



# 数据、 大数据、 数据科学

吴喜之

# 数据科学的前途

- 数据科学家的平均年薪为118709美元
- 而程序员的平均年薪为64537美元
- 麦肯锡公司的一份研究预测称，到2018年，在`具有深入分析能力的人才`方面，美国可能面临着14万到19万的缺口
- 而`可以利用大数据分析来做出有效决策的经理和分析师`缺口则会达到150万。

# 你擅长数学, 会用Python编程, 而且还对某个行业了如指掌?

- 如果你拥有这样的技能组合, 那你就有可能当上数据科学家.
- LinkedIn的投票结果显示: 统计分析和数据挖掘是最大的求职法宝.

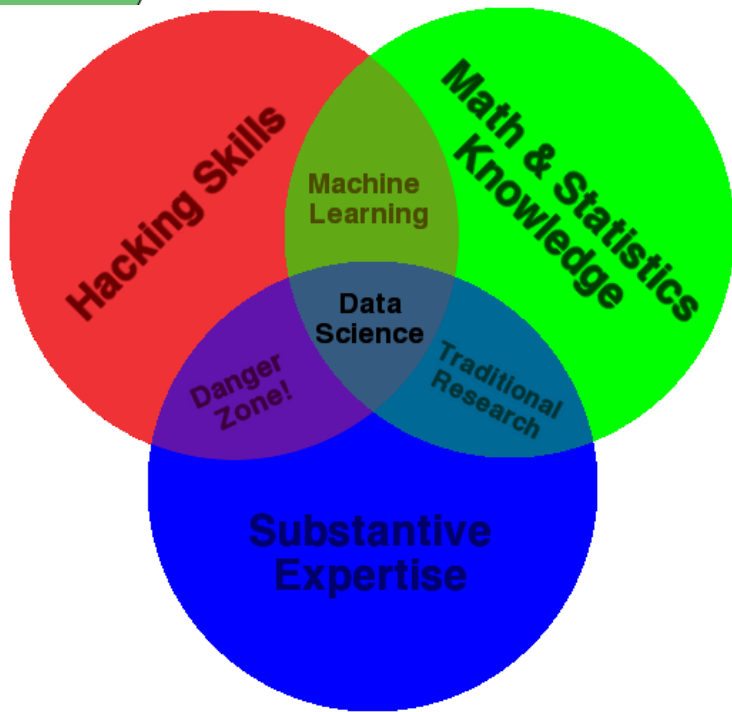
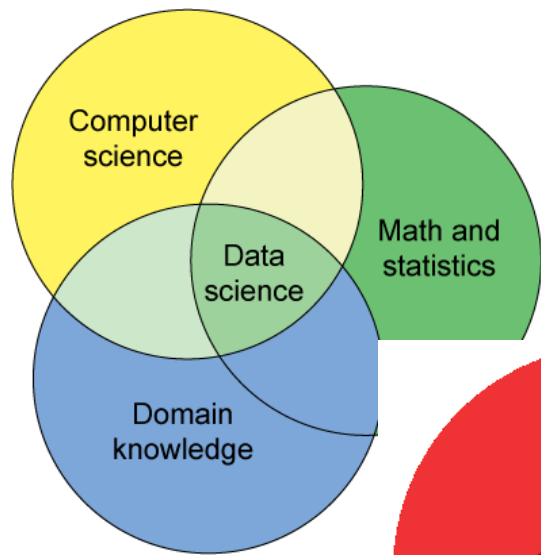
# 数据科学火爆的原因

- 尽管像谷歌、亚马逊、Netflix和Uber这样的高科技公司都有自己的数据科学团队
- 但那些非高科技公司, 比如Neiman Marcus、沃尔玛、Clorox和Gap, 它们现在也需要使用这方面的人才,
- 数据科学专业人才可以挖掘新的信息, 帮助公司开源节流.

