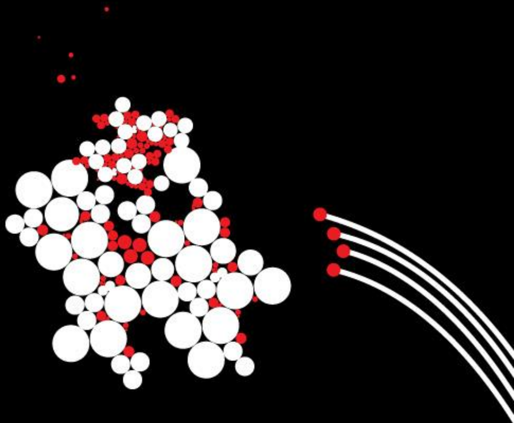
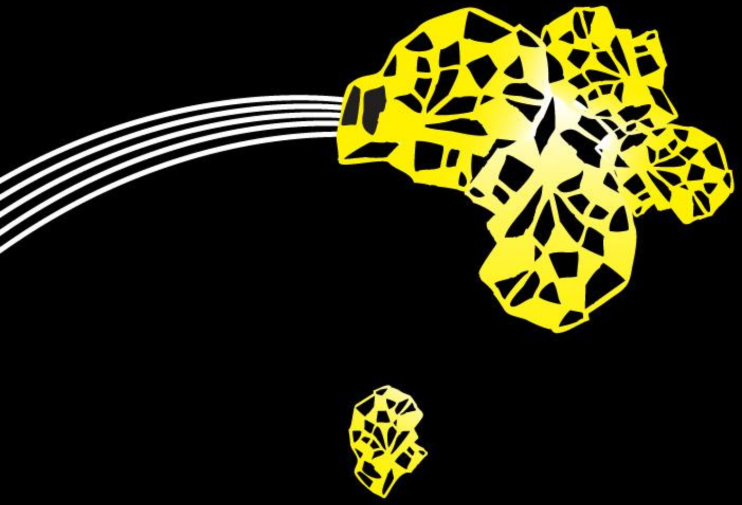
UNIVERSITY OF TWENTE.

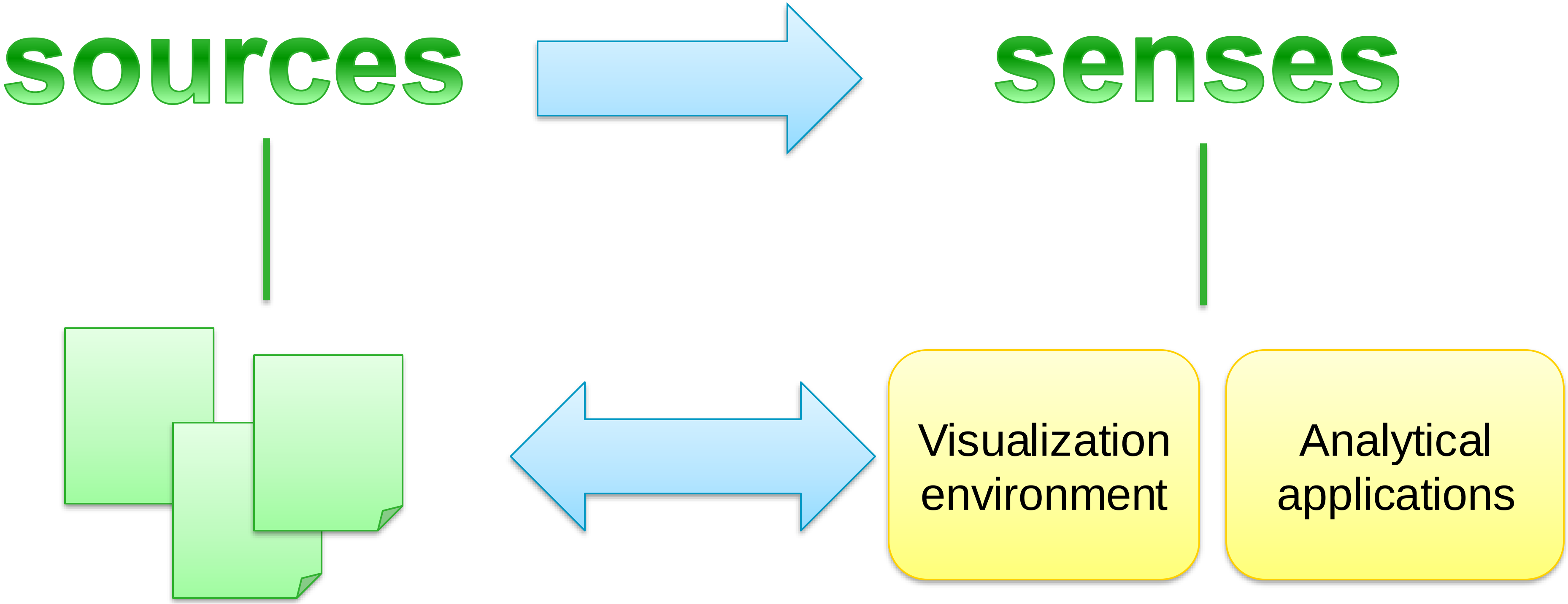


Data Science
Topic DEP
Data Exploration and Preparation

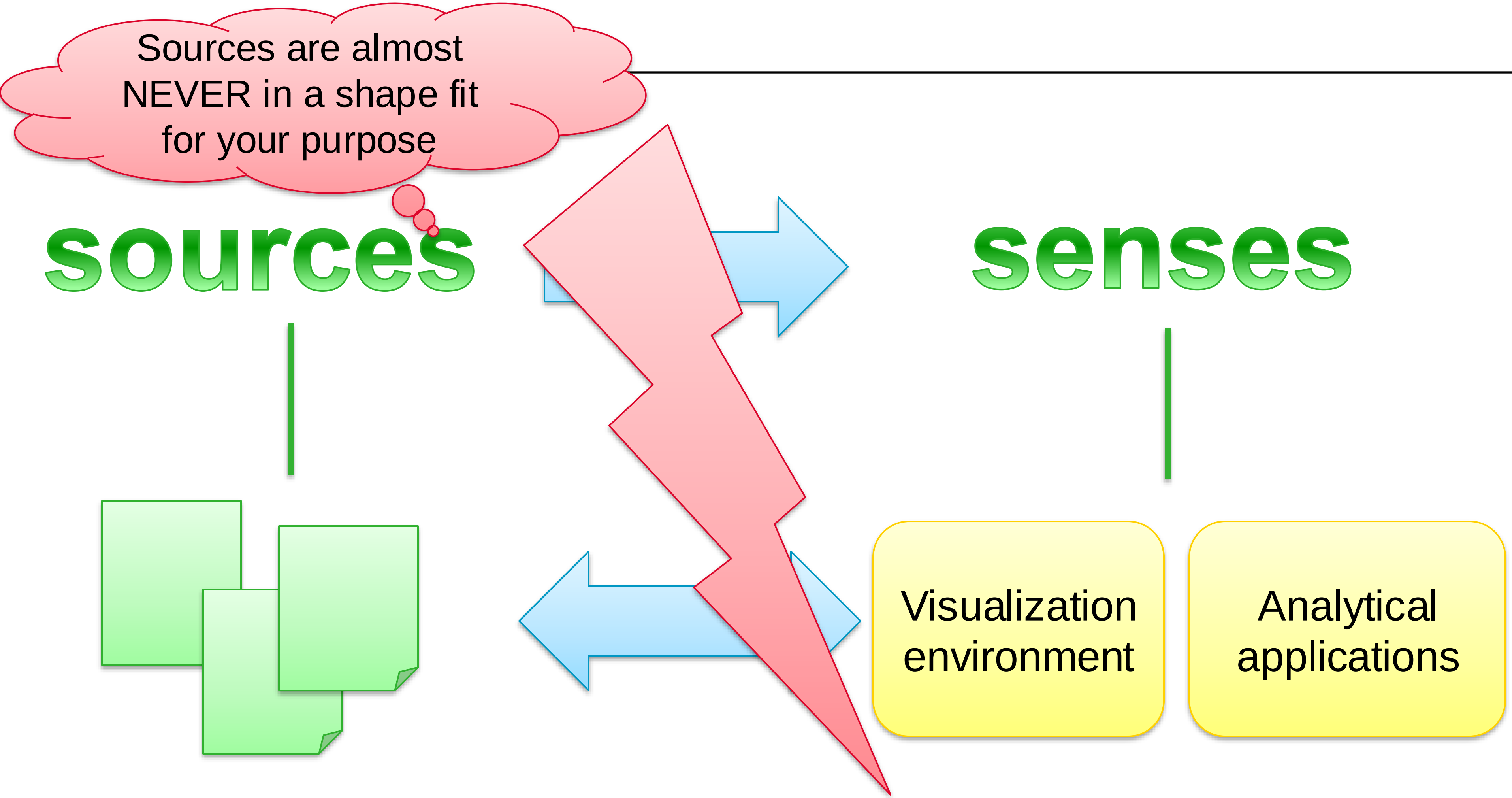
MAURICE VAN KEULEN
(FAIZAN AHMED, CHINTAN AMRIT)



DATA: FROM SOURCES TO SENSES

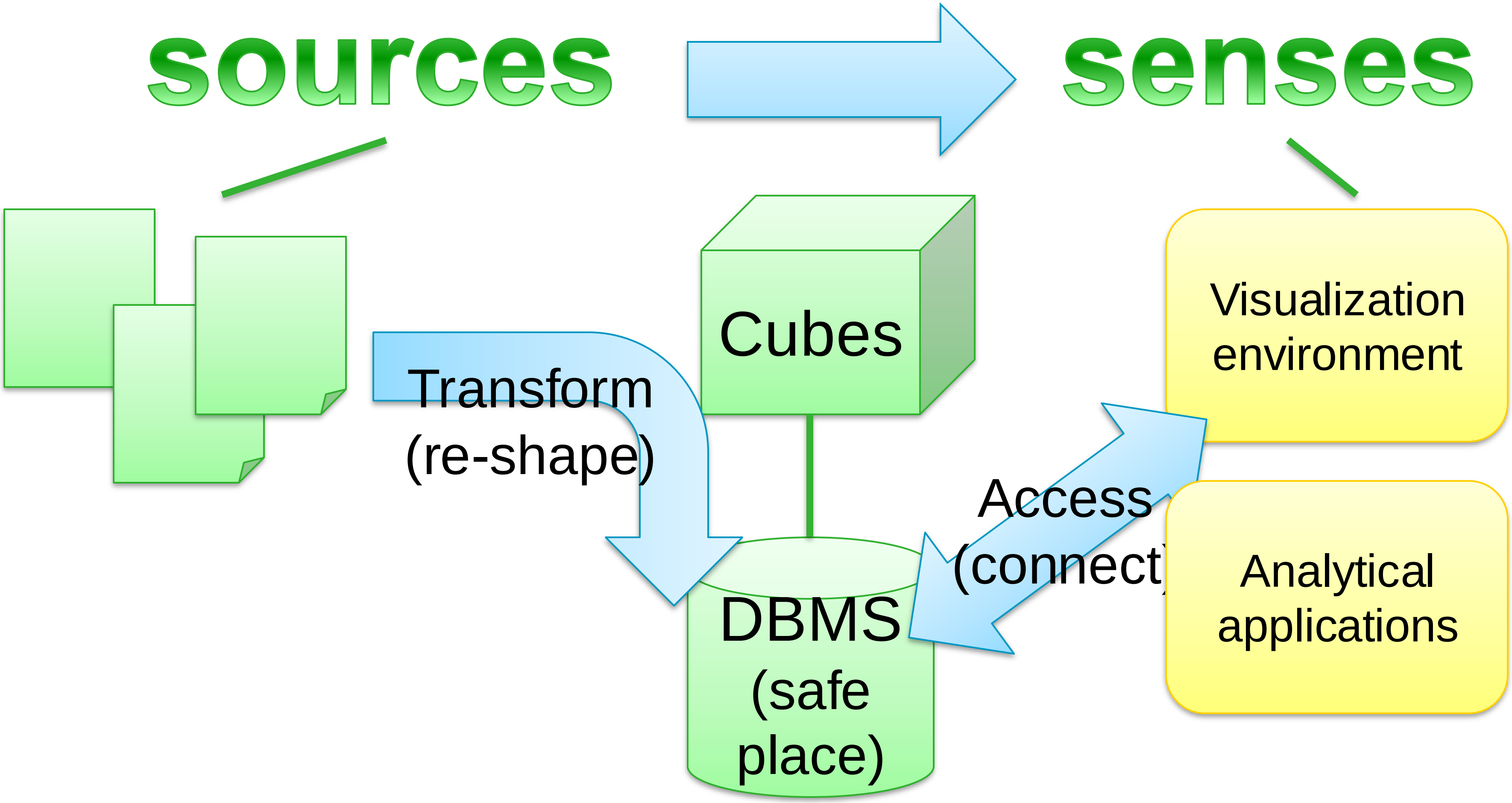


DATA: FROM SOURCES TO SENSES



Sources are almost NEVER in a shape fit for your purpose

DATA: FROM SOURCES TO SENSES



sources

senses

Cubes

Transform (re-shape)

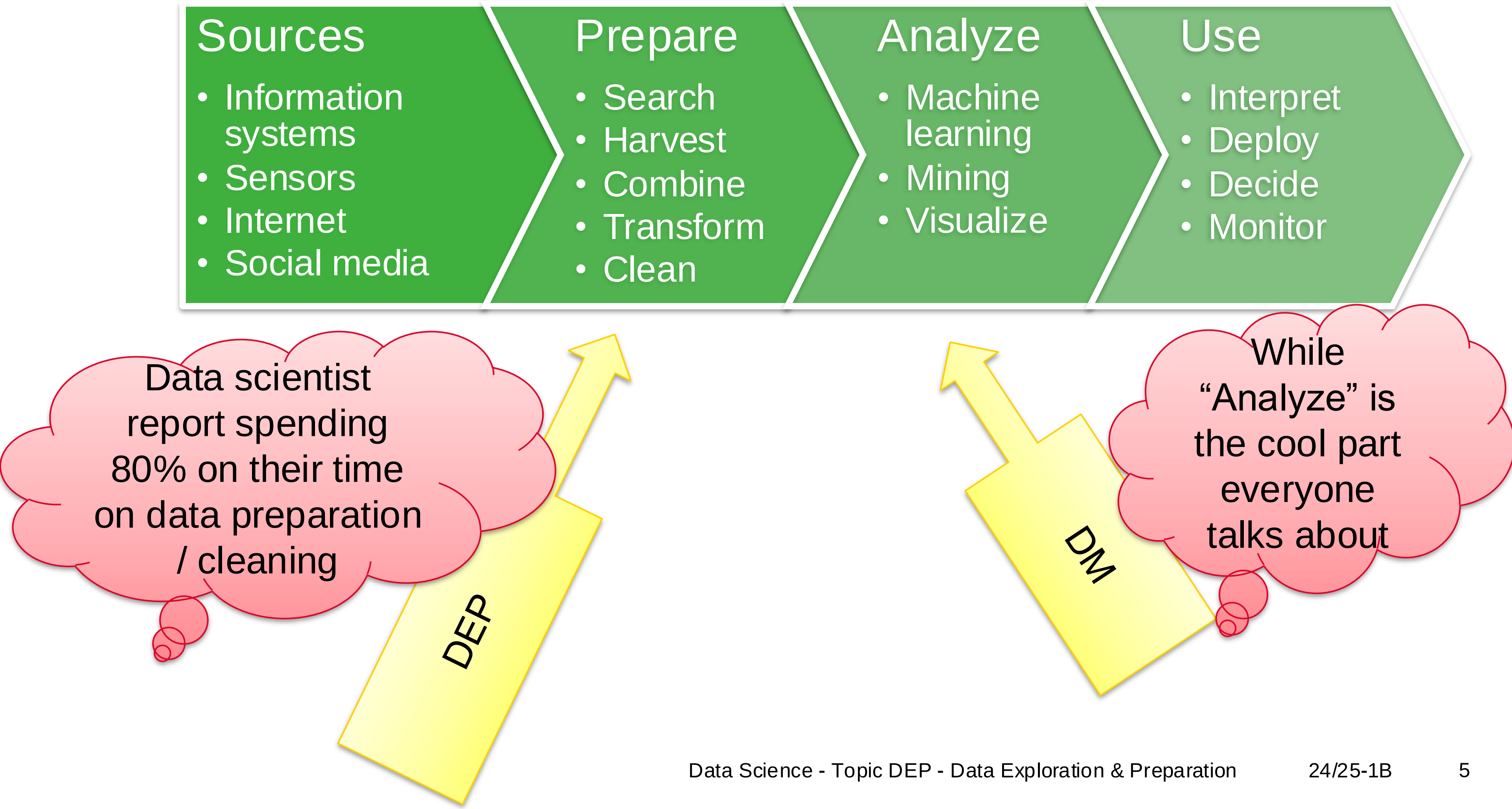
DBMS (safe place)

Access (connect)

