# Dirty Data in the Newsroom

## Comparing Data Preparation in Journalism and Data Science

ACM CHI Conference on Human Factors in Computing Systems
April 23-28, 2023, Hamburg, Germany

**Stephen Kasica**
University of British Columbia
Vancouver, Canada

Charles Berret
Linköping University
Norrköping, Sweden

Tamara Munzner
University of British Columbia
Vancouver, Canada
@tamara@vis.social
@tamaramunzner

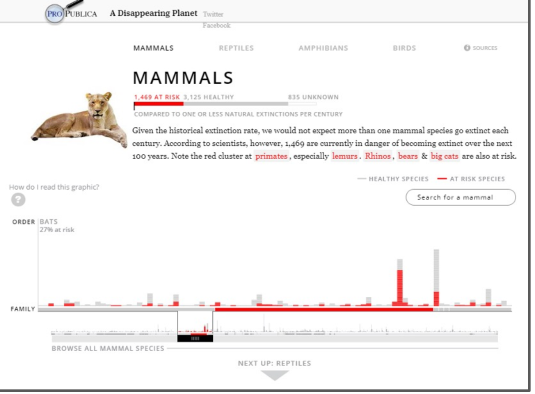
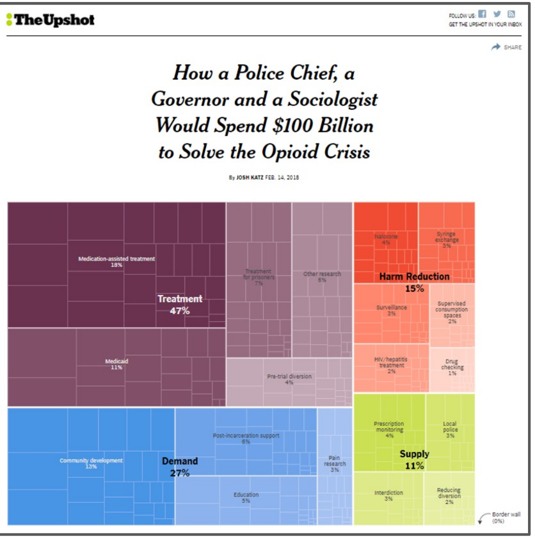
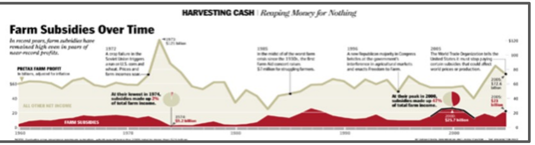
LINKÖPING UNIVERSITY

# Data Preparation

- Getting data ready for analysis or visualization
  - Includes: wrangling, cleaning, munging, gathering, integrating, etc.

- Time-consuming process in data science
  - Up to 80% of someone's time


Image by rawpixel.com on Freepik


ProPublica
A Disappearing Planet
MAMMALS
REPTILES
AMPHIBIANS
BIRDS
MAMMALS
1,469 AT RISK
3,125 HEALTHY
835 UNKNOWN
COMPARED TO ONE OR LESS NATURAL EXTINCTIONS PER CENTURY
Given the historical extinction rate, we would not expect more than one mammal species go extinct each century. According to scientists, however, 1,469 are currently in danger of becoming extinct over the next 100 years. Note the red cluster at primates, especially lemurs. Rhinos, bears & big cats are also at risk.
HEALTHY SPECIES
AT RISK SPECIES
How do I read this graphic?
Search for a mammal
ORDER
BATS
27% at risk
FAMILY
BROWSE ALL MAMMAL SPECIES
NEXT UP: REPTILES

TheUpshot
How a Police Chief, a Governor and a Sociologist Would Spend $100 Billion to Solve the Opioid Crisis
By JOSH KATZ FEB. 14, 2018
Treatment
47%
Medication-assisted treatment
18%
Medicaid
11%
Harm Reduction
15%
Demand
27%
Supply
11%
Border wall
(0%)

HARVESTING CASH
Farm Subsidies Over Time

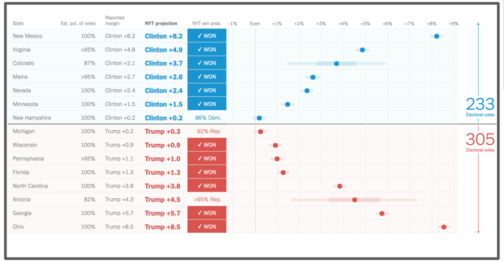
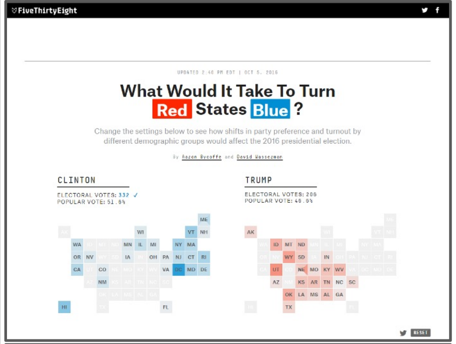
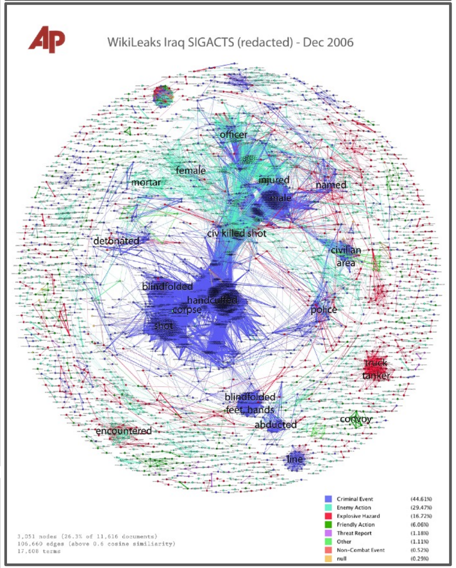
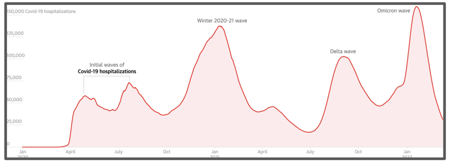
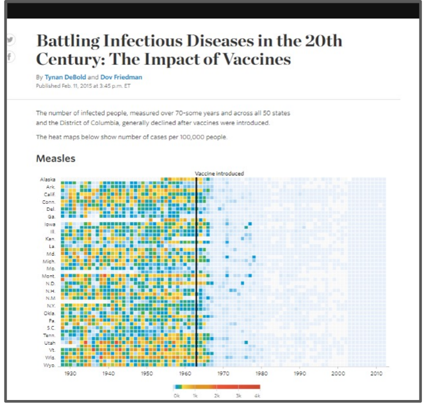
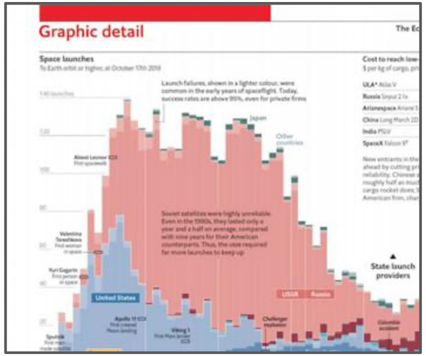
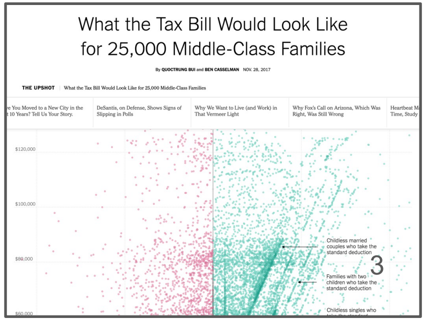
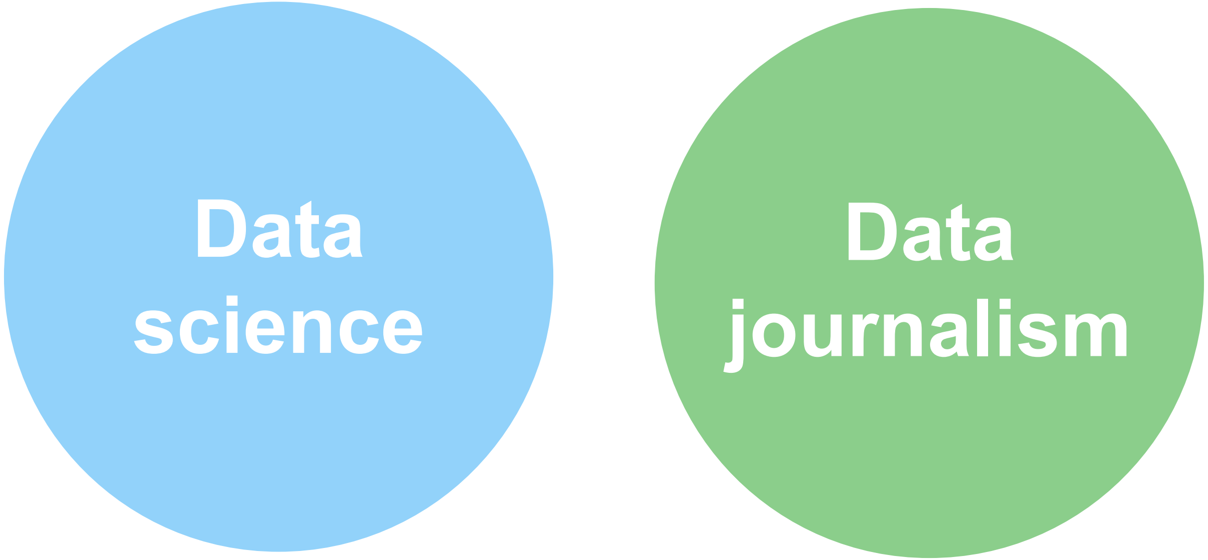
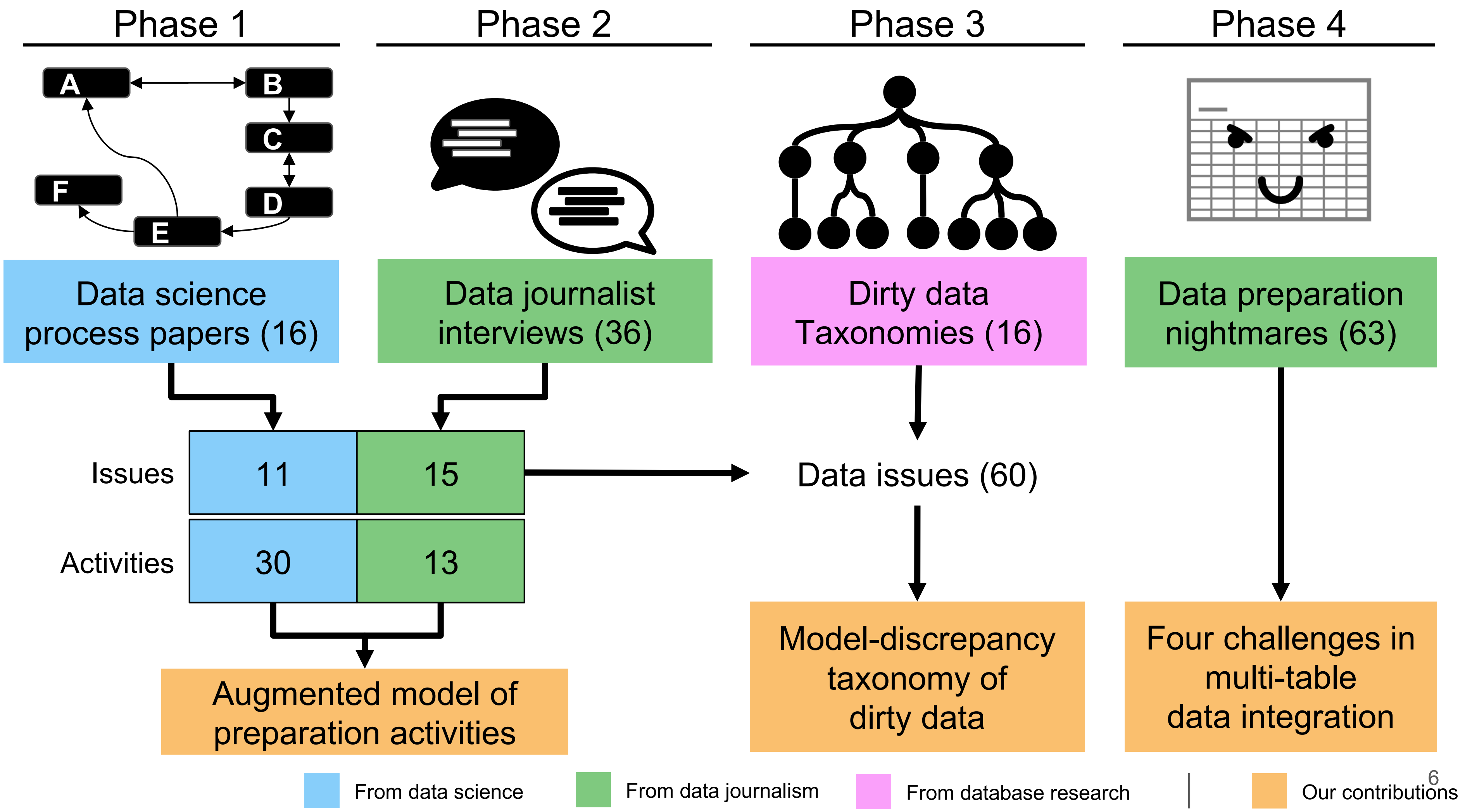
The New York Times
"All the News That's Fit to Print"
Late Edition
NEW YORK, WEDNESDAY, APRIL 8, 2020
How Outbreak Kept New York A Step Behind
Infighting, Delays and Unheeded Warnings
How the Coronavirus Toll Grew
BLACK AMERICANS BEAR THE BRUNT AS DEATHS CLIMB
Statistics From Certain Cities and States Highlight Entrenched Inequality
As Limits Ease, Wuhan Limps Into New Life
President Ousts Official Policing Relief Spending
In Wisconsin, Sense of Strain Shadows Vote

233
305

FiveThirtyEight
What Would It Take To Turn Red States Blue?
Change the settings below to see how shifts in party preference and turnout by different demographic groups would affect the 2016 presidential election.
CLINTON
TRUMP

AP
WikiLeaks Iraq SIGACTS (redacted) - Dec 2006
officer
female
mortar
injured
male
named
detonated
civ killed shot
civilian
area
blindfolded
handcuffed
corpse
shot
police
truck
tanker
blindfolded
feet hands
abducted
convoy
encountered
lime
Criminal Event
Enemy Action
Explosive Hazard
Friendly Action
Threat Report
Other
Non-Combat Event
null

Winter 2020-21 wave
Omicron wave
Delta wave
Initial waves of Covid-19 hospitalizations

Battling Infectious Diseases in the 20th Century: The Impact of Vaccines
By Tynan DeBold and Dov Friedman
The number of infected people, measured over 70-some years and across all 50 states and the District of Columbia, generally declined after vaccines were introduced.
The heat maps below show number of cases per 100,000 people.
Measles
Vaccine introduced

Graphic detail
Space launches

What the Tax Bill Would Look Like for 25,000 Middle-Class Families




Illustration by Barry Blitt, *Vanity Fair*




**How closely does research on data scientists apply to data journalists, with regards to data preparation?**



Phase 1
Phase 2
Phase 3
Phase 4
A
B
C
D
E
F
Data science process papers (16)
Data journalist interviews (36)
Dirty data Taxonomies (16)
Data preparation nightmares (63)
Issues
Activities
11
15
30
13
Data issues (60)
Augmented model of preparation activities
Model-discrepancy taxonomy of dirty data
Four challenges in multi-table data integration
From data science
From data journalism
From database research
Our contributions

# Contributions

1

Augmented model of preparation activities

2

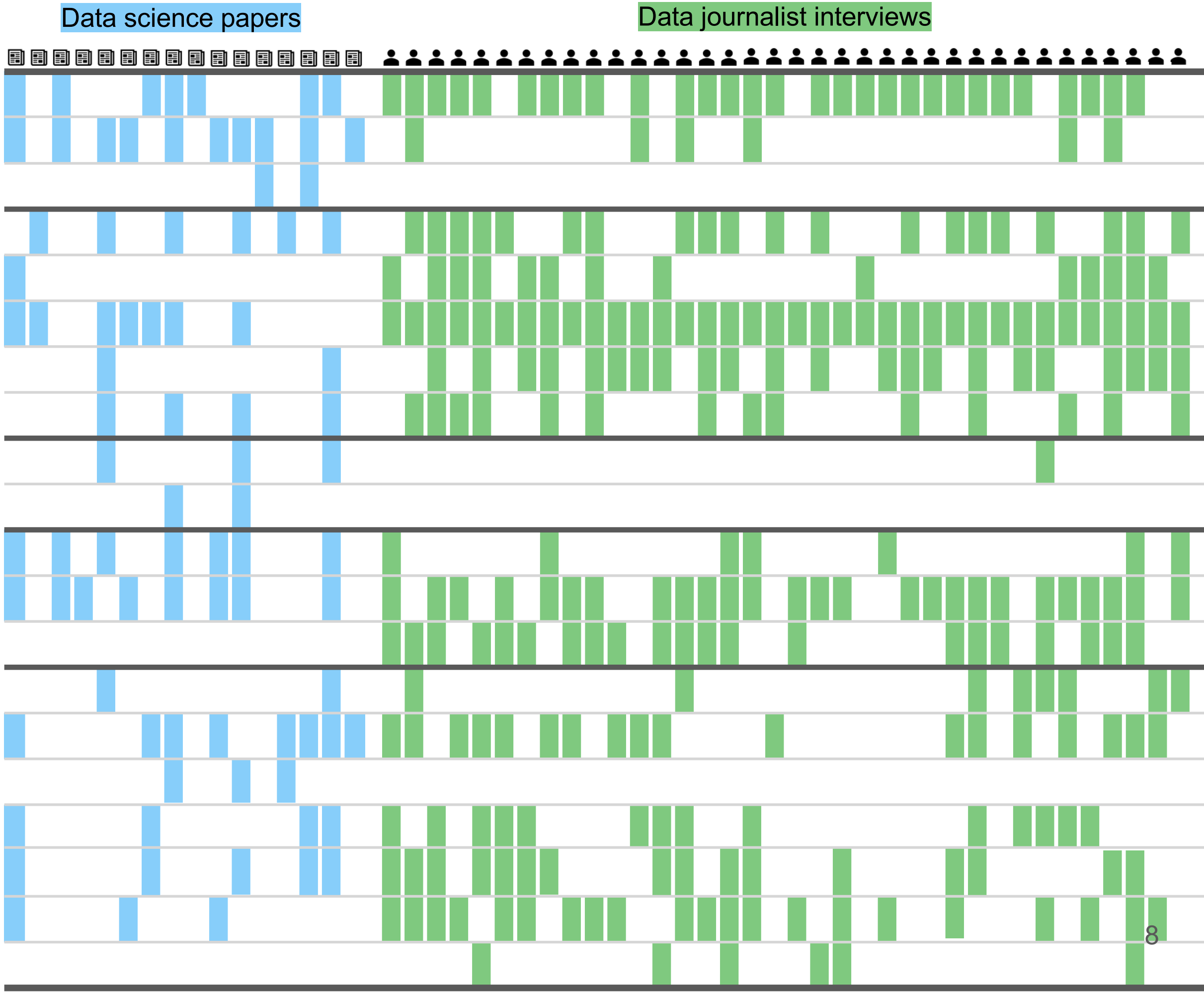
Model-discrepancy taxonomy of dirty data

3

Challenges in multi-table data integration

# Augmented model of prep. activities

| Crisan Model | | Our analysis |
|---|---|---|
| Prepare | Initiate | Establish goals |
| | | Make a plan |
| | | Test proof of concept |
| | Gather | Locate existing datasets |
| | | Collect new data |
| | | Integrate multiple datasets |
| | | Parse documents |
| | | Request datasets |
| | Create | Impute |
| | | Synthesize |
| | Profile | Assess quality |
| | | Understand semantics |
| | | Verify transformation |
| | Wrangle | Aggregate data |
| | | Transform data schema |
| | | Label data |
| | | Normalize values |
| | | Remove data |
| | | Standardize values |
| | | Identify items |



# Contributions

1

Augmented model of preparation activities

2

Model-discrepancy taxonomy of dirty data

3

Challenges in multi-table data integration

# Model-discrepancy taxonomy of dirty data

- Consider data as a design artifact
  - Dirty data = discrepancy in mental models
- Extend issue analysis to incorporate database literature
  - Analyze 16 taxonomies on dirty data: cluster 330 issues → 45 DB issues
- Combine into synthesis set of 60 issues
- Categorize into new model-discrepancy taxonomy
  - Data qualities axis
    - Existing qualities: completeness, accuracy
    - New qualities: form, granularity, relation, semantics
  - Data objects axis: table, attribute, item, value
- More details in the paper

# Contributions

1

Augmented model of preparation activities

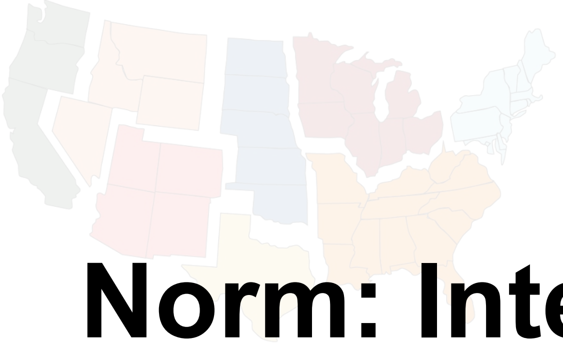
2

Model-discrepancy taxonomy of dirty data

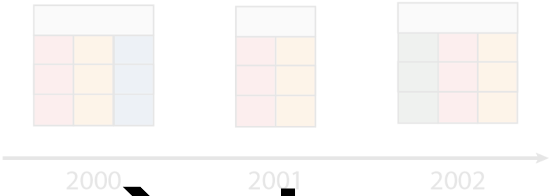
3

Challenges in multi-table data integration

# Four integration challenges



Regional



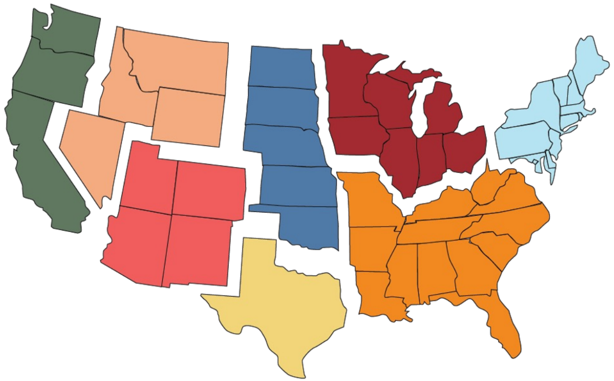

Diachronic

**Norm: Integrate → clean**

**Findings: Clean → integrate**

Fragmented

Disparate

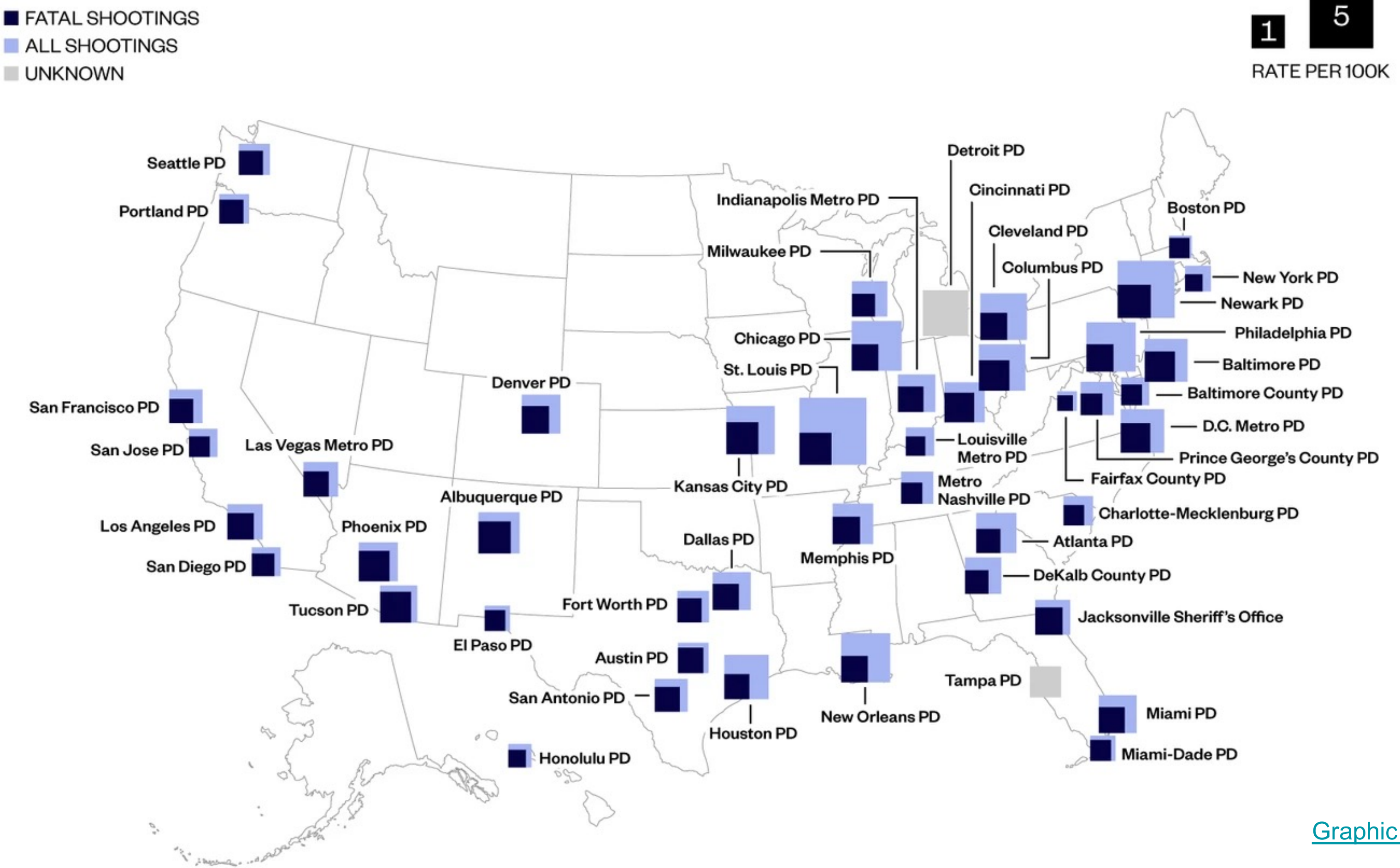


# Regional datasets

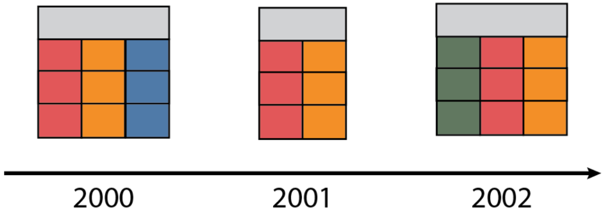
Tables with inconsistencies due to

independent, spatially dispersed data sources

# Regional: Police shootings in the United States



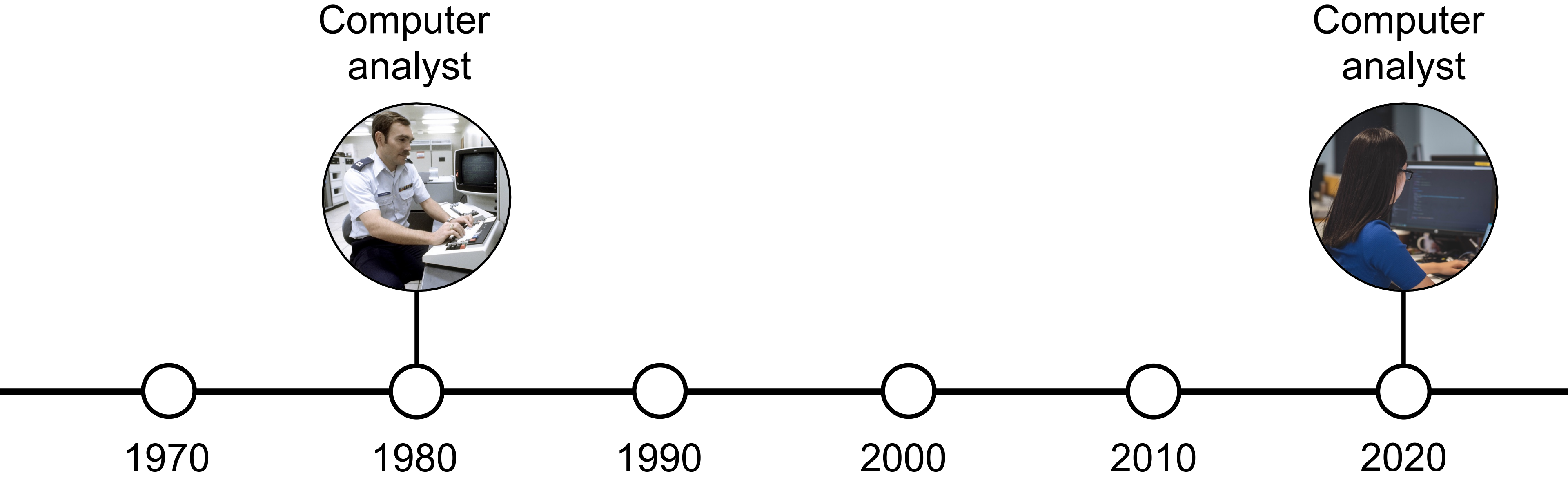

Graphic by *Vice*




# Diachronic datasets

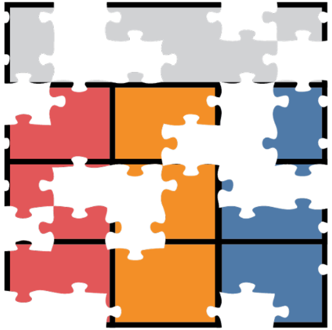
Tables on the same phenomena that evolve over time

# Diachronic: Economic data from Bureau of Labor Statistics





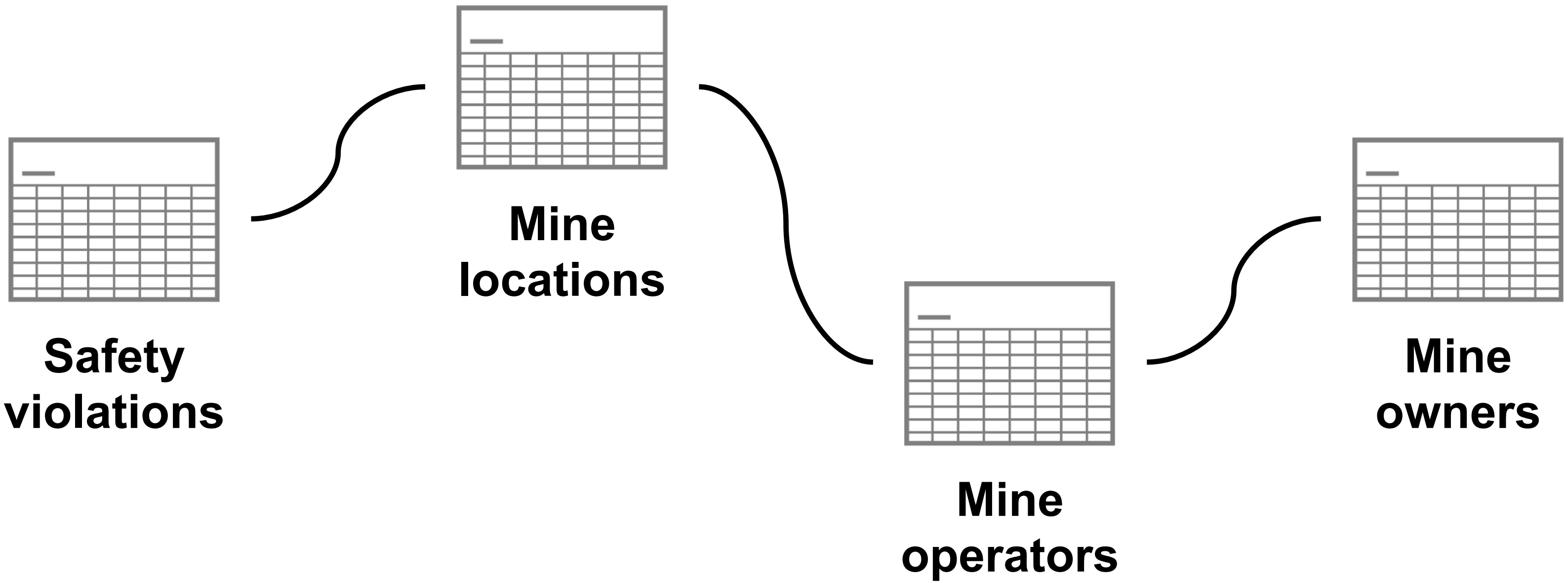

Photos by Boyd Belcher/US Air Force and Mimi Thain/Unsplash

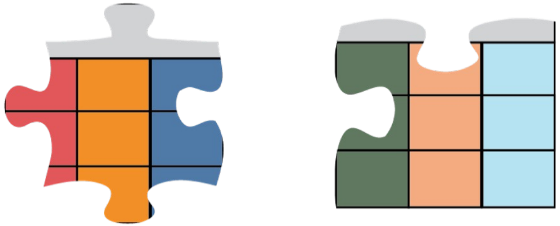


# Fragmented datasets

Tables on a similar topic that contain different yet related items.

# Fragmented: Unpaid mine safety violations





# Disparate datasets

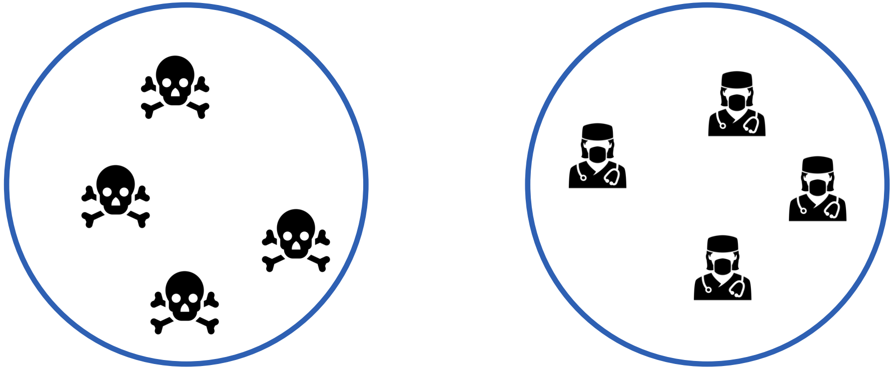
Tables that are topically dissimilar and seemingly unrelated.

# Disparate: Opioid overdoses

Death records

Healthcare workers dying from opioid overdoses

Healthcare workers



THE SPOKESMAN-REVIEW

Spokane, Washington Est. May 19, 1883

Washington Idaho

NEWS > SPOKANE

**Washington nurses, health care workers are dying of opioid overdoses**

Sun., Feb. 4, 2018

Icons by Minh Do and Sascha Elmers, Noun Project

# Dirty Data in the Newsroom

## Comparing Data Preparation in Journalism and Data Science

ACM CHI Conference on Human Factors in Computing Systems
April 23-28, 2023, Hamburg, Germany

**Stephen Kasica**
University of British Columbia
Vancouver, Canada

Charles Berret
Linköping University
Norrköping, Sweden

Tamara Munzner
University of British Columbia
Vancouver, Canada
@tamara@vis.social
@tamaramunzner

**Contributions:**

- Augmented model of preparation activities
- New model-discrepancy taxonomy of dirty data
- Four challenges in multi-table data integration