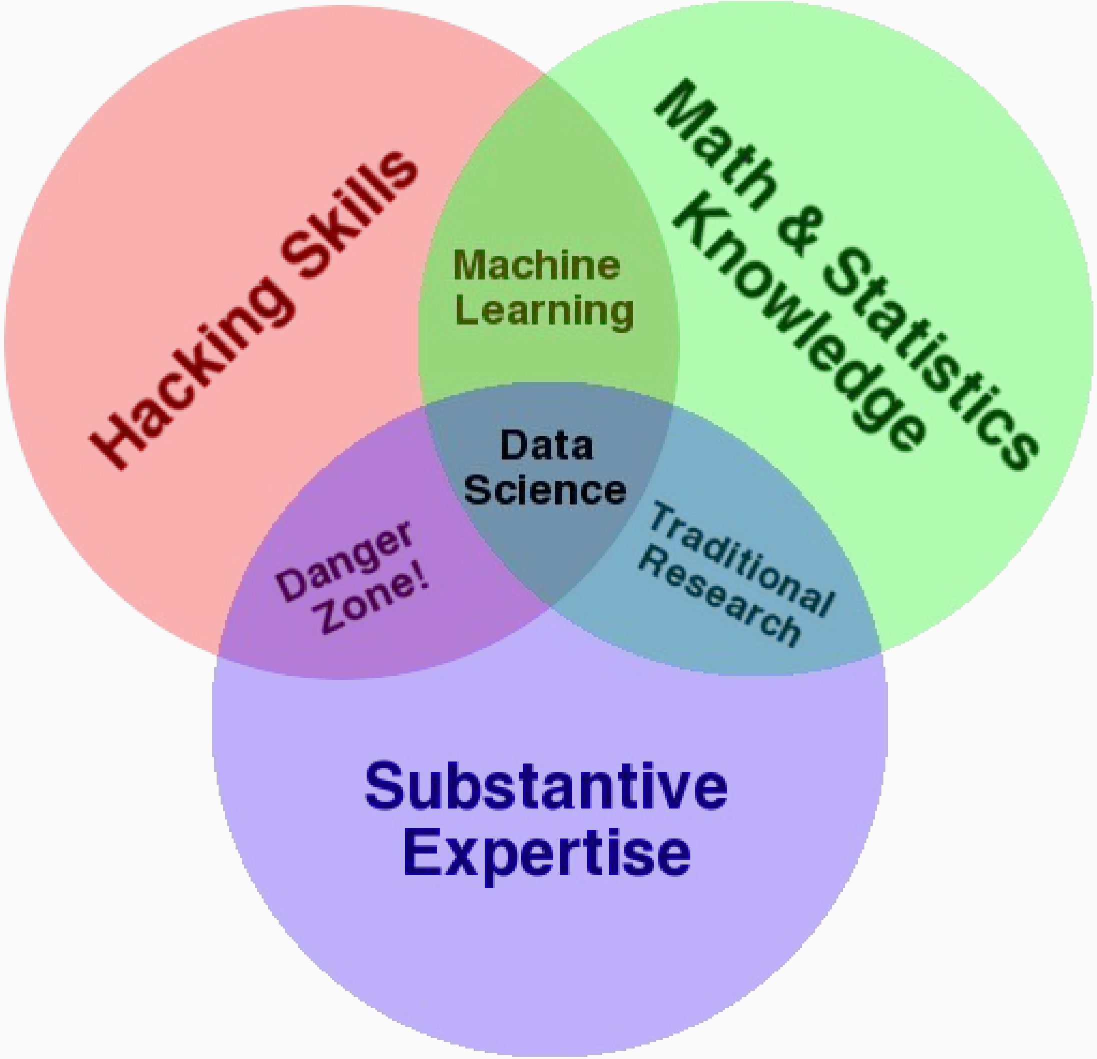
Welcome to dotDATA

Wayne Eternika, President
Steve Mandala, Vice President

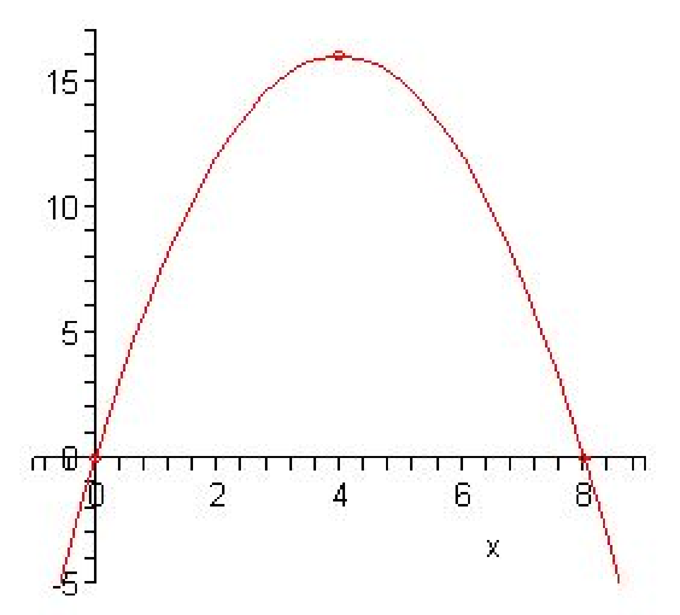
Data
Science

WHAT IS DATA SCIENCE?

Mission:
Help You Converge



Motto: Keep it Parabolic!!!



