10 Years Ago

Data Science:
What Actuaries (DON'T) Need to Know

What actuaries **need** to know

Data Science in Insurance

Why it's more important now than ever before

What is Data Science?

What is a Data Scientist?

Where to Find Data Scientists for Your Insurance Company

What actuaries **DON'T need** to know

A pragmatic approach for practitioners

Jeremy Achin
CEO & Co-founder, DataRobot Inc.

Questions?

DataRobot

Additional questions, comments, queries:
jeremy@datarobot.com

Data Science:
What Actuaries (DON'T) Need to Know

What actuaries **need** to know

Data Science in Insurance

Why it's more important now than ever before

What is Data Science?

What is a Data Scientist?

Where to Find Data Scientists for Your Insurance Company

What actuaries **DON'T need** to know

Jeremy Achin
CEO & Co-founder, DataRobot Inc.

10 Years Ago

TOWERS PERRIN
TILLINGHAST

What is Predictive Modeling?

Casualty Actuaries of the Northeast
Spring 2005
Sturbridge, MA
March 23, 2005

Presented by Christopher Monsour, FCAS, MAAA

https://www.casact.org/community/affiliates/cane/0305/monsour.pdf

What actuaries **need** to know

Data Science in Insurance

Why it's more important now than ever before

What is Data Science?

What is a Data Scientist?

Where to Find Data Scientists for Your Insurance Company

Recruit and Hire Existing Data Scientists

Other Data Scientist Candidates

What about a team?

Insurance companies will need Data Scientists to stay competitive
and
Actuaries are their best (maybe only viable) hope

Data Science in Insurance

Why it's more important now than ever before

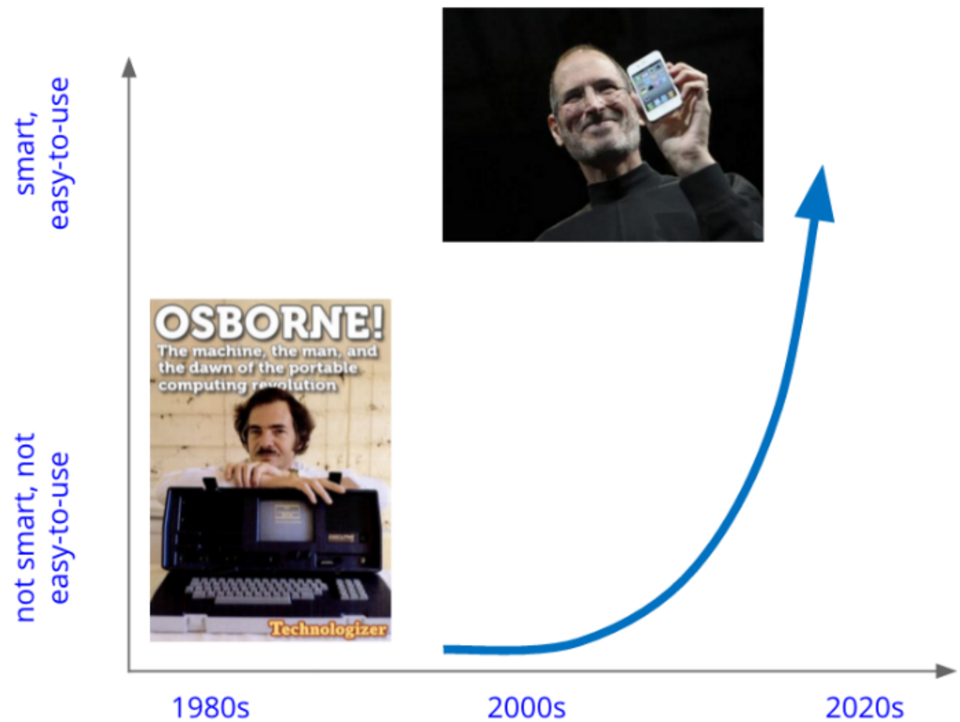
Data is everywhere and Data Science generates value from data

"Data is an emerging asset class" – World Economic Forum

"90% of the data in the world today has been created in the last two years alone"

Absolutely insane amount of computation power

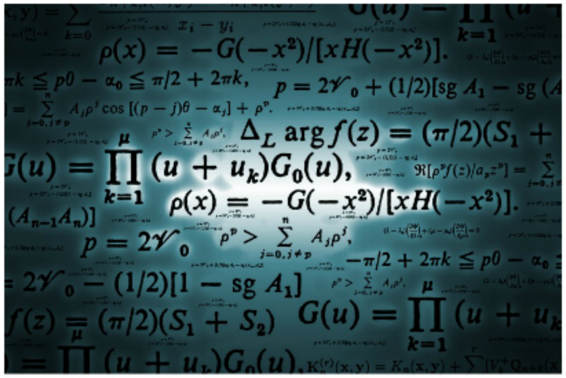
Increasingly inexpensive and smart storage & computational environment



1982 Osborne PC weighs 100 times as much, has 500 times the volume, costs 10 times as much – with 1/100 of processing speed, 1/100000 memory of a typical 2010 smart phone.

Next generation tools, platforms, and approaches to data science

Traditional Approach

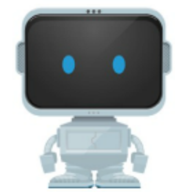


- Ivy league approach - only for the chosen ones
- Focused on activities - detached from outcomes
- Assumption based: model selection is based on modeler's understanding of the world?
- Development is costly and limited
- Heavy dependence on programming

Modern Approach

open source programming

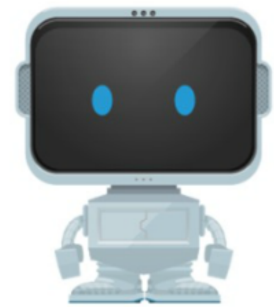
social network of coders



automated solutions

- Common man approach - for everyone
- Focused on business outcome
- Validation based: model selected if it predicts well in real world
- Development is crowd sourced, peer reviewed
- Automated solutions

Modern Approach



open source programming

social network of coders

automated solutions

- Common man approach - for everyone
- Focused on business outcome
- Validation based: model selected if it predicts well in real world
- Development is crowd sourced, peer reviewed
- Automated solutions

Completed • $25,000 • 634 teams

Liberty Mutual Group - Fire Peril Loss Cost

Tue 8 Jul 2014 – Tue 2 Sep 2014 (14 months ago)

Leaderboard

1. DataRobot
2. Ivanhoe
3. barisumog
4. datalab.se
5. paulperry
6. Mark & Dmitriy
7. tryhard
8. Leustagos and Titericz
9. Gauss, Anshul, and

Predict expected fire losses for insurance policies

A Fortune 100 company, Liberty Mutual Insurance has provided a wide range of insurance products and services designed to meet our customers' ever-changing needs for over 100 years.

Within the business insurance industry, fire losses account for a significant portion of total property losses. High severity and low frequency, fire losses are inherently volatile, which makes modeling them difficult. In this challenge, your task is to predict the target, a transformed ratio of loss to total insured value, using the provided information. This will enable more accurate identification of each policyholder's risk exposure and the ability to tailor the insurance coverage for their specific operation.

Because we seek to tap innovation both inside and outside the company, certain eligible Liberty Mutual employees are encouraged to participate in this challenge for development purposes. Refer to the competition rules for the full details.

Completed • $10,000

Allstate Claim Prediction Challenge

Wed 13 Jul 2011 – Wed 12 Oct 2011 (4 years ago)

A key part of insurance is charging each customer the appropriate price for the risk they represent.

Risk varies widely from customer to customer, and a deep understanding of different risk factors helps predict the likelihood and cost of insurance claims. The goal of this competition is to better predict Bodily Injury Liability Insurance claim payments based on the characteristics of the insured customer's vehicle.

Many factors contribute to the frequency and severity of car accidents including how, where and under what conditions people drive, as well as what they are driving.

Bodily Injury Liability Insurance covers other people's bodily injury or death for which the insured is responsible. **The goal of this competition is to predict Bodily Injury Liability Insurance claim payments based on the characteristics of the insured's vehicle.**

Leaderboard

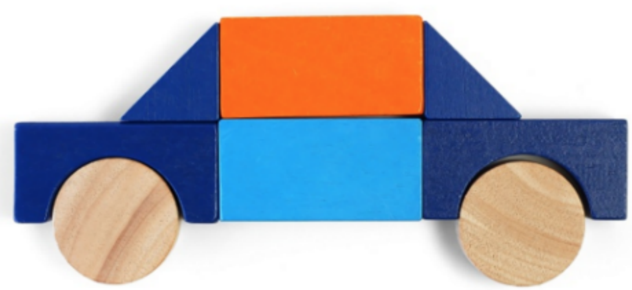
1. Matt C
2. Owen

Completed • $50,000 • 1,568 teams

Allstate Purchase Prediction Challenge

Tue 18 Feb 2014 – Mon 19 May 2014 (18 months ago)

Predict a purchased policy based on transaction history



As a customer shops an insurance policy, he/she will receive a number of quotes with different coverage options before purchasing a plan. This is represented in this challenge as a series of rows that include a customer ID, information about the customer, information about the quoted policy, and the cost. Your task is to predict the purchased coverage options using a limited subset of the total interaction history. If the eventual purchase can be predicted sooner in the shopping window, the quoting process is shortened and the issuer is less likely to lose the customer's business.

Using a customer's shopping history, can you predict what policy they will end up choosing?

Homesite
HOME INSURANCE

$20,000 • 332 teams

Homesite Quote Conversion

Merger and 1st Submission Deadline

Mon 9 Nov 2015 Mon 8 Feb 2016 (2 months to go)

Which customers will purchase a quoted insurance plan?

Before asking someone on a date or skydiving, it's important to know your likelihood of success. The same goes for quoting home insurance prices to a potential customer. Homesite, a leading provider of homeowners insurance, does not currently have a dynamic conversion rate model that can give them confidence a quoted price will lead to a purchase.

Using an anonymized database of information on customer and sales activity, including property and coverage information, Homesite is challenging you to predict which customers will purchase a given quote. Accurately predicting conversion would help Homesite better understand the impact of proposed pricing changes and maintain an ideal portfolio of customer segments.

Started: 7:29 pm, Monday 9 November 2015 UTC
Ends: 11:59 pm, Monday 8 February 2016 UTC (91 total days)
Points: this competition awards standard ranking points
Tiers: this competition counts towards tiers

Deloitte.

Completed • $70,000 • 37 teams

As the World Churns

Tue 22 Oct 2013 – Sat 21 Dec 2013 (23 months ago)

Leaderboard

1. Leustagos & Gxav
2. Dmitry Efimov
3. ivo and BreakfastPirate
4. Michael Jahrer & Jeong-Yoon Lee
5. J. A. Guerrero
6. FAndy & Sen
7. agdavis
8. An apple a day

This competition is private-entry. You can view but not participate.

Predict which customers will leave an insurance company in the next 12 months.

Understanding customer loyalty is an important part of any business. The ability to predict ahead of time when a customer is likely to churn can enable early intervention processes to be put in place, and ultimately a reduction in customer churn. This competition seeks a solution for predicting which current customers of an insurance company will leave in 12 months time, and when.

This competition is now closed to new entrants.

Started: 7:59 pm, Tuesday 22 October 2013 UTC
Ended: 11:59 pm, Saturday 21 December 2013 UTC (60 total days)
Points: this competition awarded standard ranking points
Tiers: this competition counted towards tiers

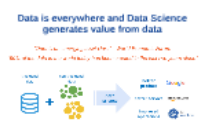
- More data
- More computation power
- Better and more accesible tools

Your competitors are doing it!
Your customers expect it!

need to know

Data Science in Insurance

Why it's more important now than ever before



- More data
- More computation power
- Better and more accesible tools

Your competitors are doing it!
Your customers expect it!

"Statistics on a Mac"

"The generalized extraction of knowledge from data"

Data → *Predictions & Insights*

Been doing this for a while now

What's different?

What is Data Science?

What is Data Science?

"Statistics on a Mac"

"The generalized extraction of knowledge from data"

Data → *Predictions & Insights*

Been doing this for a while now

What's different?

"Statistics on a Mac"

"The generalized extraction of knowledge from data"

Data → Predictions & Insights

Been doing this for a while now

What's different?

