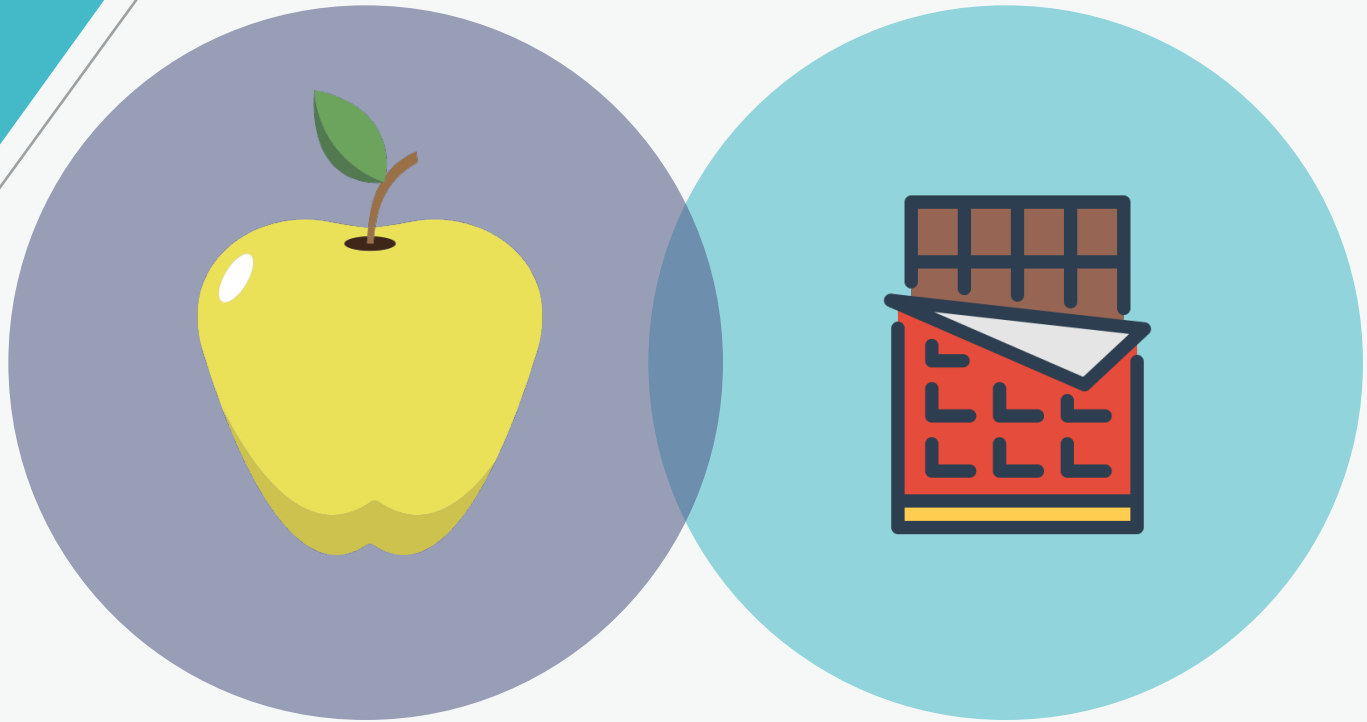

Nima Shahbazi, Co-Founder of

Deepnify

# Reducing Demand Forecast Error with Deep Learning



@Deepnify
nima@deepnify.com

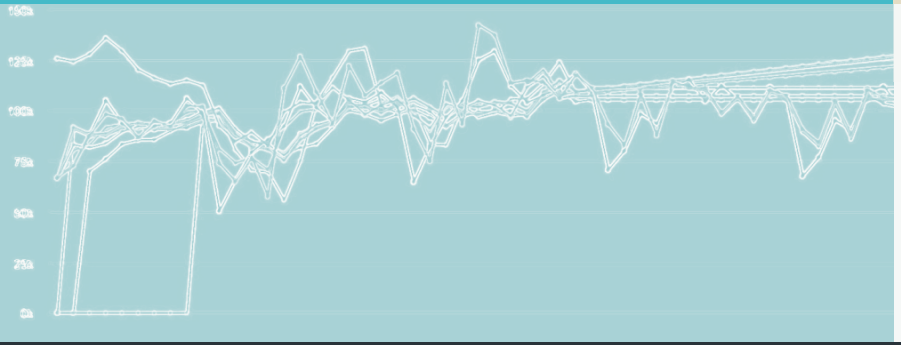


# 20% Waste



@Deepnify
nima@deepnify.com

Every fractional change in forecast error impacts food waste or stockouts.



So local store managers manually adjust orders based on local demand drivers and new information.



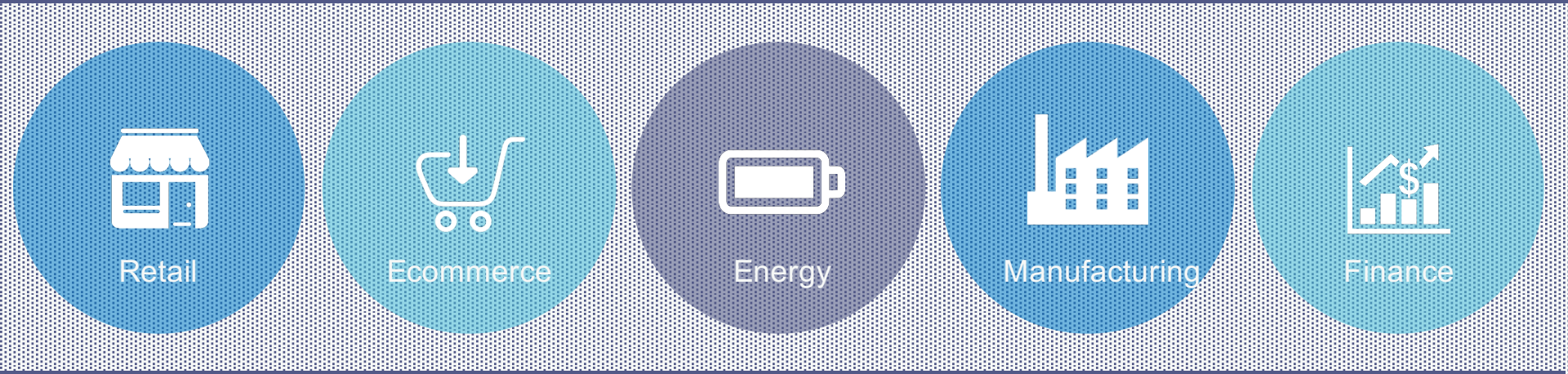
@Deepnify
nima@deepnify.com

# $15B



@Deepnify
nima@deepnify.com

# Forecast error is expensive across industries





@Deepnify
nima@deepnify.com


kaggle

#19



@Deepnify
nima@deepnify.com

# Zillow announces $1M prize for anyone who can improve the algorithm for its Zestimate

, 2017 at 9:58 am

67 Tweet Share 251 Reddit Email

GeekWire Summit: Tickets here!


