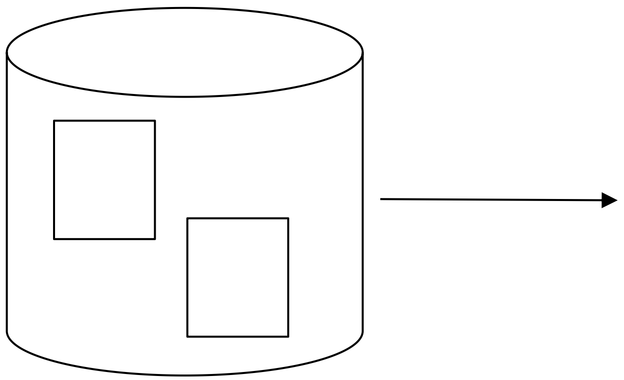
Data Exploration

AnHai Doan

Motivation

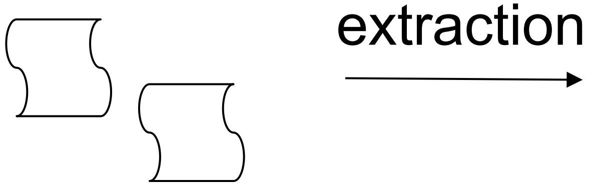


X

id	name	loc
x_1	Apple	CA
x_2	IBM	NY

news articles

data extraction



Y

id	cname	address	rev
y_1	IBM Corp	CA	25
y_2	Apple Inc	CA	51
y_3	GE	NY	351

X

id	name	loc
x_1	Apple	CA
x_2	IBM	NY

Y

id	cname	address	rev
y_1	IBM Corp	CA	25
y_2	Apple Inc	CA	51
y_3	GE	NY	351

data cleaning: GE revenue: 351 ⟶ 35.1

schema matching: name = cname
loc = address

schema merging:

$$\left.\begin{array}{l} \text{X(name, loc)} \\ \text{Y(cname, address, rev)} \end{array}\right\} \text{Z(name, loc, rev)}$$

data matching:

M

xid	yid
x_1	y_2
x_2	y_1

data merging: for name, return the longer string from X.name and Y.cname
for loc, return X.loc

schema mapping: Z = select merge_name(X.name, Y.cname), X.loc, Y.rev
from X, Y, M
where X.id = M.xid and Y.id = M.yid

