



DS363: Design and Learning with Data

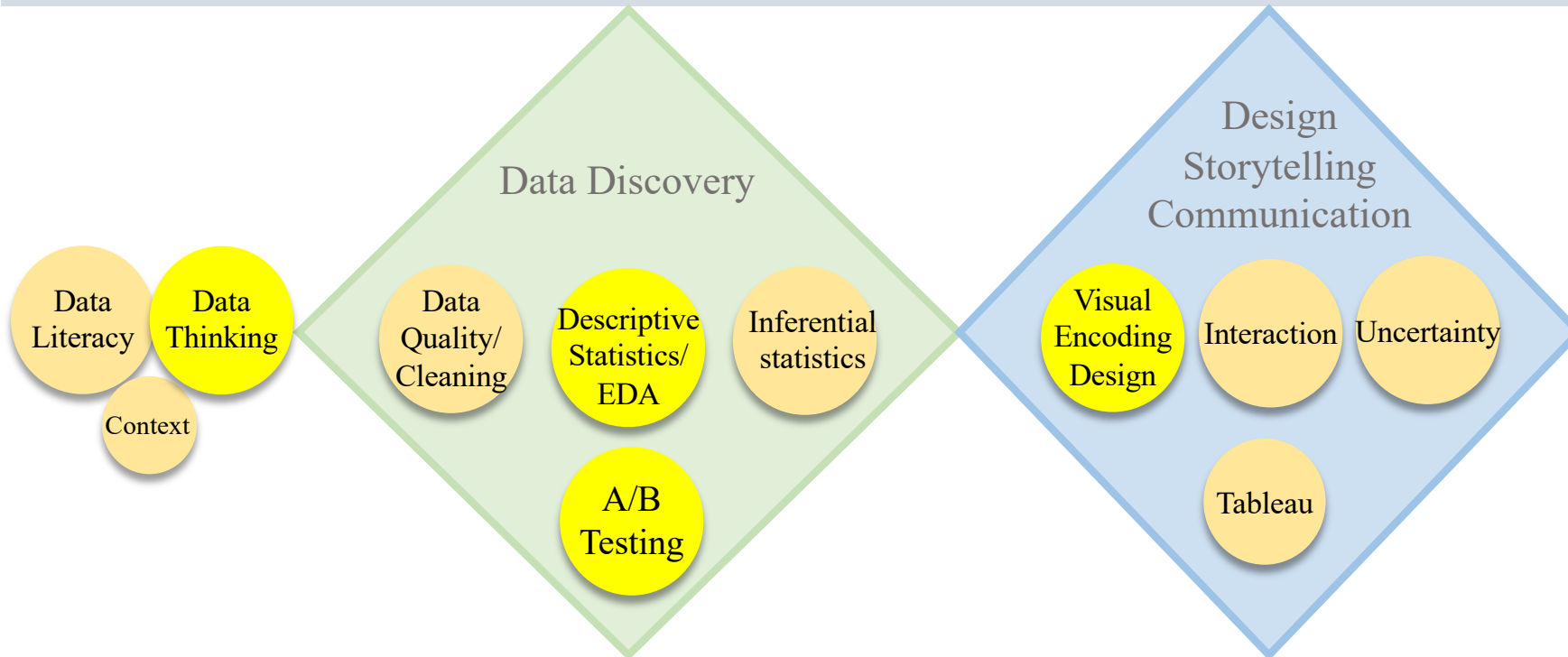
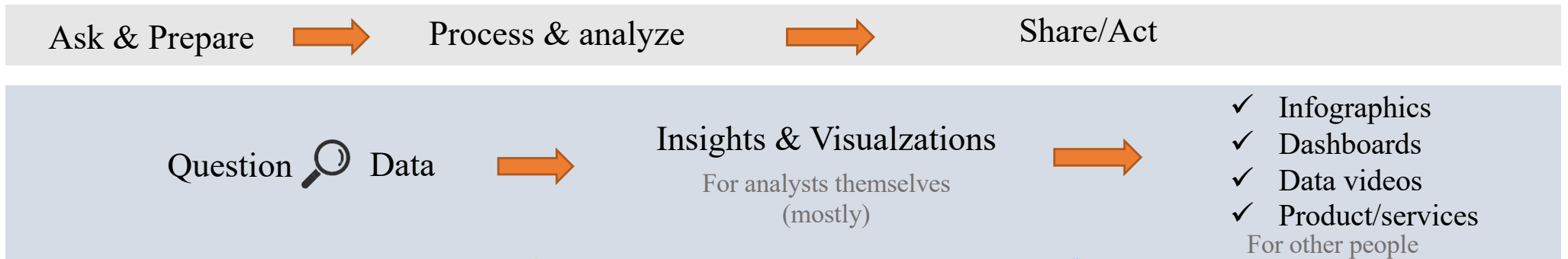
---

# Course Review

Wan Fang

Southern University of Science and Technology

# Design and Learning with Data



# Data Storytelling

- The ability to effectively communicate insights from a dataset using narratives and visualizations.



# Data Visualization

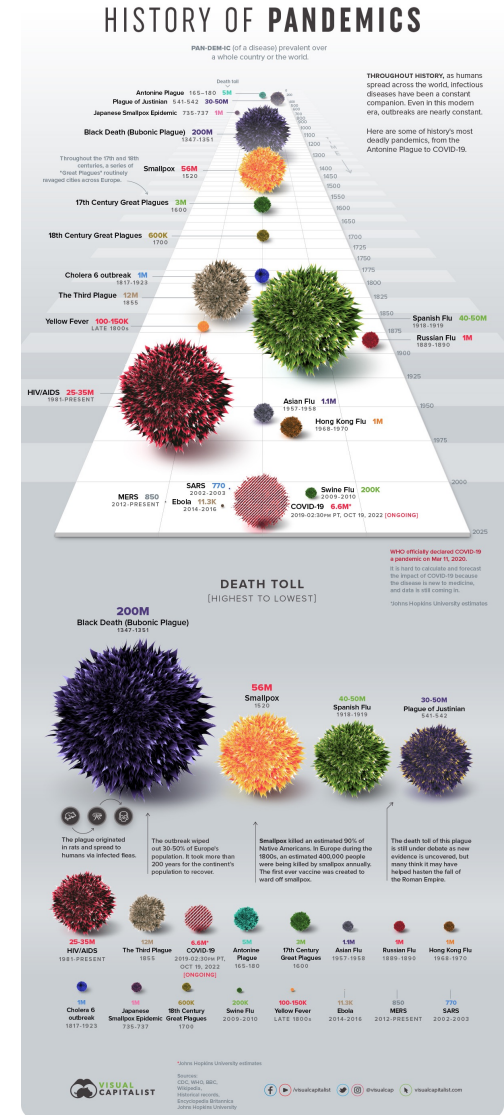
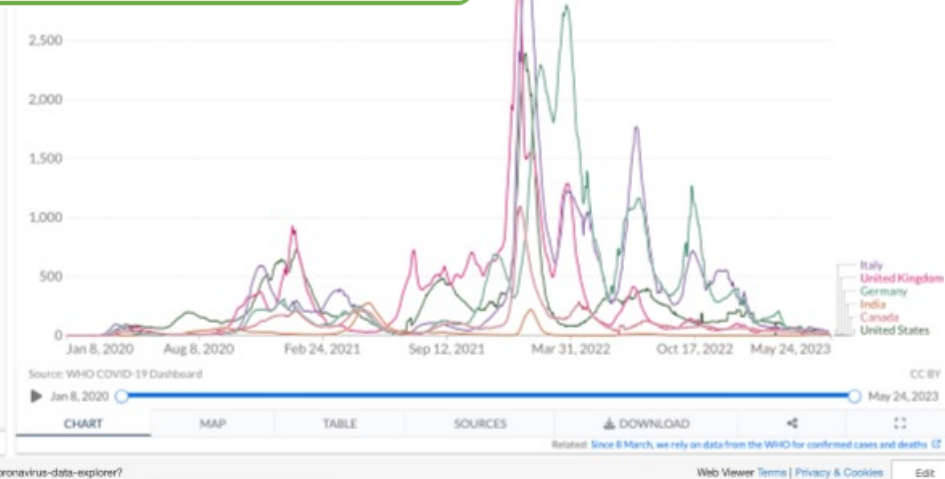
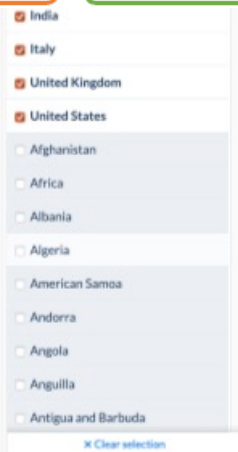
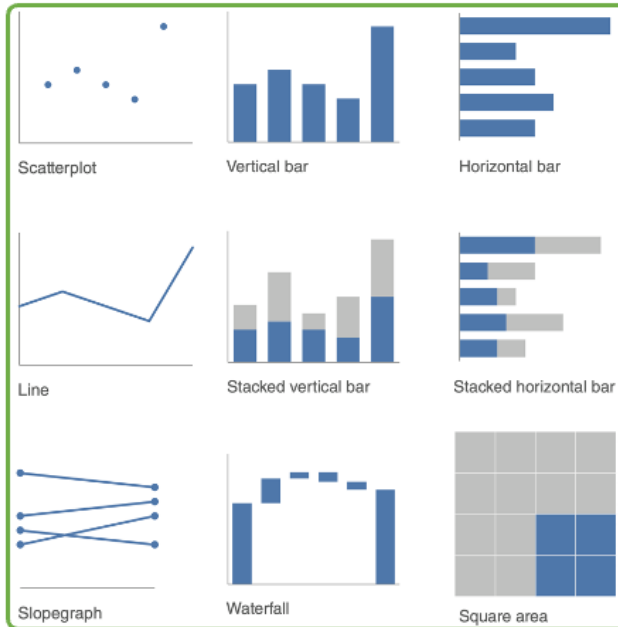