Traditional Analytics	Data Science
Structured Data	Structured & Unstructured Data
Use what data is readily available (or what IT is willing to give you)	Go get any data that may be of value
Small/medium size data	Size of data not an obstacle
Shallow understanding of the data	Deep understanding of the data
IT is heavily involved	IT as little as possible
Projects take a long time/ Limited bandwith	Faster turnaround through better tools and more automation → Higher bandwith
Limited tool set (Regression, GLM)	More/Better tools (Modern statistical approaches, machine learning)

Traditional Analytics	***Data Science***
Structured Data	*Structured & Unstructured Data*
Use what data is readily available (or what IT is willing to give you)	*Go get any data that may be of value*
Small/medium size data	*Size of data not an obstacle*
Shallow understanding of the data	*Deep understanding of the data*

IT is heavily involved

Projects take a long time/ Limited bandwith

IT as little as possible

Faster turnaround through better tools and more automation

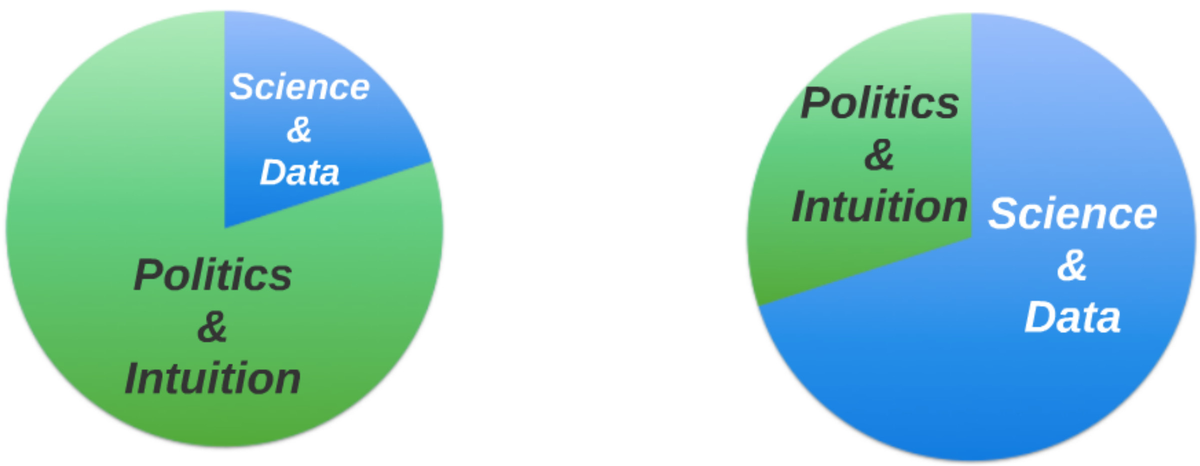
Higher bandwith

Limited tool set (Regression, GLM)

More/Better tools (Modern statistical approaches, machine learning)

End Result: Dramatic increase in quality and quantity of actionable predictions & insights

How decisions are made



Without Data Science

With Data Science

now than ever before

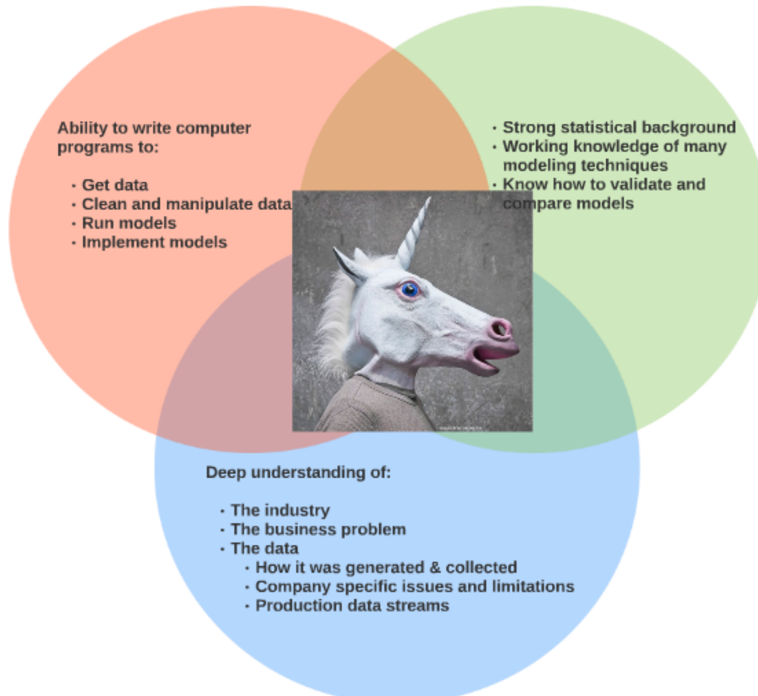
What is Data Science?

What is a Data Scientist?

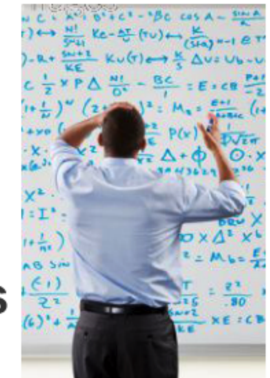
"The Sexiest Job of the 21st Century"

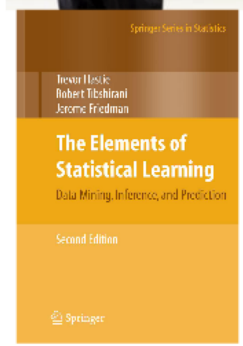
"Statistician from San Francisco"

Hacking Skills



Math & Stats





Domain Expertise

Hacking Skills

Ability to write computer programs to:

- Get data
- Clean and manipulate data
- Run models
- Implement models

Deep understanding of:

Math & Stats

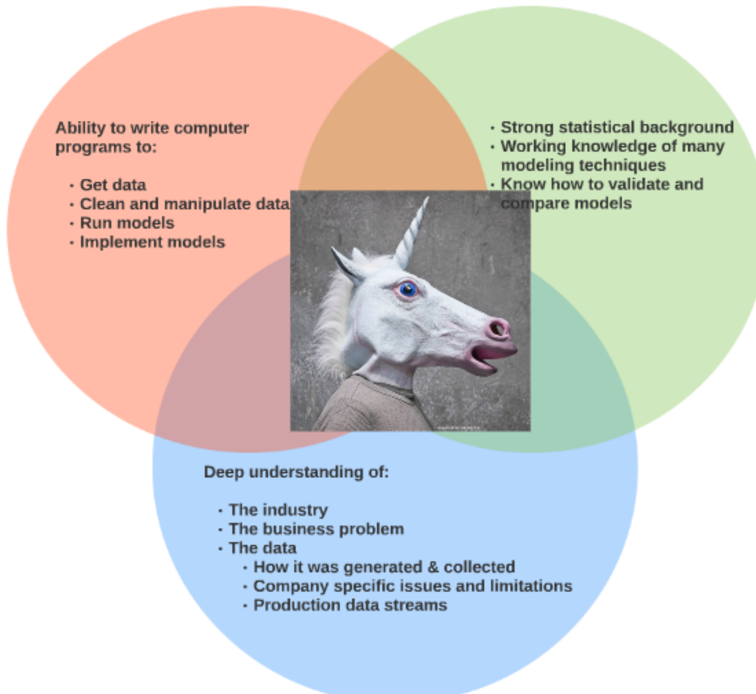
- **Strong statistical background**
- **Working knowledge of many modeling techniques**
- **Know how to validate and compare models**

Deep understanding of:

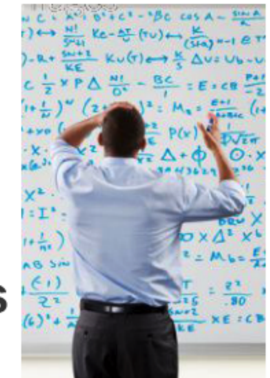
- **The industry**
- **The business problem**
- **The data**
    - **How it was generated & collected**
    - **Company specific issues and limitations**
    - **Production data streams**

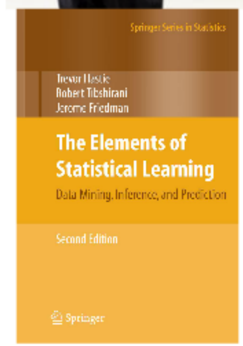
Domain Expertise

Hacking Skills



Math & Stats





Domain Expertise

“Unicorns are Lame” *-quote by: nobody ever.*

Hiring Real Unicorns is Expensive! $$$$$$$$$$$$

Also, Many People Pretending to be Unicorns.

End Result: Dramatic increase in quality and quantity of actionable predictions & insights

What is a Data Scientist?

"The Sexiest Job of the 21st Century"

"Statistician from San Francisco"

Recruit and Hire Existing Data Scientists

Challenges

- *Many imposters posing as Data Scientists*
- *Takes time to learn industry & business knowledge*
- *Takes even more time to aquire company data knowledge*
- *Data Scientists are prohibitively expensive ($ > hiring manages salary)*

to Find Data

Can you think o
insurance c
employee th
make a good
for a Data S

Where to Find Data Scientists for Your Insurance Company

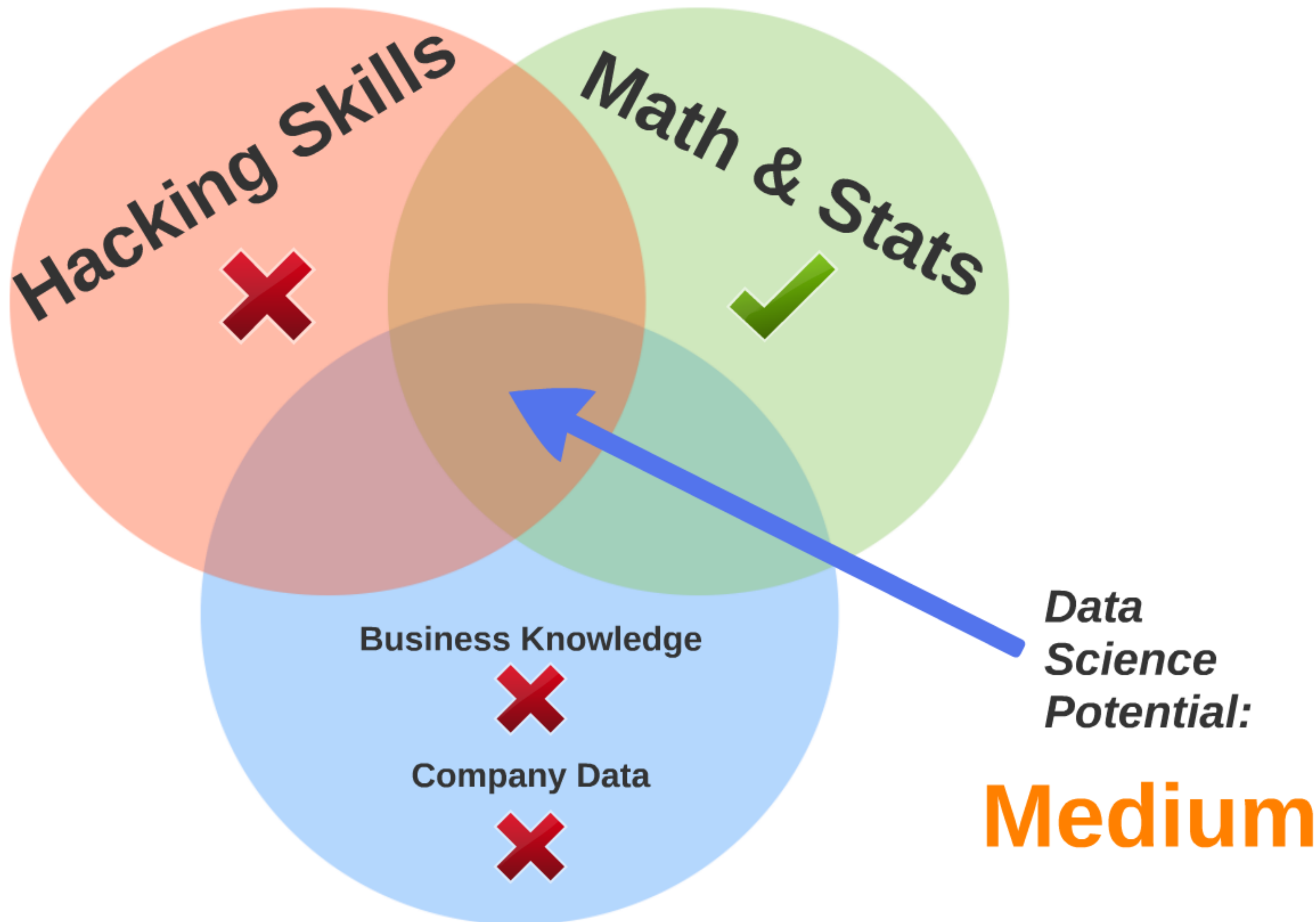
Recruit and Hire Existing Data Scientists

Challenges

- *Many imposters posing as Data Scientists*
- *Takes time to learn industry & business knowledge*
- *Takes even more time to aquire company data knowledge*
- *Data Scientists are prohibitively expensive ($ > hiring manages salary)*

