION AND PREDICTION

PRINCIPLE 1

DATA QUALITY SHOULD BE CONTROLLED VIA **FEEDBACK**, NOT **ENFORCEMENT**.

PRINCIPLE 2

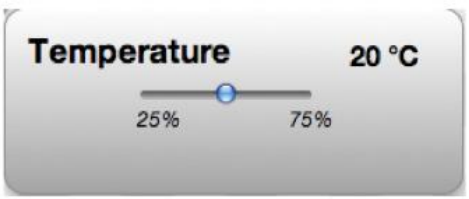
FRICTION MERITS **EXPLANATION**.

PRINCIPLE 3

ANNOTATION SHOULD BE EASIER THAN OMISSION OR SUBVERSION.

[HELLERSTEIN 2008]

FRICTION AND PREDICTION



[HELLERSTEIN 2008]

FRICTION AND PREDICTION

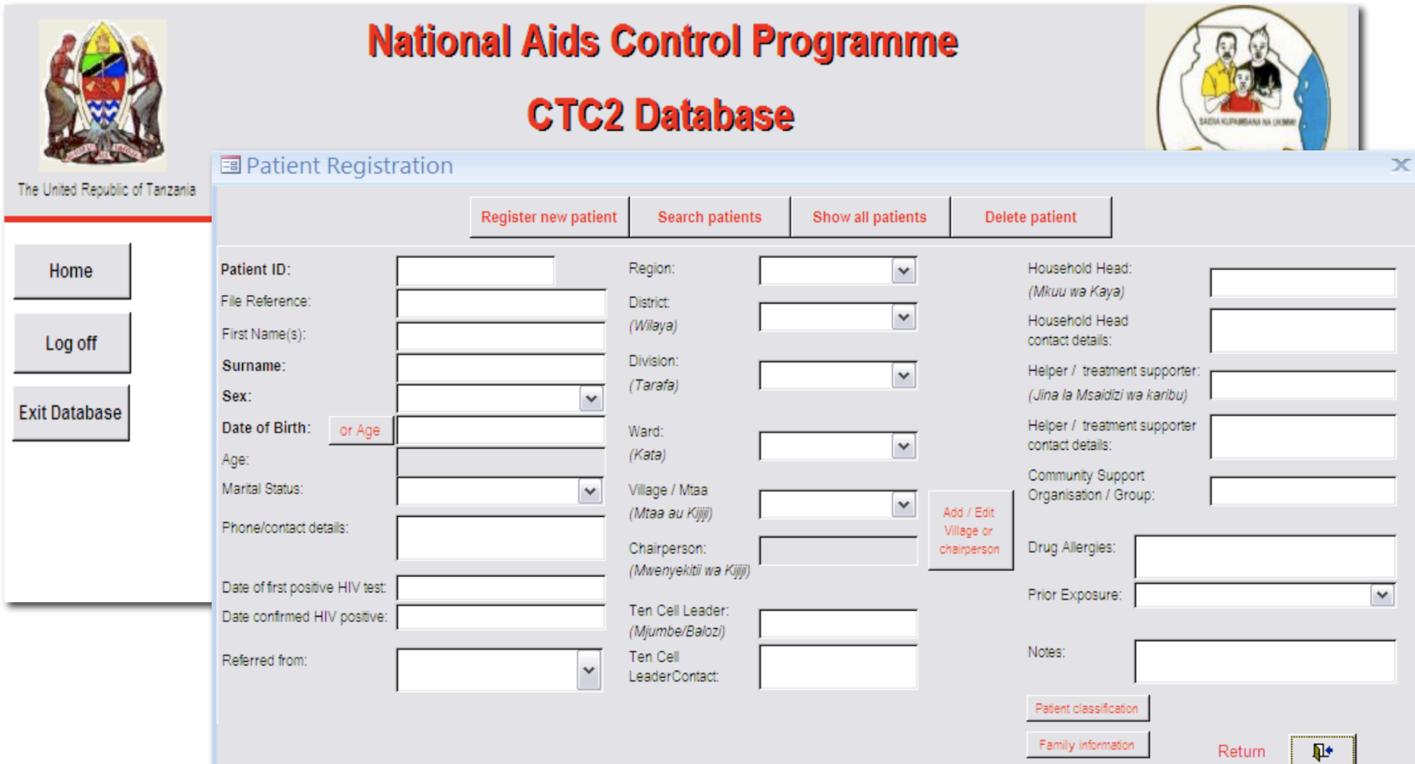
This value seems low.
Are you sure?

TEMPERATURE -60 ºC

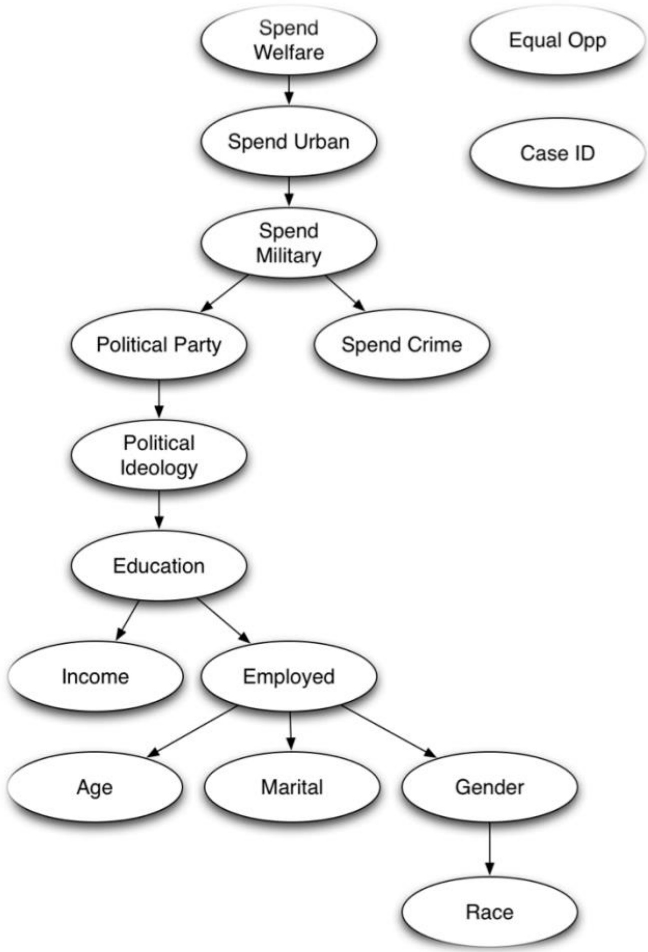
Sensor disabled.

USHER

[Chen et al. 2010]



MS Access data entry forms for Tanzanian HIV/AIDS monitoring



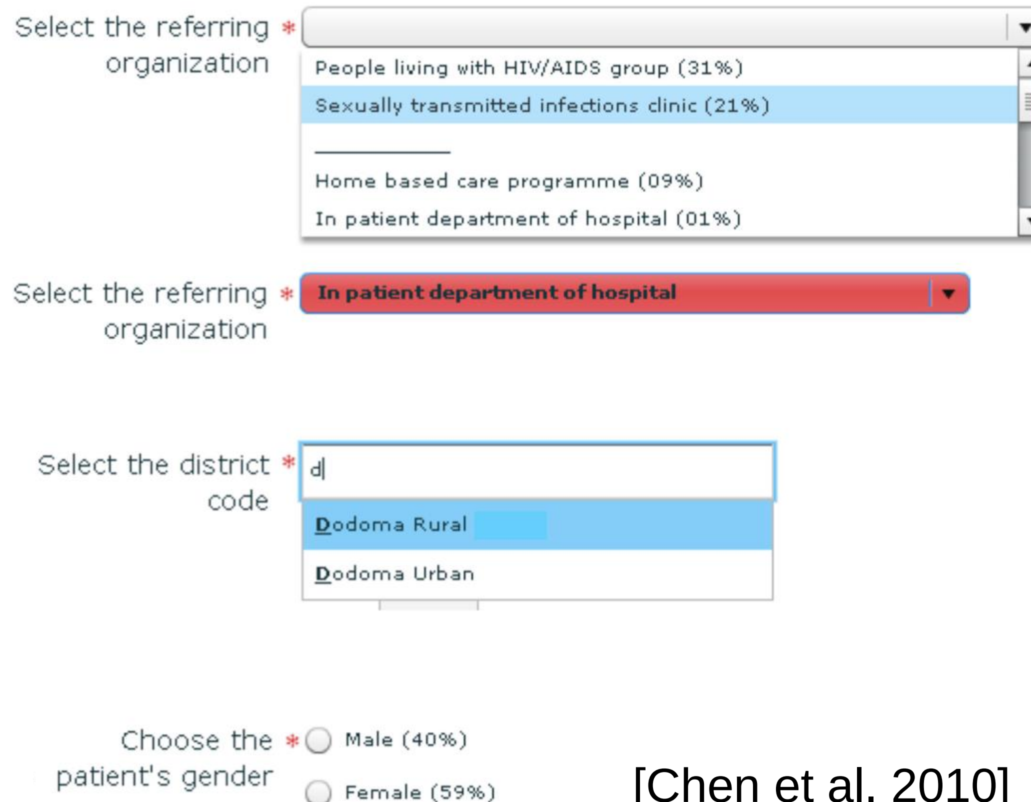
BUILD A MODEL to predict dependencies and relationships between questions.

[Chen et al. 2010]

DYNAMIC ORDERING

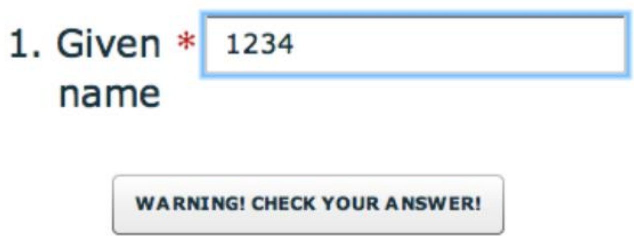
ALWAYS ASK THE MOST APPROPRIATE NEXT QUESTION

SUGGEST THE MOST LIKELY ANSWERS



[Chen et al. 2010]

SMART RE-ASKING AND SUGGESTIONS



FRICTION

AUTOMATING ~~CONSTRAINTS~~

[Chen et al. 2010]

