TOWARDS IMPACTFUL DATA SCIENCE

BY BRIAN D'ALESSANDRO

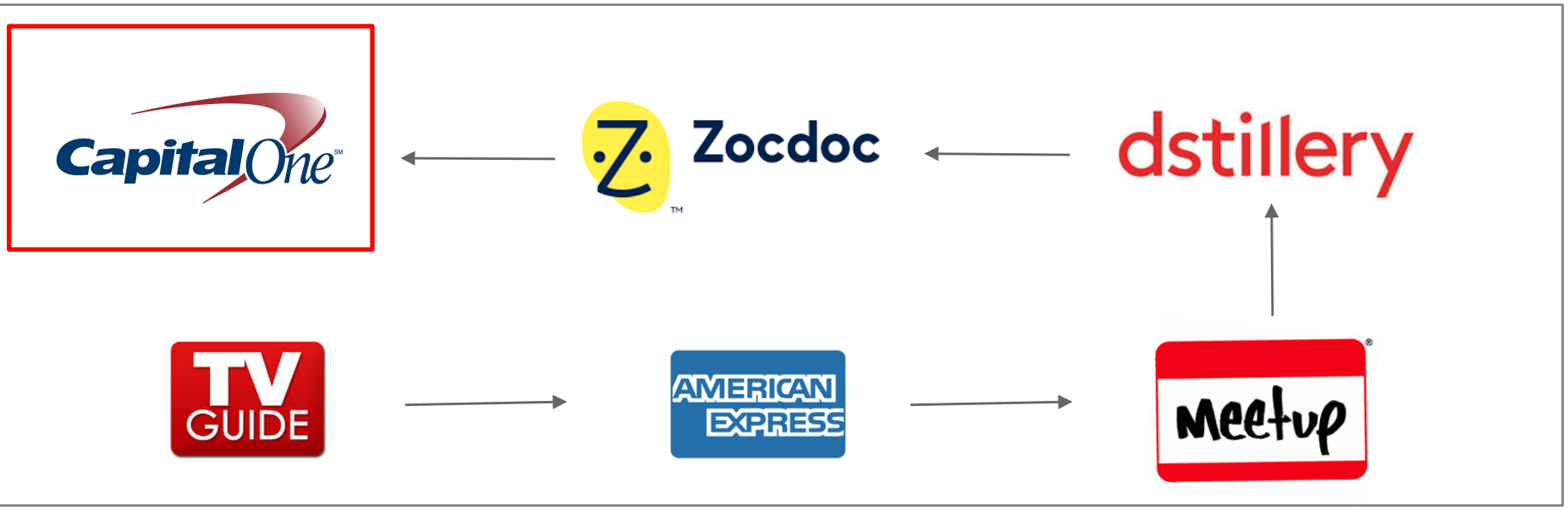
PROLOGUE

THE SPEAKER CHRONOLOGY

Sr Director Data Science, Capital One*

18 years leading Data Science & Analytics across multiple industries

Professor of Data Science at NYU.



**All statements in this presentation represent my own views and opinions and don't necessarily represent the opinions, views or state of any past and present employer.*

WHAT'S IN IT FOR YOU?

Data Scientist: You have to answer to someone, and ultimately justify your salary. Learn to manage yourself (and others) through a lens that leads to business impact.

Non-Data Scientist: Most of DS work may seem like hieroglyphics, but you're a stakeholder. But you need to understand that managing DS is different than most other disciplines, and if you manage it well, rewards will follow.

