

# An Introduction To Big Data

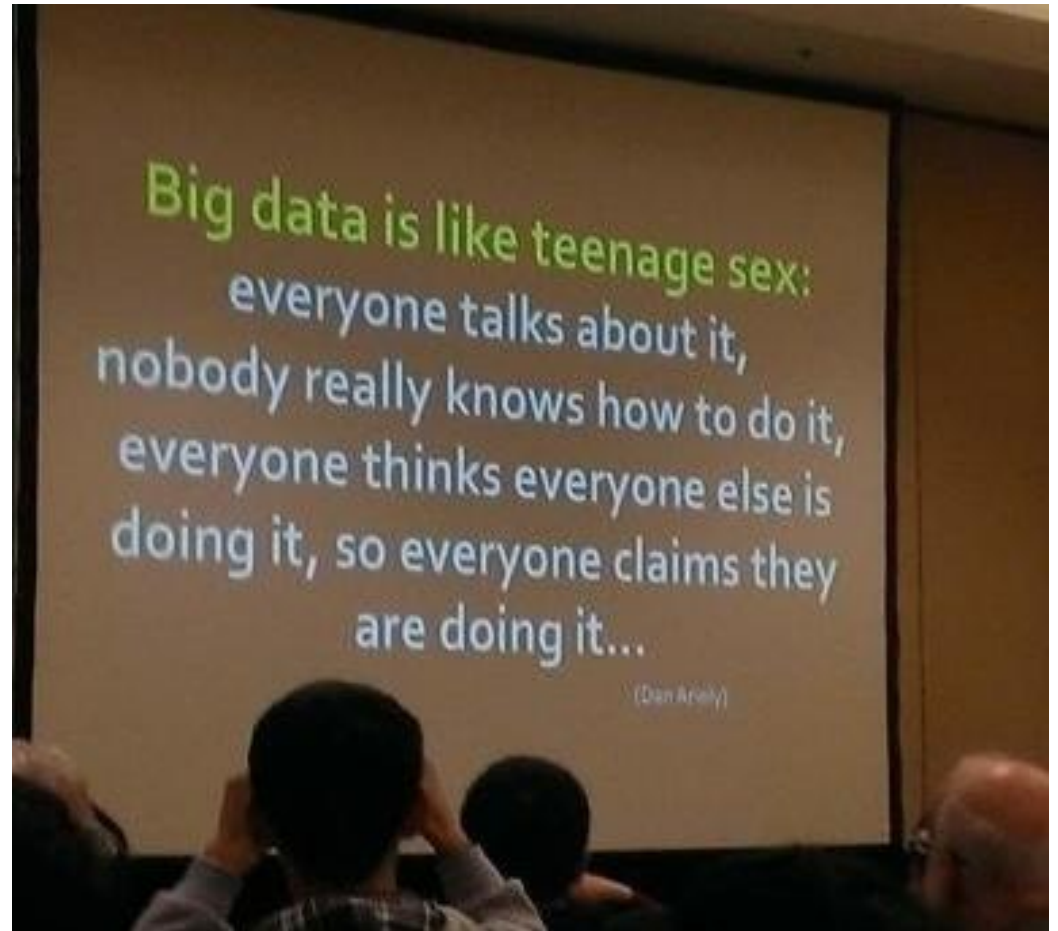
叶邦宇

蚂蚁金融服务集团支付宝基础数据部

bangyu.yby@alipay.com

# What is Big Data

- 众说纷纭！
- 规模 and 价值









# task1



# task2

- MightyTV公司推出其APP，采用机器学习等技术可收集Facebook上10位以上好友推荐的视频并智能匹配“最想看的视频”。该公司希望通过这种方式解决夫妻等一起观看视频的问题。因为目前他们很多人已因为无法统一该看什么而不再一起视频

