TRICKS OF THE TRADE:

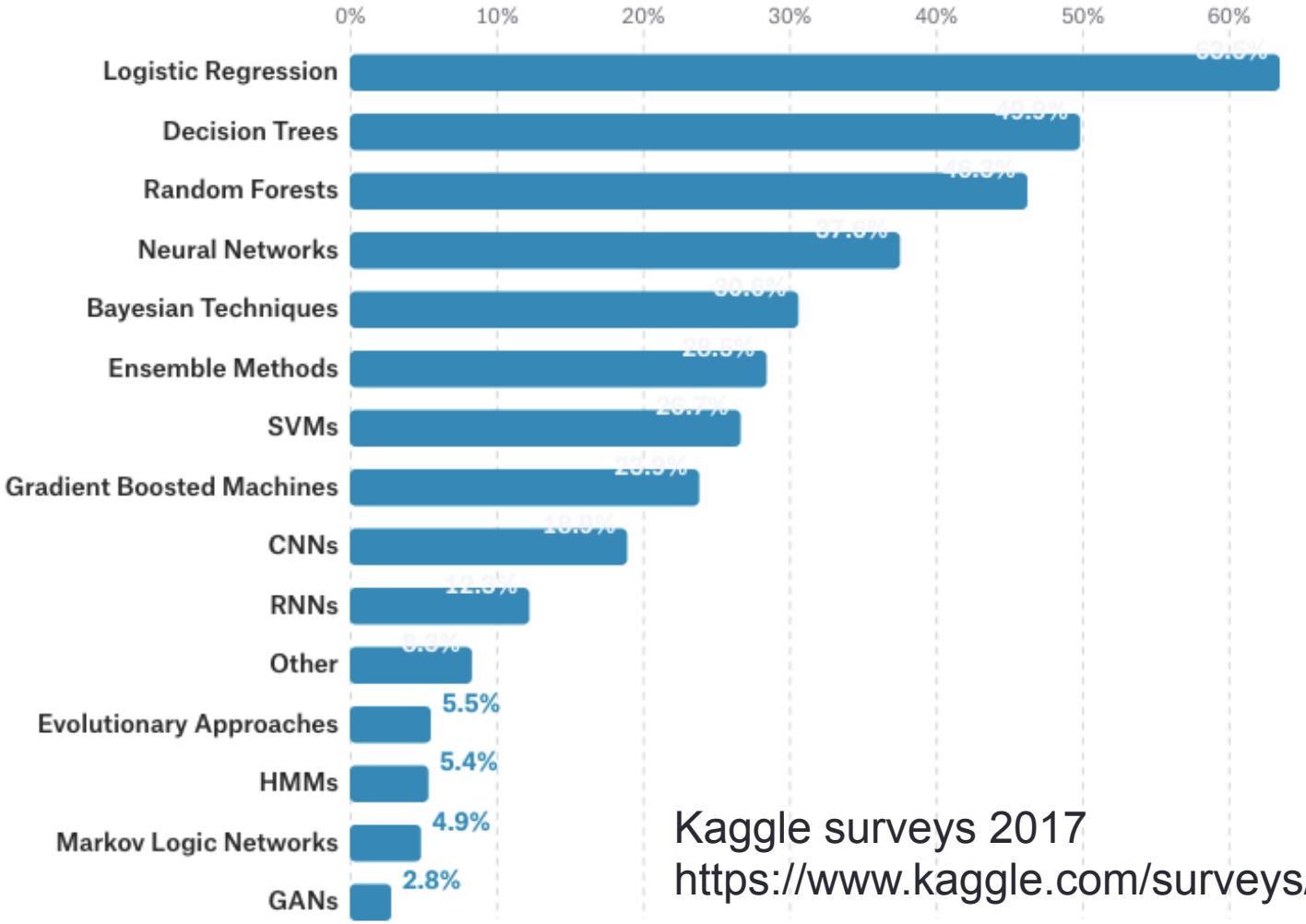
Machine learning in the real world

Burning questions

- Which model to use?
- How should I improve my models?
  - Diagnosis and error analysis

- How to build a machine learning startup?

Which model?



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

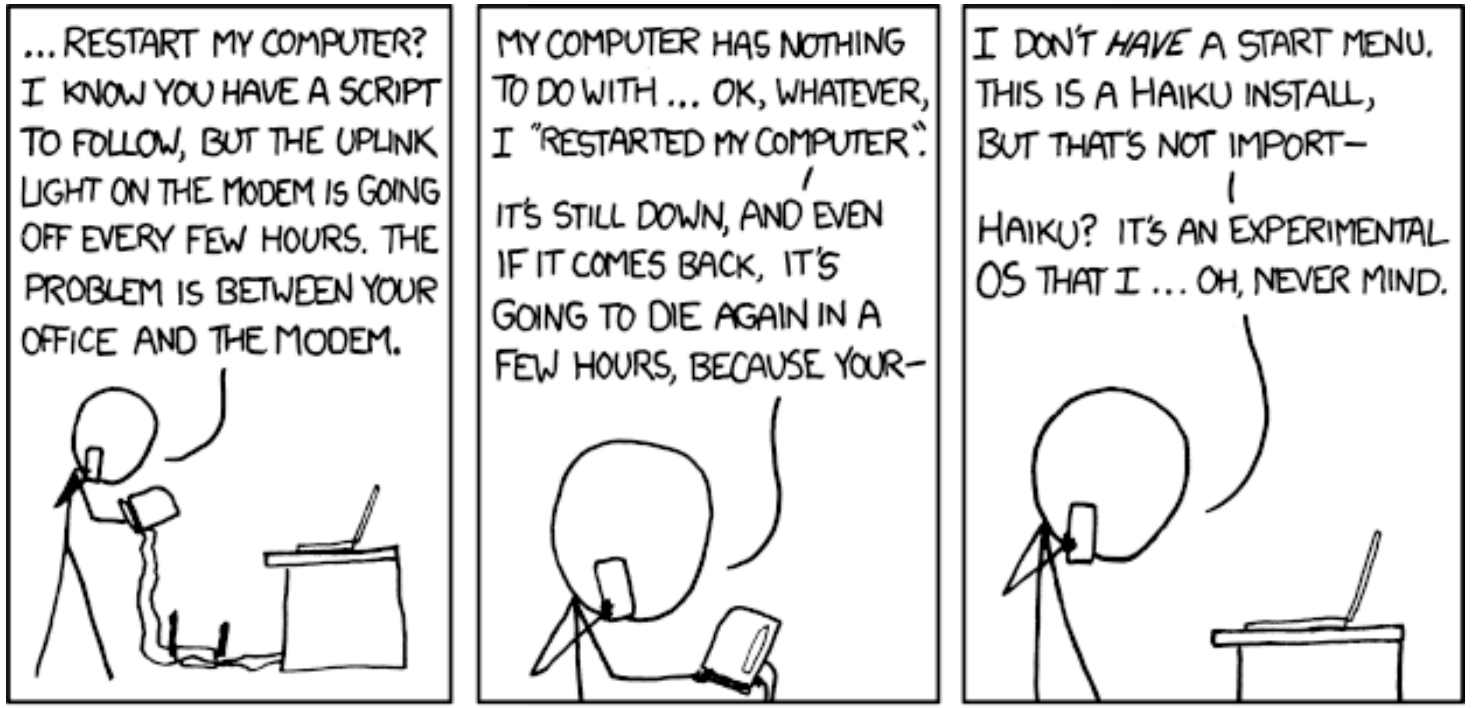
Cautionary notes

- “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.”