, much-debated, computer-generated home valuation tool,
Seattle-based real estate media company for 11 years. And
ce tag on whether the algorithm that assigns a price tag

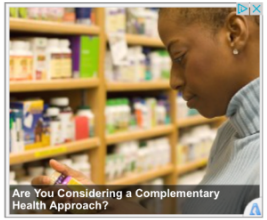


Bloomberg | Two Sigma Unleashes 730,000 Data Scientists in Algo Contest

# Two Sigma Unleashes 730,000 Data Scientists in Algo Contest

Simone Foxman and Saijel Kishan
December 1, 2016, 10:45 AM EST *Updated on* December 1, 2016, 4:48 PM EST

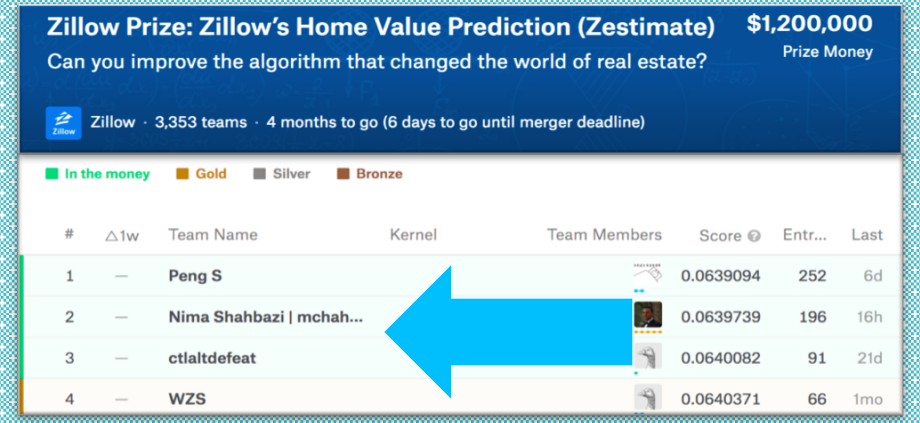
- Winners of machine-learning competition to divide $100,000 pot
- Hedge fund seeks to recruit talented data scientists

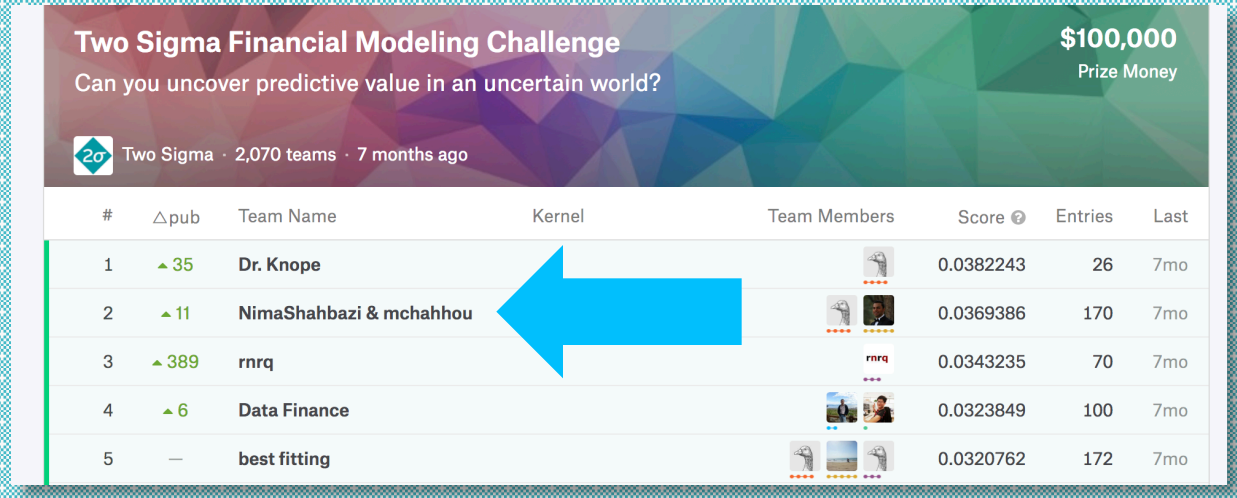



How do you tap the savvy of 730,000 data scientists for financial markets?

Offer them $100,000 in prize money and a chance to impress one of the most successful hedge funds in the world.

That's what $38 billion Two Sigma is doing with its new partner Kaggle, a community of data scientists who collaborate and compete in writing machine-learning algorithms. In a contest starting Thursday at noon, Kaggle's quants will have three months to create a predictive trading model from four gigabytes of financial data provided by Two Sigma. Cash prizes go to the seven best performers. Two Sigma, which competes with Silicon


Zillow Prize: Zillow's Home Value Prediction (Zestimate)
$1,200,000
Prize Money
Can you improve the algorithm that changed the world of real estate?
Zillow · 3,353 teams · 4 months to go (6 days to go until merger deadline)
In the money
Gold
Silver
Bronze
# △1w Team Name Kernel Team Members Score Entr... Last
1 — Peng S 0.0639094 252 6d
2 — Nima Shahbazi | mchah... 0.0639739 196 16h
3 — ctlaltdefeat 0.0640082 91 21d
4 — WZS 0.0640371 66 1mo


Two Sigma Financial Modeling Challenge
$100,000
Prize Money
Can you uncover predictive value in an uncertain world?
Two Sigma · 2,070 teams · 7 months ago
# △pub Team Name Kernel Team Members Score Entries Last
1 ▲35 Dr. Knope 0.0382243 26 7mo
2 ▲11 NimaShahbazi & mchahhou 0.0369386 170 7mo
3 ▲389 rnrq 0.0343235 70 7mo
4 ▲6 Data Finance 0.0323849 100 7mo
5 — best fitting 0.0320762 172 7mo

# Deepnify: AI-Powered Forecasting for Enterprise

ENTREPRENEUR

TRENDING Home Capital | NAFTA | Taxes | Bombardier | Canadian dollar | Family Finance

## Revolution AI: First cohort underway at NextAI; 20 teams set out to develop marketable AI solutions

Denise Deveau | March 10, 2017 | Last Updated: Mar 14 9:47 AM ET
More from Denise Deveau

Republish
Reprint

From left, Dan Yang, Noel Webb and David Vradenburg at the Rotman School of Business. Peter J. Thompson / National Post

_NextAI
@_NextAI

Follow

The winner of the 2017 People's Choice Global Impact Award goes to @deepnify! Congratulations to co-founders @thekristalee & @NeemaShahbazi!

