DATA to AI Lab

LIDS, MIT

Kalyan Veeramachaneni

Over the past 7-8 years

Data Scientist: The Sexiest Job of the 21st Century

by Thomas H. Davenport and D.J. Patil

FROM THE OCTOBER 2012 ISSUE

Help Wanted: Black Belts in Data

Starting salaries for data scientists have gone north of $200,000

by **Rodrigo Orihuela** and **Dina Bass**
June 4, 2015, 1:07 PM EDT
Updated on June 4, 2015, 2:00 PM EDT
From **BloombergBusinessweek** | Subscribe | Reprints

Artificial intelligence (AI)

2016: the year AI came of age

Google and Amazon brought AI into the home and DeepMind built a computer that could outsmart humans at Go. Will 2017 hold similar advancements?

Connection between AI and Data Science?

- All applications of AI that you see have been developed by
  - Collecting data
  - Learning models from them
  - Using those models to act
  - Take for example – alphago
    - “The system's neural networks were initially bootstrapped from human gameplay expertise. AlphaGo was initially trained to mimic human play by attempting to match the moves of expert players from recorded historical games, using a database of around 30 million moves”
- Predictive analytics

And The Winner Is: Big Data Oscar Picks

This year’s Best Picture will be 12 Years A Slave, according to Academy Award prognosticators at Farsite. Will big data get it right?

Academic Research

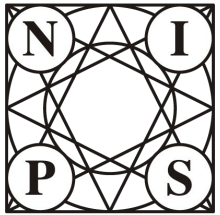
ICML @ Sydney

Thirty-fourth International Conference on Machine Learning

Since **1980**

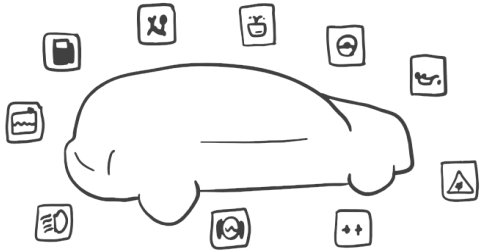
Since **1995**

Conference on Neural Information Processing Systems



Since **1986**

Numerous Data Science Problems



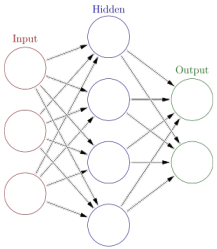
Tremendous increase in rate at which we are encountering data science problems.
The challenge is not to solve just one problem, but to overcome the bottlenecks that prevent us from solving many!

Build AI products faster

8M Vehicle Loans

Age, Original Balance, Interest Rate, Loan Remaining Months, Credit Score

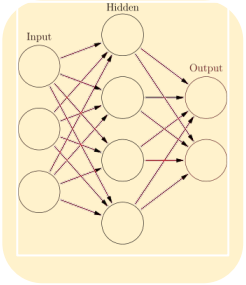
Age	Score	...	Default
32	678		y
21	786		n
67	776		n

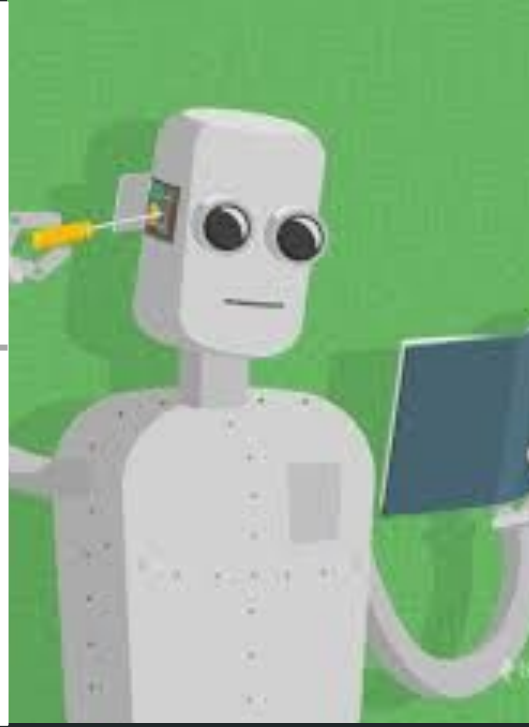


Training

Usage

Age	Score	....	Default
25	524	...	?