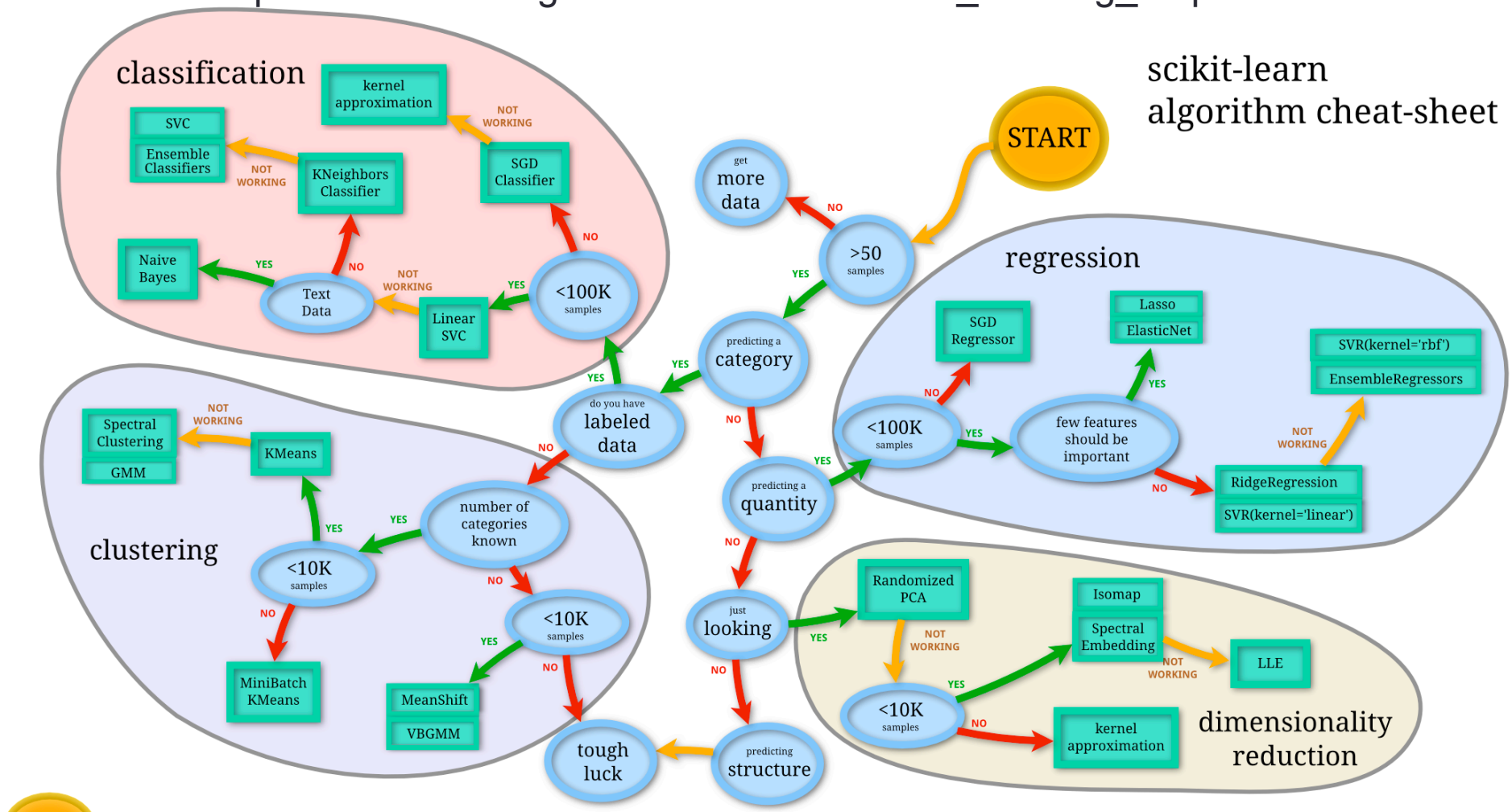
That's not so helpful…

Tech Support



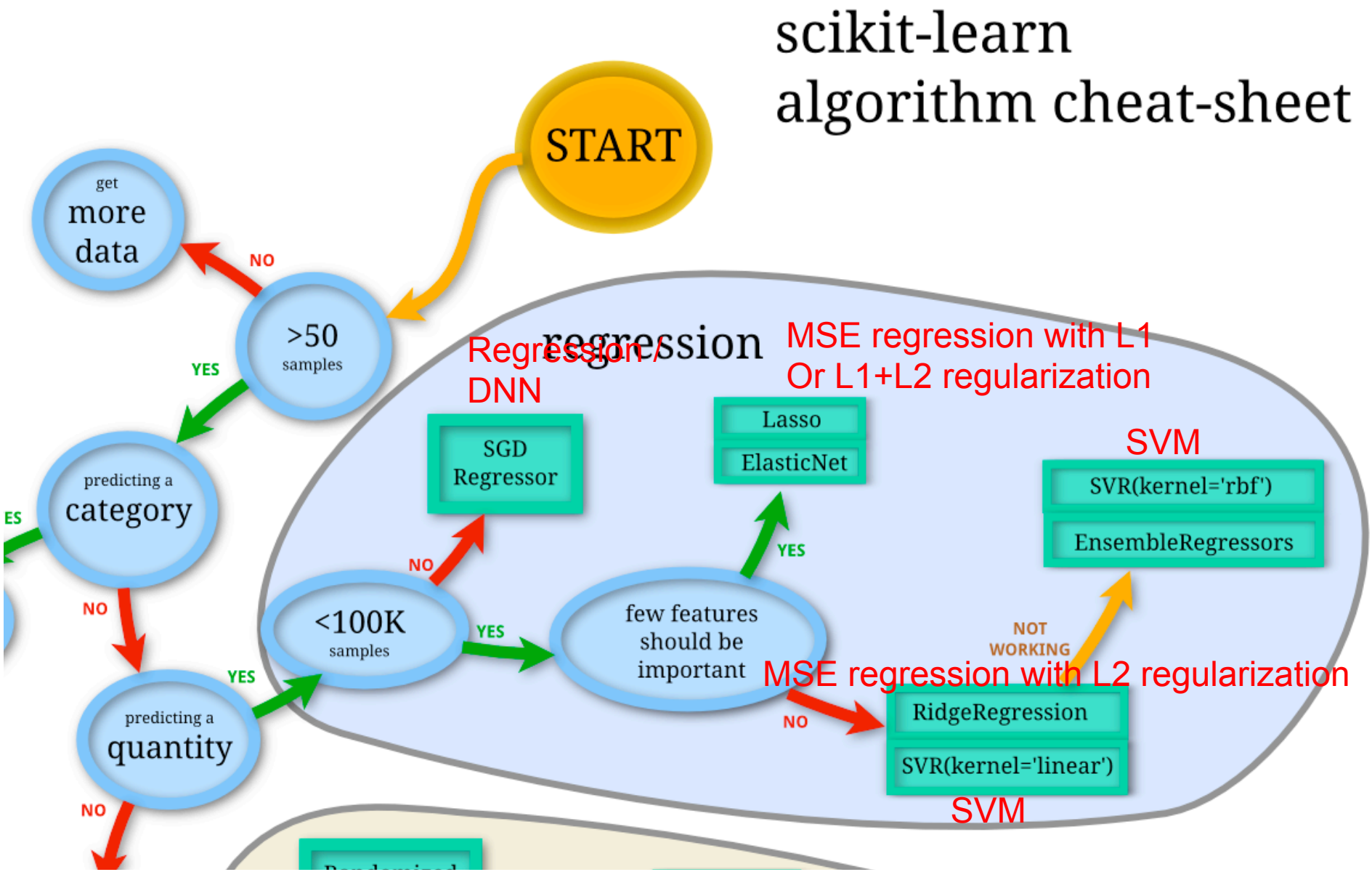
https://xkcd.com/806/

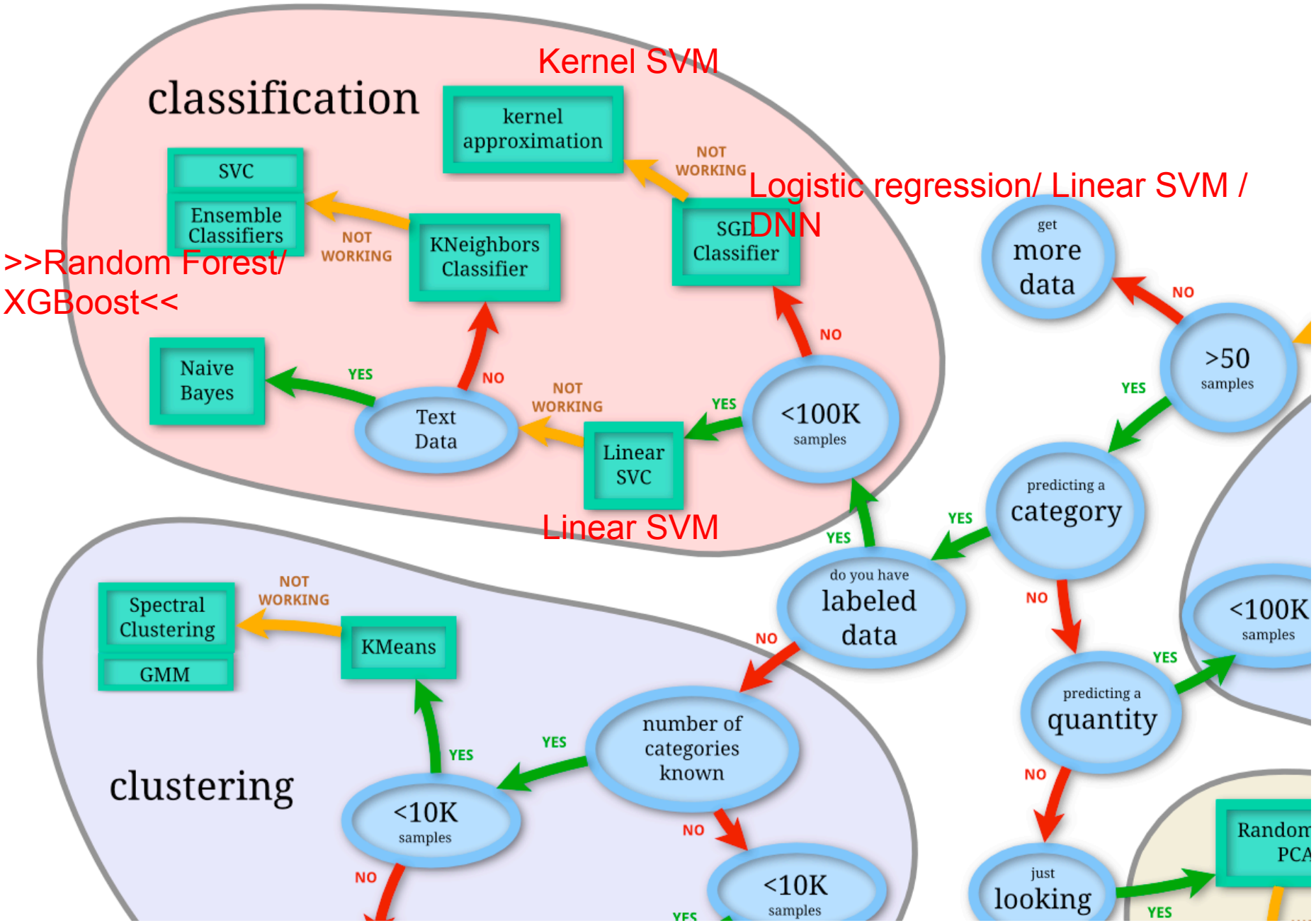
http://scikit-learn.org/stable/tutorial/machine_learning_map/index.html

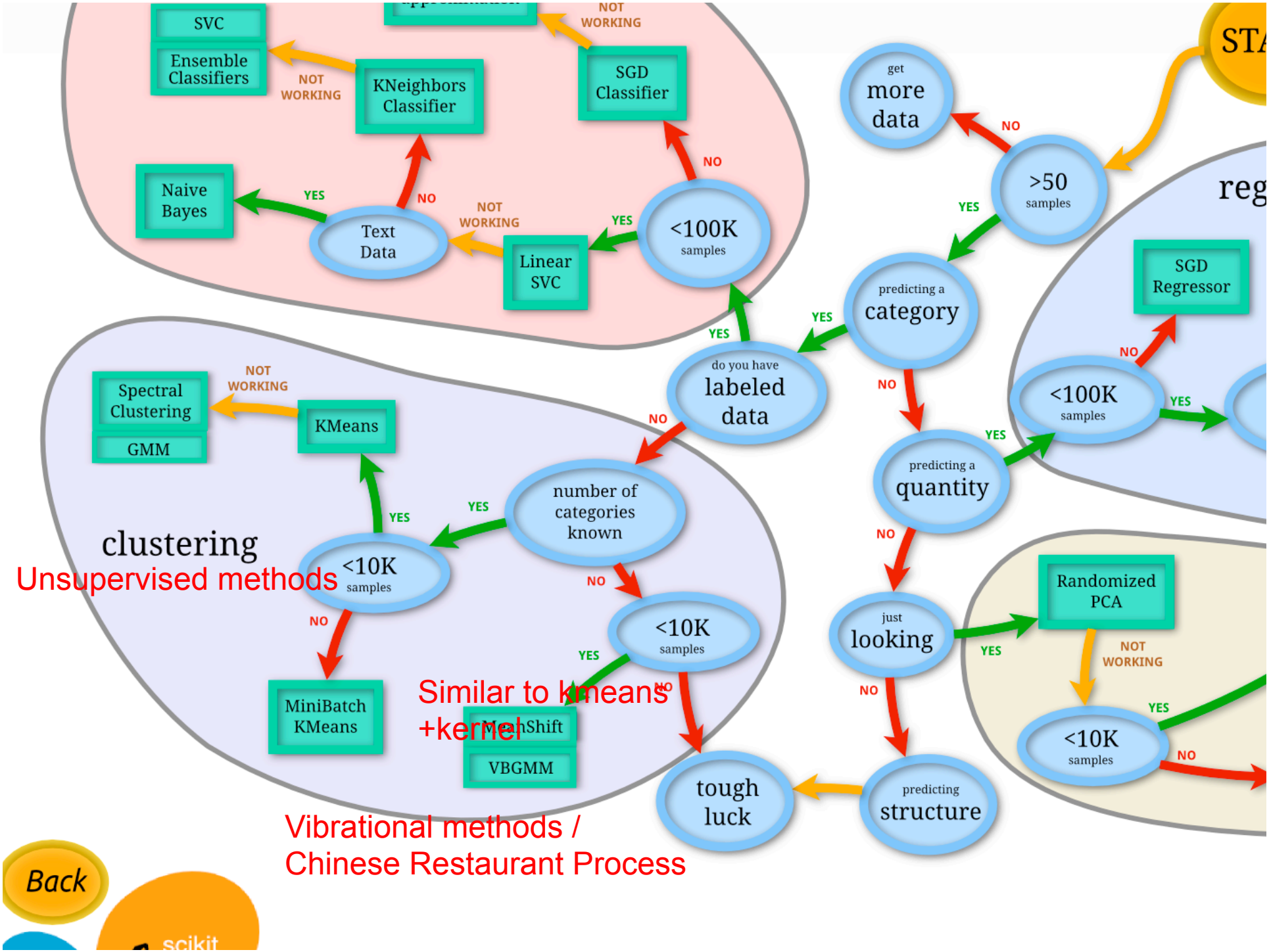


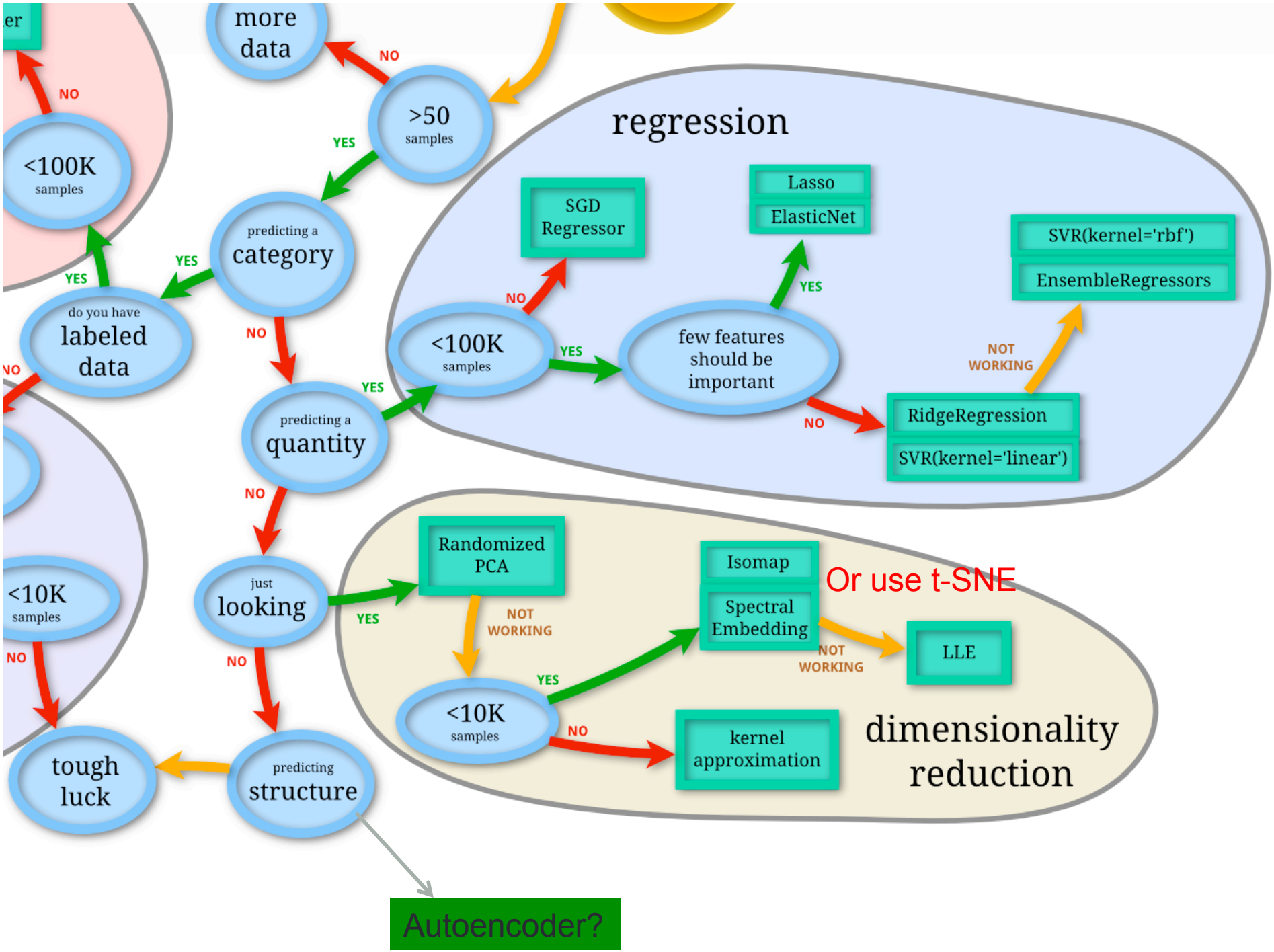
Back

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
YES
NO
SGD Regressor
few features should be important
YES
NO
Lasso
ElasticNet
NOT WORKING
SVM
SVR(kernel='rbf')
EnsembleRegressors
MSE regression with L2 regularization
RidgeRegression
SVR(kernel='linear')
SVM
Regression / DNN
MSE regression with L1
Or L1+L2 regularization


classification
Kernel SVM
kernel approximation
NOT WORKING
SVC
Ensemble Classifiers
NOT WORKING
KNeighbors Classifier
SGD Classifier
Logistic regression/ Linear SVM / DNN
>>Random Forest/ XGBoost<<
Naive Bayes
YES
NO
Text Data
NOT WORKING
Linear SVC
YES
<100K samples
NO
Linear SVM
get more data
NO
>50 samples
YES
predicting a category
YES
do you have labeled data
YES
NO
<100K samples
Spectral Clustering
GMM
NOT WORKING
KMeans
YES
number of categories known
YES
NO
predicting a quantity
YES
clustering
<10K samples
NO
NO
<10K samples
NO
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
YES
get more data
NO
>50 samples
YES
predicting a category
YES
do you have labeled data
NO
Spectral Clustering
GMM
NOT WORKING
KMeans
NO
number of categories known
YES
YES
<10K samples
NO
MiniBatch KMeans
<10K samples
YES
NO
MeanShift
VBGMM
clustering
Unsupervised methods
Similar to kmeans +kernel
Vibrational methods / Chinese Restaurant Process
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
Back
scikit


