JOURNEY TO KAGGLE COMPETITION

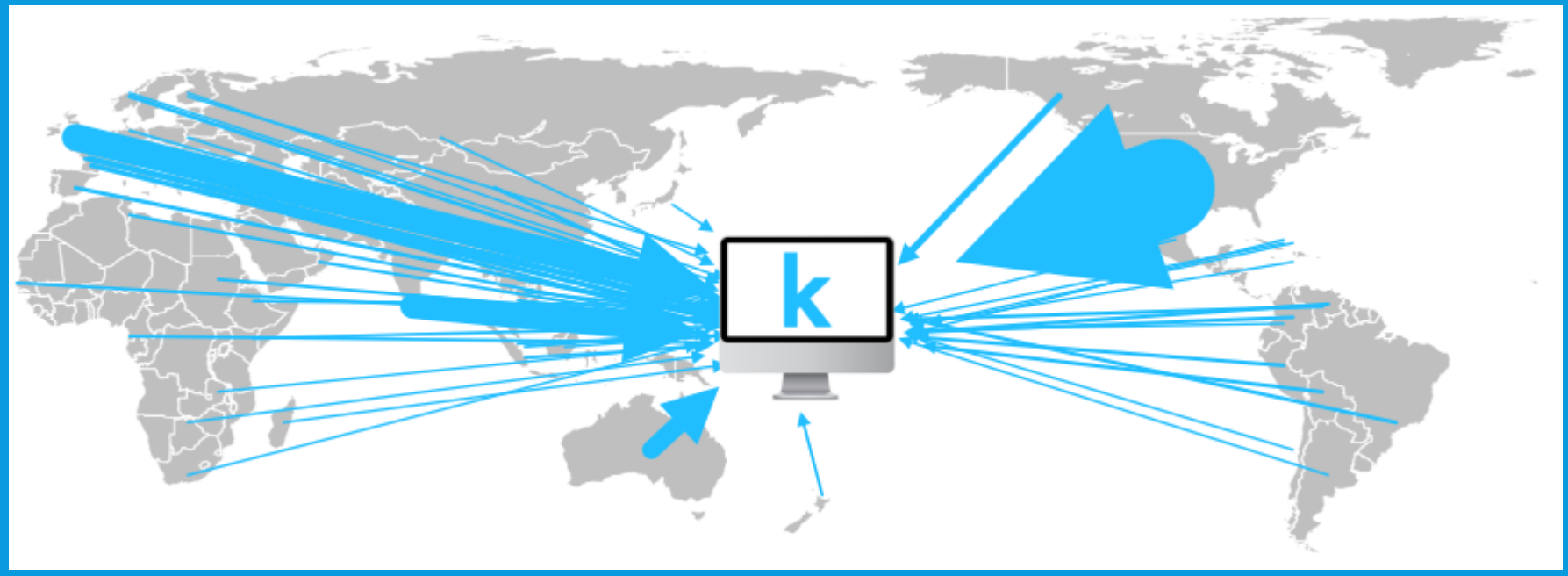
Kyu Cho

JOURNEY TO KAGGLE COMPETITION

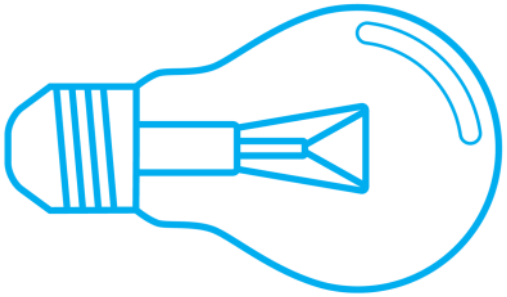
Kyu Cho

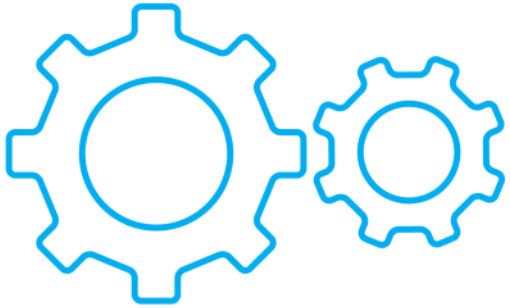
WHAT IS KAGGLE

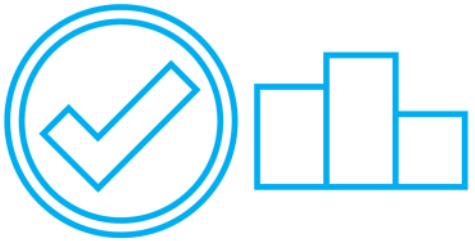
- Kaggle is a platform for predictive modelling and analytics competitions on which companies and researchers post their data and statisticians and data miners from all over the world compete to produce the best models



How Kaggle Works







1

Users create predictive models,

2

submit these to Kaggle,

3

and are scored on their accuracy.