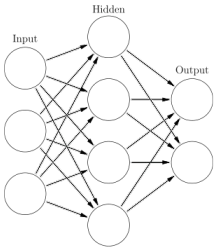


Building AI products

8M Vehicle Loans

Age, Original Balance, Interest Rate, Loan Remaining Months, Credit Score

Age	Score	...	Default
32	678		y
21	786		n
67	776		n

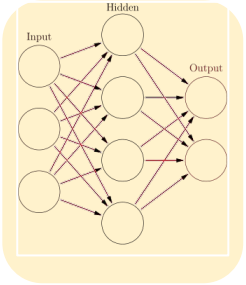


Training

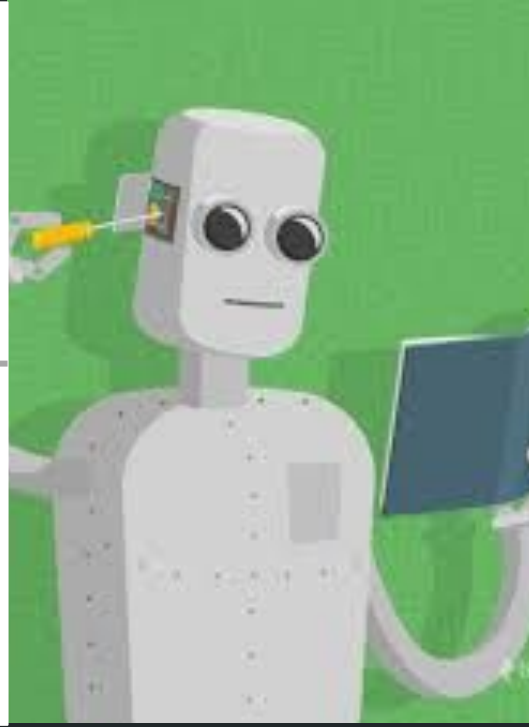
Usage

Age	Score	....	Default
25	524	...	?

New user



No

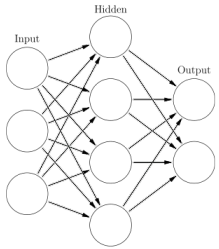


Building AI products

8M Vehicle Loans

Age, Original Balance, Interest Rate, Loan Remaining Months, Credit Score

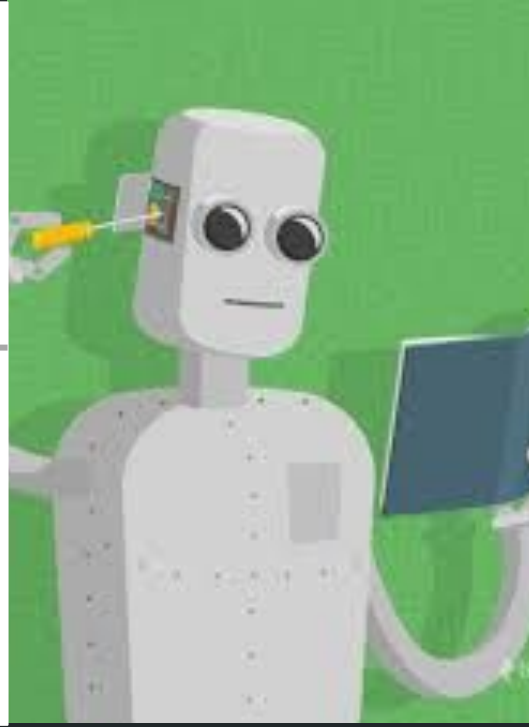
Age	Score	...	Default
32	678		y
21	786		n
67	776		n



Training

But we can train many more models for different outcomes from the same data ..

When is a user most likely to refinance ?
When is a user most likely to buy next car?
And so on…..



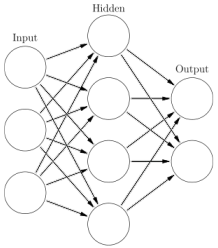
Building AI products

To unlock the potential of machine learning

8M Vehicle Loans

Age, Original Balance, Interest Rate, Loan Remaining Months, Credit Score

Age	Score	...	Default
32	678		y
21	786		n
67	776		n



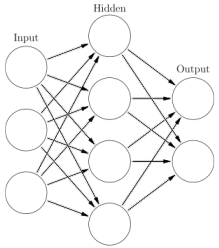
Automatically choose and learn a model

To unlock the potential of machine learning

8M Vehicle Loans

Age, Original Balance, Interest Rate, Loan Remaining Months, Credit Score

Age	Score	...	Default
32	678		y
21	786		n
67	776		n

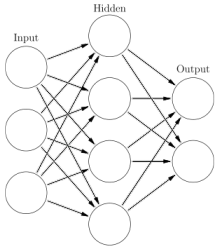


Automatically form patterns from historical data

- Number of times customer was delayed in payments
- Rate of change of user's salary profile
- Rate of change of user's credit score

To unlock the potential of machine learning

Age	Score	...	Default
32	678		y
21	786		n
67	776		n



Automatically formulate questions

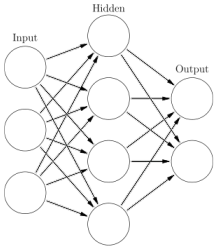
Change from loan default prediction to predict refinancing

To unlock the potential of machine learning

8M Vehicle Loans

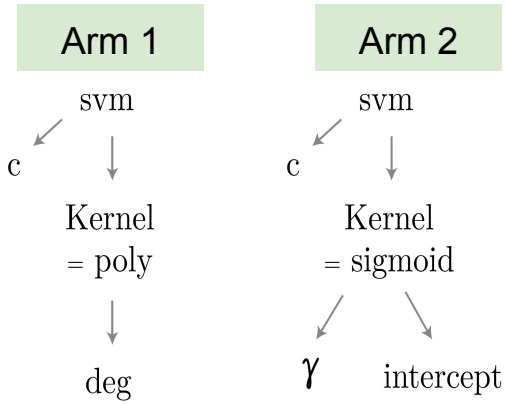
Age, Original Balance, Interest Rate, Loan Remaining Months, Credit Score

