

基于Python的 数据分析与可视化

导论

本讲提纲

1. 为什么会有这门课？

课程目标、内容与考核方式

3. 为什么用Python？

人生苦短，Python是岸

2. 什么是数据分析与可视化？

从概率统计到数据科学

4. 我该怎么上手？

环境配置和语法元素

为什么会有这门课？

课程目标

1

能力素养目标：培养学生关于数据分析与可视化的知识素养，能够针对具体问题确定合理的数据分析策略和可视化表达方式

2

应用技能目标：使学生掌握一门编程语言，理解不同场景和具体问题，能够灵活地完成数据分析任务，了解几类高级图表的设计与实现

3

职业发展目标：积累数据分析与可视化经验，帮助学生打造自身核心竞争力，为今后的职场生涯提供帮助

课程考核

课堂参与

$$20\text{分} = 2\text{分} \times 10\text{次}$$

- 最后一周为期末考试，前十周每周参与课堂教学的同学可获得2分，总计20分，允许请假一次。

课后作业

$$30\text{分} = 10\text{分} \times 3\text{次}$$

- 每一章开始布置作业，每一章结束提交作业。每次作业10分，总计30分。

结课考试

$$50\text{分}$$

- 上机考试

什么是数据分析与可视化？

从概率统计到数据科学

分类数据分析的历史回顾



Lectio Magistralis
prof. Alan Agresti

Some historical highlights
in the development
of categorical data methodology

*Rome, 22 October 2015
Istat, Aula Magna*

<https://www.bilibili.com/video/av56090638?from=search&seid=11677121380145512780>

统计学简史

- 分布统计最初产生于研究对国家，特别是对其经济以及人口的描述。现代统计主要起源于研究**总体**（population），**方差**（variation）和**简化数据**（reduction of data）。
- John Graunt (1620 - 1674) 注意到在非瘟疫时期，一个大城市每年死亡数有统计规律，而且出生儿的性别比为1.08，即每生13个女孩就有14个男孩。大城市的死亡率比农村地区要高。在考虑了已知原因的死亡及不知死亡年龄的情况下，Graunt估计出了六岁之前儿童的死亡率，并相当合理地估计出了母亲的死亡率为1.5%。因此，他从杂乱无章的材料中得出了重要的结论——**生命表**。
 - Edmund Harley (哈雷) (1656 - 1742) 利用了Breslau的记有死亡年龄的数据，改进了Graunt的生命表并引进了死亡率的定义。瑞士数学家 Leonhard Euler (欧拉) (1707 - 1783) 提出了平稳生命表的概念。John De Witt(625 - 1672)等人最早讨论退休金和人寿保险的方案。Thomas Robert Malthus (马尔萨斯) (1766 - 1834)，Alfred James Lotke (1880 - 1949)，Ronald Aylmer Fisher (费歇) (1890 - 1962)，及 William Feller (费勒) (1906 - 1970) 等人用渐趋复杂的数学来研究生命表的理论，这对人类及其它总体的动方学描述具有显著意义。
- William Petty (1623 - 1687) 是Graunt同时代的经济学家及朋友。他认为需要建立中央统计部来利用人口统计学的知识；由行政区利用列出记录年龄，性别，婚姻状况等细节的记录表格来收集数据；要有出生，死亡，婚姻，收入，教育和商业等方面的数据。
- 当时在研究诸如死亡等时间序列时，Graunt注意到了随机的起伏；但他仅以机械的术语加以描述一把这些与钟表运动的忽动忽停相联系。实际上，这种不规则的变化也影响赌博和天文学。因此，其后进一步导致了**随机误差的误差分布**概念的出现。

统计学简史

- 赌博是引起注意的第一个概率事件模型：硬币与骰子。
- Abraham de Moivre (棣美佛) (1667-1754) 导出了对二项分布的一个近似。
- Pierre simon Laplace (拉普拉斯) (1749 - 1827) 导出了对男子出生比例的类似的渐近公式。
- Jacob Bernoulli (伯努利) (1654 - 1705) 以弱大数定律支持了对大样本均值的使用。
- Tomas simpson (辛普森) (1710 - 1761) 计算了同分布随机变量和的精确分布，同样也支持了对大样本均值的使用。
- 在天文学中，要对一些运动星体位置的未知参数进行估计，通常某种意义上“最好的”估计都是来源于一些注定不和谐的观察值，因为只要观察值在数量上超过参数，就会产生度量误差。Roger Cotes (1682 - 1716) , Thomas Bayes (贝叶斯) (1702 - 1761) , Euler , Johann Tobias Mayer (1723 - 1762) , Rudger Josif Boskovic (1711 - 1787) , Laplace , 和 Adrien Marie Legendre (勒让德) (1752 - 1833) 都在研究这个问题。后来被 Friedrich Gauss (高斯) (1777 - 1855) 解决。John Michell (米歇尔) (1724 - 1793) 用统计方法证明了双星的存在。
- 然而，认定现代统计理论是由精算科学，人口学和天文学的需要而发展来的观点是不正确的；事实上，它是由心理学，医学，人体测量学，遗传学和农业的需要发展出来的。

统计学简史

- 直到1830年，几乎所有的经验分布都是关于一维误差或一个非数值变量。在1830年之后，天文学家和社会学家 Adolphe Jacques Quetelet(1796 - 1874)使得诸如身高体重之类的度量值的变量的经验分布通俗化。他在生物统计研究中大量利用了理论**二项分布和正态分布**。
- 后来 Ladislaus von Bortkiewicz (1868 - 1931) 报告了在普鲁士兵团中由马踢造成的受伤事故，发现和**Poisson (泊松) 分布**有关。在计算血红细胞数目上，Poisson分布也被 Ernst Abbe (1840 - 1905) 所用。从那时起，该分布被大量地用于计数的试验中。
- 在生物学上，统计方法使得Johann Gregor Mendel (孟德尔) (1822 - 1884) 认识到某些主要遗传基因的存在，它们在0，1和2三个水平显现，其中水平0 (双隐性) 能和水平1和2区别开来。他能确定有相同或不相同的水平的个体之间交配的结果，而且提出了某些生物学事件等价于掷一个硬币的模型；他能**对任意交配的结果给出概率并用实验来验证其假设**。
- 在较早的医学统计中，Philippe Pinel (1745-1826) 和 Pierre Charles alexandre Louis (1787 - 1872) 开始了建立**疾病分类**的困难课题；这些工作人员**保存了精确和完整的所有病例的记录，并且能给出和预后有关的统计数字**。Louis能有**利用跟踪调查的方法反驳以当时广泛滥用的放血疗法**。他的三个学生是值得一提的：Jules Gavarret (1808 - 1890) 写了一本医学统计的教科书；书中有应用Simeon Dents Poisson (1780 - 1840) 理论来对两个比例进行检验的许多应用；Oliver Wendell Holmes (1809 - 1894) 和他的不知名的数学顾问对一系列分娩热病例给出了有趣的分析，证明该病是传染的，这优于任何十九世纪的类似研究。

统计学简史

- 更直接的原动力来自于遗传学（确切地说是优生学）。
- Francis Galton (1822—1911) 在1886年研究了两代豌豆重量之间的相关时发现了Y关于一个正态变量X的线性回归及类似于椭圆的等概率线。从此，**多元正态分布**就经常出现在文献之中；而两个和三个变量的正态分布在Laplace时就已经知道了。该联合分布能够由互相独立的正态随机变量的线性变换而构造，例如Giovanni Antonio amedeo Plana (1781 - 1863) 和Irenee - Jules Bravais (1811 - 1863) 和Irenee - Jules Bravais (1811 - 1863) 所做，而且，反过来它能分解为互相独立的正整随机变量的积，如 Auguste Bravais (1820 - 1884) 在最小二乘理论上导出了一般形式的多元正态分布；Arthur Cayley (1821 - 1895) 化简为平方和并确定了该常数值。
- 正态分布在理论统计中扮演了一个非常重要角色。有许多理由来说明这一点；一般来说，如果一个模型包含着正态分布的几个非平凡特性，则它必须具备所有的特性。
- 在 1895年，Karl Pearson (皮尔森) (1857 - 1936) 认识到更理论的统计分布的需要，并且得到作为微分方程 (**Pearson方程组**) 解的密度函数；和另外一些统计学家一样，Andrei Andreevic Markov (马尔科夫) (1856 - 1922) 不愿意用Pearson分布方程组，因为即使得了皮尔森曲线作为一个极限分布，也没有明显的模型来产生它们。Markov进一步证明Pearson χ^2 统计量为样本尺寸乘以Wilhelm Hector Richard albrecht Lexis (1837—1914) 的**离散系数**。Walter Frank Raphael Weldon (1860 - 1906) 利用取独立初第二项变量和的方法得到二项变量的联合分布。许多作者，比如 Alexander Craig Aitken (1895 - 1967)，已经参与了发展该思想；但是许多其它思想已经被用来获得联合分布。在Karl Pearson的方法不能产生更多的联合分布之后，Sergei Natanovic Bernstein (1880 - 1968) 认为一个更具有生产价值的方法可能存在于随机过程的领域中。

统计学简史

- Karl Pearson时代 , 1890 - 1920
- 到1920年为止的英国生物统计学派的主要成就为
 - (i) 收集并化简了许多经验数据 ; (ii) 定义了具有多重和总相关系数 ρ 的**联合正态分布** , 还定义了估计**误差的联合分布** ; (iii) 关于拟合度的 **χ^2 检验** , 比较观察分布和理论分布 , 包括由 Herbert Edward Soper (1865 - 1930) 引进的条件Poisson变量 ; (iv) 分析**列联表** , 特别是利用 **χ^2 统计量** ; (v) 当边缘分布充分细分时由最大似然法估计 ρ ; (vi) 当边缘分布没有充分定义时估计 ρ ; (vii) 由一个统一的参数估计系统来描述一组曲线 , 即矩方法 ; (viii) 利用正态定理解决**遗传选择问题** ; (ix) 通往独立性一般定理的某些进展 ; (x) 通往估计和检验估计精确性的一个理论的进展 ; (xi) 构造了适当的表。
- 与此同时 , 在法国的 Félix Edouard Justin Emile Borel (1871 - 1956) , Maurice Frenchet (1878-1973) 及 Jules Henri Poincaré (1854 - 1912) , 和在俄国的 Alekandr Aleksandrovic Chuprov (1874 - 1926) , A.A. Markov 和 Vsevolod Ivanovic Romanovsky (1879 - 1954) 作出许多贡献 , 特别是把数据的数学处理严格化。

统计学简史

- R.A.FISHER时代，1921-1936
- 所有的皮尔森的方法都可以应用于大样本，而且可以对方差作出较精确的估计。但对于出现在实际应用中的小样本，这些方法就未必奏效了。William Sealy Gosset (1876—1937) 因此导出了一个检验；按照R.A.Fisher的建议该检验在作了一个变换之后成为现在熟和的t-检验。
- Fisher以其四篇值得纪念的论文开创了一个新纪元；相关系数估计的精确分布；协调一致了Mendelian和生物统计对遗传学的不同方法；正确解释了列联表；估计和推断的一般定理。在1920年之后，在Rothamsted实验室，Fisher发展了有广泛应用价值的方差分析和试验与分析的理论。Fisher有很强的数学功底，特别是在组合论 (combinatorics) 方面，他能吸引其他数学家作为助手。他对应用领域的选择是很幸运的；研究结果都能立即应用并有明显的经济效应；能够有效地简化假设，比如误差的正态性和独立性；和一些顽固的教条斗争；试验的花费都很低；没有伦理问题。许多重要的步骤用来发展上面提到的Pearson学派的工作的一些分支。在 (iii) 和 (iv) 中的许多重要问题被解决了；给出了正确的自由度；K.Pearson已经为该目标前进了一段；在 (vii) 中Fisher发明了更有效的方法来估计；他拒绝了用矩方法来确定分布；在(Vi)中Fisher和Yate发表了统计表。
- 在误差分布和互相独立性的假设使其能用正交变换来保持线性和二次型之间的独立性，这样就可合理地利用t-检验和F-检验。Fisher看到农业试验能利用更复杂的设计。于是，行列表的影响能够按地理因素（行和列）及处理来分别分析。这能推广到n维Latin方的应用，把处理用于Latin方相应的位置上；该方法通过实行Graeco - Latin方来实现。Fisher及其助手和同事研究了设计问题，缺损值问题，非正交性等等；这些人包括Maurice Stevenson Bartlett (1910 -)，William Gemmell Cochran (1909 - 1982)，他们后来在美国特别有影响。此外，还有David John Finney (1917 -)，Joseph Oscar Irwin (1898-1982)，Kenneth Mather (1911—)，及 Frank Yates (1890 - 1972)。

