数据思维与实践

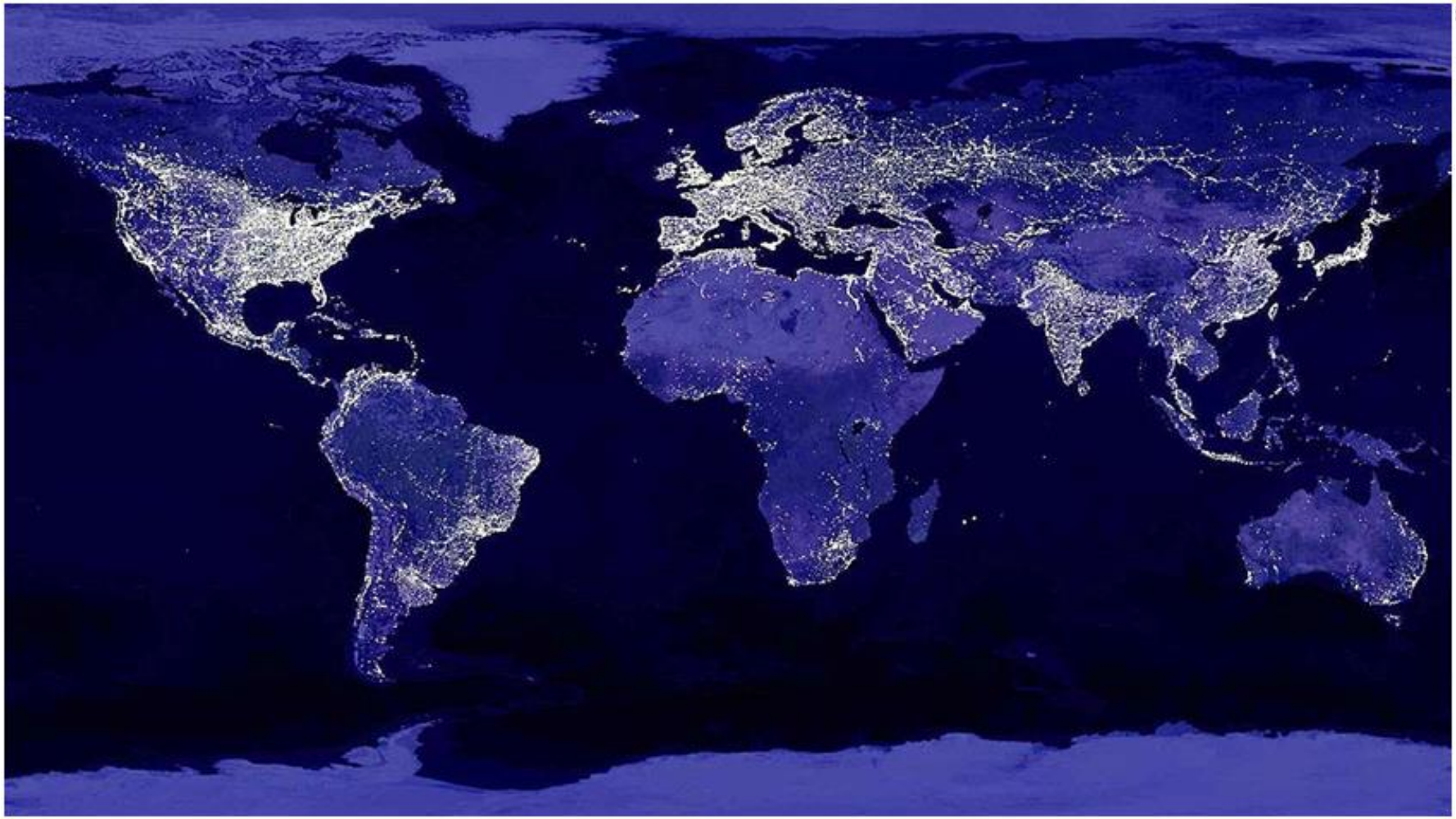
王 伟

华东师范大学

数据科学与工程学院

全民数字素养与技能培训基地

开篇实例：NASA 从太空中拍摄城市夜间亮度



来源：https://www.nasa.gov/

Satellite Reveals New Views of China

VIIRS light data
by NASA

当年的摩拜单车


mobike 摩拜单车
351192
总骑行距离（公里）
42563
总减排量（千克）
1917/2843
被骑行 / 总运营车辆
153/5163
实时 / 总骑行次数
昨天
今天
骑行次数
1500
1000
500
0
0:00
2:00
4:00
6:00
8:00
10:00
12:00
14:00
16:00
18:00
20:00
22:00
性别比例
年龄比例
骑行比例
长宁区
静安区
卢湾区
东华大学
徐汇区
上海体育馆
漕河泾
上海植物园
华东理工大学
三林镇
北股家宅
S20
复旦大学
附属儿科医院
外环高速
华泾镇
乔里康
莘庄立交
为民村

数据点亮城市

【摩拜单车】《在城市中的一天》（深圳）

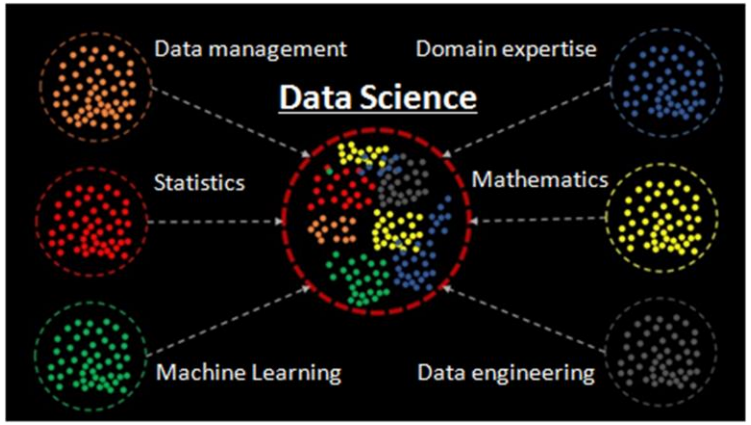
视频选集 (4/18)　自动连播

P1 北京	01:35
P2 上海	01:35
P3 广州	01:25
P4 深圳	01:35
P5 成都	01:36
P6 长沙	01:32
P7 南京	03:43
P8 杭州	00:39
P9 福州	01:32
P10 济南	01:35

https://www.bilibili.com/video/BV19W411J7s6

数据思维与实践

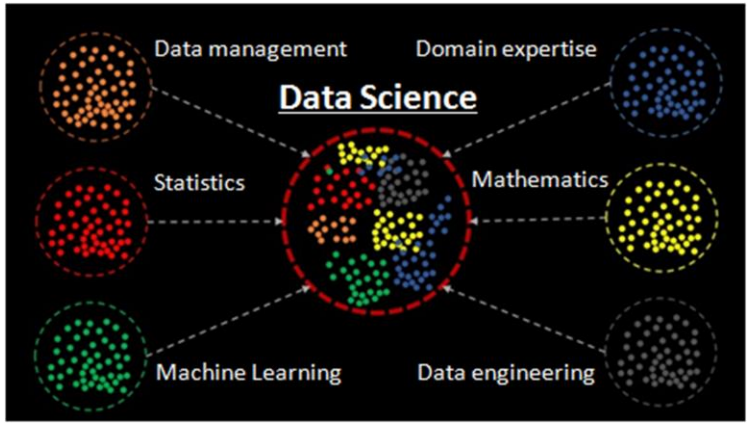
第01讲 数据科学与数据思维入门



- 从数据思维到第四范式
- 数据科学与工程
- 数据科学实践场景：开源数字王国
- 初识 Python 重要扩展库

数据思维与实践

第01讲 数据科学与数据思维入门



- **从数据思维到第四范式**
- 数据科学与工程
- 数据科学实践场景：开源数字王国
- 初识 Python 重要扩展库

智能时代 I

ALL Systems Go

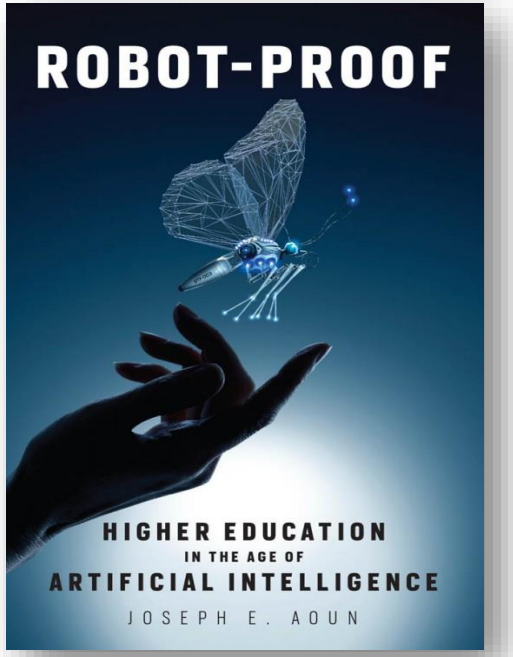
At last — a computer program that can beat a champion Go player

***Nature* 2016.01**

DARK FACTORY

The robotics revolution is changing what machines can do. Where do humans fit in?

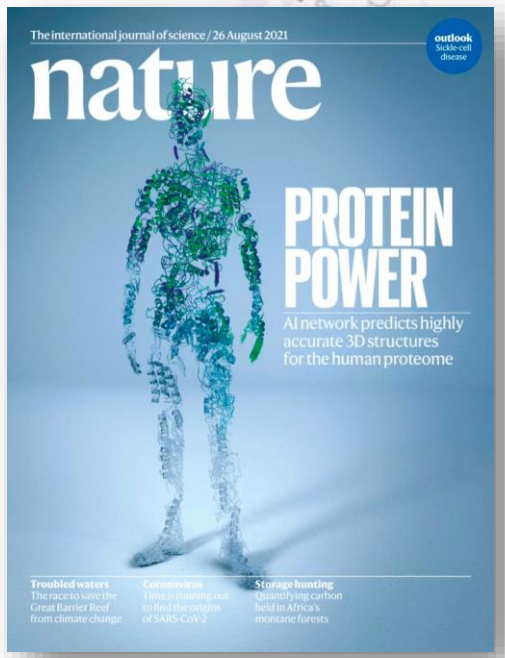
***The New Yorker* 2017.10**



Robot-Proof

Higher Education in the Age of Artificial Intelligence

***MIT Press* 2017.08**



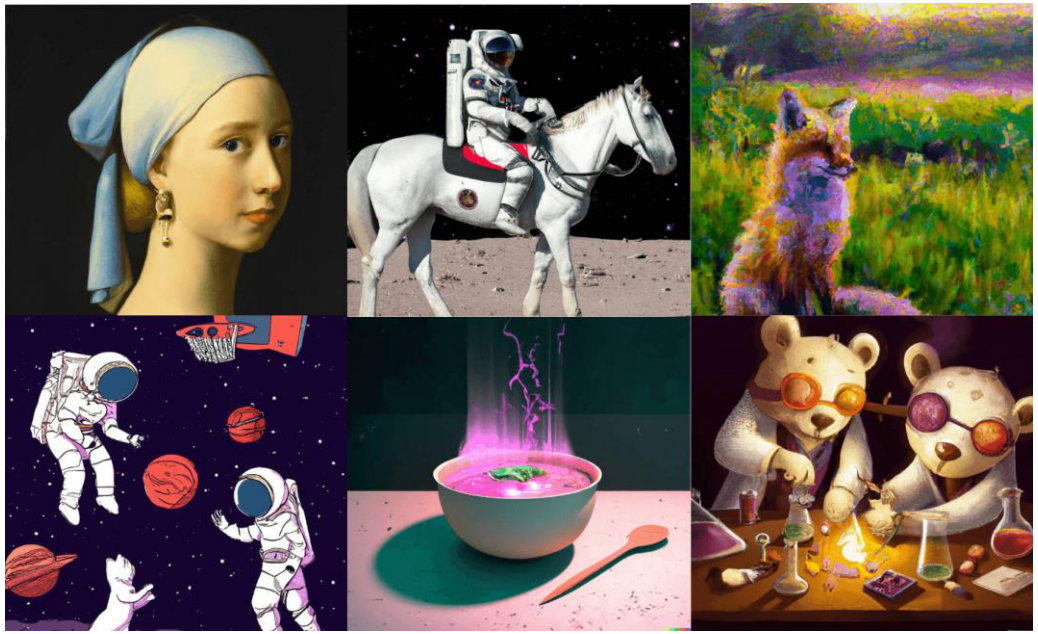
AlphaFold

AlphaFold, software that can predict the 3D shape of proteins, is already changing biology.