Age	Score	...	Default
32	678		y
21	786		n
67	776		n



Automatically choose and learn a model

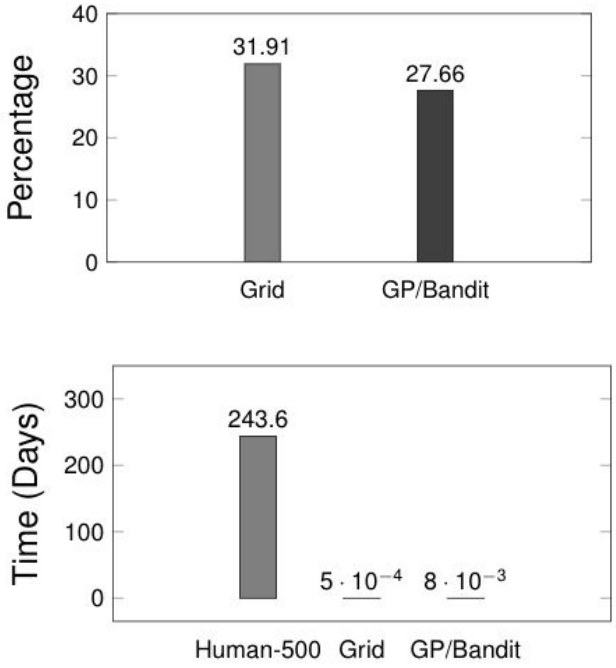
ATM – Automatically choose a model



	UCBI				Random			
	10		100		10		100	
Dataset	mean	std	mean	std	mean	std	mean	std
cell	0.869	$4 \cdot 10^{-3}$	0.875	$3 \cdot 10^{-3}$	0.854	$1.884 \cdot 10^{-2}$	0.874	$4.741 \cdot 10^{-3}$
mooc_nlp5	0.620	0.000	0.634	$1 \cdot 10^{-2}$	0.576	$4.391 \cdot 10^{-2}$	0.636	$9.382 \cdot 10^{-3}$
mooc_dropout	0.916	$6 \cdot 10^{-3}$	0.924	$2 \cdot 10^{-3}$	0.902	$2.838 \cdot 10^{-2}$	0.924	$1.476 \cdot 10^{-3}$
diabetes	0.773	$1.3 \cdot 10^{-2}$	0.787	$8 \cdot 10^{-3}$	0.767	$1.335 \cdot 10^{-2}$	0.786	$6.556 \cdot 10^{-3}$
ionosphere	0.934	$1.18 \cdot 10^{-2}$	0.949	$6 \cdot 10^{-3}$	0.916	$2.331 \cdot 10^{-2}$	0.945	$9.002 \cdot 10^{-3}$
letters	0.947	$1.72 \cdot 10^{-3}$	0.961	$6 \cdot 10^{-3}$	0.940	$2.467 \cdot 10^{-2}$	0.964	$3.563 \cdot 10^{-3}$

Pick an arm using Multi Armed bandit. Tune the hyperparameters using Gaussian Processes. Circa 2014.

ATM - Open source release and comparing to humans



Tested on 420 publicly available datasets
3 million models trained and counting
Compared against human baselines on OpenML
Ready to use!

ATM: A distributed, collaborative, scalable system for automated machine learning , Proceedings of IEEE Big data conference, 2017

ATM open source library - 2017

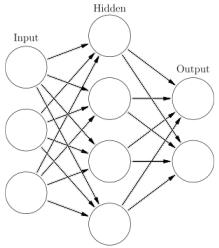
http://bit.ly/MIT-HDI

To unlock the potential of machine learning

8M Vehicle Loans

Age, Original Balance, Interest Rate, Loan Remaining Months, Credit Score

Age	Score	...	Default
32	678		y
21	786		n
67	776		n



Automatically form patterns from historical data

- Number of times customer was delayed in payments
- Rate of change of user's salary profile
- Rate of change of user's credit score

