Week 3 A03

Reading Discussion

- **The Six Divisions (GDS1-6)**

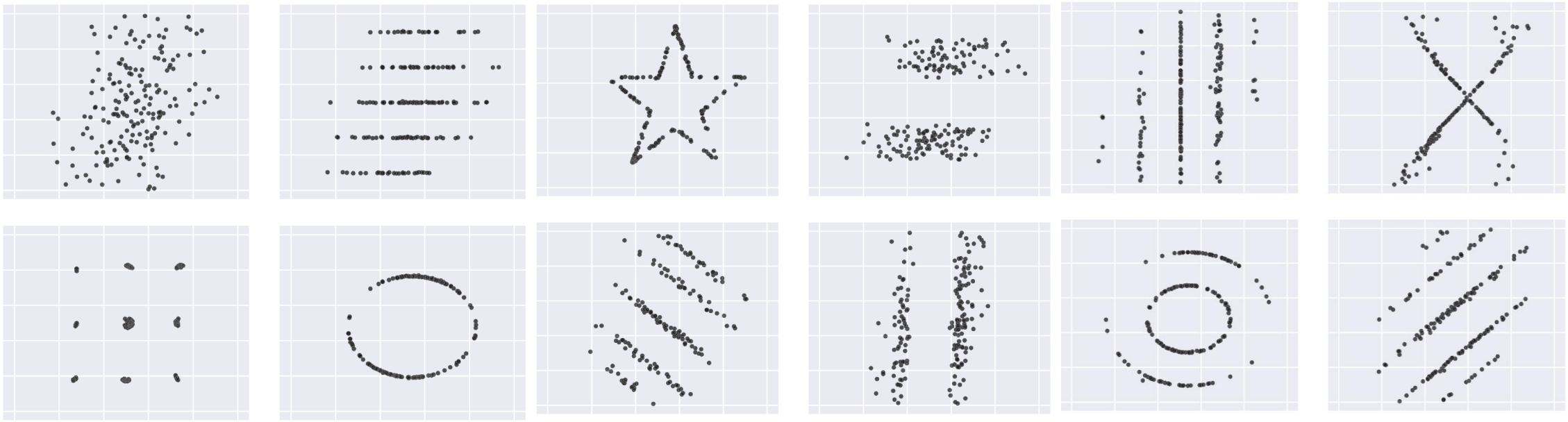
The activities of Greater Data Science are classified into 6 divisions:

- 1. Data Exploration and Preparation
- 2. Data Representation and Transformation
- 3. Computing with Data
- 4. Data Modeling
- 5. Data Visualization and Presentation
- 6. Science about Data Science

GDS1: Data Exploration

- Key idea: Exploratory Data Analysis (EDA)
- Why EDA is necessary?
- What does EDA look like?

Why EDA is necessary?

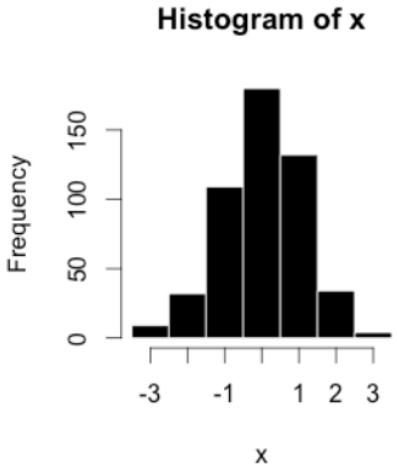


While different in appearance, each has the same summary statistics (mean, std. deviation, and Pearson's corr.) to 2 decimal places.