A PROPOSITION

Doing “Data Science” is easy…

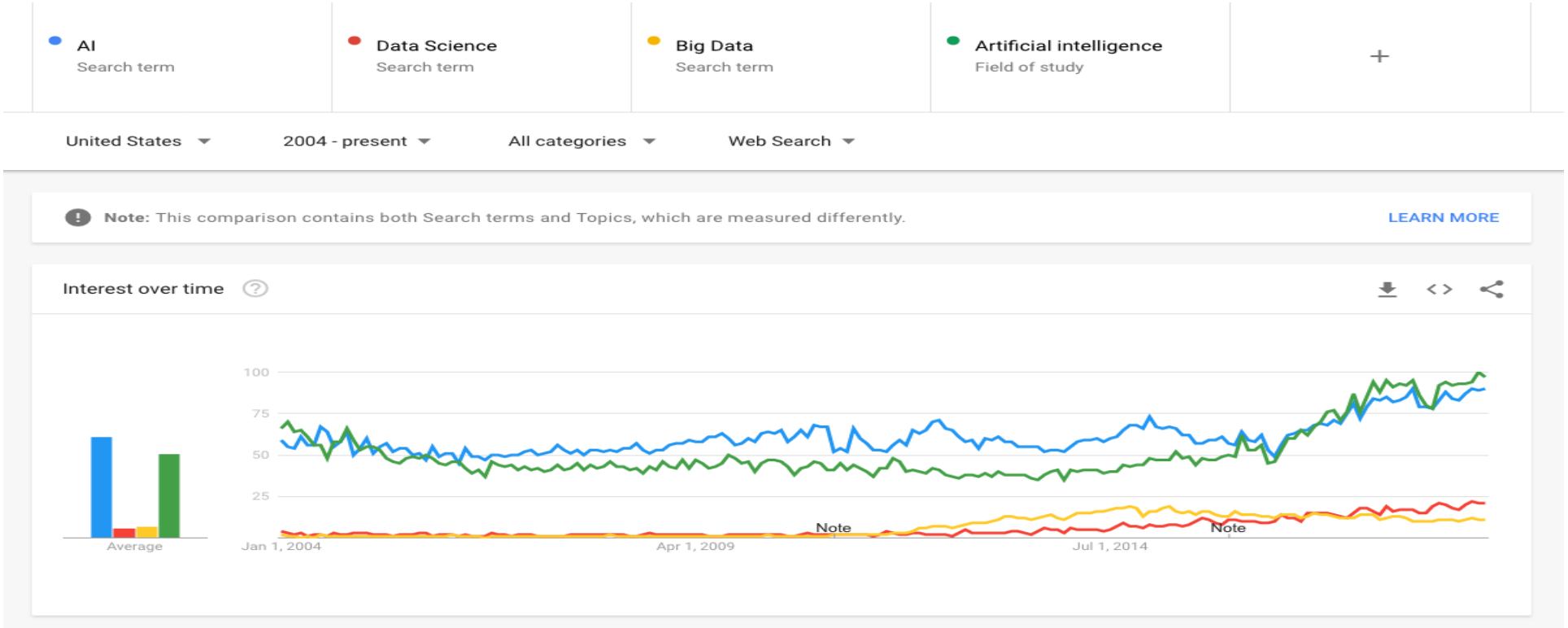
```
from sklearn.tree import DecisionTreeClassifier
tree = DecisionTreeClassifier(criterion="entropy")
tree.fit(train_df.drop([label], axis=1), train_df[label])
```

Creating measurable value through Data Science is hard!

TROUBLE IN PARADISE

STATING THE OBVIOUS

A.I. and associated concepts are getting increasingly popular (*source: Google Trends*)



BIG BUSINESSES HAVE BOUGHT INTO THE HYPE

>90% of surveyed Fortune 100 executives are increasing the pace of investment in Big Data & AI technologies, and the investments are large.

Investment in Big Data/AI	2018	2019
Greater than $500M	12.7%	21.1%
$50M -- $500M	27.0%	33.9%
Under $50M	60.3%	45.0%

The amount of investment is staggering!

Investment in Disruptive Technologies	2017	2018	2019
AI/Machine Learning	68.9%	90.4%	96.4%
Cloud Computing	85.2%	80.8%	90.5%
Digital Technologies	78.7%	64.4%	77.4%
FinTech Solutions	45.9%	54.8%	47.6%
Blockchain	37.7%	53.4%	41.7%

And it is diverse.

Source: New Vantage Partners "Big Data and AI Executive Survey 2019" http://newvantage.com/wp-content/uploads/2018/12/Big-Data-Executive-Survey-2019-Findings-122718.pdf

BUT ARE THEY SATISFIED?

The result of this investment is often questionable. In particular, businesses have a hard time incorporating AI/Big Data applications into their work, contributing to unclear impact.

Business Adoption of Big Data/AI a Challenge	2018	2019
Yes	64.7%	77.1%
No	35.3%	22.9%

The vast majority find adoption to be a challenge.

Measurable Results from Big Data/AI	2017	2018	2019
Yes	48.4%	73.2%	62.2%
No | Too Early To Tell	51.6%	26.8%	37.8%

And still too many aren't sure of the impact.

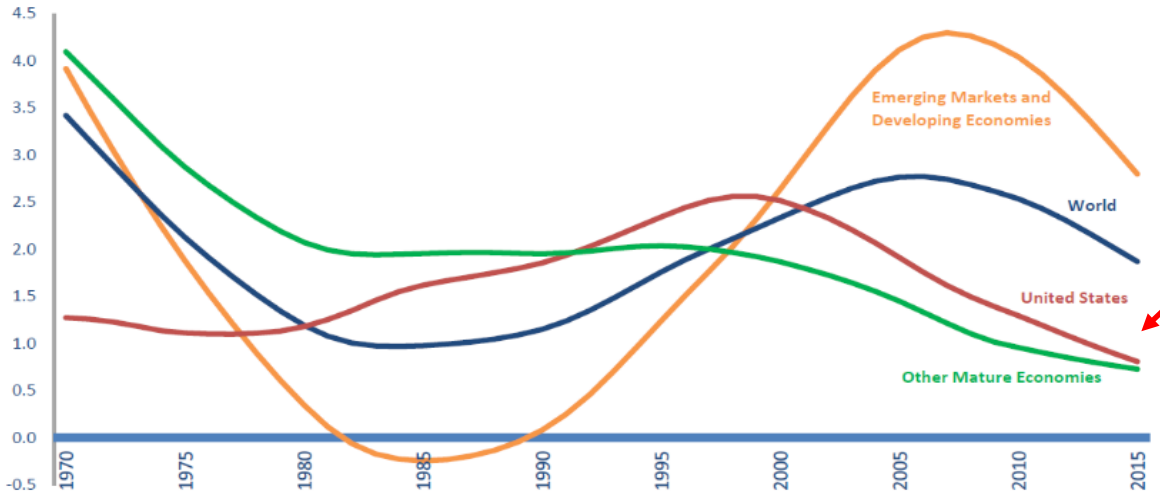
Source: New Vantage Partners "Big Data and AI Executive Survey 2019" http://newvantage.com/wp-content/uploads/2018/12/Big-Data-Executive-Survey-2019-Findings-122718.pdf

