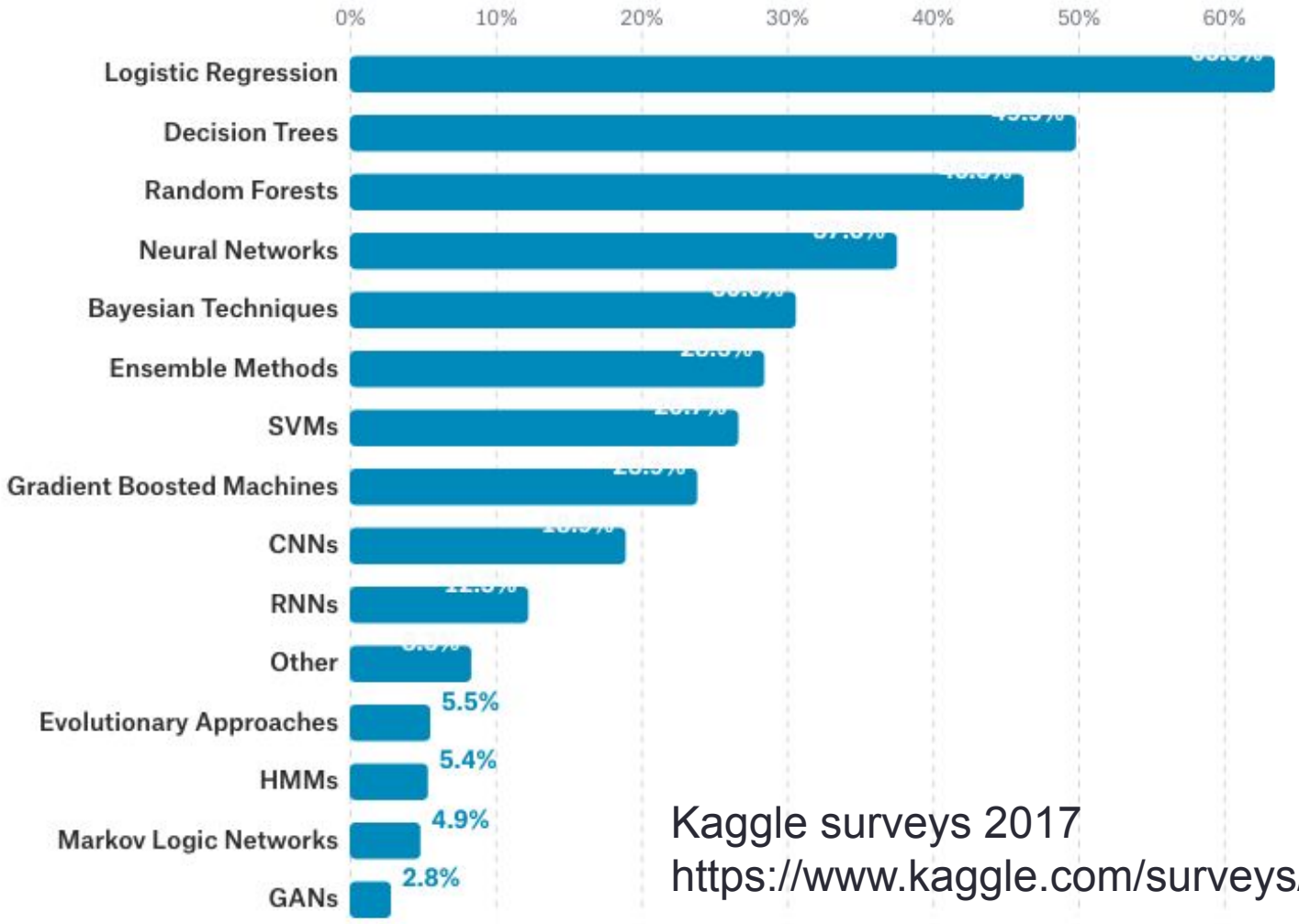
TRICKS OF THE TRADE:

Machine learning in the real world

Burning questions

- Which model to use?
- How should I improve my models?
  - Diagnosis and error analysis
- What does a data scientist do at work?

Which model?



Kaggle surveys 2017
https://www.kaggle.com/surveys/2017

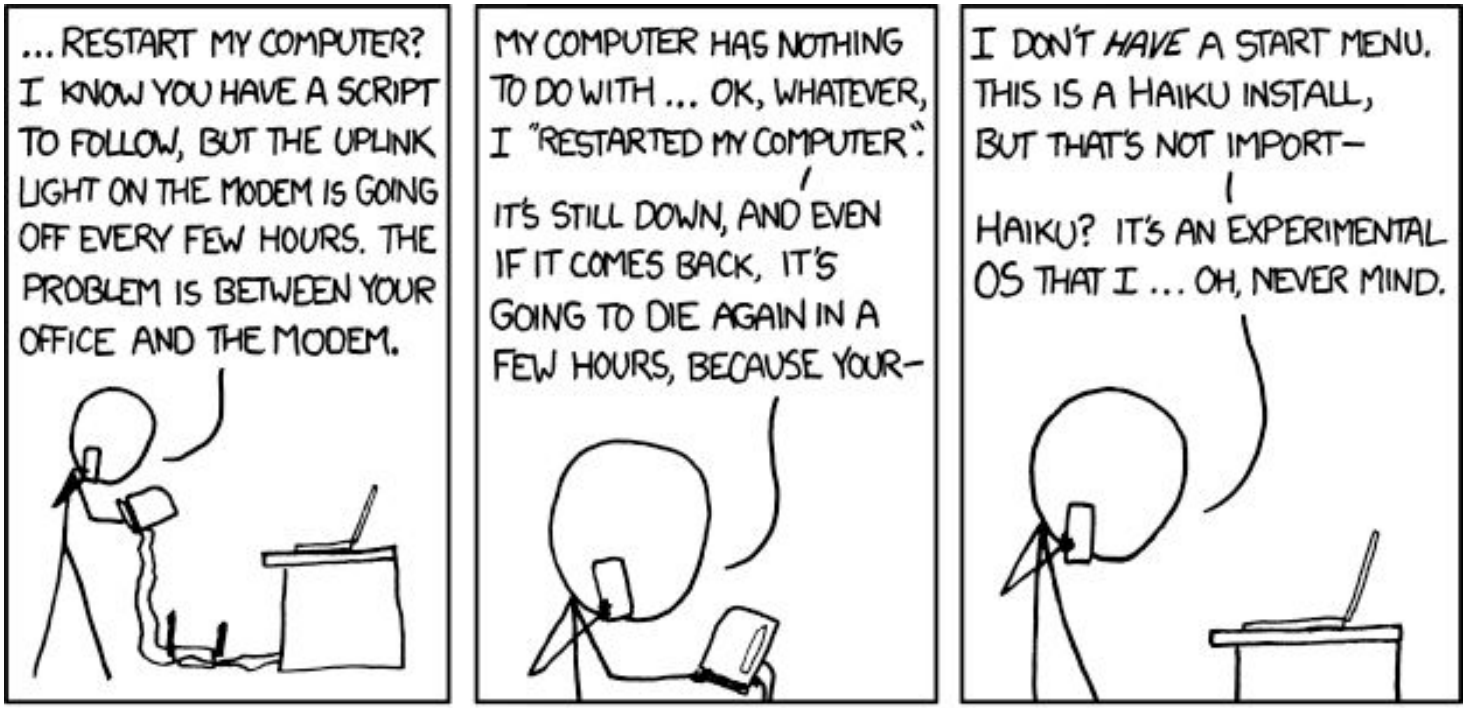
Cautionary notes

- “There is no free lunch.”
- The **“No Free Lunch”** theorem states that there is no one model that works best for every problem.
- Depends on
  - Nature of the task
  - Nature of the data
  - Amount of data
- Which model is the best?
  - Try it on your problem.

“Deep learning is not magical.”

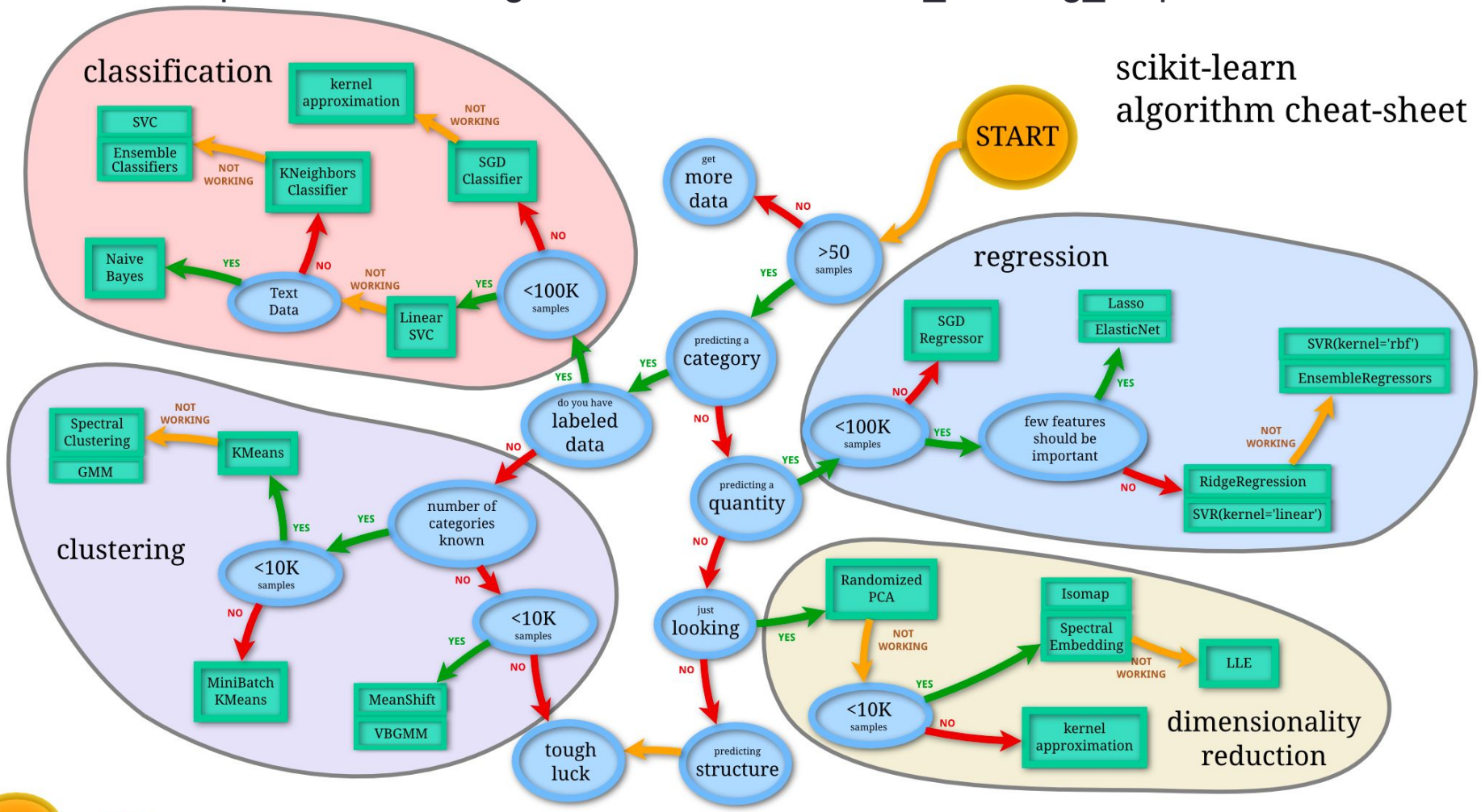
That's not so helpful…

Tech Support



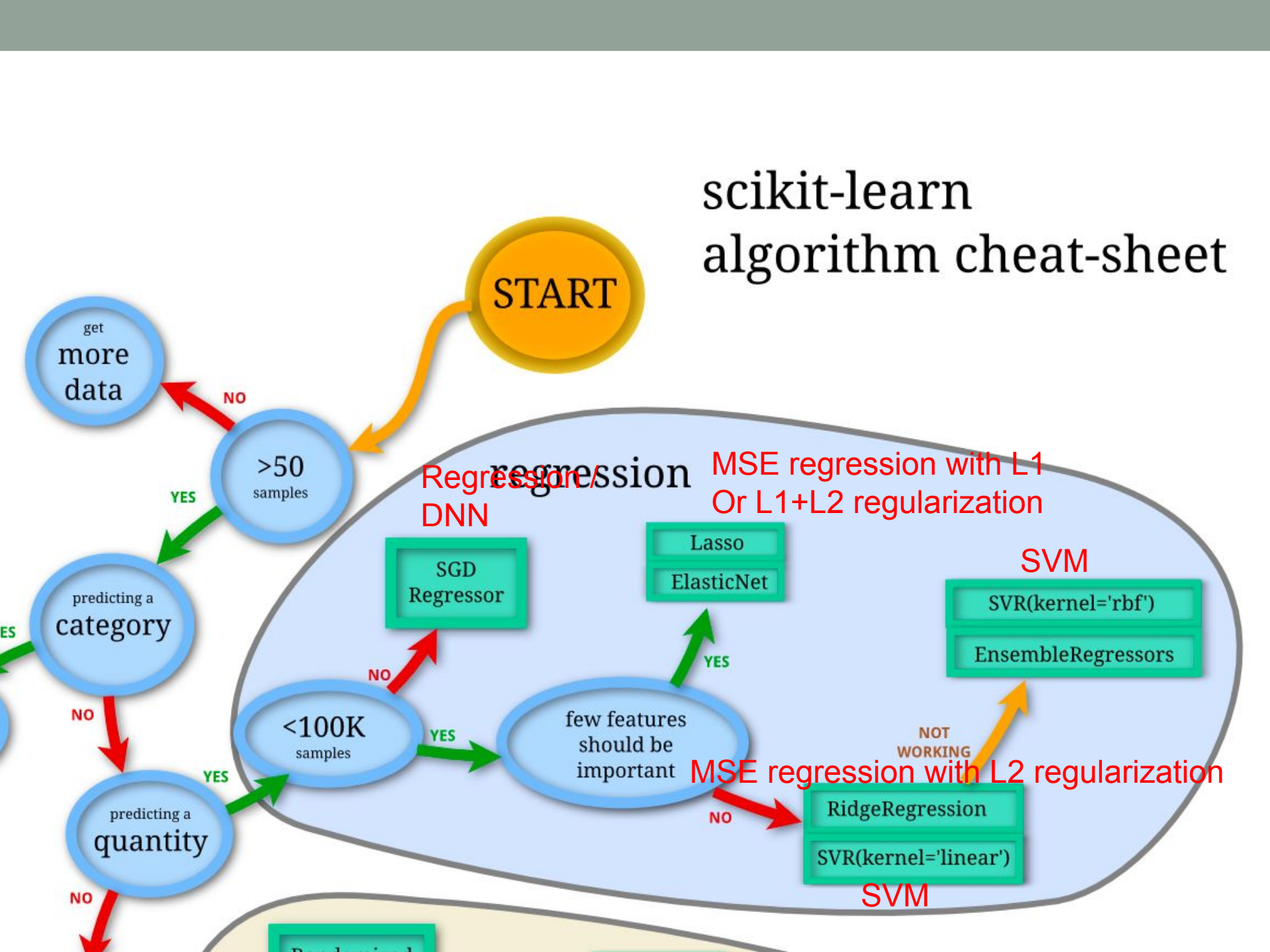
https://xkcd.com/806/

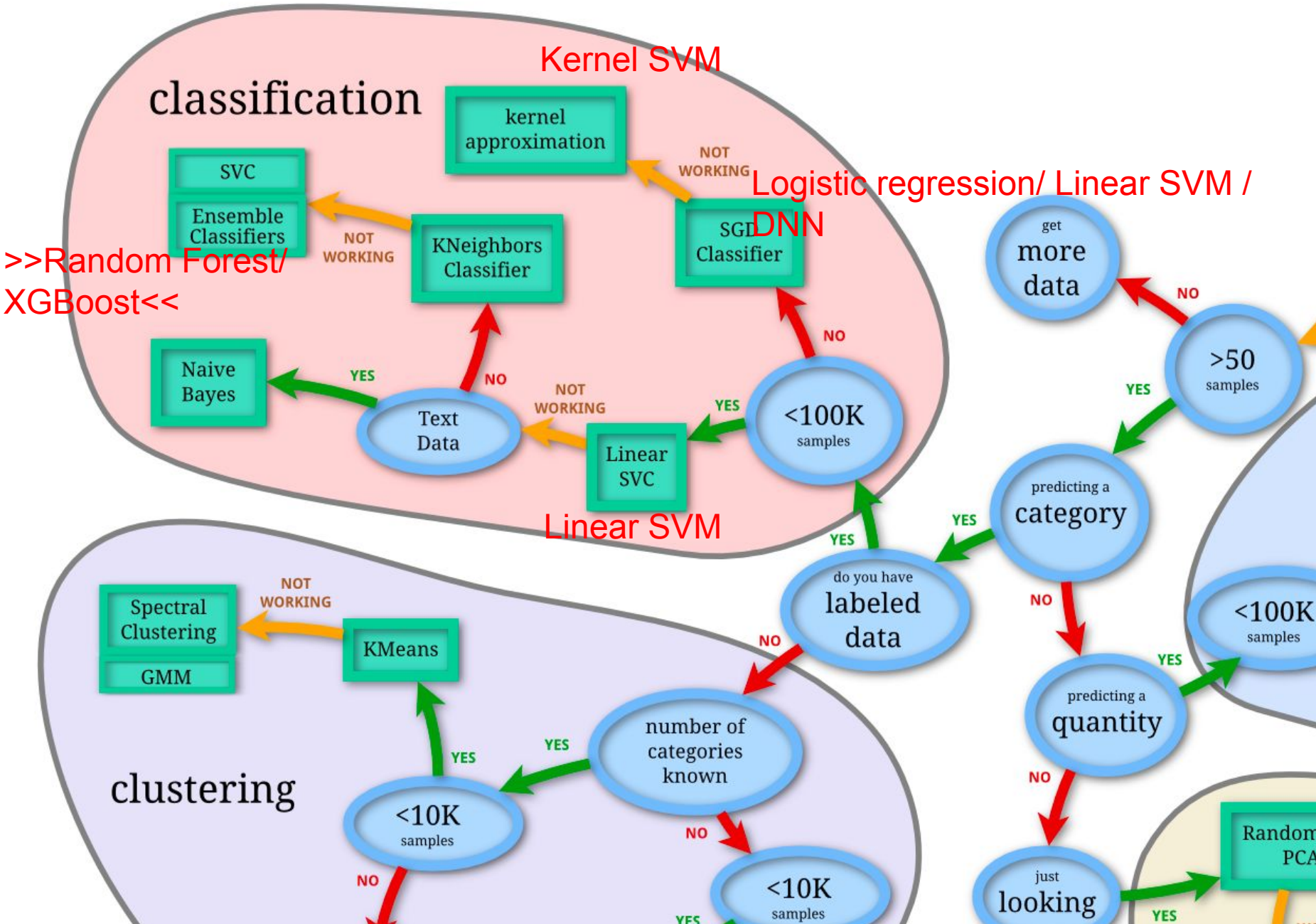
http://scikit-learn.org/stable/tutorial/machine_learning_map/index.html