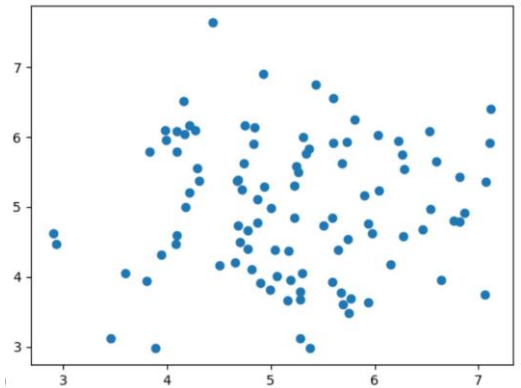
$\bar{x}$=54.02, $\bar{y}$= 48.09, sd(x) = 14.52, sd(y) = 24.79, Pearson's r = +0.32

Same Stats, Different Graphs: Generating Datasets with Varied Appearance and Identical Statistics through Simulated Annealing (autodesk.com)

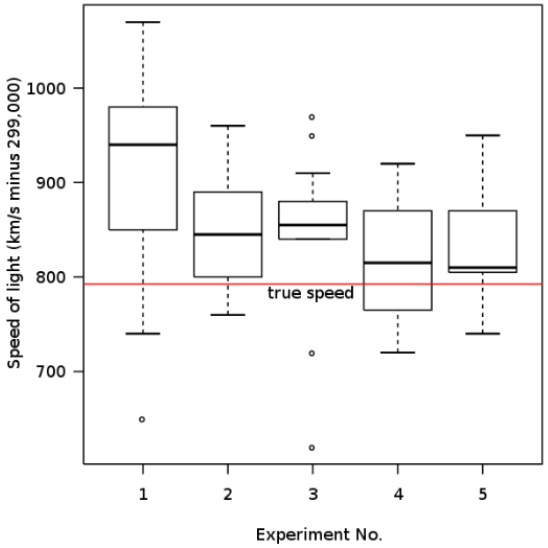
Typical Types of Plots



Histogram



Scatter plot



Box plot

.....

.....

GDS1: Data preparation

- Key idea: data cleansing

- Why data cleansing is necessary?

Why data cleansing is necessary?

- Filtering unwanted observations, e.g., duplicate elimination
- Fixing structural errors, e.g., “True” vs “T”, “Female” vs “F”
- Detecting outliers
- Handling missing values (NA)
- Validation: asking yourself if this data is reasonable or garbage?..

GDS2: Data Representation and Transformation

- Key ideas: modern databases and mathematical representations
- Modern databases: data manipulation e.g., SQL
- Mathematical representations: for computing purposes

A Typical Example of Data Transformation

Type
AA
AB
CD
AA

Onehot encoding →

Type	AA_Onehot	AB_Onehot	CD_Onehot
AA	1	0	0
AB	0	1	0
CD	0	0	1
AA	0	0	0

Human-Readable

Machine-Readable

Why is it called one-hot encoding:

a one-hot is a group of bits among which the legal combinations of values are only those with a single high (1) bit and all the others low (0).

GDS3: Computing with Data

- Key ideas: Algorithms and packages/libraries

- Algorithms: convergence errors (e.g., iterative methods)

- Packages/libraries:

  ① Python:

  numpy (scientific computing), seaborn, matplotlib, scikit-learn…

  ② R:

  ggplot2, tidyverse, R markdown…

Example of Convergence Errors

```
import numpy as np
from numpy.linalg import inv
```

```
mat_a = np.array([[1,2],[3,4]])
inv_a = np.linalg.inv(mat_a)
close_to_I = mat_a @ inv_a
print(mat_a,'\n x \n', inv_a, '\n = \n', close_to_I)
print(np.allclose(np.dot(mat_a, inv_a), np.eye(2)))
print('\nA 2 by 2 identity matrix: \n', np.eye(2))
```