Visualization environment

Analytical applications

DATA SCIENCE PROCESS



Sources
- Information systems
- Sensors
- Internet
- Social media

Prepare
- Search
- Harvest
- Combine
- Transform
- Clean

Analyze
- Machine learning
- Mining
- Visualize

Use
- Interpret
- Deploy
- Decide
- Monitor

Data scientist report spending 80% on their time on data preparation / cleaning

While “Analyze” is the cool part everyone talks about

CRISP-DM

CROSS-INDUSTRY STANDARD PROCESS FOR DATA MINING

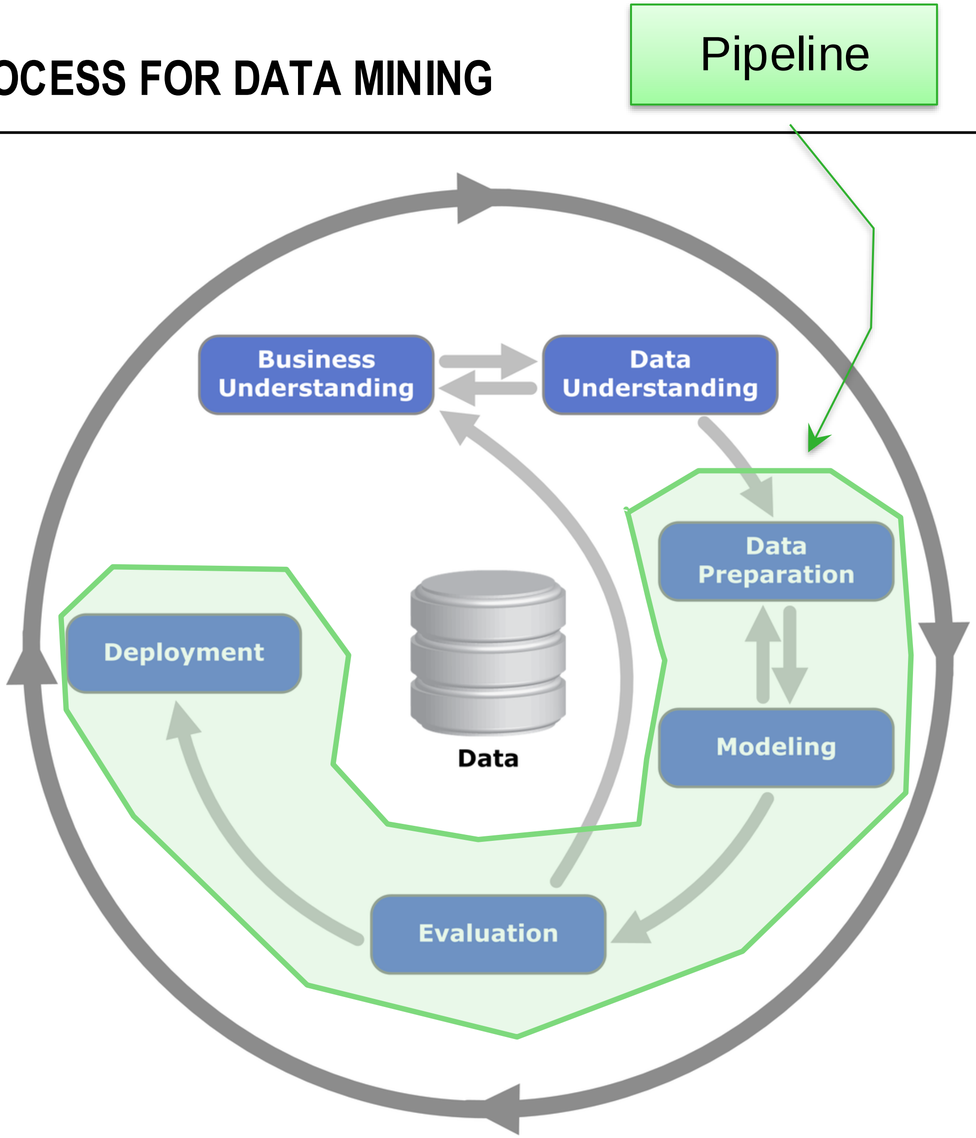
Explore: Understand the process that created the data & how results are to be used

Develop pipeline:

- Clean & transform
- Train model
- Evaluate performance & mistakes

Deploy:

- Integrate in workflow



WHY CUBES

- The **cube** is a *generic* shape for data that fits *analytical purposes*
- A dataset collection often contains many related cubes
  - Each focusing on one or more *facts*
  - Related through shared standardized *dimensions*
- Data is an *asset*
  - It should not live in files transferred by email or download
  - It should live in a safe place: a DBMS
  - Data is something you *connect* to

Example: CBS StatLine: cubes with access API

METHOD

Exploration & preparation can be done in any programming language or with any ETL / wrangling tool

1. Design cube (star schema)
   a) Determine questions the data should answer
   b) Envision tabular reports that may answer those questions
   c) Determine for each question and report, the fact, the dimensions, and granularity
   d) Combine into one star schema
   e) Formulate what one row in fact table means
2. Design associated table structure (UML)
3. Create (empty) tables in database (SQL)
4. Prepare data and fill tables (SQL)

In parallel:
Data exploration of source data

STUDY MATERIAL

Multidimensional modeling

- **Bookchapter**:
  C.S. Jensen, T.B. Pedersen, C. Thomsen, "Fundamental Concepts". **Chapter 2** in "Multidimensional Databases and Data Warehousing". 2010.
  Access: through UT library
  https://ut.on.worldcat.org/oclc/664723898
  - **Note**: You can do **without this book** and rely on slides and practice only; provided as reference because it nicely and slowly explains all the basic concepts with many examples, so if you don't understand something, go read the chapter.
- **ChatGPT**

DATABASES

PERSPECTIVE

A database can also be seen as a kind of ***cloud*** for data

- A **database** is a possibly large collection of data
  - that has to be **exchanged/shared**, **searched**, **corrected/supplemented**, etc.
  - and that **under no circumstances** may **get lost or corrupted** in any way
- A DBMS is software that manages databases, allows these actions, and makes sure your data is safe

- “Information is an asset”
- Availability, reliability, performance, scalability, security

THE DATA IS OFTEN STRUCTURED IN TABLES

THE PRIMARY ‘SHAPE’

Table consists of

- **Records**: Rows in the table
- **Attributes**: Columns in the table

Instance data: The ‘real’ data in the table, the contents

Flight

Number	From	To
KL123	AMS	VIE
OS45	VIE	AMS
KL234	AMS	BRU
NW678	AMS	NYJ
:	:	:

Airport

Code	City
AMS	Amsterdam
BRU	Brussels
VIE	Vienna
NYJ	New York
:	:

Schema: Description of the table structure

Flight(**number**:STRING, from:STRING, to:STRING)

Airport(**code**:STRING, city:STRING)

CONCEPT “KEY”

Key: collection of one or more attributes that

- **Uniquely determine** a record in the table
- Primary key: one ‘most important’ key
- Surrogate key: artificially added code or number to function as a key

Foreign key: attribute(s) in a table that form a **reference** to the (primary key of) one or more records in another relation.

THE DATA IS OFTEN STRUCTURED IN TABLES

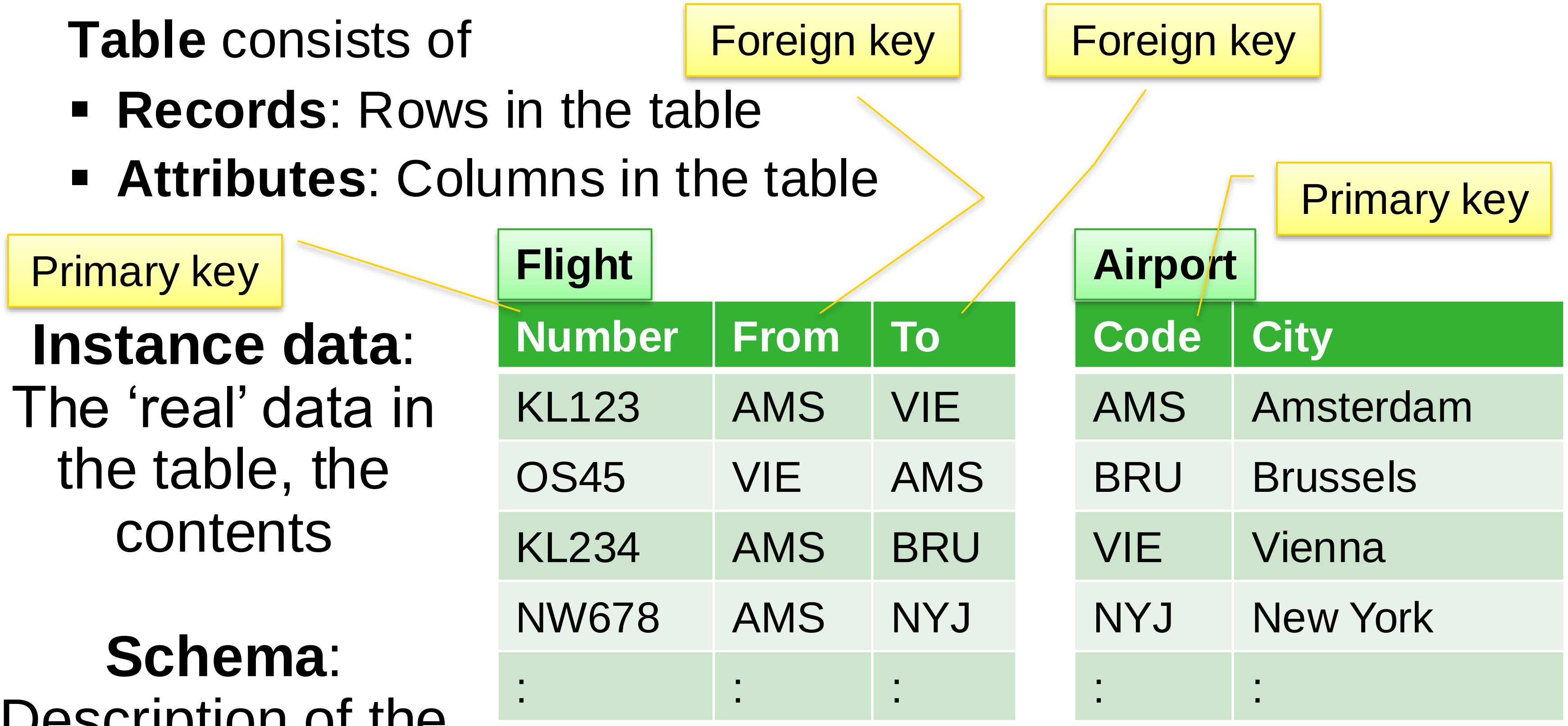
THE PRIMARY ‘SHAPE’

Table consists of

- **Records**: Rows in the table
- **Attributes**: Columns in the table

Instance data: The ‘real’ data in the table, the contents

Schema: Description of the table structure


Primary key

Foreign key

Foreign key

Primary key

Flight

Number	From	To
KL123	AMS	VIE
OS45	VIE	AMS
KL234	AMS	BRU
NW678	AMS	NYJ
:	:	:

Airport

Code	City
AMS	Amsterdam
BRU	Brussels
VIE	Vienna
NYJ	New York
:	:

Flight(**number**:STRING, from:STRING, to:STRING)

Airport(**code**:STRING, city:STRING)

DATABASE SERVER AND DATABASE CLIENT

Database Server

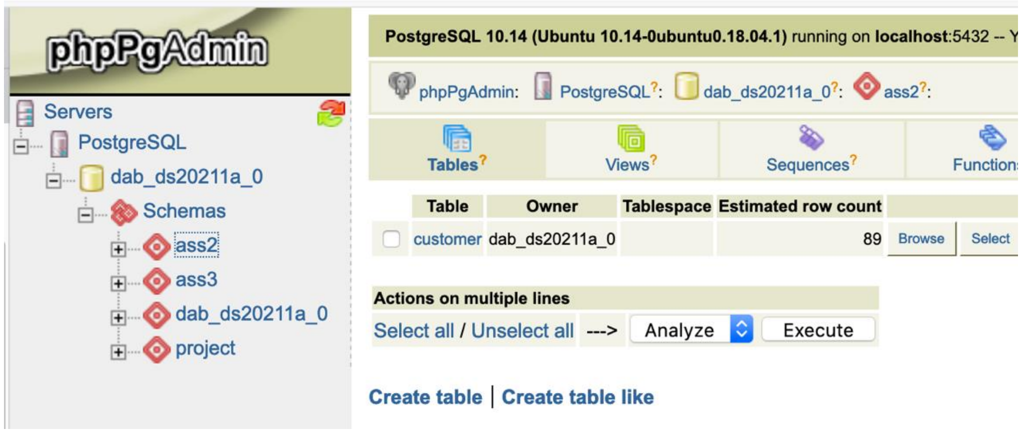
- This is the computer running the DBMS software (Database Management System)
- It runs in the background serving (SQL) requests and keeping your data safe
- We use PostgreSQL pre-installed on bronto.ewi.utwente.nl

Database client

- A tool accessing the database server
- We use PhpPgAdmin for database administration.
- We use R for data cleaning / transformation
- We use Tableau for data visualization
- All are DB clients connecting in a standard way to the server

DATABASE STUFF PRE-INSTALLED

- The database server (PostgreSQL) and database management tool (PhpPgAdmin) are pre-installed on bronto.ewi.utwente.nl
- Each group has their own database
- You need credentials (username / password) for this, which you can obtain from DAB.


phpPgAdmin
Servers
PostgreSQL
dab_ds20211a_0
Schemas
ass2
ass3
dab_ds20211a_0
project

PostgreSQL 10.14 (Ubuntu 10.14-0ubuntu0.18.04.1) running on localhost:5432 -- Y

phpPgAdmin: PostgreSQL?: dab_ds20211a_0?: ass2?:
Tables?
Views?
Sequences?
Function

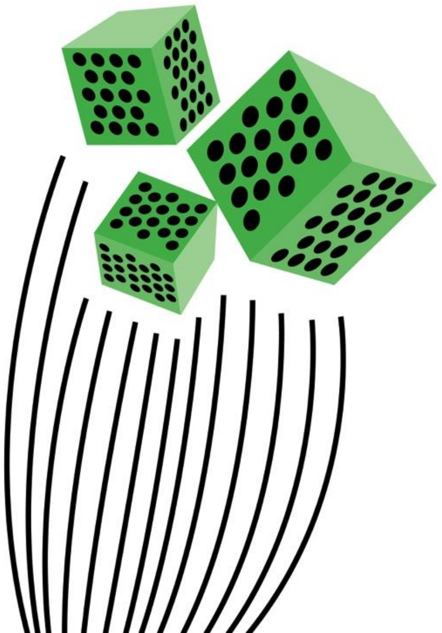
	Table	Owner	Tablespace	Estimated row count		
□	customer	dab_ds20211a_0		89	Browse	Select

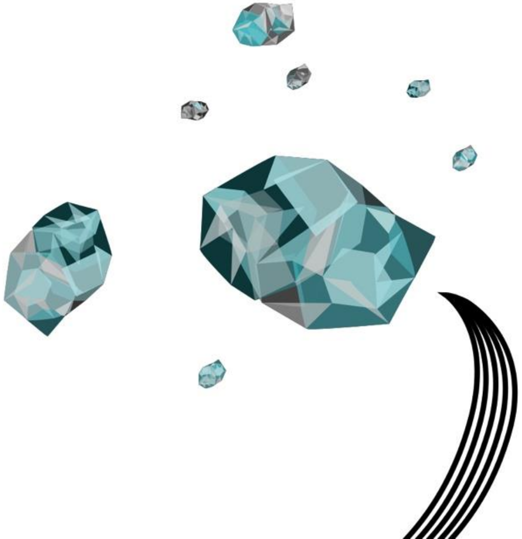
Actions on multiple lines

Select all / Unselect all ---> Analyze Execute
Create table | Create table like

UNIVERSITY OF TWENTE.