统计学简史

- NEYMAN - PEARSON时代 , 1937 - 1949
- Jerzy Neyman (1894 - 1981) 及 Egon sharpe Pearson (1895—1980) 在一系列的杰出的文章中澄清了推断理论 , 特别是有关显著性检验的基本原理一其合理性以往是常被批评。早期的显著性检验为关于二项变量之间或均值之间的 , 它们被 K.Pearson推广至 χ^2 检验 , 被R.A.Fisher推广到F-检验 , 推广了Student T-检验。Neyman和 E.S.Fearson看出 , 为了更有效 , 应该考虑与待检验的零假设相对应的备选假设。他们在这样的检验中设立两种误差并因素导致了他们的基本引理 , 似然比检验 , 及势的概念 ; 他们顺便验证了大多数常见的显著性检验的应用 ; 他们还引进了置信区间 ; 但是他们的体系从未被Fisher所承认。Neyman和Pearson的工作影响了许多人 , 特别是美国人。
- 现代 , 统计变得越来越数学化了。为了解对分布和推断理论的一般描述 , 需要测度论 ; Fourier分析成为研究波动最自然的工具 ; 在分析方差的推断上 , 和在具对称性的设计以及在诸如 Graeco Latin方及 Steiner三元体的特别结构的代数的推断上需要应用群论和数论。组合理论能用于编码理论和有限几何。因此统计数学成为纯粹数学的一部分 , 并且因其在各种领域的广泛应用而被研究。因为通常的统计检验已经彻底地研究了 , 而且往往被置身于某些具体应用领域的实际工作者所应用 , 在统计学研究人员和实际工作者之间出现了一定距离 ; 但是这种现象在其他开拓性的领域中也能看到。

统计学简史

- **电子计算机已经带来了巨大的变化。**数据，比如海洋学中水面的高度，电磁能（特别是无线电波）的流量，工业过程的状态，生物的状态，都能用计算机**收集**；没有计算机这些是不实际或不可能的。计算机节省了大量人力，特别是在输出**重复计算**上，例如在计算多元分析的相关系数和其它检验统计量时。由于计算软件包可用于所有通常的检验，特别是关于方差分析，则节省更多。**高速计算**使得有可能运用匹配和排列检验。当分布不能写成一个封闭的分析公式时，显著性水平也能计算；另外，计算机能用 Monte Carlo方法计算每一个事件的概率或近似显著性水平。由于利用软件包很方便，有时导致对统计问题欠考虑而产生的结论，特别是在多重比较上。
- 模型在统计和科学工作中的作用现已被广泛承认；虽然基于应用领域的经验和知识，模型的选择在某种程度上是美学上的和任意的；但是一旦模型被选定，所有的推断都是数学的，用不着进一步的假设或原则。所用的推断体系在某种程度上也是任意的；备选体系已经被大量研究。基于信仰的推断不再扮演重要的角色。**贝叶斯模型**在 Fisher时代曾一度失色，之后又被更广泛地应用。**信息论**已被引进；多数统计推断看来仍然以和Neyman-Pearson理论一致的方式来运作，运用在K.Pearson和Fisher时代引人的检验。
- **现代国家的增长的能力和兴趣要求以低花费收集更多的数据。**Antlers Nicolai Kiaer (1838 - 1919)有远见地建议**概率抽样应补充到人口普查方法**中。这样的抽样已经在其被Prasanta chandra Mahalanobis (1893 - 1972)引进之后成为在印度和其它地方的标准实践。被Andrew Shewhart(1891 - 1967)所推广的工业质量控制方法也有类似的意图。
- 许多新的分支或专门化和应用已经被发展了：**决策论，时间序列，多元分析，经济计量学，博奕论，临床试验，非参数推断，序贯分析，数学生物分类学……**概率统计及其应用正在继续发展和扩大。

数据分析

数据分析是指用适当的统计分析方法对收集来的大量数据进行分析，提取有用信息和形成结论而对数据加以详细研究和概括总结的过程。这一过程也是质量管理体系的支持过程。在实用中，数据分析可帮助人们作出判断，以便采取适当行动。现代数据分析的数学基础在20世纪早期就已确立，但直到计算机的出现才使得实际操作成为可能，并使得数据分析得以推广。数据分析是数学与计算机科学相结合的产物。

Data Analysis

Data analysis is a process of inspecting, cleansing, transforming, and modeling data with the goal of discovering useful information, informing conclusions, and supporting decision-making. Data analysis has multiple facets and approaches, encompassing diverse techniques under a variety of names, and is used in different business, science, and social science domains. In today's business world, data analysis plays a role in making decisions more scientific and helping businesses operate more effectively.

Process of Data Analysis

- 1 **The process of data analysis**
- 1.1 **Data requirements**
- 1.2 **Data collection**
- 1.3 **Data processing**
- 1.4 **Data cleaning**
- 1.5 **Exploratory data analysis**
- 1.6 **Modeling and algorithms**
- 1.7 **Data product**
- 1.8 **Communication**

Eight types of quantitative messages

- 1. Time-series**
- 2. Ranking**
- 3. Part-to-whole**
- 4. Deviation**
- 5. Frequency distribution**
- 6. Correlation**
- 7. Nominal comparison**
- 8. Geographic or geospatial**

#	Task	Examples
1	Retrieve Value	<ul style="list-style-type: none"> - <i>What is the mileage per gallon of the Ford Mondeo?</i> - <i>How long is the movie Gone with the Wind?</i>
2	Filter	<ul style="list-style-type: none"> - <i>What Kellogg's cereals have high fiber?</i>- <i>What comedies have won awards?</i> - <i>Which funds underperformed the SP-500?</i>
3	Compute Derived Value	<ul style="list-style-type: none"> - <i>What is the gross income of all stores combined?</i> - <i>How many manufacturers of cars are there?</i>
4	Find Extremum	<ul style="list-style-type: none"> - <i>What director/film has won the most awards?</i> - <i>What Marvel Studios film has the most recent release date?</i>
5	Sort	<ul style="list-style-type: none"> - <i>Order the cars by weight.</i>- <i>Rank the cereals by calories.</i>
6	Determine Range	<ul style="list-style-type: none"> - <i>What is the range of film lengths?</i>- <i>What is the range of car horsepowers?</i> - <i>What actresses are in the data set?</i>
7	Characterize Distribution	<ul style="list-style-type: none"> - <i>What is the distribution of carbohydrates in cereals?</i>- <i>What is the age distribution of shoppers?</i>
8	Find Anomalies	<ul style="list-style-type: none"> - <i>Are there exceptions to the relationship between horsepower and acceleration?</i>- <i>Are there any outliers in protein?</i>
9	Cluster	<ul style="list-style-type: none"> - <i>Are there groups of cereals w/ similar fat/calories/sugar?</i>- <i>Is there a cluster of typical film lengths?</i>
10	Correlate	<ul style="list-style-type: none"> - <i>Is there a correlation between carbohydrates and fat ?</i> - <i>Do different genders have a preferred payment method?</i>
11	Contextualization	<ul style="list-style-type: none"> - <i>Are there groups of restaurants that have foods based on my current caloric intake?</i>

Data Mining

Data mining is the process of discovering patterns in large data sets involving methods at the intersection of machine learning, statistics, and database systems.[1] Data mining is an interdisciplinary subfield of computer science and statistics with an overall goal to extract information (with intelligent methods) from a data set and transform the information into a comprehensible structure for further use. Data mining is the analysis step of the "knowledge discovery in databases" process, or KDD. Aside from the raw analysis step, it also involves database and data management aspects, data pre-processing, model and inference considerations, interestingness metrics, complexity considerations, post-processing of discovered structures, visualization, and online updating. The difference between data analysis and data mining is that data analysis is used to test models and hypotheses on the dataset, e.g., analyzing the effectiveness of a marketing campaign, regardless of the amount of data; in contrast, data mining uses machine-learning and statistical models to uncover clandestine or hidden patterns in a large volume of data.

Data Mining

The term "data mining" is in fact a misnomer, because the goal is the extraction of patterns and knowledge from large amounts of data, not the extraction (mining) of data itself.[7] It also is a buzzword[8] and is frequently applied to any form of large-scale data or information processing (collection, extraction, warehousing, analysis, and statistics) as well as any application of computer decision support system, including artificial intelligence (e.g., machine learning) and business intelligence. The book Data mining: Practical machine learning tools and techniques with Java[9] (which covers mostly machine learning material) was originally to be named just Practical machine learning, and the term data mining was only added for marketing reasons.[10] Often the more general terms (large scale) data analysis and analytics – or, when referring to actual methods, artificial intelligence and machine learning – are more appropriate.

Data Mining

The actual data mining task is the semi-automatic or automatic analysis of large quantities of data to extract previously unknown, interesting patterns such as groups of data records (cluster analysis), unusual records (anomaly detection), and dependencies (association rule mining, sequential pattern mining). This usually involves using database techniques such as spatial indices. These patterns can then be seen as a kind of summary of the input data, and may be used in further analysis or, for example, in machine learning and predictive analytics. For example, the data mining step might identify multiple groups in the data, which can then be used to obtain more accurate prediction results by a decision support system. Neither the data collection, data preparation, nor result interpretation and reporting is part of the data mining step, but do belong to the overall KDD process as additional steps.

The related terms data dredging, data fishing, and data snooping refer to the use of data mining methods to sample parts of a larger population data set that are (or may be) too small for reliable statistical inferences to be made about the validity of any patterns discovered. These methods can, however, be used in creating new hypotheses to test against the larger data populations.

Data Science

Data science is a multi-disciplinary field that uses scientific methods, processes, algorithms and systems to extract knowledge and insights from structured and unstructured data. Data science is the same concept as data mining and big data: "use the most powerful hardware, the most powerful programming systems, and the most efficient algorithms to solve problems".

Data science is a "concept to unify statistics, data analysis, machine learning and their related methods" in order to "understand and analyze actual phenomena" with data. It employs techniques and theories drawn from many fields within the context of mathematics, statistics, information science, and computer science. Turing award winner Jim Gray imagined data science as a "fourth paradigm" of science (empirical, theoretical, computational and now data-driven) and asserted that "everything about science is changing because of the impact of information technology" and the data deluge.

Data Science

In 2012, when Harvard Business Review called it "The Sexiest Job of the 21st Century", the term "data science" became a buzzword. It is now often used interchangeably with earlier concepts like business analytics, business intelligence, predictive modeling, and statistics. Even the suggestion that data science is sexy was paraphrasing Hans Rosling, featured in a 2011 BBC documentary with the quote, "Statistics is now the sexiest subject around." Nate Silver referred to data science as a sexed up term for statistics. In many cases, earlier approaches and solutions are now simply rebranded as "data science" to be more attractive, which can cause the term to become "dilute[d] beyond usefulness." While many university programs now offer a data science degree, there exists no consensus on a definition or suitable curriculum contents. To its discredit, however, many data-science and big-data projects fail to deliver useful results, often as a result of poor management and utilization of resources.

