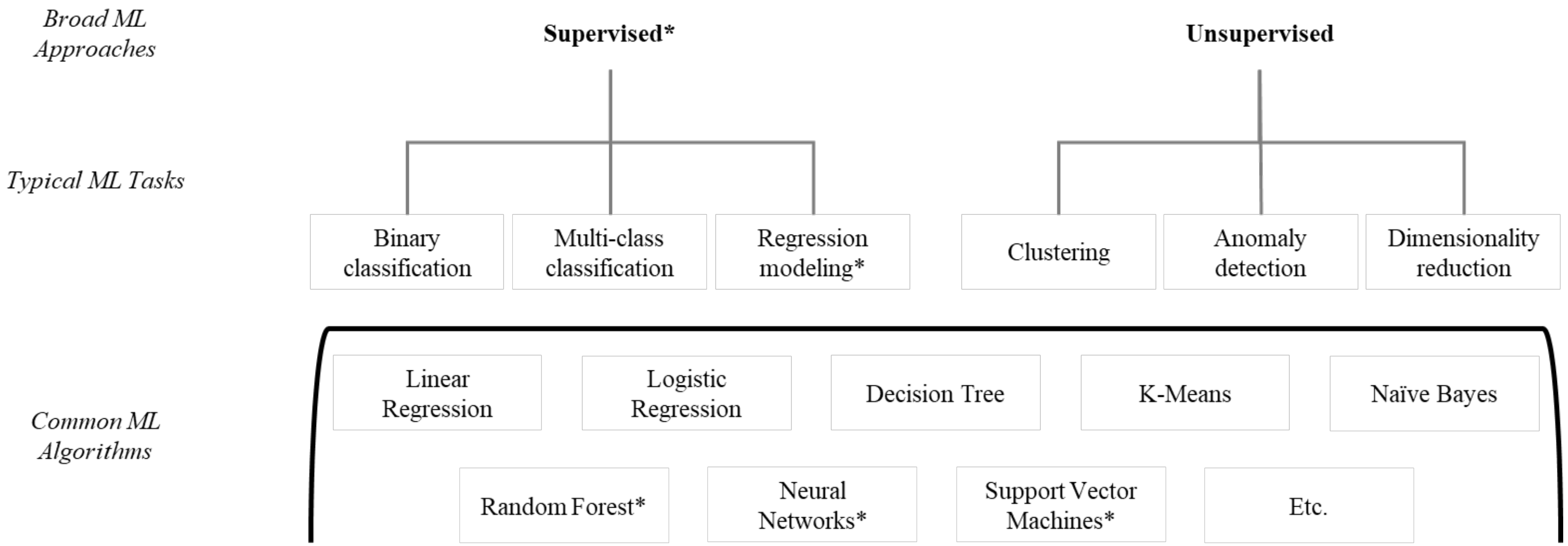
GARBAGE IN, GARBAGE OUT

Steven Hyde, Eric Bachura, & Joseph Harrison

ML Basics

- Unsupervised – Exploratory, Inductive
- Supervised – Confirmatory, Deductive



Traditional Statistics vs Machine Learning

- Statistical Inference
- Small sample size
    - More assumption
- Lots of human effort
- Abstract constructs
- Messy real world data
- Theory driven

- Predictive Accuracy
- Large sample sizes
    - Very few assumption
- No human effort
- Simple constructs
- Tightly controlled data
- Data driven

For Example:

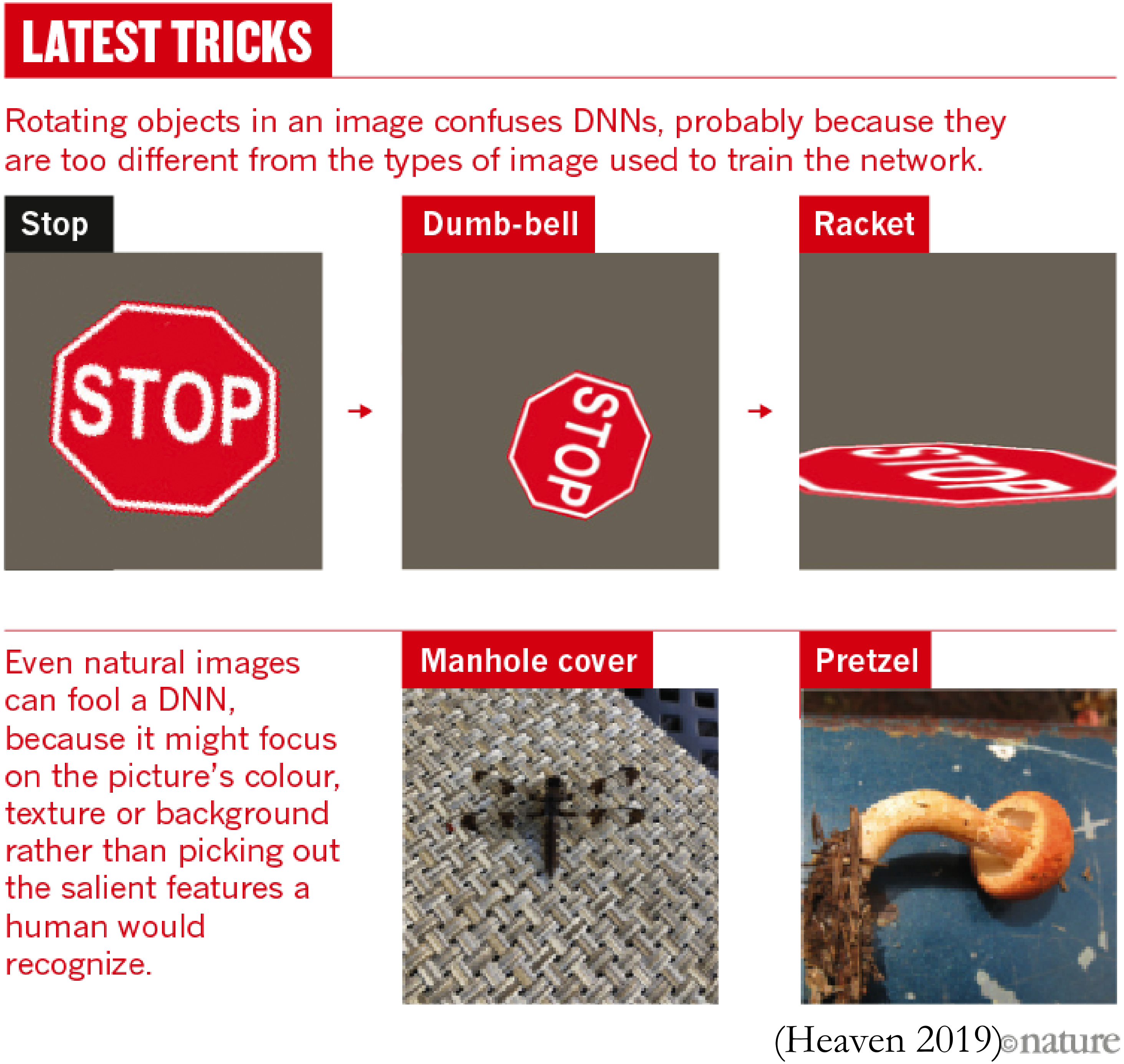
- Examines the impact of CEO narcissism on firm strategy and performance (Chatterjee & Hambrick, 2007). Measured through the prominence of CEO photograph; CEO prominence in company press releases; proportion of first-person singular pronouns by CEO; measure of relative pay. N=111

Vs

- Identify hand-written numbers, with a training set in the tens of thousands (Qiao et al., 2018)

Black Box Problem

- "Theories without methodological implications are likely to be little more than idle speculation with minimal empirical import. And methods without theoretical substance can be sterile, representing technical sophistication in isolation." Van Maanen and colleagues (2007: 1146) state,



