大数据分析与应用

2020 春

Chapter 0 课程介绍——学习意义

当前互联网数据的特点（3V）

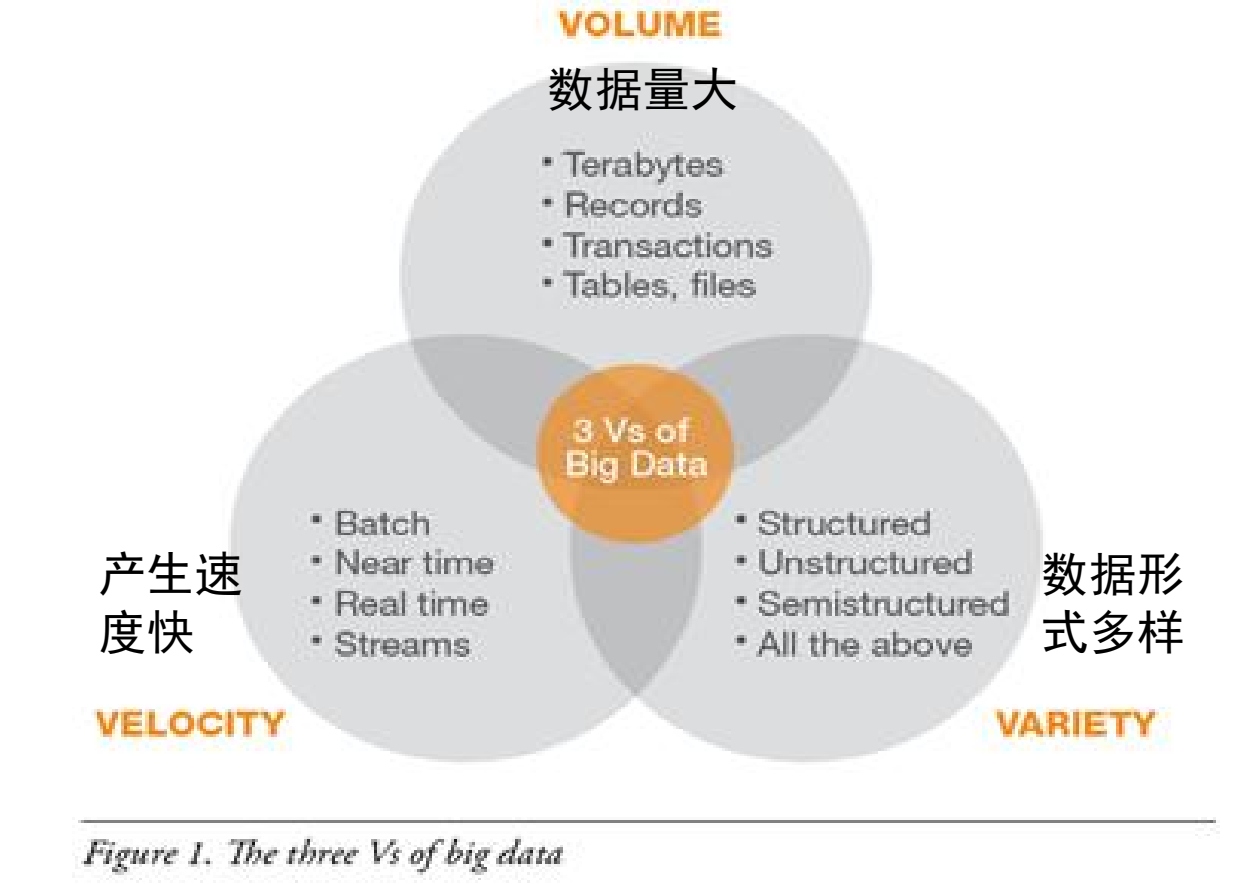


Figure 1. The three Vs of big data

面临的问题

- 从计算机的角度：单计算机单数据库无法处理；1）数据量；2）读写速度；3）非结构化；
- 从使用者角度：超过人类的认知能力，需要从中提取更高层次的知识；举例：推荐系统；