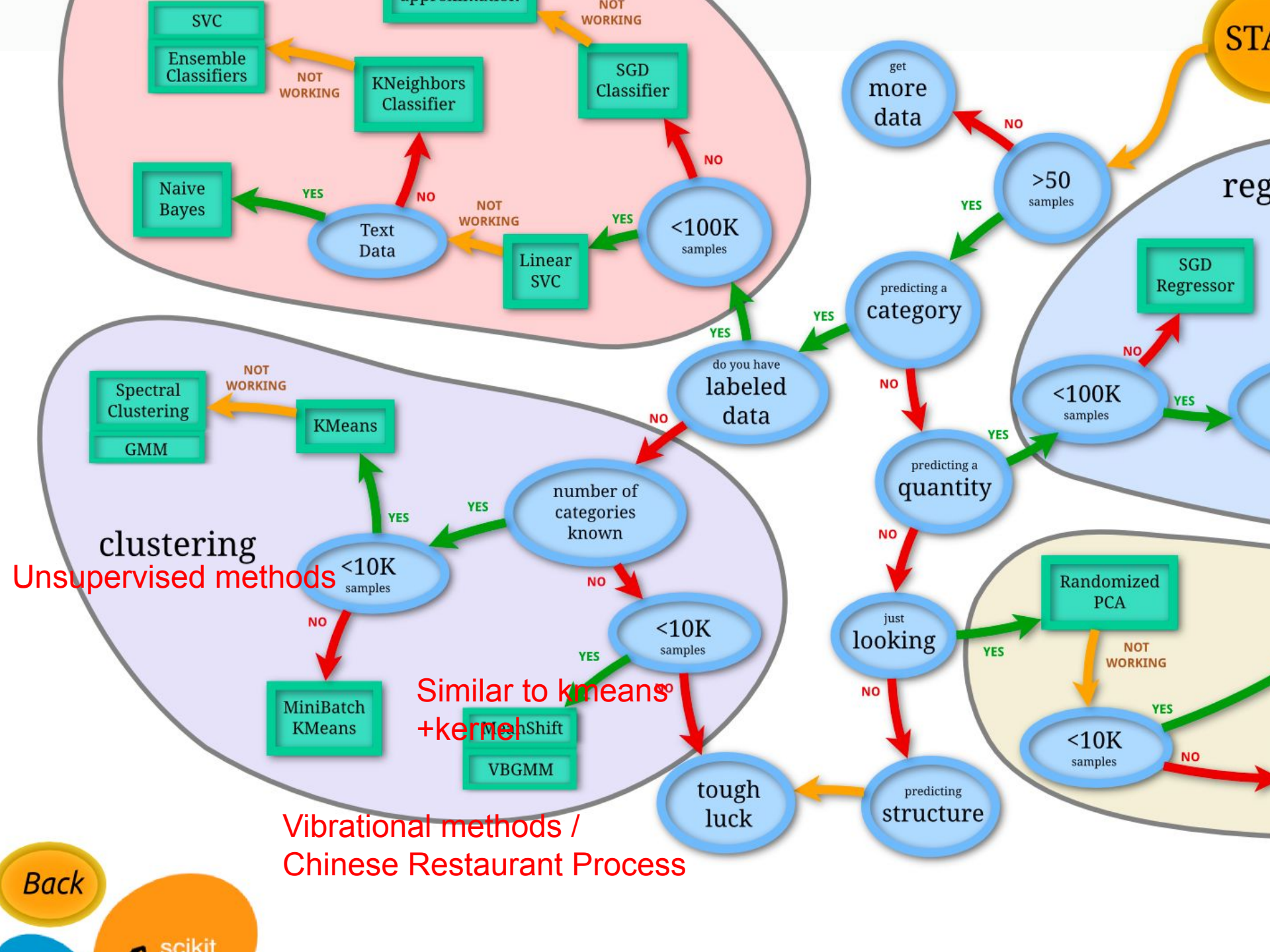


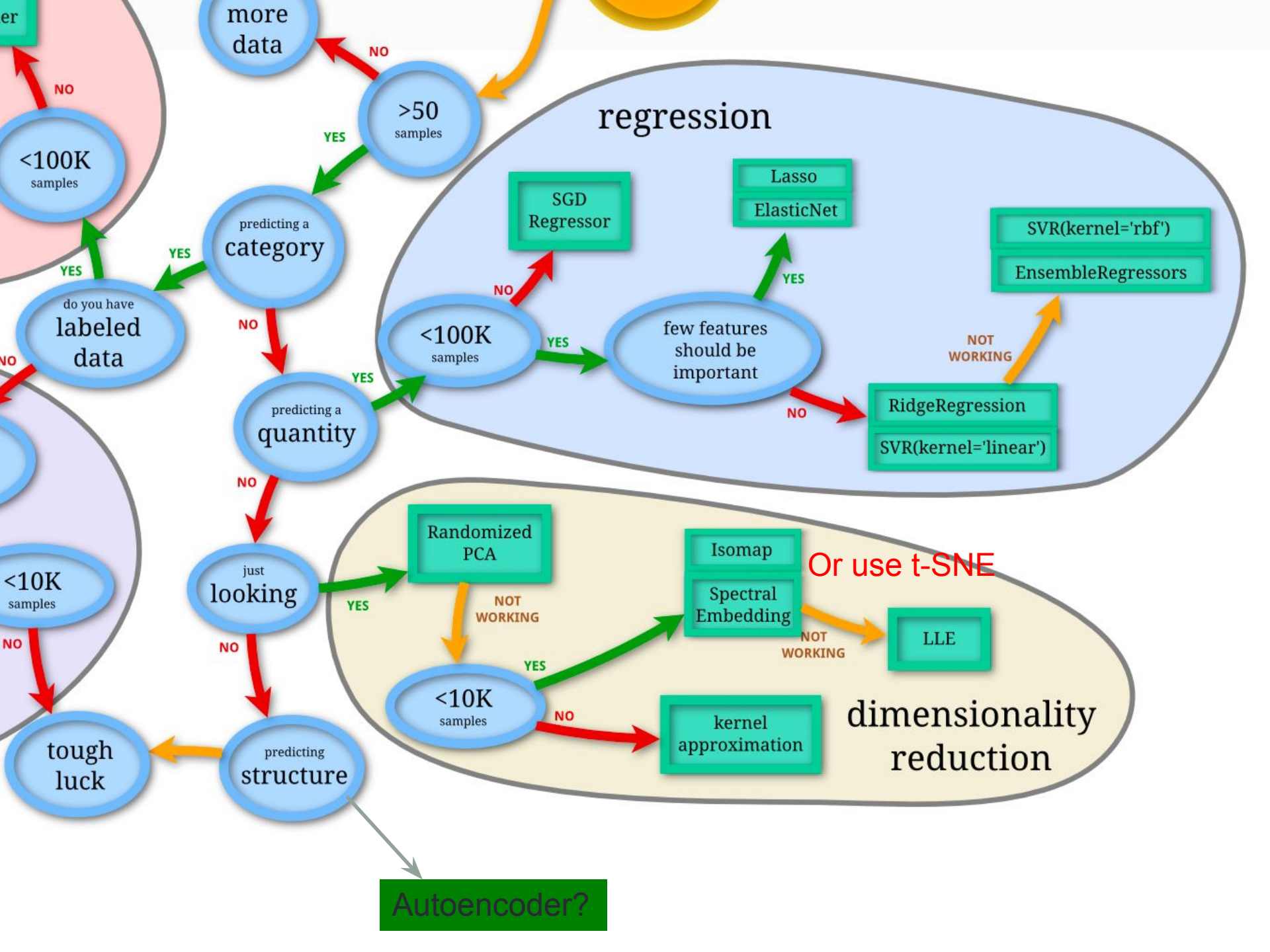
Back

Note: treat 100k, 10k samples as a guideline.
These numbers can go bigger or smaller depending on feature dimension and number of classes.


scikit-learn
algorithm cheat-sheet
START
get more data
NO
>50 samples
YES
predicting a category
NO
predicting a quantity
YES
NO
<100K samples
NO
YES
SGD Regressor
few features should be important
YES
NO
Lasso
ElasticNet
NOT WORKING
RidgeRegression
SVR(kernel='linear')
SVR(kernel='rbf')
EnsembleRegressors
regression
Regression /
DNN
MSE regression with L1
Or L1+L2 regularization
SVM
MSE regression with L2 regularization
SVM


Kernel SVM
classification
kernel approximation
NOT WORKING
SVC
Ensemble Classifiers
NOT WORKING
KNeighbors Classifier
SGD Classifier
Logistic regression/ Linear SVM / DNN
>>Random Forest/ XGBoost<<
NO
Naive Bayes
YES
NO
NOT WORKING
YES
Text Data
NOT WORKING
Linear SVC
<100K samples
Linear SVM
get more data
NO
>50 samples
YES
predicting a category
YES
YES
do you have labeled data
NO
NO
<100K samples
YES
predicting a quantity
NO
Spectral Clustering
NOT WORKING
GMM
KMeans
YES
YES
number of categories known
clustering
<10K samples
NO
NO
<10K samples
just looking
YES
Random
PCA
YES


SVC
Ensemble Classifiers
NOT WORKING
KNeighbors Classifier
NOT WORKING
SGD Classifier
NO
Naive Bayes
YES
Text Data
NO
NOT WORKING
Linear SVC
YES
<100K samples
NO
get more data
NO
>50 samples
YES
predicting a category
YES
do you have labeled data
YES
NO
Spectral Clustering
GMM
NOT WORKING
KMeans
YES
number of categories known
YES
NO
clustering
Unsupervised methods
<10K samples
NO
MiniBatch KMeans
<10K samples
YES
NO
MeanShift
VBGMM
Similar to kmeans +kernel
tough luck
predicting structure
NO
just looking
NO
predicting a quantity
YES
NO
<100K samples
YES
NO
SGD Regressor
Randomized PCA
YES
NOT WORKING
<10K samples
YES
NO
Vibrational methods / Chinese Restaurant Process
Back
scikit


more data
NO
>50 samples
YES
<100K samples
NO
predicting a category
YES
do you have labeled data
YES
NO
predicting a quantity
YES
NO
just looking
NO
predicting structure
tough luck
<10K samples
NO
regression
SGD Regressor
NO
<100K samples
YES
few features should be important
YES
Lasso
ElasticNet
NO
RidgeRegression
SVR(kernel='linear')
NOT WORKING
SVR(kernel='rbf')
EnsembleRegressors
Randomized PCA
YES
NOT WORKING
<10K samples
YES
Isomap
Spectral Embedding
NOT WORKING
LLE
NO
kernel approximation
dimensionality reduction
Or use t-SNE
Autoencoder?

"Do we Need Hundreds of Classifiers to Solve Real World Classification Problems?"

We evaluate **179 classifiers** arising from **17 families** (discriminant analysis, Bayesian, neural networks, support vector machines, decision trees, rule-based classifiers, boosting, bagging, stacking, random forests and other ensembles, generalized linear models, nearest-neighbors, partial least squares and principal component regression, logistic and multinomial regression, multiple adaptive regression splines and other methods), implemented in Weka, R (with and without the caret package), C and Matlab, including all the relevant classifiers available today. We use **121 data sets**, which represent **the whole UCI** data base (excluding the large-scale problems) and other own real problems, in order to achieve significant conclusions about the classifier behavior, not dependent on the data set collection. **The classifiers most likely to be the bests are the random forest** (RF) versions, the best of which (implemented in R and accessed via caret) achieves 94.1% of the maximum accuracy overcoming 90% in the 84.3% of the data sets. However, the difference is not statistically significant with the second best, the SVM with Gaussian kernel implemented in C using LibSVM, which achieves 92.3% of the maximum accuracy. A few models are clearly better than the remaining ones: random forest, SVM with Gaussian and polynomial kernels, extreme learning machine with Gaussian kernel, C5.0 and avNNet (a committee of multi-layer perceptrons implemented in R with the caret package). The random forest is clearly the best family of classifiers (3 out of 5 bests classifiers are RF), followed by SVM (4 classifiers in the top-10), neural networks and boosting ensembles (5 and 3 members in the top-20, respectively).

Most data sets are small (<1000) in this paper

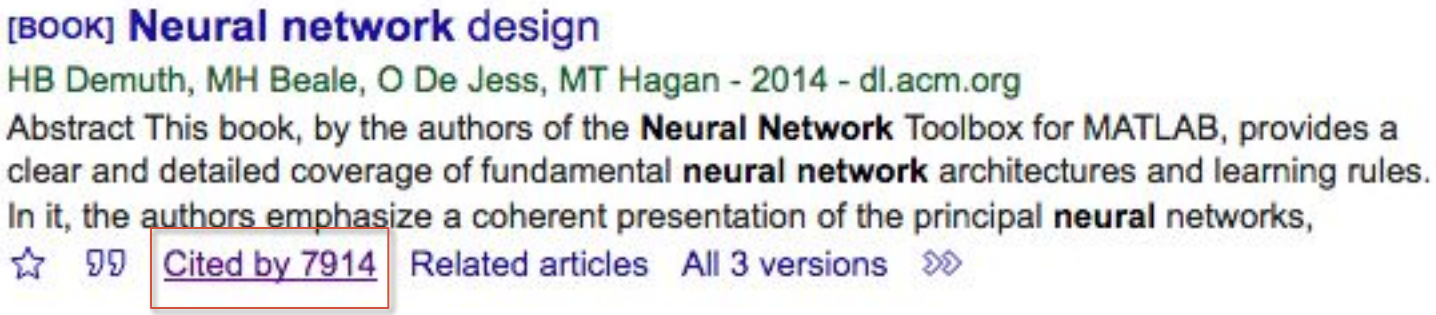
http://jmlr.org/papers/volume15/delgado14a/delgado14a.pdf

Do your literature review

- Check for the closest task
  - Does not need to be in the same domain
    - Ex: Strings of DNA -> NLP!
    - Most advances in many fields came from cross-fertilization
- Reading checklist
  - Amount of data
    - Number of classes
    - Data/classes
  - Features
  - Models
- Look at multiple papers

Literature review tricks

- Search forward and backward
  - Citations
  - Cited by



Reading between the lines

- If you don't understand something, re-read
  - Average of 5+ times to understand a paper completely
- Print it out, keep a pen/pencil at hand, and break down equations
- Try to explain it with what you already know

Elastic net Loss function $\min_w \frac{1}{2n_{samples}}||Xw - y||_2^2 + \alpha\rho||w||_1 + \frac{\alpha(1-\rho)}{2}||w||_2^2$