1 Retweet 7 Likes

1 7


kaggle

ROSSMANN

kaggle

Weston

Loblaws



@Deepnify
nima@deepnify.com


Su
sale

# Principles of Forecasting

Creating a Validation Strategy

- Error metric and objective function
- Validation performance

Reducing Error

- Reducing error with Entity Embedding
- Reducing error with Stacking models

# Evaluation Metric and Objective Function

Question:

Will minimizing the squared error yield the same result as minimizing the absolute error?

- When minimizing an error, we must decide how to penalize these errors
- $MSE = \frac{1}{n}\sum_{i=1}^{n}(Y_i' - Y_i)^2$ $\quad\quad$ $MAE = \frac{1}{n}\sum_{i=1}^{n}|Y_i' - Y_i|$
- $Y'$ *is a vector on n prediciton and* $Y$ *is a vector of observed value.*


@Deepnify
nima@deepnify.com

# Evaluation Metric and Objective Function

```
0.0003 0.0001 0.0002 0.0004 50000 0.0002 0.0004 0.0003 0.0001 0.0003
```

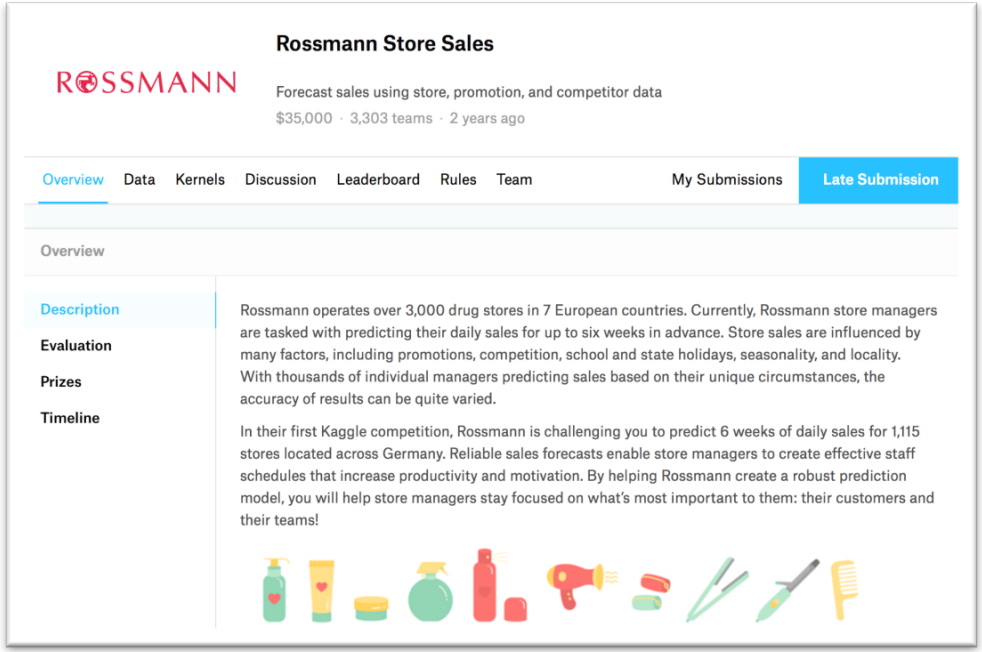
The MAD solution is $\alpha$ = 0.0003. The MSE solution is 5000.00023

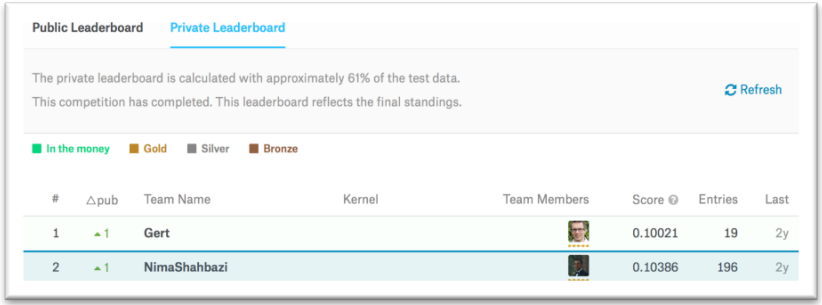
Answer:

We may not even able to beat a simple model if we chose a wrong objective Function!


@Deepnify
nima@deepnify.com

# Predicting Daily Sales



# Data

You are provided with historical sales data for 1,115 Rossmann stores. The task is to forecast the "Sales" column for the test set. Note that some stores in the dataset were temporarily closed for refurbishment.