MACROECONOMISTS ARE TAKING NOTICE

MIT IDE RESEARCH BRIEF VOL. 2018.01

AI AND THE MODERN PRODUCTIVITY PARADOX: A CLASH OF EXPECTATIONS AND STATISTICS

Erik Brynjolfsson, Daniel Rock, and Chad Syverson



Source: The Conference Board Total Economy Database™ (Adjusted version), November 2016.
Notes: Trend growth rates are obtained using HP filter, assuming a l=100.

Figure 2. Smoothed Average Annual Labor Productivity Growth (Percent) by Region

"We see the effects of transformative technologies everywhere, except for in the productivity statistics."

Measured productivity has declined by half over the past decade...

real income has stagnated since the late 1990's."

ONE SIMPLE EXPLANATION (SAYS THE ECONOMISTS)

"Increasing productivity takes time, both to develop needed, complementary technologies ...as well as to change practices around the new technologies."

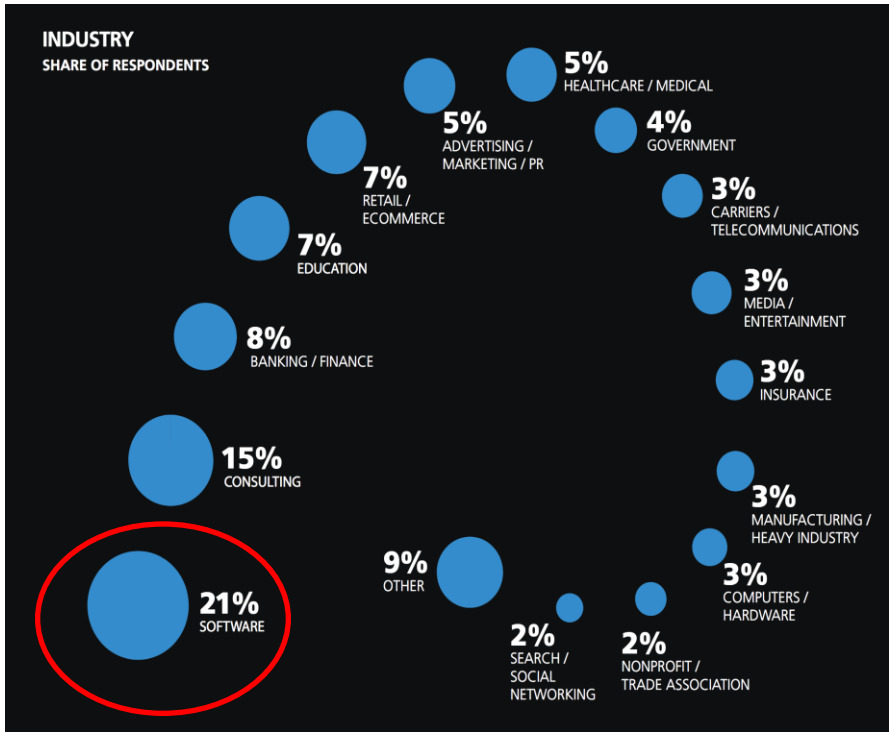
OUR EXEC FRIENDS POSSIBLY AGREE

Considering what inhibits data driven impact, complementary technologies doesn't seem to be the bottleneck. **People seem to be getting in the way!**

Biggest Challenge to Business Adoption	2018	2019
Lack of organizational alignment/agility	25.0%	40.3%
Cultural resistance	32.5%	23.6%
Technology solutions	15.2%	5.0%
Understanding data as an asset	30.0%	13.9%
Executive leadership	7.5%	7.0%

Source: New Vantage Partners "Big Data and AI Executive Survey 2019" http://newvantage.com/wp-content/uploads/2018/12/Big-Data-Executive-Survey-2019-Findings-122718.pdf

FOOD FOR THOUGHT



The field of Data Science has been largely tech driven, and …

Leading tech companies have…

- Agile driven cultures
- Fewer ML/data governance regulations
- Easier measurement opportunities* (massive scale, faster outcomes)

Its not just about being clever. They have structural advantages for doing DS well.

Image Source: https://www.oreilly.com/data/free/files/2017-data-science-salary-survey.pdf

**This is not to suggest that measurement is always easy, but imagine setting up an AB test for a life-insurance underwriting model, where the event you're measuring (death) is over a 30 year term? Or consider the financial and ethical risks of AB testing a credit lending model, with a control group being a randomly chosen population segment.*

BRIDGING THE GAP

Most of the companies in the Fortune 100 succeeded without AI, Big Data and Data Science. They are built on years of legacy management practices (that generally work) and they may not have the muscle memory for managing Data Science. How do we bridge this gap?