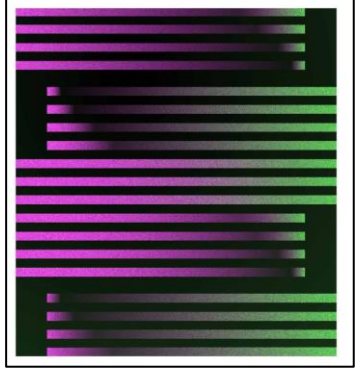
***Nature* 2021.08**

智能时代 II

DALL·E



ChatGPT: Optimizing Language Models for Dialogue



vitalik.eth
@VitalikButerin · Follow

I can easily see many jobs in the next 10-20 years changing their workflow to "human describes, AI builds, human debugs".

Noah Smith captured the human-AI dynamic succinctly in his "sandwich" workflow:

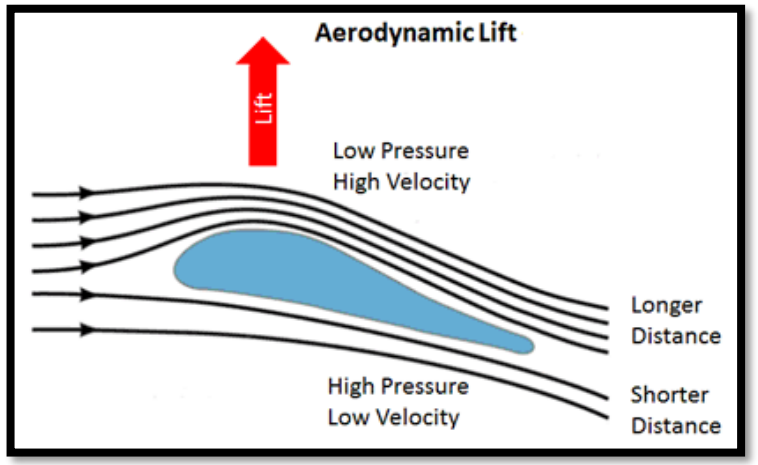
> *"This is a three-step process. First, a human has a creative impulse, and gives the AI a prompt. The AI then generates a menu of options. The human then chooses an option, edits it, and adds any touches they like."*

智能时代的我们：**人类描述** → **AI 构建** → **人类调试**

不被 AI 取代的必须： AI 不知道的事情（**专有**），理解 AI 不能理解的东西（**解释性**），制作 AI 还不知道的东西（**创造性**）。

航天学的启示

- 是**航空技术**而不是“人工飞行”
- 空气动力学、驾驶舱与仪表盘
- 技术智能也不应该是“人工的”，应该是**增强人类能力的智能**



智能时代下的教育

人类增强智能 = 人脑智能 + 技术智能

从面向“知识”到面向“能力”的转变

- **基本素养的提升**
  - **数字素养**（Digital literacy）、**数据素养**（Data literacy ）、**人文素养**（Human literacy）
- **核心能力的提升**
  - **学习能力、问题求解能力、信息获取能力、分析推理能力、决策能力** 等等
- **综合认知的提升**
  - **系统性思维**（System thinking）、**数据思维**（Data thinking）
  - **设计思维**（Design thinking）、**批判性思维**（Critical thinking）
  - **认知敏捷性**（Cognitive agility）、**创业精神**（Entrepreneurship）

我们身边的实例


上海 - 普陀区
18 ℃ 阴
天气：多云转中雨
温度：29℃ ~ 19 ℃
风况：南风 ≤3级
2022-04-22 07:55:22
杏园抗疫动态信息屏
小区感染者报告日期
2022-04-19
2022-04-16
2022-04-13
2022-04-05
2022-04-03
2022-04-02
2022-03-22
*信息来自上海市卫健委
封控楼栋标识
生活必需品团购信息
开团时间 开团内容 团购价格 到货时间 状态
2022-04-18 蔬菜菌菇套餐 85.51元 21日22点 正在分发，请关注微信短信
2022-04-17 蘑菇 10.06元 21日22点 正在分发，请关注微信短信
2022-04-16 猪肉 99.6元 优先保供政府 待退款
2022-04-14 猪肉 99.6元 20日已部分到货 50份发放完毕，17份待退款
*信息来自杏园徐建-蔬菜等必需品购买群

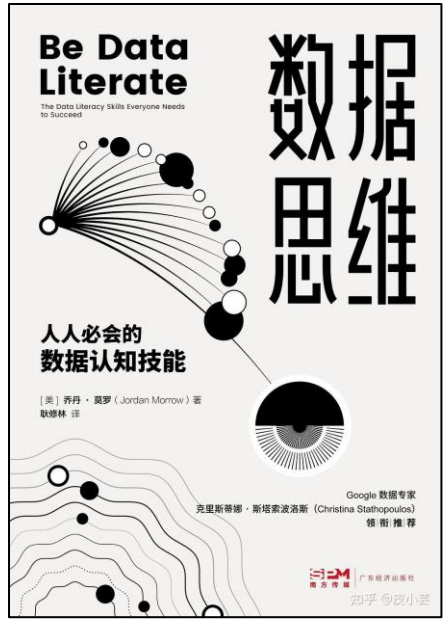

累计完成
7
核酸次数
4
物资发放次数
*统计由4月1日起
核酸日期
2022-04-20
2022-04-16（接4.12未完成部分）
2022-04-12
2022-04-09
2022-04-06
2022-04-04
2022-04-02
2022-04-01
发放日期 发放物品
2022-04-05 蔬菜
2022-04-09 蔬菜+肉
2022-04-11 连花清瘟+口罩
2022-04-19 米+油+菜+腊肠
居委电话：021-62868953
街道电话：021-62160359
市民热线：12345
杏园加油，我们必胜！
* 以上信息来源于业主群和网络信息，信息更新可能存在延迟，请以实际情况和官方通报为准。本动态屏仅限杏园内部参考使用，如


700
12
40.86%
14.07%
97.41%
99.24%
99.93%
99.80%


东区跑步群
10,802.57
1,732
65
57
打卡人次
跑步健儿
坚持周次
跑步打卡周数TOP20
跑步健儿姓名墙
积跬步时间轴

数据思维（Data Thinking）

- **数据思维**（Data Thinking）或 **数据素养**（Data Literacy）主要包括四个方面：
  - **Reading data（RD）：** 阅读数据资料和信息的能力；
  - **Working with data（WD）：** 用数据开展工作或活动的能力；
  - **Analyzing data（AD）：** **分析数据的能力**；
  - **Communicating with data（CD）：** 用数据进行表达、对话和沟通交流的能力。

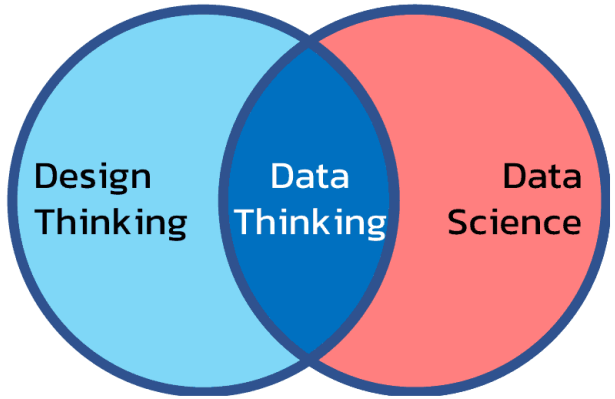


数据分析（AD）的四个层次

- **描述性分析**（Descriptive analysis）：发生了什么（过去与现在）
- **诊断性分析**（Diagnostic analysis ）：发生的原因（动因与洞察）
- **预测性分析**（Predictive analysis）：将要发生什么（趋势与可能）
- **指导性分析**（Prescriptive analysis ）：应该做什么（决策与优化）

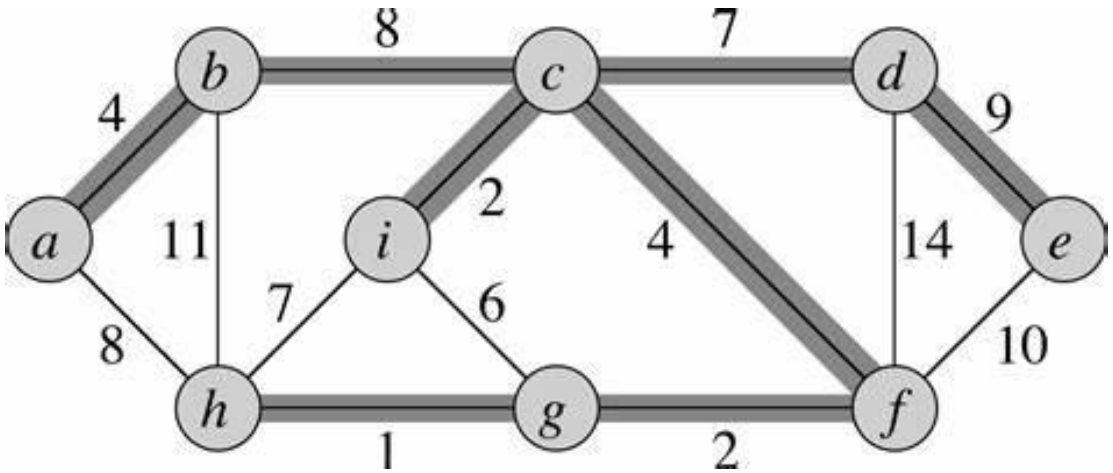
如何提升数据思维?