如何成为数据科学家

一、数据科学家的起源

"数据科学" (DataScience) 起初叫" datalogy "。最初在1966年由Peter Naur提出，用来代替"计算机科学" (丹麦人，2005年图灵奖得主，丹麦的计算机学会的正式名称就叫Danish Society of Datalogy，他是这个学会的第一任主席。Algol 60是许多后来的程序设计语言，包括今天那些必不可少的软件工程工具的原型。图灵奖被认为是“计算科学界的诺贝尔奖”。)

1996年，International Federation of Classification Societies (IFCS)国际会议召开。数据科学一词首次出现在会议 (Data Science, classification, and related methods) 标题里。

1998年，C.F. Jeff Wu做出题为“统计学=数据科学吗？”的演讲，建议统计改名数据的科学统计数据的科学家。（吴教授于1987年获得COPSS奖，2000年在台湾被选为中研院院士，2004年作为第一位统计学者当选美国国家工程院院士，也是第一位华人统计学者获此殊荣。）

2002年，国际科学理事会：数据委员会科学和技术 (CODATA) 开始出版数据科学杂志。

2003年，美国哥伦比亚大学开始发布数据科学杂志，主要内容涵盖统计方法和定量研究中的应用。

2005年，美国国家科学委员会发表了"Long-lived Digital Data Collections: Enabling Research and Education in the 21st Century"，其中给出数据科学家的定义：信息科学与计算机科学家，数据库和软件工程师，领域专家，策展人和标注专家，图书管理员，档案员等数字数据管理收集者都可以成为数据科学家。它们主要任务是：进行富有创造性的查询和分析。

如何成为数据科学家

二、数据科学家的能力框架

Thomas H. Davenport (埃森哲战略变革研究院主任) 和 D.J. Patil (美国科学促进会科学与技术政策研究员 , 为美国国防部服务) 的话来总结数据科学家需要具备的能力 :

数据科学家倾向于用探索数据的方式来看待周围的世界。 (好奇心)

把大量散乱的数据变成结构化的可供分析的数据 , 还要找出丰富的数据源 , 整合其他可能不完整的数据源 , 并清理成结果数据集。 (问题分体整理能力)

新的竞争环境中 , 挑战不断地变化 , 新数据不断地流入 , 数据科学家需要帮助决策者穿梭于各种分析 , 从临时数据分析到持续的数据交互分析。 (快速学习能力)

数据科学家会遇到技术瓶颈 , 但他们能够找到新颖的解决方案。 (问题转化能力)

当他们有所发现 , 便交流他们的发现 , 建议新的业务方向。 (业务精通)

他们很有创造力的展示视觉化的信息 , 也让找到的模式清晰而有说服力。 (表现沟通能力)

他们会把蕴含在数据中的规律建议给 Boss , 从而影响产品 , 流程和决策。 (决策力)

如何成为数据科学家

三、数据科学家所需硬件技能

《数据之美 Beautiful Data》的作者Jeff Hammerbacher在书中提到，对于 Facebook 的数据科学家“我们发现传统的头衔如商业分析师、统计学家、工程师和研究科学家都不能确切地定义我们团队的角色。该角色的工作是变化多样的：在任意给定的一天，团队的一个成员可以用 Python 实现一个阶段的处理管道流、设计假设检验、用工具R在数据样本上执行回归测试、在 Hadoop 上为数据密集型产品或服务设计和实现算法，或者把我们分析的结果以清晰简洁的方式展示给企业的其他成员。为了掌握完成这多方面任务需要的技术，我们创造了数据科学家这个角色。”

- (1) 计算机科学。一般来说，数据科学家大多要求具备编程、计算机科学相关的专业背景。简单来说，就是对处理大数据所必需的Hadoop、Mahout等大规模并行处理技术与机器学习相关的技能。
- (2) 数学、统计、数据挖掘等。除了数学、统计方面的素养之外，还需要具备使用SPSS、SAS等主流统计分析软件的技能。个人建议从python入手，拥有丰富的statistical libraries，NumPy，SciPy.org，Python Data Analysis Library，matplotlib: python plotting。
- (3) 数据可视化（Visualization）。信息的质量很大程度上依赖于其表达方式。对数字罗列所组成的数据中所包含的意义进行分析，开发Web原型，使用外部API将图表、地图、Dashboard等其他服务统一起来，从而使分析结果可视化，这是对于数据科学家来说十分重要的技能之一。
- (4) 跨界为王。麦肯锡认为未来需要更多的“translators”，能够在IT技术，数据分析和商业决策之间架起一座桥梁的复合型人才是最被人需要的。”

如何成为数据科学家

数据科学家研究



数字化数据、处理速度和数据处理能力的激增，加上分析数据的新工具——激发了人们对数据科学的巨大兴趣。各种规模的机构都青睐有能力将数据的价值转化为迈向商业利益的预见性见识的人才。那些数据是指由手机探测器、社交媒体，监督者，医学成像，智能网络等产生。除了机会越来越多，数据科学的需求也超过了人才的供应，而且未来五年都会是这样。