DATA SCIENCE AND RISK

THE RISKY BUSINESS OF DATA SCIENCE

Most organizations expect data and AI to **enable productivity and business impact**. Many organizations fail to create such impact because they haven't developed the capabilities to manage risks that are unique to Data Science and AI projects.

A.Hypothesis Risk
Are you answering the right questions with the right assumptions?

B.Signal Risk
Do you have the right data and hypotheses to find the signal if it exists?

C.Execution Risk
Are you able to operationalize your results in a testable way?

HYPOTHESIS RISK (W/ EXAMPLES)

Are you asking the right questions and making the right assumptions? Most businesses have some form of this (almost every product and investment decision is prone to this).

Wrong questions:

- Which AI platforms should we buy to gain competitive advantage?
- Should our data science team test out Multi-Armed Bandits?

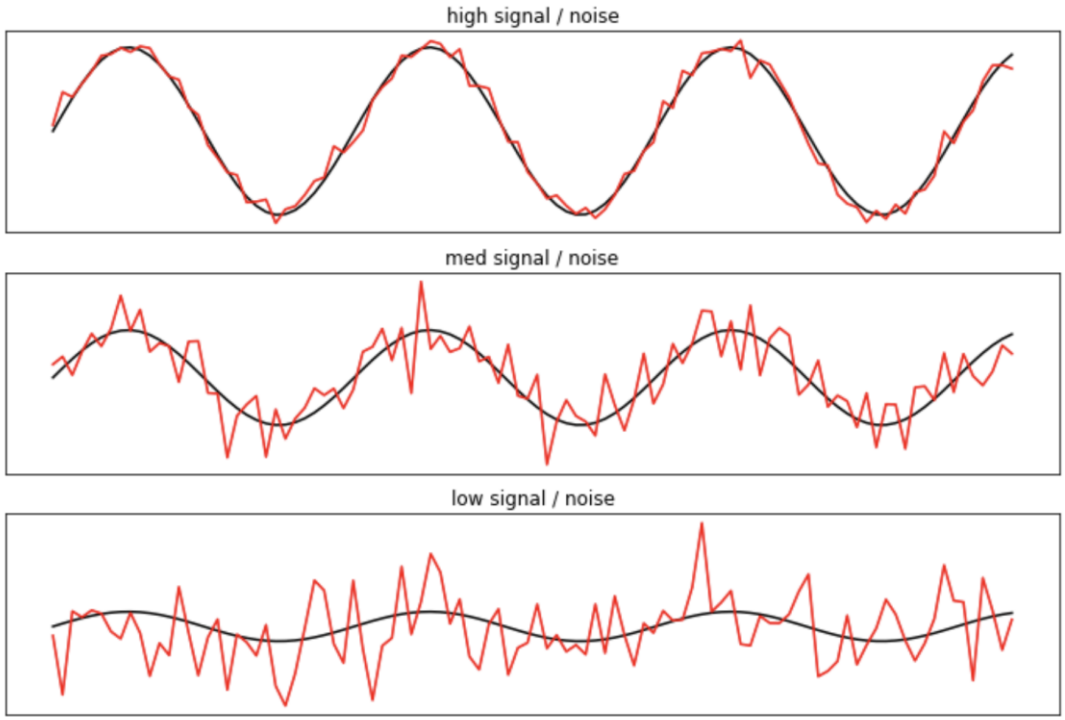
Right questions:

- How do I improve my insurance underwriting process so that we can gain more customers while reducing exposure?
- Can we get better marketing results with more dynamic creative optimization?

**Unique to Data Science*

SIGNAL RISK

If predictive signal exists, do you have the data and analytic capabilities to find it?



Everyone wants to be here

This is your best likely outcome for most problems

Too many analysts find themselves here

EXECUTION RISK

Data does not create value alone. Changing your actions because of data is what creates value, but changing your (company's) actions can be quite difficult.

- Can your model be deployed with existing technologies in a reasonable amount of time?
- Does your model or recommendation violate any existing regulations and ethics rules or technical constraints?
- If your model/recommendation requires change management, who will manage this change to ensure adoption?

**Unique to Data Science*

MY CLAIM …

Signal risk is new risk that is quite unique to data science, and outside of R&D organizations, many managerial frameworks aren't well equipped to hedge it (even Agile!!!!).

A.Hypothesis Risk

Are you answering the right questions with the right assumptions?

B.Signal Risk

Do you have the right data and hypotheses to find the signal if it exists?

C.Execution Risk

Are you able to operationalize your results in a testable way?