# 科学

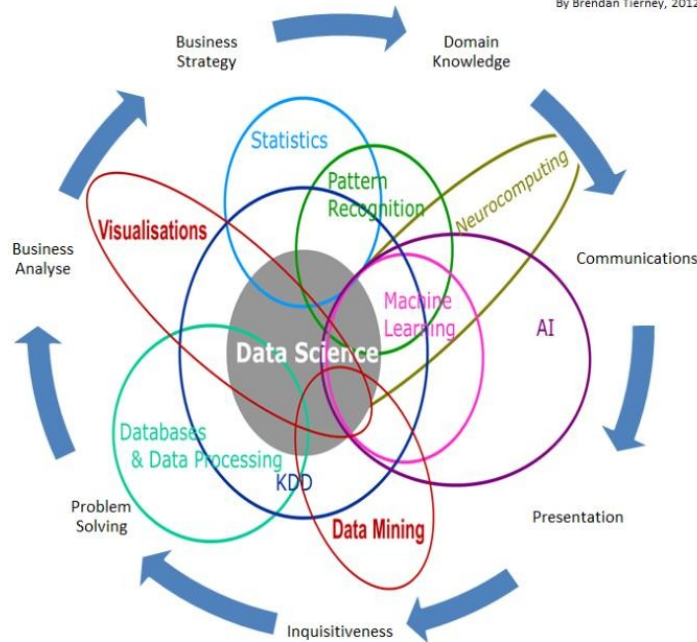
- 科学意味着没有权威.
- 任何科学研究的目的是基于数据颠覆旧的理论

# 数据科学



## Data Science Is Multidisciplinary

By Brendan Tierney, 2012



# 数据科学：数学的逻辑

- 不是具体方法和公式
- 而是基本的逻辑
- 数学逻辑是各种学科中最严格的逻辑
- 文理分科造成没有逻辑的法官和没有逻辑的文章……





# 数据科学：统计的批判性思维

- 不是你们在《统计学》或《数理统计学》课本中学到的70-100年前的知识
- 不是基于无法验证的假定而形成的假设检验和区间估计等
- 是近20年发展的并仍然在发展的最新的机器学习方法





# 数据科学： 计算机科学

- 不是一两个盗版傻瓜软件+点鼠标
- 网络漫游能力+泛型编程能力+数据库管理+可视化+非格式化流动数据处理能力+.....

### 2015 primary programming language:

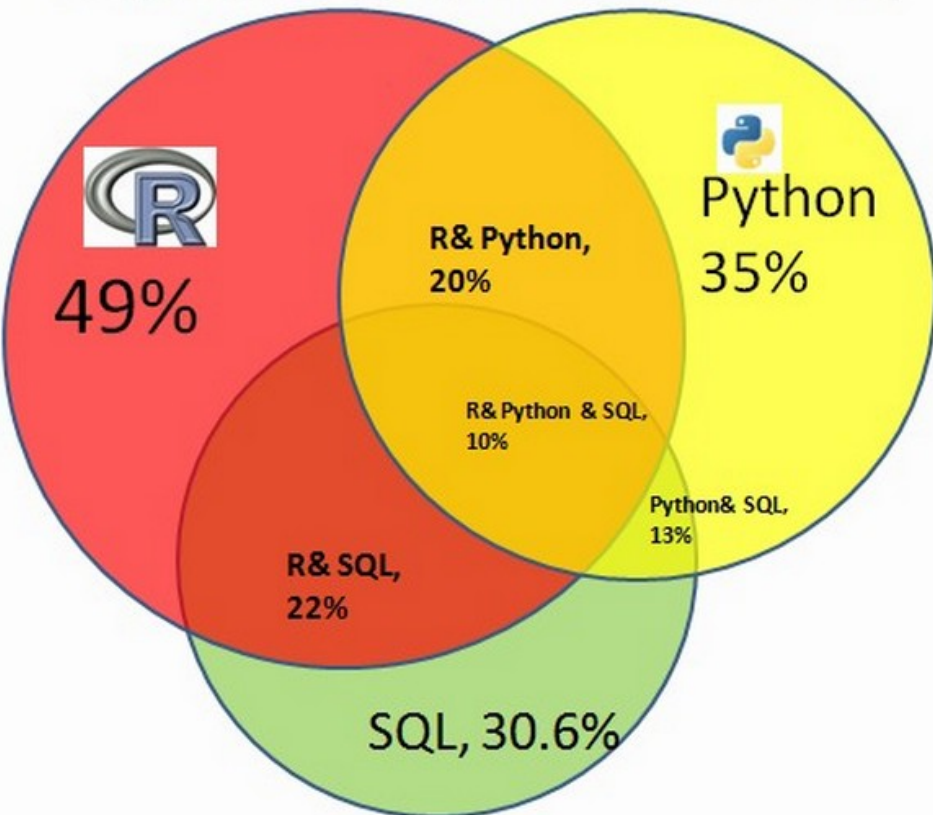
R (and its packages) (263)	 51% (of 2015 votes)
Python (including scikit-learn and other libraries) (151)	 29%
Other (Java, MATLAB, SAS, Scala, etc ) (89)	 17%
none (9)	 1.8%

