Python for Data ~~Analysis~~ Science

Ling Liu
lingliu@swufe.edu.cn
SWUFE
Week 1
2020-2021-1

Today’s Topics

- Introduction to …
  - data science
  - the course
  - Python

WHAT IS DATA SCIENCE

Data Science

A central challenge of data science is to make reliable conclusions using partial information

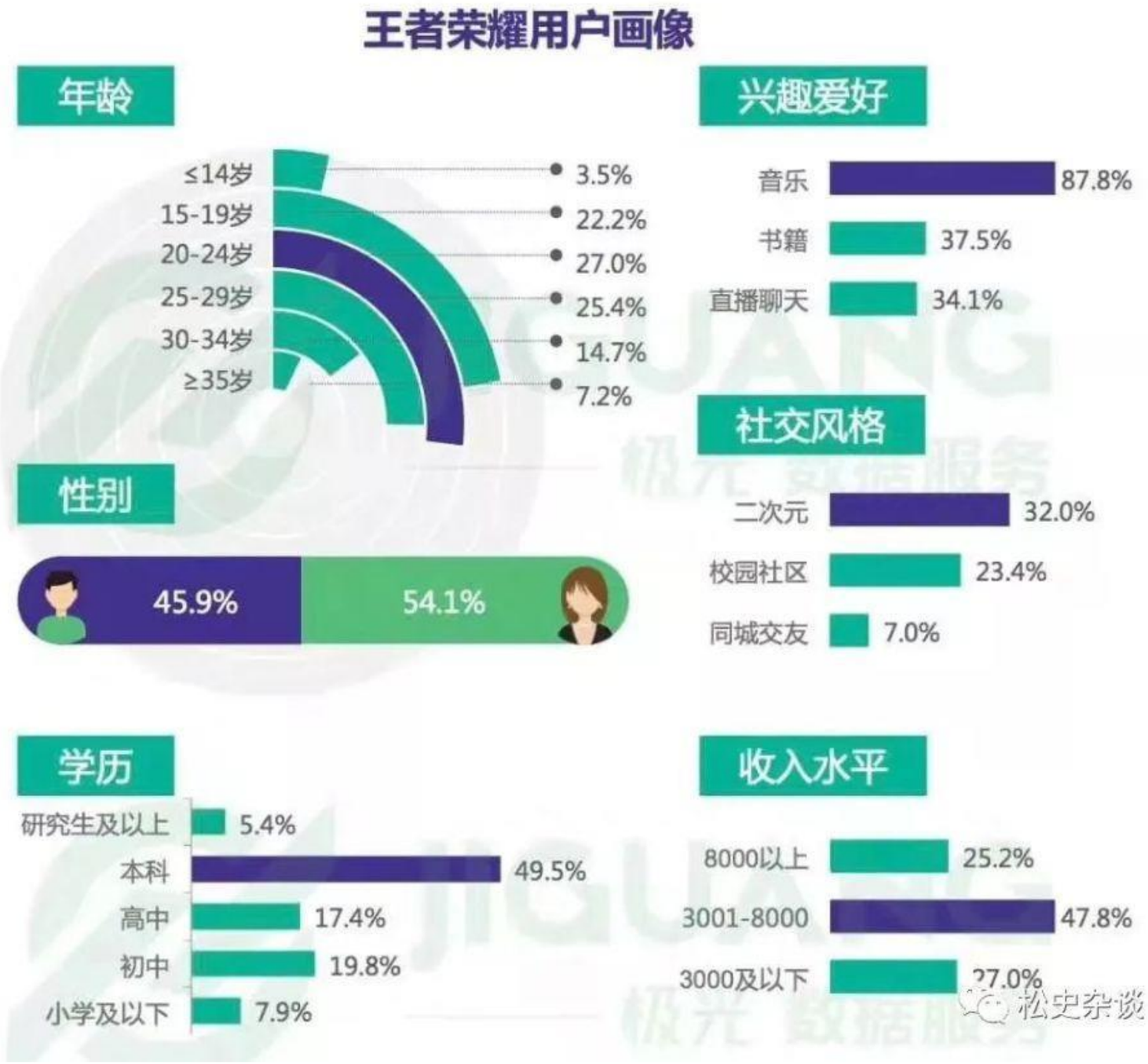
Some Examples

1. The average height of all human being?
2. Common baby names in the most recent 5 years?
3. Is eating chocolate healthy?
4. How Family Background Influences Student Achievement?
5. Can social media sentiment predict stock market movement?
6. Does pollution have impact on people’ s expected life?

