Data Cleaning

DATA 201: Thinking With Data
Winter 2022

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

Tuesday, March 22, 2022

UNIVERSITY OF CALGARY

What is Data Cleaning?

UNIVERSITY OF CALGARY

Why Data Cleaning?

UNIVERSITY OF CALGARY

“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

UNIVERSITY OF CALGARY

Where does “dirty” data come from?

UNIVERSITY OF CALGARY

Sources of Error

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

UNIVERSITY OF CALGARY

Sources of Error

- Data entry
- Measurement
- Distillation
- Data integration

UNIVERSITY OF CALGARY

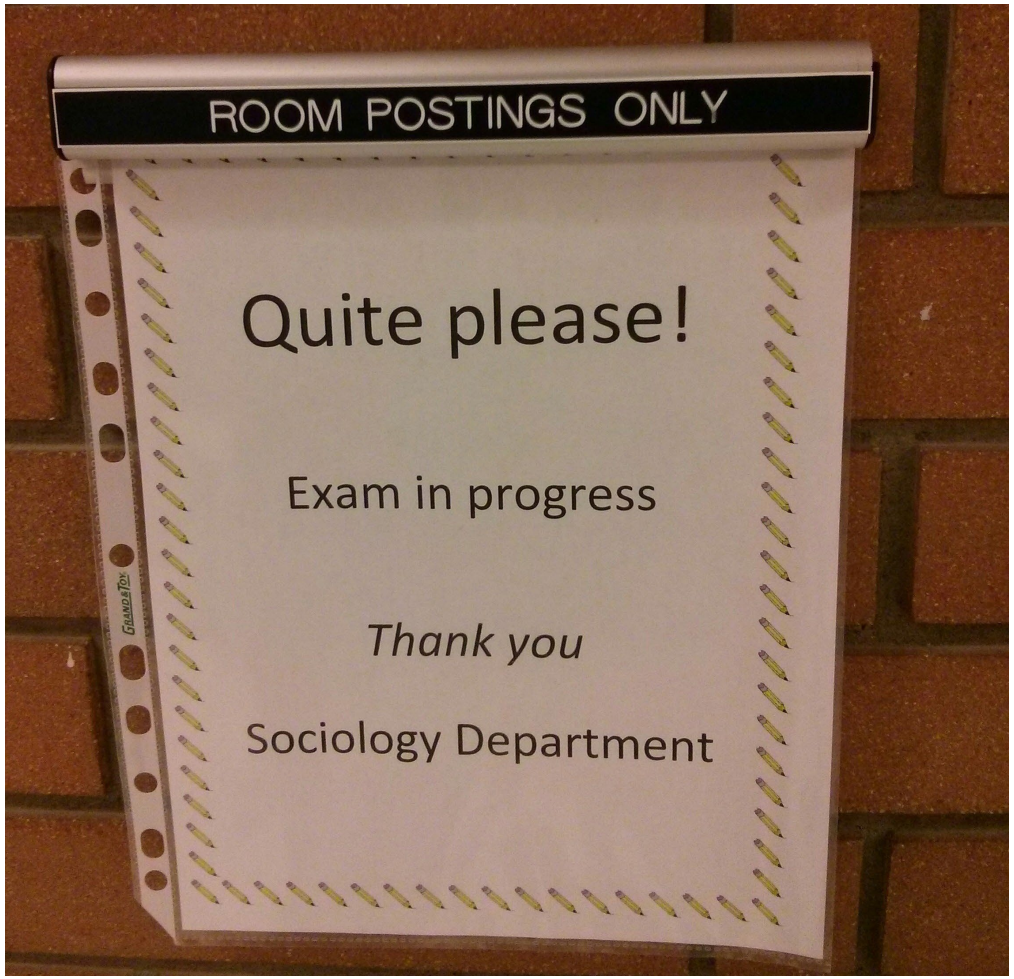
Data Entry Errors

UNIVERSITY OF CALGARY

Data Entry Errors

UNIVERSITY OF CALGARY

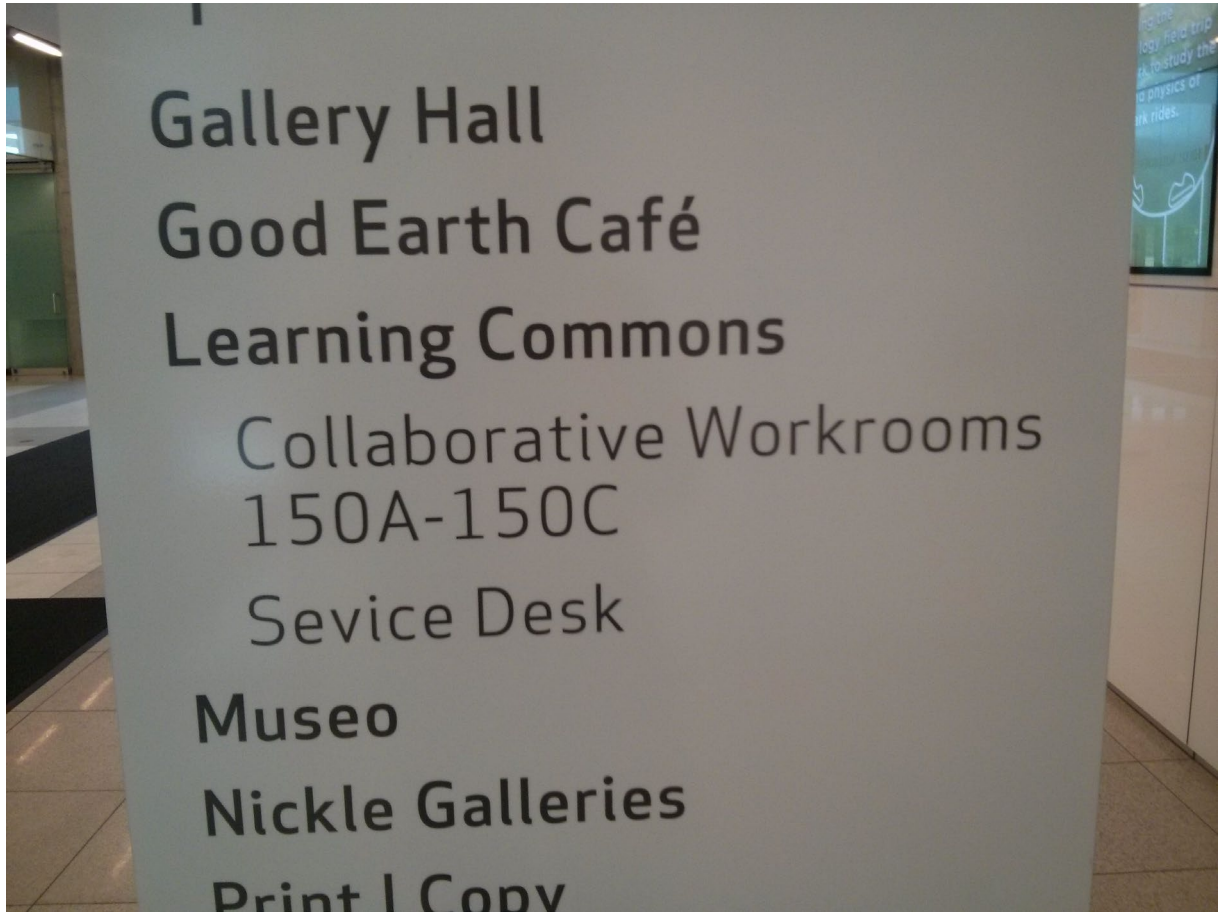
Data Entry Errors



Science B

UNIVERSITY OF CALGARY

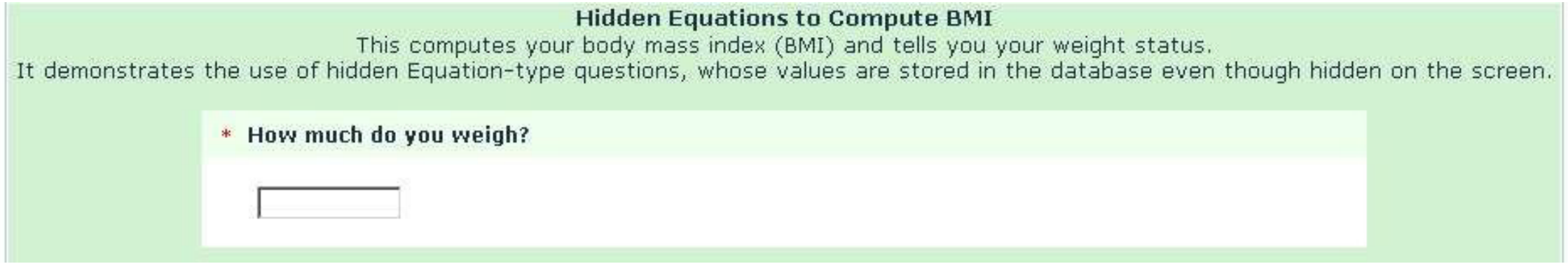
Data Entry Errors



Taylor Family
Digital Library

UNIVERSITY OF CALGARY

Data Entry Errors



UNIVERSITY OF CALGARY

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:

00 : 00 : 00

Start Time:

11 : 43 AM

Distance:

5.62 mi

Calories:

Training Plan:

None

Add An Activity

Activity Details

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

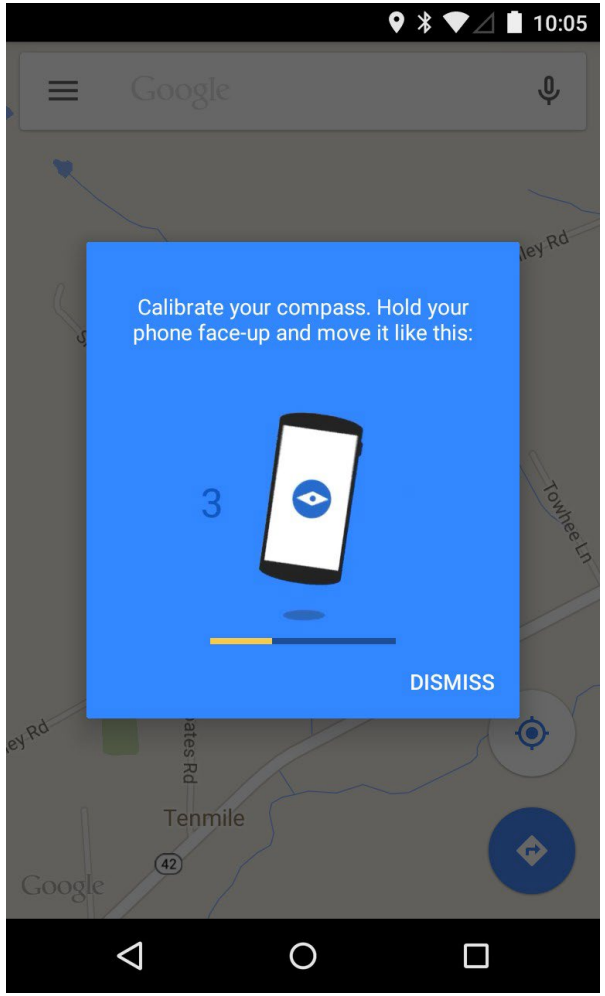
UNIVERSITY OF CALGARY

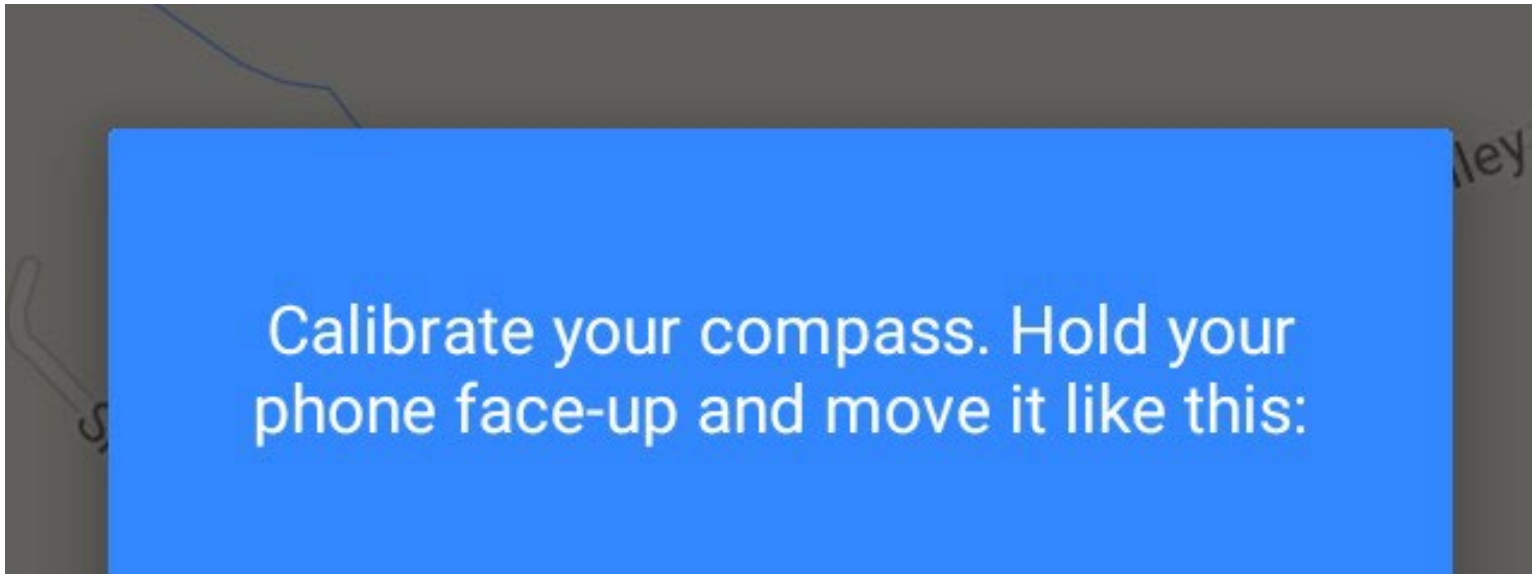
Measurement Errors

UNIVERSITY OF CALGARY

Measurement Errors

Measurement Errors





UNIVERSITY OF CALGARY

Measurement Errors

UNIVERSITY OF CALGARY

Distillation Errors

UNIVERSITY OF CALGARY

Distillation Errors

0.345413 => 0.35

Goods and Services Tax => GST

UNIVERSITY OF CALGARY

Distillation Errors

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

UNIVERSITY OF CALGARY

Data Integration Errors

UNIVERSITY OF CALGARY