- 提高**对数据及其应用的意识**，懂得数据是资产也是生产力要素；
- 掌握**数据处理和分析的基本方法**，从数据中提炼出有价值的信息；
- 了解和掌握**数据分析方法和建模的使用条件**；
- 能够看懂数据处理的结果，**能从统计和数据科学角度进行认识**；
- **能与专业领域的分析认识有机的结合**起来，如**科学研究**。

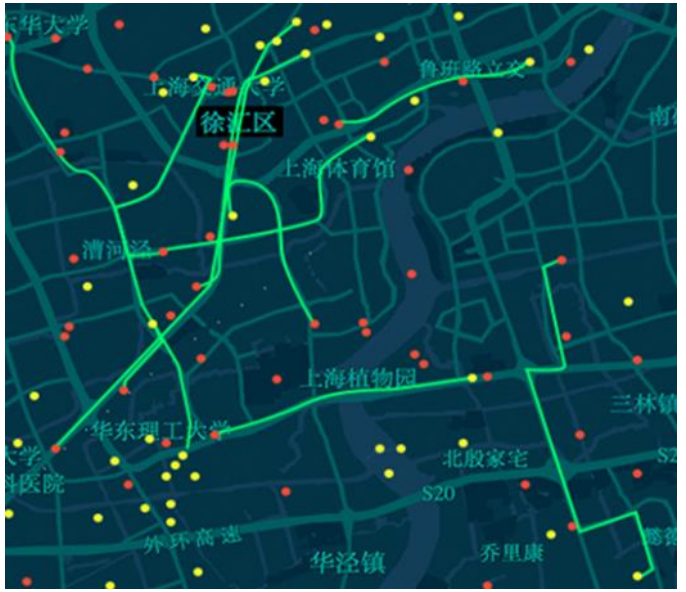


数据思维实例

行程设计：任意给出旅行的起点和终点，如何给出一个行程建议，使得在某些指标上“最短”？



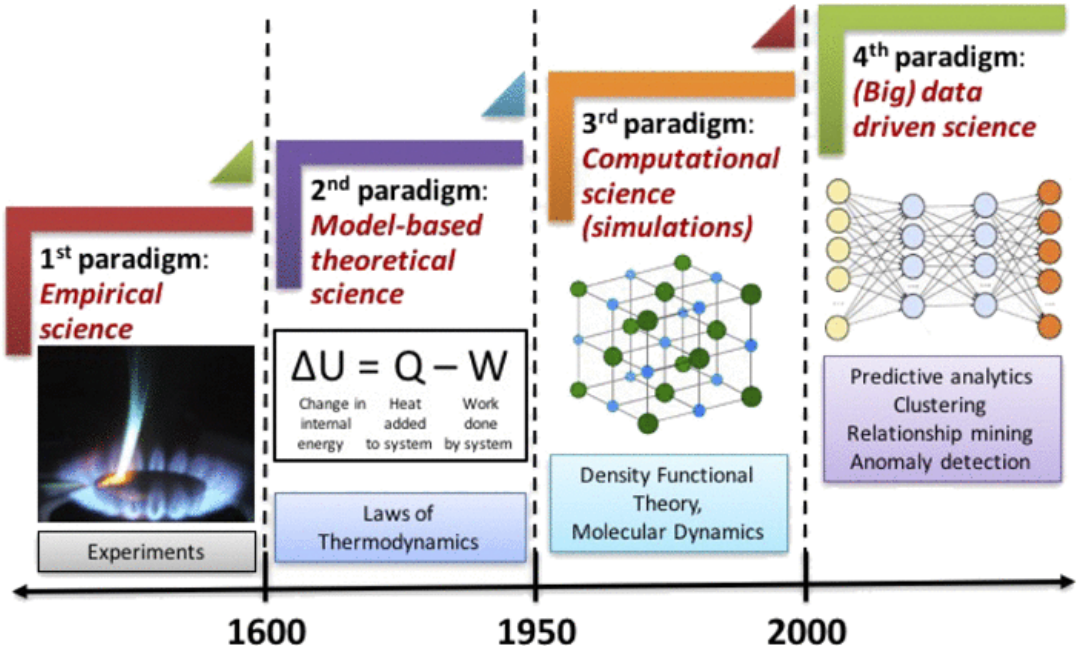
Dijkstra 算法或者动态规划算法来求解

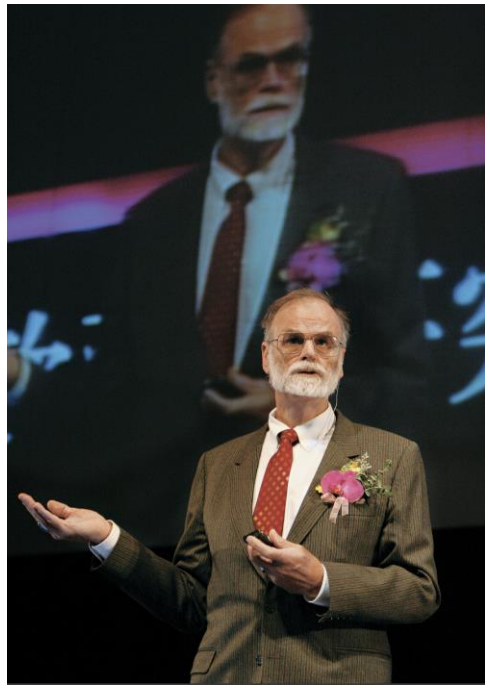


记录物理世界人们旅行的选择，构建数据模型，对最受欢迎的路线进行排序

科学研究的四类范式

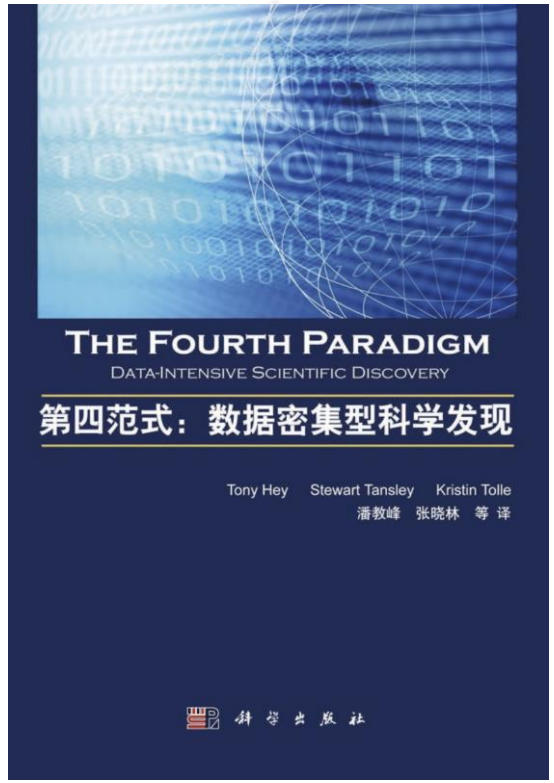
范式（Paradigm）的概念最初由美国著名科学哲学家 **Thomas Samuel Kuhn** 于 1962 年在《科学革命的结构》中提出来。指的是常规科学所赖以运作的**理论基础和实践规范**，是从事某一科学的科学家群体所共同遵从的**世界观和行为方式**。