### Files

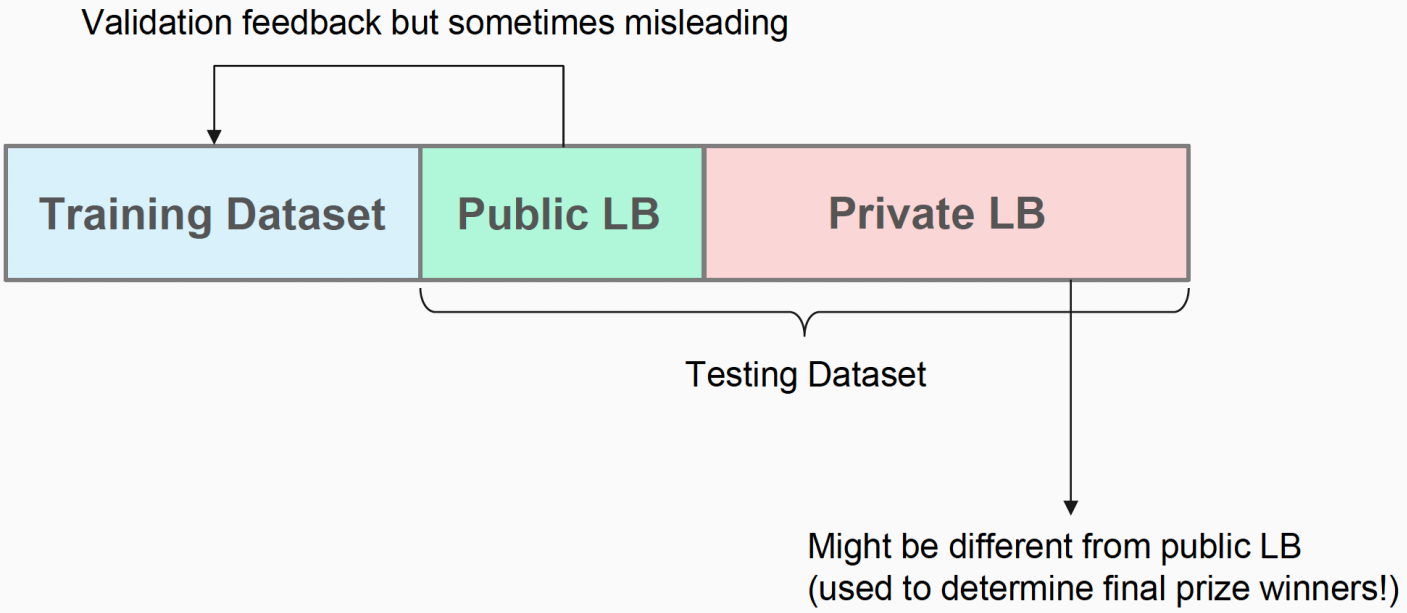
- **train.csv** - historical data including Sales
- **test.csv** - historical data excluding Sales
- **sample_submission.csv** - a sample submission file in the correct format
- **store.csv** - supplemental information about the stores

### Data fields

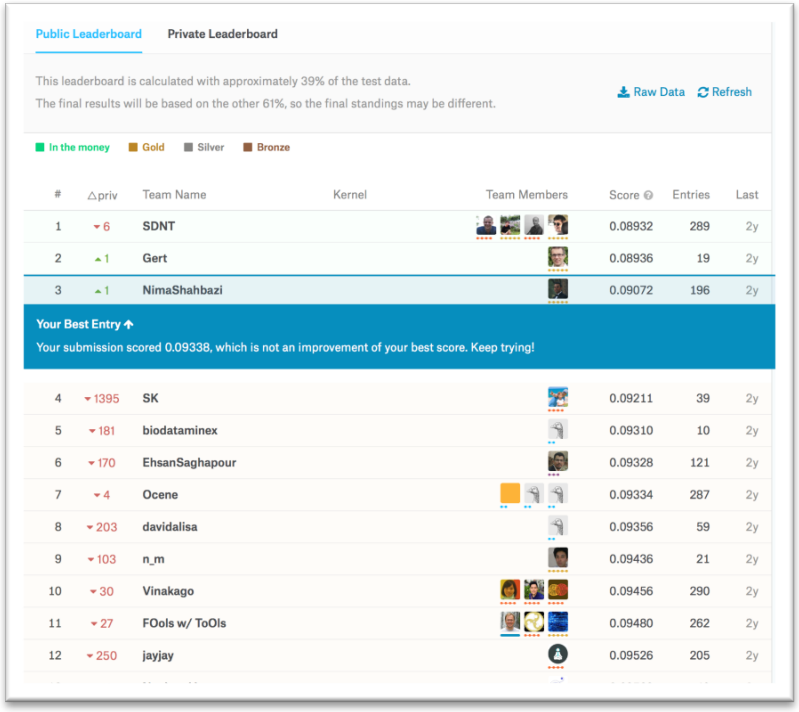
Most of the fields are self-explanatory. The following are descriptions for those that aren't.

- **Id** - an Id that represents a (Store, Date) duple within the test set
- **Store** - a unique Id for each store
- **Sales** - the turnover for any given day (this is what you are predicting)
- **Customers** - the number of customers on a given day
- **Open** - an indicator for whether the store was open: 0 = closed, 1 = open
- **StateHoliday** - indicates a state holiday. Normally all stores, with few exceptions, are closed on state holidays. Note that all schools are closed on public holidays and weekends. a = public holiday, b = Easter holiday, c = Christmas, 0 = None
- **SchoolHoliday** - indicates if the (Store, Date) was affected by the closure of public schools
- **StoreType** - differentiates between 4 different store models: a, b, c, d
- **Assortment** - describes an assortment level: a = basic, b = extra, c = extended
- **CompetitionDistance** - distance in meters to the nearest competitor store
- **CompetitionOpenSince[Month/Year]** - gives the approximate year and month of the time the nearest competitor was opened
- **Promo** - indicates whether a store is running a promo on that day
- **Promo2** - Promo2 is a continuing and consecutive promotion for some stores: 0 = store is not participating, 1 = store is participating
- **Promo2Since[Year/Week]** - describes the year and calendar week when the store started participating in Promo2
- **PromoInterval** - describes the consecutive intervals Promo2 is started, naming the months the promotion is started anew. E.g. "Feb,May,Aug,Nov" means each round starts in February, May, August, November of any given year for that store

# Data


Validation feedback but sometimes misleading
Training Dataset
Public LB
Private LB
Testing Dataset
Might be different from public LB
(used to determine final prize winners!)

# Validation Strategy



Public Leaderboard | Private Leaderboard

This leaderboard is calculated with approximately 39% of the test data.
The final results will be based on the other 61%, so the final standings may be different.