Data Integration Errors

- Data often comes from multiple sources

UNIVERSITY OF CALGARY

Data Integration Errors

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

UNIVERSITY OF CALGARY

Data Integration Errors

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

UNIVERSITY OF CALGARY

Preventing Errors

UNIVERSITY OF CALGARY

Integrity Constraints

UNIVERSITY OF CALGARY

Integrity Constraints

Temperature _____

UNIVERSITY OF CALGARY

Integrity Constraints

Temperature _____ °C

UNIVERSITY OF CALGARY

Integrity Constraints

*Temperature _____ °C

*required

UNIVERSITY OF CALGARY

Integrity Constraints

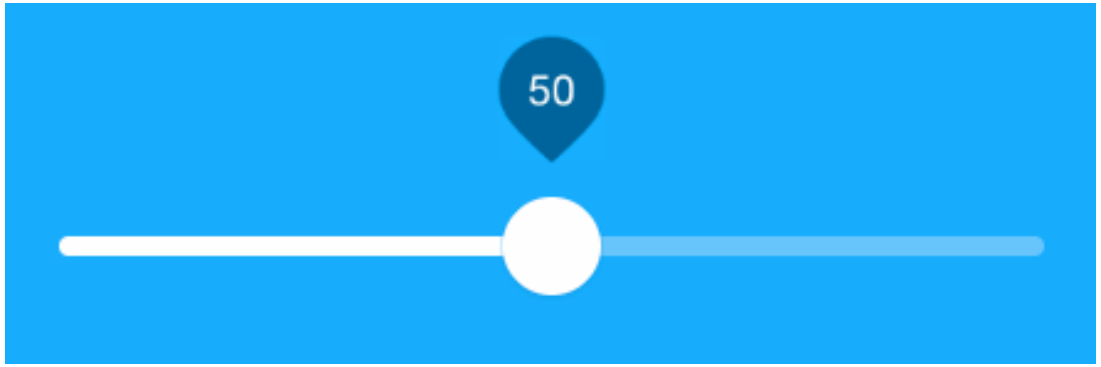
*Temperature _____ °C (-50 °C to 50 °C)