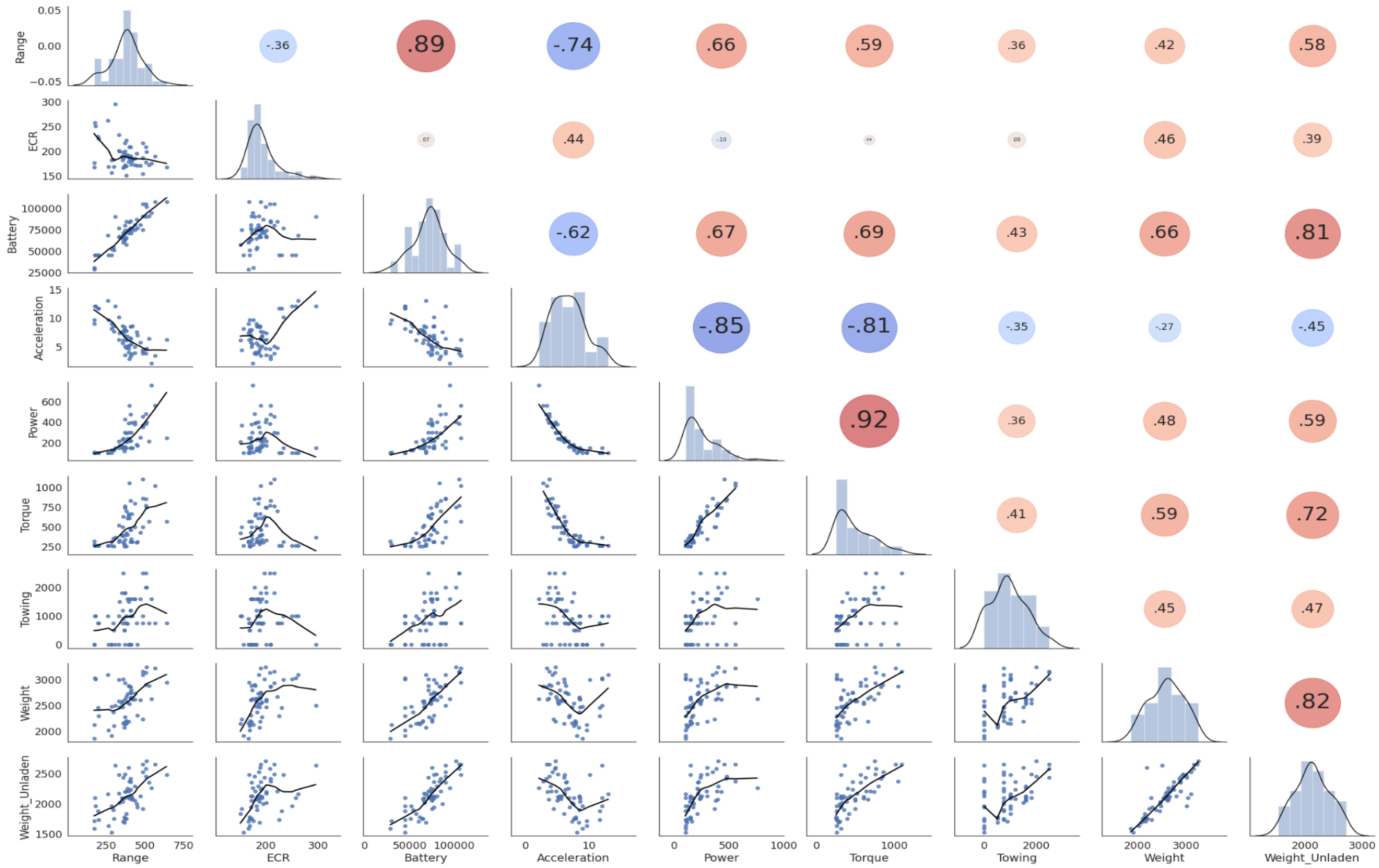
```
[[1 2]
 [3 4]]
 x
 [[-2.   1. ]
 [ 1.5 -0.5]]
 =
 [[1.0000000e+00 0.0000000e+00]
 [8.8817842e-16 1.0000000e+00]]
True

A 2 by 2 identity matrix:
 [[1. 0.]
 [0. 1.]]
```