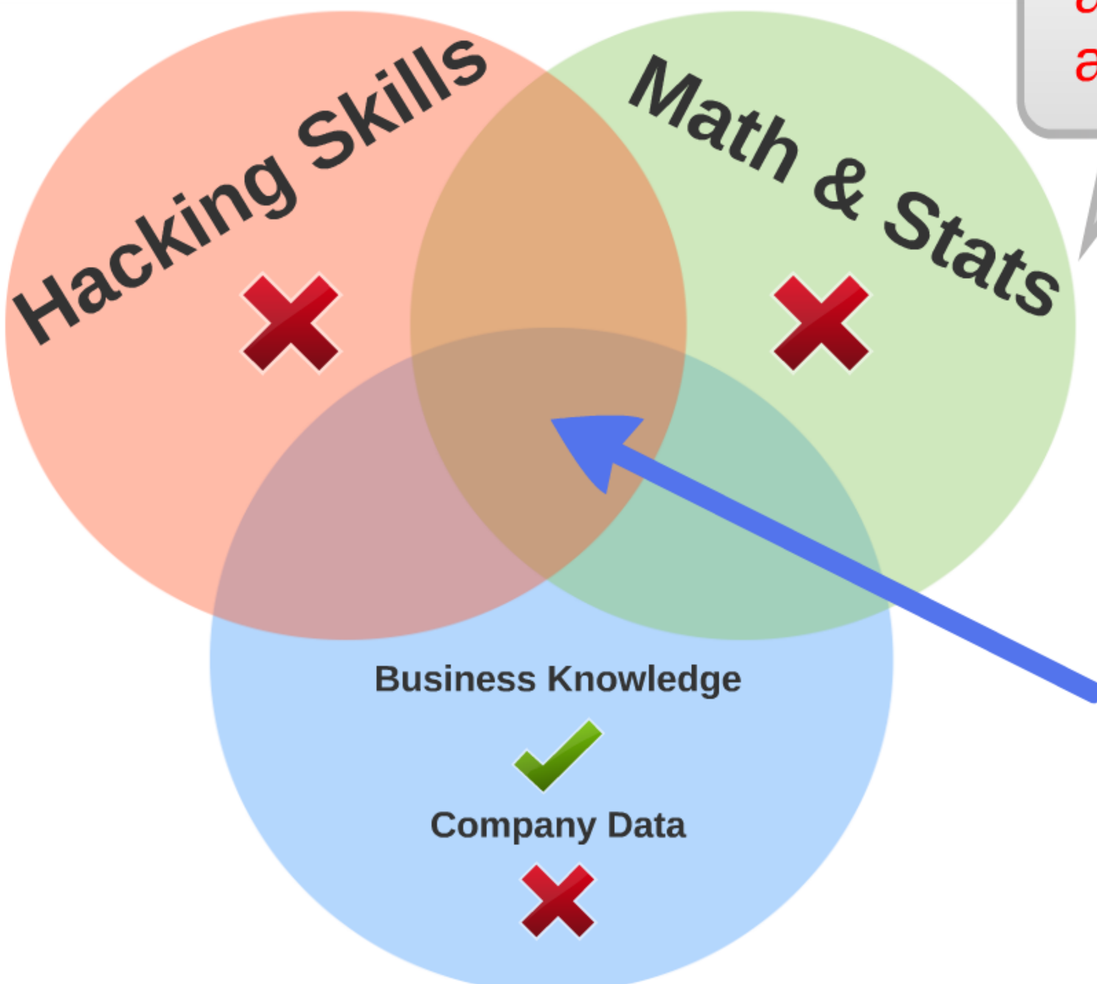
Other Data Scientist Candidates

Candidate #1: Statistician



Candidate #2: MBA

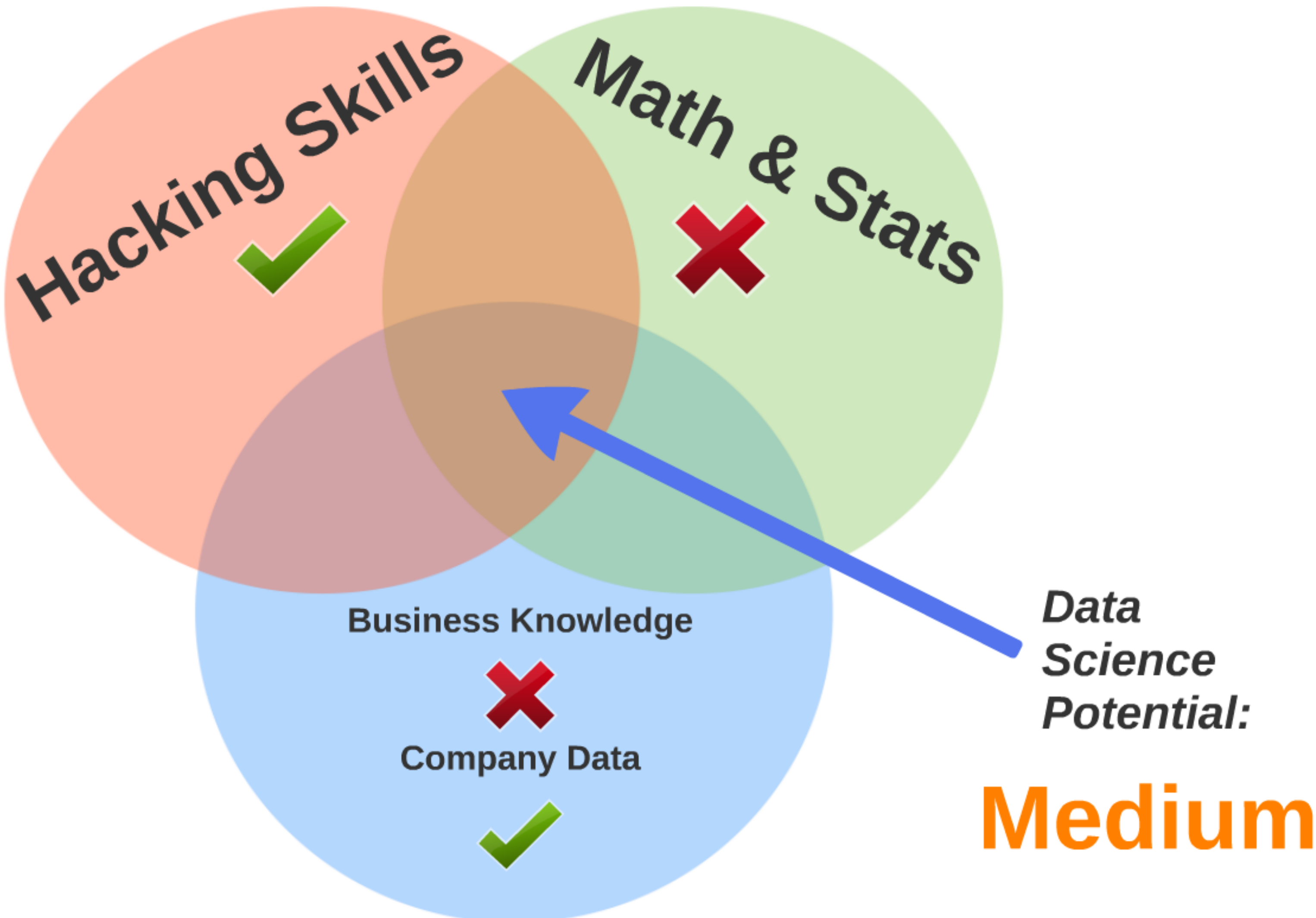
90% of MBA's think their math skills are above average for an MBA.



Data Science Potential:

Medium

Candidate #3: The IT Data Specialist

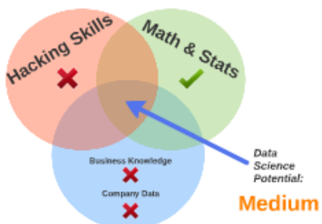


- *Takes time to learn industry & business knowledge*
- *Takes even more time to aquire company data knowledge*
- *Data Scientists are prohibitively expensive ($ > hiring manages salary)*

What about a team?

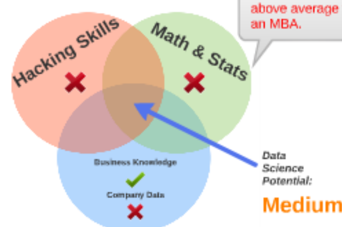
Candidate #4: Statistician, MBA, IT

Candidate #1: Statistician

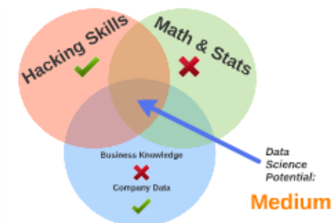


Candidate #2: MBA

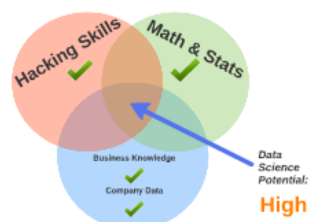
90% of MBA's think their math skills are above average for an MBA.



Candidate #3: The IT Data Specialist



Candidate #4: Team

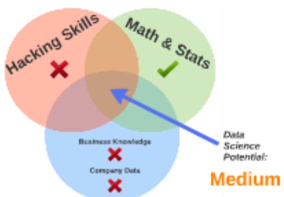


- *Takes time to learn industry & business knowledge*
- *Takes even more time to aquire company data knowledge*
- *Data Scientists are prohibitively expensive ($ > hiring manages salary)*

What about a team?

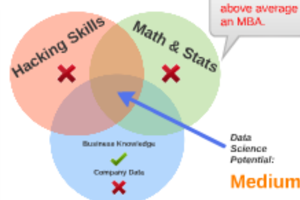
Candidate #4: Statistician, MBA, IT

Candidate #1: Statistician

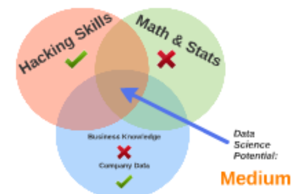


Candidate #2: MBA

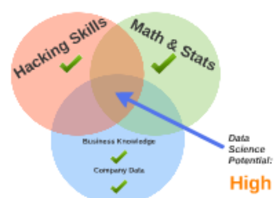
90% of MBA's think their math skills are above average for an MBA.



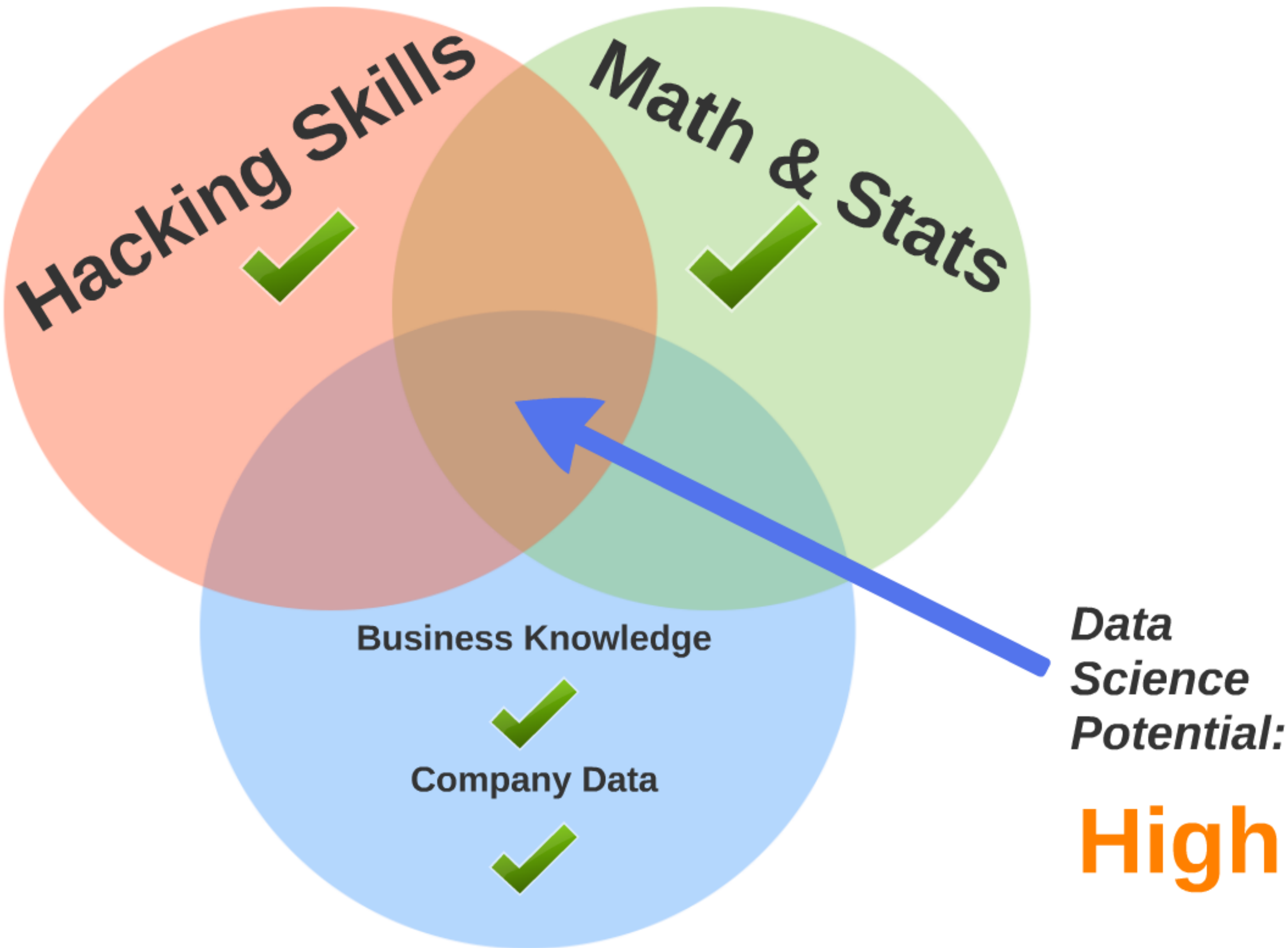
Candidate #3: The IT Data Specialist



Candidate #4: Team



Candidate #4: Team

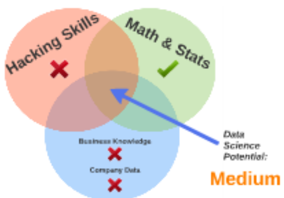


- *Takes time to learn industry & business knowledge*
- *Takes even more time to aquire company data knowledge*
- *Data Scientists are prohibitively expensive ($ > hiring manages salary)*

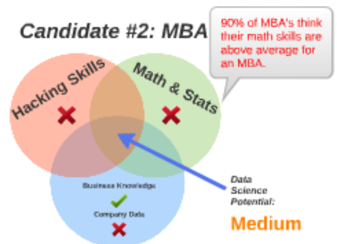
What about a team?

