

DATA CLEANING & DATA MANIPULATION

WESLEY WILLETT

VISUAL ANALYTICS

1 OCT 2014

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

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”

Step 1: Activity/Equipment Type Step 2: Add a Map **Step 3: Additional Details**

Date of Activity: Duration: : :

! **Oops! You forgot to enter a duration for this activity.**

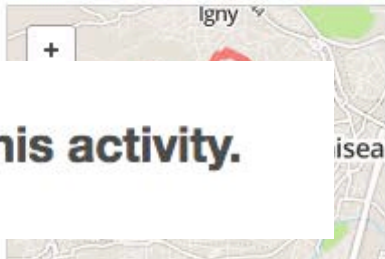
mi

Training Plan:

Average Heart Rate (optional): bpm

Add An Activity

Activity Details



Activity Type: Running
Equipment Type: None
Route: None
Distance: 5.62 mi.
Duration: --:--

MEASUREMENT ERRORS

SENSOR ISSUES

MALFUNCTIONS

PLACEMENT

INTERFERENCE

MISCALIBRATION



DISTILLATION

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 INCONSISTENCIES

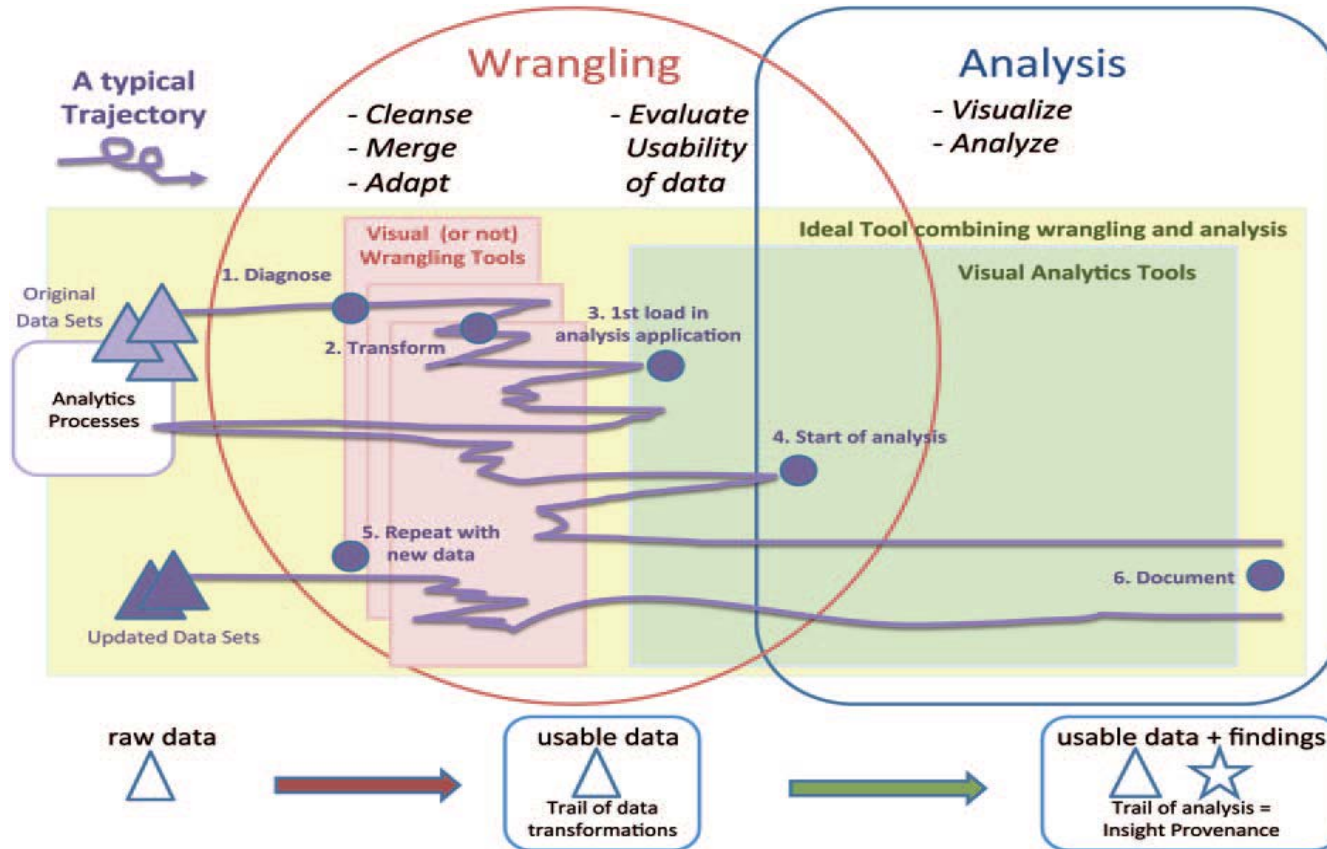
WHY IS THIS IMPORTANT?

MOST 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



SOME DATA QUALITY ISSUES

MISSING DATA

MISSED MEASUREMENTS, REDACTED ITEMS, INCOMPLETE FORMS, ETC.

