



USING PYTHON FOR DATA SCIENCE COMPETITIONS

CHRIS CHEN

- ▶ Sr. Data Analyst - Shaw Communications, Canada
- ▶ Data Architect - China Telecom
- ▶ Top 0.1% @KAGGLE
- ▶ yy.chen.ca@gmail.com
- ▶ https://www.linkedin.com/profile/preview?locale=en_US&trk=prof-0-sb-preview-primary-button

<https://www.linkedin.com/in/chris-chen-44414332>

ChrisCC

A HUGE WAVE OF MESSY DATA IS APPROACHING



Verified
account

KAGGLER



?

Highest†
314th

Current†
325th
/416,440



16,480.6 points

Joined 8 months ago

†Ranking method changed 13 May 2015 (?)

Profile

Results

Scripts

Forum

Account

Activity

Edit Profile

TOP 10%



11th/2225

TOP 10%



25th/2236

TOP 25%



235th/985

WHY?

KAGGLE – MAKING DATA SCIENCE A SPORT

- ▶ Kaggle是目前世界上最有影响力的数据建模和数据分析竞赛平台
- ▶ 企业及研究机构：发表亟待解决的问题并提供数据
- ▶ 统计学者/数据挖掘专家：参与竞赛以产生最佳模型
- ▶ 合作伙伴：FACEBOOK, NASA, AMAZON, WALMART, WIKIPEDIA, CAPITALONE.....
- ▶ 以及，来自全球范围的40万数据科学家



LEARN FROM THE BEST

- ▶ 绝顶高手
- ▶ 论坛
- ▶ Scripts

WORK WITH COOL DATASETS



Heritage Health Prize: 预测哪些患者会在未来一年内住院治疗。



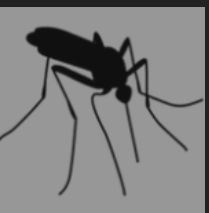
GE Flight Quest: 设计一种可规模化的算法，用于生成实时的飞行路线信息以帮助飞行员更有效安全的完成飞行。



Springleaf Marketing Response: 预测客户对邮件营销的响应程度。



Facebook Recruiting VI: 甄别出拍卖网站参与者中的机器人。










West Nile Virus Prediction: 预测在何时、何地，当地蚊子会被检测出携带有西尼罗河病毒。



Allen AI Science Challenge: 你的算法会比8岁的小学生更聪明吗？

当然少不了奖金.....

		Western Australia Rental Prices  Predict rental prices for properties across Western Australia	6.7 days 59 teams \$100,000
		The Allen AI Science Challenge Is your model smarter than an 8th grader?	2 months 382 teams \$80,000
		The Winton Stock Market Challenge Join a multi-disciplinary team of research scientists	2 months 499 teams \$50,000
		Rossmann Store Sales Forecast sales using store, promotion, and competitor data	20 days 2844 teams 1804 scripts \$35,000
		Prudential Life Insurance Assessment Can you make buying life insurance easier?	2 months 61 teams 32 scripts \$30,000

以及，工作机会



★ Manager - Cognitive Technologies - Advanced Analytics Enablement

Deloitte Consulting · National
posted 10 days ago

1,082
views



★ Quantitative Developer

AssetMetrix · Munich, Germany
posted 10 days ago

638
views



Gartner

★ Managing VP - Head of Data Science

Gartner · Stamford, CT
posted 20 days ago

936
views



★ Data Scientist

1st Merchant Funding · Miami
posted 21 days ago

1,239
views



★ Director of Sports Modelling

Paddy Power Plc · Dublin, Ireland
posted 21 days ago

788
views

工具?