哪些人是数据学科从业者？他们需要什么技能？为什么他们如此不同？

三分之二的人相信数据科学家的人才需求超过了的供应

未来5年，对数据科学家的需求：



只有12%的人将商业情报人员作为新数据科学家的最佳来源

数据科学家人才的最佳来源

34%

学计算机科学的学生

27%

除计算机科学的专业人员

24%

除计算机科学的其它领域的学生

当今的商业情报专业人士

12%

其他
2%

为了调整，各部分比率加起来可能没有百分之百

缺少培训和资源是组织机构里数据科学家的最大障碍

在我们组织里数据科学的最大障碍是：

32%

雇员得不到恰当的技能和培训

32%

缺少预算和资源

14%

错误的组织结构

10%

缺少必要的工具和技术

9%

不足的行政支持

3%

其他

数据科学家比商业情报人员可能有更高的学位



商业情报人员不可避免地在大学里学习商业

数据科学家有多种的背景，特别是高难度学科



商业情报人才

数据科学家

数据科学家相信新技术将创造更多需要数据科学家的需求

数据科学家通过利用
自动化处理数据
分析工作，降低了
对数据科学家的需求



83% 数据科学家因为
开发潜在价值，
增加了对数据
科学家的需求

数据科学家的特征



大
数
据
科
学
家

在处理不完整的数据时
我感到很舒适

我要完整的数据

我的数据文件通常
是非常散乱的

我的数据文件夹通常很干净

我研究数据看
它能告诉我什么

我汇报数据说了什么

我的资料组非常大
管理好它是项挑战

即使我的资料组很大，
也很容易管理

我的发现驱动生
产和运作决定

我的发现测量了过去的表现



普
通
的
数
据
科
学
家

10% 大数据
科学

65% 中等
科学

25% 普通数据科学

数据科学家更可能投入到数据围绕的生活圈



数据科学家跟谁一起工作？

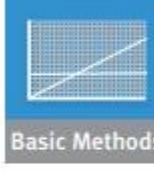


EMC²

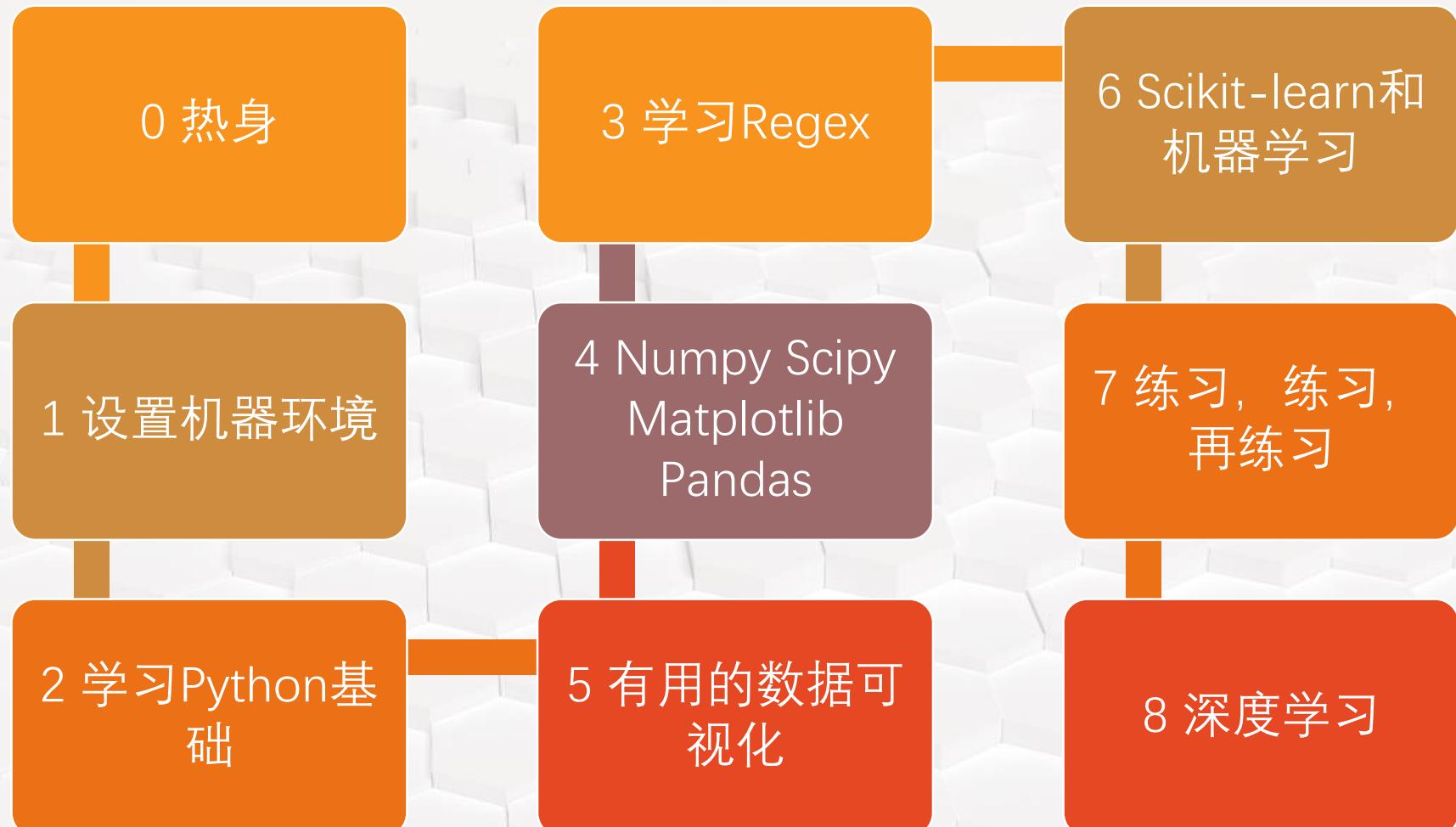
SOURCE: EMC Data Scientist Study, 2011

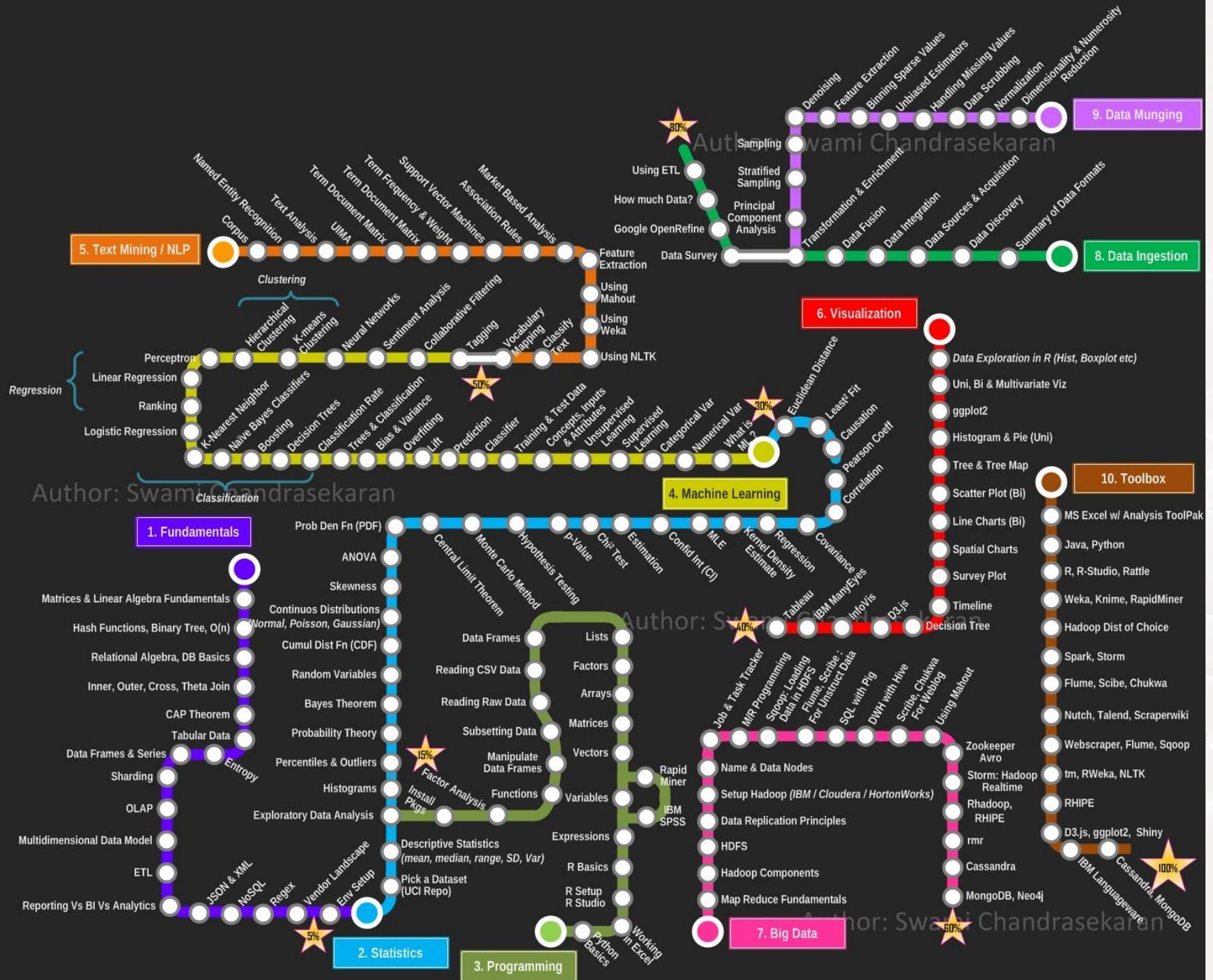
新商业时代 编译：心中那片天

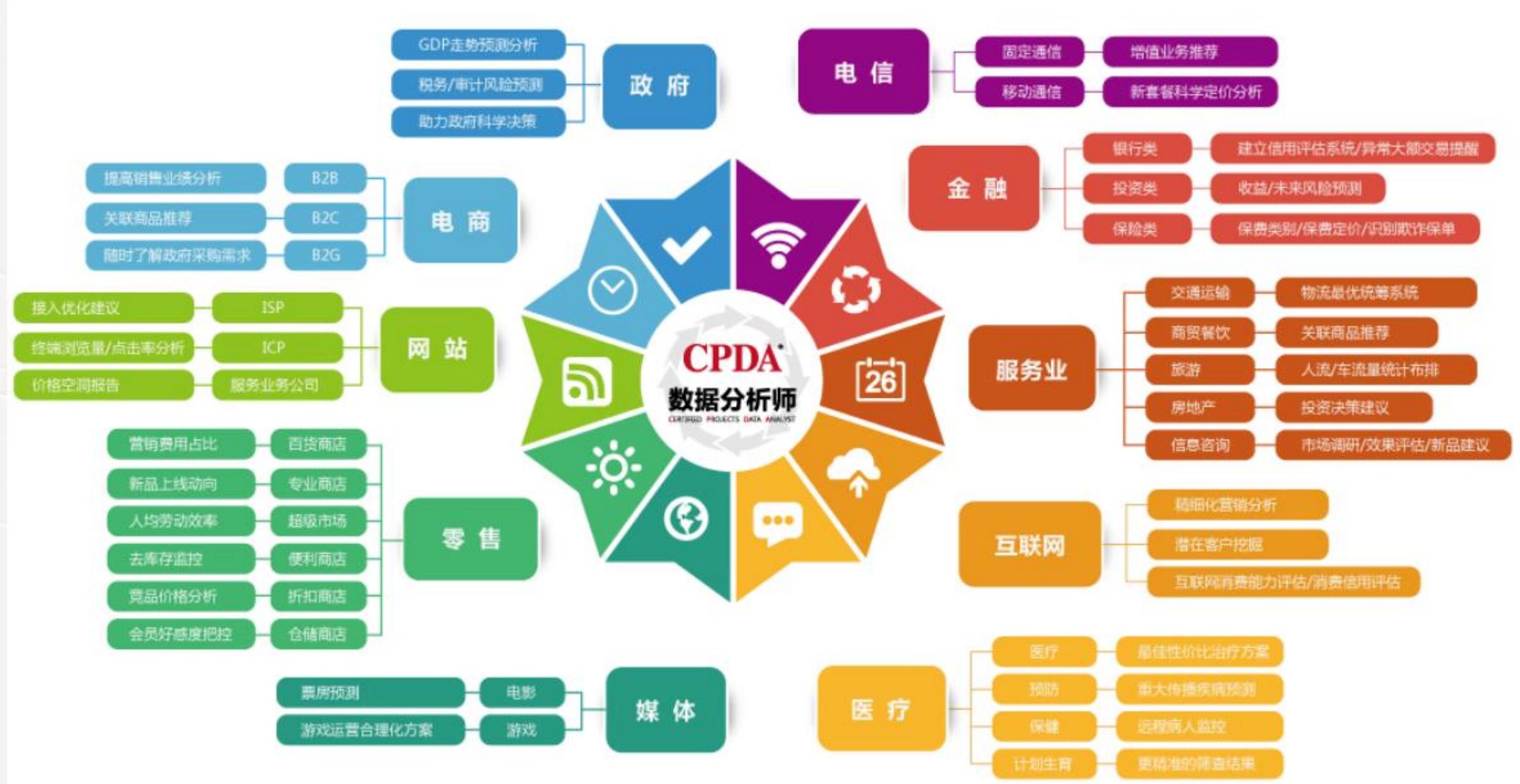
如何成为数据科学家

 Introduction	Big Data Overview	State of the practice in analytics	The role of the Data Scientist	Big Data Analytics in industry verticals
Introduction to Big Data Analytics				
 Analytics Lifecycle	Key roles for a successful analytics project	Main phases of the lifecycle	Developing core deliverables for stakeholders	
End-to-end data analytics lifecycle				
 Basic Methods	Introduction to R	Analyzing and exploring data with R	Statistics for model building and evaluation	
Using R to execute basic analytics methods				
 Adv. Methods	K-Means Clustering	Association Rules	Linear and Logistic Regression	Naïve Bayesian Classifier
Advanced analytics and statistical modeling for Big Data – Theory and Methods				
 Tools	Decision Trees	Time Series Analysis	Text Analysis	
Advanced analytics and statistical modeling for Big Data – Technology and Tools				
 Lab	Using MapReduce/Hadoop for analyzing unstructured data	Hadoop ecosystem of tools	In-database Analytics	MADlib and Advanced SQL Techniques
Advanced analytics and statistical modeling for Big Data – Technology and Tools				
How to operationalize an analytics project				
Creating the Final Deliverables				
Data Visualization Techniques				
Hands-on Application of Analytics Lifecycle to a Big Data Analytics Problem				
Endgame, or Putting it all together				

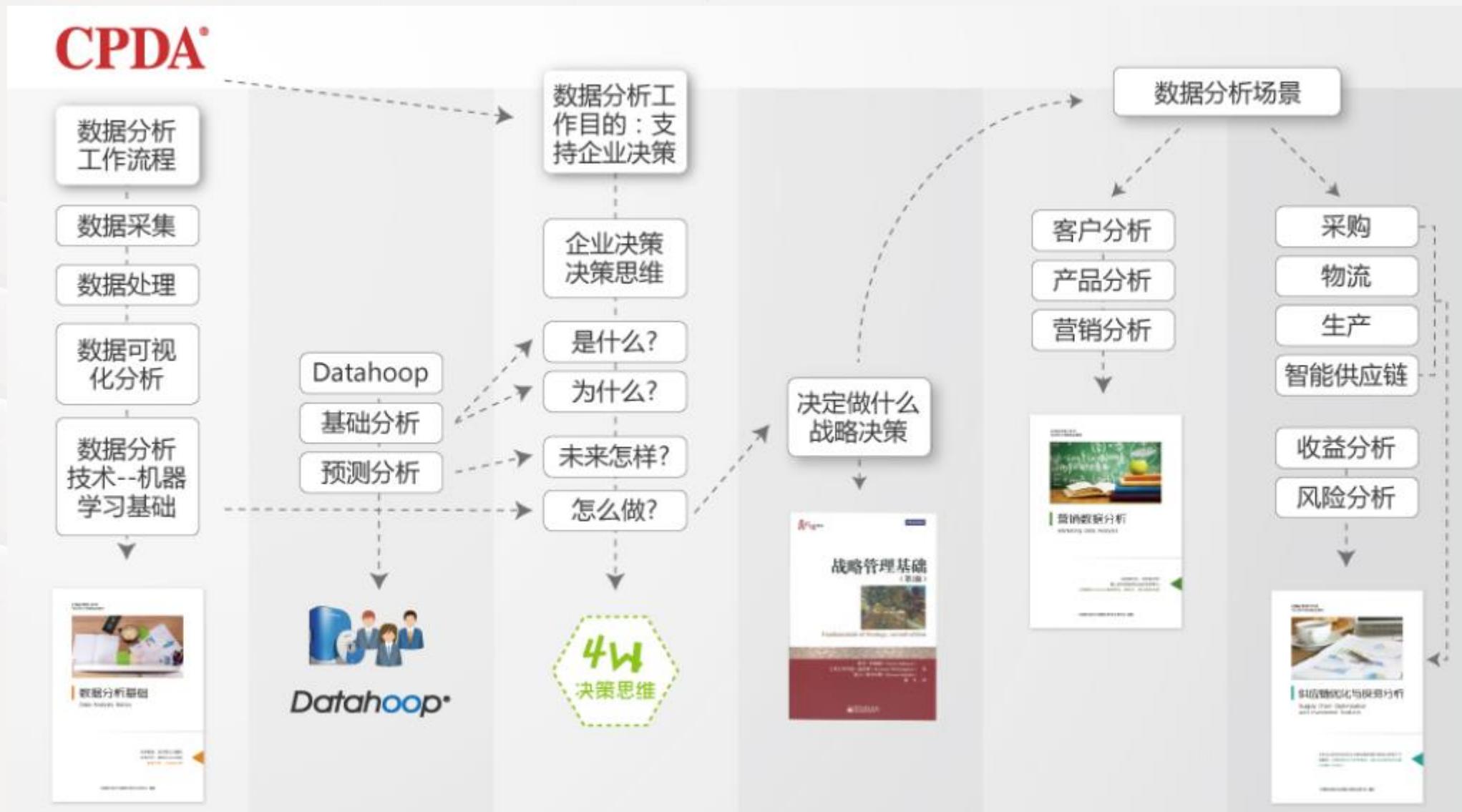
如何成为数据科学家







CPDA



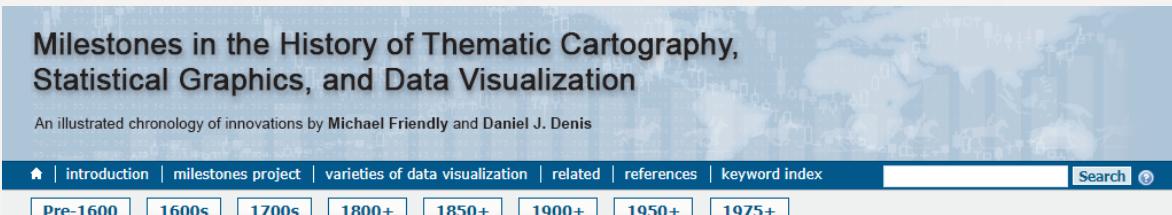
数据可视化的前世今生

Milestones in the History of Thematic Cartography, Statistical Graphics, and Data Visualization