ERRONEOUS VALUES

MISPELLINGS, 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



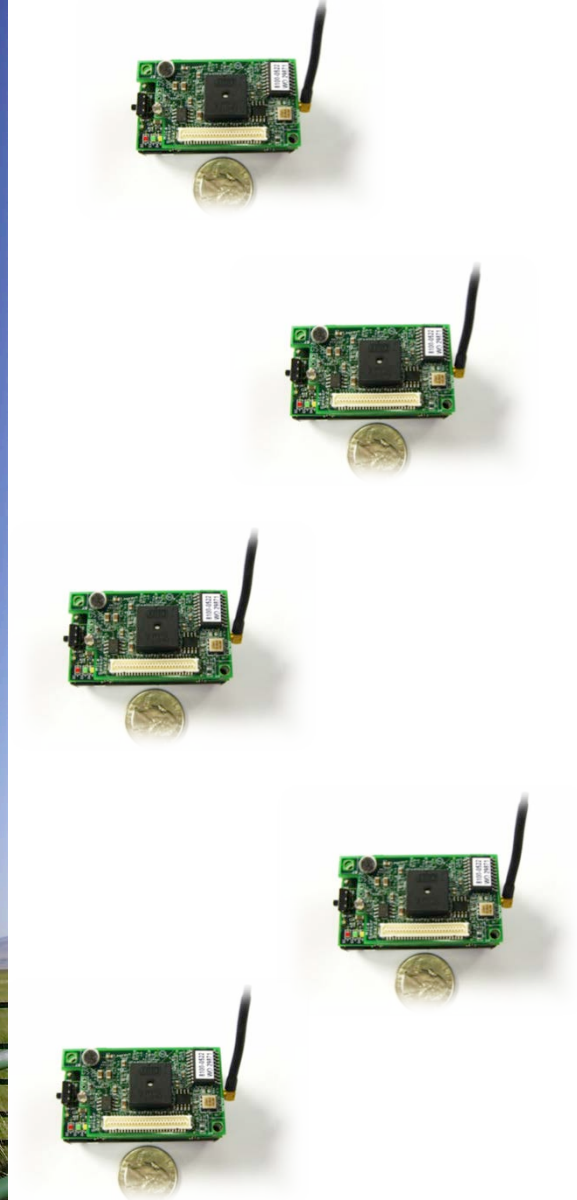
CHECK SENSORS BEFORE DEPLOYMENT
(AND PERIODICALLY REVALIDATE THEM)

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

°C

INTEGRITY CONSTRAINTS

Temperatures must be between
-50°C and 50°C.

TEMPERATURE

-60 °C

INTEGRITY CONSTRAINTS

TEMPERATURE °C

INTEGRITY CONSTRAINTS DO NOT PREVENT BAD DATA

ENFORCING CONSTRAINTS LEADS TO FRUSTRATION

FRICION 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.

FRICITION AND PREDICTION

PRINCIPLE 1

DATA QUALITY SHOULD BE CONTROLLED VIA FEEDBACK, NOT ENFORCEMENT.

PRINCIPLE 2

FRICITION MERITS EXPLANATION.

PRINCIPLE 3

ANNOTATION SHOULD BE EASIER THAN OMISSION OR SUBVERSION.

FRICITION AND PREDICTION



FRICION AND PREDICTION

This value seems low.
Are you sure?

TEMPERATURE

-60 °C

Sensor disabled.

USHER

[Chen et al. 2010]

The image shows a screenshot of a web-based data entry form titled "National Aids Control Programme CTC2 Database Patient Registration". The form is designed for patient registration and includes various fields for personal and medical information. It features a navigation menu on the left with buttons for "Home", "Log off", and "Exit Database". The main form area has tabs for "Register new patient", "Search patients", "Show all patients", and "Delete patient". The registration form includes fields for Patient ID, File Reference, First Name(s), Surname, Sex, Date of Birth (with an "or Age" option), Age, Marital Status, Phone/contact details, Date of first positive HIV test, Date confirmed HIV positive, Referred from, Region, District (Wilaya), Division (Tarafa), Ward (Kata), Village / Mtaa (Mtaa au Kiji), Chairperson (Mwenyekiti wa Kiji), Ten Cell Leader (Mjumba/Balazi), Ten Cell Leader Contact, Household Head (Mkuu wa Kaya), Household Head contact details, Helper / treatment supporter (Jina la Msaidizi wa karibu), Helper / treatment supporter contact details, Community Support Organisation / Group, Drug Allergies, Prior Exposure, and Notes. There are also buttons for "Add / Edit Village or chairperson", "Patient classification", "Family information", and "Return".

National Aids Control Programme
CTC2 Database

Patient Registration

Register new patient | Search patients | Show all patients | Delete patient

Home
Log off
Exit Database

The United Republic of Tanzania

Patient ID:

File Reference:

First Name(s):

Surname:

Sex:

Date of Birth: or Age:

Age:

Marital Status:

Phone/contact details:

Date of first positive HIV test:

Date confirmed HIV positive:

Referred from:

Region:

District (Wilaya):

Division (Tarafa):

Ward (Kata):

Village / Mtaa (Mtaa au Kiji):

Chairperson (Mwenyekiti wa Kiji):

Ten Cell Leader (Mjumba/Balazi):

Ten Cell Leader Contact:

Household Head (Mkuu wa Kaya):

Household Head contact details:

Helper / treatment supporter (Jina la Msaidizi wa karibu):

Helper / treatment supporter contact details:

Community Support Organisation / Group:

Drug Allergies:

Prior Exposure:

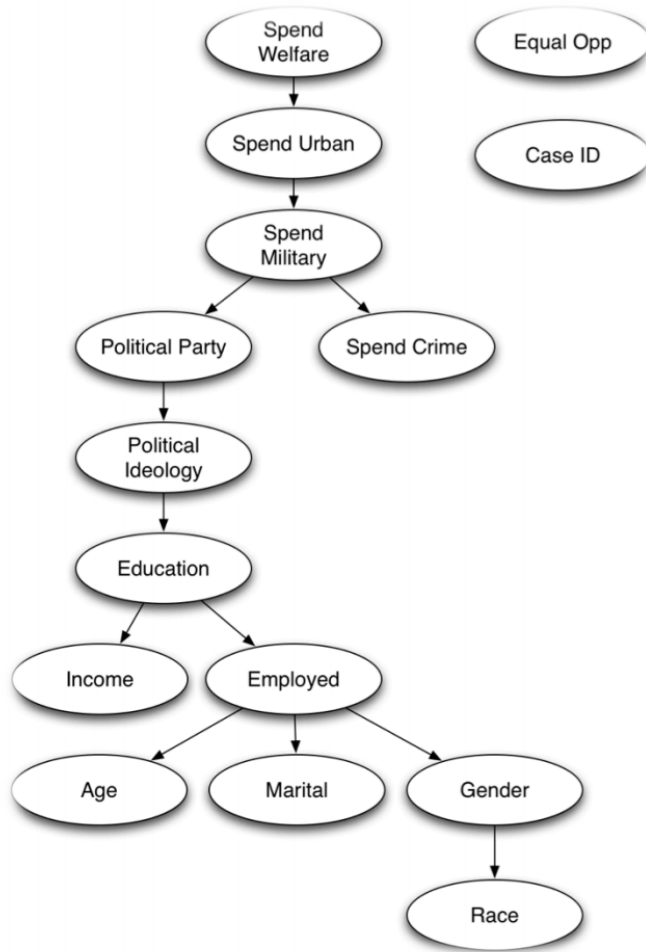
Notes:

Add / Edit Village or chairperson

Patient classification
Family information

Return

MS Access data entry forms for Tanzanian HIV/AIDS monitoring



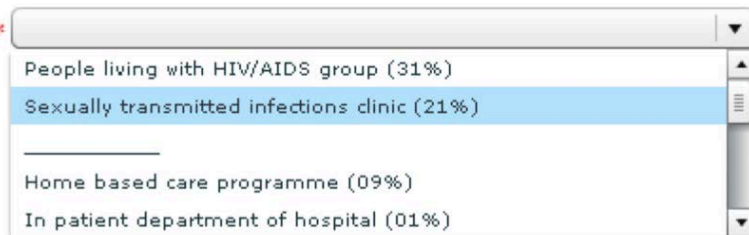
BUILD A MODEL to predict dependencies and relationships between questions.

DYNAMIC ORDERING

ALWAYS ASK THE
MOST APPROPRIATE
NEXT QUESTION

SUGGEST THE MOST
LIKELY ANSWERS

Select the referring *
organization



A dropdown menu with a white background and a grey border. The menu is open, showing a list of options. The first option is "Sexually transmitted infections clinic (21%)", which is highlighted in blue. Other options include "People living with HIV/AIDS group (31%)", "Home based care programme (09%)", and "In patient department of hospital (01%)".

- People living with HIV/AIDS group (31%)
- Sexually transmitted infections clinic (21%)
- Home based care programme (09%)
- In patient department of hospital (01%)

Select the referring *
organization



A dropdown menu with a red background and a grey border. The menu is closed, and the selected option is "In patient department of hospital".

In patient department of hospital

Select the district *
code



A dropdown menu with a white background and a blue border. The menu is open, showing a list of options. The first option is "Dodoma Rural", which is highlighted in blue. Other options include "Dodoma Urban".

- Dodoma Rural
- Dodoma Urban

Choose the *
patient's gender

Male (40%)

Female (59%)

[Chen et al. 2010]

SMART RE-ASKING AND SUGGESTIONS

1. Given * 1234
name

WARNING! CHECK YOUR ANSWER!

FRICION

AUTOMATING CONSTRAINTS

| --NA--

Birere

Kabuyanda

Kikagati

Mwizi

| Nyakitunda

DETECTING ERRORS

DATA AUDITING AND ERROR DETECTION

LOOK FOR OUTLIERS / ANOMALIES

EXAMINE DATA TYPES

SCHEMA CHECKING

VALIDATE WITH OTHER DATA

OTHER HEURISTICS

HISTORICALLY – MORE FOCUS ON AUTOMATED APPROACHES

“PROFILING” DATA

**UNDERSTANDING WHAT ASSUMPTIONS
YOU CAN MAKE ABOUT DATA**

**INTERACTIVELY IDENTIFYING
DATA QUALITY ISSUES**

AN EXAMPLE

FROM THE SNOW

IMDb Find Movies, TV shows, Celebrities and more

Movies TV News Videos

Rotten Tomatoes by Flixster

Search movies, actors, critics

MOVIES

AVAILABLE NOW DIGITAL

The Hunger Games (2012)

TOMATOMETER 84% All Critics | Top

Average Rating: 7.2/10
Reviews Counted: 257
Fresh: 217 | Rotten: 40

MY RATING

WANT TO SEE IT NOT INTERESTED

PLAY TRAILER

Tickets & Showtimes Add to Netflix

Like 29k
Tweet 4,466
Like 352
+1 838

MOVIE INFO

Every year in the ruins of what was once North America, 16 Panem forces each of its twelve districts to send a teenage Hunger Games. A twisted punishment for a past uprising of intimidation tactic, The Hunger Games are a nationally tele fight with one another until one survivor remains. Pitted ag have prepared for these Games their entire ... More

PG-13, 2 hr, 22 min. In Theaters
Drama, Mystery & Suspense, Science Fiction, Box Office & Fantasy
Lionsgate

Directed By: Gary Ross
Written By: Suzanne Collins, Gary Ross, Billy Ray

Friend Ratings

★★★★★ March 27, 2012

Jon Whetstone

The Hunger Games Trailer & Photos

More Photos (39)

More Trailers at

The Hunger Ga
April 29, 2011

Cast

Jennifer Lawrence
Katniss Everdeen

Josh Hutcherson
Peeta Melark

Liam Hemsworth

Woody Harris

Now Playing
In 6 theaters near San Francisco, CA. [Change location.](#)

The Hunger Ga

PG-13 142 min - Action

Your rating: 7.6
Ratings: 7.6/10
Reviews: 1,170

Set in a future where the twelve districts to fight Katniss Everdeen volunteered place for the latest match

Director: Gary Ross
Writers: Gary Ross (screenplay), and 2 more
Stars: Jennifer Lawrence, Hemsworth

Watch Trailer + W

98 photos | 23 videos | 9081 news articles | full cast

7 nominations [See more awards](#)

Related Videos

Music Video
The Hunger Games

Trailer
The Hunger Games

See all 23

People who liked this also liked...