DETECTING ERRORS

LOOK FOR OUTLIERS / ANOMALIES

EXAMINE DATA TYPES

SCHEMA CHECKING

VALIDATE WITH OTHER DATA

OTHER HEURISTICS

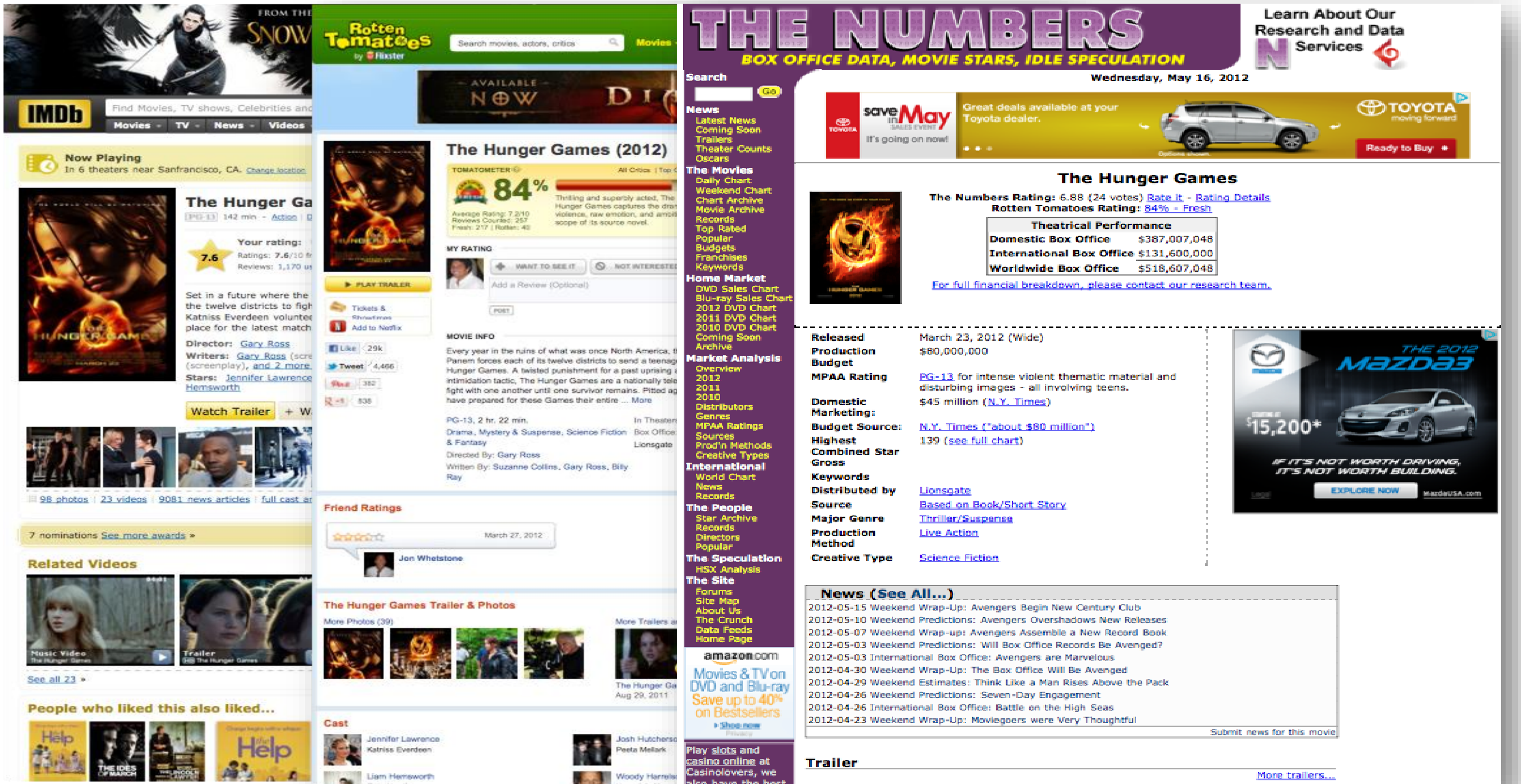
HISTORICALLY – MORE FOCUS ON AUTOMATED APPROACHES

“PROFILING” DATA

<u>UNDERSTANDING</u> WHAT ASSUMPTIONS YOU CAN MAKE ABOUT DATA

<u>INTERACTIVELY</u> IDENTIFYING DATA QUALITY ISSUES

AN EXAMPLE



IMDb
Now Playing
In 6 theaters near Sanfrancisco, CA.
The Hunger Ga
Your rating:
7.6
Director: Gary Ross
Writers: Gary Ross (screenplay), and 2 more
Stars: Jennifer Lawrence
Hemsworth
Watch Trailer
98 photos | 23 videos | 9081 news articles | full cast
7 nominations See more awards »
Related Videos
See all 23 »
People who liked this also liked...
Rotten Tomatoes
The Hunger Games (2012)
TOMATOMETER
84%
MY RATING
PLAY TRAILER
MOVIE INFO
Every year in the ruins of what was once North America, the Panem forces each of its twelve districts to send a teenag
Friend Ratings
March 27, 2012
Jon Whetstone
The Hunger Games Trailer & Photos
Cast
Jennifer Lawrence
Katniss Everdeen
Josh Hutcherson
Peeta Mellark
Liam Hemsworth
Woody Harrelson
THE NUMBERS
BOX OFFICE DATA, MOVIE STARS, IDLE SPECULATION
Learn About Our Research and Data Services
Wednesday, May 16, 2012
Search
News
Latest News
Coming Soon
Trailers
Theater Counts
Oscars
The Movies
Daily Chart
Weekend Chart
Chart Archive
Movie Archive
Records
Top Rated
Popular
Budgets
Franchises
Keywords
Home Market
DVD Sales Chart
Blu-ray Sales Chart
2012 DVD Chart
2011 DVD Chart
2010 DVD Chart
Coming Soon
Archive
Market Analysis
Overview
2012
2011
2010
Distributors
Genres
MPAA Ratings
Sources
Prod'n Methods
Creative Types
International
World Chart
News
Records
The People
Star Archive
Records
Directors
Popular
The Speculation
HSX Analysis
The Site
Forums
Site Map
About Us
The Crunch
Data Feeds
Home Page
The Hunger Games
The Numbers Rating: 6.88 (24 votes) Rate it - Rating Details
Rotten Tomatoes Rating: 84% - Fresh
Theatrical Performance
Domestic Box Office $387,007,048
International Box Office $131,600,000
Worldwide Box Office $518,607,048
For full financial breakdown, please contact our research team.
Released March 23, 2012 (Wide)
Production Budget $80,000,000
MPAA Rating PG-13 for intense violent thematic material and disturbing images - all involving teens.
Domestic Marketing: $45 million (N.Y. Times)
Budget Source: N.Y. Times ("about $80 million")
Highest Combined Star Gross 139 (see full chart)
Keywords
Distributed by Lionsgate
Source Based on Book/Short Story
Major Genre Thriller/Suspense
Production Method Live Action
Creative Type Science Fiction
News (See All...)
2012-05-15 Weekend Wrap-Up: Avengers Begin New Century Club
2012-05-10 Weekend Predictions: Avengers Overshadows New Releases
2012-05-07 Weekend Wrap-up: Avengers Assemble a New Record Book
2012-05-03 Weekend Predictions: Will Box Office Records Be Avenged?
2012-05-03 International Box Office: Avengers are Marvelous
2012-04-30 Weekend Wrap-Up: The Box Office Will Be Avenged
2012-04-29 Weekend Estimates: Think Like a Man Rises Above the Pack
2012-04-26 Weekend Predictions: Seven-Day Engagement
2012-04-26 International Box Office: Battle on the High Seas
2012-04-23 Weekend Wrap-Up: Moviegoers were Very Thoughtful
Submit news for this movie
Trailer
More trailers...

Title	Release Date	MPAA Rating	Distributor	Rotten Tomatoes Rating	IMDB Rating
The Land Girls	Jun 12, 1998	R	Gramercy		6.1
First Love, Last Rites	Aug 7, 1998	R	Strand		6.9
I Married a Strange Person	Aug 28, 1998		Lionsgate		6.8
Slam	Oct 9, 1998	R	Trimark	62	3.4
Mississippi Mermaid	Jan 15, 1999		MGM		
Following	Apr 4, 1999	R	Zeitgeist		7.7
Foolish	Apr 9, 1999	R	Artisan		3.8
Pirates	Jul 1, 1986	R		25	5.8
Duel in the Sun	Dec 31, 2046			86	7
Tom Jones	Oct 7, 1963			81	7
Oliver!	Dec 11, 1968		Sony Pictures	84	7.5
To Kill A Mockingbird	Dec 25, 1962		Universal	97	8.4
Tora, Tora, Tora	Sep 23, 1970				
Hollywood Shuffle	Mar 1, 1987			87	6.8
Over the Hill to the Poorhouse	Sep 17, 2020				
Wilson	Aug 1, 2044				7
Darling Lili	Jan 1, 1970				6.1
The Ten Commandments	Oct 5, 1956			90	2.5
12 Angry Men	Apr 13, 1957		United Artists		8.9
Twelve Monkeys	Dec 27, 1995	R	Universal		8.1
1776	Nov 9, 1972	PG	Sony/ Columbia	57	7

Title	Release Date	MPAA Rating	Distributor	Rotten Tomatoes Rating	IMDB Rating
The Land Girls	Jun 12, 1998	R	Gramercy		6.1
First Love, Last Rites	Aug 7, 1998	R	Strand		6.9
I Married a Strange Person	Aug 28, 1998		Lionsgate		6.8
Slam	Oct 9, 1998	R	Trimark	62	3.4
Mississippi Mermaid	Jan 15, 1999		MGM		
Following	Apr 4, 1999	R	Zeitgeist		7.7
Foolish	Apr 9, 1999	R	Artisan		3.8
Pirates	Jul 1, 1986	R		25	5.8
Duel in the Sun	Dec 31, 2046			86	7
Tom Jones	Oct 7, 1963			81	7
Oliver!	Dec 11, 1968		Sony Pictures	84	7.5
To Kill A Mockingbird	Dec 25, 1962		Universal	97	8.4
Tora, Tora, Tora	Sep 23, 1970				
Hollywood Shuffle	Mar 1, 1987			87	6.8
Over the Hill to the Poorhouse	Sep 17, 2020				
Wilson	Aug 1, 2044				7
Darling Lili	Jan 1, 1970				6.1
The Ten Commandments	Oct 5, 1956			90	2.5
12 Angry Men	Apr 13, 1957		United Artists		8.9
Twelve Monkeys	Dec 27, 1995	R	Universal		8.1
1776	Nov 9, 1972	PG	Sony/ Columbia	57	7

Arnolds Park	Oct 19, 2007	PG-13	The Movie Partners
Sweet Sweetback's Baad Asssss Song	Jan 1, 1971		
And Then Came Love	Jun 1, 2007	Not Rated	Fox Meadow
Around the World in 80 Days	Oct 17, 1956	PG	United Artists
Barbarella	Oct 10, 1968		Paramount Pictures
Barry Lyndon	1975		Warner Bros.
Barbarians, The	March, 1987		
Babe	Aug 4, 1995	G	Universal
Boynton Beach Club	Mar 24, 2006	R	Wingate Distribution
Baby's Day Out	Jul 1, 1994	PG	20th Century

Bad Boys	Apr 7, 1995	6.6	53929
Body Double	Oct 26, 1984	6.4	9738
The Beast from 20,000 Fathoms	Jun 13, 1953		
Beastmaster 2: Through the Portal of Time	Aug 30, 1991	3.3	1327
The Beastmaster	Aug 20, 1982	5.7	5734
Ben-Hur	Dec 30, 2025	8.2	58510
Ben-Hur	Nov 18, 1959	8.2	58510
Benji	Nov 15, 1974	5.8	1801
Before Sunrise	Jan 27, 1995	8	39705

SOME DATA QUALITY ISSUES

MISSING DATA	MISSED MEASUREMENTS, REDACTED ITEMS, INCOMPLETE FORMS, ETC.
ERRONEOUS VALUES	MISSPELLINGS, OUTLIERS, "SPURIOUS INTEGRITY", ETC.
ENTITY RESOLUTION	DIFFERENT VALUES, ABBREVS., 2+ ENTRIES FOR THE SAME THING?
TYPE CONVERSION	E.G., ZIP CODE OR PLACE NAME TO LAT-LON
DATA INTEGRATION	MISMATCHES AND INCONSISTENCIES WHEN COMBINING DATA

DETECTION METHODS

+ CAN IDENTIFY POTENTIAL ANOMALIES

- HARD TO KNOW IF THEY'RE REALLY ANOMALOUS OR HOW TO CORRECT THEM

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

MISSING AND IMPOSSIBLE VALUES

1. LOOK AT EMPTY/MISSING VALUES
2. LOOK AT IMPOSSIBLE VALUES

   Gender = 3

   Heart Rate = 0

   Unlikely Dates (e.g. "01/01/0001")

