Cleaning Cybersecurity's Dirty Data Projects:

Exploring Solutions to Common Challenges in Applying Machine Learning to Cybersecurity

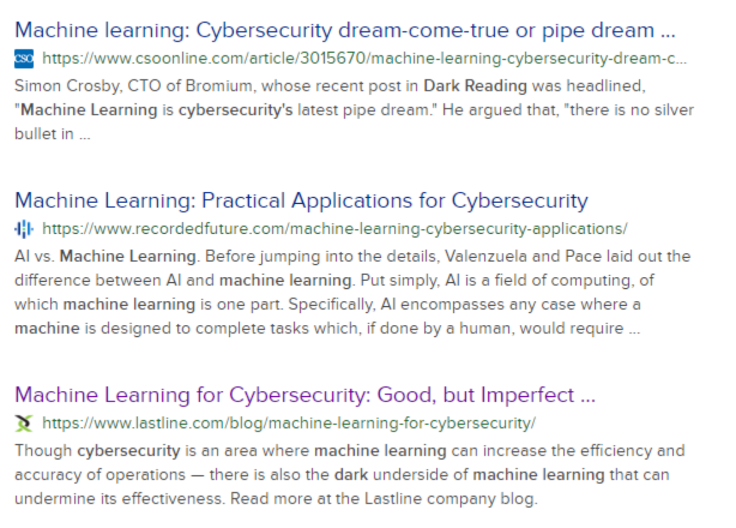
Name: Jorly

Date: May 27, 2019

Disclaimer

The statements made in this presentation does not represent the thoughts, intentions, plans or strategies of my employer. It is solely my opinion.

Inspiration for this talk



Machine learning: Cybersecurity dream-come-true or pipe dream ...

https://www.csoonline.com/article/3015670/machine-learning-cybersecurity-dream-c...

Simon Crosby, CTO of Bromium, whose recent post in **Dark Reading** was headlined, "**Machine Learning** is **cybersecurity's** latest pipe dream." He argued that, "there is no silver bullet in ...

Machine Learning: Practical Applications for Cybersecurity

https://www.recordedfuture.com/machine-learning-cybersecurity-applications/

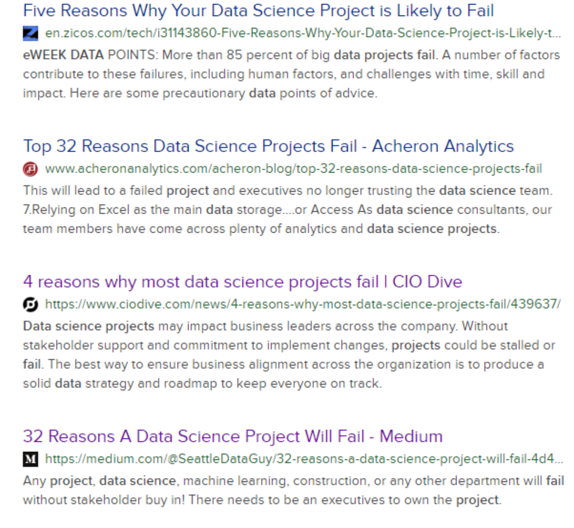
AI vs. **Machine Learning**. Before jumping into the details, Valenzuela and Pace laid out the difference between AI and **machine learning**. Put simply, AI is a field of computing, of which **machine learning** is one part. Specifically, AI encompasses any case where a **machine** is designed to complete tasks which, if done by a human, would require ...

Machine Learning for Cybersecurity: Good, but Imperfect ...

https://www.lastline.com/blog/machine-learning-for-cybersecurity/

Though **cybersecurity** is an area where **machine learning** can increase the efficiency and accuracy of operations — there is also the **dark** underside of **machine learning** that can undermine its effectiveness. Read more at the Lastline company blog.

Inspiration for this talk



Five Reasons Why Your Data Science Project is Likely to Fail

en.zicos.com/tech/i31143860-Five-Reasons-Why-Your-Data-Science-Project-is-Likely-t...

eWEEK DATA POINTS: More than 85 percent of big **data projects fail**. A number of factors contribute to these failures, including human factors, and challenges with time, skill and impact. Here are some precautionary **data** points of advice.

Top 32 Reasons Data Science Projects Fail - Acheron Analytics

www.acheronanalytics.com/acheron-blog/top-32-reasons-data-science-projects-fail

This will lead to a failed **project** and executives no longer trusting the **data science** team. 7.Relying on Excel as the main **data** storage....or Access As **data science** consultants, our team members have come across plenty of analytics and **data science projects**.

