

# Data Cleaning

---

**DATA 201: Thinking With Data**  
**Winter 2021**

Jonathan Hudson, Ph.D  
Instructor  
Department of Computer Science  
University of Calgary

**Monday, January 18, 2021**



# What is Data Cleaning?

# Why Data Cleaning?

**“I spend more than half of my time integrating, cleansing and transforming data without doing any actual analysis.  
Most of the time I’m lucky if I get to do any analysis.  
Most of the time once you transform the data you just do an average... the insights can be scarily obvious.  
It’s fun when you get to do something somewhat analytical.”**

**–Kandel et al. 2012**

**Where does “dirty”  
data come from?**

# Sources of Error

---

- What are some of your ideas on where errors in stored data comes from?

# Sources of Error

---

- Data entry
- Measurement
- Distillation
- Data integration

# Data Entry Errors



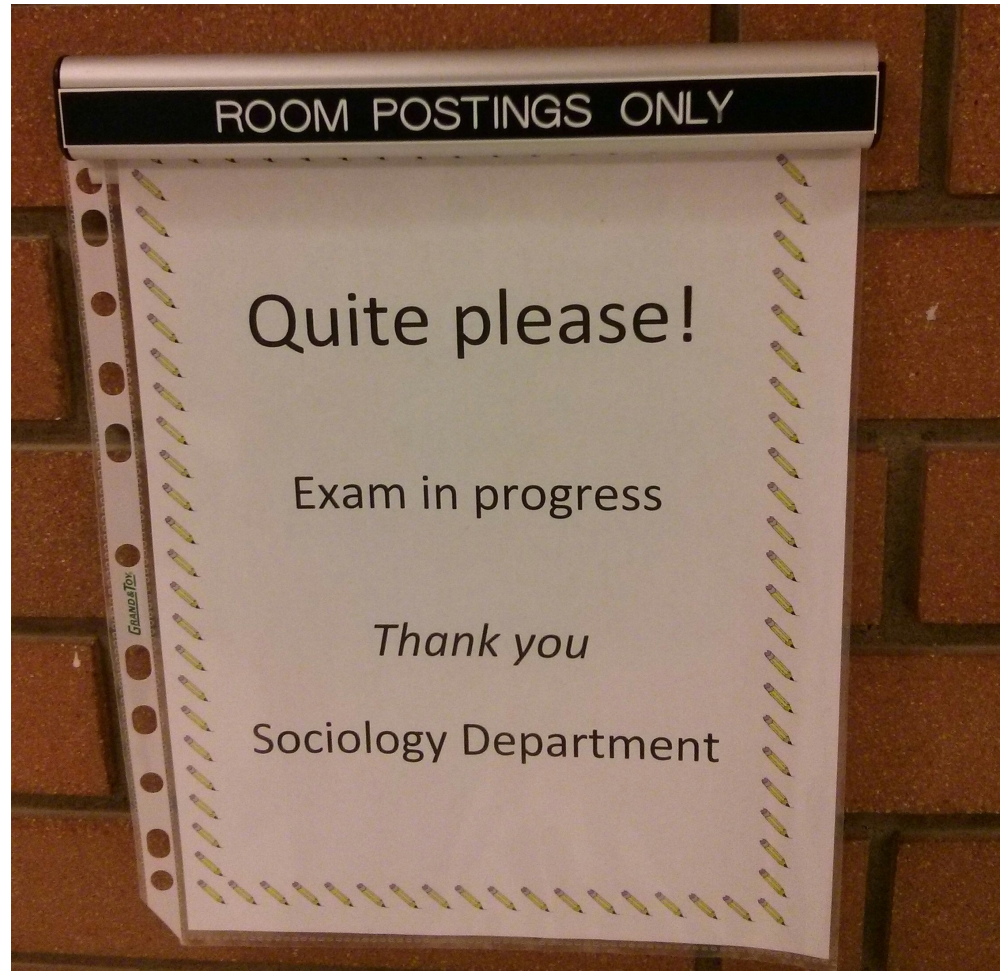
# Data Entry Errors

---



# Data Entry Errors

---



Science B



# Data Entry Errors

---



# Data Entry Errors

---

## Hidden Equations to Compute BMI

This computes your body mass index (BMI) and tells you your weight status.

