Structured Data

Data Science Project: An Inductive Learning Approach

Prof. Dr. Filipe A. N. Verri

About these slides

These slides are companion material for the book

Data Science Project: An Inductive Learning Approach

Prof. Dr. Filipe A. N. Verri

`https://leanpub.com/dsp`

All intellectual content comes from the book and is not AI-generated. Slides were produced with the assistance of Claude Code.

> *Like families, tidy datasets are all alike, but every messy dataset is messy in its own way.*
>
> *— Hadley Wickham, Tidy Data*

Overview

Contents

- Data types
- Database normalization
- Tidy data
- Bridging normalization and tidiness
- Data semantics and interpretation

Objectives

- Understand common data types and formats
- Associate data format and semantics
- Enable the reader to perform data tasks well

Data types

Stevens’ types

Data type	**Allowed operations**
Nominal	$=$
Ordinal	$=, <$
Interval	$=, <, +, -$
Ratio	$=, <, +, -, \times, \div$

Nominal:	colors
Ordinal:	small $<$ medium $<$ large
Interval:	temperature in Celsius
Ratio:	weight in kilograms

Limitations of Stevens’ types

- Do not exhaust all possibilities (e.g., probabilities)
- Data types are not always evident from the data alone
- Same data can be interpreted differently depending on context
- Data scientists must be aware of types and their consequences
- Good assumptions and hypotheses are a key part of the methodology

Database normalization

Relational algebra — Concepts

- **Relation** — table with rows (tuples) and columns (attributes)
- **Projection** $\pi_{A,C}(X)$ — select only specified columns
- **Join** $S \bowtie T$ — combine relations on common attributes
- **Functional dependency** $U \to V$ — if tuples agree on U, they agree on V
- **Multi-valued dependency** $U \twoheadrightarrow V$ — $R = R[UV] \bowtie R[UW]$
- **Join dependency** $*\{X_1, \dots, X_n\}$ — $R = \bowtie \{R[X_1], \dots, R[X_n]\}$

Normal forms

Progressive conditions to reduce redundancy:

- **1NF** — all attributes are atomic
- **2NF** — 1NF + non-prime attributes fully depend on primary key
- **3NF** — 2NF + no transitive dependencies
- **BCNF** — every functional dependency is a result of keys
- **4NF** — every multi-valued dependency is a result of keys
- **PJNF** — key dependencies imply all join dependencies

Example: 2NF vs 3NF

2NF (redundant “course credits”):

Student	Course	Credits	Grade
Alice	Math	4	A
Alice	Physics	3	B
Bob	Math	4	B
Bob	Physics	3	A

3NF (separate tables):

Course	Credits
Math	4
Physics	3

Student	Course	Grade
Alice	Math	A
Alice	Physics	B
Bob	Math	B
Bob	Physics	A

PJNF and invalid joins

- Relation $R[ABC]$ with primary key ABC (no non-trivial FDs)
- Constraint: if (a, b, c'), (a, b', c), $(a', b, c) \in R$ then $(a, b, c) \in R$
- Join dependency $*\{AB, AC, BC\}$ not implied by key
- Must split into $R_1[AB]$, $R_2[AC]$, $R_3[BC]$
- Careless joins may produce invalid tuples

Tidy data

Tidy data

A standardized way to organize data values[1]:

- Each **value** belongs to a variable and an observation
- Each **variable** (column) = same attribute across units
- Each **observation** (row) = same unit across attributes
- **Observational unit** = individual entity being measured

Concept	**Structure**	**Contains**	**Across**
Variable	Column	Same attribute	Units
Observation	Row	Same unit	Attributes

[1] H. Wickham (2014). **"Tidy Data".** In: *Journal of Statistical Software* 59.10, pp. 1–23. DOI: `10.18637/jss.v059.i10`.

Messy vs tidy

Messy

	2019	2020
Brazil	100	200
USA		400

Tidy

Country	Year	Cases
Brazil	2019	100
Brazil	2020	200
USA	2019	
USA	2020	400

Tidy data: variables and observations are clear from the table itself.

Problem: headers are values

Messy (Pew Forum)

Religion	<$10k	$10-20k	…
Agnostic	27	34	…
Atheist	12	27	…
Buddhist	27	21	…

Tidy

Religion	Income	Freq.
Agnostic	<$10k	27
Agnostic	$10-20k	34
…	…	…
Atheist	<$10k	12
Atheist	$10-20k	27
…	…	…

Table becomes longer but narrower.

