



数据科学基础

郎 波

北京航空航天大学软件开发环境国家重点实验室

2020.2



课程基本信息

- 课程组织

- 利用PPT与相关材料自主学习+讨论

- 课程考核

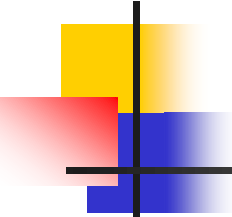
- 平时学习表现+大作业
- 大作业: report+presentation

- 教师

- Email: langbo@buaa.edu.cn
- Tel: 82317656

- 助教

- SY1906418 李坤浩



第1章 概述

- 数据科学基础内涵
- 课程架构
- 课程主要内容



数据

- 数据，产生于自然科学、社会科学以及其他领域的实验、观测和仿真等过程。例如， 图像，文本，生物学中的DNA序列等
- 数据是自然和生命的一种表示形式，数据还记录了人类的行为，包括工作、生活和社会发展
- 海量数据为理解人类所生存的世界提供了巨大的机遇



科学

- 科学是关于自然界、人类社会规律的事实、原理、方法和观念的**知识体系**以及**创建这个知识体系的社会活动**。科学的任务是发现规律，提出理论，认识世界，解释世界
 - 科学是人们研究自然、社会、思维的本质及其规律所获得的一种知识体系
 - 科学不仅是一种知识体系，它还是产生知识体系的一个活动，一个过程

科学发展的四个范式

■ Jim Gray将人类科学的发展定义成为四个“范式” (Paradigm)



Jim Gray

- 实验科学（实验归纳）：以记录和描述自然现象为主，案例如钻木取火
- 理论科学（模型推演）：简化实验模型，去掉一些复杂的干扰，只留下关键因素，然后通过演算进行归纳总结，案例如牛顿三定律、麦克斯韦方程组等
- 计算科学（仿真模拟）：对复杂现象进行模拟仿真，推演出越来越多复杂的现象，其典型案例如模拟核试验、天气预报等
- 数据科学（数据密集型科学发现）：随着数据量的高速增长，计算机将不仅仅能做模拟仿真，还能进行分析总结，得到理论。