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

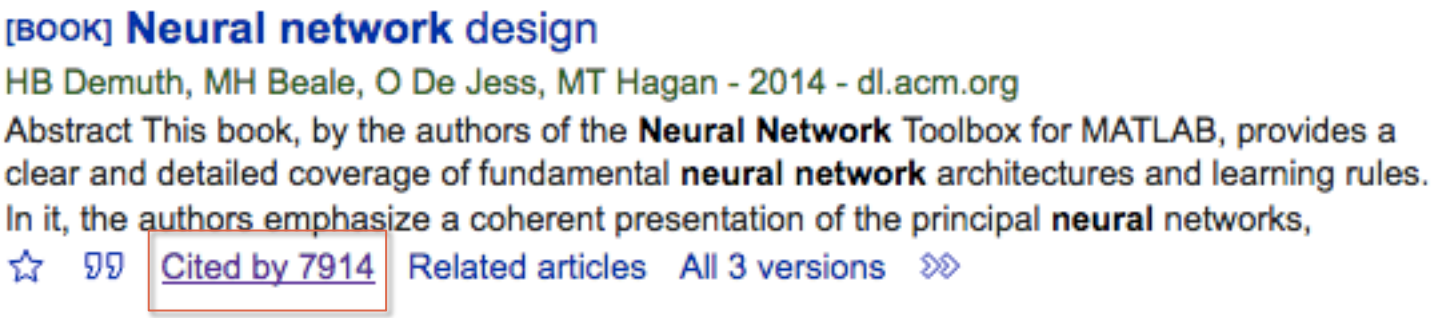
http://jmlr.org/papers/volume15/delgado14a/delgado14a.pdf

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 check list
  - 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

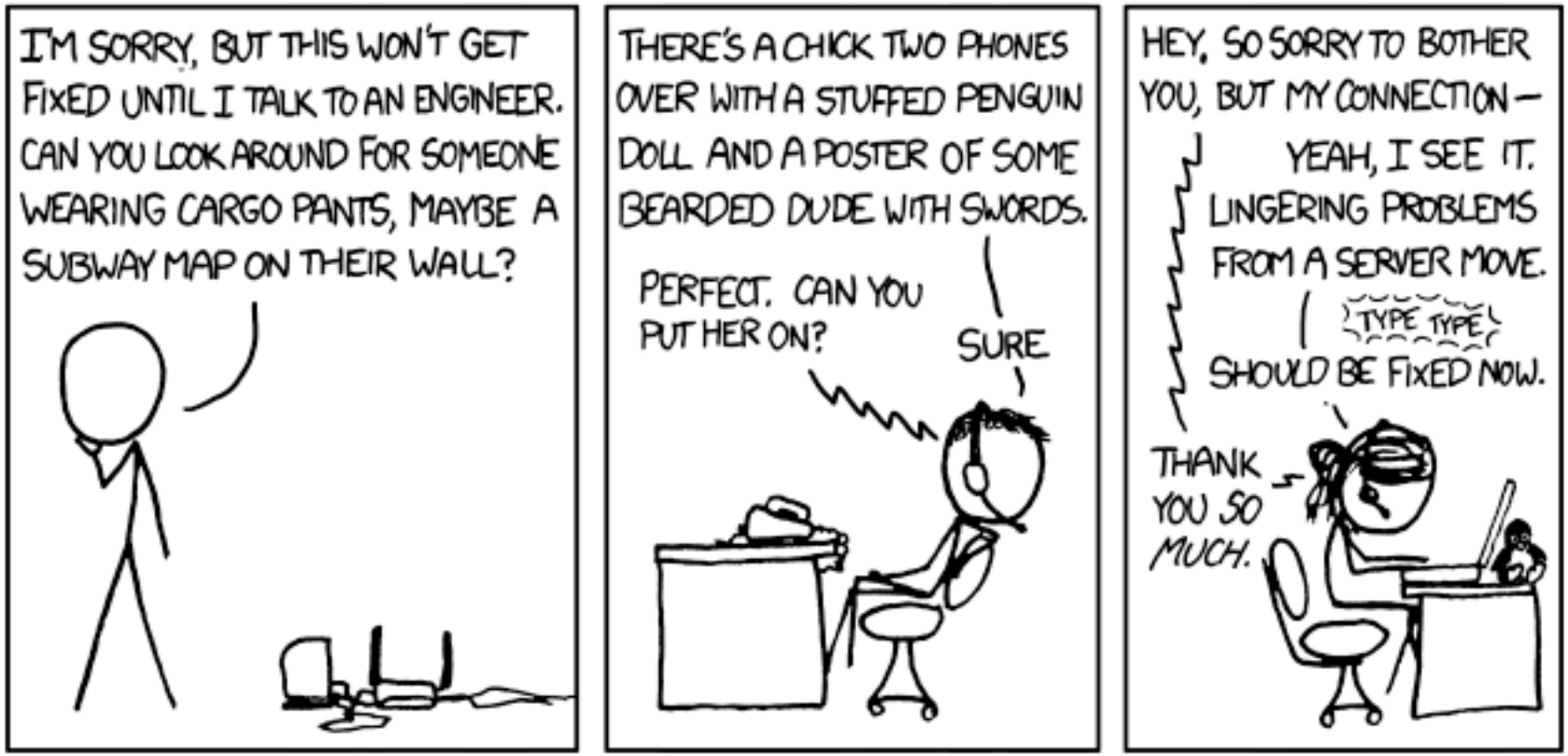
- If you don't understand something, re-read
  - Average of 5+ times to understand a paper completely
- Print it out, keep a pen/pencil at hand, and break down equations
- Try to explain it with what you already know

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?

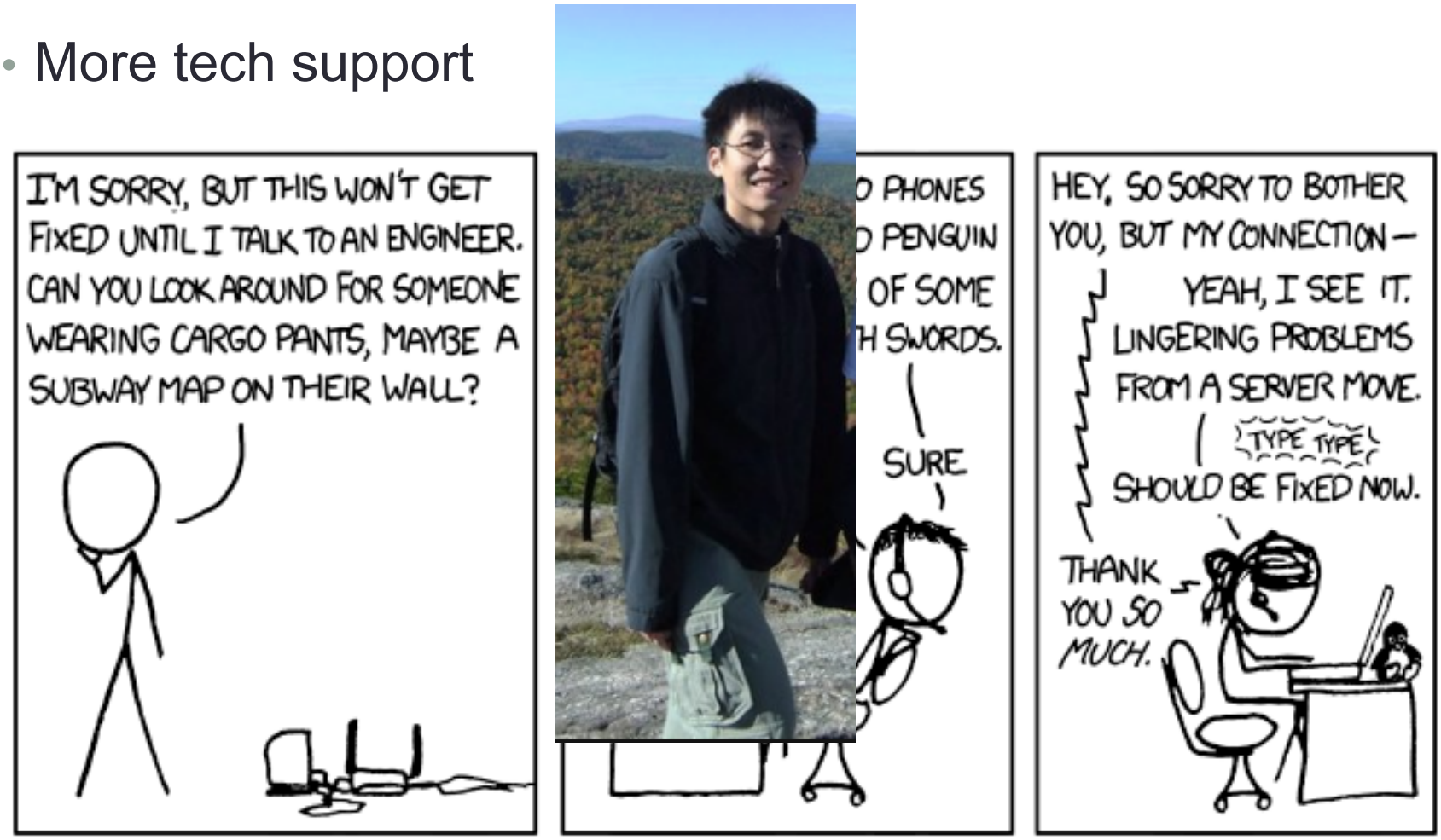
- More tech support



https://xkcd.com/806/

How to improve my results?

- More tech support



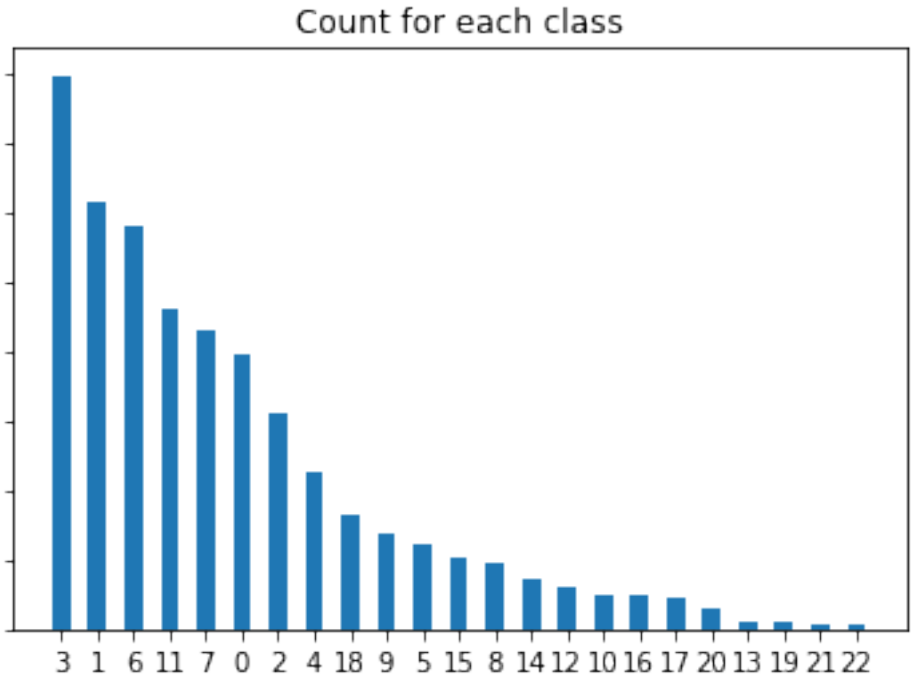
https://xkcd.com/806/

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?