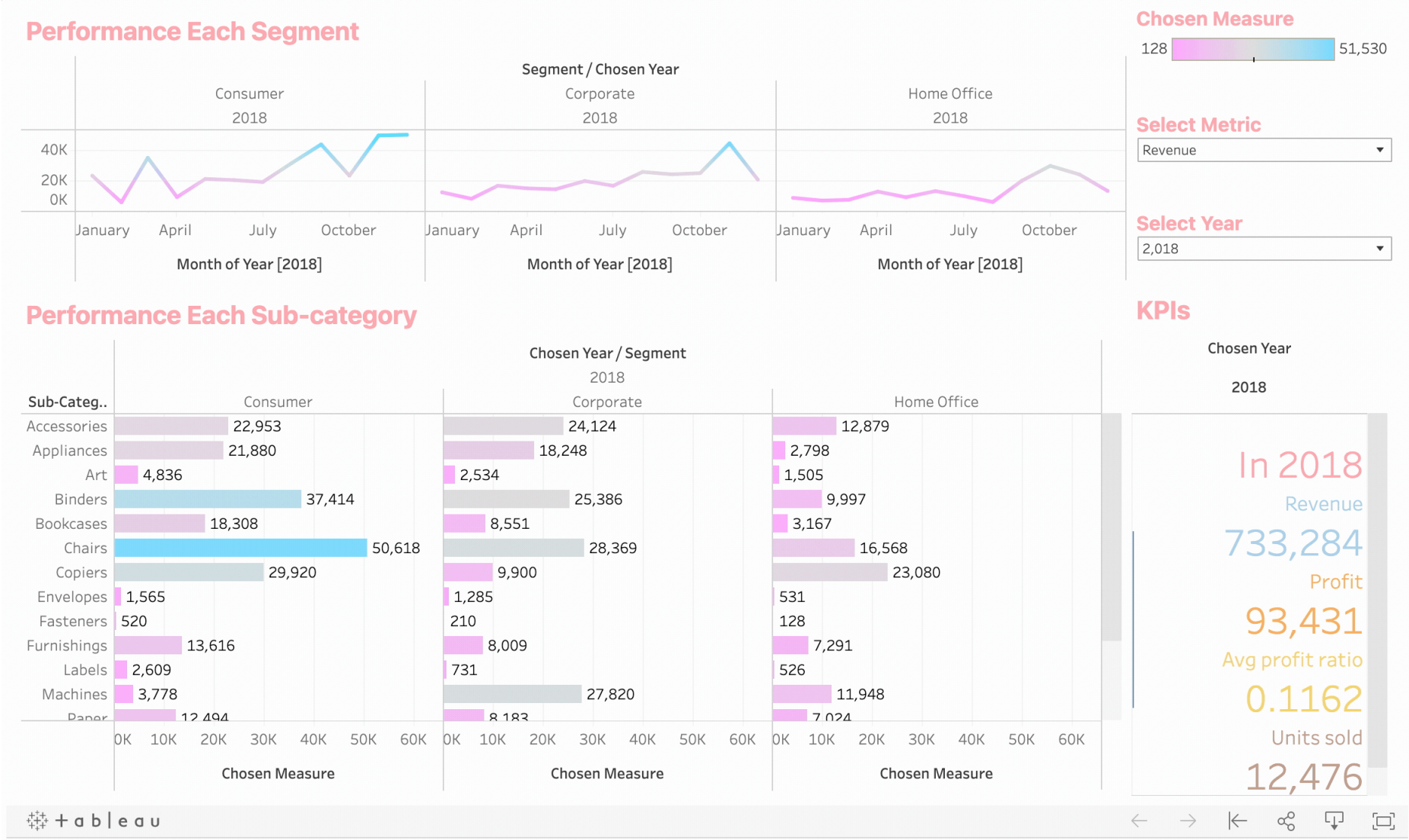
GDS4: Data Visualization and Presentation

- Key ideas: “Extreme EDA” and dynamic dashboards
- “Extreme EDA”: overlaps multiple plots in a graph. i.e., putting everything in one place.
- Dynamic dashboards: for monitoring data processing pipelines that access streaming or widely distributed data.

A Comprehensive Plot



Dynamic Dashboard



[Design A Dynamic Data Dashboard with Tableau | by Cheng | Towards Data Science]