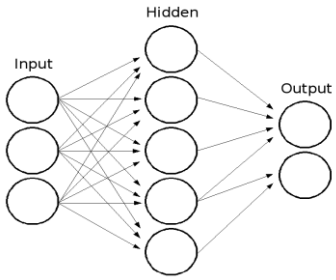
Competition Mechanics

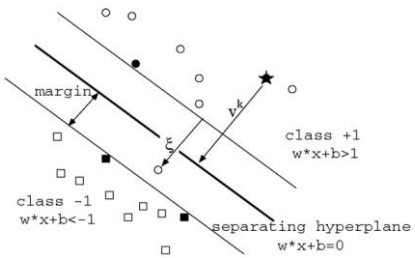
Training dataset		
Age	Income	Default
58	$ 95,824.00	TRUE
73	$ 20,708.00	FALSE
59	$ 82,152.00	FALSE
66	$ 25,334.00	FALSE
39	$ 35,952.00	FALSE
78	$ 51,754.00	FALSE
76	$ 76,479.00	TRUE
71	$ 96,614.00	TRUE
22	$ 27,701.00	FALSE
57	$ 35,841.00	FALSE

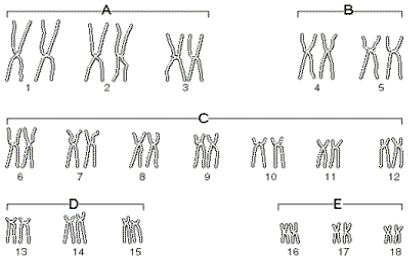
Test dataset		
Age	Income	Default
73	$ 53,445.00	
61	$ 36,679.00	
47	$ 90,422.00	
44	$ 79,040.00	
46	$ 67,104.00	
30	$ 69,992.00	
75	$ 78,139.00	
28	$ 66,058.00	
24	$ 75,240.00	
54	$ 89,503.00	

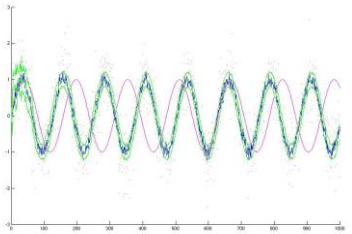
Competitions are judged on objective criteria

Different users - different techniques.









- neural networks
- logistic regression
- support vector machine
- decision trees
- ensemble methods
- adaBoost
- Bayesian networks

- genetic algorithms
- random forest
- Monte Carlo methods
- principal component analysis
- Kalman filter
- evolutionary fuzzy modelling

#	Team Name	RMSE	Entries	Latest Submission
1	PEW *	0.640871	130	6:00pm, Monday 1 November 2010
2	UriB *	0.646554	118	9:33am, Saturday 30 October 2010
3	Just For Fun *	0.649665	11	2:34am, Thursday 2 September 2010
4	Old Dogs With New Tricks *	0.649922	87	7:49am, Tuesday 2 November 2010
5	JohnL *	0.652753	11	10:10am, Thursday 7 October 2010
6	PunyPetunias *	0.65485	52	12:04pm, Tuesday 21 September 2010
7	ulvund *	0.655488	52	8:59pm, Thursday 28 October 2010
8	Diogo *	0.655815	85	5:57pm, Monday 1 November 2010
9	Jasonb *	0.656661	50	9:43am, Saturday 23 October 2010
10	ChessMaster *	0.65683	44	6:53pm, Friday 17 September 2010