Chapter 0 课程介绍——学习意义

从数据中获取知识的场景：

- 商品推荐：从商品销售数据中预测用户购买行为，及时推荐商品
- 舆情分析：从文本数据中提取公众对政策的意见与反应
- 声音识别：从语音数据中得到对应的文字制作字幕
- 趋势预测：从用电量数据中预测未来的用电量得到发电计划
- 自动决策：从传感器数据中得到自动驾驶下一秒的行为
- 客户细分：从消费数据中得到客户的偏好，区分不同类型的客户
- 。。。

Chapter 0 课程介绍——一般规律

1. 提出要求定义问题
2. 收集数据
3. 从数据中获得规律
4. 利用规律解决问题

■学习目标

✓面向机器：掌握分布式数据计算平台及其编程初步：1）Spark；2）函数式编程

✓面向数据：数据挖掘算法

✓面向业务：推荐系统，基本方法在具体实践中的应用

课程目录

- 1.数据预处理（1）
- 2.分布式计算简介（2）
- 3.关联分析（3）
- 4.推荐系统I（4）
- 5.分类方法（5-7）
- 6.回归方法（8）
- 7.聚类方法（9-10）
- 8.推荐系统II（11-12）
- 9.案例与总结（13）

- 安装单机开发平台（0）
- 实验1.分布式计算平台（14）
- 实验2.关联规则方法（14）
- 实验3.分类与回归方法（15）
- 实验4.聚类方法（15）
- 实验5.推荐系统（16）

- 课程设计 （<=4人/组）

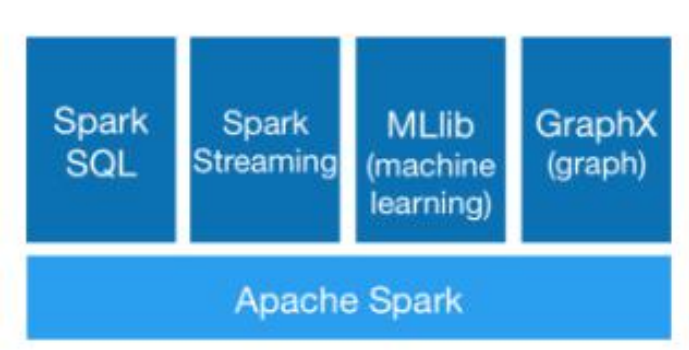
本课程使用的软件

■Python：计算机语言
■Anaconda：基于Python的科学计算平台

- 下载提示：搜索anaconda 清华镜像 选择对应操作系统的版本
- https://mirrors.tuna.tsinghua.edu.cn/anaconda/archive/