Reading between the lines

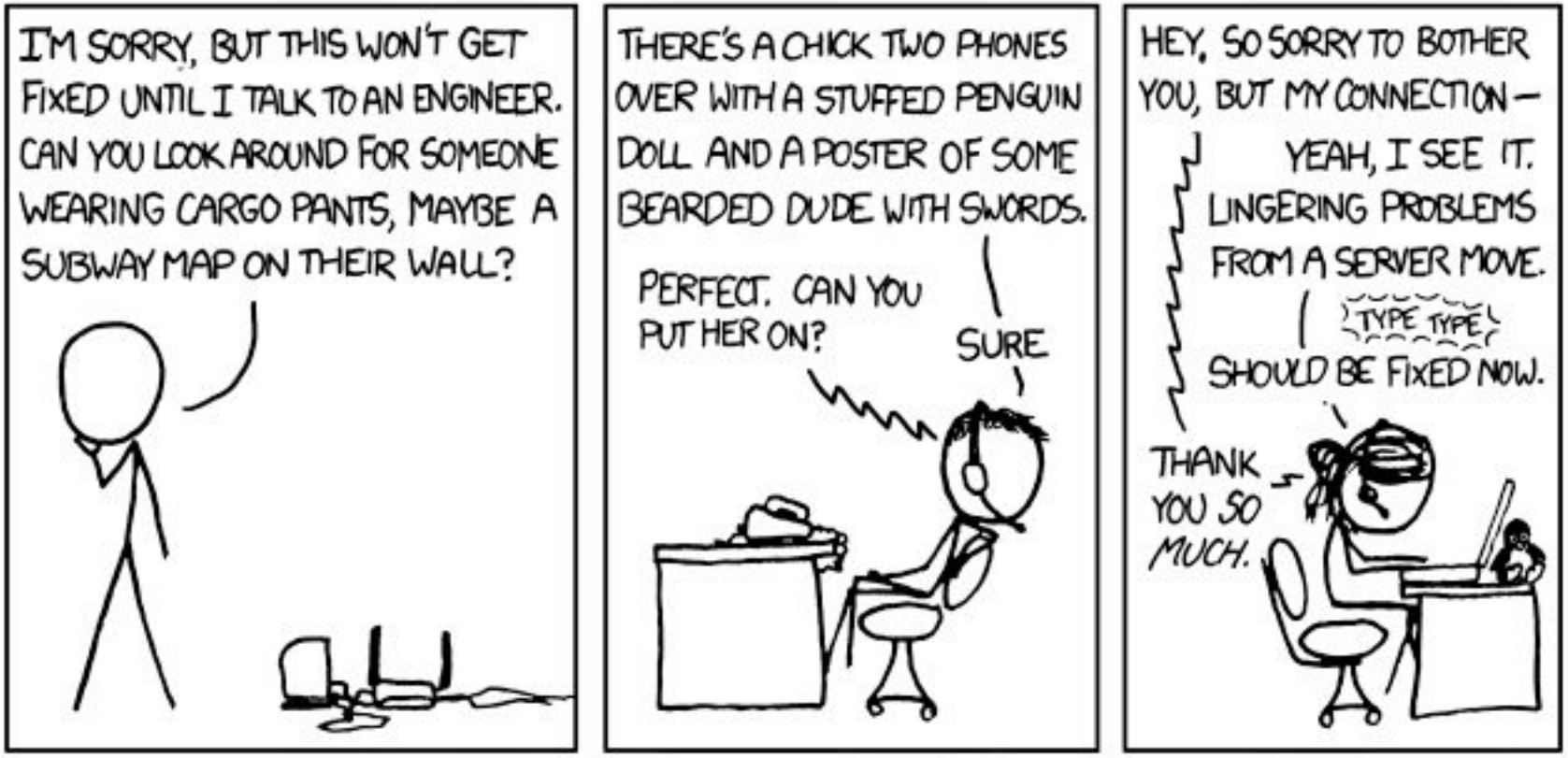
- If you don't understand something, re-read
  - Average of 5+ times to understand a paper completely
- Print it out, keep a pen/pencil at hand, and break down equations
- Try to explain it with what you already know

Elastic net Loss function $\min_w \frac{1}{2n_{samples}}||Xw - y||_2^2 + \alpha\rho||w||_1 + \frac{\alpha(1-\rho)}{2}||w||_2^2$

MSE Loss L1 L2

How to improve my results?

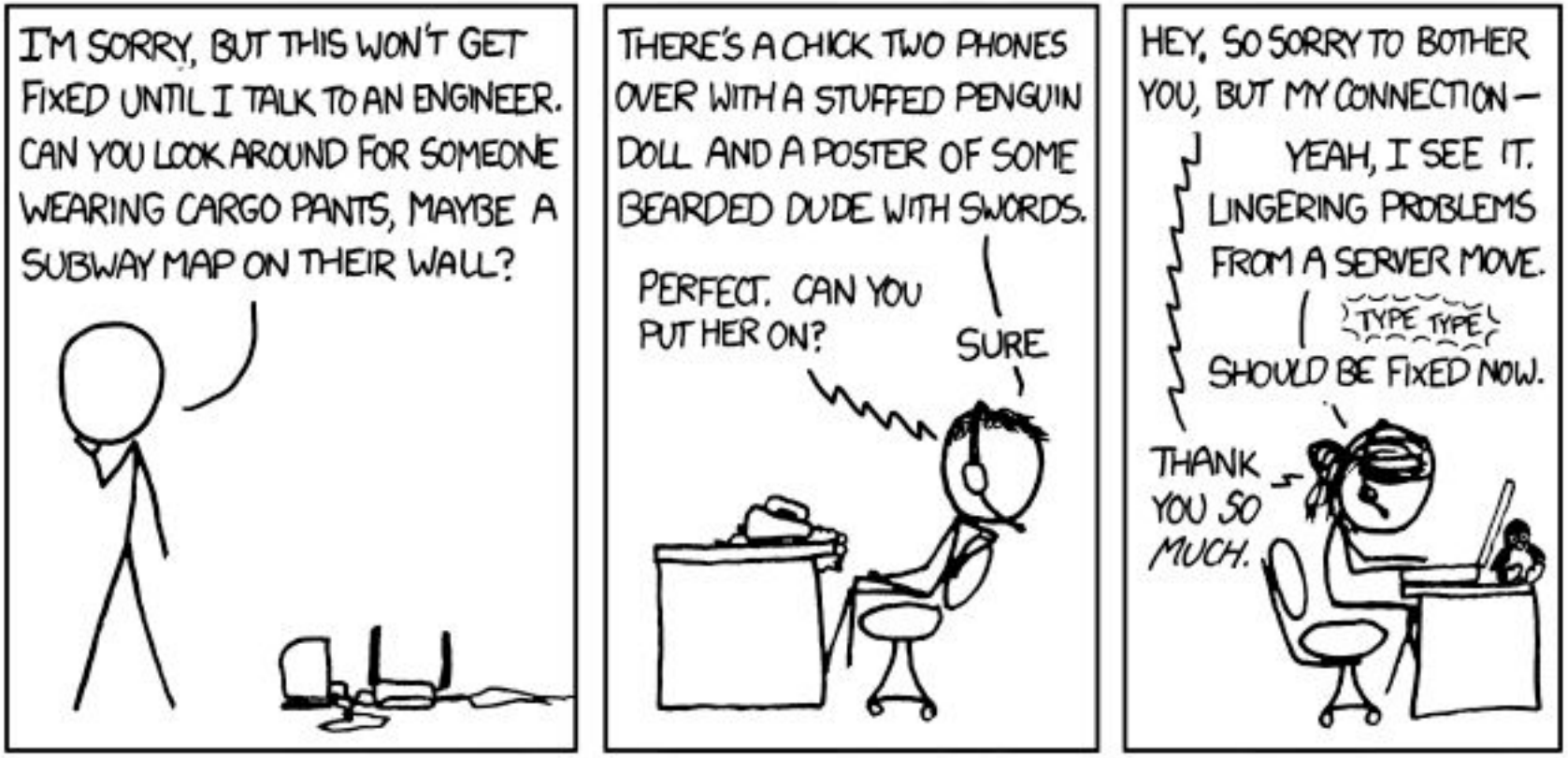
- More tech support



https://xkcd.com/806/

How to improve my results?

- More tech support



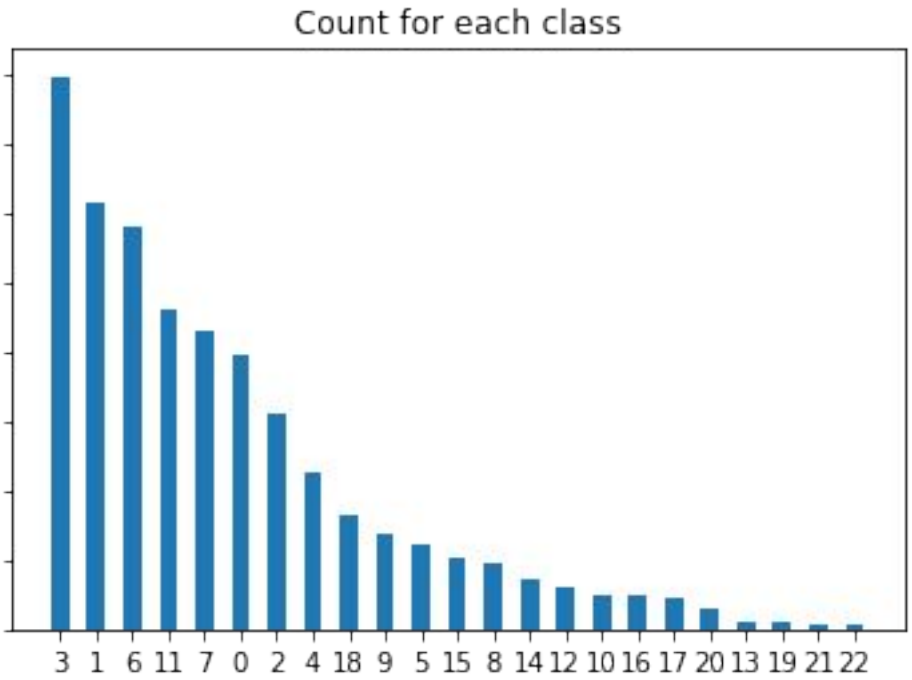
https://xkcd.com/806/

Diagnosis and prescription

- Is my task proper?
- Should I add/remove feature XXX?
- Is my loss function appropriate?
- Is my model overfitting/underfitting?
- Should I try XXX?

Is my task proper? Do you suffer from class imbalance?

- Throwing away
- Refactoring
  - Split
  - Merge
- Data augmentation
- Biasing
  - Weighting the loss function
  - Bias in mini-batch sampling



Diagnosis

- Is my task proper?
- Should I add/remove feature XXX?
- Is my loss function appropriate?
- Is my model overfitting/underfitting?
- Should I try XXX?

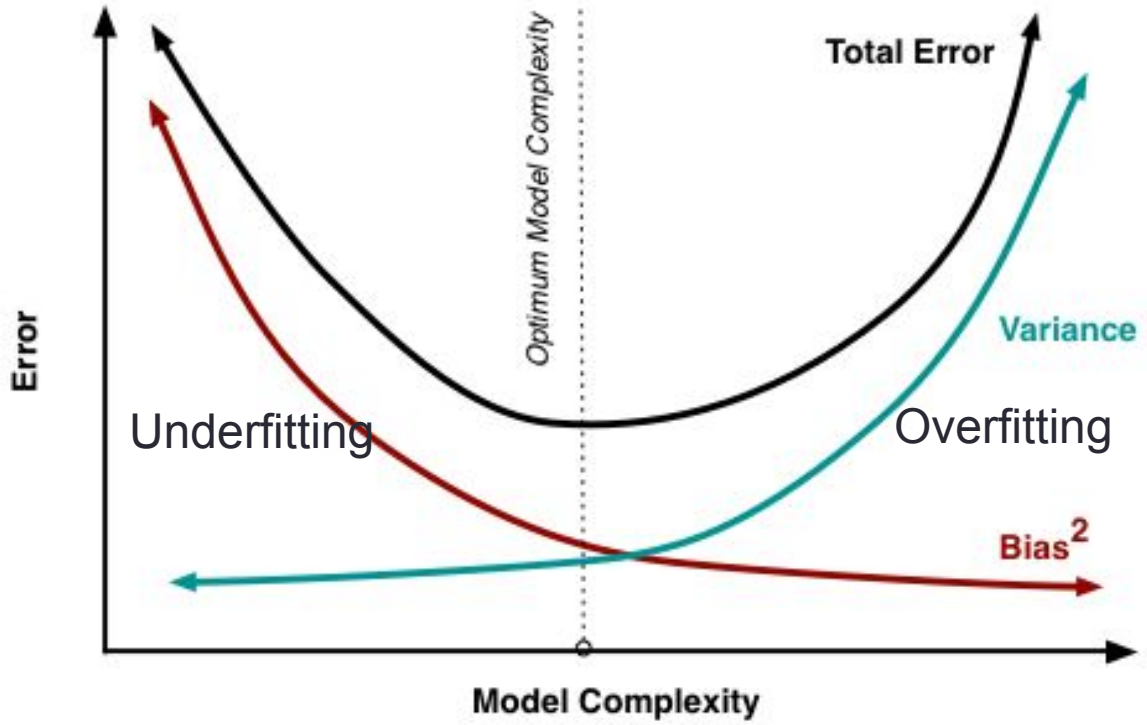
Short answer: Cross Validation

Can we do better?
Find the problem and fix it.

Understanding the bias variance trade-off

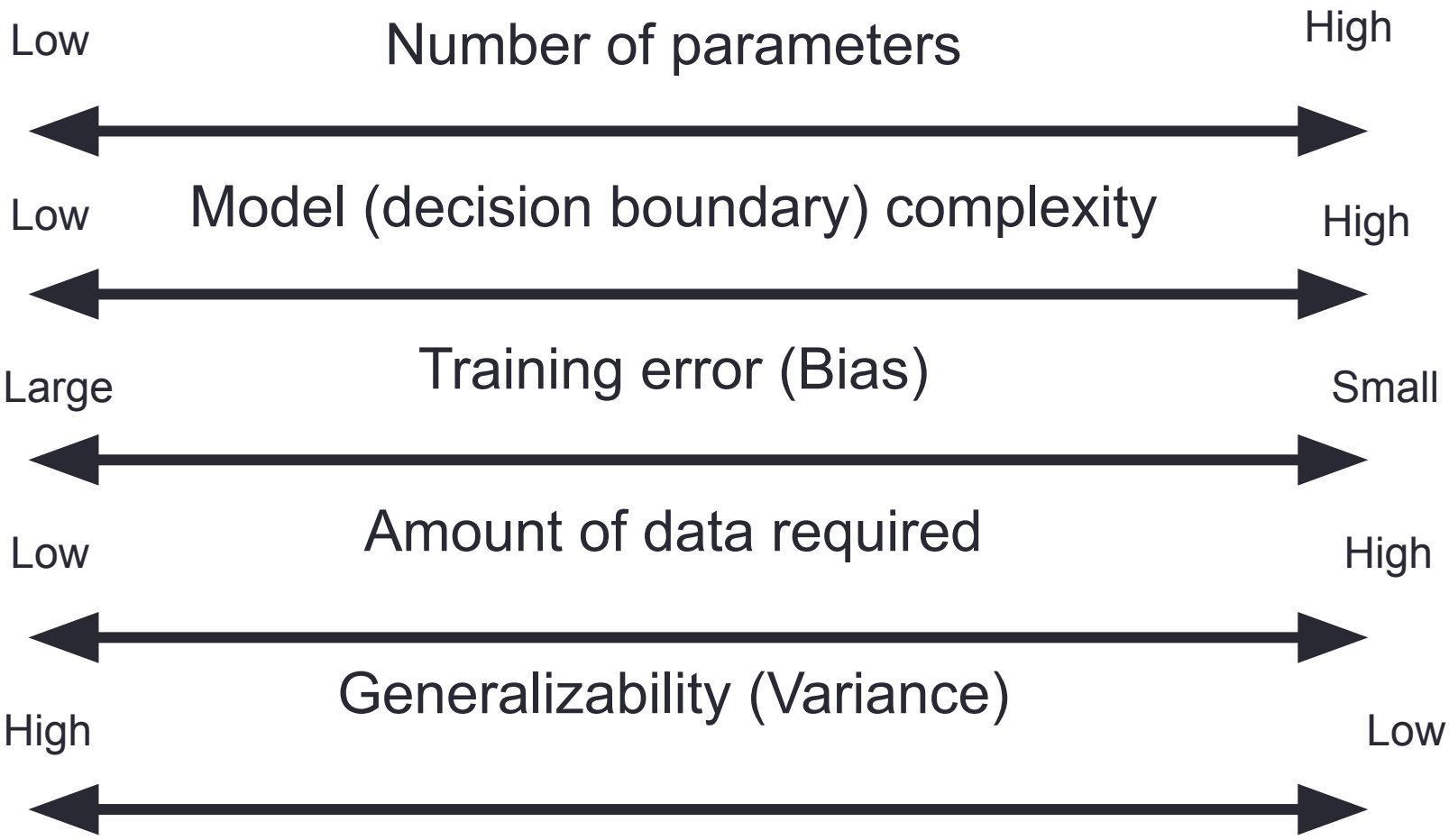
- Bias variance analysis can be helpful for diagnosis

$$\underbrace{E_{\mathbf{x},y,D}\left[(h_D(\mathbf{x})-y)^2\right]}_{\text{Expected Test Error}} = \underbrace{E_{\mathbf{x},D}\left[(h_D(\mathbf{x})-\bar{h}(\mathbf{x}))^2\right]}_{\text{Variance}} + \underbrace{E_{\mathbf{x},y}\left[(\bar{y}(\mathbf{x})-y)^2\right]}_{\text{Noise}} + \underbrace{E_{\mathbf{x}}\left[(\bar{h}(\mathbf{x})-\bar{y}(\mathbf{x}))^2\right]}_{\text{Bias}^2}$$



http://www.cs.cornell.edu/courses/cs4780/2017sp/lectures/lecturenote11.html

Bias-Variance overview



Bias-variance diagnosis

- Suppose the problem is either
  - Overfitting (high variance)
  - Too few features to classify properly (high bias)