THE MANY SHAPES OF DATA





THE DATA IS OFTEN STRUCTURED IN TABLES

THE PRIMARY ‘SHAPE’

Table consists of

- **Records**: Rows in the table
- **Attributes**: Columns in the table

Foreign key

Foreign key

Primary key

Primary key

Instance data: The ‘real’ data in the table, the contents

Schema: Description of the table structure

Flight

Number	From	To
KL123	AMS	VIE
OS45	VIE	AMS
KL234	AMS	BRU
NW678	AMS	NYJ
:	:	:

Airport

Code	City
AMS	Amsterdam
BRU	Brussels
VIE	Vienna
NYJ	New York
:	:

Flight(**number**:STRING, from:STRING, to:STRING)

Airport(**code**:STRING, city:STRING)

DATA IS ALMOST NEVER IN THE DESIRED SHAPE

Even if it is a nice table the rows and columns are not as you desire

EXAMPLE: %SCHOOLING IN THE WORLD

HTTP://BARROLEE.COM

Suppose I want to

- Analyze data on percentage of population who go to school
- … in the different countries
- ... male vs. female
- ... different kinds of schools
- ... over the years

Percentage of population is the **fact**

And these are the **dimensions**

This is a **representation** of a **cube**
(there are more possible representations)

EXAMPLE: %SCHOOLING IN THE WORLD

THIS SHAPE WOULD BE MY TARGET SHAPE FOR THIS DATA

%Schooling	Country	Continent	Sex	School kind	Completeness	Year
11	Albania	Europe	Male	Primary	Yes	2013
12	Albania	Europe	Female	Primary	Yes	2013
8	Albania	Europe	Male	Secondary	Yes	2013
9	Albania	Europe	Female	Secondary	Yes	2013
19	Brazil	South America	Male	Primary	Yes	2013
23	Brazil	South America	Female	Primary	Yes	2013
2	Brazil	South America	Male	Secondary	Yes	2013
1	Brazil	South America	Female	Secondary	Yes	2013
1	Brazil	South America	Male	Primary	No	2013
1	Brazil	South America	Female	Primary	No	2013

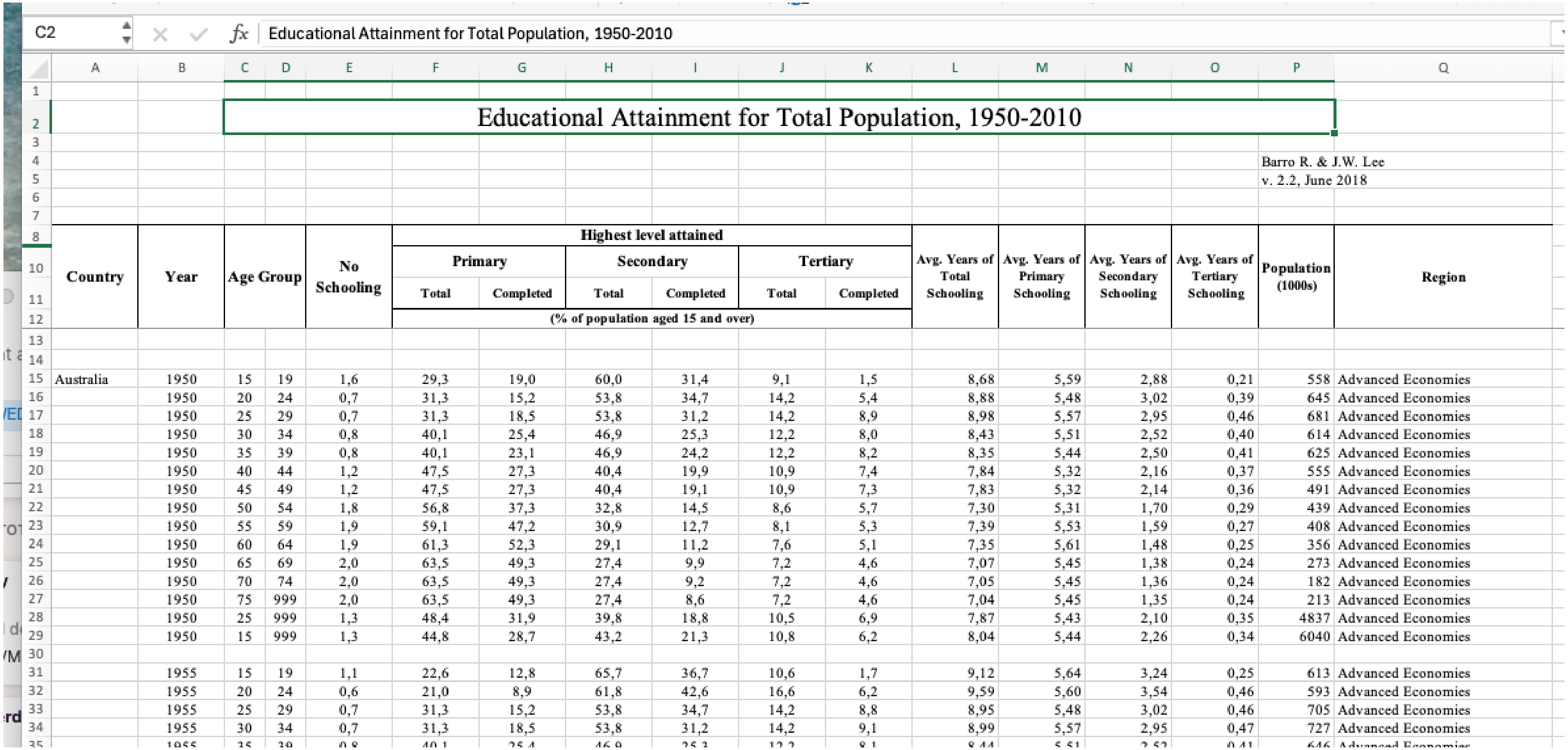
fact

5 dimensions: Continent is a
grouping of countries

dimensions

EXAMPLE: %SCHOOLING IN THE WORLD

THIS IS WHAT THE SOURCE DATA LOOKS LIKE



C2 Educational Attainment for Total Population, 1950-2010

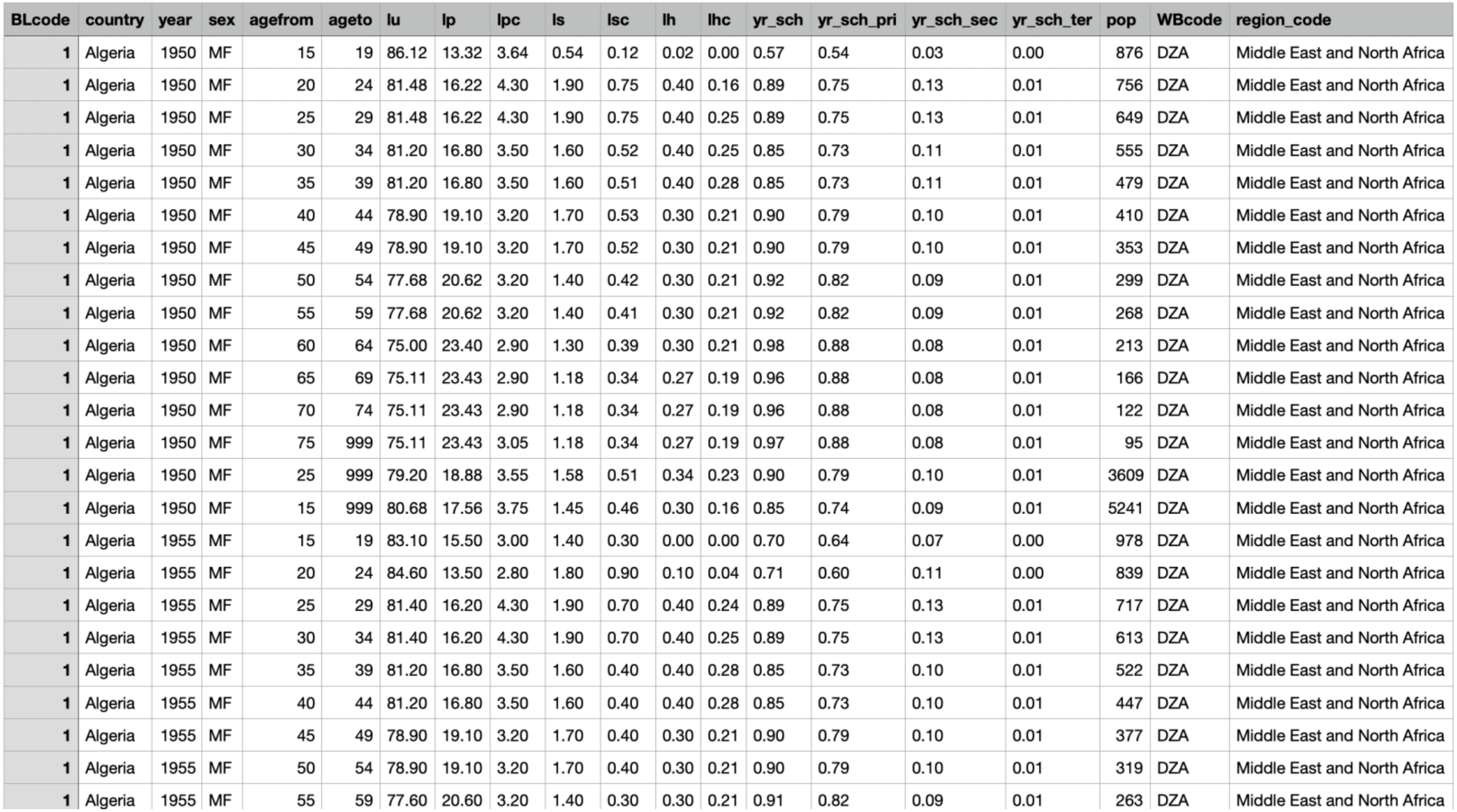
Educational Attainment for Total Population, 1950-2010

Barro R. & J.W. Lee
v. 2.2, June 2018

Country	Year	Age Group		No Schooling	Highest level attained						Avg. Years of Total Schooling	Avg. Years of Primary Schooling	Avg. Years of Secondary Schooling	Avg. Years of Tertiary Schooling	Population (1000s)	Region
					Primary		Secondary		Tertiary							
					Total	Completed	Total	Completed	Total	Completed						
					(% of population aged 15 and over)											
Australia	1950	15	19	1,6	29,3	19,0	60,0	31,4	9,1	1,5	8,68	5,59	2,88	0,21	558	Advanced Economies
	1950	20	24	0,7	31,3	15,2	53,8	34,7	14,2	5,4	8,88	5,48	3,02	0,39	645	Advanced Economies
	1950	25	29	0,7	31,3	18,5	53,8	31,2	14,2	8,9	8,98	5,57	2,95	0,46	681	Advanced Economies
	1950	30	34	0,8	40,1	25,4	46,9	25,3	12,2	8,0	8,43	5,51	2,52	0,40	614	Advanced Economies
	1950	35	39	0,8	40,1	23,1	46,9	24,2	12,2	8,2	8,35	5,44	2,50	0,41	625	Advanced Economies
	1950	40	44	1,2	47,5	27,3	40,4	19,9	10,9	7,4	7,84	5,32	2,16	0,37	555	Advanced Economies
	1950	45	49	1,2	47,5	27,3	40,4	19,1	10,9	7,3	7,83	5,32	2,14	0,36	491	Advanced Economies
	1950	50	54	1,8	56,8	37,3	32,8	14,5	8,6	5,7	7,30	5,31	1,70	0,29	439	Advanced Economies
	1950	55	59	1,9	59,1	47,2	30,9	12,7	8,1	5,3	7,39	5,53	1,59	0,27	408	Advanced Economies
	1950	60	64	1,9	61,3	52,3	29,1	11,2	7,6	5,1	7,35	5,61	1,48	0,25	356	Advanced Economies
	1950	65	69	2,0	63,5	49,3	27,4	9,9	7,2	4,6	7,07	5,45	1,38	0,24	273	Advanced Economies
	1950	70	74	2,0	63,5	49,3	27,4	9,2	7,2	4,6	7,05	5,45	1,36	0,24	182	Advanced Economies
	1950	75	999	2,0	63,5	49,3	27,4	8,6	7,2	4,6	7,04	5,45	1,35	0,24	213	Advanced Economies
	1950	25	999	1,3	48,4	31,9	39,8	18,8	10,5	6,9	7,87	5,43	2,10	0,35	4837	Advanced Economies
	1950	15	999	1,3	44,8	28,7	43,2	21,3	10,8	6,2	8,04	5,44	2,26	0,34	6040	Advanced Economies
	1955	15	19	1,1	22,6	12,8	65,7	36,7	10,6	1,7	9,12	5,64	3,24	0,25	613	Advanced Economies
	1955	20	24	0,6	21,0	8,9	61,8	42,6	16,6	6,2	9,59	5,60	3,54	0,46	593	Advanced Economies
	1955	25	29	0,7	31,3	15,2	53,8	34,7	14,2	8,8	8,95	5,48	3,02	0,46	705	Advanced Economies
	1955	30	34	0,7	31,3	18,5	53,8	31,2	14,2	9,1	8,99	5,57	2,95	0,47	727	Advanced Economies
	1955	35	39	0,8	40,1	25,4	46,9	25,3	12,2	8,1	8,44	5,51	2,52	0,41	646	Advanced Economies

EXAMPLE: %SCHOOLING IN THE WORLD

OR RATHER LOOK AT THE CSV-FILE



BLcode	country	year	sex	agefrom	ageto	lu	lp	lpc	ls	lsc	lh	lhc	yr_sch	yr_sch_pri	yr_sch_sec	yr_sch_ter	pop	WBcode	region_code
1	Algeria	1950	MF	15	19	86.12	13.32	3.64	0.54	0.12	0.02	0.00	0.57	0.54	0.03	0.00	876	DZA	Middle East and North Africa
1	Algeria	1950	MF	20	24	81.48	16.22	4.30	1.90	0.75	0.40	0.16	0.89	0.75	0.13	0.01	756	DZA	Middle East and North Africa
1	Algeria	1950	MF	25	29	81.48	16.22	4.30	1.90	0.75	0.40	0.25	0.89	0.75	0.13	0.01	649	DZA	Middle East and North Africa
1	Algeria	1950	MF	30	34	81.20	16.80	3.50	1.60	0.52	0.40	0.25	0.85	0.73	0.11	0.01	555	DZA	Middle East and North Africa
1	Algeria	1950	MF	35	39	81.20	16.80	3.50	1.60	0.51	0.40	0.28	0.85	0.73	0.11	0.01	479	DZA	Middle East and North Africa
1	Algeria	1950	MF	40	44	78.90	19.10	3.20	1.70	0.53	0.30	0.21	0.90	0.79	0.10	0.01	410	DZA	Middle East and North Africa
1	Algeria	1950	MF	45	49	78.90	19.10	3.20	1.70	0.52	0.30	0.21	0.90	0.79	0.10	0.01	353	DZA	Middle East and North Africa
1	Algeria	1950	MF	50	54	77.68	20.62	3.20	1.40	0.42	0.30	0.21	0.92	0.82	0.09	0.01	299	DZA	Middle East and North Africa
1	Algeria	1950	MF	55	59	77.68	20.62	3.20	1.40	0.41	0.30	0.21	0.92	0.82	0.09	0.01	268	DZA	Middle East and North Africa
1	Algeria	1950	MF	60	64	75.00	23.40	2.90	1.30	0.39	0.30	0.21	0.98	0.88	0.08	0.01	213	DZA	Middle East and North Africa
1	Algeria	1950	MF	65	69	75.11	23.43	2.90	1.18	0.34	0.27	0.19	0.96	0.88	0.08	0.01	166	DZA	Middle East and North Africa
1	Algeria	1950	MF	70	74	75.11	23.43	2.90	1.18	0.34	0.27	0.19	0.96	0.88	0.08	0.01	122	DZA	Middle East and North Africa
1	Algeria	1950	MF	75	999	75.11	23.43	3.05	1.18	0.34	0.27	0.19	0.97	0.88	0.08	0.01	95	DZA	Middle East and North Africa
1	Algeria	1950	MF	25	999	79.20	18.88	3.55	1.58	0.51	0.34	0.23	0.90	0.79	0.10	0.01	3609	DZA	Middle East and North Africa
1	Algeria	1950	MF	15	999	80.68	17.56	3.75	1.45	0.46	0.30	0.16	0.85	0.74	0.09	0.01	5241	DZA	Middle East and North Africa
1	Algeria	1955	MF	15	19	83.10	15.50	3.00	1.40	0.30	0.00	0.00	0.70	0.64	0.07	0.00	978	DZA	Middle East and North Africa
1	Algeria	1955	MF	20	24	84.60	13.50	2.80	1.80	0.90	0.10	0.04	0.71	0.60	0.11	0.00	839	DZA	Middle East and North Africa
1	Algeria	1955	MF	25	29	81.40	16.20	4.30	1.90	0.70	0.40	0.24	0.89	0.75	0.13	0.01	717	DZA	Middle East and North Africa
1	Algeria	1955	MF	30	34	81.40	16.20	4.30	1.90	0.70	0.40	0.25	0.89	0.75	0.13	0.01	613	DZA	Middle East and North Africa
1	Algeria	1955	MF	35	39	81.20	16.80	3.50	1.60	0.40	0.40	0.28	0.85	0.73	0.10	0.01	522	DZA	Middle East and North Africa
1	Algeria	1955	MF	40	44	81.20	16.80	3.50	1.60	0.40	0.40	0.28	0.85	0.73	0.10	0.01	447	DZA	Middle East and North Africa
1	Algeria	1955	MF	45	49	78.90	19.10	3.20	1.70	0.40	0.30	0.21	0.90	0.79	0.10	0.01	377	DZA	Middle East and North Africa
1	Algeria	1955	MF	50	54	78.90	19.10	3.20	1.70	0.40	0.30	0.21	0.90	0.79	0.10	0.01	319	DZA	Middle East and North Africa
1	Algeria	1955	MF	55	59	77.60	20.60	3.20	1.40	0.30	0.30	0.21	0.91	0.82	0.09	0.01	263	DZA	Middle East and North Africa

EXAMPLE: %SCHOOLING IN THE WORLD

WHAT RESHAPING (DATA TRANSFORMATION) NEEDS TO BE DONE?

Source has

- More attributes and rows than needed
- Data in different attributes of the same row, that I want to have on separate rows
- Data is in different files that I want in one table

One goal of data preparation: reshaping the structure

DATA IS ALMOST NEVER IN THE DESIRED SHAPE

Even if it is a nice table the contents (values) are not as you desire

DATA SEMANTICS: EXAMPLE

DB of department 1

enr	name	salary

DB of department 2

enr	name	salary

Data warehouse

enr	name	salary

What could be an obstacle for a simple union of these tables?

- Situations
- Exceptions
- Semantical differences

DATA SEMANTICS: EXAMPLE

CONTINUED

DB of department 1

enr	name	salary
3	M. van Keulen	100.000
4	R. Pieper	100.000
5	H. Blanken	200.000

DB of department 2

enr	name	salary
3	Keulen, M. van	3.781,50
6	Pieper, R.	18.907,51
9	Blanken, B.	7.563,00
12	Poel. M.	5.673,25
15	Vet, P. van der	NULL

THERE IS MORE TO SHAPE THAN STRUCTURE

There is more to shape than the **structure** of the data

- ➢The **contents** can also be in a wrong ‘shape’

Contents

- What do the rows and columns really mean?
- What have people put in them? (exceptional cases)
- Missing values, inconsistent values, wrong values, …

- ➢Problems with **data quality** are often much much time-consuming to solve than re-shaping the structure

DATA EXPLORATION: DISCOVER WRONG ‘SHAPE’ EARLY

How can we know that there are data quality problems in the data?