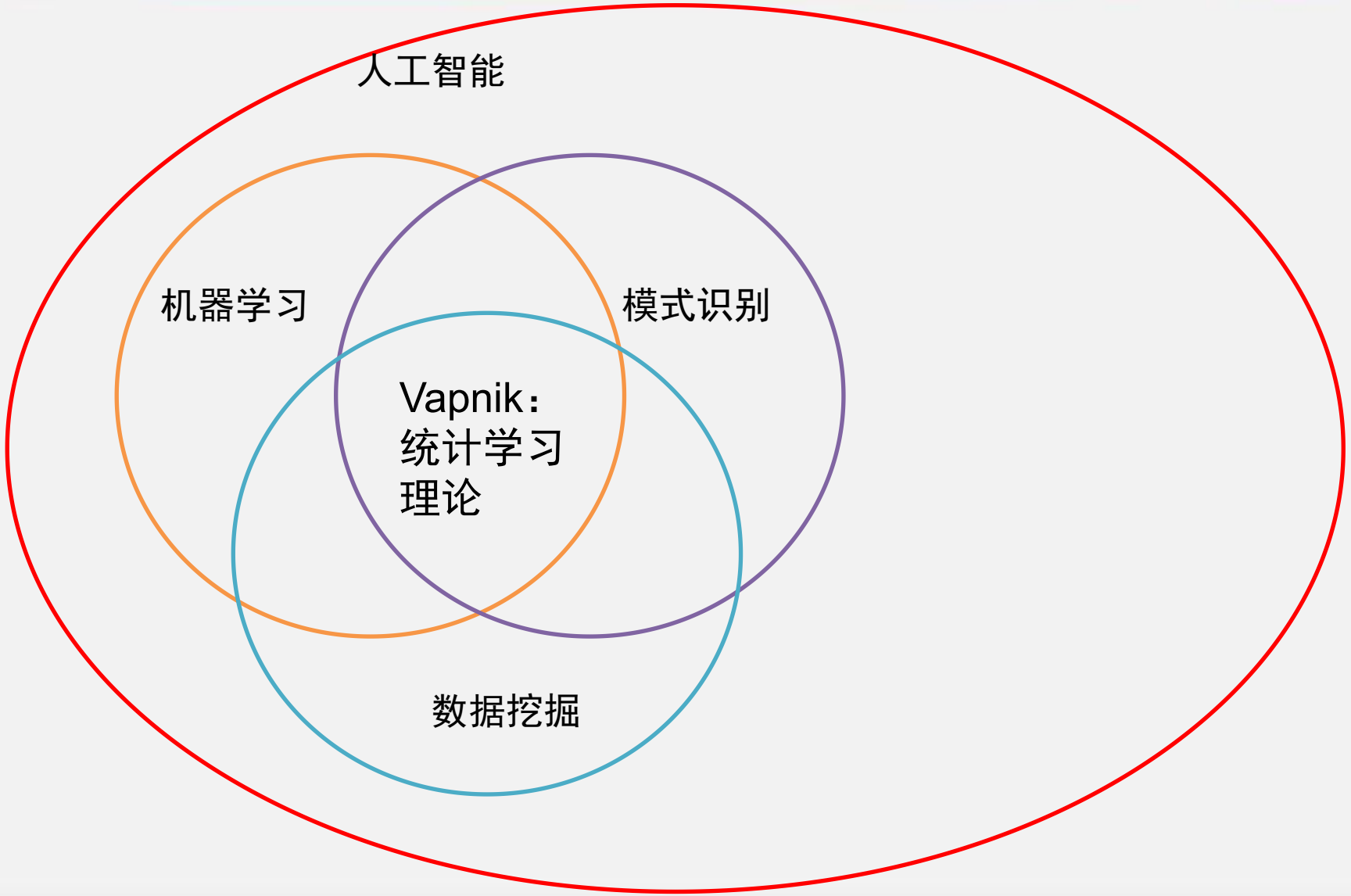
■生产平台

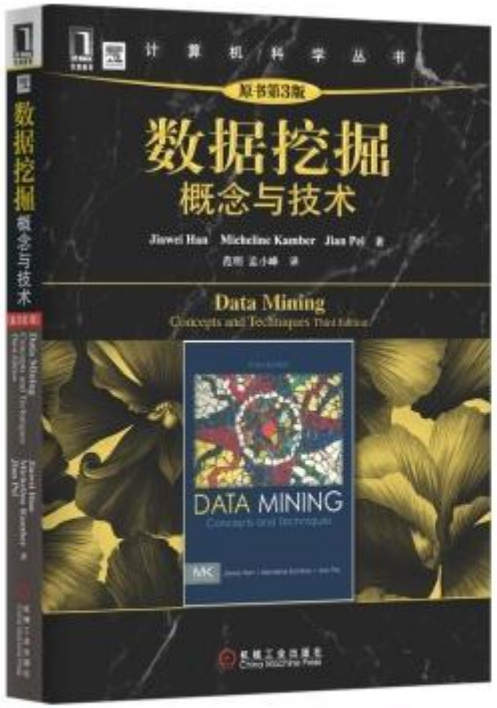
- Linux+Spark+pyspark



几个学科之间的联系

- 数据挖掘更注重在现有数据库系统所承载的数据基础上的挖掘（计算机学会数据库与信息检索分会）
- 模式识别倾向于对自然数据（语音图像）的处理(自动化学会)
- 机器学习是人工智能的子学科（计算机学会人工智能分会）








