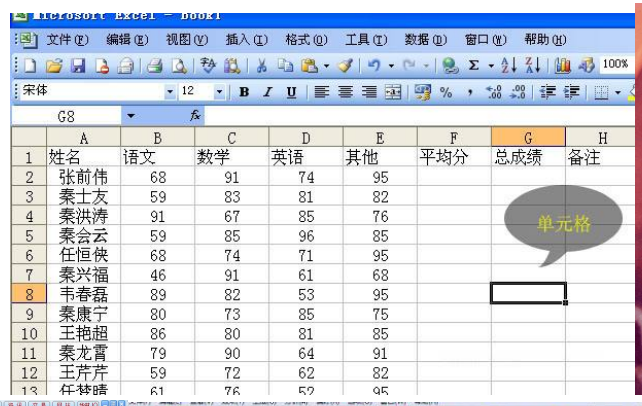
Outlines

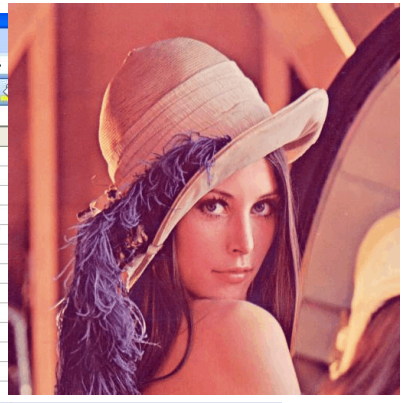
Some Examples of Data

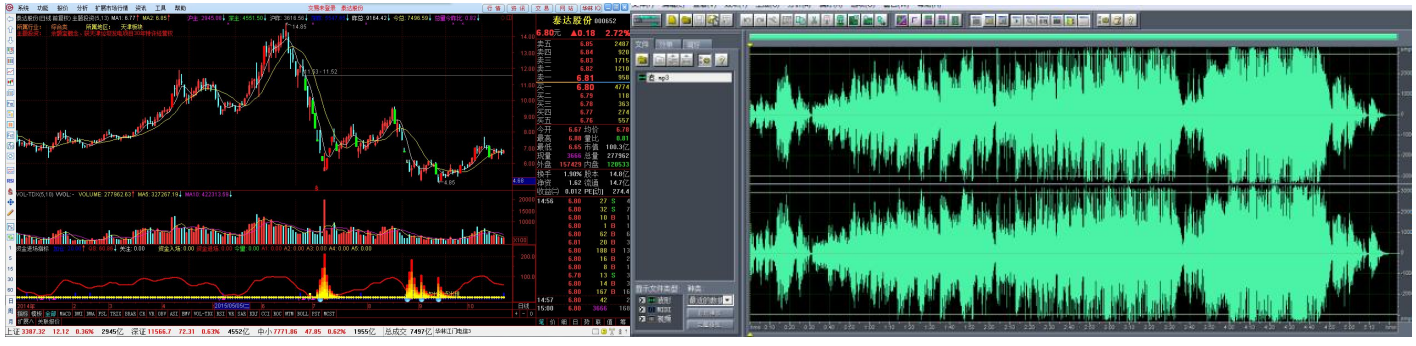
Can you give some examples of data ?

实验数据分析表				
项目	加药量(ppm)			COD_{Cr} (mg/L)
	硫酸亚铁	双氧水	复合碱	
原水	/	/	/	321
1	2000	1000	2800	186
2	3000	1500	3000	157
3	4000	2000	3100	110
4	5000	2500	3500	78

实验数据仅对该批水样负责



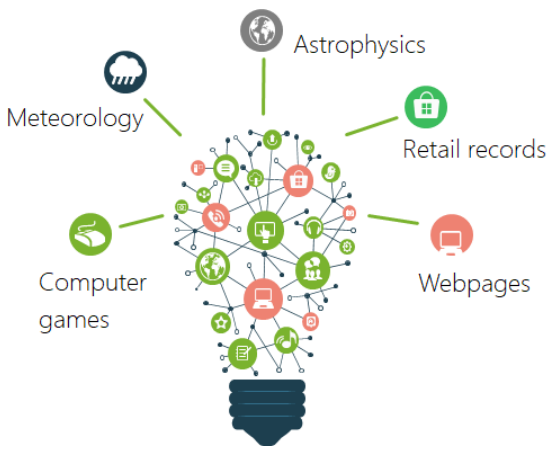




Table, 1D signal (audio, stock price), 2D signal (image), 3D signal (video), etc.

Big Data : 5 Big "V"

- Volume : KB, MB, GB (10^9 bytes), TB, PB, EB (10^{18} bytes), ZB, YB, exponential growth (about 120%/year)
- Variety : different sources from business to industry, different types
- Veracity : Noisy data with errors and inconsistency, redundant information contained in the data, need to retrieve useful information
- Velocity : fast speed for data generation and information transfer, need for realtime processing
- Value : business values for product recommendations and trading ; social values for precision medicine, public health, traffic control, etc.



What is data science

- Retrieve information from data with the help of computational power
- Transfer the information into knowledge
- Two perspectives of data sciences :
  - Study science with the help of data : bioinformatics, astrophysics, geosciences, etc.
  - Use scientific methods to exploit data : statistics, machine learning, data mining, pattern recognition, data base, etc.

Study Science with the Help of Data

A pioneering work of data science : Kepler's Laws

开普勒：分析数据产生价值

行星	周期（年）	平均距离	周期^2/距离^3
水星	0.241	0.39	0.98
金星	0.615	0.72	1.01
地球	1.00	1.00	1.00
火星	1.88	1.52	1.01
木星	11.8	5.20	0.99
土星	29.5	9.54	1.00
天王星	84.0	19.18	1.00
海王星	165	30.06	1.00

Scientific Study of Data

- Grabbing data : business and industrial problem, professional areas
- Storing data : engineering problem, computer science, electronic engineering
- Analyzing data (**key problem**) : scientific problem, mathematics, statistics, computer science

Data Analysis

- Ordinary data types :
  - Table : classical data (could be treated as matrix)
  - Set of points : mathematical description
  - Time series : text, audio, stock prices, DNA sequences, etc.
  - Image : 2D signal (or matrix equivalently, e.g., pixels), MRI, CT, supersonic imaging
  - video : 2D in space and 1D in time (another kind of time series)
  - Webpage and newspaper : time series with spatial structure
  - Network : relational data, graph (nodes and edges)
- Basic assumption : the data are generated from an underlying model, which is unknown in practice
  - Set of points : probability distribution
  - Time series : stochastic processes, e.g., Hidden Markov Model (HMM)
  - Image : random fields, e.g., Gibbs random fields
  - Network : graphical models, Beyesian models

Difficulties

- Huge volume of data
- Extremely high dimensions
  - Curse of dimensionality : the model complexity and computational complexity become exponentially increasing with the growth of dimension
  - Solutions :
    - Make use of prior information
    - Restrict to simple models
    - Make use of special structures, e.g., sparsity, low rank, smoothness
    - Dimensionality reduction, e.g., PCA, LDA, etc.
- Complex variety of data
- Large noise : data are always contaminated with noises

Solution - Algorithms

- Algorithms are in the interdisciplinary part of computer science and mathematics : establish mathematical models, solve it numerically, implement it in the computer languages
- Reduce the algorithmic complexity, with the help of techniques from mathematics or computer science
- Distributional and parallel computing : divide-and-conquer, e.g., MapReduce, GPU
- IEEE 2006 top 10 algorithms in data mining : C4.5, K-Means, SVM, Apriori, EM, PageRank, NaiveBayes, K-Nearest Neighbors, AdaBoost, CART