

# 数据结构与算法 (Python)

## 课程引言

谢正茂 [webg@pku.edu.cn](mailto:webg@pku.edu.cn)



2021春季数算B-2班

北京大学计算机系网络所

March 13, 2021



Valid until 3/16 and will update upon joining group

- 课程定位
- 编程语言和教材的选择
- 时代背景：大数据的来龙去脉

- 这门课讲些什么，讲到什么程度？
- 目标：希望同学们学完之后有哪些收获
  - 写程序处理数据、分析数据的能力，能把“大数据”工具用在各自的领域。
  - 进一步学习人工智能/机器学习，很好的起点。
- 举例说明：这门课对我今后的哪些工作有帮助

# 课程主要内容

## ● 课程基本框架

- Python 入门
- 算法基本概念, 算法复杂度
- 线性表
- 递归与动态规划
- KMP 算法
- 排序与查找
- 树及算法
- 图及算法

● 比《计算概论》讲的窄、但更加深入算法。

● 比《数算 A》课时少: 内容少、更容易。  
《数算 A》专门有《程序设计实习》课程配套。

● 供计算机以外学院（简称“外院”）同学学习的  
算法课程。

● 教材、参考书。

- 基本沿用教材的脉络讲, 少数难点使用参考  
书中的内容。



# “算法”课程从限选变成必修

- 需求激增，选课人数翻番！
- 近年来发生的大数据、互联网 +，给这个领域带来了大量的需求和工作岗位。
- 大学之后找工作/创业都是不错的选项。
  - 除研发类的少数岗位以外，“外院”的同学都能胜任。
  - 算法类、产品，“外院”同学还有某些优势。
- 做研究：大数据时代背景下，数据驱动的研究。
  - 在大数据的时代背景下，不仅是“理工科”，“文科”也需要用数据说话。
  - 最难想象的是，在“考古领域”大数据都能带来重大的发现<sup>a</sup>。
  - 写程序获取、处理、分析数据成为“文理工”共同的基础技能。
- 申请计算机相关专业，有一定帮助。



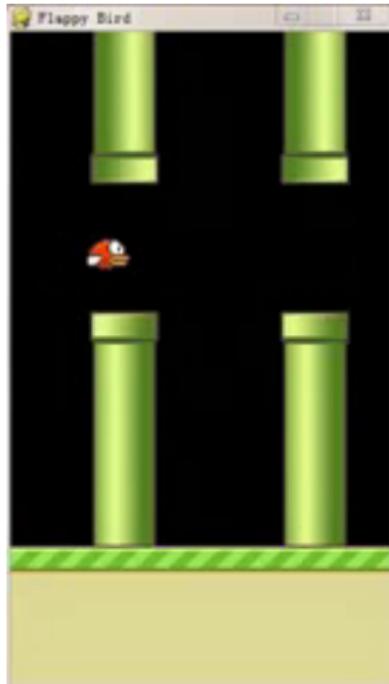
- 程序设计语言是对数据结构和算法的表达。
- PASCAL 语言 ==> C/C++, JAVA, Python
- 高级语言 → 汇编语言 → 机器语言
- 高级语言里面也分谁更“高级”，高级是抽象的层次。
  - “高级” 与否无关好坏，根据具体的工作进行选择。
  - 每种语言都有自己流行的领域。在一个领域越流行，开源的资源越多。
  - 抽象的层次高，屏蔽了底层的细节，开发起来省力；但有些功能实现不了。
  - 如果计算机专业的话，最好对 C/C++ 有所了解。
  - 如果从事硬件开发的化，需要了解汇编。

# 为什么选 Python?

- 代码短小精悍，干净整洁
  - 没有变量声明，不需要花括号 begin/end，也没有分号，比 java 短 80%，比 C 短 98%
- 解释执行，上手就玩，编程小白福音
  - 不用焚香沐浴安装 GB 级别的开发环境 compile/build，可以随问秒答，边玩边改
- “包装内附带电池”
  - 自带大量运行库，网络、数据库、图形图像、GUI、压缩加密一应俱全，几行代码建网站
- 功能无比强大，开发左右逢源，最酷的网络应用都是用它
  - Google/Youtube/Instagram/豆瓣……；NASA 也用它哦
- 搞大数据和 AI 的人们也爱它
  - 有各种面向大数据处理的数据模型、数值分析、机器学习、空间分析等 Python 工具随时恭候

# Python 坐稳人工智能时代的头牌语言

- 机器学习“全家桶”scikit-learn
- Google 开源的 AI 系统 Tensorflow
- Python 可以调用 C++ 的代码
- 160 行 Python 代码可以让 AI 从游戏视频中学习玩 Flappybird