*required

UNIVERSITY OF CALGARY

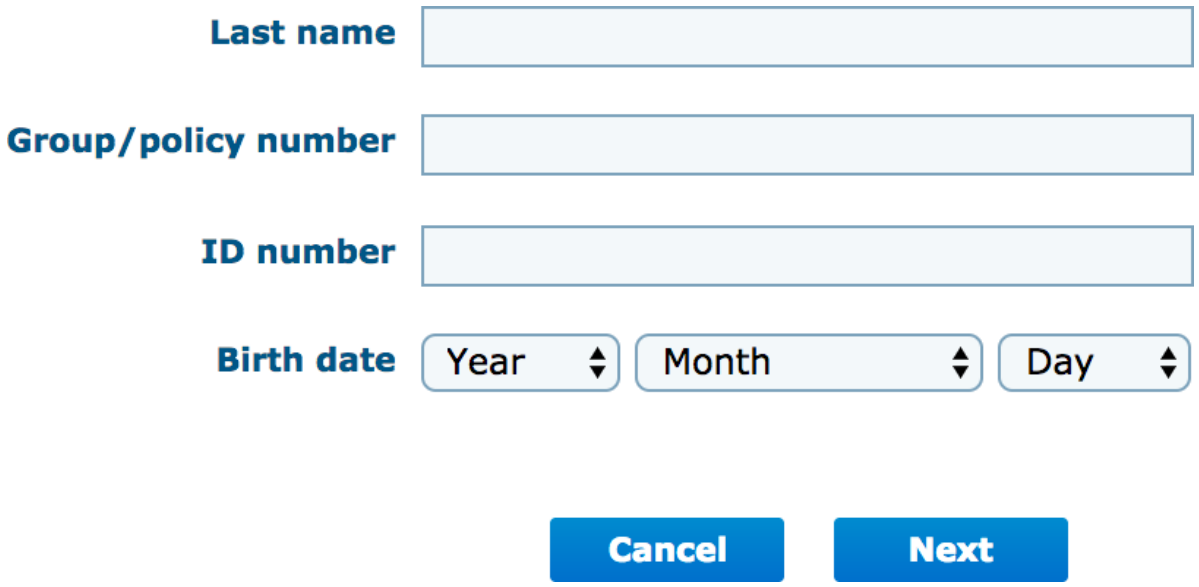
Integrity Constraints

Temperature



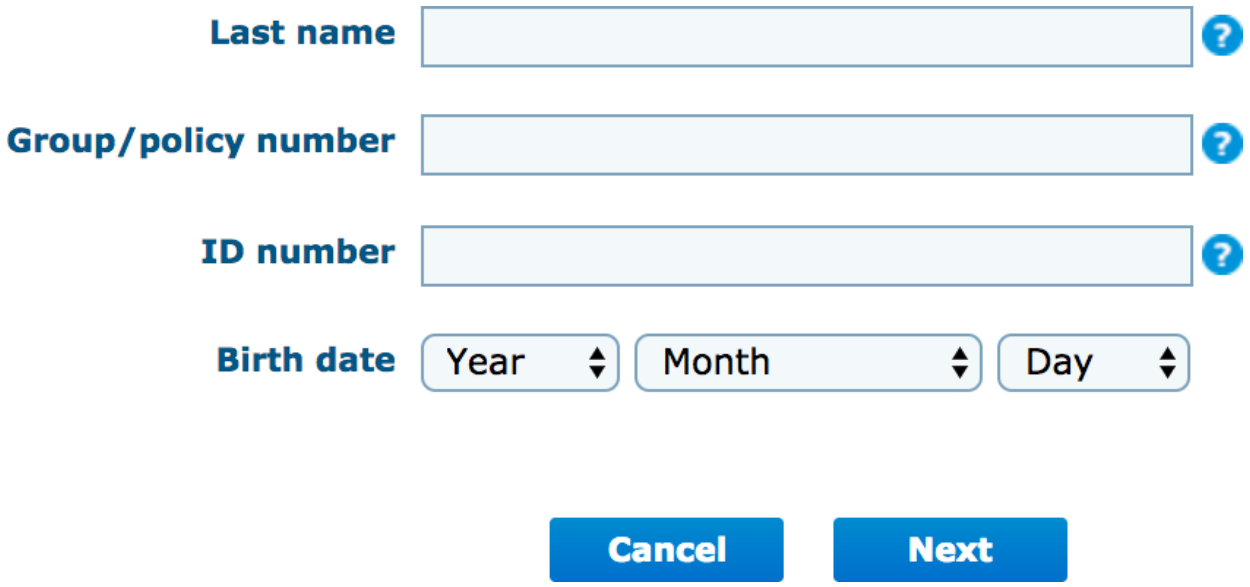
UNIVERSITY OF CALGARY

Integrity Constraints



UNIVERSITY OF CALGARY

Integrity Constraints



UNIVERSITY OF CALGARY

Integrity Constraints

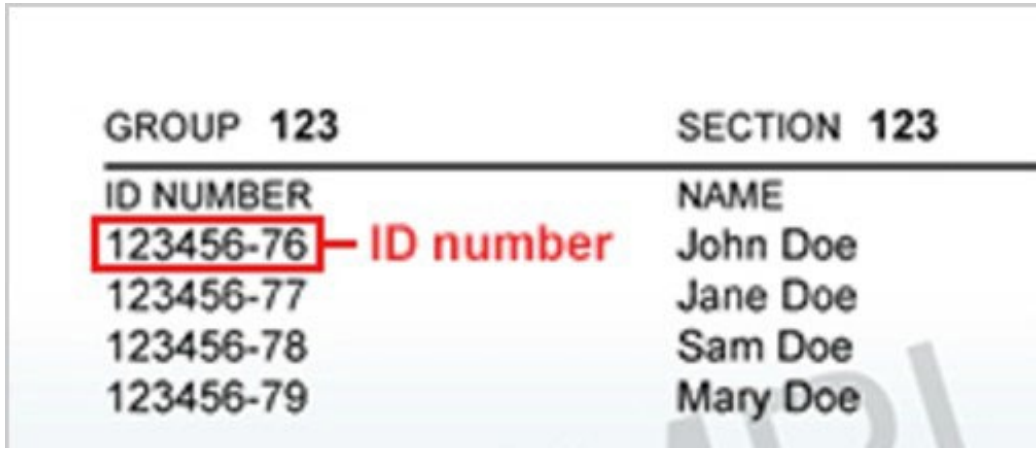
Last name

Group/policy number

ID number

Birth date Year Month Day

Cancel Next



UNIVERSITY OF CALGARY

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

UNIVERSITY OF CALGARY

Integrity Constraints

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

UNIVERSITY OF CALGARY

Minimize Sensor Errors