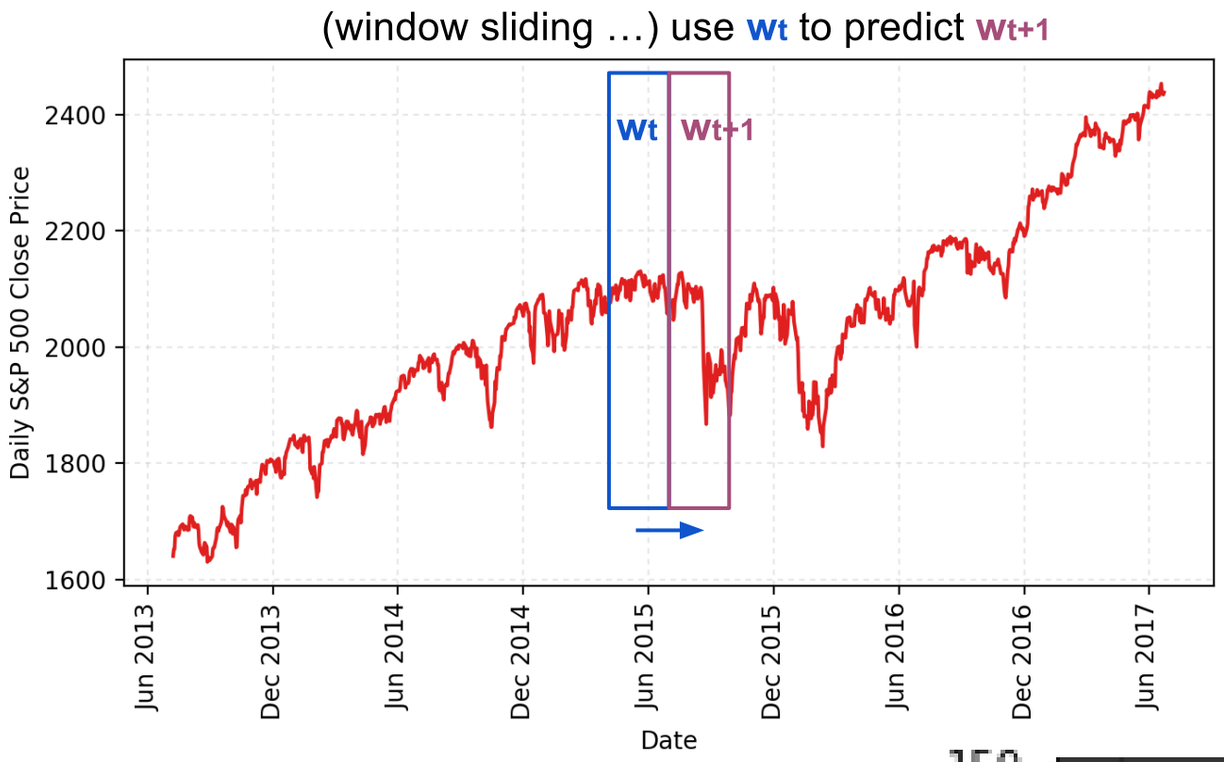
Aims of Data Analysis

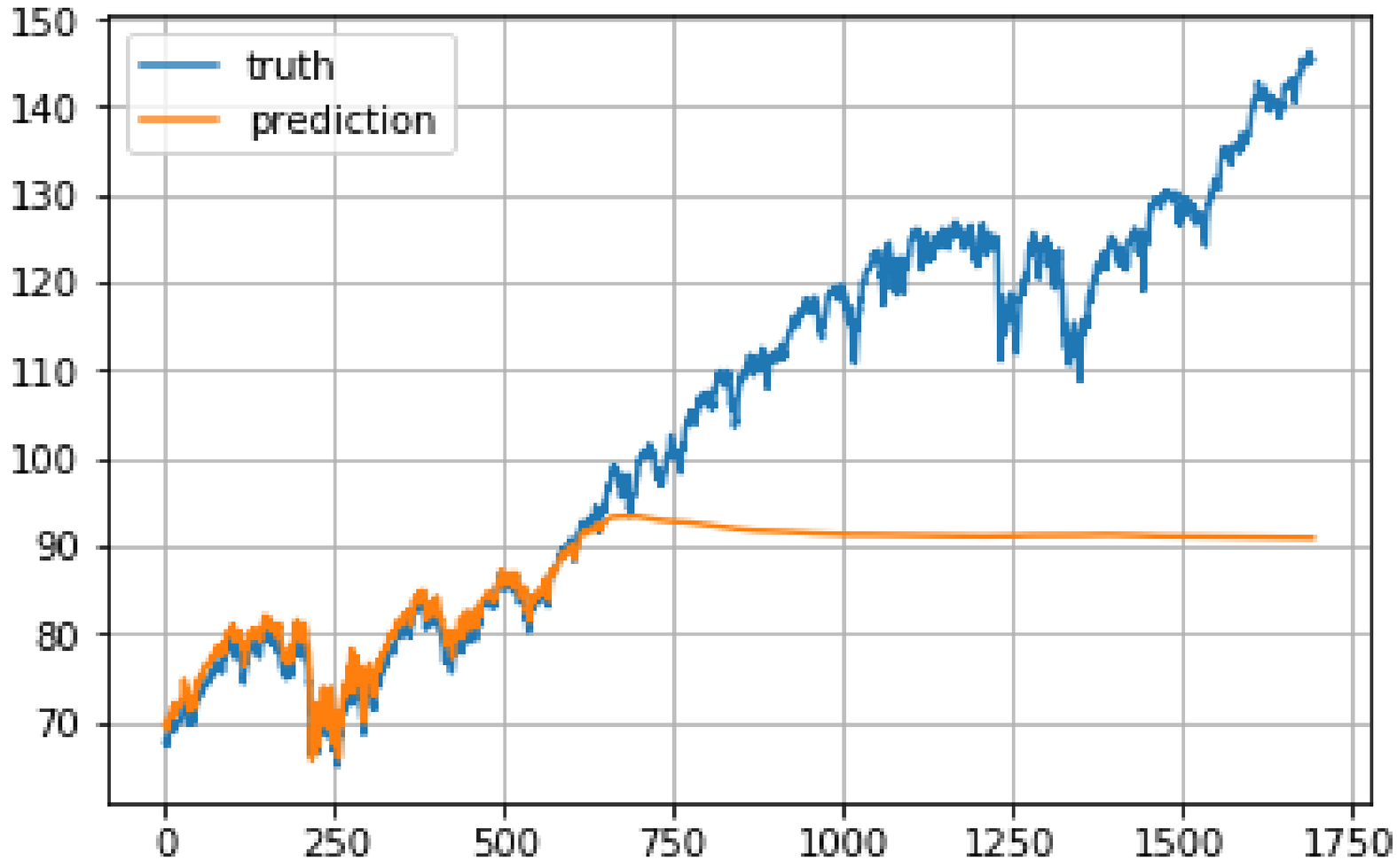
- Exploration
- Prediction/Inference
- Cause Effect Analysis