Candidate #4: Statistician, MBA, IT

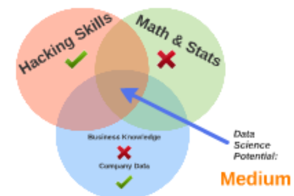
Candidate #1: Statistician



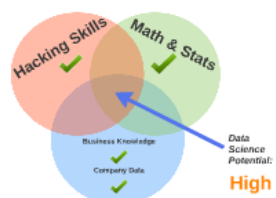
Candidate #2: MBA



Candidate #3: The IT Data Specialist

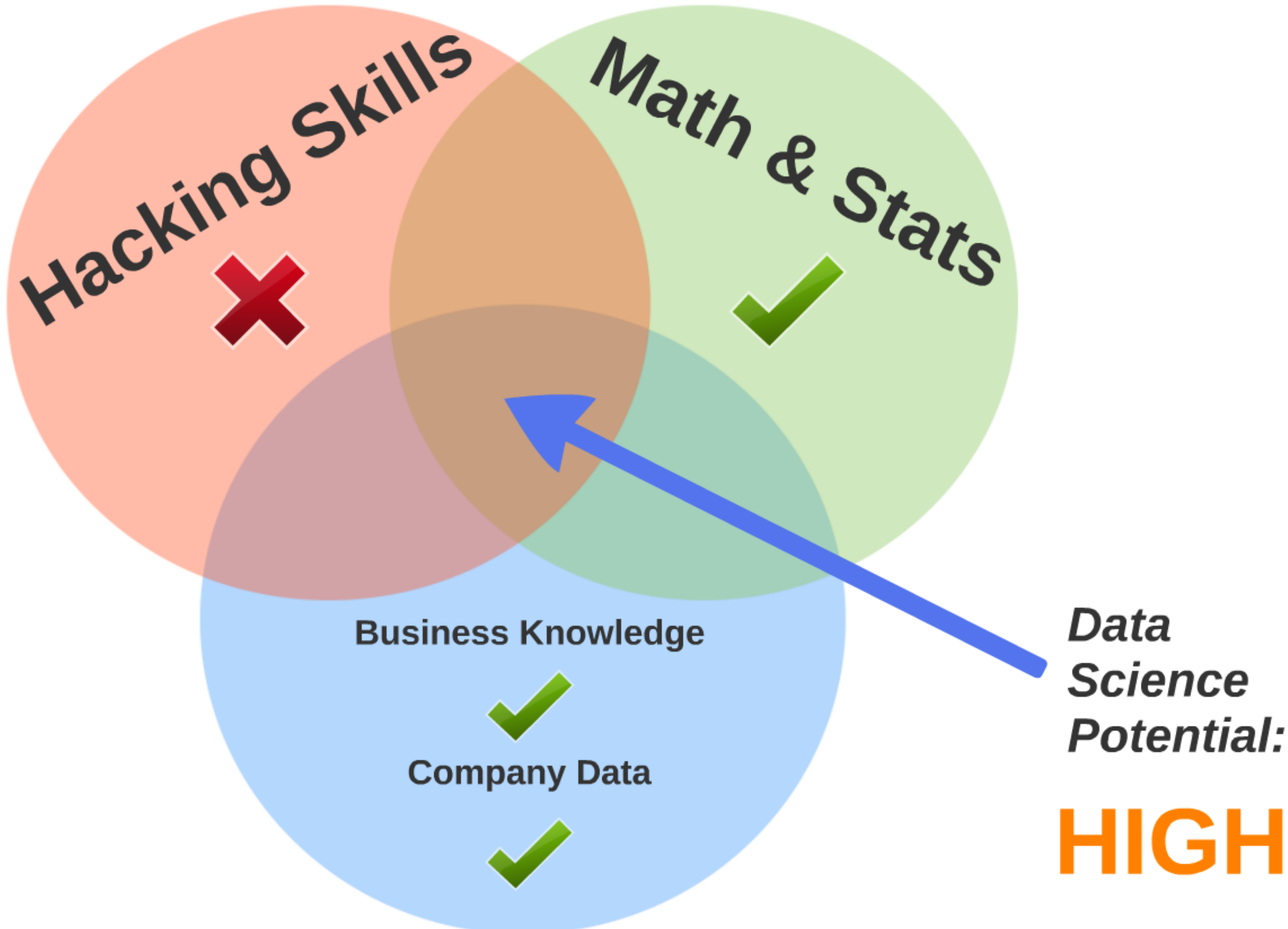


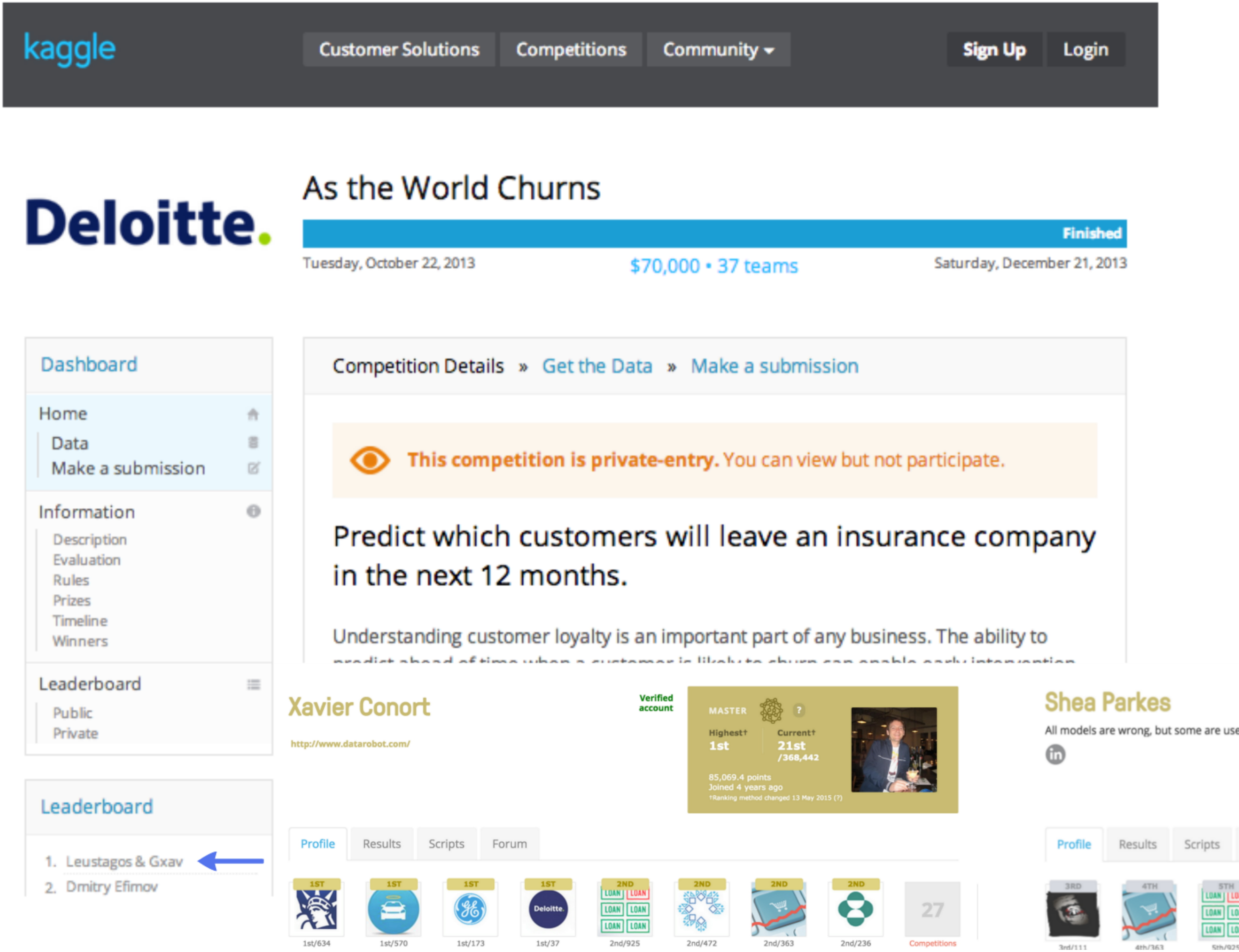
Candidate #4: Team



Can you think of any other insurance company employee that would make a good candidate for a Data Scientist?

Candidate #5: Actuary




kaggle
Customer Solutions
Competitions
Community
Sign Up
Login
Deloitte.
As the World Churns
Finished
Tuesday, October 22, 2013
$70,000 • 37 teams
Saturday, December 21, 2013
Dashboard
Home
Data
Make a submission
Information
Description
Evaluation
Rules
Prizes
Timeline
Winners
Leaderboard
Public
Private
Competition Details » Get the Data » Make a submission
This competition is private-entry. You can view but not participate.
Predict which customers will leave an insurance company in the next 12 months.
Understanding customer loyalty is an important part of any business. The ability to
Leaderboard
1. Leustagos & Gxav
2. Dmitry Efimov
Xavier Conort
Verified account
http://www.datarobot.com/
MASTER
Highest† 1st
Current† 21st /368,442
85,069.4 points
Joined 4 years ago
Profile
Results
Scripts
Forum
1st/634
1st/570
1st/173
1st/37
2nd/925
2nd/472
2nd/363
2nd/236
27
Competitions
Shea Parkes
All models are wrong, but some are usefu
Profile
Results
Scripts

Make a submission

Information

- Description
- Evaluation
- Rules
- Prizes
- Timeline
- Winners

Leaderboard

- Public
- Private

This competition is private-entry. You can view but not particip

Predict which customers will leave an insurance in the next 12 months.

Understanding customer loyalty is an important part of any business. The

Leaderboard

1. Leustagos & Gxav
2. Dmitry Efimov

Xavier Conort

Verified account

http://www.datarobot.com/

MASTER ?

Highest† 1st

Current† 21st /368,442

85,069.4 points
Joined 4 years ago
†Ranking method changed 13 May 2015 (?)

Profile | Results | Scripts | Forum

1ST	1ST	1ST	1ST	2ND	2ND	2ND	2ND	27
1st/634	1st/570	1st/173	1st/37	2nd/925	2nd/472	2nd/363	2nd/236	Competitions

Owen

Verified account

Working for DataRobot but I am not a robot.

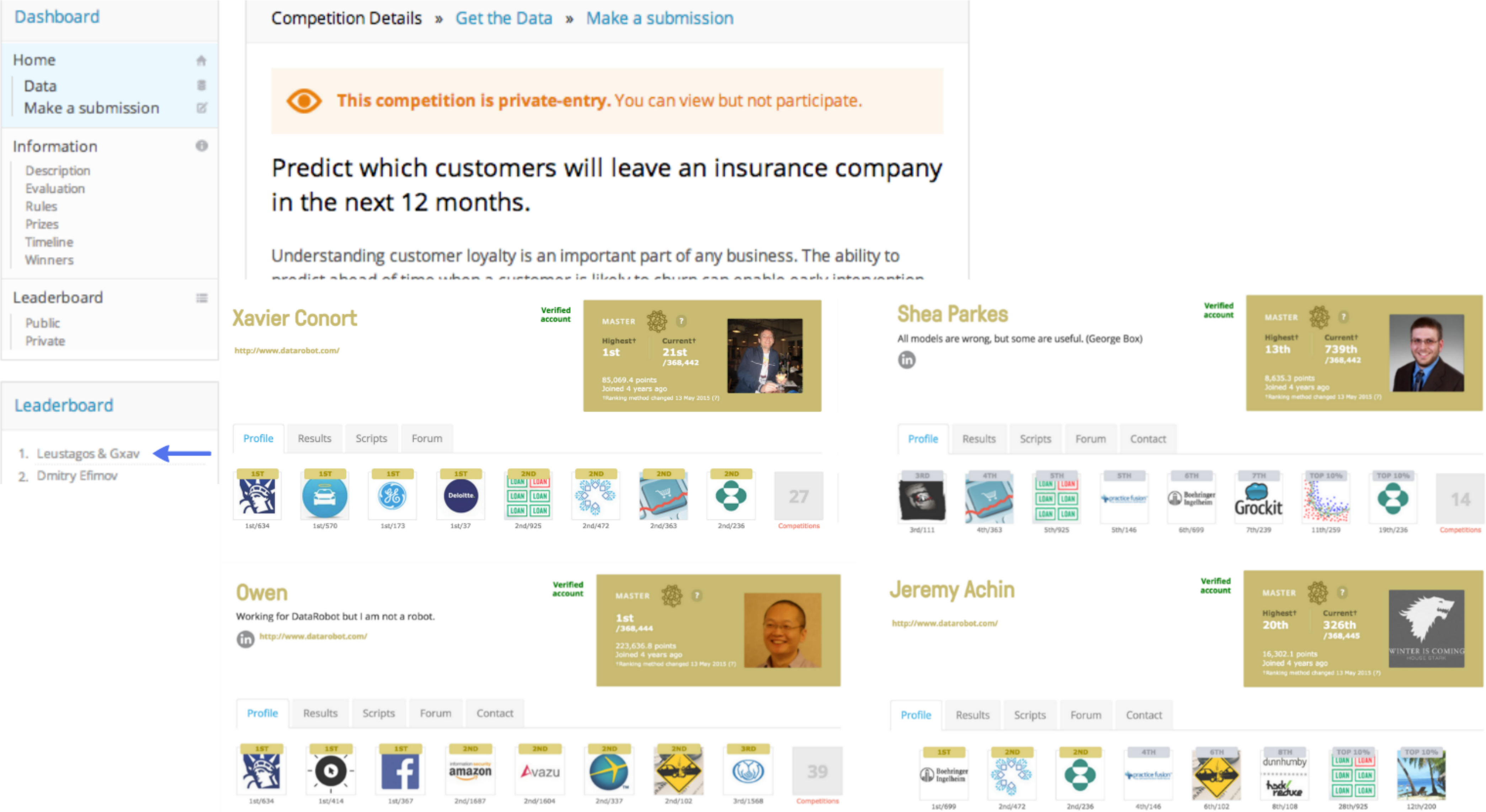
http://www.datarobot.com/

MASTER ?

1st /368,444

223,636.8 points
Joined 4 years ago
†Ranking method changed 13 May 2015 (?)

Profile | Results | Scripts | Forum | Contact



Dashboard
Home
Data
Make a submission
Information
Description
Evaluation
Rules
Prizes
Timeline
Winners
Leaderboard
Public
Private
Competition Details » Get the Data » Make a submission
This competition is private-entry. You can view but not participate.
Predict which customers will leave an insurance company in the next 12 months.
Understanding customer loyalty is an important part of any business. The ability to
Leaderboard
1. Leustagos & Gxav
2. Dmitry Efimov
Xavier Conort
http://www.datarobot.com/
Verified account
MASTER
Highest† 1st
Current† 21st /368,442
85,069.4 points
Joined 4 years ago
Profile
Results
Scripts
Forum
1st/634
1st/570
1st/173
1st/37
2nd/925
2nd/472
2nd/363
2nd/236
27
Competitions
Shea Parkes
All models are wrong, but some are useful. (George Box)
Verified account
MASTER
Highest† 13th
Current† 739th /368,442
8,635.3 points
Joined 4 years ago
Profile
Results
Scripts
Forum
Contact
3rd/111
4th/363
5th/925
5th/146
6th/699
7th/239
11th/259
19th/236
14
Competitions
Owen
Working for DataRobot but I am not a robot.
http://www.datarobot.com/
Verified account
MASTER
1st /368,444
223,636.8 points
Joined 4 years ago
Profile
Results
Scripts
Forum
Contact
1st/634
1st/414
1st/367
2nd/1687
2nd/1604
2nd/337
2nd/102
3rd/1568
39
Competitions
Jeremy Achin
http://www.datarobot.com/
Verified account
MASTER
Highest† 20th
Current† 326th /368,445
16,302.1 points
Joined 4 years ago
Profile
Results
Scripts
Forum
Contact
1st/699
2nd/472
2nd/236
4th/146
6th/102
8th/108
28th/925
12th/200

Summary

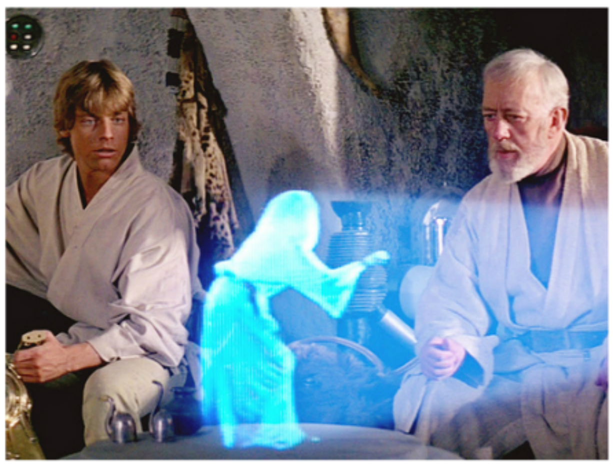
Candidate	Data Science Potential
Statistician	Medium
MBA	Low
IT	Medium
Team	  
Actuary	HIGH

Team

Actuary

HI

Insurance companies will need Data Scientists to stay competitive and Actuaries are their best (maybe only viable) hope



Where to Find Data Scientists for Your Insurance Company

Recruit and Hire Existing Data Scientists

Challenges

- *Many imposters posing as Data Scientists*
- *Takes time to learn industry & business knowledge*
- *Takes even more time to aquire company data knowledge*
- *Data Scientists are prohibitively expensive ($ > hiring managers salary)*

Other Data Scientist Candidates

What about a team?