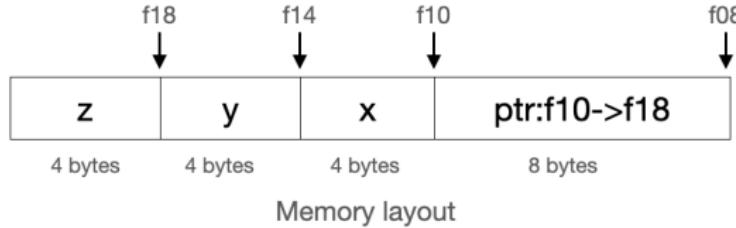


# C 与 Python 的“非全面”比较：关于指针

```
1 #include <stdio.h>
2
3 int main()
4 {
5     int z, y, x;
6     int *ptr = &x;
7     printf("ptr is %p\n", ptr);
8     printf("ptr's address is %p\n", &ptr);
9     printf("sizeof(x) is %lu\n", sizeof(x));
10    printf("sizeof(ptr) is %lu\n", sizeof(ptr));
11    *ptr++ = 5;
12    *ptr++ = 6;
13    *ptr = 7;
14    printf("ptr is %p\n", ptr);
15    printf("ptr's address is %p\n", &ptr);
16    printf("x = %d\n", x);
17    printf("y = %d\n", y);
18    printf("z = %d\n", z);
19
20 }
```

```
(base) rmbp13:code xiezhengmao$ ./pointer
ptr is 0x7ffecbc8f10
ptr's address is 0x7ffecbc8f08
sizeof(x) is 4
sizeof(ptr) is 8
ptr is 0x7ffecbc8f18
ptr's address is 0x7ffecbc8f08
x = 5
y = 6
z = 7
```

- 11-13， 绕过变量直接访问内存
- 在地址空间中可以随意移动
- 把（虚拟）内存暴露给程序员，因而支持一些强大的功能。
- 同时也是坑人王！包揽 95% 的程序 bug
- 地址越界、野指针、程序后门
- 强制人像机器一样考虑问题



# 一个常见的 bug

```
diff --git a/main.c b/main.c
index b6742a9..c5ae788 100644
--- a/main.c
+++ b/main.c
@@ -4,7 +4,7 @@ const char* foo = "David";

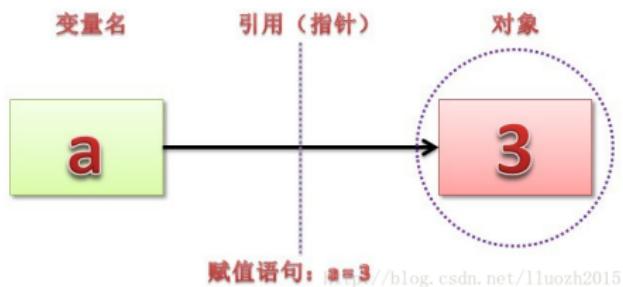
int main()
{
-    char* name = (char*) malloc (strlen(foo));
+    char* name = (char*) malloc (strlen(foo) + 1);
        strcpy(name, foo);
        return 0;
}
```

- `strcpy()` functions copy the string `src` to `dst` (including the terminating '\0' character.)
- '\0' 写在了没有分配的空间上
- 地址越界 `bug`: 不一定马上出错/在不同的平台上表现不一样
- 最难找的 `bug`: 潜伏期长, `bug` 和程序 `crash` 在不同地方
- 专门为他开发 `debug` 工具: <https://dmalloc.com/>
- 用起“指针”来如履薄冰

# C 与 Python 的“非全面”比较：指针的使用

- 争论：Python 中有没有指针？
- 感觉不到指针 vs. 所有可赋值的东西都是指针
- 变量只是指针/引用/标签：identity of object
- 封装：把对指针的直接操作藏起来（傻瓜化）
- id() 实际上就是内存地址，但语言极力避免用户直接操作地址。

```
>>> l=[1, 2, 3]
>>> ll=l
>>> ll[1]='David'
>>> l
[1, 'David', 3]
>>> id(l), id(ll)
(140381553307392, 140381553307392)
>>> import _ctypes
>>> _ctypes.PyObj_FromPtr(140381553307392)
[1, 'David', 3]
>>> hex(id(l))
'0x7fad209e2f00'
```



# Python 哲学

- rule 3: Simple is better than complex.
- rule 7: Readability counts.
- 把所有东西都拿出来 vs. 尽可能的“藏”起来
- 专注于问题本身，而少管与“计算机”相关的东西。
- 与之相反，“体系结构”专业专门研究与“计算机”相关的东西。
  - 从逻辑门开始，研究寄存器、加法器、乘法器、指令集、CPU、内存、南桥北桥、硬盘，.....
  - 体系机构教研室，开发了国产处理器：北大众志；军用保密芯片。