Exploration

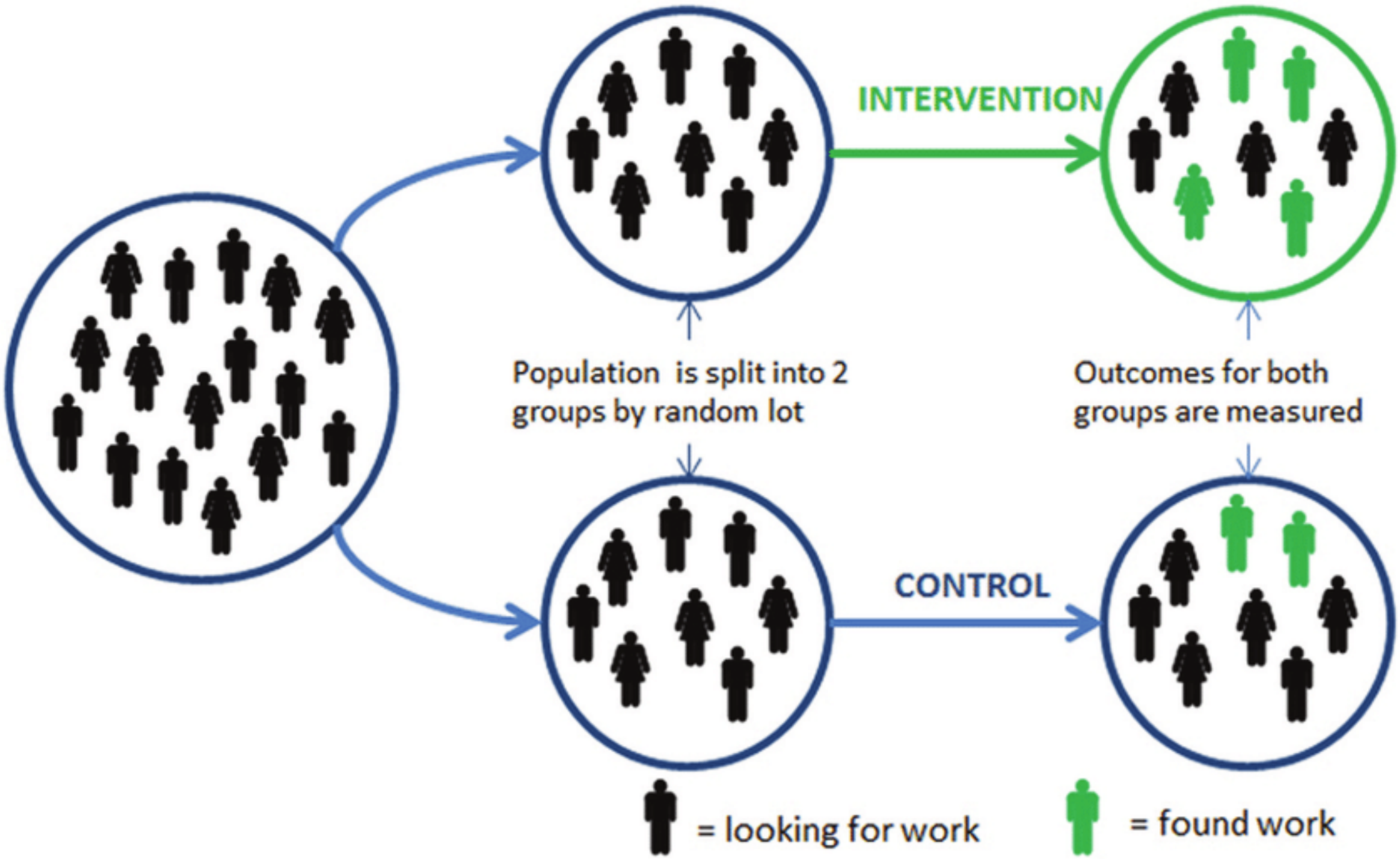


Prediction

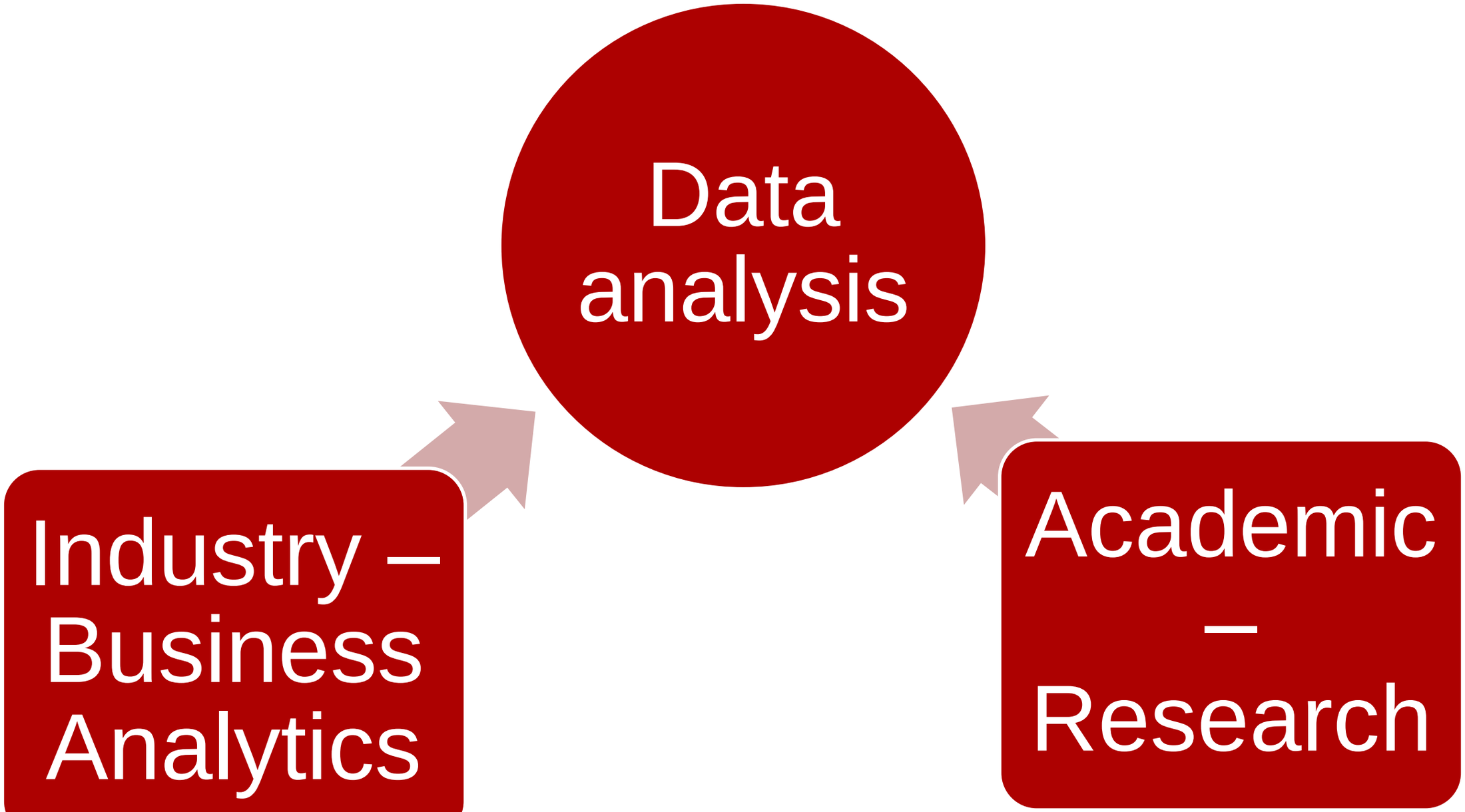

(window sliding …) use Wt to predict Wt+1
Wt
Wt+1
Daily S&P 500 Close Price
2400
2200
2000
1800
1600
Jun 2013
Dec 2013
Jun 2014
Dec 2014
Jun 2015
Dec 2015
Jun 2016
Dec 2016
Jun 2017
Date


truth
prediction
150
140
130
120
110
100
90
80
70
0
250
500
750
1000
1250
1500
1750

