# DATA QUALITY & WRANGLING

UNIVERSITY OF
CALGARY

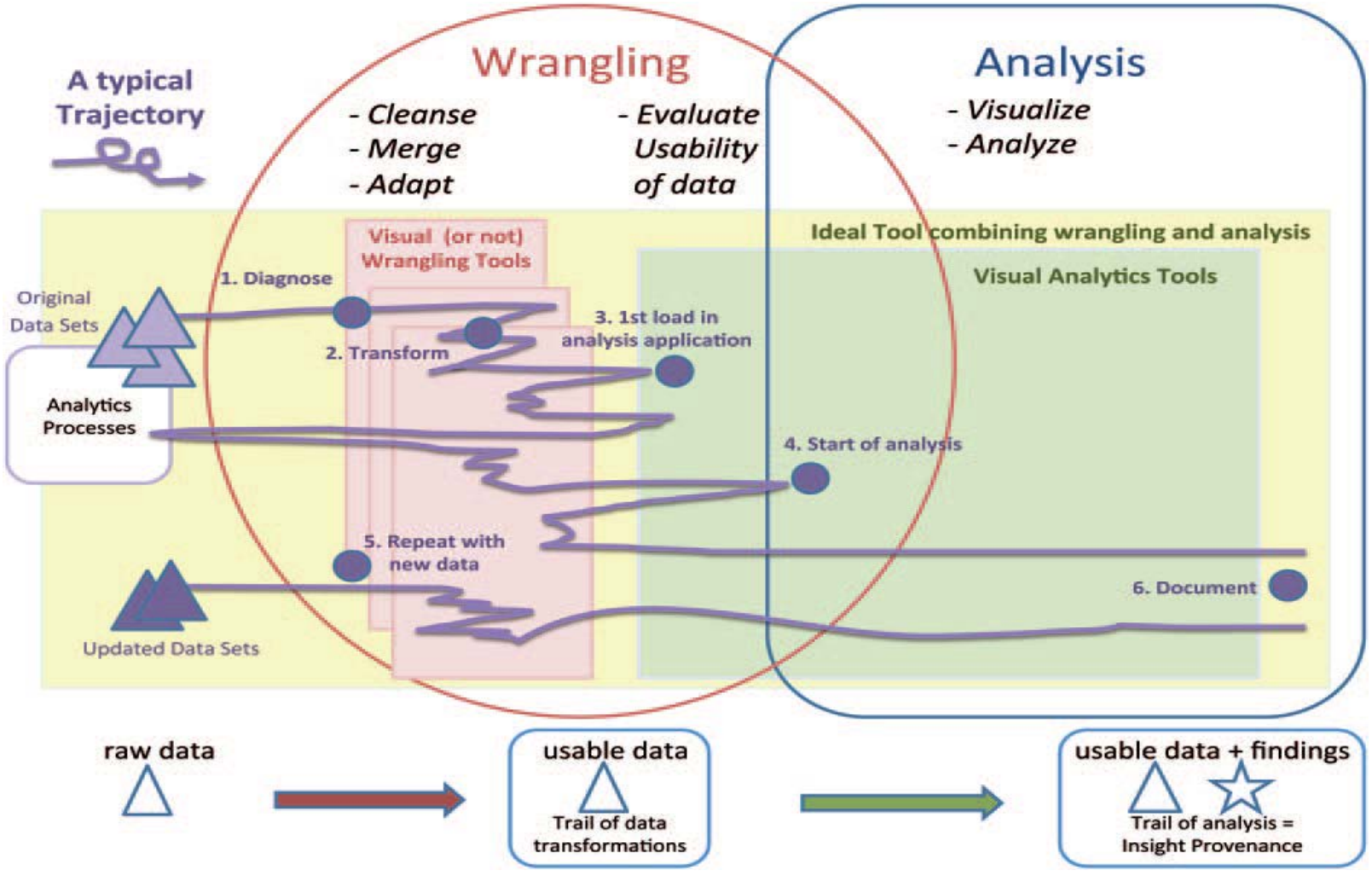
# WHY IS THIS IMPORTANT?

# 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 et al. 2012]

# ANALYSIS TRAJECTORIES



# COMMON OPERATIONS

Transforming data into usable formats

MORE ON THIS LATER!

Correcting and removing errors

Removing formatting

Connecting and resolving data

# WHAT IS "DIRTY DATA"?

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

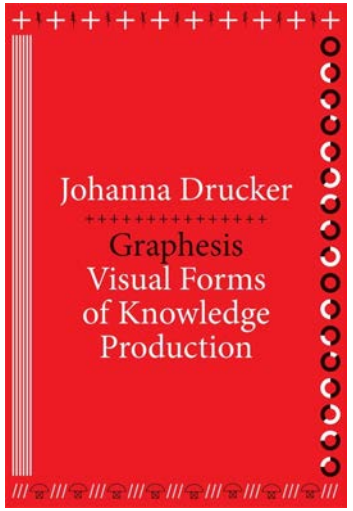
# “DATA” as “CAPTA”

**data (n.)**
1640s, "a fact given or granted," classical plural of *datum*, from Latin *datum* "(thing) given," neuter past participle of *dare* "to give" (from PIE root *do- "to give").

… from Latin *captus*, past participle of *capere* "to take, hold, seize" (from PIE root *kap- "to grasp").

JOHANNA DRUCKER

Also check out her excellent 2014 book ***Graphesis.***



# DATA ISN'T TRUTH

## ...IT'S AN INCOMPLETE, MESSY, AND POTENTIALLY INACCURATE REFLECTION OF THE WORLD

"GARBAGE IN, GARBAGE OUT"