Raw Data | Refresh

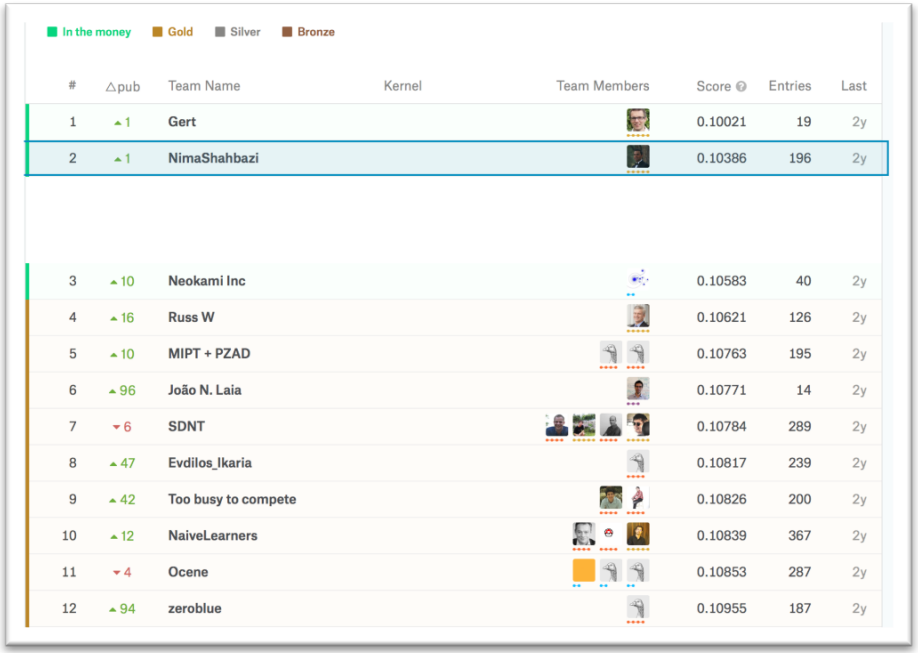
In the money | Gold | Silver | Bronze

| # | △priv | Team Name | Kernel | Team Members | Score | Entries | Last |
|---|---|---|---|---|---|---|---|
| 1 | ▼6 | SDNT | | | 0.08932 | 289 | 2y |
| 2 | ▲1 | Gert | | | 0.08936 | 19 | 2y |
| 3 | ▲1 | NimaShahbazi | | | 0.09072 | 196 | 2y |

**Your Best Entry ↑**

Your submission scored 0.09338, which is not an improvement of your best score. Keep trying!

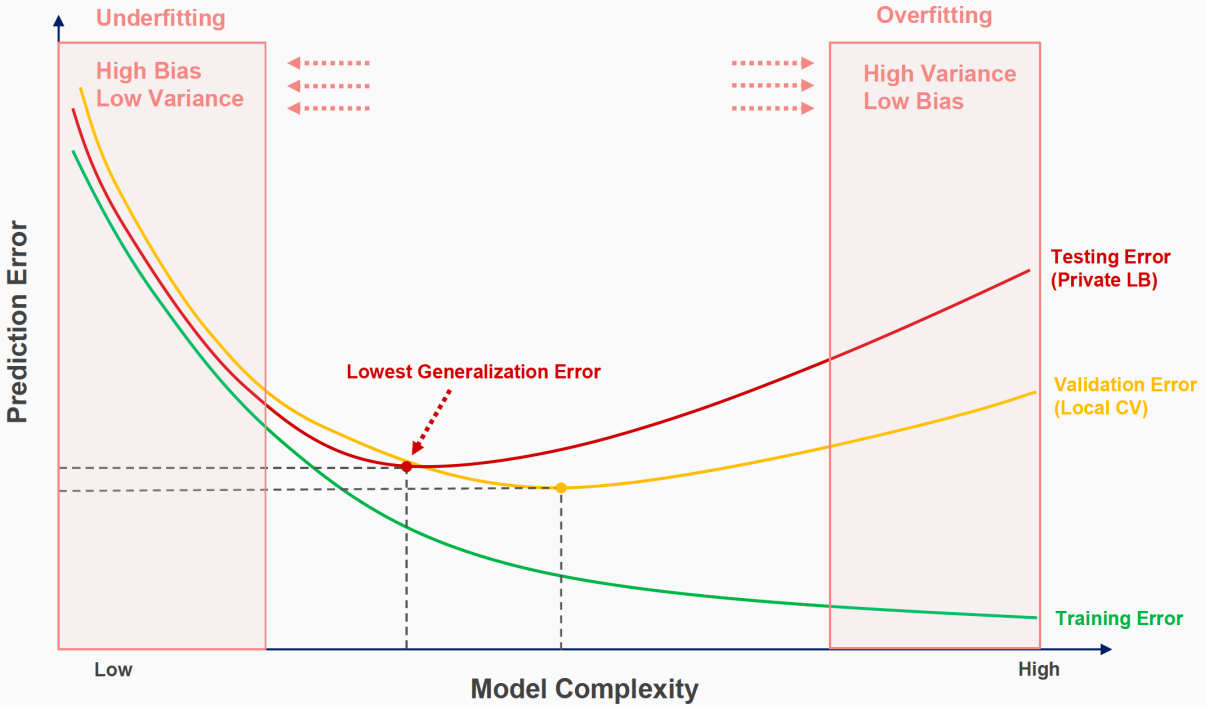
| # | △priv | Team Name | Kernel | Team Members | Score | Entries | Last |
|---|---|---|---|---|---|---|---|
| 4 | ▼1395 | SK | | | 0.09211 | 39 | 2y |
| 5 | ▼181 | biodataminex | | | 0.09310 | 10 | 2y |
| 6 | ▼170 | EhsanSaghapour | | | 0.09328 | 121 | 2y |
| 7 | ▼4 | Ocene | | | 0.09334 | 287 | 2y |
| 8 | ▼203 | davidalisa | | | 0.09356 | 59 | 2y |
| 9 | ▼103 | n_m | | | 0.09436 | 21 | 2y |
| 10 | ▼30 | Vinakago | | | 0.09456 | 290 | 2y |
| 11 | ▼27 | FOols w/ ToOls | | | 0.09480 | 262 | 2y |
| 12 | ▼250 | jayjay | | | 0.09526 | 205 | 2y |