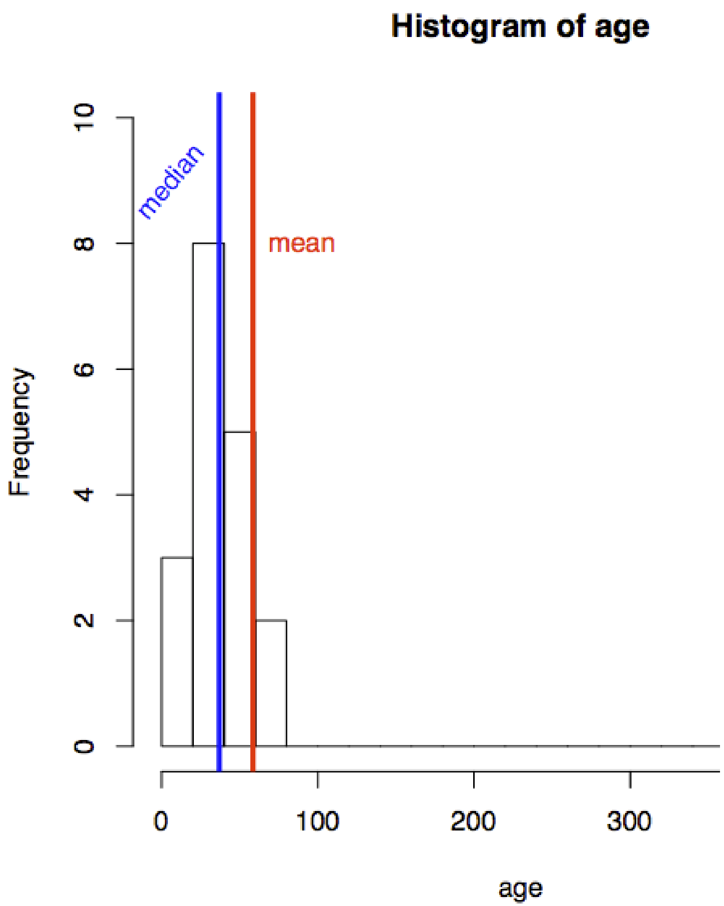
JUST SORTING THE DATA CAN HELP HIGHLIGHT ISSUES LIKE THESE

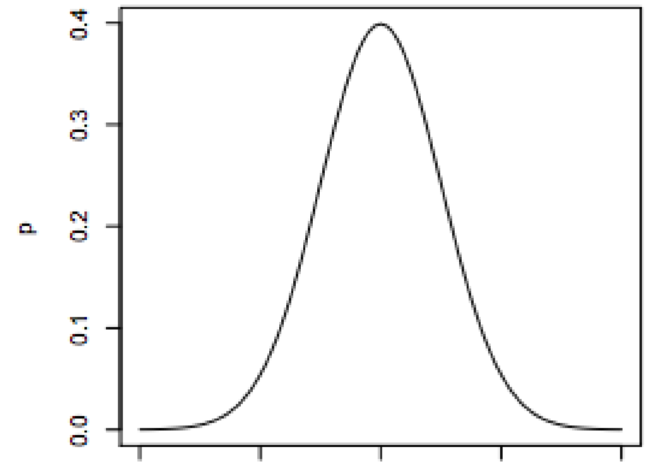
OUTLIER DETECTION

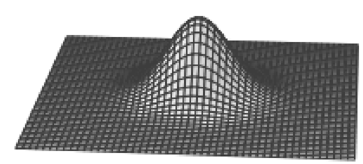
1. EXAMINE DISTRIBUTIONS
2. MODEL DATA AND LOOK FOR RESIDUALS
3. PARTITION DATA

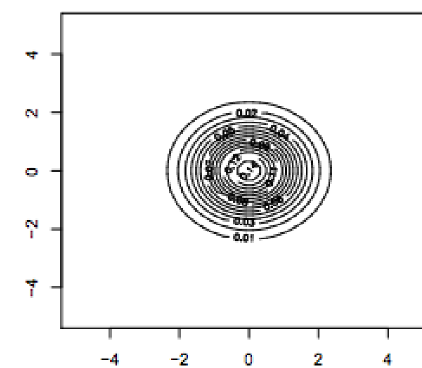
FOR ONE DATA DIMENSION OR MULTIPLE DIMENSIONS

EXAMINE DISTRIBUTIONS









DETECTING DUPLICATES

Title	Name
Ben-Hur	Anand Vaskar
Ben Hur	Anand Vaskkar
BEN-HUR	A. Vaskar
Ben-Hur (1959 film)	Vaskar, Anand

THESE *MIGHT* ALL BE THE SAME

SOME USEFUL DISTANCE METRICS

LEVENSHTEIN (“STRING-EDIT”) DISTANCE

How many edits do I need to change one value into another?

Ben-Hur
Ben Hur

DISTANCE = 1

Anand Vaskar
Anand Vaskkar

DISTANCE = 1

SOME USEFUL DISTANCE METRICS

LEVENSHTEIN ("STRING-EDIT") DISTANCE

How many edits do I need to change one value into another?

Ben-Hur
Ben-Hur (1959 film)

DISTANCE = 12

Anand Vaskar
Vaskar, Anand

DISTANCE = 12

SOME USEFUL DISTANCE METRICS

SOUNDEX / METAPHONE

How similar do they sound?

Ben-Hur	Anand Vaskar
Ben-Hurr	Anand Vaskkar
Been Her	Ahnund Vachkar

SOME USEFUL DISTANCE METRICS

“FINGERPRINTING” METHODS

Strip away unimportant details.

(e.g., remove punctuation, capitals, and sort)

Anand Vaskar ➞ anand vaskar

Vaskar, Anand ➞ anand vaskar

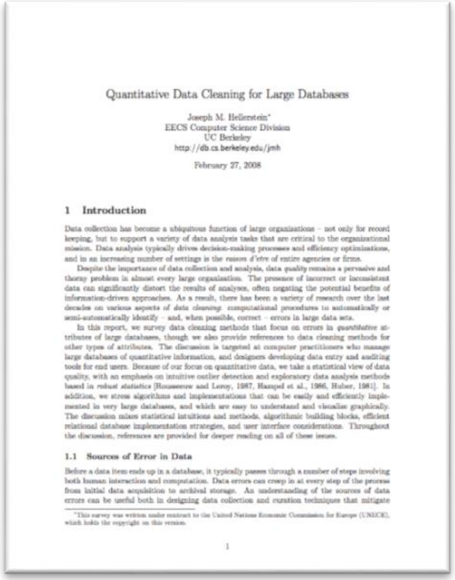
AND MANY MORE

STRING/KEY COMPARISONS

DISTANCE METRICS FOR NUMERIC DATA

e.g., HAMPEL X84 (UNIVARIATE), MAHALANOBIS (MULTIVARIATE)

"Quantitative Data Cleaning for Large Databases"
Hellerstein (2008)


Quantitative Data Cleaning for Large Databases

Joseph M. Hellerstein*
EECS Computer Science Division
UC Berkeley
http://db.cs.berkeley.edu/jmh

February 27, 2008

1 Introduction

Data collection has become a ubiquitous function of large organizations – not only for record keeping, but to support a variety of data analysis tasks that are critical to the organizational mission. Data analysis typically drives decision-making processes and efficiency optimizations, and in an increasing number of settings is the *raison d'etre* of entire agencies or firms.

Despite the importance of data collection and analysis, data *quality* remains a pervasive and thorny problem in almost every large organization. The presence of incorrect or inconsistent data can significantly distort the results of analyses, often negating the potential benefits of information-driven approaches. As a result, there has been a variety of research over the last decades on various aspects of *data cleaning*: computational procedures to automatically or semi-automatically identify – and, when possible, correct – errors in large data sets.

In this report, we survey data cleaning methods that focus on errors in *quantitative* attributes of large databases, though we also provide references to data cleaning methods for other types of attributes. The discussion is targeted at computer practitioners who manage large databases of quantitative information, and designers developing data entry and auditing tools for end users. Because of our focus on quantitative data, we take a statistical view of data quality, with an emphasis on intuitive outlier detection and exploratory data analysis methods based in *robust statistics* [Rousseeuw and Leroy, 1987, Hampel et al., 1986, Huber, 1981]. In addition, we stress algorithms and implementations that can be easily and efficiently implemented in very large databases, and which are easy to understand and visualize graphically. The discussion mixes statistical intuitions and methods, algorithmic building blocks, efficient relational database implementation strategies, and user interface considerations. Throughout the discussion, references are provided for deeper reading on all of these issues.

1.1 Sources of Error in Data

Before a data item ends up in a database, it typically passes through a number of steps involving both human interaction and computation. Data errors can creep in at every step of the process from initial data acquisition to archival storage. An understanding of the sources of data errors can be useful both in designing data collection and curation techniques that mitigate

*This survey was written under contract to the United Nations Economic Commission for Europe (UNECE), which holds the copyright on this version.

1

DECIDING HOW TO FIX PROBLEMS

YOU CAN DO ALMOST ALL OF
THIS IN **SQL …** BUT IT'S A LOT OF WORK

DECIDING HOW TO FIX PROBLEMS

WHICH DUPLICATE TO KEEP?

OUTLIERS: KEEP, REMOVE, OR REPAIR?