X

id	name	loc
x_1	Apple	CA
x_2	IBM	NY

Y

id	cname	address	rev
y_1	IBM Corp	CA	25
y_2	Apple Inc	CA	51
y_3	GE	NY	351

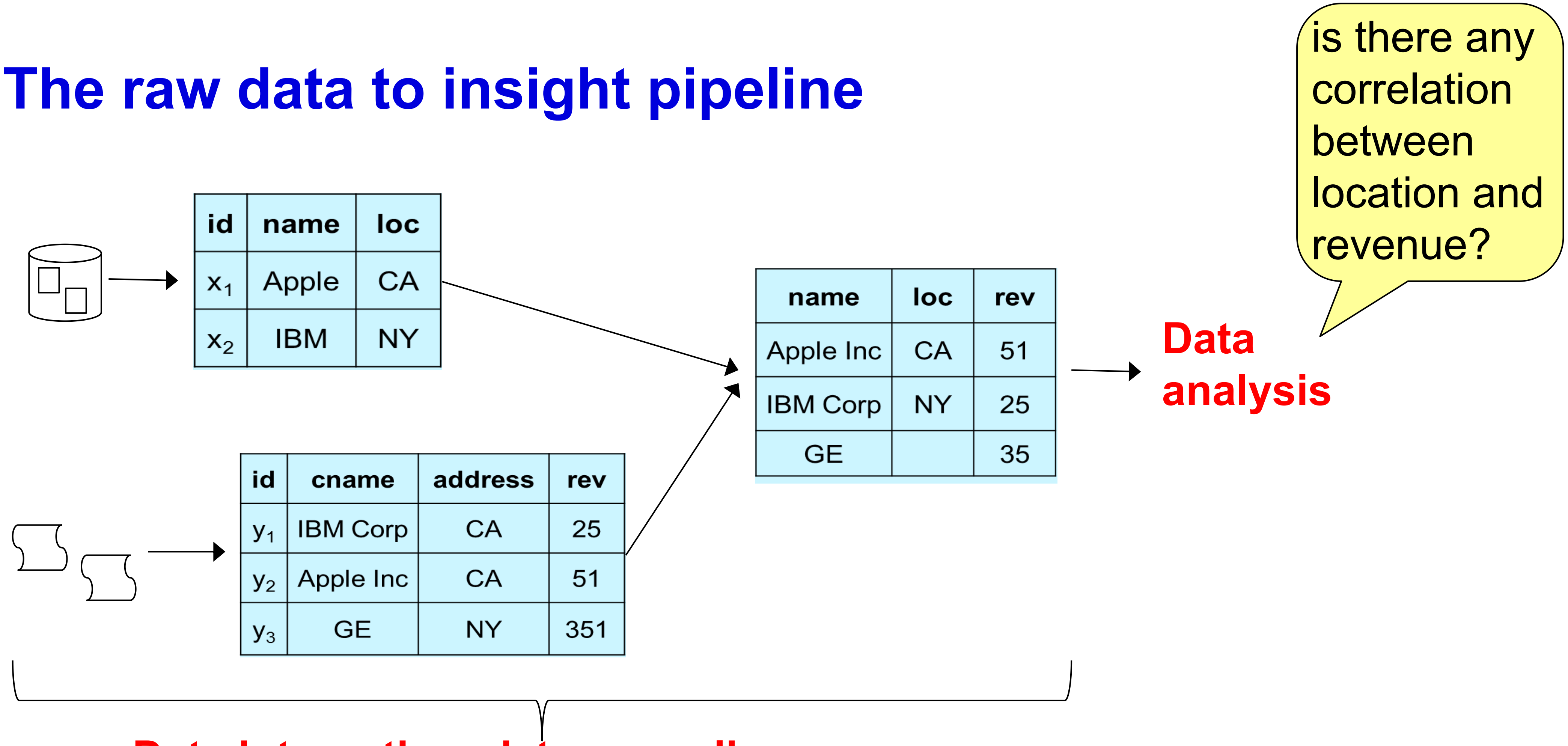
→

Z

name	loc	rev
Apple Inc	CA	51
IBM Corp	NY	25

Another Example

- **The raw data to insight pipeline**



id	name	loc
x_1	Apple	CA
x_2	IBM	NY

id	cname	address	rev
y_1	IBM Corp	CA	25
y_2	Apple Inc	CA	51
y_3	GE	NY	351

name	loc	rev
Apple Inc	CA	51
IBM Corp	NY	25
GE		35

Data analysis

is there any correlation between location and revenue?

Data integration, data wrangling, …

Three Goals for Data Exploration

- **Understand the data**
  - basic characteristics, structure
- **Find interesting stuff**
- **Find quality issues**

- **These help decide what to do later, influence downstream actions**

- **Data exploration is also often known as exploratory data analysis (EDA)**

Basic Operations

- **Browse and query**
- **Visualize**
- **Profile (mostly with automatic programs)**
  - compute statistics
  - detect more stuff (e.g., meta-data such as keys) about the data
  - find data quality problems

- **Two papers to read**
  - a book chapter called “data exploration”
  - a survey paper: “profiling relational data: a survey”
  - will provide them later