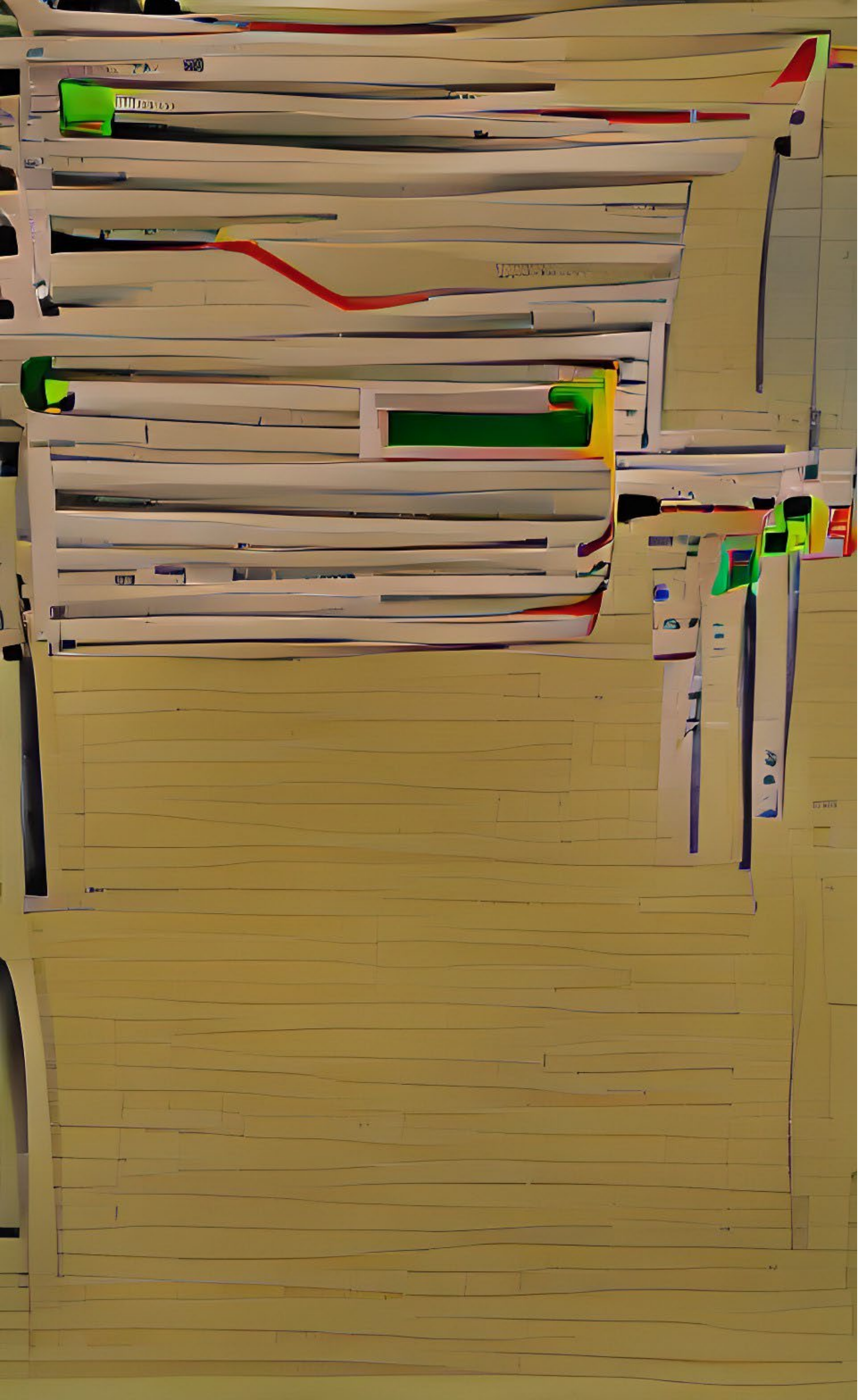
(Heaven 2019)