# 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

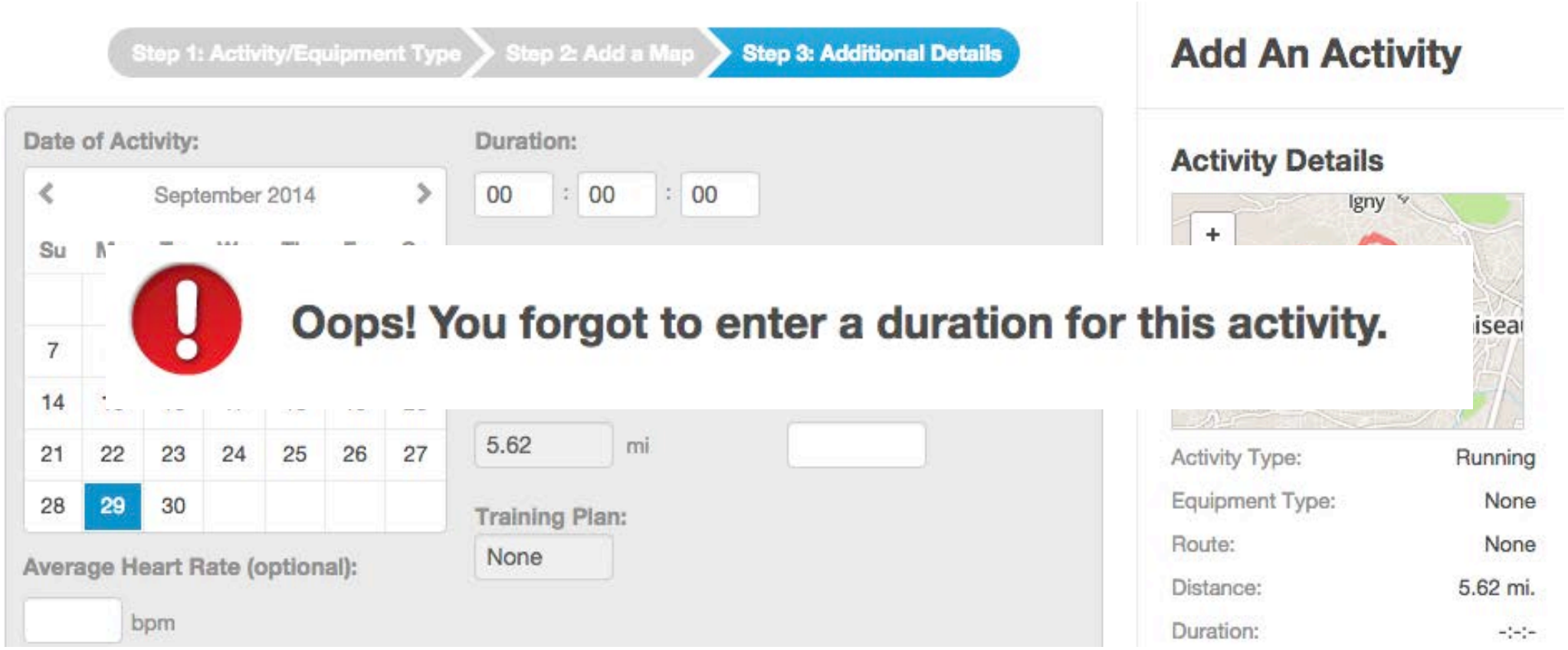
Misunderstanding data or conventions

“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, Cucumbers
➡️ 1985, $2, “Apples,Oranges,Cucumbers”

# DATA INTEGRATION ERRORS

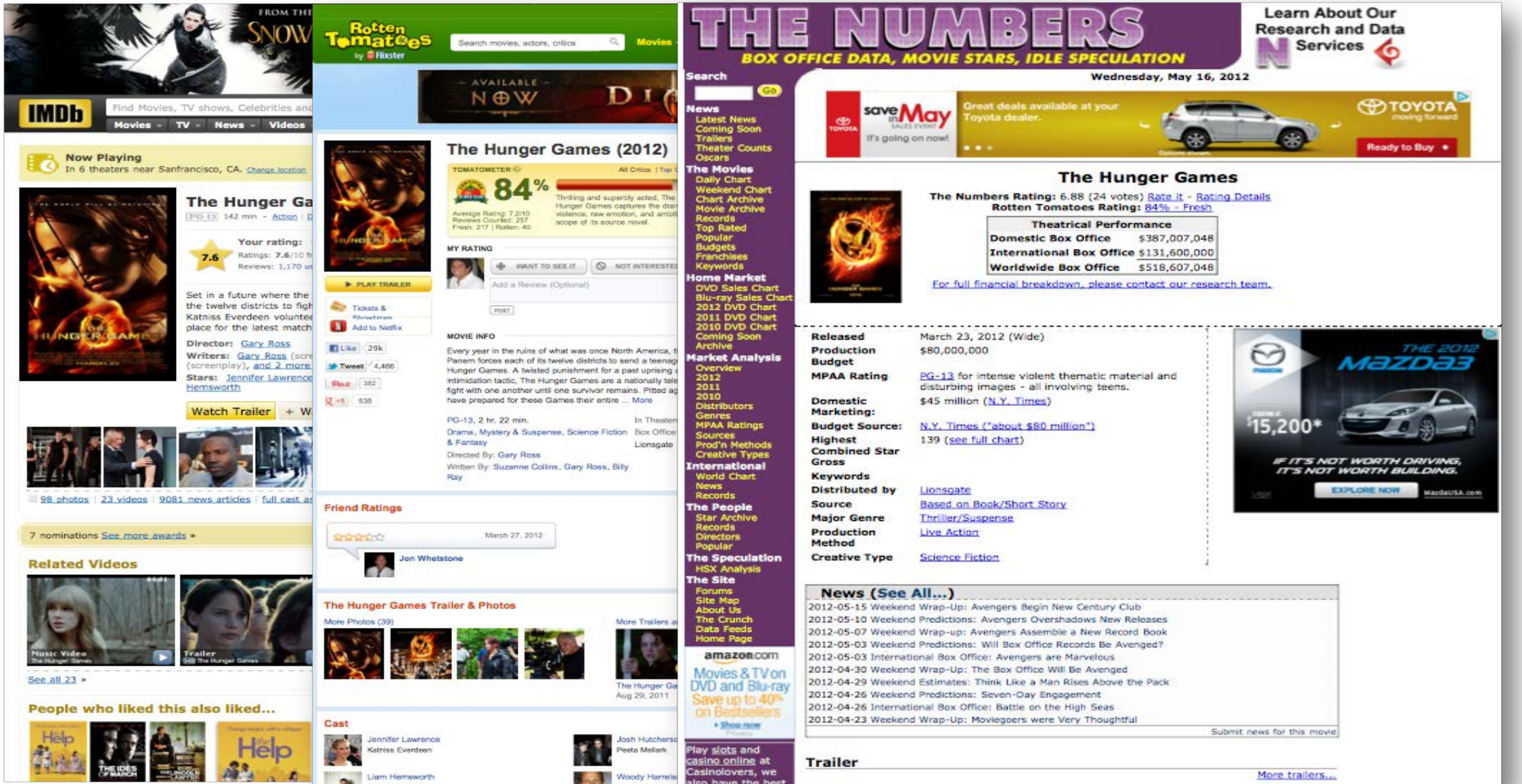
Data often comes from multiple sources

Schemas change over time

Data is often coerced from one type to another

Can lead to data loss, duplication, and other inconsistencies

# AN EXAMPLE


IMDb
Now Playing
In 6 theaters near Sanfrancisco, CA.
The Hunger Ga
Your rating:
7.6
Watch Trailer
Related Videos
People who liked this also liked...
Rotten Tomatoes
The Hunger Games (2012)
84%
MOVIE INFO
Friend Ratings
The Hunger Games Trailer & Photos
Cast
Jennifer Lawrence
Katniss Everdeen
THE NUMBERS
BOX OFFICE DATA, MOVIE STARS, IDLE SPECULATION
Learn About Our Research and Data Services
Wednesday, May 16, 2012
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
Trailer

| 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 | 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 |