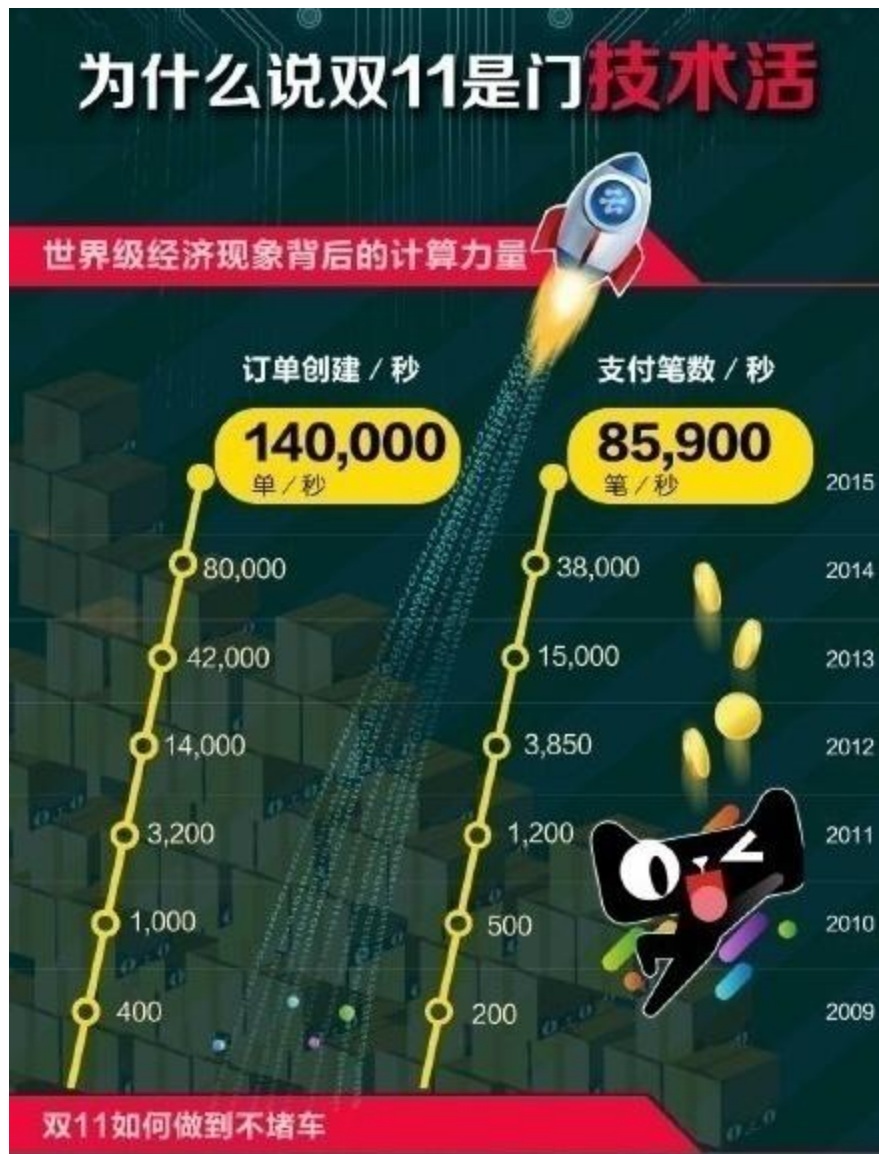


# task3

- 某公司将机器学习图像映射技术应用于诊断疾病。这家公司将医学影像输入到机器中，让它学习其中的规律。对于初期疾病的判断，机器的准确率是专家小组的三倍，而且疾病误判的几率也被大幅降低。



# task4





# task5

- 淘宝数据平台显示，购买最多的文胸尺码为B罩杯。B罩杯占比达41.45%，其中又以75B的销量最好。
- 其次是A罩杯，购买占比达25.26%.
- C罩杯只有8.96%。
- 在文胸颜色中，黑色最为畅销。
- 以省市排名，胸部最大的是新疆妹子。