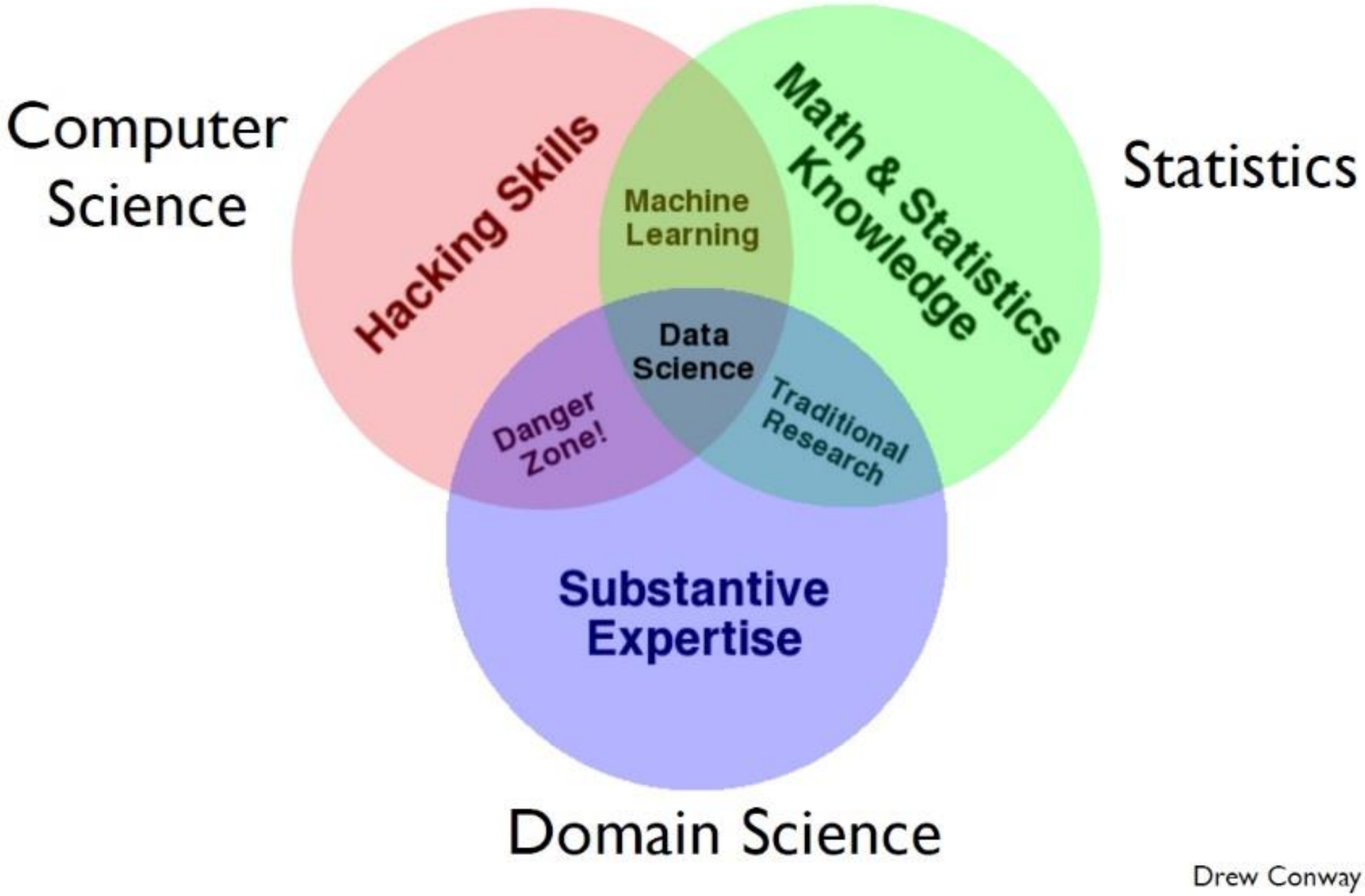
Explanation - Causal Effect



Applications

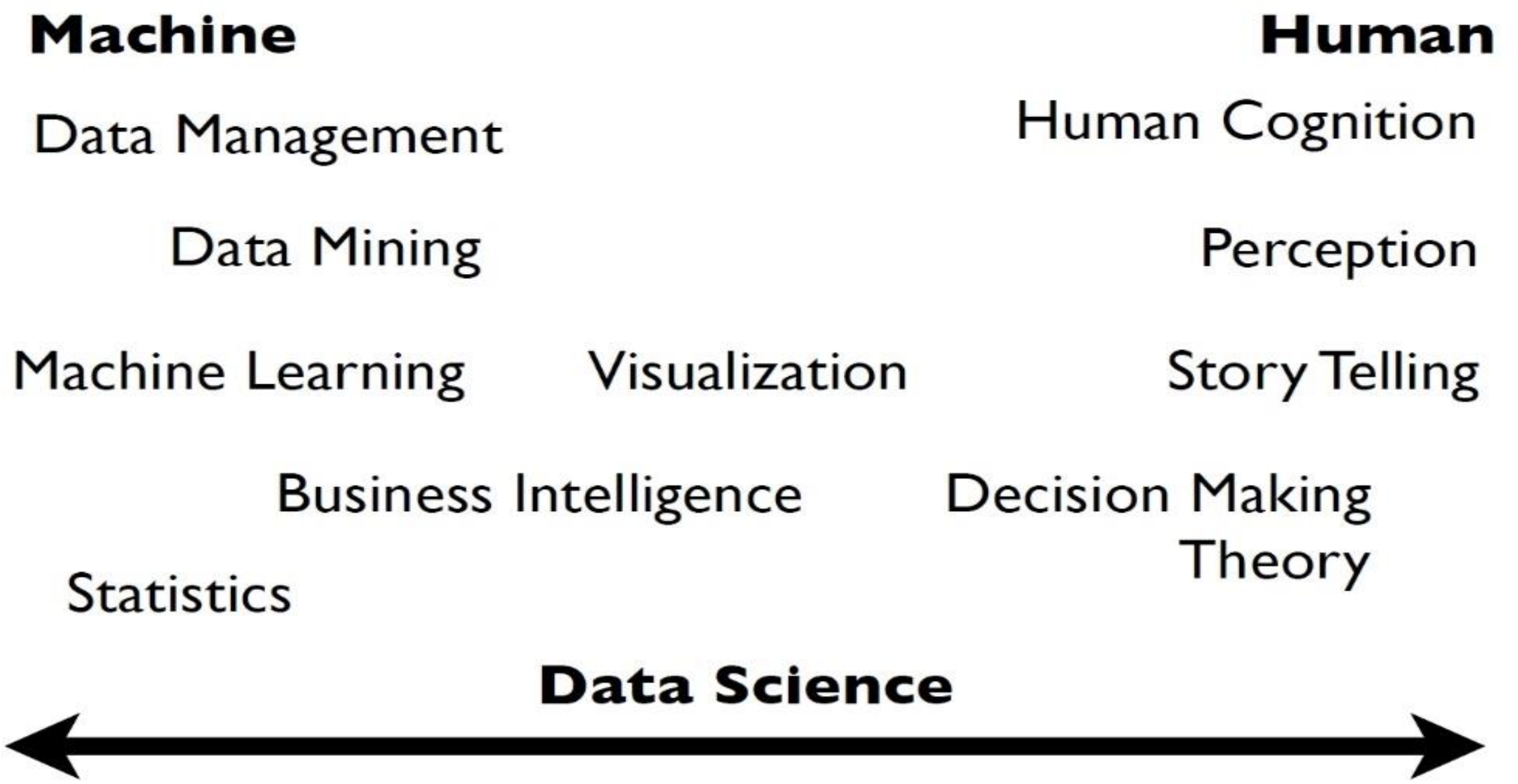


Data Science



from: Harvard CS 109 Data Science.
http://cs109.github.io/2014/index.html

Data Science



Inspired by Daniel Keim, "Visual Analytics: Definition, Process, and Challenges"

from: Harvard CS 109 Data Science.
http://cs109.github.io/2014/index.html

WHY PYTHON

Why Python

- Have advantages in every stages in data analysis
- Simple to use
  - script language
  - simple syntax
- Ecosystem – scientific computing
  - Open source
- Generalization
- Problem：
  - a little bit slower than C/C++

‘Hello world’ in pure machine code

```
b8    21 0a 00 00   #moving "!\n" into eax
a3    0c 10 00 06   #moving eax into first memory location
b8    6f 72 6c 64
a3    08 10 00 06
b8    6f 2c 20 57
a3    04 10 00 06
b8    48 65 6c 6c
a3    00 10 00 06
b9    00 10 00 06
ba    10 00 00 00
bb    01 00 00 00
b8    04 00 00 00
cd    80
b8    01 00 00 00
cd    80
```

6502 Assembly [edit]

```
; goodbyeworld.s for C= 8-bit machines, ca65 assembler format.
; String printing limited to strings of 256 characters or less.

a_cr    = $0d           ; Carriage return.
bsout   = $ffd2         ; C64 KERNAL ROM, output a character to current device.
                        ; use $fded for Apple 2

        .code

        ldx #0          ; Starting index 0 in X register.
printnext:
        lda text,x      ; Get character from string.
        beq done        ; If we read a 0 we're done.
        jsr bsout       ; Output character.
        inx             ; Increment index to next character.
        bne printnext   ; Repeat if index doesn't overflow to 0.
done:
        rts             ; Return from subroutine.

        .rodata

text:
        .byte   "Hello world!", a_cr, 0
```

JAVA

```java
public class HelloWorld
{
    public static void main (String[] args)
    {
        System.out.println("Hellold!");
    }
}
```

C++

```cpp
#include <iostream>
using namespace std;

int main()
{
    cout << "Hello, World!";
    return 0;
}
```

Python

```python
print('hello world')
```

Comparison of data analysis tools

Excel

SPSS

SAS

R

Python

OUR COURSE

Why choose (xiang) this (bu) course (kai) ?