Competitions are judged based on predictive accuracy

Transforming Data Into Insight For Making Better Decisions

Analysis

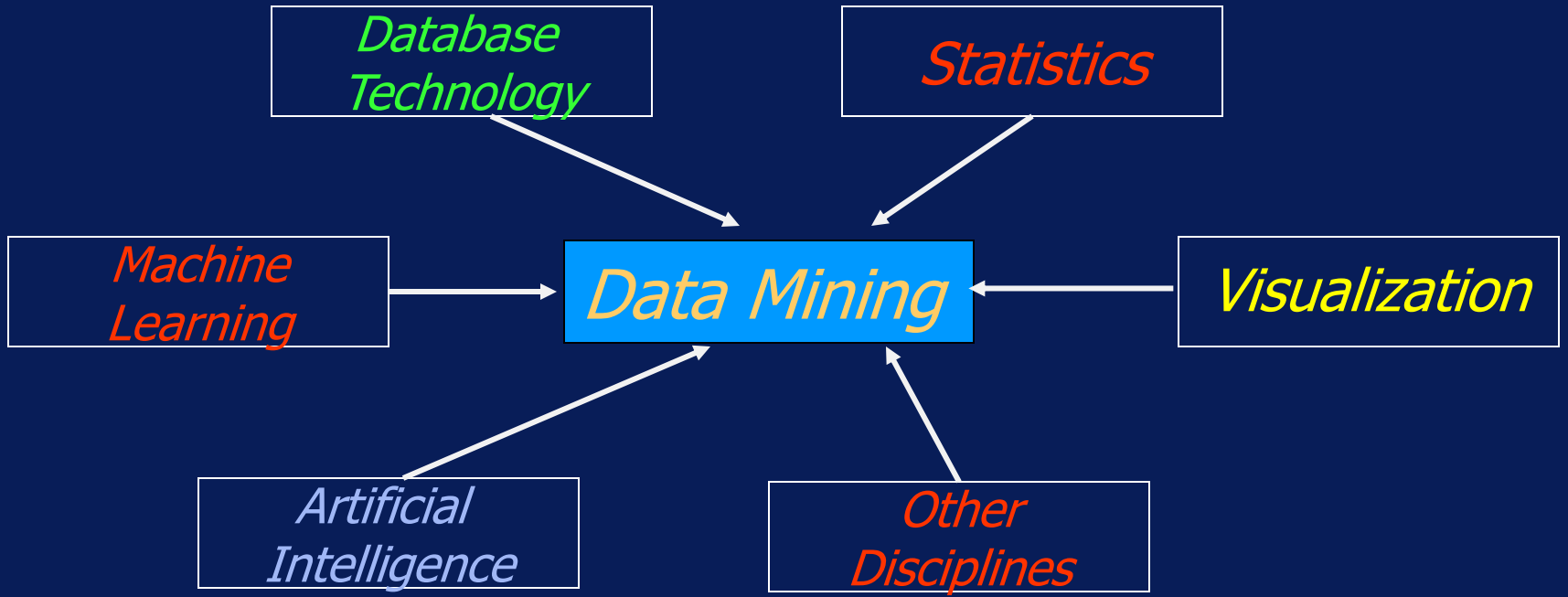
Data Driven Decision

Data

Insight

Action

Multidisciplinary Field



2. Regression

Case study: Predicting house prices

Models

- Linear regression
- Regularization: Ridge (L2), Lasso (L1)

Algorithms

- Gradient descent
- Coordinate descent

Concepts

- Loss functions, bias-variance tradeoff, cross-validation, sparsity, overfitting, model selection

3. Classification

Case study: Analyzing sentiment

Models

- Linear classifiers (logistic regression, SVMs, perceptron)
- Kernels
- Decision trees

Algorithms

- Stochastic gradient descent
- Boosting

Concepts

- Decision boundaries, MLE, ensemble methods, random forests, CART, online learning

4. Clustering & Retrieval

Case study: Finding documents

Models

- Nearest neighbors
- Clustering, mixtures of Gaussians
- Latent Dirichlet allocation (LDA)

Algorithms

- KD-trees, locality-sensitive hashing (LSH)
- K-means
- Expectation-maximization (EM)

Concepts

- Distance metrics, approximation algorithms, hashing, sampling algorithms, scaling up with map-reduce