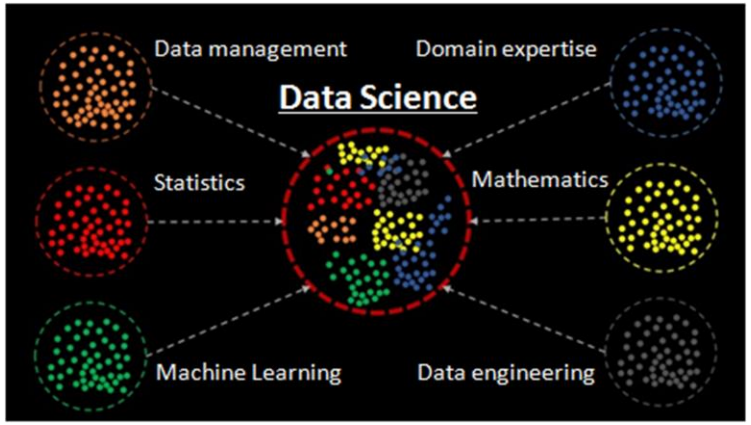
Jim Gray

第四范式：数据密集型科学



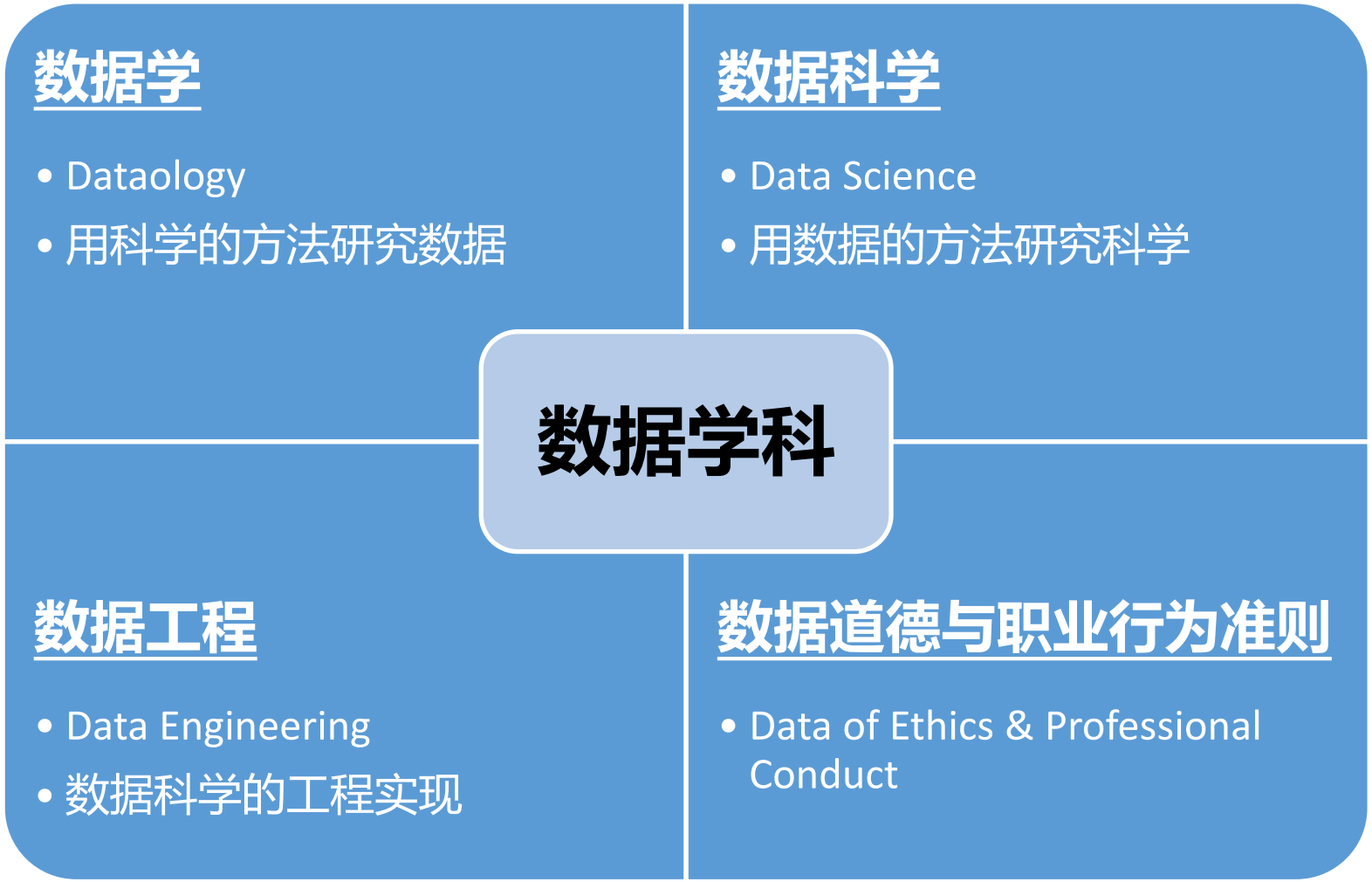
数据思维与实践

第01讲 数据科学与数据思维入门



- 从数据思维到第四范式
- **数据科学与工程**
- 数据科学实践场景：开源数字王国
- 初识 Python 重要扩展库

数据科学与工程的基本内涵



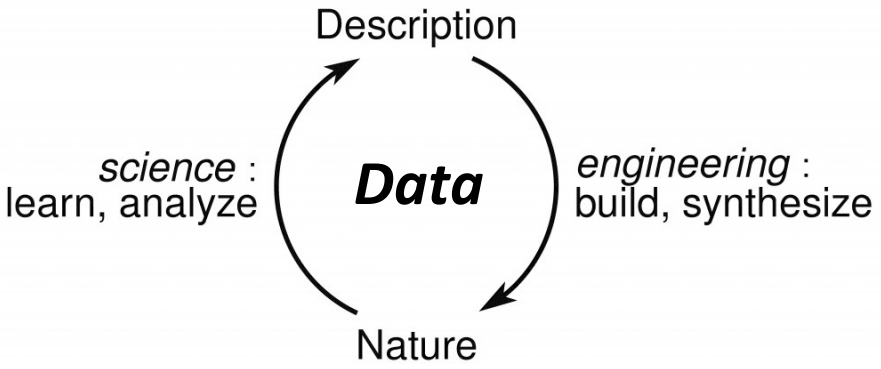
数据科学（Data Science）

- **数据科学**就是以数据为中心的，利用数据来开展：

❶ **理解世界** —— **科学方面**

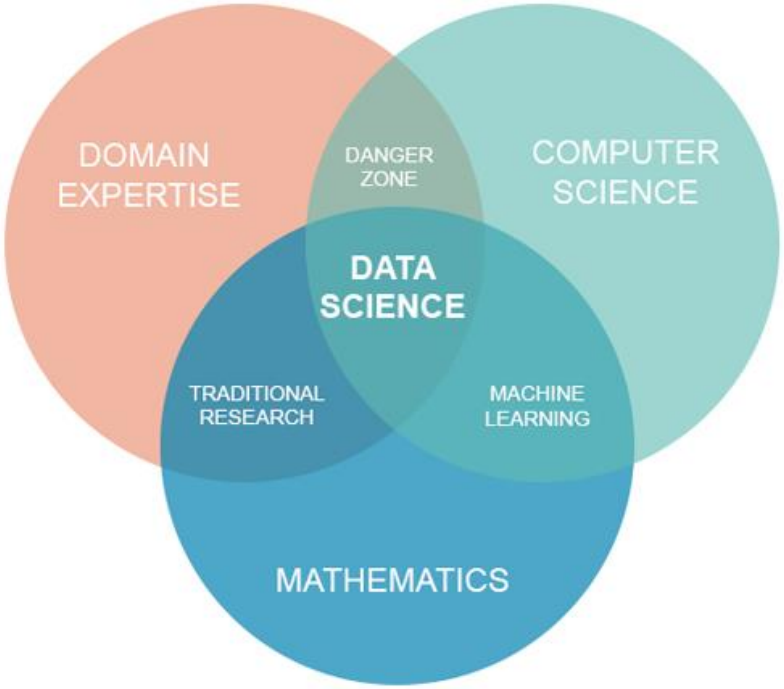
&

❷ **问题求解** —— **工程方面**



数据科学的基本内涵

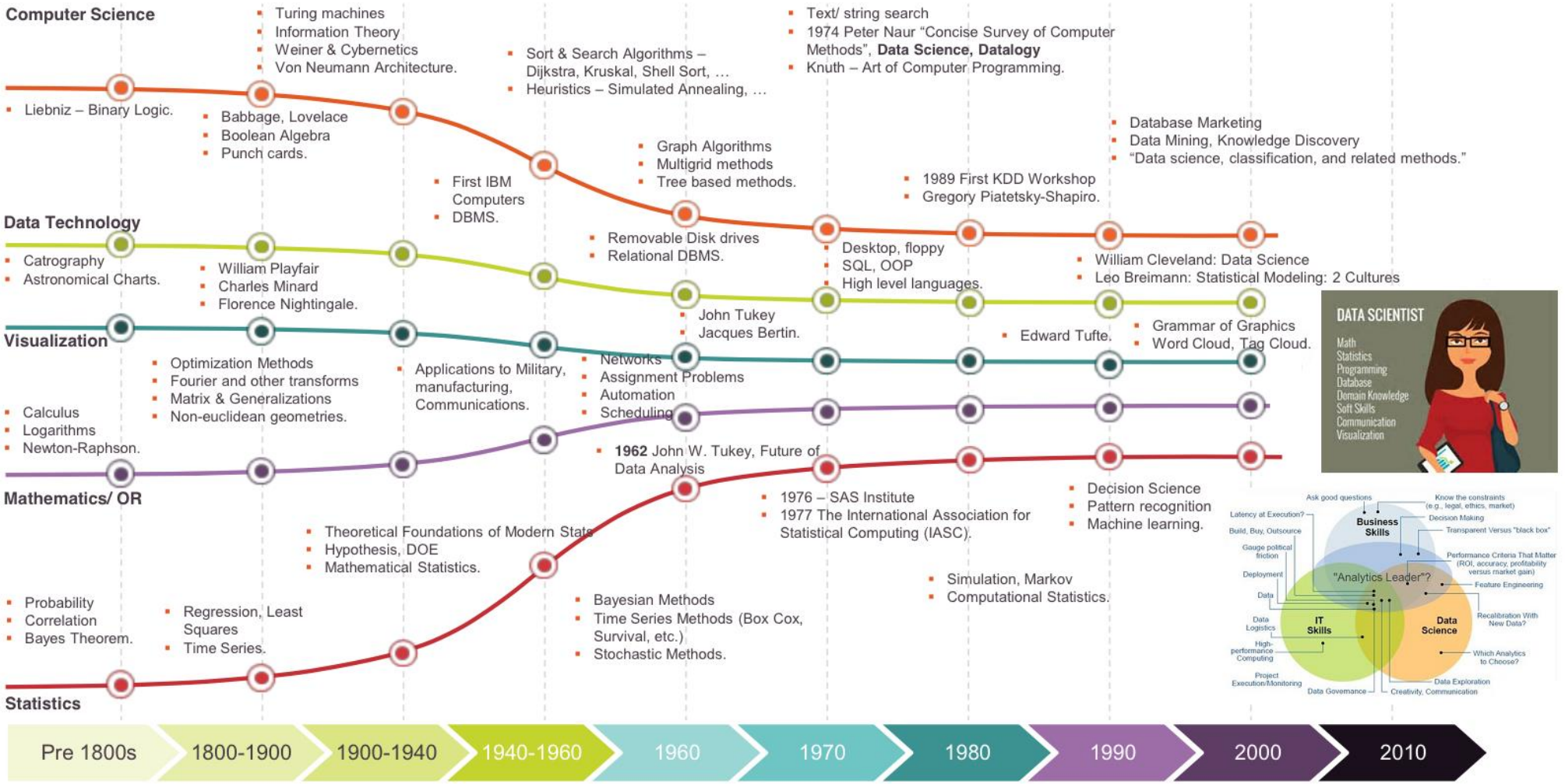
Data science is ***interdisciplinary***



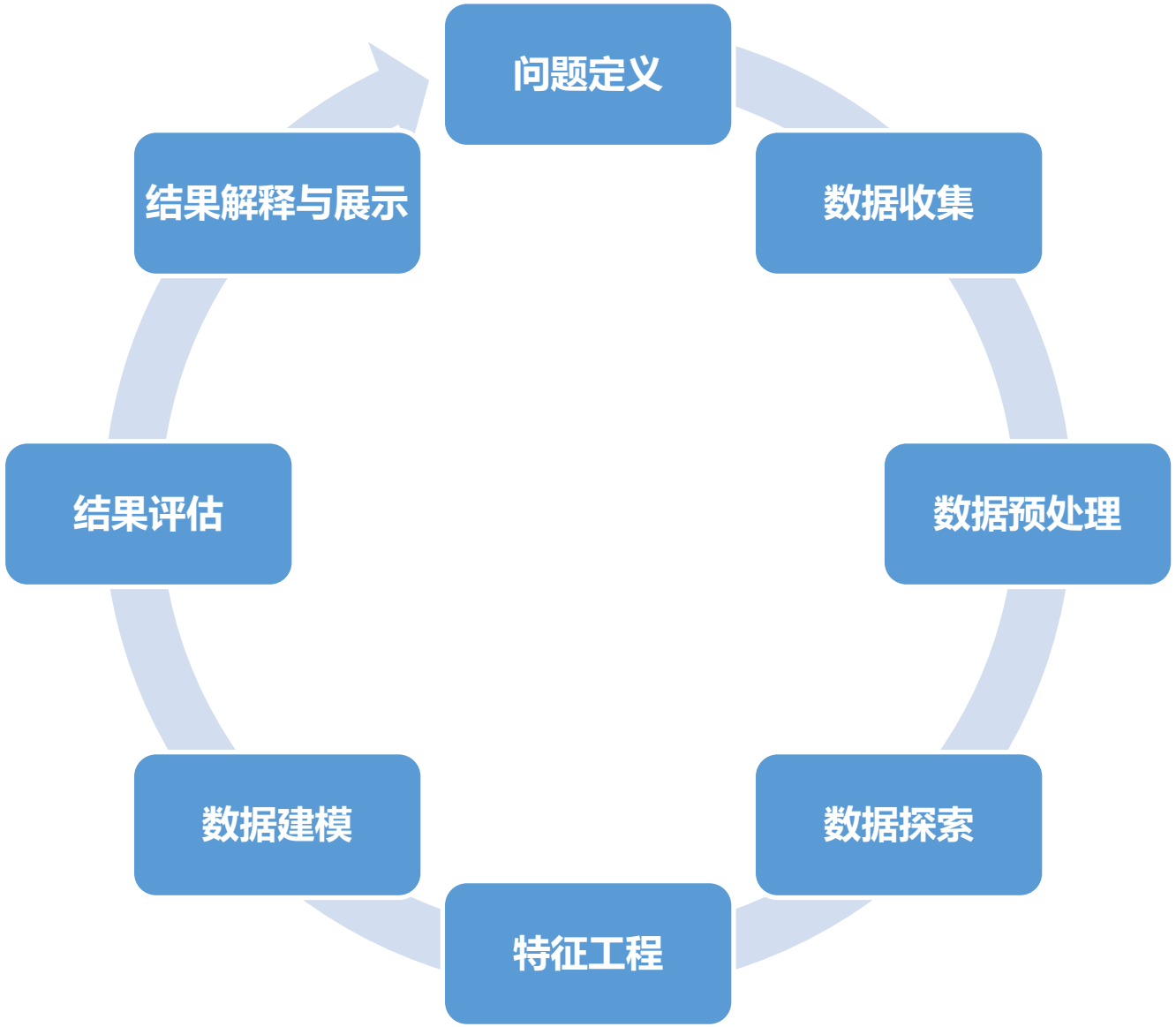
Drew Conway's Venn Diagram of Data Science