4 reasons why most data science projects fail | CIO Dive

https://www.ciodive.com/news/4-reasons-why-most-data-science-projects-fail/439637/

Data science projects may impact business leaders across the company. Without stakeholder support and commitment to implement changes, **projects** could be stalled or **fail**. The best way to ensure business alignment across the organization is to produce a solid **data** strategy and roadmap to keep everyone on track.

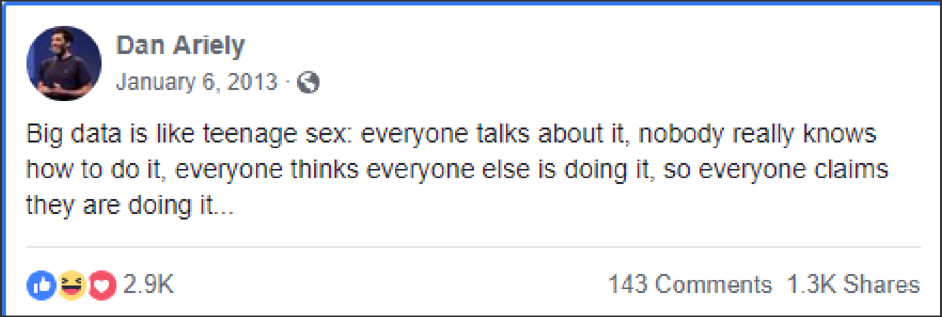
32 Reasons A Data Science Project Will Fail - Medium

https://medium.com/@SeattleDataGuy/32-reasons-a-data-science-project-will-fail-4d4...

Any **project**, **data science**, machine learning, construction, or any other department will **fail** without stakeholder buy in! There needs to be an executives to own the **project**.

What is Machine Learning?

This joke also applies to machine learning:



Dan Ariely

January 6, 2013 ·

Big data is like teenage sex: everyone talks about it, nobody really knows how to do it, everyone thinks everyone else is doing it, so everyone claims they are doing it...

2.9K 143 Comments 1.3K Shares

What is Machine Learning?

There are many definitions, I will share you one that was mentioned by Andrew Ng in his Machine Learning course:

> "A computer program is said to learn from experience E with respect to some class of tasks T and performance measure P, if its performance at tasks in T, as measured by P, improves with experience E." *Tom M. Mitchell*

What is Machine Learning?

A handwriting recognition learning problem:

* Task T: recognizing and classifying handwritten words within images
* Performance measure P: percent of words correctly classifier
* Training experience E: a database of handwritten words with given classifications

Image reference: Dataaspirant

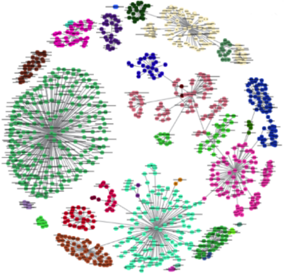
Vague Business Problem

- Business problems are not specific or well-defined
- Assumptions are easily made
- Not yet consulting with subject matter experts

Image reference:
CG Magazine

Disregard for Minimum Viable Product(s) (MVPs)

Some include models, dashboards, web applications, etc.



Item	Costumer	Price	Product
Htc Mobile	Alan B	$438	B-ZTF
Samsung Mobile	Joe Smith	$980	C-687
Apple Mobile	Patrick J	$680	E-597
LG Mobile	Anna Han	$340	Z-423
Nokia Mobile	Mark Mule	$230	W-225

Image references:
Cambridge Intelligence
JotForm
appsbuilder

Measuring the Successes of Outcomes is an Afterthought

- Difficulties in determining success
  - faster detection/response
  - reducing analyst workload
- Choosing appropriate success metrics
- Responsibilities in making these decisions

Image reference:
Pinterest

Nonstop Data Issues

- Determining appropriate (contextual) data sources
- Limited expertise and documentation
- Difficulties in aggregating multiple mountains of data
- Storage and computation issues
- Dynamic (behaviors, fields, etc.) changes in the data
- Learning from skewed data (99% good, 1% bad)

SEE, I TOLD YOU THAT BIG DATA WAS TOO SCARY
BIG DATA
KRONOS
Workforce Innovation That Works™
CREATED BY marketoon studios
KRONOS.COM