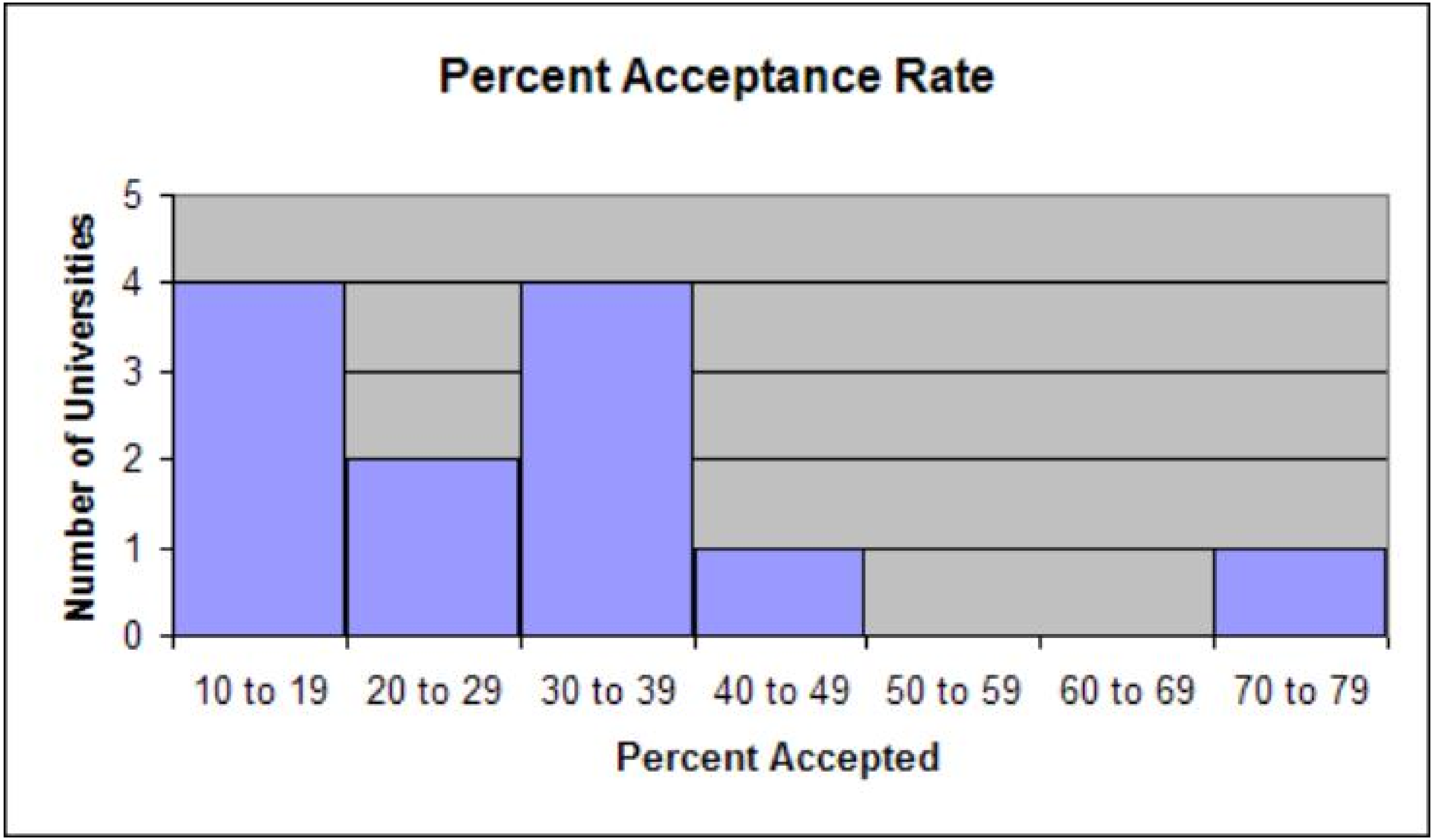
Browse and Query

- **The simplest operations**
- **Yet difficult to do well today**
  - there is no good tool to browse a large table
  - no good tool to query large tables (e.g., 5G) quickly
  - no tool to allow users to ask NL queries (though they start appearing)
- **Today most folks in industry still use Excel**
  - show an example