BADLY-STORED DATES, ADDRESSES, OR KEYS MAY NEED TO BE PARSED MANUALLY

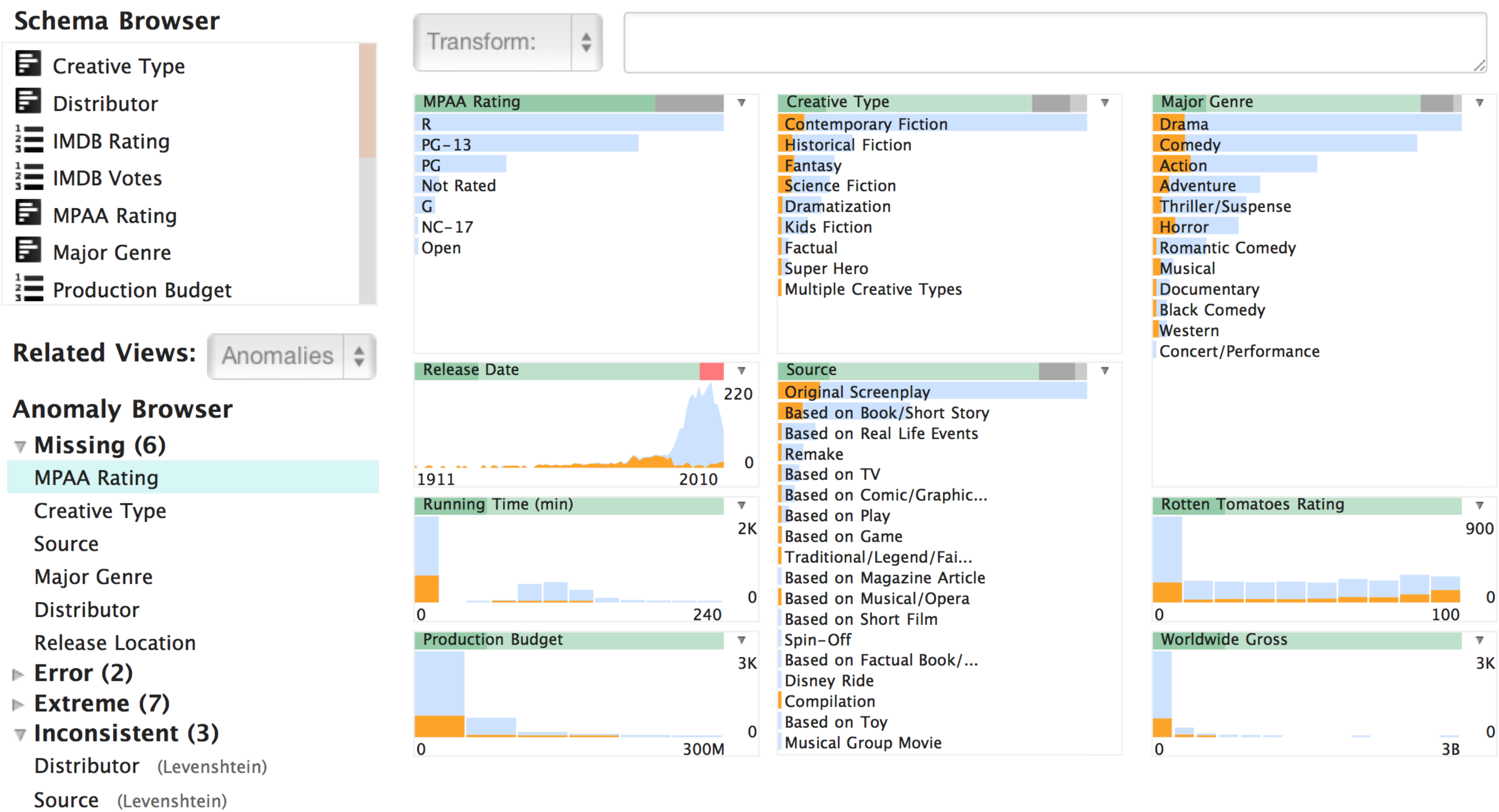
DECIDING HOW TO FIX PROBLEMS

FUZZY MATCHING SYSTEMS

MACHINE LEARNING TO DETECT/RESOLVE ERRORS

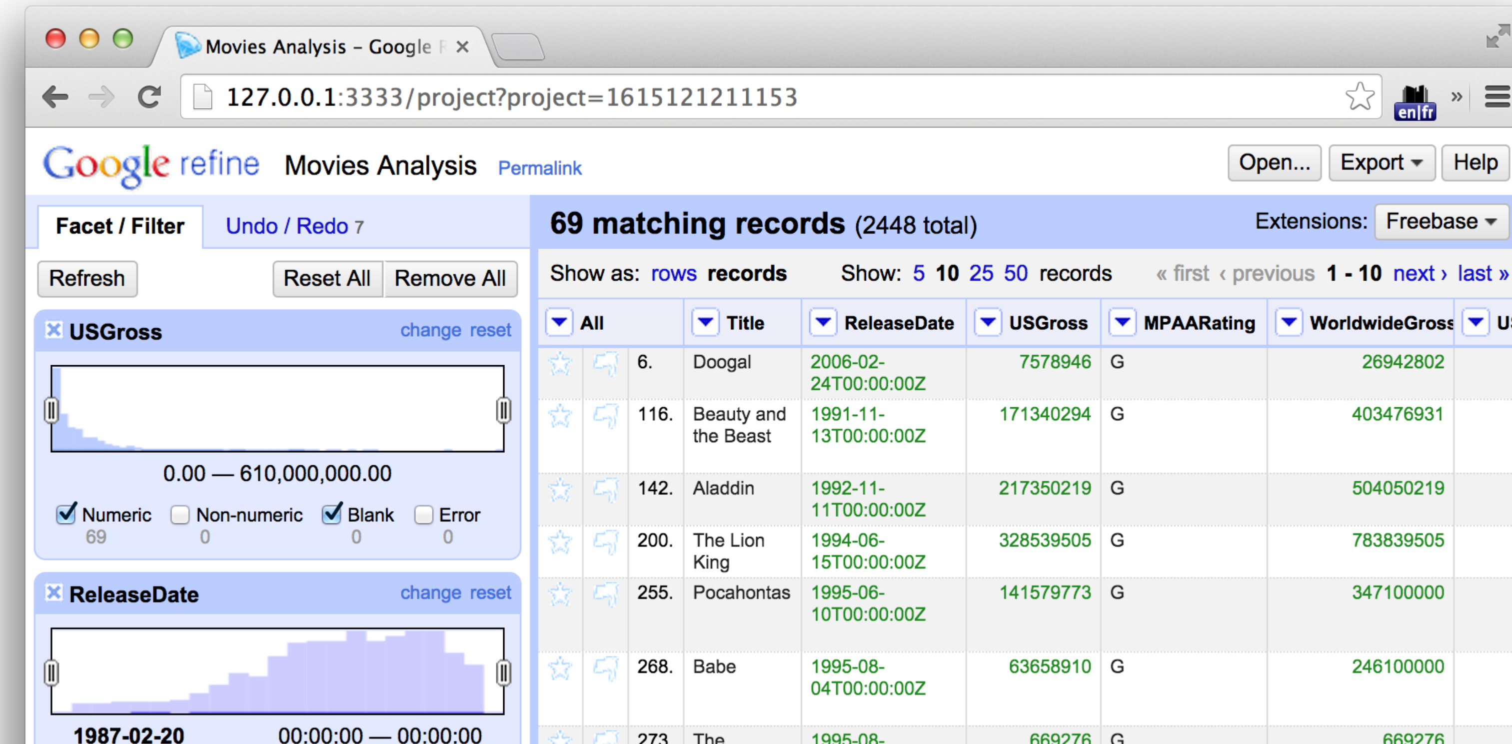
USUALLY REQUIRES HUMAN JUDGMENT
(ESPECIALLY FOR NEW DATA)

INTERACTIVE PROFILING



PROFILER [KANDEL ET AL. 2012]

PROFILING IN OPEN REFINE



SOME APPROACHES FOR IMPROVING DATA QUALITY

TOOLS FOR MANIPULATING AND CLEANING DATA

"WRANGLING" DATA

CLEANING AND TRANSFORMING DATASETS TO MAKE IT <u>POSSIBLE</u> TO ANALYZE AND VISUALIZE THEM

COMMON OPERATIONS

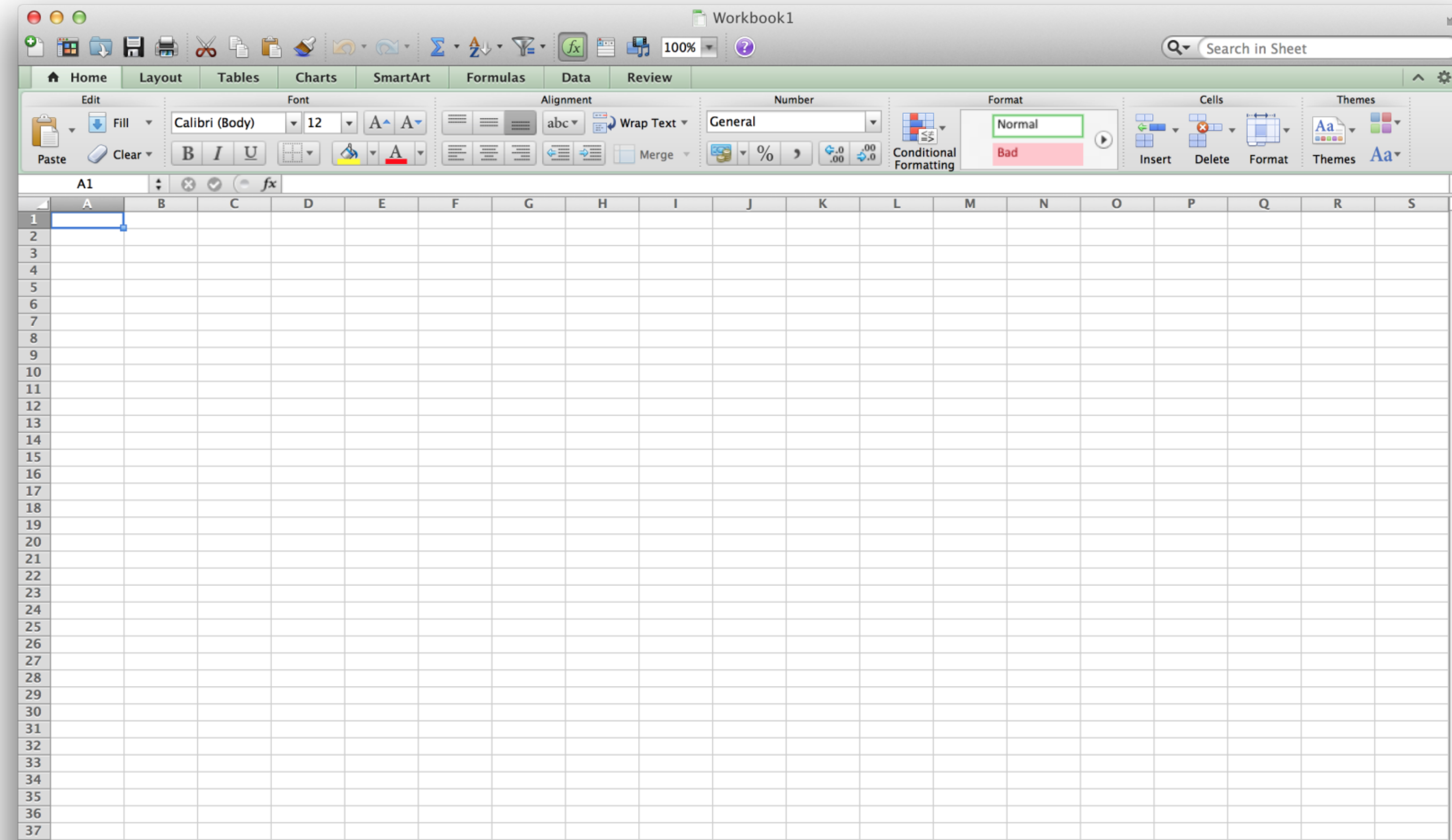
CORRECTING AND REMOVING ERRORS

CHANGING FORMATS

REMOVING FORMATTING

CONNECTING AND RESOLVING DATA

SPREADSHEETS



TRANSFORMATIONS ARE TIME-CONSUMING

"I spend more than half 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."

"Most of the time once you transform the data, the insights can be scarily obvious."

[Kandel 2012]

- Corrections
- Courts
- Crime Type
- Criminal Justice Data Improvement Program
- Employment and Expenditure
- Federal
- Law Enforcement
- Victims

New Releases

- FY 2011 Current Solicitations
- National Corrections Reporting Program, 2009 - Statistical Tables (update)
- Characteristics of Suspected Human Trafficking Incidents, 2008-2010
- Jail Inmates at Midyear 2010 - Statistical Tables
- Justice Assistance Grant (JAG) Program, 2010
- Workplace Violence, 1993-2009
- Punitive Damage Awards in State Courts, 2005
- Jails in Indian Country, 2009

MORE NEW RELEASES

Other Releases

A Dialogue Between the Bureau of Justice Statistics and Key Criminal Justice Data Users

In 2008 the Bureau of Justice Statistics (BJS) convened a multidisciplinary workshop for professionals who use justice statistics. Participants and representatives from academia, court systems, victim advocacy, and law enforcement communities provided feedback about how they use BJS statistical information and recommended ways BJS could optimize the value of data it collects and publishes. A Dialogue Between BJS and Key Criminal Justice Data Users is now available.

Announcements

BJS Visiting Fellows

Lynn A. Addington, Ph.D., Janet L. Lauritsen, Ph.D., and Avinash Bhati, Ph.D., are Visiting Fellows at the Bureau of Justice Statistics (BJS). They will conduct research designed to enhance the analytical approach and usability of specific BJS data collections. Visit the BJS Fellows page for additional information about Professor Addington, Professor Lauritsen, Mr. Bhati, and the BJS Visiting Fellows Program.

Stay Connected

JUSTSTATS | RSS | GOV Delivery

Interested in statistics?

Subscribe to JUSTSTATS

Get email notices of new crime and justice statistical materials as they become available from BJS, the FBI, and OJJDP.

Sign up

Once you subscribe, you will receive an email notification from JUSTSTATS when

Data Analysis Tools

Data Online
Dynamic interface that allows users to construct and download custom tables.

Crime and Justice Electronic Data Abstract spreadsheets
Aggregated data from a wide variety of published sources, intended for analytic use.

Federal Criminal Case Processing Statistics - FCCPS
The Federal Criminal Case Processing Statistics (FCCPS) tool permits an on-line analysis of suspects and defendants processed across stages of the Federal criminal justice system.