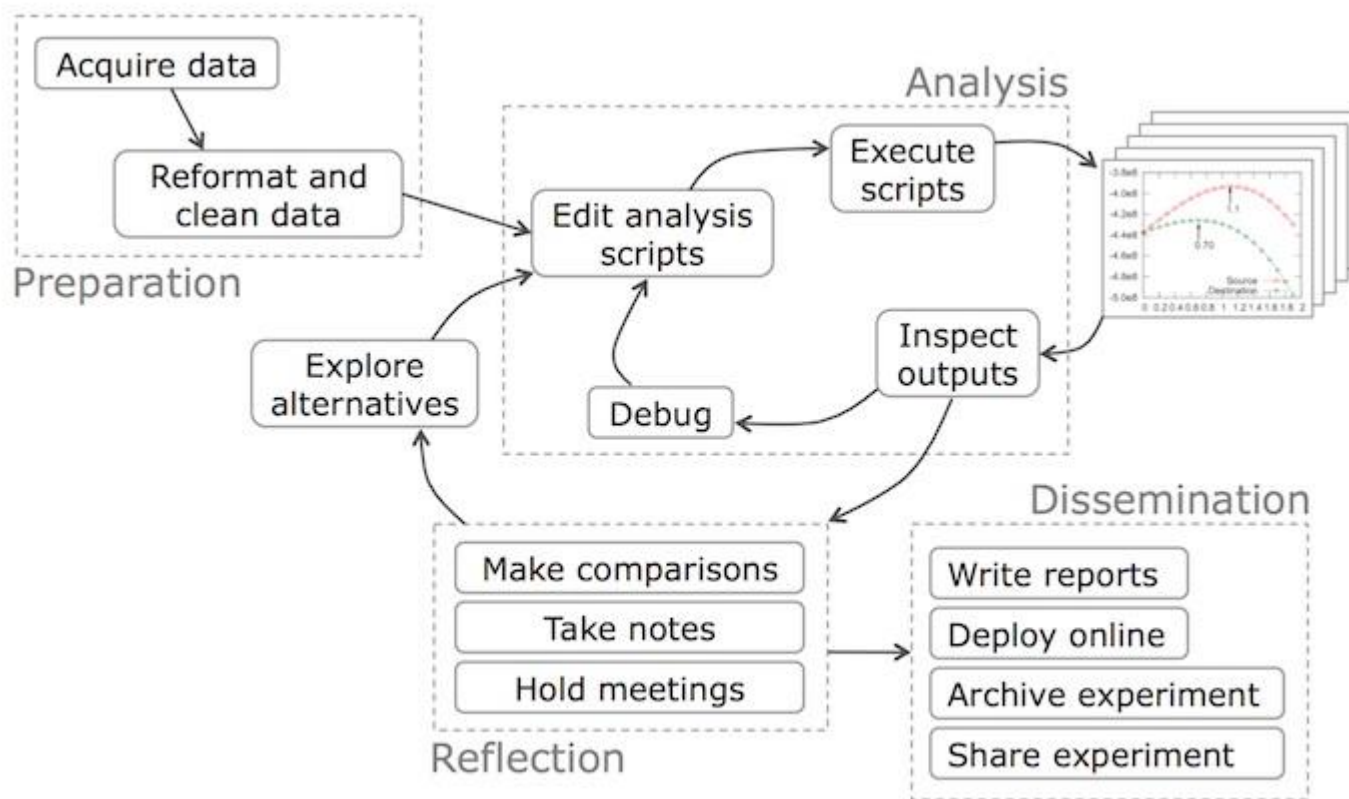


数据科学

■ 数据科学

- 数据科学是处理数据的科学，是通过数据计算揭示自然界和人类行为规律的科学，包含了处理数据的理论、方法与过程
- Data science is the study of the computational principles, methods, and systems for extracting knowledge from data

Data Science Workflow

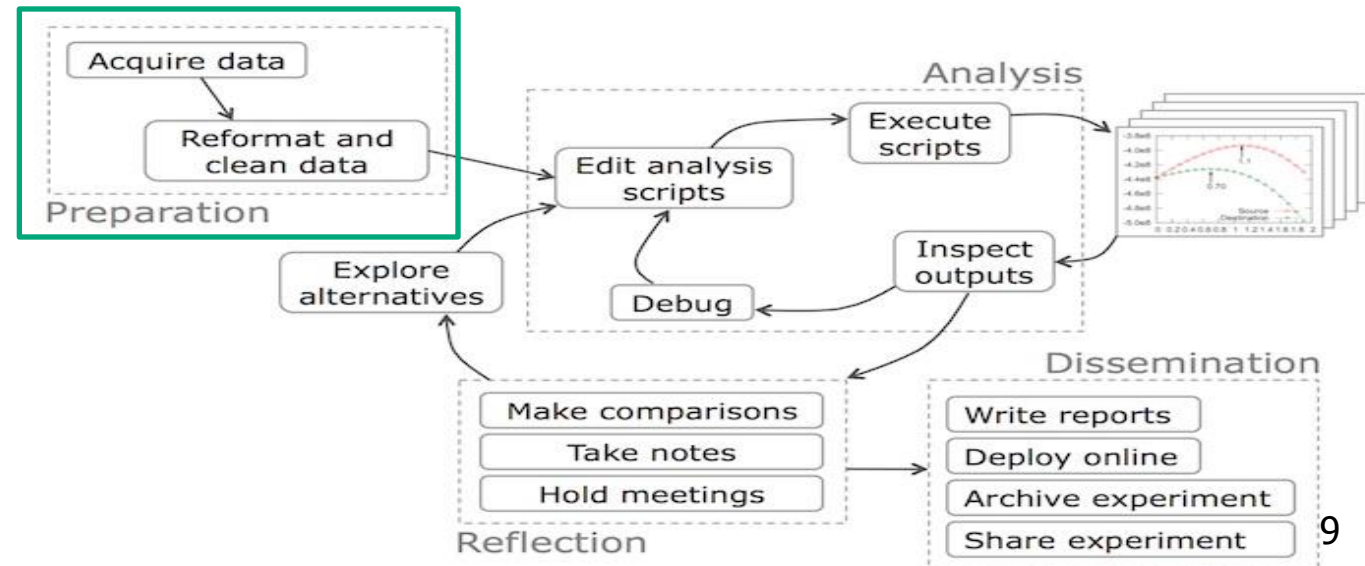


Philip Jia Guo, Software Tools to Facilitate Research Programming, Phd. Thesis, Stanford University , 2012.5,