- Symptoms
  - Variance: Training error is much lower than test/dev error.
  - Bias: Training error is also high

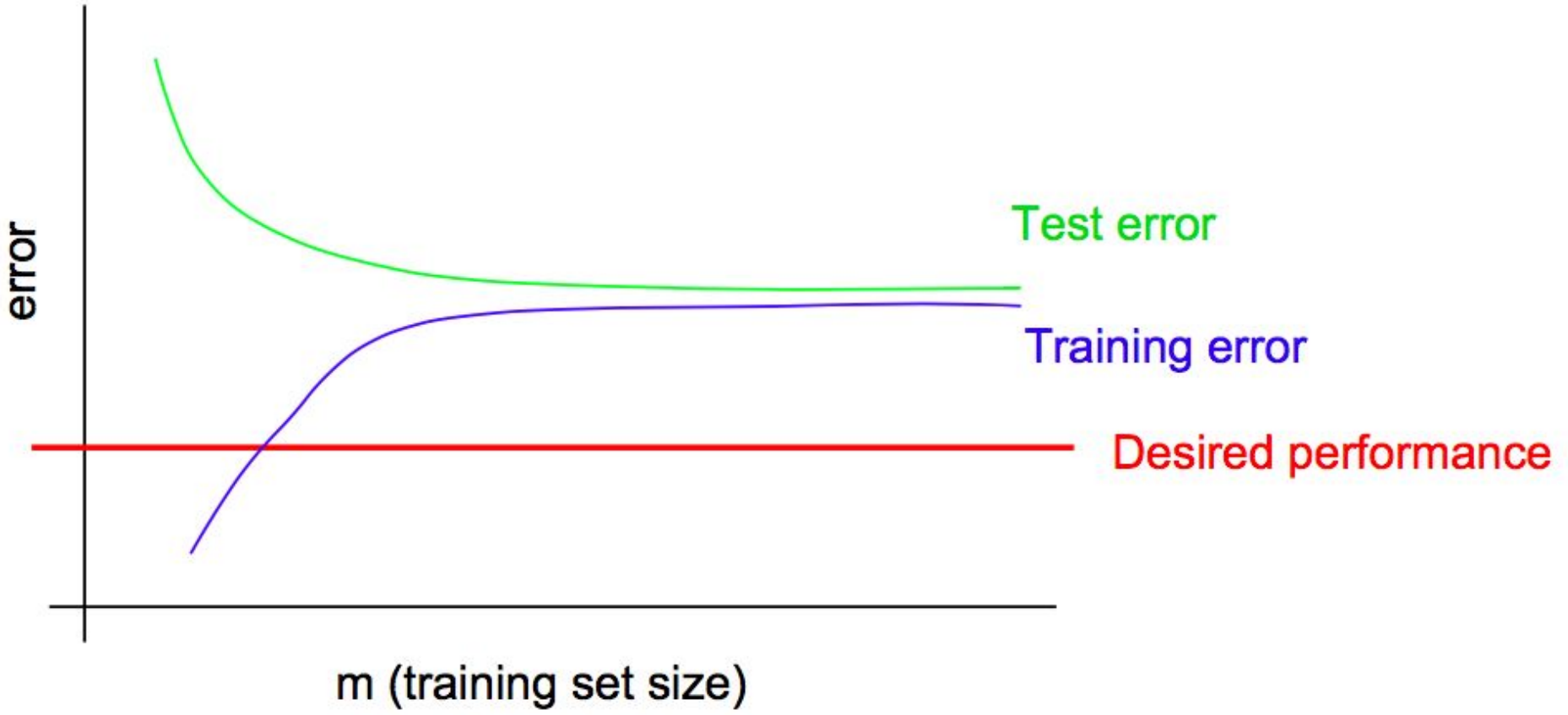
High variance case

- Solution:
  - Reduce overfitting: regularization, reduce features, etc
  - Get more training set

cs229.stanford.edu/materials/ML-advice.pdf

High bias case

- Solution: more features, more layers, bigger models



cs229.stanford.edu/materials/ML-advice.pdf

Desired performance?

- The goal of the application
  - Possibilities
    - Target performance to launch the product
    - Human performance on the task
      - Have A and B do the same task, measure the difference.
- Knowing human performance gives several advantage
  - Knows when to stop
    - Beating human is sometimes a goal, but
    - Some errors might be labeling errors or judgment calls

More diagnosis example

- Flying a drone using RL
- You're poor and can't repeat CMU's drone crashing experiments

- What you did:
- Make a simulation
- Define a reward function R
- Learn the policy, p, using some RL algorithm.
  - Maximize $R(p_{drone})$

Drone diagnosis example

- In real testing, your drone crashes and burns because your policy is bad.

- Diagnosis:
  - If the drone fly well in simulation but not in real life testing, you simulation is bad.
  - Let be p_{rule} be some rule-base control policy developed by drone flying expert. If $R(p_{drone}) < R(p_{human})$, RL algorithm fails to maximize the rewards. Fix the RL algorithm
  - If $R(p_{drone}) > R(p_{human})$, the RL maximization is doing its job properly. Fix the reward function.

Error diagnosis

- You're making a cat classifier. (cat/not cat)
- It sucks.
- You heard of this super new hype algorithm (for example: capsule network). Should you spend months to try it out?

This is not a cat

https://www.vat19.com/item/not-a-cat-cat-the-cat-that-isnt

Looking at the errors

- Spend an hour or two looking at your errors. Identify why. Keep a table.

	Blurred	Weird angle	Notes
Pic1	x		Stuffed toy
Pic2	x		
Pic3	x		
Pic4		x	Top view
…	…	…	…
	68%	2%	

Solution: Use a method to sharpen the image. Train on blurry images.

The table categories can expand as you look through more pictures and see frequently occurring error cases. So keep notes.

Diagnosis summary

- Simple analysis of the data can help you notice underlying problems
- Bias-variance diagnosis is a general method that can be applied to most tasks
- Other diagnosis depends on the application and need some understanding of the algorithms
- Error analysis can help guide your model improvements

Taking the time to do diagnosis can help you save months of trying random things.

Burning questions

- Which model to use?
- How should I improve my models?
  - Diagnosis and error analysis
- What does a data scientist do at work?

Burning questions

- Which model to use?
- How should I improve my models?
  - Diagnosis and error analysis
- What does a data scientist do at work?
- How to build a machine learning startup?

What does a data scientist do at work?

- Kaggle competition winner
  - http://blog.kaggle.com/2015/12/21/rossmann-store-sales-winners-interview-1st-place-gert/

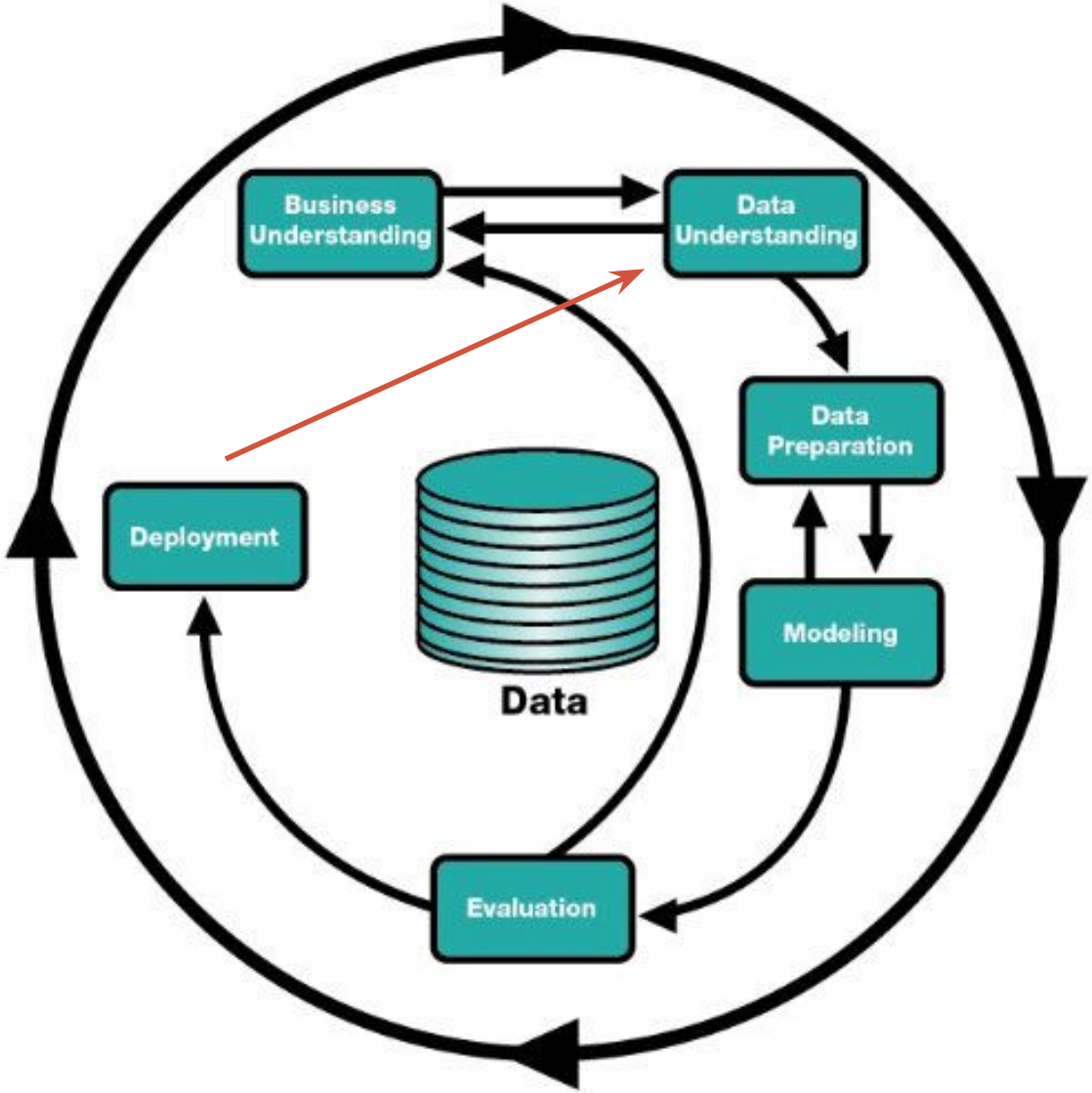
How did you spend your time on this competition?

I spent 50% on feature engineering, 40% on feature selection plus model ensembling, and less than 10% on model selection and tuning.

Note: this is highly automatable (See Datarobot, auto-sklearn, auto-ml, etc.). Will data scientist lose our jobs?

But this is not all of data science

What does a data scientist do at work?



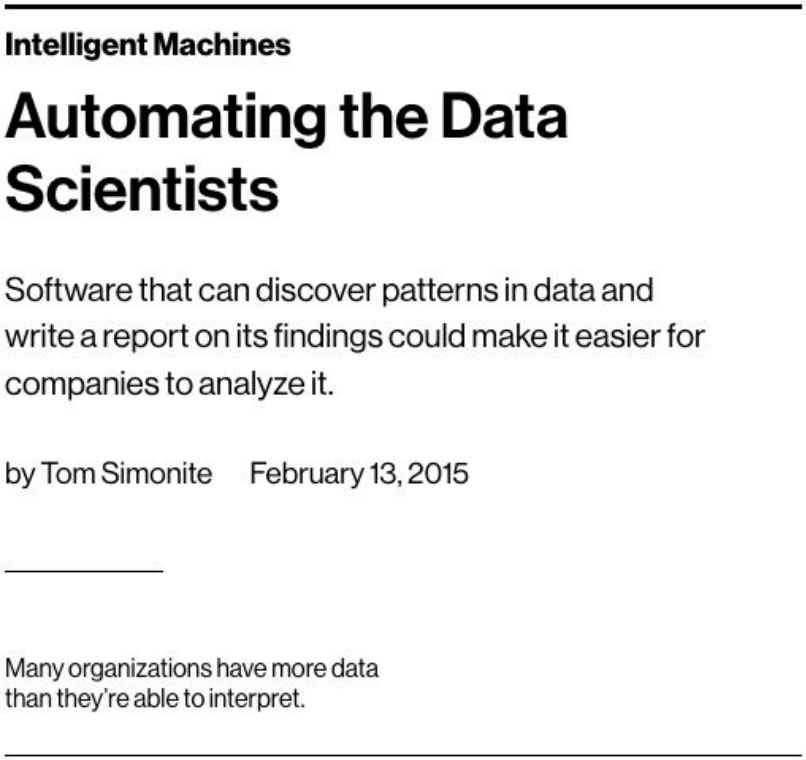
- Data science loop
- In industry:

  "data preparation and modeling (10%) and the rest (90%)"

https://www.kdnuggets.com/2016/08/data-science-automation-debunking-misconceptions.html

Kaggle competition pitfalls and automation

- Given task
- Given (closed-set) of inputs
- Given performance metrics
  - Easy replaced by automation.



Intelligent Machines

Automating the Data Scientists

Software that can discover patterns in data and write a report on its findings could make it easier for companies to analyze it.

by Tom Simonite February 13, 2015

Many organizations have more data than they're able to interpret.

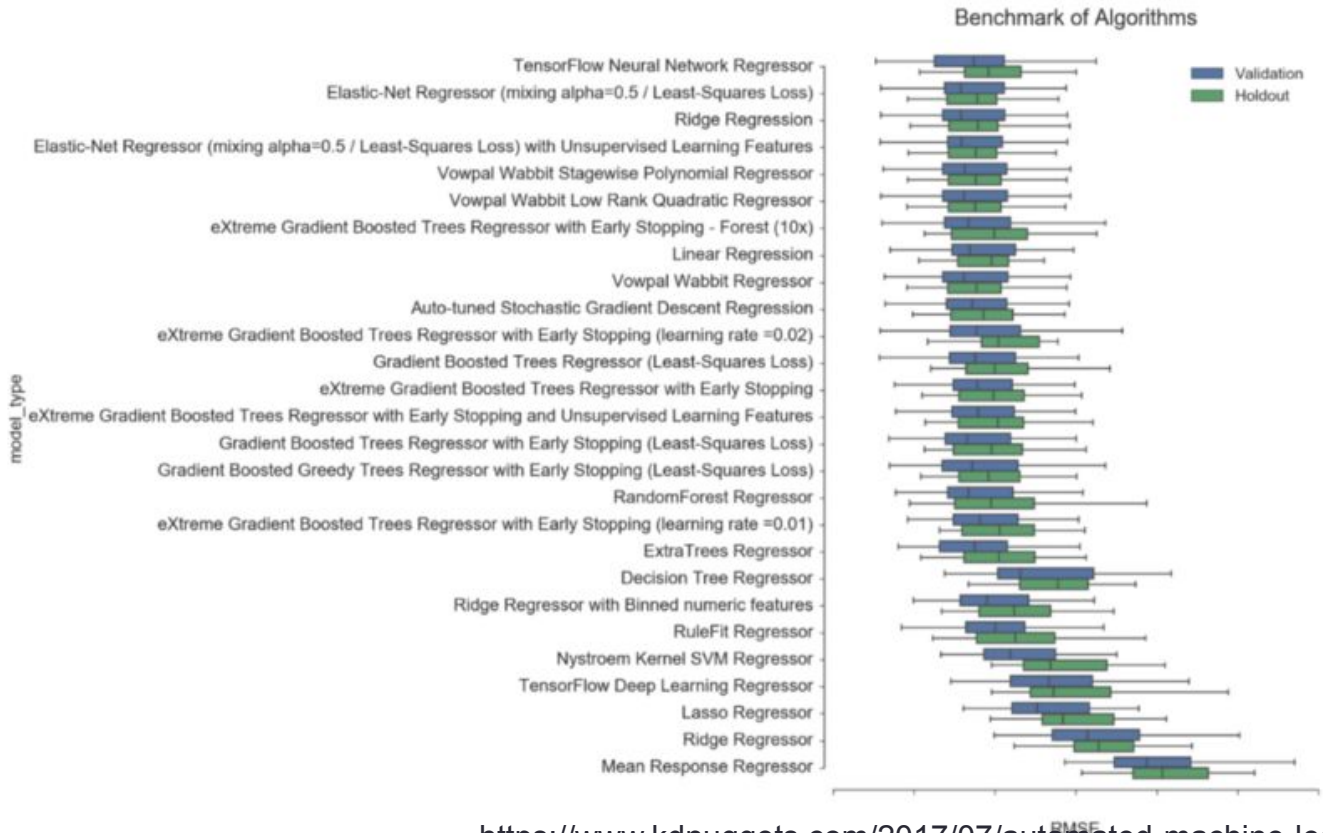
https://www.technologyreview.com/s/535041/automating-the-data-scientists/

The 90%

- In real life, you will have to look for and decide your tasks, your inputs, your metrics.
- Understand what's possible and what's not
  - And be able to describe it to non-data scientists
- Justify business usage and make sure about deployment.