Teamwork

What we are

Connection to resources
Connection to connections
Learning support + framework
Social support

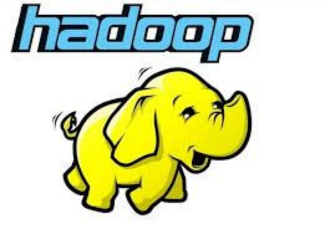
What we aren't

Short clips

(First 38 seconds) https://www.youtube.com/watch?v=HPzXlFp4rKE

(First 3 minutes) https://www.youtube.com/watch?v=8pHzROP1D-w

Technology Behind it:



Where are you going in the datasphere?

Languages in Demand

Be a polyglot

Breakout Groups

Doodle

Standby for Doodle Polls

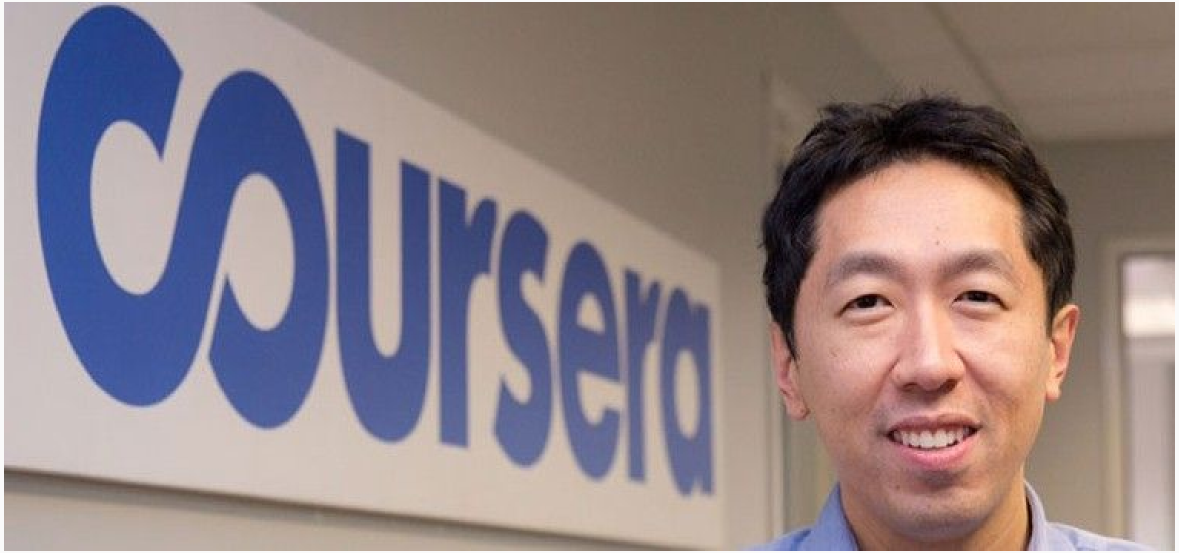
Expect exposure to things you haven’t thought about

Be a leader, be a member, but just be active

Scheduled times will result from polling as well

Breakout Groups are in addition to general meetings

Who is he?



A little thing called Machine Learning

We push MOOCs!

Like dealers push drugs.

Abundance of Information

coursera

edX

udemy

UDACITY

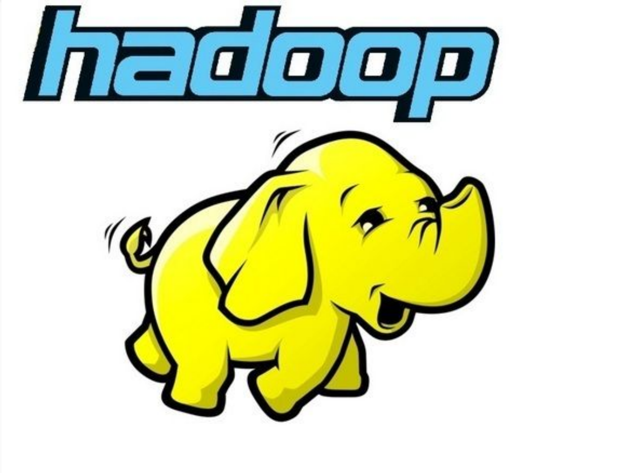
Internships

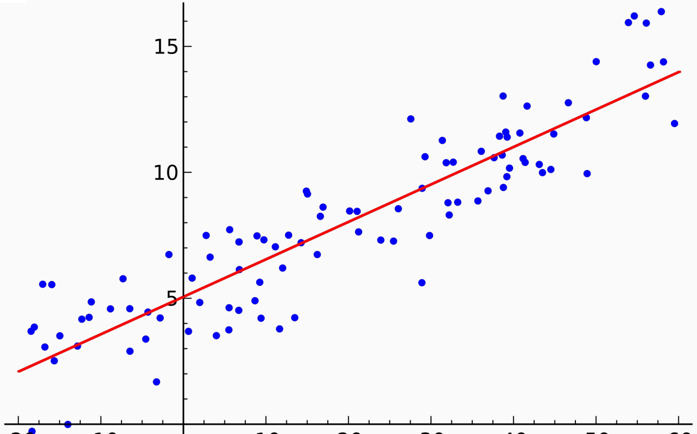
...just today

Recently ..