# PREVENTING ERROR

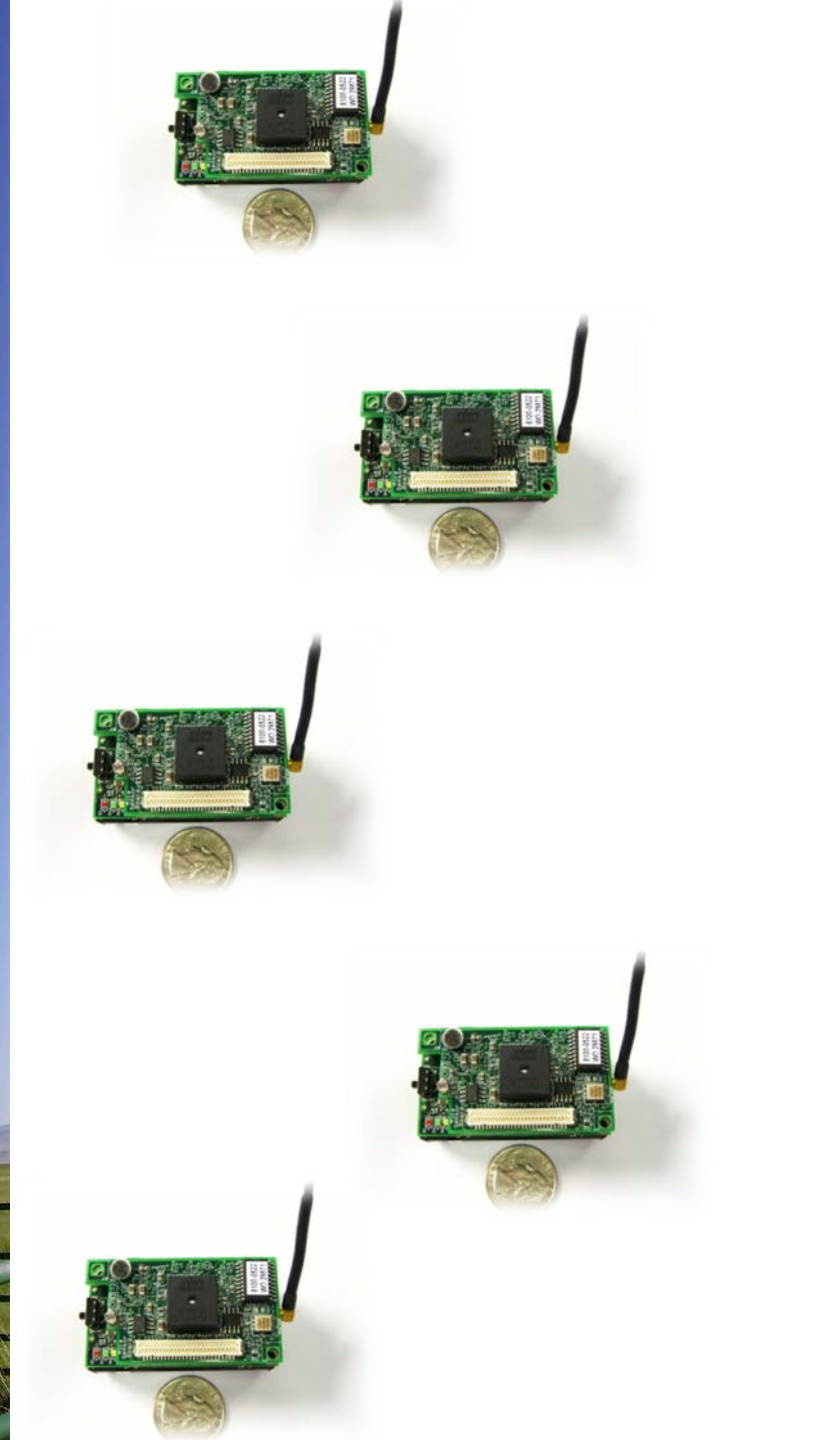
CATCHING DIRTY DATA AT THE SOURCE

# MINIMIZING SENSOR ERROR

**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 -50°C and 50°C.

TEMPERATURE -60 °C

# 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

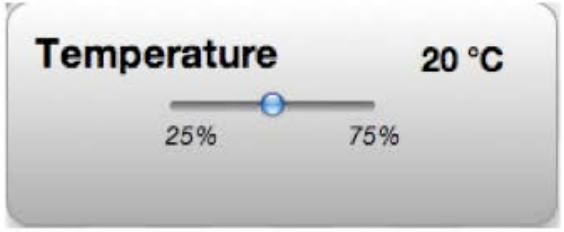
**PRINCIPLE 1** Data quality should be controlled via **feedback**, not **enforcement**.

**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

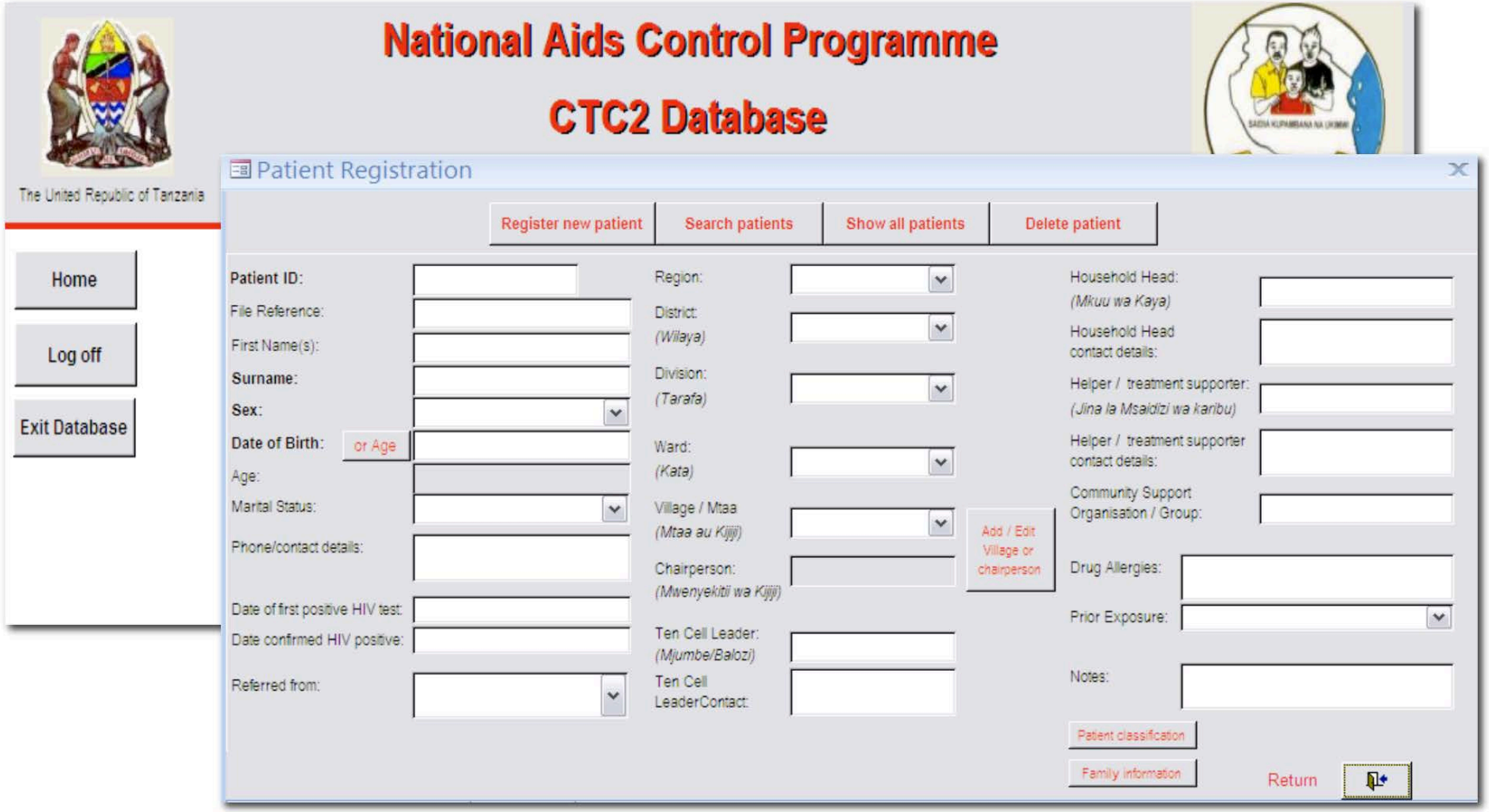
This value seems low.
Are you sure?