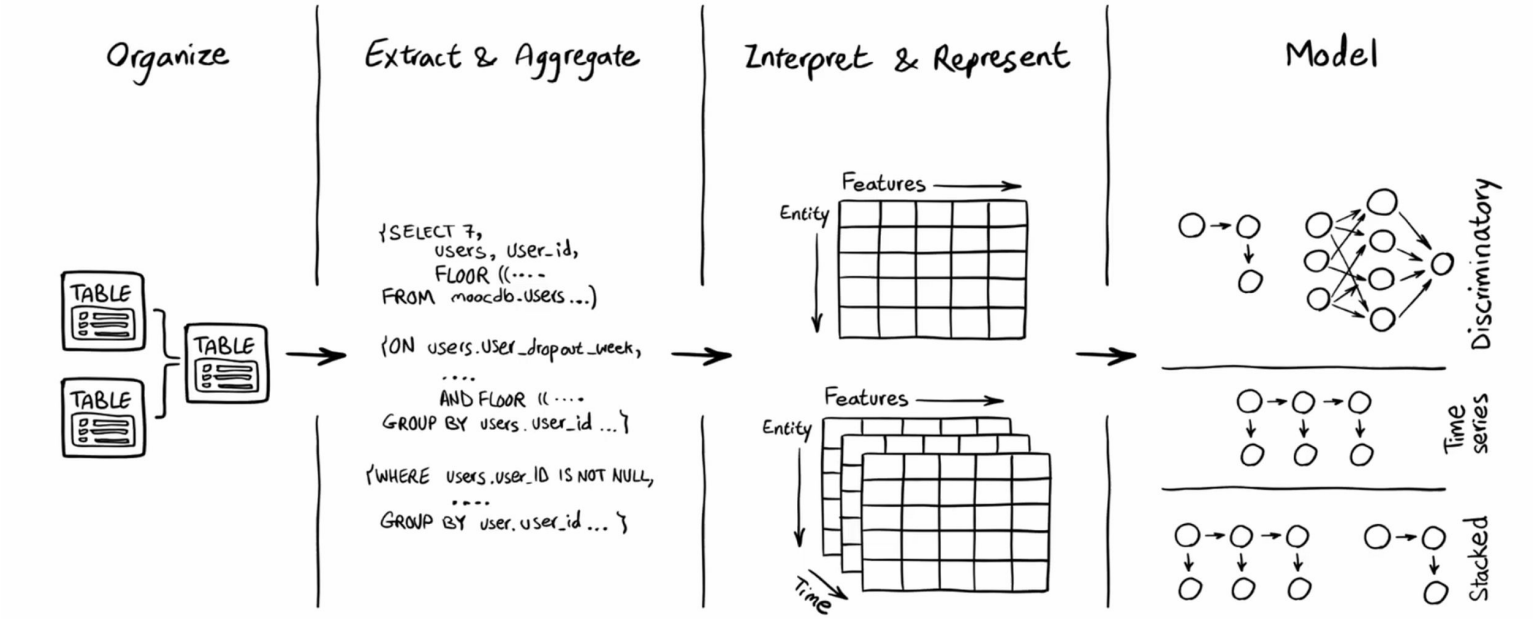
The Quintessential Matrix

FEATURIZED DATA

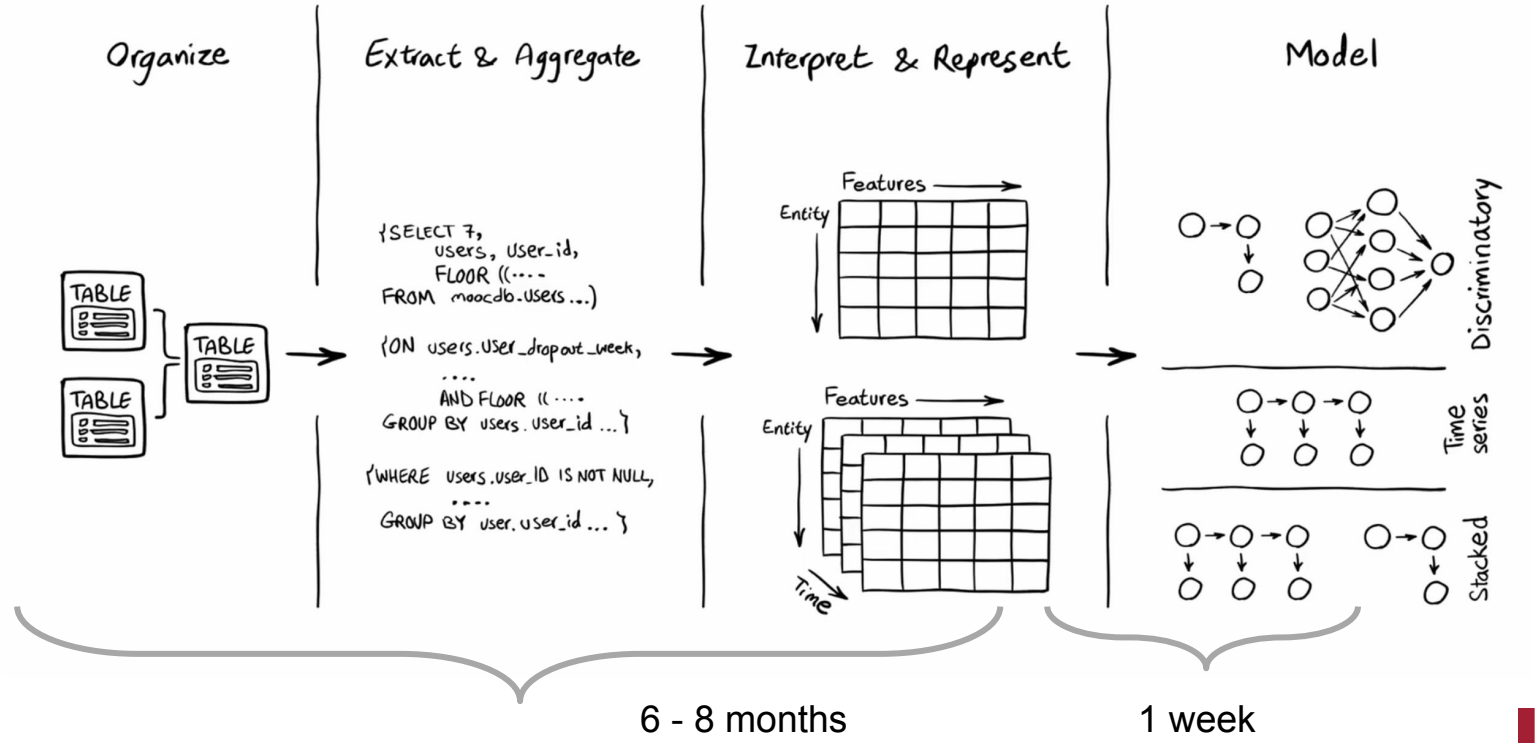
Id	X_2	X_3	X_4	...	...	...	X_n	Y

First column is id – project(accenture), car(Jaguar) – training example.
Each column is a feature. Last column is the label.