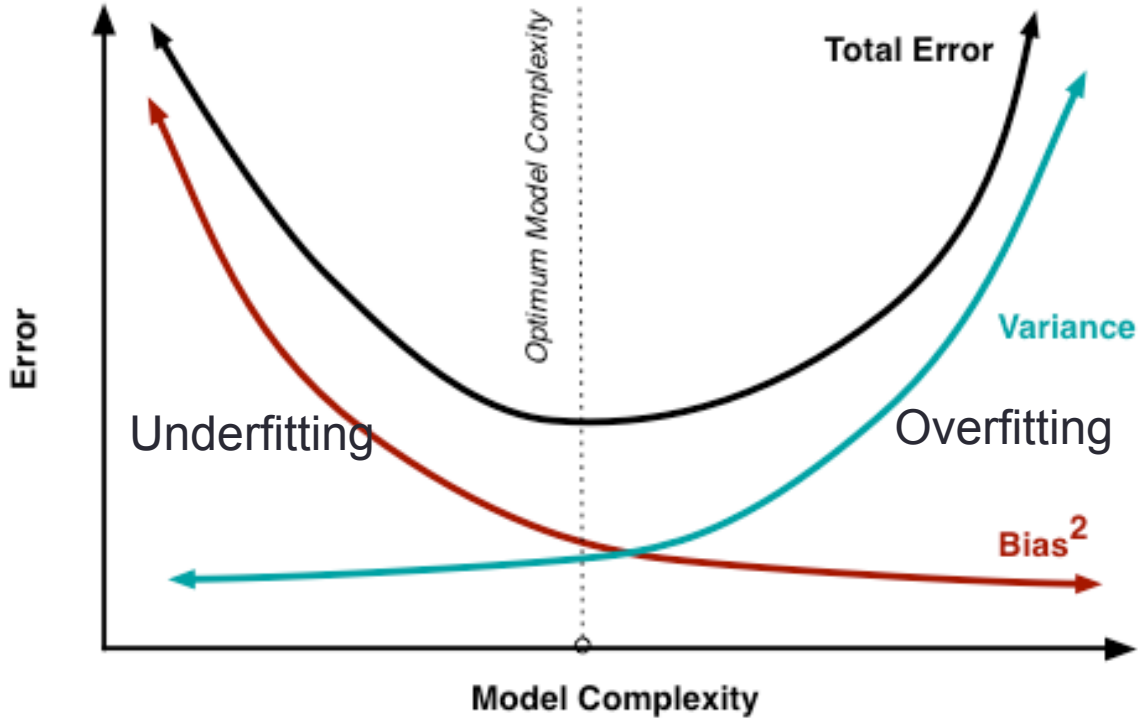
Short answer: Cross Validation

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

Understanding the bias variance trade-off

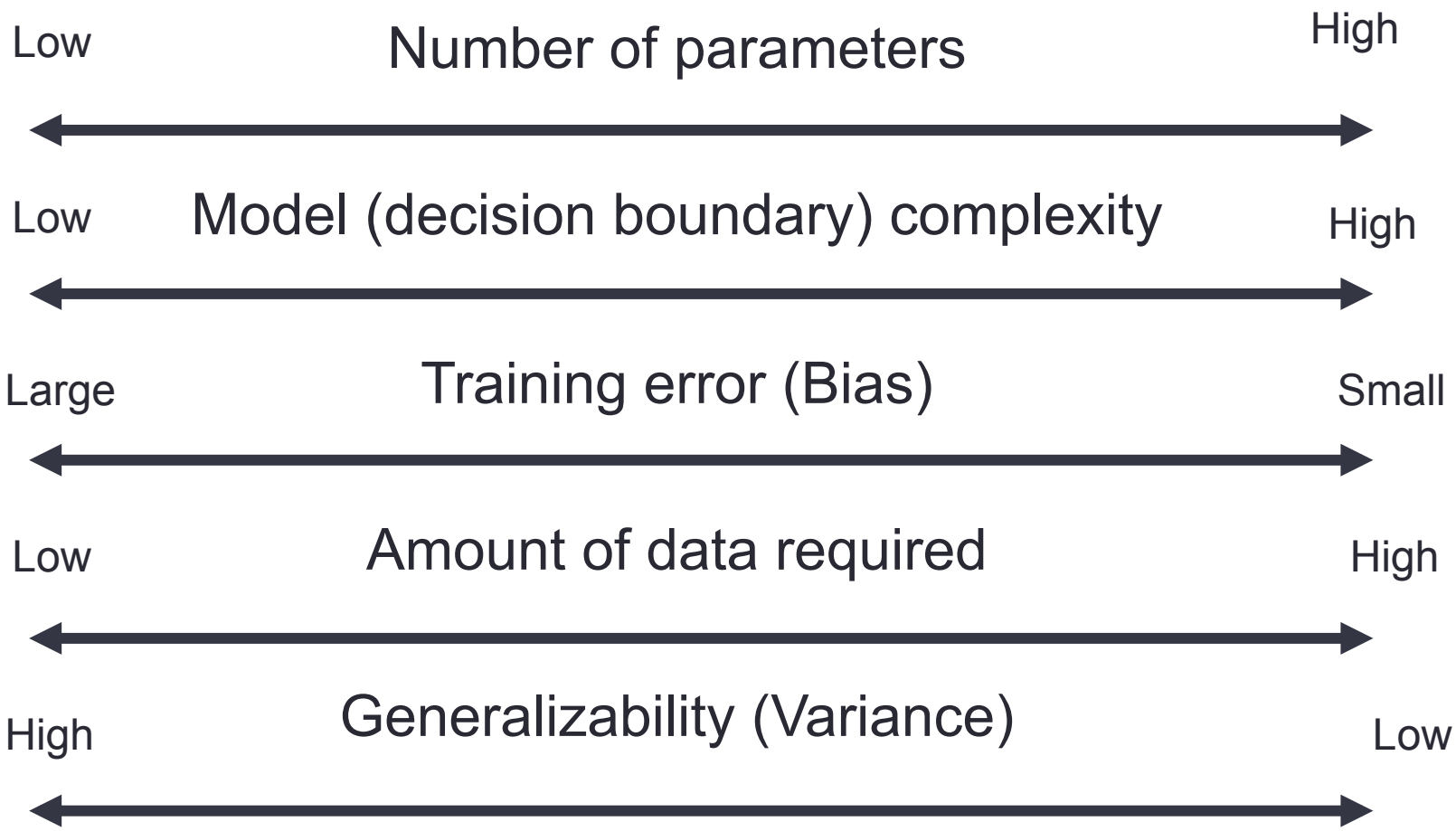
- Bias variance analysis can be helpful for diagnosis

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



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

Bias-Variance overview



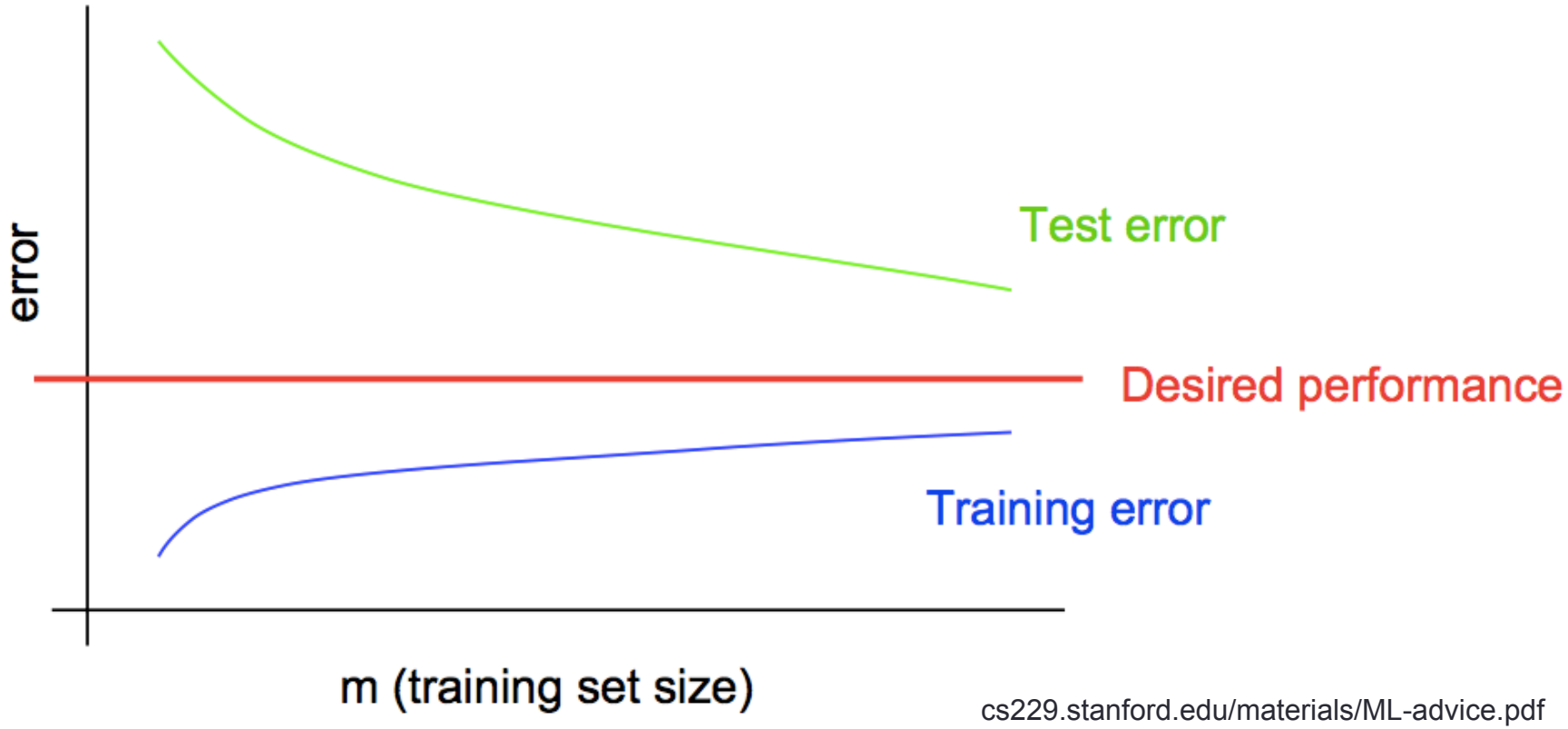
Bias-variance diagnosis

- 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

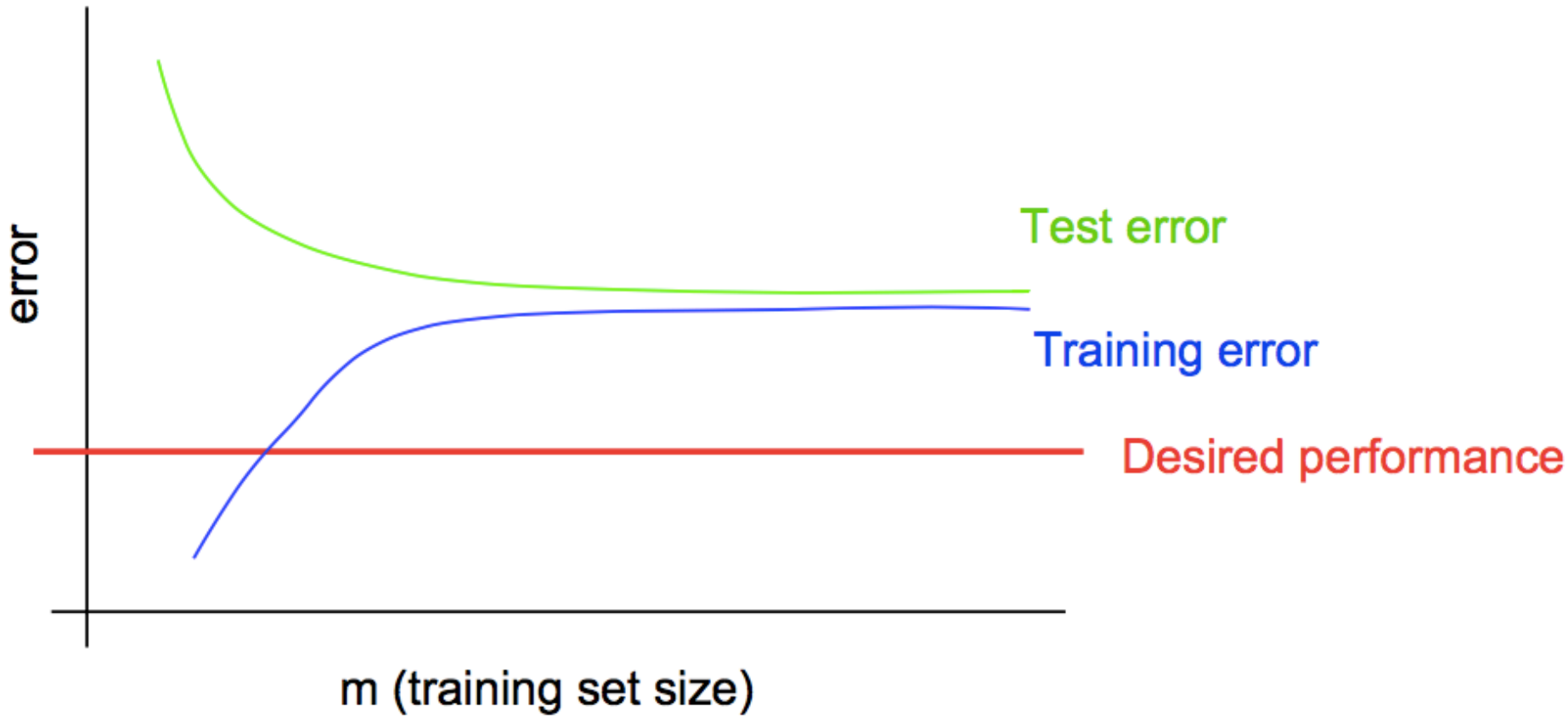
- Solution:
  - Reduce overfitting: regularization, reduce features, etc
  - Get more training set



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

- How to build a machine learning startup?