# task6

- True&Co网站正利用大数据帮助女性寻找号码更合适的胸罩。统计数据显示，大多数女性都戴错了胸罩的号码，为此这家网站试图帮助解决这个问题。用户只要填写网站上的调查问卷，它就可以根据答案做出反应，并通过计算给出正确型号的胸罩。该公司的内部品牌甚至会基于用户的反馈和公司收集到的数据开发和设计新式胸罩

# task7

- 在加拿大多伦多的一家医院，针对早产婴儿，每秒钟有超过3000次的数据读取。通过这些数据分析，医院能够提前知道哪些早产儿出现问题并且有针对性地采取措施，避免早产婴儿夭折。

# task8

← → ↻ www.amazon.cn

汽车用品 >

游戏、影视、乐器 >

亚马逊 Cloud Drive >


全部商品分类

限时热门抢购

**¥ 48.00**

距结束 01:08:04


全部秒杀商品



好奇

**¥ 133.00**

立即查看



户外/出  
低至 1!

立即查看

## ☆ 为您推荐

Canon



¥ 8,399.00



¥ 2,280.00



¥ 6,199.00  
买赠闪迪16G高速



¥ 3,299.00  
购买 1 件售价立



# task9

- 微软发布了在线网站CaptionBot.ai，可以自动对任何图片加描述

- I am not very confident, but I think it's a little girl holding a teddy bear and they seem 🙄🙄



- I think it's a woman sitting on a chair in front computer

I think it's a woman sitting on a chair in front of a computer.



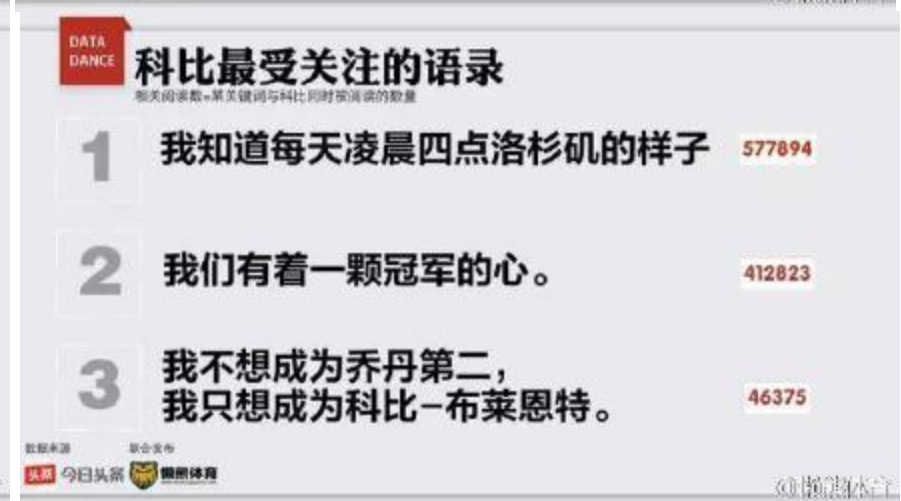
- I am not very confident, but I think it's a group of people standing next to a man in a military uniform and they seem 🤔🤔🤔.

I am not really confident, but I think it's a group of people standing next to a man in a military uniform and they seem 🤔🤔🤔.





# task10



# conclusion

- 1, 家庭生活
- 2, 医疗
- 3, 电子商务
- 4, 游戏/娱乐
- 5, 科学研究
- 6, 国民政策制定/总统选举
- .....