企业批量购书

数据挖掘 概念与技术（原书第3版）[Data Mining Concepts and Techniques Third Edition]

数据挖掘领域里程碑意义的经典著作！中文版、影印版同步上市！决战大数据时代！IT技术人员不得不读

[美] Jiawei Han，[美] Micheling Kamber，[美] Jian Pei 等 著；范明，孟小峰 译

京 东 价：**¥54.90** [7折] [定价：~~¥79.00~~] (降价通知)

累计评价 3452

领　　券：券 100-5 200-16

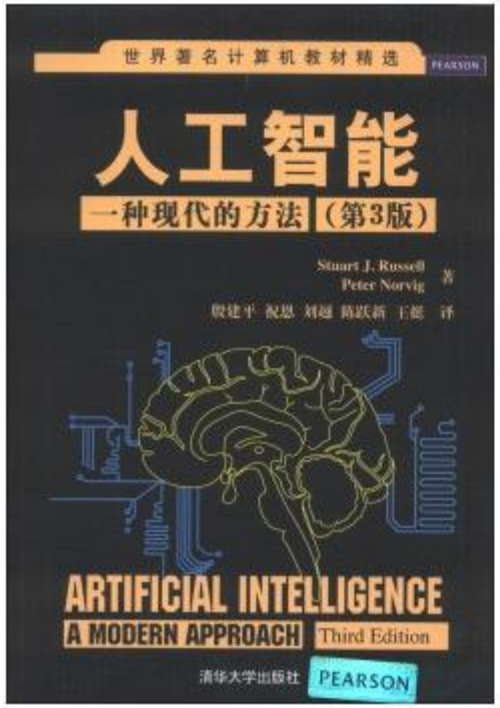
排　　名：自营 计算机与互联网销量榜 第 80 位

配 送 至：浙江杭州市西湖区 **有货**，支持 79免运费 | 货到付款

商品好评度高于行业均值

服　　务：由 京东 发货并提供售后服务。23:00前完成下单,预计**明天(02月23日)**送达

选择系列：概念与技术（原书第3版）　技术与工程实践



世界著名计算机教材精选·人工智能：一种现代的方法（第3版）[Artificial Intelligence: a Modern Approach, Third Edition]

[美]罗素（Stuart J.Russell），[美]诺维格（Peter Norvig）著；殷建平，祝恩，刘越 等 译

京 东 价：**¥102.20** [8折] [定价：~~¥128.00~~] (降价通知)

累计评价 296

领　　券：券 100-5 200-16

排　　名：自营 大中专教材教辅销量榜 第 295 位

配 送 至：浙江杭州市西湖区 **有货**，支持 79免运费 | 货到付款

商品好评度低于行业均值

服　　务：由 京东 发货并提供售后服务。23:00前完成下单,预计**明天(02月23日)**送达

白条分期：30天免息　¥34.58×3期　¥17.55×6期　¥9.03×12期　¥4.77×24期





机器学习【首届京东文学奖-年度新锐入围作品】

周志华 著

京 东 价 ¥88.00 [10折] [定价 ~~¥88.00~~] (降价通知)