D E Shaw & Co

Workshops





Workshops

WACM

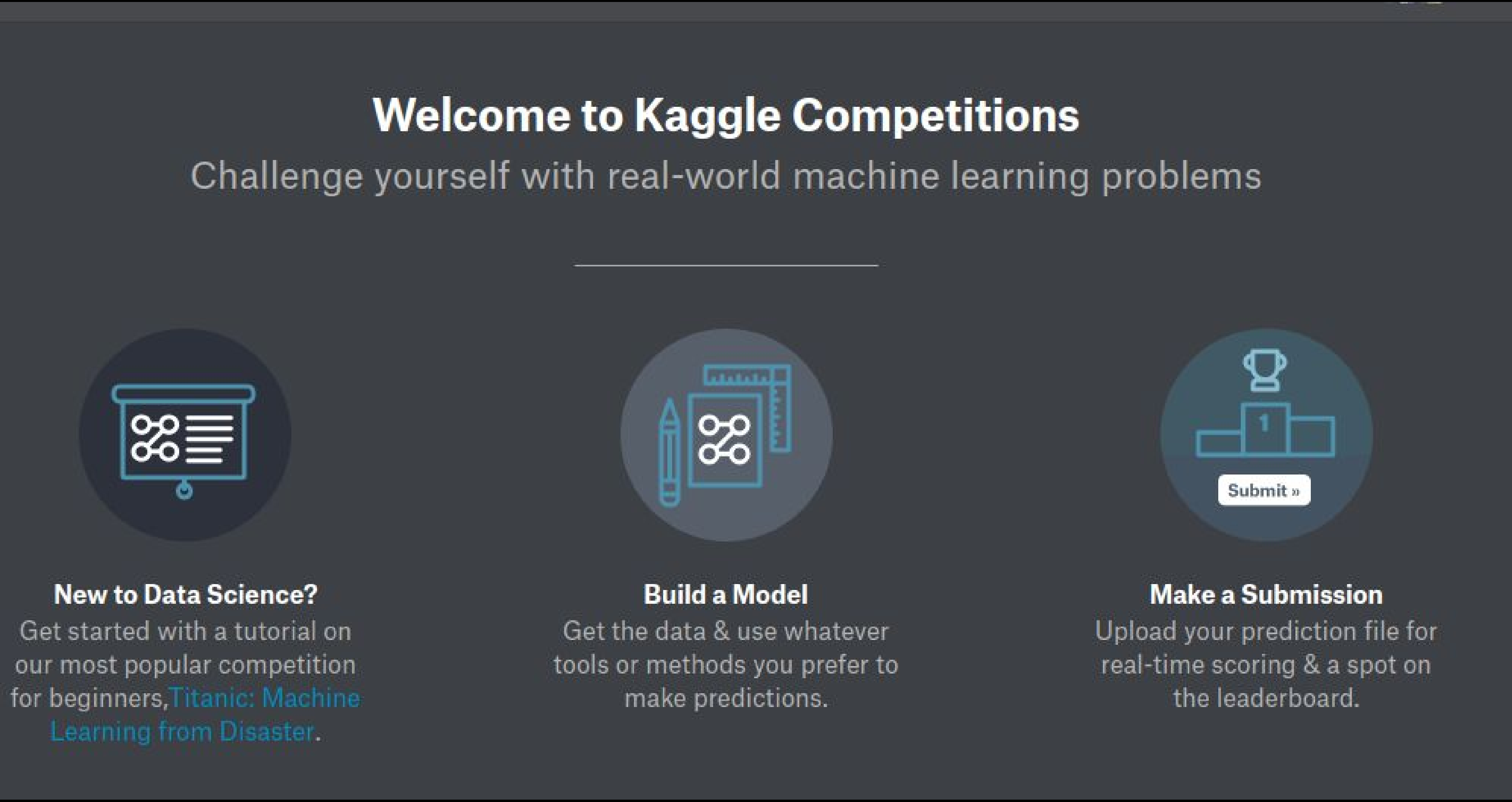
UW-Madison Chapter of acm-w

DoIT

upl

Projects

kaggle™



Welcome to Kaggle Competitions

Challenge yourself with real-world machine learning problems

New to Data Science?