MORE DATA ANALYSIS TOOLS

Special Topics

- Deaths in Custody
- Drugs and Crime
- Homicide Trends
- Intimate Partner Violence
- Reentry Trends

MORE SPECIAL TOPICS

BJS Partners

- Federal Bureau of Investigation

State	2004	2005	2006	2007	2008			
Alabama	4029.3	3900	3937	3974.9	4081.9			
Alaska	3370.9	3615	3582	3373.9	2928.3			
Arizona	5073.3	4827	4741.6	4502.6	4087.3			
Arkansas	4033.1	4068	4021.6	3945.5	3843.7			
California	3423.9	3321	3175.2	3032.6	2940.3			
Colorado	3918.5	4041	3441.8	2991.3	2856.7			
Connecticut	2684.9	2579	2575	2470.6	2490.8			
Delaware	3283.6	3118	3474.5	3427.1	3594.7			
District of Columbia	4852.8	4490	4653.9	4916.3	5104.6			
Florida	4182.5	4013	3986.2	4088.8	4140.6			
Georgia	4223.5	4145	3928.8	3893.1	3996.6			
Hawaii	4795.5	4800	4219.9	4119.3	3566.5			
Idaho	2781	2697	2386.9	2264.2	2116.5			
Illinois	3174.1	3092	3019.6	2935.8	2932.6			
Indiana	3403.6	3460	3464.3	3386.5	3339.6			
Iowa	2904.8	2845	2870.3	2648.6	2440.5			
Kansas	4015.5	3806	3858.5	3693.8	3397			
Kentucky	2540.2	2531	2621.9	2524.6	2677.1			
Louisiana	4419.1	3696	4088.5	4196.1	3880.2			
Maine	2413.7	2419	2546.1	2448.3	2463.7			
Maryland	3640.7	3551	3481.2	3431.5	3516			
Massachusetts	2468.2	2358	2396	2399.2	2402			
Michigan	3066.1	3098	3226	3057.8	2945.7			
Minnesota	3041.6	3088	3088.8	3045	2858.1			
Mississippi	3481.1	3274	3213	3137.8	2941.7			
Missouri	3900.1	3929	3828.4	3828.2	3663.6			
Montana	2936.1	3146	2863.4	2863.6	2720.9			
Nebraska	3519.6	3432	3364.9	3142.8	2878.3			
Nevada	4210	4246	4099.6	3785.1	3456.4			

Year	Property Crime Rate				
Reported crime in Alabama					
2004	4029.3				
2005	3900				
2006	3937				
2007	3974.9				
2008	4081.9				
Reported crime in Alaska					
2004	3370.9				
2005	3615				
2006	3582				
2007	3373.9				
2008	2928.3				
Reported crime in Arizona					
2004	5073.3				
2005	4827				
2006	4741.6				
2007	4502.6				
2008	4087.3				

Year	Property Crime Rate				
Reported crime in Alabama					
2004	4029.3				
2005	3900				
2006	3937				
2007	3974.9				
2008	4081.9				
Reported crime in Alaska					
2004	3370.9				
2005	3615				
2006	3582				
2007	3373.9				
2008	2928.3				
Reported crime in Arizona					
2004	5073.3				
2005	4827				
2006	4741.6				
2007	4502.6				
2008	4087.3				

Year	Property Crime Rate				
Reported crime in Alabama					
2004	4029.3				
2005	3900				
2006	3937				
2007	3974.9				
2008	4081.9				
Reported crime in Alaska					
2004	3370.9				
2005	3615				
2006	3582				
2007	3373.9				
2008	2928.3				
Reported crime in Arizona					
2004	5073.3				
2005	4827				
2006	4741.6				
2007	4502.6				
2008	4087.3				

Year	Property Crime Rate
Reported crime in Alabama	
2004	4029.3
2005	3900
2006	3937
2007	3974.9
2008	4081.9
Reported crime in Alaska	
2004	3370.9
2005	3615
2006	3582
2007	3373.9
2008	2928.3
Reported crime in Arizona	
2004	5073.3
2005	4827
2006	4741.6
2007	4502.6
2008	4087.3

Year	Property Crime Rate				
Reported crime in Alabama					
2004	4029.3				
2005	3900				
2006	3937				
2007	3974.9				
2008	4081.9				
Reported crime in Alaska					
2004	3370.9				
2005	3615				
2006	3582				
2007	3373.9				
2008	2928.3				
Reported crime in Arizona					
2004	5073.3				
2005	4827				
2006	4741.6				
2007	4502.6				
2008	4087.3				

State	2004	2005	2006	2007	2008			
Alabama	4029.3	3900	3937	3974.9	4081.9			
Alaska	3370.9	3615	3582	3373.9	2928.3			
Arizona	5073.3	4827	4741.6	4502.6	4087.3			
Arkansas	4033.1	4068	4021.6	3945.5	3843.7			
California	3423.9	3321	3175.2	3032.6	2940.3			
Colorado	3918.5	4041	3441.8	2991.3	2856.7			
Connecticut	2684.9	2579	2575	2470.6	2490.8			
Delaware	3283.6	3118	3474.5	3427.1	3594.7			
District of Columbia	4852.8	4490	4653.9	4916.3	5104.6			
Florida	4182.5	4013	3986.2	4088.8	4140.6			
Georgia	4223.5	4145	3928.8	3893.1	3996.6			
Hawaii	4795.5	4800	4219.9	4119.3	3566.5			
Idaho	2781	2697	2386.9	2264.2	2116.5			
Illinois	3174.1	3092	3019.6	2935.8	2932.6			
Indiana	3403.6	3460	3464.3	3386.5	3339.6			
Iowa	2904.8	2845	2870.3	2648.6	2440.5			
Kansas	4015.5	3806	3858.5	3693.8	3397			
Kentucky	2540.2	2531	2621.9	2524.6	2677.1			
Louisiana	4419.1	3696	4088.5	4196.1	3880.2			
Maine	2413.7	2419	2546.1	2448.3	2463.7			
Maryland	3640.7	3551	3481.2	3431.5	3516			
Massachusetts	2468.2	2358	2396	2399.2	2402			
Michigan	3066.1	3098	3226	3057.8	2945.7			
Minnesota	3041.6	3088	3088.8	3045	2858.1			
Mississippi	3481.1	3274	3213	3137.8	2941.7			
Missouri	3900.1	3929	3828.4	3828.2	3663.6			
Montana	2936.1	3146	2863.4	2863.6	2720.9			
Nebraska	3519.6	3432	3364.9	3142.8	2878.3			
Nevada	4210	4246	4099.6	3785.1	3456.4			

Year	Property Crime Rate				
Reported crime in Alabama					
2004	4029.3				
2005	3900				
2006	3937				
2007	3974.9				
2008	4081.9				
Reported crime in Alaska					
2004	3370.9				
2005	3615				
2006	3582				
2007	3373.9				
2008	2928.3				
Reported crime in Arizona					
2004	5073.3				
2005	4827				
2006	4741.6				
2007	4502.6				
2008	4087.3				

State	Year	Property Crime Rate
	Reported crime in Alabama	
	2004	4029.3
	2005	3900
	2006	3937
	2007	3974.9
	2008	4081.9
	Reported crime in Alaska	
	2004	3370.9
	2005	3615
	2006	3582
	2007	3373.9
	2008	2928.3
	Reported crime in Arizona	
	2004	5073.3
	2005	4827
	2006	4741.6
	2007	4502.6
	2008	4087.3

State	Year	Property Crime Rate
	Reported crime in Alabama	
	2004	4029.3
	2005	3900
	2006	3937
	2007	3974.9
	2008	4081.9
	Reported crime in Alaska	
	2004	3370.9
	2005	3615
	2006	3582
	2007	3373.9
	2008	2928.3
	Reported crime in Arizona	
	2004	5073.3
	2005	4827
	2006	4741.6
	2007	4502.6
	2008	4087.3

State	Year	Property Crime Rate
	Reported crime in Alabama	
	2004	4029.3
	2005	3900
	2006	3937
	2007	3974.9
	2008	4081.9
	Reported crime in Alaska	
	2004	3370.9
	2005	3615
	2006	3582
	2007	3373.9
	2008	2928.3
	Reported crime in Arizona	
	2004	5073.3
	2005	4827
	2006	4741.6
	2007	4502.6
	2008	4087.3
	Reported crime in Arkansas	

State	Year	Property Crime Rate
Alabama	Reported crime in Alabama	
	2004	4029.3
	2005	3900
	2006	3937
	2007	3974.9
	2008	4081.9
	Reported crime in Alaska	
	2004	3370.9
	2005	3615
	2006	3582
	2007	3373.9
	2008	2928.3
	Reported crime in Arizona	
	2004	5073.3
	2005	4827
	2006	4741.6
	2007	4502.6
	2008	4087.3
	Reported crime in Arkansas	

State	Year	Property Crime Rate
Alabama	Reported crime in Alabama	
Alabama	2004	4029.3
Alabama	2005	3900
Alabama	2006	3937
Alabama	2007	3974.9
Alabama	2008	4081.9
	Reported crime in Alaska	
	2004	3370.9
	2005	3615
	2006	3582
	2007	3373.9
	2008	2928.3
	Reported crime in Arizona	
	2004	5073.3
	2005	4827
	2006	4741.6
	2007	4502.6
	2008	4087.3
	Reported crime in Arkansas	

State	Year	Property Crime Rate
Alabama	Reported crime in Alabama	
Alabama	2004	4029.3
Alabama	2005	3900
Alabama	2006	3937
Alabama	2007	3974.9
Alabama	2008	4081.9
	Reported crime in Alaska	
	2004	3370.9
	2005	3615
	2006	3582
	2007	3373.9
	2008	2928.3
	Reported crime in Arizona	
	2004	5073.3
	2005	4827
	2006	4741.6
	2007	4502.6
	2008	4087.3
	Reported crime in Arkansas	

State	Year	Property Crime Rate
Alabama	2004	4029.3
Alabama	2005	3900
Alabama	2006	3937
Alabama	2007	3974.9
Alabama	2008	[illegible]
	Reported crime in Alaska	
	2004	[illegible]
	2005	[illegible]
	2006	[illegible]
	2007	[illegible]
	2008	[illegible]
	Reported crime in Arizona	
	2004	[illegible]
	2005	4827
	2006	4741.6
	2007	4502.6
	2008	4087.3
	Reported crime in Arkansas	

REPEAT X 50

State	Year	Property Crime Rate
Alabama	2004	4029.3
Alabama	2005	3900
Alabama	2006	3937
Alabama	2007	3974.9
Alabama	2008	4081.9
Alaska	2004	3370.9
Alaska	2005	3615
Alaska	2006	3582
Alaska	2007	3373.9
Alaska	2008	2928.3
Arizona	2004	5073.3
Arizona	2005	4827
Arizona	2006	4741.6
Arizona	2007	4502.6
Arizona	2008	4087.3
Arkansas	2004	4033.1
Arkansas	2005	4068
Arkansas	2006	4021.6
Arkansas	2007	3945.5
Arkansas	2008	3843.7
California	[illegible]	[illegible]
California	[illegible]	[illegible]
California	2006	3175.2

RESHAPE (‘PIVOT’) THE TABLE

State	2004	2005	2006	2007	2008			
Alabama	4029.3	3900	3937	3974.9	4081.9			
Alaska	3370.9	3615	3582	3373.9	2928.3			
Arizona	5073.3	4827	4741.6	4502.6	4087.3			
Arkansas	4033.1	4068	4021.6	3945.5	3843.7			
California	3423.9	3321	3175.2	3032.6	2940.3			
Colorado	3918.5	4041	3441.8	2991.3	2856.7			
Connecticut	2684.9	2579	2575	2470.6	2490.8			
Delaware	3283.6	3118	3474.5	3427.1	3594.7			
District of Columbia	4852.8	4490	4653.9	4916.3	5104.6			
Florida	4182.5	4013	3986.2	4088.8	4140.6			
Georgia	4223.5	4145	3928.8	3893.1	3996.6			
Hawaii	4795.5	4800	4219.9	4119.3	3566.5			
Idaho	2781	2697	2386.9	2264.2	2116.5			
Illinois	3174.1	3092	3019.6	2935.8	2932.6			
Indiana	3403.6	3460	3464.3	3386.5	3339.6			
Iowa	2904.8	2845	2870.3	2648.6	2440.5			
Kansas	4015.5	3806	3858.5	3693.8	3397			
Kentucky	2540.2	2531	2621.9	2524.6	2677.1			
Louisiana	4419.1	3696	4088.5	4196.1	3880.2			
Maine	2413.7	2419	2546.1	2448.3	2463.7			
Maryland	3640.7	3551	3481.2	3431.5	3516			
Massachusetts	2468.2	2358	2396	2399.2	2402			
Michigan	3066.1	3098	3226	3057.8	2945.7			
Minnesota	3041.6	3088	3088.8	3045	2858.1			
Mississippi	3481.1	3274	3213	3137.8	2941.7			
Missouri	3900.1	[illegible]	[illegible]	[illegible]	[illegible]			
Montana	2936.1	[illegible]	[illegible]	[illegible]	[illegible]			
Nebraska	3519.6	[illegible]	[illegible]	[illegible]	[illegible]			
Nevada	4210	4246	4099.6	3785.1	3456.4			