Candidate #4: Statistician, MBA, IT

Candidate #1: Statistician

+

Candidate #2: MBA

+

Candidate #3: The IT Data Specialist

Candidate #4: Team

Can you think of any other insurance company employee that would make a good candidate for a Data Scientist?

Candidate #5: Actuary

Summary

Insurance companies will need Data Scientists to stay competitive
and
Actuaries are their best (maybe only viable) hope

What actuaries **need** to know

Data Science in Insurance

Why it's more important now than ever before

What is Data Science?

What is a Data Scientist?

Where to Find Data Scientists for Your Insurance Company

Recruit and Hire Existing Data Scientists

What about a team?

Other Data Scientist Candidates

Insurance companies will need Data Scientists to stay competitive
and
Actuaries *are* their best (maybe only viable) hope

What actuaries **DON'T need** to know

Mechanics underlying complex tools, platforms and algorithms

A pragmatic approach for practitioners

Learning Data Science with **automated platform** and with **business outcome in focus**

Once you are familiar with real world experience & practical knowledge...

Take advantage of free educational resources out there

Call to action

Take the unicorn by the horn

Mechanics underlying complex tools, platforms and algorithms

We use complex, automated tools successfully everyday to make our life simple and productive
... without knowing the mechanics.



Complex mechanics of Data Science ...

No need to learn programming to get started

Problem solving with automated platform

Quick learning & smart application of advanced algorithms

Activity	Tool/ Platform/ Source	Learning Focus
Data manipulation & General programming	R, python, tamr, Paxata.	- data manipulation - key statistical packages - key visualization packages
Visualizations	Qlik, tableau	- visualize data by drag & drop - productize your solution
Automated Modeling, Machine Learning	rapidminer, sas, DataRobot	- defining the right question - interpreting results - running experiments using the automated platform

We use complex, automated tools successfully everyday to make our life simple and productive

... without knowing the mechanics.

Quiz # 1

An increasing number of people are using this device for a vast number of activities e.g. education, medical practice, business communication, flight tracking, special assistance communication and even for actuarial work.

Quiz # 2

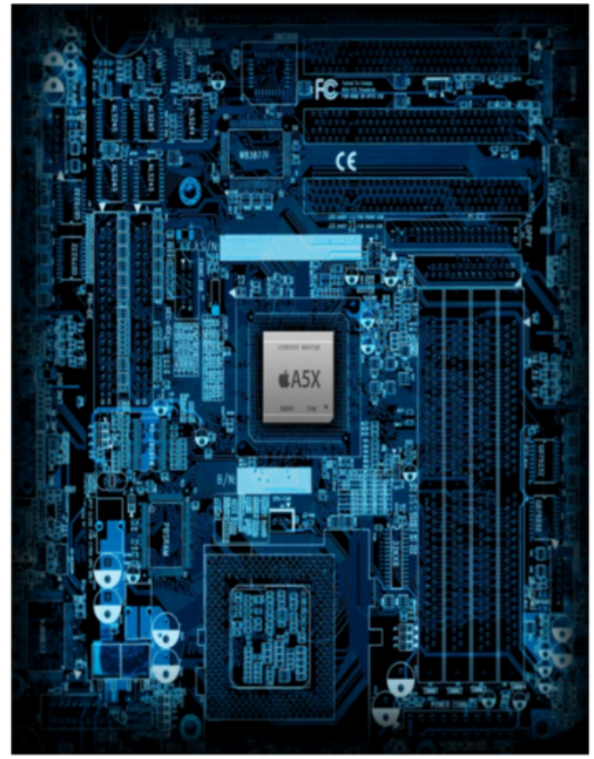
Hundreds of developers wrote millions of lines of C++ codes to build this platform that actuaries use extensively.

e:

Quiz # 3

Quiz # 1

An increasing number of people are using this device for a vast number of activities e.g. education, medical practice, business communication, flight tracking, special assistance communication and even for actuarial work.




27

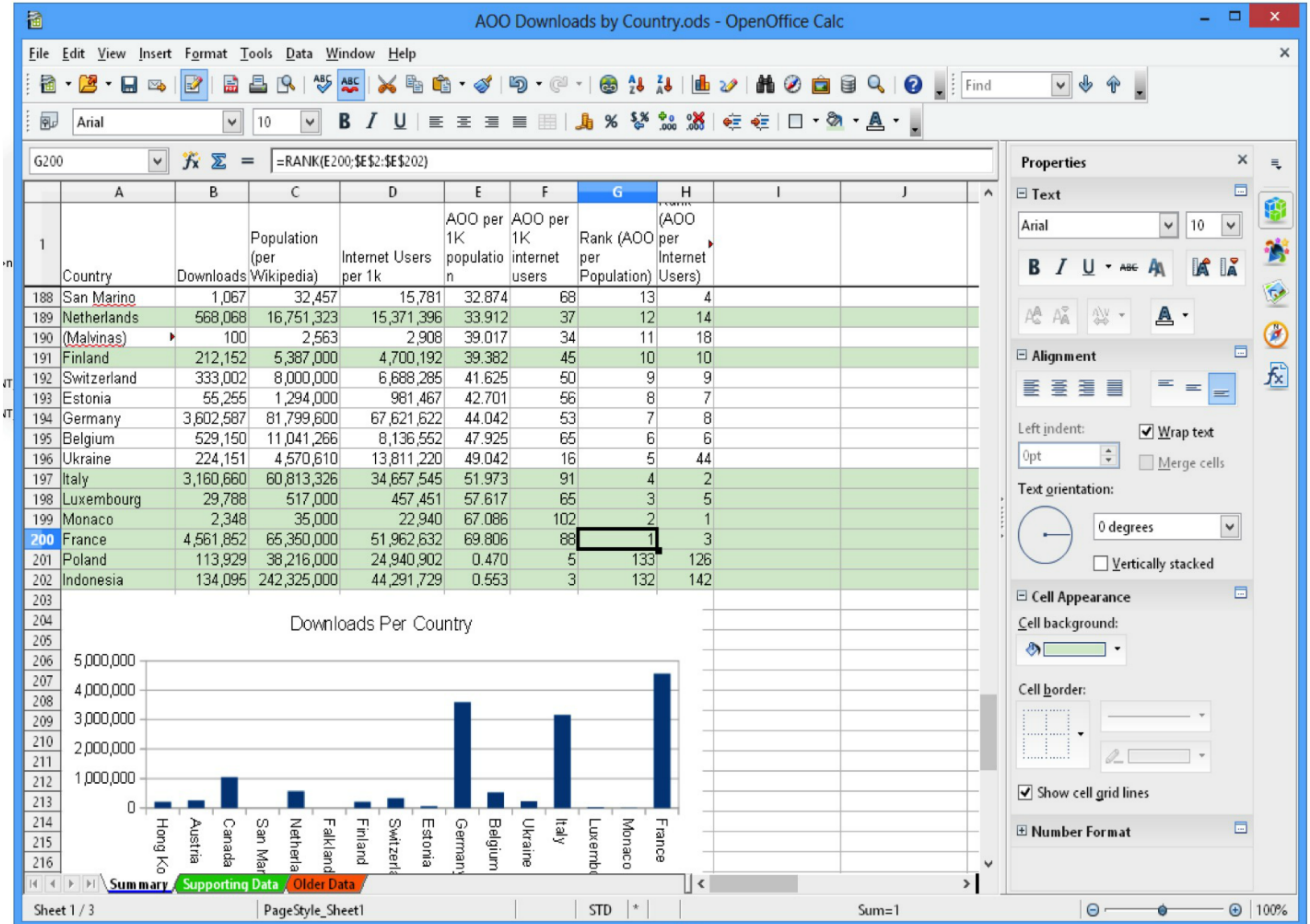
Quiz # 2

```
sal_Bool TTBasic::Compile( SbModule* p )
{               p->SetComment( ((TestToolObj*)pTestObject)->GetRevision(p->GetSource()) );
                SbModule* pOldModule = GetCompileModule();
                SetCompileModule( p );
                p->SetSource( ((TestToolObj*)pTestObject)->PreCompile(p->GetSource()) );
                SetCompileModule( pOldModule );
                if ( ((TestToolObj*)pTestObject)->WasPrecompilerError() )
                                return sal_False;
                return MyBasic::Compile( p );
}
const String TTBasic::GetSpechialErrorText()
{               String nErrorText;
                if ( pTestObject && IS_ERROR() && GetErrorCode() == GET_ERROR()->nError )
                                nErrorText = GenRealString( GET_ERROR()->aText );
                                nErrorText.AppendAscii( ": " );
                                nErrorText += String::CreateFromInt64( GET_ERROR()->nError );
                }
                else { nErrorText = GetErrorText();}
                return nErrorText;
}
void TTBasic::ReportRuntimeError( AppBasEd *pEditWin )
{               SbxVariableRef aDummy = new SbxVariable;
                aDummy->SetUserData( 24 );  // ID_MaybeAddErr
                ((TestToolObj*)pTestObject)->Notify( pTestObject->GetBroadcaster(), SbxHint( SBX_HINT_DATAWANTED, aDummy ) );
                aDummy->SetUserData( 18 );  // ID_ExceptLog
                ((TestToolObj*)pTestObject)->Notify( pTestObject->GetBroadcaster(), SbxHint( SBX_HINT_DATAWANTED, aDummy ) );
                MyBasic::ReportRuntimeError( pEditWin );
}
void TTBasic::DebugFindNoErrors( sal_Bool bDebugFindNoErrors )
{               ((TestToolObj*)pTestObject)->DebugFindNoErrors( bDebugFindNoErrors );
}
```

Hundreds of developers wrote millions of lines of C++ codes to build this platform that actuaries use extensively.

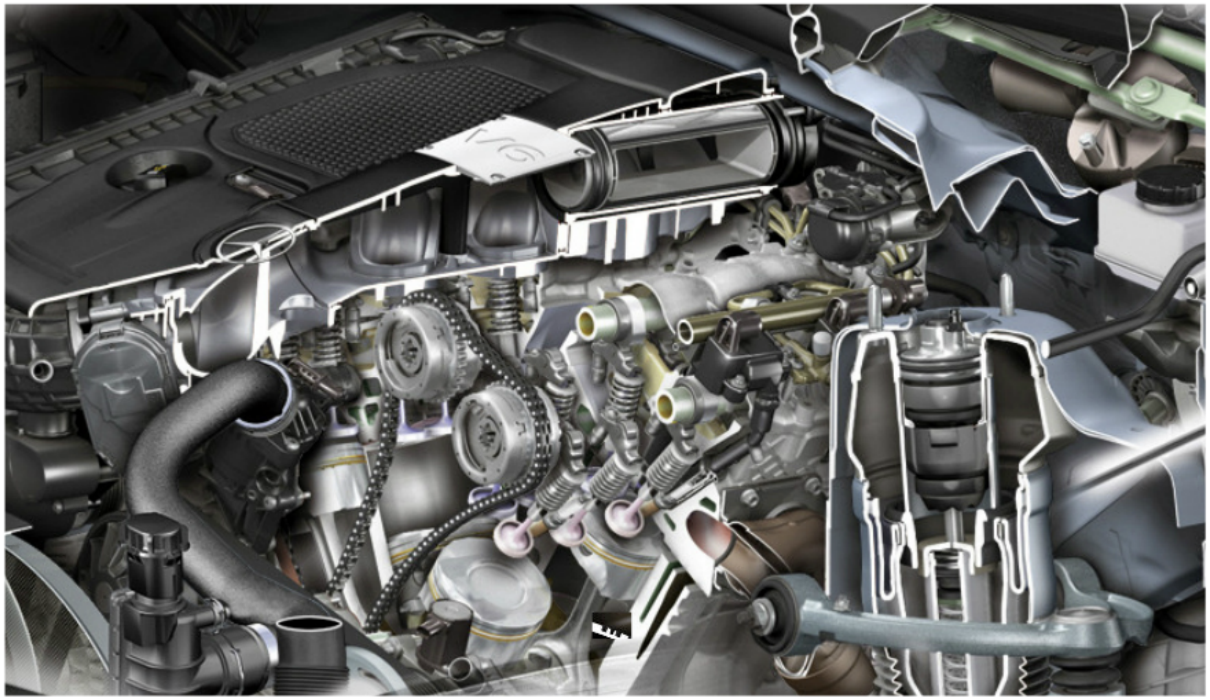
```
namespace formula
{
                //==================================================================
                //= ResourceManager
                //==================================================================
  ::osl::Mutex  ResourceManager::s_aMutex;
  sal_Int32                     ResourceManager::s_nClients = 0;
                ResMgr*     ResourceManager::m_pImpl = NULL;

                //------------------------------------------------------------------
                void ResourceManager::ensureImplExists()
                {
                                if (m_pImpl)
                                                return;

                                ::com::sun::star::lang::Locale aLocale = Application::GetSettings().GetUILocale();

                                ByteString sFileName("for");

                                m_pImpl = ResMgr::CreateResMgr(sFileName.GetBuffer(), aLocale);
                }

                //------------------------------------------------------------------
                ::rtl::OUString ResourceManager::loadString(sal_uInt16 _nResId)
                {
                                ::rtl::OUString sReturn;

                                ensureImplExists();
                                if (m_pImpl)
                                                sReturn = String(ResId(_nResId,*m_pImpl));

                                return sReturn;
                }
                //------------------------------------------------------------------
  ::rtl::OUString ResourceManager::loadString( sal_uInt16 _nResId, const sal_Char* _pPlaceholderAscii, const ::rtl::OUString& _rReplace )
  {
    String sString( loadString( _nResId ) );
    sString.SearchAndReplaceAscii( _pPlaceholderAscii, _rReplace );
    return sString;
  }
  //------------------------------------------------------------------
  void ResourceManager::registerClient()
  {
                ::osl::MutexGuard aGuard(s_aMutex);
                ++s_nClients;
  }

  //------------------------------------------------------------------
  void ResourceManager::revokeClient()
  {
                ::osl::MutexGuard aGuard(s_aMutex);
                if (!--s_nClients && m_pImpl)
                {
                                delete m_pImpl;
                                m_pImpl = NULL;
                }
  }
  ResMgr*           ResourceManager::getResManager()
  {
    ensureImplExists();
                                return m_pImpl;
  }

//..................................................................
} // formula
//..................................................................
```