你的刀够快吗？



XGBOOST, VOWPAL WABBIT, PANDAS, SCIKIT-LEARN, THEANO

数据？

来自各大公司以及研究机构



不管是大数据， 还是小数据



ZIP ~ 6GB



ZIP ~ 1MB

都是人民群众喜闻乐见的

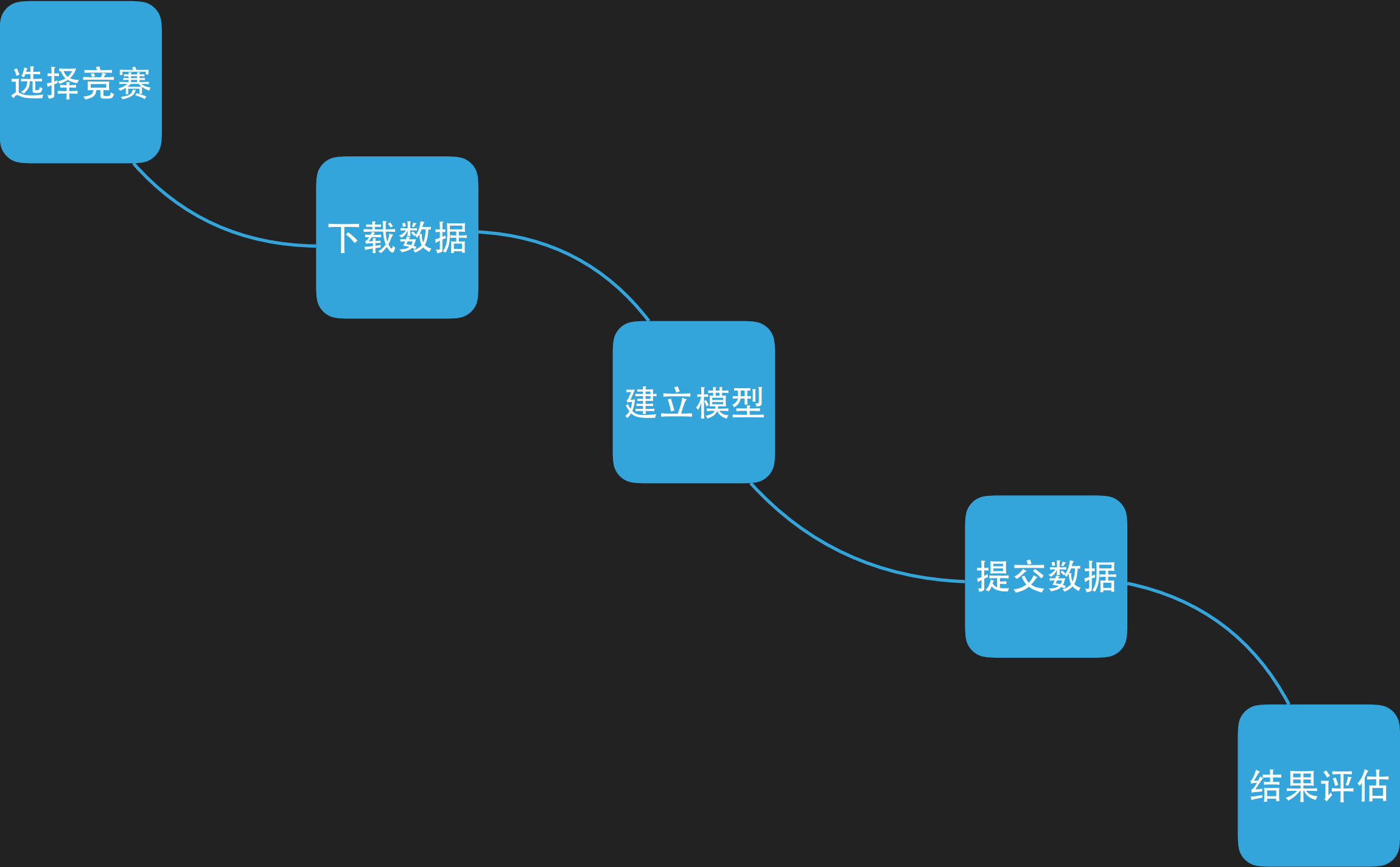
- ▶ Customer Data
- ▶ Log Files
- ▶ Timeseries
- ▶ Web Pages
- ▶ Images
- ▶ Reddit Comments

以及



如何参与？

基本流程



有用的资源

- ▶ Kaggle Forum
- ▶ Kaggle Blog
 - ▶ <http://blog.kaggle.com>
- ▶ Kaggle Competition: Where and how to begin
 - ▶ <http://www.analyticsvidhya.com/blog/2015/06/start-journey-kaggle/>
- ▶ Feature Engineering
 - ▶ <http://machinelearningmastery.com/discover-feature-engineering-how-to-engineer-features-and-how-to-get-good-at-it/>
- ▶ Kaggle Ensembling Guide
 - ▶ <http://mlwave.com/kaggle-ensembling-guide/>





QUESTIONS?