RESHAPE ('PIVOT') THE TABLE

ONLY NOW ARE WE READY FOR ANALYSIS

State	2004	2005	2006	2007	2008			
Alabama	4029.3	3900	3937	3974.9	4081.9			
Alaska	3370.9	3615	3582	3373.9	2928.3			
Arizona	5073.3	4827	4741.6	4502.6	4087.3			
Arkansas	4033.1	4068	4021.6	3945.5	3843.7			
California	3423.9	3321	3175.2	3032.6	2940.3			
Colorado	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]			
Connecticut	[illegible]	2579	[illegible]	[illegible]	[illegible]			
Delaware	3283.6	3118	3474.5	3427.1	3594.7			
District of Columbia	4852.8	4490	4653.9	4916.3	5104.6			
Florida	4182.5	[illegible]	3986.2	[illegible]	[illegible]			
Georgia	4223.5	[illegible]	[illegible]	[illegible]	[illegible]			
Hawaii	4795.5	4800	4219.9	4119.3	3566.5			
Idaho	2781	2697	2386.9	2264.2	2116.5			
Illinois	3174.1	3092	[illegible]	[illegible]	[illegible]			
Indiana	3403.6	3460	[illegible]	[illegible]	[illegible]			
Iowa	2904.8	2845	[illegible]	[illegible]	[illegible]			
Kansas	4015.5	3806	3858.5	3693.8	3397			
Kentucky	2540.2	2531	2621.9	2524.6	2677.1			
Louisiana	4419.1	3696	4088.5	4196.1	3880.2			
Maine	2413.7	2419	2546.1	2448.3	2463.7			
Maryland	3640.7	3551	3481.2	3431.5	3516			
Massachusetts	2468.2	2358	2396	2399.2	2402			
Michigan	3066.1	3098	3226	3057.8	2945.7			
Minnesota	3041.6	3088	3088.8	3045	2858.1			
Mississippi	3481.1	3274	3213	3137.8	2941.7			
Missouri	3900.1	3929	3828.4	3828.2	3663.6			
Montana	2936.1	3146	2863.4	2863.6	2720.9			
Nebraska	3519.6	3432	3364.9	3142.8	2878.3			
Nevada	4210	4246	4099.6	3785.1	3456.4			

SPREADSHEETS

+ FAMILIAR
+ VISUAL

- TEDIOUS
- TIME-CONSUMING
- REPETITIVE

State	2004	2005	2006	2007	2008
Alabama	4029.3	3900	3937	3974.9	4081.9
Alaska	3370.9	3615	3582	3373.9	2928.3
Arizona	5073.3	4827	4741.6	4502.6	4087.3
Arkansas	4033.1	4068	[illegible]	[illegible]	[illegible]
California	3423.9	3321	3175.2	3032.6	2940.3
Colorado	3918.5	4041	3441.8	2991.3	2856.7
Connecticut	2684.9	2579	[illegible]	[illegible]	[illegible]
Delaware	3283.6	3118	[illegible]	[illegible]	[illegible]
District of Columbia	4852.8	4490	[illegible]	[illegible]	[illegible]
Florida	4182.5	4013	[illegible]	[illegible]	[illegible]
Georgia	4223.5	4145	[illegible]	[illegible]	[illegible]
Hawaii	4795.5	4800	[illegible]	[illegible]	[illegible]
Idaho	2781	2697	[illegible]	[illegible]	[illegible]
Illinois	3174.1	3092	[illegible]	[illegible]	[illegible]
Indiana	3403.6	3460	[illegible]	[illegible]	[illegible]
Iowa	2904.8	2845	[illegible]	[illegible]	[illegible]
Kansas	4015.5	3806	[illegible]	[illegible]	[illegible]
Kentucky	2540.2	2531	2621.9	2524.6	2677.1
Louisiana	4419.1	3696	4088.5	4196.1	3880.2
Maine	2413.7	2419	2546.1	2448.3	2463.7
Maryland	3640.7	3551	3481.2	3431.5	3516
Massachusetts	2468.2	2358	2396	2399.2	2402

```
from wrangler import dw
import sys

w = dw.DataWrangler()

# Split data repeatedly on newline  into  rows
w.add(dw.Split(column="data", result="row", on="\n", max=0)

# Split data repeatedly on ','
w.add(dw.Split(column="data", c

# Delete empty rows
w.add(dw.Filter(row=dw.Row(cond

# Extract from split after 'in
w.add(dw.Extract(column="split"

# Fill extract  with values from above
w.add(dw.Fill(column="extract", direction="down"))

# Delete  rows where split1 is null
```

SCRIPTS

+ REUSABLE
+ SCALABLE

- HARD
- TEDIOUS
- TIME-CONSUMING

INTERACTIVE DATA CLEANING

Trifacta Wrangler
https://www.trifacta.com/

Wrangler (Stanford HCI Group)
http://vis.stanford.edu/wrangler/

OpenRefine (formerly Google Refine)
http://openrefine.org/

INTERACTIVE DATA CLEANING BY EXAMPLE

```
Reported crime in Alabama,
,
2004,4029.3
2005,3900
2006,3937
2007,3974.9
2008,4081.9
,
Reported crime in Alaska,
,
2004,3370.9
2005,3615
2006,3582
2007,3373.9
2008,2928.3
,
Reported crime in Arizona,
,
2004,5073.3
2005,4827
2006,4741.6
2007,4502.6
2008,4087.3
,
Reported crime in Arkansas,
,
2004,4033.1
2005,4068
2006,4021.6
2007,3945.5
2008,3843.7
,
Reported crime in California,
,
2004,3423.9
2005,3321
2006,3175.2
```

[(http://vimeo.com/19185801)](http://vimeo.com/19185801)

WRANGLER [KANDEL ET AL. 2011]

[redacted]

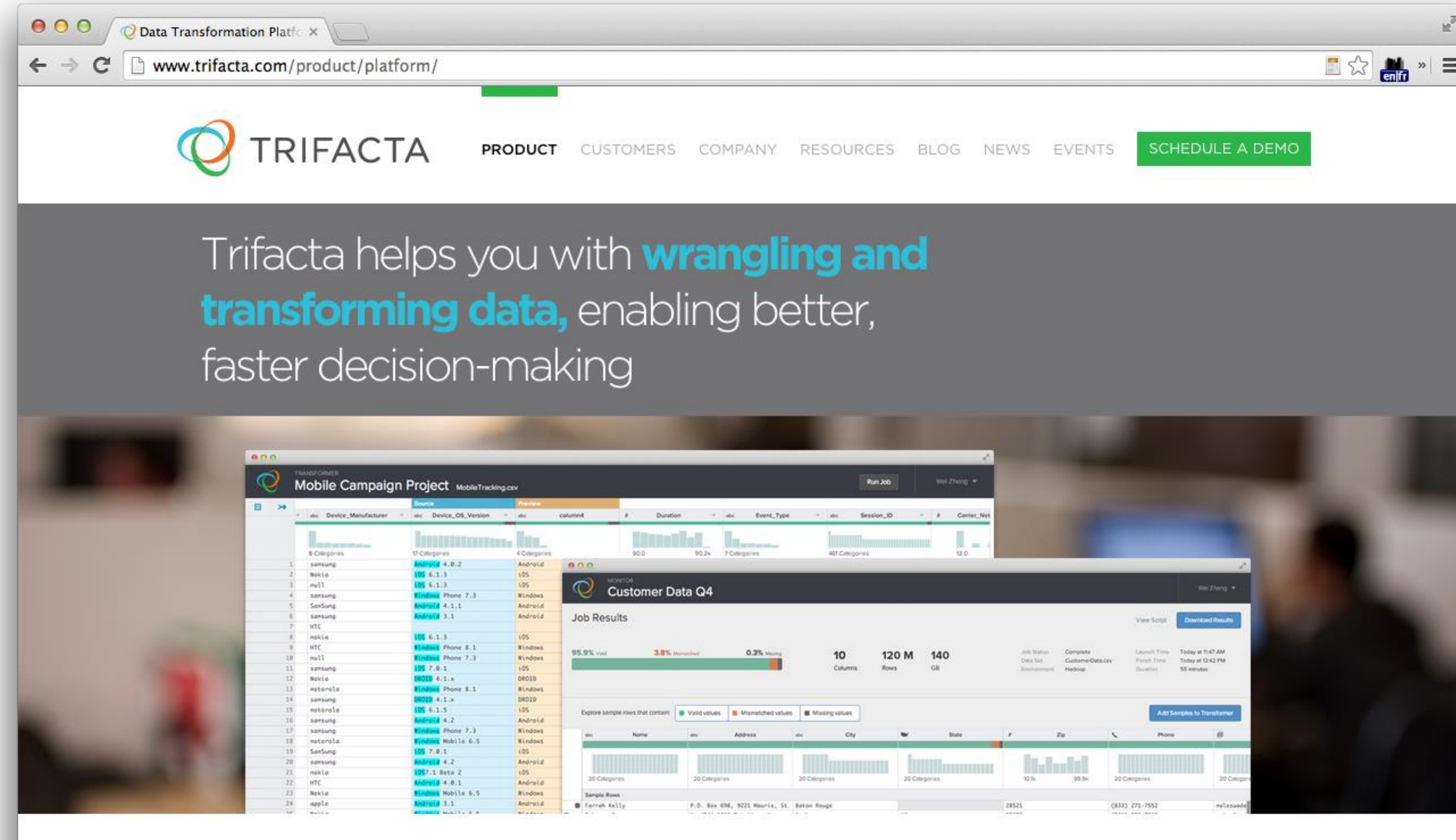
	split	extract	split1
1	2004	Alabama	4029.3
2	2005	Alabama	3900
3	2006	Alabama	3937
4	2007	Alabama	3974.9
5	2008	Alabama	4081.9
6	2004	Alaska	3370.9
7	2005	Alaska	3615
8	2006	Alaska	3582
9	2007	Alaska	3373.9
10	2008	Alaska	2928.3
11	2004	Arizona	5073.3
12	2005	Arizona	4827
13	2006	Arizona	4741.6
14	2007	Arizona	4502.6
15	2008	Arizona	4087.3
16	2004	Arkansas	4033.1
17	2005	Arkansas	4068
18	2006	Arkansas	4021.6
19	2007	Arkansas	3945.5
20	2008	Arkansas	3843.7
21	2004	California	3423.9
22	2005	California	3321
23	2006	California	3175.2
24	2007	California	3032.6
25	2008	California	2940.3

WRANGLER [KANDEL ET AL. 2011]