Data Science Workflow

■ Preparation

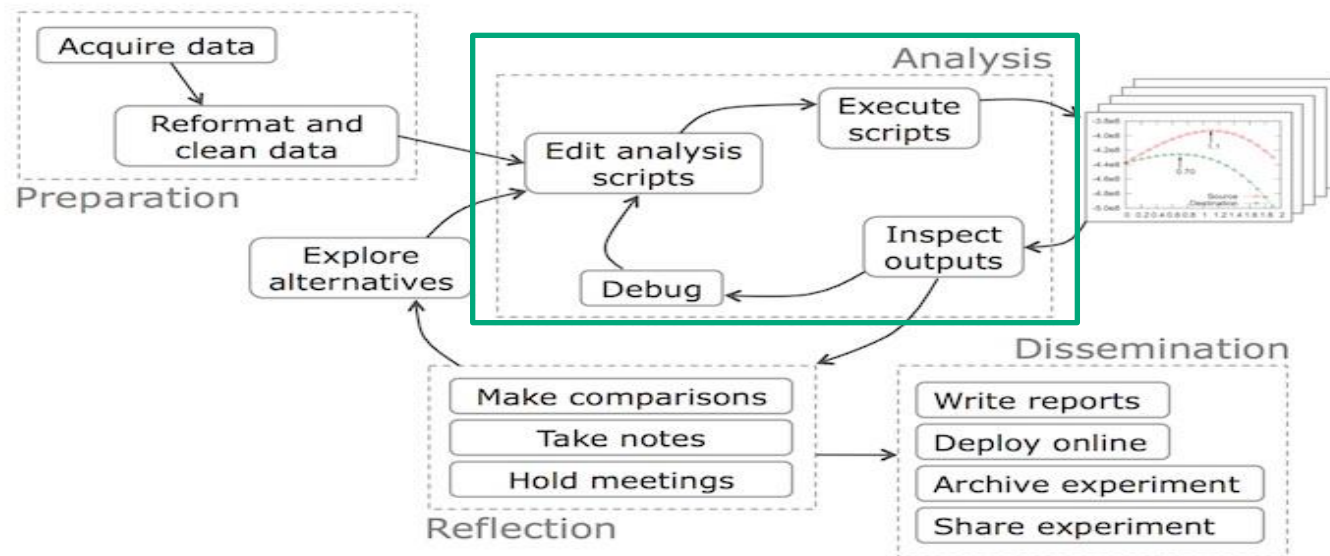
- Acquire data
- Reformat : change data into a form that is amenable to computation.
- Clean: remove noise, correct inconsistencies
- Data integration: merges data from multiple sources
- Data reduction: dimensionality reduction, numerosity reduction