Help

The Ides of March

The Way, Way Back

Help

THE NUMBERS

BOX OFFICE DATA, MOVIE STARS, IDLE SPECULATION

Learn About Our Research and Data Services

Wednesday, May 16, 2012

save May in sales event! It's going on now!

Great deals available at your Toyota dealer.

TOYOTA moving forward

Ready to Buy

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 Box

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...

THE 2012 Mazda3

Starting at \$15,200*

IF IT'S NOT WORTH DRIVING, IT'S NOT WORTH BUILDING.

EXPLORE NOW

mazdausa.com

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

MISPELLINGS, 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
Inconsistent	Missing value	Find NULL/empty values
	Measurement units	Clustering Euclidean Distance
		Outlier Detection z-score, Hampel X84
	Misspelling	Clustering Levenshtein Distance
Incorrect	Ordering	Clustering Atomic Strings
	Representation	Clustering Structure Extraction
	Special characters	Clustering Structure Extraction
	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
	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

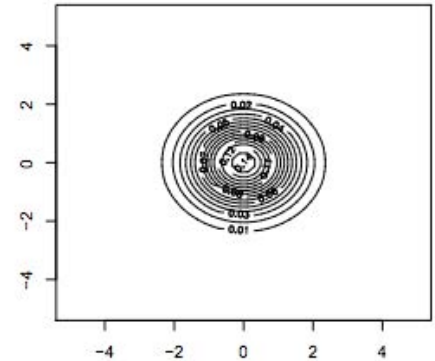
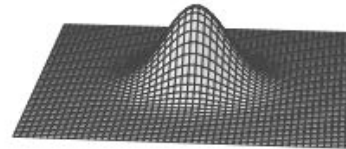
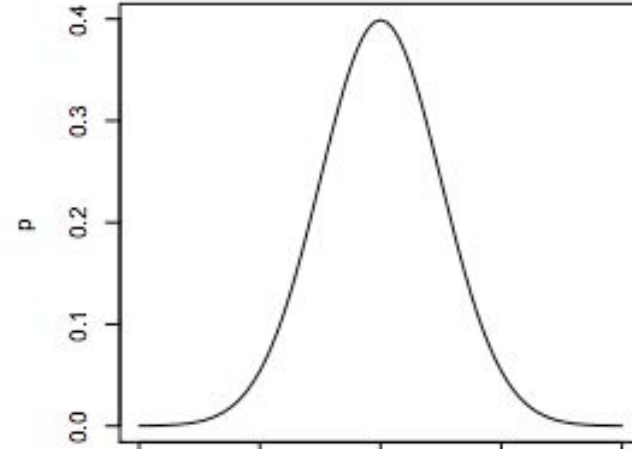
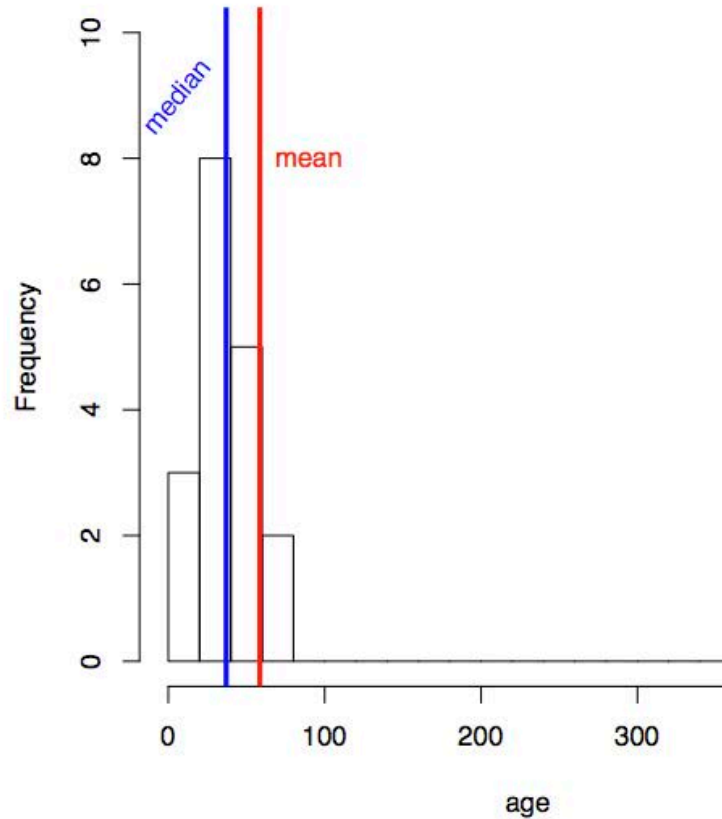
OUTLIER DETECTION

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

FOR **ONE DATA DIMENSION** OR **MULTIPLE DIMENSIONS**

EXAMINE DISTRIBUTIONS

Histogram of age



DETECTING DUPLICATES

Title

Ben-Hur

Ben Hur

BEN-HUR

Ben-Hur (1959 film)

Name

Anand Vaskar

Anand Vaskkar

A. Vaskar

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

Ben-Hurr

Been Her

Anand Vaskar

Anand Vaskkar

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. Hellerstein¹
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 reason *d’être* 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, when negating the potential benefits of information-driven approaches. As a result, there has been a variety of research over the last decade 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 covers 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

¹This survey was written under contract to the United Nations Economic Commission for Europe (UNECE), which holds the copyright on this version.