Welcome!

Aim: help you from applying Python to social science, economics, finance and other areas.

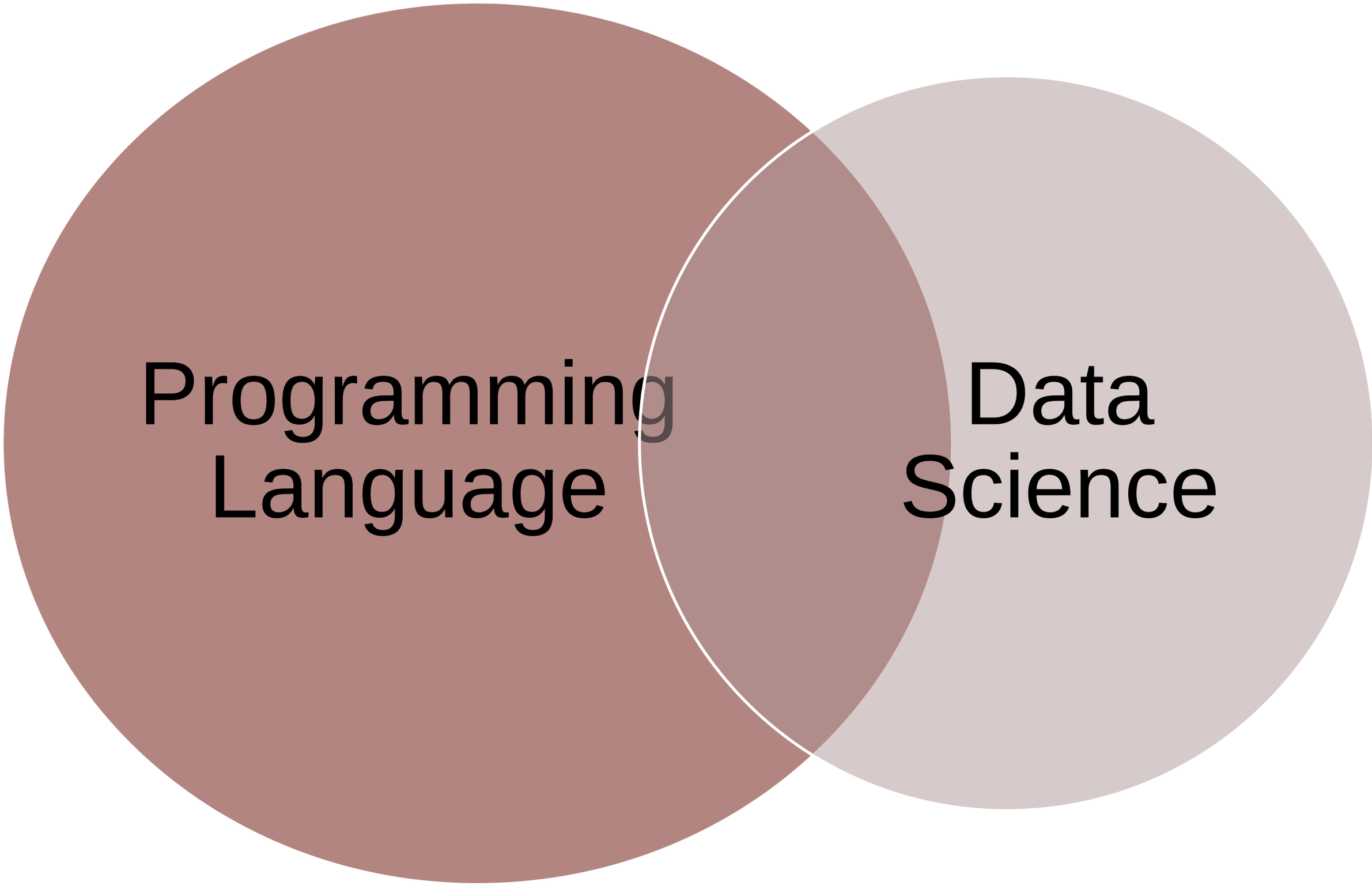
Hope:

1. Think from a data scientist perspective
2. Being given a research question and some help with the models, you will be able to write code to collect, clean and analyze data, as well as report the results.

Learning Objectives

- Programming thinking
  - Writing a program will become your "go-to" solution for data analysis tasks
- Data science thinking
- Get to use Python
  - Including experience with relevant libraries for data manipulation, scientific computing, and visualization.

The course



About me

- 刘凌
- lingliu@swufe.edu.cn
- Something you need to know
  - Chinese/English Slides
  - Rapid speech(will not be improved in the foreseeable future), T_T

放弃了大家还是好朋友。。系列

- 如果…你想水过公选课, 请退课
- 如果…你想拿高分但是没有时间花在这门课上，或者有太多其他需要做的事情，或者…或者是任何一个你不愿意费力气的理由，不要选了

- 这是一门非常有挑战的课程！
  - “此课比专业课还辛苦！”---这是真的！
  - 进度快、要求高、给分低

讲了这么久，总该要个证明吧。。

要不就算了。。系列

- 如果你完全没学过统计、概率、计量、机器学习或数据挖掘中的任意一门课程，要不就算了？
- 如果你一点编程都没学过，要不也算了。。
- 只是为了简历上有个“编程”课
- 非常不喜欢编程

请自学。。系列

- 如果…你已经非常了解并熟练掌握Python, pandas,但想增强大数据分析的能力,比如mapreduce和hadoop,很遗憾本课程并不涉及这些内容
- 如果… 你想学习机器学习、深度学习，这个课程也不适合你

也许还挺适合你。。系列

- 学过Python或任意一门编程语言，以后需要多用编程
- 有一定的自学能力
- 想为未来升学、工作学些技能

What do they say?