# Four categories of Big Data

- 数据采集      采集/爬虫/抓取
- 数据存储      数据存储/分布式系统/数据库
- 数据挖掘      数据挖掘/机器学习/推荐系统
- 数据计算      MapReduce/Spark
- 数据可视化      Data Visualization

# Four categories of Big Data

- 数据采集      采集/爬虫/抓取
- 数据存储      数据存储/分布式系统/数据库
- 数据挖掘      数据挖掘/机器学习/推荐系统
- 数据计算      MapReduce/Spark
- 数据可视化      Data Visualization



# Subjects...

- Artificial Intelligence 人工智能
- Data Mining, 数据挖掘
- Machine Learning, 机器学习
- Natural Language Processing 自然语言处理
- Recommender System, 推荐系统
- Social Network, 社交网络
- Search Engine, 搜索引擎

# 僵尸粉



# 判断一个粉丝是不是僵尸粉

## 特征

- 转发多，原创少
- 关注多，粉丝少
- 内容基本上都是广告
- 头像基本上都是美女图片
- .....

# 判断一个粉丝是不是僵尸粉

- if(粉丝数<=10&&关注数>=1000)  
    if(原创数<=100&&转发数>=1000)  
        if(广告所占比例>=90%)  
            .....
- 基于规则的方法，有什么缺点？ ？ ？

# Why we need machine learning

- WHY





# Why we need machine learning

- Drawbacks of Rule-based (if else if else....)

Not elegant

Hard to maintain

Lack of explanations of probability

# Machine learning

$$Y = 0.1 * \text{转发数} - 0.2 * \text{粉丝数} + 0.3 * \text{关注数} + 0.4$$

# 人工智能思潮

- 基于规则的专家系统
- 基于数据的机器学习

# Recommender System

- Amazon:

“查看此商品的顾客也查看了”

“购买此商品的顾客也购买了”

购买此商品的顾客也同时购买



数据挖掘导论(完整版)

陈封能 (Pang ...

★★★★☆ 93

平装

¥48.70



计算机科学丛书:机器学习

米歇尔 (Mitt ...

★★★★☆ 108

平装

¥29.60



数据挖掘十大算法

吴信东

★★★★☆ 18

平装

¥34.30



数据挖掘:实用机器学习工具  
与技术(原书第3版)

威滕 (Ian H ...

★★★★☆ 5

平装

¥60.80



大数据·互联网大规模数据  
挖掘与分布式处理

Anand Raj ...

★★★★☆ 104

平装

¥36.60

# Recommender System

- 喜欢还是不喜欢？ 喜欢是多喜欢？

# Recommender System

- **rating**
  - rating=1代表非常讨厌
  - rating=2代表讨厌
  - rating=3表示一般般
  - rating=4表示喜欢
  - rating=5表示非常喜欢



# Recommender System

	美人鱼	疯狂动物城	叶问3
A:	1	2	4
B:	5	4	2
C:	5	5	1
D:	1	1	5
E:	1	2	4

# Recommender System

	美人鱼	疯狂动物城	叶问3
A:	1	2	?
B:	5	4	2
C:	5	5	1
D:	1	1	5
E:	1	2	4

# User-based

- $\text{Rating}(D, \text{叶问3})=5$
- $\text{Rating}(E, \text{叶问3})=4$

# User-based

- $\text{Rating}(D, \text{叶问3}) = 5$
- $\text{Rating}(E, \text{叶问3}) = 4$
- $\text{Prediction}(A, \text{叶问3})$
- $= (5 + 4) / 2 = 4.5$

# 反思

- 相关性 与 因果性

# Recommender System

- 推荐系统严格依赖于场景和业务
- 考虑淘宝推荐衣服 & 美团网推荐酒店



# Recommender System

- What to do right now and next?
  - 1, Cold start /Data sparsity problem
  - 2, Scalability of model
  - 3, Online-learning
  - 4, Explanations in recommender systems
  - 5, Attacks and protections

.....