	A	B	C
Category 1	15%	22%	42%
Category 2	40%	36%	20%
Category 3	35%	17%	34%
Category 4	30%	29%	26%
Category 5	55%	30%	58%
Category 6	11%	25%	49%

Table

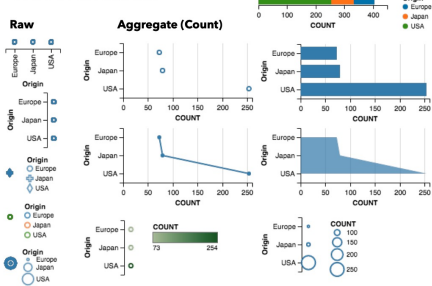
	A	B	C
Category 1	15%	22%	42%
Category 2	40%	36%	20%
Category 3	35%	17%	34%
Category 4	30%	29%	26%
Category 5	55%	30%	58%
Category 6	11%	25%	49%

Heatmap

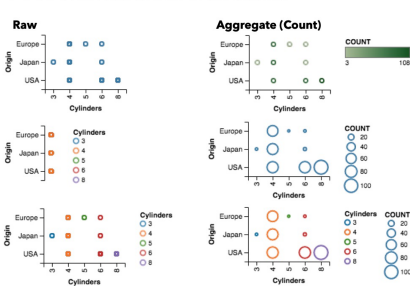


# Dimensional Visualization of Data

## 1D: Nominal

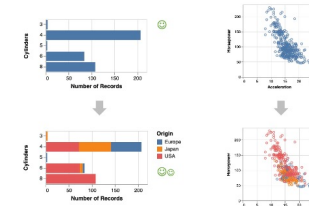


## 2D: Nominal x Nominal



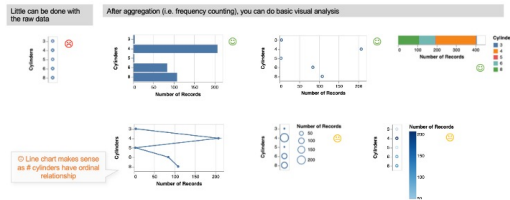
## 3D ANY

- Each visualization can accommodate 1-2 extra columns with color or size encodings. Why not explore higher-dimensions?



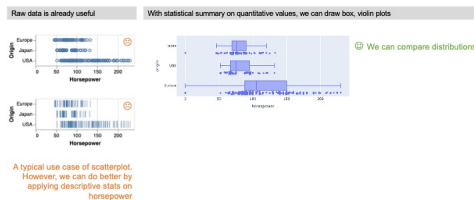
## 1D Ordinal

- When you are interested in a single column containing ordinal values (i.e., counting and ranking are allowed)
  - E.g., # of cylinders column of the car dataset



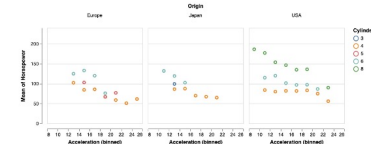
## 2D Nominal x Quantitative

- If you are interested in how one nominal and one quantitative columns
  - E.g., origin and horsepower columns of the car dataset

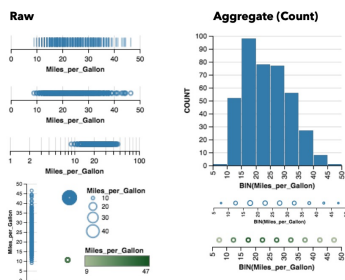


## Higher Dimension

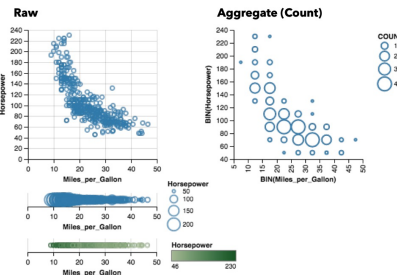
- Single charts usually cannot accommodate larger than 5 dimensions.
  - However, we can use composite charts.
  - For example, we have used scatterplot matrix in the previous tutorial.



## 1D: Quantitative



## 2D: Quantitative x Quantitative





# The Importance of Context

## 6 basic problem types

- Making predictions
- Categorizing things
- Spotting something unusual
- Identifying themes
- Discovering connections
- Finding patterns

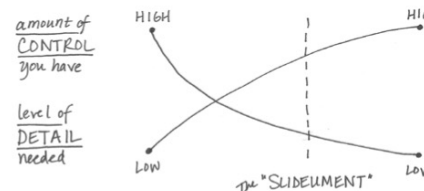
## Craft effective questions

- SMART methodology
  - Specific — does the question address the problem? Does it have a context?
  - Measurable — does it give the answer that can be measured?
  - Action-oriented — will the info that we get help us devise an action plan?
  - Relevant — is it about a particular problem we are trying to solve?
  - Time-bounded — are the answers relevant to the specific time being studied?

## Who, What, and How

- ***To whom are you communicating?***
  - It is important to have a good understanding of who your audience is and how they perceive you. This can help you to **identify common ground** that will help you ensure they hear your message.
- ***What do you want your audience to know or do?***
  - You should be clear how you want your audience to act and take into account how you will communicate to them and the overall tone that you want to set for your communication.
- ***How can you use data to help make your point?***
  - It's **only after** you can concisely answer these first two questions that you're ready to move forward with the third.

LIVE PRESENTATION . . . . . WRITTEN DOC or EMAIL



### Prompting action

Here are some action words to help act as thought starters as you determine what you are asking of your audience:

accept | agree | begin | believe | change | collaborate | commence | create | defend | desire | differentiate | do | empathize | empower | encourage | engage | establish | examine | facilitate | familiarize | form | implement | include | influence | invest | invigorate | know | learn | like | persuade | plan | promote | pursue | recommend | receive | remember | report | respond | secure | support | simplify | start | try | understand | validate

### Ignore the nonsupporting data?

You might assume that showing only the data that backs up your point and ignoring the rest will make for a stronger case. I do not recommend this. Beyond being misleading by painting a one-sided story, this is very risky. A discerning audience will poke holes in a story that doesn't hold up or data that shows one aspect but ignores the rest. The right amount of context and supporting and opposing data will vary depending on the situation, the level of trust you have with your audience, and other factors.

# Data X

- **Data science vs Data analytics**
- **Data ecosystem / Data life cycle / Data Privacy & Ethics**
- **Data Integrity / Data & Analytics Skills**

## Descriptive analytics

- looks at data to examine, understand, and describe something that's already happened.

## Diagnostic analytics

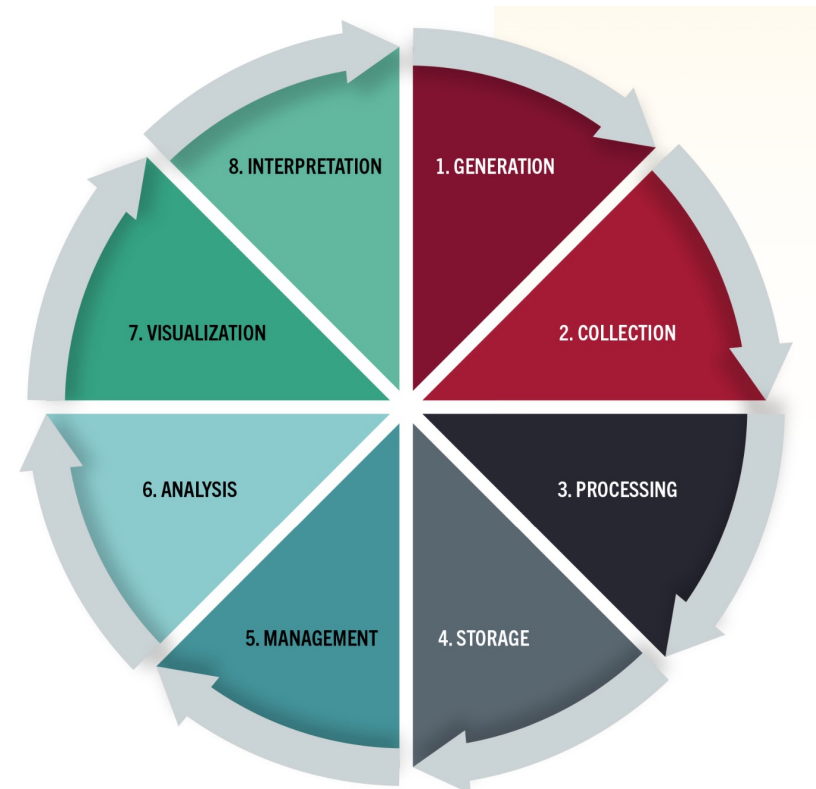
- goes deeper than descriptive analytics by seeking to understand the “why” behind what happened.