It demonstrates the use of hidden Equation-type questions, whose values are stored in the database even though hidden on the screen.

\* How much do you weigh?

# Data Entry Errors

Step 1: Activity/Equipment Type

Step 2: Add a Map

Step 3: Additional Details

## Date of Activity:

September 2014

Su	Mo	Tu	We	Th	Fr	Sa
	1	2	3	4	5	6
7	8	9	10	11	12	13
14	15	16	17	18	19	20
21	22	23	24	25	26	27
28	29	30				

## Average Heart Rate (optional):

 bpm

## Duration:

 :  : 

## Start Time:

 :  

## Distance:

 mi

## Calories:

## Training Plan:

## Add An Activity

### Activity Details



Activity Type: Running

Equipment Type: None

Route: None

Distance: 5.62 mi.

Duration: --:--

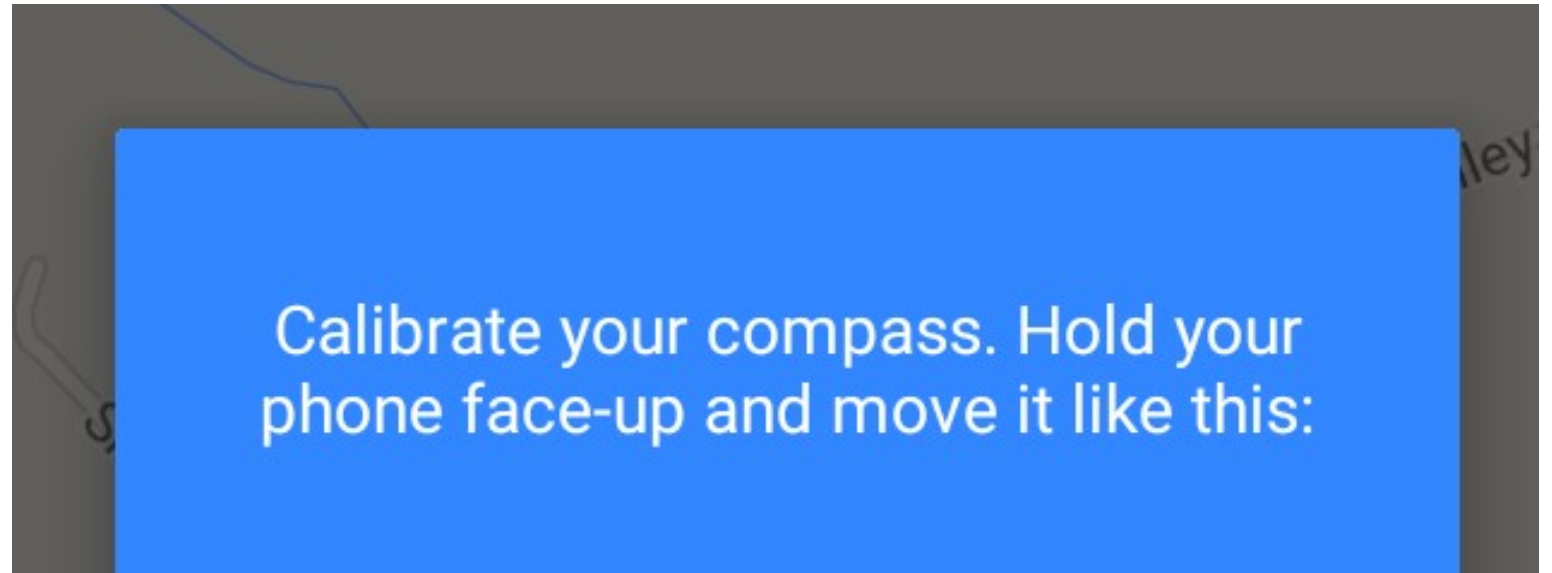
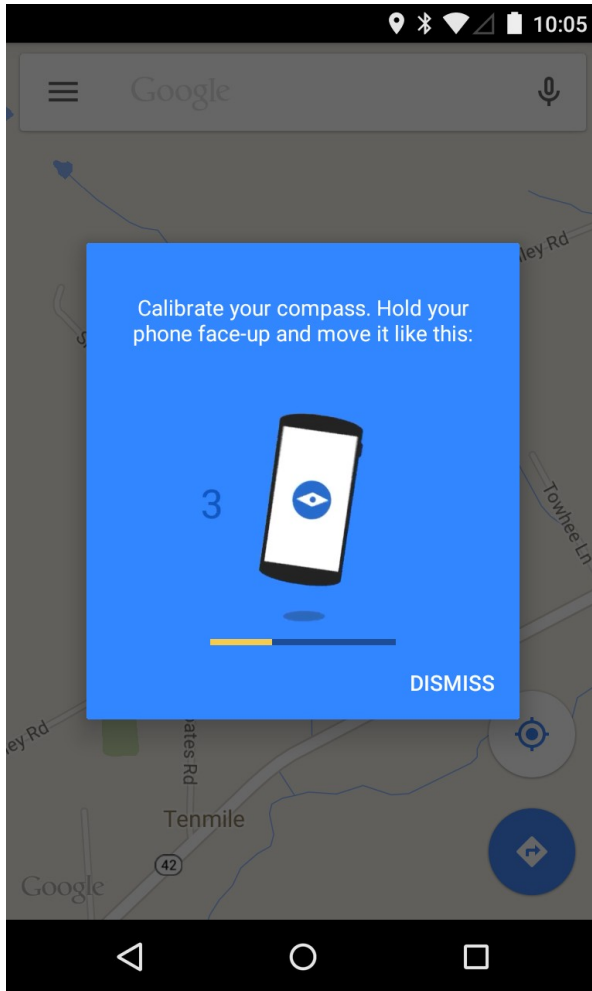
# Measurement Errors