TEMPERATURE -60 ºC

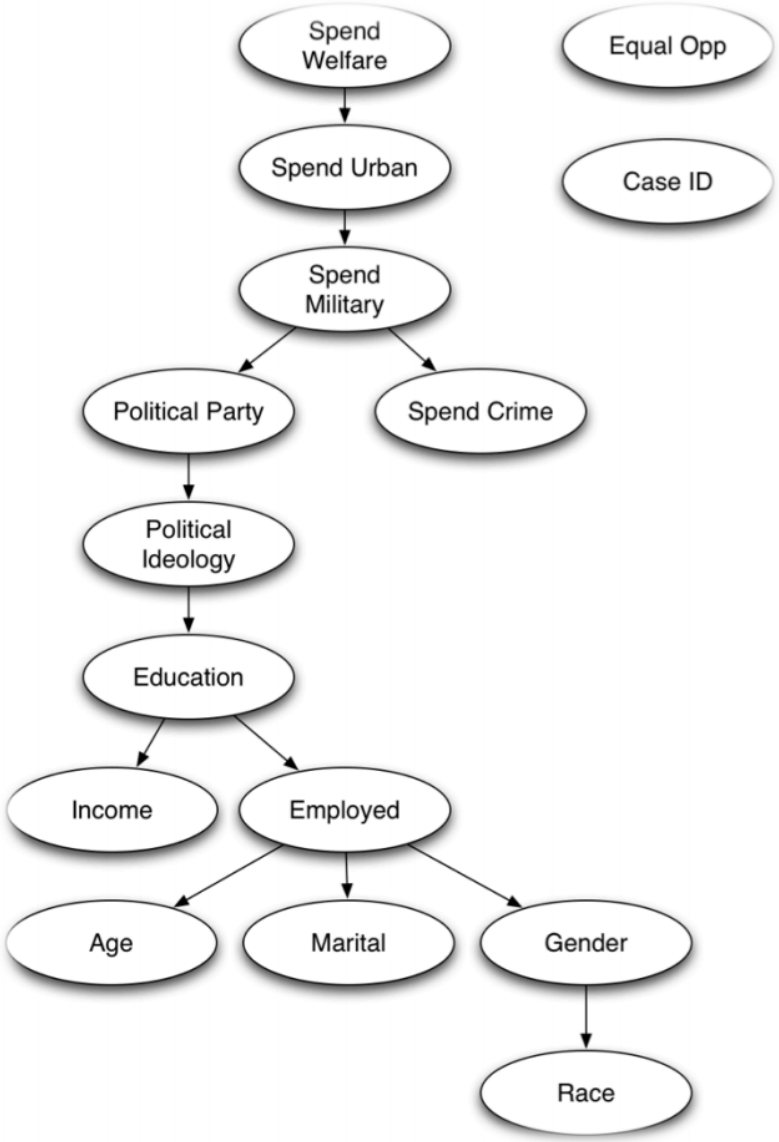
```
Sensor
disabled.
```

# USHER

## [Chen et al. 2010]




MS Access data entry forms for Tanzanian HIV/AIDS monitoring



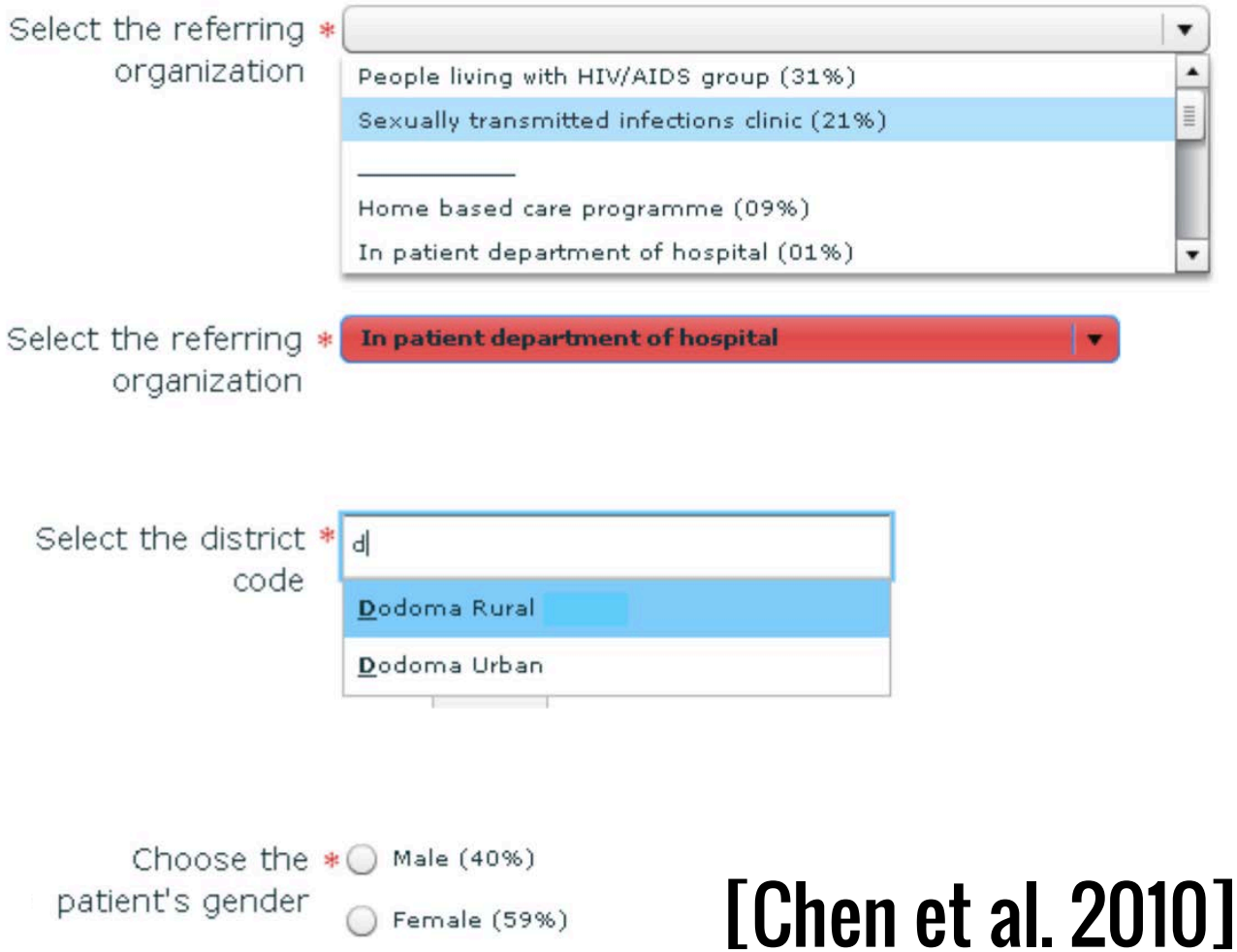

**BUILD A MODEL** to predict dependencies and relationships between questions.

[Chen et al. 2010]

# DYNAMIC ORDERING

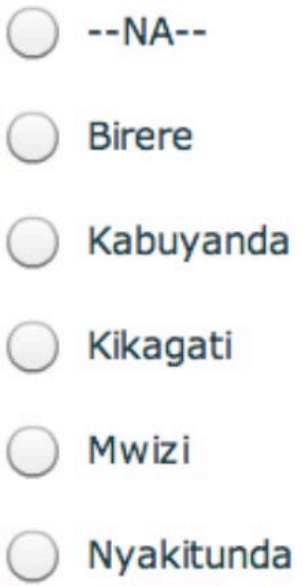
Always ask the most appropriate next question

Suggest the most likely answers




[Chen et al. 2010]

# SMART RE-ASKING AND SUGGESTIONS


FRICTION

AUTOMATING CONSTRAINTS

[Chen et al. 2010]

# DETECTING ERRORS

# 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**

# 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., Zip code or place name to lat-lon |
| DATA INTEGRATION | Mismatches and inconsistencies when combining data |

# MISSING AND IMPOSSIBLE VALUES

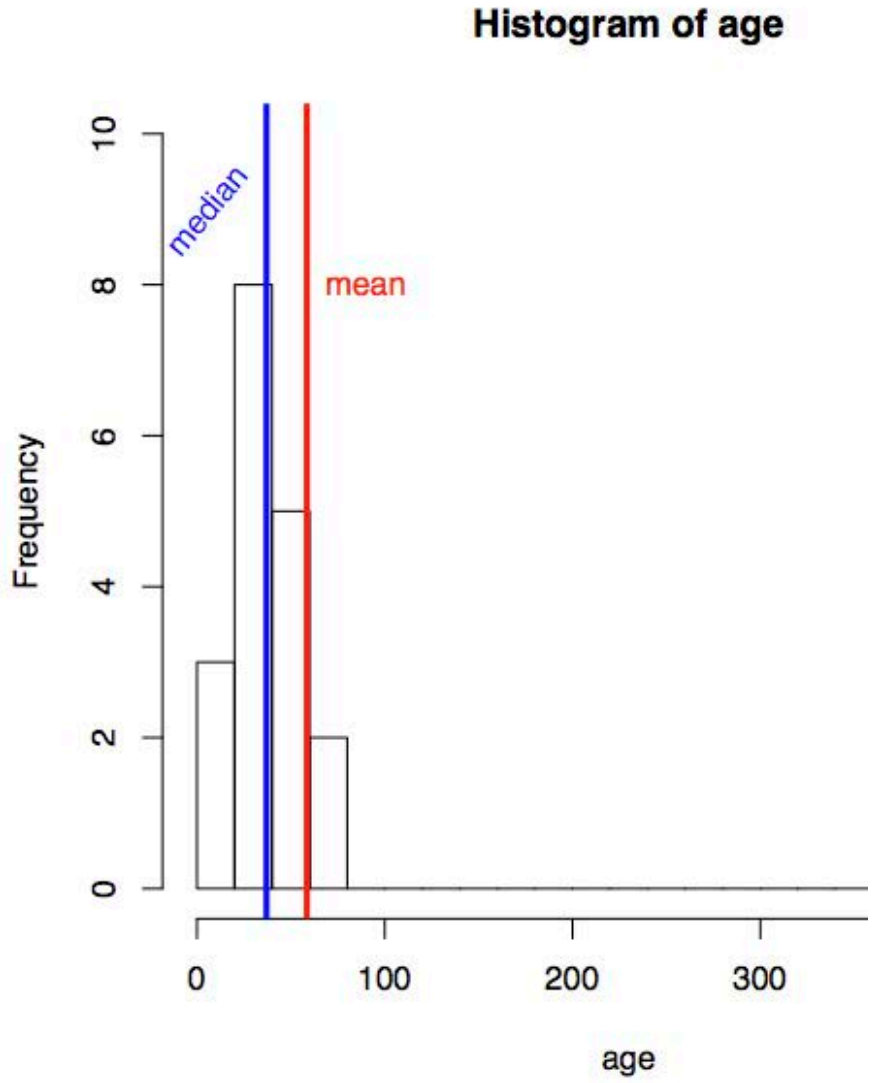
1. LOOK AT EMPTY/MISSING VALUES
2. COMPARE DATA TYPES
   - "xx" in an INT or DATE column
3. LOOK AT IMPOSSIBLE VALUES
   - Gender = 3
   - Heart Rate = 0
   - Unlikely Dates (e.g. "01/01/0001")

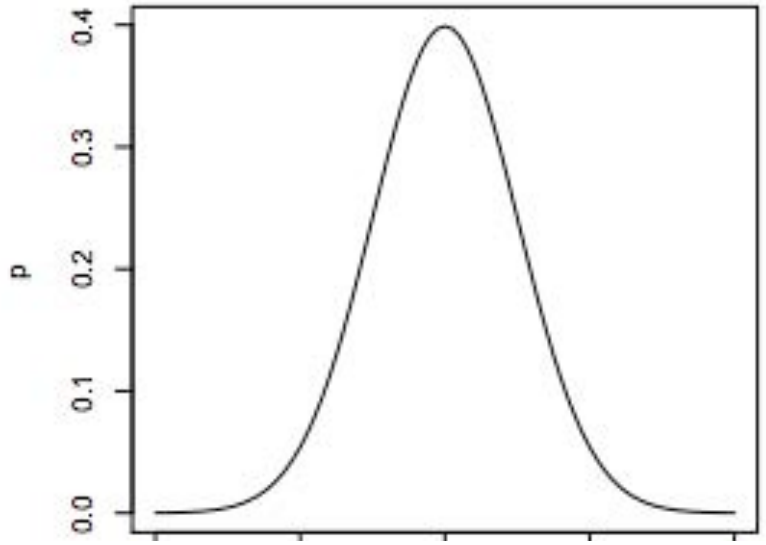
# OUTLIER DETECTION

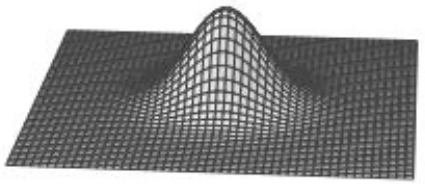
1. Examine distributions
2. Model data and look for residuals
3. Partition data

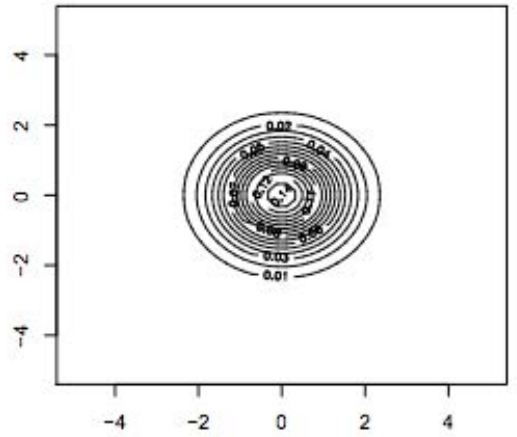
For one data dimension or multiple dimensions

# EXAMINE DISTRIBUTIONS





# DETECTING DUPLICATES

| **Title** |
| --- |
| 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 | 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

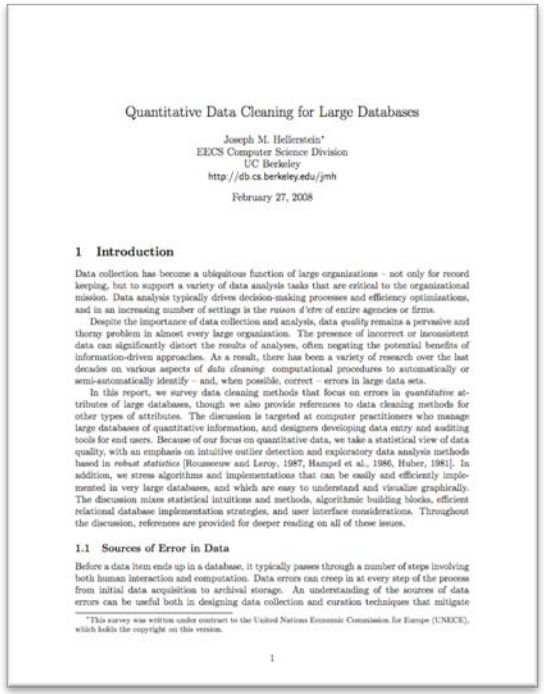
STRING/KEY COMPARISONS

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

# RULE-BASED 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 |

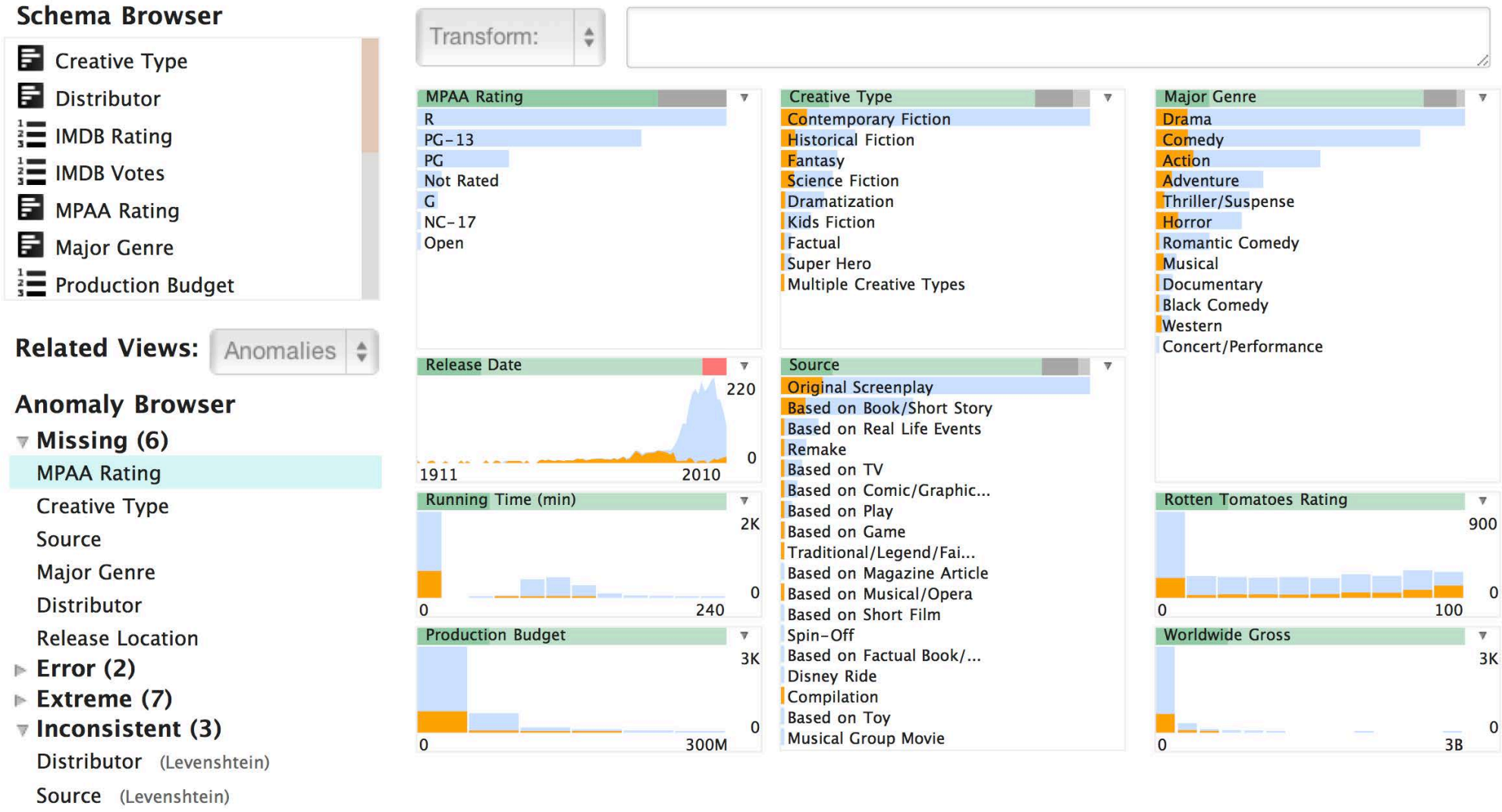
# “PROFILING” DATA

**Exploratory analysis** and **critical thinking** during data extraction and processing

**Understanding** what assumptions you can make about data

**Interactively** identifying and fixing data quality issues early

# INTERACTIVE PROFILING




PROFILER [KANDEL ET AL. 2012]

# 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

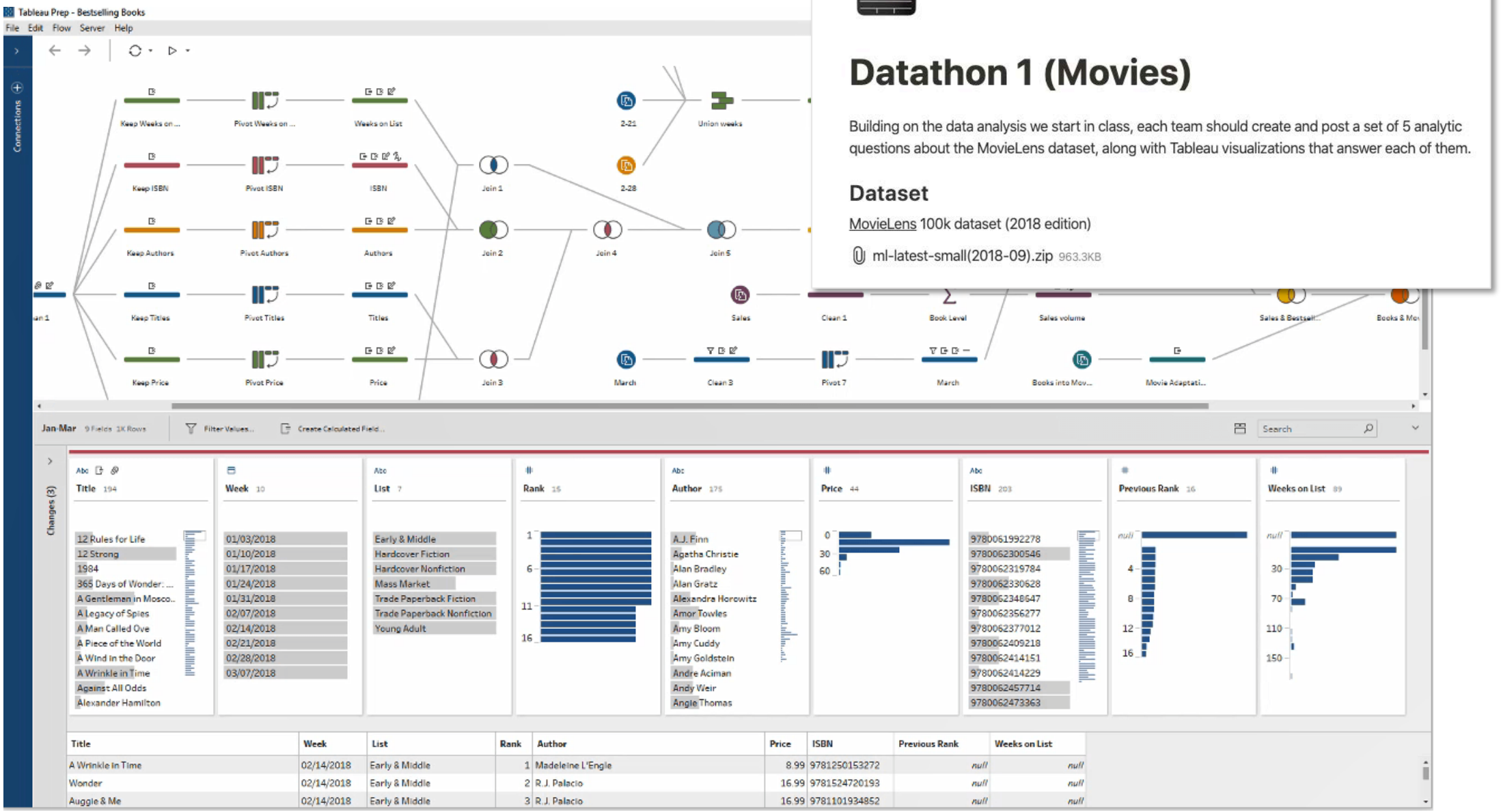
Fuzzy matching systems

Machine learning to detect/resolve errors

**Usually requires human judgment**
(especially for new data)

# LET'S TRY SOME WRANGLING

# TABLEAU PREP




## Datathon 1 (Movies)

Building on the data analysis we start in class, each team should create and post a set of 5 analytic questions about the MovieLens dataset, along with Tableau visualizations that answer each of them.