UNIVERSITY OF CALGARY

Minimize Sensor Errors

- Check sensors before deployment

UNIVERSITY OF CALGARY

Minimize Sensor Errors

- Check sensors before deployment
- Periodically revalidate equipment

UNIVERSITY OF CALGARY

Minimize Sensor Errors

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

UNIVERSITY OF CALGARY

Minimize Sensor Errors

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

UNIVERSITY OF CALGARY

Minimize Sensor Errors

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

UNIVERSITY OF CALGARY

Double Data Entry

UNIVERSITY OF CALGARY

Double Data Entry

- Perform all data entry twice

- Identify mismatches then discard or repair

UNIVERSITY OF CALGARY

Detecting Errors

UNIVERSITY OF CALGARY

Data Auditing & Error Detection

UNIVERSITY OF CALGARY

Data Auditing & Error Detection

- Look for outliers / anomalies

UNIVERSITY OF CALGARY

Data Auditing & Error Detection

- Look for outliers / anomalies
- Examine data types

UNIVERSITY OF CALGARY

Data Auditing & Error Detection

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

UNIVERSITY OF CALGARY

Data Auditing & Error Detection

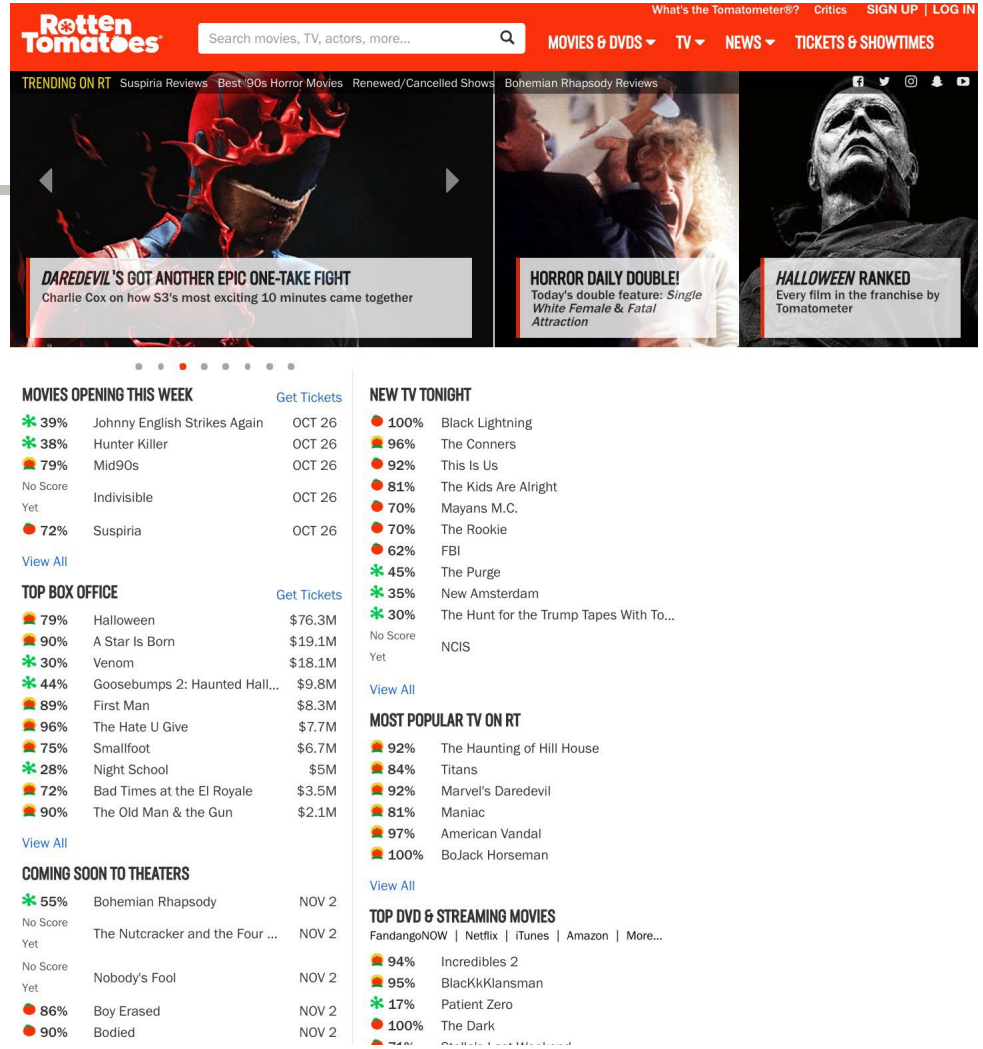
- Look for outliers / anomalies
- Examine data types
- Schema checking
- Validate with other data