Image reference:
socedo

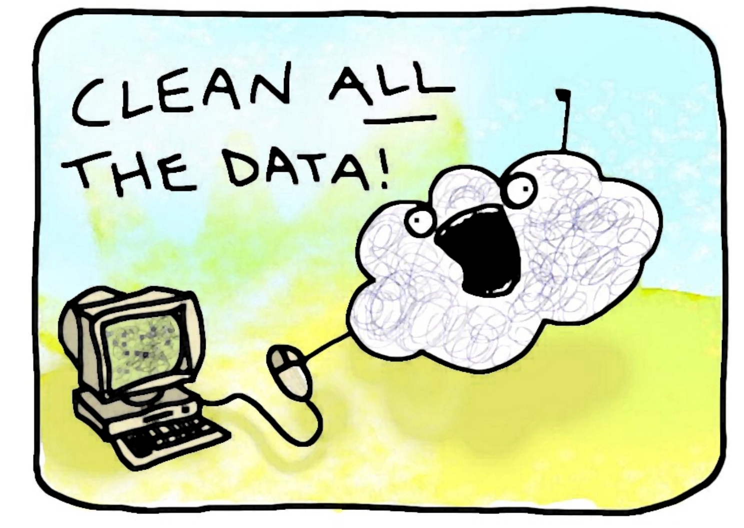
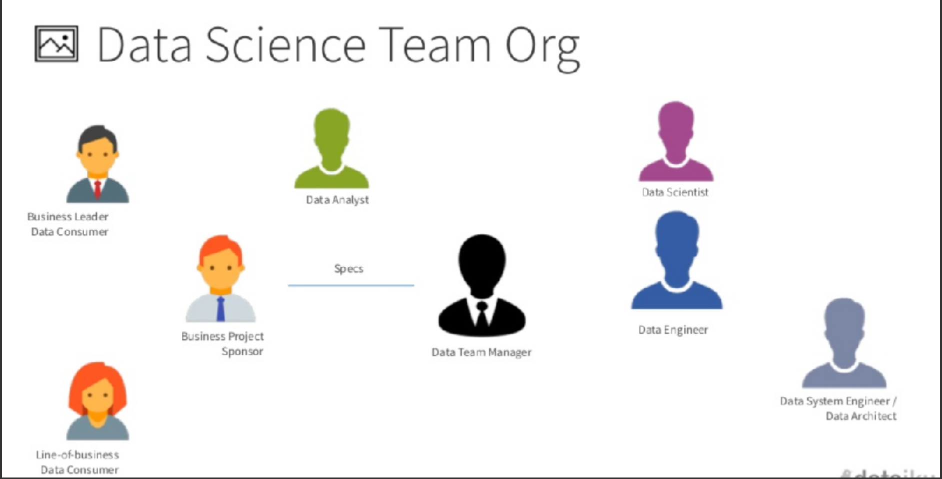