### Dataset

MovieLens 100k dataset (2018 edition)

ml-latest-small(2018-09).zip 963.3KB

# A FEW MORE TOOLS

# SPREADSHEETS

| State | 2004 | 2005 | 2006 | 2007 | 2008 |
|---|---|---|---|---|---|
| Alabama | 4029.3 | 3900 | 3937 | 3974.9 | 4081.9 |
| Alaska | 3370.9 | 3615 | 3582 | [illegible] | [illegible] |
| Arizona | 5073.3 | 4827 | 4741.6 | 4502.6 | 4587.3 |
| 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 | [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 |

+ FAMILIAR
+ VISUAL

- TEDIOUS
- TIME-CONSUMING
- REPETITIVE

# SCRIPTS

```
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="

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

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

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

# Delete  rows where split1 is null
```

+ REUSABLE
+ SCALABLE

- HARD
- TEDIOUS
- TIME-CONSUMING

# WRANGLING WITH JUPYTER/PANDAS

Iteratively **plot**, **filter**, and **modify** dataframes.

Really handy & powerful – you have already done a bit of this.

# USEFUL METHODS

## DataFrame.fillna()

Replace empty values.

```
meanAge = np.mean(df.Age)
df.Age = df.Age.fillna(meanAge)
```

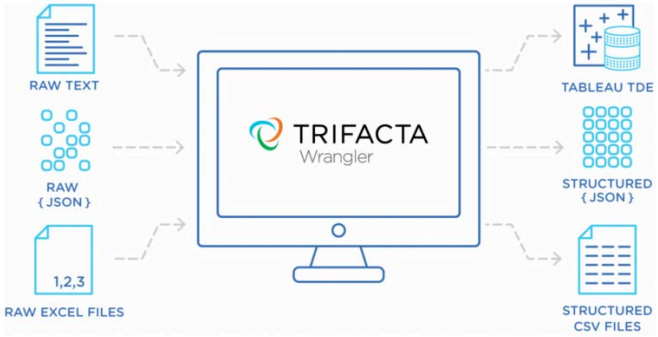
## DataFrame.interpolate()

Fill gaps in time series data.

## DataFrame.replace()

Regular expression replace across whole dataframes.

# 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/

# WRANGLER

# 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

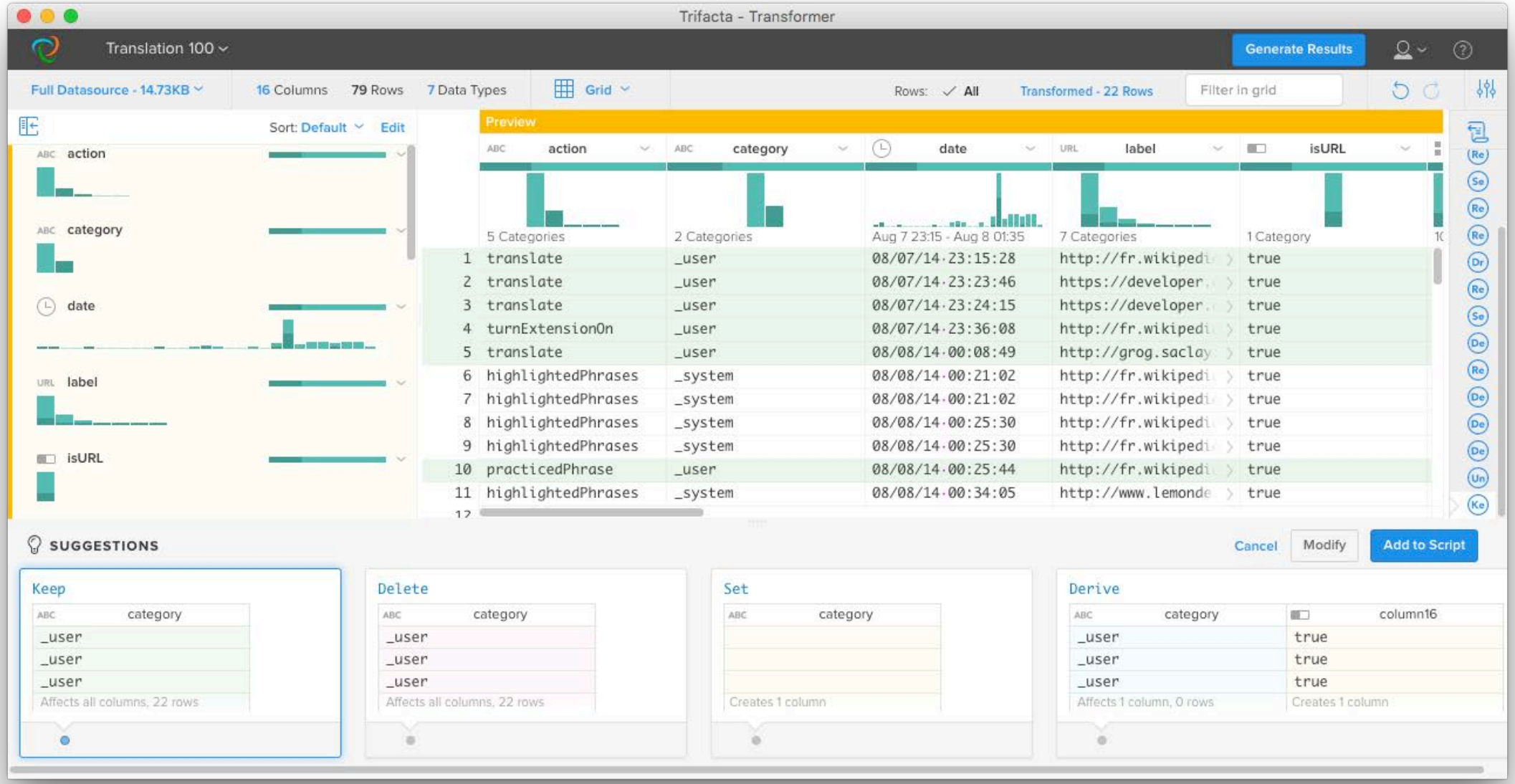
| | split | extract | split1 |
|---|---|---|---|
| 1 | 2004 | Alabama | 4029.3 |
| 2 | 2005 | Alabama | 3900 |
| 3 | 2006 | Alabama | 3937 |
| 4 | 2007 | Alabama | 3974.9 |
| 5 | 2008 | Alabama | 4081.9 |
| 6 | 2004 | Alaska | 3370.9 |
| 7 | 2005 | Alaska | 3615 |
| 8 | 2006 | Alaska | 3582 |
| 9 | 2007 | 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