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

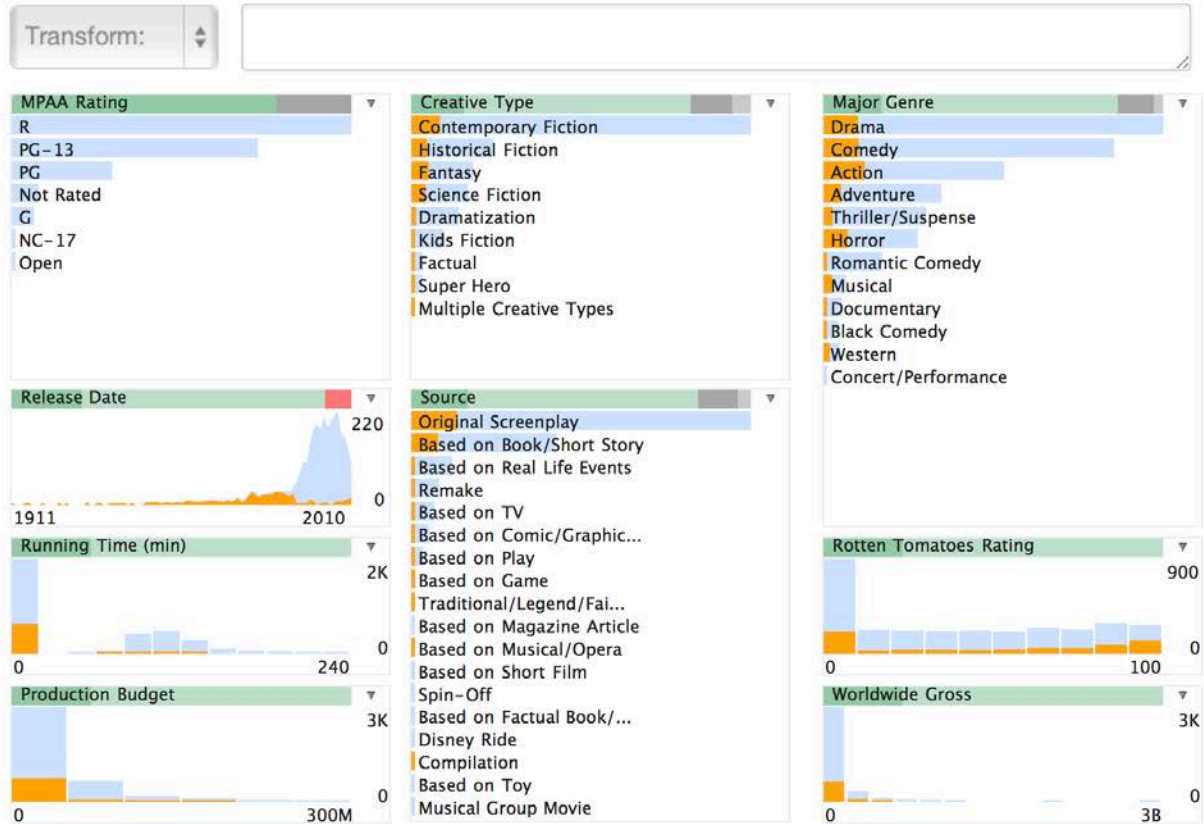
Schema Browser

- Creative Type
- Distributor
- IMDB Rating
- IMDB Votes
- MPAA Rating
- Major Genre
- Production Budget

Related Views: Anomalies

Anomaly Browser

- Missing (6)
 - MPAA Rating
 - Creative Type
 - Source
 - Major Genre
 - Distributor
 - Release Location
- Error (2)
- Extreme (7)
- Inconsistent (3)
 - Distributor (Levenshtein)
 - Source (Levenshtein)



PROFILING IN OPEN REFINE

Movies Analysis - Google

127.0.0.1:3333/project?project=1615121211153

Google refine Movies Analysis Permalink

Open... Export Help

Facet / Filter Undo / Redo 7

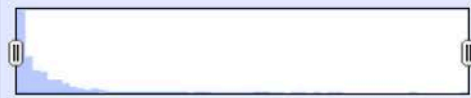
Refresh Reset All Remove All

69 matching records (2448 total) Extensions: Freebase

Show as: rows records Show: 5 10 25 50 records « first < previous 1 - 10 next > last »

All	Title	ReleaseDate	USGross	MPAARating	WorldwideGross	USI
6.	Doogal	2006-02-24T00:00:00Z	7578946	G	26942802	
116.	Beauty and the Beast	1991-11-13T00:00:00Z	171340294	G	403476931	
142.	Aladdin	1992-11-11T00:00:00Z	217350219	G	504050219	
200.	The Lion King	1994-06-15T00:00:00Z	328539505	G	783839505	
255.	Pocahontas	1995-06-10T00:00:00Z	141579773	G	347100000	
268.	Babe	1995-08-04T00:00:00Z	63658910	G	246100000	
273.	The	1995-08-	669276	G	669276	