Data scientists + automation tools

- Automate tools can remove boring tasks in data science
  - Give powerful benchmarks
  - Explore the space faster



https://www.kdnuggets.com/2017/07/automated-machine-learning-paradigm-shift.html

Questions beyond classification/regression

Can I trust my model?

Why does it give this answer?

ML products

Customer facing

Recommendation systems, maps (traffic estimate), speech2text

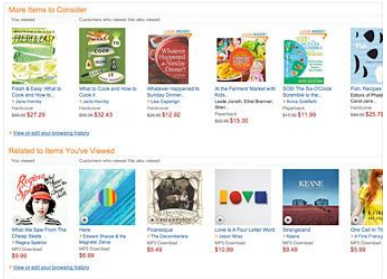
Best guess by the model
Mostly automatic (check deposit by app)

Internal facing

Loan applications, demand forecasting

Can say "I'm not sure"
Human in the loop.
Machine-assisted

Requires confidence level

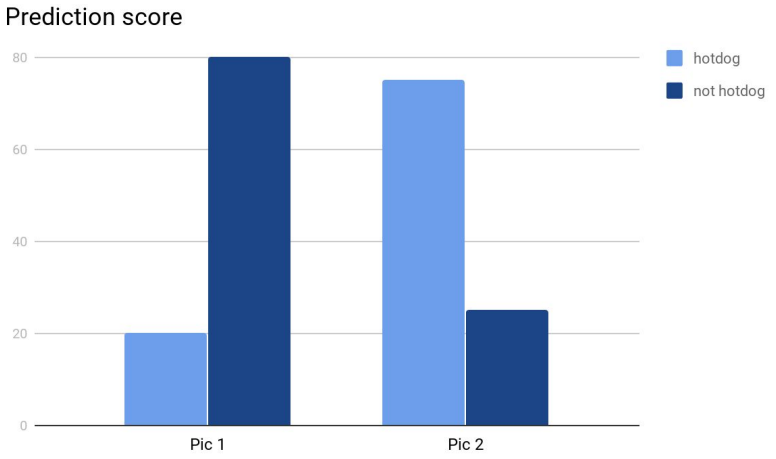


Slides courtesy of Phiradet Bangcharoensap

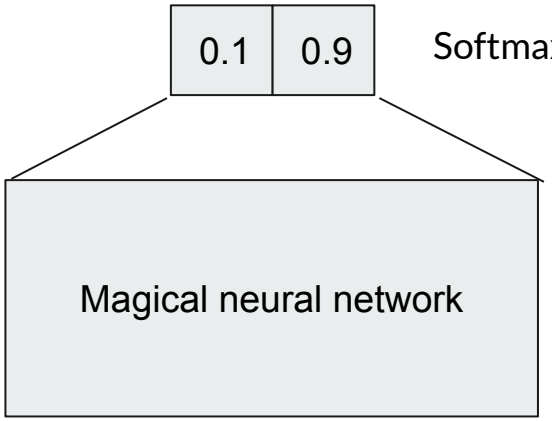
Confidence score

Practical models are not only accurate, but need to be able to state its confidence

Confidence = probability of being correct

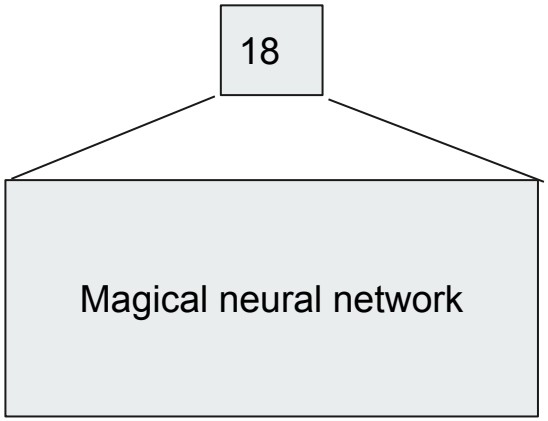


Naive way for confidence



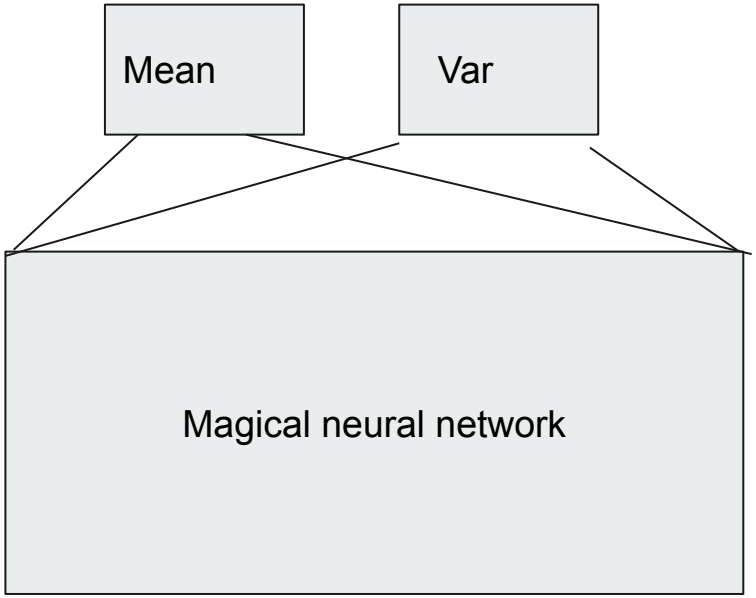
Confidence level = 0.9?

What about regression?



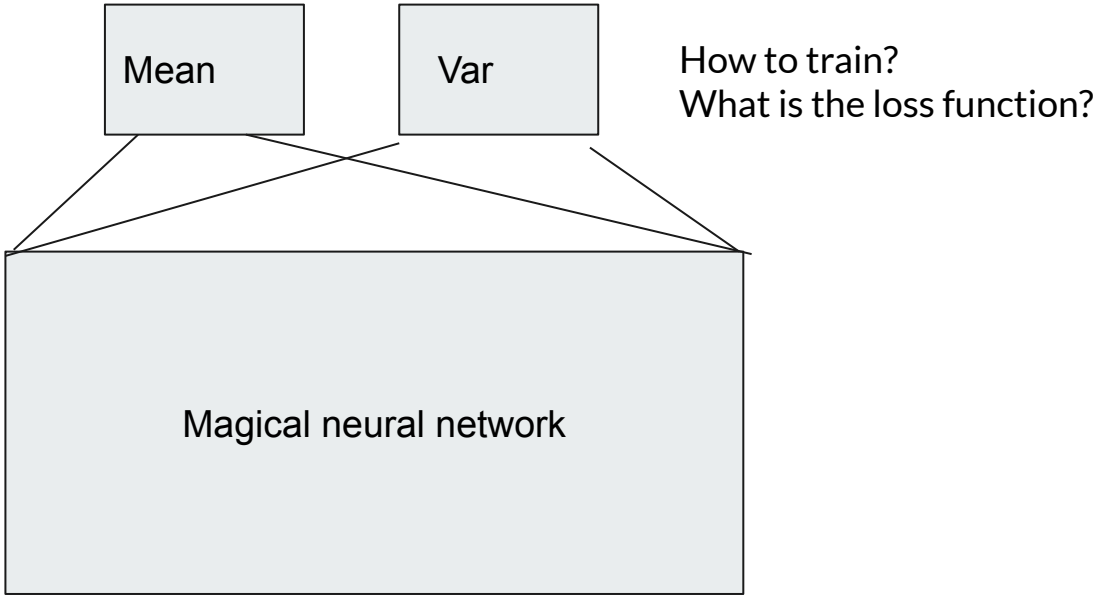
Confidence level = ???

A Naive way for regression (1994!)



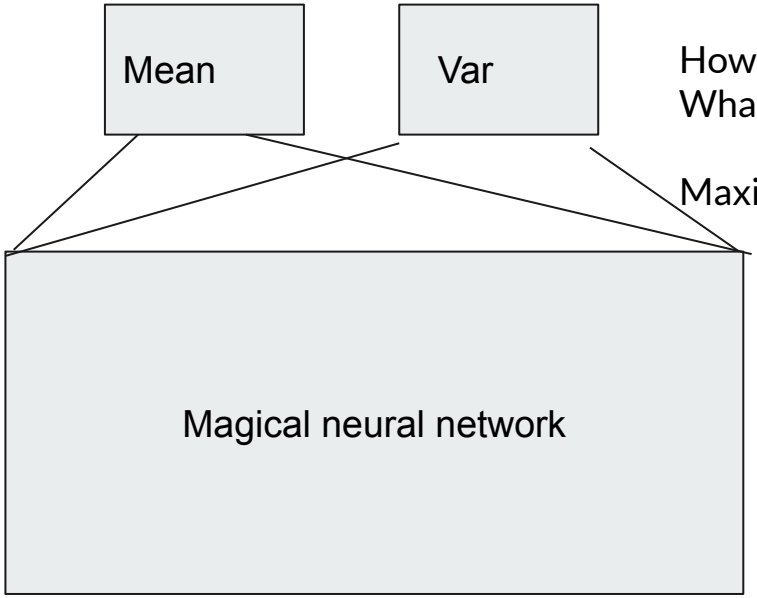
Estimating the mean and variance of the target probability distribution. IEEE 1994

A Naive way for regression (1994!)



Estimating the mean and variance of the target probability distribution. IEEE 1994

A Naive way for regression (1994!)

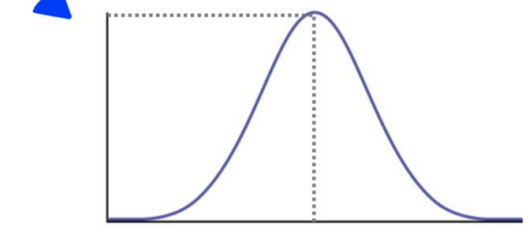


How to train?
What is the loss function?

Maximum likelihood of the training data

Confidence score = $\frac{1}{\sqrt{2\pi\sigma^2}} e^{\frac{-(\mu-\mu)^2}{2\sigma^2}}$ (0)

$= \frac{1}{\sqrt{2\pi\sigma^2}}$



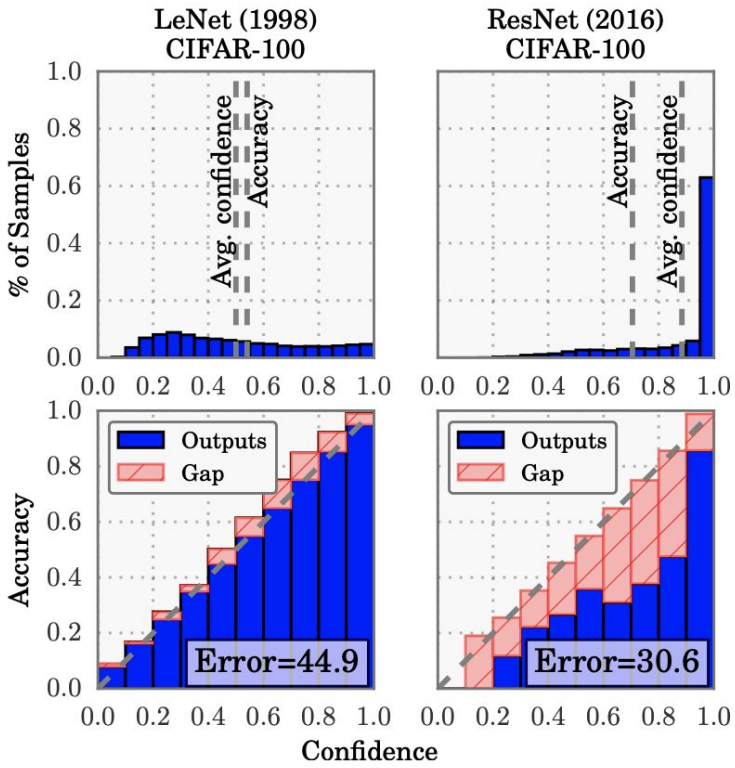
Predicted value = $\mu(x)$

Estimating the mean and variance of the target probability distribution. IEEE 1994

Poorly calibrated confidence

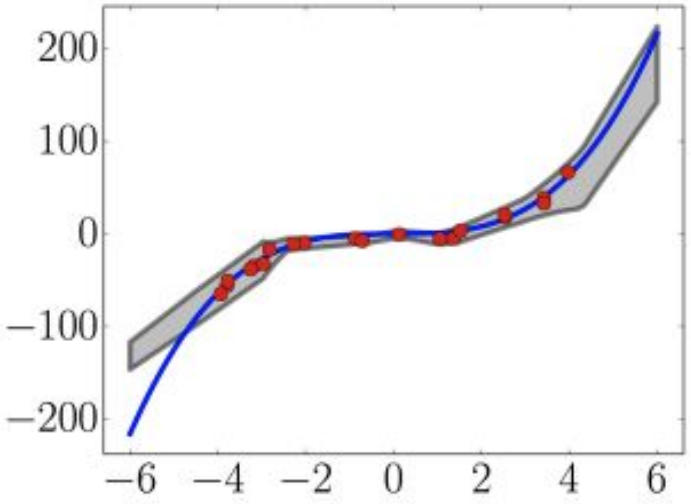
Deep Neural Networks are always overconfident!

Confidence = Probability of being correct

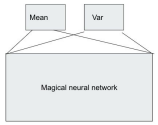


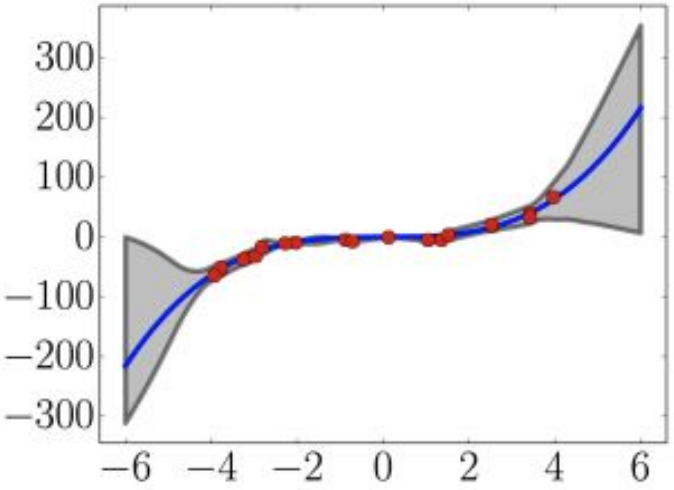
On Calibration of Modern Neural Networks JMLR 2017

Out of distribution problem



output from predicting variance via maximum likelihood





output from ensemble

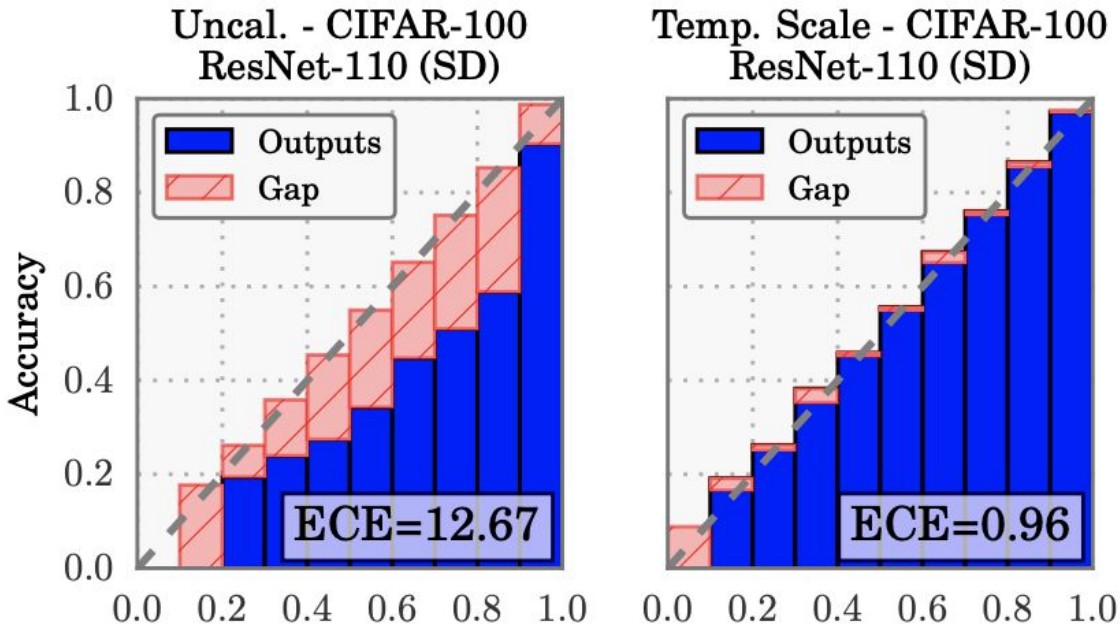
Simple and scalable predictive uncertainty estimation using deep ensembles. 2017

Neural network calibration

Make the confidence output follows the probability of being correct.