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/

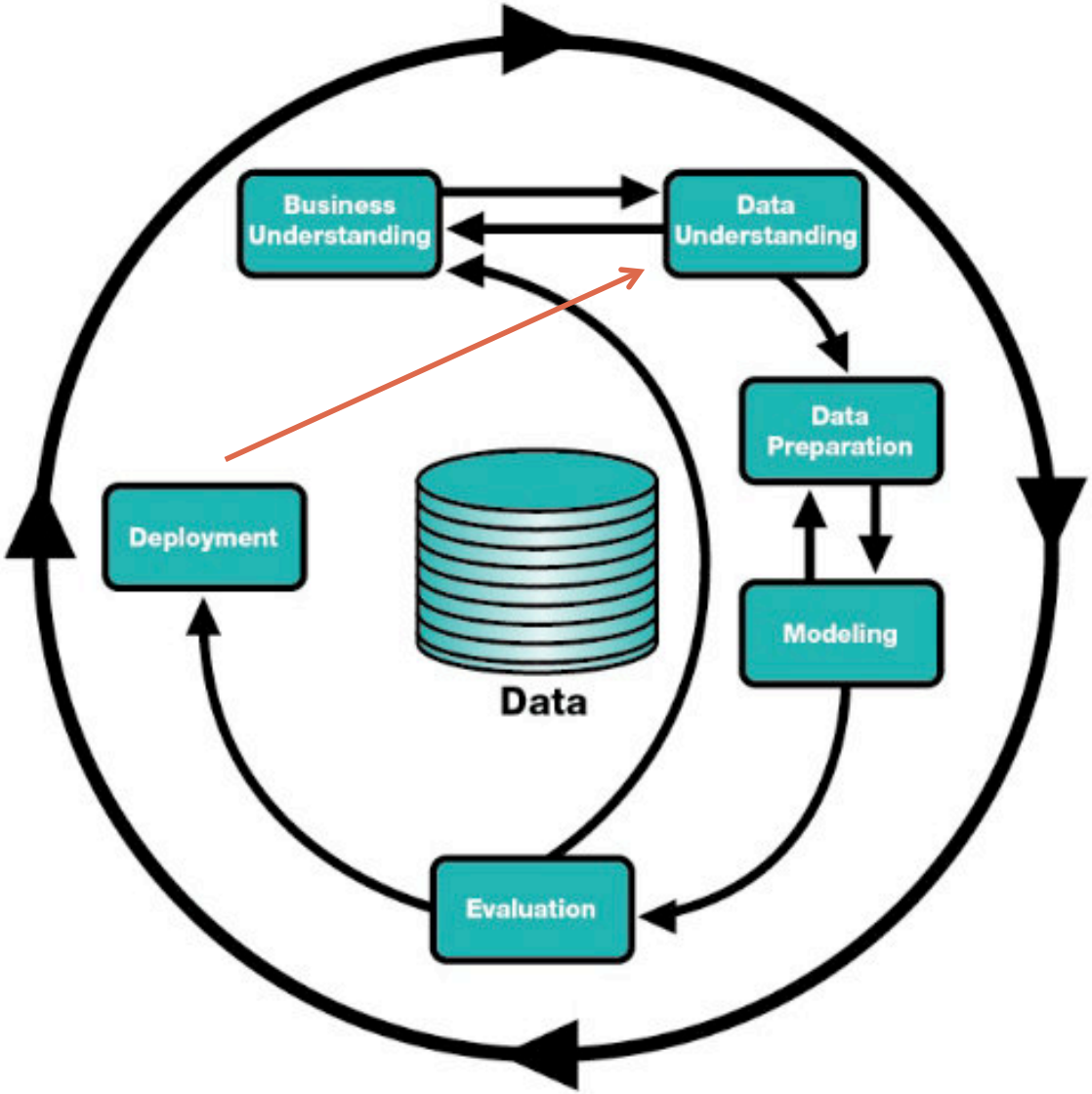
How did you spend your time on this competition?

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). Will data scientist lose our jobs?

But this is not all of data science

What does a data scientist do at work?



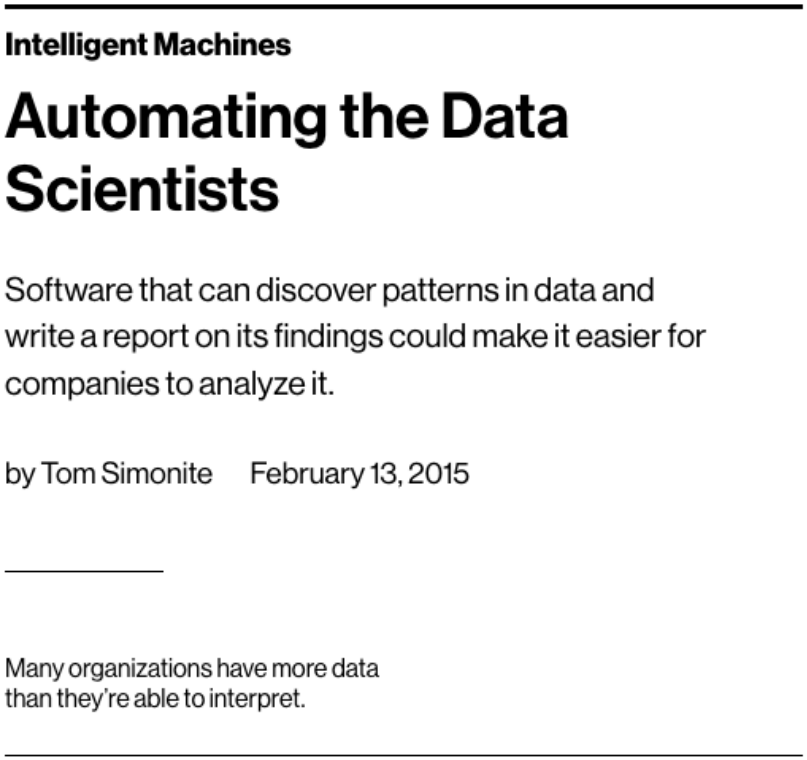
- Data science loop
- In industry:
  "data preparation and modeling (10%) and the rest (90%)"

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.

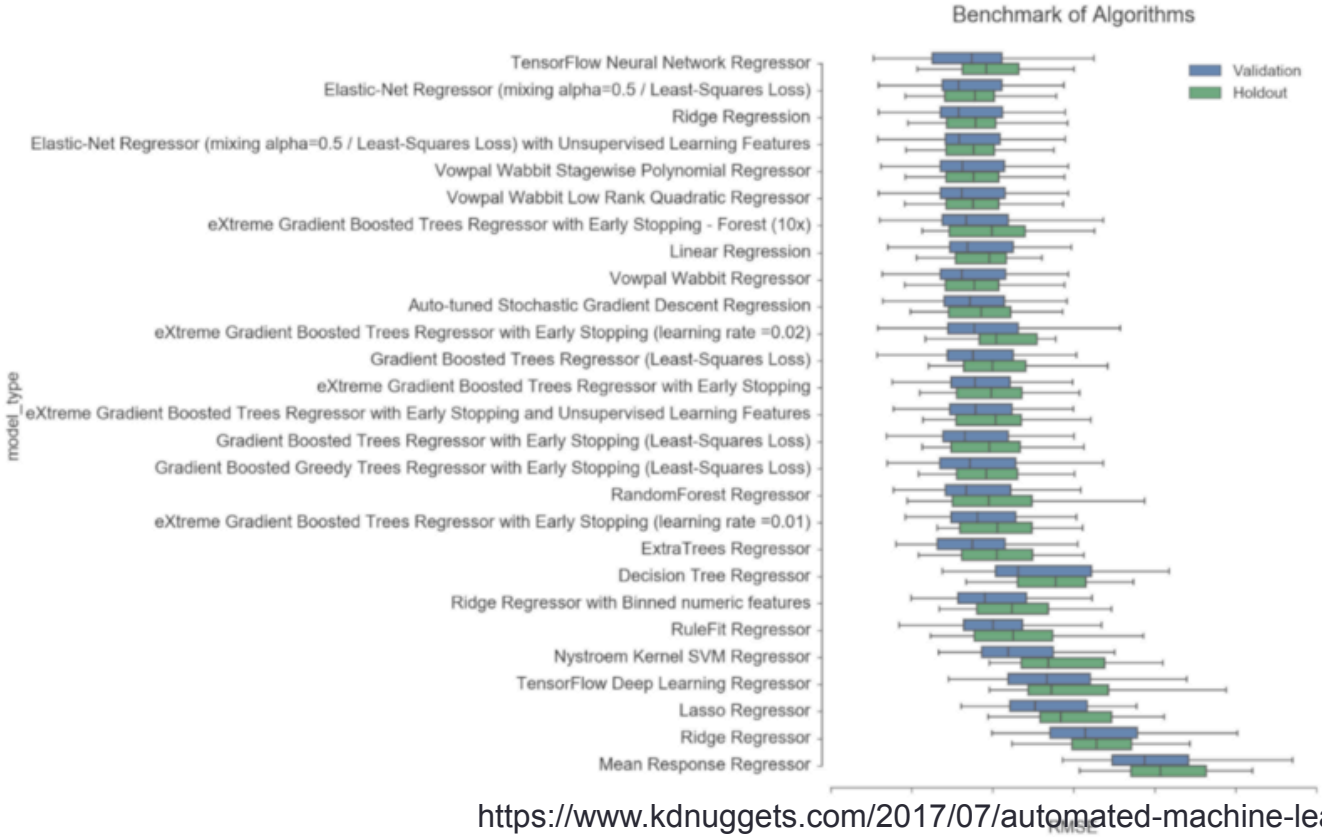
https://www.technologyreview.com/s/535041/automating-the-data-scientists/

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



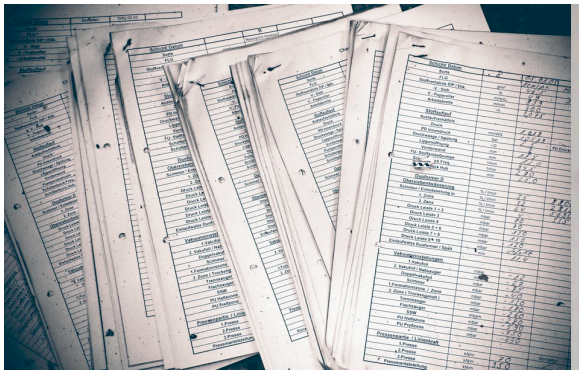
https://www.kdnuggets.com/2017/07/automated-machine-learning-paradigm-shift.html

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?

Build a machine learning application

- What are needed for Machine Learning in business applications?



Data

Business goal

Team

https://www.techemergence.com/roi-of-machine-learning-conse

Business goal

Pick the task

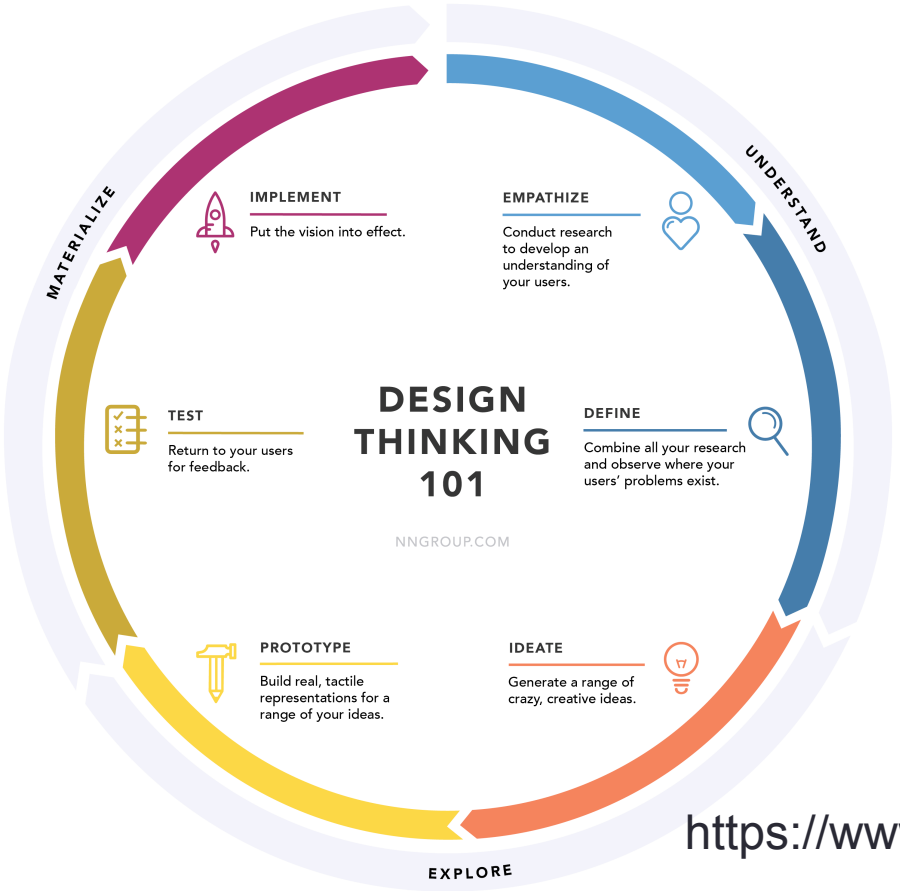
Pick the problem

- The goal can dictate the collection of data and the machine learning methods
- Treat it just like any other business plan
  - Machine Learning tasks are harder to evaluate the potential benefits
  - As machine learning gets automated, the bottleneck now is in **management, implementation, and business imagination**.