### 2014 primary programming language:

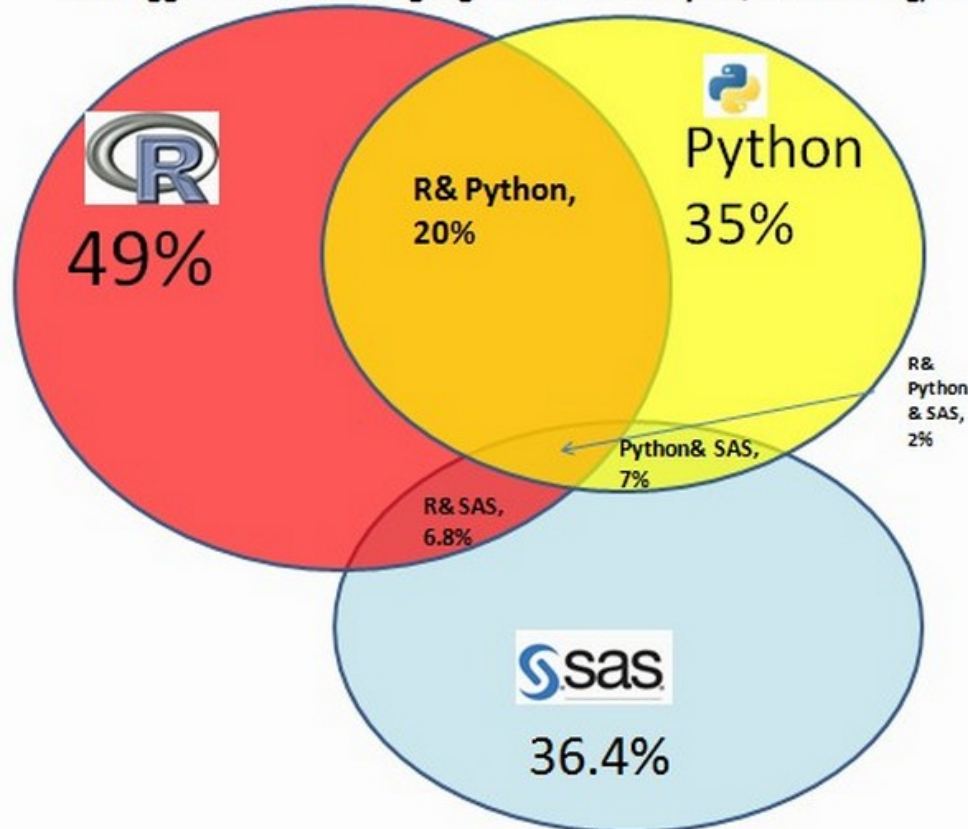
R (and its packages) (237)	 46% (of 2014 votes)
Python (including scikit-learn and other libraries) (117)	 23%
Other (Java, MATLAB, SAS, Scala, etc ) (118)	 23%
none (40)	 7.8%

