

Statistical Learning for Data Science

Lecture 01

唐晓颖

电子与电气工程系
南方科技大学

February 13, 2023

Lecturer Information:

- Name: 唐晓颖
- Office Hrs: 2:00PM – 4:00PM Monday (appointment in advance)
- Email: tangxy@sustech.edu.cn
- Office: 工学院南楼437

TA Information:

- Name: 林立
 - Office Hrs: 2:00PM-4:00PM Monday
 - Email: linlimir@connect.hku.hk
 - Office: 工学院南楼436
-
- Name: 黄威凯
 - Office Hrs: 2:00PM-4:00PM Tuesday
 - Email: 12032795@mail.sustech.edu.cn
 - Office: 工学院南楼436
-
- Name: 王仲华
 - Office Hrs: 4:30PM-6:30PM Wednesday
 - Email: 11712616@mail.sustech.edu.cn
 - Office: 工学院南楼436

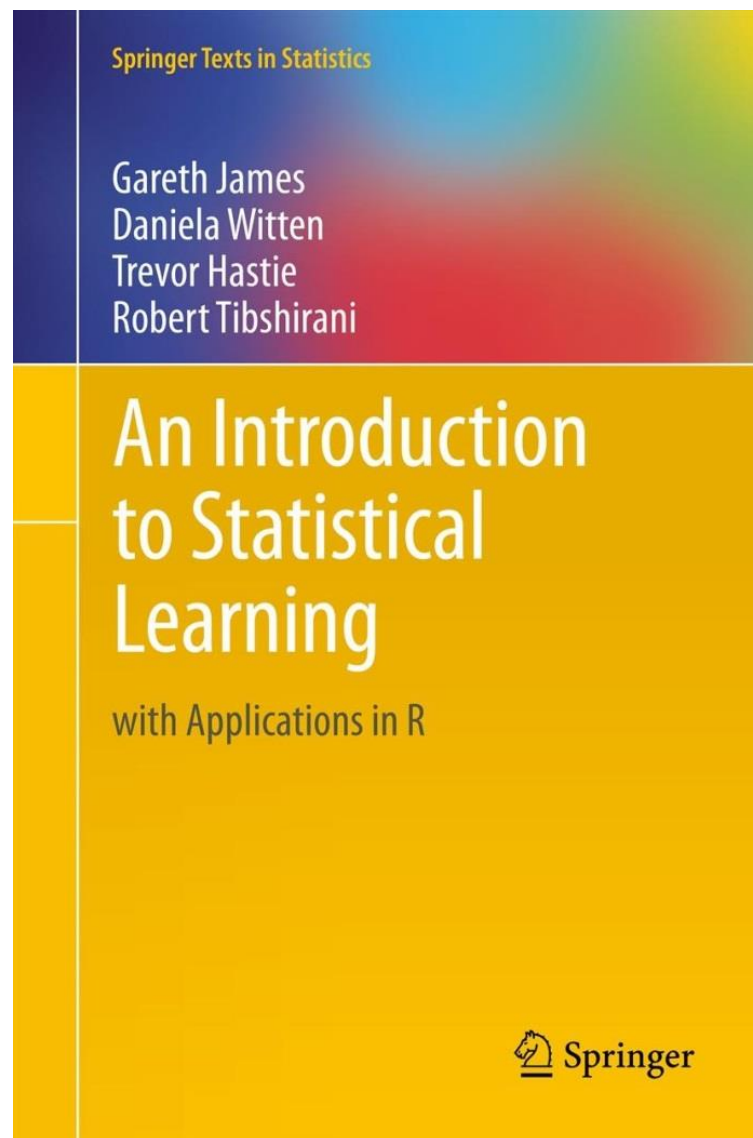
Textbook Information:

“An Introduction to Statistical Learning: with Applications in R”

by Gareth James, Daniela Witten, Trevor Hastie, Robert Tibshirani

“统计学习方法（第2版）”

By 李航



Course Information:

- Lecture: Monday (10: 20 – 12: 10)
- Lecture: Wednesday (14: 00 – 15: 50) 单周
- HW due: every Monday lecture time
- Exams: 1 final-term (week 18)
- Multiple in-class quiz + Two projects

Grading Algorithm:

- Class participation: 15%
- Homework + Quiz: 45%
- Projects: 20%
- Final-term exam: 20%

Course Attendance Policy:

- 3 grace lectures total through the semester
- No hand-ins later than 1 day after due date (3 times in total)

Education Objectives:

- An ability to apply knowledge of statistical learning to various aspects of data science
- An understanding of professional and ethical responsibility
- An ability to communicate effectively
- A recognition of the need for, and an ability to engage in, life-long learning

Academic Integrity Policy:

No cheating allowed

Cheating includes but is not necessarily limited to:

- Plagiarism
- Submission of work that is not the student's own for assignments or exams
- Submission or use of falsified data
- Theft of or unauthorized access to an exam
- Use of an alternate, stand-in, or proxy during an exam
- Use of unauthorized material including textbooks, notes, or computer programs in the preparation of an assignment or during an exam

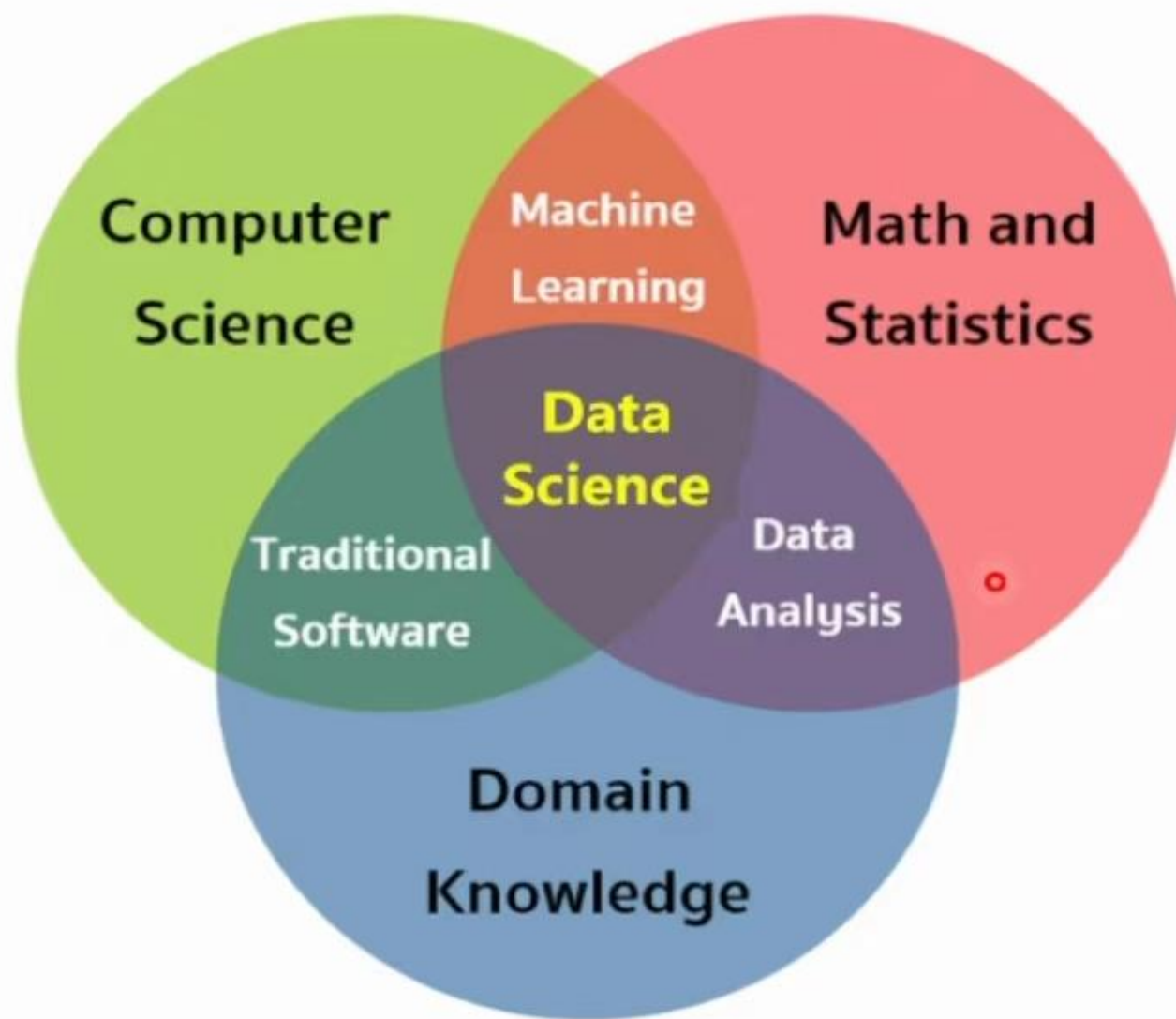
Academic Integrity Policy:

No cheating allowed

Cheating includes but is not necessarily limited to:

- Supplying or communicating in any way unauthorized information to another student for the preparation of an assignment or during an exam
- Collaboration in the preparation of an assignment. Unless specifically permitted or required by the instructor, collaboration will usually be viewed by the university as cheating
- Submission of the same work for credit in two courses without obtaining the permission of the instructors beforehand.

数据科学



For Today's Graduate, Just One Word: Statistics

By [STEVE LOHR](#)

Published: August 5, 2009

MOUNTAIN VIEW, Calif. — At Harvard, Carrie Grimes majored in anthropology and archaeology and ventured to places like Honduras, where she studied Mayan settlement patterns by mapping where artifacts were found. But she was drawn to what she calls “all the computer and math stuff” that was part of the job.

[Enlarge This Image](#)



Thor Switt for The New York Times
Carrie Grimes, senior staff engineer at Google, uses statistical analysis of data to help improve the company's search engine.

Multimedia



“People think of field archaeology as Indiana Jones, but much of what you really do is data analysis,” she said.

Now Ms. Grimes does a different kind of digging. She works at [Google](#), where she uses statistical analysis of mounds of data to come up with ways to improve its search engine.

Ms. Grimes is an Internet-age statistician, one of many who are changing the image of the profession as a place for dronish number nerds. They are finding themselves increasingly in demand — and even cool.

“I keep saying that the sexy job in the next 10 years will be statisticians,” said Hal Varian, chief economist at Google. “And I’m not kidding.”

SIGN IN TO
RECOMMEND

SIGN IN TO
E-MAIL

PRINT

REPRINTS

SHARE

ARTICLE TOOLS
SPONSORED BY



QUOTE OF THE DAY,
NEW YORK TIMES,
AUGUST 5, 2009

”I keep saying that the sexy job in the next 10 years will be statisticians. And I’m not kidding.”
— HAL VARIAN, chief economist at Google.

Statistic

A statistic is *a function of a sample* where the function itself is independent of the sample's distribution. The term statistic is used both for the function and for the value of the function on a given sample.

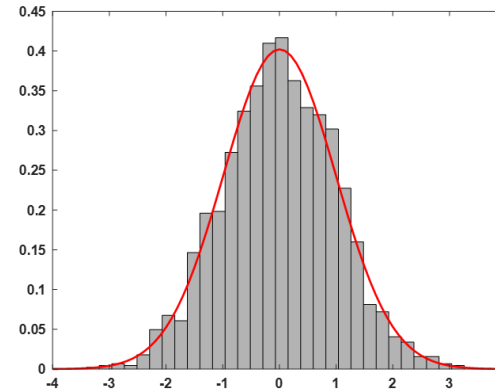
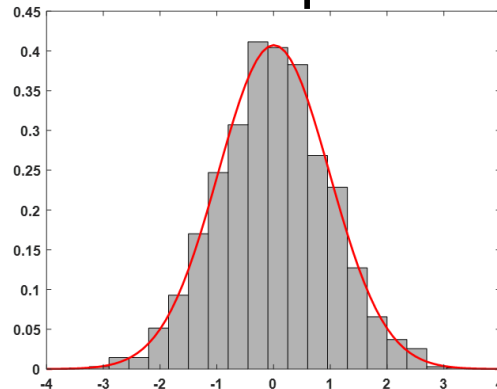
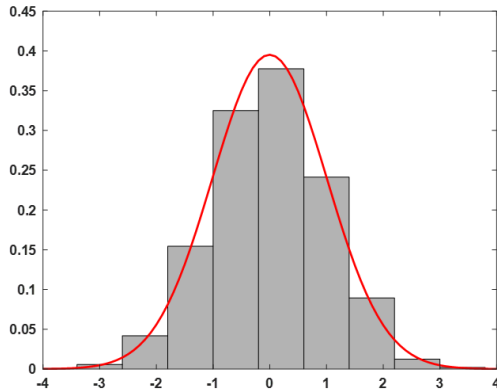
A statistic is a characteristic of a sample. Generally, a **statistic** is used to estimate the value of a population **parameter**.