More ***Union*** than *Intersection*

数据科学的形成过程

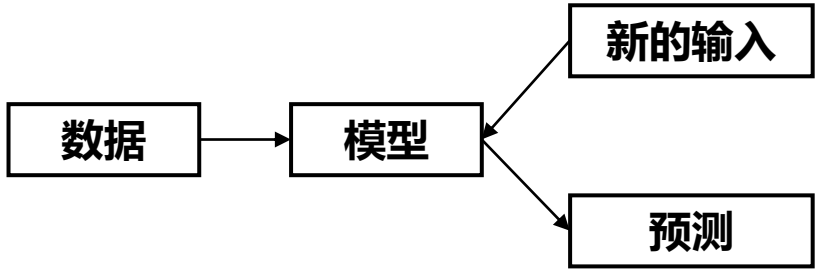

Computer Science
Liebniz – Binary Logic.
Babbage, Lovelace
Boolean Algebra
Punch cards.
Turing machines
Information Theory
Weiner & Cybernetics
Von Neumann Architecture.
Sort & Search Algorithms –
Dijkstra, Kruskal, Shell Sort, …
Heuristics – Simulated Annealing, …
Text/ string search
1974 Peter Naur "Concise Survey of Computer Methods", Data Science, Datalogy
Knuth – Art of Computer Programming.
Graph Algorithms
Multigrid methods
Tree based methods.
First IBM Computers
DBMS.
1989 First KDD Workshop
Gregory Piatetsky-Shapiro.
Database Marketing
Data Mining, Knowledge Discovery
"Data science, classification, and related methods."
Data Technology
Catrography
Astronomical Charts.
William Playfair
Charles Minard
Florence Nightingale.
Removable Disk drives
Relational DBMS.
Desktop, floppy
SQL, OOP
High level languages.
William Cleveland: Data Science
Leo Breimann: Statistical Modeling: 2 Cultures
Visualization
John Tukey
Jacques Bertin.
Edward Tufte.
Grammar of Graphics
Word Cloud, Tag Cloud.
Optimization Methods
Fourier and other transforms
Matrix & Generalizations
Non-euclidean geometries.
Applications to Military, manufacturing, Communications.
Networks
Assignment Problems
Automation
Scheduling
Calculus
Logarithms
Newton-Raphson.
Mathematics/ OR
1962 John W. Tukey, Future of Data Analysis
1976 – SAS Institute
1977 The International Association for Statistical Computing (IASC).
Decision Science
Pattern recognition
Machine learning.
Theoretical Foundations of Modern Stats
Hypothesis, DOE
Mathematical Statistics.
Simulation, Markov
Computational Statistics.
Probability
Correlation
Bayes Theorem.
Regression, Least Squares
Time Series.
Bayesian Methods
Time Series Methods (Box Cox, Survival, etc.)
Stochastic Methods.
Statistics
DATA SCIENTIST
Math
Statistics
Programming
Database
Domain Knowledge
Soft Skills
Communication
Visualization
Business Skills
IT Skills
Data Science
"Analytics Leader"?
Pre 1800s
1800-1900
1900-1940
1940-1960
1960
1970
1980
1990
2000
2010

数据科学过程



数据建模

- **数据**是人类经验的数字化形式。
- **数据建模**：捕捉数据的本质特征，根据数据的特征形成模型。
  - **按任务属性**：分类模型、聚类模型、推荐模型、……
  - **按数据属性**：图像模型、语音模型、文字模型、……
- **数据思维**是一种通过数据驱动决策的思维模式，包括：
  - 数据、模型、算力和业务模式