# 计算机语言的亲和性

KDnuggets 2014 Poll: Languages used for Analytics/Data Mining

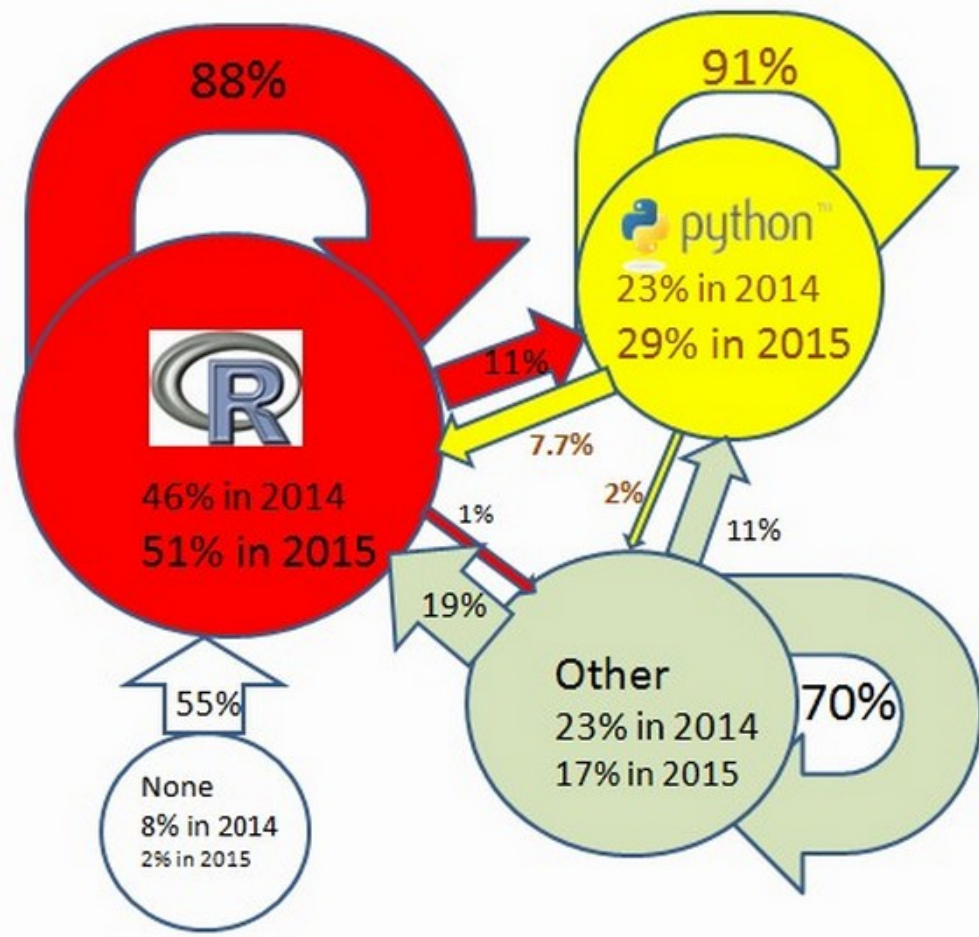


KDnuggets 2014 Poll: Languages used for Analytics/Data Mining, 2



# 计算机语言的 转换和亲和性

Primary Analytics, Data Mining, Data Science  
Languages in 2014 vs 2015



# 数据科学：产生数据的领域知识

- 只有明白目标领域知识的人才能了解数据的意义以及指导数据分析的方向并判断数据分析结果的可信性
- 没有领域知识的人主导数据分析肯定会误导(无论他/她有什么地位、权威背景或头衔——包括院士或诺贝尔奖得主)
- 用数据来说话，其他一切都是废话！

**数据科学就那么多基本内容  
基于这些，就可以扩展到无限！**

- **不要迷信那些炒作的“新名词”，“新概念”**
- **用自己的大脑，用常识判断**
- **跟风就意味着永远当跟随者，绝对不会有出息**

# 炒作没有人管， 但自己不能糊涂

- 原料：数据（小或大）及有关领域知识
- 思维：基于数据的批判性思维，而非基于主观经验、权威或者知识，也不是迎合、取宠式思维
- 工具：泛型编程能力+计算机系统
- 个人：快速自学的能力及对数据分析的爱好

# 是不是做数据科学家的材料？

- 与专长于任何特定编程语言相比，泛型编程技巧远远更加重要
- 最重要的素质就是能够快速学习东西。在如今这个时代，技术的发展突飞猛进，语言会很快过时，新的语言则将迅速普及。因此，学东西很快的人，会比单独领域的专家更有前途

# 是不是做数据科学家的材料？

- 每天花大量时间来编程，分析控制面板上的数据，获得相关信息，如果你对这样的工作感兴趣，那么你可能就适合干这一行.
- 如果仅仅是想拿高工资，那么你可能会觉得这样的日子过起来苦不堪言.

# 是不是做数据科学家的材料？

- 真正适合干这一行的人，会在业余时间里编写程序，分析数据，而他们这样做只是为了自娱自乐.
- 如果你爱的并不是数据本身，而是它可以给你带来的高薪，那么你会发现，自己很难与那样的人竞争.
- 每个人都应该学会热爱数据，即便只是为了自己事业前途着想，也该这样做.