促销信息 跨自营/店铺满减 每满100元，可减50元现金 详情 >>

跨自营/店铺满减 每满100元，可减30元现金 详情 >>

以上促销可在购物车任选其一

增值业务 礼品包装

排　　名 自营 计算机与互联网销量榜 第 3 位

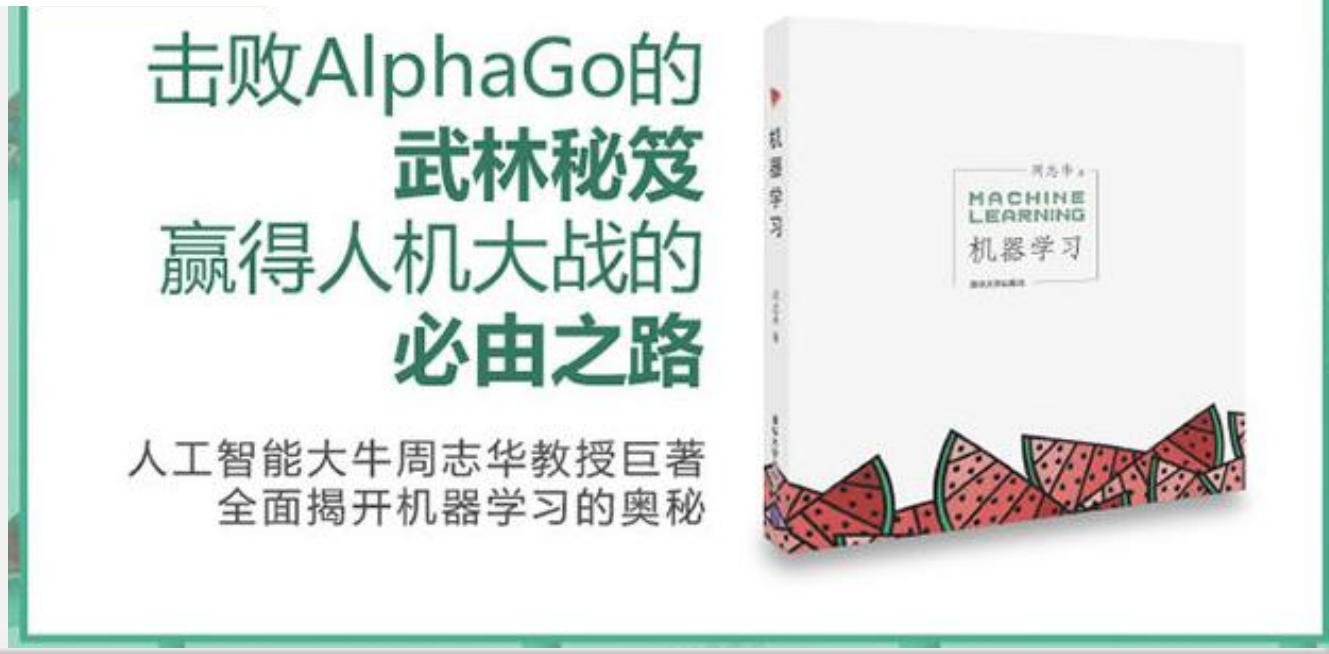
配 送 至 北京朝阳区三环以内 有货

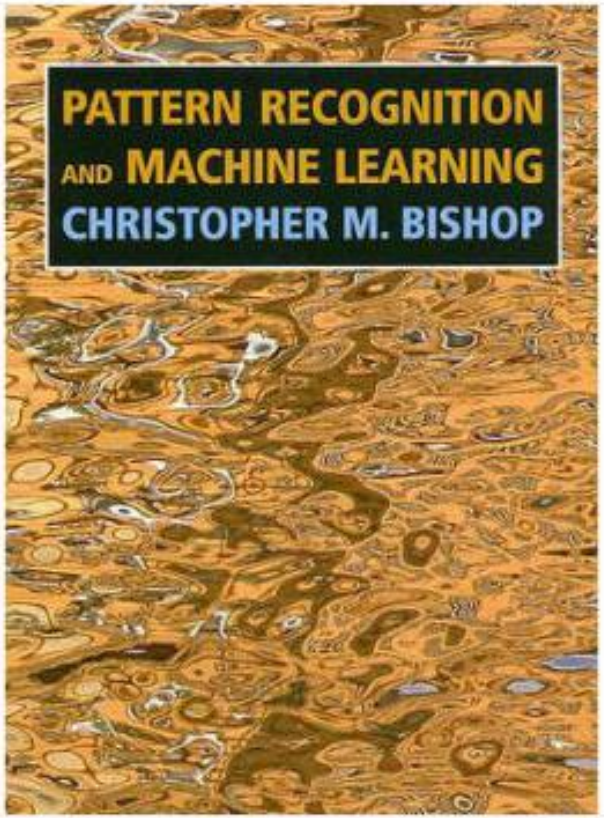
由 京东 发货，并提供售后服务. 23:00前下单,预计明天(03月01日)送达