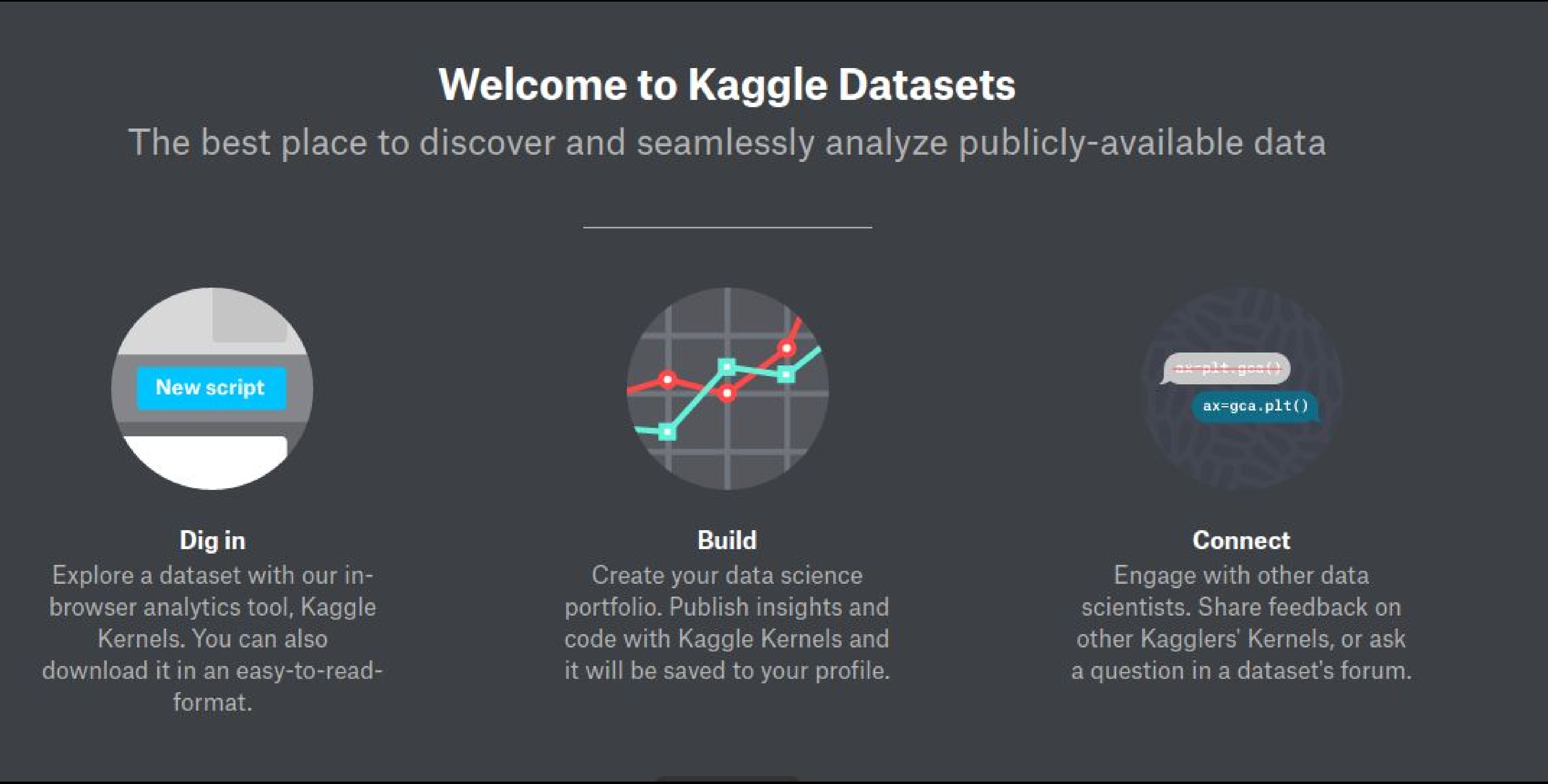
Get started with a tutorial on our most popular competition for beginners,Titanic: Machine Learning from Disaster.

Build a Model

Get the data & use whatever tools or methods you prefer to make predictions.

Make a Submission

Upload your prediction file for real-time scoring & a spot on the leaderboard.



Welcome to Kaggle Datasets

The best place to discover and seamlessly analyze publicly-available data

Dig in

Explore a dataset with our in-browser analytics tool, Kaggle Kernels. You can also download it in an easy-to-read-format.