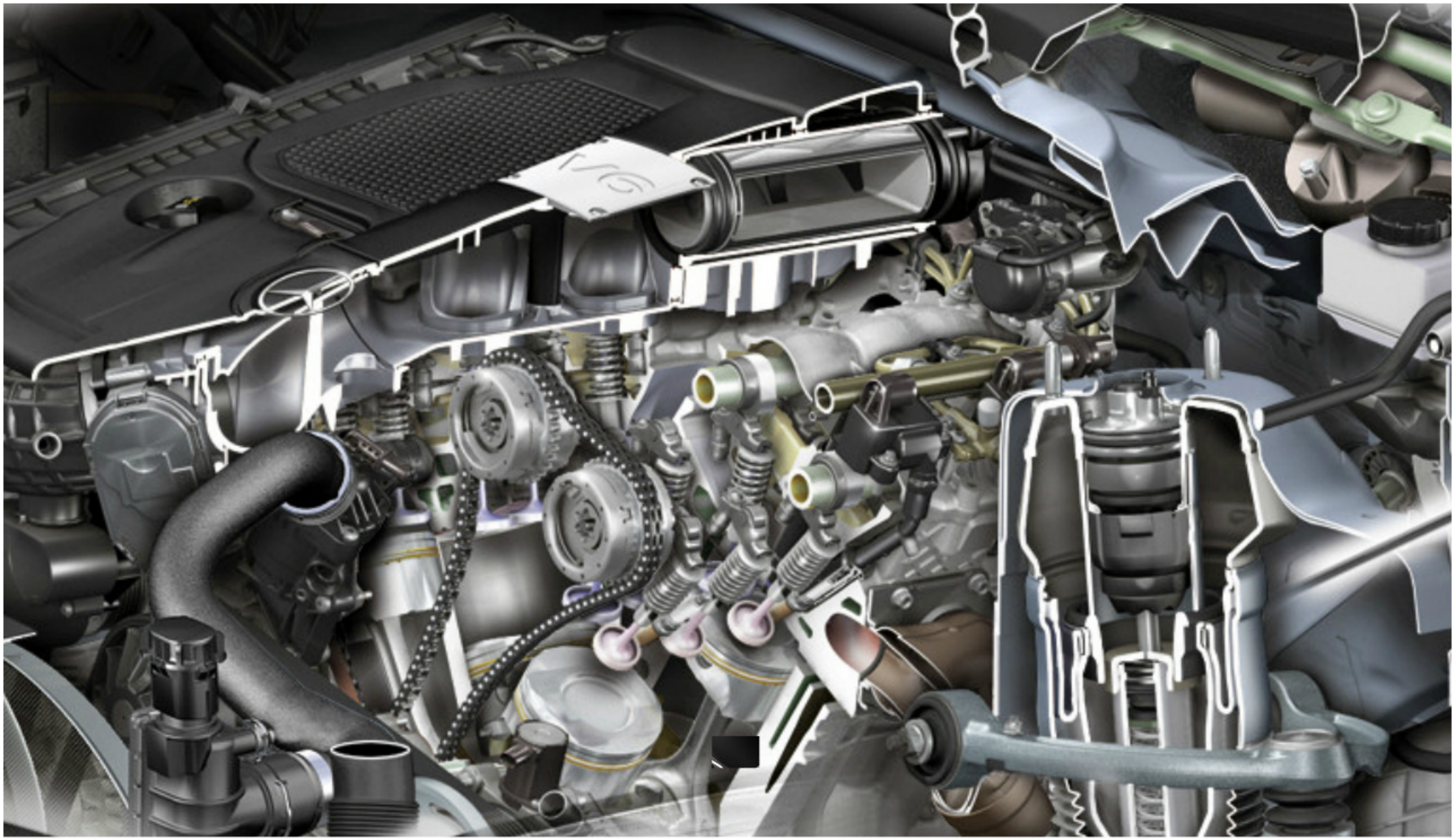
NoEr



	Country	Downloads	Population (per Wikipedia)	Internet Users per 1k	AOO per 1K population	AOO per 1K internet users	Rank (AOO per Population)	Rank (AOO per Internet Users)
188	San Marino	1,067	32,457	15,781	32.874	68	13	4
189	Netherlands	568,068	16,751,323	15,371,396	33.912	37	12	14
190	(Malvinas)	100	2,563	2,908	39.017	34	11	18
191	Finland	212,152	5,387,000	4,700,192	39.382	45	10	10
192	Switzerland	333,002	8,000,000	6,688,285	41.625	50	9	9
193	Estonia	55,255	1,294,000	981,467	42.701	56	8	7
194	Germany	3,602,587	81,799,600	67,621,622	44.042	53	7	8
195	Belgium	529,150	11,041,266	8,136,552	47.925	65	6	6
196	Ukraine	224,151	4,570,610	13,811,220	49.042	16	5	44
197	Italy	3,160,660	60,813,326	34,657,545	51.973	91	4	2
198	Luxembourg	29,788	517,000	457,451	57.617	65	3	5
199	Monaco	2,348	35,000	22,940	67.086	102	2	1
200	France	4,561,852	65,350,000	51,962,632	69.806	88	1	3
201	Poland	113,929	38,216,000	24,940,902	0.470	5	133	126
202	Indonesia	134,095	242,325,000	44,291,729	0.553	3	132	142

Quiz # 3





We use complex, automated tools successfully everyday to make our life simple and productive

... without knowing the mechanics.

Quiz # 1

An increasing number of people are using this device for a vast number of activities e.g. education, medical practice, business communication, flight tracking, special assistance communication and even for actuarial work.

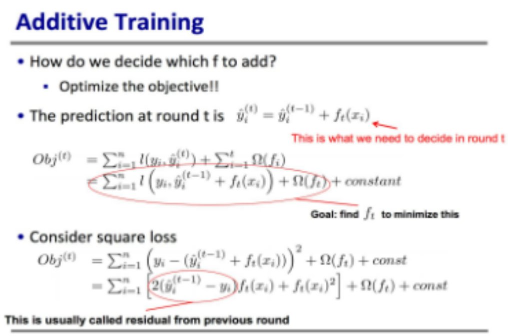
Quiz # 2

Hundreds of developers wrote millions of lines of C++ codes to build this platform that actuaries use extensively.

e:

Quiz # 3

Complex mechanics of Data Science ...



Additive Training

- How do we decide which f to add?
  - Optimize the objective!!
- The prediction at round t is $\hat{y}_i^{(t)} = \hat{y}_i^{(t-1)} + f_t(x_i)$

This is what we need to decide in round t

$$Obj^{(t)} = \sum_{i=1}^{n} l(y_i, \hat{y}_i^{(t)}) + \sum_{i=1}^{t} \Omega(f_i)$$
$$= \sum_{i=1}^{n} l\left(y_i, \hat{y}_i^{(t-1)} + f_t(x_i)\right) + \Omega(f_t) + constant$$

Goal: find f_t to minimize this

- Consider square loss

$$Obj^{(t)} = \sum_{i=1}^{n} \left(y_i - (\hat{y}_i^{(t-1)} + f_t(x_i))\right)^2 + \Omega(f_t) + const$$
$$= \sum_{i=1}^{n} \left[2(\hat{y}_i^{(t-1)} - y_i) f_t(x_i) + f_t(x_i)^2\right] + \Omega(f_t) + const$$

This is usually called residual from previous round

Ready to use, off-the-self open source implementation ...

Free Python and R implementation:

And with documentation:

mplex mechanics of Data Scienc

Additive Training

- How do we decide which f to add?
  - Optimize the objective!!
- The prediction at round t is $\hat{y}_i^{(t)} = \hat{y}_i^{(t-1)} + f_t(x_i)$

This is what we need to decide in round t

$$
\begin{aligned}
Obj^{(t)} &= \sum_{i=1}^{n} l(y_i, \hat{y}_i^{(t)}) + \sum_{i=1}^{t} \Omega(f_i) \\
&= \sum_{i=1}^{n} l\left(y_i, \hat{y}_i^{(t-1)} + f_t(x_i)\right) + \Omega(f_t) + constant
\end{aligned}
$$

Goal: find f_t to minimize this

- Consider square loss

$$
\begin{aligned}
Obj^{(t)} &= \sum_{i=1}^{n} \left(y_i - (\hat{y}_i^{(t-1)} + f_t(x_i))\right)^2 + \Omega(f_t) + const \\
&= \sum_{i=1}^{n} \left[2(\hat{y}_i^{(t-1)} - y_i) f_t(x_i) + f_t(x_i)^2\right] + \Omega(f_t) + const
\end{aligned}
$$

This is usually called residual from previous round

Chen, Singh, Taskar, Guestrin

Lemma 3.1,

$$
\begin{aligned}
&2p_i \sum_{s} \|P(y_s|\mathbf{x}, y_s = k) - P(y_s|\mathbf{x})\|_{tv} \\
&= 2p_i[d(t,t,k) + \sum_{s \neq t} d(s,t,k)] \\
&\leq 2p_i[d(t,t,k) + \sum_{s \neq t} d(t,t,k) \prod_{(a,b)\in\mathcal{Q}(s,t)} \alpha_{b,a}] \\
&= 2p_i(1-p_i)[1 + \sum_{s \neq t} \prod_{(a,b)\in\mathcal{Q}(s,t)} \alpha_{b,a}]
\end{aligned}
$$

Here the inequality is given by Corollary 3.1 ($d(s,t,k) \leq d(t,t,k) \prod_{(a,b)\in\mathcal{Q}(s,t)} \alpha_{b,a}$), and the last equality is given by Lemma 3.3 ($d(t,t,k) = 1-p_i$). Recall that $\mathbf{H}_{ii} = p_i(1-p_i)$, we have proved Theorem 4.1. □

The calculation of γ can be performed using dynamic programming. To explain the algorithm clearly, let us define an auxiliary message variable

$$\beta_{s\to t} \triangleq \alpha_{t,s} \sum_{h \in \mathcal{V}(\{s,t\};s)} \prod_{(a,b)\in\mathcal{Q}(h,s)} (\alpha_{b,a} + 1) \quad (16)$$

We can rearrange $\gamma_i^{(n)}$ in Theorem 4.1 in terms of β, where the index i satisfies $\mu_i = \mathbb{1}(y_t = k)$, as

$$\gamma_i^{(n)}(\mathbf{y}, \mathbf{x}) = 2\Big(1 + \sum_{s:(s,t)\in\mathcal{E}} \beta_{s\to t}\Big).$$

Similarly for $\gamma_i^{(e)}$ in Theorem 4.2, where the index i satisfies $\mu_i = \mathbb{1}(y_t = k_1, y_{t'} = k_2)$, we have

$$\gamma_i^{(e)}(\mathbf{y}, \mathbf{x}) = 2\Big(3 + \sum_{s:(s,t)\in\mathcal{E}, s\neq t'} \beta_{s\to t} + \sum_{s:(s,t')\in\mathcal{E}, s\neq t} \beta_{s\to t'}\Big).$$

The calculation of β can be done efficiently using a message passing algorithm with the following update.

$$\beta_{s\to t} \leftarrow \alpha_{t,s}\Big(1 + \sum_{h:(h,s)\in\mathcal{E}, h\neq t} \beta_{h\to s}\Big) \quad (17)$$

In the case of linear chain CRF, our problem is reduced to the calculation of $\beta_{t+1\to t} \triangleq \sum_{s=t}^{n} \prod_{i=t}^{s-1} \alpha_{i,i+1}$ and $\beta_{t-1\to t} \triangleq \sum_{s=1}^{t} \prod_{i=s+1}^{t} \alpha_{i,i-1}$, and the message updates in Eq. 17 correspond to a forward-backward algorithm using the following recursion formula:

$$\beta_{t+1\to t} = \alpha_{t,t+1}(1 + \beta_{t+2\to t+1})$$

Algorithm 1 Gradient Boosting for CRF

```
repeat
  for U ∈ {N, E} do
    for y, x ∈ D in parallel do
      {inference of p_i, γ_i are done using dynamic programming}
      Infer G_i(y, x) ← μ_i(y) − p_i,
            H_ii(y, x) ← p_i(1 − p_i) for each i ∈ U
      Infer γ_i(y, x) using dynamic programming
            for each i ∈ U
    end for
    for [c] ⊂ U in parallel do
      {We use [c] to enumerate over set of equivalent index defined by C in U}
      δ_c ← argmin_{δ∈F_N} Ω(φ_i + δ)
            + Σ_{i∈[c]} Σ_{y,x∈D} [G_i(y, x)δ(y, x)
            + γ_i(y, x)H_ii(y, x)δ²(y, x)]
      φ_c ← φ_c + εδ_c
    end for
  end for
until convergence
```

Corollary 4.1. *When $\mathcal{U}$ is the index set of node potentials, $\gamma_i = 2n$ satisfies Eq. (6), where n is the number of nodes in the CRF.*

Based on Theorem 4.1 and 4.2, we can get an efficient gradient boosting algorithm for CRF (GBCRF), which is presented in Algorithm 1. Here ϵ is a shrinkage term used to avoid overfitting. Our algorithm adaptively estimates γ via the mixing rate calculation at each iteration. At the beginning, when each variable is nearly independent from each other, we will have a γ that is close to 2 (and thus the updates are aggressive). γ increases as the variables become dependent on each other (resulting in more conservative updates).