Data Science Workflow

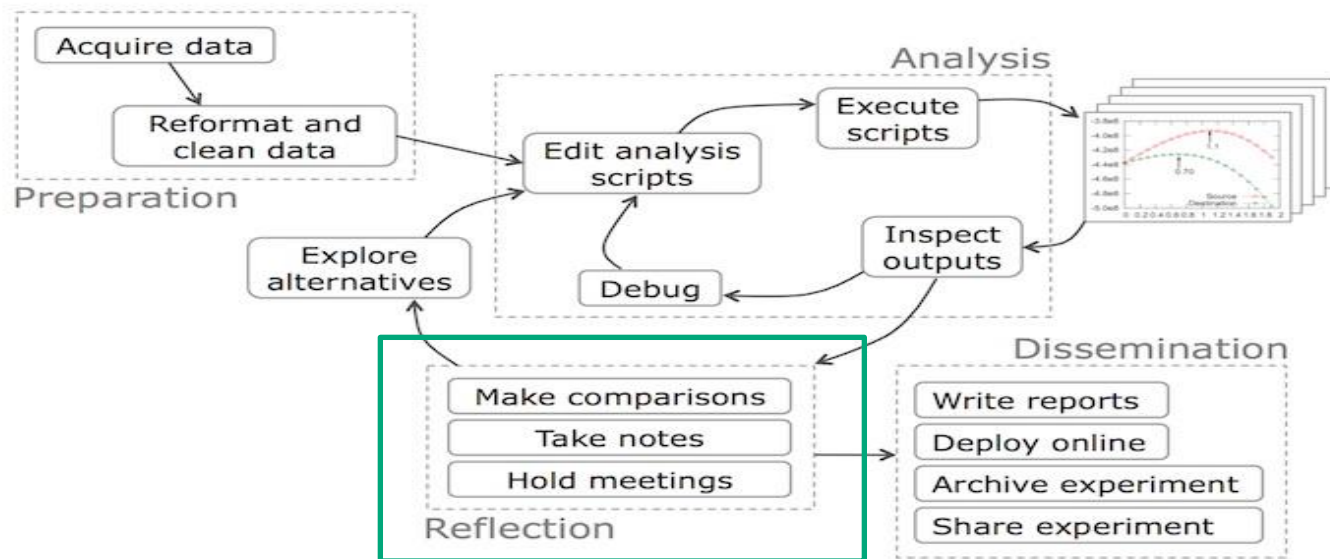
■ Analysis

- writing, executing, and refining computer programs to analyze and obtain insights from data.



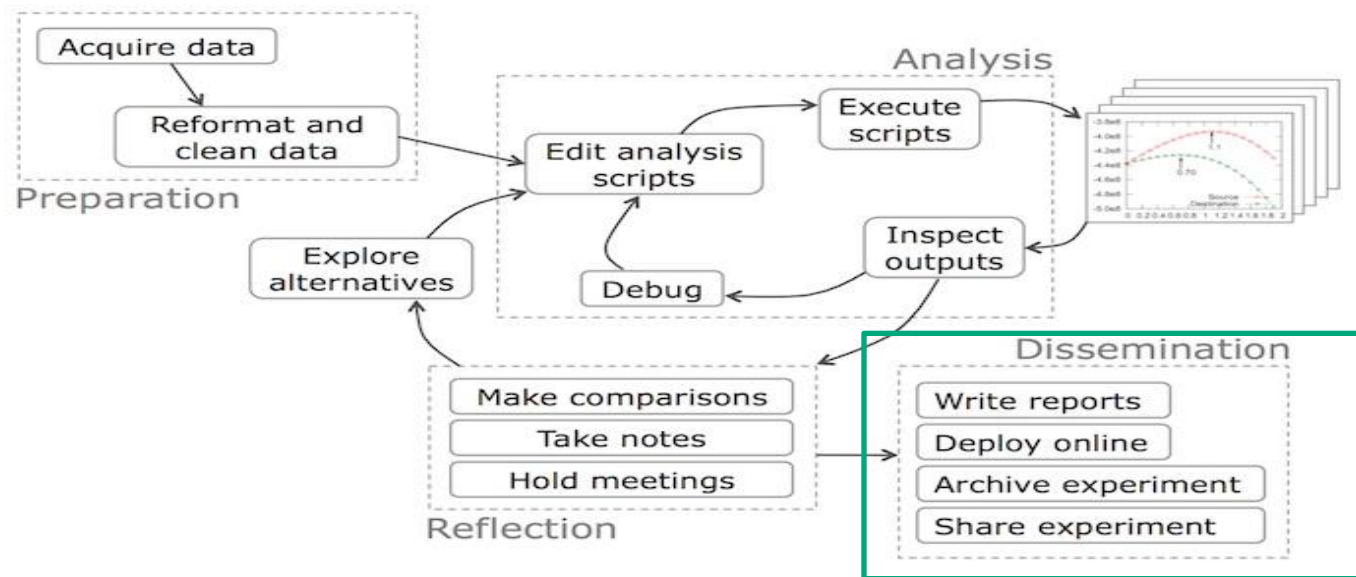
Data Science Workflow

- Reflection: thinking and communicating about the outputs of analyses.



Data Science Workflow

- Dissemination: disseminating results for decision making





Theoretical Foundations of Data Science

- **Data science is a broad discipline**, covering everything from the experimental design phase and the data collection stage, all the way to the data analysis process and the inevitable decision making phase at the very end of the “**data to knowledge to action**” paradigm.
- As such, **data science is likely to grow into a new discipline, with TFoDS being at its heart**, since theoretical foundations are of paramount importance in the development of any new scientific field.

---Workshop on “Theoretical Foundations of Data Science (TFoDS)” (NSF sponsored)



Theoretical Foundations of Data Science

- **TFoDS are necessary in all aspects of data science**, from the generation and collection of data to the analysis and the decision making processes.
- **TFoDS should have strong interfaces to application domains**. Algorithms developed in a vacuum for theoretical purposes only will typically fail to take into account the peculiarities and incompleteness properties of real data.



