Overview of Data Science

STAT1013 Data Science Toolbox

by Ben Dai (CUHK-STAT)
on January 9, 2023

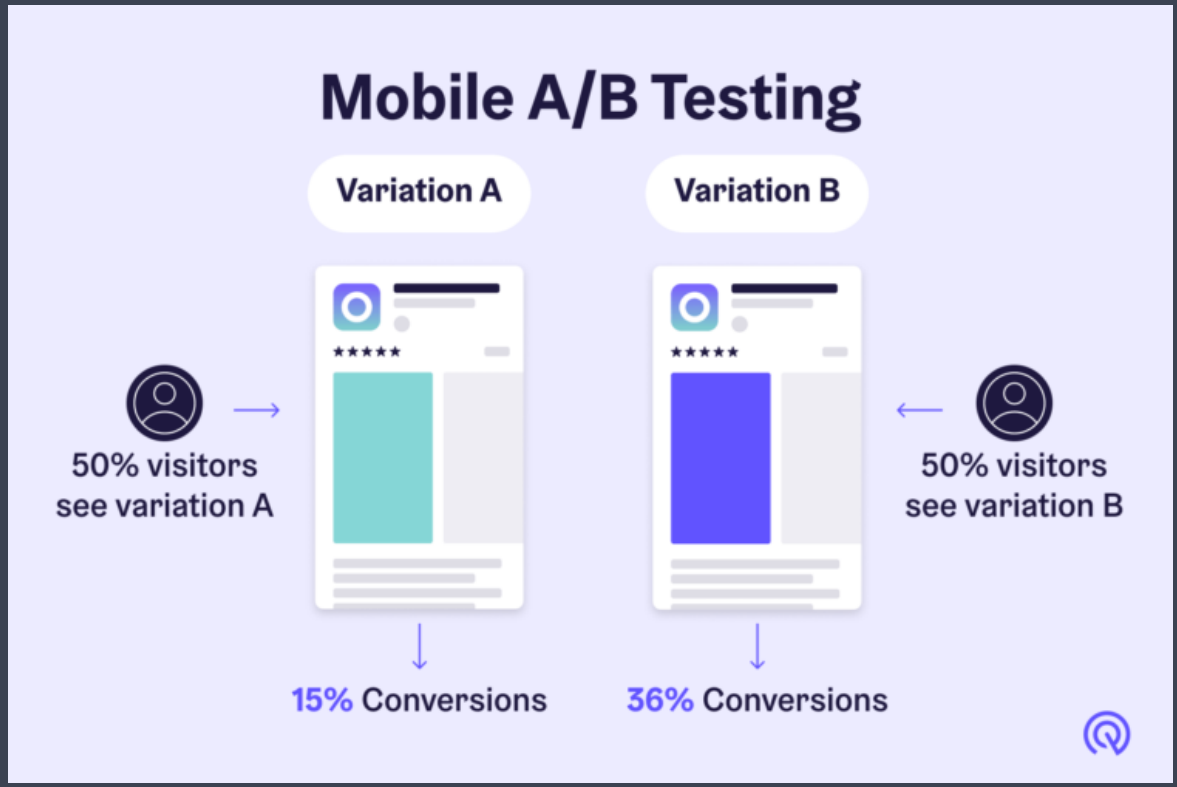
» What Is Data Science?

Data science is the domain of study that deals with Big Data using **STAT**/**ML** tools to:

* uncover hidden patterns
  → Statistics, Exploratory data analysis (EDA), and Data Visualization
* support business decision-making
  → Statistical inference, hypothesis testing, A/B test
* learning from data to make prediction
  → Machine learning methods, prediction

» Illustrative projects

A/B test website to increase future business gains



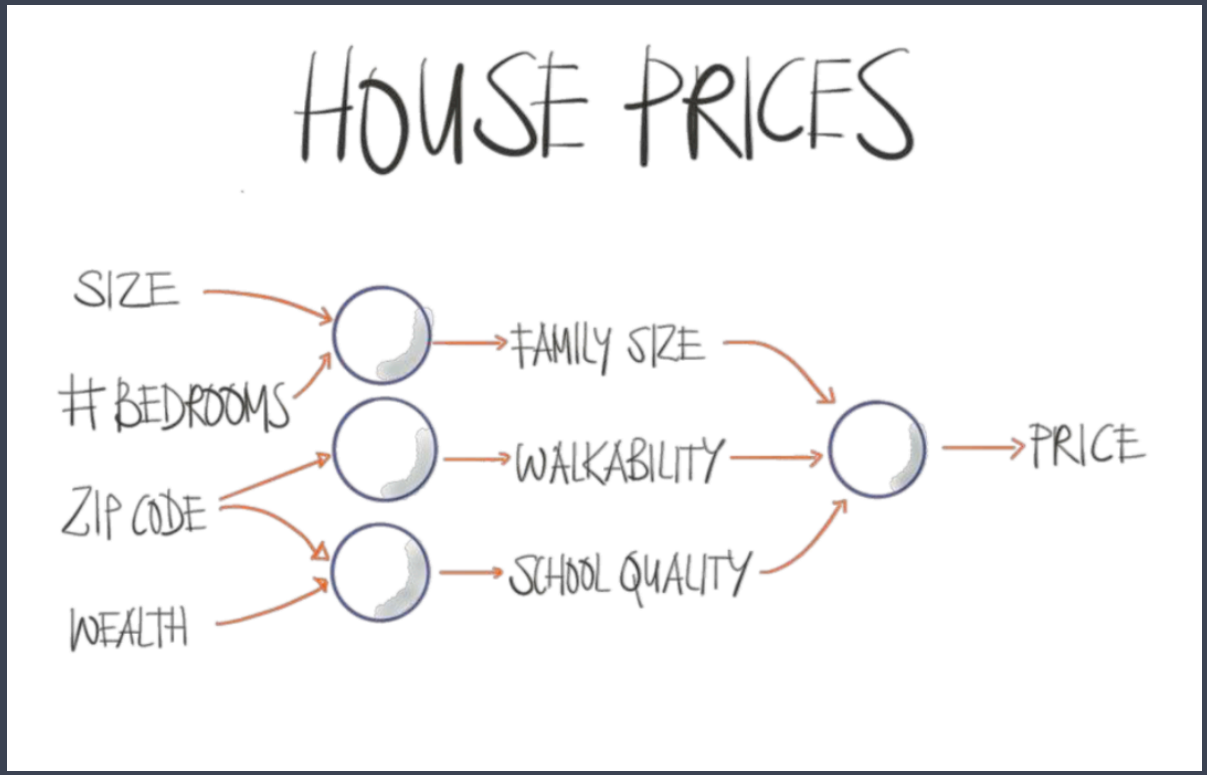
Decision A (15%) vs B (36%) → which one is better?