Relation to LogitBoost: Our algorithm can be viewed as a generalization of multi-class classification using LogitBoost [Friedman et al., 1998]. When $\mathcal{E} = \emptyset$ in Eq. (2), our model degenerates to the LogiBoost model. In this case, the variables in each position are independent, the estimation of γ is 2, and Algorithm 1 is exactly equivalent to LogitBoost. When the variables are dependent on each other, which is common in structured prediction, our model estimates the dependency level via the Markov Chain mixing rate to guide the boosting objective in each iteration.

Ready to use, off-the-self open source implementation ...

Free Python and R Implementation:
https://github.com/dmlc/xgboost

And with documentation:

https://github.com/dmlc/xgboost/blob/master/doc/python/python_intro.md

Ready to use, off-the-self open source implementation ...

Free Python and R Implementation:
https://github.com/dmlc/xgboost

And with documentation:

https://github.com/dmlc/xgboost/blob/master/doc/python/python_intro.md

https://github.com/dmlc/xgboost/blob/master/R-package/vignettes/xgboostPresentation.Rmd

How to use: complete Step-by-step real life illustration:
https://github.com/dmlc/xgboost/blob/master/demo/kaggle-otto/understandingXGBoostModel.Rmd

No need to learn programming to get started

```python
PACK_LEN1 = Struct('!BB').pack
PACK_LEN2 = Struct('!BBH').pack
PACK_LEN3 = Struct('!BBQ').pack
PACK_CLOSE_CODE = Struct('!H').pack
MSG_SIZE = 2 ** 14


class WebSocketError(Exception):
    """WebSocket protocol parser error."""

    def __init__(self, code, message):
        self.code = code
        super().__init__(message)


def WebSocketParser(out, buf):
    while True:
        fin, opcode, payload = yield from parse_frame(buf)

        if opcode == OPCODE_CLOSE:
            if len(payload) >= 2:
                close_code = UNPACK_CLOSE_CODE(payload[:2])[0]
                if close_code not in ALLOWED_CLOSE_CODES and close_code < 3000:
                    raise WebSocketError(
                        CLOSE_PROTOCOL_ERROR,
```

```r
#' p <- progress_estimated(3, min_time = 3)
#' for (i in 1:3) p$pause(0.1)$tick()$print()
#'
#' \dontrun{
#' p <- progress_estimated(10, min_time = 3)
#' for (i in 1:10) p$pause(0.5)$tick()$print()
#' }
progress_estimated <- function(n, min_time = 0) {
  Progress$new(n, min_time = min_time)
}

#' @importFrom R6 R6Class
Progress <- R6::R6Class("Progress",
  public = list(
    n = NULL,
    i = 0,
    init_time = NULL,
    stopped = FALSE,
    stop_time = NULL,
    min_time = NULL,

    initialize = function(n, min_time = 0, ...) {
      self$n <- n
      self$min_time <- min_time
      self$begin()
```

Problem solving with automated platform

Quick learning & smart application of advanced algorithms

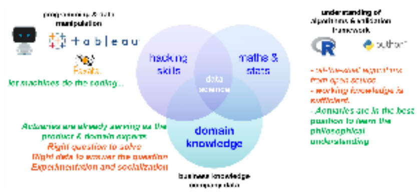
Activity	Tool/ Platform/ Source	Learning Focus
Data manipulation & General programming	R python tamr Paxata	- data manipulation - key statistical packages - key visualization packages
Visualizations	Qlik tableau	- visualize data by drag & drop - productize your solution
Automated Modeling, Machine Learning	rapidminer sas DataRobot	- defining the right question - interpreting results - running experiments using the automated platform

A pragmatic approach for practitioners

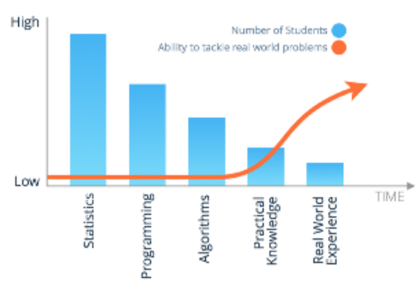
Learning Data Science with **automated platform** and with **business outcome in focus**

Revisting data science venn-diagram

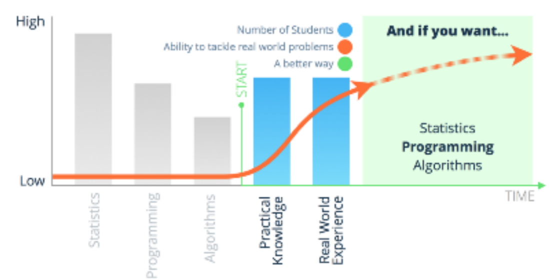
Augmentation
Man & machine working together



Traditional Data Science Education Approach
Long learning cycle, higher drop outs



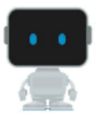
Pragmatic education with modern automation
Fast learning cycle, less dropouts



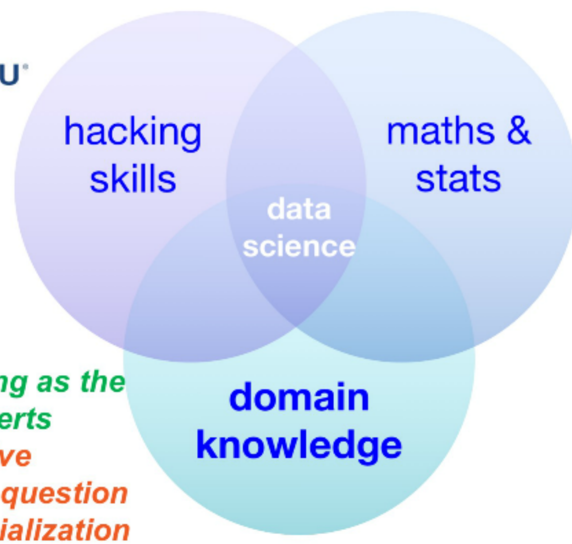
Revisting data science venn-diagram

Augmentation
Man & machine working together

programming & data manipulation



let machines do the coding...



understanding of algorithms & validation framework

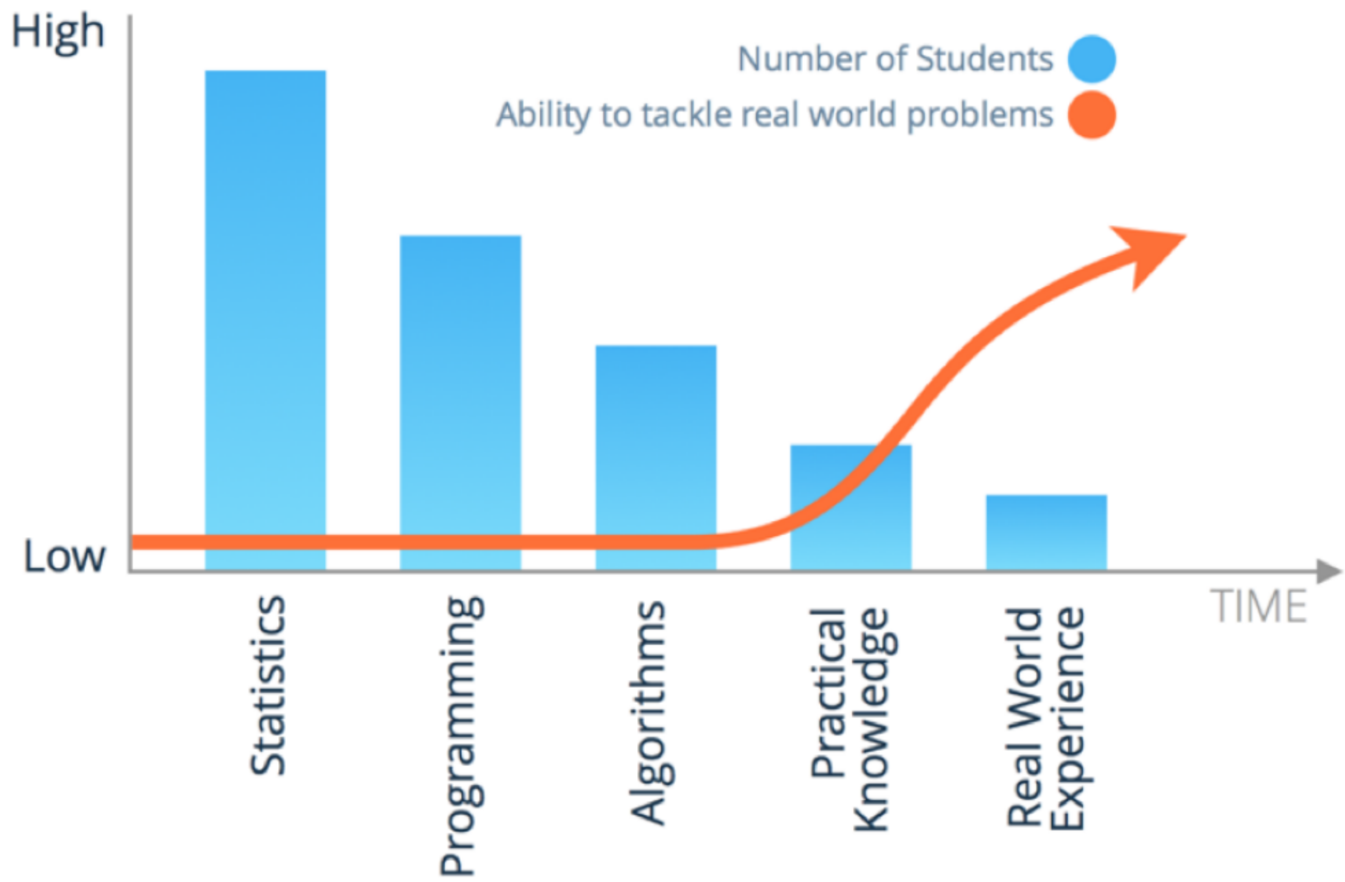
- off-the-shelf algorithms from open source.
- working knowledge is sufficient.
- Actuaries are in the best position to learn the philosophical understanding

Actuaries are already serving as the product & domain experts
Right question to solve
Right data to answer the question
Experimentation and socialization

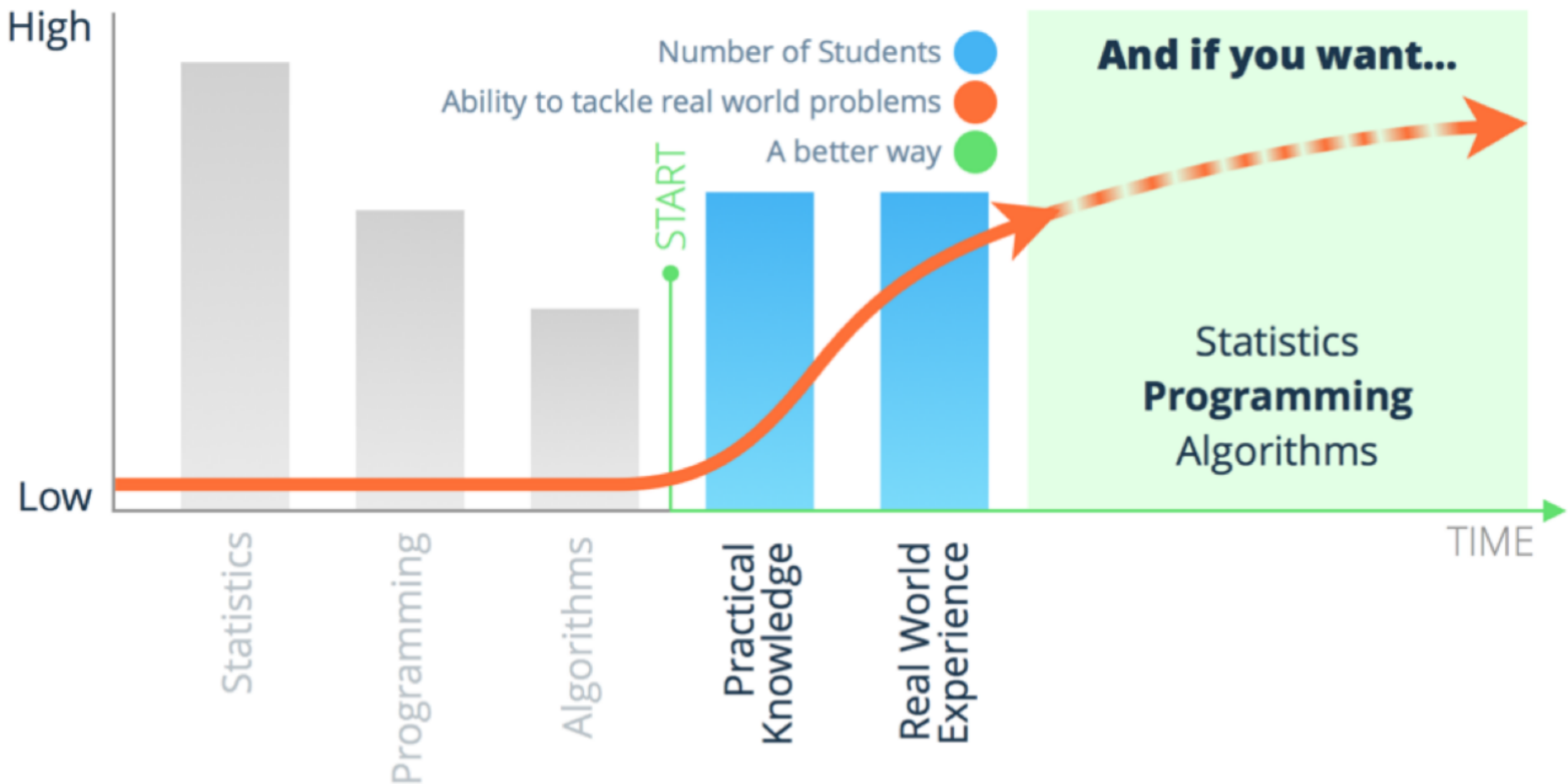
business knowledge
company data

Traditional Data Science Education Approach

Long learning cycle, higher drop outs



Pragmatic education with modern automation
Fast learning cycle, less dropouts

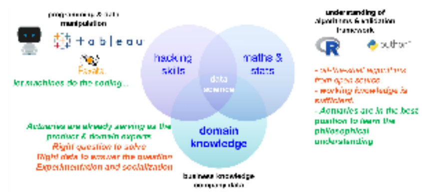


A pragmatic approach for practitioners

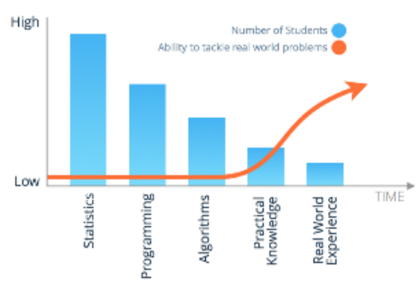
Learning Data Science with **automated platform** and with **business outcome in focus**

Revisting data science venn-diagram

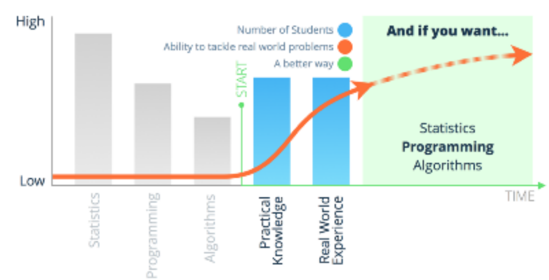
Augmentation
Man & machine working together



Traditional Data Science Education Approach
Long learning cycle, higher drop outs

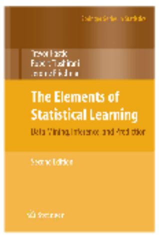


Pragmatic education with modern automation
Fast learning cycle, less dropouts



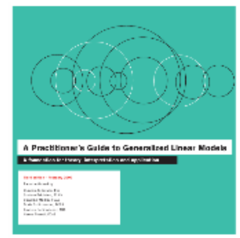
Once you are familiar with real world experience & practical knowledge...