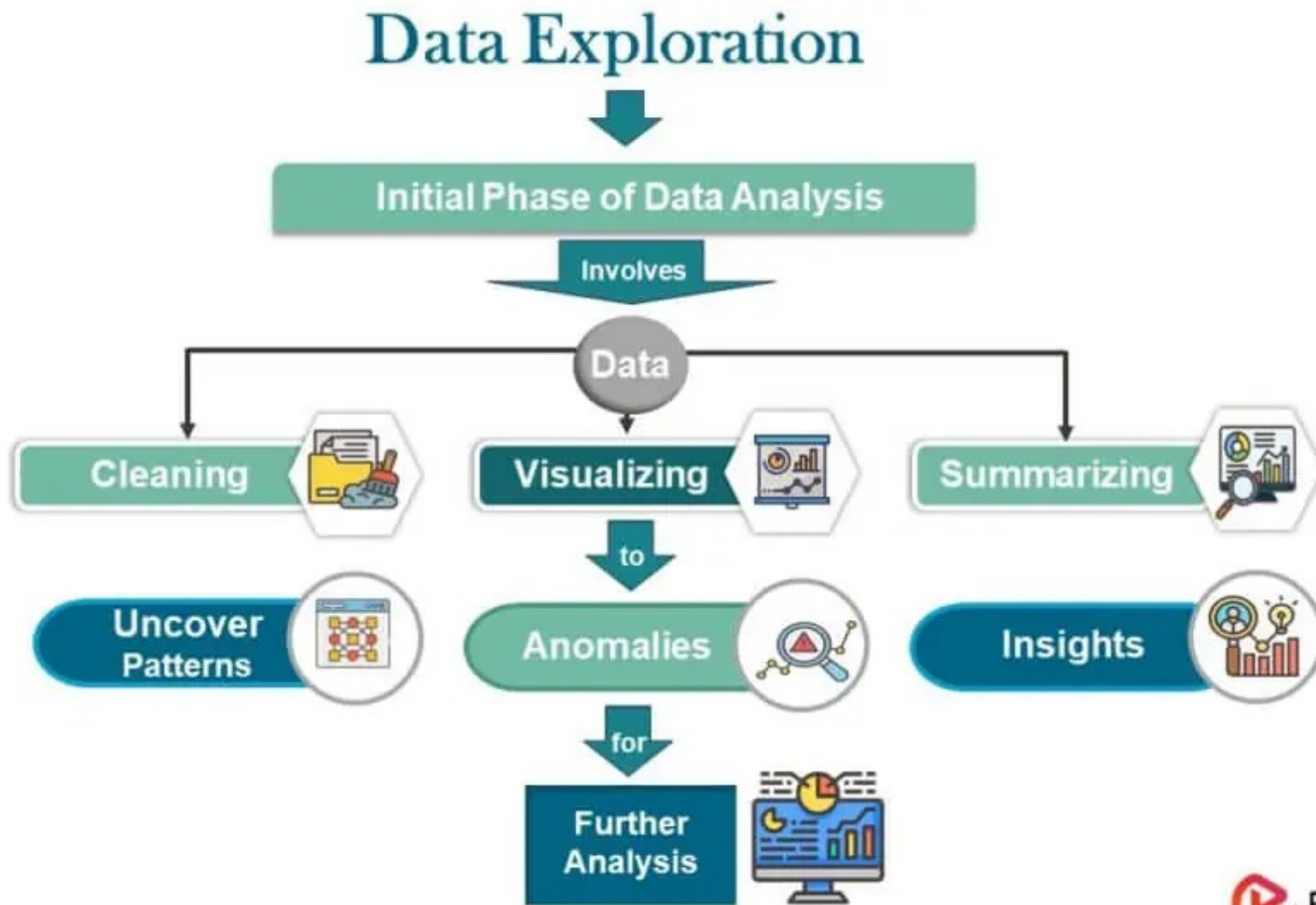
## Predictive analytics

- relies on historical data, past trends, and assumptions to answer questions about what will happen in the future.

## Prescriptive analytics

- identifies specific actions an individual or organization should take to reach future targets or goals.







# Big Data Quality Assessment Framework

Dimensions	Elements	Indicators
1) Availability	1) Accessibility	<ul style="list-style-type: none"> <li>Whether a data access interface is provided</li> <li>Data can be easily made public or easy to purchase</li> </ul>
	2) Timeliness	<ul style="list-style-type: none"> <li>Within a given time, whether the data arrive on time</li> <li>Whether data are regularly updated</li> <li>Whether the time interval from data collection and processing to release meets requirements</li> </ul>
2) Usability	1) Credibility	<ul style="list-style-type: none"> <li>Data come from specialized organizations of a country, field, or industry</li> <li>Experts or specialists regularly audit and check the correctness of the data content</li> <li>Data exist in the range of known or acceptable values</li> </ul>
3) Reliability	1) Accuracy	<ul style="list-style-type: none"> <li>Data provided are accurate</li> <li>Data representation (or value) well reflects the true state of the source information</li> <li>Information (data) representation will not cause ambiguity</li> </ul>
	2) Consistency	<ul style="list-style-type: none"> <li>After data have been processed, their concepts, value domains, and formats still match as before processing</li> <li>During a certain time, data remain consistent and verifiable</li> <li>Data and the data from other data sources are consistent or verifiable</li> </ul>
	3) Integrity	<ul style="list-style-type: none"> <li>Data format is clear and meets the criteria</li> <li>Data are consistent with structural integrity</li> <li>Data are consistent with content integrity</li> </ul>
	4) Completeness	<ul style="list-style-type: none"> <li>Whether the deficiency of a component will impact use of the data for data with multi-components</li> <li>Whether the deficiency of a component will impact data accuracy and integrity</li> </ul>
4) Relevance	1) Fitness	<ul style="list-style-type: none"> <li>The data collected do not completely match the theme, but they expound one aspect</li> <li>Most datasets retrieved are within the retrieval theme users need</li> <li>Information theme provides matches with users' retrieval theme</li> </ul>
5) Presentation Quality	1) Readability	<ul style="list-style-type: none"> <li>Data (content, format, etc.) are clear and understandable</li> <li>It is easy to judge that the data provided meet needs</li> <li>Data description, classification, and coding content satisfy specification and are easy to understand</li> </ul>

garbage in, garbage out

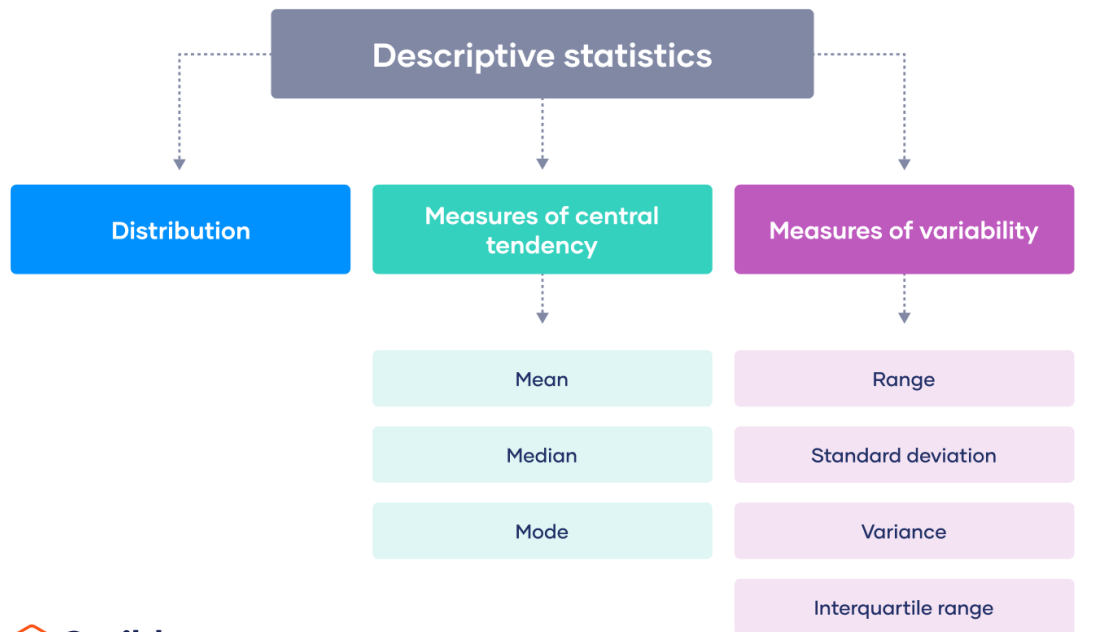
[garbage in, garbage out] 