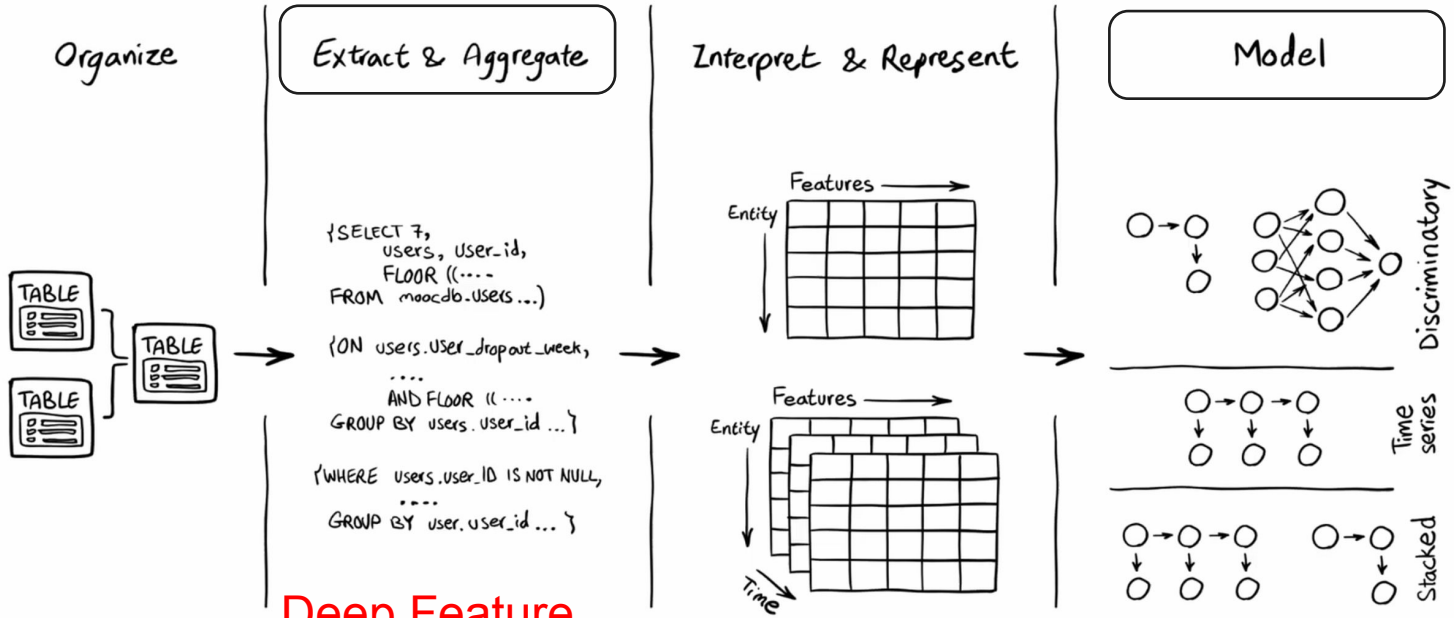
So we started the process of "Data Science"



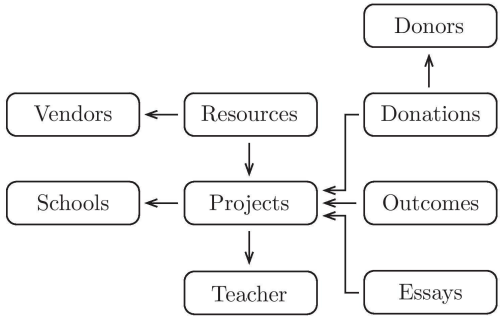
So we started the process of "Data Science"



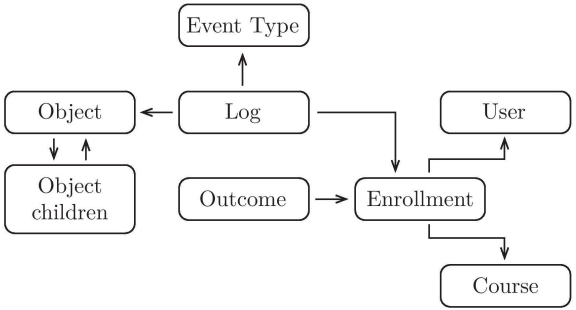
The data science process



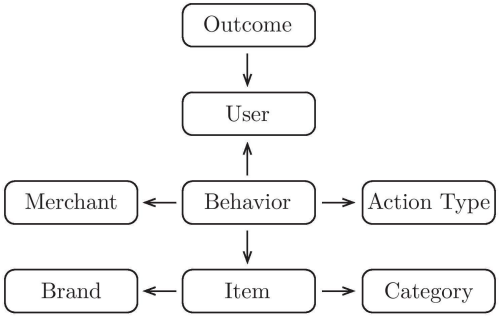
Testing our automated feature engineering on KAGGLE Competitions