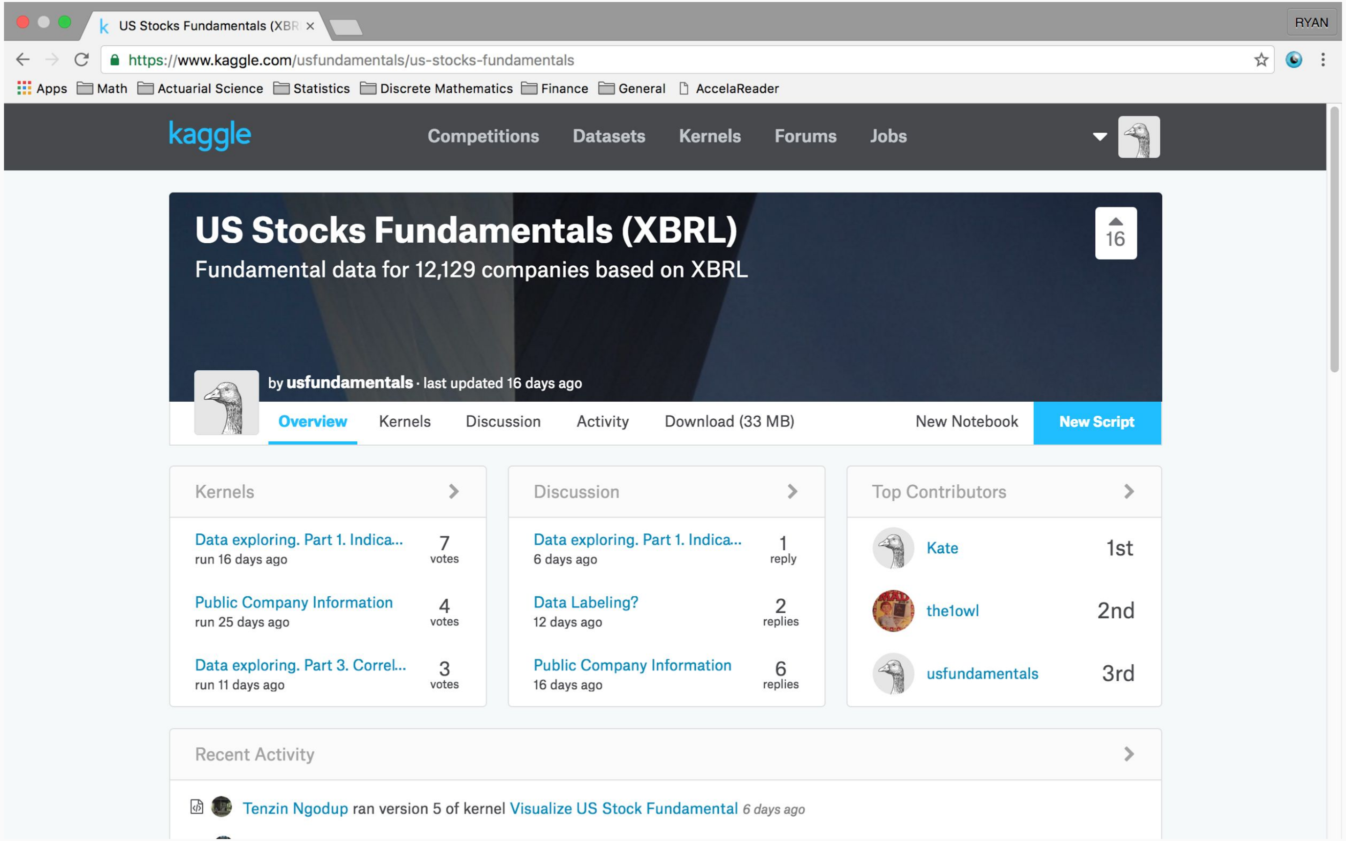
Build

Create your data science portfolio. Publish insights and code with Kaggle Kernels and it will be saved to your profile.

Connect

Engage with other data scientists. Share feedback on other Kagglers' Kernels, or ask a question in a dataset's forum.

Domain: Stocks



US Stocks Fundamentals (XBRL) – https://www.kaggle.com/usfundamentals/us-stocks-fundamentals

US Stocks Fundamentals (XBRL)

Fundamental data for 12,129 companies based on XBRL

by usfundamentals · last updated 16 days ago

Overview | Kernels | Discussion | Activity | Download (33 MB) | New Notebook | New Script

Kernels

Data exploring. Part 1. Indica... run 16 days ago	7 votes
Public Company Information run 25 days ago	4 votes
Data exploring. Part 3. Correl... run 11 days ago	3 votes

Discussion

Data exploring. Part 1. Indica... 6 days ago	1 reply
Data Labeling? 12 days ago	2 replies
Public Company Information 16 days ago	6 replies

Top Contributors

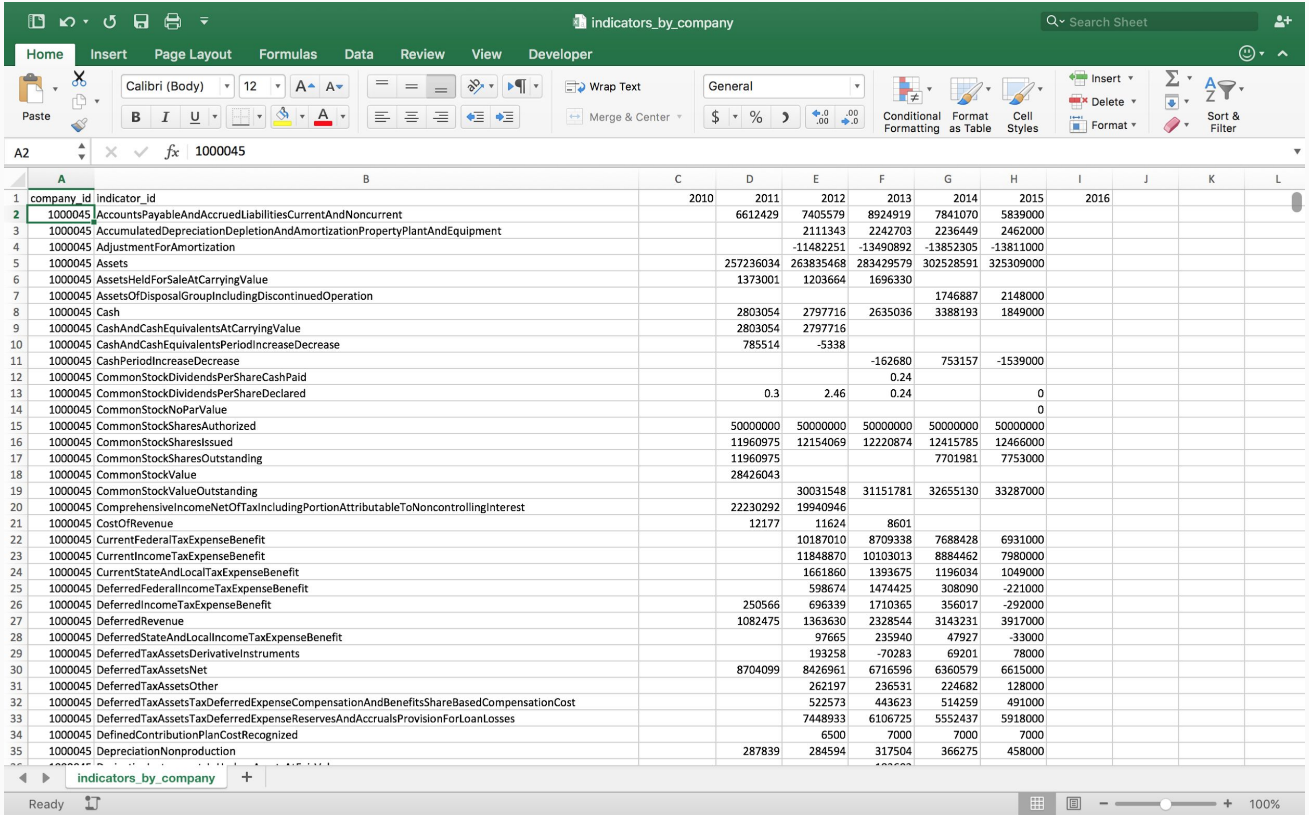
Kate	1st
the1owl	2nd
usfundamentals	3rd

Recent Activity

Tenzin Ngodup ran version 5 of kernel Visualize US Stock Fundamental *6 days ago*