**DEFINITION**

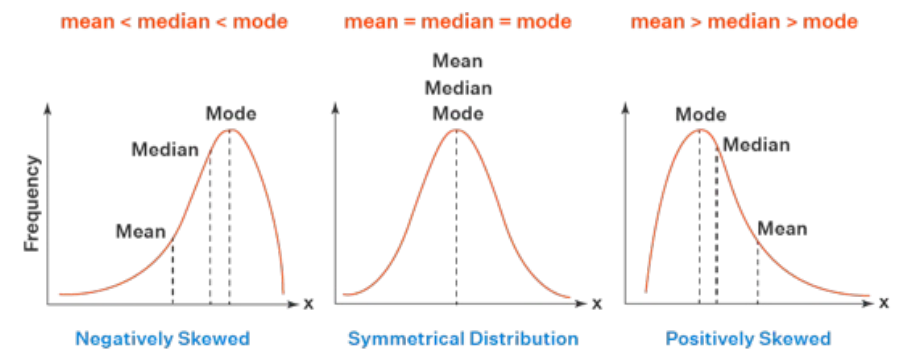
used to express the idea that in computing and other fields, incorrect or poor-quality input will produce faulty output.

Descriptive  
statistics

# Descriptive Statistics



 Scribbr



Statistical graphics

# Univariate/Multivariate Graphical

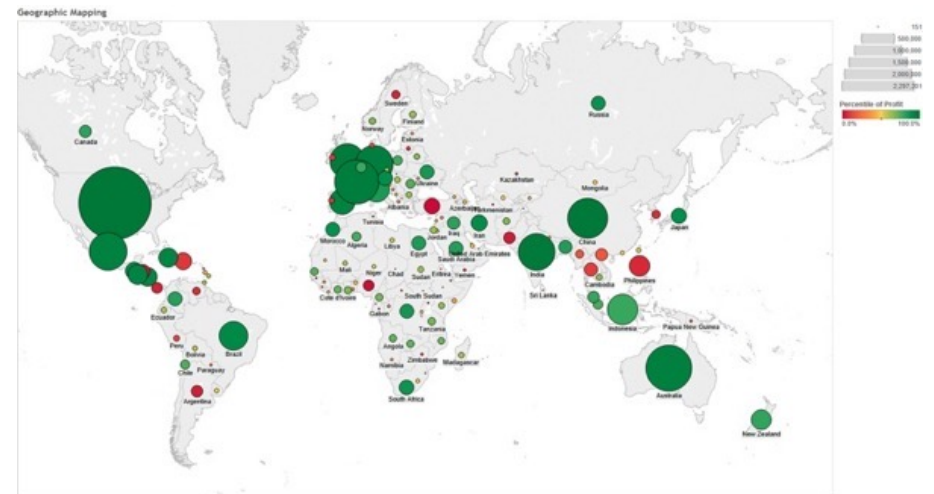
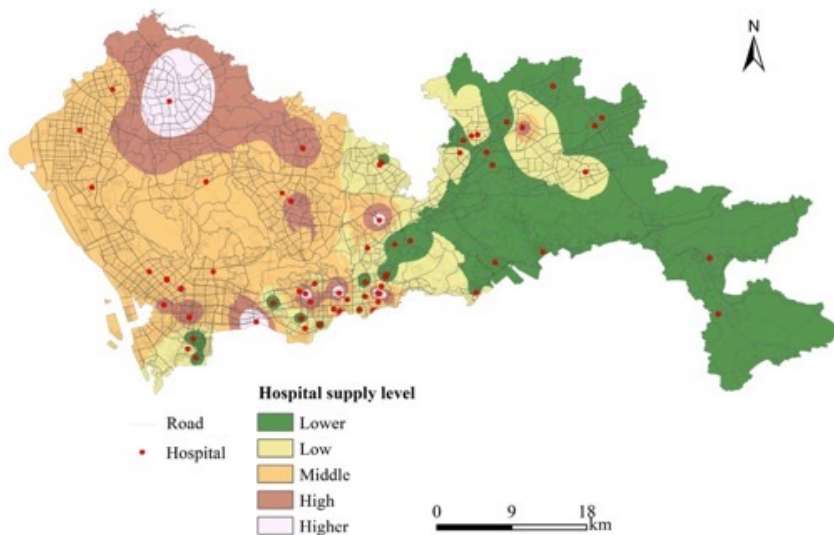
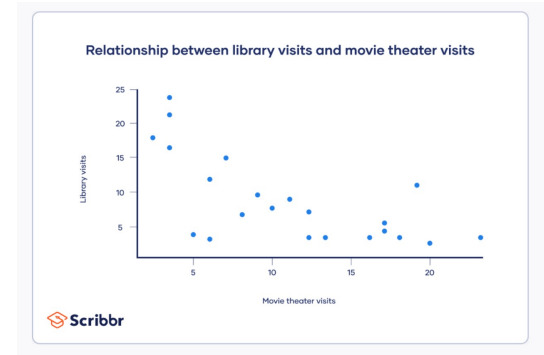
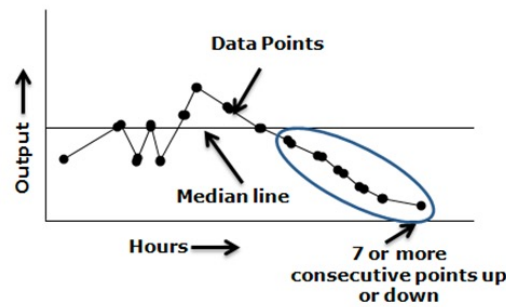
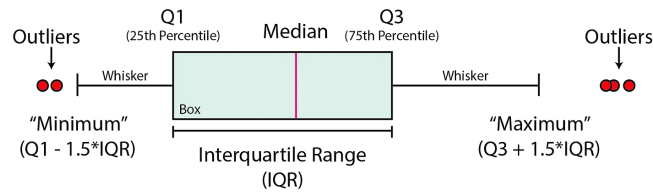
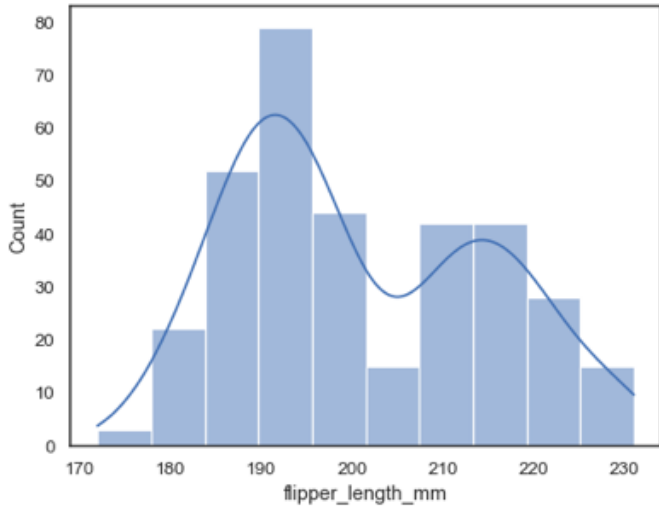


Figure 4. Spatial distribution of road network and hospital supply level in Shenzhen.

Descriptive  
Statistics/  
EDA

# With Python

Exploratory Data Analysis | Deepnote

deepnote.com/workspace/fang-wan-9f34fd68-36fc-4869-b2cb-bda9010c988b/project/Exploratory-Data-Analysis-4216a0b0-3afc-4a2a-854d-2535d81a10a1/

Fang WAN  
sophie.fwan@gmail.com

Search...