- Have a critical attitude
- Go actively in search for them: data exploration
  - Tool: Summary statistics & Data visualization
  - Identify patterns (distributions, skew, …)
  - Find outliers
  - Test assumptions (uniqueness, dependencies, …)
  - Check for common problems (missings, …)
  - Ask domain expert for explanations & reasons (know more about the process that creates the data)

COMMON SUMMARY STATISTICS & VISUALIZATIONS

EXCERPT OF DATA QUALIT METRICS

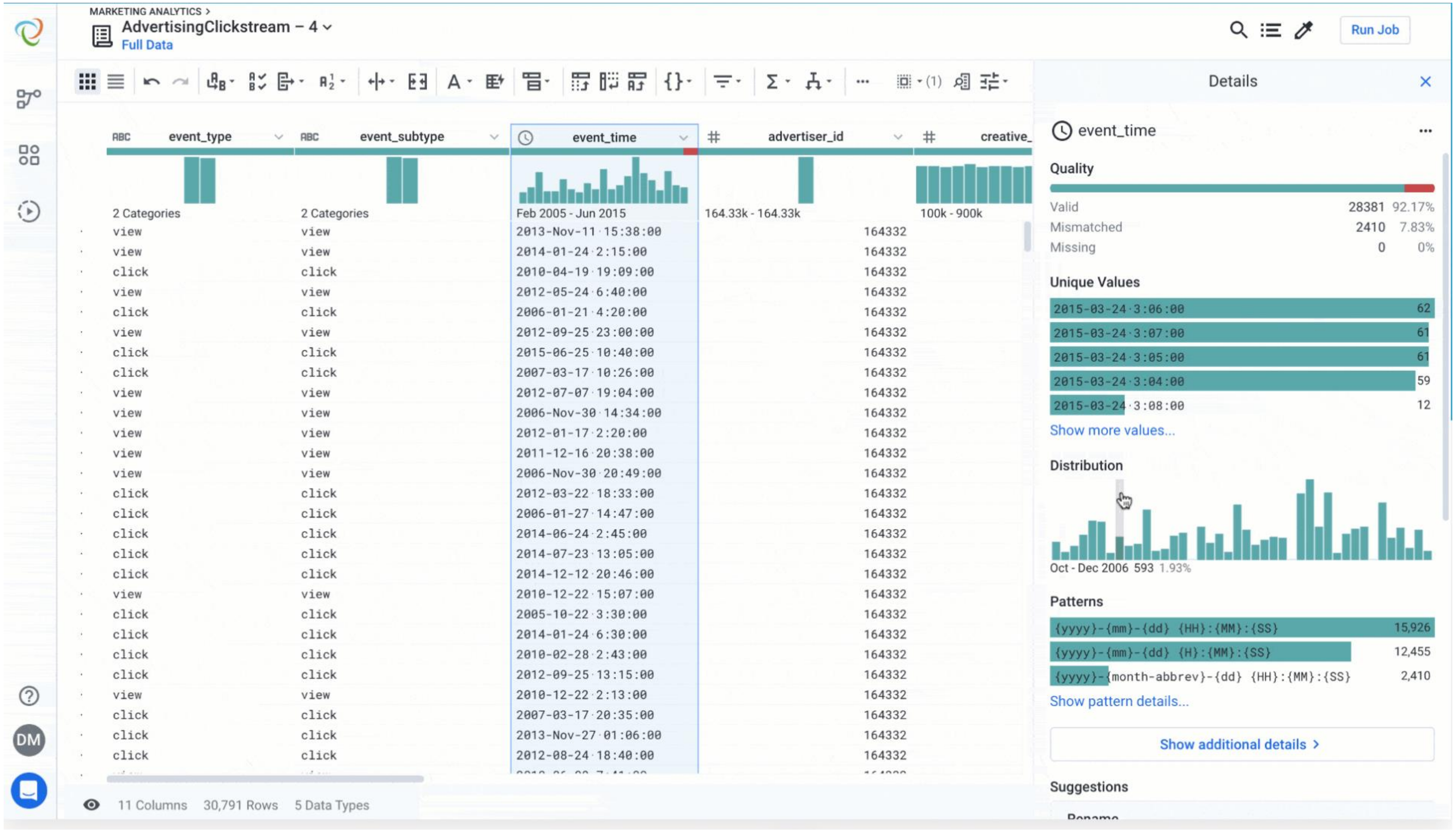
Per attribute

- Basic: Range, Mean / Median, Standard deviation, Uniqueness, #missings
- Advanced: Distribution (histogram), Skewness/Kurtosis (asymmetry & peakiness), Percentiles, Outliers, Cross-tabulation, temporal/spatial patterns

Between attributes

- Correlation & covariance
- Assumptions: inclusion (keys), multi-attribute uniqueness, semantic dependencies

EXAMPLE: DATA WRANGLING TOOL ‘TRIFACTA’



MARKETING ANALYTICS >

AdvertisingClickstream – 4

Full Data

ABC event_type	ABC event_subtype	event_time	# advertiser_id	# creative_
2 Categories	2 Categories	Feb 2005 - Jun 2015	164.33k - 164.33k	100k - 900k
view	view	2013-Nov-11·15:38:00	164332	
view	view	2014-01-24·2:15:00	164332	
click	click	2010-04-19·19:09:00	164332	
view	view	2012-05-24·6:40:00	164332	
click	click	2006-01-21·4:20:00	164332	
view	view	2012-09-25·23:00:00	164332	
click	click	2015-06-25·10:40:00	164332	
click	click	2007-03-17·10:26:00	164332	
view	view	2012-07-07·19:04:00	164332	
view	view	2006-Nov-30·14:34:00	164332	
view	view	2012-01-17·2:20:00	164332	
view	view	2011-12-16·20:38:00	164332	
view	view	2006-Nov-30·20:49:00	164332	
click	click	2012-03-22·18:33:00	164332	
click	click	2006-01-27·14:47:00	164332	
click	click	2014-06-24·2:45:00	164332	
click	click	2014-07-23·13:05:00	164332	
click	click	2014-12-12·20:46:00	164332	
view	view	2010-12-22·15:07:00	164332	
click	click	2005-10-22·3:30:00	164332	
click	click	2014-01-24·6:30:00	164332	
click	click	2010-02-28·2:43:00	164332	
click	click	2012-09-25·13:15:00	164332	
view	view	2010-12-22·2:13:00	164332	
click	click	2007-03-17·20:35:00	164332	
click	click	2013-Nov-27·01:06:00	164332	
click	click	2012-08-24·18:40:00	164332	

11 Columns 30,791 Rows 5 Data Types

Details

event_time

Quality

Valid	28381	92.17%
Mismatched	2410	7.83%
Missing	0	0%

Unique Values

2015-03-24·3:06:00	62
2015-03-24·3:07:00	61
2015-03-24·3:05:00	61
2015-03-24·3:04:00	59
2015-03-24·3:08:00	12

Show more values...

Distribution

Oct - Dec 2006 593 1.93%

Patterns

{yyyy}-{mm}-{dd} {HH}:{MM}:{SS}	15,926
{yyyy}-{mm}-{dd} {H}:{MM}:{SS}	12,455
{yyyy}-{month-abbrev}-{dd} {HH}:{MM}:{SS}	2,410

Show pattern details...

Show additional details >

Suggestions

Rename

ATTRIBUTE TYPES & FORMATS

Not every analysis method can be applied to any data. Some have limitations regarding attribute types:

- Continuous vs. Discrete
  - Continuous: real numbers, coordinates, time
  - Discrete: integer, nominal, ordinal

Nominal: limited set of ‘labels’ or ‘categories’

- Example: Male, Female

Ordinal: same but with an order

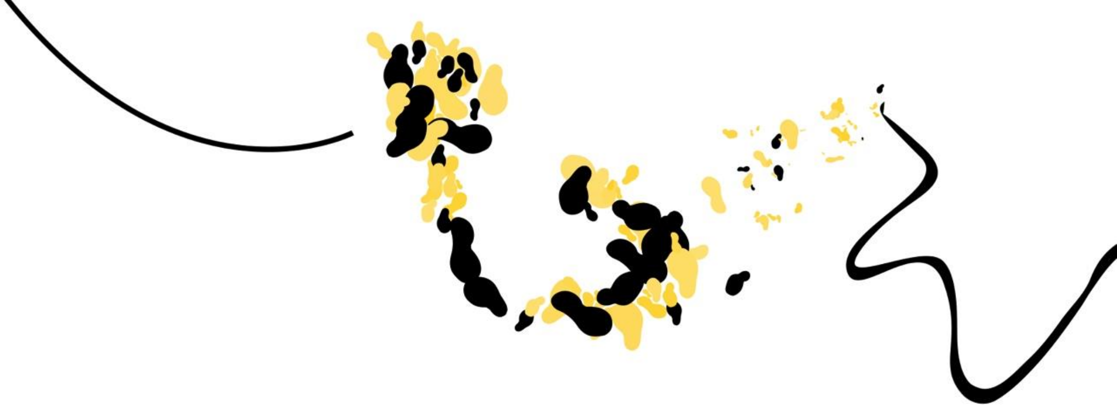
- Example: Very Low, Low, Medium, High, Very High

ATTRIBUTE TYPES (CONTINUED)

In programming languages, databases and tools, variables/attributes always have a type.

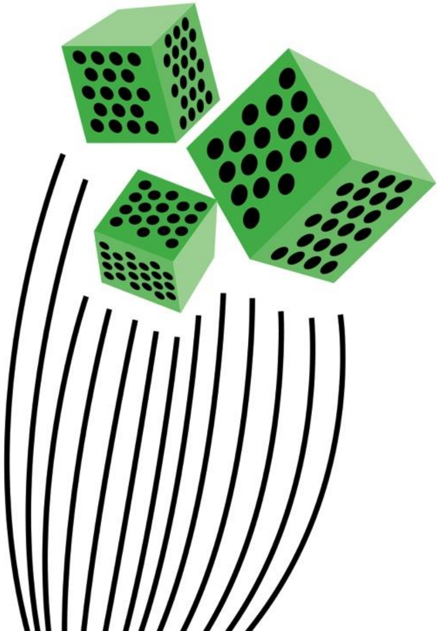
Some often occurring

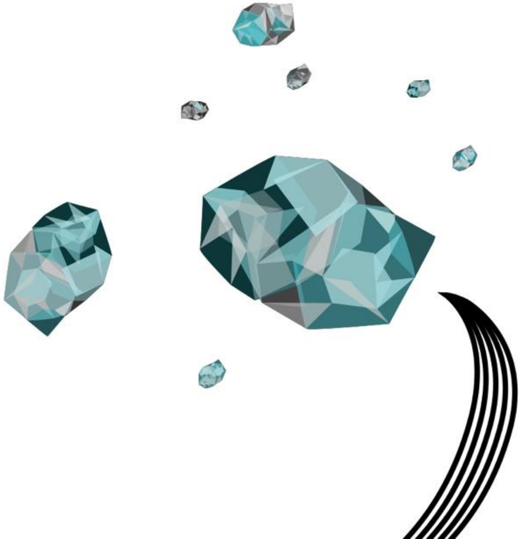
- Integer: whole numbers upto certain maximum
- Float/double: real numbers of certain precision
- Date/Time/DateTime
- String: sequence of characters with certain length
  (in databases: “characted varying” or “varchar” or “text”)
- Boolean: true or false
- Char: one character