平心而论，这次课程论文是我本学期收获最大的一项作业。因为对这个课题一开始就很感兴趣，看了相关论文总是想着自己能不能做做看，哪怕是复刻也好。但是难度也是有的，参考文献都没有提供具体 MD&A 的文本处理方法，没有词典，我只能自己编，没有代码（当然，这是肯定的哈哈哈），我只能自己写。3 个月以前，我还对编程既崇拜而又一无所知，现在我已经能自己装包搜索写（拷）代码，好好地研究问题了。

这次研究的前期准备非常耗时间，一开始我以为年报 pdf 转 txt 非常简单，之后才发现很困难，所有 pdf 转换了大概有三天，中间断了四五次，完成的时候很感恩。

对于这次课程论文，我现在有以下几个心得：

1. 研究最重要的是思路，而不是工具。不要做工具的奴隶。研究过程我克服的困难大多是技术性的。研究的几个核心困难，我没有完全克服（也是因为有点超出能力范围了），比如词典到底要编成什么样才算好，我计分的算法到底合不合理（我觉得还是有改进的空间），为什么我回归出来效果不好，怎么把拟合优度做得像参考文献一样好，这些问题其实才是研究的核心问题。虽然我在论文里讨论过一点，但是还有深入和改进的空间。
2. 检索能力非常重要。我在研究里遇到的 95%以上的问题都是以前没有遇到过的，但是我都强行解决了，全部是靠搜索出来的。搜索（尤其用英文）可以找到很多好东西。国外在开源方面真的太强大啦。
3. 培养对知识的好奇心。好奇好奇，再好奇！

给未来修读这门课的同学的建议：

如果任务很重就不要选了，反正慎重吧，但是选了就要坚持把任务做完，其实也没有那么难。我觉得这门课是对 python 技能的很好巩固与拓展机会，能很好弥补学校 python 基础课不够的地方，让 python 真正对你的学习、学术起到帮助作用。

程序其实就是得花时间写，要是有毅力，自己多写肯定是最好的，也没必要选课。但如果不能，那倒强烈推荐选课，这样就可以“被动”写大作业了哈哈哈哈。

啊!!! 一定一定要尽早做
配时间太少了，直接导致

这门课中真的非常非常感谢老师，因为自己一直想学习一些数据处理的技能，所以就义无反顾的选了这么课，虽然过程中伴随的许多后悔，但你的代码一行一行敲出来，结果不在报错、程序运行出结果的那一刻!!! 那有多么快乐你知道吗？!! 一定要感受!! 公选嘛，就要学一些有用的，我一点都不后悔，我特别特别感谢老师，老师教的好，知识多，爱学生，还有最后的大作业真的很 push 啊，可以逼你自学很多东西!!

还有一定
的难受，一定

还有，遇
这时候善用百

```
14  ----------------------------------------------------------
15  收获与总结（为了保持论文的完整性，所以把与论文无关的个人感悟类summary放到了这里）
16
17  总的来说，收获多、遇到的困难更多。
18  遇到的第一个问题就是数据的获取，数据实在是太多了，最开始用的是resset数据库，结果格式问题
    比较大，本来已经接近完成的代码，又因为换了CSMAR数据库的数据重新再写一遍。
19  除此之外，原数据文件太大，导致电脑内存不够，上网查到可以每次只读一点，于是开始分块读取，
    算是基本上解决了这个问题。还有数据合并等等一系列问题。
20
21  收获相对来说就更加丰厚了，以前从未想过能处理这么大的数据量，以前的编程只停留在简单的算法
    设计上，而这次相当于是处理一个实际问题，实战经验提升飞快，也帮助我复习了python（另一个py
    thon课要考试）而整个过程和自己最开始所想的相差太多，事情没有我想象的那么简单，数据处理这
    一步耗费了太多时间和精力，导致原定计划完全完成不了，真就计划赶不上变化。
22
23  对未来同学的建议：
24  如果不想努力，慎选！
25  如果只是想水个学分，别选！
26  如果想真正学到东西，练习自己的python，一定要选！
27  但是如果认真完成，收获绝对超乎你的想象。
28
```

Arrangements

Python basics and text analysis (5-7 weeks)

- string, list, control flow, function, etc
- Data collection
- Text analysis

Data cleaning (3-4 weeks)

- pandas, matplotlib, seaborn, numpy
- Data merge, pivot table, etc

Data analysis，models, research (6-7 weeks)

- Statsmodels, linearmodel, sci-kit learn
- Machine learning, statistics
- final projects

Course materials

- 《深入浅出数据科学》
  - 作者： [美] 斯楠·奥兹德米尔（Sinan Ozdemir），译者： 张星辰，编辑： 王峰松，人民邮电出版社，978-7-115-48126-9
- 《Python数据科学手册》
  - 人民邮电出版社，美国O’reilly出版社，Jake VanderPLas著，陶俊杰，陈晓丽译。
  - 英文免费在线书籍和代码参考（https://jakevdp.github.io/PythonDataScienceHandbook/）

课程地点

- J 212
- I401B

Course logistics

- 课程中心: 课件、作业提交、相关资料下载
- 我(那也许能坚持下去)的个人主页
  - https://ling60.github.io/Python-for-Data-Analaysis/