Phase 1: Pre Implementation Data Handling

- Is ML needed?
- Establish ground truth
  - Data source aligns with construct at right level
  - Training data validated
  - Sufficient variance
  - Data is large enough
- Theoretical pre-pruning

Theoretical pre-pruning

- Traditional approach- kitchen sink first, prune based on data
- Put theory first
- Don't give the algorithm the "background color" if you don't want it to use it.
  - Be as inclusive as possible, but only with features for which prior theory or logic would suggest there is a conceptual link



Phase 2: ML Implementation

- Select appropriate ML technique for construct
  - Binary, categorical, or continuous
  - Algorithm works with construct
    - Horse race if unknown

Phase 3: Post-Implementation

- Report everything
  - Hyperparameters too
- Refine your model

Demonstration: Need for Affiliation

- McClelland's need are subconscious trait-like motive
- Affiliation is the need to establishing and maintaining relationships (McClelland, 1987; McClelland and Boyatzis, 1982)
- Thematic Apperception Test (TAT)

Training data

- Quarterly earnings calls of the SP 500, 2008-2016
  - 700 randomly selected transcripts and 364 from the manual
  - 2 independent coders used Winter's (1994) scoring manual
  - Training set (80%) and hold out (20%)

Feature selection – Pre pruning

- LIWC, over 90 categories, representing on average 86%
- Pruned model:

Motivation
(drives, affiliation, achievement, power, reward, risk, authentic, comparisons, number, and quantifiers)

Goal Attainment
(tone, affect words, positive emotions, negative emotions, anxiety, anger, sadness, swear words, discrepancy, tentative, certainty, and clout)

Relations to others
(personal pronouns, social, family, friend, work, leisure, home, sexual, prepositions, assent, and total words count)

ML Techniques

- Scikit-learn, regression task
- Horse race: NN*, RF, SVM, & LIWC

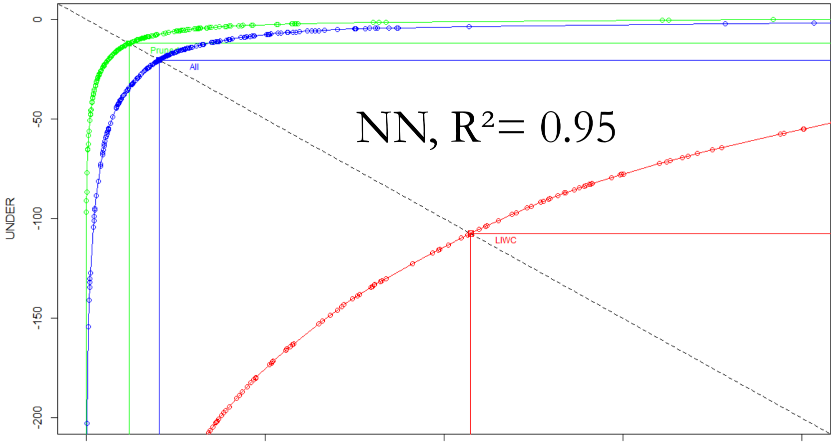
Type	All Features	Pruned Features
NN	Structure: Multi-layer perceptron, 2 hidden layers with 36 nodes each Solver: Stochastic gradient descent Activation: Hyperbolic tangent function Maximum iterations: 10000	Structure: Multi-layer perceptron, 2 hidden layers with 18 nodes each Solver: Optimized gradient descent Activation: Rectified linear unit Maximum iterations: 10000
RF	Structure: 1000 trees Decision criterion: MAE Features considered on splitting: All available	Structure: 1000 trees Decision Criterion: MSE Features considered on splitting: Pre-pruned features
SVM	Kernel: Polynomial Regularization value: 1	Kernel: Polynomial Regularization value: 10