EXAMPLE: ADDING SIGNAL RISK TO THE MIX

Your busy product manager colleague is in charge of developing your e-commerce websites search engine. She gives direction to both an engineering team and a data science team. She wants to see what each team can do in 3 months towards the broad goal of building a search engine to drive e-commerce sales

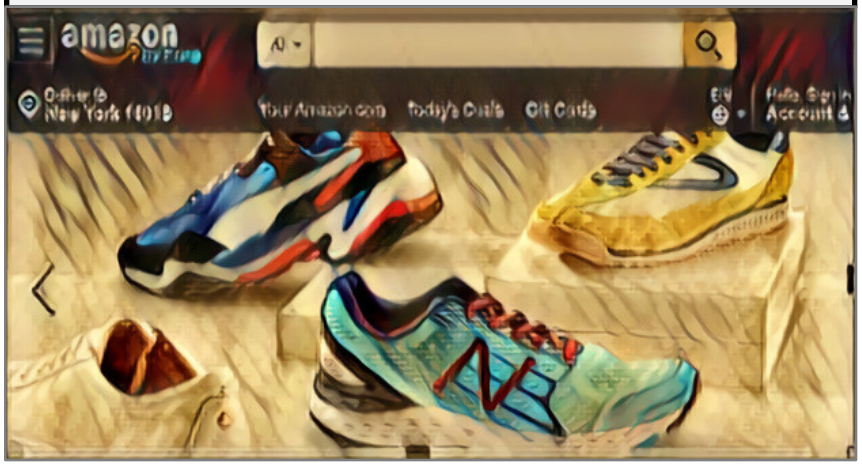
To Software Engineers:
Build a functional web UI based search engine that takes in generic input and outputs a sorted list of relevant products.

To Data Scientists:
Develop a way to rank the results returned from a search engine to maximize visitor purchase conversion rate.

AFTER 3 MONTHS – DEMO TIME

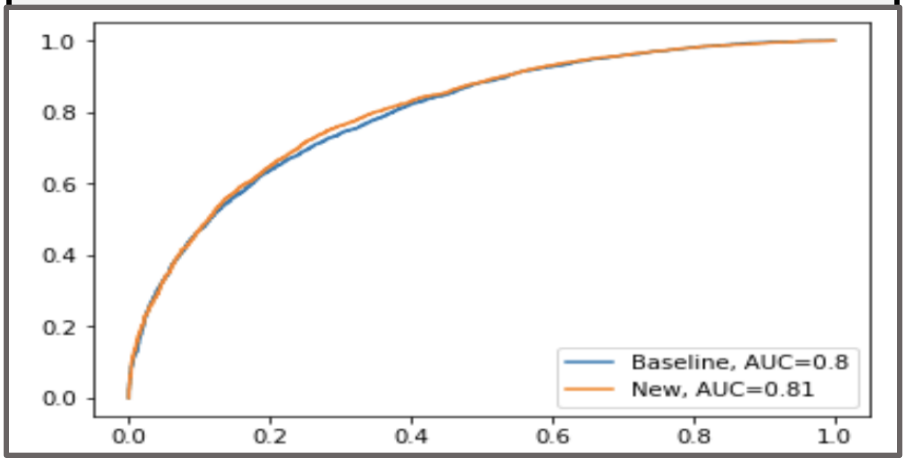
Software Engineers:

- Indexed all products into an ElasticSearch database
- Built a front-end, responsive UI
- Built a search microservice so the UI can ping the ElasticSearch



Data Scientists:

- Consolidated multiple event data streams, and flagged several issues with the data to fix
- Built ETL pipelines to process logs
- Tested multiple algorithms, finding that the best model doesn't beat a heuristic baseline.



AFTER 3 MONTHS – DEMO TIME

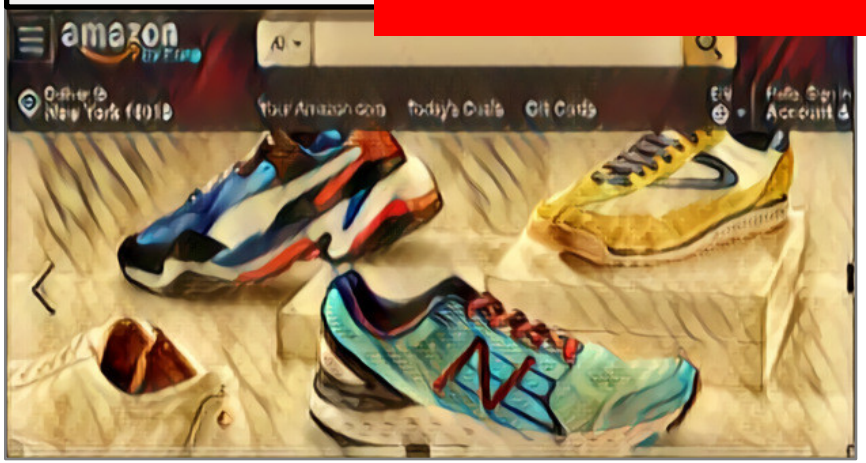
Software Engineers:

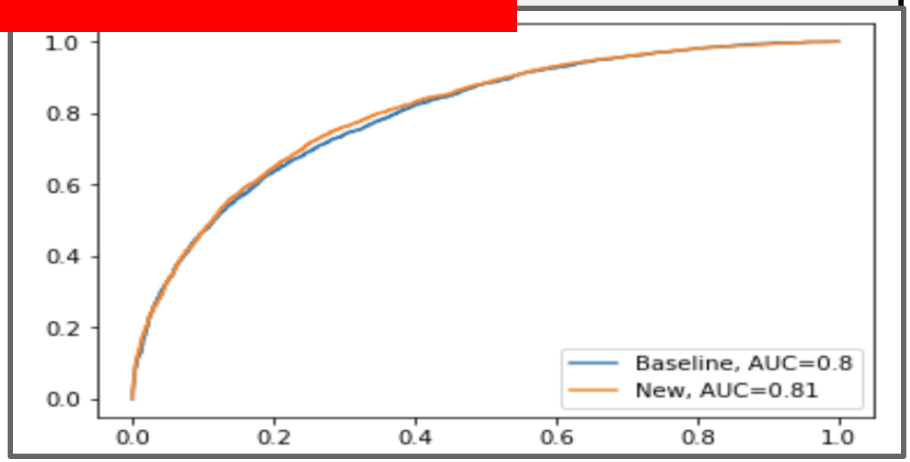
- Indexed all products into an ElasticSearch database
- Built a front-end, responsive UI
- Built a search microservice so the UI can ping the

Data Scientists:

- Consolidated multiple event data streams, and flagged several issues with the data to fix
- Built ETL pipelines to process logs
- rithms, finding
- doesn't beat a

Who is the likely hero here?





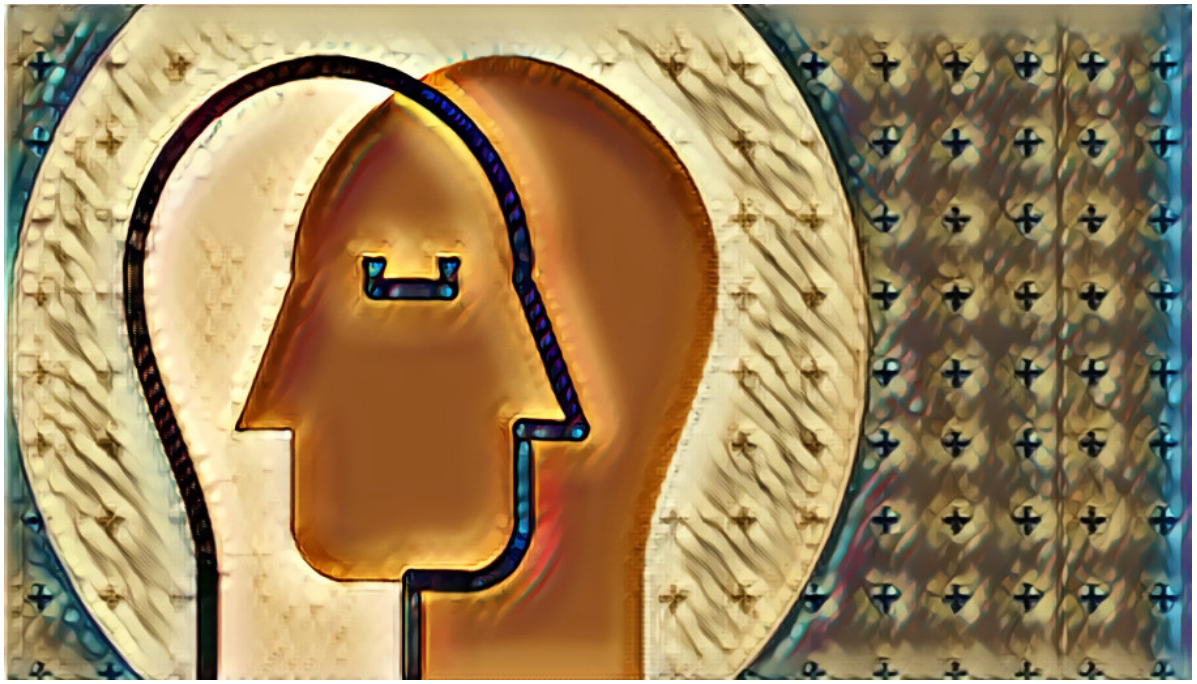
BUILDING BETTER PROCESSES

TOWARDS MORE IMPACTFUL DATA SCIENCE

Creating impact likely won't come from starting an AI Strategy, managed under some Center of Excellence. Organizations need the right processes and technology, and these should be designed and chosen specifically to target the associated risks with a Data Science project.

START WITH A LITTLE EMPATHY

Data Science as a discipline is still new and heavily unstandardized. Doing it well requires deep technical knowledge in technology and math, and most of all, common management frameworks fail to capture the unique dynamics of a Data Science project.



Empathy =

- Aligning DS goals with business goals
- Implementing more accountable DS processes
- Adopting transparent and explainable AI/ML

Much of this is achievable with processes that respect intrinsic risks of Data Science

BE SINGLE MINDED

Data Science teams shouldn't operate in silos. They should interface with business stakeholders as much as possible, and aim for the following:

- Create shared goals/KPIs, accountability
- Understand the 'day-in-the-life' of line workers
- Learn all operational, legal and technical constraints

Avoiding business silos ensures the right questions get asked, solutions are feasible and builds understanding of signal risk.

