

# **Introduction to Data, Pandas and SQL**

## **Part A – Data and Databases**

Pavlos Protopapas

# Lecture Outline

---

## **Part A: Data and Databases**

What is data and how can we store it?

## **Part B: Pandas and SQL**

Tools to inspect data

# Lecture Outline

---

## **Part A: Data and Databases**

What is data and how can we store it?

## **Part B: Pandas and SQL**

Tools to inspect data

# The Data Science Process

**Ask an interesting question**

Get the Data

Explore the Data

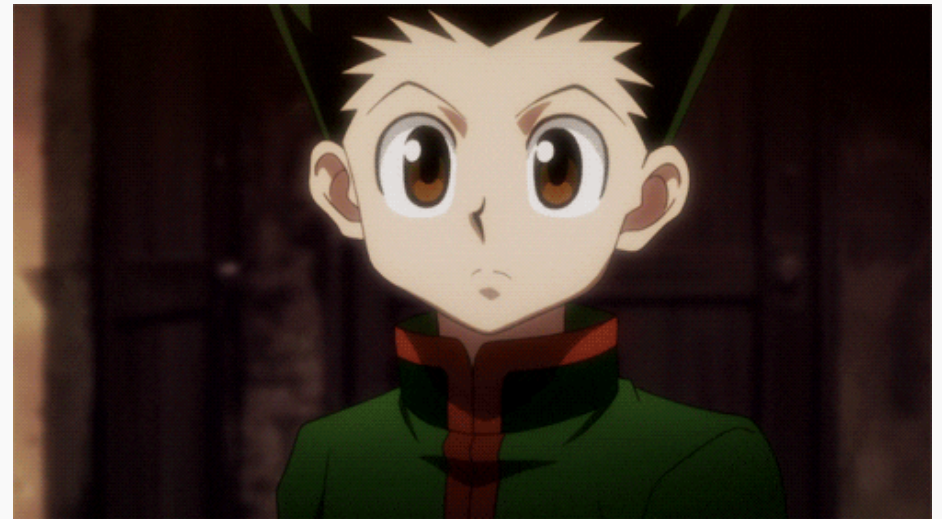
Model the Data

Communicate/Visualize the  
Results

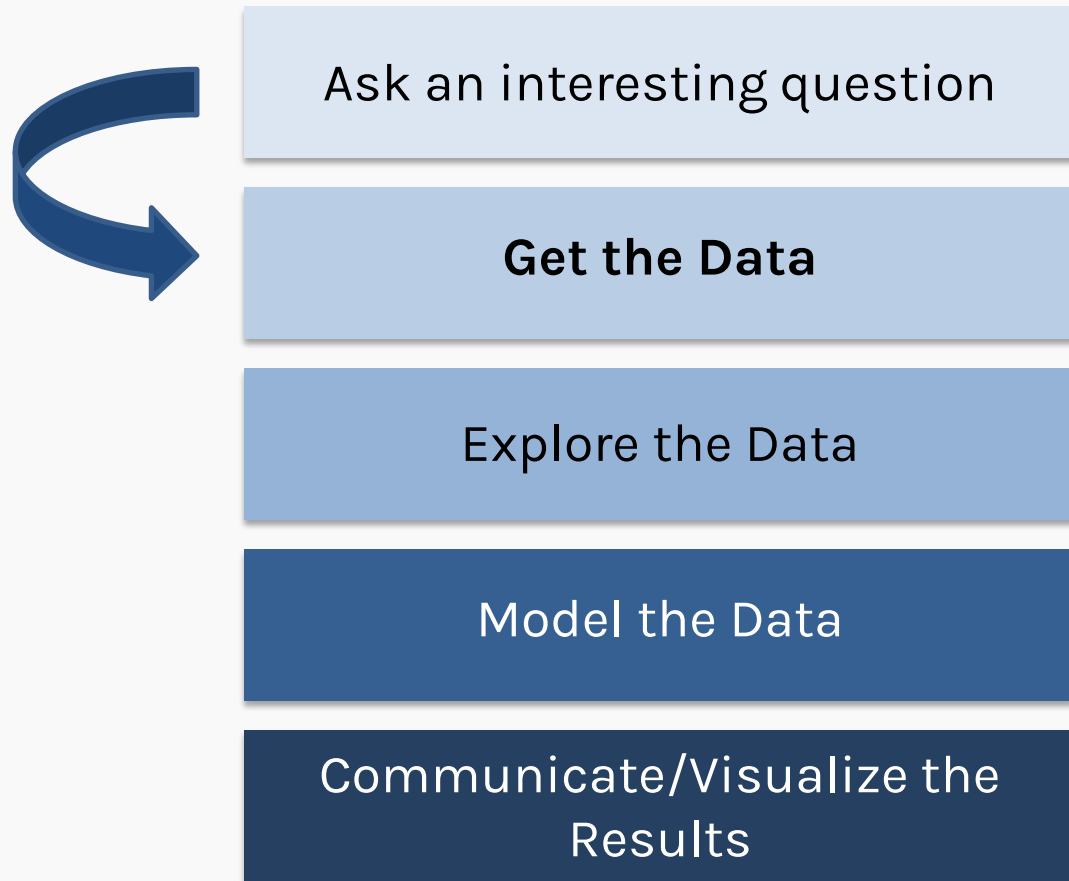
What is the scientific goal?

What would you do if you had all of the  
data?

What do you want to predict or estimate?