# What should we do?

- 怎么更快处理数据？
- 怎么利用数据？

# 冷静思考

1，真的需要大数据吗？

很多时候简单的规则和统计即可

2，真的需要很牛逼的算法吗？

很多时候，重要的不是算法牛，而是数据多

# 研究现状

- 相关牛人:

@明风

@吴甘沙-驭势科技

@Andrew-Xia

@尹绪森

# 研究现状

- 相关牛人:

@余凯\_西二旗民工

@老师木

@南大周志华

@王斌\_IIEIR

@phunter\_lau

@黄萱菁

# 就业形势

## [内推] 算法工程师

待遇水平：	年薪30~50万
公司部门：	爱奇艺 / 系统架构部
所在城市：	北京
详细地址：	北京市海淀区海淀北一街2号鸿城拓展大厦11层
发布信息：	bullud于2013-11-21

### 机会吸引力：

爱奇艺系统架构部X-Team寻找优秀工程师，协同解决音视频处理、信息检索、模式识别、自然语言处理、机器学习、数据挖掘等技术在实际应用中遇到的各种挑战性问题，研发改变行业和生活的系统级产品。

### 机会详情：

我们将现代信息科技的成果恰到好处的注入到产品中，拥有良好的基本素质和相关专业知识的您，将有助您这份工作上获得成功。因此，

我们期望优秀的您：

1. 拥有计算机、电子工程、自动化、数学、物理等相关方向硕士或博士学位
2. 已熟练掌握 C/C++，Java，Python语言中的任意一种，并熟悉基础性数据结构和相关算法
3. 熟练掌握了机器学习、机器视觉、信号处理、模式识别等任一方面的基础知识（Plus）。
4. 拥有音频识别、音视频处理、人脸识别、文本挖掘、语音识别、场景检测等任一方面的项目经验（Plus）。
5. 具备良好的英语文献阅读能力，学习能力、团队协作能力。



# 就业形势

## [内推] 数据挖掘工程师

待遇水平：年薪20~40万  
公司部门：百度 / 百度国际化事业部  
所在城市：北京  
发布信息：BigHuge于2013-12-20

---

### 机会详情：

#### 工作职责：

- 1.建设和挖掘结构化数据，用以支撑国际化事业部推荐及其他产品
- 2.有大规模数据挖掘与机器学习系统经验，能实现数据采集、分析和挖掘，产出对数据分布规律、变化趋势、关联关系的知识
- 3.结合具体产品，设计合理的策略和算法对用户数据和内容数据进行分析，提升产品效果

#### 职位要求：

- 1.热爱互联网，对推荐技术、数据挖掘、探索解决问题有浓厚的兴趣
  - 2.良好的逻辑能力，良好的学习能力，有不拘一格的灵活思路
  - 3.扎实的机器学习/数据挖掘理论和技术基础，有2年以上的相关研究或工程经验尤佳
  - 4.精通Java,PHP或者Python等程序设计语言，对数据结构有深刻的理解和掌握
  - 5.了解Map-Reduce, MPI等分布式计算框架，具备相关开发能力的尤佳
  - 6.注重团队协作，有良好的沟通能力
-

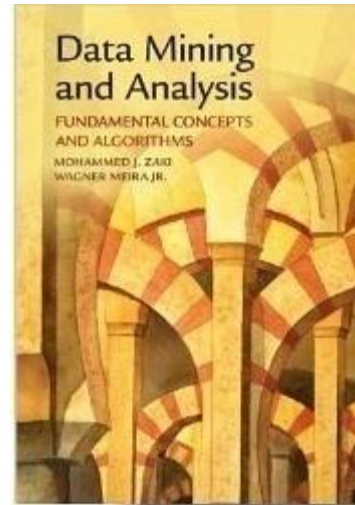
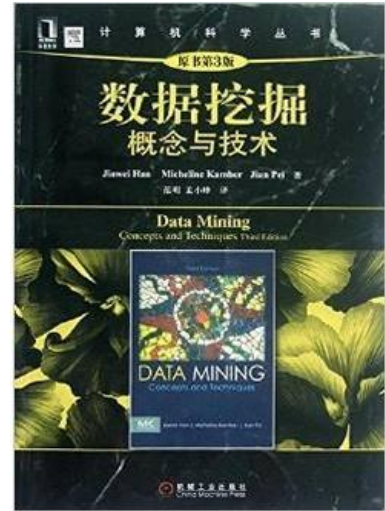
# How to be a data scientist