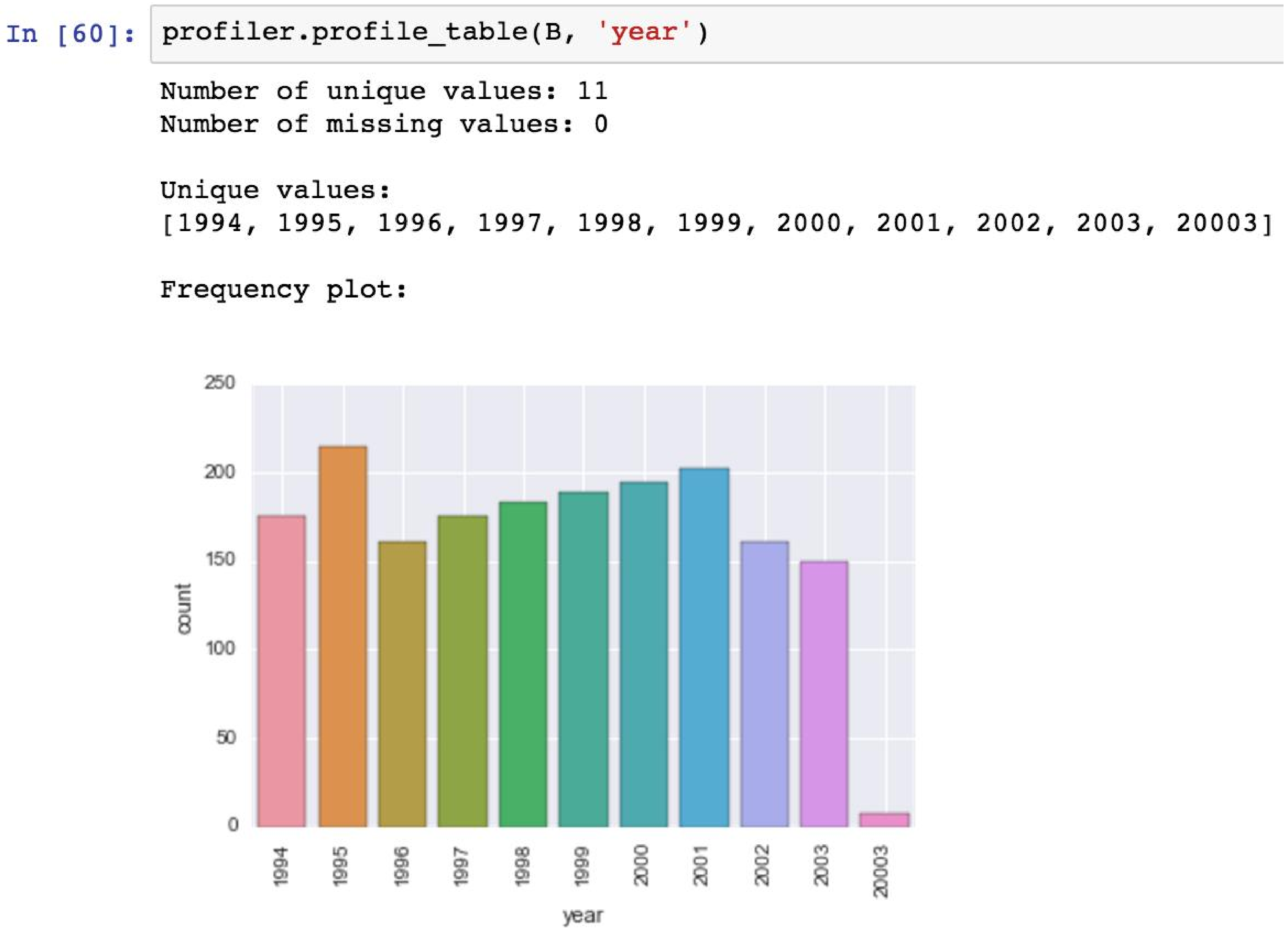
Visualize

- **Why?**
  - can display a lot of information at once (dense information)
  - can leverage human eyes, which are very good at detecting patterns and anomalies (thousands of years of evolution)
- **Basic types of visualization**
  - visualizing a single attribute (aka column)
  - visualizing several attributes
  - visualizing a large number of attributes
- **Visualizing a single attribute**
  - histogram and more (see chapter)