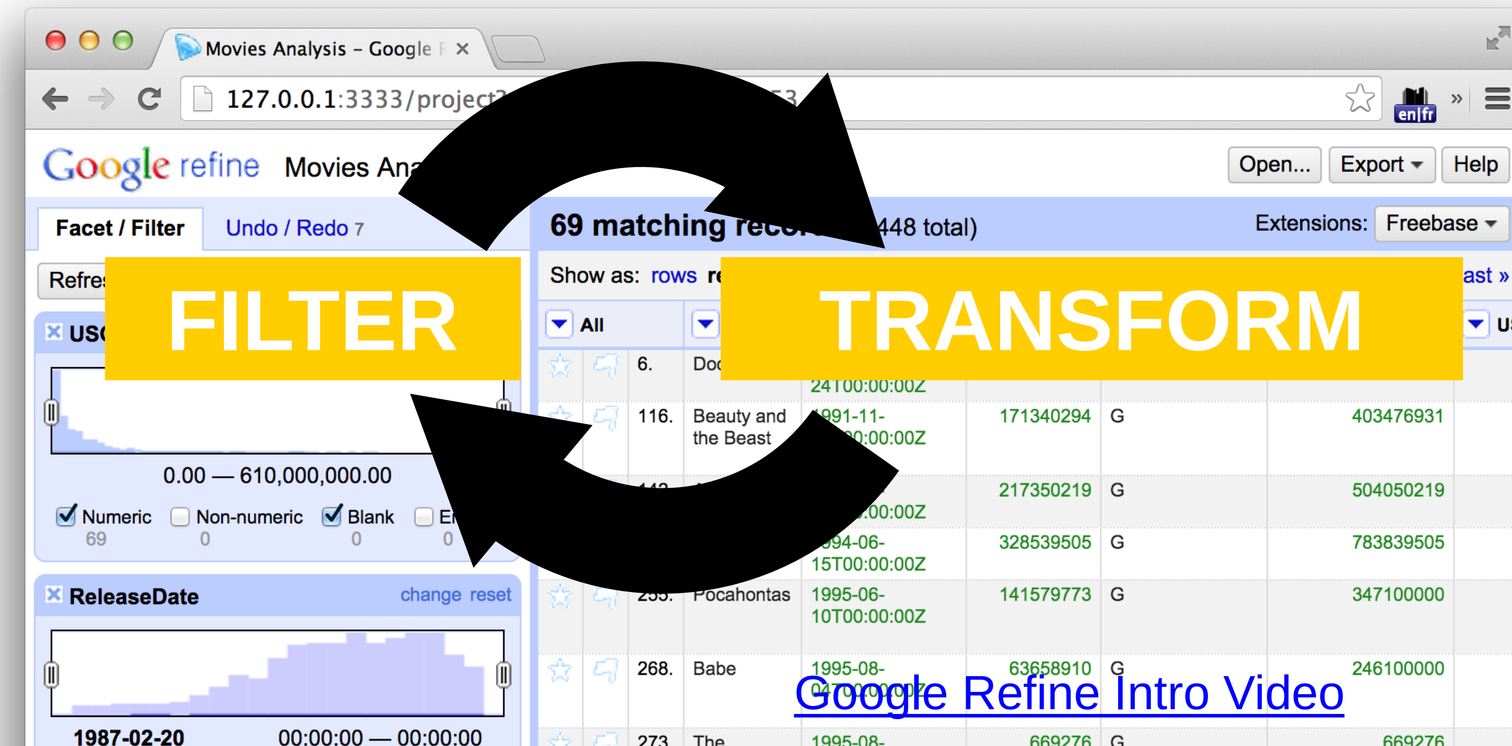
```
from wrangler import dw
import sys

if(len(sys.argv) < 3):
	sys.exit('Error: Please include an input and output file.  Example python script.py
input.csv output.csv')

w = dw.DataWrangler()

# Split data repeatedly on newline  into  rows
w.add(dw.Split(column=["data"],
		table=0,
		status="active",
		drop=True,
		result="row",
		update=False,
		insert_position="right",
		row=None,
		on="\n",
		before=None,
		after=None,
		ignore_between=None,
		which=1,
		max=0,
		positions=None,
		quote_character=None))
```

WRANGLER [KANDEL ET AL. 2011]

RESEARCH → PRODUCTS



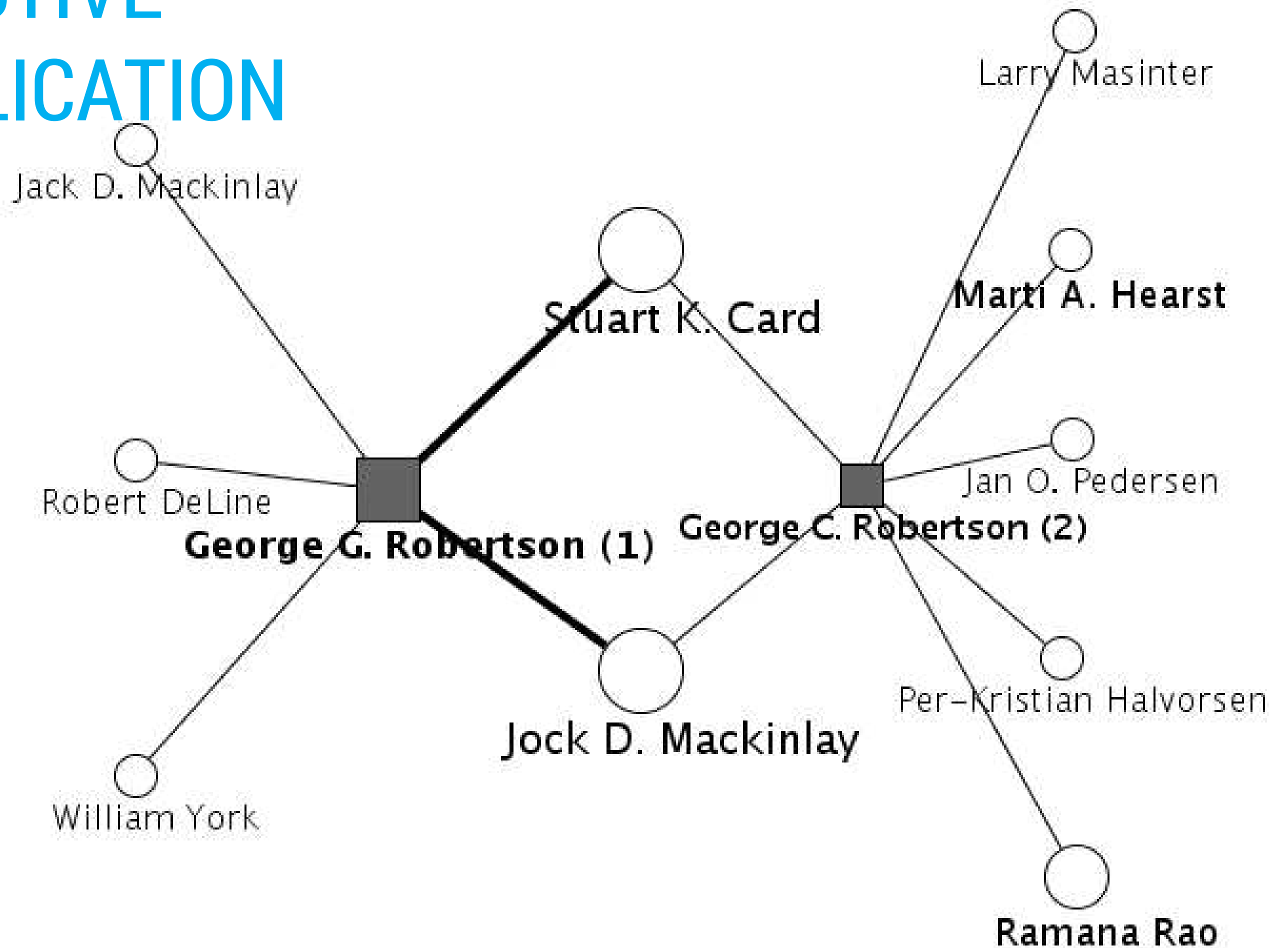
DATA CLEANING IN GOOGLE REFINE



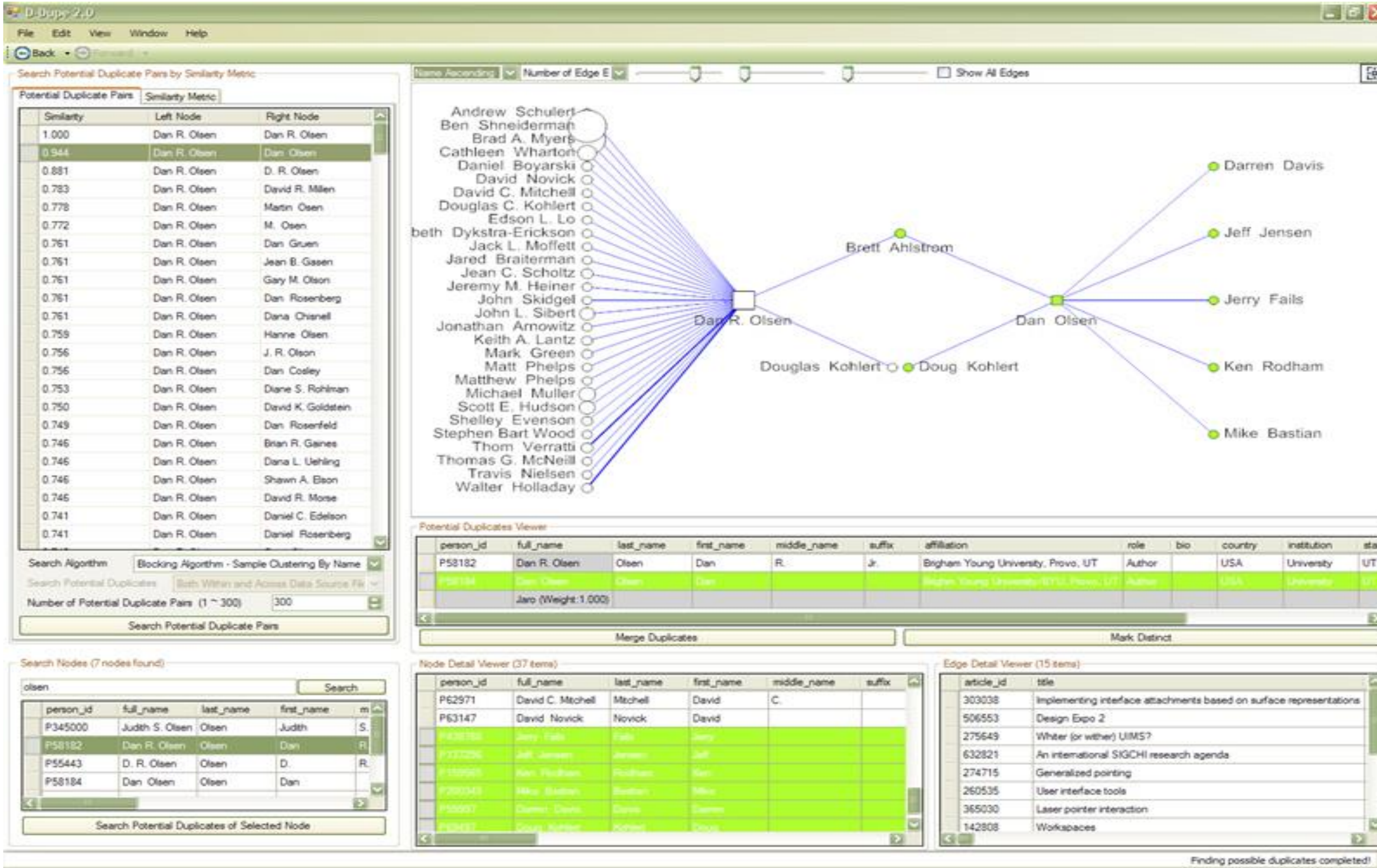
[Google Refine Intro Video](#)

THERE ARE LOTS OF OTHER SPECIALIZED TOOLS

INTERACTIVE DE-DUPLICATION



D-DUPE [BILGIC ET AL. 2008]



D-DUPE [BILGIC ET AL. 2008]

REFERENCES

"Quantitative Data Cleaning for Large Databases"

Hellerstein (2008)

Quantitative Data Cleaning for Large Databases

Joseph M. Hellerstein*
EECS Computer Science Division
UC Berkeley
http://db.cs.berkeley.edu/jmh

February 27, 2008

1 Introduction

Data collection has become a ubiquitous function of large organizations – not only for record keeping, but to support a variety of data analysis tasks that are critical to the organizational mission. Data analysis typically drives decision-making processes and efficiency optimizations, and in an increasing number of settings is the *raison d'etre* of entire agencies or firms.