# The Data Science Process



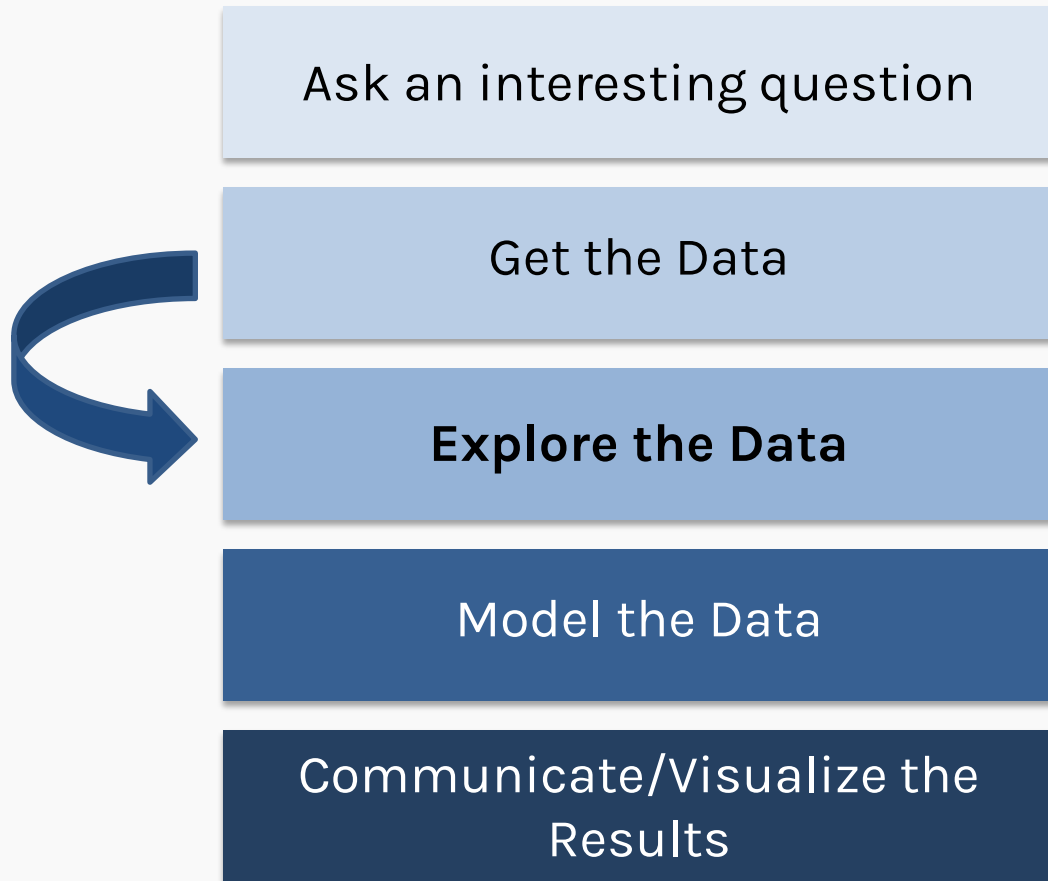
How were the data sampled?

Which data are relevant?

Are there privacy issues?



# The Data Science Process

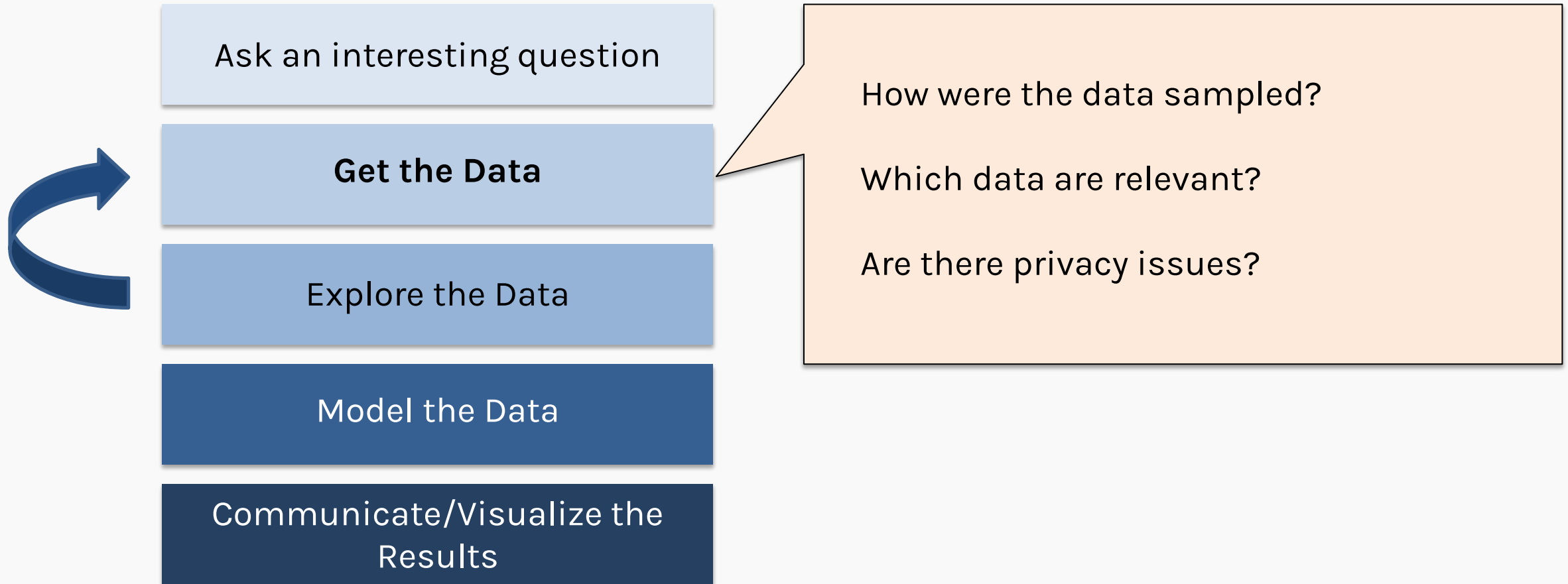


Plot the data.