Histogram Example



Histogram Example



```
In [60]: profiler.profile_table(B, 'year')
```

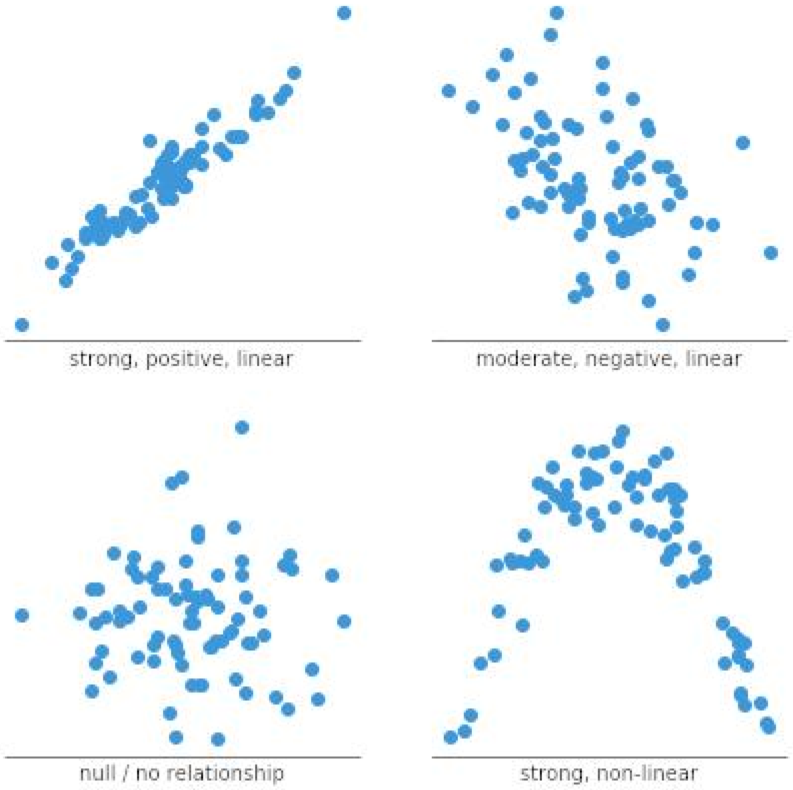
```
Number of unique values: 11
Number of missing values: 0

Unique values:
[1994, 1995, 1996, 1997, 1998, 1999, 2000, 2001, 2002, 2003, 20003]

Frequency plot:
```

Visualizing Multiple Attributes

- **Usually 2-4 attributes at the same time**
- **Scatterplot is a well-known technique**