USGross change reset




0.00 — 610,000,000.00

Numeric Non-numeric Blank Error

69 0 0 0

ReleaseDate change reset



1987-02-20 00:00:00 — 00:00:00

**SOME APPROACHES FOR
IMPROVING DATA QUALITY**

**TOOLS FOR MANIPULATING
AND CLEANING DATA**

“WRANGLING” DATA

**CLEANING AND TRANSFORMING DATASETS TO MAKE IT
POSSIBLE TO ANALYZE AND VISUALIZE THEM**

COMMON OPERATIONS

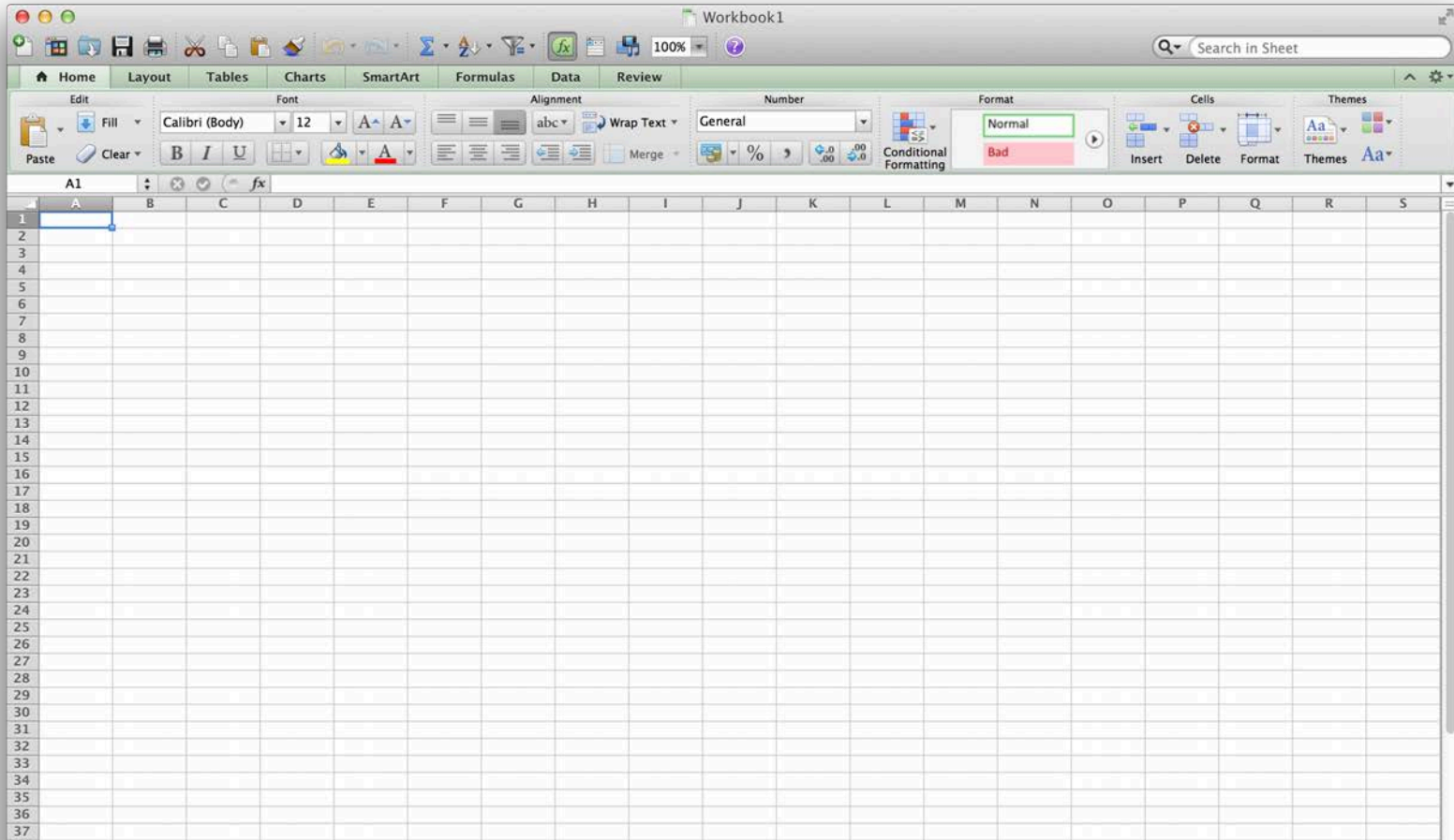
CORRECTING AND REMOVING ERRORS

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](#)

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

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

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.

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](#)



- [Intimate Partner Violence](#)
- [Reentry Trends](#)

[▶ MORE SPECIAL TOPICS](#)

BJS Partners

- [Federal Bureau of Investigation](#)

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

GOAL

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
	2008	4087.3

CREATE 'STATE' COLUMN

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
	2008	4087.3

DELETE EMPTY ROWS

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
	Reported crime in Arkansas	

EXTRACT STATE NAME

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
	Reported crime in Arkansas	

EXTRACT STATE NAME

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
	Reported crime in Arkansas	

FILL DOWN

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
	Reported crime in Arkansas	

DELETE ROW

State	Year	Property Crime Rate
Alabama	2004	4029.3
Alabama	2005	3900
Alabama	2006	3937
Alabama	2007	3974.9
Alabama	2008	4081.0
Reported crime in Alaska		
	2004	
Reported crime in Arizona		
	2004	
	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		
California		
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	3900	3900	3900	3900			
Montana	2936.1	3100	3100	3100	3100			
Nebraska	3519.6	3400	3400	3400	3400			
Nevada	4210	4246	4099.6	3785.1	3456.4			

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					
Florida					
Georgia					
Hawaii					
Idaho					
Illinois					
Indiana					
Iowa					
Kansas					
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

**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			
Arkansas	4033.1	4068			
California	3423.9	3321			
Colorado	3918.5	4041	3441.8	2991.3	2856.7
Connecticut	2684.9	2579			
Delaware	3283.6	3118			
District of Columbia	4852.8	4490			
Florida	4182.5	4013			
Georgia	4223.5	4145			
Hawaii	4795.5	4800			
Idaho	2781	2697			
Illinois	3174.1	3092			
Indiana	3403.6	3460			
Iowa	2904.8	2845			
Kansas	4015.5	3806			
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

SPREADSHEETS

- + FAMILIAR
- + VISUAL

- TEDIOUS
- TIME-CONSUMING
- REPETITIVE


```
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",
```