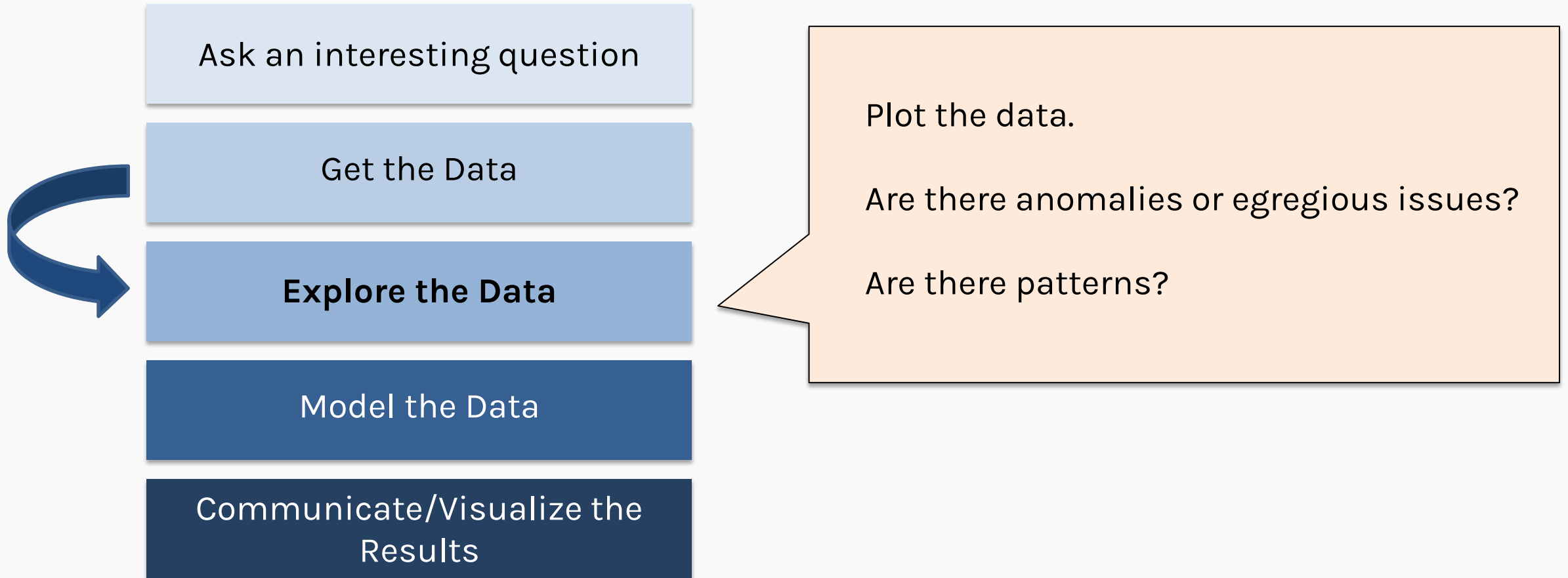
Are there anomalies or egregious issues?

Are there patterns?