Theoretical Foundations of Data Science

- There are obvious **core areas that underpin data science**, include linear algebra and optimization (**Math**), programming languages, data structures and complexity theory (**CSE**), information theory and signal processing (**ECE**), probabilistic models, statistical inference (**Stat**).
- TFoDS will be intrinsically inter-disciplinary



数据科学基础

■ 数据科学基础——课程切入点

- 数据科学是**以数据为对象**，利用数据处理的理论、方法与技术**进行数据计算**，获得自然界和人类行为的**规律**
- **数据计算**是在数据对象上执行算法的过程，是发现规律的关键
- 数据的特性与表达，数据计算空间的构成与特性，计算空间的基本变换（计算），数据计算理论与方法的基本思想等，是**数据科学基础**的重要组成部分



数据模型与数据空间

- 数据的表达即数据模型
 - 把数据编码成某种数学对象，使人们能够在其上高效执行某些计算，并获得对于数据产生领域有用的答案
- 数据表达形式或数据模型的结构，是数据计算驱动的
- 数据科学中的数据空间，由数据对象构成，数据的表达形式决定了的数据空间的形态



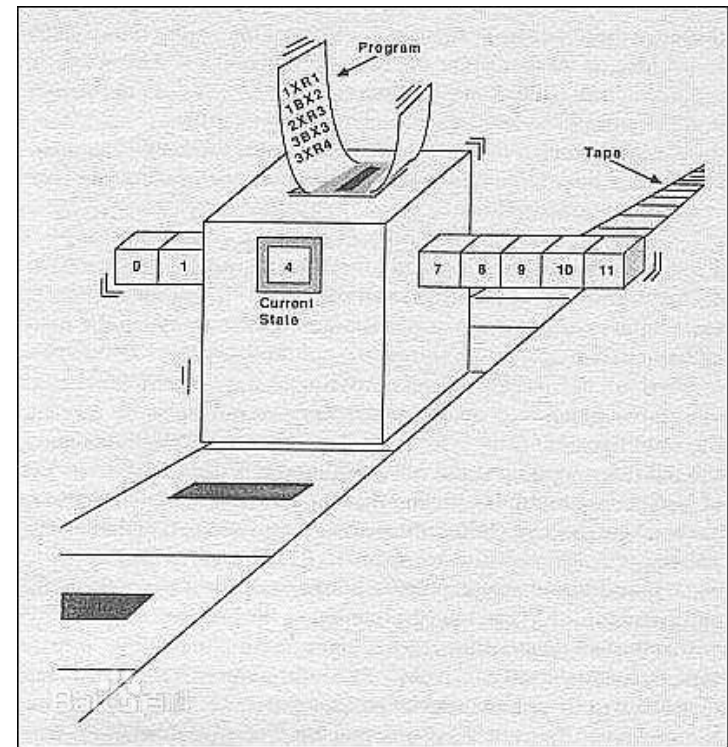
数据模型

- 数据模型需要满足的条件

- 数据表示（数据结构）要**具有充分的灵活性（柔性）**，能够扑捉到数据有意义的特性
- 数据表示（数据结构）**要充分的结构化**，使得计算能够在数据上实施，并且能够在合理的时间内得到有价值的结果