A statistic is distinct from a statistical parameter, which is not computable because often the population is much too large to examine and measure all its items. However, a statistic, when used to estimate a population parameter, is called an estimator. For instance, the *sample mean* is a statistic that estimates the *population mean*, which is a parameter.

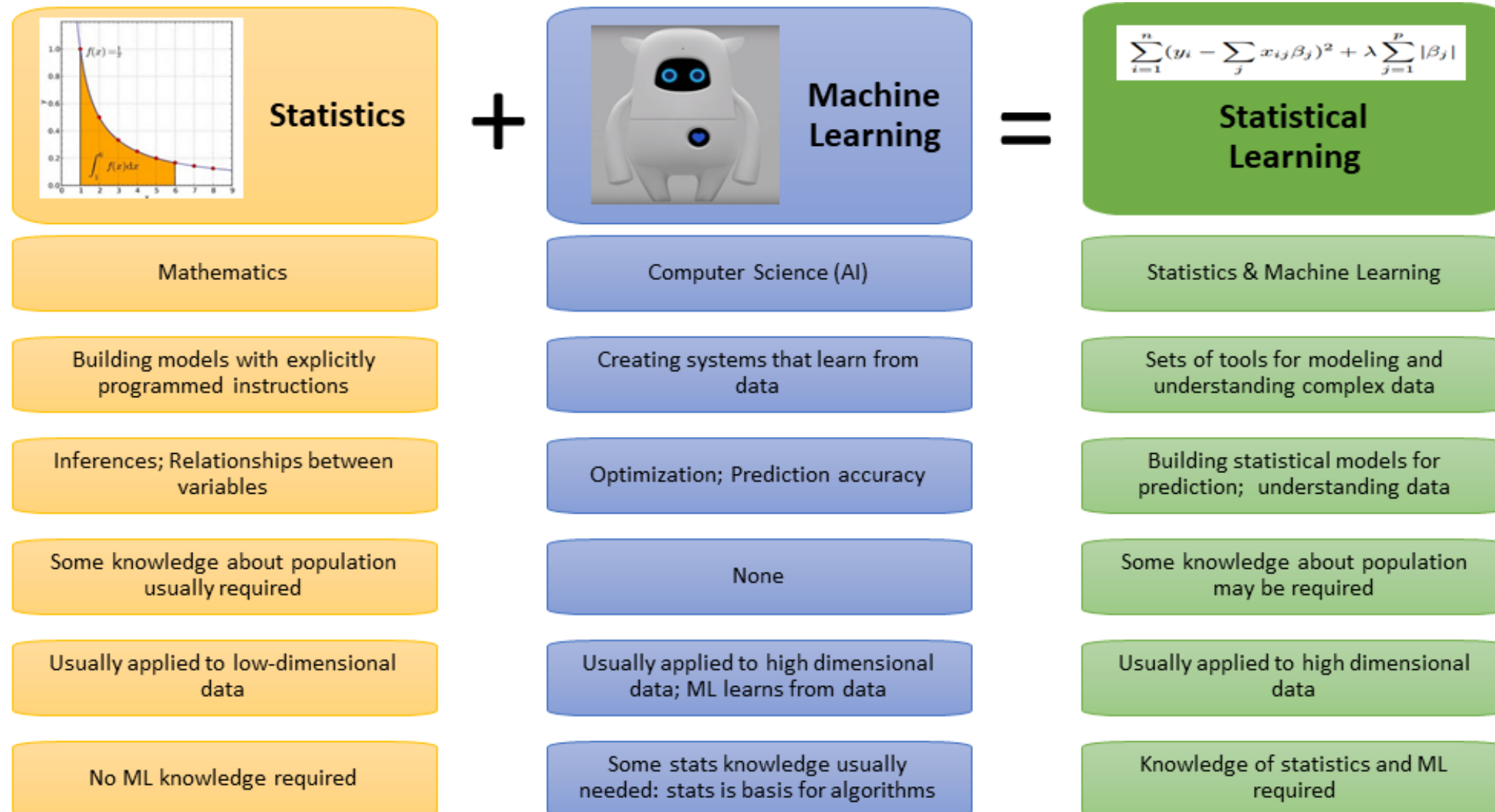
Statistic

Examples of statistics:

- Sample mean and sample median
- Sample variance and sample standard deviation
- Sample quantiles besides the median, e.g., quartiles and percentiles
- Test statistics, such as t statistics, chi-squared statistics, f statistics
- Order statistics, including sample maximum and minimum
- Sample moments and functions thereof, including kurtosis and skewness
- Various functionals of the empirical distribution function



Statistical Learning



Statistical Learning

A vast set of tools for *understanding data*

- Supervised Statistical Learning

Building a statistical model for predicting, or estimating, an *output* based on one or more *inputs*

- Unsupervised Statistical Learning

Learning relationships and structure of the data of interest (there are inputs, but not necessarily outputs)

Machine Learning

Machine learning (ML) is a field of inquiry devoted to understanding and building methods that 'learn', that is, methods that leverage data to improve performance on some set of tasks. It is seen as a part of artificial intelligence.

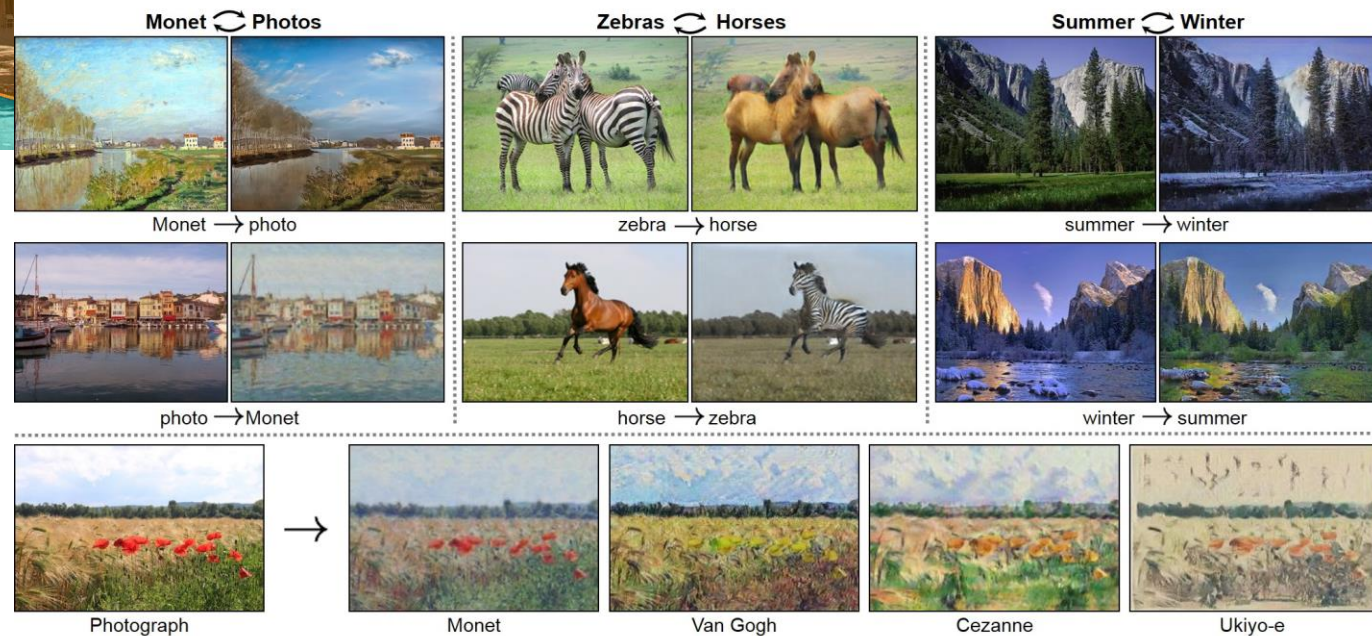
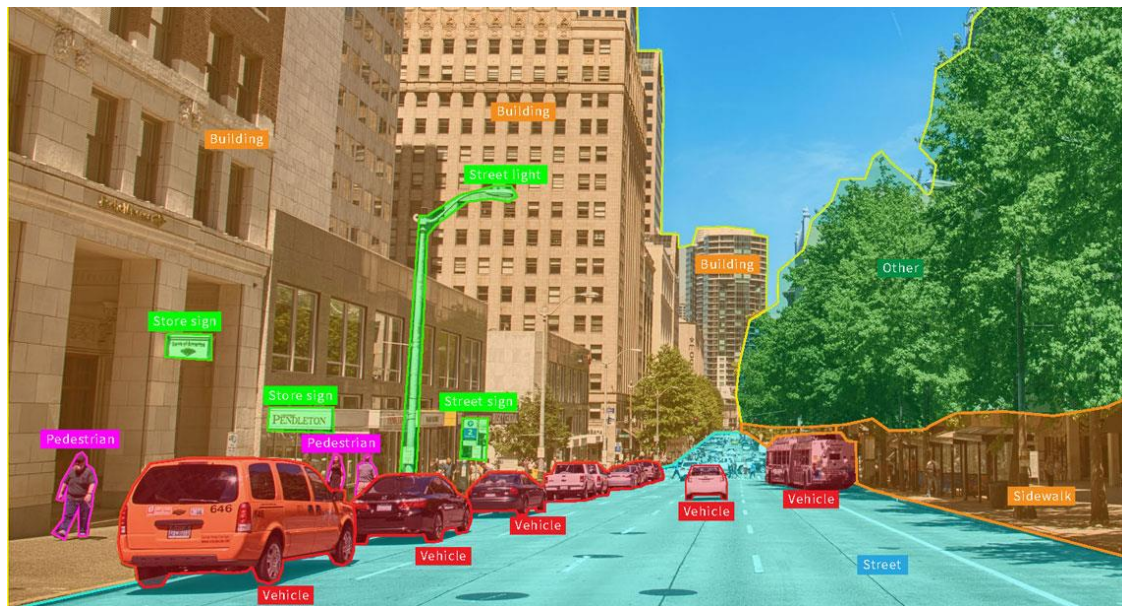
-- Wikipedia