- 1, 数学（基础）  
高等数学 线性代数 概率论
- 2, 编程语言（基础）  
首选：Python！ Python！  
其他：C++ /Java /R
- 3, 算法和模型（理论）  
机器学习 数据挖掘 推荐系统 自然语言处理
- 4, 开源库（工具）  
单机：Scikit Learn  
分布式：Spark
- 5, 实操（实践）  
阿里巴巴大数据竞赛 /Kaggle竞赛等

# References

## 数据挖掘（入门）

- Jiawei Han  
《Data Mining: Concepts and Techniques》  
《数据挖掘：概念与技术》，机械工业出版社
- 《Data Mining and Analysis:  
Fundamental Concepts and Algorithms》



# References

## 机器学习（入门）

- 南大周志华老师的《机器学习》

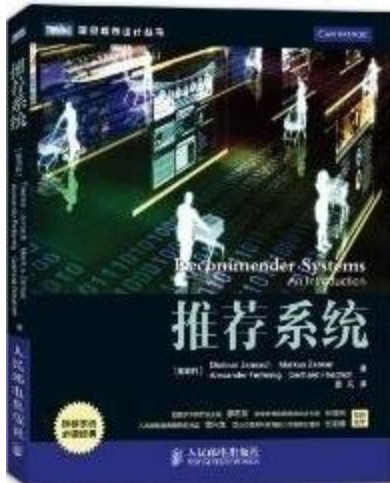


# References

## 推荐系统（入门）

《Recommender Systems: An Introduction》

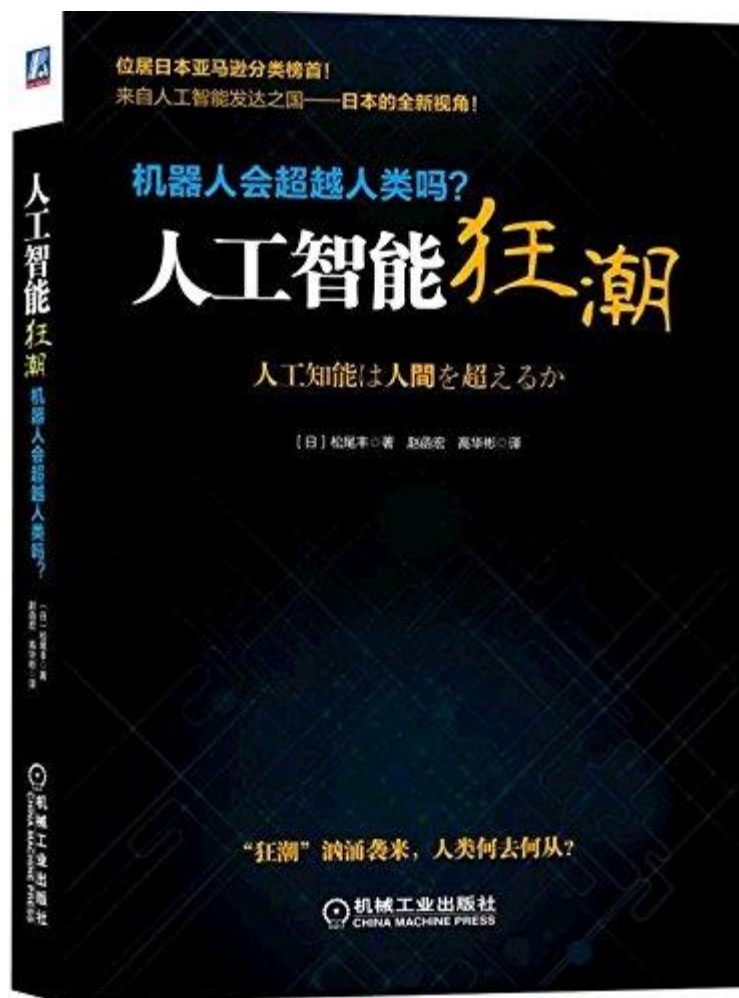
《推荐系统》，人民邮电出版社



# 休闲读物



# 休闲读物





# 计算机专业学生的出路

- I 技术类
- II 非技术类





# 计算机专业学生的出路

- I 技术类
- 程序猿/软件工程师/码农/算法工程师/研发工程师