An illustrated chronology of innovations by Michael Friendly and Daniel J. Denis

Home | introduction | milestones project | varieties of data visualization | related | references | keyword index | Search | Help

Pre-1600 1600s 1700s 1800+ 1850+ 1900+ 1950+ 1975+

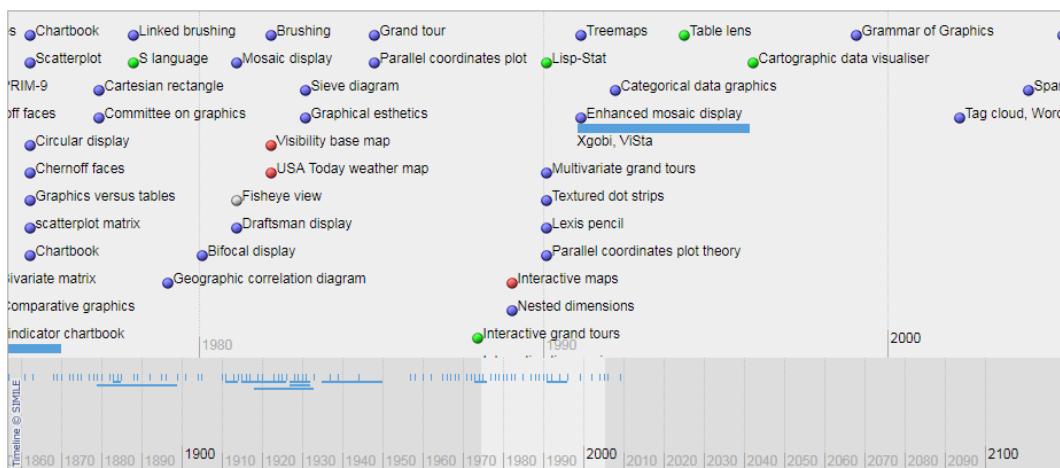


Timeline

This page provides a graphic overview of the events in the history of data visualization that we call "milestones." These milestones are shown below in the form of an *interactive timeline*. The timeline is divided into two *vertical sections*. You can *drag each section left or right* to see milestones of different time periods. You can also click one of the links at the bottom of the timeline to jump to a particular epoch.

Each of the milestone's in the timeline can be clicked to reveal its summary that includes both a link to its full details and a category to which it belongs. The category can also be clicked to initiate a search of other milestone's based on that category.

Item categories: ● Cartography ● Statistics and graphics ● Technology ● Other

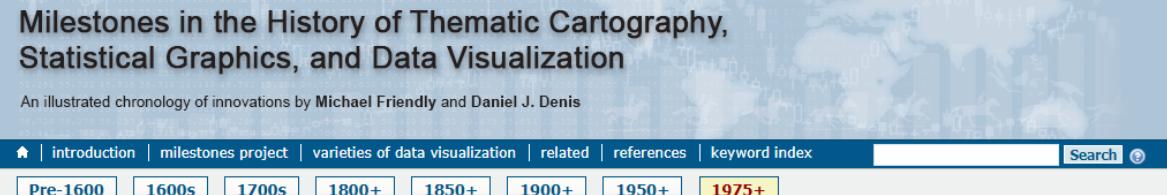


Milestones in the History of Thematic Cartography, Statistical Graphics, and Data Visualization

An illustrated chronology of innovations by Michael Friendly and Daniel J. Denis

Home | introduction | milestones project | varieties of data visualization | related | references | keyword index | Search | Help

Pre-1600 1600s 1700s 1800+ 1850+ 1900+ 1950+ 1975+



1975-present: High-D data visualization

It is harder to provide a succinct overview of the most recent developments in data visualization, because they are so varied, have occurred at an accelerated pace, and across a wider range of disciplines. It is also more difficult to highlight the most significant developments (and because we have focused on the earlier history), so there are presently areas and events unrepresented here.

With this disclaimer, a few major themes stand out

- the development of a variety of highly interactive computer systems and more importantly,
 - new paradigms of direct manipulation for visual data analysis (linking, brushing, selection, focusing, etc.)
 - new methods for visualizing high-dimensional data (grand tour, scatterplot matrix, parallel coordinates plot, etc.)
 - the invention of new graphical techniques for discrete and categorical data (fourfold display, sieve diagram, mosaic plot, etc.), and analogous extensions of older ones (diagnostic plots for generalized linear models, mosaic matrices, etc.) and
 - the application of visualization methods to an ever-expanding array of substantive problems and data structures.
- These developments in visualization methods and techniques arguably depended on advances in theoretical and technological infrastructure. Some of these are: (a) large-scale software engineering; (b) extensions of classical linear statistical modeling to wider domains; (c) vastly increased computer processing speed and capacity, allowing computationally intensive methods and access to massive data problems.

In turn, the combination of these themes and advances now provides some solutions for earlier problems.

Jump to Milestone...

1975 Office of Management and Budget

Measuring 50 years of economic change

Chartbook

Added: 2008-07-17

用好的方法诠释统计数据



<https://www.bilibili.com/video/av15387839?from=search&seid=15785884901541150373>

数据可视化

数据可视化是关于数据视觉表现形式的科学技术研究。其中，这种数据的视觉表现形式被定义为，一种以某种概要形式抽提出来的信息，包括相应信息单位的各种属性和变量。可视化仍是一个处于不断演变之中的概念，其边界在不断地扩大。主要指的是技术上较为高级的技术方法，而这些技术方法允许利用图形、图像处理、计算机视觉以及用户界面，通过表达、建模以及对立体、表面、属性以及动画的显示，对数据加以可视化解释。与立体建模之类的特殊技术方法相比，数据可视化所涵盖的技术方法要广泛得多。

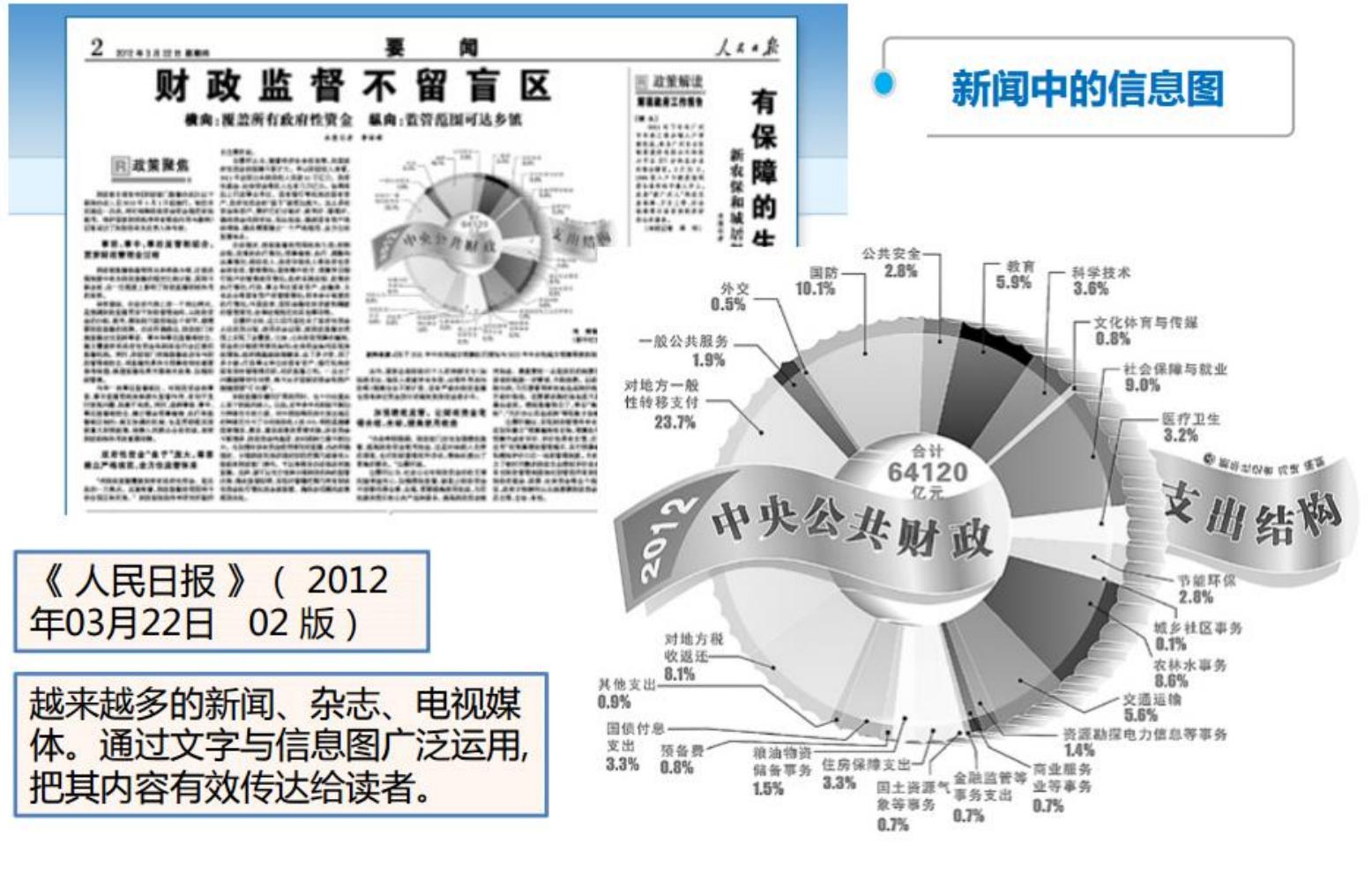
数据可视化的五个类别



信息图

信息图是最早出现的可视化。传统意义上，信息图被看作是帮助人们更好地理解特定内容的视觉元素，如统计图表、示意图。现在，信息图表是指信息、数据、知识等的可视化表达，通常会具有一定创造性、趣味性，便于阅读。

最早也是最广泛的应用是在新闻传媒领域，现在已然超越了新闻传媒领域并在其他一些产业都被广泛采用，比如，科技视觉化，产品设计，教育事业，IT，商业交流和娱乐业。



科学可视化

随着计算机的出现和计算机技术的发展，最先提出可视化需求的是科学研究界。

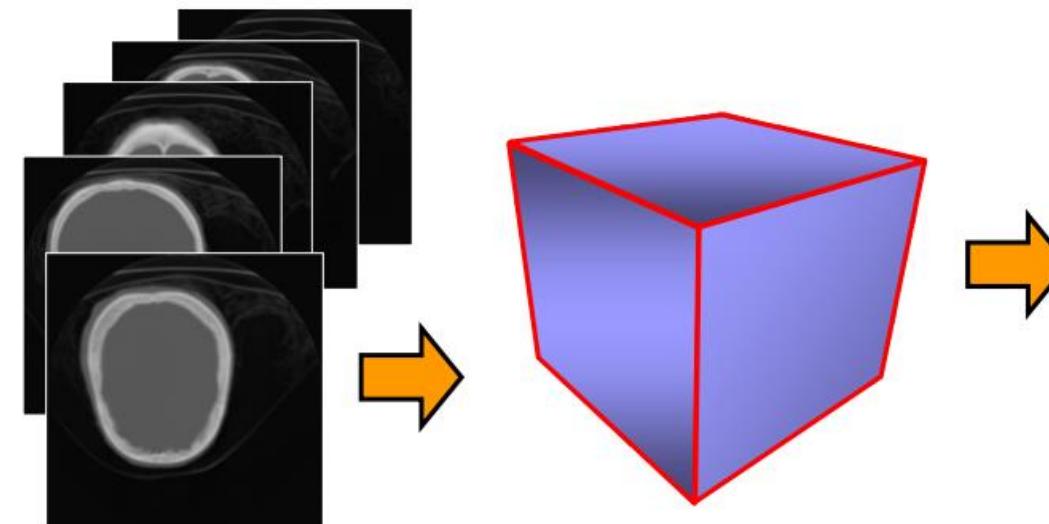
1987年，由布鲁斯·麦考梅克
(Bruce H. McCormick) 等3位所编写 的美国国家科学基金会报告