Example Data Set

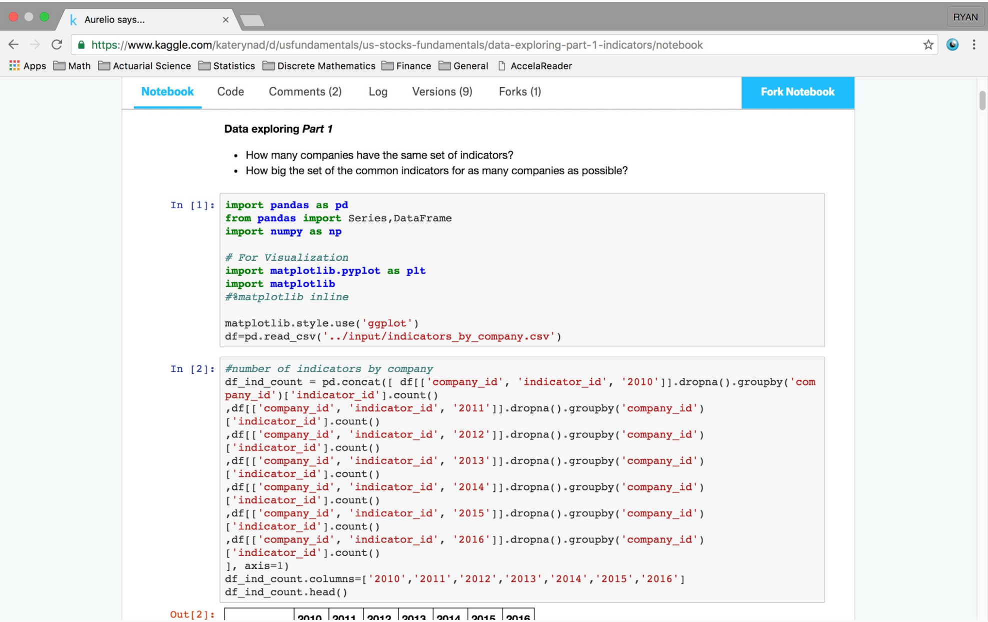
Where are our business majors?



A2 1000045

	A	B	C	D	E	F	G	H	I
1	company_id	indicator_id	2010	2011	2012	2013	2014	2015	2016
2	1000045	AccountsPayableAndAccruedLiabilitiesCurrentAndNoncurrent		6612429	7405579	8924919	7841070	5839000	
3	1000045	AccumulatedDepreciationDepletionAndAmortizationPropertyPlantAndEquipment			2111343	2242703	2236449	2462000	
4	1000045	AdjustmentForAmortization			-11482251	-13490892	-13852305	-13811000	
5	1000045	Assets		257236034	263835468	283429579	302528591	325309000	
6	1000045	AssetsHeldForSaleAtCarryingValue		1373001	1203664	1696330			
7	1000045	AssetsOfDisposalGroupIncludingDiscontinuedOperation					1746887	2148000	
8	1000045	Cash		2803054	2797716	2635036	3388193	1849000	
9	1000045	CashAndCashEquivalentsAtCarryingValue		2803054	2797716				
10	1000045	CashAndCashEquivalentsPeriodIncreaseDecrease		785514	-5338				
11	1000045	CashPeriodIncreaseDecrease				-162680	753157	-1539000	
12	1000045	CommonStockDividendsPerShareCashPaid				0.24			
13	1000045	CommonStockDividendsPerShareDeclared		0.3	2.46	0.24		0	
14	1000045	CommonStockNoParValue						0	
15	1000045	CommonStockSharesAuthorized		50000000	50000000	50000000	50000000	50000000	
16	1000045	CommonStockSharesIssued		11960975	12154069	12220874	12415785	12466000	
17	1000045	CommonStockSharesOutstanding		11960975			7701981	7753000	
18	1000045	CommonStockValue		28426043					
19	1000045	CommonStockValueOutstanding			30031548	31151781	32655130	33287000	
20	1000045	ComprehensiveIncomeNetOfTaxIncludingPortionAttributableToNoncontrollingInterest		22230292	19940946				
21	1000045	CostOfRevenue		12177	11624	8601			
22	1000045	CurrentFederalTaxExpenseBenefit			10187010	8709338	7688428	6931000	
23	1000045	CurrentIncomeTaxExpenseBenefit			11848870	10103013	8884462	7980000	
24	1000045	CurrentStateAndLocalTaxExpenseBenefit			1661860	1393675	1196034	1049000	
25	1000045	DeferredFederalIncomeTaxExpenseBenefit			598674	1474425	308090	-221000	
26	1000045	DeferredIncomeTaxExpenseBenefit		250566	696339	1710365	356017	-292000	
27	1000045	DeferredRevenue		1082475	1363630	2328544	3143231	3917000	
28	1000045	DeferredStateAndLocalIncomeTaxExpenseBenefit			97665	235940	47927	-33000	
29	1000045	DeferredTaxAssetsDerivativeInstruments			193258	-70283	69201	78000	
30	1000045	DeferredTaxAssetsNet		8704099	8426961	6716596	6360579	6615000	
31	1000045	DeferredTaxAssetsOther			262197	236531	224682	128000	
32	1000045	DeferredTaxAssetsTaxDeferredExpenseCompensationAndBenefitsShareBasedCompensationCost			522573	443623	514259	491000	
33	1000045	DeferredTaxAssetsTaxDeferredExpenseReservesAndAccrualsProvisionForLoanLosses			7448933	6106725	5552437	5918000	
34	1000045	DefinedContributionPlanCostRecognized			6500	7000	7000	7000	
35	1000045	DepreciationNonproduction		287839	284594	317504	366275	458000	