Despite the importance of data collection and analysis, data *quality* remains a pervasive and thorny problem in almost every large organization. The presence of incorrect or inconsistent data can significantly distort the results of analyses, often negating the potential benefits of information-driven approaches. As a result, there has been a variety of research over the last decades on various aspects of *data cleaning*: computational procedures to automatically or semi-automatically identify – and, when possible, correct – errors in large data sets.

In this report, we survey data cleaning methods that focus on errors in *quantitative* attributes of large databases, though we also provide references to data cleaning methods for other types of attributes. The discussion is targeted at computer practitioners who manage large databases of quantitative information, and designers developing data entry and auditing tools for end users. Because of our focus on quantitative data, we take a statistical view of data quality, with an emphasis on intuitive outlier detection and exploratory data analysis methods based in *robust statistics* [Rousseeuw and Leroy, 1987, Hampel et al., 1986, Huber, 1981]. In addition, we stress algorithms and implementations that can be easily and efficiently implemented in very large databases, and which are easy to understand and visualize graphically. The discussion mixes statistical intuitions and methods, algorithmic building blocks, efficient relational database implementation strategies, and user interface considerations. Throughout the discussion, references are provided for deeper reading on all of these issues.

1.1 Sources of Error in Data

Before a data item ends up in a database, it typically passes through a number of steps involving both human interaction and computation. Data errors can creep in at every step of the process from initial data acquisition to archival storage. An understanding of the sources of data errors can be useful both in designing data collection and curation techniques that mitigate

*This survey was written under contract to the United Nations Economic Commission for Europe (UNECE), which holds the copyright on this version.

TIDY DATA PRINCIPLES

Tidy Data

Hadley Wickham
RStudio

Abstract

A huge amount of effort is spent cleaning data to get it ready for analysis, but there has been little research on how to make data cleaning as easy and effective as possible. This paper tackles a small, but important, component of data cleaning: data tidying. Tidy datasets are easy to manipulate, model and visualise, and have a specific structure: each variable is a column, each observation is a row, and each type of observational unit is a table. This framework makes it easy to tidy messy datasets because only a small set of tools are needed to deal with a wide range of un-tidy datasets. This structure also makes it easier to develop tidy tools for data analysis, tools that both input and output tidy datasets. The advantages of a consistent data structure and matching tools are demonstrated with a case study free from mundane data manipulation chores.

Keywords: data cleaning, data tidying, relational databases, R.

TIDY DATA

= data structured to facilitate analysis

labelled columns

labelled rows

	treatmenta	treatmentb
John Smith	—	2
Jane Doe	16	11
Mary Johnson	3	1

= data structure

TIDY DATA

Data semantics

variables
= column names

observations
= rows

name	trt	result
John Smith	a	—
Jane Doe	a	16
Mary Johnson	a	3
John Smith	b	2
Jane Doe	b	11
Mary Johnson	b	1

values

TIDY DATA

- Variables are columns
- Observations are rows
- Each observational unit in one table

In addition: put fixed variables first and then measured variables last

If you order, do so by the first variable

MESSY DATA - EXAMPLES

Column headers = values, not variables

religion	<$10k	$10-20k	$20-30k	$30-40k	$40-50k	$50-75k
Agnostic	27	34	60	81	76	137
Atheist	12	27	37	52	35	70
Buddhist	27	21	30	34	33	58
Catholic	418	617	732	670	638	1116
Don't know/refused	15	14	15	11	10	35
Evangelical Prot	575	869	1064	982	881	1486
Hindu	1	9	7	9	11	34
Historically Black Prot	228	244	236	238	197	223
Jehovah's Witness	20	27	24	24	21	30
Jewish	19	19	25	25	30	95

MESSY DATA - EXAMPLES

Better (most of the time)

Process to produce this
= melting

religion	income	freq
Agnostic	<$10k	27
Agnostic	$10-20k	34
Agnostic	$20-30k	60
Agnostic	$30-40k	81
Agnostic	$40-50k	76
Agnostic	$50-75k	137
Agnostic	$75-100k	122
Agnostic	$100-150k	109
Agnostic	>150k	84
Agnostic	Don't know/refused	96

YOU!

This table is good for data entry but not analysis.
How do we tidy it up?

year	artist	track	time	date.entered	wk1	wk2	wk3
2000	2 Pac	Baby Don't Cry	4:22	2000-02-26	87	82	72
2000	2Ge+her	The Hardest Part Of ...	3:15	2000-09-02	91	87	92
2000	3 Doors Down	Kryptonite	3:53	2000-04-08	81	70	68
2000	98^0	Give Me Just One Nig...	3:24	2000-08-19	51	39	34
2000	A*Teens	Dancing Queen	3:44	2000-07-08	97	97	96
2000	Aaliyah	I Don't Wanna	4:15	2000-01-29	84	62	51
2000	Aaliyah	Try Again	4:03	2000-03-18	59	53	38
2000	Adams, Yolanda	Open My Heart	5:30	2000-08-26	76	76	74

year	artist	time	track	date	week	rank
2000	2 Pac	4:22	Baby Don't Cry	2000-02-26	1	87
2000	2 Pac	4:22	Baby Don't Cry	2000-03-04	2	82
2000	2 Pac	4:22	Baby Don't Cry	2000-03-11	3	72
2000	2 Pac	4:22	Baby Don't Cry	2000-03-18	4	77
2000	2 Pac	4:22	Baby Don't Cry	2000-03-25	5	87
2000	2 Pac	4:22	Baby Don't Cry	2000-04-01	6	94
2000	2 Pac	4:22	Baby Don't Cry	2000-04-08	7	99
2000	2Ge+her	3:15	The Hardest Part Of ...	2000-09-02	1	91
2000	2Ge+her	3:15	The Hardest Part Of ...	2000-09-09	2	87
2000	2Ge+her	3:15	The Hardest Part Of ...	2000-09-16	3	92
2000	3 Doors Down	3:53	Kryptonite	2000-04-08	1	81
2000	3 Doors Down	3:53	Kryptonite	2000-04-15	2	70
2000	3 Doors Down	3:53	Kryptonite	2000-04-22	3	68
2000	3 Doors Down	3:53	Kryptonite	2000-04-29	4	67
2000	3 Doors Down	3:53	Kryptonite	2000-05-06	5	66

MESSY DATA - EXAMPLES

Multiple variables in one column

country	year	m014	m1524	m2534	m3544	m4554	m5564	m65	mu	f014
AD	2000	0	0	1	0	0	0	0	—	—
AE	2000	2	4	4	6	5	12	10	—	3
AF	2000	52	228	183	149	129	94	80	—	93
AG	2000	0	0	0	0	0	0	1	—	1
AL	2000	2	19	21	14	24	19	16	—	3
AM	2000	2	152	130	131	63	26	21	—	1
AN	2000	0	0	1	2	0	0	0	—	0
AO	2000	186	999	1003	912	482	312	194	—	247
AR	2000	97	278	594	402	419	368	330	—	121
AS	2000	—	—	—	—	1	1	—	—	—

FIRST WE MELT

How do we do this…?

country	year	m014	m1524	m2534	m3544	m4554	m5564	m65	mu	f014
AD	2000	0	0	1	0	0	0	0	—	—
AE	2000	2	4	4	6	5	12	10	—	3
AF	2000	52	228	183	149	129	94	80	—	93
AG	2000	0	0	0	0	0	0	1	—	1
AL	2000	2	19	21	14	24	19	16	—	3
AM	2000	2	152	130	131	63	26	21	—	1
AN	2000	0	0	1	2	0	0	0	—	0
AO	2000	186	999	1003	912	482	312	194	—	247
AR	2000	97	278	594	402	419	368	330	—	121
AS	2000	—	—	—	—	1	1	—	—	—

country	year	column	cases
AD	2000	m014	0
AD	2000	m1524	0
AD	2000	m2534	1
AD	2000	m3544	0
AD	2000	m4554	0
AD	2000	m5564	0
AD	2000	m65	0
AE	2000	m014	2
AE	2000	m1524	4
AE	2000	m2534	4
AE	2000	m3544	6
AE	2000	m4554	5
AE	2000	m5564	12
AE	2000	m65	10
AE	2000	f014	3

NEXT: SPLIT COLUMNS

country	year	sex	age	cases
AD	2000	m	0-14	0
AD	2000	m	15-24	0
AD	2000	m	25-34	1
AD	2000	m	35-44	0
AD	2000	m	45-54	0
AD	2000	m	55-64	0
AD	2000	m	65+	0
AE	2000	m	0-14	2
AE	2000	m	15-24	4
AE	2000	m	25-34	4
AE	2000	m	35-44	6
AE	2000	m	45-54	5
AE	2000	m	55-64	12
AE	2000	m	65+	10
AE	2000	f	0-14	3

MESSY DATA - EXAMPLES

Multi observational units in the same table

year	artist	track	time	date.entered	wk1	wk2	wk3
2000	2 Pac	Baby Don't Cry	4:22	2000-02-26	87	82	72
2000	2Ge+her	The Hardest Part Of ...	3:15	2000-09-02	91	87	92
2000	3 Doors Down	Kryptonite	3:53	2000-04-08	81	70	68
2000	98^0	Give Me Just One Nig...	3:24	2000-08-19	51	39	34
2000	A*Teens	Dancing Queen	3:44	2000-07-08	97	97	96
2000	Aaliyah	I Don't Wanna	4:15	2000-01-29	84	62	51
2000	Aaliyah	Try Again	4:03	2000-03-18	59	53	38
2000	Adams, Yolanda	Open My Heart	5:30	2000-08-26	76	76	74

TIDYER & MORE SPACE EFFICIENT

id	artist	track	time
1	2 Pac	Baby Don't Cry	4:22
2	2Ge+her	The Hardest Part Of ...	3:15
3	3 Doors Down	Kryptonite	3:53
4	3 Doors Down	Loser	4:24
5	504 Boyz	Wobble Wobble	3:35
6			
7			
8	Aaliyah	I Don't Wanna	4:15
9	Aaliyah	Try Again	4:03
10	Adams, Yolanda	Open My Heart	5:30
11	Adkins, Trace	More	3:05
12	Aguilera, Christina	Come On Over Baby	3:38
13	Aguilera, Christina	I Turn To You	4:00
14	Aguilera, Christina	What A Girl Wants	3:18
15	Alice Deejay	Better Off Alone	6:50

id	date	rank
1	2000-02-26	87
1	2000-03-04	82
1	2000-03-11	72
1	2000-03-18	77
1	2000-03-25	87
2	2000-09-02	91
2	2000-09-09	87
2	2000-09-16	92
3	2000-04-08	81
3	2000-04-15	70
3	2000-04-22	68
3	2000-04-29	67
3	2000-05-06	66

BUT not all tools work well across multiple tables

MORE EXAMPLES HERE

Tidy Data

Hadley Wickham
RStudio

Abstract

A huge amount of effort is spent cleaning data to get it ready for analysis, but there has been little research on how to make data cleaning as easy and effective as possible. This paper tackles a small, but important, component of data cleaning: data tidying. Tidy datasets are easy to manipulate, model and visualise, and have a specific structure: each variable is a column, each observation is a row, and each type of observational unit is a table. This framework makes it easy to tidy messy datasets because only a small set of tools are needed to deal with a wide range of un-tidy datasets. This structure also makes it easier to develop tidy tools for data analysis, tools that both input and output tidy datasets. The advantages of a consistent data structure and matching tools are demonstrated with a case study free from mundane data manipulation chores.

Keywords: data cleaning, data tidying, relational databases, R.

CSVKIT

csvkit 1.0.2

About

build passing | dependencies up-to-date | coverage 88% | downloads no longer available | pypi v1.0.2 | license MIT

python 2.7, 3.3, 3.4, 3.5, 3.6

csvkit is a suite of command-line tools for converting to and working with CSV, the king of tabular file formats.

It is inspired by pdftk, gdal and the original csvcut tool by Joe Germuska and Aaron Bycoffe.

If you need to do more complex data analysis than csvkit can handle, use agate.

Important links: