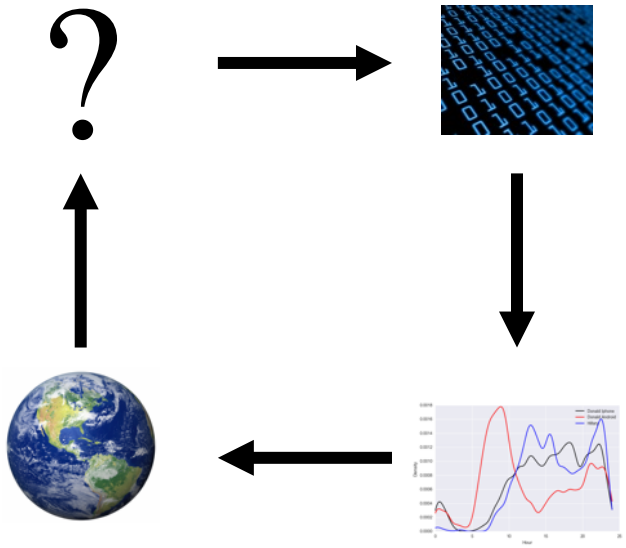
Data Science 100

Lecture 4: Data Wrangling

Slides by:

Joe Hellerstein

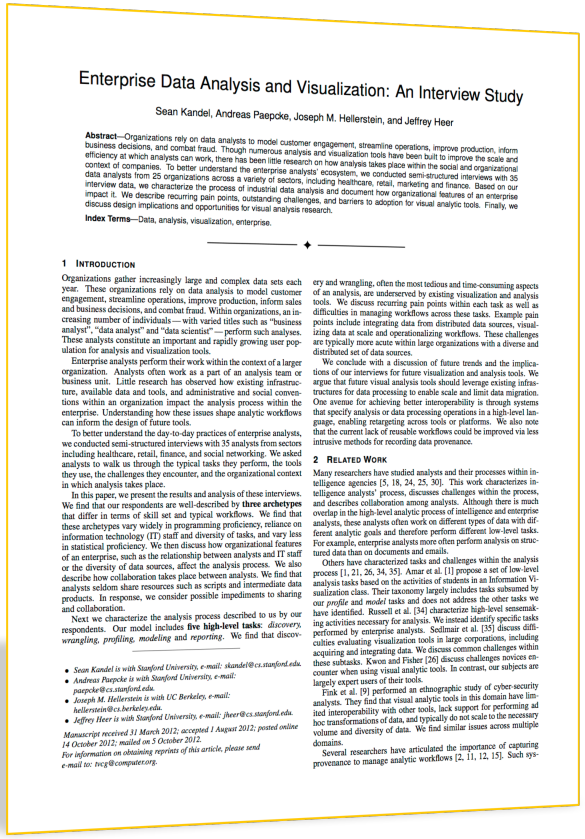
hellerstein@berkeley.edu



So you want to be a data scientist…

How will you spend your time?

Enterprise Data Analysis and Visualization: An Interview Study



Enterprise Data Analysis and Visualization: An Interview Study

Sean Kandel, Andreas Paepcke, Joseph M. Hellerstein, and Jeffrey Heer

Abstract—Organizations rely on data analysts to model customer engagement, streamline operations, improve production, inform business decisions, and combat fraud. Though numerous analysis and visualization tools have been built to improve the scale and efficiency at which analysts can work, there has been little research on how analysis takes place within the social and organizational context of companies. To better understand the enterprise analysts' ecosystem, we conducted semi-structured interviews with 35 data analysts from 25 organizations across a variety of sectors, including healthcare, retail, marketing and finance. Based on our interview data, we characterize the process of industrial data analysis and document how organizational features of an enterprise impact it. We describe recurring pain points, outstanding challenges, and barriers to adoption for visual analytic tools. Finally, we discuss design implications and opportunities for visual analysis research.

Index Terms—Data, analysis, visualization, enterprise.

1 Introduction

Organizations gather increasingly large and complex data sets each year. These organizations rely on data analysis to model customer engagement, streamline operations, improve production, inform sales and business decisions, and combat fraud. Within organizations, an increasing number of individuals — with varied titles such as "business analyst", "data analyst" and "data scientist" — perform such analyses. These analysts constitute an important and rapidly growing user population for analysis and visualization tools.

Enterprise analysts perform their work within the context of a larger organization. Analysts often work as a part of an analysis team or business unit. Little research has observed how existing infrastructure, available data and tools, and administrative and social conventions within an organization impact the analysis process within the enterprise. Understanding how these issues shape analytic workflows can inform the design of future tools.

To better understand the day-to-day practices of enterprise analysts, we conducted semi-structured interviews with 35 analysts from sectors including healthcare, retail, finance, and social networking. We asked analysts to walk us through the typical tasks they perform, the tools they use, the challenges they encounter, and the organizational context in which analysis takes place.

In this paper, we present the results and analysis of these interviews. We find that our respondents are well-described by **three archetypes** that differ in terms of skill set and typical workflows. We find that these archetypes vary widely in programming proficiency, reliance on information technology (IT) staff and diversity of tasks, and vary less in statistical proficiency. We then discuss how organizational features of an enterprise, such as the relationship between analysts and IT staff or the diversity of data sources, affect the analysis process. We also describe how collaboration takes place between analysts. We find that analysts seldom share resources such as scripts and intermediate data products. In response, we consider possible impediments to sharing and collaboration.

Next we characterize the analysis process described to us by our respondents. Our model includes **five high-level tasks**: *discovery, wrangling, profiling, modeling* and *reporting*. We find that discovery and wrangling, often the most tedious and time-consuming aspects of an analysis, are underserved by existing visualization and analysis tools. We discuss recurring pain points within each task as well as difficulties in managing workflows across these tasks. Example pain points include integrating data from distributed data sources, visualizing data at scale and operationalizing workflows. These challenges are typically more acute within large organizations with a diverse and distributed set of data sources.

We conclude with a discussion of future trends and the implications of our interviews for future visualization and analysis tools. We argue that future visual analysis tools should leverage existing infrastructures for data processing to enable scale and limit data migration. One avenue for achieving better interoperability is through systems that specify analysis or data processing operations in a high-level language, enabling retargeting across tools or platforms. We also note that the current lack of reusable workflows could be improved via less intrusive methods for recording data provenance.

2 Related Work

Many researchers have studied analysts and their processes within intelligence agencies [5, 18, 24, 25, 30]. This work characterizes intelligence analysts' process, discusses challenges within the process, and describes collaboration among analysts. Although there is much overlap in the high-level analytic process of intelligence and enterprise analysts, these analysts often work on different types of data with different analytic goals and therefore perform different low-level tasks. For example, enterprise analysts more often perform analysis on structured data than on documents and emails.

Others have characterized tasks and challenges within the analysis process [1, 21, 26, 34, 35]. Amar et al. [1] propose a set of low-level analysis tasks based on the activities of students in an Information Visualization class. Their taxonomy largely includes tasks subsumed by our *profile* and *model* tasks and does not address the other tasks we have identified. Russell et al. [34] characterize high-level sensemaking activities necessary for analysis. We instead identify specific tasks performed by enterprise analysts. Sedlmair et al. [35] discuss difficulties evaluating visualization tools in large corporations, including acquiring and integrating data. We discuss common challenges within these subtasks. Kwon and Fisher [26] discuss challenges novices encounter when using visual analytic tools. In contrast, our subjects are largely expert users of their tools.

Fisk et al. [9] performed an ethnographic study of cyber-security analysts. They find that visual analytic tools in this domain have limited interoperability with other tools, lack support for performing ad hoc transformations of data, and typically do not scale to the necessary volume and diversity of data. We find similar issues across multiple domains.

Several researchers have articulated the importance of capturing provenance to manage analytic workflows [2, 11, 12, 15]. Such sys-

- *Sean Kandel is with Stanford University, e-mail: skandel@cs.stanford.edu.*
- *Andreas Paepcke is with Stanford University, e-mail: paepcke@cs.stanford.edu.*
- *Joseph M. Hellerstein is with UC Berkeley, e-mail: hellerstein@cs.berkeley.edu.*
- *Jeffrey Heer is with Stanford University, e-mail: jheer@cs.stanford.edu.*

Manuscript received 31 March 2012; accepted 1 August 2012; posted online 14 October 2012; mailed on 5 October 2012.
For information on obtaining reprints of this article, please send e-mail to: tvcg@computer.org.

Interview study of 35 analysts

25 companies
Healthcare
Retail, Marketing
Social networking
Media
Finance, Insurance

Various titles
Data analyst
Data scientist
Software engineer
Consultant
Chief technical officer

Kandel et al. "Enterprise Data Analysis and Visualization: An Interview Study. IEEE Visual Analytics Science & Technology (VAST), 2012
http://db.cs.berkeley.edu/papers/vast12-interview.pdf

“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’ at all…

… Most of the time once you transform the data ... the insights can be scarily obvious.”

“Once you play with the data you realize you made an assumption that is completely wrong. It’s really useful, it’s not just a waste of time, even though you may be banging your head.”

“In practice it tends not to be just data prep, you are learning about the data at the same time, you are learning about what assumptions you can make.”

Big Data Borat
@BigDataBorat

Following

In Data Science, 80% of time spent prepare data, 20% of time spent complain about need for prepare data.

Data Wrangling

Aka Data Prep, Data Munging, Data Transformation

*Assessing and transforming raw data to make it **fit for use***

Data Wrangling

Aka Data Prep, Data Munging, Data Transformation

*Assessing and transforming raw data to make it **fit for use***

This is how you “get your head in the game”

- Understand what you have
- Assess strengths and weaknesses of your data
- Hypothesize about what to do with your data
- Get it ready

Nobody will know your data as well as you do while wrangling

- Not even the “you” of a few days later

Discussion

When is data “dirty”?

How does that happen?

Today: Data Unboxing and Wrangling

Basic tools:

- UNIX command line
- SublimeText editor

Trifacta: a free visual data wrangling tool

- Born at Berkeley/Stanford
- Codifies some good practices you can also follow “by hand”

Later: Python’s Pandas library

You may choose to use other tools too

- Good data scientists maintain a well-stocked toolbox

A bit of background before we jump in

Stages of Wrangling

Raw: Data ingestion & discovery ("unboxing")

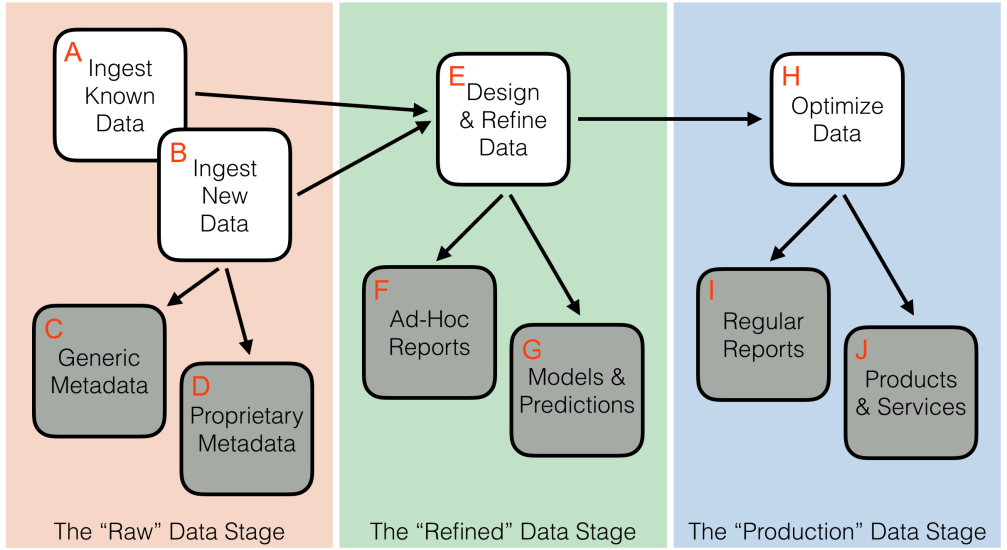
- What: Exploratory *ad hoc* analysis
- Who: The individual wrangler

Refined: Curating data for reuse

- What: Data warehousing, canonical models
- Who: Data curators, IT engineers, actuaries, etc.

Production: Ensuring feeds and workflows

- What: Recurrent, automated use cases:
  - Traditional (e.g. reporting) + New (e.g. recommenders)
- Who: Often involves SW engineers and IT/ops folks



Rattenbury, et al. "Data Wrangling: Techniques and Concepts for Agile Analytics". To appear, O'Reilly Media, 2017.

Today

We will focus on the “Raw→Refined” stage

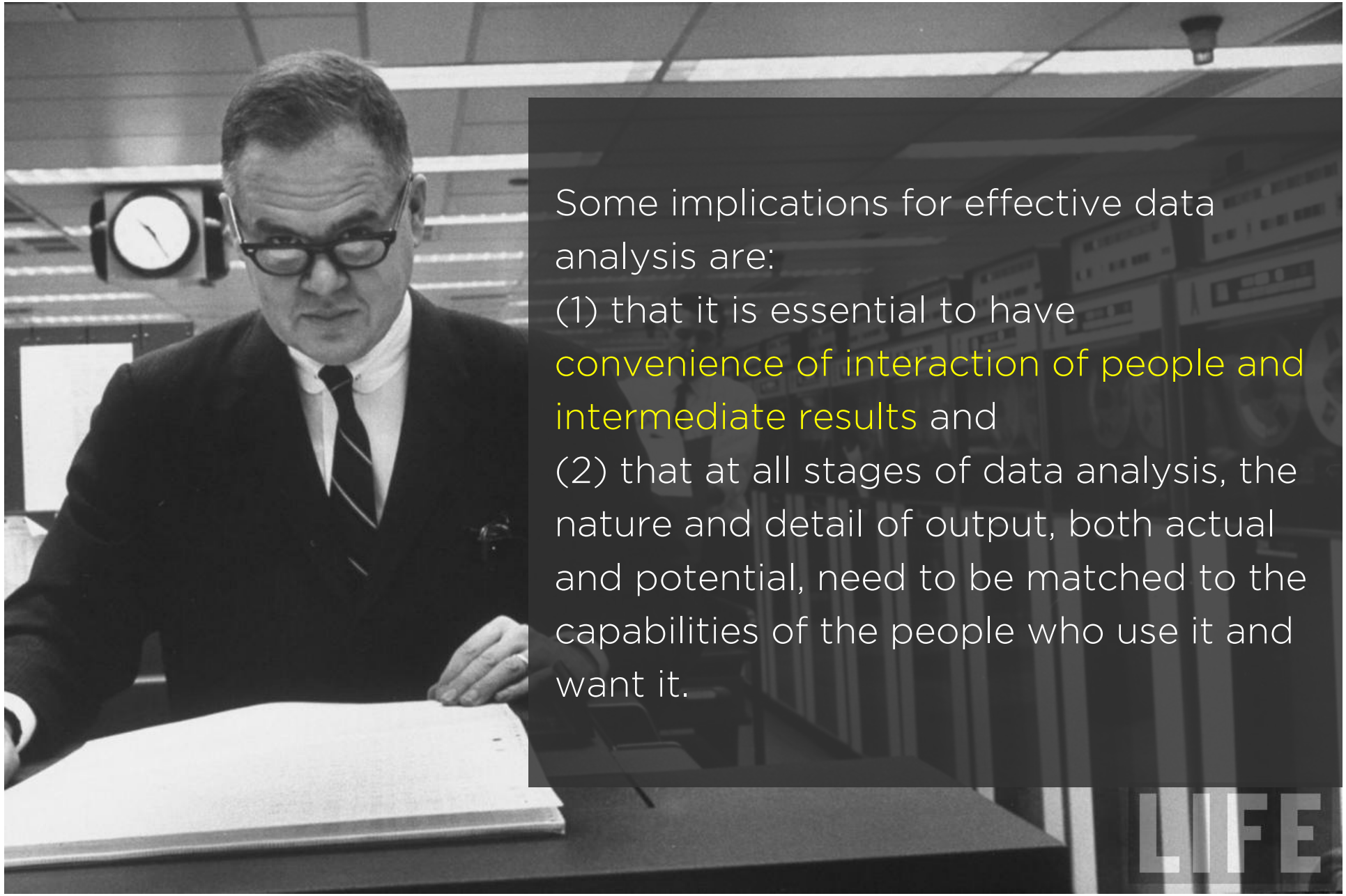
- Unboxing
- Transformation to analytics-ready structure
- Assessment/mitigation of quality issues

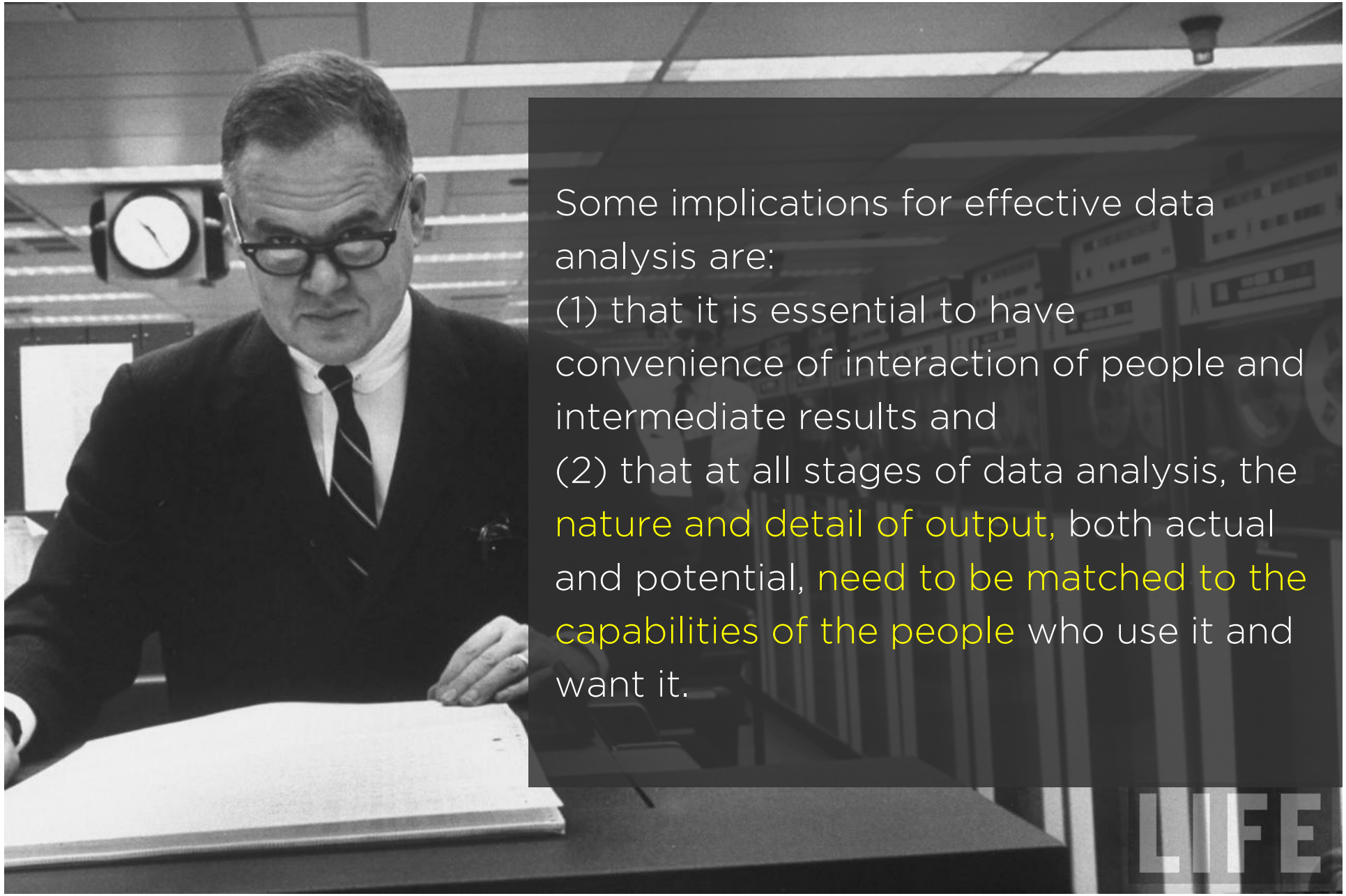
This is just a taste

- Soon you will learn to do this in Python & Pandas

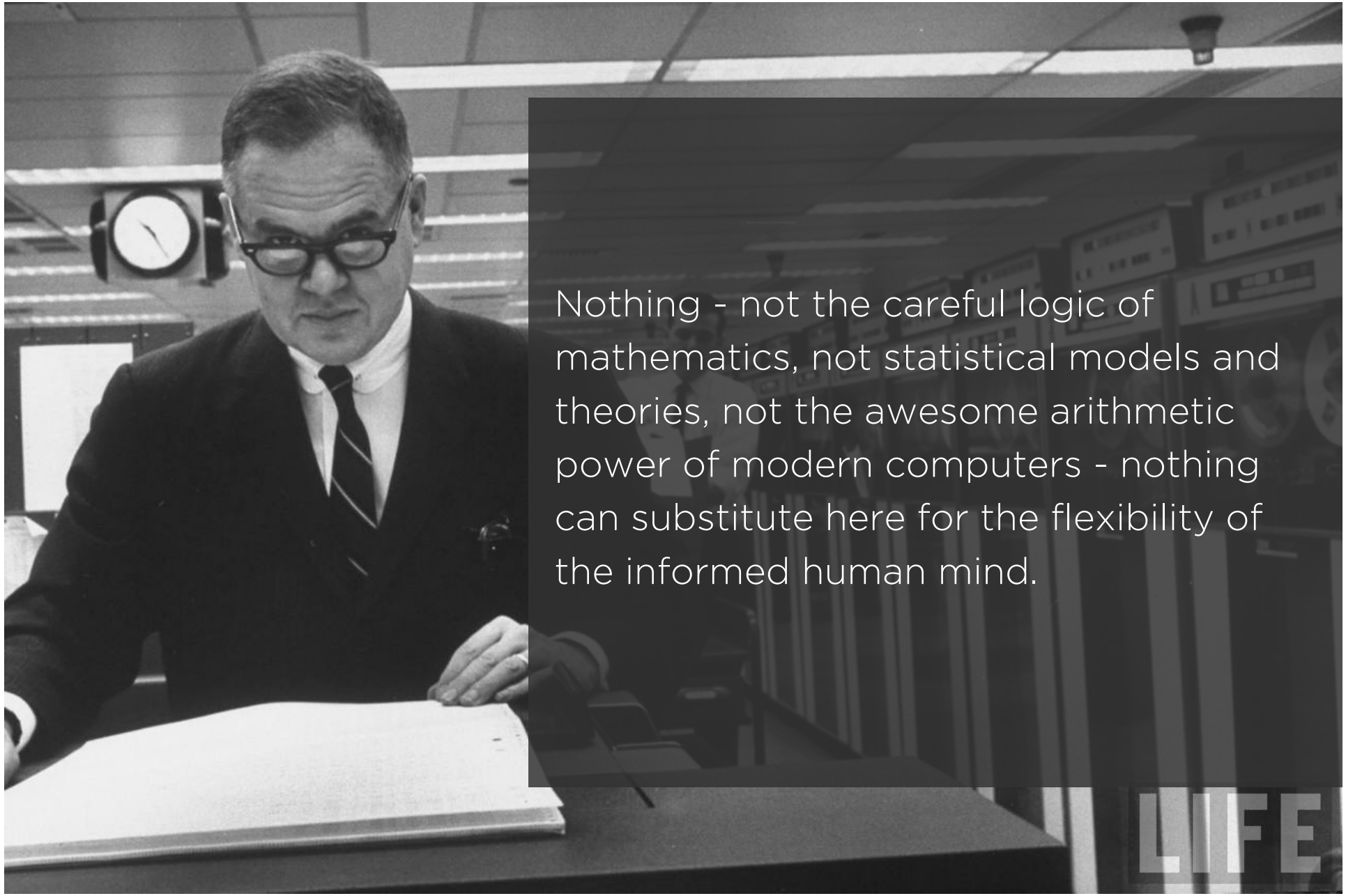
More on this in future lectures!

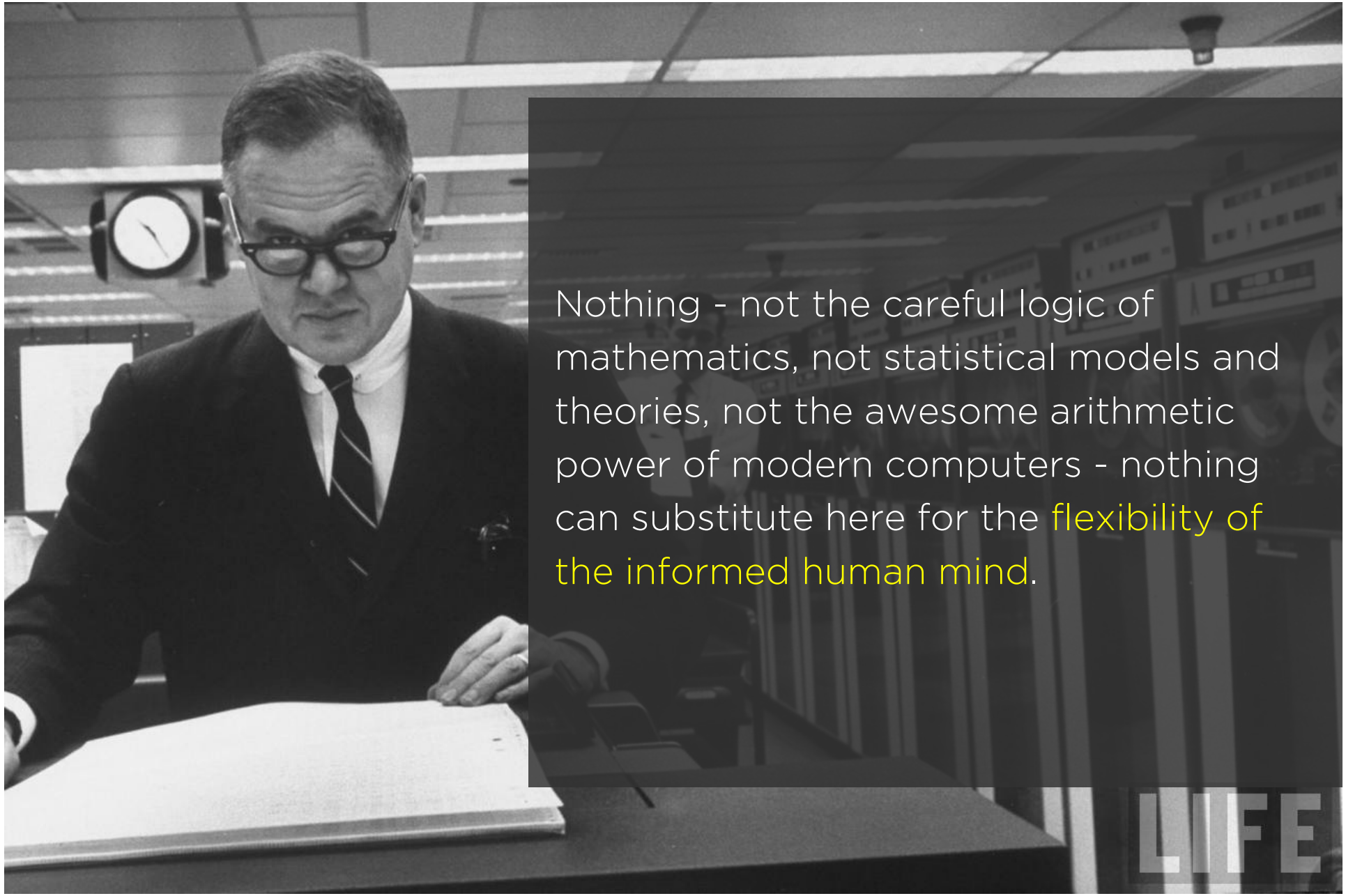
Data Analysis & Statistics, Tukey 1965


Some implications for effective data analysis are:
(1) that it is essential to have convenience of interaction of people and intermediate results and
(2) that at all stages of data analysis, the nature and detail of output, both actual and potential, need to be matched to the capabilities of the people who use it and want it.
LIFE



Some implications for effective data analysis are:
(1) that it is essential to have convenience of interaction of people and intermediate results and
(2) that at all stages of data analysis, the nature and detail of output, both actual and potential, need to be matched to the capabilities of the people who use it and want it.


Nothing - not the careful logic of mathematics, not statistical models and theories, not the awesome arithmetic power of modern computers - nothing can substitute here for the flexibility of the informed human mind.
LIFE


Nothing - not the careful logic of mathematics, not statistical models and theories, not the awesome arithmetic power of modern computers - nothing can substitute here for the flexibility of the informed human mind.
LIFE

Unboxing Data

- What do I have here?
- What do I want to do with it?

These questions rarely have pat answers.

- Typically *contextual* and *user-driven*
- Typically subject to *iterative* cycles of wrangling and analysis

Rough Guide to Wrangling Issues

an outline for today's lecture

- Structure: the "shape" of a data file
- Granularity: how fine/coarse is each datum
- Faithfulness: how well does the data capture "reality"
- Temporality: how is the data situated in time
- Scope: how (in)complete is the data

Many of these are ***subjective*** qualities! Depend on *context*.

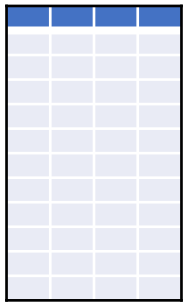
(Data) Science is a human process.

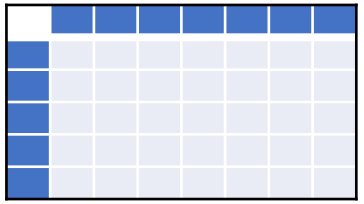
Structure: Rectangular Data

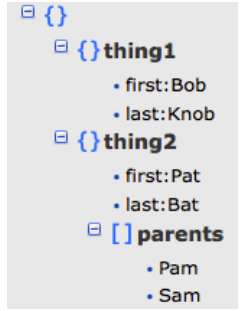
Natural for data analysis: easy to access, filter, tabulate

Two main variants

1. Relations (a.k.a. tables, data-frames)
   - Manipulate with Relational Algebra
2. Matrices
   - Manipulate with Linear Algebra

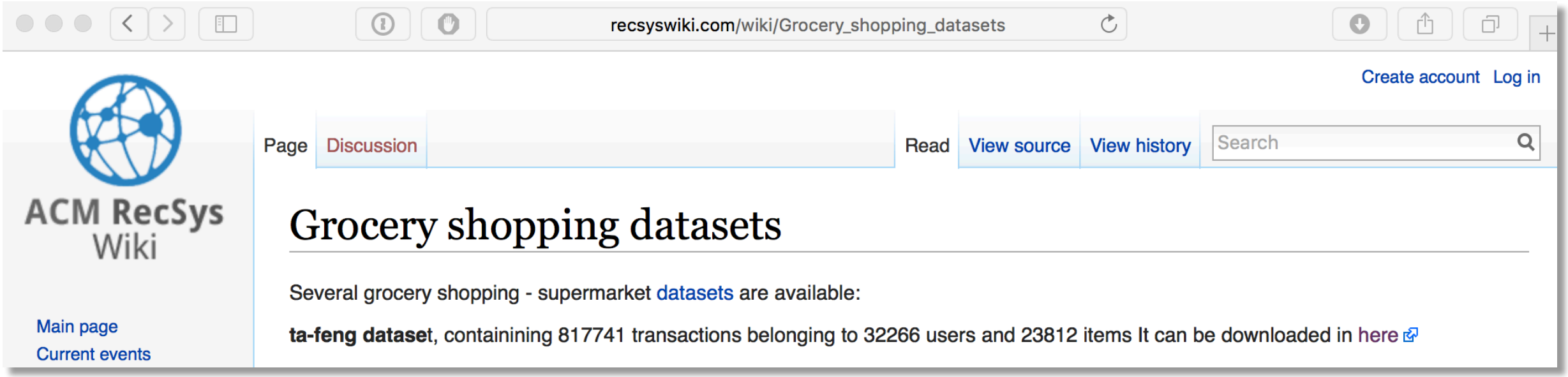






There are non-rectangular structures as well, e.g. JSON, XML.
For analysis, you will typically need to convert these to rectangular structures.

Without further ado, some data

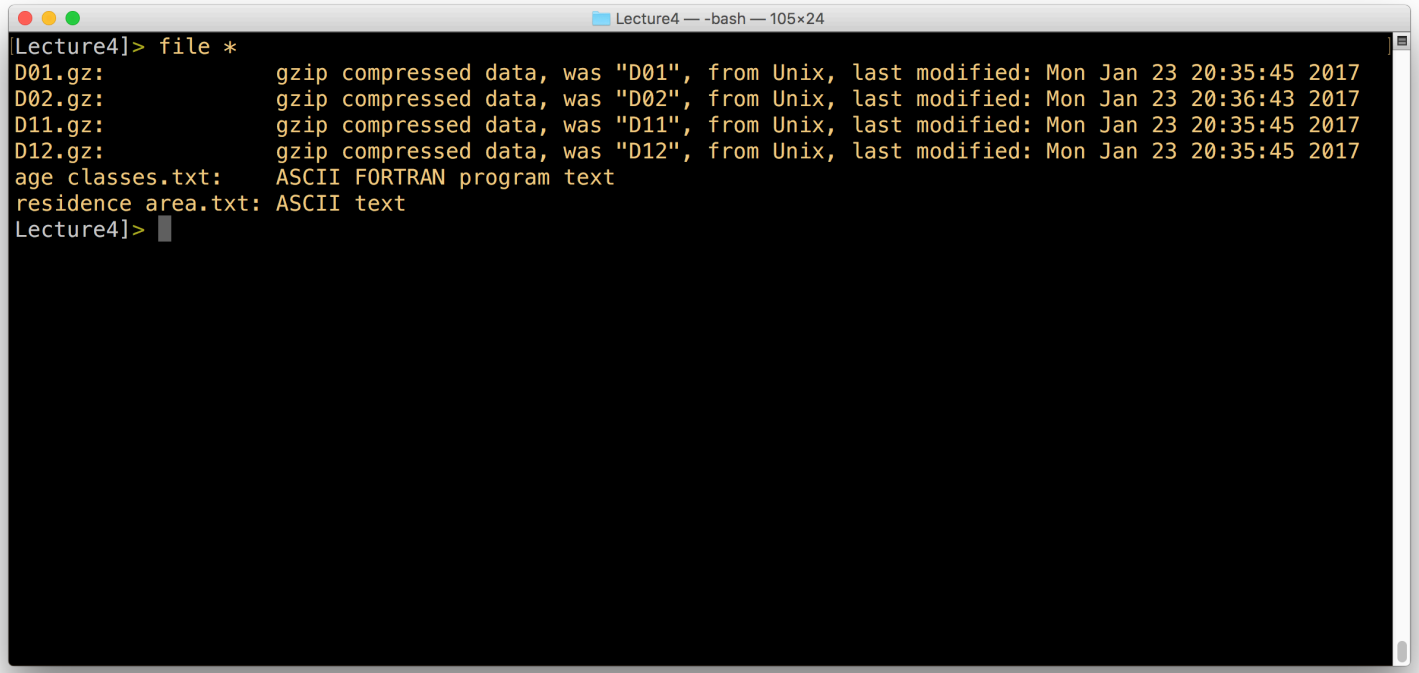


Gently modified for pedagogical purposes here.

Original dataset here.

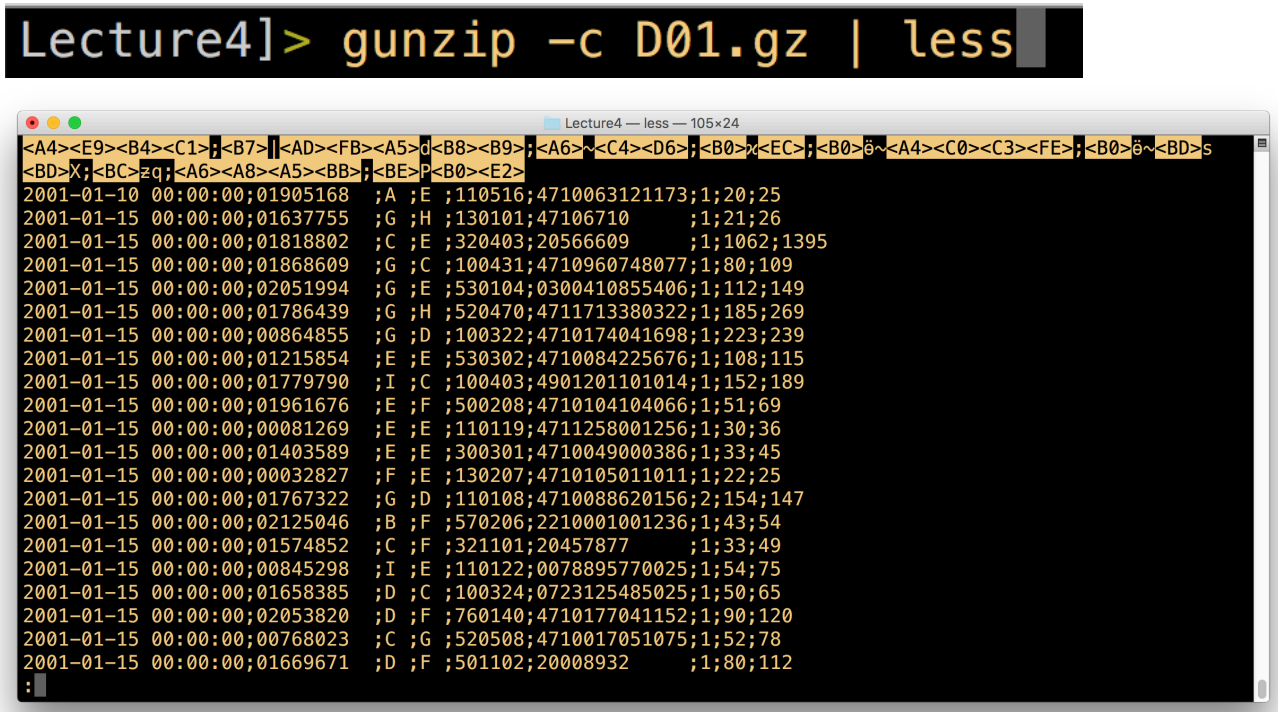
Unboxing with UNIX Command Line

- File metadata
  - `ls -lh`
  - `file`
  - `wc`
- File (de)compression
  - `gunzip`, `zip`, `bzip`, etc.
- `stdout` and the pipe
- File content:
  - `cat`
  - `head`
  - `tail`
  - `less`
  - <ctrl>-C



Structure Questions: A Checklist

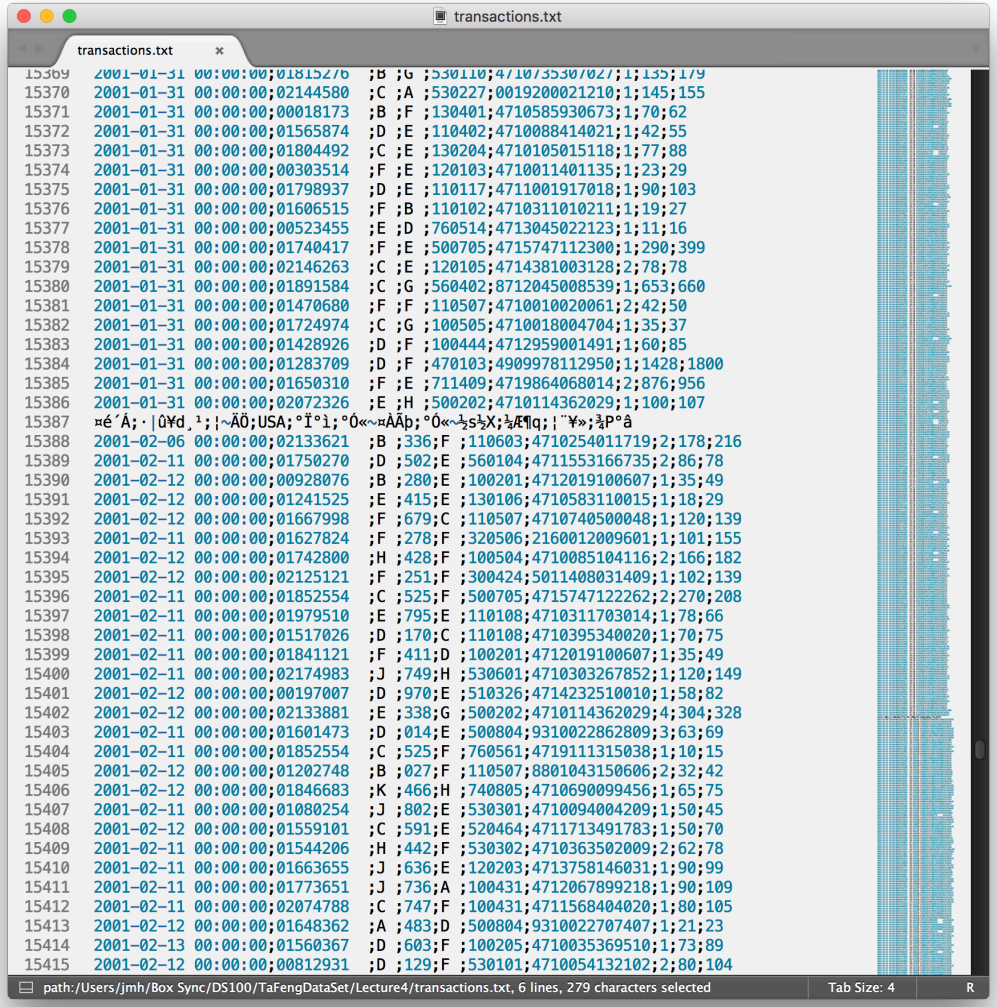
- Coarse structure
  - Is the data structured as a collection of *records*?
  - How are the individual records delimited in the dataset?
  - How are the record *fields* delimited from one another?
  - Do all records in the dataset contain the same fields?
  - How to access the same fields across records?
    By position? Name?
- Are the records nested?
  - No: one atomic (singular) value per field
  - Yes: collection of 0-to-many values in a field
- Encoding
  - How are values encoded? Strings? Codes? Binary?
  - How complex are the individual values?
    - Primitive: numbers and short strings?
    - Unstructured data: natural language text, audio, video
  - Can you decode?



Modern Text Editor: SublimeText

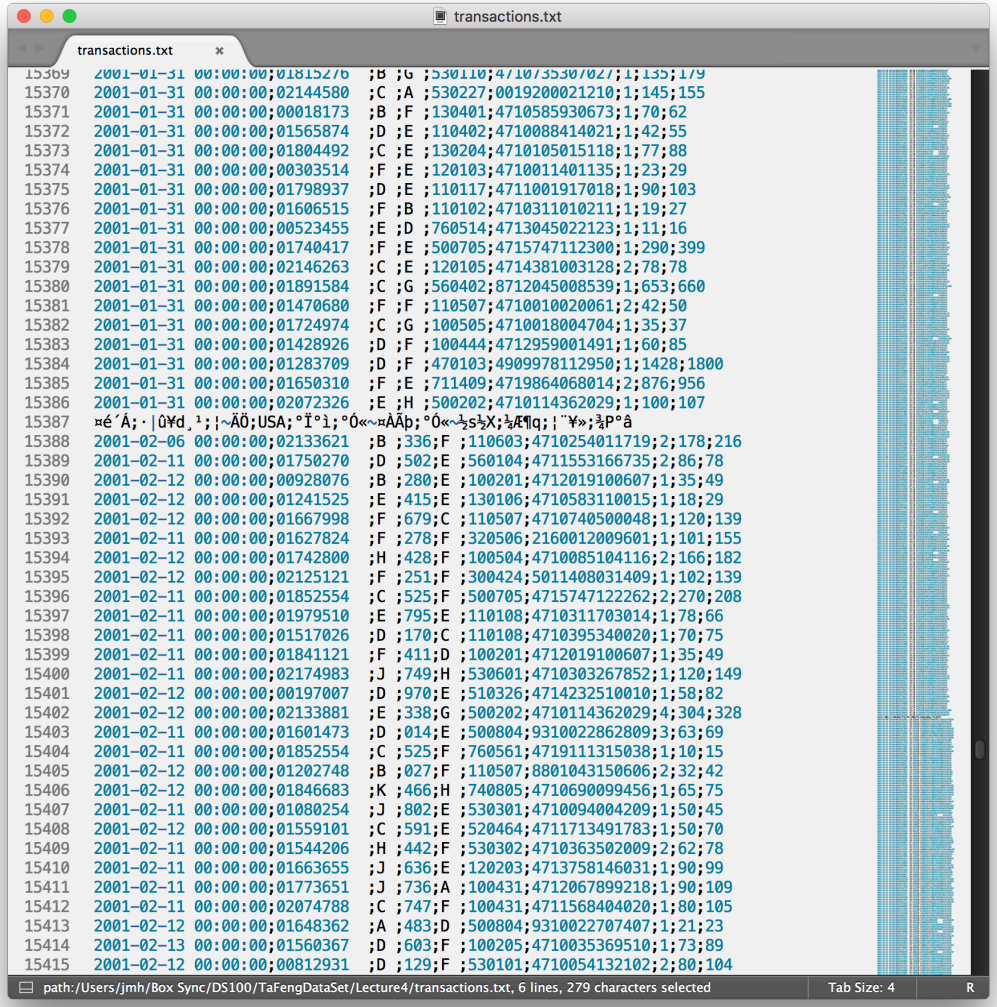
- Assessing
  - Coloring
  - Minimap
- Transformation
  - Do not destroy!
  - Names/Versions?

More on this soon!



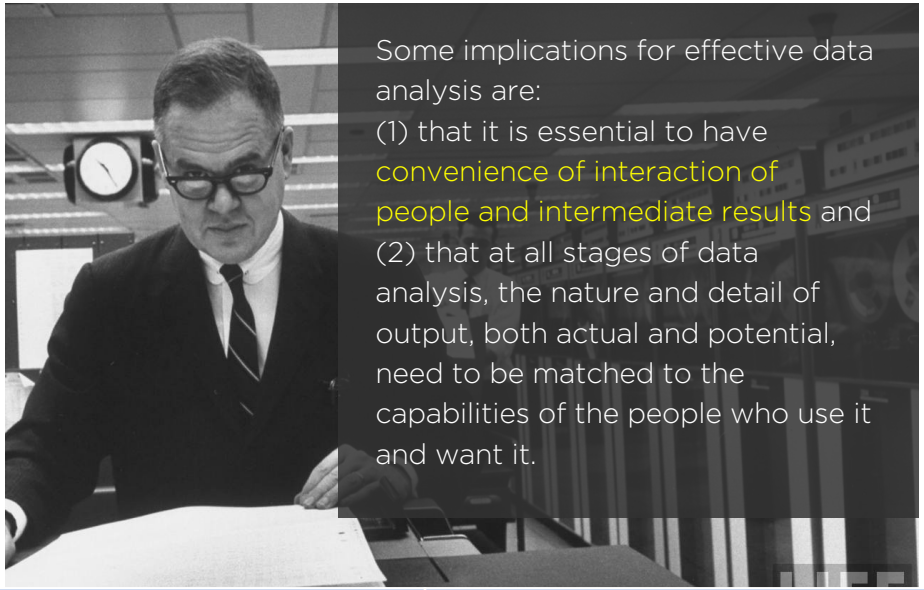
Structure Questions: A Checklist

- Coarse structure
  - Is the data structured as a collection of records?
  - How are the individual records delimited in the dataset?
  - How are the record *fields* delimited from one another?
  - Do all records in the dataset contain the same fields?
  - How to access the same fields across records?
    By position? Name?
- Is the data nested?
  - No: one atomic (singular) value per field
  - Yes: collection of 0-to-many values in a field
- Encoding
  - How are values encoded? Strings? Codes? Binary?
  - How complex are the individual values?
    - Primitive: numbers and short strings?
    - Unstructured data: natural language text, audio, video
  - Can you decode?



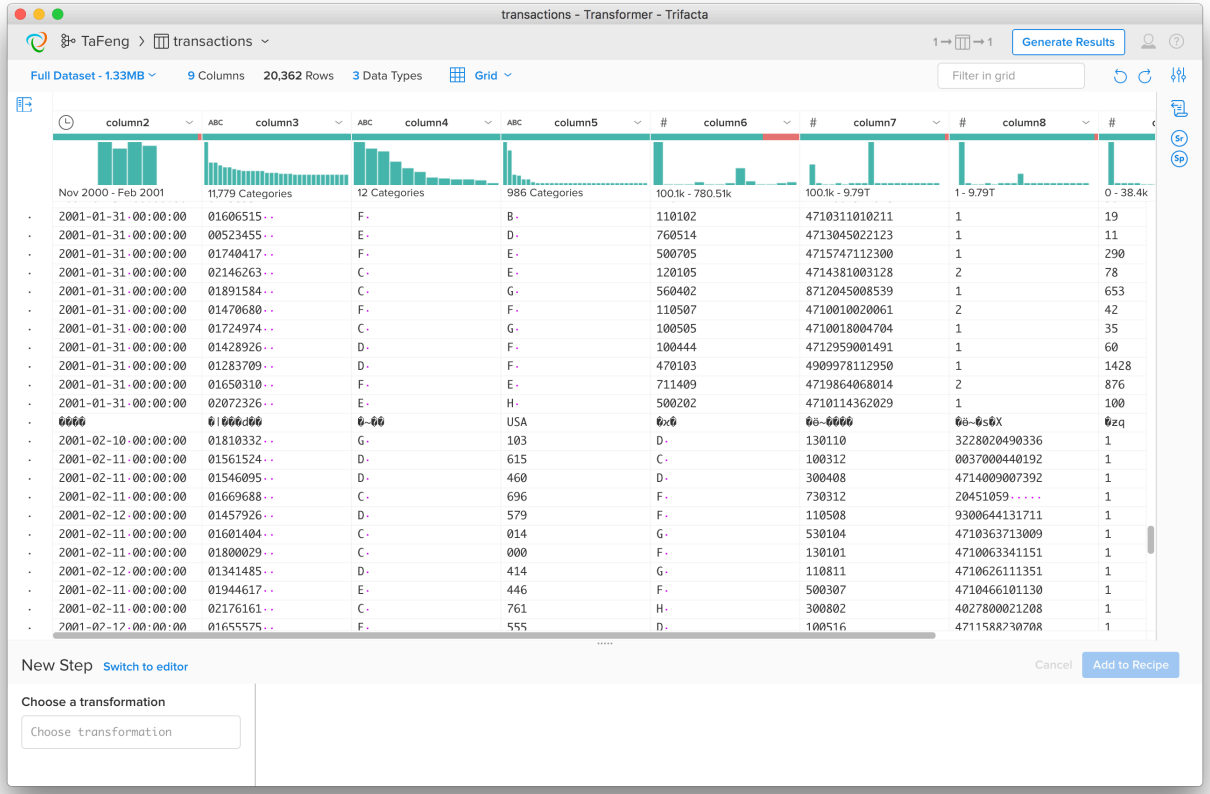
Unboxing with Trifacta

- Visual Profiling + Transformation
  - Iterative and ongoing
- Predictive Interaction
- Transformation language: Wrangle
  - Superset of relational algebra
  - Wrangle (and Pandas) both related to SQL
  - Generates a “recipe” (script)
    - You can run it on data to generate new data
- Scale & Interoperability
  - Paid version works with “big data”
    - Spark/Hadoop



Structure Questions: A Checklist

- Coarse structure
  - Is the data structured as a collection of records?
  - How are the individual records delimited in the dataset?
  - How are the record *fields* delimited from one another?
  - Do all records in the dataset contain the same fields?
  - How to access the same fields across records?
    By position? Name?
- Is the data nested?
  - No: one atomic (singular) value per field
  - Yes: collection of 0-to-many values in a field
- Encoding
  - How are values encoded? Strings? Codes? Binary?
  - How complex are the individual values?
    - Primitive: numbers and short strings?
    - Unstructured data: natural language text, audio, video
  - Can you decode?



So Far: Assessing Structure

- Basic tools
  - UNIX commands
  - Text editors
- Simple data transformation
  - Edit-in-place, character-by-character or Find/Replace
    - Save versions of files
  - Scripted transformations
    - Save scripts, inputs, outputs

Value Encodings: Primitive Types

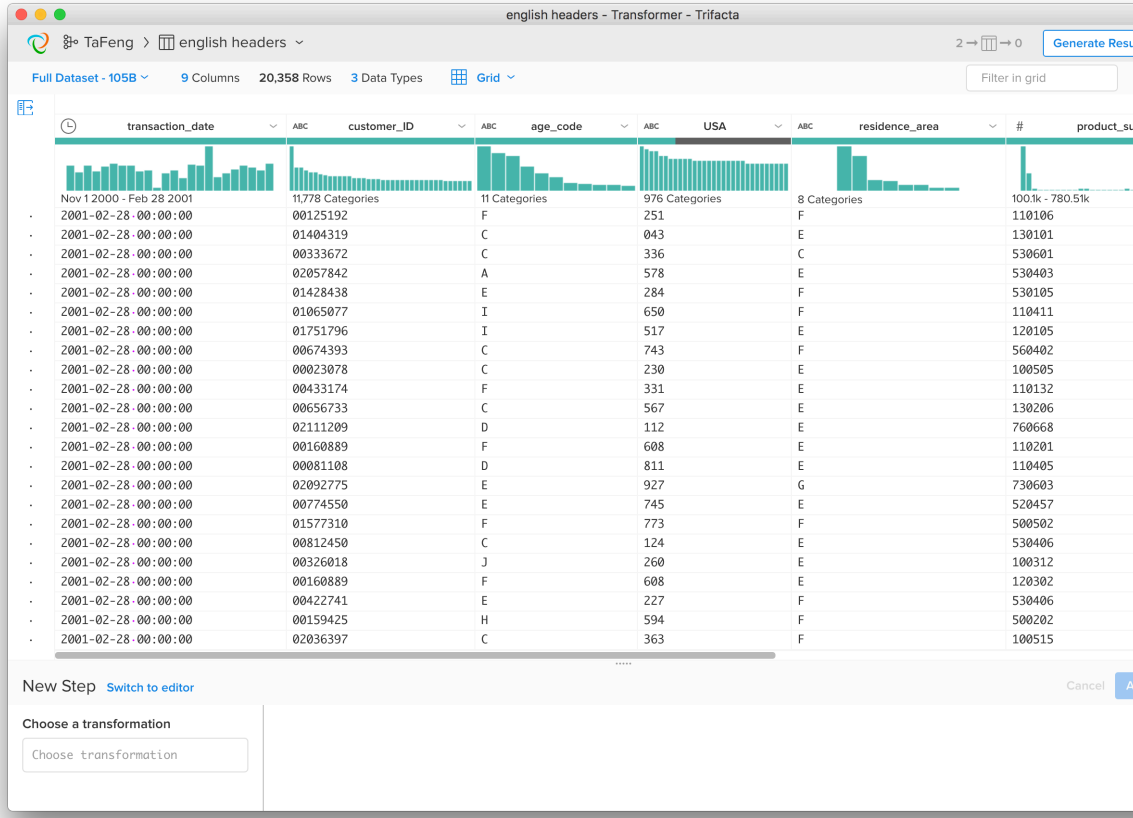
Numbers and Strings

We'll spend time here today

Common special-case "primitives"

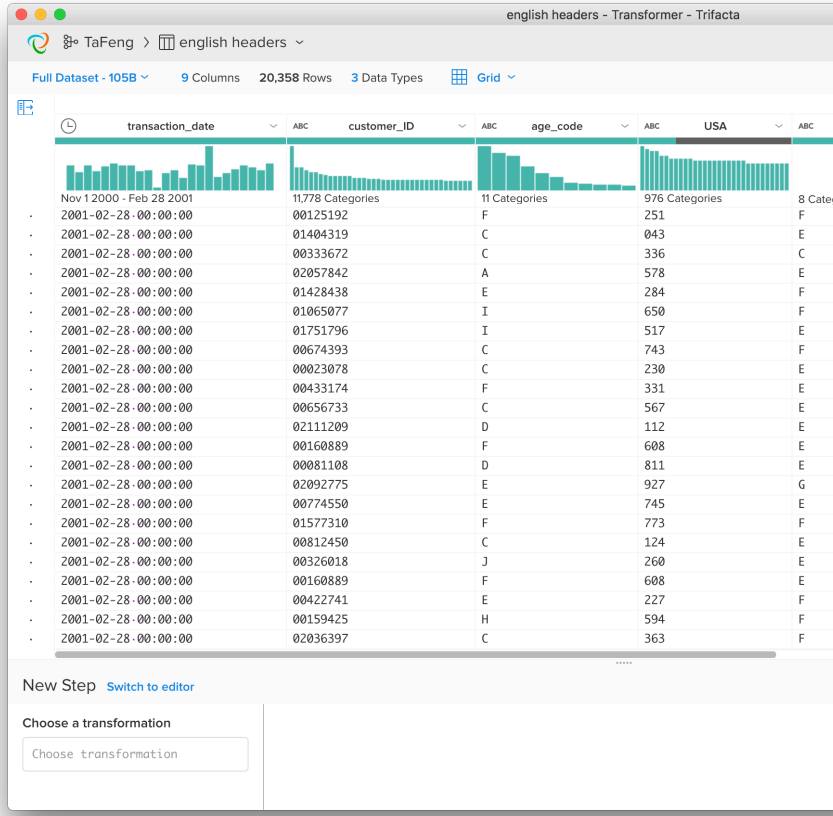
- Date/Time
- Geolocations

Space and time can be surprisingly subtle/messy!



Primitive Type Semantics

- Categorical #1: Nominal
  - E.g. political party affiliation
  - Note: Sometimes even numeric data is nominal
- Categorical #2: Ordinal
  - E.g. education level (none, HS, College, Post-College)
  - Ordered, but not particularly *quantitative*
- Quantitative: amounts, measures
  - Negative or positive
  - Arithmetic “makes sense” (e.g. difference, ratio)
  - Integers vs. Rational numbers



Who cares?

- Affects how you interpret the data
- E.g. In visualization or summarization

In Sum: Tools and Structure

Coarse structure

- Identifying data “shape”, records/fields and delimiters
- Value encodings
- Categorical, Ordinal, Quantitative

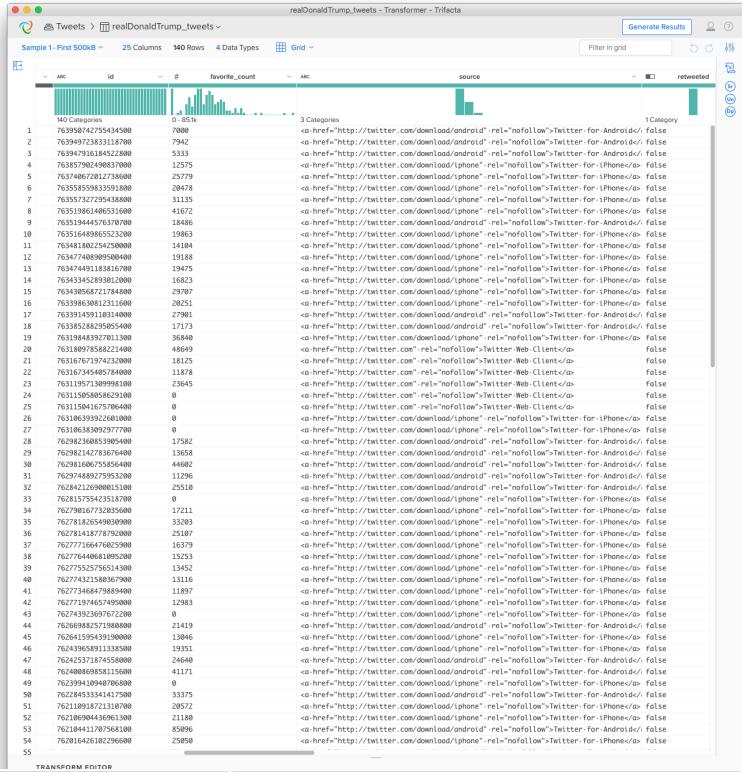
Basic tools

- UNIX commands
- Text editors

Simple data transformation

- Edit-in-place, character-by-character or Find/Replace
- Scripted transformations

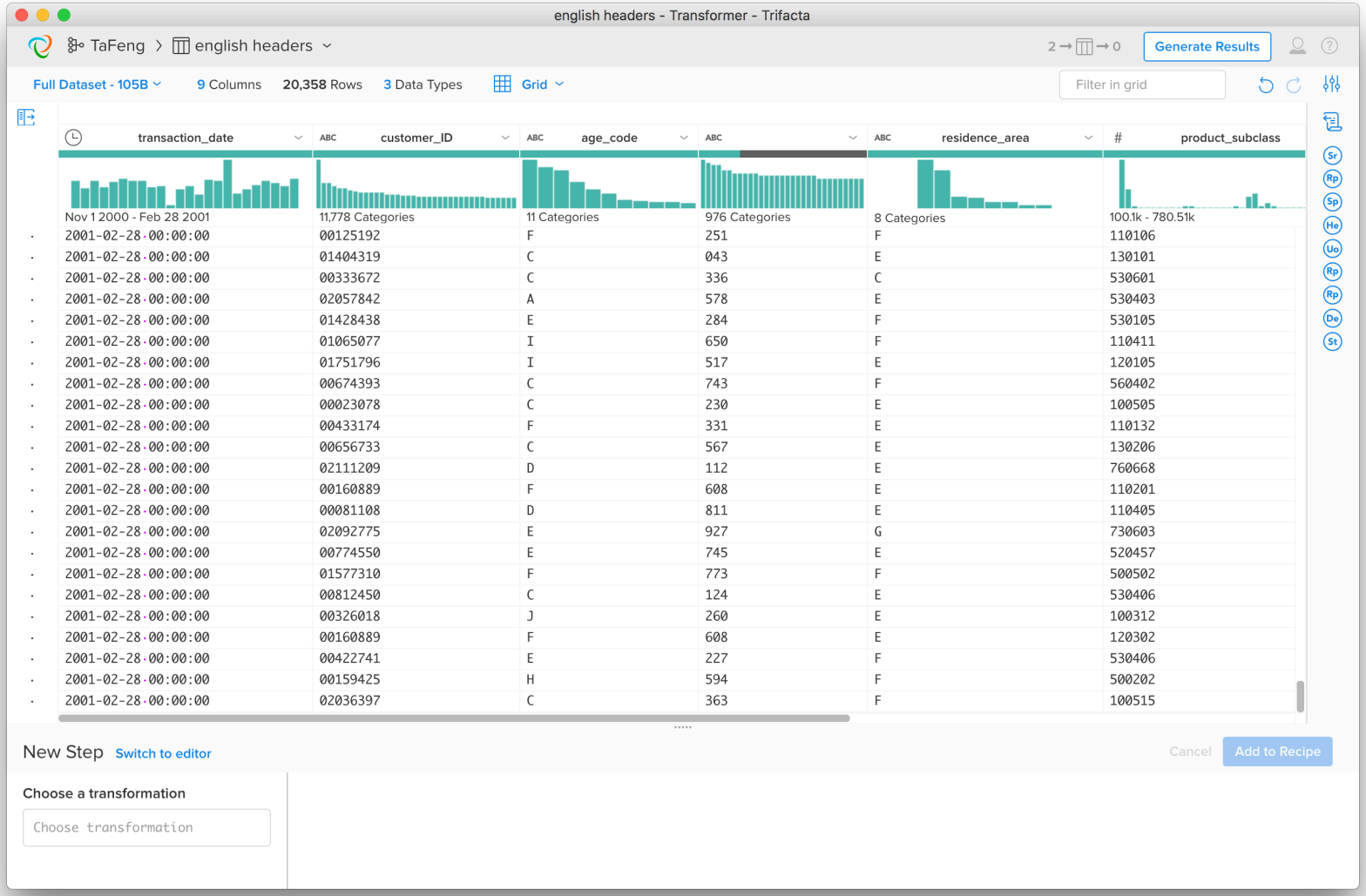
Assessing Granularity



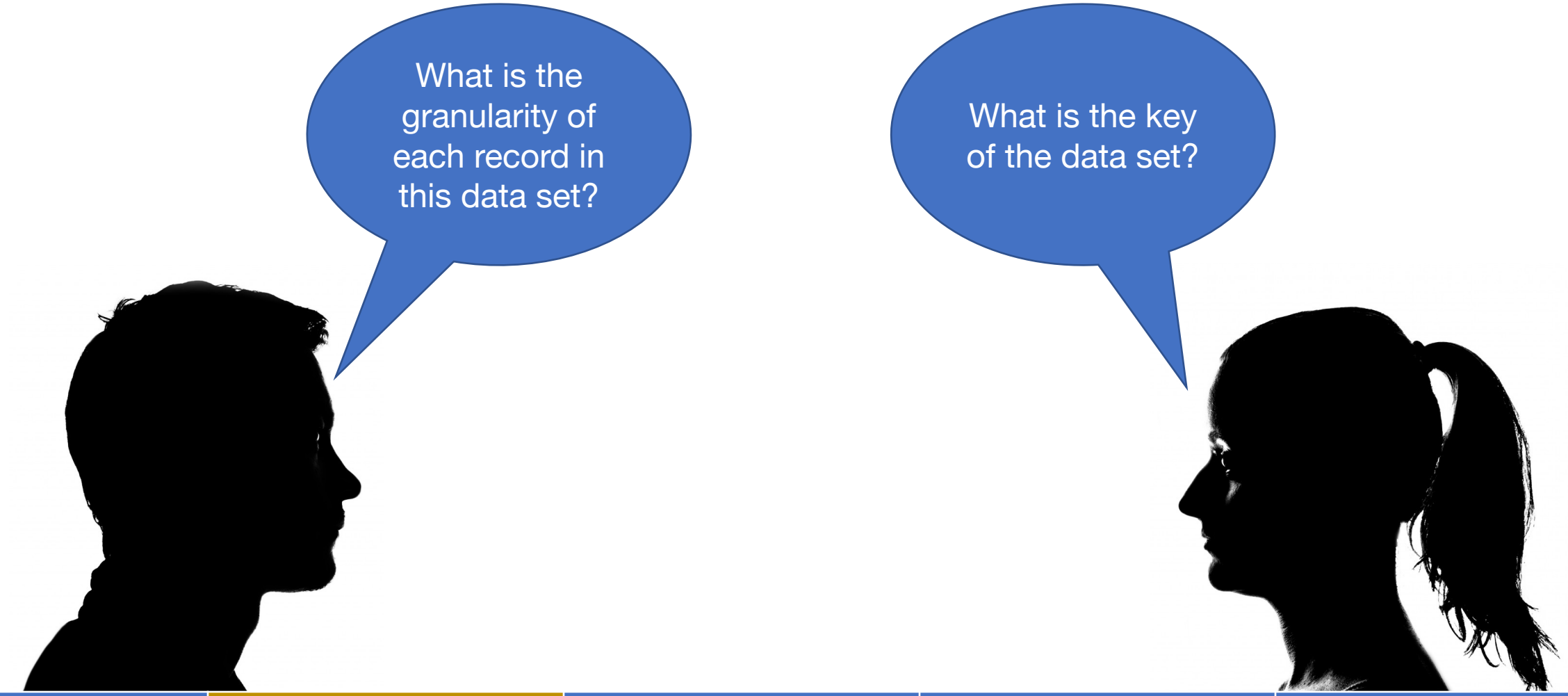
What does each TaFeng record represent?

Theories?

Justification?



Assessing Granularity



Key (“primary key”)

Given a relation, a key attribute uniquely determines the values in each record

- E.g. an *identifier* like `transaction_id`
- Can be *composite*: e.g. `(City, State)`

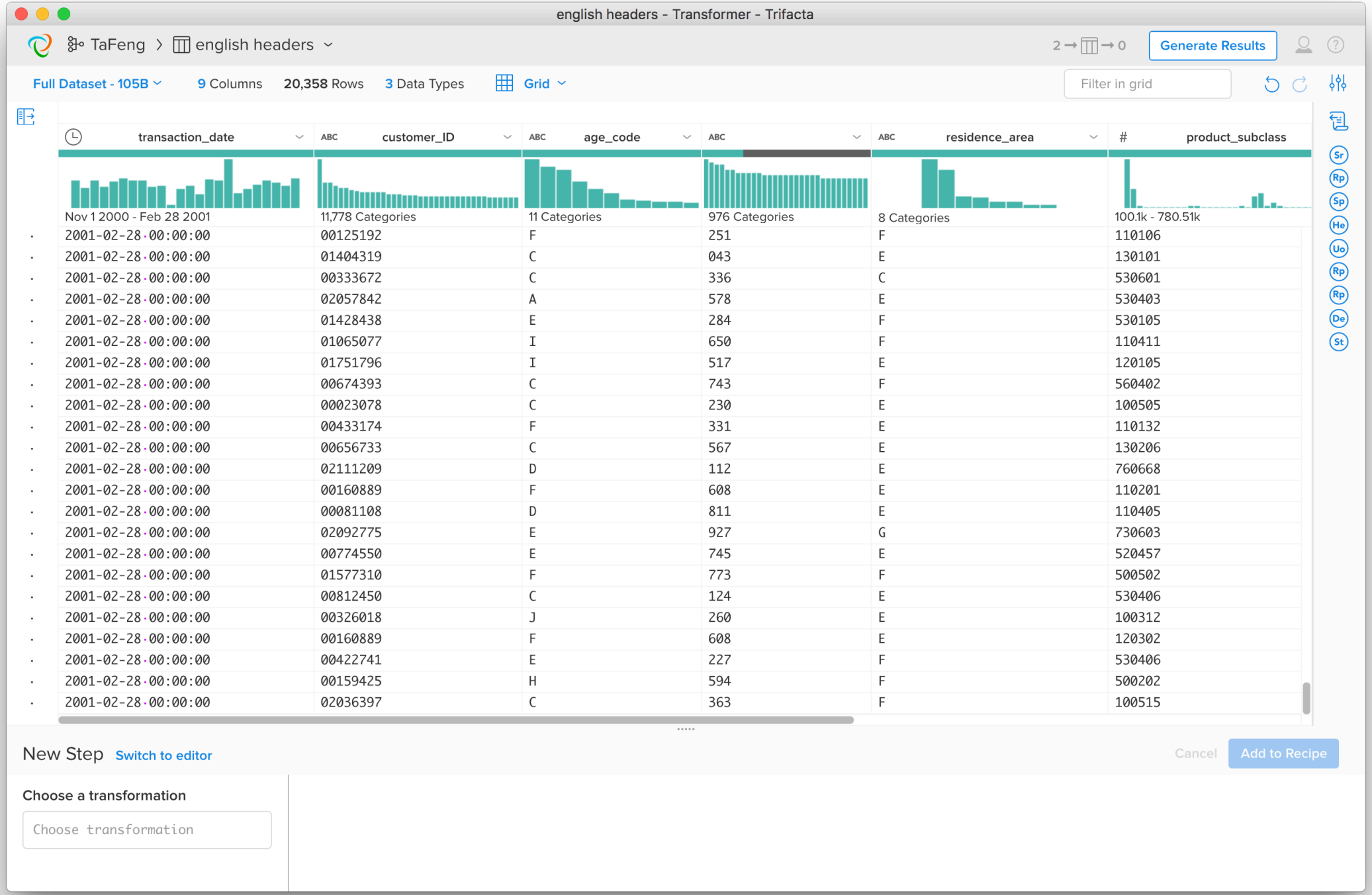
Each value occurs **at most once** in the key column(s).

Hence the semantics (meaning) of the key determines granularity

- `customer_id?` `transaction_time?`
- `(age_code, day_of_week)`

Be careful!

- Goofy key choice? Goofy data.
- Real-world data may have “noisy”, duplicated keys to clean up


english headers - Transformer - Trifacta
TaFeng
english headers
2 → 0
Generate Results
Full Dataset - 105B
9 Columns
20,358 Rows
3 Data Types
Grid
Filter in grid
transaction_date
customer_ID
age_code
residence_area
product_subclass
Nov 1 2000 - Feb 28 2001
11,778 Categories
11 Categories
976 Categories
8 Categories
100.1k - 780.51k
2001-02-28 00:00:00 00125192 F 251 F 110106
2001-02-28 00:00:00 01404319 C 043 E 130101
2001-02-28 00:00:00 00333672 C 336 C 530601
2001-02-28 00:00:00 02057842 A 578 E 530403
2001-02-28 00:00:00 01428438 E 284 F 530105
2001-02-28 00:00:00 01065077 I 650 F 110411
2001-02-28 00:00:00 01751796 I 517 E 120105
2001-02-28 00:00:00 00674393 C 743 F 560402
2001-02-28 00:00:00 00023078 C 230 E 100505
2001-02-28 00:00:00 00433174 F 331 E 110132
2001-02-28 00:00:00 00656733 C 567 E 130206
2001-02-28 00:00:00 02111209 D 112 E 760668
2001-02-28 00:00:00 00160889 F 608 E 110201
2001-02-28 00:00:00 00081108 D 811 E 110405
2001-02-28 00:00:00 02092775 E 927 G 730603
2001-02-28 00:00:00 00774550 E 745 E 520457
2001-02-28 00:00:00 01577310 F 773 F 500502
2001-02-28 00:00:00 00812450 C 124 E 530406
2001-02-28 00:00:00 00326018 J 260 E 100312
2001-02-28 00:00:00 00160889 F 608 E 120302
2001-02-28 00:00:00 00422741 E 227 F 530406
2001-02-28 00:00:00 00159425 H 594 F 500202
2001-02-28 00:00:00 02036397 C 363 F 100515
New Step
Switch to editor
Cancel
Add to Recipe
Choose a transformation
Choose transformation

Joining tables with keys

- Keys in our data
  - `age classes.txt`: `code`
  - `residence area.txt`: `code`
  - `transactions.txt`: `transaction_id`

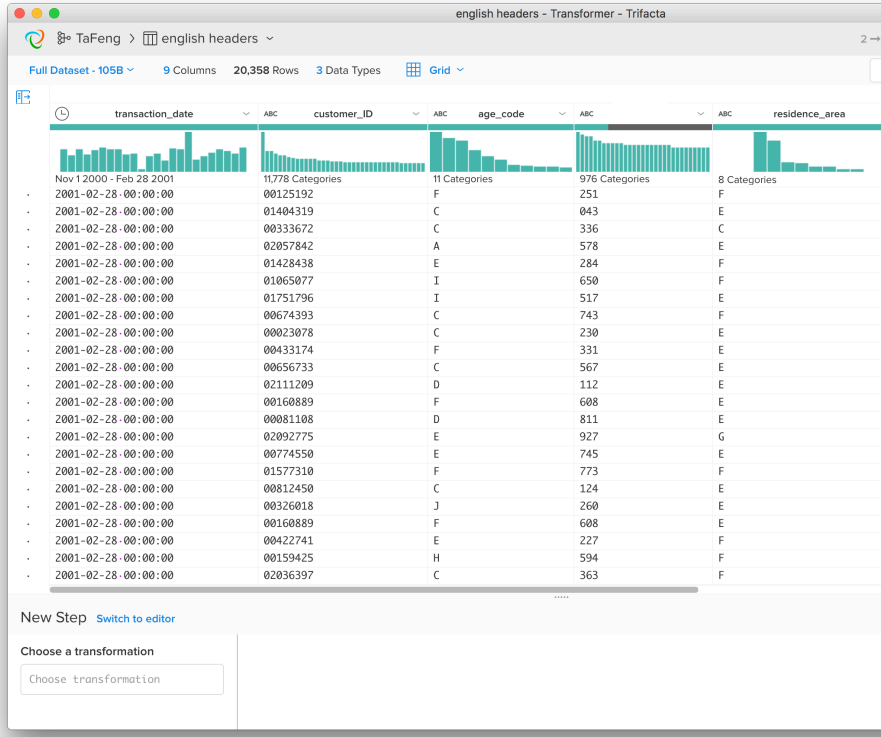
- Transactions.txt also has *foreign keys*
  - i.e. attributes that reference the key of another table
  - `age_code` is a foreign key to `age classes.txt`
  - `residence_area` is a foreign key to `residence area.txt`

- Joining on a foreign key is a “lookup”
  - At most one match for each row of `TaFengTransactions.txt`

More on this in future lectures!

Granularity Questions: a Checklist

- What kind of thing does each record represent?
- Do all records capture granularity at the same level?
- What alternative interpretations of the records are there?
  - E.g. Is this a list of transactions?
    Or a list of customers?
- What kinds of aggregation is possible/desirable?
  - From individual people to demographic groups?
  - From individual events to totals across time or regions?
  - Hierarchies (city/county/state, second/minute/hour/day)



So Far: Assessing Granularity

Identifying (primary) keys can help assess Granularity

- Each record is the information *per key*

Foreign Keys are pointers to individual rows in other data sets

- To lookup a row in another table: join the foreign key to its primary key

Assessing Faithfulness

Theme: the Faithfulness of a record can only be evaluated in context

- Application context
- Context in your data set
  - Across records

Students

id: integer	DOB: date	GPA: float	Risk: float
123457	01/16/1997	3.2	465
123458	01/24/2017	2.7	28
123457	01/16/2002	5.0	27
123459	03/22/1996	3.6	31
123460	06/13/1997	2.2	43

Faithfulness Across Records: Outliers

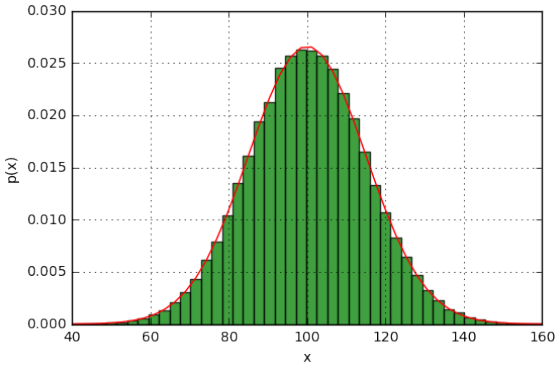
What is an “outlier”?

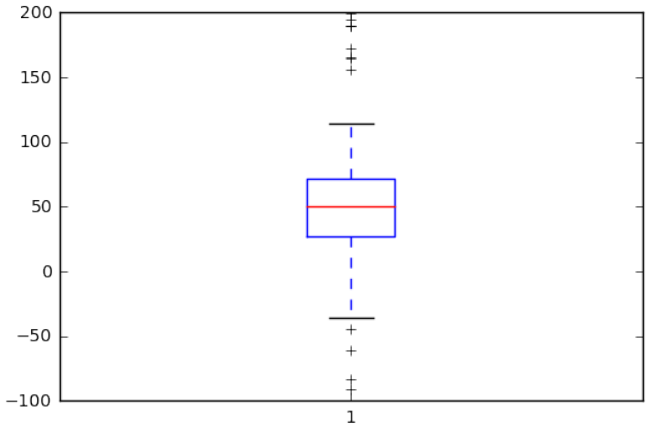
More on this in future lectures!

- A value that is “far” from the “center”

Distribution-based definition

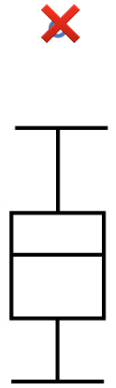
- Center (e.g. average, median)
- Spread (e.g. standard deviation, IQR)





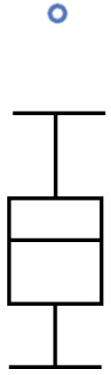
What to do with Outliers?

Delete. (“trimming”)



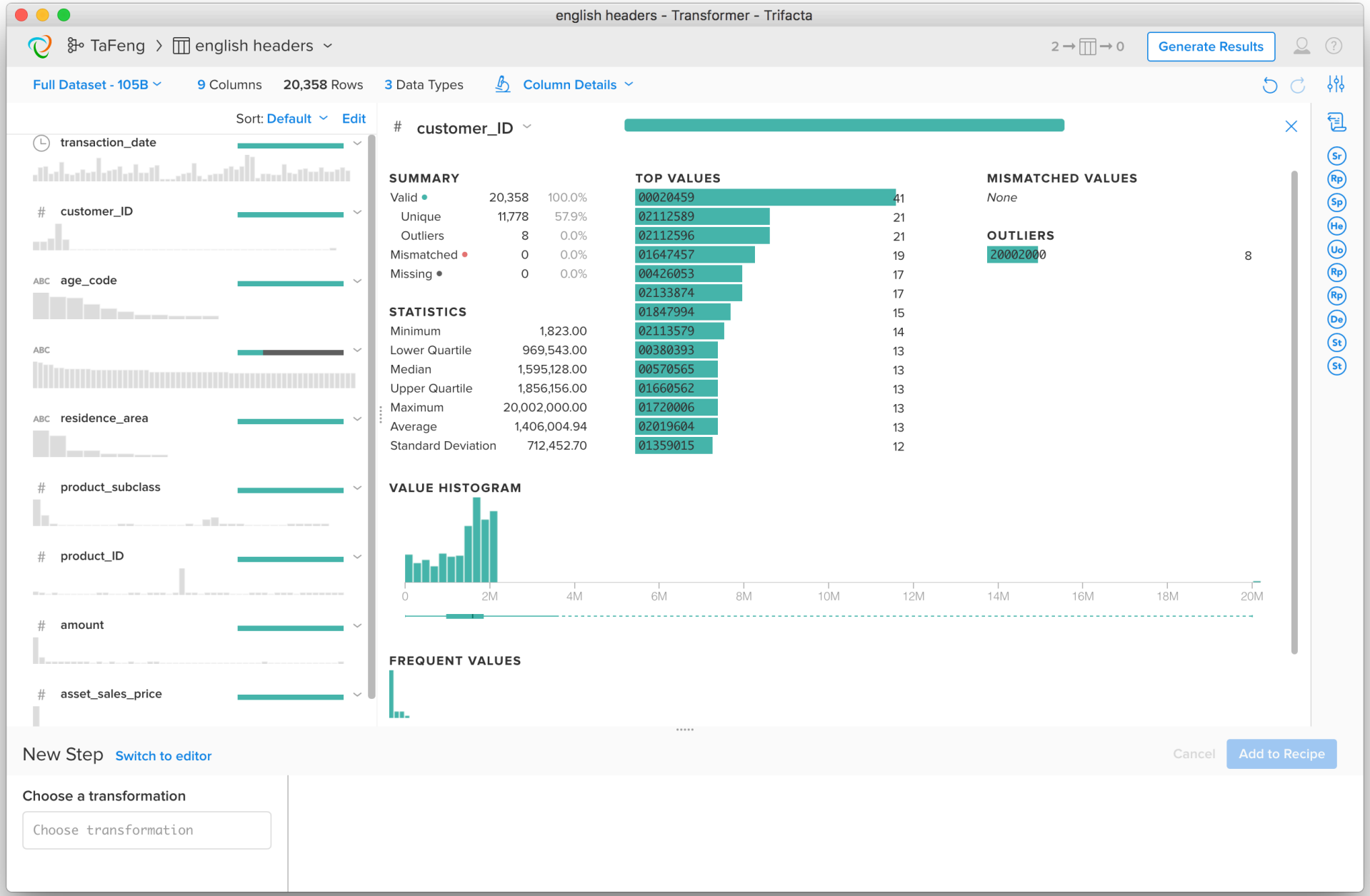
Set to a default

- E.g. the nearest non-outlier (“Winsorizing”)

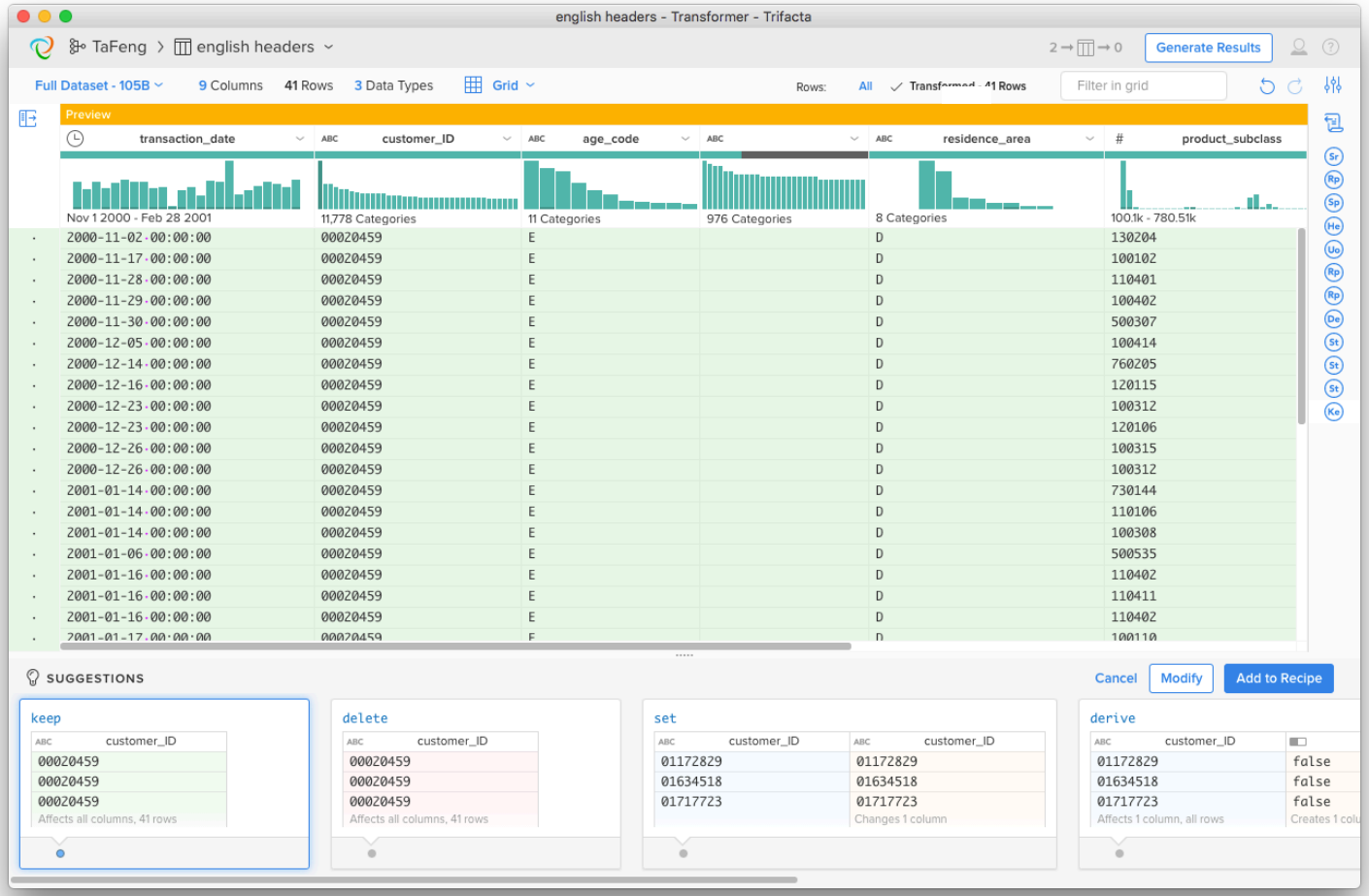


Good Hygiene:

- Leave the original column
- Derive an indicator column to flag presence of outlier
- Derive a clean column for your use


english headers - Transformer - Trifacta
TaFeng
english headers
Generate Results
Full Dataset - 105B
9 Columns
20,358 Rows
3 Data Types
Column Details
Sort: Default
Edit
transaction_date
customer_ID
age_code
residence_area
product_subclass
product_ID
amount
asset_sales_price
customer_ID
SUMMARY
Valid 20,358 100.0%
Unique 11,778 57.9%
Outliers 8 0.0%
Mismatched 0 0.0%
Missing 0 0.0%
STATISTICS
Minimum 1,823.00
Lower Quartile 969,543.00
Median 1,595,128.00
Upper Quartile 1,856,156.00
Maximum 20,002,000.00
Average 1,406,004.94
Standard Deviation 712,452.70
TOP VALUES
00020459 41
02112589 21
02112596 21
01647457 19
00426053 17
02133874 17
01847994 15
02113579 14
00380393 13
00570565 13
01660562 13
01720006 13
02019604 13
01359015 12
MISMATCHED VALUES
None
OUTLIERS
20002000 8
VALUE HISTOGRAM
0
2M
4M
6M
8M
10M
12M
14M
16M
18M
20M
FREQUENT VALUES
New Step
Switch to editor
Cancel
Add to Recipe
Choose a transformation
Choose transformation

Assessing Faithfulness Within Records: Dependencies and Correlations



Functional Dependencies (FDs)

More on this in future!

- Generalization of Keys
- Attribute A *determines* Attribute B
  - `customer_id -> age_code`
  - i.e. age_code = f(customer_id)
    - Function represented as lookup table
- More generally, a set of columns determines another set of columns
  - `{transaction_time,customer_id} -> {age_code, residence_area}`
- Primary Keys are special FDs
  - Right-hand-size is the set of *all* attributes in the relation

Correlations

More on this in future!

Dependence (i.e. lack of independence!) between 2 random variables

Think of the attributes in a relational schema

- An *instance* of that relation was generated from some real-world process
- Each column of that relation is a “random variable” generated by the process

A Functional Dependency is a “deterministic” correlation

Correlations are more general: statistical relationships

- `amount` and `sales_price` are correlated

What to do about bad FDs/Correlations

- Cleaning Noisy FDs: `customer_id -> age_code` (kinda)
  - For a few customer IDs, there are multiple values of age_code
  - Set offending right-hand-side values to all match
  - Set offending left-hand-side values to NULL
- Cleaning Correlations: `height` correlated with `weight`
  - Some rows don't seem to follow the correlation (how do we decide?)
  - Can *impute* a likely value for one side or the other
  - Can set one side or the other to NULL
- Don't forget Good Hygiene!
  - Leave the original column
  - Derive new columns (indicators and/or cleaned data)

Faithfulness Questions: A Checklist

- Type-specific Faithfulness checks
  - Are dates and addresses legal/reasonable?
  - Numeric codes legal?: E.g. phone numbers, credit cards, social-security numbers etc.
  - Can you validate network endpoints? Email addresses, IP addresses, social network names
  - Need to deduplicate named entities: misspellings, acronyms (UCB vs Berkeley)
- Checks for data entry problems
  - Frequency outliers for common default entry values (00000, 1234567)
  - Misspellings (compare to dictionaries)
  - Sensor drift – often timeseries-based
  - “Curbstoning” in surveys
- Quantitative dirty data
  - Outliers, FDs, Correlations
- Check distribution of inaccuracies
  - Do inaccuracies seem to affect a large fraction of records?
  - Are they concentrated in a particular subset of records?

Tricky!

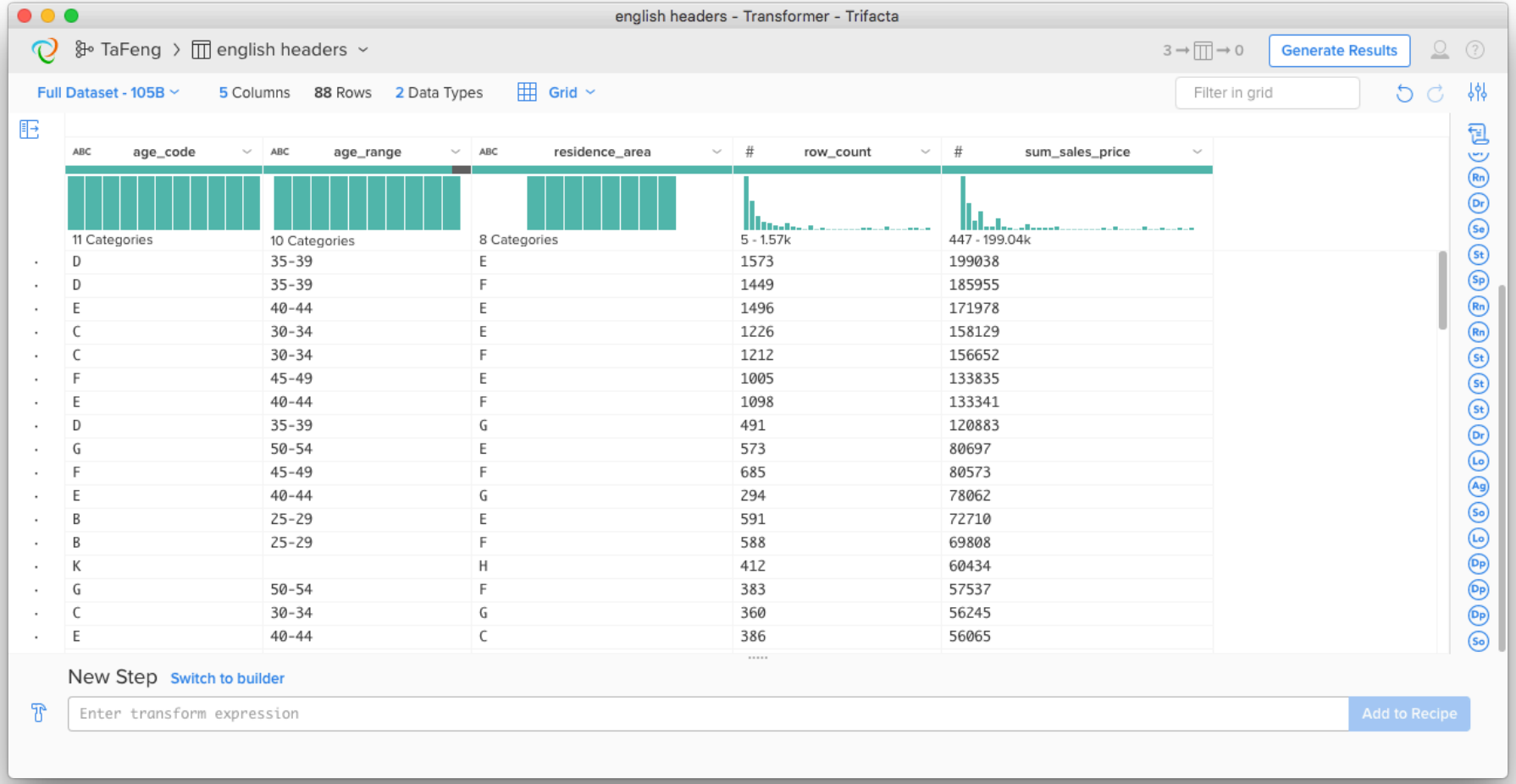
Summing Up: Faithfulness

- Outliers
- Functional Dependencies & Correlations
- Good Data Cleaning Hygiene
  - Don’t overwrite: use indicators and derived columns

Granularity *Transformations*

- We can coarsen the granularity by picking new keys and “rolling up”
  - GroupBy and Aggregation
- GroupBy
  - Choose a new primary key (the group-by columns)
  - Result will have one row per distinct values of this primary key
- Aggregation
  - Summary (rollup) results per group
  - E.g. count(), or aggregation functions on attributes (sum(x), average(x), stdev(x) etc.)

Finalizing Ta Feng



What About Non-Rectangular Data

- Nested Data
- Free Text (i.e. for humans)
- Example that has both: Twitter feed

Unnesting Nested Data Types

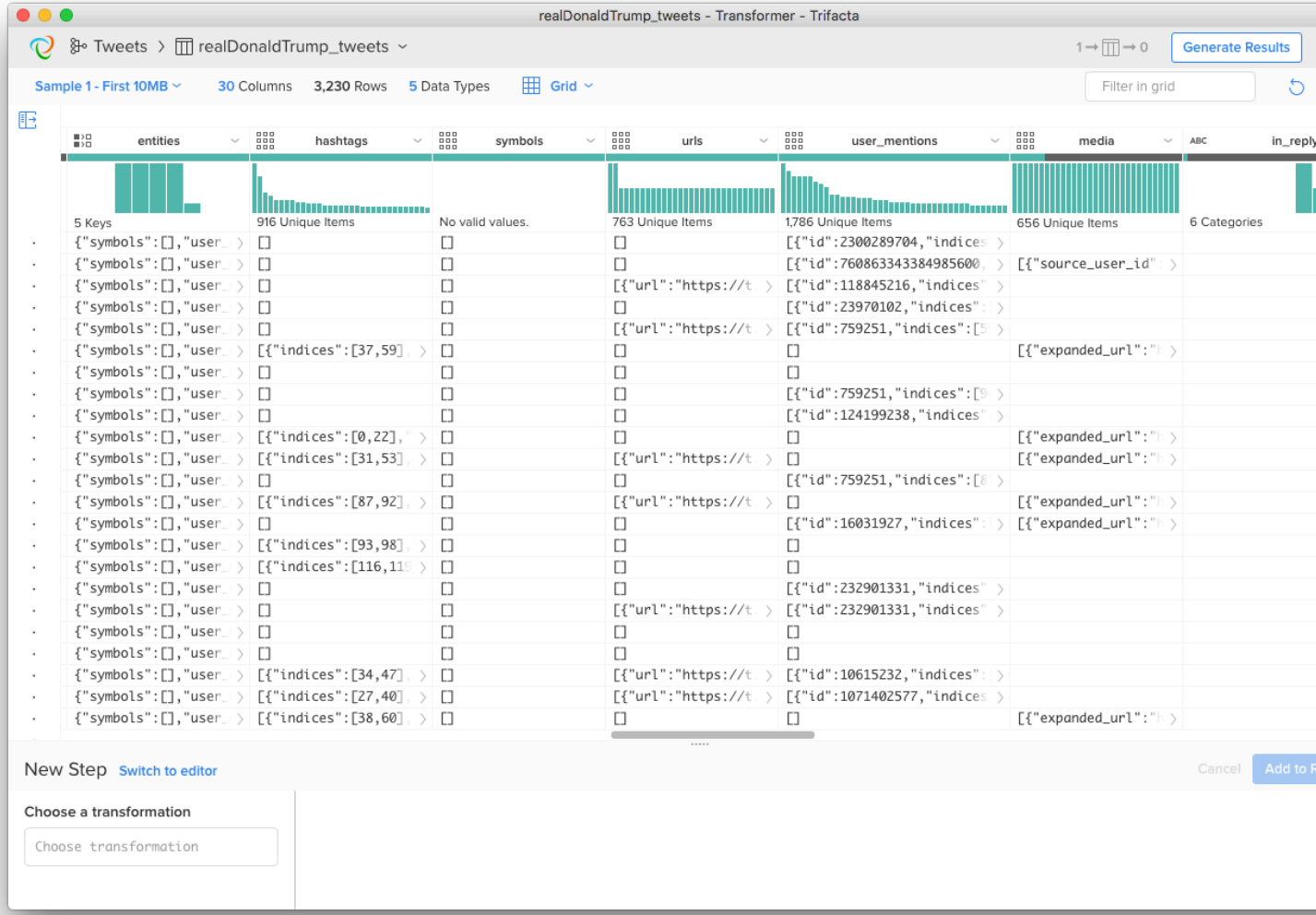
More on this in future lectures!

Maps

- A.k.a. dictionaries, hashes
- A set of key:val pairs

Arrays

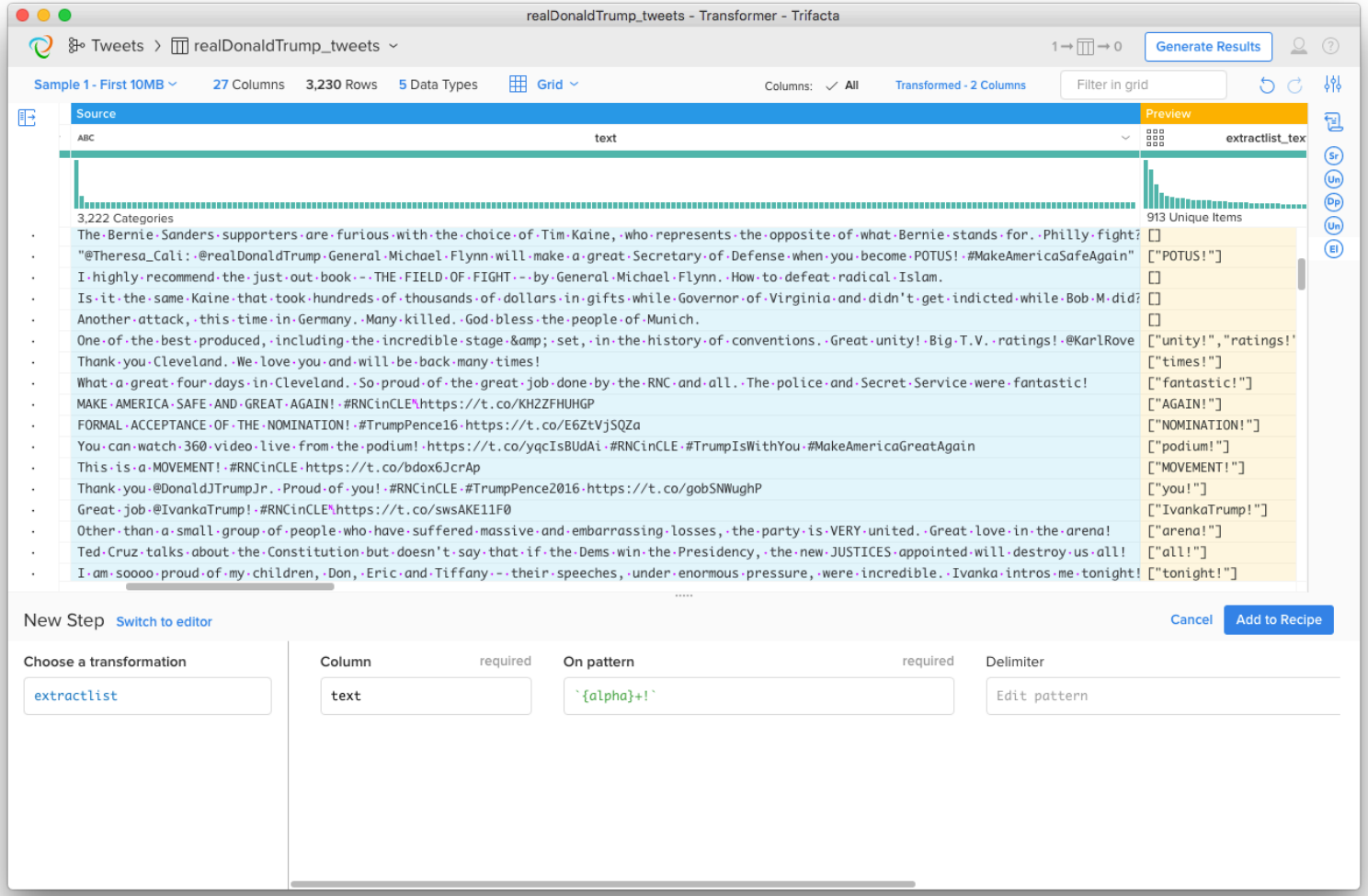
- Or lists



More on this in future lectures!

Playing with Text I: String Manipulation

Regular expression (regex) extraction



Playing with Text II: Natural Language

Natural Language Processing: a slippery slope

- Entity Resolution
  - IBM vs. International Business machines
- Named Entity Recognition
  - Steve Jobs munched on an apple as he announced new jobs at Apple.
- Sentiment
  - My windows box crashed yet again.
- Etc. Etc.
- Hot topic today: Dialog systems:
  - Alexa, where did I leave my keys?
    - I don't know. It's a pity you don't have a tracker on your key fob.
  - A what?
    - A bluetooth tracker – there are variety of them for sale
  - Stop trying to get me to buy stuff!
    - Sorry about that; I know you've been trying to save this month...
    - …but if you had one I could locate your keys.

Playing with Text II: Natural Language

- Some simple things
  - Term frequencies
  - Simple Sentiment analysis

Quick Note: Assessing Temporality

- Often two kinds of time in data
  - Time of data entry
  - Time of a recorded phenomenon being “true”
    - E.g. A physical time of an event happening
    - E.g. An “effective” time, e.g. date that a subscription will start
- Often more
- Time is tricky!
  - Periodicities (recurring patterns in Days of the week, or Months of the year)
  - Non-uniform hierarchy of units (# days in a month, # of days in a year, etc.)
  - Time zones are complex: especially daylight savings (summer) time
  - Clocks can be skewed
  - Relativity: true perception of event may vary (yep!)
- Assess timestamps in data carefully!!

Quick Note: Assessing Scope

- Do you have all the data you need
  - Missing columns
    - Join in external data
  - Missing rows/values?
    - Look for sequential patterns with breaks
    - Is this a good sample?
      - Granularity
      - Extent
    - How to *impute* reasonable stand-in values

More on this in future lectures!

- Can be quite application specific/subjective!

Looking Back: Data Transformations

Modifying structure

- Splitting rows and columns (“splitrows”, “split”)
- Unnesting (“flatten”)

Hiding Things

- Rows (“keep”, “delete” a.k.a. “select”)
- Columns (“drop”, “aggregate”)

Adding Things

- “derive” a.k.a. “map”, “apply”

Changing Things

- “replace”/”set”

Text manipulation

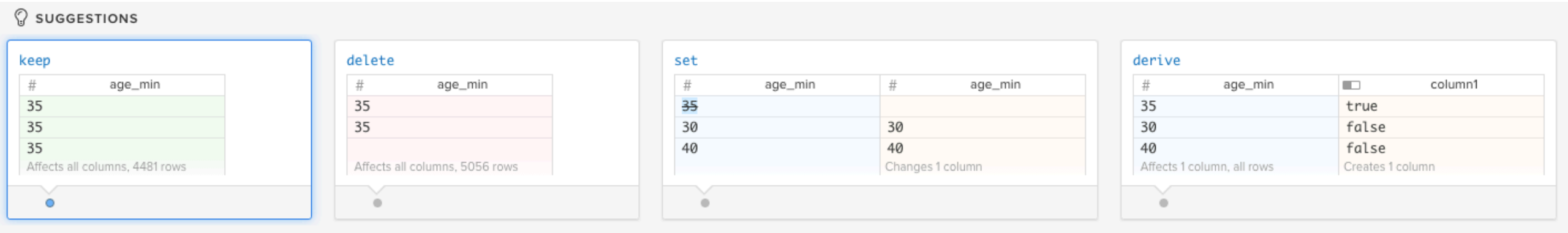
- “extract”, “replace”, “set”

GroupBy/Aggregate arithmetic

- “aggregate”, a.k.a. “group by”, “reduce”

Multi-table Operations

- “join”, “lookup”
- “union”



Looking Back: Bigger Picture

*What **decisions** did we make along the way?*

- Data that got hidden
  - Choice of columns to drop
  - Values we cleaned: indicators, cleaned columns
  - Filtering of rows (vs. indicators)
- Data that got added
  - Other derived columns: calculations, splits, extractions
  - Joins, Unions
- Changes in granularity:
  - Coarser: grouping keys and aggregates
  - Finer: unnesting

*How might the data **wrangling** influence the **analyses** we can do?*

- Sometimes we won't figure this out until analysis begins
- And we loop back to wrangling!

NSFW

Never Stop Fiercely Wrangling

Your analysis depends on it.

More fun:

https://github.com/Quartz/bad-data-guide