Take advantage of free educational resources out there



Machine Learning
(Coursera)
Andrew Ng
https://www.coursera.org/course/ml

http://statweb.stanford.edu/~tibs/ElemStatLearn/



Udacity (http://www.udacity.com)
CS 101: Introduction to Computer Science
Dave Evans *(Beginner Python)*

CS 212: Design Of Computer Programs
Peter Norvig *(Intermediate Python)*

An Introduction To Statistical Learning

http://www-bcf.usc.edu/~gareth/ISL/

Coursera (http://www.coursera.com)
Roger Peng
Jeff Leek

The swirl package
A Guided Introduction to Learning R
http://swirlstats.com/

python™

Udacity (http://www.udacity.com)

CS 101: Introduction to Computer Science

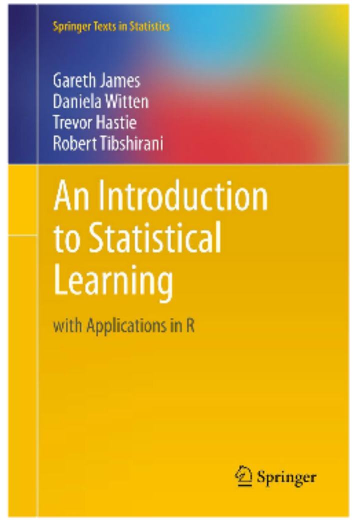
Dave Evans *(Beginner Python)*

CS 212: Design Of Computer Programs

Peter Norvig *(Intermediate Python)*

UDACITY

An Introduction To
Statistical Learning



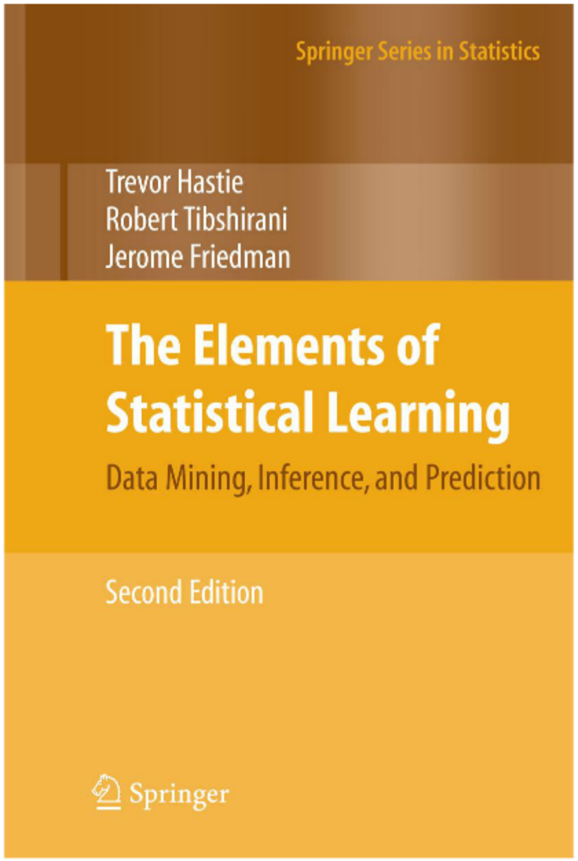
http://www-bcf.usc.edu/~gareth/ISL/

Coursera (http://www.coursera.com)
Roger Peng
Jeff Leek

Johns Hopkins Bloomberg School of Public Health

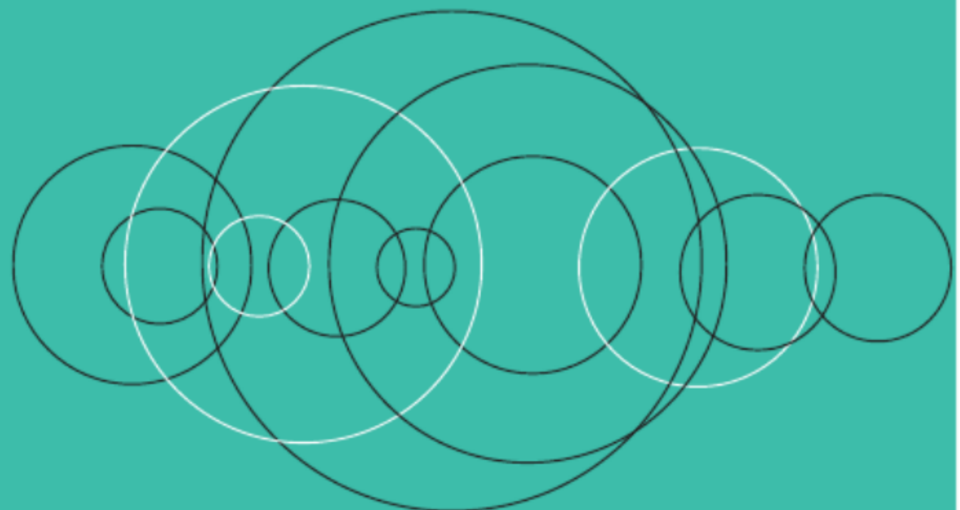
The swirl package
A Guided Introduction to Learning R
http://swirlstats.com/

{swirl} statistics with interactive R learning: an R package



Machine Learning (Coursera)
Andrew Ng
https://www.coursera.org/course/ml

http://statweb.stanford.edu/~tibs/ElemStatLearn/



A Practitioner's Guide to Generalized Linear Models

A foundation for theory, interpretation and application

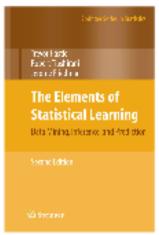
Third edition – February 2007

Paper authored by:

Duncan Anderson, FIA
Sholom Feldblum, FCAS
Claudine Modlin, FCAS
Doris Schirmacher, FCAS
Ernesto Schirmacher, ASA
Neeza Thandi, FCAS

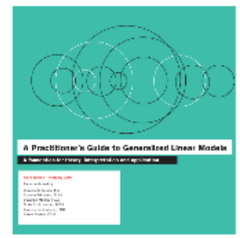
Once you are familiar with real world experience & practical knowledge...

Take advantage of free educational resources out there



Machine Learning
(Coursera)
Andrew Ng
https://www.coursera.org/course/ml

http://statweb.stanford.edu/~tibs/ElemStatLearn/



python™

Udacity (http://www.udacity.com)
CS 101: Introduction to Computer Science
Dave Evans *(Beginner Python)*

CS 212: Design Of Computer Programs
Peter Norvig *(Intermediate Python)*

UDACITY

An Introduction To Statistical Learning

http://www-bcf.usc.edu/~gareth/ISL/

Coursera (http://www.coursera.com)
Roger Peng
Jeff Leek

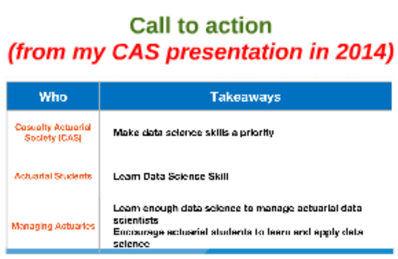
JOHNS HOPKINS

The swirl package
A Guided Introduction to Learning R
http://swirlstats.com/

{swirl}

Call to action

Take the unicorn by the horn

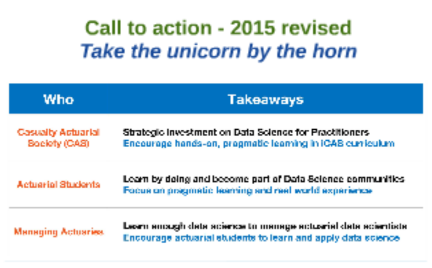


Call to action
(from my CAS presentation in 2014)

Who	Takeaways
Casualty Actuarial Society (CAS)	Make data science skills a priority
Actuarial Students	Learn Data Science Skill
Managing Actuaries	Learn enough data science to manage actuarial data scientists Encourage actuarial students to learn and apply data science

Strategic Investment in Actuarial Data Scientist

CAS EXPANDS INTO SPECIALTY CREDENTIALS; NEW CAS INSTITUTE TO LAUNCH CREDENTIAL IN PREDICTIVE ANALYTICS AND DATA SCIENCE



Call to action - 2015 revised
Take the unicorn by the horn

Who	Takeaways
Casualty Actuarial Society (CAS)	Strategic Investment on Data Science for Practitioners Encourage hands-on, pragmatic learning in iCAS curriculum
Actuarial Students	Learn by doing and become part of Data Science communities Focus on pragmatic learning and real world experience
Managing Actuaries	Learn enough data science to manage actuarial data scientists Encourage actuarial students to learn and apply data science

Call to action
(from my CAS presentation in 2014)

Who	Takeaways
Casualty Actuarial Society (CAS)	Make data science skills a priority
Actuarial Students	Learn Data Science Skill
Managing Actuaries	Learn enough data science to manage actuarial data scientists Encourage actuarial students to learn and apply data science

Strategic Investment in Actuarial Data Scientist

CAS EXPANDS INTO SPECIALTY CREDENTIALS; NEW CAS INSTITUTE TO LAUNCH CREDENTIAL IN PREDICTIVE ANALYTICS AND DATA SCIENCE

11/16/2015 —

Arlington, VA, November 16, 2015 – The Casualty Actuarial Society (CAS) announces the creation of The CAS Institute, an organization offering new credentials and specialized professional education for quantitative professionals looking to remain current in their field. The CAS Institute will develop a curriculum for each of its offered specialty areas, initially covering advanced topics such as predictive analytics and data science. CAS President Bob Miccolis, FCAS, formally introduced The CAS Institute today during the CAS Annual Meeting in Philadelphia.

The CAS Institute is a subsidiary of the CAS and brings the rigorous CAS educational standards to a wider community of quantitative specialists seeking to earn specialized, in-demand credentials and quality professional education to address talent demands, both today and in the future. The CAS will continue to offer actuarial credentials, while The CAS Institute will offer specialty credentials.

Call to action - 2015 revised

Take the unicorn by the horn

Who	Takeaways
Casualty Actuarial Society (CAS)	Strategic investment on Data Science for Practitioners Encourage hands-on, pragmatic learning in iCAS curriculum
Actuarial Students	Learn by doing and become part of Data Science communities Focus on pragmatic learning and real world experience
Managing Actuaries	Learn enough data science to manage actuarial data scientists Encourage actuarial students to learn and apply data science

Call to action
Take the unicorn by the horn

Call to action
(from my CAS presentation in 2014)

Who	Takeaways
Casualty Actuarial Society (CAS)	Make data science skills a priority
Actuarial Students	Learn Data Science Skill
Managing Actuaries	Learn enough data science to manage actuarial data scientists Encourage actuarial students to learn and apply data science

Strategic Investment in Actuarial Data Scientist

CAS EXPANDS INTO SPECIALTY CREDENTIALS; NEW CAS INSTITUTE TO LAUNCH CREDENTIAL IN PREDICTIVE ANALYTICS AND DATA SCIENCE

Call to action - 2015 revised
Take the unicorn by the horn

Who	Takeaways
Casualty Actuarial Society (CAS)	Strategic Investment on Data Science for Practitioners Encourage hands-on, pragmatic learning in iCAS curriculum
Actuarial Students	Learn by doing and become part of Data Science communities Focus on pragmatic learning and real world experience
Managing Actuaries	Learn enough data science to manage actuarial data scientists Encourage actuarial students to learn and apply data science

What actuaries **DON'T need** to know

Mechanics underlying complex tools, platforms and algorithms

A pragmatic approach for practitioners

Learning Data Science with **automated platform** and with **business outcome in focus**

Once you are familiar with real world experience & practical knowledge...

Take advantage of free educational resources out there

Call to action

Take the unicorn by the horn

Data Science:
What Actuaries (DON'T) Need to Know

What actuaries **need** to know

Data Science in Insurance

Why it's more important now than ever before

What is Data Science?

What is a Data Scientist?

Where to Find Data Scientists for Your Insurance Company

What actuaries **DON'T need** to know

Mechanics underlying complex tools, platforms and algorithms

A pragmatic approach for practitioners

Learning Data Science with automated platform and with business outcome in focus

Once you are familiar with real world experience & practical knowledge...

Take advantage of free educational resources out there

Call to action

Take the unicorn by the horn

Jeremy Achin
CEO & Co-founder, DataRobot Inc.

Questions?

DataRobot

Additional questions, comments, queries:

jeremy@datarobot.com