数据分析（建模）的方法

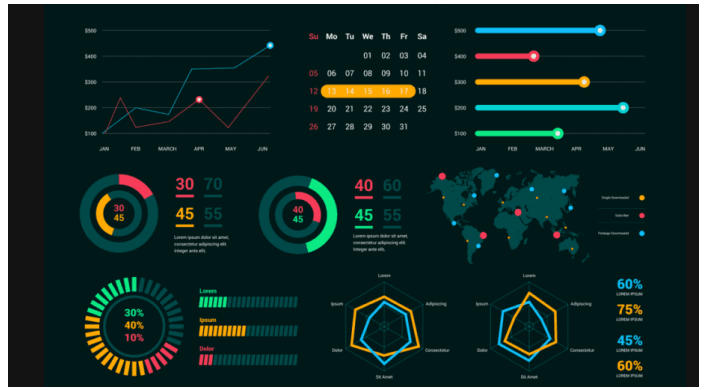
常用方法

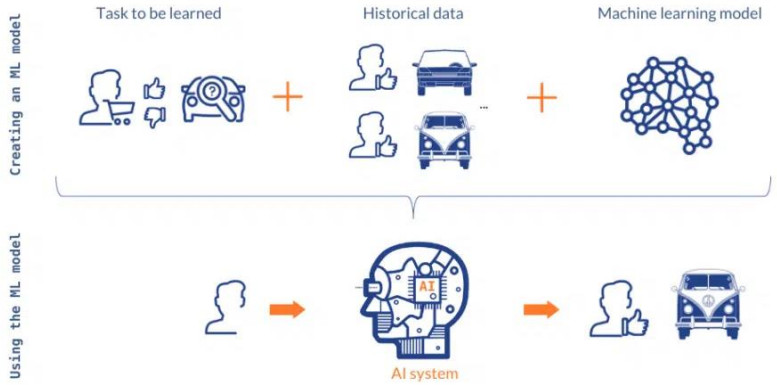
统计方法

可视化方法

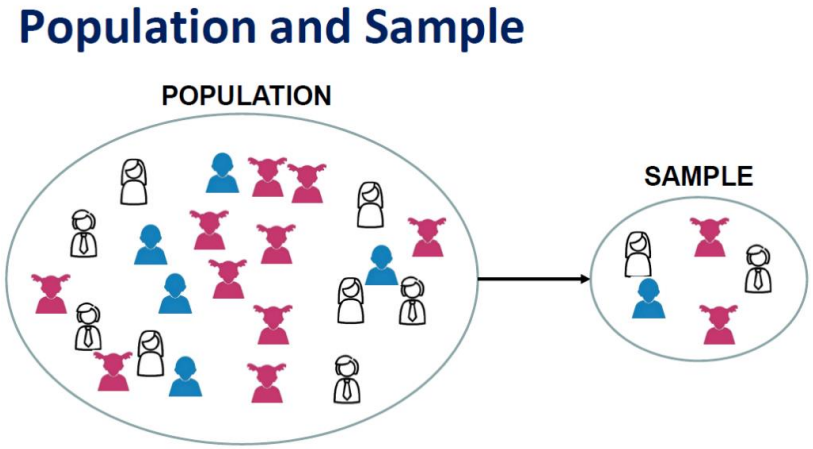
机器学习方法

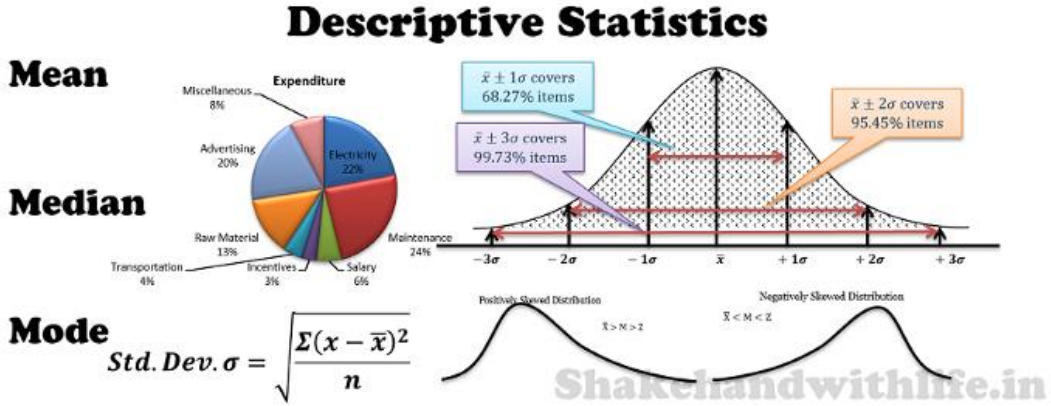
A statistical model

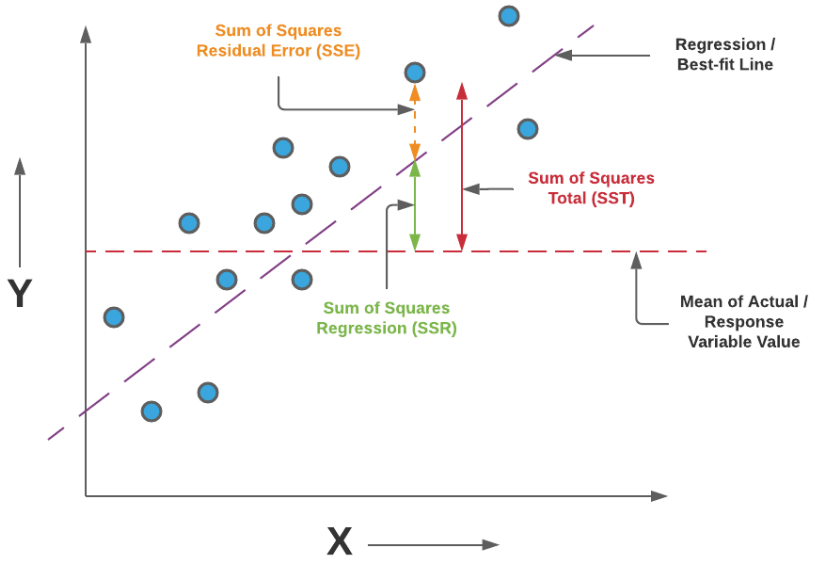


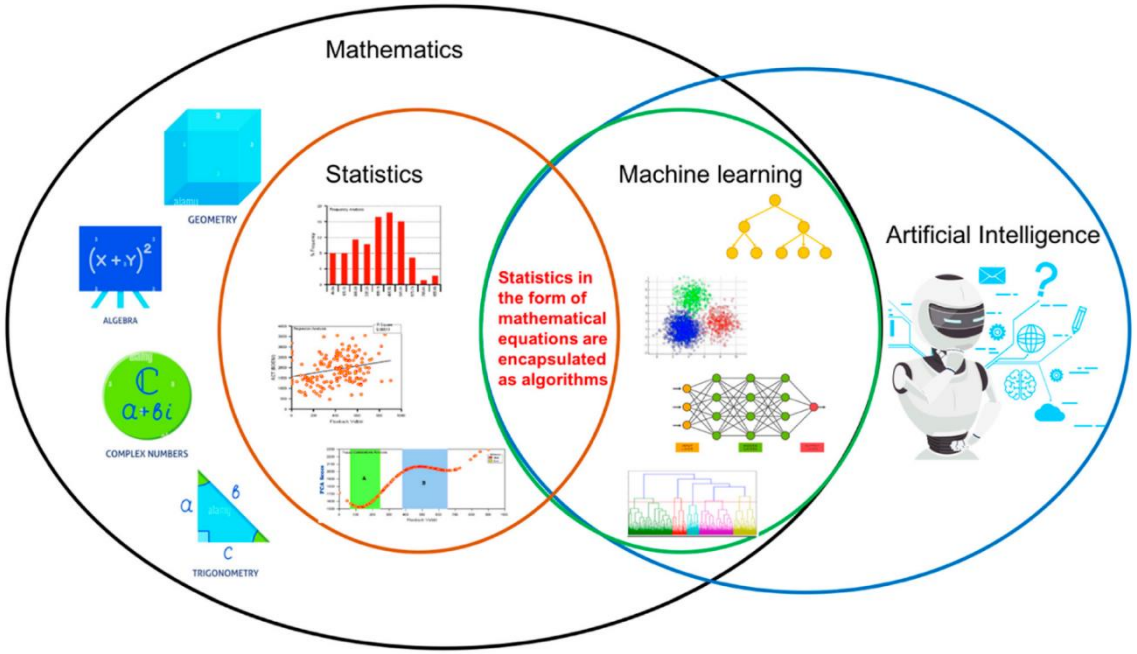


统计方法


Population and Sample
POPULATION
SAMPLE


Descriptive Statistics
Mean
Median
Mode
Std. Dev. σ =
Expenditure
Miscellaneous 8%
Advertising 20%
Electricity 22%
Raw Material 13%
Maintenance 24%
Transportation 4%
Incentives 3%
Salary 6%
x̄ ± 1σ covers 68.27% items
x̄ ± 2σ covers 95.45% items
x̄ ± 3σ covers 99.73% items
−3σ
−2σ
−1σ
x̄
+1σ
+2σ
+3σ
Positively Skewed Distribution
Negatively Skewed Distribution
Shakehandwithlife.in


Sum of Squares
Residual Error (SSE)
Regression /
Best-fit Line
Sum of Squares
Total (SST)
Sum of Squares
Regression (SSR)
Mean of Actual /
Response
Variable Value
Y
X

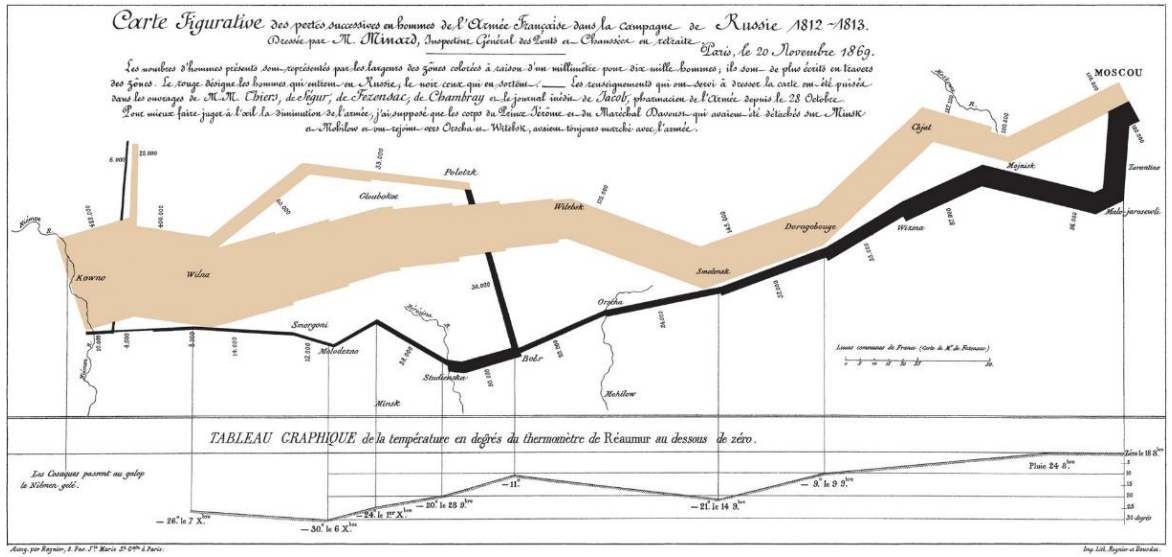

Mathematics
Statistics
Machine learning
Artificial Intelligence
GEOMETRY
ALGEBRA
COMPLEX NUMBERS
TRIGONOMETRY
Statistics in the form of mathematical equations are encapsulated as algorithms

可视化方法

- **数据可视化**是利用计算机图形图像处理技术，将数据转换为图形或者图像，在屏幕上显示出来进行交互处理的理论方法和技术。
- 数据可视化不仅是一种工具和技术，同时是一种**表达数据的方式**，它是对现实世界的抽象表达。

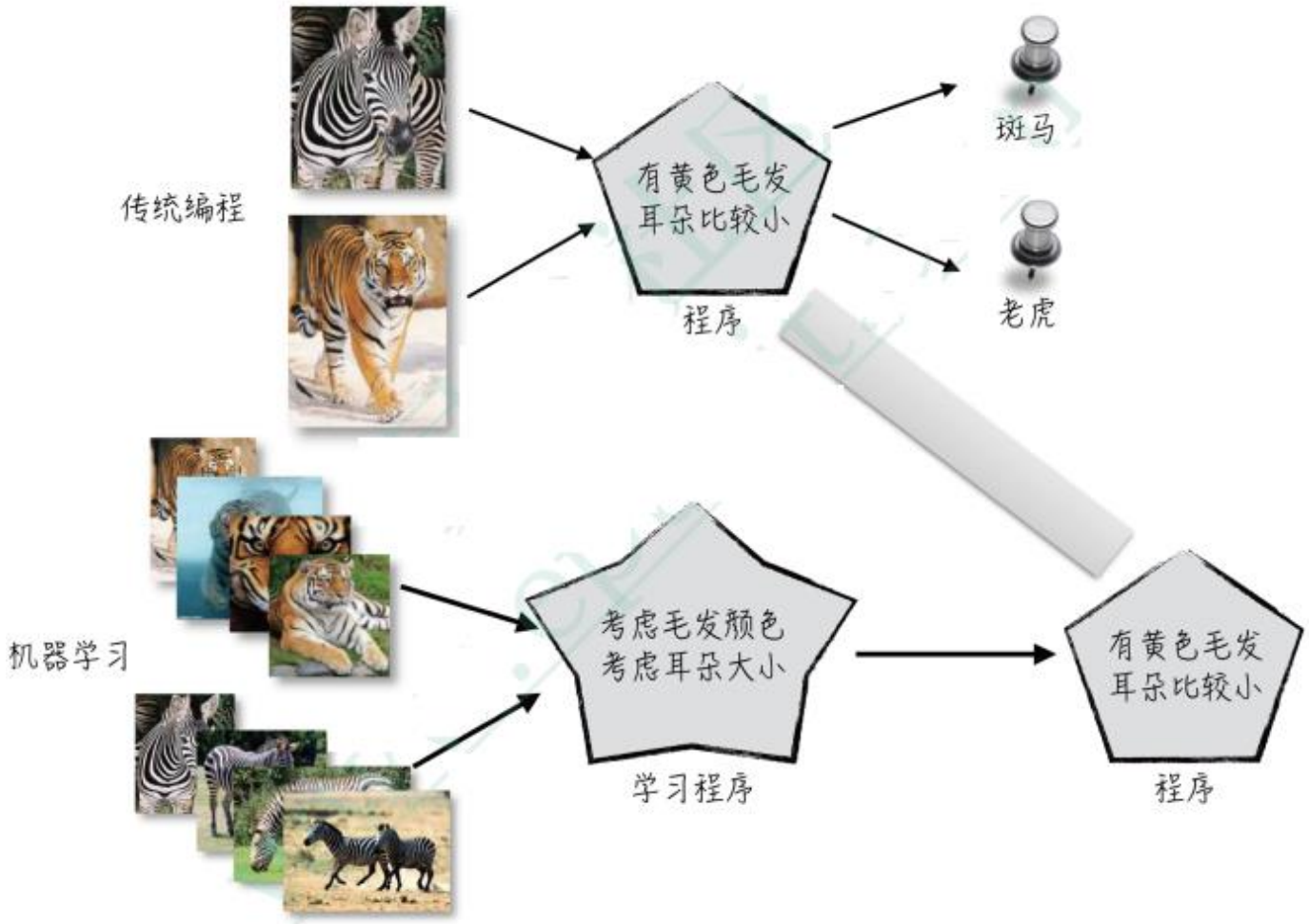
Charles Joseph Minard
(1781 – 1870)

《1812 – 1813 对俄战争中法军人力持续损失示意图》：图中透过两个维度呈现了六种数据：**军队人数**、**距离**、**温度**、**经纬度**、**移动方向**、和 **时-地关系**。