In the money | Gold | Silver | Bronze

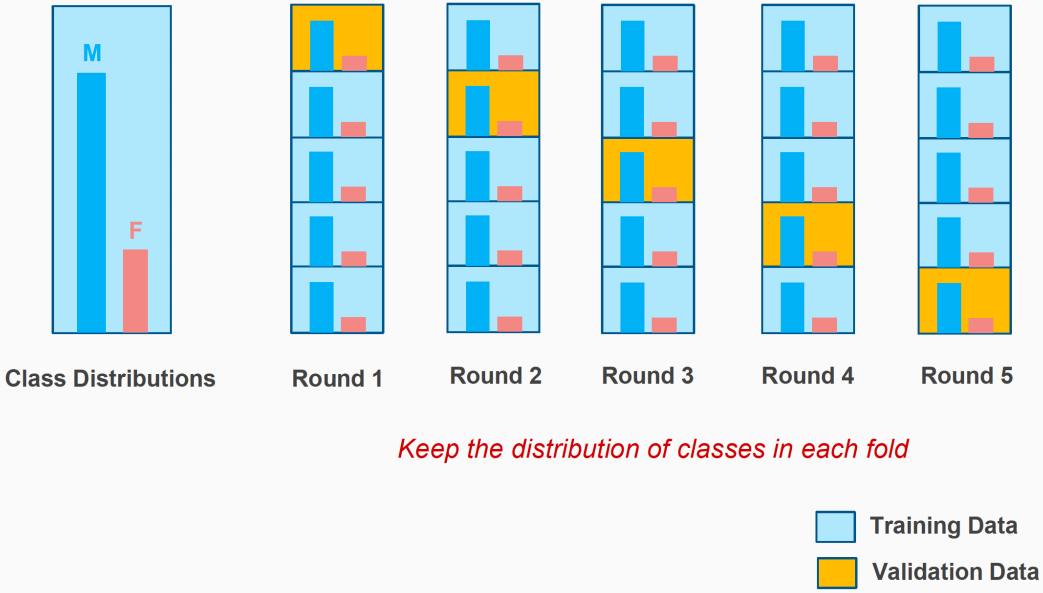
| # | △pub | Team Name | Kernel | Team Members | Score | Entries | Last |
|---|---|---|---|---|---|---|---|
| 1 | ▲1 | Gert | | | 0.10021 | 19 | 2y |
| 2 | ▲1 | NimaShahbazi | | | 0.10386 | 196 | 2y |
| 3 | ▲10 | Neokami Inc | | | 0.10583 | 40 | 2y |
| 4 | ▲16 | Russ W | | | 0.10621 | 126 | 2y |
| 5 | ▲10 | MIPT + PZAD | | | 0.10763 | 195 | 2y |
| 6 | ▲96 | João N. Laia | | | 0.10771 | 14 | 2y |
| 7 | ▼6 | SDNT | | | 0.10784 | 289 | 2y |
| 8 | ▲47 | Evdilos_Ikaria | | | 0.10817 | 239 | 2y |
| 9 | ▲42 | Too busy to compete | | | 0.10826 | 200 | 2y |
| 10 | ▲12 | NaiveLearners | | | 0.10839 | 367 | 2y |
| 11 | ▼4 | Ocene | | | 0.10853 | 287 | 2y |
| 12 | ▲94 | zeroblue | | | 0.10955 | 187 | 2y |

# Validation Strategy



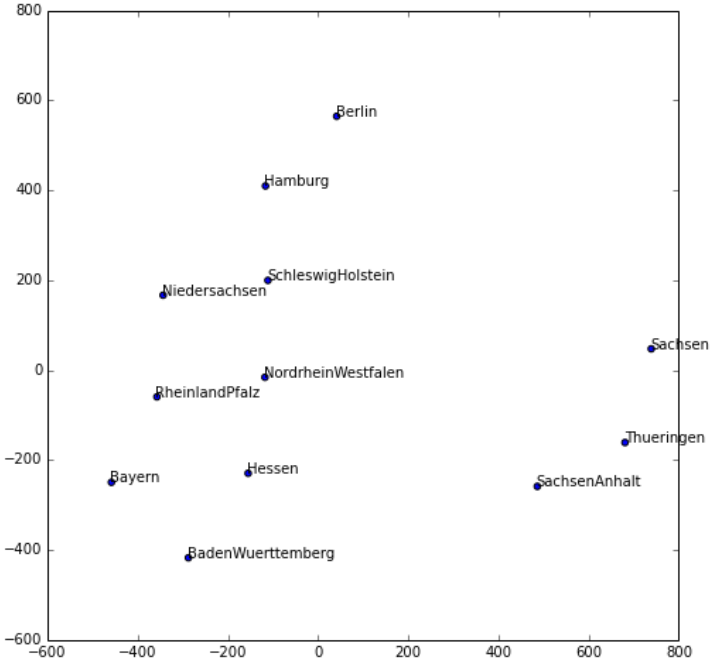
# Validation Strategy

Hold out, K-Fold, Stratified K-Fold, Time Series Fold …



# Entity Embedding

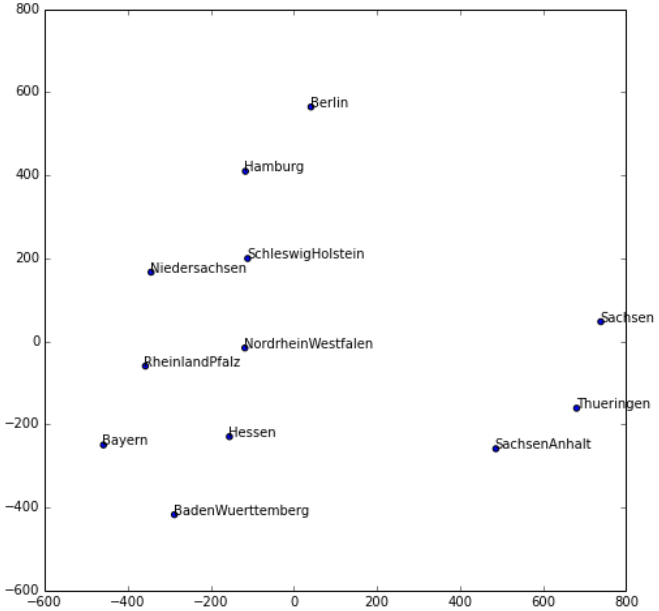
- Learn Deep NN to predict Sales.
- Embedding layer: Vector of 6 values per State.
- German States 2D T-SNE embedding [2]
- Algorithm does not know anything about German geography!
- It is inspired by semantic embedding in the natural language processing domain [3]



# Entity Embedding

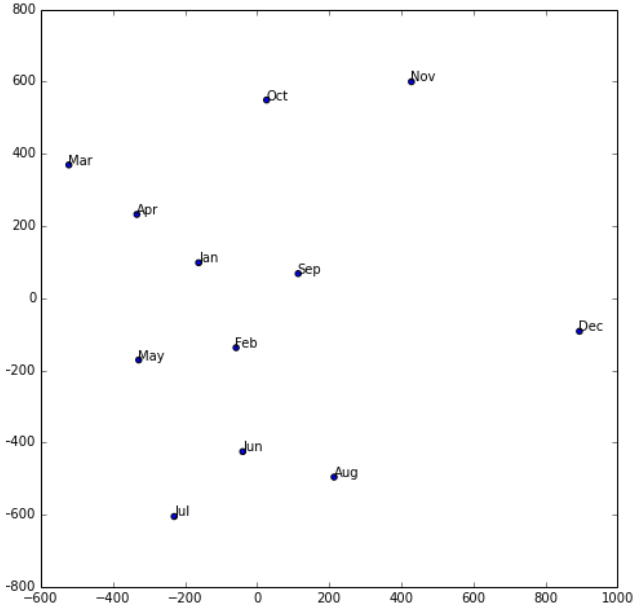
Real German States and learned embedding look surprisingly similar!