# The Data Science Process

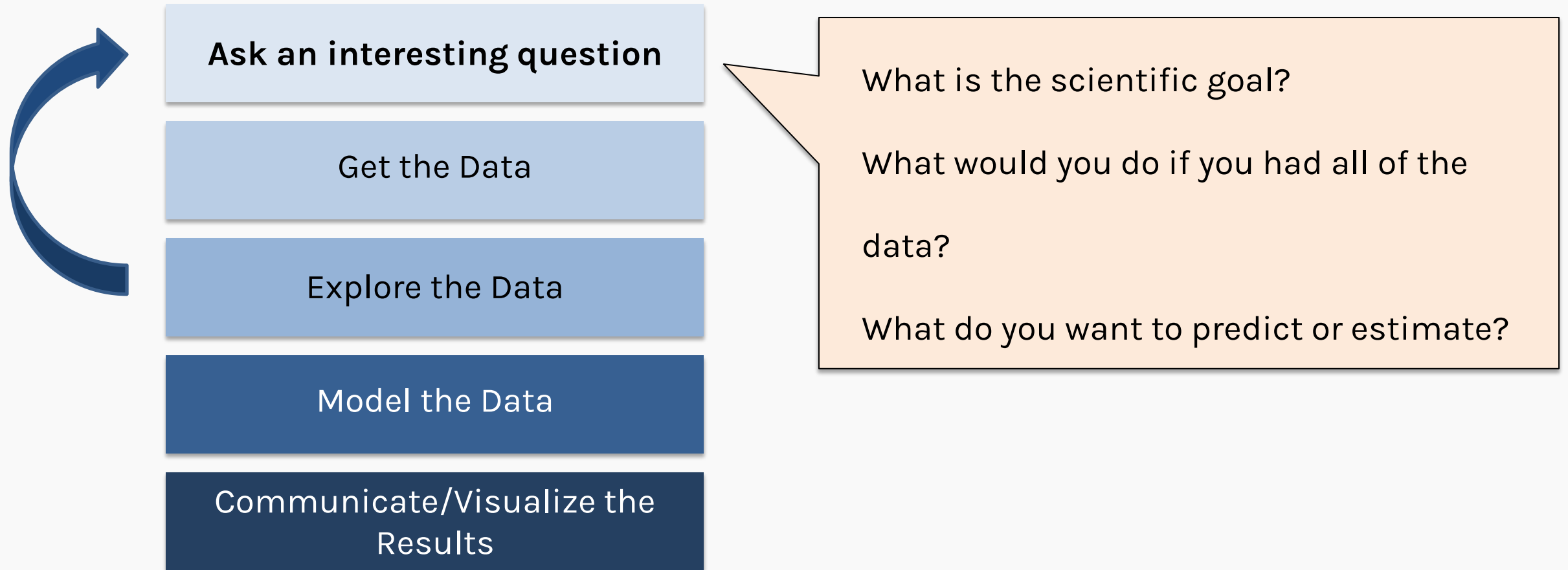


# The Data Science Process

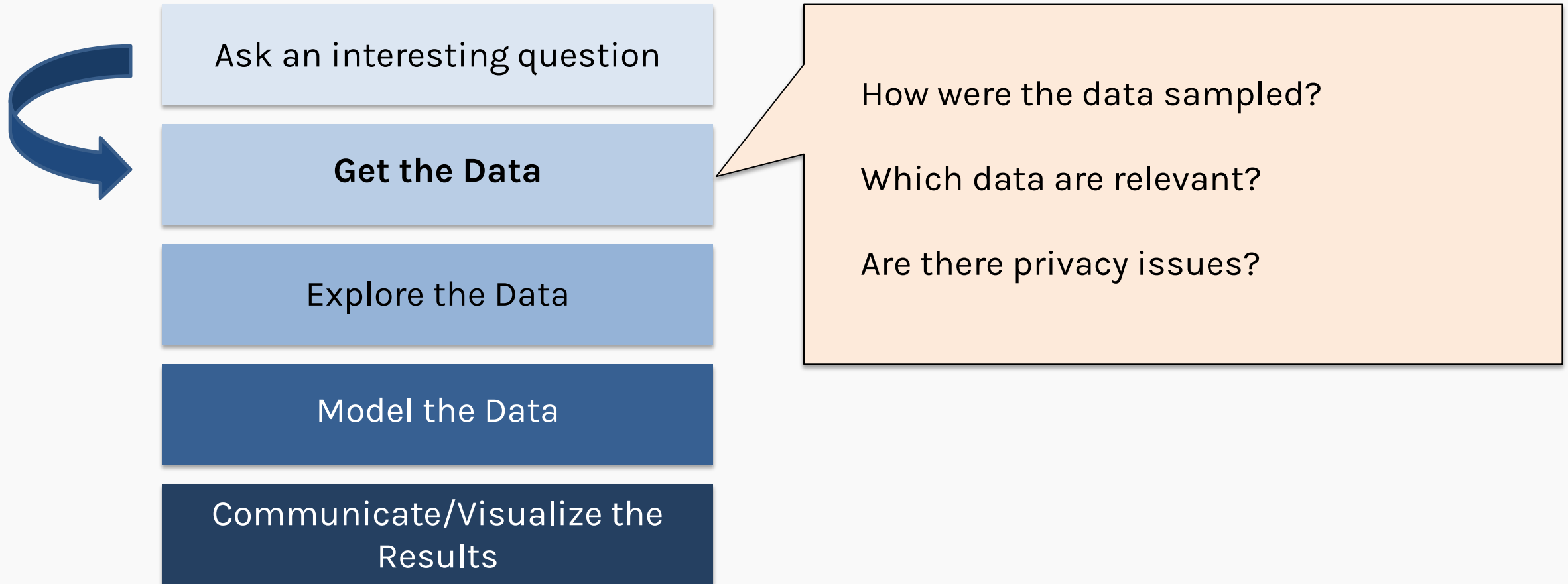




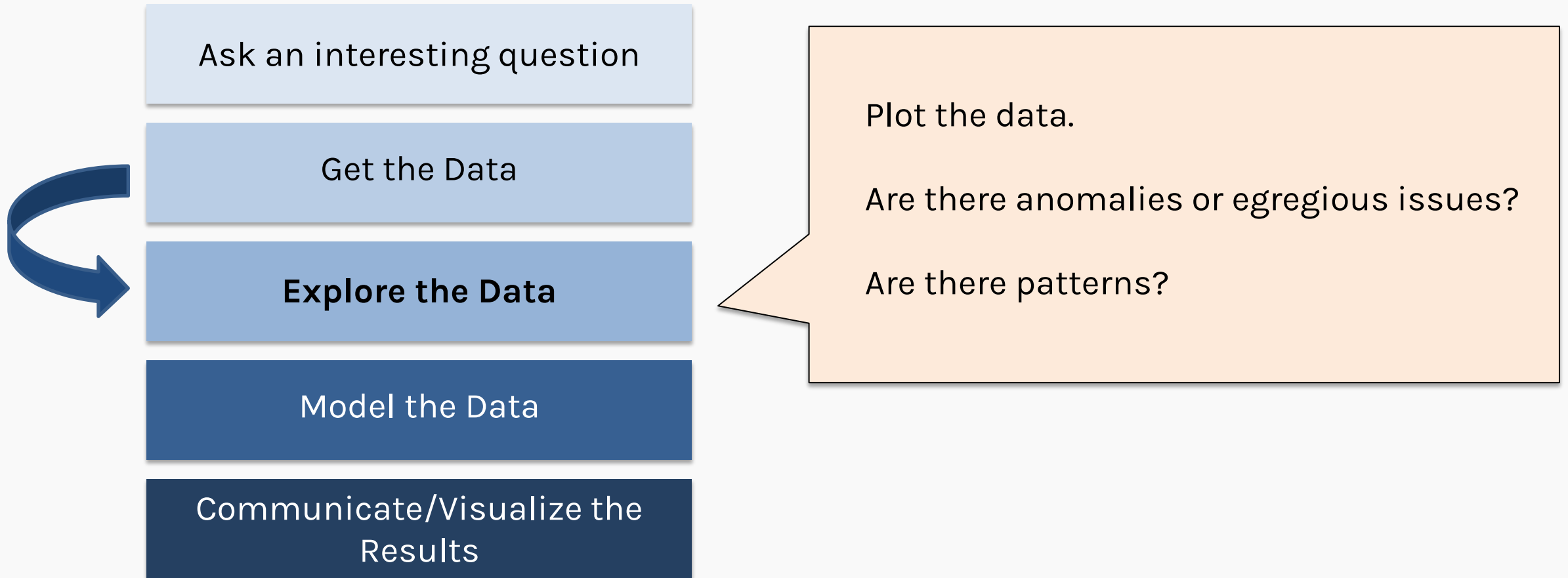
# The Data Science Process



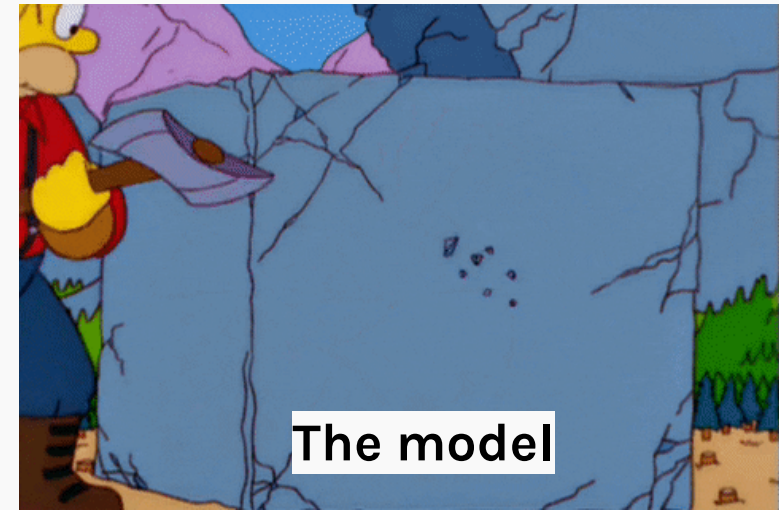
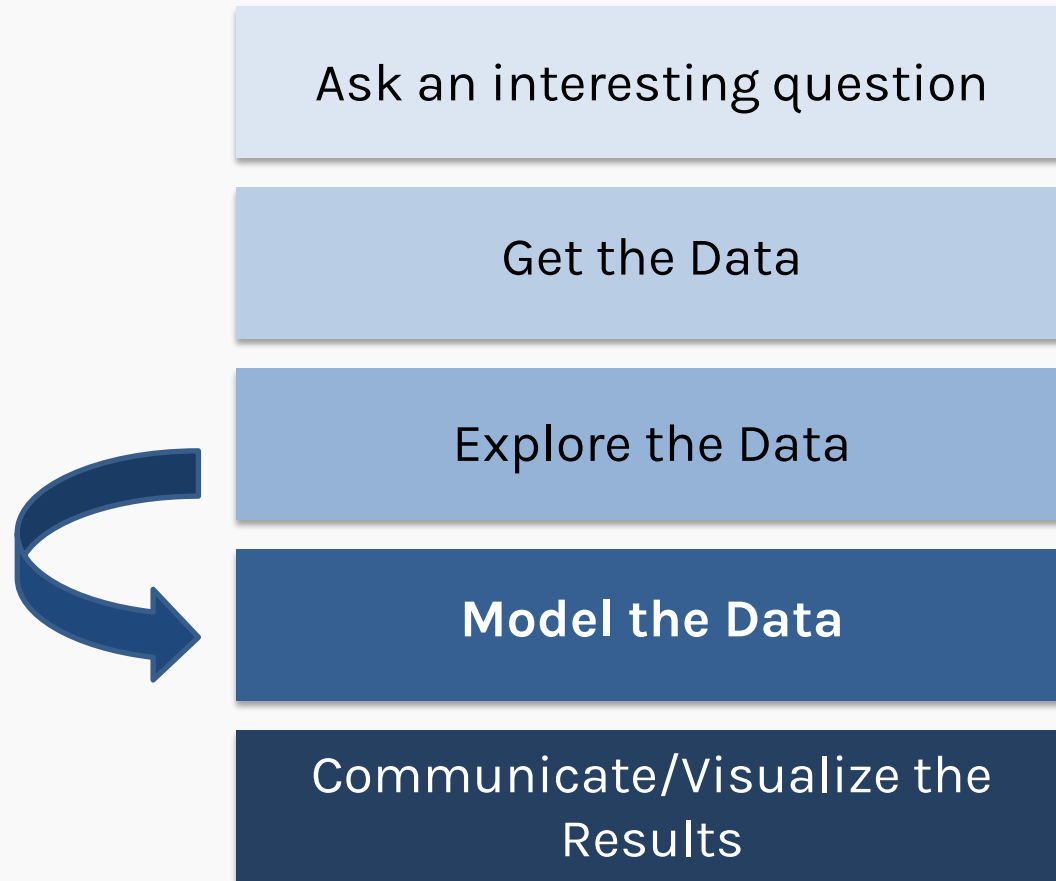
# The Data Science Process



# The Data Science Process



# The Data Science Process

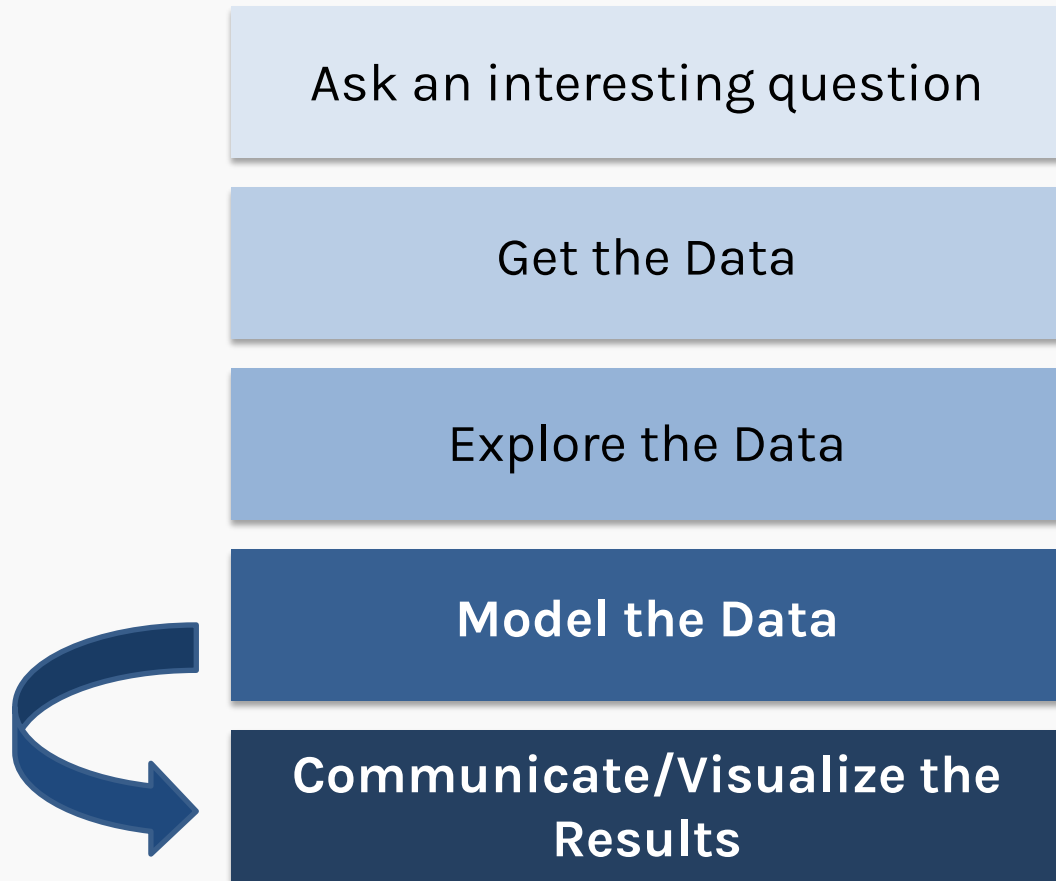


Build a model.

Fit the model.

Validate the model.

# The Data Science Process



What did we learn?

Do the results make sense?

Can we effectively tell a story?

# What is data?

---

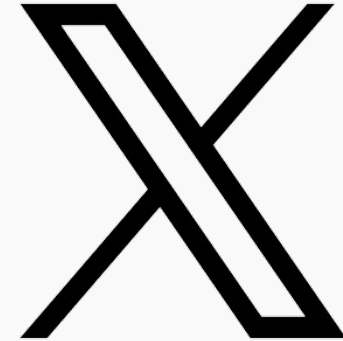
**Datum** A single piece of information, which can be treated as an observation

**Data** The plural of datum; multiple observations

**Dataset** A homogenous collection of data (each datum must have the same focus)

# What is data?

Everything can be data! Just requires making observations.