CUBES

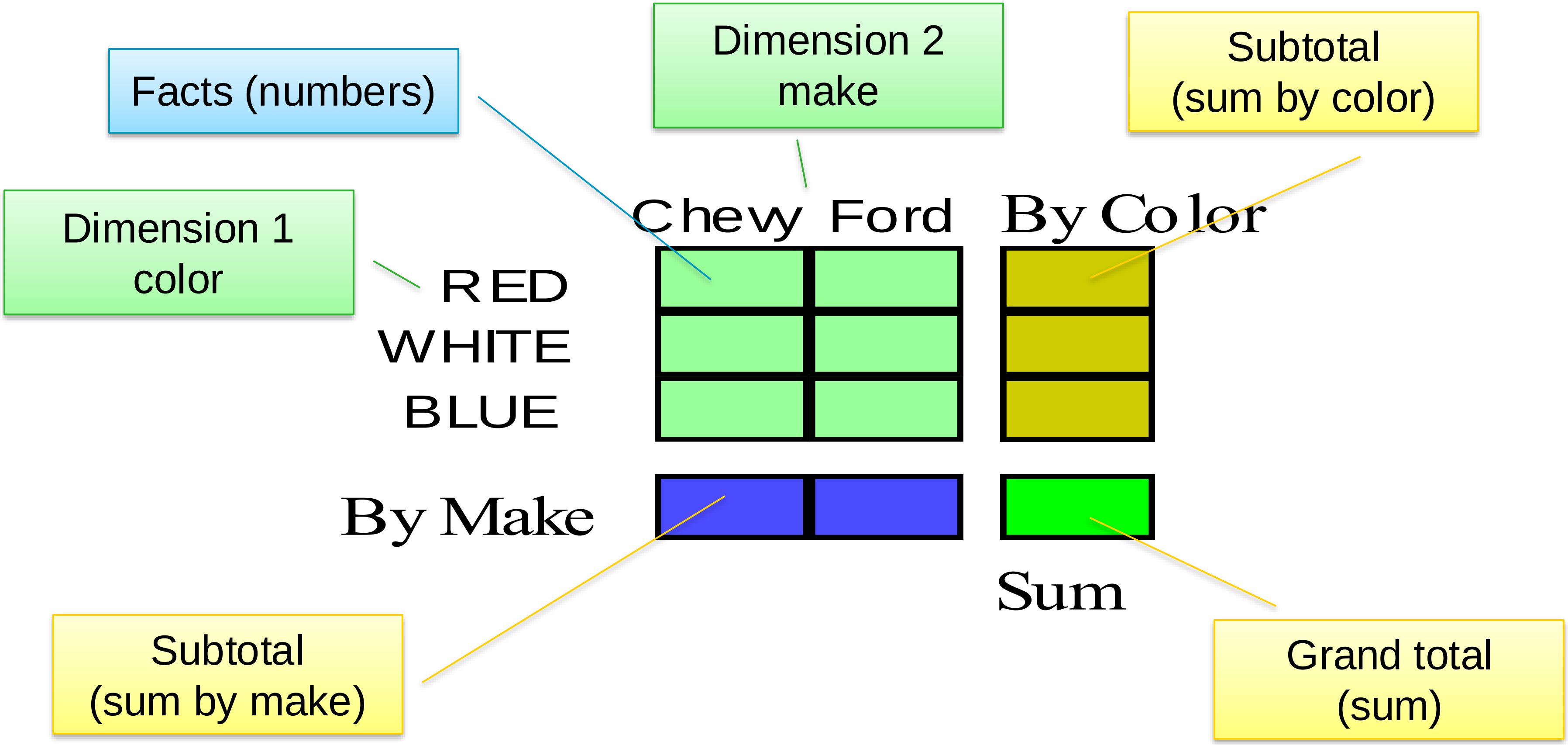
GENERIC SHAPE SUITABLE FOR ANALYTICS





SPREADSHEET

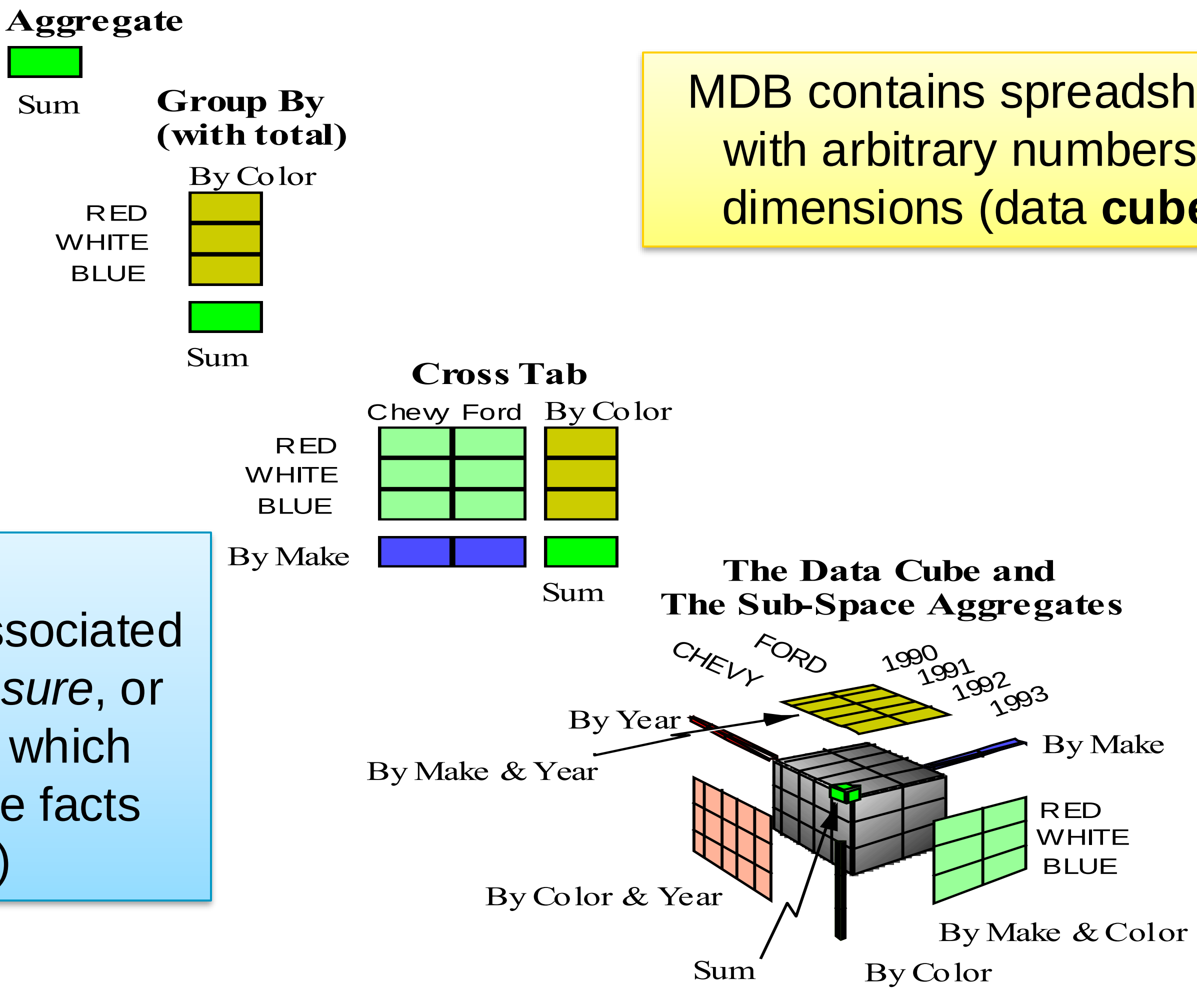
IS A CUBE WITH TWO DIMENSIONS



CUBE = MULTI-DIMENSIONAL DATABASE

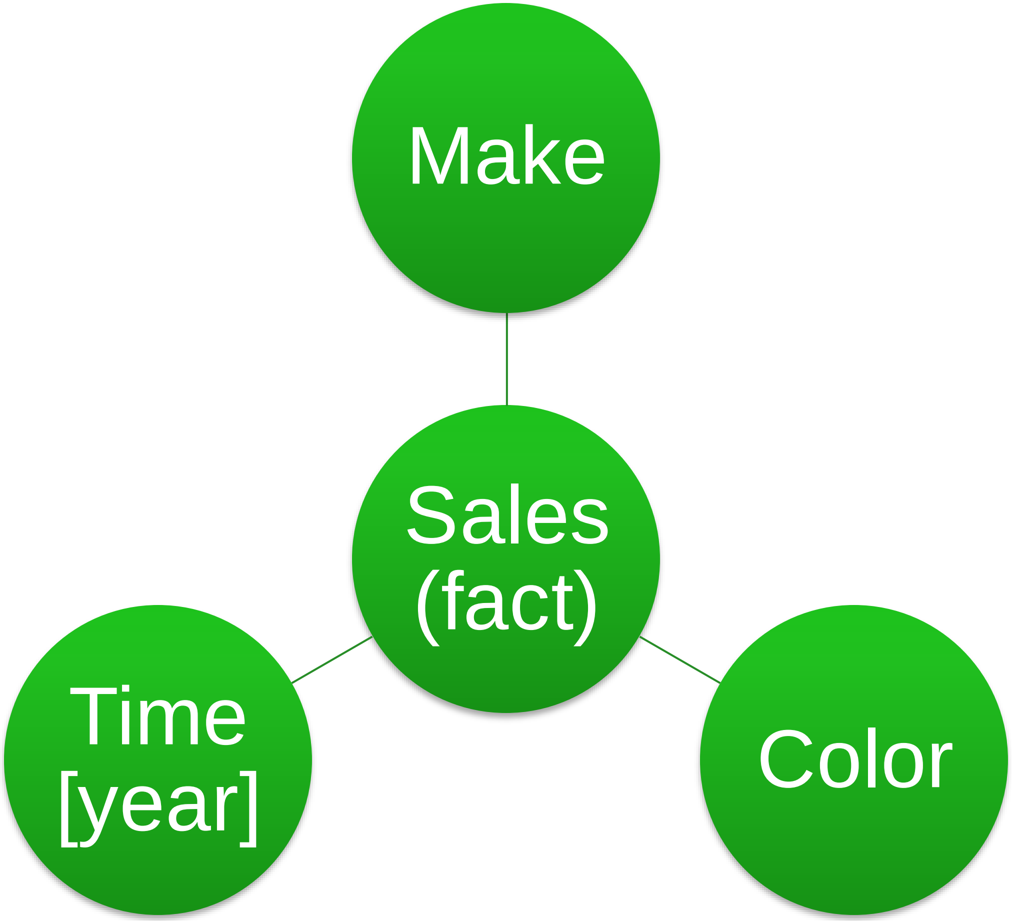
= MULTI-DIMENSIONAL SPREADSHEET

MDB contains spreadsheets with arbitrary numbers of dimensions (data **cube**s)



Data is either
- a **fact** with associated numerical *measure*, or
- a **dimension** which characterize the facts (mostly textual)

STAR SCHEMA (CONCEPTUAL DESIGN)

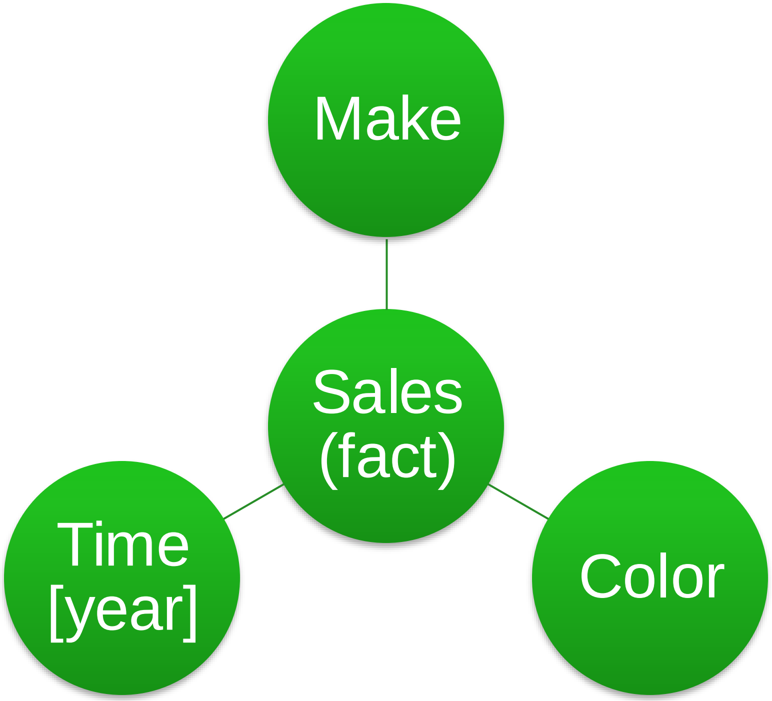


WHAT DOES THIS STAR SCHEMA MEAN

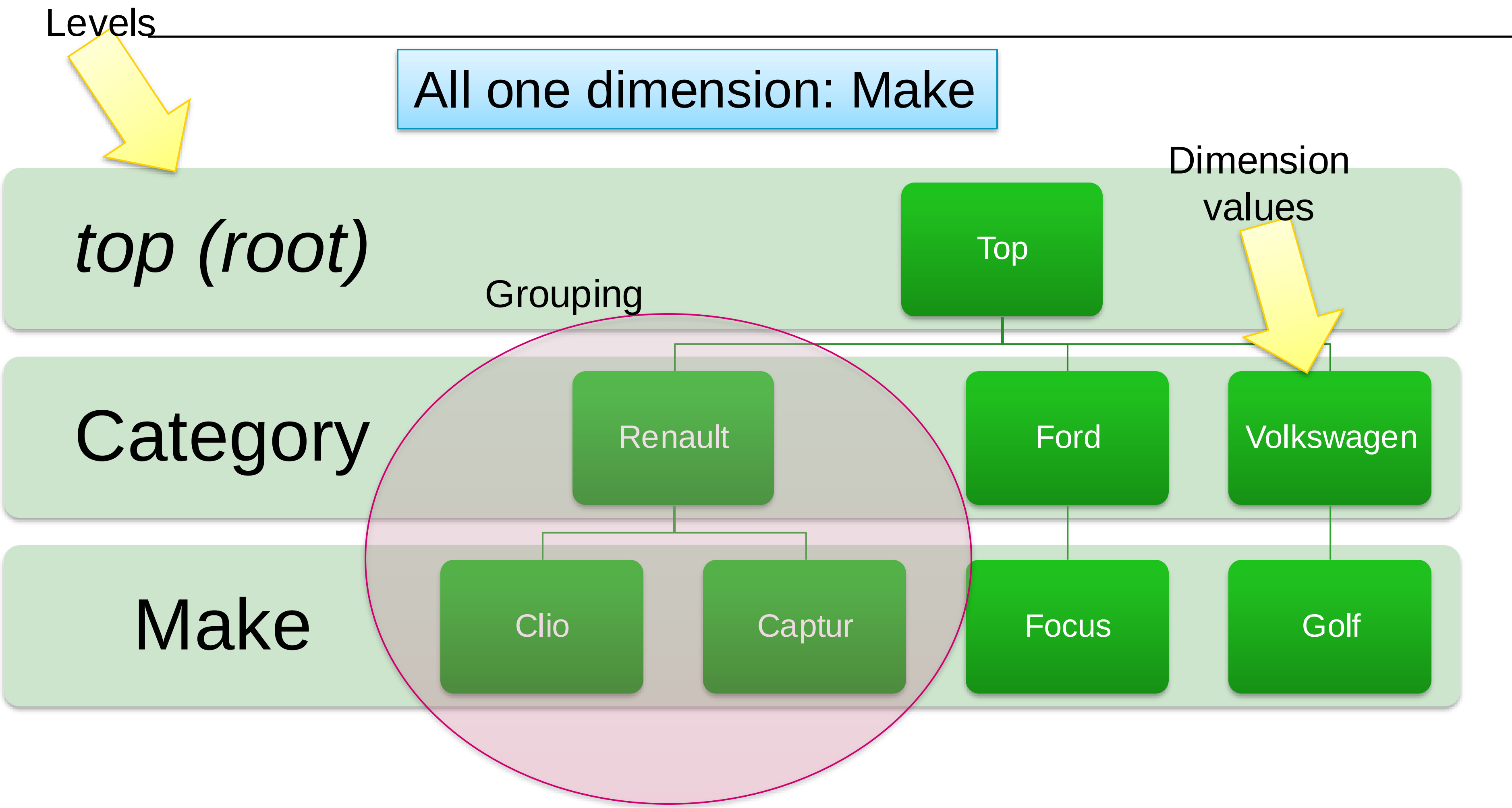
In one spreadsheet / table

- One row of sales
  Per combination of
  Make, Time Unit, Color
- Attributes for
  sales (fact: amount)
  make (dim: name, category)
  time (dim: year)
  color (dim: color)
- For each dimension value
  there are multiple facts!

➢ More detail outside in!



DIMENSIONS MAY HAVE “GROUPINGS”



FACTS HAVE A MEASURE AND GRANULARITY

Fact has two components

- Numerical property (**measure**)
- Combination formula (e.g., aggregate like SUM)

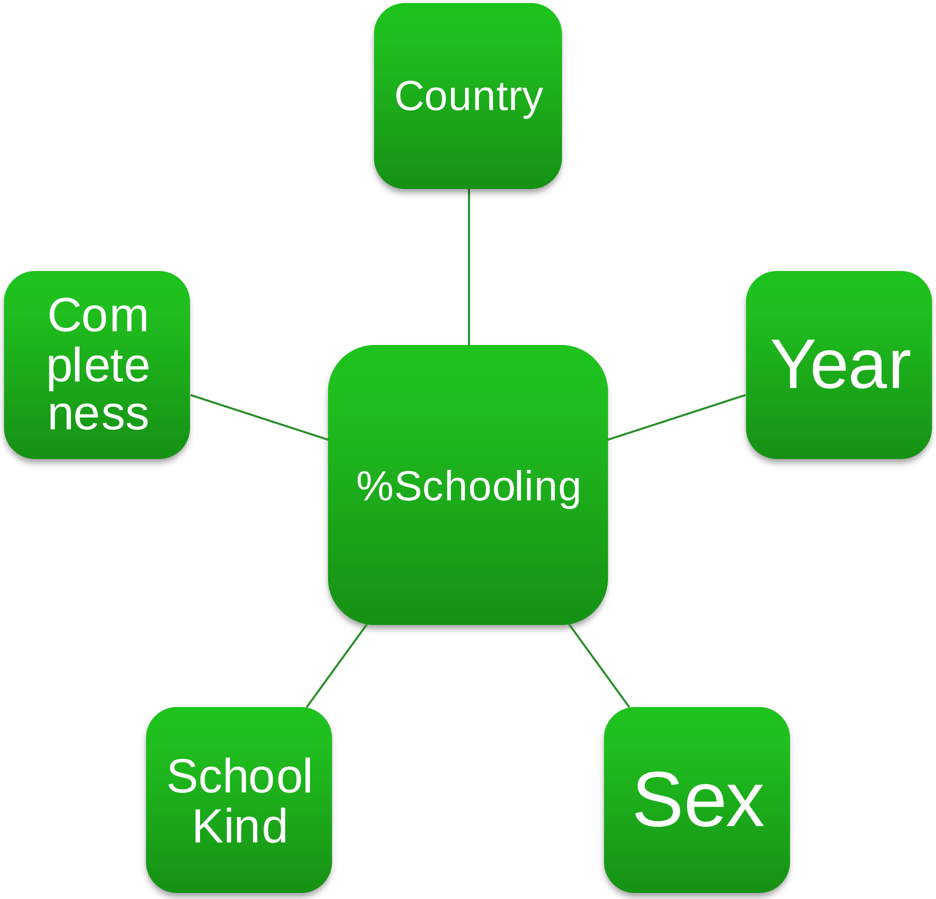
Facts have a certain **granularity**

- **Sales** by *month* by *make* by *color*
  (**fact** by *dim1* by *dim2* by *dim3* …)
- **Sales** by *year* by *category*

Second is **coarser**; first is **finer**

CONCEPTUAL DESIGN

EXAMPLE: %SCHOOLING IN THE WORLD



Target shape

Not a separate dimension, but grouping on Country

LOGICAL DESIGN: FULLY INLINED

INLINING A DIMENSION: IN THE SAME TABLE AS THE FACT

%Schooling	Country	Continent	Sex	School kind	Completeness	Year
11	Albania	Europe	Male	Primary	Yes	2013
12	Albania	Europe	Female	Primary	Yes	2013
8	Albania	Europe	Male	Secondary	Yes	2013
9	Albania	Europe	Female	Secondary	Yes	2013
19	Brazil	South America	Male	Primary	Yes	2013
23	Brazil	South America	Female	Primary	Yes	2013
2	Brazil	South America	Male	Secondary	Yes	2013
1	Brazil	South America	Female	Secondary	Yes	2013
1	Brazil	South America	Male	Primary	No	2013
1	Brazil	South America	Female	Primary	No	2013

Fact

Dim 1

Dim 2

Dim 3

Dim 4

Dim 5

This is a different **representation** of the “same cube”

LOGICAL DESIGN: PARTIALLY INLINED

COUNTRY DIMENSION AS A SEPARATE TABLE

%Schooling	CountryID	Sex	School kind	Completeness	Year
11	1	Male	Primary	Yes	
12	1	Female	Primary	Yes	
8	1	Male	Secondary	Yes	
9	1	Female	Secondary	Yes	
19	2	Male	Primary	Yes	
23	2	Female	Primary	Yes	
2	2	Male	Secondary	Yes	
1	2	Female	Secondary	Yes	2013
1	2	Male	Primary	No	2013
1	2	Female	Primary	No	2013