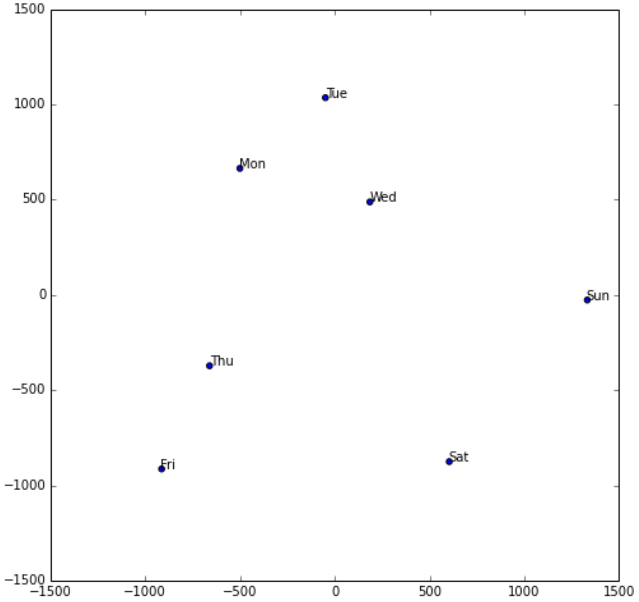
Embedding maps states close together when they have a similar interaction of features.

# Entity Embedding

Month and day of week 2D T-SNE embedding [2]

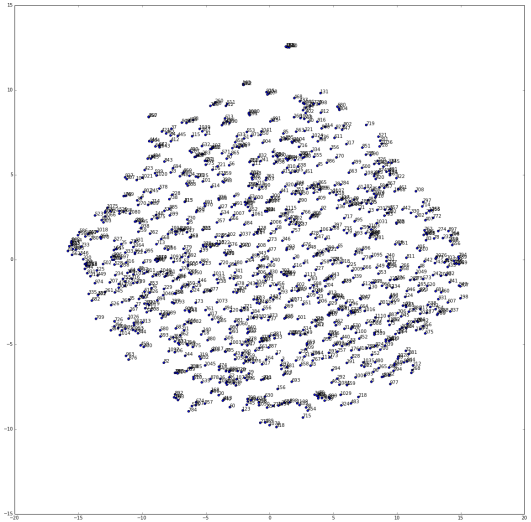


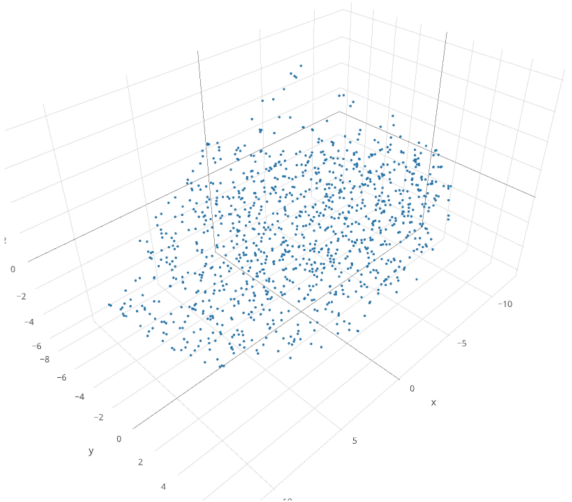



# Entity Embedding

Stores 2D and 3D T-SNE embedding [2]

Applies to many other problems to find the hidden relations among entities based on their interaction