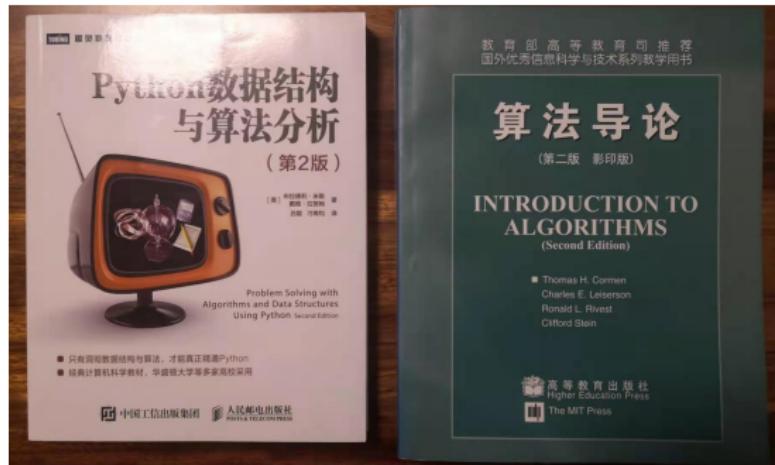
# 为什么 C 会那样“不友好”？

- 与底层硬件 (CPU) 靠的最近：C 代码和汇编指令有简单的对应关系，可以手工翻译。  
比如说：指针操作、程序跳转 *Goto* 语句
- 新的硬件出来，最先支持的高级语言肯定是 C
  - 硬件：奔腾、AMD、高通枭龙、华为麒麟、苹果 M1、北大众志、龙芯
  - 指令集：x86、MIPS、Sparc、Alpha、ARM
- Python 语言也是由 C 实现的<sup>1</sup>。
- Windows、MacOS、Linux 的内核都主要用 C 实现的
- 隔壁班同学提问：既然 C++ 包括了 C，我们为什么不直接讲 C++?  
:) 其实它们的位置完全不一样。而 Python 和 C++ 的位置更近。  
今年来 C++ 快速更新，C++11/14/20 大量借鉴了 Python 的一些特性，而 C 保持相对稳定。

地址	机器指令	汇编指令
0x0804857a	894442404	mov dword [esp + 0x4], eax
0x0804857e	c70424b28604.	mov dword [esp], 0x80486b2
0x08048585	e8eafdffff	call sym.imp.scant

<sup>1</sup><https://www.python.org/downloads/source/>

# 主要教材与参考书



- 教材 `pythonds`<sup>2</sup>，线上版本比印刷版本内容更加新。
- 源代码：<https://github.com/RunestoneInteractive/pythonds>

<sup>2</sup><https://runestone.academy/runestone/books/published/pythonds/index.html>

# 如何学习一门程序设计语言？

- 教程 vs. 手册
  - 手册包含了语言完整的特性，面面俱到，并详细阐述技术细节；一般都是大部头。
  - 教程教人快速上手，丰俭由人，有所偏重；一般都是中短篇幅。
- 学会之后需要有机会经常练习，否则忘的很快。
- 看书之外，使用开发工具自带的帮助，随用随查，更方便/更常用。
- 初学的时候，一两页的“*Cheat sheet*”也许能够解决你大部分的需要。

- 天网搜索引擎
  - 北大天网由北京大学网络实验室研究开发，是国家重点科技攻关项目“中文编码和分布式中英文信息发现”的研究成果。北大天网于 1997 年 10 月 29 日正式在 CERNET 上向广大互联网用户提供 Web 信息搜索及导航服务，是国内第一个基于网页索引搜索的搜索引擎。
- Web Informall
  - 中国互联网页信息博物馆，从 2002 年开始对中国的互联网网页进行增量搜集、存储、展示。网络爬虫每天的数据搜集能力达到了三千五百万网页。
- 区块链（联盟链）的应用场景设计

# 关于课程成绩

- 满分 100 分
  - 期末理论笔试: 40%
  - 平时课堂 + 作业: 30%
  - 上机编程: 30%
- 附加分 3 分
  - 微信答疑 3 分

# 大数据的来龙去脉

- 大数据与这门课有什么关系吗？
- 首先，没有大数据，各位今天也不会坐在这里。大数据让不同的行业都有了利用数据的需求。
- 大数据、互联网、人工智能，这些当前最热的名词内部都有着很强的联系。
- 大数据为时代的进步按下了加速键，了解大数据的来龙去脉有利于我们对时代的理解。
- 后面会介绍一些大数据案例，大家走马观花看一下，希望带来一点感性的认识。