* What if A (20%) vs B (30%)?
* What if A (25%) vs B (26%)?

Source: `https://appradar.com/academy/app-store-listing-a-b-testing`

» Illustrative projects

ML predict house sales prices



* Which feature is the most important to house prices?

Prediction Given new house info, can you predict the house prices?

Source: `https://www.kaggle.com/competitions/house-prices-advanced-regression-techniques/discussion/111538`

» Before that ...

We need training:

- [ ] Python basic language
- [ ] Python libraries: numpy, scipy, pandas, ...
- [ ] Markdown to share your code and report
- [ ] Statistics/ML background

» Basic Python language

Colab Basic Python usage

» Types of Data

Types_data

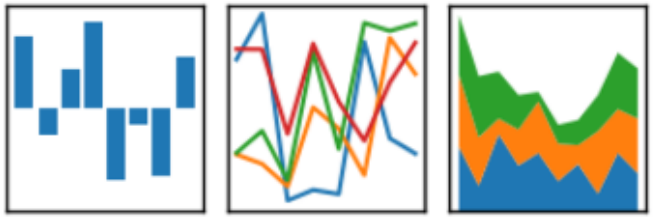
```
## Categorical data
>>> ['Bonjour', 'Hello', 'Hallo']
>>> [0, 1, 3, 4]

## Numerical data -> continuous data
>>> [0.12, 0.32, 1.324, 1.34]

## Numerical data -> ordinal data
>>> [1, 6, 3, 5]
```

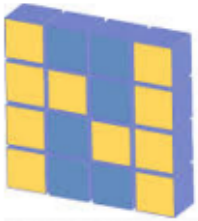
- Order matters? Continuous or discrete
  - Order matters + Continuous $\rightarrow$ Continuous data
  - Order matters + discrete $\rightarrow$ Ordinal data
  - Order does not matter $\rightarrow$ Categorical data

» Types of Data

pandas

$$y_{it} = \beta' x_{it} + \mu_i + \epsilon_{it}$$





NumPy

python™

Pandas dtype	Python type	NumPy type	Usage
object	str or mixed	string_, unicode_, mixed types	Text or mixed numeric and non-numeric values
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

» Exercise

data_example

```
In [1]: from sklearn.datasets import load_iris

In [2]: data = load_iris(as_frame=True)

In [3]: data.frame
Out[3]:
     sepal length (cm)  sepal width (cm)  petal length (cm)  petal width (cm)  target
0                  5.1               3.5                1.4               0.2       0
1                  4.9               3.0                1.4               0.2       0
2                  4.7               3.2                1.3               0.2       0
3                  4.6               3.1                1.5               0.2       0
4                  5.0               3.6                1.4               0.2       0
..                 ...               ...                ...               ...     ...
145                6.7               3.0                5.2               2.3       2
146                6.3               2.5                5.0               1.9       2
147                6.5               3.0                5.2               2.0       2
148                6.2               3.4                5.4               2.3       2
149                5.9               3.0                5.1               1.8       2

[150 rows x 5 columns]

In [4]: df = pd.read_csv('https://raw.githubusercontent.com/datasciencedojo/datasets/master
/titanic.csv')

In [5]: df.sample(5)
Out[5]:
PassengerId  Survived  Pclass   Sex   Age  SibSp  Parch      Ticket     Fare Cabin Embarked
        453         0       1  male  30.0      0      0      113051  27.7500  C111        C
        227         1       2  male  19.0      0      0   SW/PP 751  10.5000   NaN        S
        490         1       3  male   9.0      1      1  C.A. 37671  15.9000   NaN        S
        614         0       3  male   NaN      0      0      370377   7.7500   NaN        Q
```

* Data types of datasets: Iris + Titanic