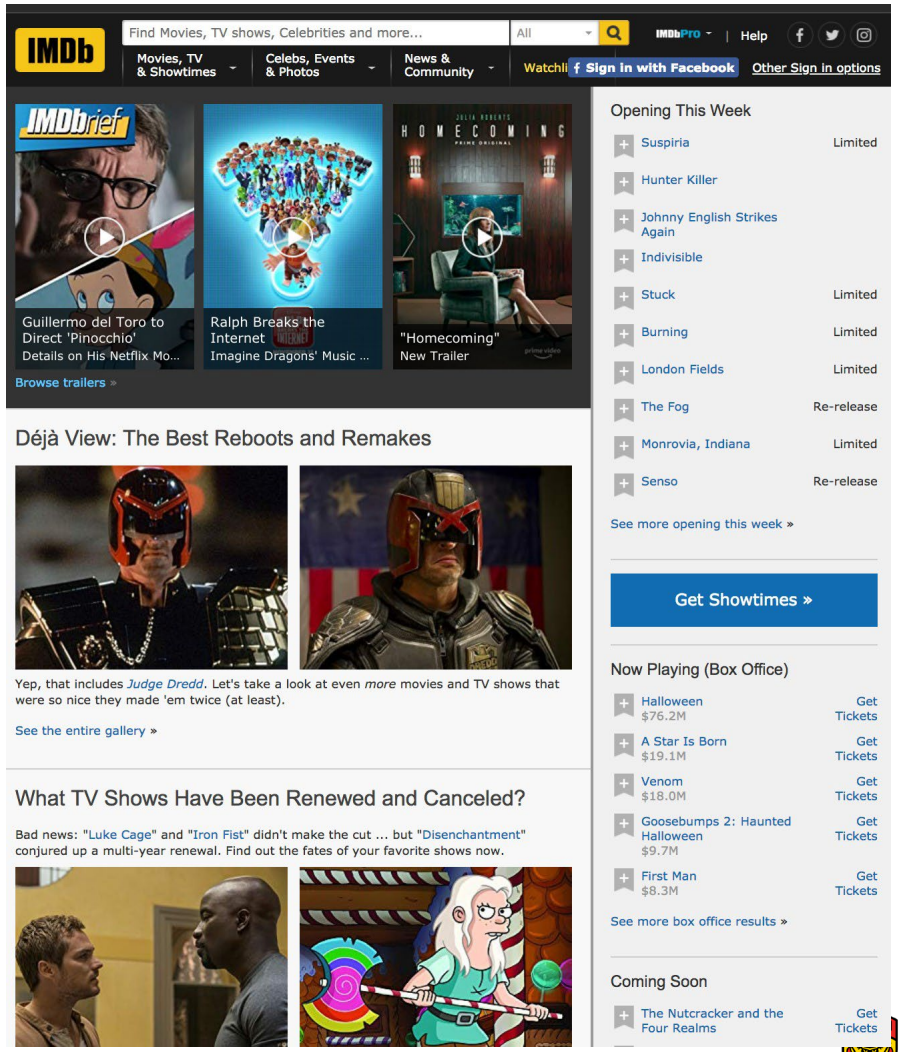
UNIVERSITY OF CALGARY

Data Auditing & Error Detection

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

UNIVERSITY OF CALGARY





UNIVERSITY OF CALGARY

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

UNIVERSITY OF CALGARY

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

UNIVERSITY OF CALGARY

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

UNIVERSITY OF CALGARY

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

UNIVERSITY OF CALGARY

Common Data Quality Issues

UNIVERSITY OF CALGARY

Common Data Quality Issues

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

UNIVERSITY OF CALGARY

Common Data Quality Issues

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

UNIVERSITY OF CALGARY

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 —

UNIVERSITY OF CALGARY

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 —

UNIVERSITY OF CALGARY

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 —

UNIVERSITY OF CALGARY

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

UNIVERSITY OF CALGARY

Detect Duplicates

UNIVERSITY OF CALGARY

Duplicates

TITLE	NAME
Ben-Hur	Anand Vaskar
Ben Hur	Anand Vaskkar
BEN-HUR	A. Vaskar
Ben-Hur (1959)	Vaskar, Anand

UNIVERSITY OF CALGARY

Levenshtein Distance

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

UNIVERSITY OF CALGARY

Levenshtein Distance

Ben-Hur
Ben Hur

Distance = 1

UNIVERSITY OF CALGARY

Levenshtein Distance

Anand Vaskar
Anand Vaskkar

Distance = ?

UNIVERSITY OF CALGARY

Levenshtein Distance

Ben-Hur
Ben Hur (1959 film)

Distance = ?

UNIVERSITY OF CALGARY

Soundex / Metaphone

How similar do they sound?

UNIVERSITY OF CALGARY

Soundex / Metaphone

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

UNIVERSITY OF CALGARY

Fingerprinting

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

UNIVERSITY OF CALGARY

Fingerprinting

Ben-Hur => ben hur

UNIVERSITY OF CALGARY

Fingerprinting

Anand Vaskar => anand vaskar

UNIVERSITY OF CALGARY

How to Fix Problems?

UNIVERSITY OF CALGARY

Questions to Think About

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

UNIVERSITY OF CALGARY

Some Ways to Fix Problems

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

UNIVERSITY OF CALGARY

Cleaning

UNIVERSITY OF CALGARY

Common Operations

- Correct and remove errors
- Change formats
- Remove formatting

UNIVERSITY OF CALGARY

Spreadsheets

+ Easily accessible

+ Visual

- Tedious

- Time-consuming

- Repetitive

UNIVERSITY OF CALGARY

Scripts

+ Reusable

+ Scalable

- Difficult

- Tedious

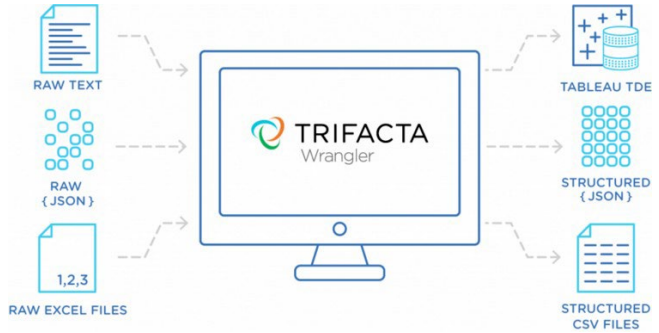
- Time-consuming

```
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
```


UNIVERSITY OF CALGARY

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
https://pages.cpsc.ucalgary.ca/~jwhudson/

UNIVERSITY OF CALGARY