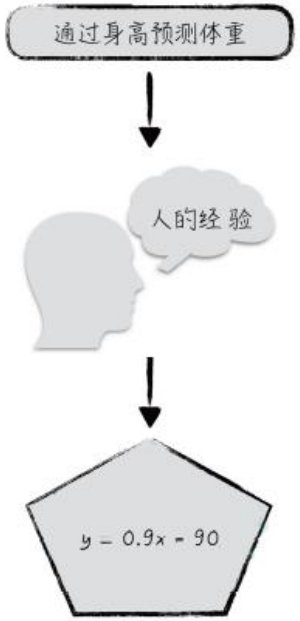


机器学习方法

从**编程的角度**来看，机器学习是一种能自动生成程序的**特殊程序**。

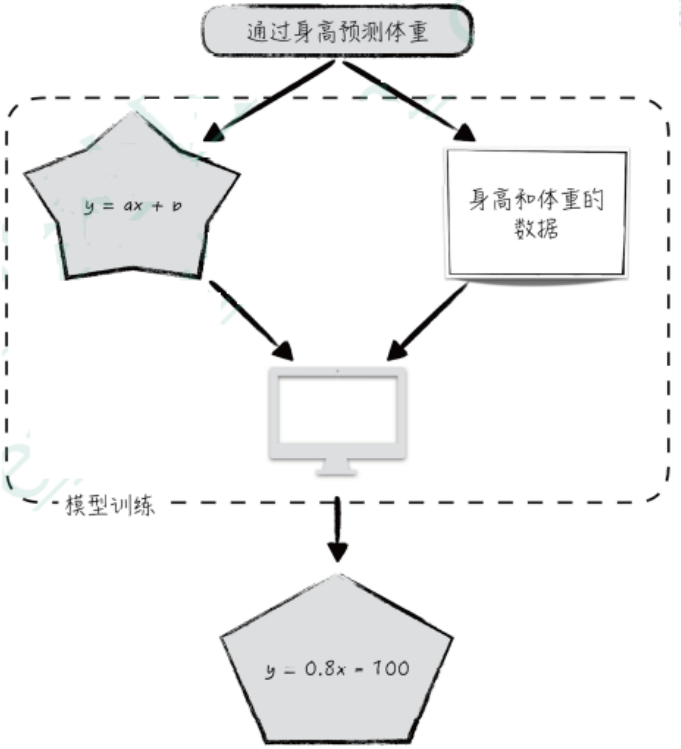


机器学习：数据驱动的问题求解



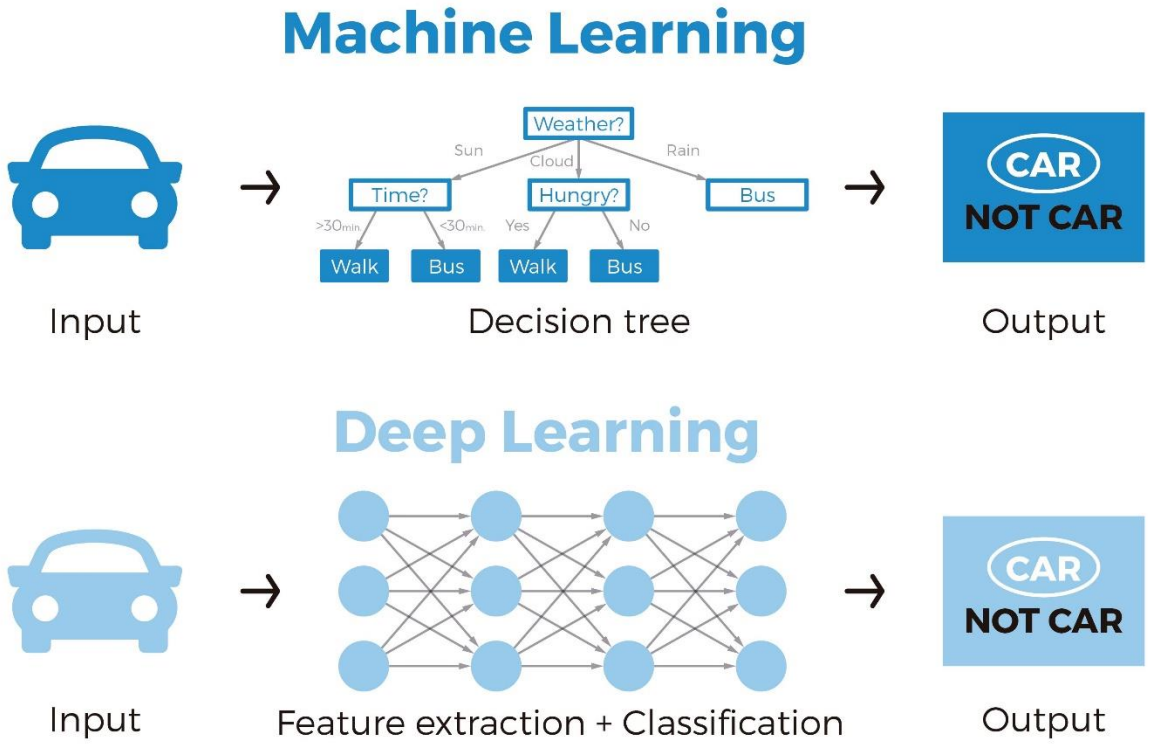
经验驱动的算法模型

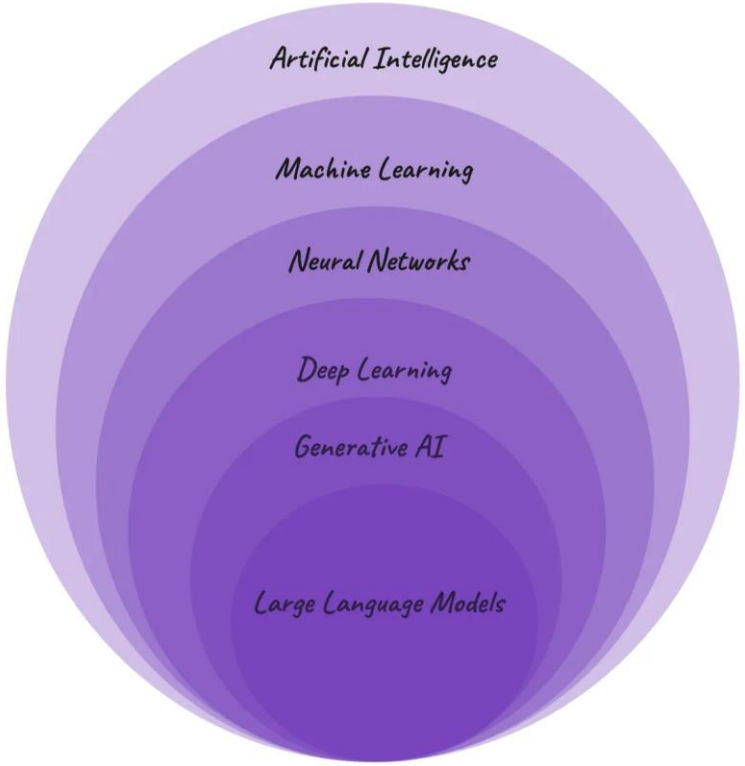
数据驱动的算法模型



经验 → 算法模型 ← 数据

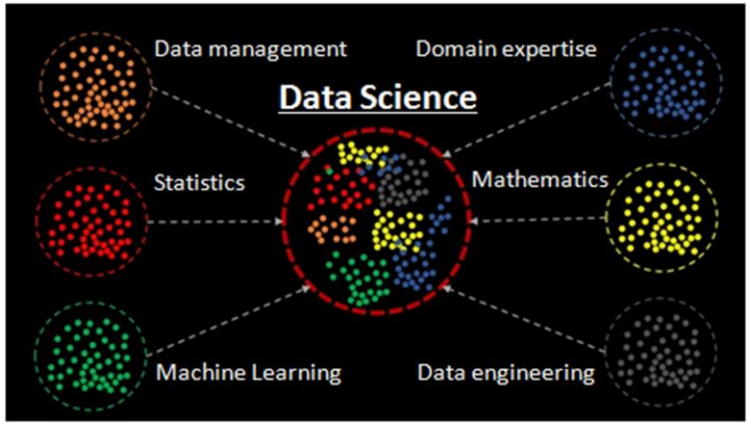
机器学习的最新发展


Machine Learning
Weather?
Sun
Cloud
Rain
Time?
Hungry?
Bus
>30min
<30min
Yes
No
Walk
Bus
Walk
Bus
CAR
NOT CAR
Input
Decision tree
Output
Deep Learning
CAR
NOT CAR
Input
Feature extraction + Classification
Output


Artificial Intelligence
Machine Learning
Neural Networks
Deep Learning
Generative AI
Large Language Models

数据思维与实践

第01讲 数据科学与数据思维入门



- 从数据思维到第四范式
- 数据科学与工程
- **数据科学实践场景：开源数字王国**
- 初识 Python 重要扩展库

开源数字王国的故事

在一个遥远的**开源数字王国（OpenKingdom）**里，住着各种各样的开源爱好者，他们喜欢写代码、并做各种好玩的开源项目。他们日以夜继地在 GitHub 上开展协作，项目也越来越多，逐渐地开始提出各种有趣的问题，这些问题甚至连国王也回答不上来。

有一天，一个名叫**李纳斯（Lee）**的年轻人来到这个王国旅游，他是一个**懂开源的数据科学家**。国王知道了，就把大家提出的各种问题向这位年轻人请教，你现在就是这位年轻的数据科学家，希望来迎接这个挑战。

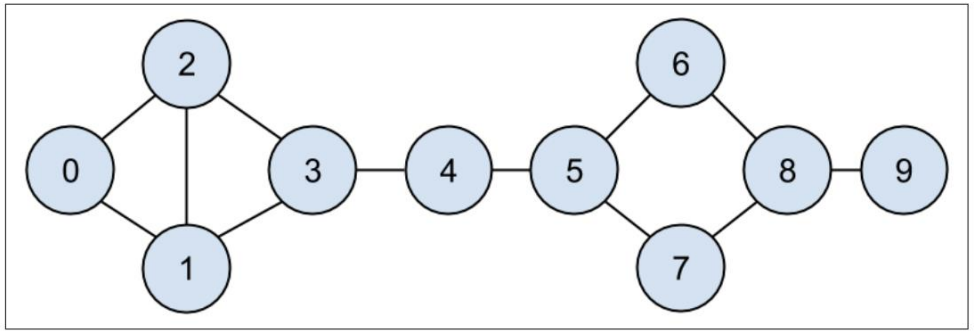
第一周的任务

OpenKingdom 有市民的所有数据，特别是开源爱好者们在 GitHub 上面的**开源活动数据**，这无疑给 Lee 提供了一个巨大的数据宝藏。

1、谁是我们这个开源数字王国中的关键联系人？

```
users = [
    { "id": 0, "name": "Hero" },
    { "id": 1, "name": "Dunn" },
    { "id": 2, "name": "Sue" },
    { "id": 3, "name": "Chi" },
    { "id": 4, "name": "Thor" },
    { "id": 5, "name": "Clive" },
    { "id": 6, "name": "Hicks" },
    { "id": 7, "name": "Devin" },
    { "id": 8, "name": "Kate" },
    { "id": 9, "name": "Klein" }
]
```

```
friendship_pairs = [(0, 1), (0, 2), (1, 2), (1, 3), (2, 3), (3, 4),
                    (4, 5), (5, 6), (5, 7), (6, 8), (7, 8), (8, 9)]
```



GitHub Social Network

OPENKINGDOM

第一周的任务

2、国王希望市民之间建立更多的联系，如何找到“有共同兴趣的开源爱好者？”

提示：朋友的朋友，也是我的朋友~

```
users = [
    { "id": 0, "name": "Hero" },
    { "id": 1, "name": "Dunn" },
    { "id": 2, "name": "Sue" },
    { "id": 3, "name": "Chi" },
    { "id": 4, "name": "Thor" },
    { "id": 5, "name": "Clive" },
    { "id": 6, "name": "Hicks" },
    { "id": 7, "name": "Devin" },
    { "id": 8, "name": "Kate" },
    { "id": 9, "name": "Klein" }
]
```

```
interests = [
    (0, "Hadoop"), (0, "Big Data"), (0, "HBase"), (0, "Java"),
    (0, "Spark"), (0, "Storm"), (0, "Cassandra"),
    (1, "NoSQL"), (1, "MongoDB"), (1, "Cassandra"), (1, "HBase"),
    (1, "Postgres"), (2, "Python"), (2, "scikit-learn"), (2, "scipy"),
    (2, "numpy"), (2, "statsmodels"), (2, "pandas"), (3, "R"), (3, "Python"),
    (3, "statistics"), (3, "regression"), (3, "probability"),
    (4, "machine learning"), (4, "regression"), (4, "decision trees"),
    (4, "libsvm"), (5, "Python"), (5, "R"), (5, "Java"), (5, "C++"),
    (5, "Haskell"), (5, "programming languages"), (6, "statistics"),
    (6, "probability"), (6, "mathematics"), (6, "theory"),
    (7, "machine learning"), (7, "scikit-learn"), (7, "Mahout"),
    (7, "neural networks"), (8, "neural networks"), (8, "deep learning"),
    (8, "Big Data"), (8, "artificial intelligence"), (9, "Hadoop"),
    (9, "Java"), (9, "MapReduce"), (9, "Big Data")
]
```