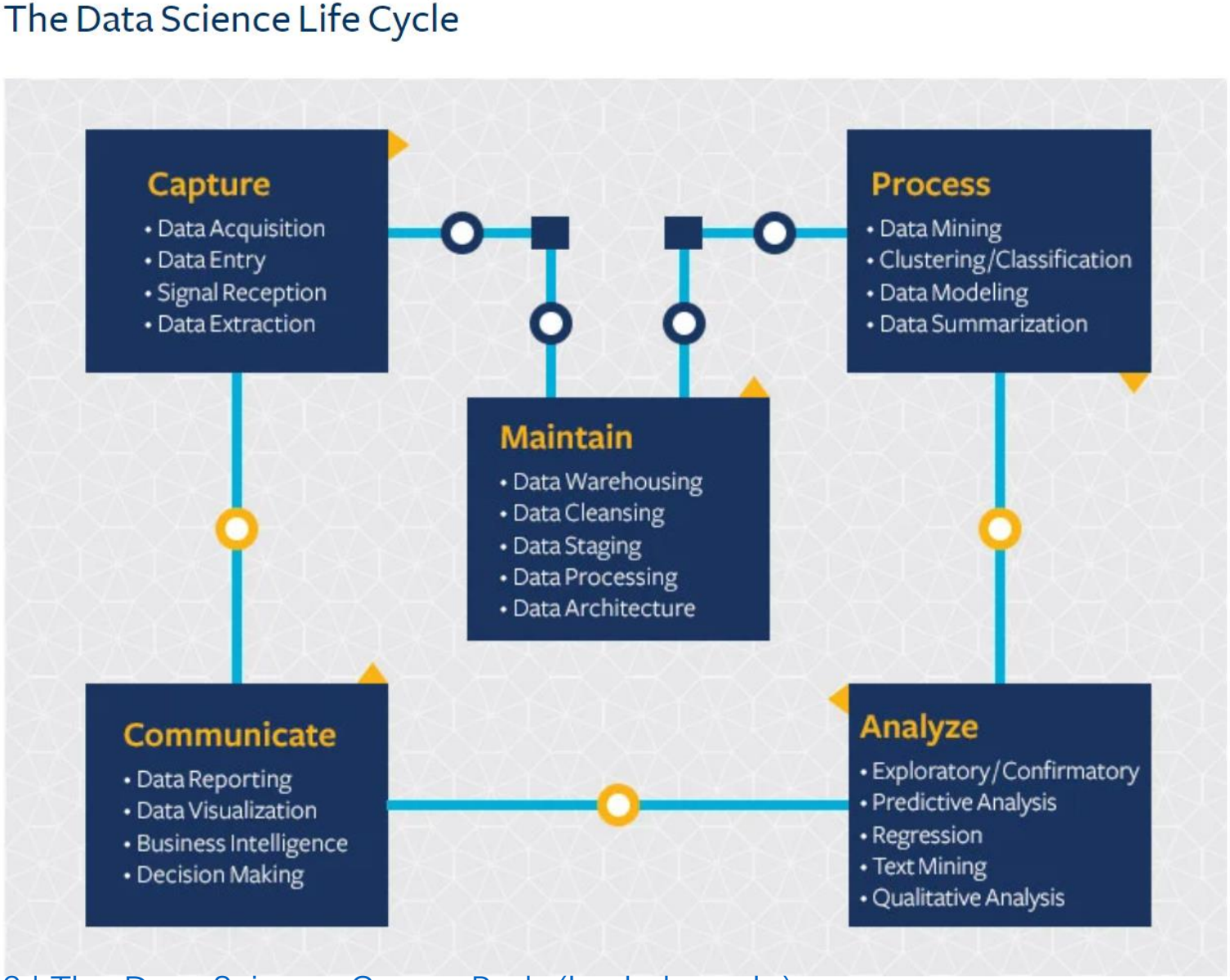
GDS5: Data Modeling

- Key ideas: Generative modeling and Predictive modeling
- Generative modeling: develop stochastic models which fit the data, and then **make inferences** about the data-generating mechanism based on the structure of those models. E.g., Naive Bayes Classifier (Spam or Ham?)
- Predictive modeling: estimate the outcome (target) from a new set of independent variables (features). E.g., modern Machine Learning.

GDS6: Science about Data Science

- Definition of science: the systematic study of the structure and behavior of the physical and natural world through **observation, experimentation, and the testing of theories against the evidence obtained**.
- Key idea: Analyzing data and modeling to maximize returns/effectiveness at each phase of the process.

Conclusion from UC Berkeley



What is Data Science? | The Data Science Career Path (berkeley.edu)