《Visualization in Scientific Computing》(“科学计算之中的可视化”)
首次提出科学可视化的需求。

这份报告之中强调了，随着计算机运算能力的迅速提升，生物、医学、气象、天文等等方面建立了规模越来越大，复杂程度越来越高的数值模型，因而，就需要高级的计算机图形学技术与方法来处理和可视化 这些规模庞大的数据集。

常见的包括医学影像数据可视化、流场数据可视化、天文粒子数据可视化等

医学影像数据可视化



天文粒子数据可视化

数值模拟：暗物质，模拟宇宙结构演化过程

模拟方法: N-body 数值模拟-140亿年宇宙演化

输出数据：300亿粒子，每个时刻1.4TB，64个
时刻，总共90TB

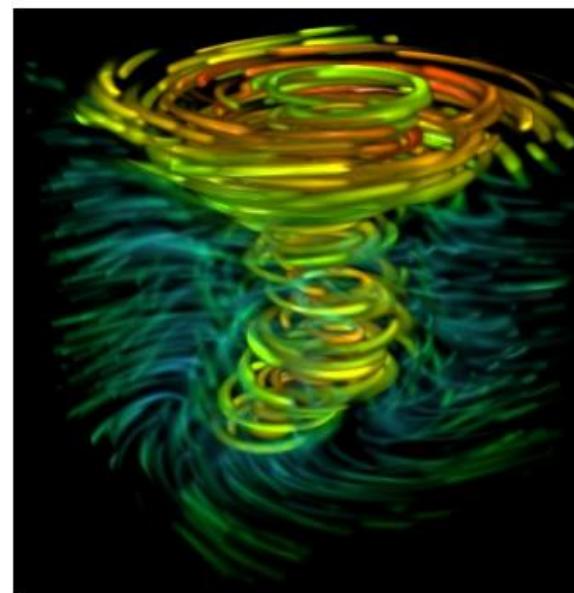


中科院超级计算中心
中科院计算机网络信息中心

流场数据可视化

流场可视化是科学可视化的重要分支，对流体研究有着巨大的意义。运用计算机图形学和图像处理技术，将流场数据转换为二维或三维图形、图像或动画进行呈现，并对其模式和相互关系进行可视化分析。

流场可视化方法在汽车、航空和航海动力学、生物、医学医疗等众多领域都已有着重要而广泛的应用。



信息可视化

information visualization是由斯图尔特·卡德 (Stuart K. Card) 等3人，于1989年，在一篇发表的论文中创造出来的

《Readings in Information Visualization: Using Vision to Think》

随着计算机应用领域越来越广泛，来自商业、财务、行政管理、数字媒体等方面大型异质性数据集合日益膨胀。信息可视化研究：从数据中提取信息，并通过图形化的方式表达出来，利用人类视觉能力，并且辅以交互手段，用以揭示数据中隐含信息的奥秘。

强大的需求造就和带动了近十几年以来信息可视化研究的炙手可热。



● 标签云—文本关键字的可视化



- 标签云（文字云）（ Tag Cloud ）是一种常用的文本可视化方法。
- 照片分享社区 Flickr 是首个使用标签云的知名网站，其标签云由网站共同创立者、界面交互设计师斯图尔特·巴特菲尔德在 2006 年设计创造。
- 标签云是关键词的视觉化描述，用于分析文本。标签一般是独立的词汇，词的热门程度又能通过改变字体大小或颜色来表现，词的布局可以有多种形式，词本身可以是超级链接，直接指向与该词相联的一系列条目。

地理信息的可视化



➤如何表现有地理位置的数据是可视化很重要的一块领域。最直接的就是把数据或者数据分析的结果形象化地表现在地图上，帮助使用者理解数据的规律和趋势。地图很多时候确实是包含地理信息的数据最有效的可视化方法。



北京除夕短信可视化
http://v.youku.com/v_show/id_XNDU1MTAyMDA4.html

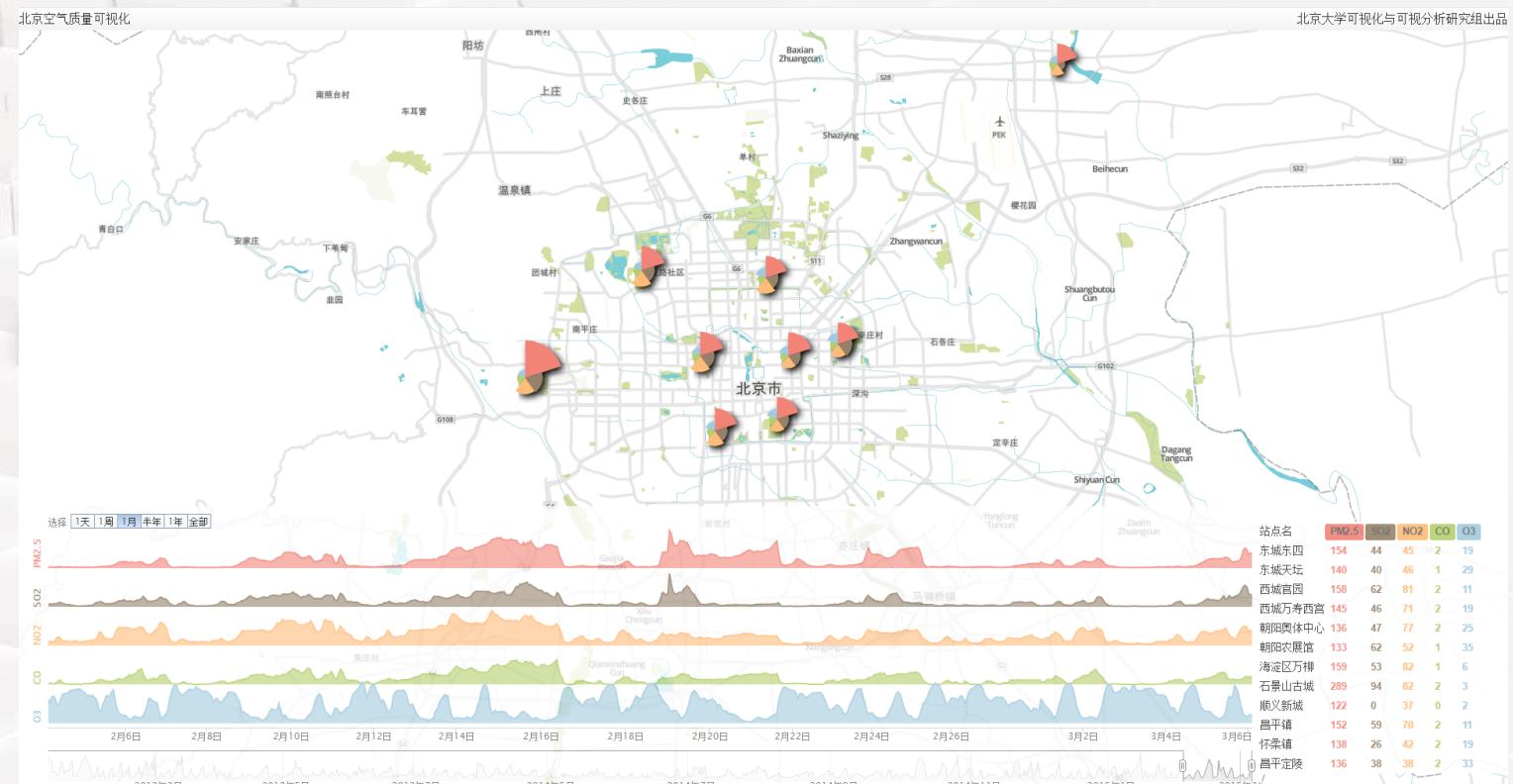


可视化分析

2005年，美国可视化与分析中心首先提出可视分析的概念

可视分析，是一门新兴的通过交互可视界面来进行分析、推理和决策的交叉学科。人们通过使用可视分析的工具和技术，从海量、动态、不确定甚至包含相互内容冲突的数据中整合信息，获取更深层的理解。可视分析允许人们检测预期的信息和探索未知的内容。

可视分析研究涉及计算机图形学、人机交互、可视化、分析、感知和认知等诸多领域。将在海量数据分析、社会关系分析、安全保障、应急事件分析处理决策等方面发挥重要作用。



<http://vis.pku.edu.cn/bjairvis/>

微博的可视分析

<http://vis.pku.edu.cn/weibova/weiboevents/index.html?>

- 微博这个名字，您一定不陌生，说不定您还是最积极的用户之一呢。
- 人类已经无法阻止微博进入我们的工作和生活。
- 但是如果您仅仅把微博用作亲朋好友交心、分享、聊天的工具，那么您仅仅看到了它的冰山一角。随着社交网络的发展，微博更成为事件传播的途径之一，各种各样的新闻和故事在微博的帮助下在微博用户的推动下更完善而壮大起来。
- 微博正在改变着媒体生态。一是信息流动模式由单极向多极转化。二是独立表达方式强化了话语权。
- 微博也在重构舆论传播格局。更快的表达。更大的影响。



from Data to Viz

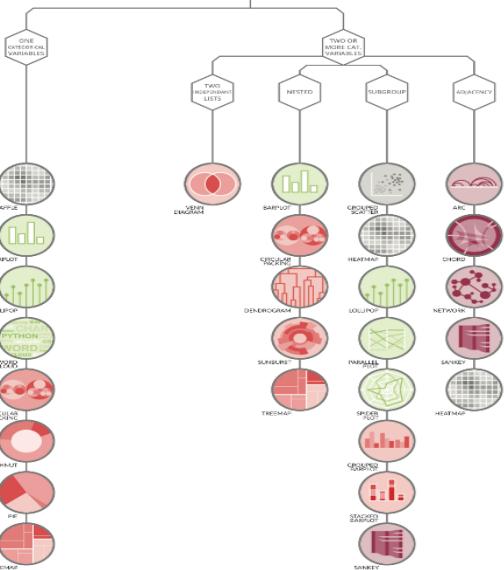
'From Data to Viz' is a classification of chart types based on input data format. It will help you find the perfect chart in three simple steps :

- 1 Identify what type of data you have.
- 2 Go to the corresponding decision tree and follow it down to a set of possible charts.
- 3 Choose the chart from the set that will suit your data and your needs best.