Machine learning is not a magic bullet

I hate to use this buzzword

- Design thinking
  - User centric thinking framework



https://www.nngroup.com/articles/design-thinking/

Example: Recommendation systems

- How can the recommendation system generates value?
- How can it be integrated into the existing UX flow?
- How can this feature differentiate you from competitors?
- What is a minimum requirement for the feature?

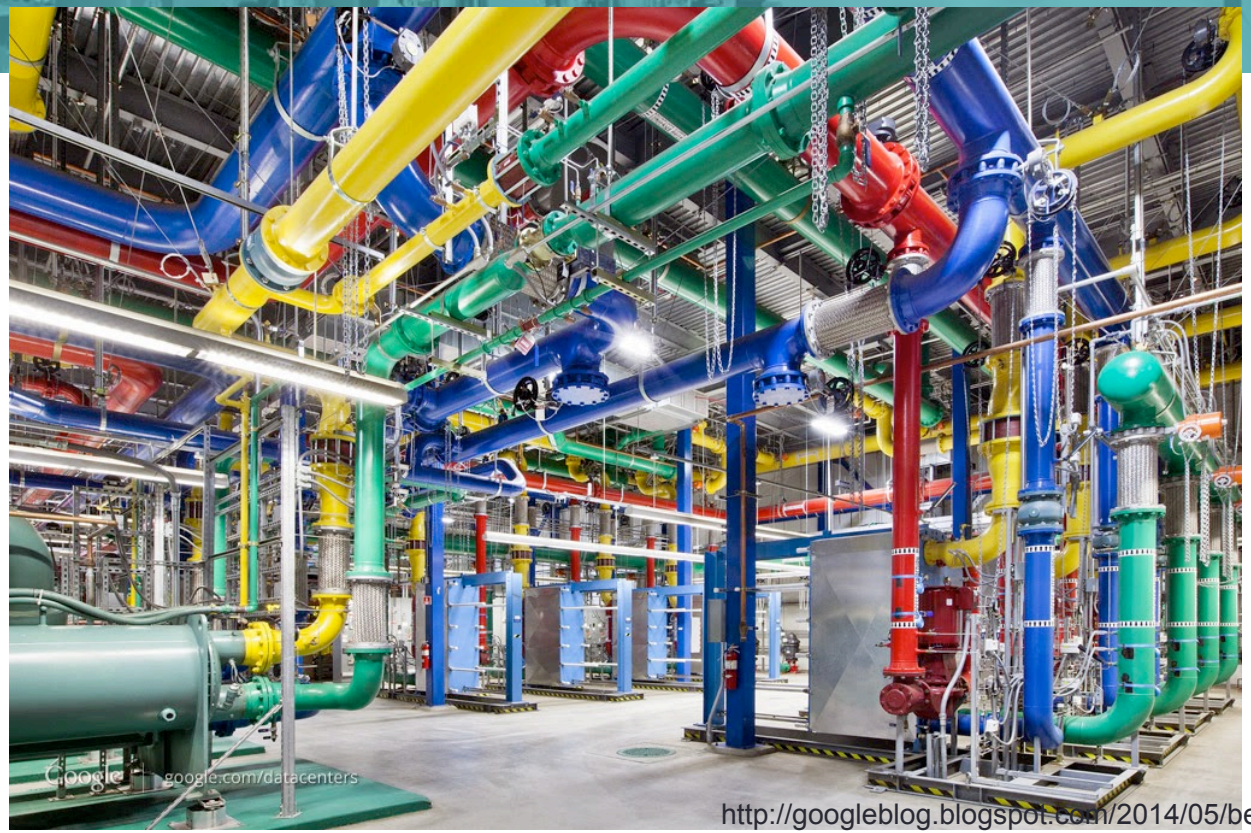
What tasks are good for machine learning?

- Tasks that requires less than 1 second to think
- ASR, Object recognition, Face recognition

- Augmenting human performance
- Healthcare diagnosis
- Self-driving cars

- Tasks at a large scale

DeepMind AI Reduces Google Data Centre Cooling Bill by 40%



http://googleblog.blogspot.com/2014/05/better-data-centers-through-machine.html

Aspects to consider

- Metrics
- Timings

Defining a metric

- Machine learning methods are evaluated based on metrics

Speech Recognition - Word Error Rate

Search – Recall rate

Machine Translation - BLEU score

Recommendation system – Click through rate

- Tasks can have multiple metrics
- Understand the relationship between the metric and the business goal

Example: Chatbots

- Task completion rate
- Number of turns to complete a task
- Satisfaction
- Word Error Rate (if speech input)

Understand the trade-offs

- Performance metrics
- Amount of computation power required
- Type of computation required
- Latency

Picking one guiding metric

- In order to develop fast, pick one summarizing metric to optimize
  - Precision + recall -> F-score
  - Weighted average of multiple metrics
- Task driven
- Instead of combining all metrics. Some metrics are optimizing metrics while some are constraints metrics
    - Must use less than X ram
    - Produce output in less than x milliseconds
    - Less than 1 false alarm per day
    - Note performance, metrics can sometimes be relaxed.
      - Easier to make something fast, than make something good.

Understand the timings

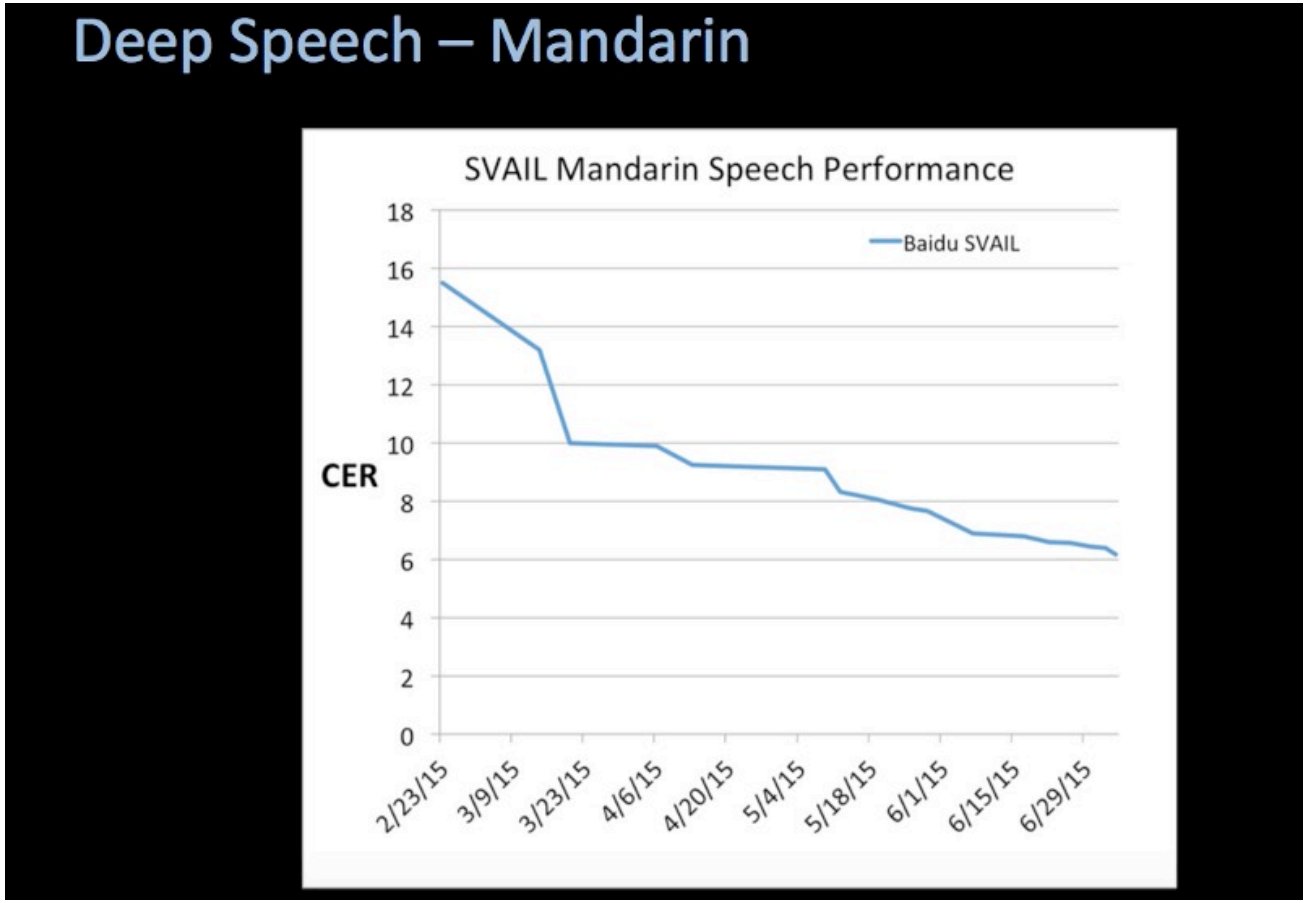
- Most data are not ready for machine learning
- Data preparation takes time

http://www.nvidia.com/object/drive-px.html

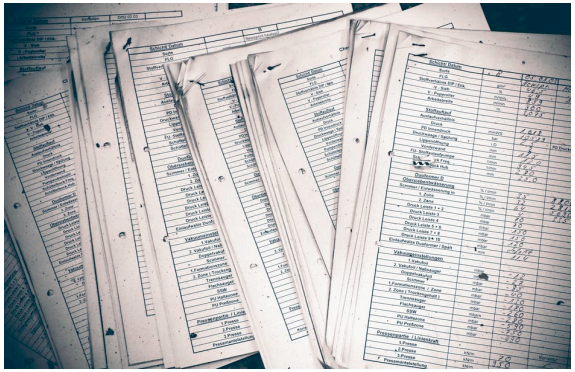
- Improving performance takes time

Performance and time spent on development

- 80/20 rule



https://medium.com/s-c-a-l-e/how-baidu-mastered-mandarin-with-deep-learning-and-lots-of-data-1d94032564a5

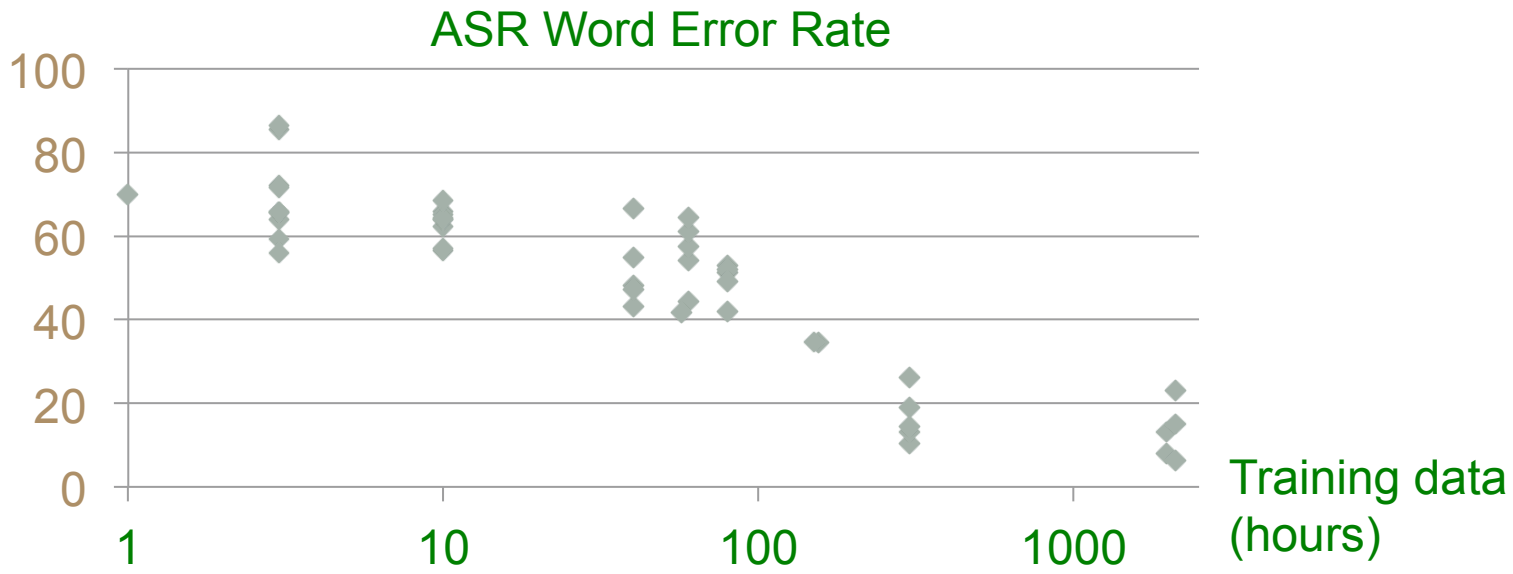


Data

Accumulate data

Machine learning uses data

- More data means better models



- You can start a machine learning project without data
- If the business plan says it's feasible

Making use of your data

- Train/dev/test split
  - 80/10/10, 90/5/5, 5-fold CV, leave one out CV, etc. for academia
- For real applications, get dev and test sets that represent your users.
  - Reflects the data you want to do well on.
  - There can be a mis-match between train and dev data. But avoid mis-match between dev and test data.
  - If no users, recruit friends to pretend to be the users.
- Example: Cat classifier.
  - Should you use ImageNet cat pictures as train/dev/test?
  - Go pretend you're a user and take cat pictures for the dev/test set.