strong, positive, linear

moderate, negative, linear

null / no relationship

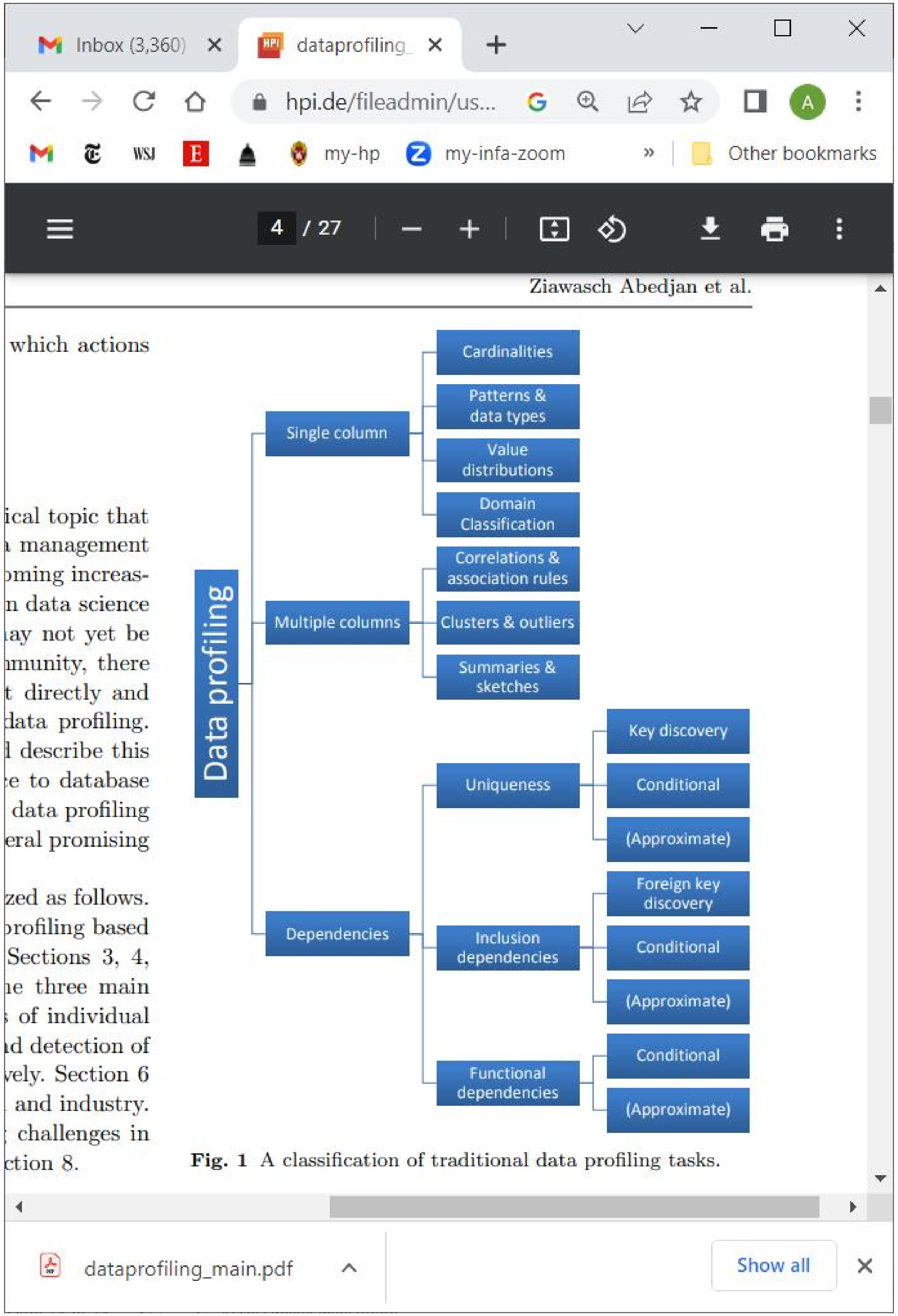
strong, non-linear

- **More in the book chapter**

Basic Operations

- **Browse and query**
- **Visualize**
- **Profile (mostly with automatic programs)**
  - compute statistics
  - detect more stuff (e.g., meta-data such as keys) about the data
  - find data quality problems

Types of Profiling



Ziawasch Abedjan et al.

which actions

ical topic that ı management oming increas- n data science ay not yet be nmunity, there t directly and data profiling. l describe this e to database data profiling eral promising

zed as follows. profiling based Sections 3, 4, he three main s of individual ıd detection of vely. Section 6 and industry. challenges in ction 8.

Fig. 1 A classification of traditional data profiling tasks.

More Types of Profiling

- Can target specifically at detecting certain data quality problems
- See the google doc “data-profiling-cleaning-examples”

More Advanced Stuff

- **30% of values in Salary column are missing, can you explain why?**
  - 95% are missing for Temp workers in the state CA
  - suggest a systematic error here

Summary of Data Exploration

- **Basic operations**
  - Browse and query
  - Visualize
  - Profile (mostly with automatic programs)
    - compute statistics
    - detect more stuff (e.g., meta-data such as keys) about the data
    - find data quality problems
- **Guidance on how to use them**
- **Systems that incorporate the above tools**