Project Excitement



Dropout Prediction



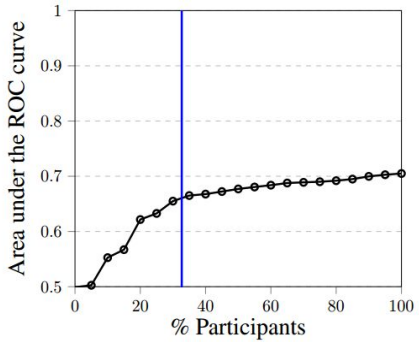
Repeat Buyer

Automation – Circa. 2015

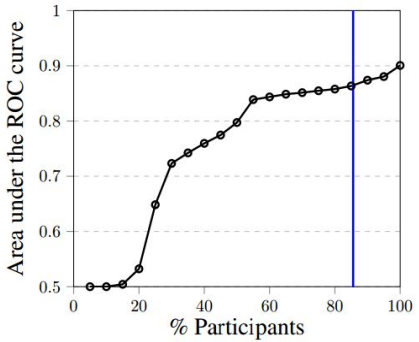
GIVEN LEARNING SEGMENTS

We can generate features, learn models and evaluate

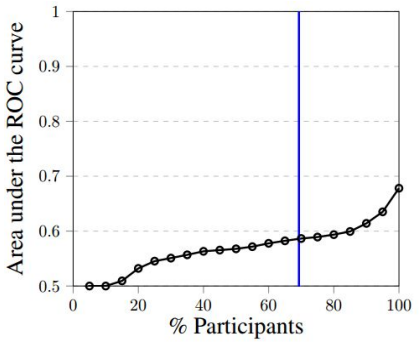
Repeat Buyer



Dropout Prediction



Project Excitement



Lines show the standing of the Deep Feature Synthesis in the competition as of May 18th , 2015

Deep Feature Synthesis: Towards Automating Data Science Endeavors, Proceedings of IEEE/ACM Data Science and Advance Analytics Conference, 2015

Automation – Circa. 2015

GIVEN LEARNING SEGMENTS

We can generate features, learn models and evaluate

Tested against
1,000
data scientists

On average
92%
of top score

Over
1,200
days saved

Deep Feature Synthesis: Towards Automating Data Science Endeavors, Proceedings of IEEE/ACM Data Science and Advance Analytics Conference, 2015

Automation – Circa. 2015

The Washington Post
Democracy Dies in Darkness

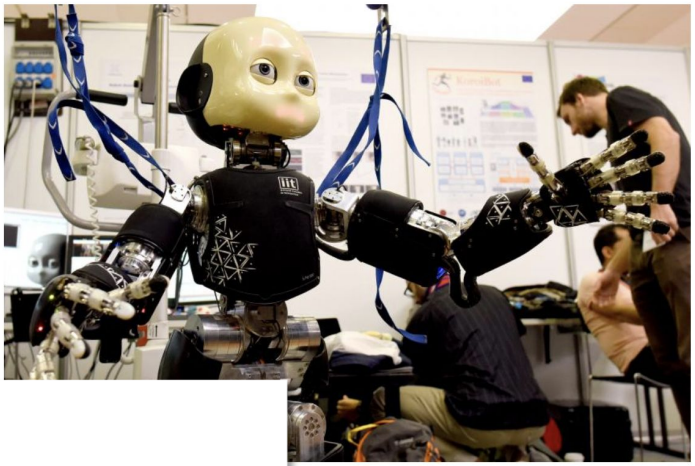
Speaking of Science

New MIT algorithm rubs shoulders with human intuition in big data analysis

By Rachel Feltman October 19, 2015

AN ALGORITHM MAY BE BETTER THAN HUMANS AT BREAKING DOWN BIG DATA

BY SEUNG LEE ON 10/19/15 AT 6:25 PM



Automating big-data analysis

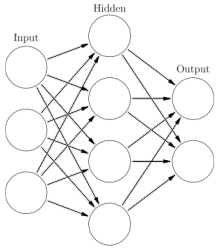
System that replaces human intuition with algorithms outperforms 615 of 906 human teams.