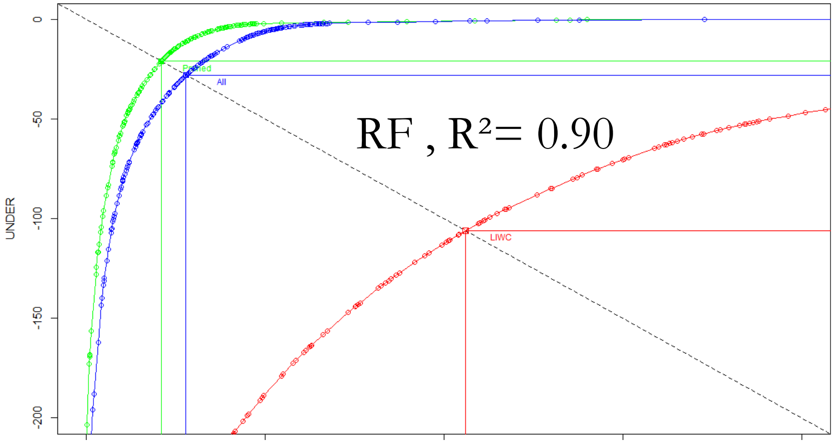
Results

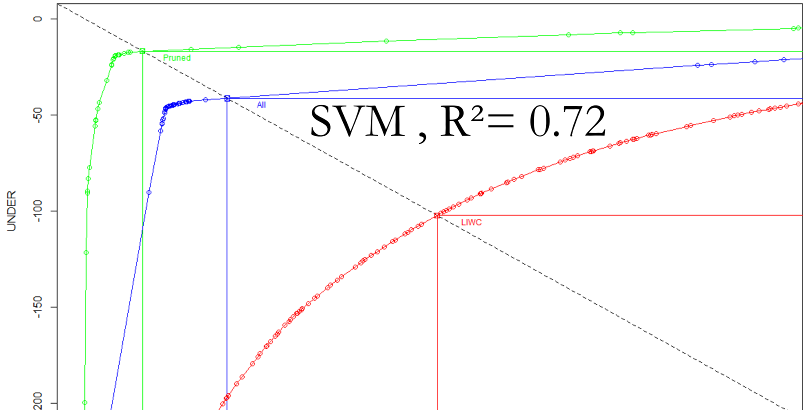
Model	R^2	RMSE	MSE	MAE
Average Performance Across ML Models				
Pre-pruned LIWC categories (Pruned)	0.879	0.590	0.435	0.251
All LIWC categories (All)	0.838	0.671	0.551	0.268
Affiliation dictionary (LIWC)	0.005	5.230	27.350	4.315
Performance of each ML Model				
NN-Pruned	0.955	0.194	0.038	0.124
NN-All	0.915	0.266	0.071	0.147
RF-Pruned	0.901	0.288	0.083	0.151
RF-All	0.886	0.309	0.095	0.176
SVM-Pruned	0.724	0.480	0.230	0.084
SVM-All	0.550	0.613	0.376	0.143

Findings

- Pre-pruned models consistently outperformed unpruned models: mean +9.7% in R^2, -19% RMSE, -35.2% MSE, & -23% MAE
  - LWIC was a poor measure (R^2 = .005, RMSE = 5.230, MSE = 27.350, MAE = 4.315)
- Order of accuracy: NN, RF, and SMV.
- Pre-Pruning more beneficial for SVM and NN, least for RF.







Post Implementation: What's next?

- Continued refinement of the model

Phase I: Pre-implementation Data Handling

Step	Considerations
1: Establish the relevance of ML for the focal construct	▪ Will the application of ML techniques improve measurement relative to simpler techniques?
2: Establish the "ground truth" of the training data	▪ Does the data source align with the construct? ▪ Has the training data been validated? ▪ Is there sufficient variance in the training data? ▪ Is the training sample sufficiently large to designate a holdout sample?
3: Apply theoretical pre-pruning	▪ What available features are conceptually linked to the construct?

Phase II: ML Implementation

Step	Considerations
4: Select appropriate ML techniques for the focal construct	▪ Is the construct binary, categorical, or continuous? ▪ Which ML algorithm most accurately predicts the construct? ▪ What hyperparameters optimize the accuracy of the model?

Phase III: Post-implementation Reporting and Refinement

Questions