# What is data?

Everything can be data! Just requires making observations.





# Obtaining Data

---

You can obtain data if

- You curate it.
- Someone else provides it- all pre-packaged for you in files.
- Someone else provides an API.
- Someone else has available content, and you try to take it (web scraping).

# Types of Data

## Types of Data

### Quantitative

Data that can be measured with numbers.  
Eg: Height, Speed

#### Discrete

Data that can take on either a finite number of values, or an infinite, but countable number of values.  
Eg: No. of students in a class

#### Continuous

Data that can take on infinite number of values.  
Eg: Blood pressure, temperature

#### Interval

Data where there is order and the difference between two values is meaningful.  
Eg: GRE Score (260-340), Temperature (°C, F)

#### Ratio

Has all the properties of an interval variable, and also has a **clear definition of 0.0**. When the variable equals 0.0, there is none of that variable.  
Eg: Age, Temperature (Kelvin)

### Qualitative

Non-numerical data that is categorical.  
Eg: Blood group, Eye color

#### Nominal

No intrinsic order to the variables.  
Eg: Gender, Ethnicity

#### Ordinal

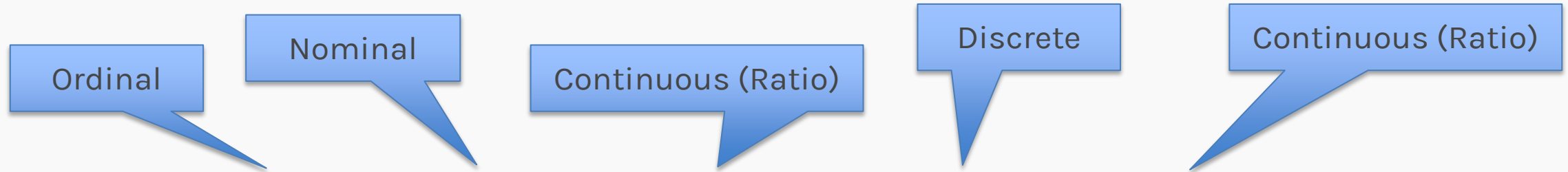
Data where order matters but not the difference between the values.  
Eg: Letter Grades, 1<sup>st</sup>, 2<sup>nd</sup> and 3<sup>rd</sup> prizes

# Types of Data

<b>quality</b>	<b>type</b>	<b>price</b>	<b>quantity</b>	<b>total</b>
High	Toy	20	5	100
Low	Book	5	3	15
Medium	Craft	12	4	48
Medium	Book	10	10	100

Guess the type of data in each column!

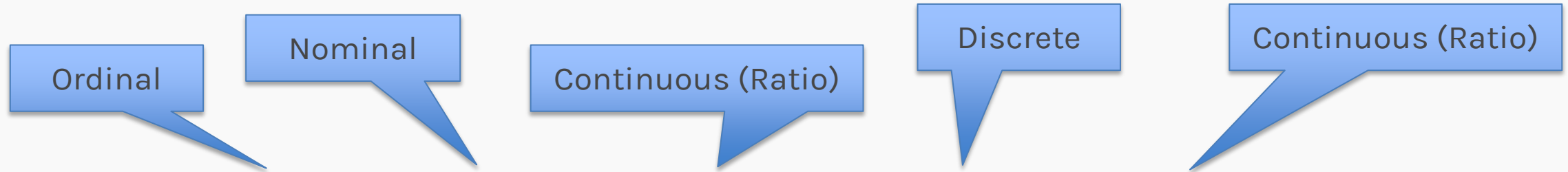
# Types of Data



quality	type	price	quantity	total
High	Toy	20	5	100
Low	Book	5	3	15
Medium	Craft	12	4	48
Medium	Book	10	10	100

Guess the type of data in each column!

# Types of Data



<b>quality</b>	<b>type</b>	<b>price</b>	<b>quantity</b>	<b>total</b>
High	Toy	20	5	100
Low	Book	5	3	15
Medium	Craft	12	4	48
Medium	Book	10	10	100

Often you must encode the data in some form to be useful. For example, you might use integers to encode the ordinal column “quality” here.

# Encoding types

Pandas dtype	Python type	NumPy type	Usage
object	str	string_, unicode_	Text
int64	int	int_, int8, int16, int32, int64, uint8, uint16, uint32, uint64	Integer numbers
float64	float	float_, float16, float32, float64	Floating point numbers
bool	bool	bool_	True/False values
datetime64	NA	datetime64[ns]	Date and time values
timedelta[ns]	NA	NA	Differences between two datetimes
category	NA	NA	Finite list of text values

# Types of Data

Table: contributors

New Record Delete Record

	id	last_name	first_name	middle_name	street_1	street_2	city	state	zip	amount	date	candidate_id
	Filter	Filter	Filter	Filter	Filter	Filter	Filter	Filter	Filter	Filter	Filter	Filter
1	1	Agee	Steven	NULL	549 Laurel ...	NULL	Floyd	VA	24091	500	2007-06-30	16
2	5	Akin	Charles	NULL	10187 Suga...	NULL	Bentonville	AR	72712	100	2007-06-16	16
3	6	Akin	Mike	NULL	181 Baywo...	NULL	Monticello	AR	71655	1500	2007-05-18	16
4	7	Akin	Rebecca	NULL	181 Baywo...	NULL	Monticello	AR	71655	500	2007-05-18	16
5	8	Aldridge	Brittni	NULL	808 Capitol...	NULL	Washington	DC	20024	250	2007-06-06	16
6	9	Allen	John D.	NULL	1052 Cann...	NULL	North Augu...	SC	29860	1000	2007-06-11	16
7	10	Allen	John D.	NULL	1052 Cann...	NULL	North Augu...	SC	29860	1300	2007-06-29	16
8	11	Allison	John W.	NULL	P.O. Box 10...	NULL	Conway	AR	72033	1000	2007-05-18	16
9	12	Allison	Rebecca	NULL	3206 Sum...	NULL	Little Rock	AR	72227	1000	2007-04-25	16

Now guess the type of data in each column and mention the data type you would encode it with.