Featuretools open source library - 2017

www.featuretools.com

To unlock the potential of machine learning

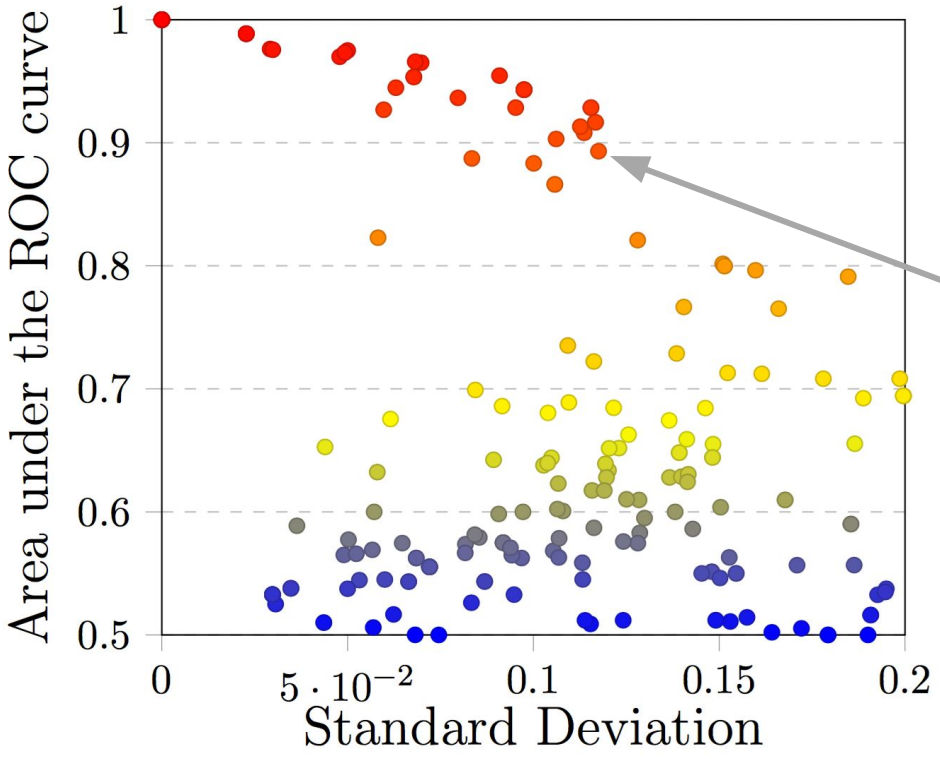
Age	Score	...	Default
32	678		y
21	786		n
67	776		n



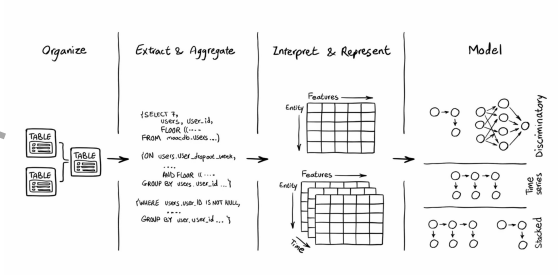
Automatically formulate questions

Change from loan default prediction to predict refinancing

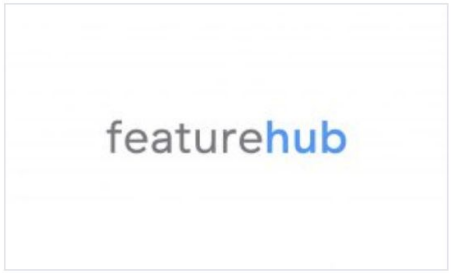
TRANE



Predict whether the mean sales volume will exceed $5000 in a 2-week window



MIT - The Human Data Interaction Project



Industrial scale problems

Industrial scale problems

Predict destination
7000 fields
> 1 years data

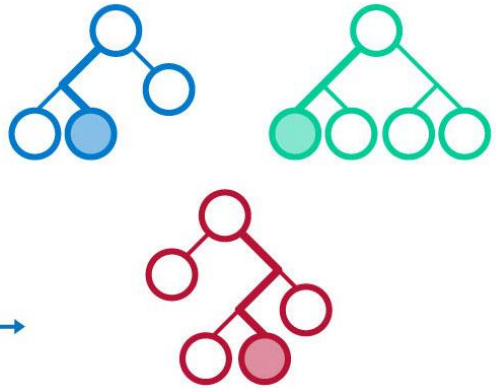
Predicting software release delays
512 fields, 5 tables
>5 years data

Improved machine learning accuracy by 2x

```
card:          "...",
amount:        76.40,
date:          9/26/16
input_method:  "swipe",
country:       "Spain"
    •
    •
    •
merchant code: 5311
```

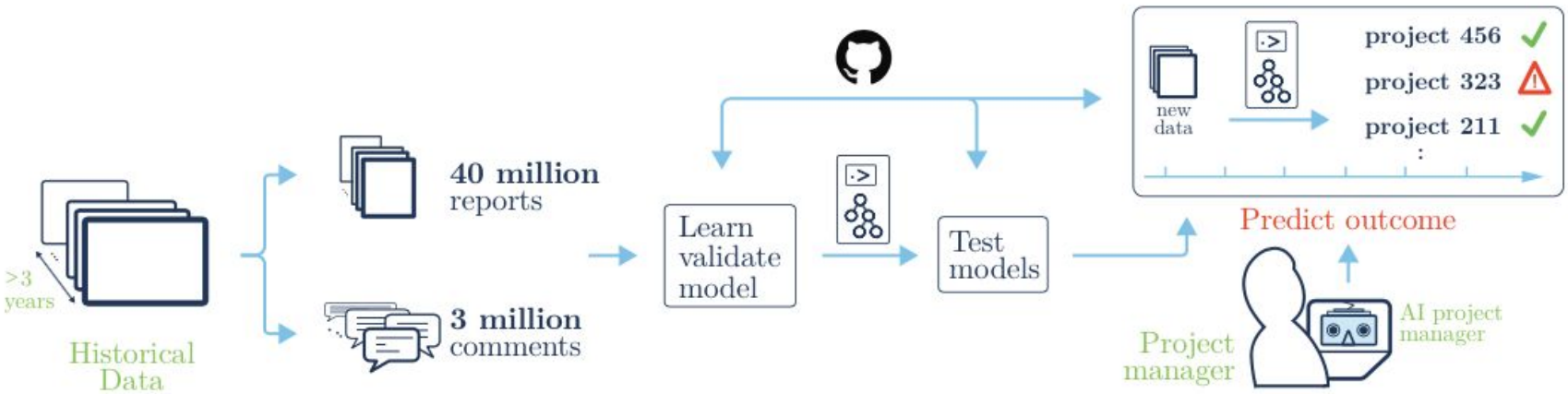
Ingest 900m transactions and identify ~100,000 examples of fraudulent purchases for training