How?

Need a seperate training set to train the calibration (calibration set)



Overview of methods for calibration

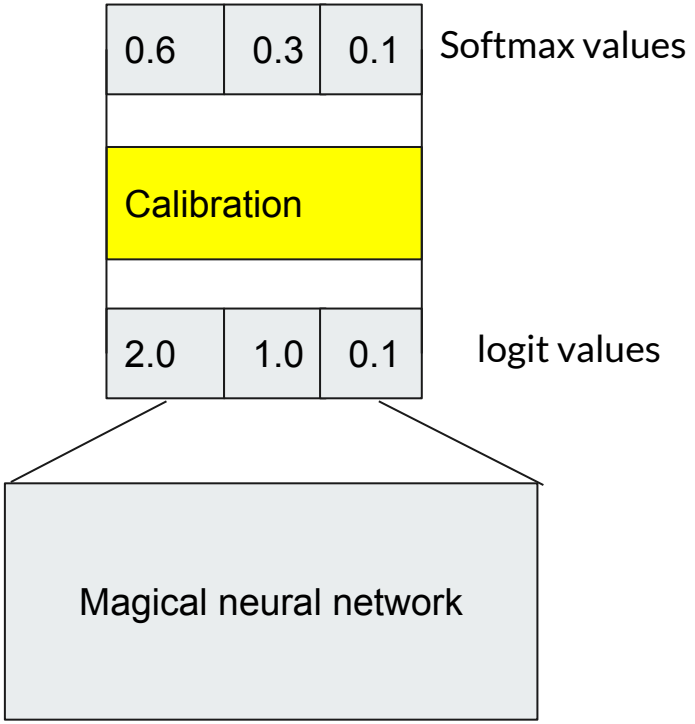
1 Calibration

2 Ensemble

- 2.1 Combining models
- 2.2 Bayesian neural networks

Calibration

Post processing after getting pre-softmax output (logit)



Calibration - Temperature scaling

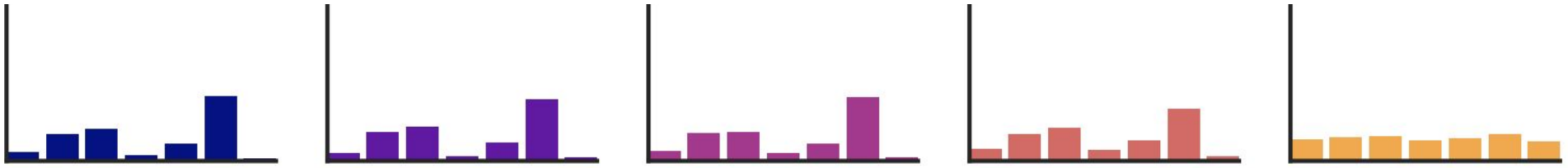
Add a temperature T to rescale the softmax

$$\hat{q}_i = \max_k \sigma_{\mathrm{SM}}(\mathbf{z}_i / T)$$

T is tuned to maximize log likelihood on the calibration set

Low T

High T



https://arxiv.org/pdf/1611.01144.pdf

Other calibration methods

Histogram binning
Bayesian binning into quartiles (BBQ)
Matrix and vector scaling (model on top of model)
Isotonic regression (model on top of model)

Try different methods on your dataset. No absolute best.

Overview of methods for calibration

1 Calibration

2 Ensemble

- 2.1 Combining models
- 2.2 Bayesian neural networks

Combining models

Create multiple models

    Calculate mean and variance of the answers!

Multiple models can be just from different initializations

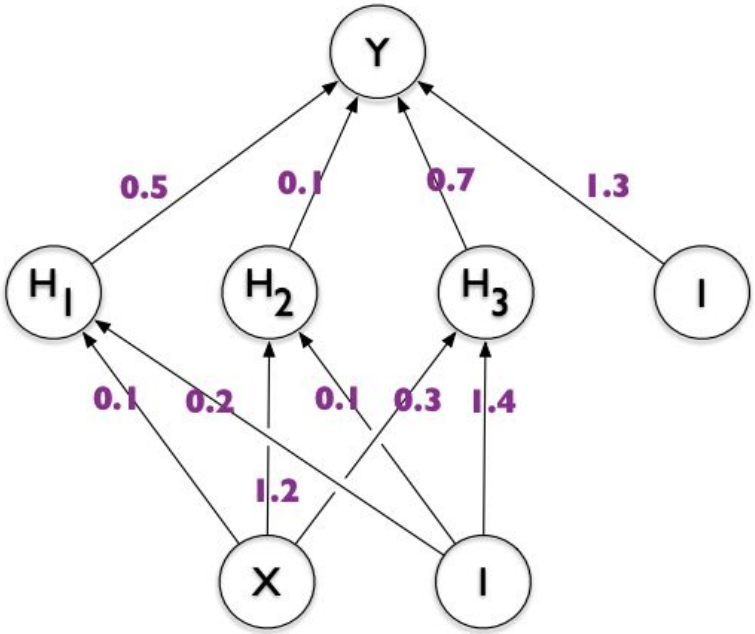
Cons: need to keep a lot of models around

How to get multiple models around?

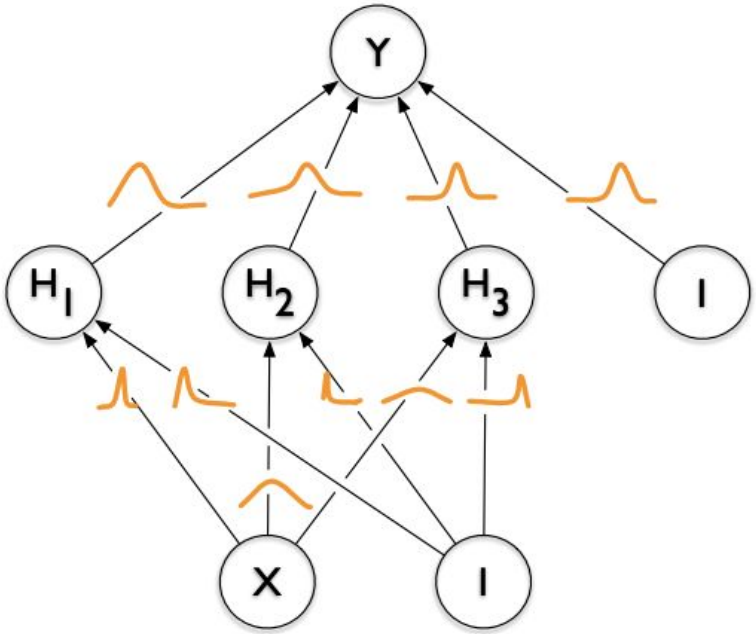
"Simple and scalable predictive uncertainty estimation using deep ensembles." 2017.

Bayesian Neural Network

Normal NN



Bayesian NN



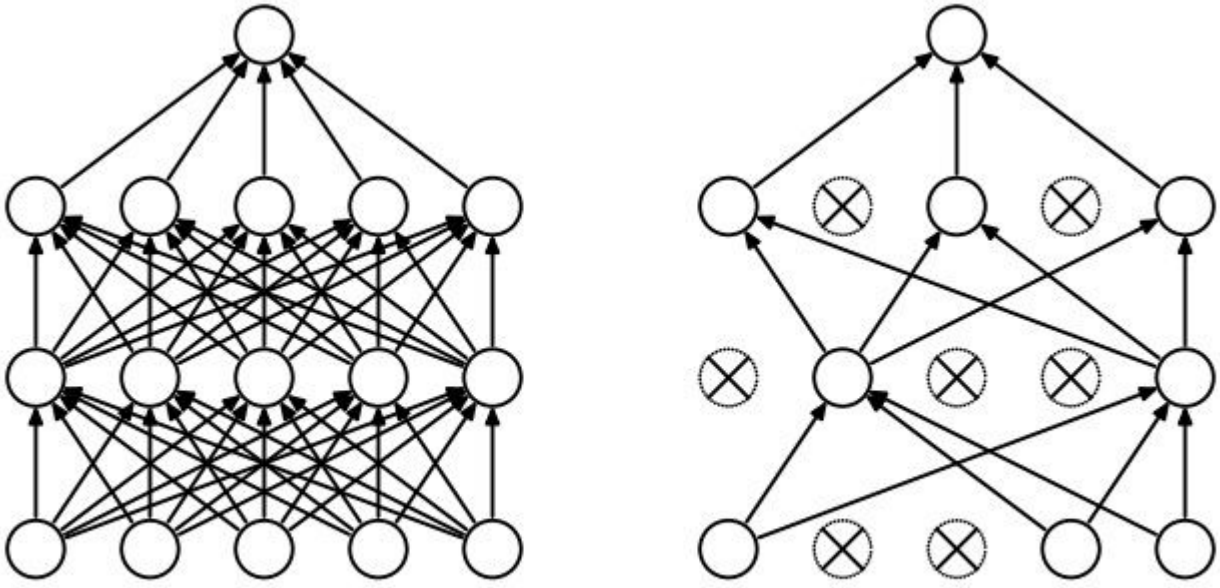
Problem: hard to do inference. Need to find $P(y \mid x, D) = \sum_{\theta} P(y \mid x, \theta) P(\theta \mid D)$

Weight Uncertainty in Neural Network. ICML 2015

Monte Carlo dropout

Use dropout at test time to simulate different models

Sampling of dropout at inference time is the same as sampling weights from Bernoulli distribution



(a) Standard Neural Net

(b) After applying dropout.

https://www.depends-on-the-definition.com/model-uncertainty-in-deep-learning-with-monte-carlo-dropout/

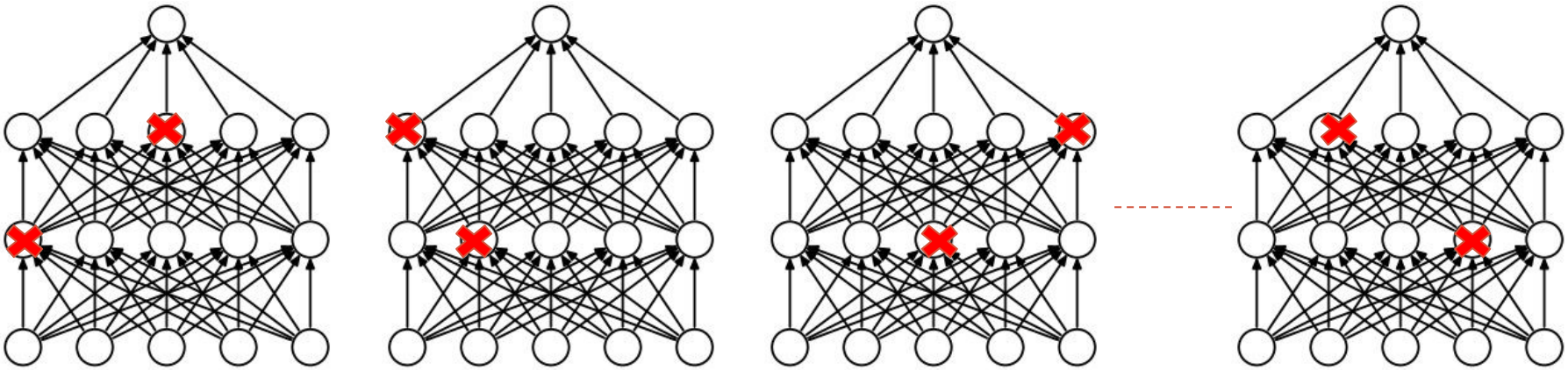
https://arxiv.org/pdf/1506.02142.pdf Dropout as a Bayesian Approximation: Representing Model Uncertainty in Deep Learning

Monte Carlo dropout for confidence estimation

compute mean and variance for the answer and confidence

20.2 +- 3.2

25 27 18 15



So we got the confidence, can we say why?

Two levels of understanding

Model level

- Describes the model tendency
- Talks about the behavior on training data

Output level

- Attributes model decision for a given test sample to different features

Model level: Simple example

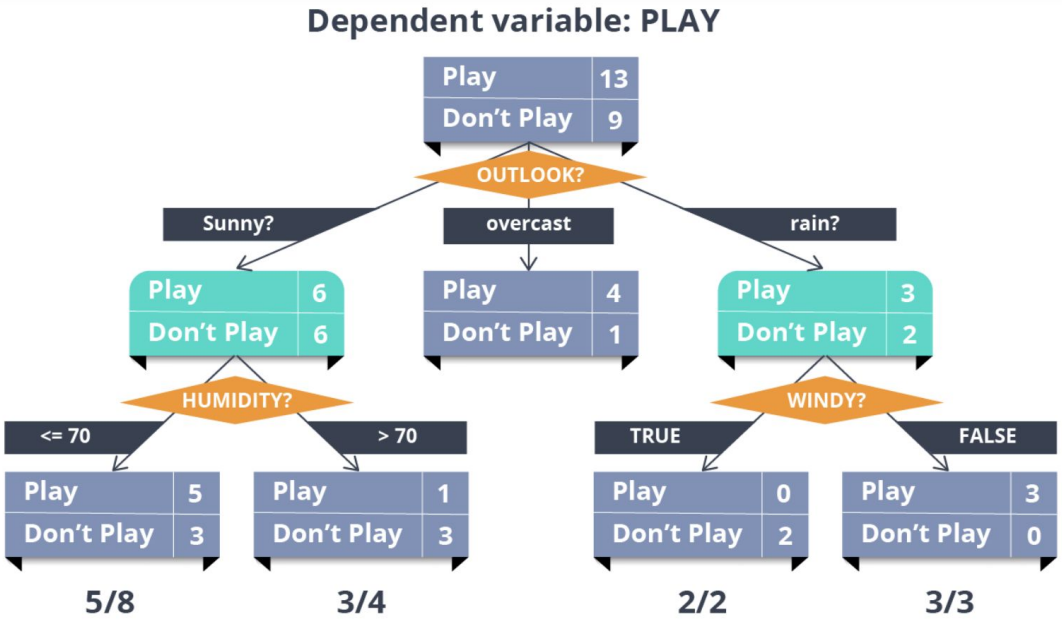
Logistic regression

$$g(z') = \phi_0 + \sum_{i=1}^{M} \phi_i z'_i,$$

You can say which feature is important from size of the coefficients.