# A Brief History of Humankind

- 七万年前人类的认知革命
- 一万两千年前人类的农业革命
- 500-300 年前的科学/工业革命
- 50 年前开始，仍在进行中的信息革命

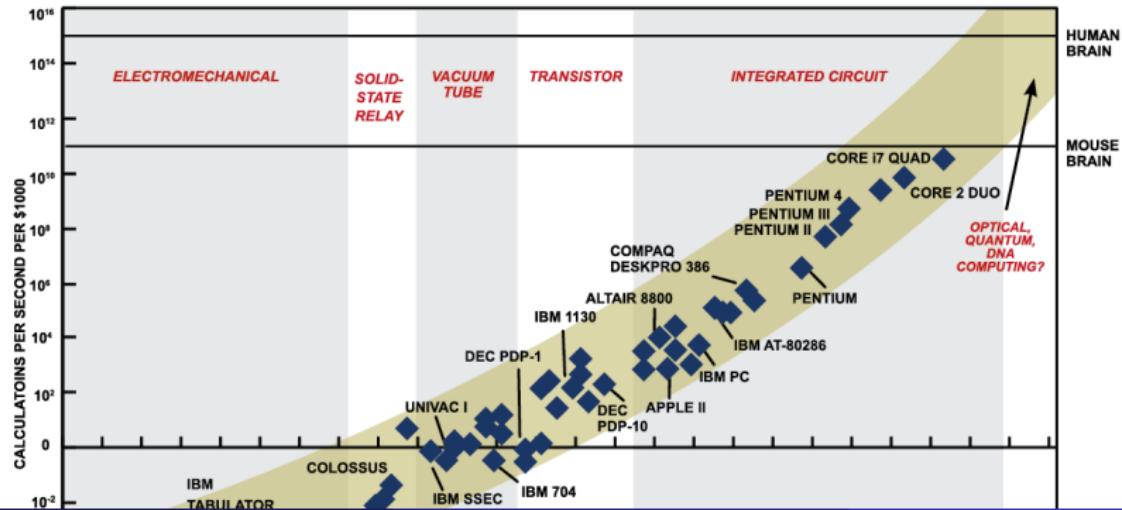
这种巨大社会变革发生的间隔越来越短，有人惊呼“奇点临近”。

- 人类科技发展到了一定程度，速度会越来越快。就好像达到“临界质量”会产生核反应！
- 信息技术的发展、信息的快速传播发挥了决定性的作用。
- 遥远而缓慢的时代：四大发明，中国 => 丝绸之路 => 西亚 => 阿拉伯商人 => 欧洲
- 当代：睡一觉起来全世界都知道了。明星的八卦、特朗普的 twitter。
- 信息的快速（有效）传播提高了人们应对危机的能力，以本次的新冠病毒为例。
  - 人口全球化流动，导致病毒的传播也全球化，一两个月内遍布全球主要地区。
  - 信息的快速传播，形成了高效的预警机制；大量的人协同工作，快速研发出疫苗，找到了针对病毒的解决方案。
  - 新冠现在看起来很凶，以后可能变成跟普通感冒一样，吃点药在家里休息几天就好。

# 摩尔定律 (英语: Moore's law)

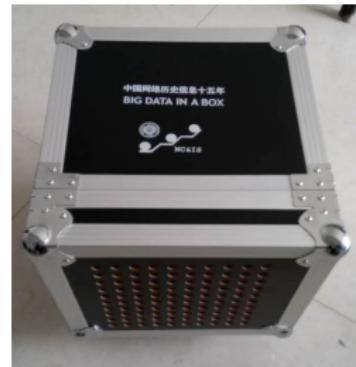
- 集成电路上可容纳的晶体管数目，约每隔 18 个月便增加一倍。
- 微处理器的性能每隔 18 个月提高一倍，或价格下降一半。
- 相同价格所买的电脑，性能每隔 18 个月增加一倍。
- v:1996 年以来的摩尔定律

由英特尔 (Intel) 创始人之一戈登 · 摩尔提出 (两年一倍)；英特尔首席执行官大卫 · 豪斯 (David House) 改为 “18 个月”。



# 摩尔定律对我们的影响. 一

- 消费电子价格
  - 电脑、手机、电视、相机、显示器
  - 通胀预期下，电子相关产品持续贬值
- 产品快速迭代/淘汰
  - BB 机
  - 智能电视内嵌 Android 操作系统
- 存储设备的数据密度 Web Informall, 持续搜集、保存中国互联网的网页。原来“汗牛充栋”，现在变成个小盒子。
- 数字消费升级
  - VCD, DVD, 高清, 超高清, 2K, 4K
  - 存储电影：硬盘大了，单个影片的文件大小也大了。比如：08 奥运开幕式视频



# 摩尔定律对我们的影响. 二

- 晶体管密度增大，设备的尺寸缩小，数量增多  
(潜在数据源)
  - 大哥大 ==> 手机
  - 无处不在的传感器：手表、家电、门锁、公共设施（摄像头）
  - 可穿戴设备（瓶颈为电池）
- 为即将到来的大数据提供了数据源
- 个人隐私？

姿态感应器  
环境光传感器  
红外传感器  
指纹传感器  
霍尔传感器  
陀螺仪  
指南针  
接近光传感器  
重力传感器  
色温传感器  
气压计  
Camera激光对焦传感器

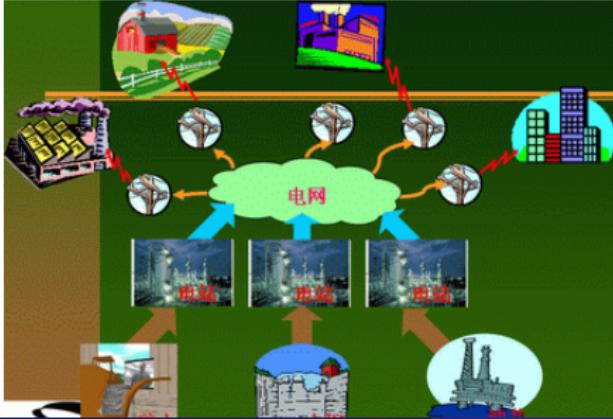
# 从“网格”到“云”