Fact | Dim 1 | Dim 2 | Dim 3 | Dim 4 | Dim 5

Dim 1

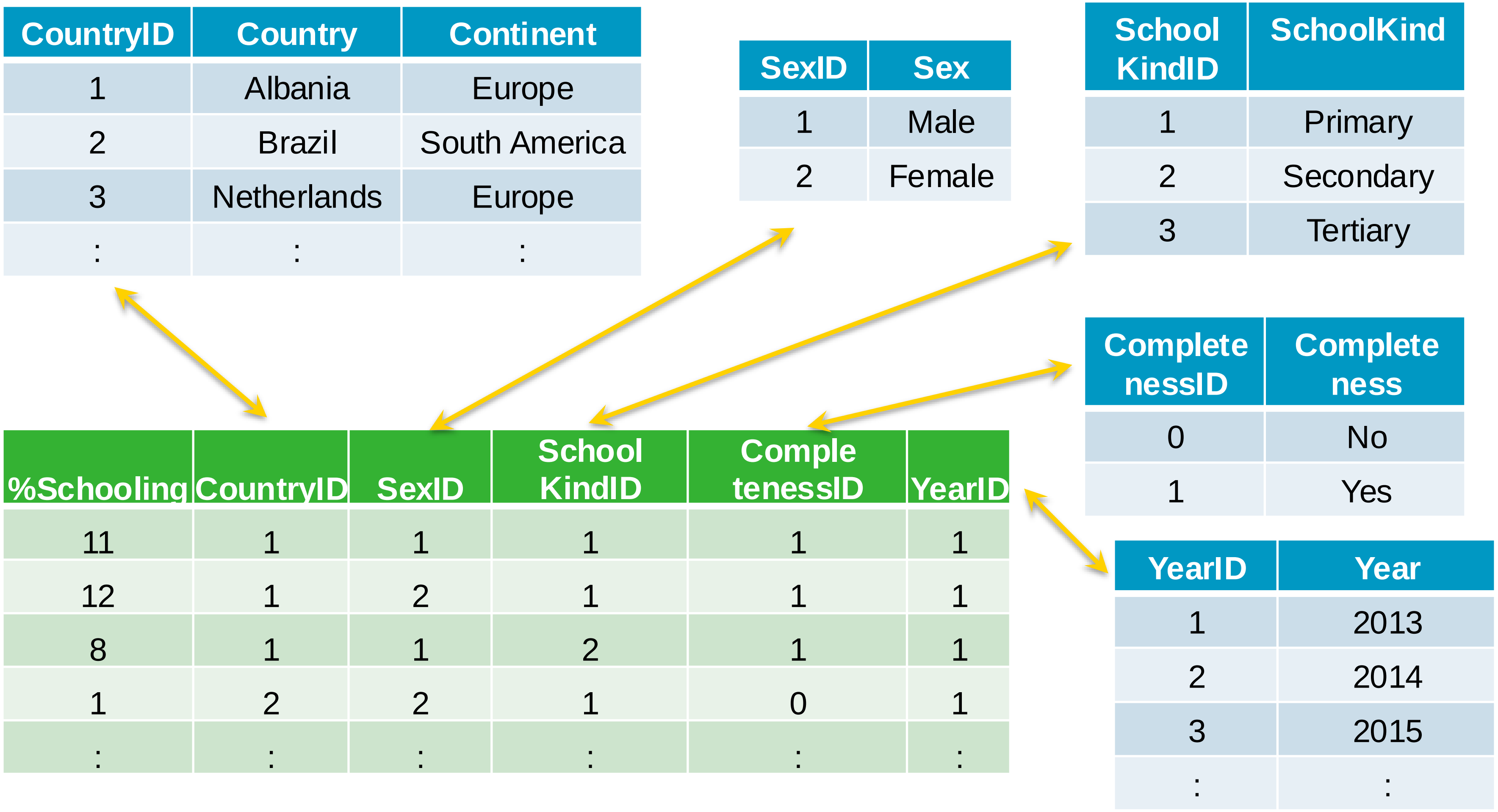
CountryID	Country	Continent
1	Albania	Europe
2	Brazil	South America
3	Netherlands	Europe
4	Ghana	Africa

Separate dimension table for “Country”

Yet another **representation** of the “same cube”

LOGICAL DESIGN: FULLY NORMALIZED

NORMALIZED: ALL DIMENSIONS ARE SEPARATE TABLES



CountryID	Country	Continent
1	Albania	Europe
2	Brazil	South America
3	Netherlands	Europe
:	:	:

SexID	Sex
1	Male
2	Female

SchoolKindID	SchoolKind
1	Primary
2	Secondary
3	Tertiary

CompletenessID	Completeness
0	No
1	Yes

%Schooling	CountryID	SexID	SchoolKindID	CompletenessID	YearID
11	1	1	1	1	1
12	1	2	1	1	1
8	1	1	2	1	1
1	2	2	1	0	1
:	:	:	:	:	:

YearID	Year
1	2013
2	2014
3	2015
:	:

MULTIDIMENSIONAL MODELING

METHOD FOR DATA PREPARATION

Start

1. Conceptual design: Design cube (star schema)
   a) Determine questions the data should answer
   b) Envision tabular reports that may answer those questions
   c) Determine for each question and report, the fact, the dimensions, and granularity
   d) Combine into one star schema
   e) Formulate what one row in fact table means
2. Logical design: Design associated table structure
3. Realization: Prepare data & Create tables & fill them

(a) Determine questions the data should answer

MULTIDIMENSIONAL MODELING EXAMPLE

ORCHARD

Large industrial orchard grows several fruits (apples, oranges, etc.) on many fields. Harvested fruits are automatically filtered for bad fruits before being sold. Orchard management wants to quickly and effectively determine the bad fields and weak fruits. Moreover, they like to analyze the effect of improvements.

Question(s)

- Determine bad fields and weak fruits
- Effects of improvements

(a) Determine questions the data should answer

(b) Envision tabular reports that may answer those questions

MULTIDIMENSIONAL MODELING EXAMPLE

ORCHARD

Large industrial orchard grows several fruits (apples, oranges, etc.) on many fields. Harvested fruits are automatically filtered for bad fruits before being sold. Orchard management wants to quickly and effectively determine the bad fields and weak fruits. Moreover, they like to analyze the effect of improvements.

Question(s)

- Determine bad fields and weak fruits
- Effects of improvements

Field	Fruit	Date	Condition	Harvest
A	Apples	1 Sep	Good	1400 kg
A	Apples	1 Sep	Bad	200 kg
A	Bananas	1 Sep	Good	800 kg
B	Apples	1 Sep	Bad	1900 kg

(c) Determine for each question and report, the fact, the dimensions, and granularity

MULTIDIMENSIONAL MODELING EXAMPLE

ORCHARD

Large industrial orchard grows several fruits (apples, oranges, etc.) on many fields. Harvested fruits are automatically filtered for bad fruits before being sold. Orchard management wants to quickly and effectively determine the bad fields and weak fruits. Moreover, they like to analyze the effect of improvements.

Field	Fruit	Date	Condition	Harvest
A	Apples	1 Sep	Good	1400 kg
A	Apples	1 Sep	Bad	200 kg
A	Bananas	1 Sep	Good	800 kg
B	Apples	1 Sep	Bad	1900 kg

Dimensions

- Field, Fruit, Date, Condition

Fact

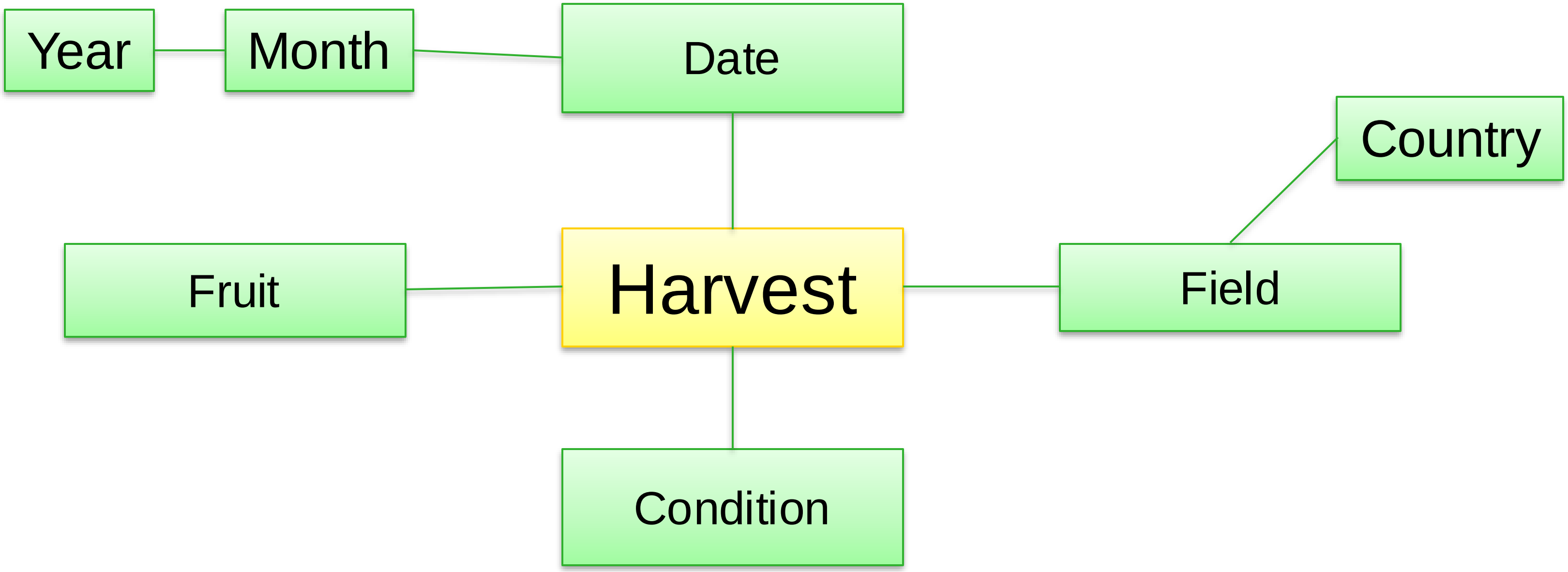
- Harvest (weight)

(d) Combine into one star schema

MULTIDIMENSIONAL MODELING EXAMPLE

ORCHARD

Large industrial orchard grows several fruits (apples, oranges, etc.) on many fields. Harvested fruits are automatically filtered for bad fruits before being sold. Orchard management wants to quickly and effectively determine the bad fields and weak fruits. Moreover, they like to analyze the effect of improvements.



(e) Formulate what one row in fact table means

MULTIDIMENSIONAL MODELING EXAMPLE

ORCHARD

Large industrial orchard grows several fruits (apples, oranges, etc.) on many fields. Harvested fruits are automatically filtered for bad fruits before being sold. Orchard management wants to quickly and effectively determine the bad fields and weak fruits. Moreover, they like to analyze the effect of improvements.

Dimensions

- Field, Fruit, Date, Condition

Fact

- Harvest (weight)

What does this mean?

- **For each** time unit (say, day), **we store** the total weight of the harvest **per** fruit **for** bad and good fruit **seperately per** field.

RULES OF THUMB

- Focus on the questions
  (don’t be distracted by source data structure)
  - Dimensions:
    look for ‘aspects’ and formulations like “per X”; determine required granularity
  - Fact:
    on what numbers would an answer/report be based?

- Checks you can do afterwards
  - Dimensions are (almost always) independent
  - For all combinations of values of the dimensions, you (potentially) have one fact
  - Can all questions be answered?

MULTIDIMENSIONAL MODELING EXAMPLE 2

AUDIO/VIDEO SALES

Director of chain of high-end audio/video shops wants to know per month and city how many sales come from customers in the same city as the shop vs. sales from customers coming from other cities. He needs this to decide if he needs to open shops in all major cities or that customers are willing to travel to go to his shops.

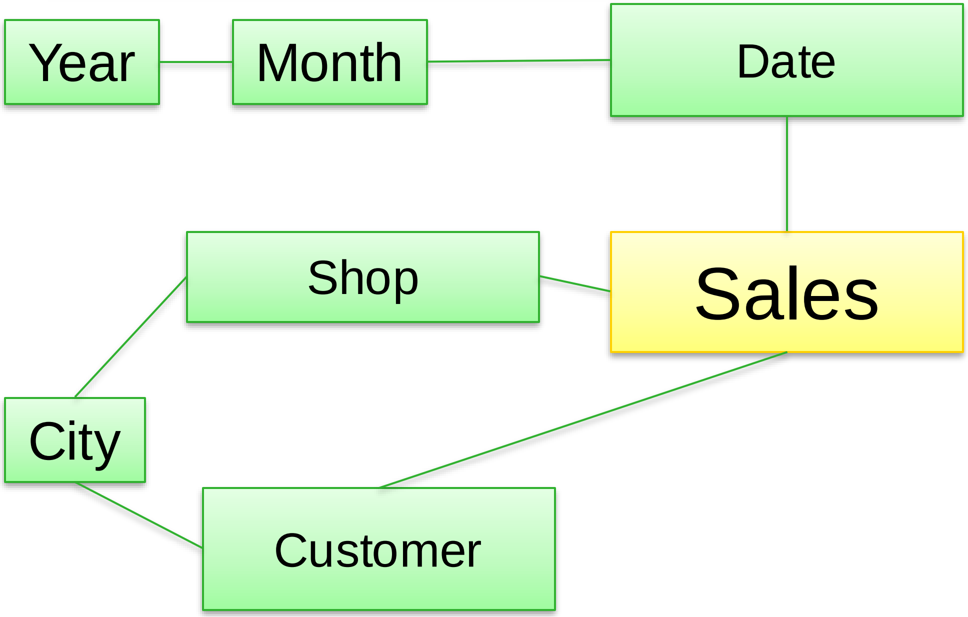
Assignment:

- I will make initial design
- You tell me what I did wrong

MULTIDIMENSIONAL MODELING EXAMPLE 2

AUDIO/VIDEO SALES

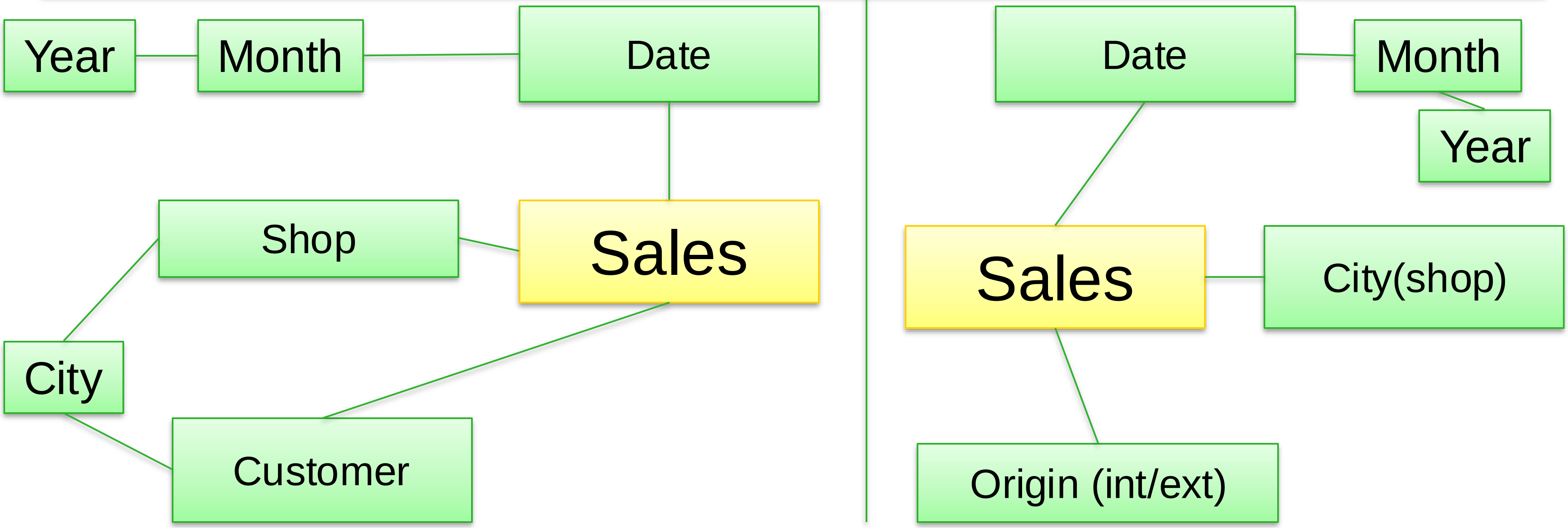
Director of chain of high-end audio/video shops wants to know per month and city how many sales come from customers in the same city as the shop vs. sales from customers coming from other cities. He needs this to decide if he needs to open shops in all major cities or that customers are willing to travel to go to his shops.



MULTIDIMENSIONAL MODELING EXAMPLE 2

AUDIO/VIDEO SALES

Director of chain of high-end audio/video shops wants to know per month and city how many sales come from customers in the same city as the shop vs. sales from customers coming from other cities. He needs this to decide if he needs to open shops in all major cities or that customers are willing to travel to go to his shops.



MULTIDIMENSIONAL MODELING EXAMPLE 2

SOFTWARE LICENCES

Company spends much money on licenses for software. You start paying when you open software and stop when you terminate it. Software use is inter-active or running (e.g., simulation), but software can also be idle. Mgmt wants to know if they pay a lot of money of started software per category that is idle for a long time.