Image reference:
redpath

Not Enough Thought in Putting Together a Data Science Team



Data Science Team Org

Business Leader
Data Consumer

Data Analyst

Data Scientist

Business Project
Sponsor

Specs

Data Team Manager

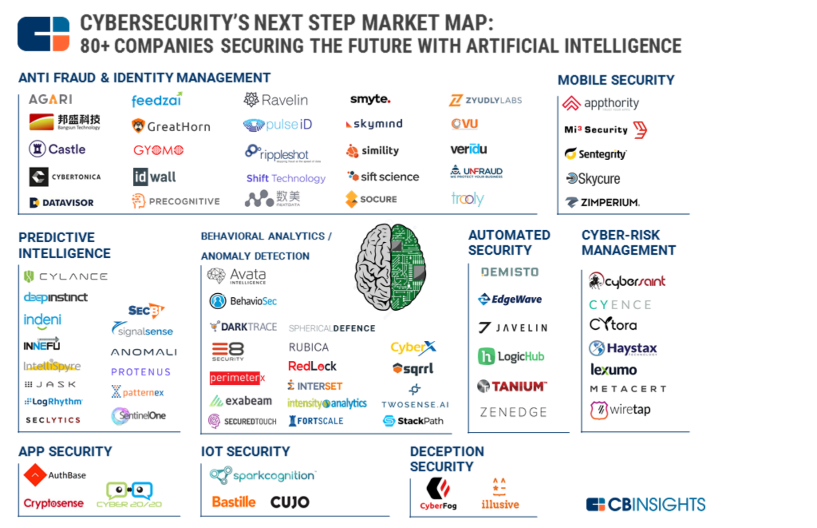
Data Engineer

Data System Engineer /
Data Architect

Line-of-business
Data Consumer

Image reference:
Dataiku

Blind Reliance on Tools to Solve Machine Learning Problems



CYBERSECURITY'S NEXT STEP MARKET MAP: 80+ COMPANIES SECURING THE FUTURE WITH ARTIFICIAL INTELLIGENCE

ANTI FRAUD & IDENTITY MANAGEMENT

AGARI
邦盛科技 Bangsun Technology
Castle
CYBERTONICA
DATAVISOR
feedzai
GreatHorn
GYOMO
idwall
PRECOGNITIVE
Ravelin
pulseiD
rippleshot
Shift Technology
数美 NEXTDATA
smyte.
skymind
simility
sift science
SOCURE
ZYUDLYLABS
CVU
veridu
UNFRAUD
trooly

MOBILE SECURITY

appthority
Mi3 Security
Sentegrity
Skycure
ZIMPERIUM.

PREDICTIVE INTELLIGENCE

CYLANCE
deepinstinct
indeni
INNEFU
IntelliSpyre
JASK
LogRhythm
SECLYTICS
SECBI
signalsense
ANOMALI
PROTENUS
patternex
SentinelOne

BEHAVIORAL ANALYTICS / ANOMALY DETECTION

Avata INTELLIGENCE
BehavioSec
DARKTRACE
E8 SECURITY
perimeterx
exabeam
SECUREDTOUCH
SPHERICALDEFENCE
RUBICA
RedLock
INTERSET
intensityanalytics
FORTSCALE
CyberX
sqrrl
TWOSENSE.AI
StackPath

AUTOMATED SECURITY

DEMISTO
EdgeWave
JAVELIN
LogicHub
TANIUM
ZENEDGE

CYBER-RISK MANAGEMENT

cybersaint
CYENCE
CYtora
Haystax
lexumo
METACERT
wiretap

APP SECURITY

AuthBase
Cryptosense
CYBER 20/20

IOT SECURITY

sparkcognition
Bastille
CUJO

DECEPTION SECURITY

CyberFog
illusive

CBINSIGHTS

Practical Tips

- Use a project intake form/front door request
- Include SMEs early on
- Build a data science team with a broad set of skills
- Use tools for documenting and tracking work
- Figure out how to integrate tools
- If possible, use agile (fail fast)

Last Thoughts

"Unfortunately, machine learning will never be a silver bullet for cybersecurity compared to image recognition or natural language processing, two areas where machine learning is thriving. There will always be a person who tries to find issues in our systems and bypass them." *Alexander Polyakov*

Reference

- Troy Hiltbrand. *Cybersecurity Plus Data Science: The Career Path of the Future?*. https://tdwi.org/articles/2018/01/16/
adv-all-cybersecurity-plus-data-science-future-career-path
aspx

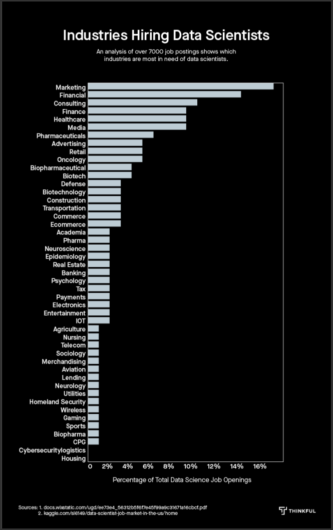
- Adam Levenson. *Insights on the Data Science Job Market: Analyzing 7k Data Science Job Descriptions.* https://www.thinkful.com/blog/
insights-on-the-data-science-job-market-analyzing-7k-data-

- Tom M. Mitchell. (1997). *Machine Learning.*

- Steve Zurier. *6 Steps for Applying Data Science to Security.* https://www.darkreading.com/analytics/6-steps-for-applying-data-science-to-security/d/d-id/1331840?image_number=1

Inspiration for this talk

In David Antzelevich's Data Science job market project, Cybersecurity was the least to hire data scientists. Larger image



Industries Hiring Data Scientists

An analysis of over 7000 job postings shows which industries are most in need of data scientists.

Sources: 1. docs.wixstatic.com/ugd/ee73e4_56312b5f6f7e45f99a9c31671a16cbcf.pdf
2. kaggle.com/sl6149/data-scientist-job-market-in-the-us/home