TAKE A PORTFOLIO APPROACH

Considering the risks on any single data science project, one can hedge by taking on multiple projects at the same time. There are a few tricks to make this easier:

- Shelve unlikely prospects as soon as possible
- Break problem down into smaller parts
- Don't get stuck in sunk cost fallacy
- Write "prototype," not "production" code

Team A

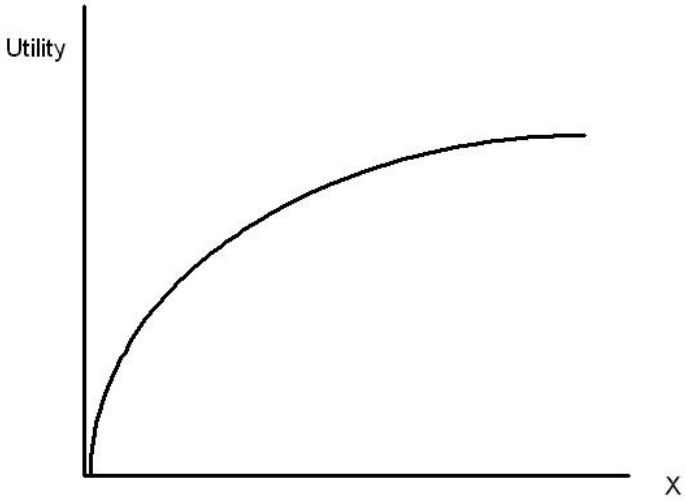
Team B

Team C

Team D

INCREASE YOUR TEAM'S AGILITY

Agile data science practices enable analysts to learn how to best solve problems that are going well, and test in realistic scenarios. On the flip side, agile DS management enables teams to fail fast when the questions are wrong or there is no signal to the problem.

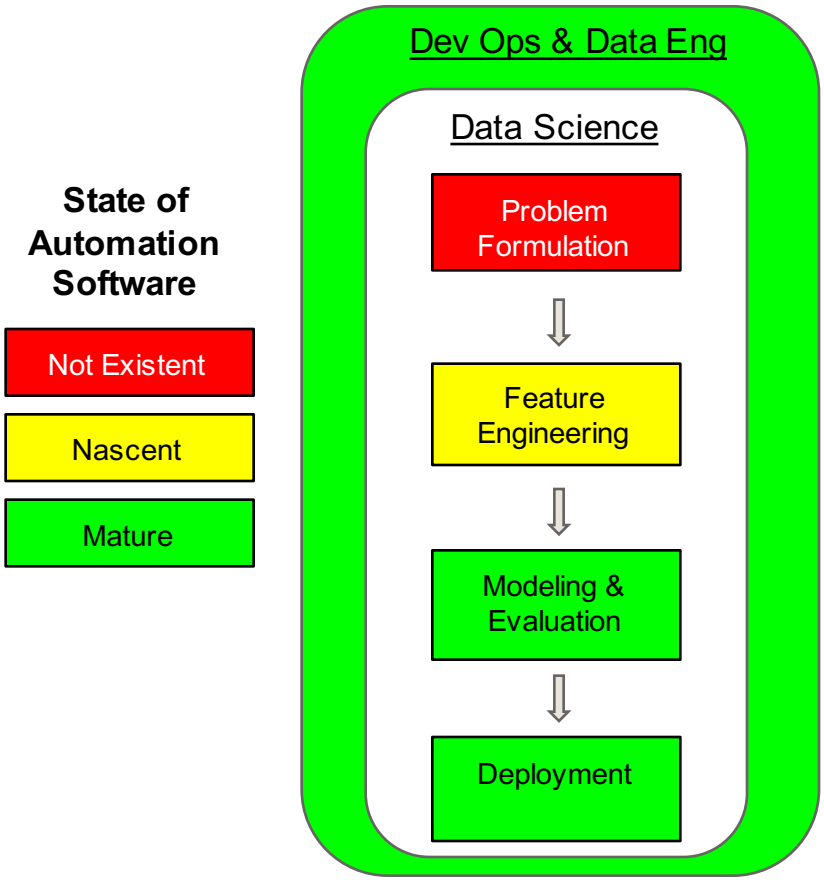


Model Complexity and
Effort Building/Implementing

A good but simple model is always better than no model!

Bias yourself towards deployment when competing against time.

AUTOMATE EVERYTHING (EVEN YOURSELF)!



Almost every step of the DS process, as well as supporting data infrastructure and working environment, are now well supported with automation software.

i. Devops (Databricks, Kubernetes, Docker, AWS, Azure, Google Cloud)
ii. ETL (Databricks, Airflow)
iii. Feature Engineering (SparkBeyond, FeatureTools)
iv. Model Selection (SparkBeyond, DataRobot, H2O.AI)
v. Model Deployment (SparkBeyond, DataRobot, AWS SageMaker)

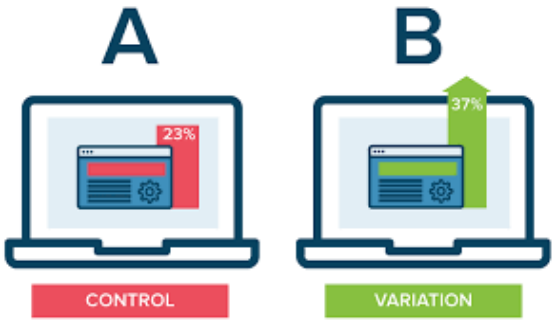
There is no better way to move faster than to use good tools! And don't fall for "not built here syndrome"