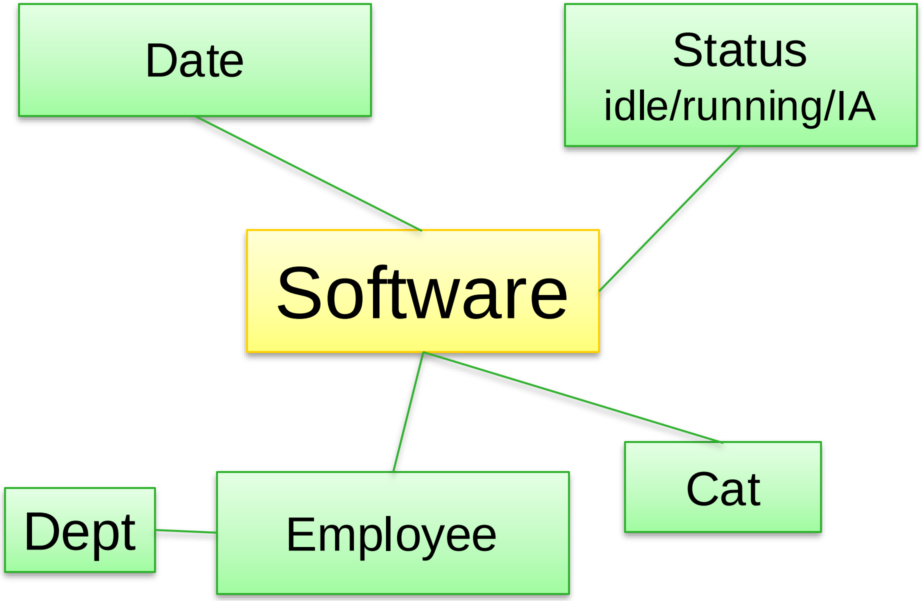
Assignment:

- I will make initial design
- You tell me what I did wrong

MULTIDIMENSIONAL MODELING EXAMPLE 2

SOFTWARE LICENCES

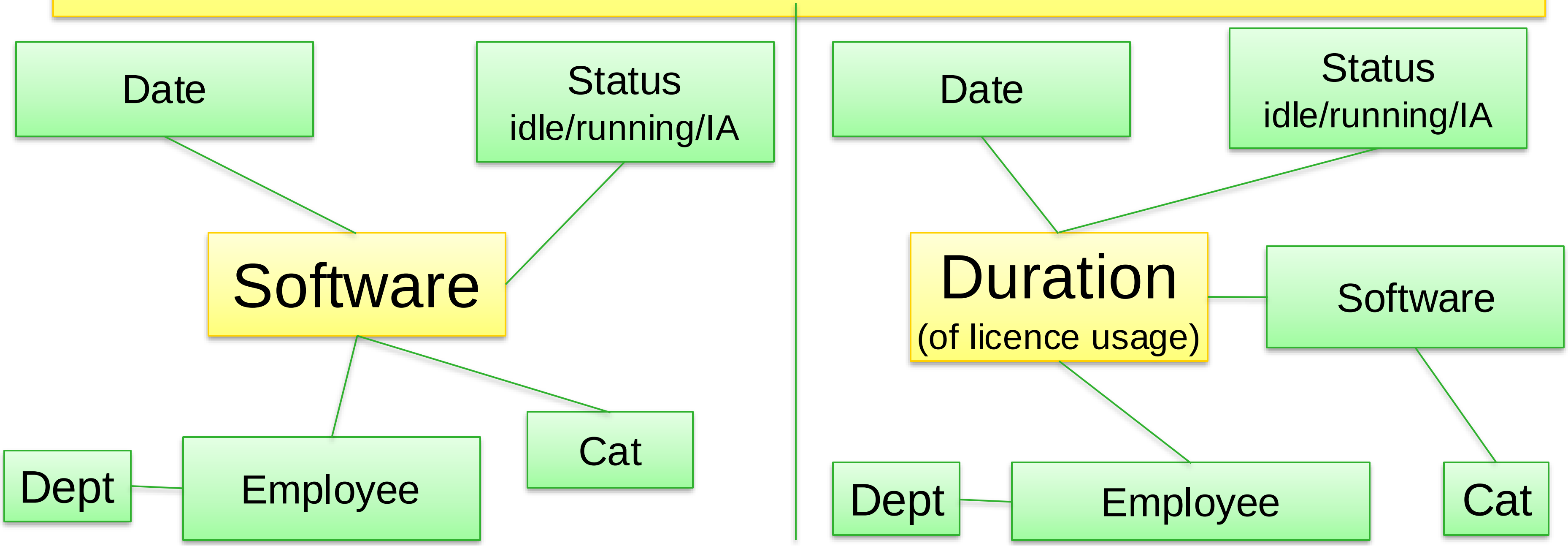
Company spends much money on licences for software. You start paying when you open software and stop when you terminate it. Software use is inter-active or running (e.g., simulation), but software can also be idle. Mgmt wants to know if they pay a lot of money of started software per category that is idle for a long time.



MULTIDIMENSIONAL MODELING EXAMPLE 2

SOFTWARE LICENCES

Company spends much money on licences for software. You start paying when you open software and stop when you terminate it. Software use is inter-active or running (e.g., simulation), but software can also be idle. Mgmt wants to know if they pay a lot of money of started software per category that is idle for a long time.



METHOD FOR DATA PREPARATION

Done

1. Conceptual design: Design cube (star schema)
   a) Determine questions the data should answer
   b) Envision tabular reports that may answer those questions
   c) Determine for each question and report, the fact, the dimensions, and granularity
   d) Combine into one star schema
   e) Formulate what one row in fact table means

Now

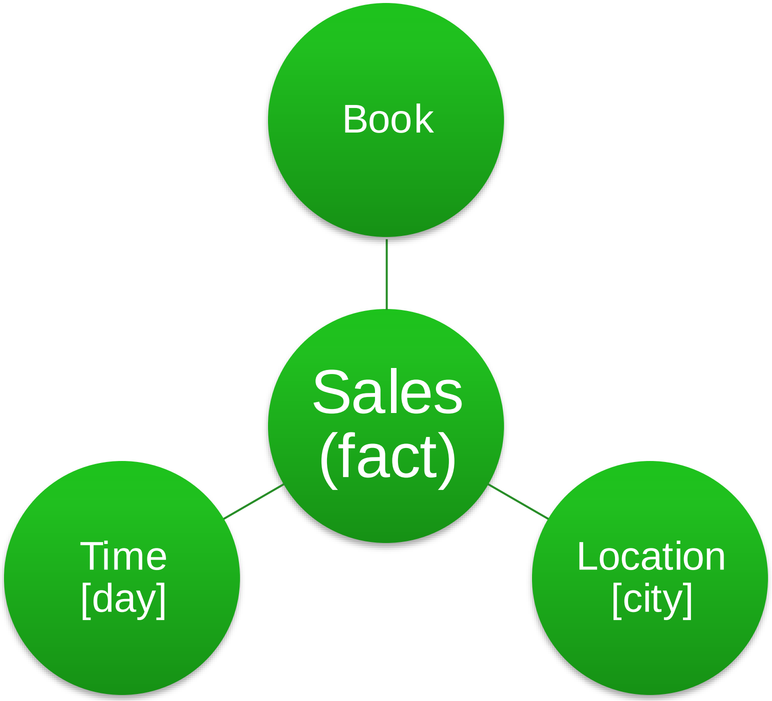
2. Logical design: Design associated table structure
3. Realization: Prepare data & Create tables & fill them

STAR SCHEMA IS THE CONCEPTUAL DESIGN

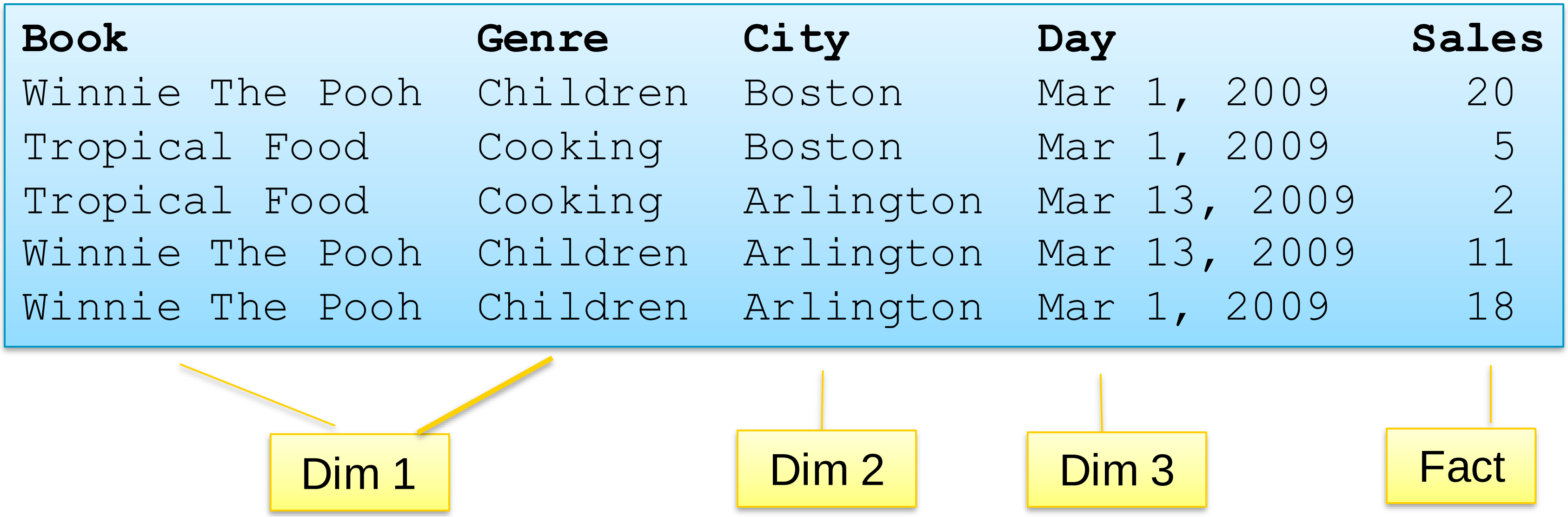
CONCEPTUAL DESIGN: FACT + DIMENSIONS ONLY

- One row of sales
  Per combination of
  Book, Time Unit, Location
- Attributes for
  sales (fact: amount)
  book (dim: name, category)
  time (dim: day)
  location (dim: city, country)
- For each dimension value
  there are multiple facts!

➢ More detail outside in!



REALISING A DATA CUBE WITH ONE TABLE



Book	Genre	City	Day	Sales
Winnie The Pooh	Children	Boston	Mar 1, 2009	20
Tropical Food	Cooking	Boston	Mar 1, 2009	5
Tropical Food	Cooking	Arlington	Mar 13, 2009	2
Winnie The Pooh	Children	Arlington	Mar 13, 2009	11
Winnie The Pooh	Children	Arlington	Mar 1, 2009	18

Dim 1

Dim 2

Dim 3

Fact

REALISING A DATA CUBE WITH ONE TABLE

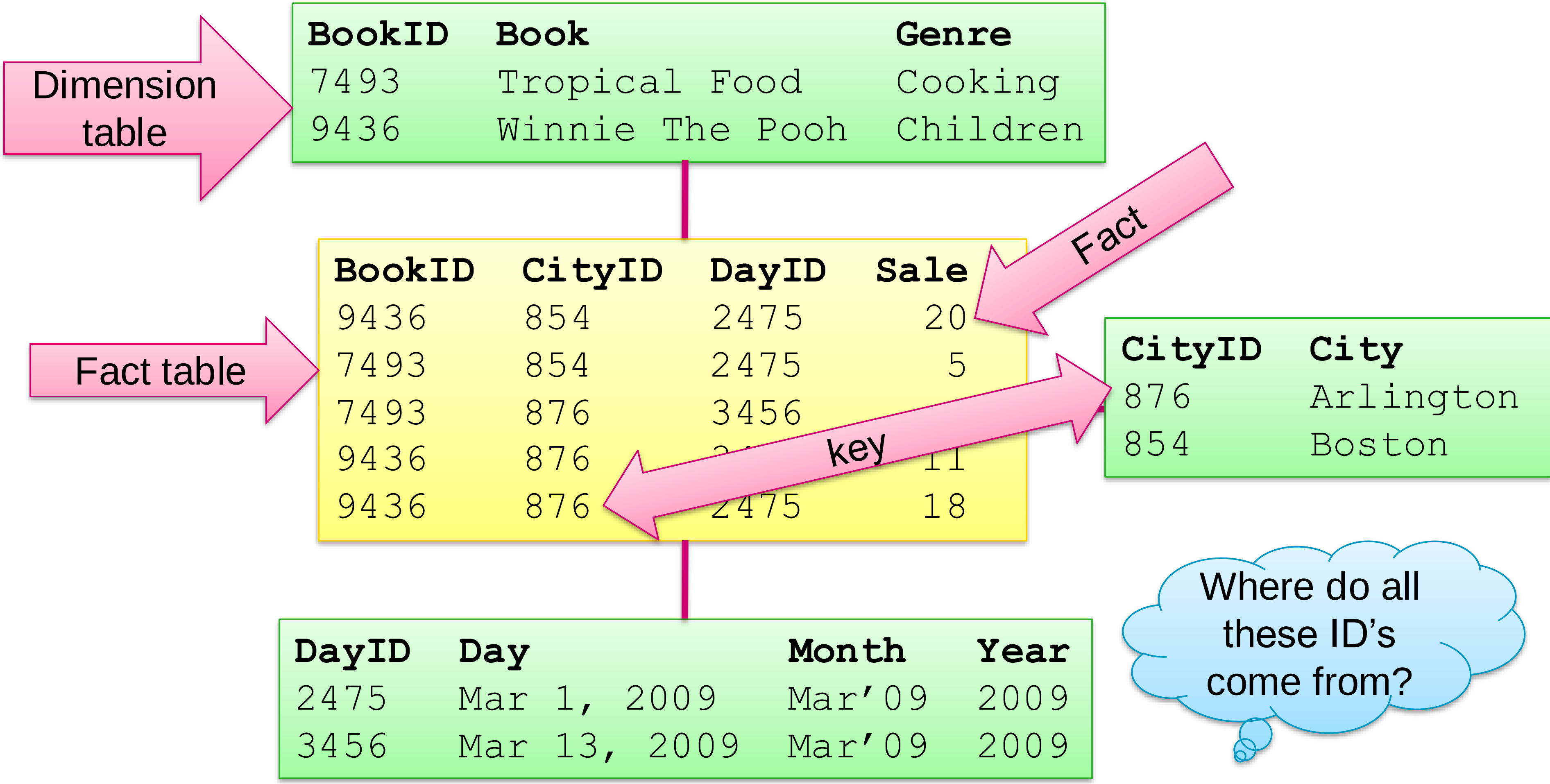
Logical design with the inlined approach (one table with 5 inlined attributes)

Sales	
Book	TEXT
Genre	TEXT
City	TEXT
Day	DATE
Sales	DOUBLE

```
Book              Genre      City       Day            Sales
Winnie The Pooh   Children   Boston     Mar 1, 2009       20
Tropical Food     Cooking    Boston     Mar 1, 2009        5
Tropical Food     Cooking    Arlington  Mar 13, 2009       2
Winnie The Pooh   Children   Arlington  Mar 13, 2009      11
Winnie The Pooh   Children   Arlington  Mar 1, 2009       18
```

REALISING A DATA CUBE WITH RELATIONAL TABLES

THIS IS EXACTLY THE SAME DATA; JUST A DIFFERENT REPRESENTATION



BookID	Book	Genre
7493	Tropical Food	Cooking
9436	Winnie The Pooh	Children

BookID	CityID	DayID	Sale
9436	854	2475	20
7493	854	2475	5
7493	876	3456	[illegible]
9436	876	[illegible]	11
9436	876	2475	18

CityID	City
876	Arlington
854	Boston

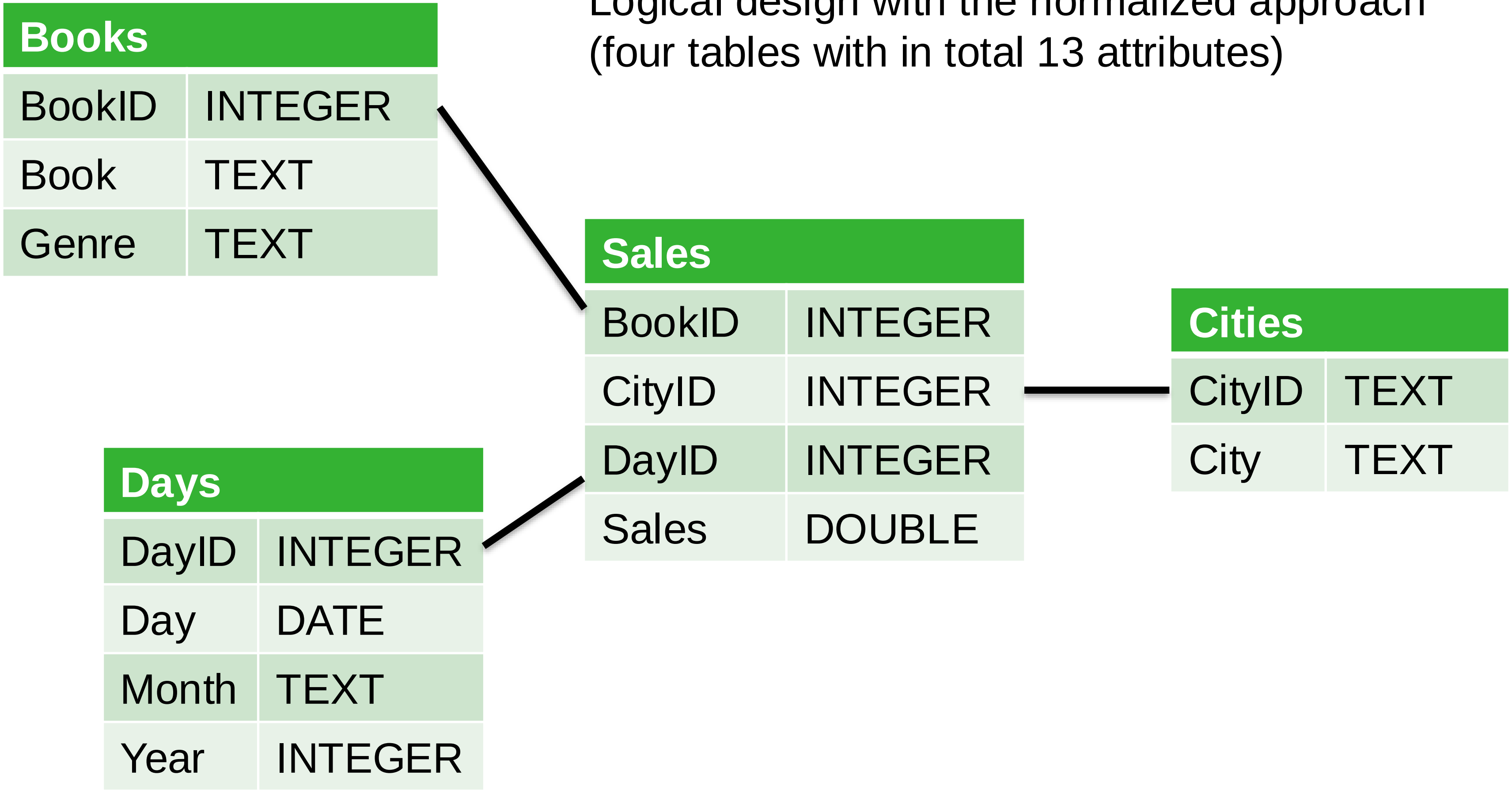
DayID	Day	Month	Year
2475	Mar 1, 2009	Mar'09	2009
3456	Mar 13, 2009	Mar'09	2009

Where do all these ID's come from?