Final Report



Data exploring *Part 1*

- How many companies have the same set of indicators?
- How big the set of the common indicators for as many companies as possible?

In [1]:

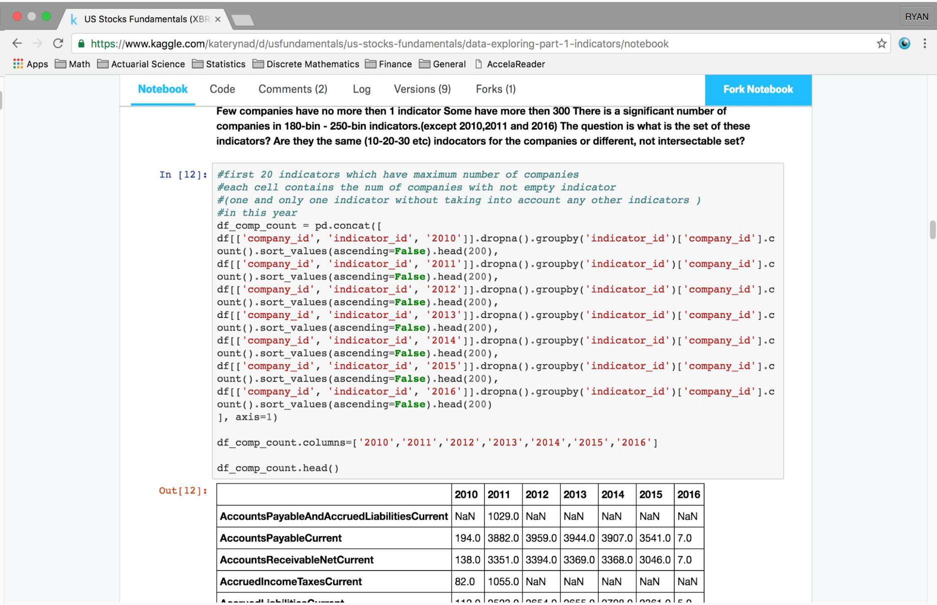
```
import pandas as pd
from pandas import Series,DataFrame
import numpy as np

# For Visualization
import matplotlib.pyplot as plt
import matplotlib
#%matplotlib inline

matplotlib.style.use('ggplot')
df=pd.read_csv('../input/indicators_by_company.csv')
```

In [2]:

```
#number of indicators by company
df_ind_count = pd.concat([ df[['company_id', 'indicator_id', '2010']].dropna().groupby('com
pany_id')['indicator_id'].count()
,df[['company_id', 'indicator_id', '2011']].dropna().groupby('company_id')
['indicator_id'].count()
,df[['company_id', 'indicator_id', '2012']].dropna().groupby('company_id')
['indicator_id'].count()
,df[['company_id', 'indicator_id', '2013']].dropna().groupby('company_id')
['indicator_id'].count()
,df[['company_id', 'indicator_id', '2014']].dropna().groupby('company_id')
['indicator_id'].count()
,df[['company_id', 'indicator_id', '2015']].dropna().groupby('company_id')
['indicator_id'].count()
,df[['company_id', 'indicator_id', '2016']].dropna().groupby('company_id')
['indicator_id'].count()
], axis=1)
df_ind_count.columns=['2010','2011','2012','2013','2014','2015','2016']
df_ind_count.head()
```

Out[2]:

	2010	2011	2012	2013	2014	2015	2016



Few companies have no more then 1 indicator Some have more then 300 There is a significant number of companies in 180-bin - 250-bin indicators.(except 2010,2011 and 2016) The question is what is the set of these indicators? Are they the same (10-20-30 etc) indocators for the companies or different, not intersectable set?

In [12]:

```
#first 20 indicators which have maximum number of companies
#each cell contains the num of companies with not empty indicator
#(one and only one indicator without taking into account any other indicators )
#in this year
df_comp_count = pd.concat([
df[['company_id', 'indicator_id', '2010']].dropna().groupby('indicator_id')['company_id'].c
ount().sort_values(ascending=False).head(200),
df[['company_id', 'indicator_id', '2011']].dropna().groupby('indicator_id')['company_id'].c
ount().sort_values(ascending=False).head(200),
df[['company_id', 'indicator_id', '2012']].dropna().groupby('indicator_id')['company_id'].c
ount().sort_values(ascending=False).head(200),
df[['company_id', 'indicator_id', '2013']].dropna().groupby('indicator_id')['company_id'].c
ount().sort_values(ascending=False).head(200),
df[['company_id', 'indicator_id', '2014']].dropna().groupby('indicator_id')['company_id'].c
ount().sort_values(ascending=False).head(200),
df[['company_id', 'indicator_id', '2015']].dropna().groupby('indicator_id')['company_id'].c
ount().sort_values(ascending=False).head(200),
df[['company_id', 'indicator_id', '2016']].dropna().groupby('indicator_id')['company_id'].c
ount().sort_values(ascending=False).head(200)
], axis=1)

df_comp_count.columns=['2010','2011','2012','2013','2014','2015','2016']

df_comp_count.head()
```