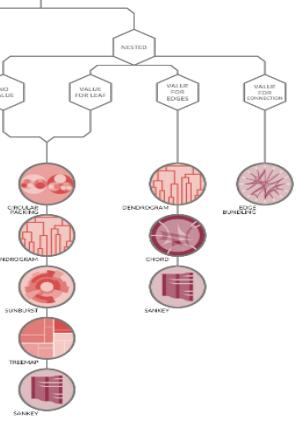
Dataviz is a world with endless possibilities and this project does not claim to be exhaustive. However it should provide you with a good starting point. For an interactive version and much more, visit:

data-to-viz.com

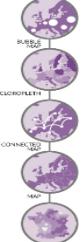
CATEGORIC



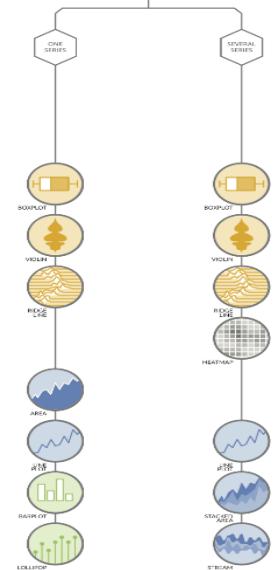
RELATIONAL



MAP



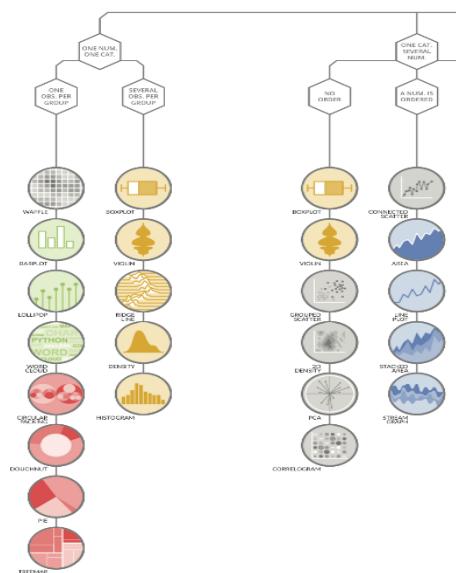
TIME SERIES



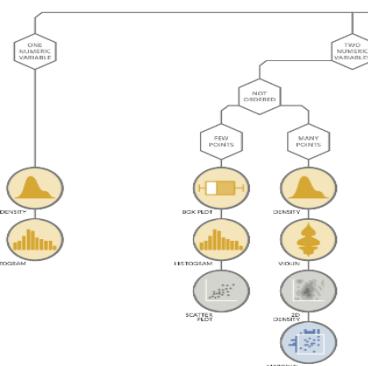
WHAT DO YOU WANT TO SHOW ?

- | | |
|---|---|
| ● Distribution
● Correlation
● Ranking
● Part of a whole | ● Evolution
● Maps
● Flow |
|---|---|

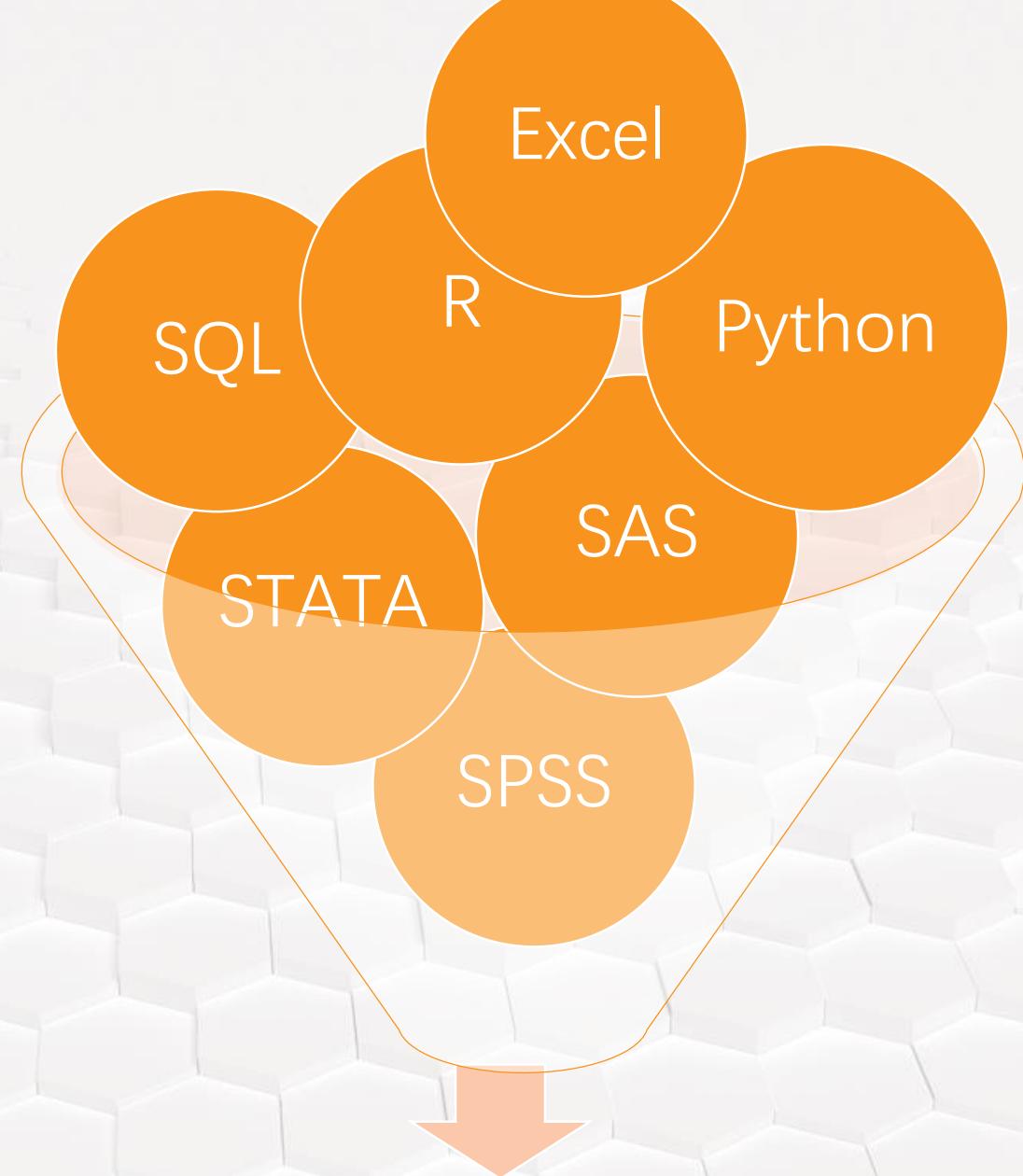
CATEGORIC AND NUMERIC



NUMERIC



为什么用Python?



我该用什么?

Mother Tongues

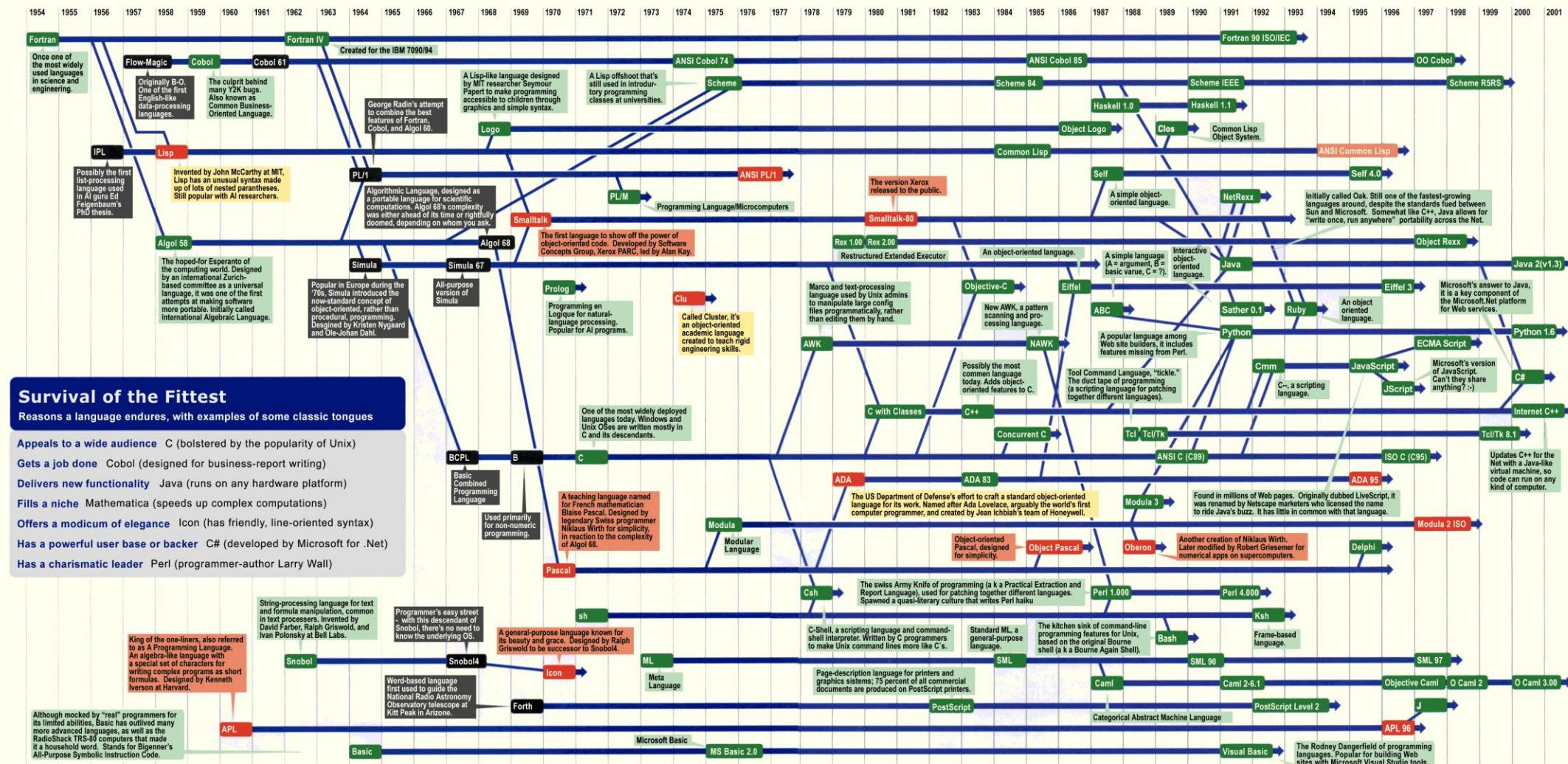
Tracing the roots of computer languages through the ages

Just like half of the world's spoken tongues, most of the 2,300-plus computer programming languages are either endangered or extinct. As powerhouses C/C++, Visual Basic, Cobol, Java and other modern source codes dominate our systems, hundreds of older languages are running out of life.

An ad hoc collection of engineers-electronic lexicographers, if you will—aim to save, or at least document the lingo of classic software. They're combing the globe's 9 million developers in search of coders still fluent in these nearly forgotten lingua frangas. Among the most endangered are Ada, APL, B (the predecessor of C), Lsp, Oberon, Smalltalk, and Simula.