# Measurement Errors

---



# Measurement Errors





# Measurement Errors

---



# Distillation Errors

# Distillation Errors

---

$0.345413 \Rightarrow 0.35$

Goods and Services Tax  $\Rightarrow$  GST

# Distillation Errors

---

2017, \$2, Apples

2017, \$2, Oranges

2017, \$2, Bananas

=> 2017, \$2, “Apples, Oranges, Bananas”

# Data Integration Errors

# Data Integration Errors

---

- Data often comes from multiple sources

# Data Integration Errors

---

- Data often comes from multiple sources
- Schemas change over time

# Data Integration Errors

---

- Data often comes from multiple sources
- Schemas change over time
- Same data may have different format from different sources



# Preventing Errors

# Integrity Constraints

# Integrity Constraints

---

Temperature 

---

# Integrity Constraints

---

Temperature \_\_\_\_\_ °C

# Integrity Constraints

---

\*Temperature \_\_\_\_\_ °C

\*required

# Integrity Constraints

---

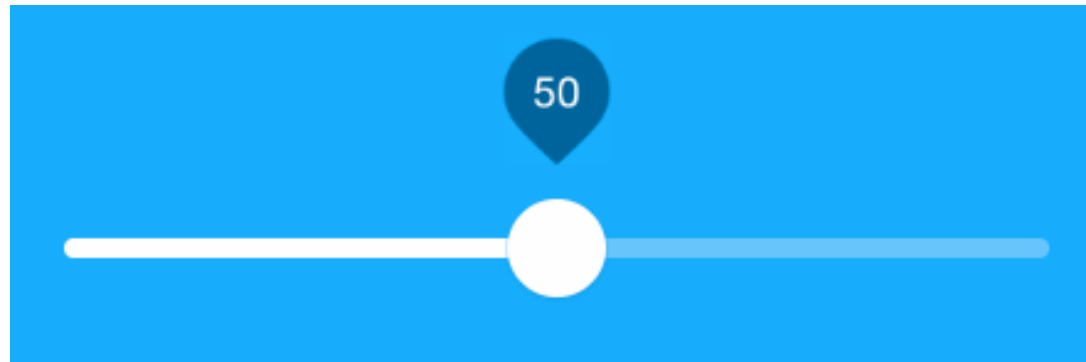
\*Temperature \_\_\_\_\_ °C (-50 °C to 50 °C)

\*required

# Integrity Constraints

---

Temperature



# Integrity Constraints

---

**Last name**

**Group/policy number**

**ID number**

**Birth date**

**Cancel**

**Next**



# Integrity Constraints

---

Last name  ?