- 网格计算背景
- 中国网格计算发展历程
- 中国教育科研网格计划ChinaGrid
- 网格计算展望



CERNET第十届学术年会, 2003, 郑州

2



谢正茂 webg@pku.edu.cn (北京大学计算机系)

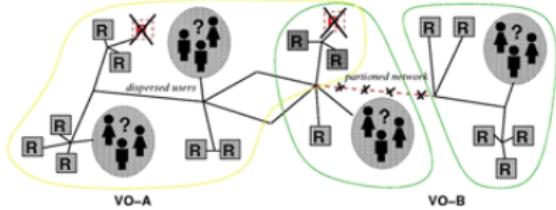
## 网格

动态多机构虚拟组织中的资源共享

和协同问题解决

(“Resource sharing & coordinated problem solving  
in dynamic, multi-institutional virtual organizations”)

引用



在网格环境中, 不论用户工作在何种“客户端”上。系统均能根据用户的实际需求, 利用开发工具和调度服务机制, 向用户提供优化聚合后的协同计算资源、并按用户的个性提供及时的服务



数据结构与算法 (Python)

March 13, 2021

24 / 52

# 人类公用资源/知识库

"Scan This Book! Kelly", Kevin(May 14th, 2006)

- 2004 年谷歌人类知识计划
  - 320 万本书, 7.50 亿篇文章
  - 2 千 5 百万首歌, 50 万部电影
  - 5 亿个图像, 3 百万个视频、电视节目和短片
  - 1000 亿网页
- 在中国, 超星扫描了 200 多个图书馆中的 130 万本中文书, 大约是 1949 年以来中文出版书籍的一半。

# Google Books 计划

<https://books.google.com/>

- 2004 年 12 月开始，目的扫描书和杂志，使用字符识别软件确认文本的字、词、句和段落，将数字化图像转化为数据化文本。
- 2013 年 4 月，扫描 3 千万本
  - 2010 年，全世界估计有 1.3 亿本书
  - 2008 年 11 月，数字化 700 万本
  - 2007 年，数字化 100 万本。
- 2007 年 9 月，发布“ My Library ”，对用户进行个性化推荐。

- 人们需要从计算机光学字符识别程序 (OCR) 无法识别的文本扫描项目中读出两个单词并输入。
  - 其中一个单词其他用户也识别过，从而可以从该用户的输入中判断注册者是人；
  - 另一个单词则是有待辨识和解疑的新词。为了保证准确度，系统会将同一个模糊单词发给五个不同的人，直到他们都输入正确后才确定这个单词是对的。
- 在这里，数据的主要用途是证明用户是人，但它也有第二个目的：破译数字化文本中不清楚的单词。

# 人类一天产生的数据量

**2940亿封电子邮件发送**

平均每个地球人每天发送42封

**6288个新移动应用**可被下载

日均下载量已达3500万

**57万6000小时视频**

上传到YouTube

**3亿5000万张照片**

上传到Facebook

如果把它们都印出来，叠起来能有  
80个埃菲尔铁塔那么高

**500TB数据**上传到Facebook

如果用2TB的硬盘储存

每年Facebook要新购65吨硬盘

**20亿小时电视与电影**

在Netflix上观看

整个因特网的流量信息可以装满

**24万亿张DVD光盘**

需要6万艘10万吨的油轮运送

**2亿3000万条tweets**

在Twitter上发布

据预测，到2020年

**每年上传的总数据为35ZB**

即35,000,000PB

**每年数据量增长60%**

其中非结构化数据增长80%

# Human Data Size in 2008



Human Genomics  
**(7000PB)**  
1GB / person  
200PB+ captured  
200% CAGR

<http://www.intel.com>  
World Wide Web  
**(~1PB)**  
<http://www.intel.com>  
<http://www.intel.com>

wiki wiki  
Wikipedia  
**(10GB)**  
100% CAGR  
i wiki wiki

Internet Archive  
**(1PB+)**

Estimated On-line RAM in Google  
**(8PB)**

Personal Digital Photos  
**(1000PB+)**  
100% CAGR

2004 Walmart Transaction DB  
**(500TB)**

Typical Oil Company  
**(350TB+)**

Merck Bio Research DB  
**(1.5TB/qtr)**

MIT Babyltalk Speech Experiment  
**(1.4PB)**

Terashake Earthquake Model of LA Basin  
**(1PB)**

One Day of Instant Messaging in 2002  
**(750GB)**

Total digital data to be created this year **270.000PB** (approx.)