Getting started 1

Integrations

Settings & members

PROJECTS +

Exploratory Data Analysis

Starter Project

PRIVATE +

1 - Introduction - Duplicate

119 days left in your Team trial

Explore

Chat & resources

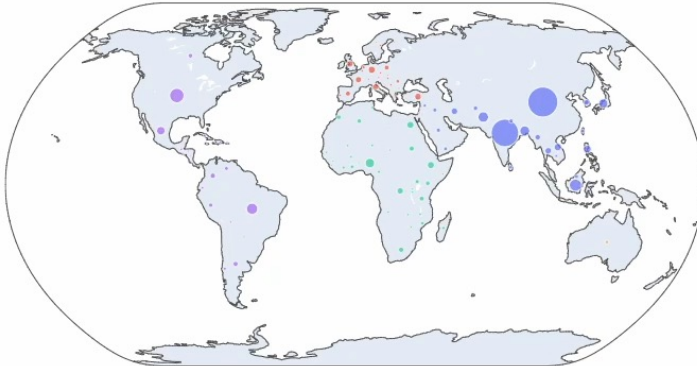
Deepnote Team

Machine offline. [Restore variables](#) from last session.

Run notebook

```
import plotly.express as px
import pandas as pd
```

```
df = px.data.gapminder().query("year==2007")
fig = px.scatter_geo(df, locations="iso_alpha", color="continent",
                    hover_name="country", size="pop",
                    projection="natural earth")
fig.show()
```



```
# read data
df = pd.read_csv('Mag6PlusEarthquakes_1900-2013.csv')
```

Project Comments History

NOTEBOOKS +

10\_minutes\_to\_pandas

earthquakes\_plotly

test

winequality\_white

INTEGRATIONS +

Connect an integration

To view its schema and query it with SQL

FILES +

Mag6PlusEarthquakes\_190...

winequality-white.csv

ENVIRONMENT

Python 3.7 Basic

RAM: 0 GB CPU: 0%

TERMINALS +

Fang's terminal #1

TABLE OF CONTENTS

Create a Heading block or a header in a Markdown cell.

VARIABLE EXPLORER 0

No variables in scope. [Restore](#)

Inferential  
statistics

# Inferential Statistics

Step 1: Write your hypotheses and plan your research design

Step 2: Collect data from a sample

Step 3: Summarize your data with descriptive statistics

Step 4: Test hypotheses or make estimates with inferential statistics

Step 5: Interpret your results

Example: Statistical hypotheses to test an effect

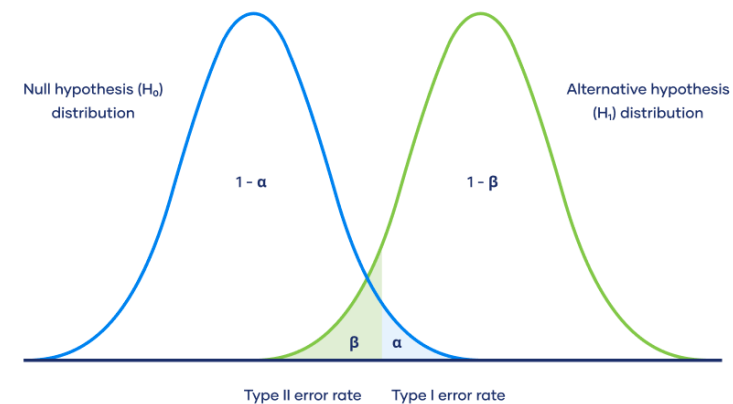
- **Null hypothesis:** A 5-minute meditation exercise will have no effect on math test scores in teenagers.
- **Alternative hypothesis:** A 5-minute meditation exercise will improve math test scores in teenagers.

Example: Statistical hypotheses to test a correlation

- **Null hypothesis:** Parental income and GPA have no relationship with each other in college students.
- **Alternative hypothesis:** Parental income and GPA are positively correlated in college students.

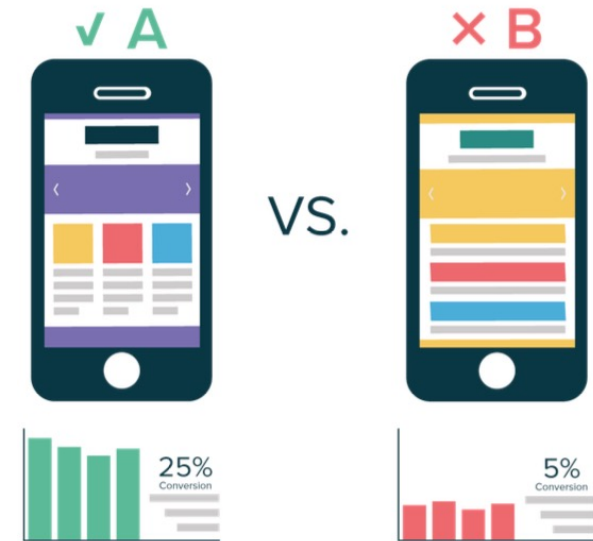
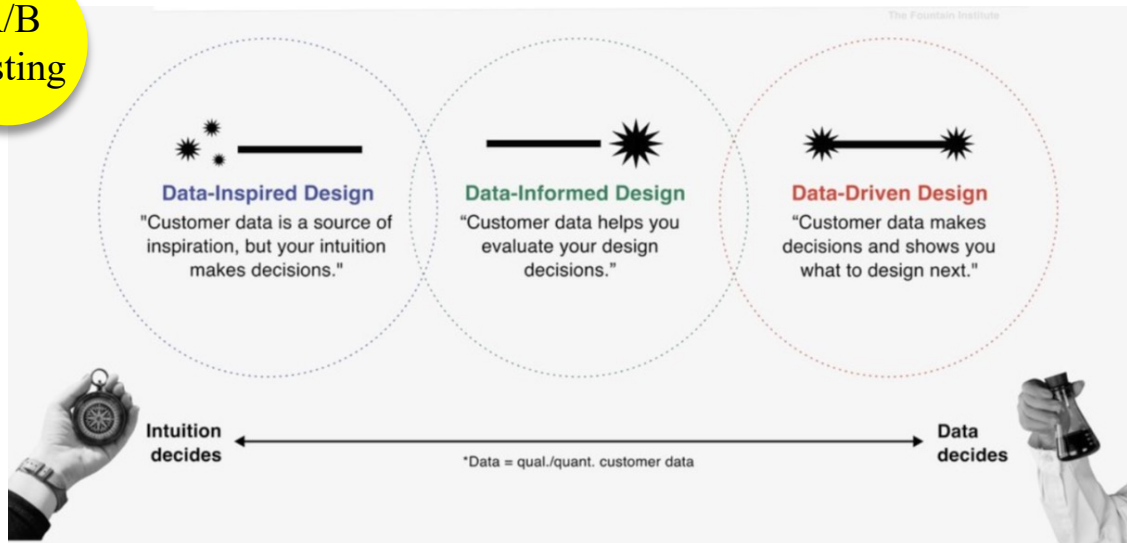
Null hypothesis is ...	True	False
Rejected	Type I error False positive Probability = $\alpha$	Correct decision True positive Probability = $1 - \beta$
Not rejected	Correct decision True negative Probability = $1 - \alpha$	Type II error False negative Probability = $\beta$

 Scribbr

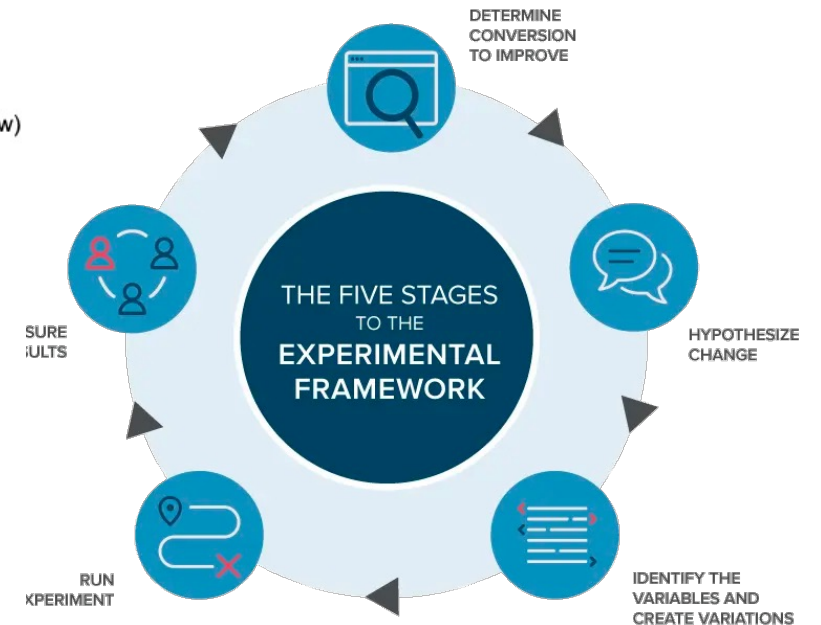




## A/B Testing



- **A/B testing is a special kind of quantitative experiments**
  - where "A" representing the old control, and "B" representing a single / multiple experimental changes
  - where the goal is to maximize single/multiple measurements of user behavior
  - A/B testing is most useful when combined with qualitative methods (e.g. observation, survey, interview)
- **A/B testing has multiple goals as listed below**
  - Finding a winner among many design variations
  - Testing hypotheses of cause (design) and effect (user behavior)
  - Gaining generalizable understanding of the task and the user population
  - Developing a reusable platform for continuously running a series of A/B tests
- **A/B testing consists of five big steps**
  1. Choose a single (or couple) quantifiable measure(s) of user behavior
  2. Construct hypotheses
  3. Create design variations
  4. Run the experiment
  5. Analyze the results



A/B Testing

**1. Equally allocate users to design variations (if possible)**  
 Don't assign majority of users to specific conditions. It will slow down the experiment by increasing the minimum sample sizes to get statistical significance.