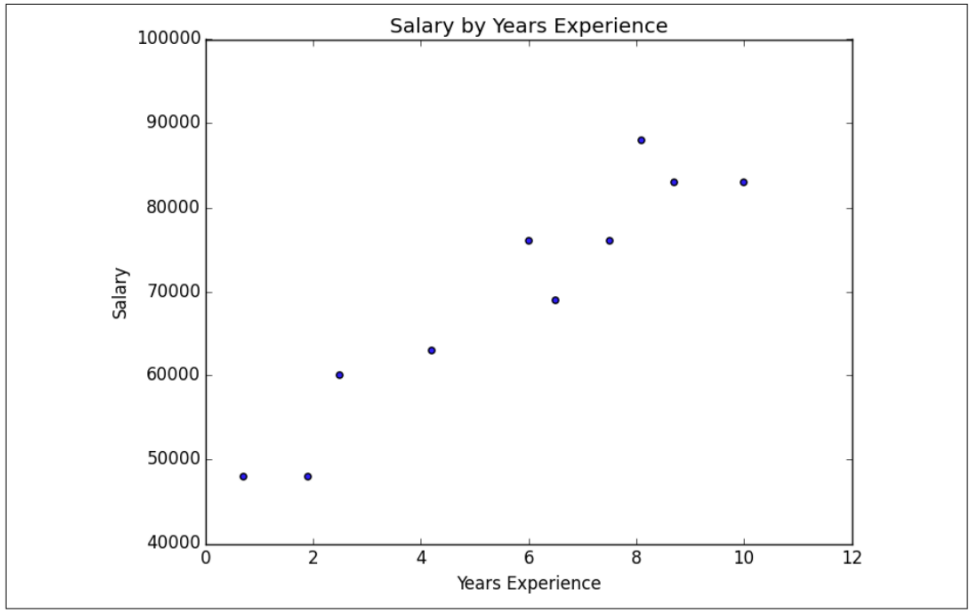
OPENKINGDOM

第一周的任务

3、收入和开源经验之间的关系。

猜想：开源工作经验越丰富，收入越高吗？

```
salaries_and_tenures = [(83000, 8.7), (88000, 8.1),
                        (48000, 0.7), (76000, 6),
                        (69000, 6.5), (76000, 7.5),
                        (60000, 2.5), (83000, 10),
                        (48000, 1.9), (63000, 4.2)]
```



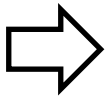
OPENKINGDOM

第一周的任务

4、开源数字王国公民最感兴趣的开源主题是哪些?

```
interests = [
    (0, "Hadoop"), (0, "Big Data"), (0, "HBase"), (0, "Java"),
    (0, "Spark"), (0, "Storm"), (0, "Cassandra"),
    (1, "NoSQL"), (1, "MongoDB"), (1, "Cassandra"), (1, "HBase"),
    (1, "Postgres"), (2, "Python"), (2, "scikit-learn"), (2, "scipy"),
    (2, "numpy"), (2, "statsmodels"), (2, "pandas"), (3, "R"), (3, "Python"),
    (3, "statistics"), (3, "regression"), (3, "probability"),
    (4, "machine learning"), (4, "regression"), (4, "decision trees"),
    (4, "libsvm"), (5, "Python"), (5, "R"), (5, "Java"), (5, "C++"),
    (5, "Haskell"), (5, "programming languages"), (6, "statistics"),
    (6, "probability"), (6, "mathematics"), (6, "theory"),
    (7, "machine learning"), (7, "scikit-learn"), (7, "Mahout"),
    (7, "neural networks"), (8, "neural networks"), (8, "deep learning"),
    (8, "Big Data"), (8, "artificial intelligence"), (9, "Hadoop"),
    (9, "Java"), (9, "MapReduce"), (9, "Big Data")
]
```

计算感兴趣词汇的个数

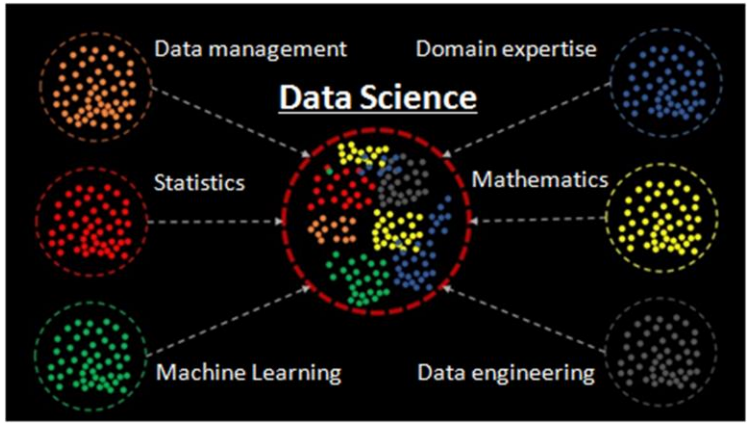


```
learning 3
java 3
python 3
big 3
data 3
hbase 2
regression 2
cassandra 2
statistics 2
probability 2
hadoop 2
networks 2
machine 2
neural 2
scikit-learn 2
r 2
```

OPENKINGDOM

数据思维与实践

第01讲 数据科学与数据思维入门



- 数据思维与第四范式
- 数据科学与工程
- 数据科学实践场景：开源数字王国
- **初识 Python 重要扩展库**

Python 重要扩展库

NumPy (Numerical Python) 是 Python 语言的一个扩展程序库，支持大量的维度数组与矩阵运算，并提供大量的数学函数库

Pandas 基于 Numpy 构建，含有使数据分析工作变得更快更简单的高级数据结构和操作工具，让以 Numpy 为中心的应用变得更加简单

Scipy 库函数类似于Matlab的工具箱，是 Python 科学计算程序的核心包，用于有效地计算 NumPy 矩阵

Matplotlib 是 Python 语言及其数值计算库 NumPy 的绘图库。它提供了一个面向对象的API，可以将绘图嵌入到使用通用GUI工具包的程序中

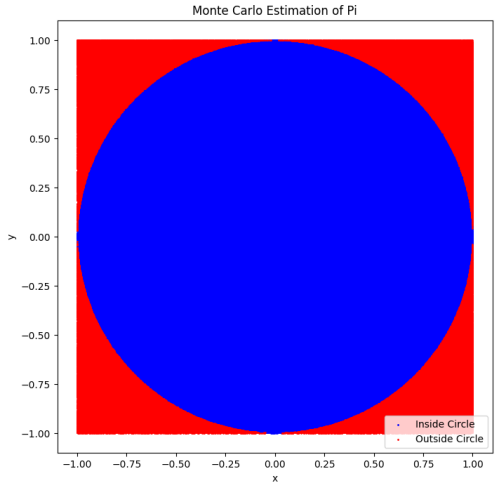
实例：蒙特卡洛法模拟求 Pi 值

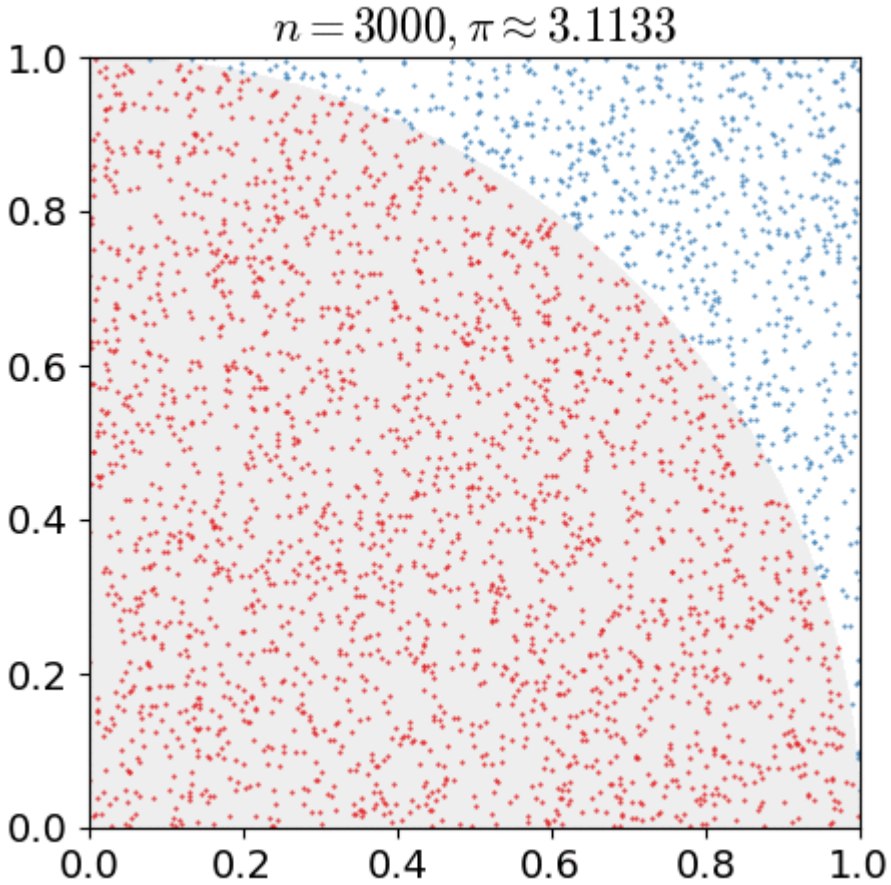
- 蒙特卡罗方法（Monte Carlo method），也称统计模拟方法，是1940年代中期提出的一种以概率统计理论为指导的数值计算方法，主要使用随机数来解决很多计算问题的方法。

$$\pi = 3 + \cfrac{1}{7+\cfrac{1}{15+\cfrac{1}{1+\cfrac{1}{292+\cfrac{1}{1+\cfrac{1}{1+\cfrac{1}{1+\ddots}}}}}}}$$

$$\pi = \sum_{k=0}^{\infty} \frac{4(-1)^k}{2k+1}$$

$$\frac{\pi}{4} = 4\arctan\frac{1}{5} - \arctan\frac{1}{239}$$






THANK
YOU

课程的目标

- **目标1：**了解数据科学全貌，建立数据思维的意识；
- **目标2：**建立数据分析与处理的思维，掌握数据科学的核心原理与关键技术；
- **目标3：**培养以数据为中心的问题求解能力，并做到初步的数据编程训练；
- **目标4：**让大家感受到数据科学的美与愉悦。

https://github.com/ECNU/Data-Thinking-and-Practice

课程安排

单元	内容	知识点	实验（Lab）
01	**数据科学与数据思维入门**	数据思维、第四范式、数据科学与工程、应用案例、Python扩展库	Python 分析案例与分析工具
02	**数学基础与实例**	线性代数、概率与统计学、假设和推断	Python 数学实例
03	**数据收集与管理**	大数据时代、数据类型、数据采集方法	Python 数据收集与管理
04	**数据探索与预处理**	数据清洗、数据集成、数据规约、数据变换	Python 数据探索与预处理
05	**数据建模与分析**	机器学习初步、预测、聚类、最邻近模型、神经网络、k均值	Python 数据建模与分析
06	**数据科学实践案例**	自然语言处理、网络分析、Titanic生存预测、客户价值分析	Python 数据科学实践案例
07	**数据科学前沿（番外）**	大数据与云计算、图形图像处理、深度学习、大语言模型与AIGC	-

课程成绩

- 平时出勤：10%
- 期中大作业：40%
- 期末大作业：50%

数据素养与数据思维矩阵

	阅读数据（RD）	问题解决（WD）	分析数据（AD）	数据沟通（CD）
描述性分析				
诊断性分析				
预测性分析				
指导性分析				