Group/policy number  ?

ID number  ?

Birth date

Cancel

Next

# Integrity Constraints

**Last name**  ?

**Group/policy number**  ?

**ID number**  ?

**Birth date** Year  Month  Day



GROUP 123	SECTION 123
ID NUMBER	NAME
123456-76	John Doe
123456-77	Jane Doe
123456-78	Sam Doe
123456-79	Mary Doe

Cancel

Next

# Prediction and Friction

---

- Use data quality measures to predict how likely a value is to be correct
- Adjust the interface to add friction when entering unlikely responses

# Integrity Constraints

---

- Integrity constraints do not prevent bad data
- Enforcing constraints may lead to frustration

# Minimize Sensor Errors

# Minimize Sensor Errors

---

- Check sensors before deployment

# Minimize Sensor Errors

---

- Check sensors before deployment
- Periodically revalidate equipment

# Minimize Sensor Errors

---

- Check sensors before deployment
- Periodically revalidate equipment
- Use redundant sensors



# Minimize Sensor Errors

---

- Check sensors before deployment
- Periodically revalidate equipment
- Use redundant sensors
- Check data against historical logs or computed models

# Minimize Sensor Errors

---

- Check sensors before deployment
- Periodically revalidate equipment
- Use redundant sensors
- Check data against historical logs or computed models
- Use common sense

# Double Data Entry

# Double Data Entry

---

- Perform all data entry twice
- Identify mismatches then discard or repair

# Detecting Errors

# Data Auditing & Error Detection

# Data Auditing & Error Detection

---

- Look for outliers / anomalies

# Data Auditing & Error Detection

---

- Look for outliers / anomalies
- Examine data types



# Data Auditing & Error Detection

---

- Look for outliers / anomalies
- Examine data types
- Schema checking

# Data Auditing & Error Detection

---

- Look for outliers / anomalies
- Examine data types
- Schema checking
- Validate with other data

# Data Auditing & Error Detection

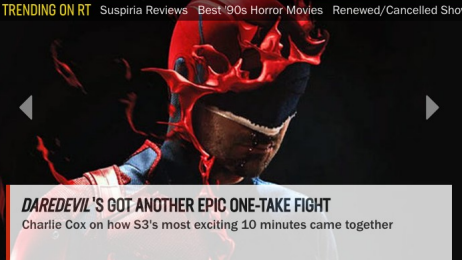
---

- Look for outliers / anomalies
- Examine data types
- Schema checking
- Validate with other data
- Common Sense

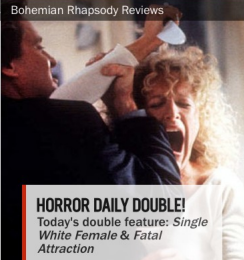
**Rotten Tomatoes** What's the Tomatometer? Critics SIGN UP | LOG IN

Search movies, TV, actors, more... MOVIES & DVDS TV NEWS TICKETS & SHOWTIMES


TRENDING ON RT [Suspiria Reviews](#) [Best '90s Horror Movies](#) [Renewed/Cancelled Shows](#) [Bohemian Rhapsody Reviews](#)



**DAREDEVIL'S GOT ANOTHER EPIC ONE-TAKE FIGHT**  
Charlie Cox on how S3's most exciting 10 minutes came together



**HORROR DAILY DOUBLE!**  
Today's double feature: *Single White Female* & *Fatal Attraction*



**HALLOWEEN RANKED**  
Every film in the franchise by Tomatometer

**MOVIES OPENING THIS WEEK** [Get Tickets](#)

39%	Johnny English Strikes Again	OCT 26
38%	Hunter Killer	OCT 26
79%	Mid90s	OCT 26
No Score	Indivisible	OCT 26
Yet		
72%	Suspiria	OCT 26

[View All](#)

**TOP BOX OFFICE** [Get Tickets](#)

79%	Halloween	\$76.3M
90%	A Star Is Born	\$19.1M
30%	Venom	\$18.1M
44%	Goosebumps 2: Haunted Hall...	\$9.8M
89%	First Man	\$8.3M
96%	The Hate U Give	\$7.7M
75%	Smallfoot	\$6.7M
28%	Night School	\$5M
72%	Bad Times at the El Royale	\$3.5M
90%	The Old Man & the Gun	\$2.1M