# 大数据在中国

## 新一轮“信息革命”

3.5ZB in 2011

1天的数据量

> 文明起始到2003年



10.77亿

移动互联网用户

中国 2016年12月



2090亿

2021年RFID标签销售量  
2011年是1200万



200PB/季度

智慧城市数据



中国某一线城市

\$8000亿

10年个人位置信息服务创造的价值



5PB/年

健康档案数据  
中国某一线城市



“数据日益成为商业的新源材料：一种与资本和劳动力并列的新经济元素。”

— *The Economist, 2010*

“信息将成为21世纪的石油”。

— *Gartner, 2010*

# 衡量数据量的方法

- 计算机基础存储单位：字节
- $1 \text{ KB} = 2^{10} = 1,024$  字节
- $1 \text{ MB} = 2^{20} = 1,048,576$  字节
- $1 \text{ GB} = 2^{30} = 1,073,741,824$  字节
- $1 \text{ TB} = 2^{40} = 1,099,511,627,776$  字节
- $1 \text{ PB} = 2^{50} = 1,125,899,906,842,624$  字节
- $1 \text{ EB} = 2^{60} = 1,152,921,504,606,846,976$  字节
- $1 \text{ ZB} = 2^{70} = 1,180,591,620,717,411,303,424$  字节



640 KB 足够所有人用了!  
——比尔·盖茨, 1981

# 美国海军上尉和他的大数据实践

- 马修 · 方丹 · 莫里
- *Matthew Fontaine Maury*
- 1806 年出生于美国弗吉尼亚
- 1824 年刚刚达到入伍年龄便进入了美国海军学校
- 1839 年，已经晋升为海军上尉的莫里在一次事故中不幸腿部致残
- 不适合于服役远航的莫里在 1842 年被任命为主管海图和仪器库的负责人



# 大航海时代的海图



六分仪



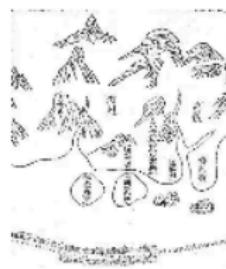
经典的书籍、教材



指南针



1459年毛罗地图



郑和航海图



哥伦布航海图

# 莫里的目标



变废  
为宝



与商船交换信息，在自愿基础上以互利互惠的合作方式开创了国际气象界公开交换环境资料的传统



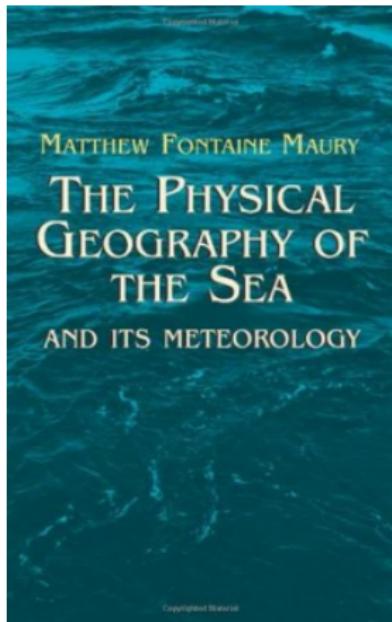
莫里

更新后的图表  
← 航海日志

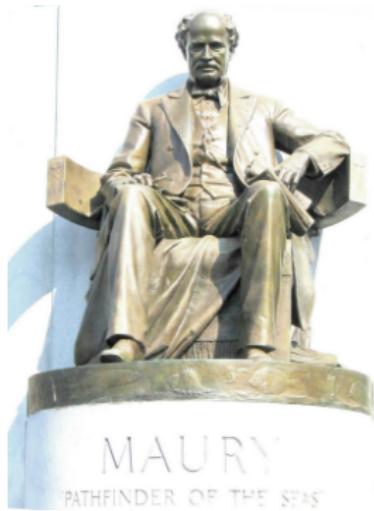


商船

# 名垂千古：海洋学的奠基人



- 1855 年，莫里出版权威著作《海洋物理地理学和气象学》，被誉为海洋学的奠基人
- 当时，他已经绘制了 120 万个数据点
- 四个国家授予了他爵士爵位，包括梵蒂冈在内的其他八个国家还颁给了他金牌奖章
- 即使到今天，美国海军颁布的导航图上仍然有他的名字



# 数据无处不在



交通数据



金融数据



物联网数据



零售数据



社交网络数据



科学数据

# 每一个智能手机...



# 典型和非典型物联网设备



# 智能设备：每时每刻收集数据



智能手机

地理位置数据  
运动数据  
环境亮度数据  
图像数据  
语音数据  
...



身体状况数据  
运动习惯数据  
实时图像数据  
...



可穿戴设备

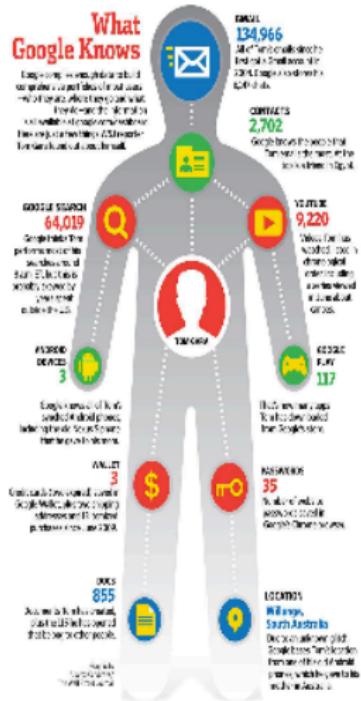


智能家居设备

# 信息时代的数字脚印



# 互联网企业比你更了解你 v:PrivacyIsWorthMore



从你的数据中了解你!