数据模型示例

- The tape in Turing machine
- Strings
- Database table





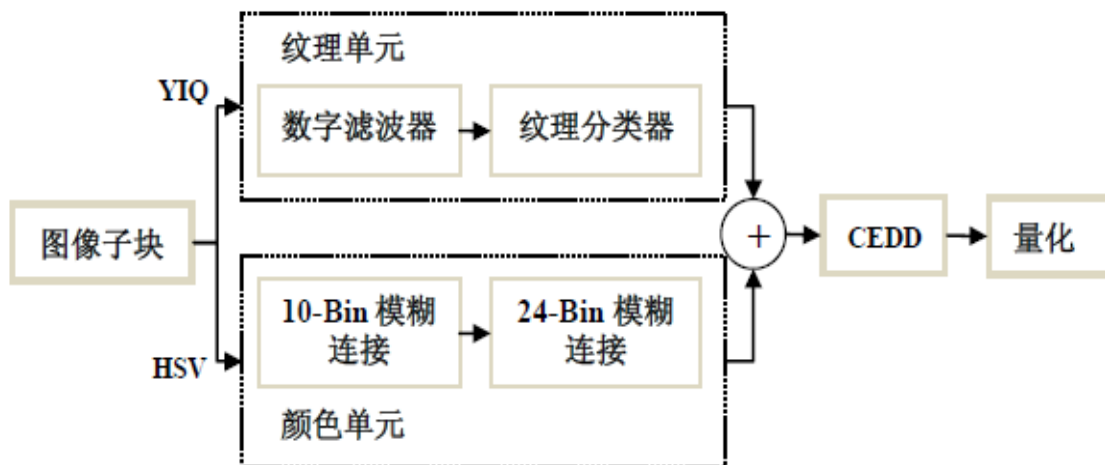
数据科学中的数据对象表示

- 特征是一个数据对象区别于其他对象特点或者特性。
- 将一类对象的多个或者多种特性结合在一起，形成一个特征向量来代表该类对象，如果是n个特性的组合，则为一个n维特征向量。

$$X = (x_1, x_2, x_3, \dots, x_n)$$

数据特征示例

- 图像特征，包括色彩、边缘、纹理、直方图等
- 图像的CEDD特征，采取颜色信息和纹理信息相结合的方式，在得到表示图像纹理信息的6维直方图后，每一维都加入24维颜色信息，最终结合颜色和纹理特征，构成144维的特征向量



两种重要的数据模型

■ 矩阵

- m 个数据要进行相关性或相似性计算，每个数据由一个 n 维特征向量描述，则 $m \times n$ 矩阵可以表达这个数据集。

- 例如 term-document matrix:

D1 = "I like databases"

D2 = "I hate databases"

	I	like	hate	databases
D1	1	1	0	1
D2	1	0	1	1

■ 图

- 描述数据对象之间的交互关系。假设有一个对象或实体的集合 V ，将有一个集合 E 描述实体间的两两交互关系，例如 web 网页的链接关系



数据科学中的数据模型与数据空间

- 数据模型
 - 向量Vectors
 - 矩阵Matrices
 - 图Graphs
- 这些数据模型可以描述最普通应用领域中的数据，并且是描述的灵活性与算法数据结构之间的最佳权衡
- 数据空间
 - 由数据特征向量构成的高维空间



数据特性与数据空间

- 数据特性分析与向量构造
- 高维空间的形态与性质
- 子空间与空间的线性变换
- 随机图与随机图上的计算



计算理论与方法的基本思想

- 机器学习

- 机器学习中的基本问题
- 过拟合和一致收敛、正则化 (regularization)
- 在线学习
- 核函数
- BOOSTING
- 优化问题

- 大数据算法



课程主要内容

- High-Dimensional Space
 - The Law of Large Numbers, The Geometry of High Dimensions, Properties of the Unit Ball, Random Projection and Johnson-Lindenstrauss Lemma
- Best-Fit Subspaces and Singular Value Decomposition (SVD)
 - Singular Vectors and Eigenvectors, Singular Value Decomposition (SVD), Best Rank-k Approximations, Applications of Singular Value Decomposition



课程主要内容

- Random Graphs
 - Random graph Model, Phase Transitions, The Giant Component, Growth Model
- Random Walks and Markov Chains
 - Stationary Distribution, Markov Chain Monte Carlo, Convergence of Random Walks



课程主要内容

- Machine Learning

- The core problem of machine learning, Overfitting and Uniform Convergence, Regularization, VC-Dimension, Stochastic Gradient Descent
- Deep learning, Clustering, Semi-supervised learning, Active learning, Multi-task learning



课程主要内容

- Algorithms for Massive Data Problems
 - Frequency Moments of Data Streams, Matrix Algorithms using Sampling
- Data property and Data processing
 - Data Attribute Types, Statistical Descriptions of Data, Measuring Data Similarity and Dissimilarity
 - Data quality, Data cleaning, Data integration, Data reduction

课程教材

■ 教材与参考书

- 教材：Avrim Blum, John Hopcroft and Ravindran Kannan, Foundations of Data Science, 2016.1
- 参考书：Jiawei Han, Micheline Kamber, Jian Pei, Data Mining: Concepts and Techniques, 3rd Edition

John E. Hopcroft



Computer Science Department
Cornell University
426 Gates Hall
Ithaca, NY 14853
jeh at cs dot cornell dot edu
(607) 255-1179