- 前端/Android & iOS/后台开发  
大数据/嵌入式/游戏开发/



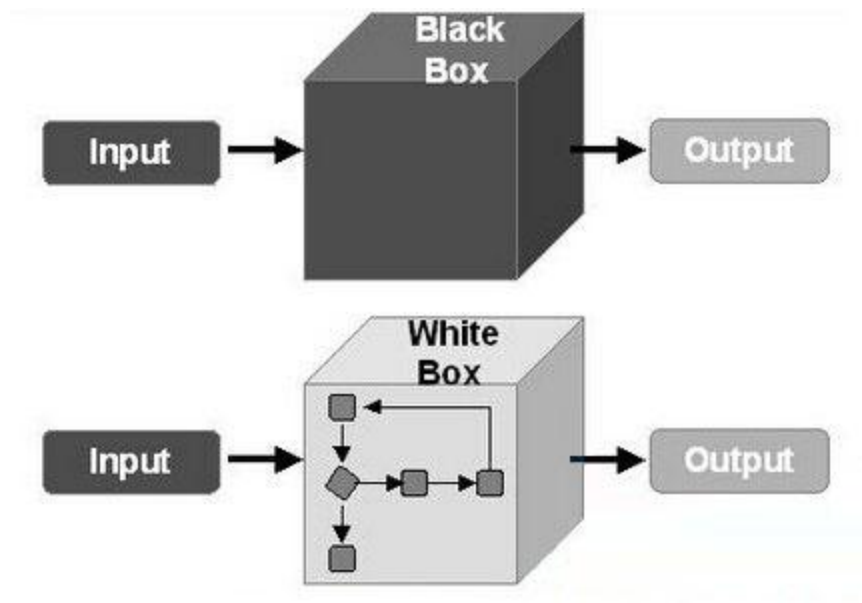
# 计算机专业学生的出路

- I 技术类
- 产品经理/产品狗/PM/Product Manager/



# 计算机专业学生的出路

- I 技术类
- 测试



# 计算机专业学生的出路

- I 技术类
- 网络工程师



# 计算机专业学生的出路

- II 非技术类
- 考研（本专业 || 跨专业）
- 老师
- 公务员
- 事业单位
- 转行
- 打字员
- 赢娶白富美/嫁给高富帅，走向人生巅峰



# 成为技术帝？ ？ ？

- I 认真学习，每门课100分。奖学金拿到手软。
- II 编程语言（高级语言+辅助语言）  
C/C++/Java                      Python/Perl/Ruby/R
- III 算法  
ACM竞赛 OJ刷题
- IV 项目
- V 大数据
- VI 准备考研
- VII 比赛
- VIII 开源社区

# 工作感言

- 基础知识很重要
- 有机会就上
- 码农是屌丝逆袭最好甚至是唯一机会
- 身体最重要，其他浮云

# 感谢

- 感谢曾台盛老师的邀请
- 感谢曾台盛、王鸿伟、杨竞菁、陈丹、Mandy、武存江、陈明玉、宋金玲、董会英、林捷、陈育明、王莲芳、林国新等老师
- 感谢本科同班同学陈承、黄炫贵
- Last but not least...



# Q&A

- Thanks A Lot
- Any questions?