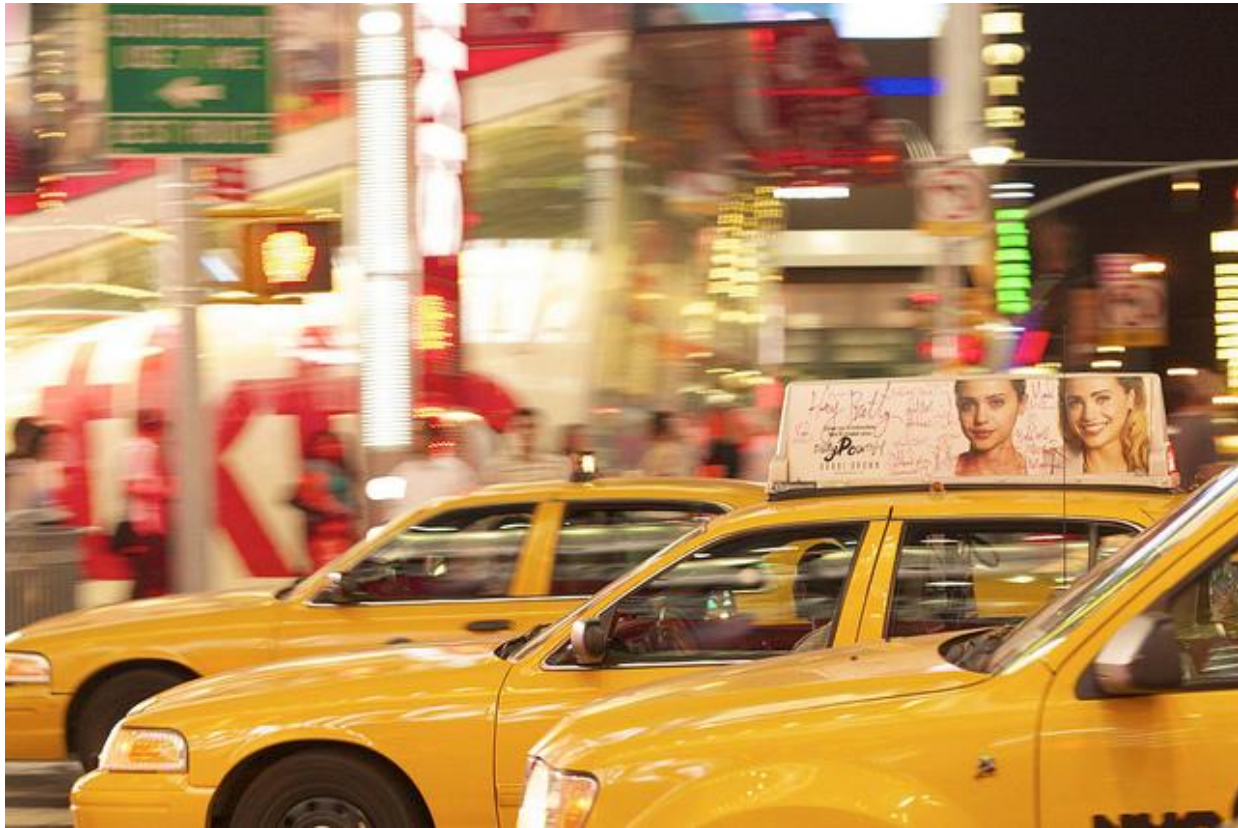


# **Good Data Gone Bad, Bad Data Gone Worse**

Renee Phillips [pgconf.eu](https://pgconf.eu) 2019



This is me.



Sakeeb Sabaaka [Creative commons 2.0 license](#)

@DataRenee

<https://2019.pgconf.eu/f>

This is a talk about how good data goes bad, and how bad data gets worse

# First, what is data?

- Pieces of information, with a format
- Facts used for making decisions
- Information stored in and/or used by a computer

Data is:

A representation of some aspect of the  
world

Next, what is good data?

Fit for its intended uses in operations,  
planning, decision making





# Why do we want good data?

- Planning
- Operations
- Looking smart
- Saving money
- Decision making
- Completing transactions



Finally, what is bad data?

Not fit for its intended uses in  
operations, planning, decision  
making





What can we assess to check if the data might work for the intended purpose?

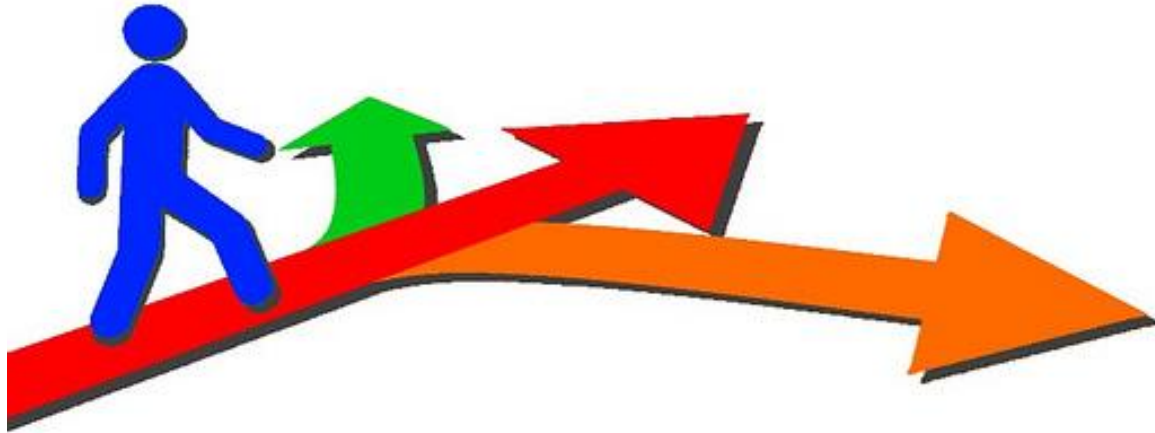
# The Six Primary Dimensions for Data Quality Assessment

1. Completeness
2. Uniqueness
3. Timeliness
4. Validity
5. Accuracy
6. Consistency

# Guidelines for quality assurance in health and health care research

1. Acquisition
2. Entry
3. Cleaning
4. Storage
5. Analysis

# 11 Things to Look At







# Assessing Data Quality

## Data Attributes

1. Accuracy
2. Completeness
3. Conformance
4. Consistency
5. Timeliness
6. Uniqueness
7. Validity

## Data Actions

1. Acquisition/Entry
2. Cleaning
3. Storage
4. Analysis

	Acquisition/ Entry	Cleaning	Storage	Analysis
Accuracy	X			
Completeness	X	X		
Conformance	X		X	
Consistency			X	
Timeliness	X		X	
Uniqueness	X		X	
Validity	X			

# Assessing Data Quality

## Data Attributes

1. Accuracy
2. Completeness
3. Conformance
4. Consistency
5. Timeliness
6. Uniqueness
7. Validity

## Data Actions

1. Acquisition/Entry
2. Cleaning
3. Storage
4. Analysis

# Accuracy

Have we stored the correct value?

# Accuracy at Entry

Signing up for  
airline rewards  
program, I entered  
my date of birth.  
Super easy.

Nice to meet you,  
Renee! Can you tell us  
more about you?

When is your birthday?

Date of birth

10 19

What is your gender?

Gender

FLYING  
BLUE  
FOR ME

Continue

# But Wait

My birthday was not in  
the month they return...



## Your personal information



In case you want to change your name or date of birth, please send an e-mail with a copy of your passport to [Flying Blue customer service](#).

Title

MS

First name(s)

RENEE

Last name

PHILLIPS

Date of birth

--10

# Ohhhh

The dreaded off by  
one error.



Nice to meet you,  
Renee! Can you tell us  
more about you?

When is your birthday?

Date of birth

December

January

February

March

April

day year

ender?

▼

FLYING  
BLUE  
FOR ME

Continue



# Just to be sure

This really isn't  
user error. August  
31 happens in  
every year...

Nice to meet you,  
Renee! Can you tell us  
more about you?

When is your birthday?

Date of birth

August



31

Please check the date entered and try again.



What is your gender?

Gender



FLYING  
BLUE  
FOR ME

Continue

# How does this happen?

- Bad UX/UI

# What can we do?

- Test the inputs when we design database
- Talk to the Front End team

# PawSense™

## catproof your computer

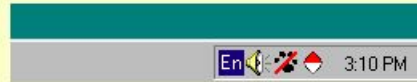


© 2000 BitBoost.com

When cats walk or climb on your keyboard, [they can enter random commands and data, damage your files, and even crash your computer.](#) This can happen whether you are near the computer or have suddenly been called away from it.

PawSense is a software utility that helps protect your computer from cats. It quickly [detects and blocks](#) cat typing, and also helps train your cat to stay off the computer keyboard.

- Every time your computer boots up, PawSense will automatically start up in the background to watch over your computer system.
- Even while you use your other software, PawSense constantly monitors keyboard activity. PawSense analyzes keypress timings and combinations to distinguish cat typing from human typing. PawSense normally recognizes a cat on the keyboard within one or two pawsteps.
- If a cat gets on the keyboard, PawSense makes a [sound that annoys cats.](#) This teaches your cat that getting on the keyboard is bad [even if humans aren't watching.](#)



<https://www.bitboost.com/pawsense/>

# How does this happen?

- Errors in entry
- Unauthorized entry

# What can we do?

- Knowledge Elicitation (talk to domain experts)
- Set database constraints
- Check anticipated vs actual values
- Maintain security

# Assessing Data Quality

## Data Attributes

1. Accuracy
2. **Completeness**
3. Conformance
4. Consistency
5. Timeliness
6. Uniqueness
7. Validity

## Data Actions

1. **Acquisition/Entry**
2. Cleaning
3. Storage
4. Analysis

# Completeness

Are there gaps between expected data and the data we have?







# How does this happen?

- Incorrect sampling
- Incomplete understanding of the business problem
- Not enough data available

# What can we do?

- Ask more questions of domain experts
- Find additional data sets

# Assessing Data Quality

## Data Attributes

1. Accuracy
2. **Completeness**
3. Conformance
4. Consistency
5. Timeliness
6. Uniqueness
7. Validity

## Data Actions

1. Acquisition/Entry
2. **Cleaning**
3. Storage
4. Analysis

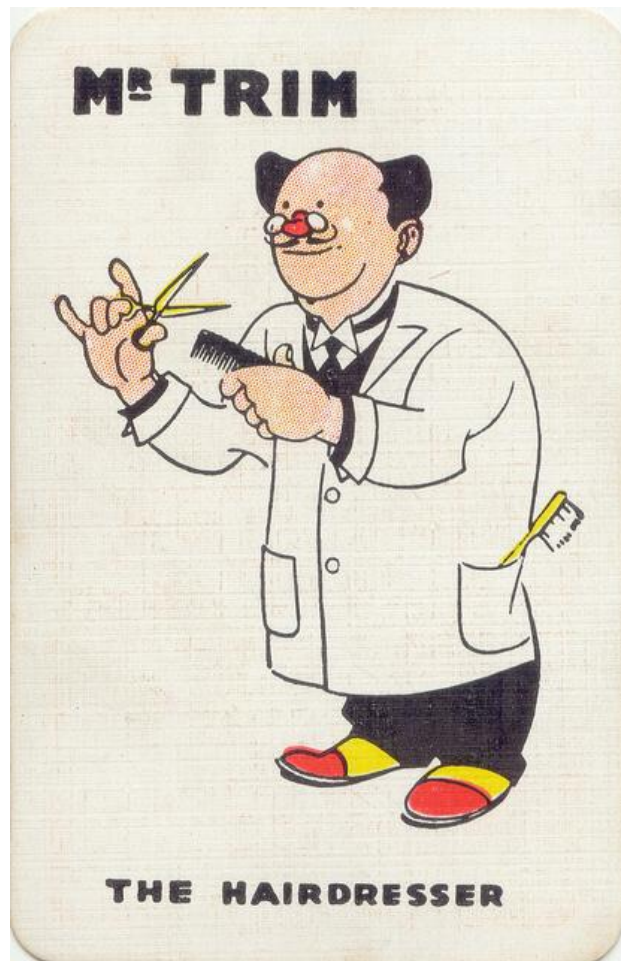
**M<sup>r</sup> TRIM**



**THE HAIRDRESSER**

Sometimes data is  
discarded for  
relevance or size





[patricia m creative commons 2.0 license](#)



[Conner McCall creative commons 2.0 license](#)

# How does this happen?

- Dataset is too large
- Dataset contains unnecessary columns



# What can we do?

- Import selectively
- Screen data carefully
- Trim and filter as appropriate

# Assessing Data Quality

## Data Attributes

1. Accuracy
2. **Completeness**
3. Conformance
4. Consistency
5. Timeliness
6. Uniqueness
7. Validity

## Data Actions

1. Acquisition/Entry
2. Cleaning
3. **Storage**
4. Analysis

# Storage



[smcgee creative commons 2.0 license](#)

Choose  
the right  
size  
storage for  
your  
database.



[United States Department of Agriculture License Creative Commons 2.0](#)

# How does this happen?

- Storage size chosen incorrectly or not updated
- Storage location or equipment chosen poorly
- Column, table, or database dropped in error

# What can we do?

- Be realistic about data needs, assess frequently
- Have backups
- Trust your users, apply alerts and process to prevent loss

# Assessing Data Quality

## Data Attributes

1. Accuracy
2. Completeness
3. Conformance
4. Consistency
5. Timeliness
6. Uniqueness
7. Validity

## Data Actions

1. Acquisition/Entry
2. Cleaning
3. Storage
4. Analysis

Dear Rich Bastard,



Dear Rich Bastard,

or maybe try

```
\pset null ' ^ \ \ _ (ツ) _ / ^ '
```

# How does this happen?

- Null is a black hole of data problems

# What can we do?

- Document the code
- Be careful with null (Go see Lætitia's talk about Null Unknown)
- Make NULL appear as something more noticeable

# Assessing Data Quality

## Data Attributes

1. Accuracy
2. Completeness
3. Conformance
4. Consistency
5. Timeliness
6. Uniqueness
7. Validity

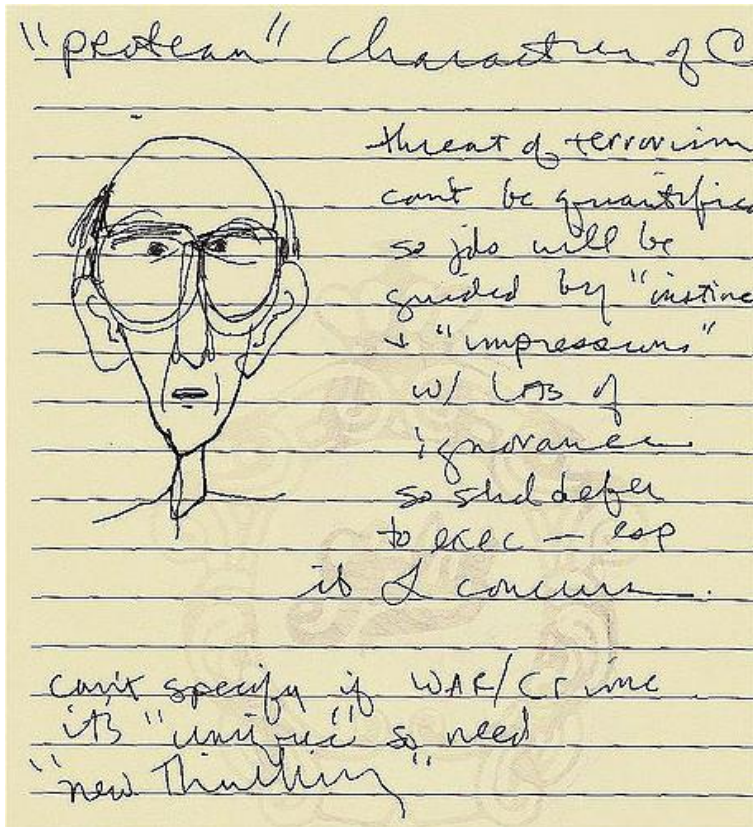
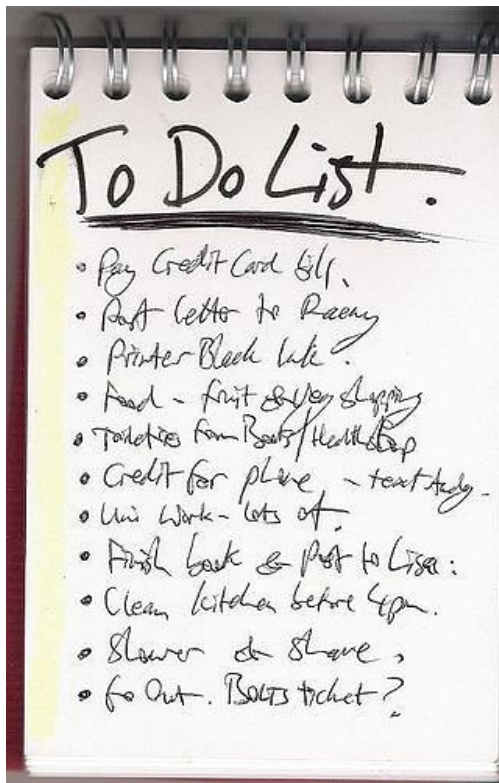
## Data Actions

1. Acquisition/Entry
2. Cleaning
3. Storage
4. Analysis

# Conformance

Is the data in a format that is expected and acceptable?

# Entry: Not to be Confused With Entropy



left [mister ebby creative commons 2.0 license](#)

right [Ann Althouse creative commons 2.0 license](#)

# How does this happen?

- Errors in entry
- Limitations in data collection/ availability
- Transforming from one data type to another

# What can we do?

- Test the inputs when we design database
- Have multiple entries of subset of data, check for consistency
- Search out additional datasets



# Assessing Data Quality

## Data Attributes

1. Accuracy
2. Completeness
3. Conformance
4. Consistency
5. Timeliness
6. Uniqueness
7. Validity

## Data Actions

1. Acquisition/Entry
2. Cleaning
3. Storage
4. Analysis



[anilmohabir creative commons 2.0](#)

-2, -1, 1, 2

2BCE, 1BCE, 1CE,

2CE

# How does this happen?

- Some databases have a year 0
- Money is a challenge

# What can we do?

- Beware of calendar challenges
- Don't use the money type
- Know what jurisdictions recognize leap seconds
- Daylight Savings
- Offsets

# Sometimes data is machine generated



[quisnovus creative commons 2.0 license](#)

# How does this happen?

- Improper machine calibration
- Improper machine reading

# What can we do?

- Set database constraints
- Ensure machine generated data is tested
- Ensure data collectors are well trained



# Assessing Data Quality

## Data Attributes

1. Accuracy
2. Completeness
3. Conformance
4. Consistency
5. Timeliness
6. Uniqueness
7. Validity

## Data Actions

1. Acquisition/Entry
2. Cleaning
3. Storage
4. Analysis

**NOTNULL**

# NOTNULL

People get creative

**NULL**

# NULL

A black hole of data quality issues  
Tony Hoar feels bad about it

# How does this happen?

- Data entry is forced to provide a response
- Radio button or check boxes provided are not sufficient to capture respondent needs

# What can we do?

- Consult domain experts
- Provide free form “other” option and/or “unknown” option
- Practice good knowledge elicitation

# Assessing Data Quality

## Data Attributes

1. Accuracy
2. Completeness
3. Conformance
4. Consistency
5. Timeliness
6. Uniqueness
7. Validity

## Data Actions

1. Acquisition/Entry
2. Cleaning
3. Storage
4. Analysis



Valid literal values for the "true" state are:

TRUE

't'

'true'

'y'

'yes'

'on'

'1'

For the "false" state, the following values can be used:

FALSE

'f'

'false'

'n'

'no'

'off'

'0'

# How does this happen?

- Text fields
- Integer fields

# What can we do?

- Use BOOLEAN data type
- Cast BOOLEAN to TEXT or INT if you must

# Consistency

Does the database only change data in expected ways?

Are there conflicts between data?



[Camille Rose](#) [creative commons 2.0 license](#)



# How does this happen?

- Selecting datasets for different areas of the database
- Databases started with inconsistency

# What can we do?

- Coalesce appropriately
- Clean data with a documented, reproducible method
- Set database constraints to check consistency



# Assessing Data Quality

## Data Attributes

1. Accuracy
2. Completeness
3. Conformance
4. Consistency
5. Timeliness
6. Uniqueness
7. Validity

## Data Actions

1. Acquisition/Entry
2. Cleaning
3. Storage
4. Analysis

Changing granularity may make analysis  
unreliable or impossible





# How does this happen?

- Data models are changed
- Database system is changed
- Data is transferred inappropriately

# What can we do?

- Exercise extreme caution when designing first data model
- Speak with domain experts
- Use new column instead of changing existing column

# Assessing Data Quality

## Data Attributes

1. Accuracy
2. Completeness
3. Conformance
4. Consistency
5. Timeliness
6. Uniqueness
7. Validity

## Data Actions

1. Acquisition/Entry
2. Cleaning
3. Storage
4. Analysis







# How does this happen?

- Improper database constraints
- Database system is changed
- Data is entered inappropriately
- Data not available in emergency

# What can we do?

- Have access to multiple datasets for emergencies
- Speak with domain experts to plan permission
- Be clear in analysis when the change happened and what that does to the analysis
- Be able to correct entries after initial input
- Use concurrency control in PostgreSQL
  - <https://www.postgresql.org/docs/current/mvcc.html>

# Timeliness

Is there more recent data that is appropriate to the task?

Is the data accessible quickly enough?

# Assessing Data Quality

## Data Attributes

1. Accuracy
2. Completeness
3. Conformance
4. Consistency
5. Timeliness
6. Uniqueness
7. Validity

## Data Actions

1. Acquisition/Entry
2. Cleaning
3. Storage
4. Analysis

# Google Maps Lost a Neighborhood. Again.

Via [Slashdot](#)



- Acquisition
- Entry
- Cleaning
- Analysis
- Consistency
- Accuracy

Really, this story is just like a greatest hits of problems.

# How does this happen?

- A newly discovered dataset is outdated
- A newly created dataset is not imported
- User is not aware of the age of data

# What can we do?

- Check provenance of data
- Actively search for additional sources
- Combine datasets where appropriate
- Identify if more data is really needed

# Assessing Data Quality

## Data Attributes

1. Accuracy
2. Completeness
3. Conformance
4. Consistency
5. Timeliness
6. Uniqueness
7. Validity

## Data Actions

1. Acquisition/Entry
2. Cleaning
3. Storage
4. Analysis





[Erinn Simon creative commons 2.0 license](#)  
[Jonathan Cristoferreti creative commons 2.0 license](#)

# How does this happen?

- Infrastructure limitations

# What can we do?

- Ensure enough storage on collector
- Offline first design

# Assessing Data Quality

## Data Attributes

1. Accuracy
2. Completeness
3. Conformance
4. Consistency
5. **Timeliness**
6. Uniqueness
7. Validity

## Data Actions

1. Acquisition/Entry
2. Cleaning
3. **Storage**
4. Analysis



[Michael Brace creative commons 2.0 license](#)

# How does this happen?

- Data set is large
- Data security prevents access
- Data is stale





# What can we do?

- Build or select better indexes
- Use partitioning
- Store less data
- Evaluate roles and users
- Update materialized views in transactions



# Assessing Data Quality

## Data Attributes

1. Accuracy
2. Completeness
3. Conformance
4. Consistency
5. Timeliness
6. Uniqueness
7. Validity

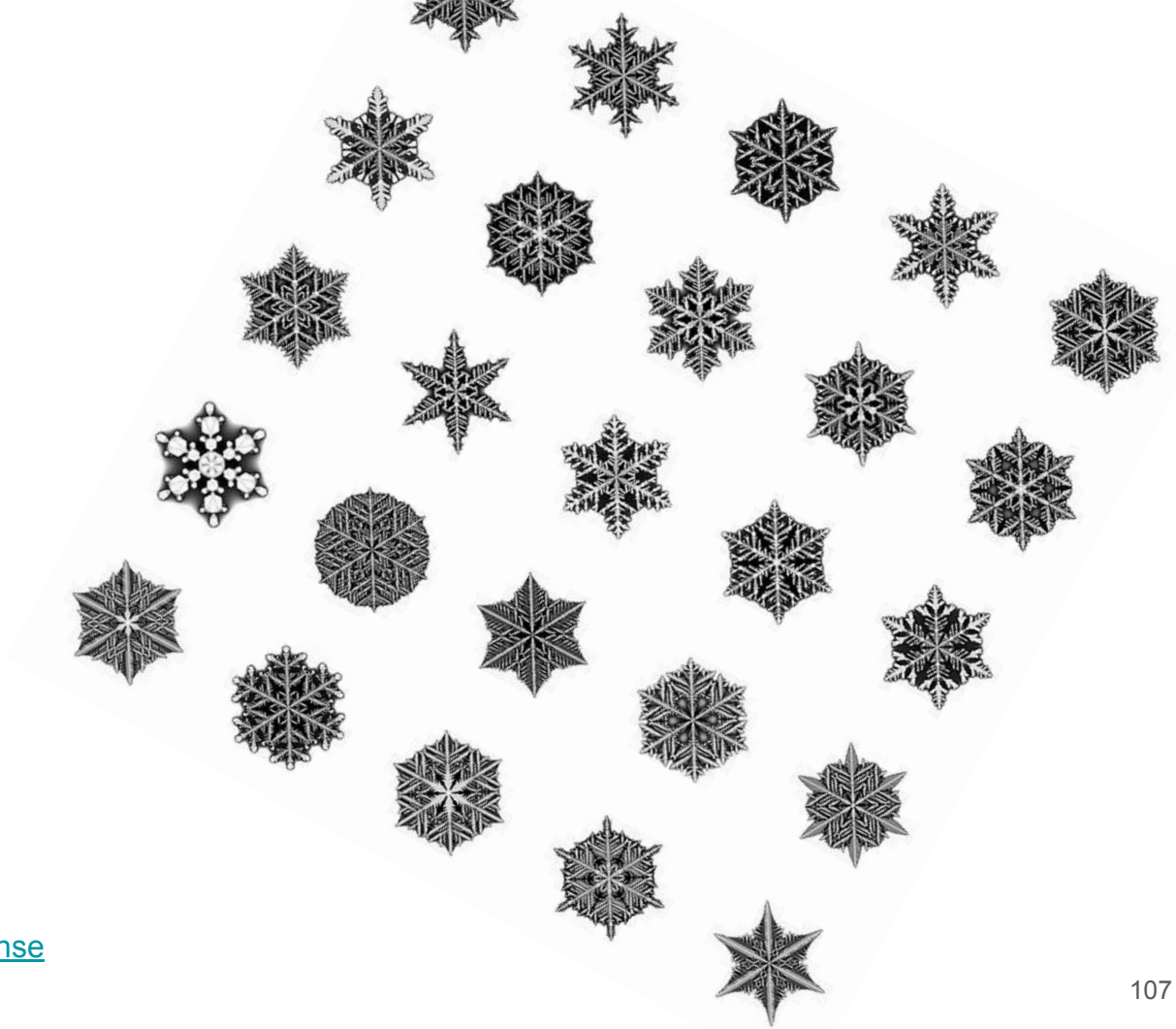
## Data Actions

1. Acquisition/Entry
2. Cleaning
3. Storage
4. Analysis

# Uniqueness

Are there duplicates in the dataset?

# Uniqueness



[matthew venn creative commons 2.0 license](#)

```
# SELECT DISTINCT fruit FROM fruits
ORDER BY fruit;
    fruit
```

```
-----
```

```
apple
```

```
banana
```

```
banananana
```

```
grape
```

```
loom
```

```
naranja
```



[sergio santos creative commons 2.0 license](#)

# How does this happen?

- Assumptions about domain
- Entry errors
- Naming things
- Duplicate entries

# What can we do?

- Consult domain experts
- Check entry
- Set primary key
- Clean based on information not instinct

# Validity

Are the format, syntax, and type correct?

Does the data have the potential to be accurate?



# Assessing Data Quality

## Data Attributes

1. Accuracy
2. Completeness
3. Conformance
4. Consistency
5. Timeliness
6. Uniqueness
7. **Validity**

## Data Actions

1. **Acquisition/Entry**
2. Cleaning
3. Storage
4. Analysis

patient | birth | temperature


-----+-----+-----

Susan | 5/12/84 | 101.4

Meg | 1/12/90 | 98.6

Julie | 1/12/90 | 97.2

Fiona | 3/31/65 | 970 

Sally | 4/3/01 | 111111 

000861 == 861

# How does this happen?

- Importing from various sources
- Improper data type selection
- Improper format changes to data
- NOTNULL
- Data Entry errors

# What can we do?

- Select correct data type
- Set value constraint on column
- Figure out a solution for NULL?

	Acquisition/ Entry	Cleaning	Storage	Analysis
Accuracy	X			
Completeness	X	X		
Conformance	X		X	
Consistency			X	
Timeliness	X		X	
Uniqueness	X		X	
Validity	X			



<https://2019.pgconf.eu/>