Enrich transactional data with ~100 historical features automatically using Deep Feature Synthesis

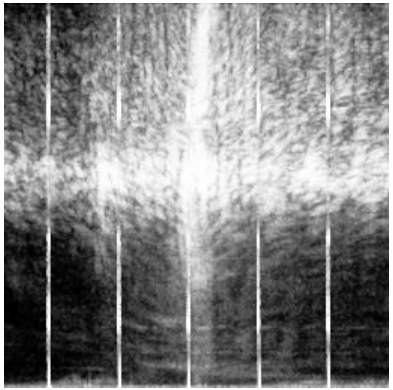


Train random forest classifier and optimize for real time deployment

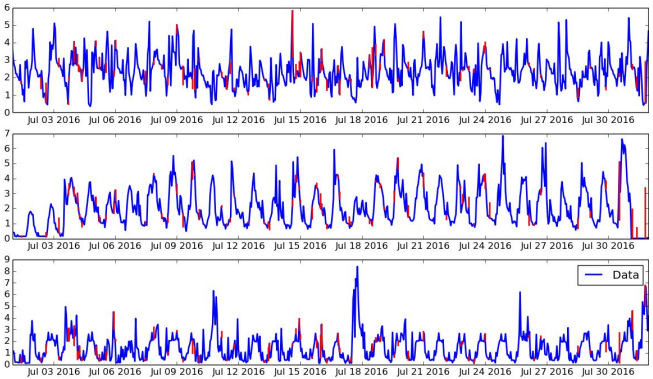
Accenture


>3 years
Historical Data
40 million reports
3 million comments
Learn validate model
Test models
new data
project 456
project 323
project 211
Predict outcome
Project manager
AI project manager

What about several other types of data?



Images



Signals

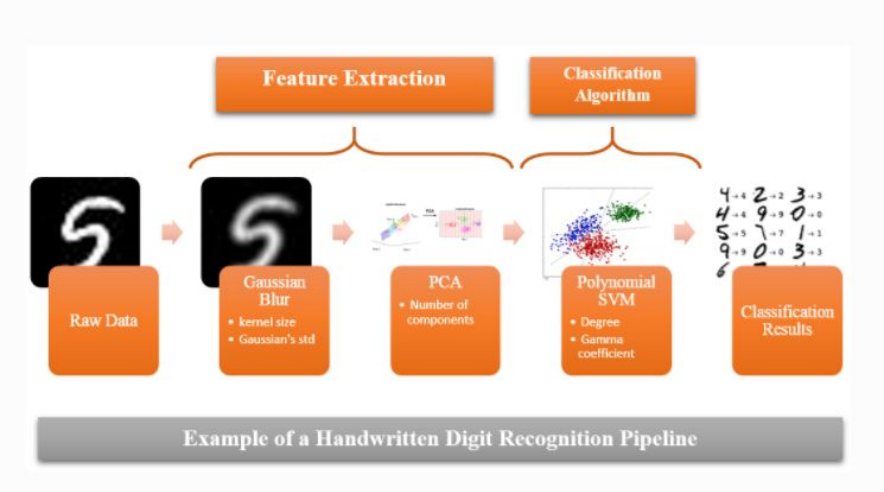
What does the future look like?

2016

Deep Mining - Much broader system

Deep Mining project aims to construct a end to end Machine Learning system automatically and for all data types.

Here is an example: The [handwritten digit recognition problem]



Our automated system – Feb 2018

MIT+FeatureLabs

A	B	C	D	E
TA2 system	**Coverage**	**% of datasets the system did best?**	**How many datasets the system was best?**	
aika (Berkeley)	50%	11%	2 out of 18	
brown	39%	11%	2 out of 18	
columbiau_uchicago	0%	0%	0 out of 18	
cra_eve	0%	0%	0 out of 18	
featurelabs_mit_btb (MIT)	78%	44%	8 out of 18	
isi (USC)	33%	0%	0 out of 18	
nyu	39%	6%	1 out of 18	
qntfy	33%	11%	2 out of 18	
sri_tpot	61%	11%	2 out of 18	
texasam_tamu	94%	22%	4 out of 18	
Tests were out of 18 "seed" datasets				
Coverage --> how many datasets the system ran - trained a model, outputted a model and NIST was able to test and score the model - without an error				
Best performance --> how many datasets did the system do better than everyone else				
		#1		
		#2		

MIT - The Human Data Interaction Project

http://bit.ly/MIT-HDI

Goal: 10 new industrial scale applications!

To receive updates, apply to be one of our partner send email to:

dailabmit@gmail.com

or

kalyanv@mit.edu