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?
ICML @ Sydney
Thirty-fourth International Conference on Machine Learning
Since 1980

KDD2017
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
Building AI products

### Training

<table>
<thead>
<tr>
<th>Age</th>
<th>Score</th>
<th>Default</th>
</tr>
</thead>
<tbody>
<tr>
<td>32</td>
<td>678</td>
<td>y</td>
</tr>
<tr>
<td>21</td>
<td>786</td>
<td>n</td>
</tr>
<tr>
<td>67</td>
<td>776</td>
<td>n</td>
</tr>
</tbody>
</table>

### Usage

<table>
<thead>
<tr>
<th>Age</th>
<th>Score</th>
<th>Default</th>
</tr>
</thead>
<tbody>
<tr>
<td>25</td>
<td>524</td>
<td>?</td>
</tr>
</tbody>
</table>
Building AI products
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.....
To unlock the potential of machine learning

Automatically choose and learn a model
To unlock the potential of machine learning

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

- Automatically formulate questions
- Change from loan default prediction to predict refinancing
To unlock the potential of machine learning

Automatically choose and learn a model
ATM – Automatically choose a model


<table>
<thead>
<tr>
<th>Dataset</th>
<th>mean</th>
<th>10^-3 std</th>
<th>mean</th>
<th>10^-3 std</th>
<th>mean</th>
<th>10^-2 std</th>
<th>mean</th>
<th>10^-2 std</th>
<th>Random mean</th>
<th>10^-2 std</th>
</tr>
</thead>
<tbody>
<tr>
<td>cell</td>
<td>0.869</td>
<td>4</td>
<td>0.875</td>
<td>3</td>
<td>0.854</td>
<td>1.884</td>
<td>0.874</td>
<td>4.741</td>
<td></td>
<td></td>
</tr>
<tr>
<td>mooc_nlp5</td>
<td>0.620</td>
<td>0</td>
<td>0.634</td>
<td>1</td>
<td>0.576</td>
<td>4.391</td>
<td>0.636</td>
<td>9.382</td>
<td></td>
<td></td>
</tr>
<tr>
<td>mooc_dropout</td>
<td>0.916</td>
<td>6</td>
<td>0.924</td>
<td>2</td>
<td>0.902</td>
<td>2.838</td>
<td>0.924</td>
<td>1.476</td>
<td></td>
<td></td>
</tr>
<tr>
<td>diabetes</td>
<td>0.773</td>
<td>1</td>
<td>0.787</td>
<td>8</td>
<td>0.767</td>
<td>1.335</td>
<td>0.786</td>
<td>6.556</td>
<td></td>
<td></td>
</tr>
<tr>
<td>ionosphere</td>
<td>0.934</td>
<td>1</td>
<td>0.949</td>
<td>6</td>
<td>0.916</td>
<td>2.331</td>
<td>0.945</td>
<td>9.002</td>
<td></td>
<td></td>
</tr>
<tr>
<td>letters</td>
<td>0.947</td>
<td>1</td>
<td>0.961</td>
<td>6</td>
<td>0.940</td>
<td>2.467</td>
<td>0.964</td>
<td>3.563</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

A distributed, multi-model, self-learning platform for machine learning
US Patent Application 14/598,628
Filed January 16, 2015
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!

**Graph 1:**
- Percentage
- Grid: 31.91
- GP/Bandit: 27.66

**Graph 2:**
- Time (Days)
- Human-500: $5 \times 10^{-4}$
- Grid: $8 \times 10^{-3}$
- GP/Bandit: $243.6$

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

- 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

<table>
<thead>
<tr>
<th>Age</th>
<th>Score</th>
<th>Default</th>
</tr>
</thead>
<tbody>
<tr>
<td>32</td>
<td>678</td>
<td>y</td>
</tr>
<tr>
<td>21</td>
<td>786</td>
<td>n</td>
</tr>
<tr>
<td>67</td>
<td>776</td>
<td>n</td>
</tr>
</tbody>
</table>
### The Quintessential Matrix

#### FEATURIZED DATA

<table>
<thead>
<tr>
<th>Id</th>
<th>$X_2$</th>
<th>$X_3$</th>
<th>$X_4$</th>
<th>...</th>
<th>...</th>
<th>...</th>
<th>$X_n$</th>
<th>Y</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

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”

Organize

Extract & Aggregate

{SELECT ?

users, user_id,

FROM moodle:users...}

{ON users.user_dropout_week,

....