# TRIFACTA DESKTOP



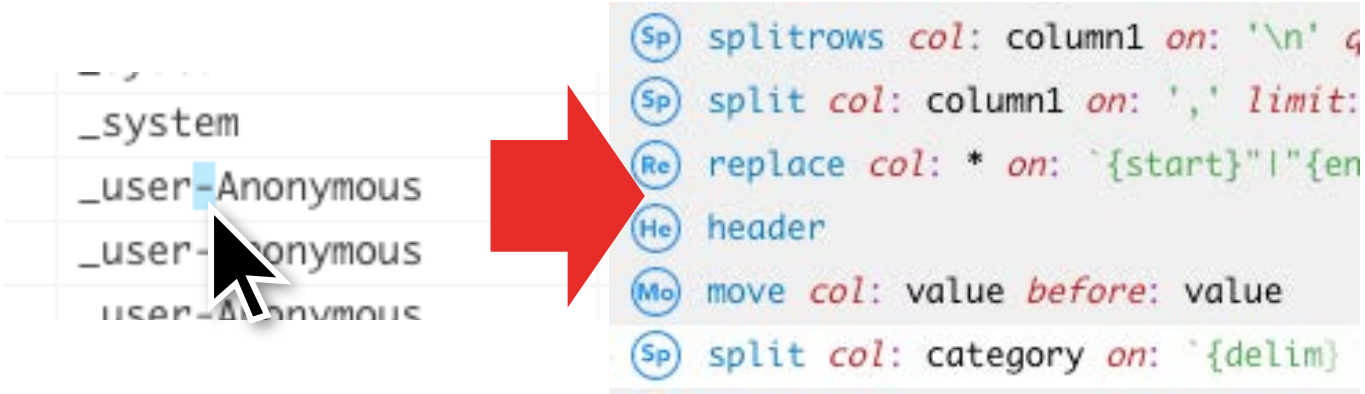
# INTERACTIVE SUPPORT FOR

Profiling
Cleaning
Reformatting
Unpacking

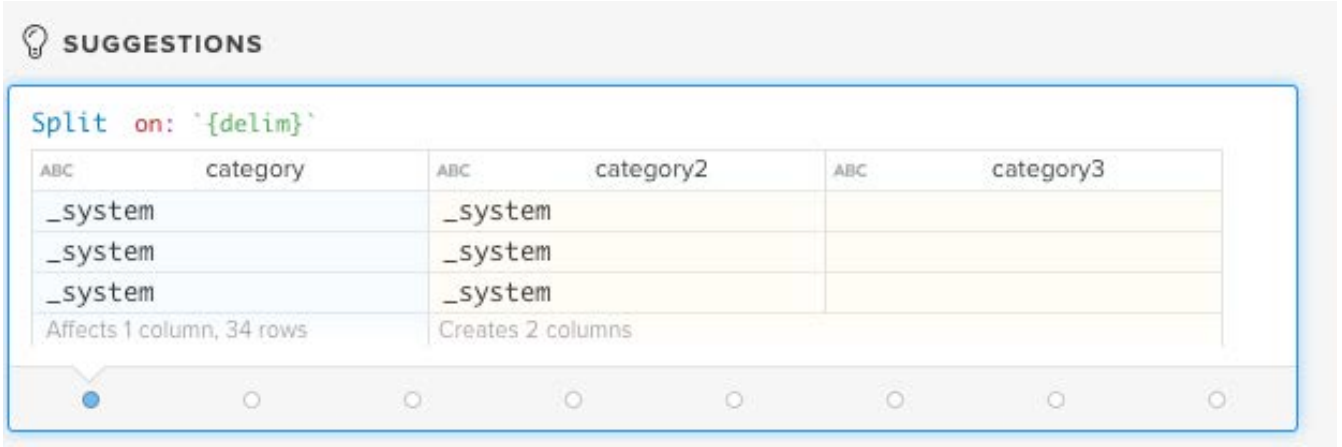
... And scaling up

# OVERVIEWS AND DATA PROFILING


# INTERACTIVELY BUILD SCRIPTS




# SMART SUGGESTIONS AND PREVIEWS



# WRANGLER TUTORIAL RESOURCES

How-to Videos & Tutorials

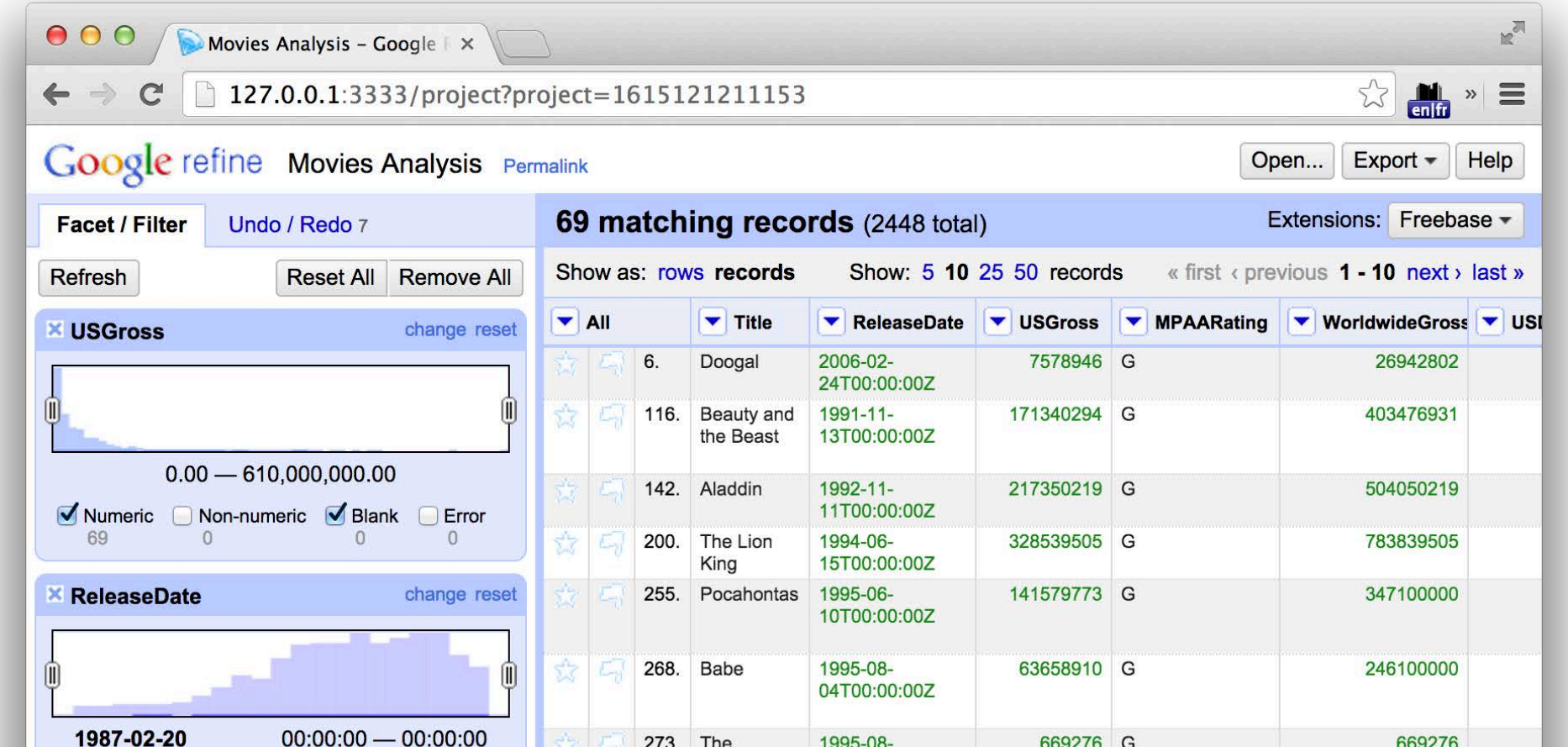
https://www.trifacta.com/support/articles/topics/125211-online-training/

Wrangle Language Documentation

https://docs.trifacta.com/display/PE/Wrangle+Language

# ANOTHER USEFUL TOOL
# OPEN REFINE

# WRANGLING IN OPEN REFINE




http://openrefine.org/

# INSTALL

https://goo.gl/A2peOM

https://github.com/OpenRefine/OpenRefine/releases

# QUESTIONS?