Why match dev/test data

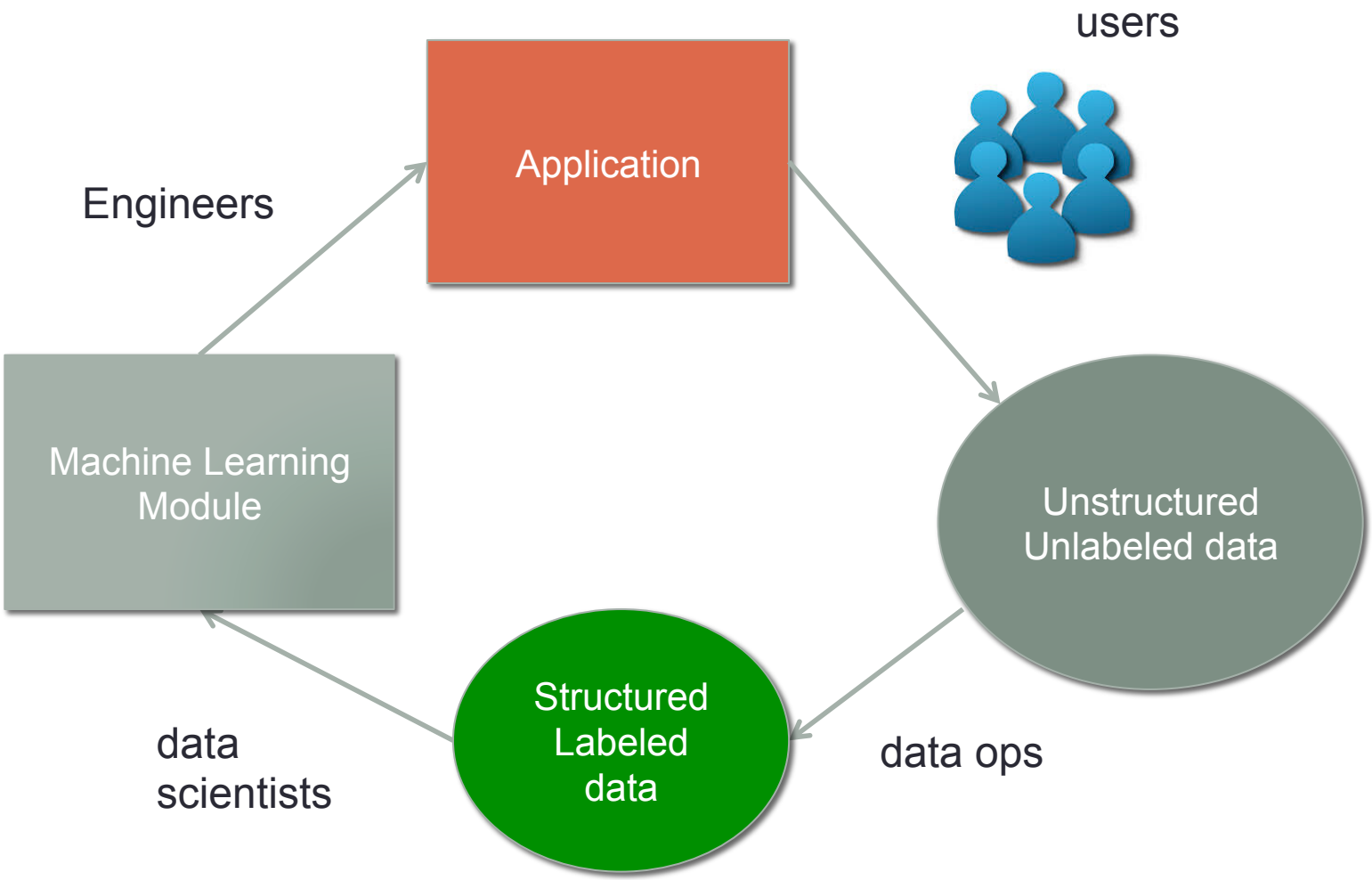
- You tune on the dev set.
- Scenario: do well on dev set, but bad on test set
- If dev and test set comes from the same distributions
  - Diagnosis: you overfit on the dev set. Get more dev data
- If dev and test set comes from different distributions
  - 3 possibilities
  - Overfit on the dev set
  - Test set is harder than the dev set
  - Dev and test set are just different. And what works on dev might not work on the test set. Wasted effort.

Dev and test set and size

- Dev - tune hyperparameters, select features, and make other decisions regarding the learning algorithm.
- Test - evaluate the performance of the algorithm, but not to make any decisions about regarding what learning algorithm or parameters to use.

- Dev – big enough to notice difference between algorithms (if you care about 0.1% difference, make sure you have enough dev set to spot it).
- Test – large enough to give confidence that your model will do well in real task

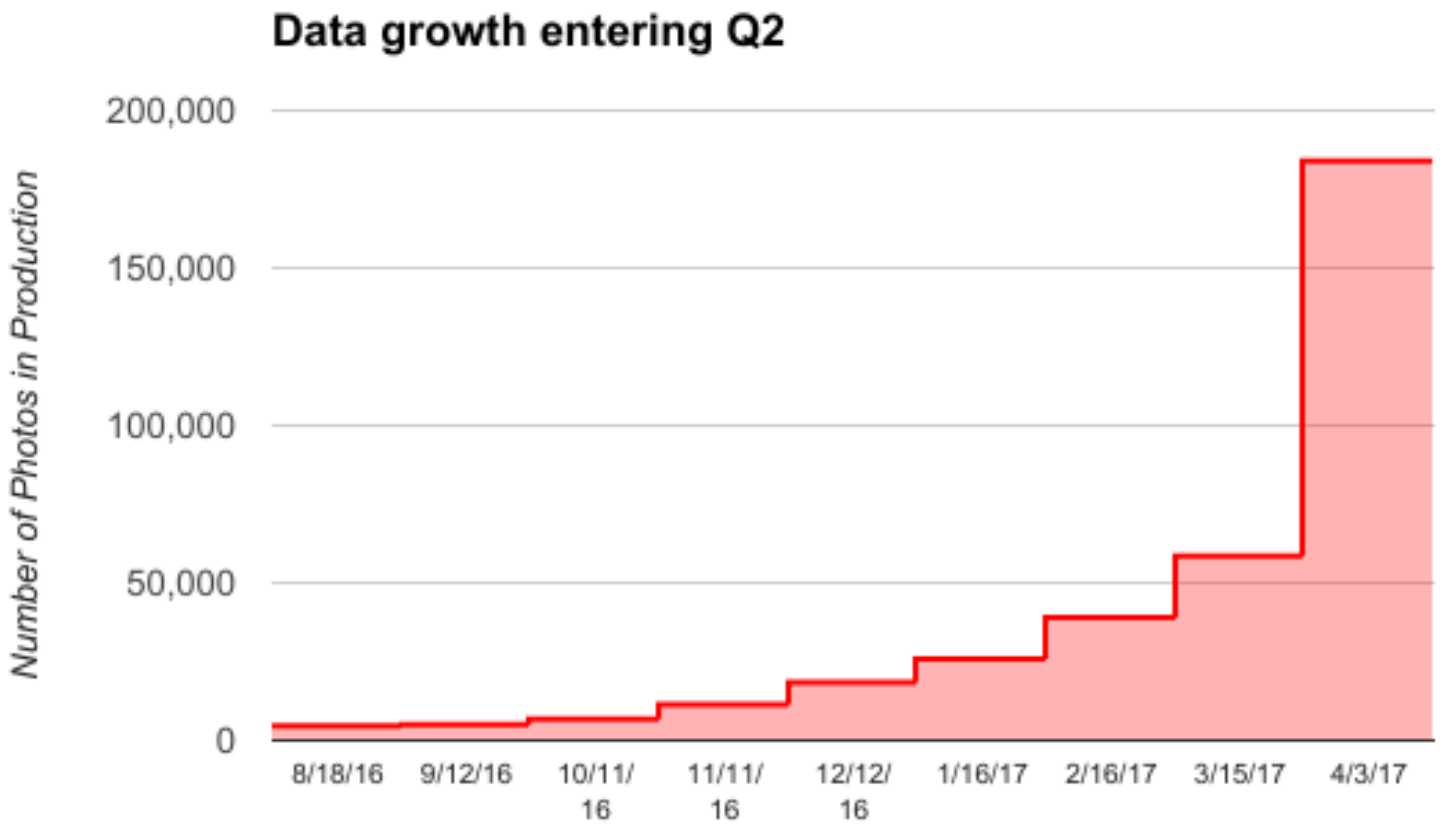
Getting into the data cycle



Build and fix

- Aim high, but launch fast, and iterate
  - Aim high: business goal must be sound, dictates what data to collect
  - Launch fast: get users and more training data. Gets error feedback fast, so you can diagnose and plan your efforts.
- Train on public domain datasets for fast training data acquisition (but dev/test must be real cases)
- Sometimes simple models are usually enough to launch
  - Logistic regression
  - Rule-based
    - Decision tress