Out[12]:

	2010	2011	2012	2013	2014	2015	2016
AccountsPayableAndAccruedLiabilitiesCurrent	NaN	1029.0	NaN	NaN	NaN	NaN	NaN
AccountsPayableCurrent	194.0	3882.0	3959.0	3944.0	3907.0	3541.0	7.0
AccountsReceivableNetCurrent	138.0	3351.0	3394.0	3369.0	3368.0	3046.0	7.0
AccruedIncomeTaxesCurrent	82.0	1055.0	NaN	NaN	NaN	NaN	NaN
AccruedLiabilitiesCurrent	112.0	2523.0	2654.0	2655.0	2708.0	2361.0	5.0

Got Your Own Project?

Help get in touch with relevant faculty & staff

2 Project-based courses. CS638 and Stat 679, on data science

Working with professors to facilitate independent projects/directed studies. Similar to Directed Reading Program in Mathematics dept

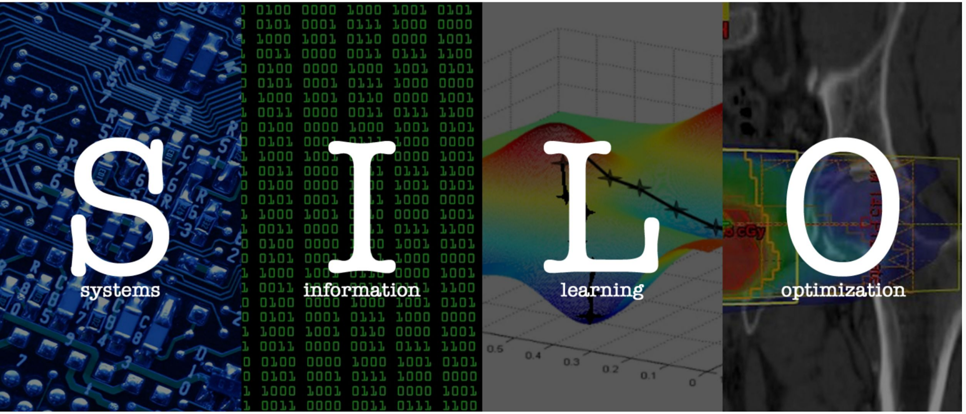
Stay Tuned!

Opportunities on Campus (and Beyond)

Seminars & Talks

CENTER for PRE+HEALTH ADVISING

GREAT LAKES BIOENERGY
RESEARCH CENTER



Department of Statistics
UNIVERSITY OF WISCONSIN–MADISON

Google

CRAY
THE SUPERCOMPUTER COMPANY

BENDYWORKS

State Farm

zendesk

Epic

AMERICAN FAMILY
INSURANCE

GE

Microsoft

singlewire
software

rigbot

OWI
OFFICE OF WATER INFORMATION

Spredfast

HARDIN
Design &
Development

centare

Capital One

REDOX^

SMARTUQ

MICHAEL BEST
& FRIEDRICH LLP

QBE

DoIT
DIVISION OF INFORMATION TECHNOLOGY
UNIVERSITY OF WISCONSIN–MADISON

GRAMMATECH

EATSTREET

Sentry

ESKER

Intact
Solutions,
Inc.

MAVERICK
Software Consulting

5NINES

PerBlue

CUNA
MUTUAL
GROUP

healthgrades

comScore

TREK

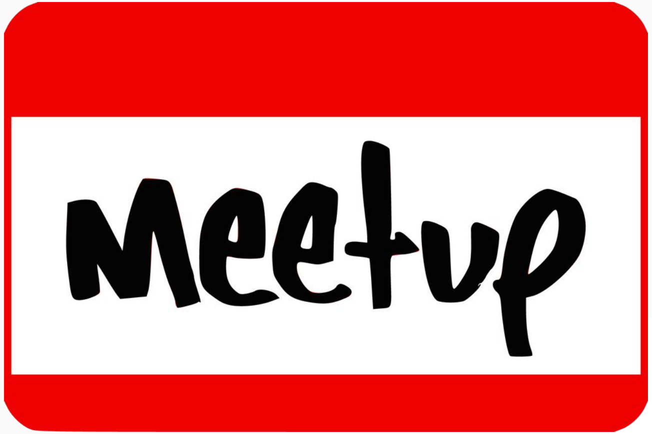
det
DIVISION OF
ENTERPRISE
TECHNOLOGY
Wisconsin Department of Administration

SSEC

X-ES
Extreme Enginee

3M Science.
Applied to Life.™

Local Meetups & Groups



Reminder

Doodle Polls

Join the org on the WIN page:
https://win.wisc.edu/organization/dotDATA

Like our FB page & join our FB group:
www.facebook.com/dotdatascience/
www.dotdatascience.org

A few more minutes ..

- Graduate Students
- Undergrads with interest in helping drive cohesive administration
- Questions? Queries? Conundrums? Ideas? Share them with us!