- 大数据的时代背景
  - 宏观经济大数据典型案例
- 编程语言和教材的选择

# 大数据宏观经济监测预警意义

- 提高经济运行及时性、全面性和准确性
  - 《十三五规划纲要》
- 提高决策的针对性、科学性和时效性
  - 《促进大数据发展行动纲要》
- 建立基于大数据的宏观经济监测体系
  - 《促进大数据发展 2017 年工作要点》
- 百家大数据机构描绘中国经济

# 目录

- 大数据的时代背景
  - 交通大数据的探索
- 编程语言和教材的选择

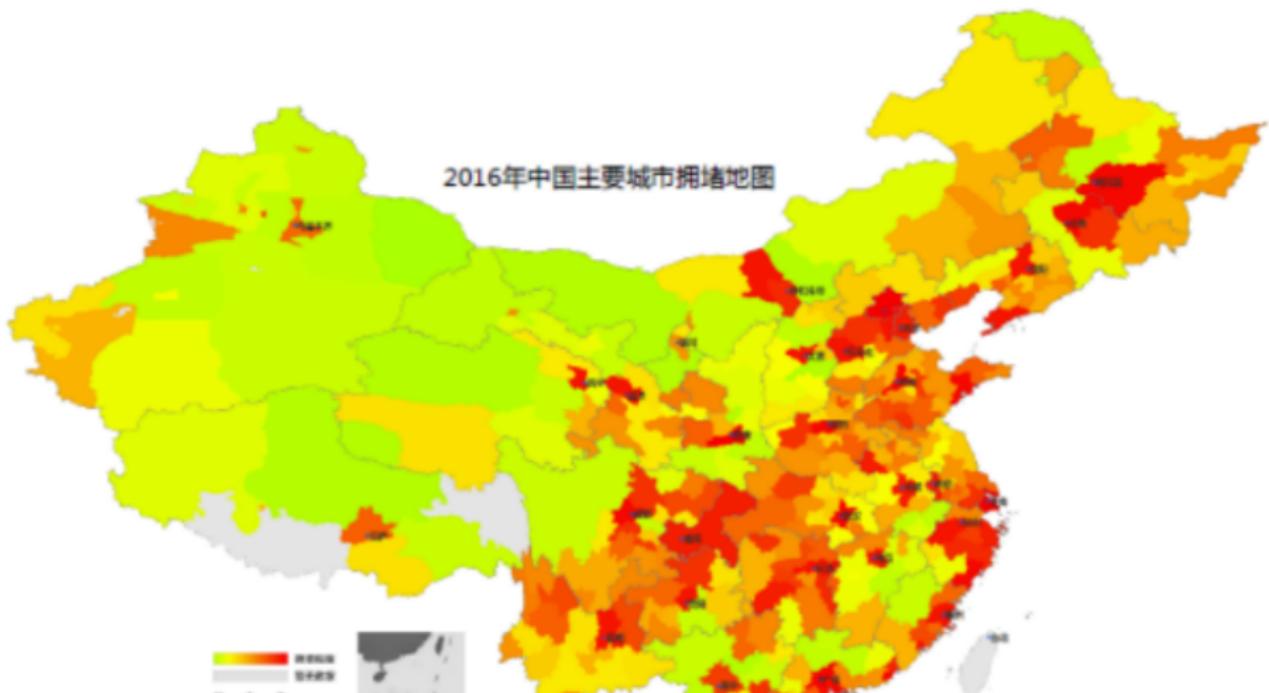
# 智慧城市-智能交通

- 济南是全国最拥堵的城市
  - 拥堵指数达到 2.173, 高峰时刻平均速度 19.89Km/h.
- 哈尔滨、北京、重庆、贵阳也都是非常拥堵的城市
  - 一些二线、三线城市日益拥堵