# 还需要什么？

- 能力，特别是学习能力，比知识更重要
- 欢迎挑战，乐于攀登
- 在怀疑中成长（马克思的座右铭：De omnibus dubitandum, 怀疑一切）
- 不要给自己贴标签（“我是学某个方向的，别的不搞”、“岁数大了，学不会了”、“对某方向会不感兴趣”）
- All I know is that I am not a Marxist---Karl Marx

开一个小玩笑：  
你属于什么民族？

你是北方人吗？

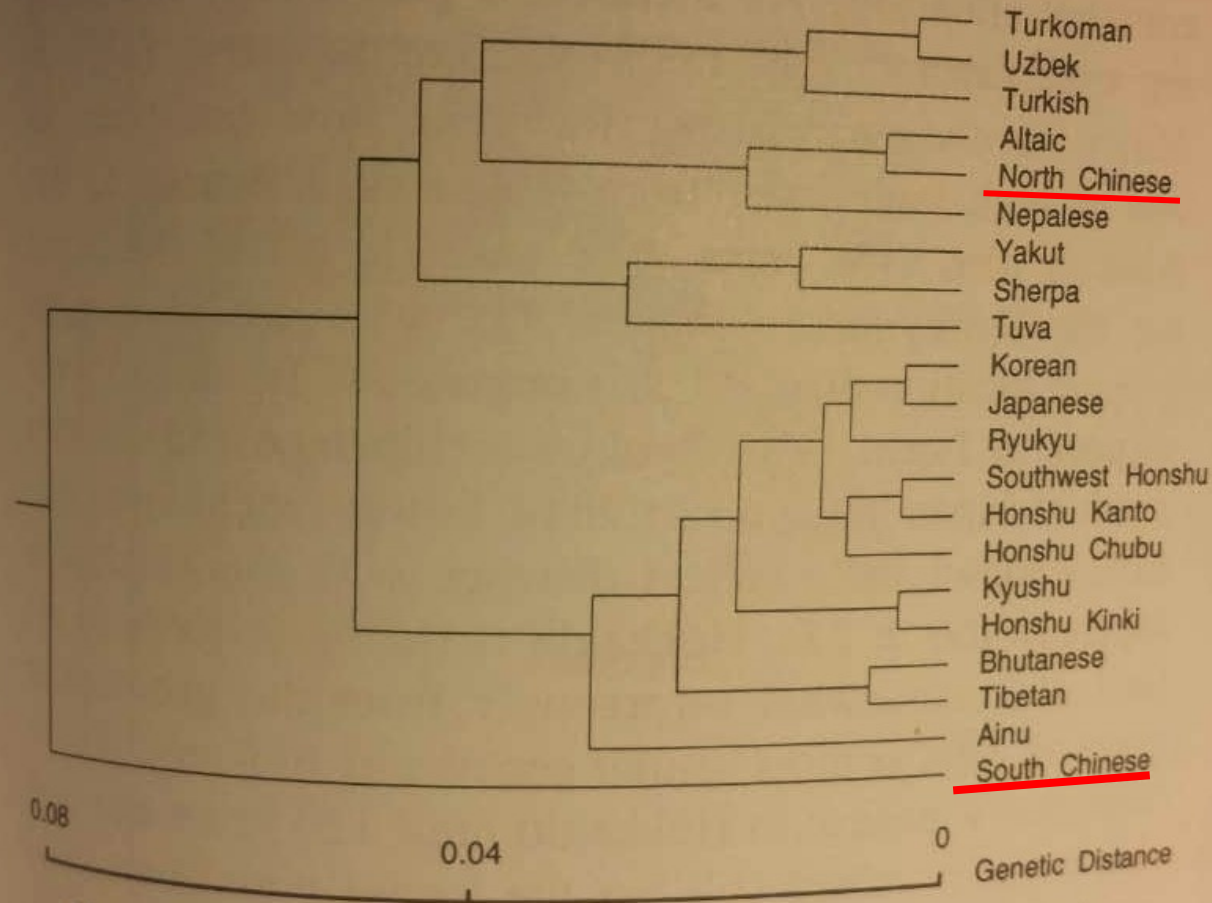


Fig. 4.12.1 Genetic tree of 21 populations from East and Central Asia.