# How do we store data?

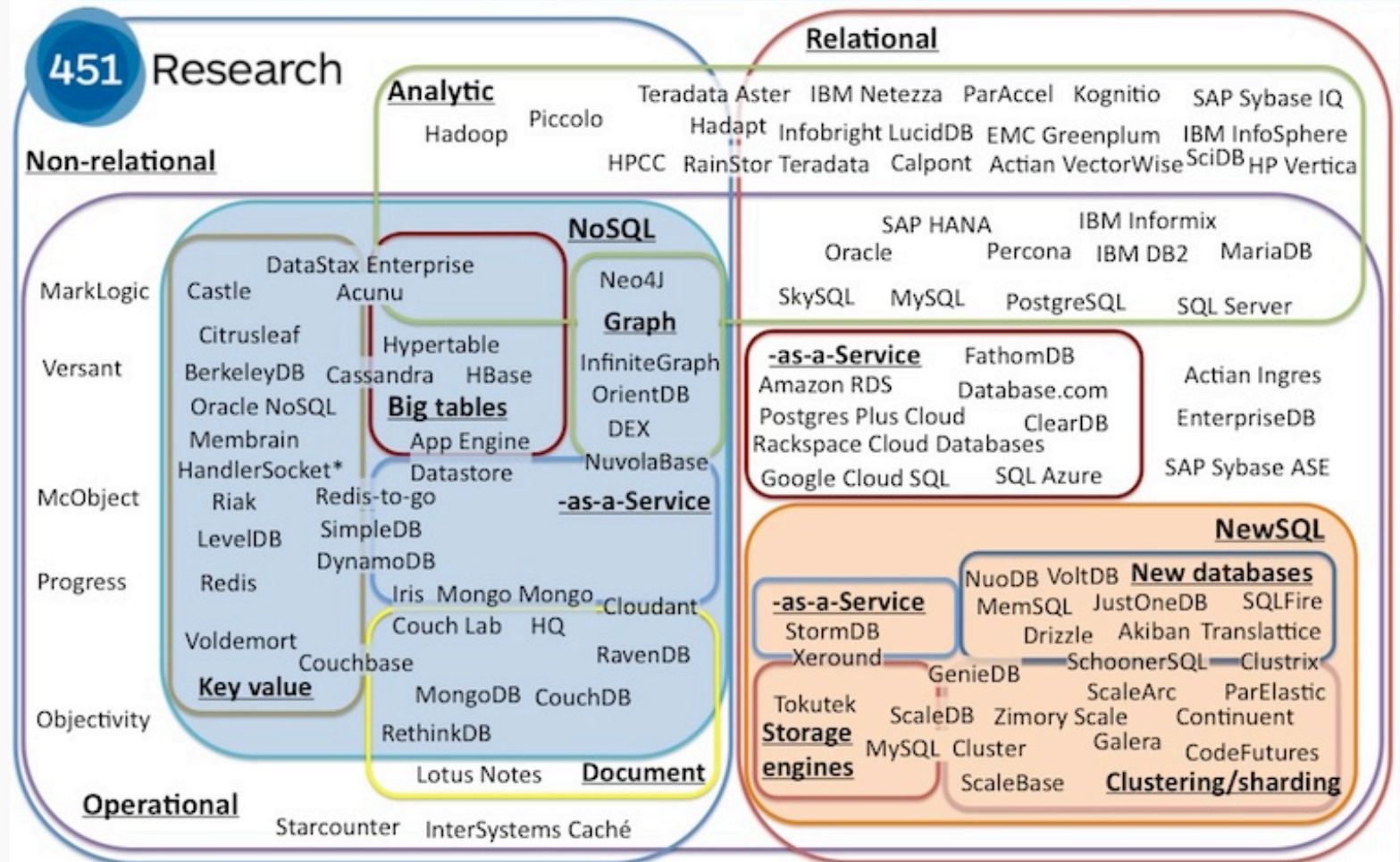
A database is an organized collection of structured information, or data, typically stored electronically in a computer system.





# Databases

## The evolving database landscape



Types of databases

# Relational Database

Relational databases  
organize data in multiple,  
related tables

Example: Pandas,  
SQL: Postgres, SQLite,  
Hbase, VoltDB

<http://www.linkedin.com/in/williamhgates>



**Bill Gates**

Greater Seattle Area | Philanthropy

## Summary

Co-chair of the Bill & Melinda Gates Foundation. Chairman, Microsoft Corporation. Voracious reader. Avid traveler. Active blogger.