重　　量 0.92kg

服务支持 自营放心购 破损包退换 闪电退款 上门换新

49元免基础运费(50kg内) 京准达 自提





模式识别与机器学习 Pattern Recognition and Machine...

原版新书-上海现货（1-2个工作日内从上海由京东快递发出）澜瑞外文全场免运费！

Christopher M Bishop & 著

京东价 ¥798.00 (降价通知)

优惠券 满588减20 满288减10

促销信息 会员特价 请登录 确认是否享受优惠

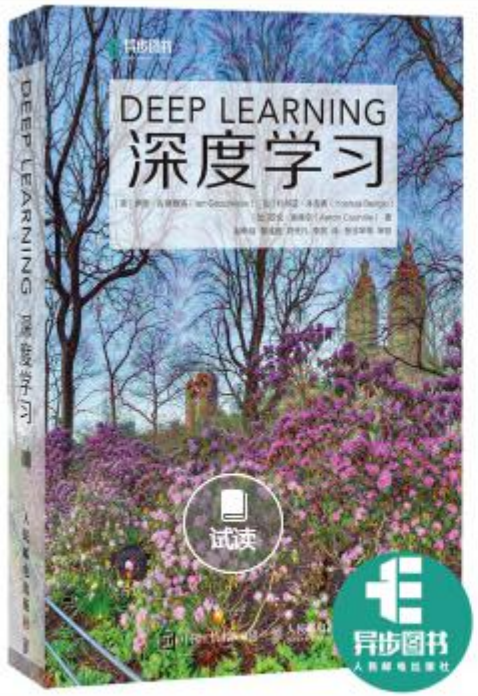
配送至 北京朝阳区三环以内 有货 支持 送运费险 | 闪电退款 免运费

由 澜瑞外文Lanree图书专营店 负责发货，并提供售后服务.

白条分期 不分期 ¥269.65起×3期 ¥136.84起×6期 ¥70.37起×12期 ¥37.16起×24期 惠

1 + -

加入购物车







深度学习 [deep learning]

AI圣经 Deep Learning中文版 长期位居美国ya马逊人工智能和机器学习类图书榜首 深度学习领域奠基性的经典畅销书 特斯拉CEO埃隆·马斯克等国内外众多专家推荐

[美] Ian，Goodfellow，[加] Yoshua，Bengio，[加] Aaron … 著

京 东 价 ¥134.00 [8折] [定价 ~~¥168.00~~] (降价通知)

累计评价 6.1万+

促销信息 加价购 满20元另加24.90元，或满30元另加19.90元，即可在购物车换购热销商品 详情 >>

增值业务 礼品包装

排　　名 自营 计算机与互联网销量榜 第 22 位

配 送 至 北京朝阳区三环以内 有货

由 京东 发货，并提供售后服务. 23:00前下单,预计明天(03月01日)送达

重　　量 1.055kg

服务支持 自营放心购 破损包退换 闪电退款 上门换新

49元免基础运费(50kg内) 京准达 自提

Data Scientist: Hottest Job in Next Decade

The ability to take data—to be able to **understand** it, to **process** it, to **extract value** from it, to **visualize** it, to **communicate** it—that's going to be a hugely important skill in the next decades. So the complimentary scarce factor is the ability to understand that data and extract value from it.

Hal Varian, Google's Chief Economist
The McKinsey Quarterly, Jan 2009