AND floor == ...

GROUP BY users.user_id ... ?}

{WHERE users.user_id IS NOT NULL,

....

GROUP BY user.user_id ... ?}

Interpret & Represent

Features

Entity

Time

Model

Discriminatory

Time series

Slacked

6 - 8 months

1 week
The data science process

Organize

Extract & Aggregate

{SELECT ?,
users, user_id,
FLOOR (????)
FROM moocab.users ...}

{ON users.user_dropout_week,
????
AND FLOOR ???
GROUP BY users.user_id ...}

{WHERE users.user_id IS NOT NULL,
????
GROUP BY user.user_id ...}

Interpret & Represent

Model

Deep Feature Synthesis

ATM

Automated
Testing our automated feature engineering on KAGGLE Competitions

Project Excitement

Dropout Prediction

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

Automating big-data analysis
System that replaces human intuition with algorithms outperforms 615 of 906 human teams.
To unlock the potential of machine learning

Automatically formulate questions

Change from loan default prediction to predict refinancing
Predict whether the mean sales volume will exceed $5000 in a 2-week window.
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

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

Predict outcome

Project manager

AI project manager

new data

project 456 ✓
project 323
project 211 ✓
What about several other types of data?

- Images
- Signals

Data analysis techniques are often used to explore and interpret data from various sources.
What does the future look like?

2016

DARPA to Build “Virtual Data Scientist” Assistants Through A.I.

A.I. will make up for the lack of data scientists.
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
<table>
<thead>
<tr>
<th>A</th>
<th>B</th>
<th>C</th>
<th>D</th>
<th>E</th>
</tr>
</thead>
<tbody>
<tr>
<td>TA2 system</td>
<td>Coverage</td>
<td>% of datasets the system did best?</td>
<td>How many datasets the system was best?</td>
<td></td>
</tr>
<tr>
<td>nika (Berkeley)</td>
<td>50%</td>
<td>11%</td>
<td>2 out of 18</td>
<td></td>
</tr>
<tr>
<td>brown</td>
<td>39%</td>
<td>11%</td>
<td>2 out of 18</td>
<td></td>
</tr>
<tr>
<td>columbiau_uchicago</td>
<td>0%</td>
<td>0%</td>
<td>0 out of 18</td>
<td></td>
</tr>
<tr>
<td>cre_eve</td>
<td>0%</td>
<td>0%</td>
<td>0 out of 18</td>
<td></td>
</tr>
<tr>
<td>featurelabes_mit Bbw (MIT)</td>
<td>78%</td>
<td>44%</td>
<td>8 out of 18</td>
<td></td>
</tr>
<tr>
<td>isi (USC)</td>
<td>33%</td>
<td>0%</td>
<td>0 out of 18</td>
<td></td>
</tr>
<tr>
<td>nyu</td>
<td>39%</td>
<td>6%</td>
<td>1 out of 18</td>
<td></td>
</tr>
<tr>
<td>qntfy</td>
<td>33%</td>
<td>11%</td>
<td>2 out of 18</td>
<td></td>
</tr>
<tr>
<td>sri_tpot</td>
<td>61%</td>
<td>11%</td>
<td>2 out of 18</td>
<td></td>
</tr>
<tr>
<td>texasam_tamu</td>
<td>94%</td>
<td>22%</td>
<td>4 out of 18</td>
<td></td>
</tr>
</tbody>
</table>

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

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