**2. Get sufficient samples to get statistical significance of at least 95% (p<0.05)**  
 300-400 samples per variation is usually considered enough. If you cannot reach the 95% (p=0.05) with 300-400 samples \* # variations, it is very unlikely that the variations have strong impact on the measured goal.

**3. Run the experiment for multiple behavioral cycles**  
 Lots of user behaviors tend to have weekly patterns (e.g. on Monday, people tend to be quite busy). To minimize the risk of weekly biases, run the experiment for at least one week (preferably two weeks or a month). However, it depends on what the task is and who the users are.

**4. Pay attention to external factors**  
 If it is Christmas or Valentine's day, an online marketing campaign might be highly affected by the external factor. While it's impossible to get a perfectly clean environment, consider external factors in your interpretation.

- A/B testing is not silver bullet. It has a lot of limitations and pitfalls.

When A/B testing is not worth	What's the problem? How can we make it worth again?
<b>You don't have meaningful traffic (i.e. # users)</b>	<ul style="list-style-type: none"> <li>Without meaningful traffic you won't be able to tell anything about statistical significance.</li> <li>Reduce # variations</li> <li>Wait until you have a big enough population</li> <li>Construct a hypothesis that you can indirectly test on a similar platform (e.g. running A/B test of your streaming platform on YouTube users), crowdsourcing platform (e.g. Amazon Mechanical Turk), or as a lab experiment</li> </ul>
<b>You don't have enough resources for running A/B test.</b>	<ul style="list-style-type: none"> <li>You will regret, "We should've finished our design first."</li> <li>Wait until you have enough time</li> <li>Consider using <a href="#">existing tools</a> for A/B testing. They are not mature yet, but still useful for marketing.</li> </ul>
<b>You don't have an informed hypothesis.</b>	<ul style="list-style-type: none"> <li>Gather more information and perform EDA</li> <li>Treat A/B test like real science. Learn from lectures, case studies, and tutorials. Ask psychologists and data scientists who have conducted randomized controlled experiments what hypothesis you can test for the given situation</li> </ul>

# Visual Encoding Design

data**voyager**

Bookmarks (17)

**Data**

Cars Change

**Fields**

- A Name
- A Origin
- Year
- # Acceleration
- # Cylinders
- # Displacement
- # Horsepower
- # Miles\_per\_Gallon
- # Weight\_in\_lbs
- # COUNT

**Wildcard Fields**

- # Quantitative Fields
- A Categorical Fields
- Temporal Fields

**Encoding** Clear

x # COUNT

y A Origin

**Mark** bar

size Drop a field here

color # Year

shape Drop a field here

detail Drop a field here

text Drop a field here

**Facet**

row Drop a field here

column Drop a field here

**Wildcard Shelves**

any Drop a field here

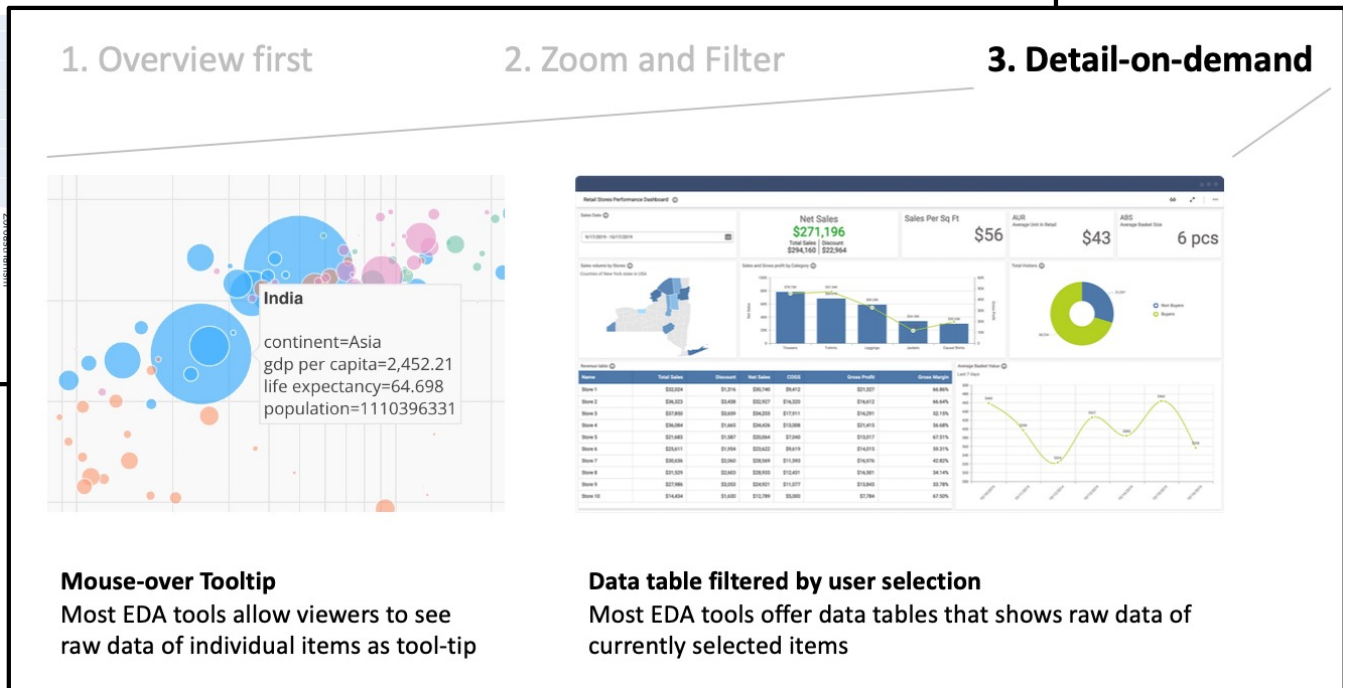
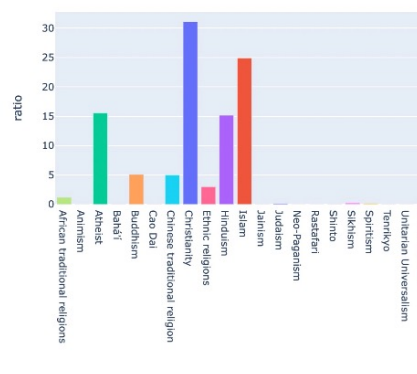
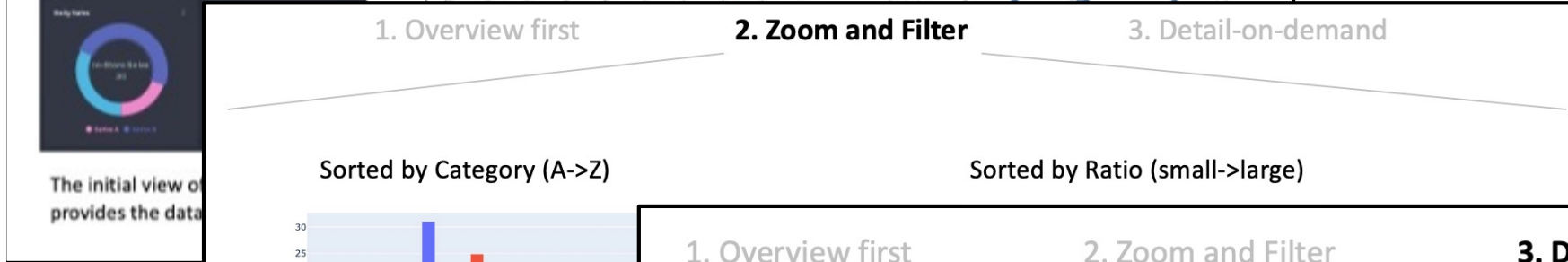
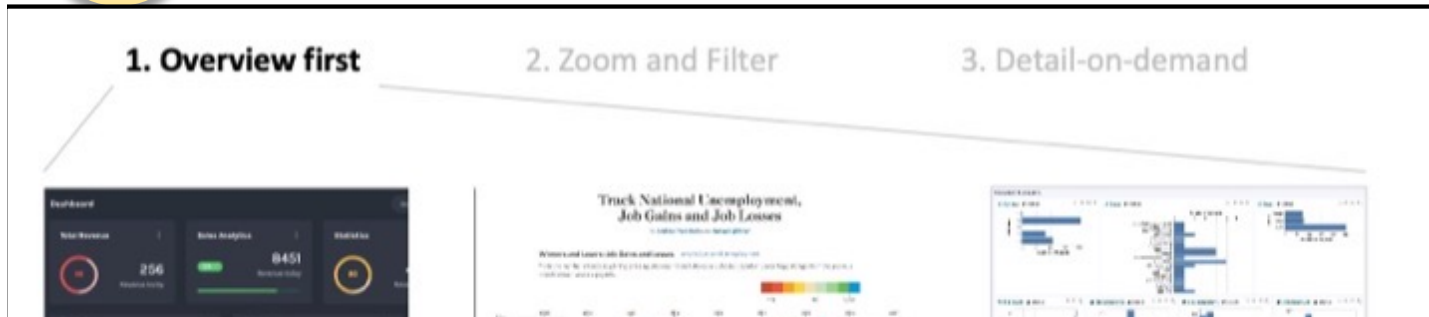
**Filter**