5. Matrix Factorization & Dimensionality Reduction

Case study: Recommending Products

Models

- Collaborative filtering
- Matrix factorization
- PCA

Algorithms

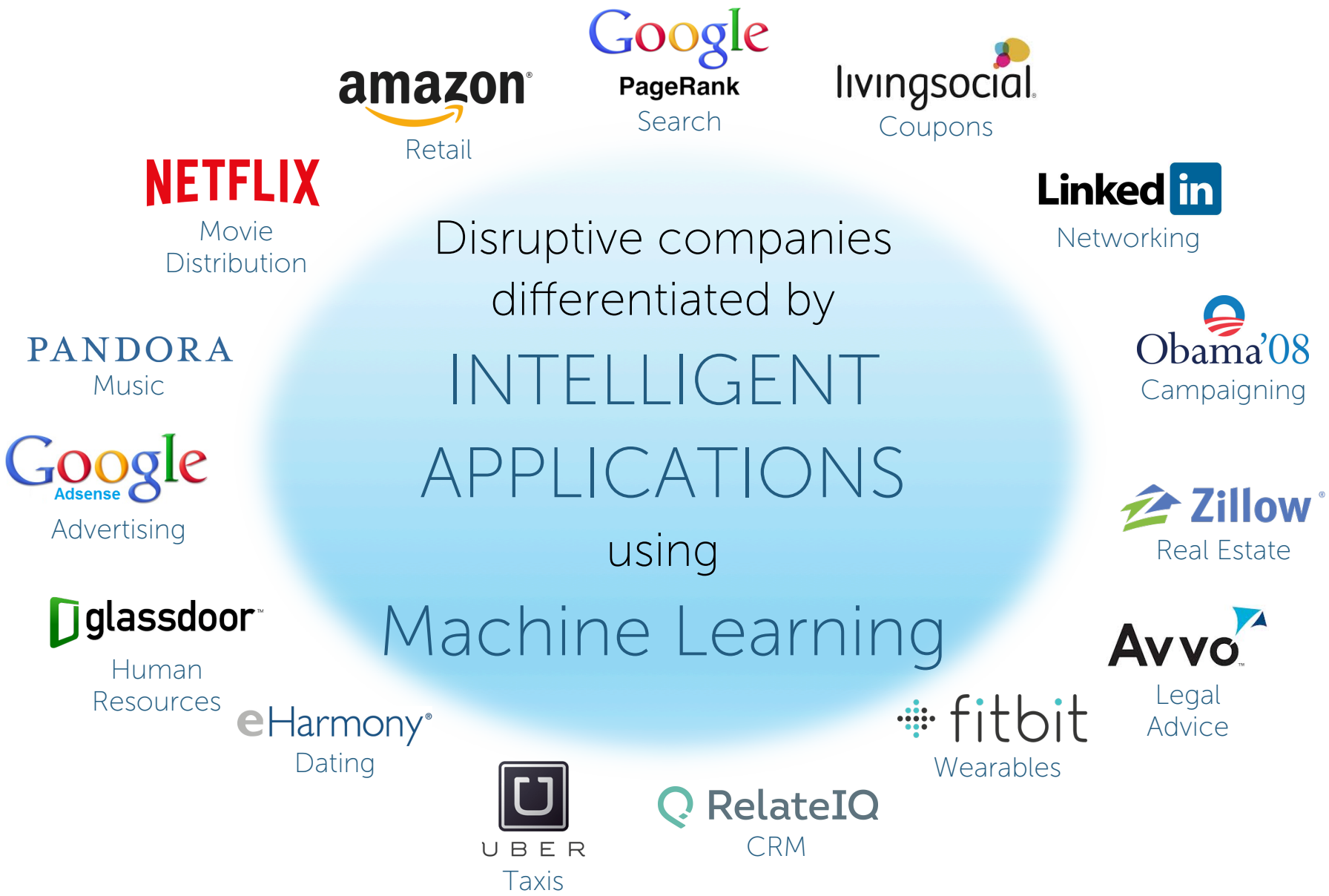
- Coordinate descent
- Eigen decomposition
- SVD

Concepts

- Matrix completion, eigenvalues, random projections, cold-start problem, diversity, scaling up

Open Source Data Mining Tools

- Python
- R
- WEKA
- KNIME
- Orange
- RapidMiner
- Rattle
- Mahout
- MlLib



REFERENCES

- http://www.washington.edu/
- http://www.kaggle.com/