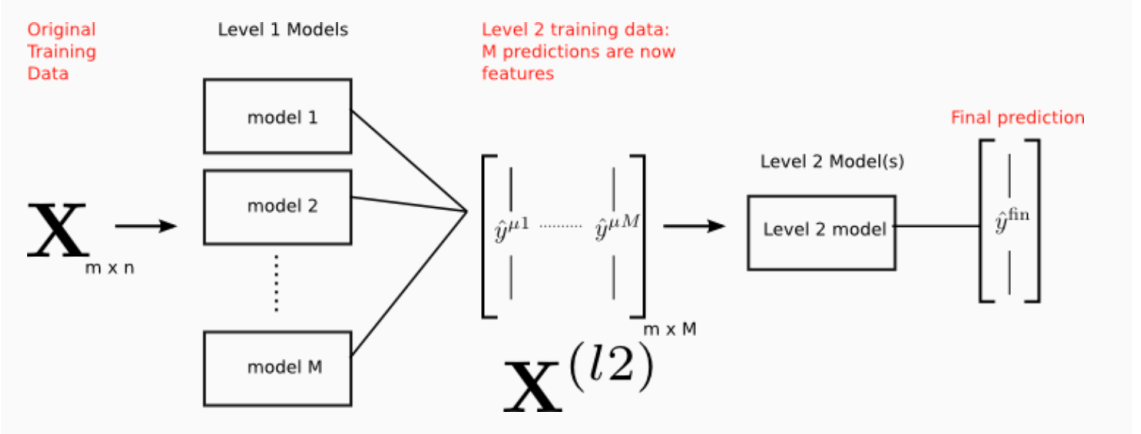



@Deepnify
nima@deepnify.com

# Improving Accuracy (Stacking)

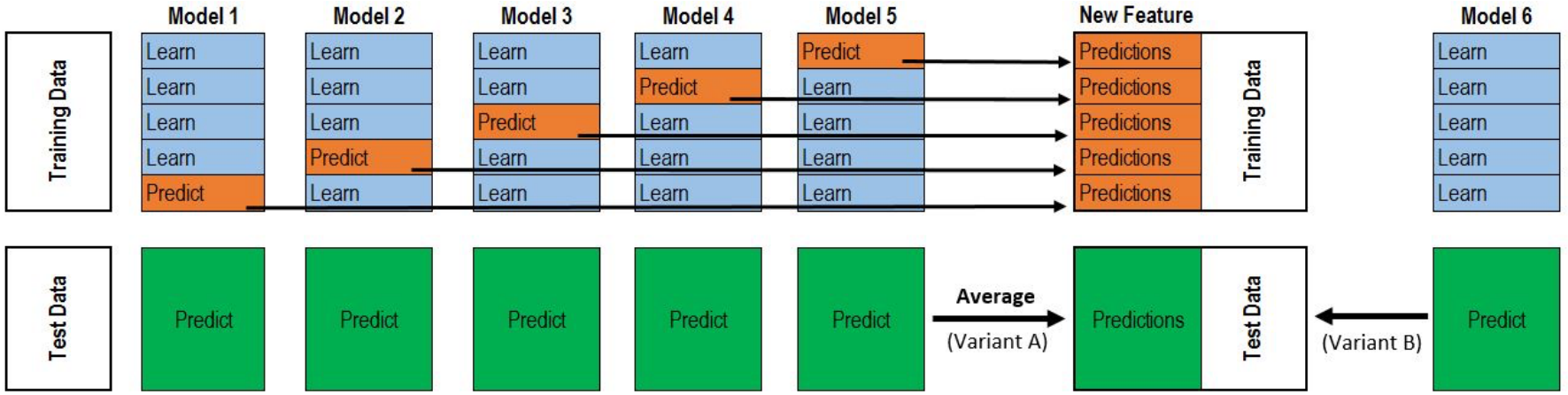
- Different models

- Deep NN
- Gradient Boosting Trees
- Random Forest
- SVM
- …



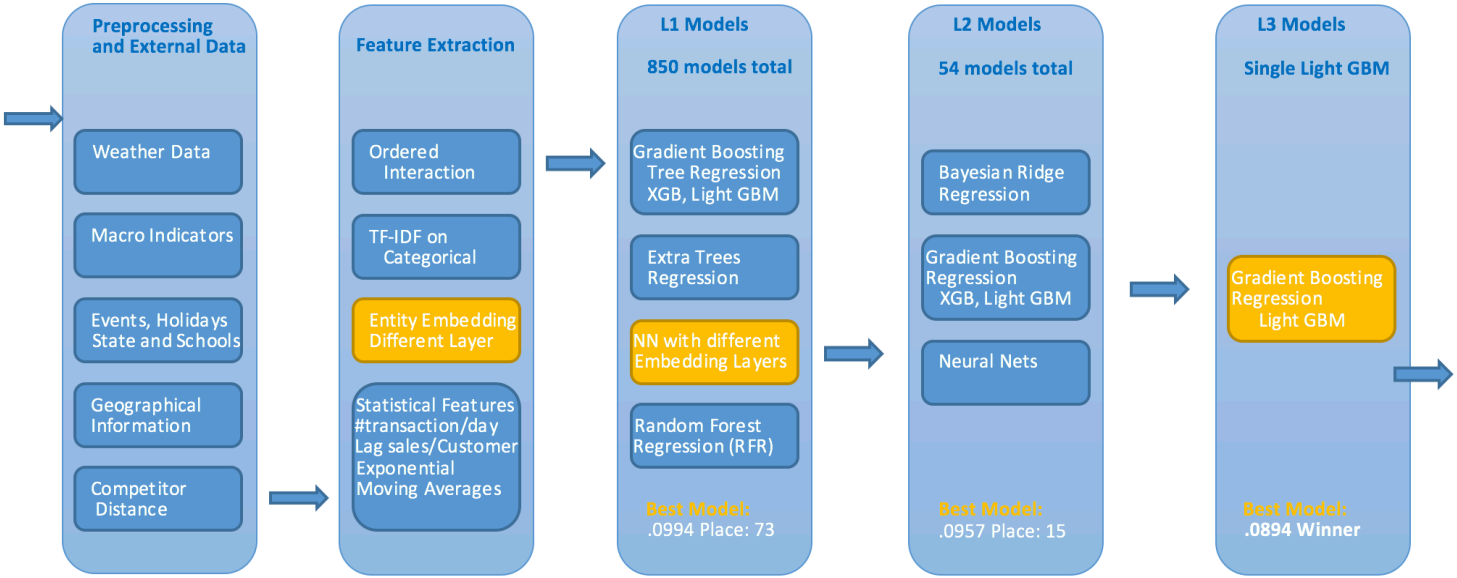
# Improving Accuracy

Stacking Generalization



# Full Pipeline

Rossmann Winning Strategy (NN Model on pre-process data: error 0.1451)



# References

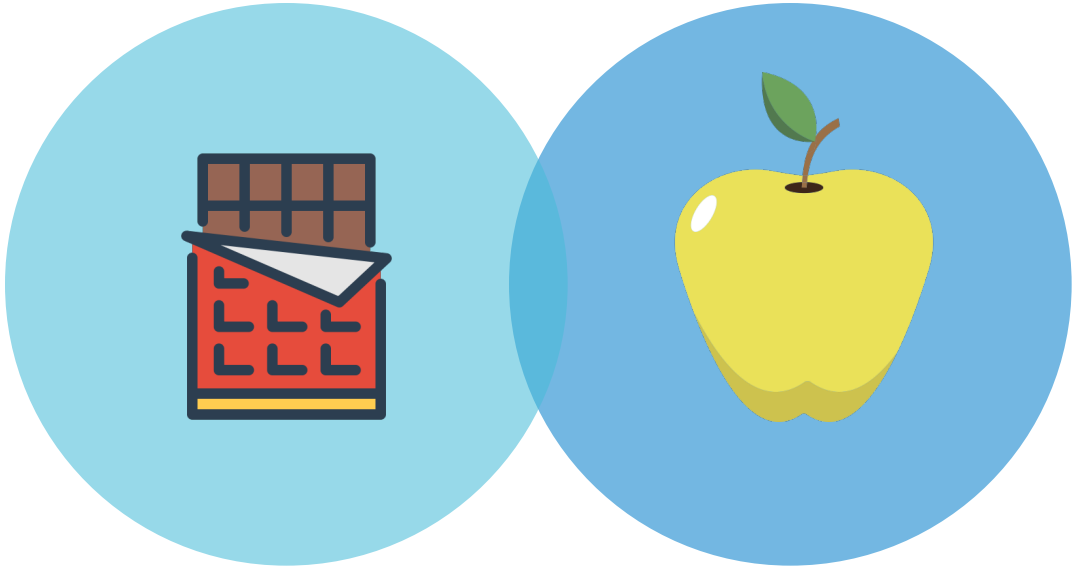
[1] Rossmann Store Sales Kaggle Competition, https://www.kaggle.com/c/rossmann-store-sales

[2] GitHub: https://github.com/entron/entity-embedding-rossmann

[3] Cheng Guo and Felix Berkhahn. 2016. Entity embeddings of categorical variables. arXiv preprint arXiv:1604.06737.



@Deepnify
nima@deepnify.com




@Deepnify
nima@deepnify.com