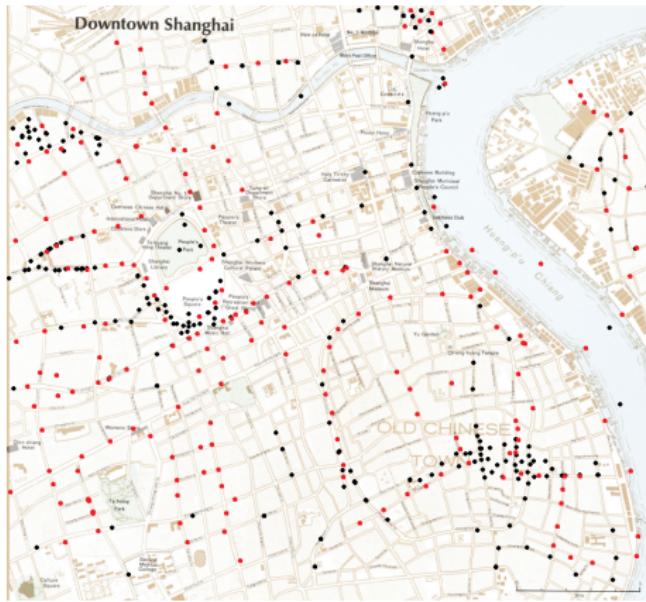
# 年度拥堵地图

- 1/3 cities 高峰时候拥堵
- 32 cities 拥堵指数超过 1.8
- 大城市和一线城市总是很拥堵



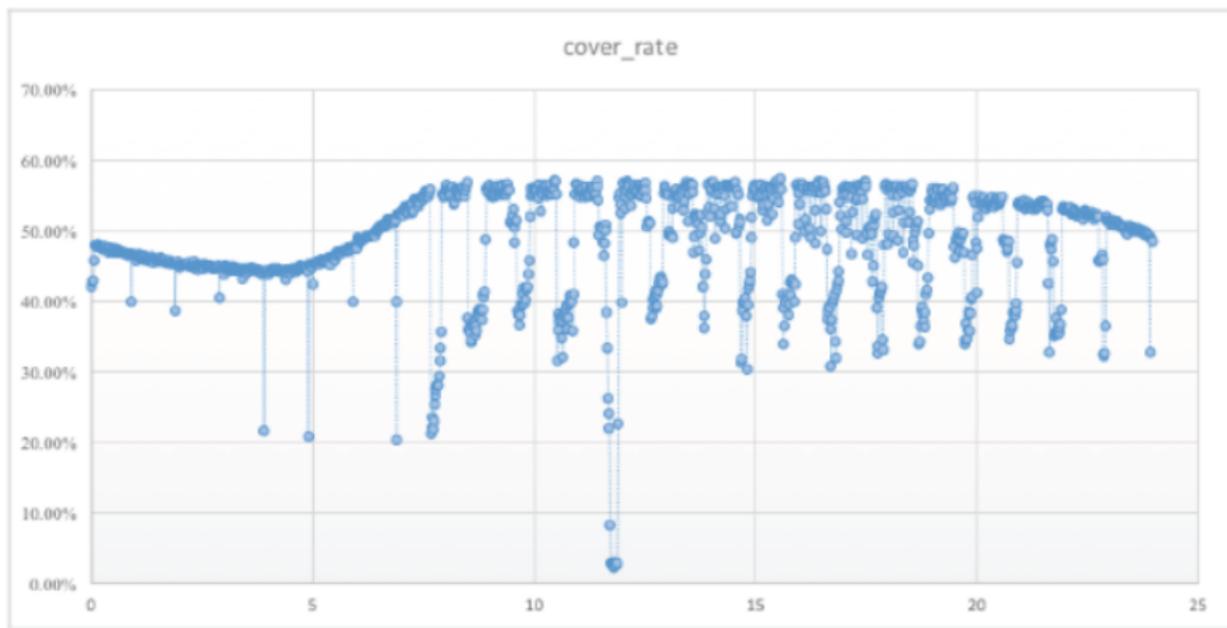
# 基于浮动车的解决方案

- 基于浮动车
  - 出租车
  - 滴滴
  - 公交、大巴、大货
  - 高德、百度地图
- 时空局限
  - 覆盖率、渗透率低
  - 喜欢聚集
  - 相似路径
  - 有规定的偏好



# 高时空覆盖

提高打车成功率；降低车辆空驶率



# 大数据下淘宝“千人千面”

2017年11月12日凌晨，  
阿里公布了淘宝天猫“双十一”购物狂欢节全天的销售额：

支付宝全天成交金额为：  
**2135亿**

比2015年的1682亿增长：

**27%**

订单：

**12.35亿**



# 大数据的价值

## 商业运营

- 古装偶像剧在上个月产生了多少次播放？上周呢？昨天呢？前一个小时呢？
- 不同渠道带来的独立用户有多少？他们的停留时间和留存率如何？
- 每一部视频的投入（版权、带宽）和产出（广告收入，付费点播）比如何？分地区分析？分终端分析？
- 过去 6 个月赵丽颖、冯绍峰在视频观众里受关注程度变化趋势如何？

- Singularity: 科技到了一定阶段后，信息的传播、全球一体化，发展会越来越快。对比一下“四大发明”的传播过程，今天晚上的东西发个 `twitter`，明天早上全世界都知道了。
  - 人类会不会进入一个类似“临界质量”的状态？
  - “临界质量”后面是什么？
- 摩尔定律：电子技术的发展，为上面的变化提供了硬件基础。
- 数字化浪潮
  - 传统媒体的数字化（书、画、建筑，人类原来的所有知识）
  - 新的数据源，不断产生数据。
- 网格计算、云计算为数据提供收集、存储、处理的工具和平台。
- “数据的开放与共享”在民间和官方之间达成共识，比如如何战胜本次的“新冠病毒”。
  - 疫情公开，政府  $\Rightarrow$  丁香园、腾讯新闻、网易...
  - 疫情防控，密切接触/易感人群，交通数据。
  - 疫情预测。
- “大数据”就是在这样的情况下发生了。

- 利用多来源、多纬度、多种类的数据去解决问题/获取答案
- 大数据的思维并不新鲜。
  - “顺藤摸瓜”: 不能直接达到目标, 通过相关联的事务逐步接近; 寻找事物之间的关联。
  - *Google Flu Trends*: 45 个查询词, 通过一个数学模型, 能够预测或者实时播报美国的流感。
- 价值在于“全”, 而不是特别强调“大”。