Drop a field here

Example	Encoding	Ordered	Useful values	Quantitative	Ordinal	Categorical	Relational
	position, placement	yes	infinite	Good	Good	Good	Good
1, 2, 3; A, B, C	text labels	optional alpha or num	infinite	Good	Good	Good	Good
	length	yes	many	Good	Good		
	size, area	yes	many	Good	Good		
	angle	yes	medium	Good	Good		
	pattern density	yes	few	Good	Good		
	weight, boldness	yes	few		Good		
	saturation, brightness	yes	few		Good		
	color	no	few (<20)			Good	
	shape, icon	no	medium			Good	
	pattern texture	no	medium			Good	
	enclosure, connection	no	infinite			Good	Good
	line pattern	no	few				Good
	line endings	no	few				Good
	line weight	yes	few		Good		

Interaction

# Ben Shneiderman's information-seeking mantra





# 生命之桥

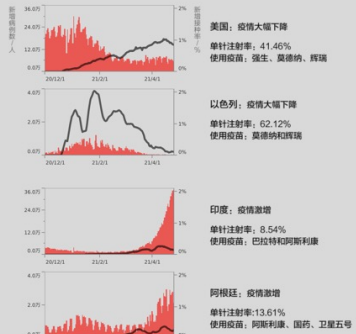
## 新冠疫苗分配的富与贫

新冠疫苗是过河之路，帮助我们渡过疫情的激流；新冠疫苗是连接之桥，将每个国家、每个生命彼此联结。  
我们用艺术可视化的方式，具有交互性的可视化海报，以横向阅读的叙事方式，展现了新冠疫苗分配不公平的情况，希望引起人们的关注，让更多人参与对公平疫苗分配之中。

请向右翻阅，横向阅读 →

### 新冠疫苗的接种对疫情的控制有重要作用

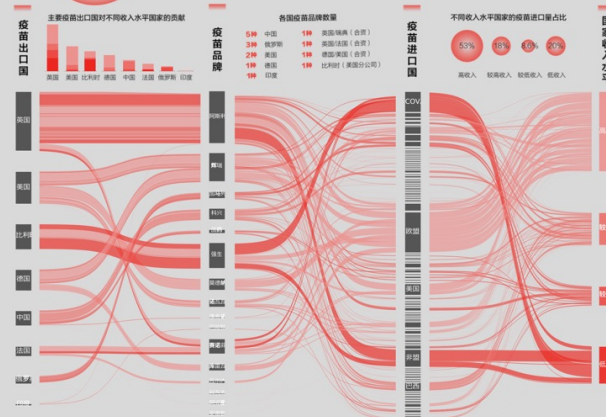
我们分别选取了疫苗接种率较高的美国和以色列，以及较低的印度和阿根廷，通过展现新增病例和疫苗接种率的关系，体现疫苗对于疫情控制的重要性。



数据来源：[https://www.kaggle.com/gpredscovid-world-vaccination-progress\\_截止2021.4.25](https://www.kaggle.com/gpredscovid-world-vaccination-progress_截止2021.4.25)  
[https://ourworldindata.org/covid-vaccinations\\_截止2021.4.25](https://ourworldindata.org/covid-vaccinations_截止2021.4.25)

### 新冠疫苗流向

我们选取了8个疫苗主要出口国，分析得出疫苗主要流向高收入国家，而低收入国家由于covax计划也获得较多疫苗。



数据来源：[https://public.tableau.com/prof/eduke\\_global\\_health\\_innovation\\_center#/v/home/COVID-19/VaccineAdvancedMarketCommitmentsbyCountry\\_16131542122100\\_截止2021.4.9](https://public.tableau.com/prof/eduke_global_health_innovation_center#/v/home/COVID-19/VaccineAdvancedMarketCommitmentsbyCountry_16131542122100_截止2021.4.9)

### 疫苗接种

据世界卫生组织统计，截至4月13日，全球共有15个国家的接种量数超1千万剂，共占全球总接种量的84%。与此同时，将近130个国家，超过25亿人口没有接种。

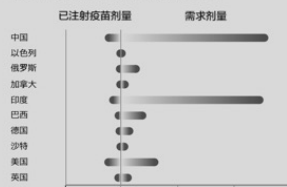


数据来源：[https://www.kaggle.com/gpredscovid-world-vaccination-progress\\_截止2021.4.25](https://www.kaggle.com/gpredscovid-world-vaccination-progress_截止2021.4.25)  
[https://ourworldindata.org/covid-vaccinations\\_截止2021.4.25](https://ourworldindata.org/covid-vaccinations_截止2021.4.25)

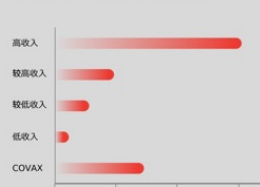
### 囤积新冠疫苗

报告显示，疫苗接种率已达到人口的90%，可以达到群体免疫，而当今世界疫苗接种情况仍不容乐观，已注射剂量远远达不到需求剂量。  
本图表中需求剂量 = 人口 \* 2 (剂量) \* 90%

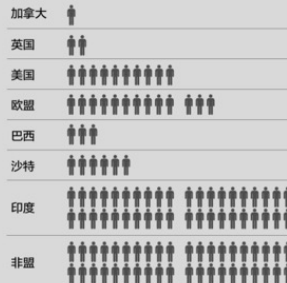
### 疫苗已注射剂量远小于需求剂量



### 不同收入水平国家的订购剂量

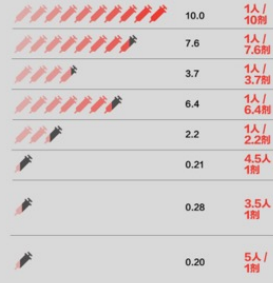


### 疫苗需求剂量对比



部分高收入国家需求小，却在大量囤积疫苗；与此同时没有一个中低收入国家拥有足够疫苗。

### 人均疫苗占有剂量对比 (基于订购剂量计算)

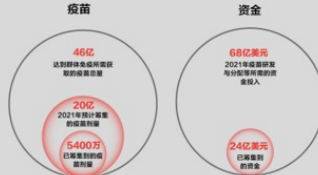


数据来源：[https://www.covid19-vaccine.live\\_截止2021.5.14](https://www.covid19-vaccine.live_截止2021.5.14)

### COVAX: 与低收入国家分享新冠疫苗的国际行动

世卫组织强烈批判在严峻的全球疫情形势下囤积疫苗的“疫苗民族主义”行为，为促进新冠疫苗的公平性，提出COVAX这一解决方案。

参与国家 190  
参与的低收入国家 92  
累计送达国家 121  
累计运送疫苗剂量 54,007,670



“ NO ONE IS SAFE UNTIL EVERYONE IS SAFE. ”

### 紧急使用清单的疫苗品牌对比

综合来看，中国国药较适合于公平分配计划。

	疫苗技术	接种剂量	有效性	价格	严重副作用	储存条件
中国	灭活	2	79%	\$44/剂	无	2°C-8°C
美国	莫德纳 mRNA	2	94%	\$25-\$37/剂	无	-25°C--15°C (6个月) 2°C-8°C (30天)
美国	强生 腺病毒载体	1	66%	\$10/剂	无	2°C-8°C (3个月)
英国/法国/德国	阿斯利康/牛津 腺病毒载体	2	70%	\$3-\$4/剂	血栓	2°C-8°C (6个月)
中国	辉瑞 mRNA	2	95%	\$20/剂	需冷藏/严重过敏	-80°C--20°C (6个月) 2°C-8°C (5天)