REALISING A DATA CUBE WITH DIMENSION TABLES

THIS IS EXACTLY THE SAME DATA; JUST A DIFFERENT REPRESENTATION

Logical design with the normalized approach (four tables with in total 13 attributes)



Books	
BookID	INTEGER
Book	TEXT
Genre	TEXT

Sales	
BookID	INTEGER
CityID	INTEGER
DayID	INTEGER
Sales	DOUBLE

Cities	
CityID	TEXT
City	TEXT

Days	
DayID	INTEGER
Day	DATE
Month	TEXT
Year	INTEGER

LOGICAL DESIGN FOR A CUBE

TWO EXTREMES: NORMALIZED AND INLINED

Normalized design

- Each dimension is a separate table
  - Attributes: dim-id, dim-att$_1$, dim-att$_2$, …
  - dim$_i$ in fact table is foreign key to dim-id

Inlined design

- dim$_i$ in fact table is directly dim-att or possibly several columns dim-att$_1$, dim-att$_2$, …

Choose inlined design for a dimension if

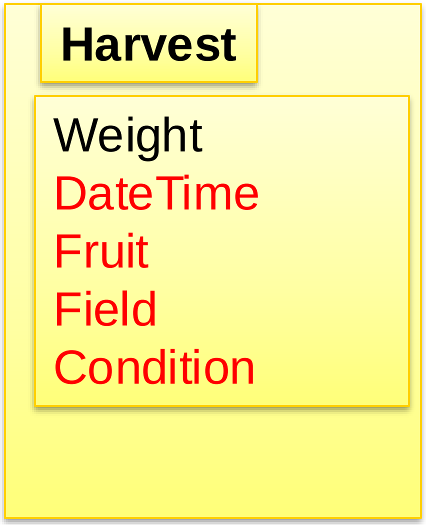
- Not many possible values; values are short strings
- Dimension is not re-used in other cubes
- Dimension value itself is an identifier (e.g., date) Grouping may be computable (e.g., month, year)

TABLE DESIGN EXAMPLE: FULLY INLINED DESIGN

STILL 4 DIMENSIONS!

Fact table

Dimensions



Harvest

Weight
DateTime
Fruit
Field
Condition

Fruit and Field are large dimensions likely to be used in other cubes as well. Therefore, these are advised to keep as separate tables.

TABLE DESIGN EXAMPLE: NORMALIZING FRUIT AND FIELD; INLINING CONDITION AND DATETIME

STILL 4 DIMENSIONS!

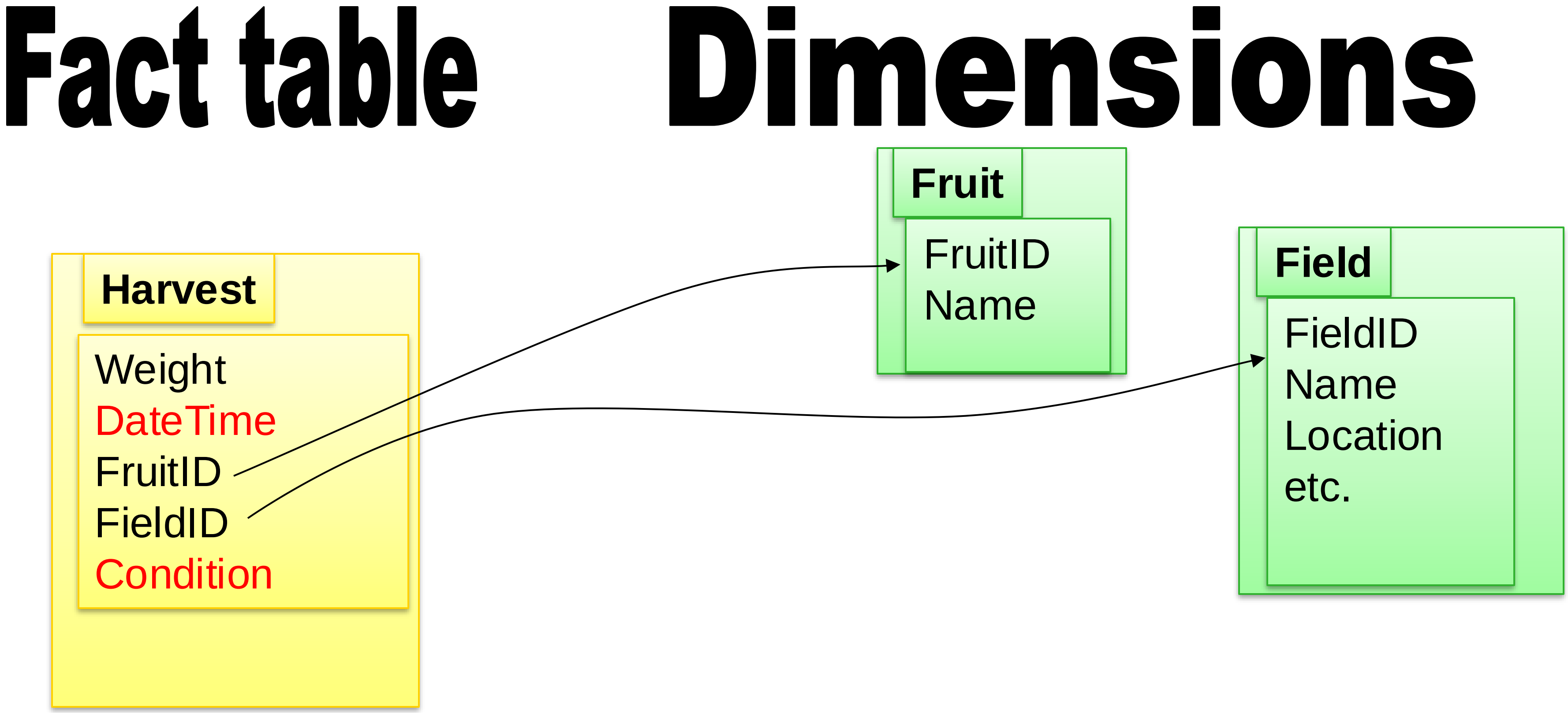
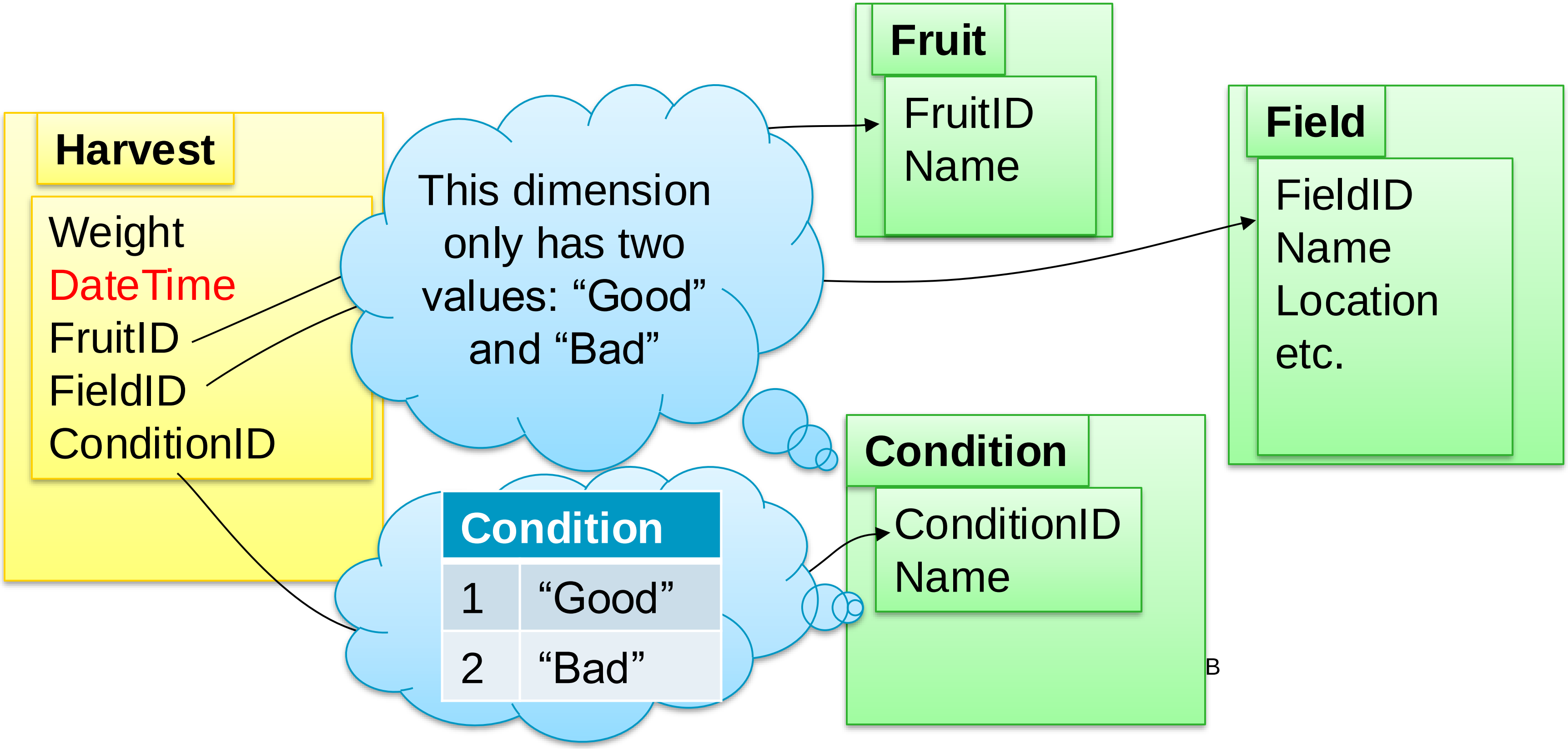


TABLE DESIGN EXAMPLE: NORMALIZING FRUIT, FIELD, AND CONDITION; INLINING DATETIME

STILL 4 DIMENSIONS!

Fact table   Dimensions



Harvest

Weight
DateTime
FruitID
FieldID
ConditionID

This dimension only has two values: “Good” and “Bad”

Condition	
1	“Good”
2	“Bad”

Fruit

FruitID
Name

Field

FieldID
Name
Location
etc.

Condition

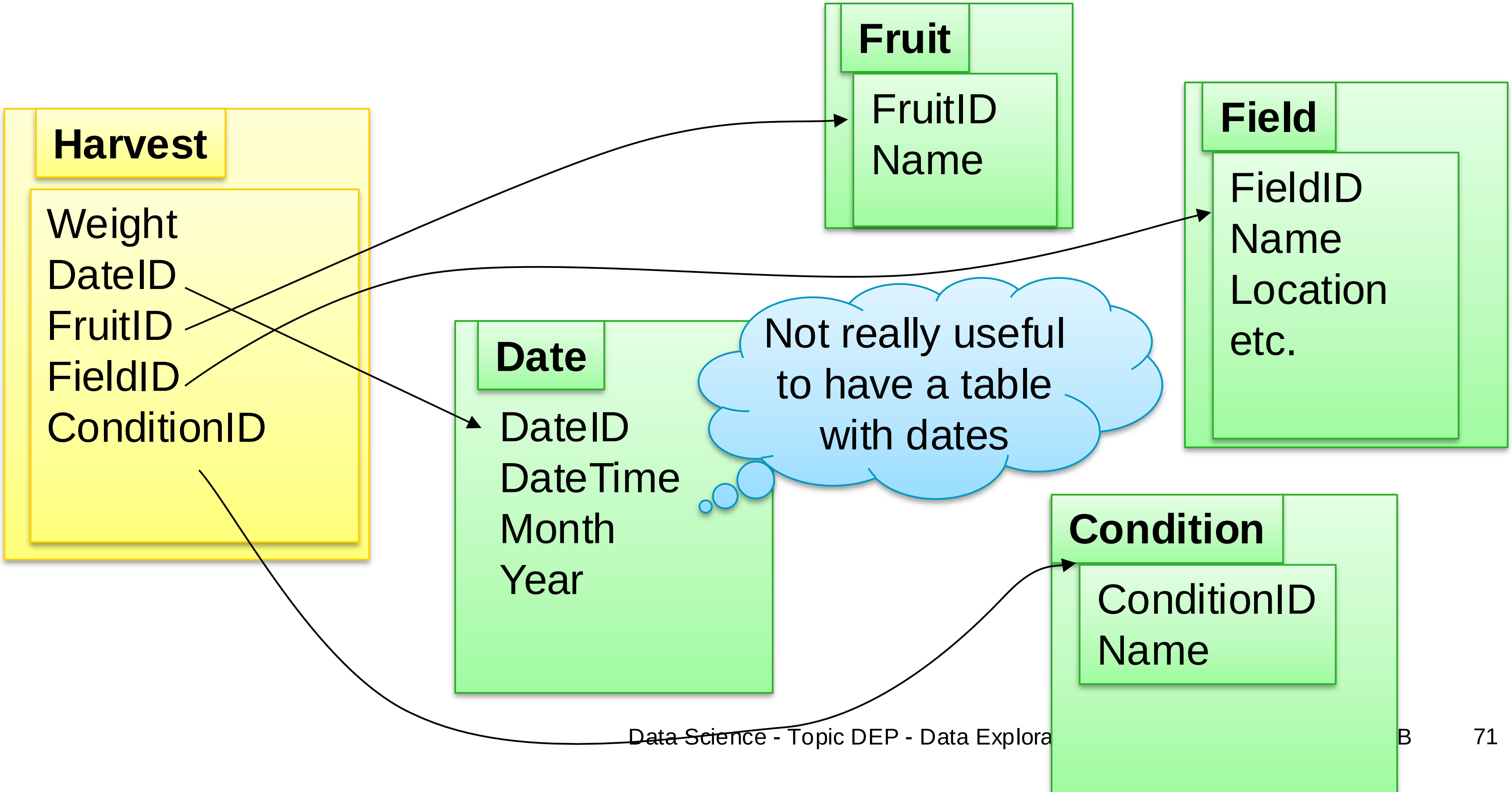
ConditionID
Name

B

TABLE DESIGN EXAMPLE: FULLY NORMALIZE DESIGN

ALL DIMENSIONS HAVE SEPARATE TABLES

Fact table **Dimensions**



Harvest

Weight
DateID
FruitID
FieldID
ConditionID

Fruit

FruitID
Name

Field

FieldID
Name
Location
etc.

Date

DateID
DateTime
Month
Year

Not really useful to have a table with dates

Condition

ConditionID
Name

METHOD FOR DATA PREPARATION

Done

1. Conceptual design: Design cube (star schema)
   - a) Determine questions the data should answer
   - b) Envision tabular reports that may answer those questions
   - c) Determine for each question and report, the fact, the dimensions, and granularity
   - d) Combine into one star schema
   - e) Formulate what one row in fact table means

Done

Now

2. Logical design: Design associated table structure
3. Realization: Prepare data & Create tables & fill them

REALIZATION OF LOGICAL DESIGN

Realize the logical design in a database

- Choose appropriate attribute types
- Create (empty) tables in the database
  - Directly with SQL commands
  - With a database adminstration client (e.g., phpPgAdmin)
  - As part of the data preparation program typically using functions from a package / library
    - Often writing a data frame to a non-existent table automatically creates the table
  - With an ETL or Data Wrangling tool

PREPARE & FILL TABLES = ACTUAL RESHAPING

Reshaping

- Reading sources
- Restructuring to match target table structure
- Data cleaning
- Writing to cube (i.e., the tables in the database)

How

- SQL (if both sources and target are databases)
- Any programming language
- ETL or Data Wrangling tool

What we do (with R or Python)

PREPARE & FILL TABLES = ACTUAL RESHAPING

Some advice on programming for data preparation

- **Small do-test steps**
  Do: Add only one or two small bits, then execute and verify the result, before continuing
  Do not: Add many steps and then don’t know where the mistake is when you receive an error
- **Read the error message carefully**
  It may contain a lot of gibberish you don’t understand, but part of it may provide clues to what is wrong
- **Google and AI’s are your friends**
  You may think Googling is not academic, but the internet is full of information on what may have caused certain errors and what you can do to fix them. AI’s like ChatGPT ‘studied’ all this information and can be your personal teacher.

TOPIC ASSIGNMENTS & NEXT LECTURES

DM also uses R or Python

Topic assignments (4)

- 1: on-paper assignment on cube concepts
- 2&3: assignments on data exploration & preparation (R or Python)
- 4 on-paper assignment on multidimensional modeling

You only need to participate if you have little experience with R or Python

Topic “zero”

- Lecture and assignments: introduction to programming

In the project, use any programming language you like; We just prepared topic assignments for R and Python

TAKE AWAY MESSAGE

Given real-world challenge with real-world data …
(especially if DEP is your primary topic for the project)

What do you do?

- Explore your source data; critically look for DQ issues
- Use the method of multidimensional modeling!
  - Think! What should the data answer => design cube
  - Do! Convert + clean data => store in cube in DBMS
- This will give you high quality data in a shape suitable for analytical purposes:
  - Visualization, Data Mining, Machine Learning, etc.