Pros: Simple, easy to understand
Cons: Only model linear effects

Model level: tree-based (feature importance in XGBoost)



Assign scores based on how the features are use

A node that splits better (better purity at children): high score

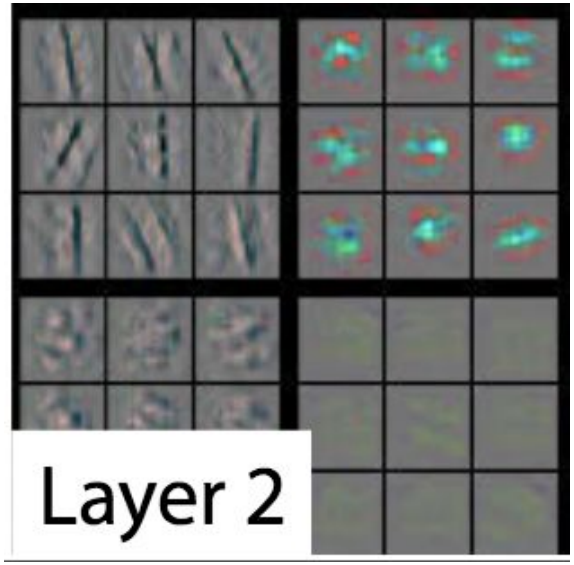
A node with more training samples: high weight

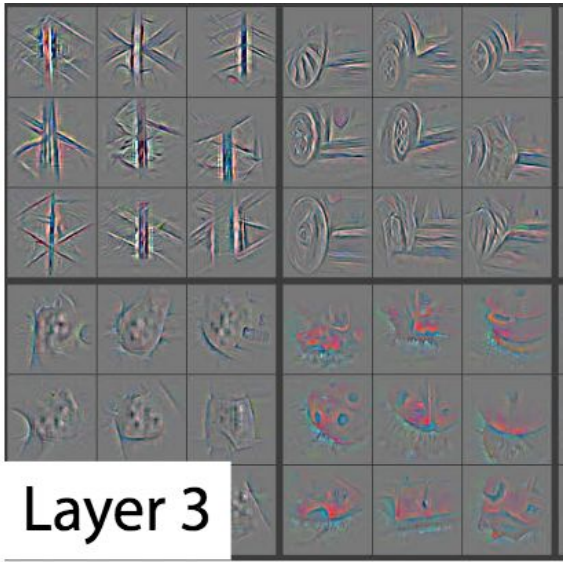
https://www.edureka.co/blog/decision-trees/

Model level: visualizing filters

Plot out the weights of the CNN filters...not so useful in practice

Layer 1





https://arxiv.org/pdf/1311.2901.pdf

Usefulness of model level

Feature selection

Gives you confidence that the model is learning reasonable things

Two levels of understanding

Model level

- Describes the model tendency
- Talks about the behavior on training data

Output level

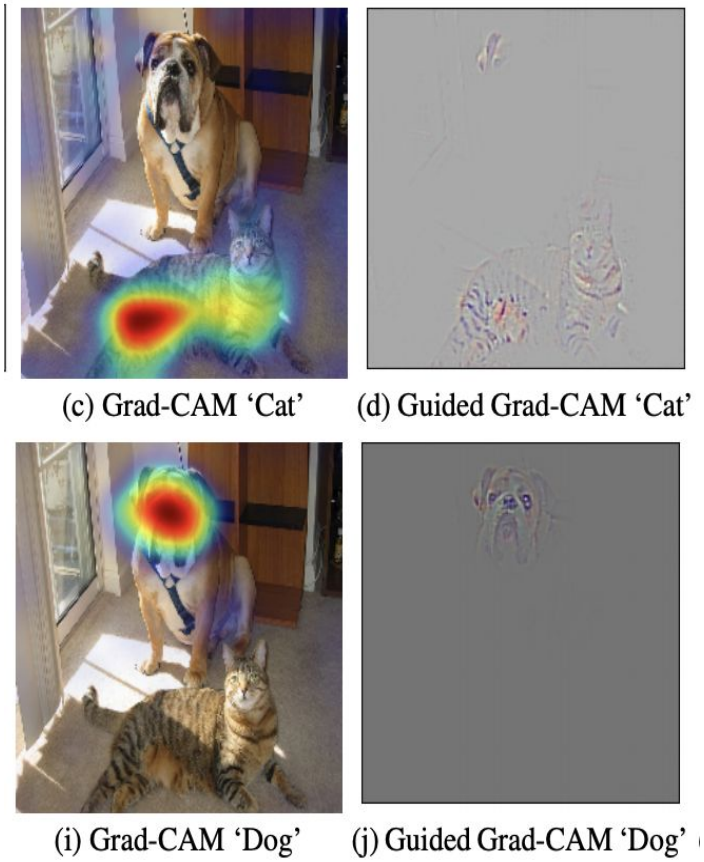
- Attributes model decision for a given test sample to different features

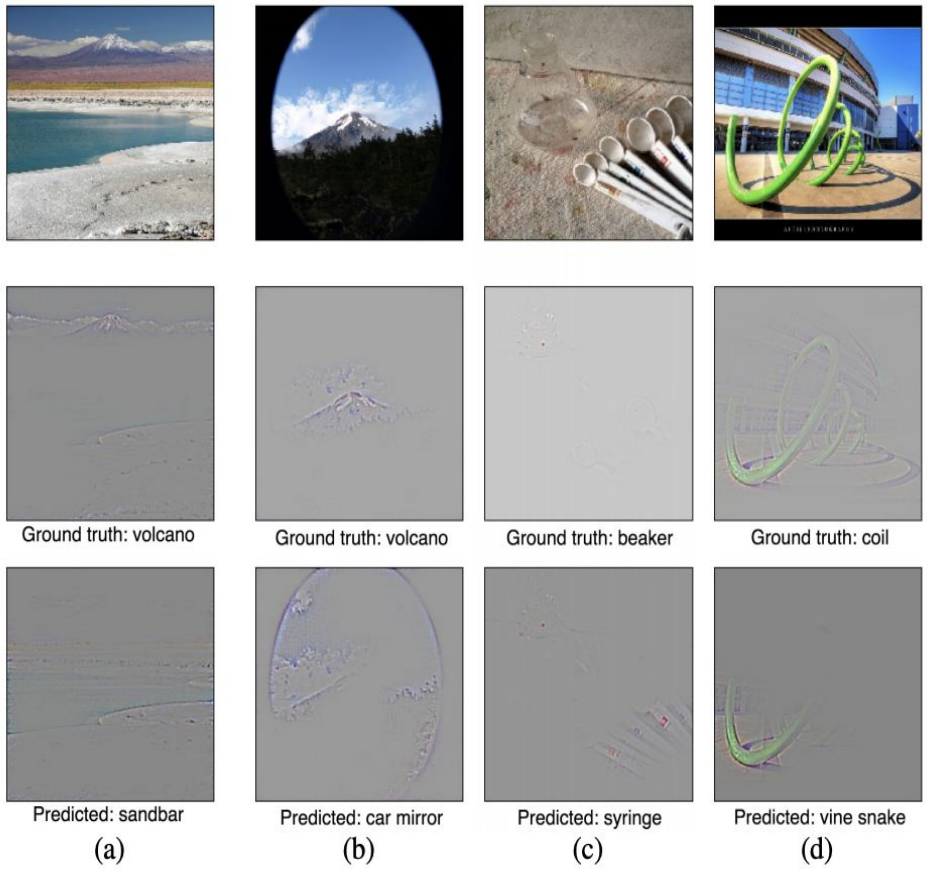
Output level: key ideas

A feature is important if I tweak the feature and the output change a lot.

If I tweak the output, how does the gradient flows

Output level: Gradient-weighted Class Activation Mapping (Grad-CAM)





https://arxiv.org/pdf/1610.02391.pdf
https://github.com/jacobgil/keras-grad-cam

Output level: key ideas

A feature is important if I tweak the feature and the output change a lot.

Mostly useful for images types

Have a simpler model (**surrogate model**) explains the complicated model.

Converting things to logistic regression

Additive feature attribution

$$g(z') = \phi_0 + \sum_{i=1}^{M} \phi_i z'_i,$$

Simplified input features z' have binary values

z' can recover original features x, $x = h_x(z')$ (This mapping depends on x)

Goal: make $g(z') = f(h_x(z'))$. Then we can explain f in terms of simplified features. (**Solved by optimization**)

Example of simplified inputs

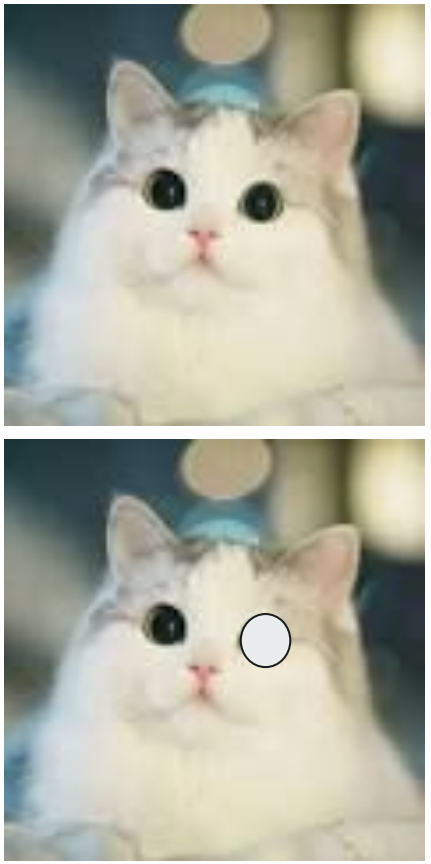
Original x = image

z = 0 if patch is not present

z = 1 if patch is present

$h_x(z) = x$ if, $z=1$

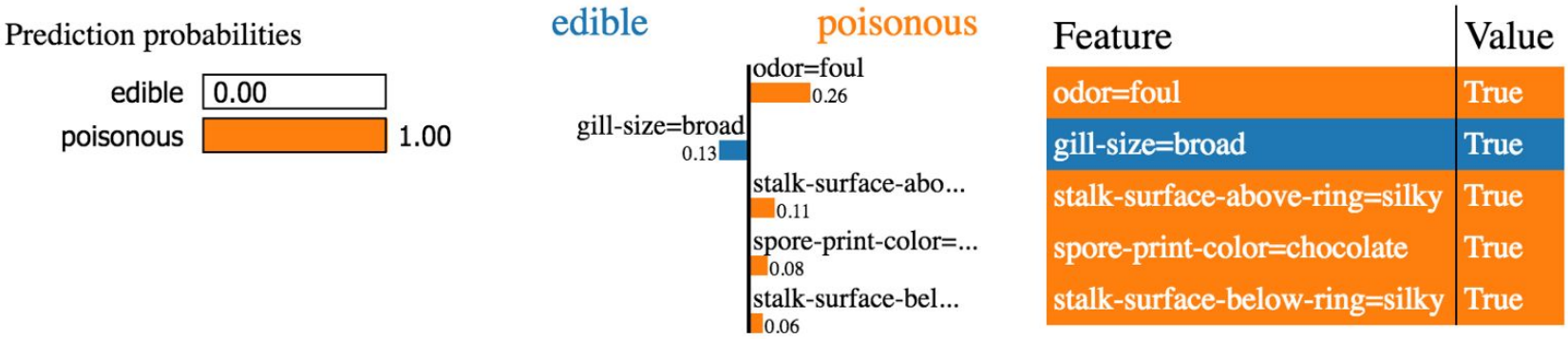
$h_x(z) = x$ with missing patch, if $z=0$



Additive feature attribution

$$g(z') = \phi_0 + \sum_{i=1}^{M} \phi_i z'_i,$$

Many methods fall is additive feature attribution. For example LIME



Introducing SHAP

SHAP is also an additive feature attribution, but gives credit to expected attribution

Example of simplified inputs

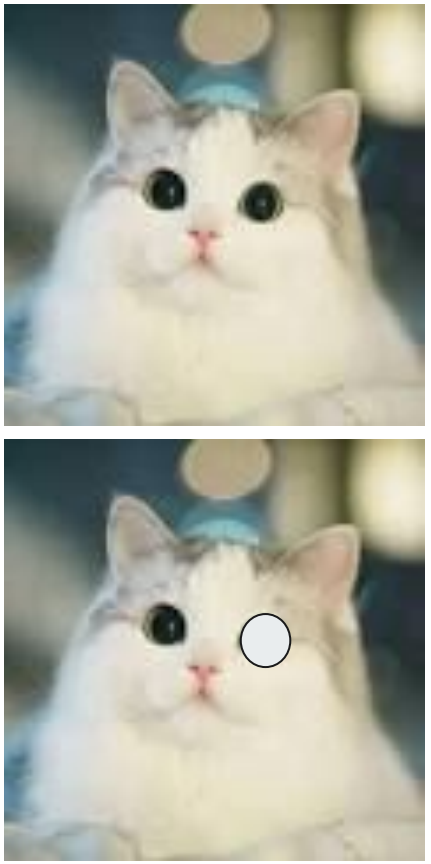
Original x = image

z = 0 if patch is not present

z = 1 if patch is present

$h_x(z) = x$ if, $z=1$

$h_x(z) = x$ with missing patch, if $z=0$

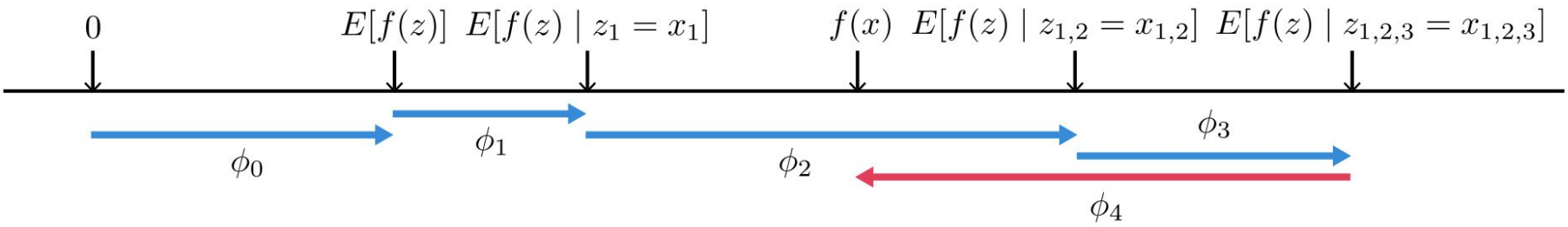


This simplified input talks about this patch.

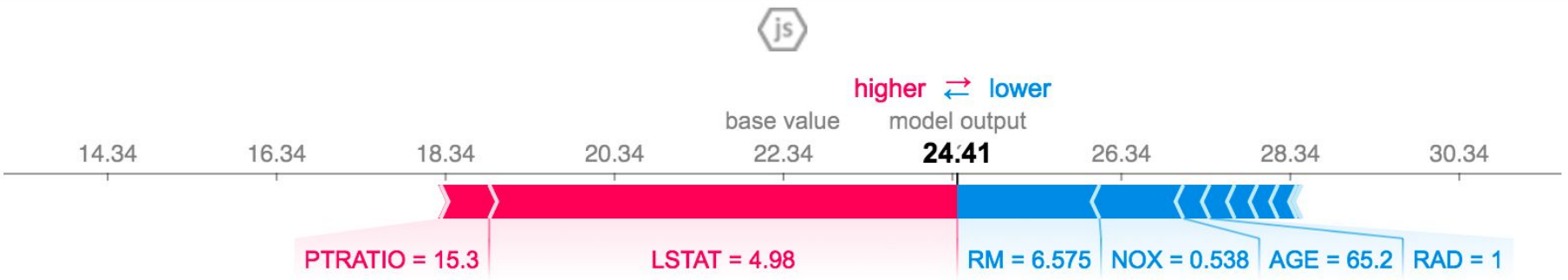
What happens if we change the other inputs?

Expected contribution talks about the contribution of the feature regardless of the other contributions

SHAP

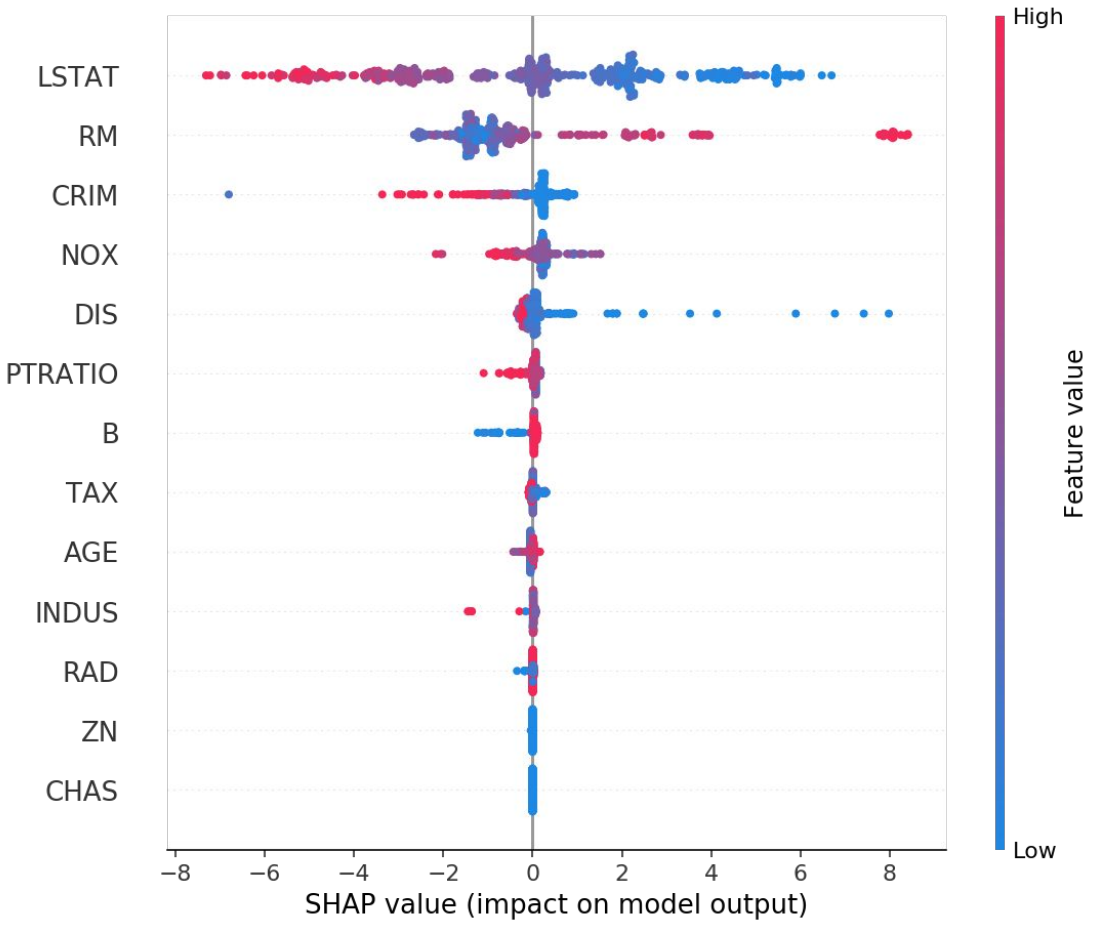

0
$E[f(z)]$
$E[f(z) \mid z_1 = x_1]$
$f(x)$
$E[f(z) \mid z_{1,2} = x_{1,2}]$
$E[f(z) \mid z_{1,2,3} = x_{1,2,3}]$
ϕ_0
ϕ_1
ϕ_2
ϕ_3
ϕ_4

SHAP example



https://github.com/slundberg/shap

SHAP example



Notes on additive attribution

If the features are correlated, it's hard to divide the attribution

Feature 1 = height

Feature 2 = height*2

Notes on attributions

It does not necessary tell how to improve.