你是南方人吗？

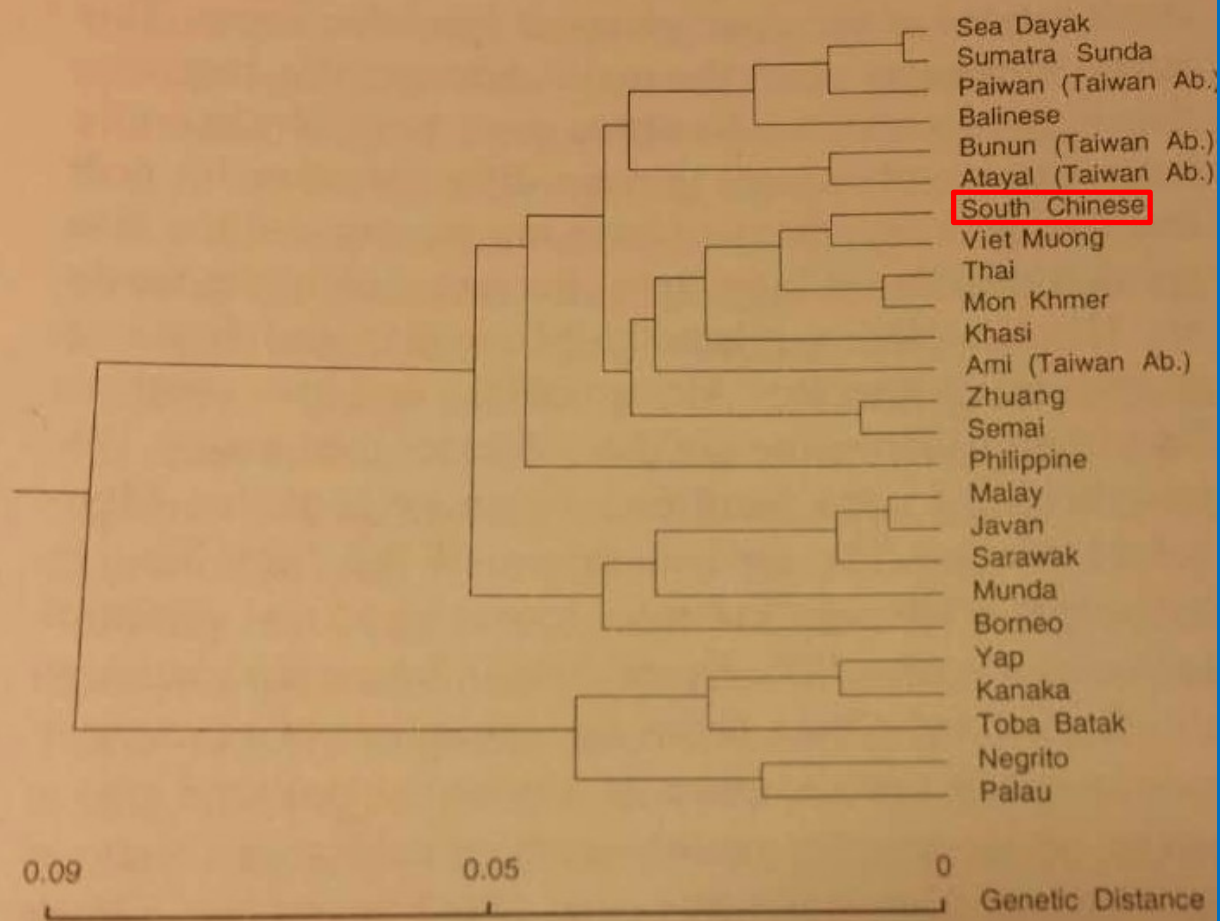


Fig. 4.13.1 A genetic tree of 25 populations in Southeast Asia. Taiwan Ab., Taiwan Aboriginal.

# 什么人最快乐？

- 被人需要(由于你的专长，能力，善良，尊重，爱心，品质，性格，智力，分享..... 而非打扮追风穿戴飚车， 也非拥有权力或钱财)
- 关心他人，能够让他人快乐的人是快乐的
- 尊敬别人的人是快乐的
- 诚实坦荡的人是快乐的
- 心胸宽阔的人是快乐的
- 要爱人如己，这是根本

谢谢！