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

```
# 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

Wrangler

vis.stanford.edu/wrangler

Wrangler (Stanford HCI Group)

<http://vis.stanford.edu/wrangler/>

Refine ^{OPEN} 

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>

WRANGLER [KANDEL ET AL. 2011]

#	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

The image displays the Trifacta website and its data processing interface. The top portion shows the website's header with the Trifacta logo, navigation links for 'PRODUCT', 'CUSTOMERS', 'COMPANY', 'RESOURCES', 'BLOG', 'NEWS', and 'EVENTS', and a prominent green 'SCHEDULE A DEMO' button. Below the header, a grey banner contains the text: 'Trifacta helps you with **wrangling and transforming data**, enabling better, faster decision-making'. The bottom portion of the image shows two overlapping screenshots of the Trifacta data processing interface. The background screenshot, titled 'Mobile Campaign Project', shows a data table with columns for 'Device_Manufacturer', 'Device_OS_Version', 'Category', 'Duration', 'Event_Type', 'Session_ID', and 'Center_Net'. It includes several bar charts and a list of data rows with columns for device manufacturer, OS version, and OS. The foreground screenshot, titled 'Customer Data Q4', shows a 'Job Results' summary with a progress bar indicating 95.9% valid, 3.8% mismatched, and 0.3% missing values. It also displays a table of customer data with columns for Name, Address, City, State, Zip, and Phone, and a 'Sample Row' for 'Farruk, Kelly'.

DATA CLEANING IN GOOGLE REFINE

The screenshot displays the Google Refine interface for a project named "Movies Analysis". The browser address bar shows "127.0.0.1:3333/project?". The main area shows "69 matching records" out of "448 total". A "Facet / Filter" section on the left includes a histogram for "US" and a "ReleaseDate" facet with a histogram. A table of data is visible, with columns for movie titles, release dates, and other attributes. Two large yellow boxes with black arrows are overlaid on the interface. The first box, labeled "FILTER", points to the "Facet / Filter" section. The second box, labeled "TRANSFORM", points to the main data table. A blue link at the bottom right reads "Google Refine Intro Video".

FILTER

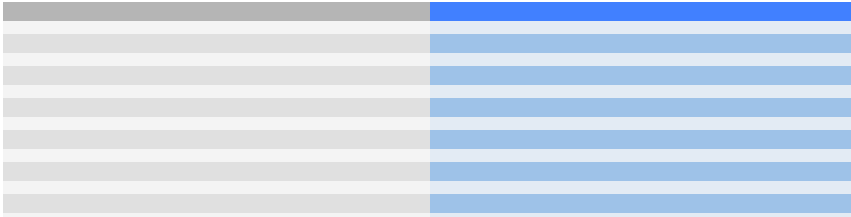
TRANSFORM

[Google Refine Intro Video](#)

**A FEW OTHER
IMPORTANT POINTS**

JOINING DATA

ADDING COLUMNS OR METADATA
FROM ANOTHER SOURCE



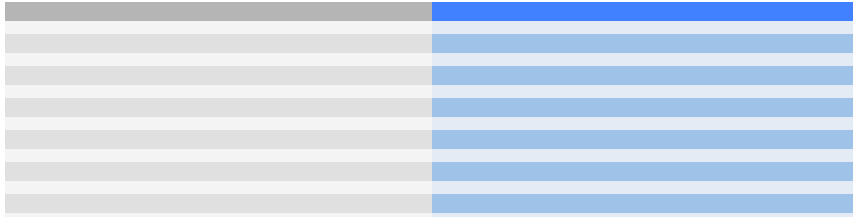
FOR EXAMPLE

NEW PATIENT FILE (+ OLD FILE)

POSTAL CODE (+ CITY INFORMATION)

JOINING DATA

ADDING COLUMNS OR METADATA
FROM ANOTHER SOURCE

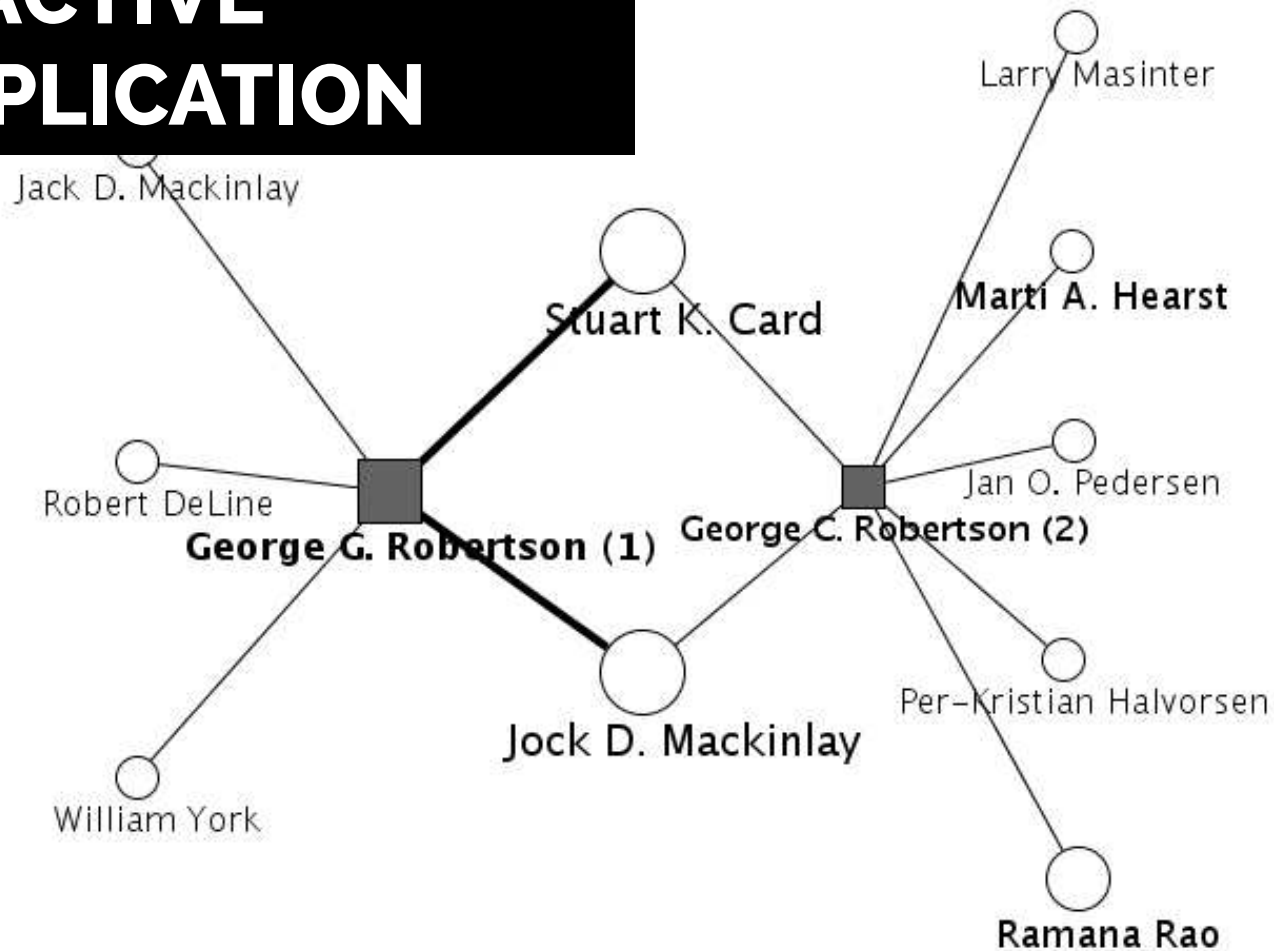


HELP VALIDATE AND CORRECT ERRORS

WILL REVISIT LATER (TIME PERMITTING)

**THERE ARE LOTS OF OTHER
SPECIALIZED TOOLS**

INTERACTIVE DE-DUPLICATION



D-DUPE [BILGIC ET AL. 2008]

D-Dupe 2.0