### 疫情无国界，新冠疫苗是生命之桥

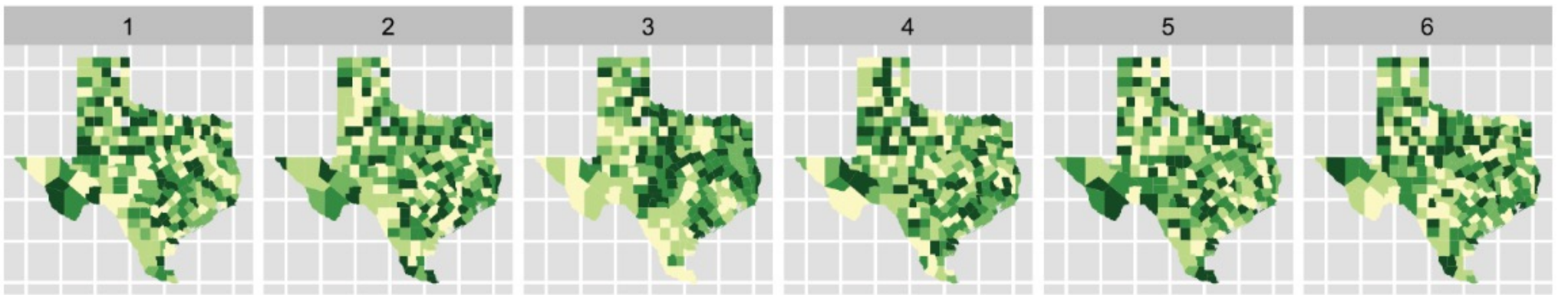
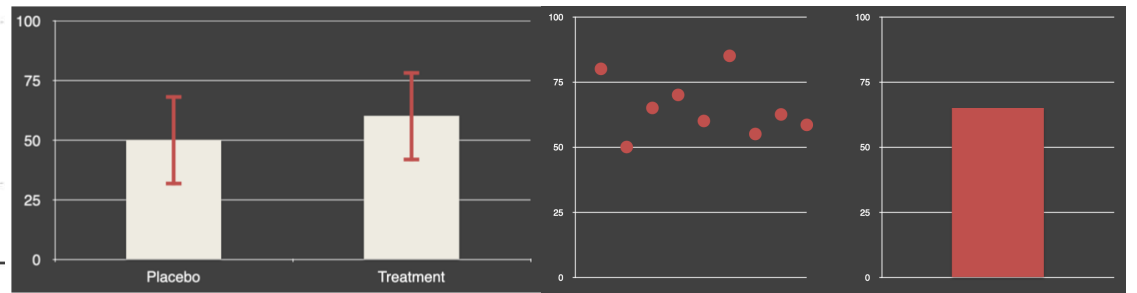
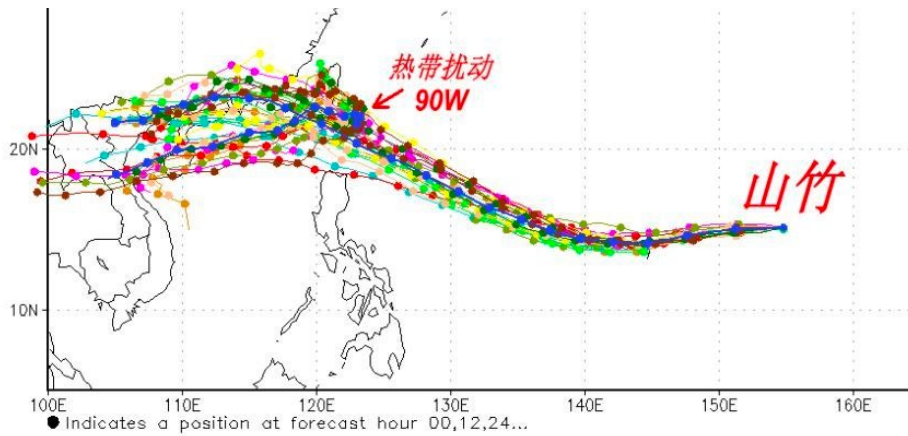
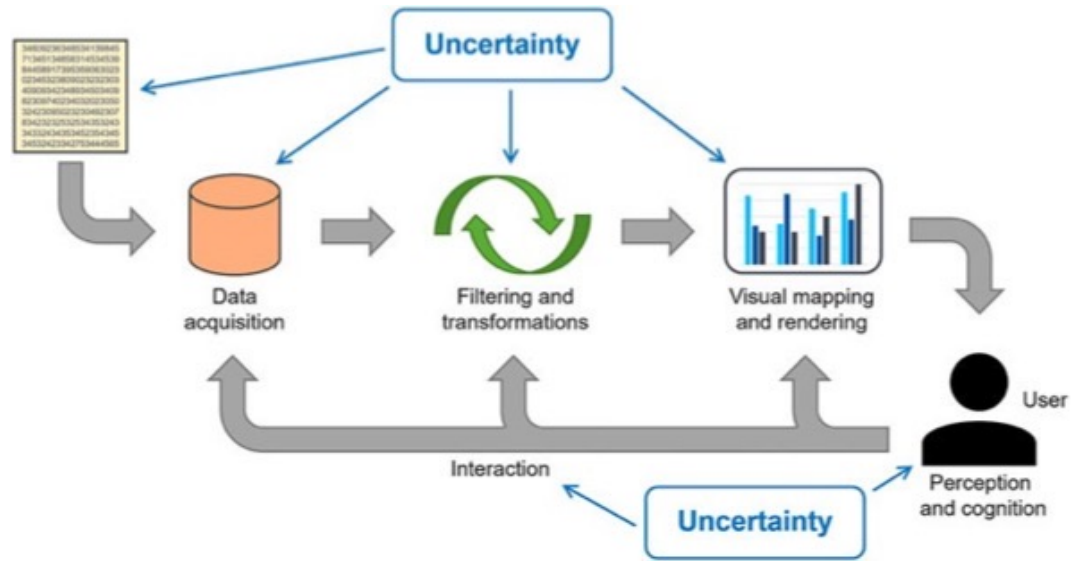
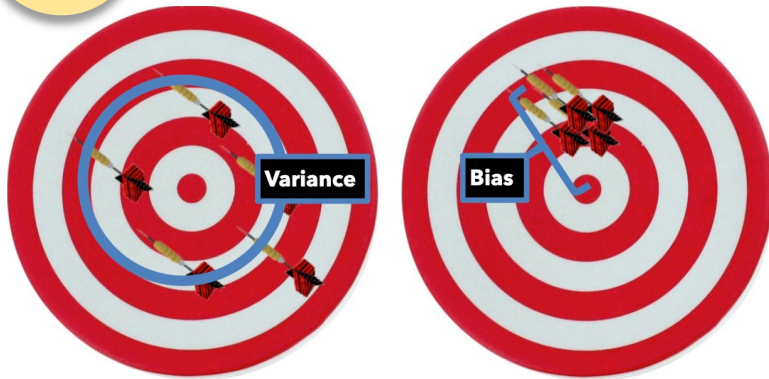
面对百年疫情，新冠疫苗的研发与分配绝不是国与企业之间的竞争，而是全人类与病毒的决战，捍卫人类生命福祉的社会理性应高于市场主导的经济逻辑。

人类是一个命运共同体，只有确保所有处于危险之中的人得到有效防护，人类才能彻底战胜病毒，真正实现整体安全。因为正如谭德塞所说，抗击疫情“必须从一个全球大家庭的角度出发”。此时此刻，我们能做的只有团结。

数据来源：[http://www.360doc.com/content/21/03/22/09/3843034\\_968229393.shtml](http://www.360doc.com/content/21/03/22/09/3843034_968229393.shtml)  
<https://news.un.org/zh/story/2021/05/1083742>



# Uncertainty



# The Takeaways

In our life, study, work

Understand the intrinsic nature of data

Make use of it for our purpose

# Time for Course Evaluation

Your participation is important!

---

# Course Evaluation

## 方法及步骤

1. 网页端：登录教务系统：<https://tis.sustech.edu.cn/> -业务办理-评教任务-2024春季学期学生评价任务。系统按课程类型设置评价任务（理论类、实验实践类、体育类、艺术类），如页面上有多个评价任务，请逐一进入并提交评价。
2. 微信端：通过微信进入“南方科技大学”微信企业号--教学质量管理平台，在“我的任务-待评”中填写并提交本学期所选课程的所有听课评价。

---

# Course Project | Final Showcase

- Final Project Showcase on **Data Storytelling**
  - **Submit** first draft before Fri Jun 07 @ 12:00 | **Present** on Fri Jun 07 in-class
  - **Submit** final version before Sunday Jun 09 @ 23:30
  - ~ 8 mins presentation regarding your data-driven story
  - ~ 2 mins Q&A





---

<https://ds363.ancorasir.com/>

**Thank you~**

Wan Fang  
Southern University of Science and Technology