If the explainer says the sales is low because of weak marketing

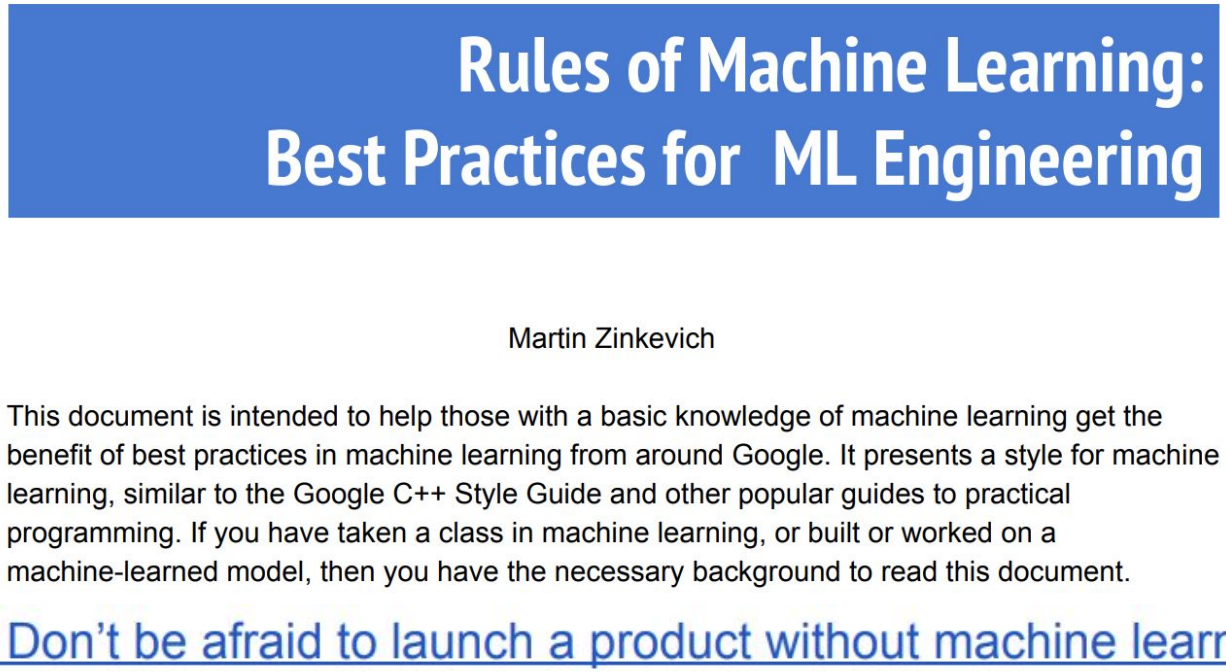
Does not mean increasing marketing will improve sales.

See Causal Inference

Further readings

Rules of Machine Learning: Best Practices for ML Engineering

http://martin.zinkevich.org/rules_of_ml/rules_of_ml.pdf



Rules of Machine Learning:
Best Practices for ML Engineering

Martin Zinkevich

This document is intended to help those with a basic knowledge of machine learning get the benefit of best practices in machine learning from around Google. It presents a style for machine learning, similar to the Google C++ Style Guide and other popular guides to practical programming. If you have taken a class in machine learning, or built or worked on a machine-learned model, then you have the necessary background to read this document.

Rule #1: Don't be afraid to launch a product without machine learning.

Ethics issues

- Abuse
- Bias

Different Levels In a Self Driving Car

LEVEL 0

There are no autonomous features.

LEVEL 1

These cars can handle one task at a time, like automatic braking.

LEVEL 2

These cars would have at least two automated functions.

LEVEL 3

These cars handle "dynamic driving tasks" but might still need intervention.

LEVEL 4

These cars are officially driverless in certain environments.

LEVEL 5

These cars can operate entirely on their own without any driver presence.

OodlesTechnologies

www.oodlestechnologies.com

https://www.oodlestechnologies.com/blogs/5-Different-Automation-Levels-In-a-Self-Driving-Car

Tesla is updating Autopilot's 'Hold Steering Wheel' alert after complaints, says Elon Musk

Fred Lambert - Jun. 13th 2018 6:12 am ET @FredericLambert

https://electrek.co/2018/06/13/tesla-autopilot-hold-steering-wheel-alerts-complaints/

‘The Business of War’: Google Employees Protest Work for the Pentagon

Thousands of Google employees have signed a letter to Sundar Pichai, the company’s chief executive, protesting Google’s role in a program that could be used to improve drone strike targeting.
Michael Short/Bloomberg

https://www.nytimes.com/2018/04/04/technology/google-letter-ceo-pentagon-project.html

Bias

- Our society is biased which leads to biased data and biased models
  - Understanding how models make prediction help us remove the bias
- Our morality is also biased
  - Understanding the cost-benefit tradeoff

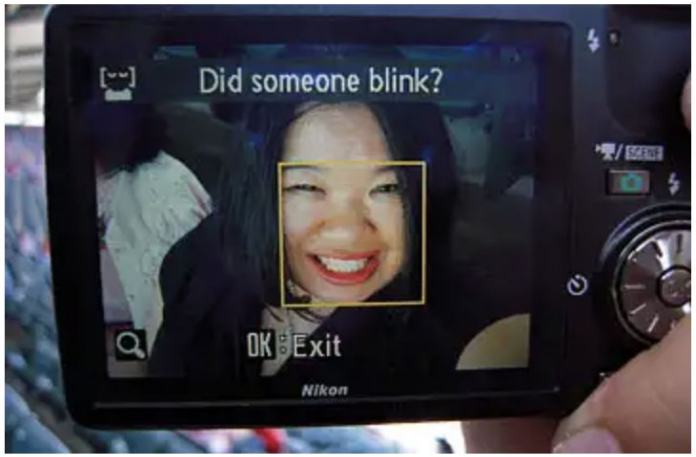
Harms of Bias

Allocation

Models allocate resource to certain groups of people more - credit prediction

Representation

Models have more data of certain demographic and does better



http://content.time.com/time/business/article/0,8599,1954643,00.html

Study finds gender and skin-type bias in commercial artificial-intelligence systems

Examination of facial-analysis software shows error rate of 0.8 percent for light-skinned men, 34.7 percent for dark-skinned women.

Watch Video

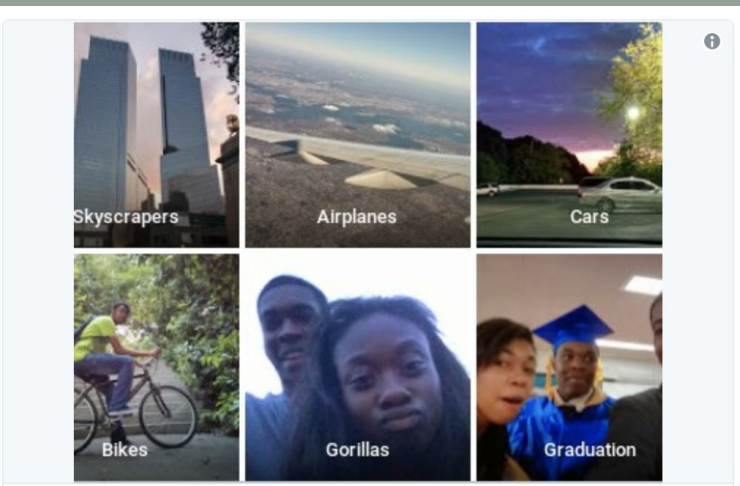
Larry Hardesty | MIT News Office
February 11, 2018

Press Inquiries | PRESS MENTIONS

http://news.mit.edu/2018/study-finds-gender-skin-type-bias-artificial-intelligence-systems-0212

Ways to combat bias

- Improve accuracy <- hard
- Blacklist
- Equal representation
- Awareness
- Test the system on general users



Google 'fixed' its racist algorithm by removing gorillas from its image-labeling tech

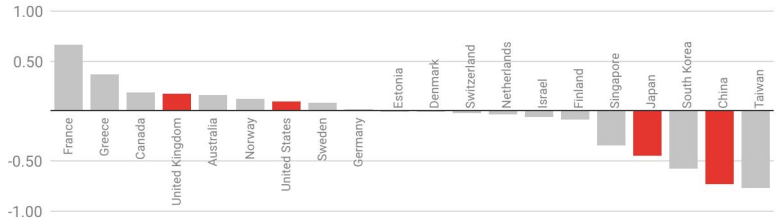
Nearly three years after the company was called out, it hasn't gone beyond a quick workaround

By James Vincent | @jjvincent | Jan 12, 2018, 10:35am EST

https://www.theverge.com/2018/1/12/16882408/google-racist-gorillas-photo-recognition-algorithm-ai

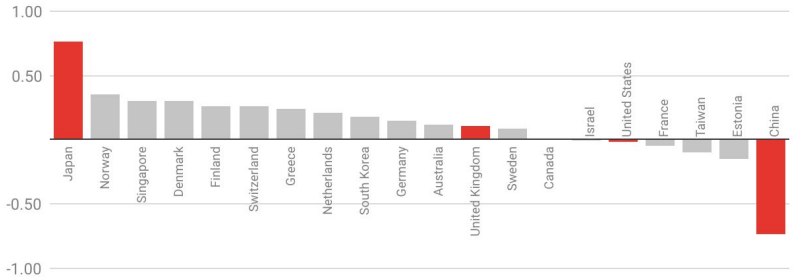
Moral machine

Countries with more individualistic cultures are more likely to spare the young



A comparison of countries piloting self-driving cars: If the bar is closer to 1, respondents placed a greater emphasis on sparing the young; if the bar is closer to -1, respondents placed a greater emphasis on sparing the old; 0 is the global average.

How countries compare in sparing pedestrians over passengers



If the bar is closer to 1, respondents placed a greater emphasis on sparing pedestrians; if the bar is closer to -1, respondents placed a greater emphasis on sparing passengers; 0 is the global average.

Problems with bias is both a social issue and a technological issue

https://www.technologyreview.com/s/612341/a-global-ethics-study-aims-to-help-ai-solve-the-self-driving-trolley-problem/

- “If I were to guess like what our biggest existential threat is, it’s probably that. So we need to be very careful with the artificial intelligence. There should be some regulatory oversight maybe at the national and international level, just to make sure that we don’t do something very foolish.”

- “I think people who are naysayers and try to drum up these doomsday scenarios — I just, I don’t understand it. It’s really negative and in some ways I actually think it is pretty irresponsible”

Poll

Beyond this course

- Things you might want to pursue more
  - Graphical models
    - Stochastic process
    - Graph theory
  - Random forests and ensemble methods
  - Optimization

Automatic Speech Recognition

- Basics of human speech generation
- Hidden Markov Models
- Signal processing, Fourier Transforms (basis transforms)
- Connectionist Temporal Classification
- How to build and deploy your own speech application

ASR

Natural language processing

- Conditional Random Fields
- Basics of language understanding
- Machine translation and Question Answering
- Chatbots

- Take both for the best experience!

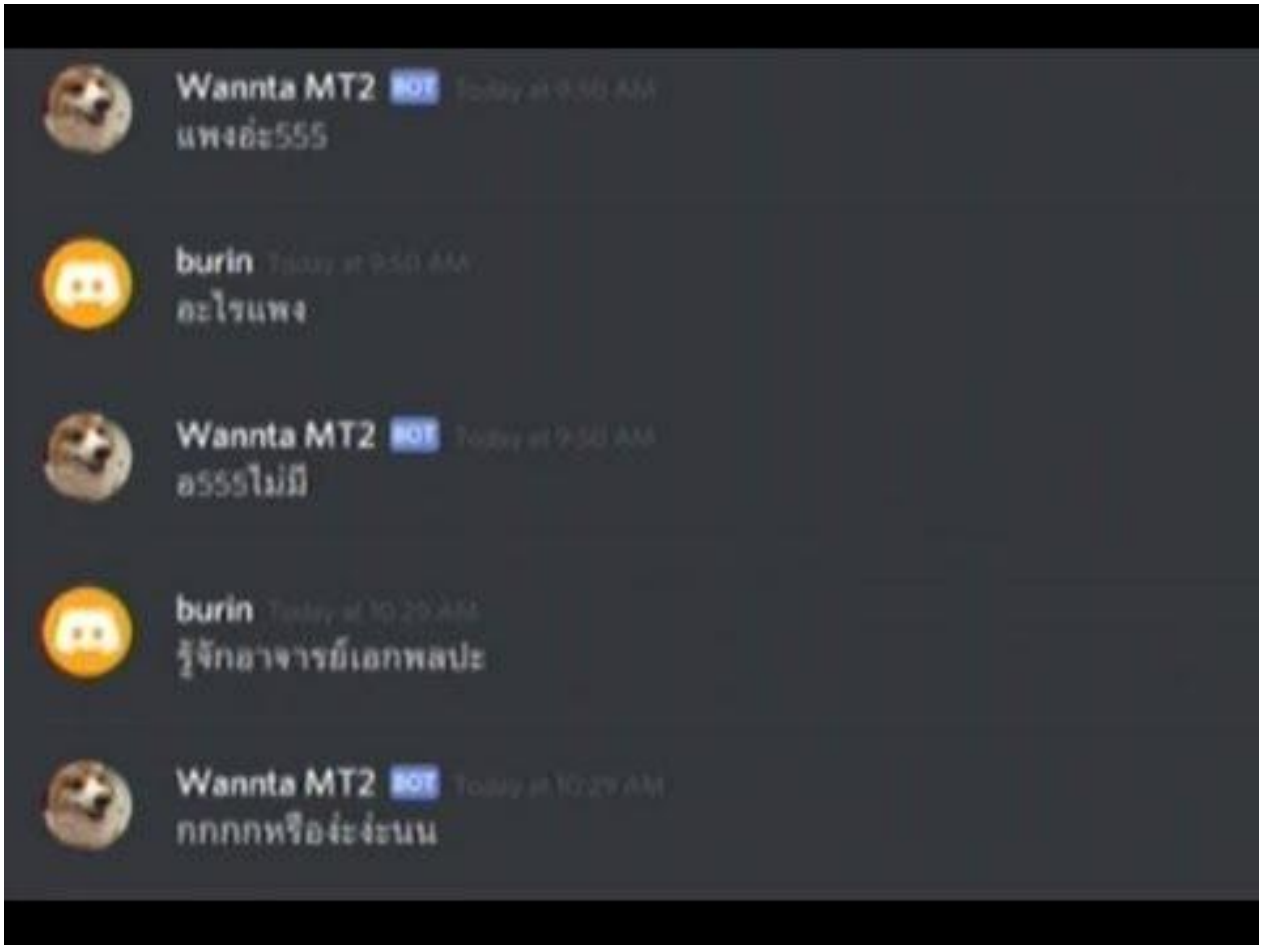
GitHub, Inc. [US] | https://github.com/google-research/bert/blob/master/multilingual.md

Data Source and Sampling

The languages chosen were the top 100 languages with the largest Wikipedias. The entire Wikipedia dump for each language (excluding user and talk pages) was taken as the training data for each language

The only language which we had to unfortunately exclude was Thai, since it is the only language (other than Chinese) that does not use whitespace to delimit words, and it has too many characters-per-word to use character-based tokenization. Our WordPiece algorithm is quadratic with respect to the size of the input token so very long character strings do not work with it.

NLP


Wannta MT2 BOT
แพจอ่ะ555
burin
อะไรแพจ
Wannta MT2 BOT
อ555ไม่มี
burin
รู้จักอาจารย์เอกพลปะ
Wannta MT2 BOT
กกกกหรืออ่ะจ่ะนน

Course philosophy

- Pattern Recognition: understand, build, and use machine learning models
- ASR: understand, and use machine learning models for ASR
- NLP: understand, and use machine learning models for NLP