File Edit View Window Help

Search Potential Duplicate Pairs by Similarity Metric

Back Forward

Search Potential Duplicate Pairs

Similarity	Left Node	Right Node
1.000	Dan R. Olsen	Dan R. Olsen
0.944	Dan R. Olsen	Dan Olsen
0.881	Dan R. Olsen	D. R. Olsen
0.783	Dan R. Olsen	David R. Millen
0.778	Dan R. Olsen	Martin Osen
0.772	Dan R. Olsen	M. Osen
0.761	Dan R. Olsen	Dan Gruen
0.761	Dan R. Olsen	Jean B. Gasen
0.761	Dan R. Olsen	Gary M. Olson
0.761	Dan R. Olsen	Dan Rosenberg
0.761	Dan R. Olsen	Dana Chianell
0.758	Dan R. Olsen	Hanne Olsen
0.756	Dan R. Olsen	J. R. Olson
0.756	Dan R. Olsen	Dan Cosley
0.753	Dan R. Olsen	Diane S. Rohlman
0.750	Dan R. Olsen	David K. Goldstein
0.749	Dan R. Olsen	Dan Rosenfeld
0.746	Dan R. Olsen	Brian R. Gaines
0.746	Dan R. Olsen	Dana L. Uehling
0.746	Dan R. Olsen	Shawn A. Elson
0.746	Dan R. Olsen	David R. Morse
0.741	Dan R. Olsen	Daniel C. Edelson
0.741	Dan R. Olsen	Daniel Rosenberg

Search Algorithm: Blocking Algorithm - Sample Clustering By Name

Search Potential Duplicates: Both Within and Across Data Source File

Number of Potential Duplicate Pairs (1 ~ 300): 300

Search Potential Duplicate Pairs

Name Ascending Number of Edge E Show All Edges

Potential Duplicates Viewer

person_id	full_name	last_name	first_name	middle_name	suffix	affiliation	role	bio	country	institution	state
P58182	Dan R. Olsen	Olsen	Dan	R.	Jr.	Bingham Young University, Provo, UT	Author		USA	University	UT
P58184	Dan Olsen	Olsen	Dan			Bingham Young University/BYU, Provo, UT	Author		USA	University	UT
	Jaro (Weight:1.000)										

Merge Duplicates Mark Distinct

Search Nodes (7 nodes found)

Search

person_id	full_name	last_name	first_name	m
P345000	Judith S. Olsen	Olsen	Judith	S.
P58182	Dan R. Olsen	Olsen	Dan	R.
P55443	D. R. Olsen	Olsen	D.	R.
P58184	Dan Olsen	Olsen	Dan	

Search Potential Duplicates of Selected Node

Node Detail Viewer (37 items)

person_id	full_name	last_name	first_name	middle_name	suffix
P62971	David C. Mitchell	Mitchell	David	C.	
P63147	David Novick	Novick	David		
P438788	Jerry Falls	Falls	Jerry		
P137296	Jeff Jensen	Jensen	Jeff		
P159565	Ken Rodham	Rodham	Ken		
P200349	Mike Bastian	Bastian	Mike		
P59997	Darren Davis	Davis	Darren		
P63497	Doug Kohlert	Kohlert	Doug		

Edge Detail Viewer (15 items)

article_id	title
303038	Implementing interface attachments based on surface representations
506553	Design Expo 2
275649	Whiter (or whither) UIMS?
632821	An international SIGCHI research agenda
274715	Generalized pointing
260535	User interface tools
365030	Laser pointer interaction
142808	Workspaces

Finding possible duplicates completed!

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'être* 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.

NEXT UP

AFTER THE BREAK

TUTORIAL 2 – CLEANING DATA

THIS AFTERNOON

**STATISTICAL ANALYSIS
(PIERRE DRAGICEVIC)**

BEFORE NEXT WEEK'S CLASS

INSTALL:



Tableau

<http://www.tableausoftware.com/>

(Instructions in email)