Tests and Examines

Course activities

10%

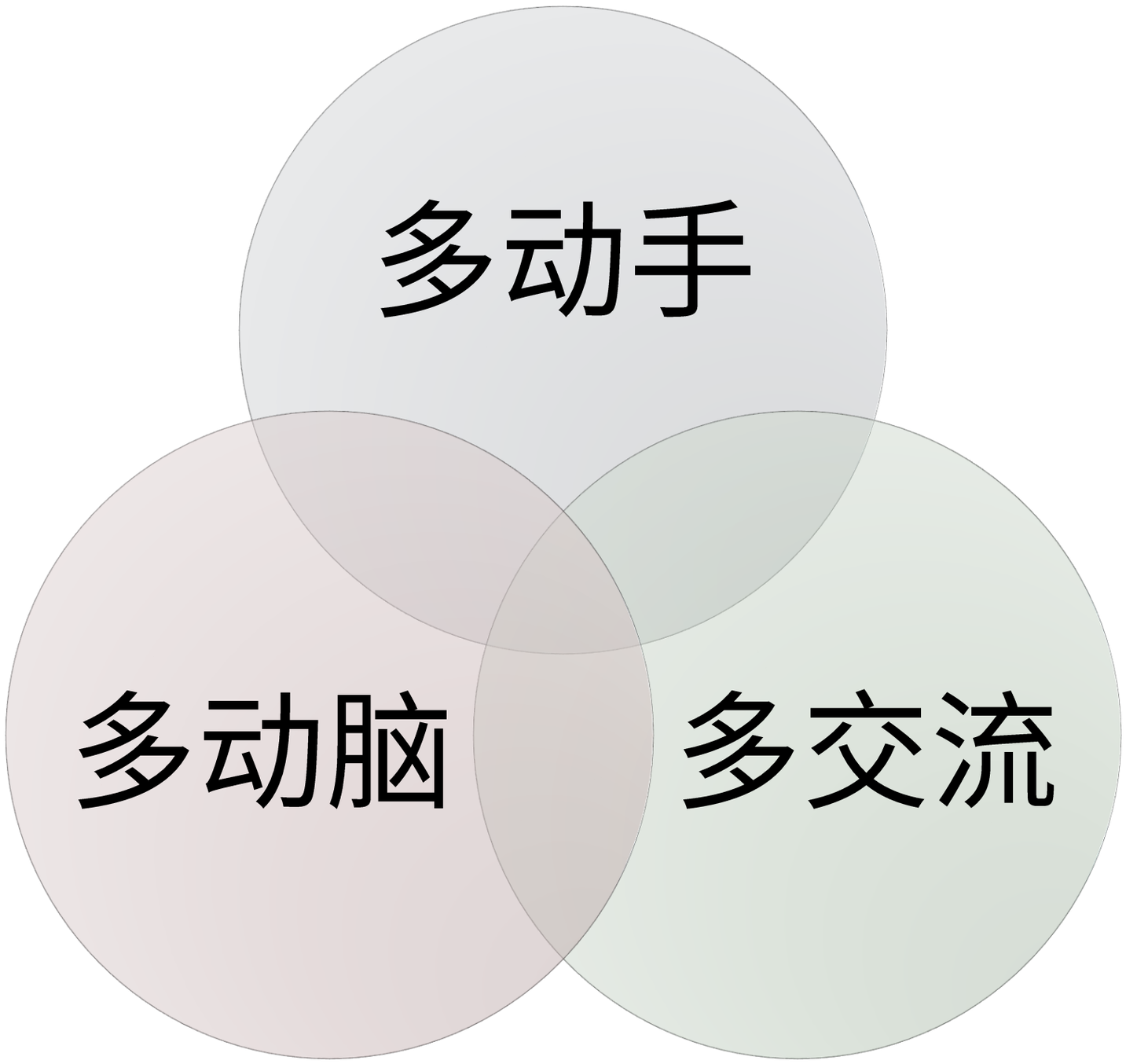
Test and homework

30%

Final Project

60%

Please…



Academic Integrity

- 诚实, 不抄袭 是做人的基准
- All your submissions must be your own work
  - Cite your source, explain any conventional action
- 不可以抄袭别人的作业/工作
  - 鼓励讨论
  - 有问题, 请提问

Any questions?

PYTHON INTRODUCTION

Install

- Please install Python3
- Install with packages
  - Anaconda: https://store.continuum.io/cshop/anaconda/
- Python stand alone
  - http://python.org/download

How to use Python

- CMD
  - ipython
- .py
- IDE
  - console(控制台)
  - editor(编辑器)

The Jupyter Notebook (Ipython Notebook)

- http://jupyter.org/
- jupyter notebook

- Top 10 Ipython Notebook Tutorials for data science and machine learning
  - http://www.kdnuggets.com/2016/04/top-10-ipython-nb-tutorials.html

IDE

- 集成开发环境
  - IPython + Text editor(notepad++, vim, etc)
  - Anaconda + Spyder
  - PyCharm
  - Enthought Canopy

PYTHON DATA TYPES

object

- 对象
- “In computer science, an object can be a variable, a data structure, or a function, and as such, is a location in memory having a value and possibly referenced by an identifier.”

Python - object

- Everything in Python is an object
- Every object has a “type”
  - int
  - float
  - string
  - bool

Data types/structure

- Int/float/string
- List/dic/set/tuple
- If you import a package, it might have their own-defined data types

Variable

```
7  b = 6
8  print(b)
9  b = 'hello'
10 print(b)
```

Variable types is changeable

Variable Naming

- x =3
- value assignment ("=")
- Only letters, numbers and underline can be used as variable names
- Digit numbers cannot be the first character of the name
- Which of the following names are legal?
  - xyz, kols, _123, 1hju, _tu&, Ad6

python built-in names

```
False      class      finally    is         return
None       continue   for        lambda     try
True       def        from       nonlocal   while
and        del        global     not        with
as         elif       if         or         yield
assert     else       import     pass
break      except     in         raise
```

variable names

- not allowed:
  - x: = 1.0
  - 1X = 1
  - x-1 = 1
  - for = 1
- You can also do this:
  -

```
x, y, z = 1, 3.1415, 'hello'
```

type

```
type(1)
type(1.4)
type('hello')
type(True)
type(None)
type('False')
```

Homework

- Install Python/Anaconda
  - Read the provided document
- Be familiar with jupyter notebook, iPython and Spyder
- try install new packages
- Try basic operations

Next week

- 机房I401B
- Python basics
  - Data type
  - Variables
- String, list, dictionary, etc