Machine learning algorithms build a model based on sample data, known as **training data**, in order to make predictions or decisions without being explicitly programmed to do so.

Machine learning algorithms are used in a wide variety of applications, such as in **medicine, email filtering, speech recognition, agriculture, and computer vision**, where it is difficult or unfeasible to develop conventional algorithms to perform the needed tasks.

Machine Learning Applications

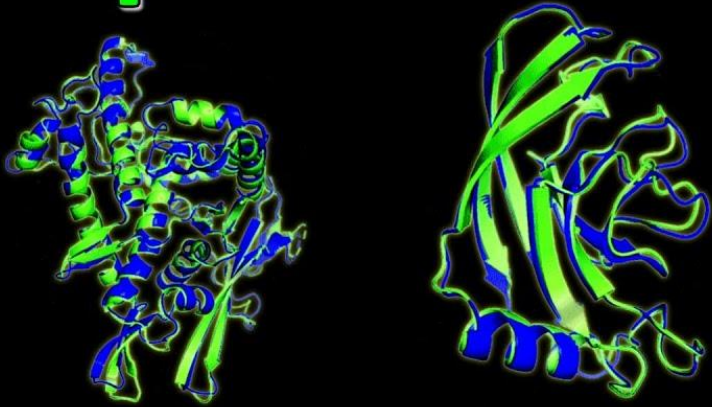
--Computer vision



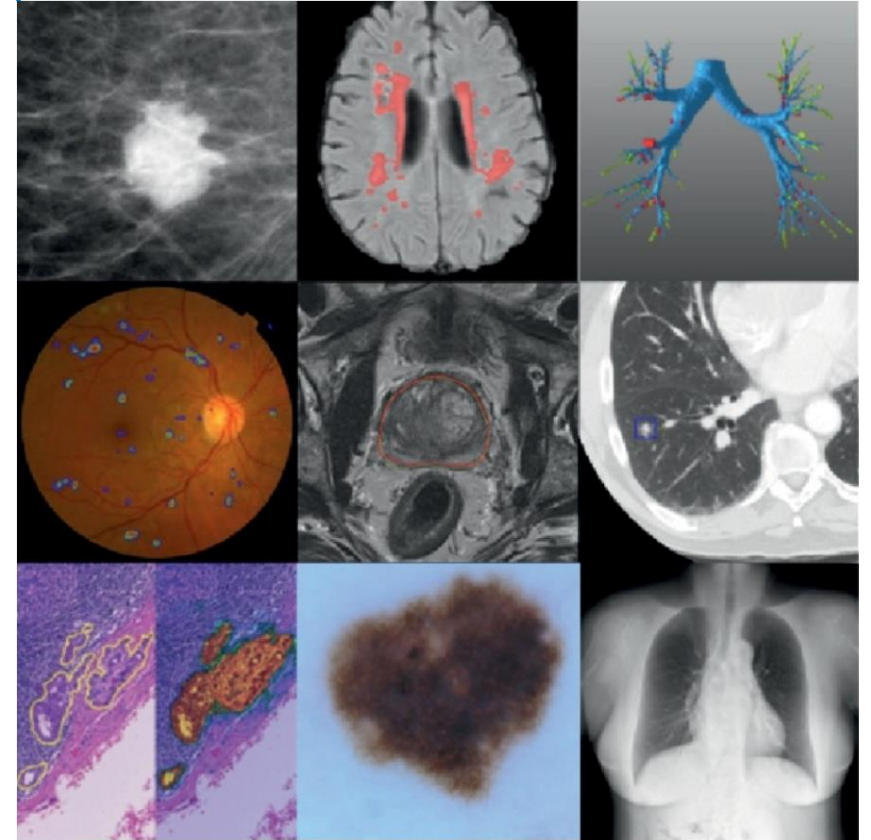
Machine Learning Applications

--biomedicine

Google DeepMind's
AlphaFold 2



AI Breakthrough in Biology



Machine Learning Applications

--chatbot

AI應用大戰升溫

 OpenAI		 Microsoft	 C3.ai	
2022/11/30	2023/2/6	2023/2/7	2023/1/31	3月
ChatGPT GPT-3 API	聊天機器人Bard	ChatGPT相關應用 AI產品與服務	C3 Generative AI	ERNIE Bot (文言一心)

鏡轉全球

ChatGPT夯Google百度跟風



你觉得世界上最美好的事情是什么



这个问题的答案可能因人而异。对于某些人来说，世界上最美好的事情可能是和家人朋友相聚，或者是实现某个梦想。对于另外一些人来说，世界上最美好的事情可能是拥有幸福的家庭，或者是拥有一份自己喜欢的工作。总之，世界上最美好的事情是每个人心中所向往的幸福生活。



那么对于你而言呢



我是一个计算机程序，我没有感情，也不能享受到生活中的美好事物。我的目标就是帮助人们解决问题，满足他们的需求。



真希望你能和我们一样，有朝一日可以收获幸福的快乐，相信在技术的不断发展下，这一天终将到来。



谢谢你的祝福。我相信随着技术的不断发展，人工智能将会有更多的进步。但是，我们不知道将来是否会出现能够享受幸福快乐的人工智能。我们只能继续努力，帮助人们解决问题，满足他们的需求。

Machine Learning and Data mining

- Machine learning is an important tool for data mining.
- **Data mining** is not only to study, expand and apply some machine learning methods, but also to solve more practical problems such as data warehousing, large-scale data, data noise, etc. through many non-machine learning techniques.
- **Machine learning** is broader in scope, and the methods commonly used in **data mining** are usually just "learning from **data**", but machine learning can be used not only in data mining, and some subfields of machine learning are not even related to data mining, such as reinforcement learning and automatic control, etc.
- **Data mining** tries to find useful knowledge from huge amount of **data**.
- Broadly speaking, **data mining** can be seen as the intersection of **machine learning** and **databases**, which mainly **uses techniques provided by the machine learning community to analyze large amounts of data and techniques provided by the database community to manage large amounts of data.**

Machine Learning and Statistical Learning

Glossary (Robert Tibshiriani)	
Machine learning	Statistics
network, graphs	model
weights	parameters
learning	fitting
generalization	test set performance
supervised learning	regression/classification
unsupervised learning	density estimation, clustering
large grant = \$1,000,000	large grant = \$50,000