[View All](#)

**COMING SOON TO THEATERS**

55%	Bohemian Rhapsody	NOV 2
No Score	The Nutcracker and the Four ...	NOV 2
Yet		
No Score	Nobody's Fool	NOV 2
Yet		
86%	Boy Erased	NOV 2
90%	Bodied	NOV 2

**NEW TV TONIGHT**

100%	Black Lightning
96%	The Conners
92%	This Is Us
81%	The Kids Are Alright
70%	Mayans M.C.
70%	The Rookie
62%	FBI
45%	The Purge
35%	New Amsterdam
30%	The Hunt for the Trump Tapes With To...
No Score	NCIS

[View All](#)

**MOST POPULAR TV ON RT**

92%	The Haunting of Hill House
84%	Titans
92%	Marvel's Daredevil
81%	Maniac
97%	American Vandal
100%	BoJack Horseman

[View All](#)


**TOP DVD & STREAMING MOVIES**

[FandangoNOW](#) | [Netflix](#) | [iTunes](#) | [Amazon](#) | [More...](#)

94%	Incredibles 2
95%	BlackKkKlansman
17%	Patient Zero
100%	The Dark

**IMDb** Find Movies, TV shows, Celebrities and more... All [IMDbPRO](#) | [Help](#) | [f](#) | [t](#) | [i](#)


Movies, TV & Showtimes Celebs, Events & Photos News & Community [Watchlist](#) [Sign in with Facebook](#) [Other Sign in options](#)




**IMDbbrief**

Guillermo del Toro to Direct 'Pinocchio' Details on His Netflix Mo...

[Browse trailers >](#)



Ralph Breaks the Internet Imagine Dragons' Music ...



"Homecoming" New Trailer

**Opening This Week**

- [+ Suspiria](#) Limited
- [+ Hunter Killer](#)
- [+ Johnny English Strikes Again](#)
- [+ Indivisible](#)
- [+ Stuck](#) Limited
- [+ Burning](#) Limited
- [+ London Fields](#) Limited
- [+ The Fog](#) Re-release
- [+ Monrovia, Indiana](#) Limited
- [+ Senso](#) Re-release

[See more opening this week >](#)

**Get Showtimes >**

**Déjà View: The Best Reboots and Remakes**

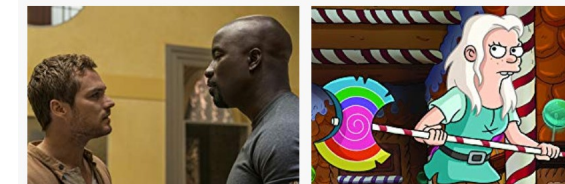


Yep, that includes *Judge Dredd*. Let's take a look at even *more* movies and TV shows that were so nice they made 'em twice (at least).

[See the entire gallery >](#)

**What TV Shows Have Been Renewed and Canceled?**

Bad news: "Luke Cage" and "Iron Fist" didn't make the cut ... but "Disenchantment" conjured up a multi-year renewal. Find out the fates of your favorite shows now.



**Now Playing (Box Office)**

+	Halloween	\$76.2M	<a href="#">Get Tickets</a>
+	A Star Is Born	\$19.1M	<a href="#">Get Tickets</a>
+	Venom	\$18.0M	<a href="#">Get Tickets</a>
+	Goosebumps 2: Haunted Halloween	\$9.7M	<a href="#">Get Tickets</a>
+	First Man	\$8.3M	<a href="#">Get Tickets</a>

[See more box office results >](#)

**Coming Soon**

+	The Nutcracker and the Four Realms	<a href="#">Get Tickets</a>
---	------------------------------------	-----------------------------

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

Arn - Tempelriddaren	Dec 25, 2007	R		
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	17
Around the World in 80 Days	Oct 17, 1956	PG	United Artists	73
Barbarella	Oct 10, 1968		Paramount Pictures	74
Barry Lyndon	1975		Warner Bros.	94
Barbarians, The	March, 1987			
Babe	Aug 4, 1995	G	Universal	98
Boynton Beach Club	Mar 24, 2006	R	Wingate Distribution	
Baby's Day Out	Jul 1, 1994	PG	20th Century Fox	21