Problem: multiple variables in one column

Messy (TB dataset)

country	year	column	cases
AD	2000	m014	0
AD	2000	m1524	0
AD	2000	m2534	1
...	...	...	...

Same number of rows, but wider.

Tidy

country	year	sex	age	cases
AD	2000	m	0–14	0
AD	2000	m	15–24	0
AD	2000	m	25–34	1
...	...	...	...	...

Problem: variables in both rows and columns

- Most complicated case of messy data
- One column contains variable names (e.g., “element” = tmax/tmin)
- Day columns (d1, d2, …) are values, not variable names
- Fix: lengthen first, then widen by variable names

Problem: multiple observational units in one table

- Common during data collection
- Example: billboard data with track info repeated for each week
- Fix: separate into one table per observational unit
- Create unique identifiers to link the tables

Problem: single unit in multiple tables

- Data split across files (e.g., one file per year)
- Table/file itself represents a variable value
- Fix: make columns compatible, combine, add origin column

Bridging normalization, tidiness, and data theory

Bridge between concepts

Relations	**Tidy data**	**Philosophy**
Entities	Observational units	Substance
Tuple	Observation	Primary substance
Primary key	Fixed variables	Univocal name
Non-prime attr.	Measured variable	Predicate

The ontological understanding of data influences how it is organized.

Tidy or not tidy?

Temperature measured by 3 sensors, 3 times a day:

Unit = measurement event

date	time	sensor	temp
01-01	00:00	1	20
01-01	00:00	2	21
01-01	00:00	3	22
...	...	...	...

Unit = time instant

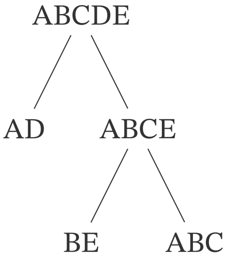
date	time	t1	t2	t3
01-01	00:00	20	21	22
01-01	08:00	21	22	23
...	...	...	...	...

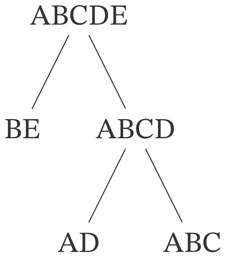
Both are tidy. Tidiness is a matter of **perspective**.

Decomposition trees

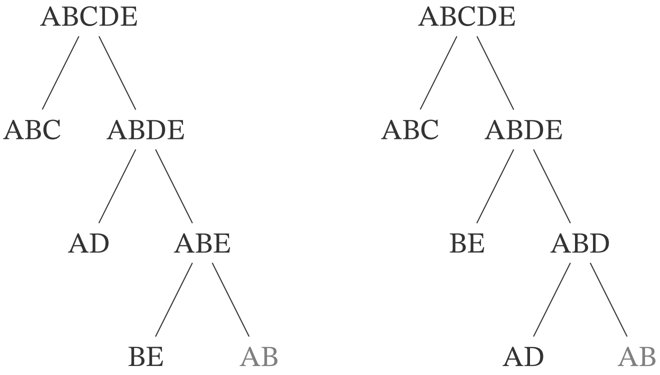
$R[ABCDE]$ with $A \to D$, $B \to E$, $AB \to C$

Valid decompositions to 3NF:





Invalid decomposition trees



$R[AB]$ is not a consequence of a functional dependency.

Change of observational unit

- Traverse decomposition tree from bottom to top with joins
- After each join, perform summarization on new observational unit
- Example: student enrollment → student summary (avg. grade, total load)
- Order of joins and summarization is crucial
- Not trivial to calculate all possible decomposition trees

Data semantics and interpretation

Data semantics and interpretation

- Beyond functional dependencies: **statistical dependencies**
- Attributes may exist in an unknown $P(A, B)$
- Important to understand relationships between observations:
  - Independent? Identically distributed? Selection bias?
  - Temporal dependence? Hidden variables?
- Wrong assumptions → wrong conclusions

Unstructured data

Unstructured data

- No predefined data model (text, images, videos)
- Can be converted to structured data (e.g., bag-of-words)
- Conversion is not always straightforward or lossless
- Out of scope of this book

Takeaways

- The choice of observational unit is not always straightforward
- Format and types must reflect what the solution will “see” in production
- Normalization (storage) and tidy data (analysis) are complementary
- Tidiness is a matter of perspective

Questions?