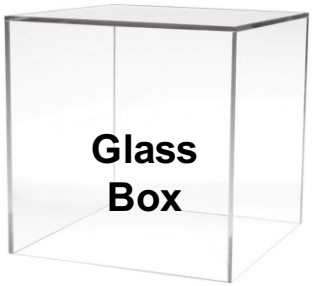
BUILD TRUST IN AI SYSTEMS

Data Scientists ultimately have to answer to someone, and that someone usually expects business impact. The key to long term value creation through Data Science is to build trust within the organization that the $$$ spent on data and scientists pays off.

Where possible, always AB test an action driven by data, and make sure there is a clear KPI that this action looks to improve.

AB Testing is often infeasible, and additionally, ML systems sometimes need to be interpretable and transparent. In these cases, a "Glass Box" approach is the best way to build trust in the model.



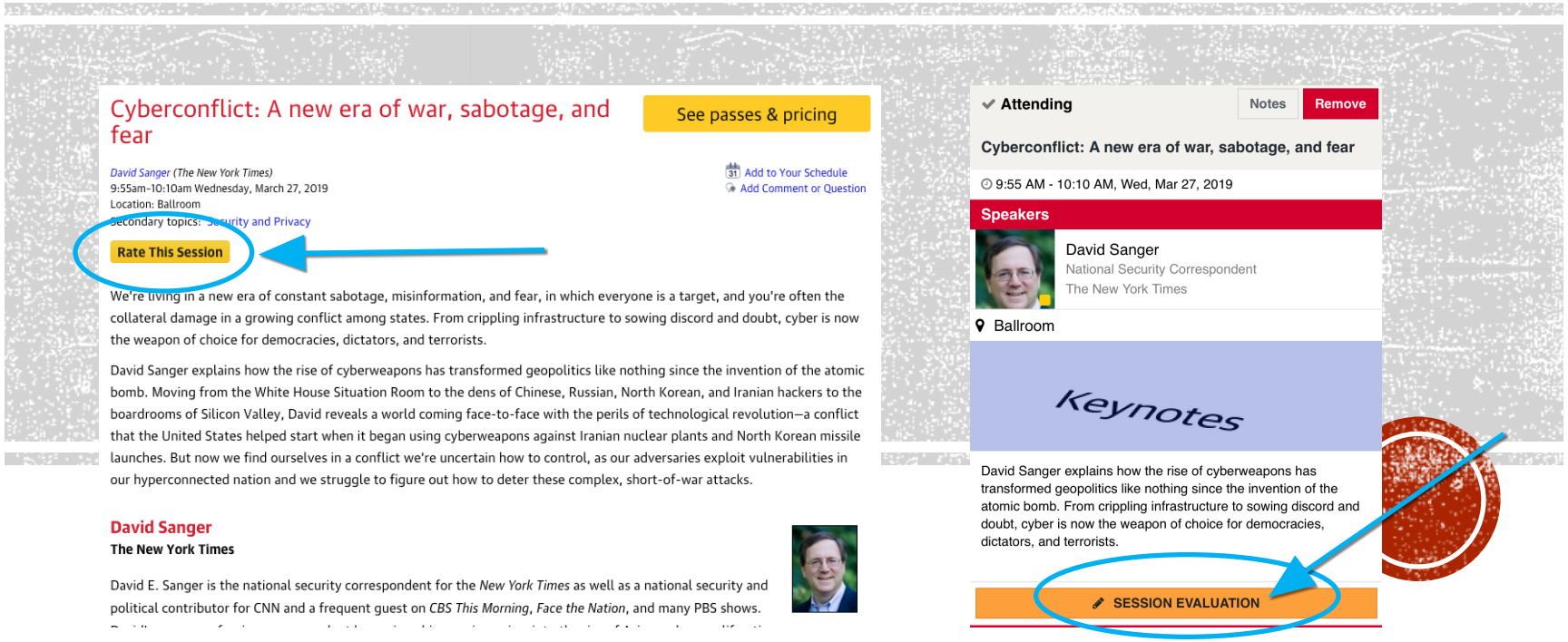


Slide inspiration :https://www.locallyoptimistic.com/post/the-blacker-the-box/

DS RISK MITIGATION: (VERBOSE) PLAYBOOK

1. Create shared goals/KPIs between data scientists and business owners (ensures DS team is working on the right projects and has accountability to the business)
2. Maintain a research portfolio (think multi-armed bandit)
3. Be agile (build and test an integrated system before developing the "best" model)
    a. Reduce research lifecycle (fail fast or succeed quickly)
    b. Invest in automation technology to move past common bottlenecks
4. Build trust in AI/ML systems (so that business owners/execs sign off)
    a. Where possible, test on KPIs
    b. Otherwise, invest in transparent & explainable decision making

RATE TODAY 'S SESSION



Session page on conference website

O'Reilly Events App