## Experience

Co-chair • Bill & Melinda Gates Foundation  
2000 – Present

Co-founder, Chairman • Microsoft  
1975 – Present

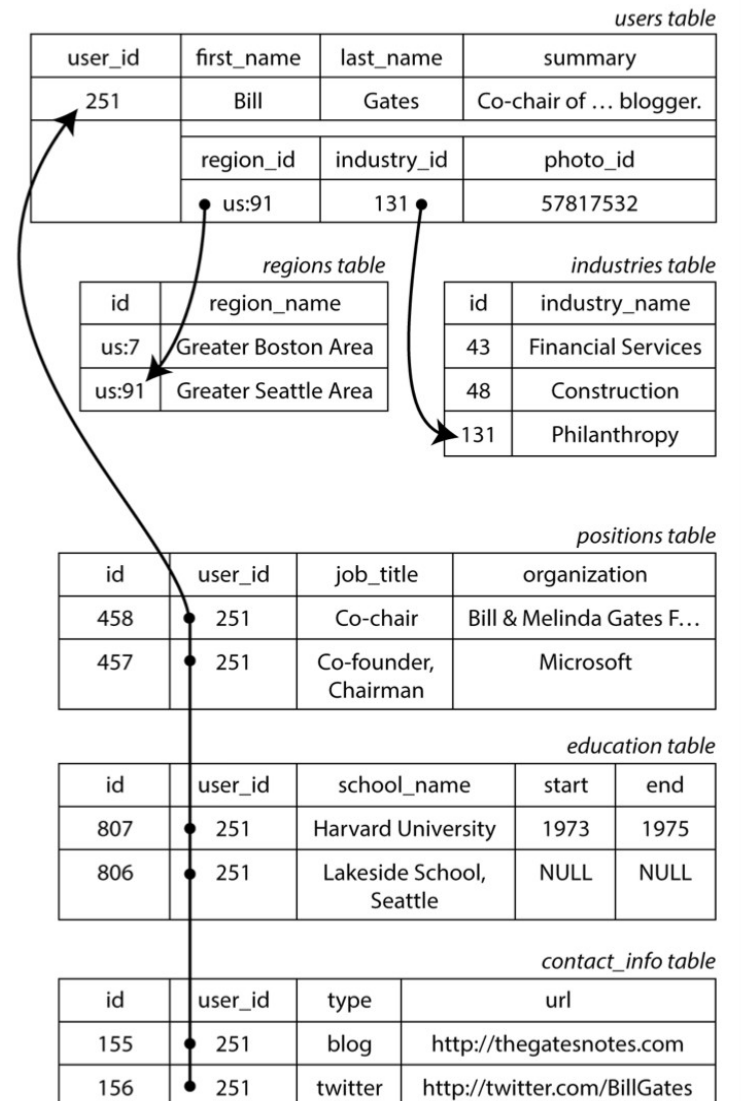
## Education

Harvard University  
1973 – 1975

Lakeside School, Seattle

## Contact Info

Blog: thegatesnotes.com  
Twitter: @BillGates



# Document Database

Document Databases use JSON/xml like documents to organize data instead of rows and columns

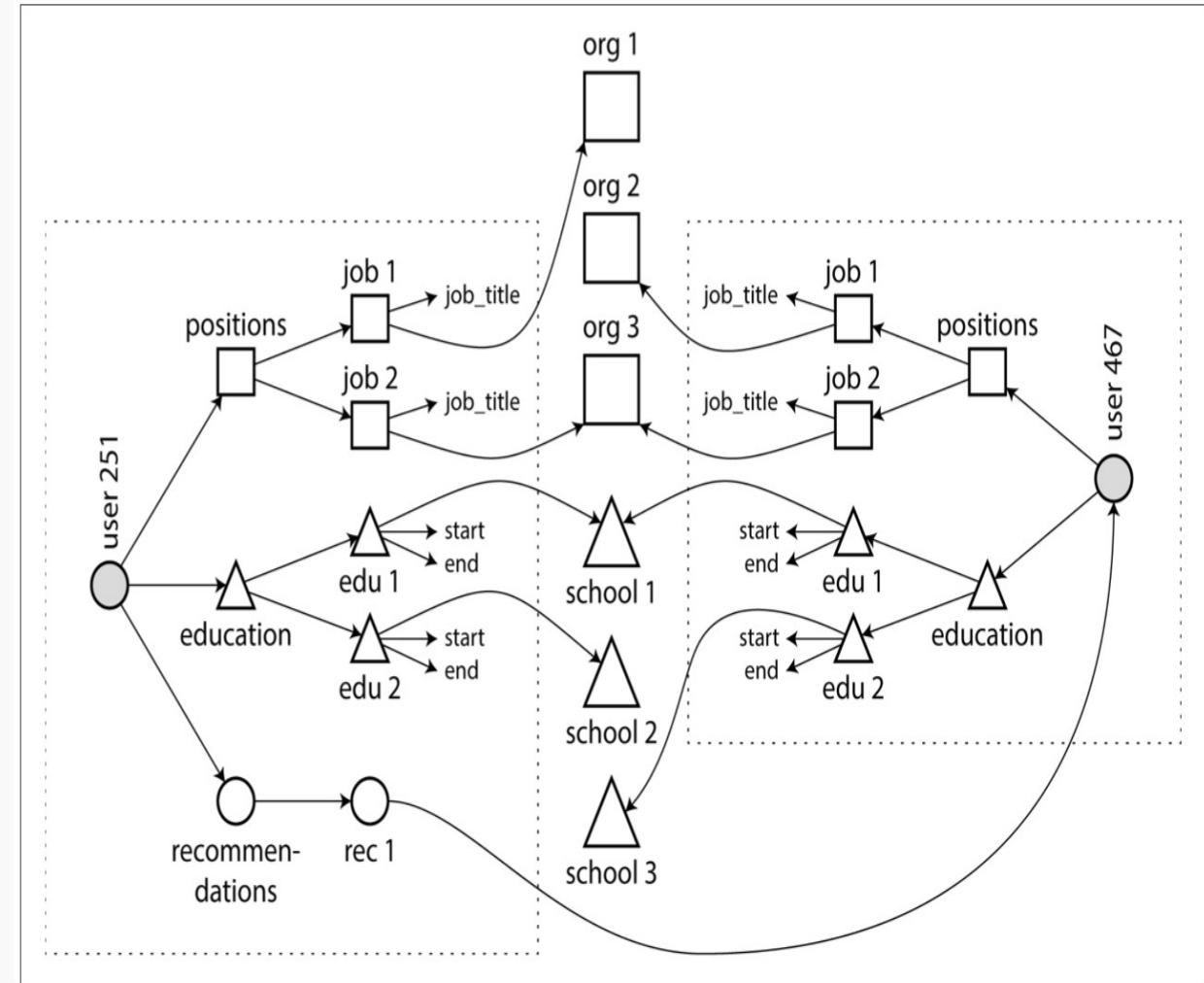
Example: MongoDB, CouchDB, Amazon DocumentDB

```
{
  "user_id": 251,
  "first_name": "Bill",
  "last_name": "Gates",
  "summary": "Co-chair of the Bill & Melinda Gates... Active blogger.",
  "region_id": "us:91",
  "industry_id": 131,
  "photo_url": "/p/7/000/253/05b/308dd6e.jpg",
  "positions": [
    {"job_title": "Co-chair", "organization": "Bill & Melinda Gates Foundation"},
    {"job_title": "Co-founder, Chairman", "organization": "Microsoft"}
  ],
  "education": [
    {"school_name": "Harvard University", "start": 1973, "end": 1975},
    {"school_name": "Lakeside School, Seattle", "start": null, "end": null}
  ],
}
```

# Graph Database

Graph Databases establish connections between data using nodes, edges and properties

Example: Neo4J



# OLTP and OLAP

**OLTP:** Online Transaction processing: look up a few records by some key, using an index. Insertion or updating based on users clicks, input. Ecommerce!

**OLAP:** Queries for business intelligence. Join lots of tables, get information about an entity of interest, such as a user, or a product. Typically, ETL (Extract-Transform-Load) scripts sit on OLTP databases to produce OLAP ones.