Code-raker Grady Booch, Rational Software's chief scientist, is working with the Computer History Museum in Silicon Valley to record and, in some cases, maintain languages by writing new compilers so our ever-changing hardware can grok the code. Why bother? "They tell us about the state of software practice, the minds of their inventors, and the technical, social, and economic forces that shaped history at the time," Booch explains. "They'll provide the raw material for software archaeologists, historians, and developers to learn what worked, what was brilliant, and what was an utter failure." Here's a peek at the strongest branches of programming's family tree. For a nearly exhaustive rundown, check out the Language List at HTTP://www.informatik.uni-freiburg.de/Java/misc/lang_list.html. - Michael Mendeno

Key	
1954	Year Introduced
Active:	thousands of users
Protected:	taught at universities; compilers available
Extinct:	usage dropping off
Endangered:	no known active users or up-to-date compilers
Lineage continues:	



Sources: Paul Boutin; Brent Hailpern, associate director of computer science at IBM Research; The Retrocomputing Museum; Todd Proebsting, senior researcher at Microsoft; Gio Wiederhold, computer scientist, Stanford University

<https://ccrma.stanford.edu/courses/250a-fall-2005/docs/ComputerLanguagesChart.png>



Python的创始人为荷兰人吉多·范罗苏姆（Guido van Rossum）。1989年圣诞节期间，在阿姆斯特丹，Guido为了打发圣诞节的无趣，决心开发一个新的脚本解释程序，作为ABC语言的一种继承。之所以选中Python（大蟒蛇的意思）作为该编程语言的名字，是取自英国20世纪70年代首播的电视喜剧《蒙提·派森的飞行马戏团》（Monty Python's Flying Circus）。

——百度百科，Python

Python发展接近三十年，确实已经成为了编程语言中的“网红”。很多程序员都喜欢Python，但不仅仅是程序员喜欢，Python这个技能也能让你在就业市场上拿到很好的offer。因为Python仍旧是目前IT就业市场最受欢迎，最热门的技术技能之一，且容易上手，学会了python，可以大幅提高IT人的自身竞争力。

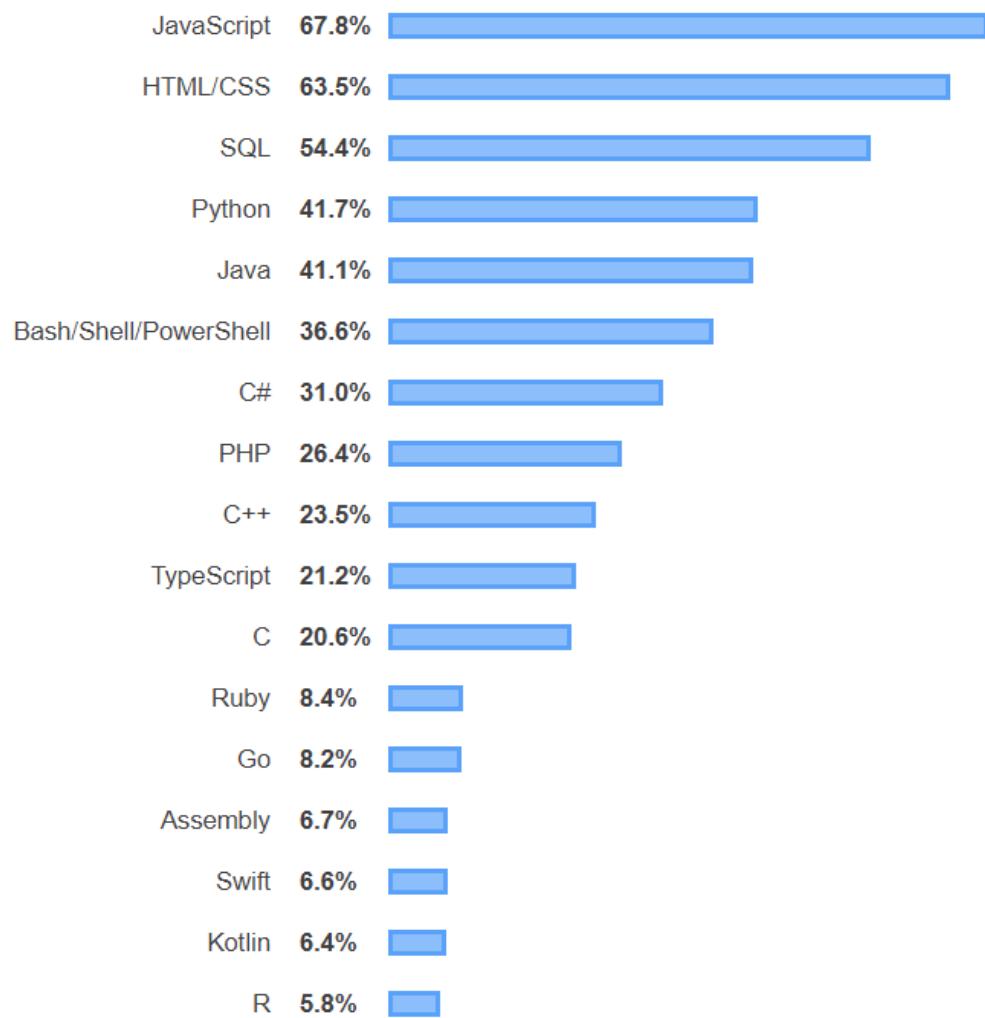


Most Popular Technologies

Programming, Scripting, and Markup Languages

All Respondents

Professional Developers



Python, the fastest-growing major programming language, has risen in the ranks of programming languages in our survey yet again, edging out Java this year and standing as the second most loved language (behind Rust).

<https://insights.stackoverflow.com/survey/2019>

Life is short, you need Python!

Python能成为如今的主流编程语言之一不是没有原因的。其中，最主要的原因大概有以下几点：

1.适合初学者

Python具有语法简单、语句清晰的特点，这就让初学者在学习阶段可以把精力集中在编程对象和思维方法上。

2.大佬都在用

Google, YouTube, Facebook, IBM, NASA, Yahoo, ACH, 和NECH只是技术领域中使用Python的几个大公司，它们也在不断招收Python工程师们。

3.应用超广泛

作为一种多才多艺的语言，从网站搭建到数据处理再到小工具小游戏的设计，都能用到Python。

4.人工智能必备

随着人工智能的兴起，Python作为一种科学语言的流行程度急剧上升。有许多机器学习库就是用Python编写的。



物理硬件性能显著提升，语言带来的性能影响趋向于减小，性能扩展，性能瓶颈用C/C++等来实现，暴露Python接口。Python相对较慢，但适合写上层逻辑；C/C++速度快，适合写底层算法

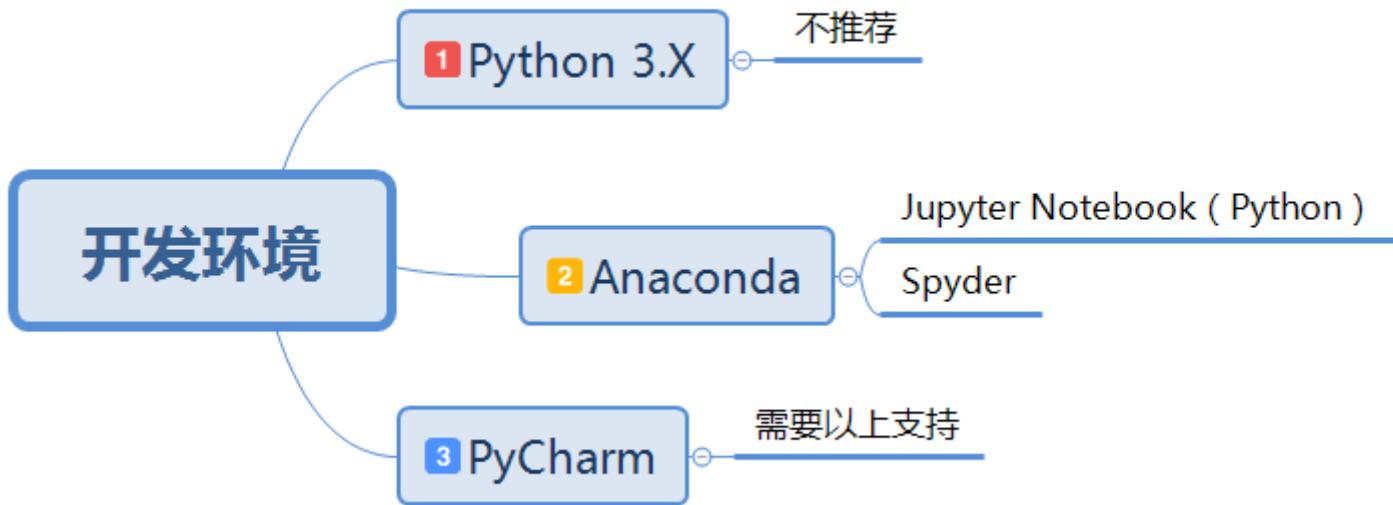


我该怎么上手？

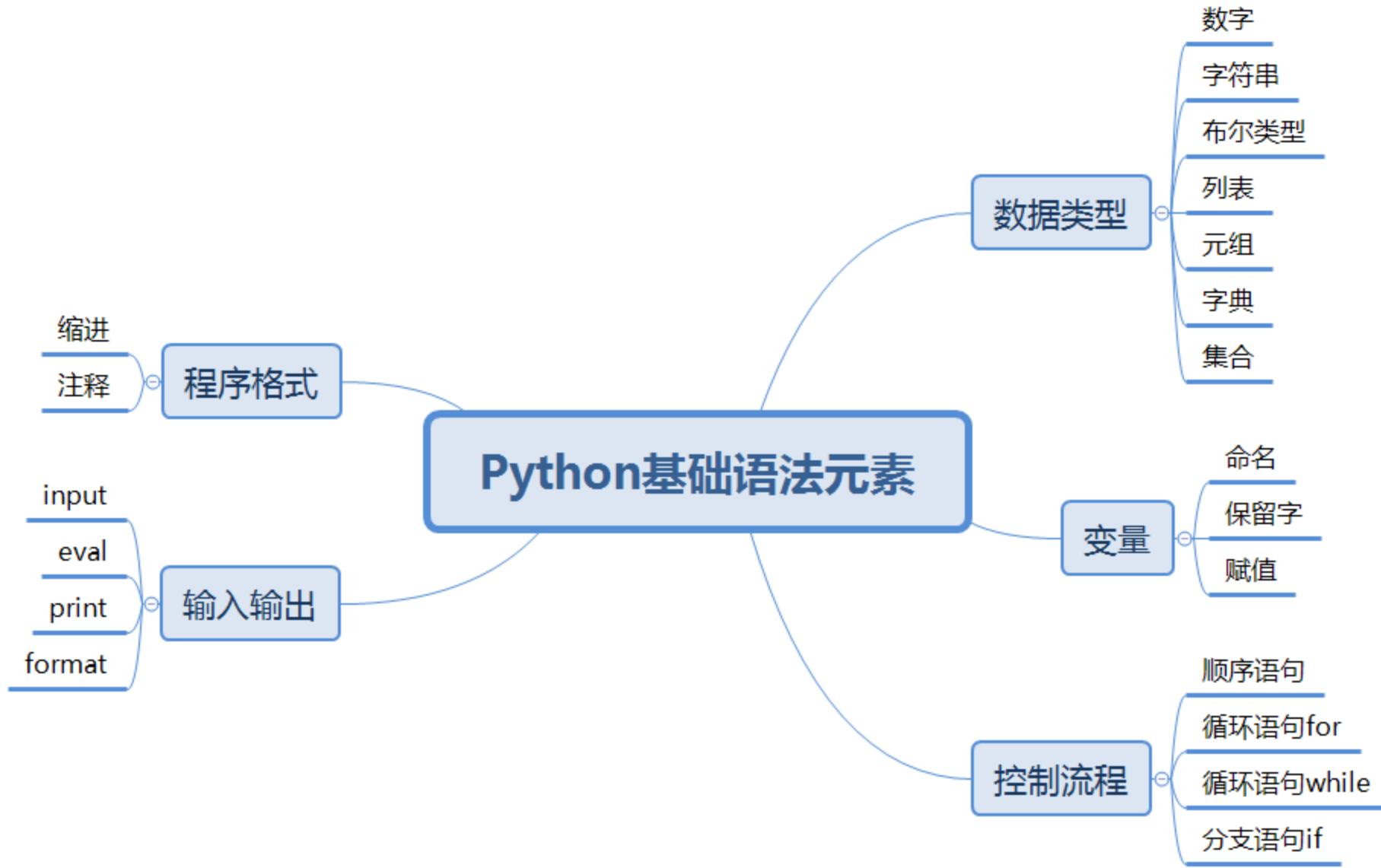
配置开
发环境

学习基
本语法

项目实
践上路



Python基本语法元素



：	<填充>	<对齐>	<宽度>	<, >	<.精度>	<类型>
引导符号	用于填充的字符 如* _等	<左对齐 >右对齐 ^居中对齐	设定输出的宽度	数字千分位分隔符 适用于整数和浮点数	浮点数：小数部分精度 字符串：最大输出长度	整数类型： b,c,d,o,x,X 浮点类型： e,E,f,%