Title	Release Date	IMDB Rating	IMDB Votes	MPAA Rating
Bad Boys	Apr 7, 1995	6.6	53929	R
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	R
Beauty and the Beast	Nov 13, 1991	3.4	354	G



# Common Data Quality Issues

# Common Data Quality Issues

---

- Missing Data —
- Erroneous Values —
- Entity Resolution —
- Type Conversion —
- Data Integration —

# Common Data Quality Issues

---

- Missing Data — Missed measurements, redacted items, incomplete forms, etc.
- Erroneous Values —
- Entity Resolution —
- Type Conversion —
- Data Integration —

# Common Data Quality Issues

---

- Missing Data — Missed measurements, redacted items, incomplete forms, etc.
- Erroneous Values — Misspellings, outliers, “spurious integrity”, etc.
- Entity Resolution —
- Type Conversion —
- Data Integration —

# Common 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 —
- Data Integration —

# Common 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., postal code or place name to lat-lon
- Data Integration —

# Common 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., postal code or place name to lat-lon
- Data Integration — Mismatches and inconsistencies when combining data

# Detect Duplicates



# Duplicates

---

TITLE

Ben-Hur

Ben Hur

BEN-HUR

Ben-Hur (1959)

NAME

Anand Vaskar

Anand Vaskkar

A. Vaskar

Vaskar, Anand

# Levenshtein Distance

---

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

# Levenshtein Distance

---

Ben-Hur  
Ben Hur

Distance = 1

# Levenshtein Distance

---

Anand Vaskar  
Anand Vaskkar

Distance = ?

# Levenshtein Distance

---

Ben-Hur  
Ben Hur (1959 film)

Distance = ?

# Soundex / Metaphone

---

How similar do they sound?

# Soundex / Metaphone

---

Ben-Hur  
Ben-Hurr  
Been Her

Anana Vaskar  
Anand Vaskkar  
Ahnund Vachkar

# Fingerprinting

---

Strip away unimportant details (e.g., punctuation and capitals)



# Fingerprinting

---

Ben-Hur => ben hur

# Fingerprinting

---

Anand Vaskar => anand vaskar

# How to Fix Problems?

# Questions to Think About

---

- Which duplicate to keep?
- Outliers: keep, remove, or repair?
- How to deal with bad formats (e.g., dates and addresses)?

# Some Ways to Fix Problems

---

- Fuzzy matching systems
- Machine learning to detect/resolve errors
- Usually requires human judgement (especially for new data)

# Cleaning

# Common Operations

---

- Correct and remove errors
- Change formats
- Remove formatting

# Spreadsheets

---

- + Easily accessible
- + Visual
- Tedious
- Time-consuming
- Repetitive





# Scripts

---

+ Reusable

+ Scalable

- Difficult

- Tedious

- Time-consuming

```
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", on=",", max=0))

# Delete empty rows
w.add(dw.Filter(row=dw.Row(conditions=dw.Empty()))))

# Extract from split after 'in '
w.add(dw.Extract(column="split", on=".*", after="in "))

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

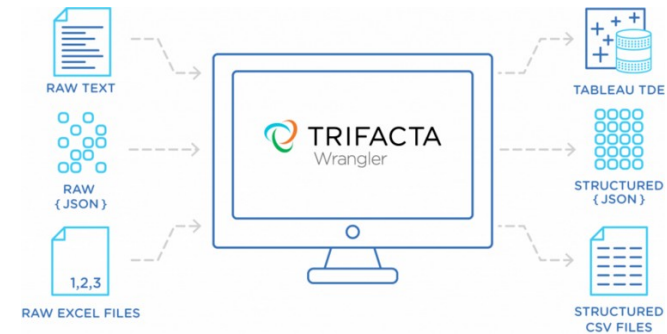
# Delete rows where split1 is null
```

# 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>



# Onward to ... Charts

---

Jonathan Hudson  
[jwhudson@ucalgary.ca](mailto:jwhudson@ucalgary.ca)  
<https://pages.cpsc.ucalgary.ca/~jwhudson/>



UNIVERSITY OF  
CALGARY