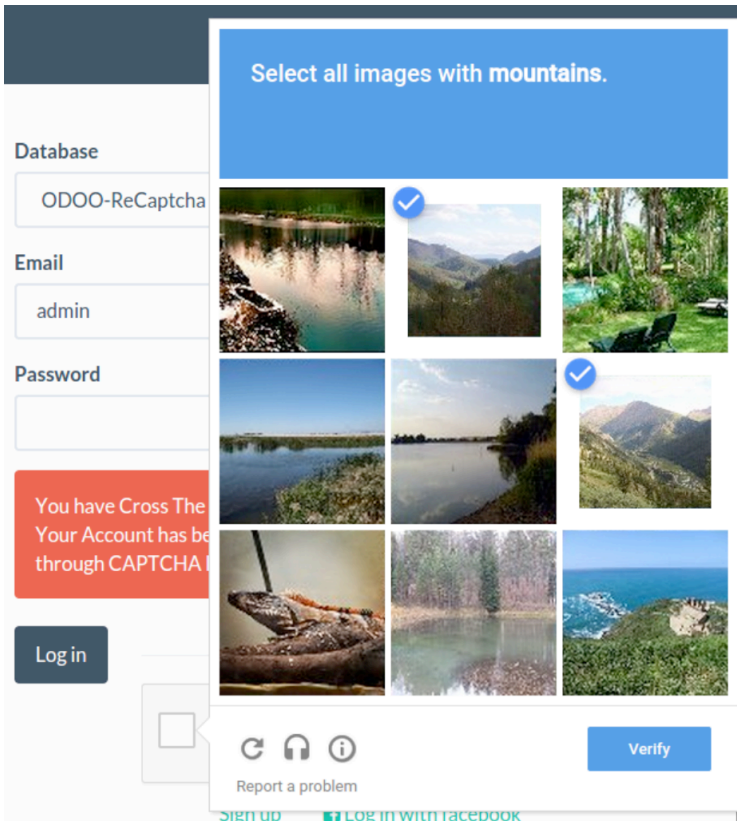
Getting into the data cycle

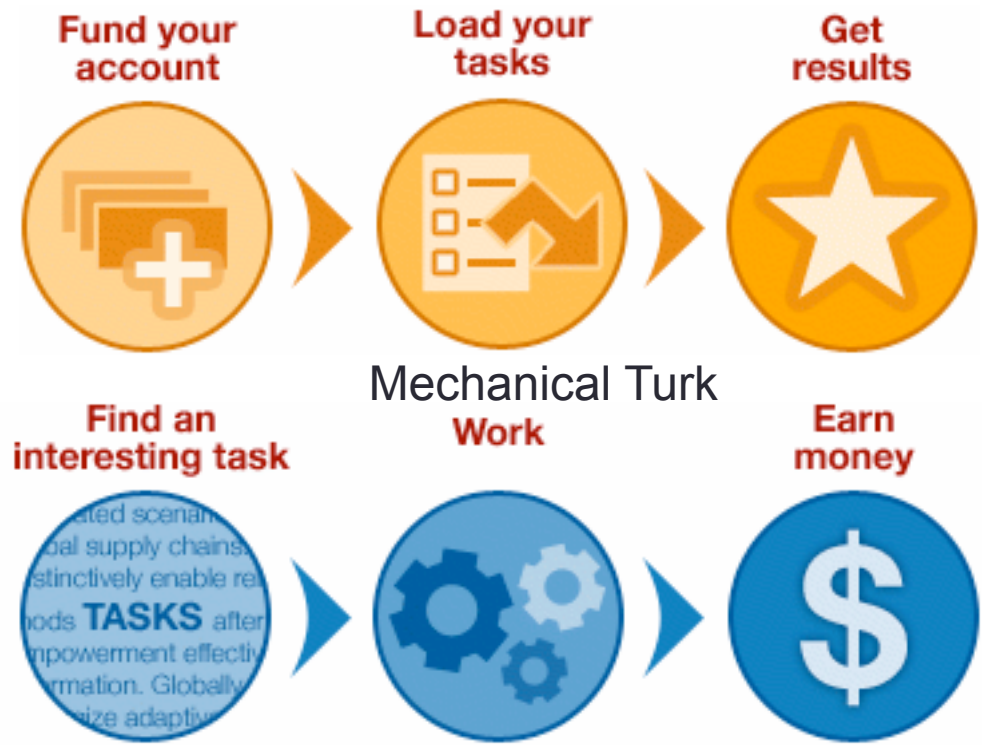


SMARTVID.IO

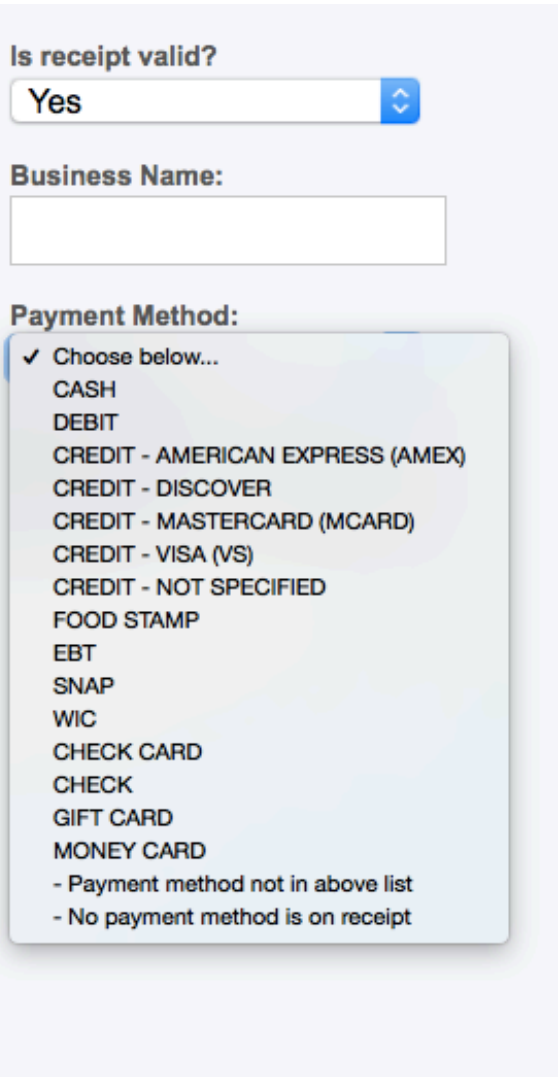
Crowdsourcing

- Distribute data labeling to the general public
- Need some incentives



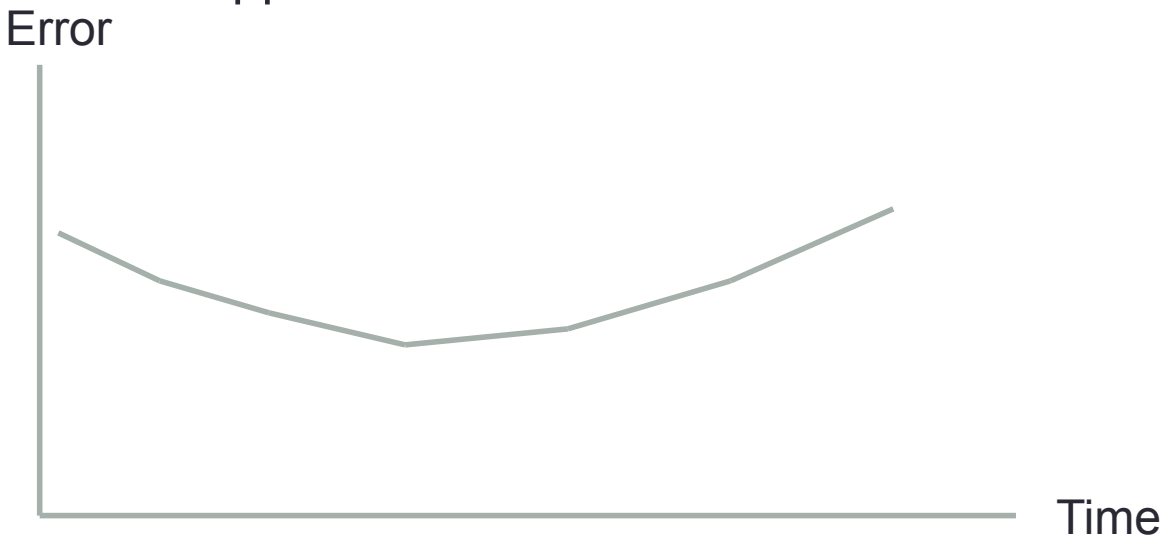


Mturk Example



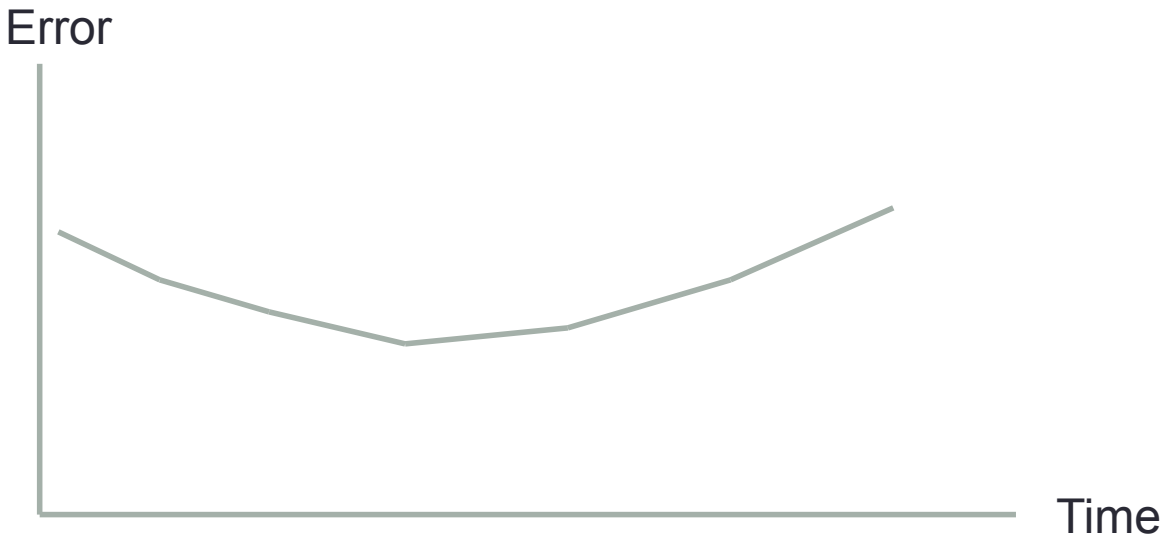
The rolling metric and datasets

- Example: You launched your cat classifier model
  - got into the data loop
  - used the new training data to train new models every month
  - then you noticed the test error keeps going up
  - What happened?



The rolling metric and datasets

- Your test no longer aligns with current usage
  - New phone cameras
  - New breed of cats(???)
  - New camera filters



The rolling dataset and metrics

- Train/dev/set must change over time to accommodate the change in usage
  - Example:
    - Test/dev: last month usage
    - Train: last year usage

- The metrics can also change. Don’t cling on the same metrics if your usage no longer aligns. This will waste your team efforts
  - New hardware, new requirements

Team

Build the team

The team

- Data science now require a diverse skill sets. Sometimes, multiple roles can be covered by one person

- Data scientists (modeling)
- Data ops/engineers (deployment, data ETL)
- Performance engineers (optimization)
- Domain expert (help guide modeling)
- Labeling team

- Compute

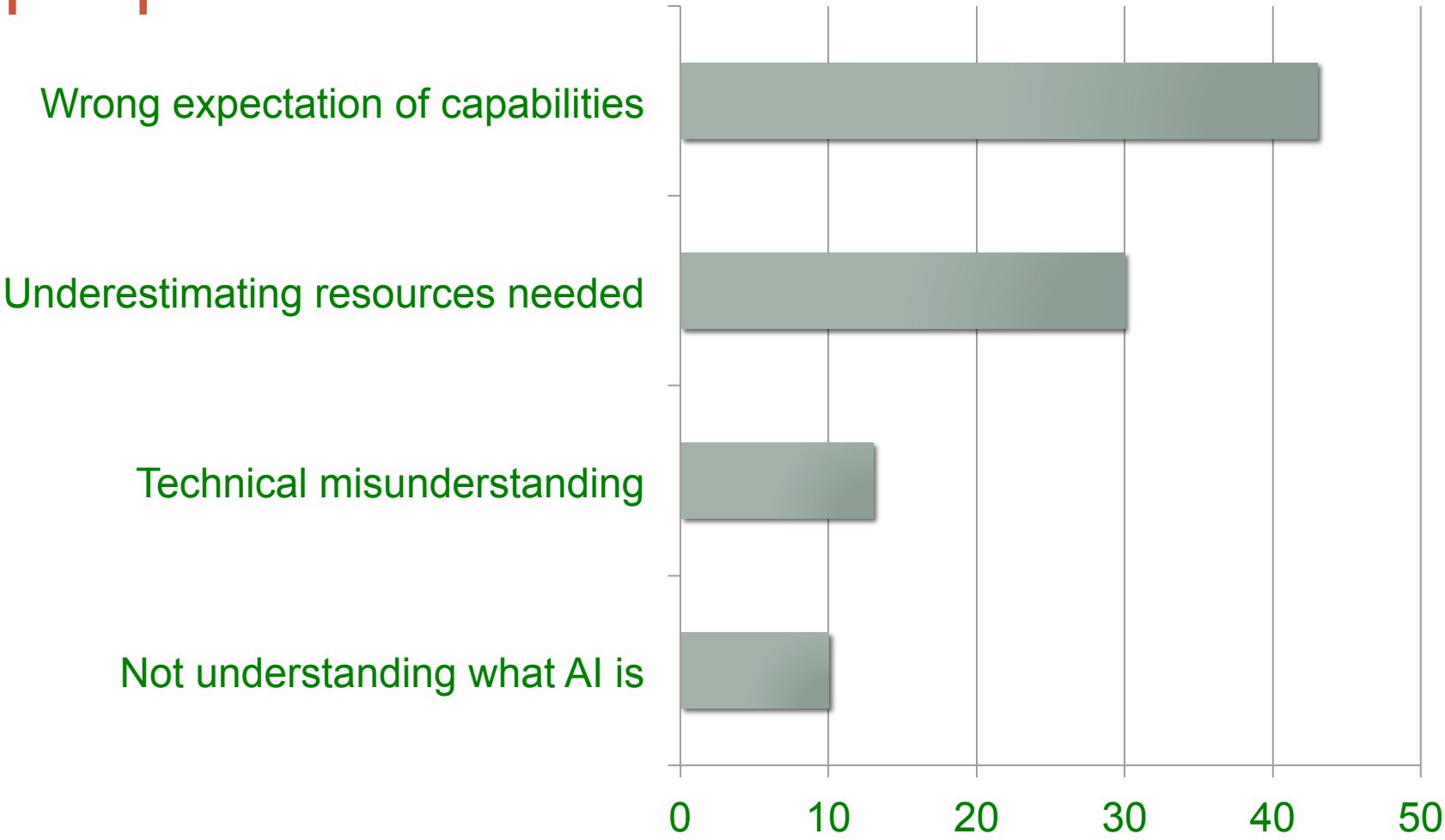
Summary

- Modeling: Picking a starting point
  - Lots of model. Pick depending on the task.
  - Always good to consult experts/papers
  - Go with simpler models first then do diagnosis
- Diagnosis: Guide your efforts efficiently when iterating
  - Bias-variance
  - Error analysis

Summary

- Be a good communicator
  - Find potential tasks that can benefit from data science
  - You will work as a team with non-data scientists
  - Practice by writing blogs/facebook posts explaining data science.

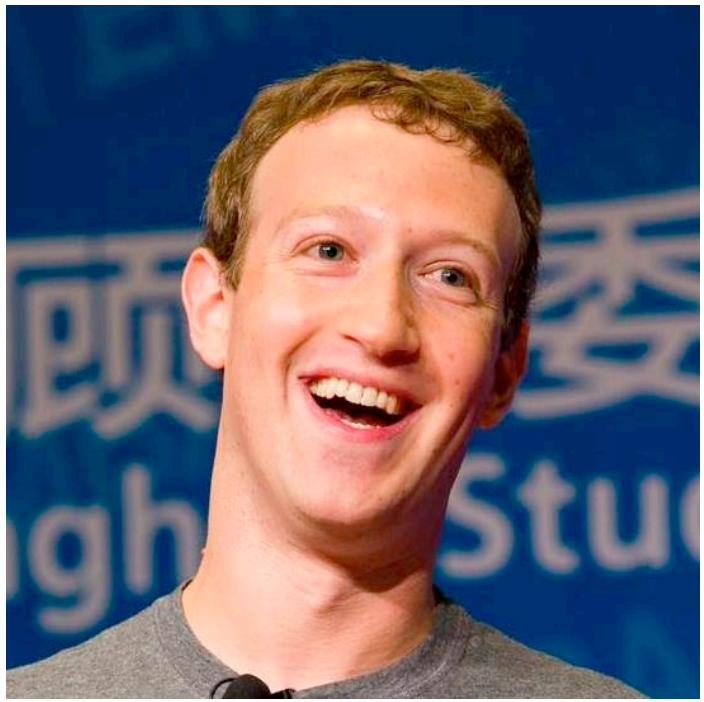
Common misconceptions business people have



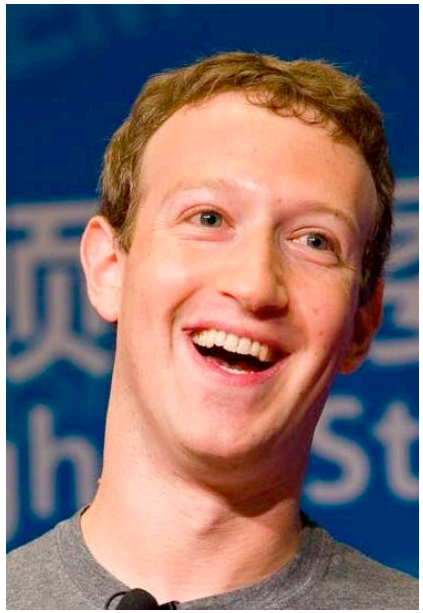
https://www.techemergence.com/roi-of-machine-learning-consensus/

- “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



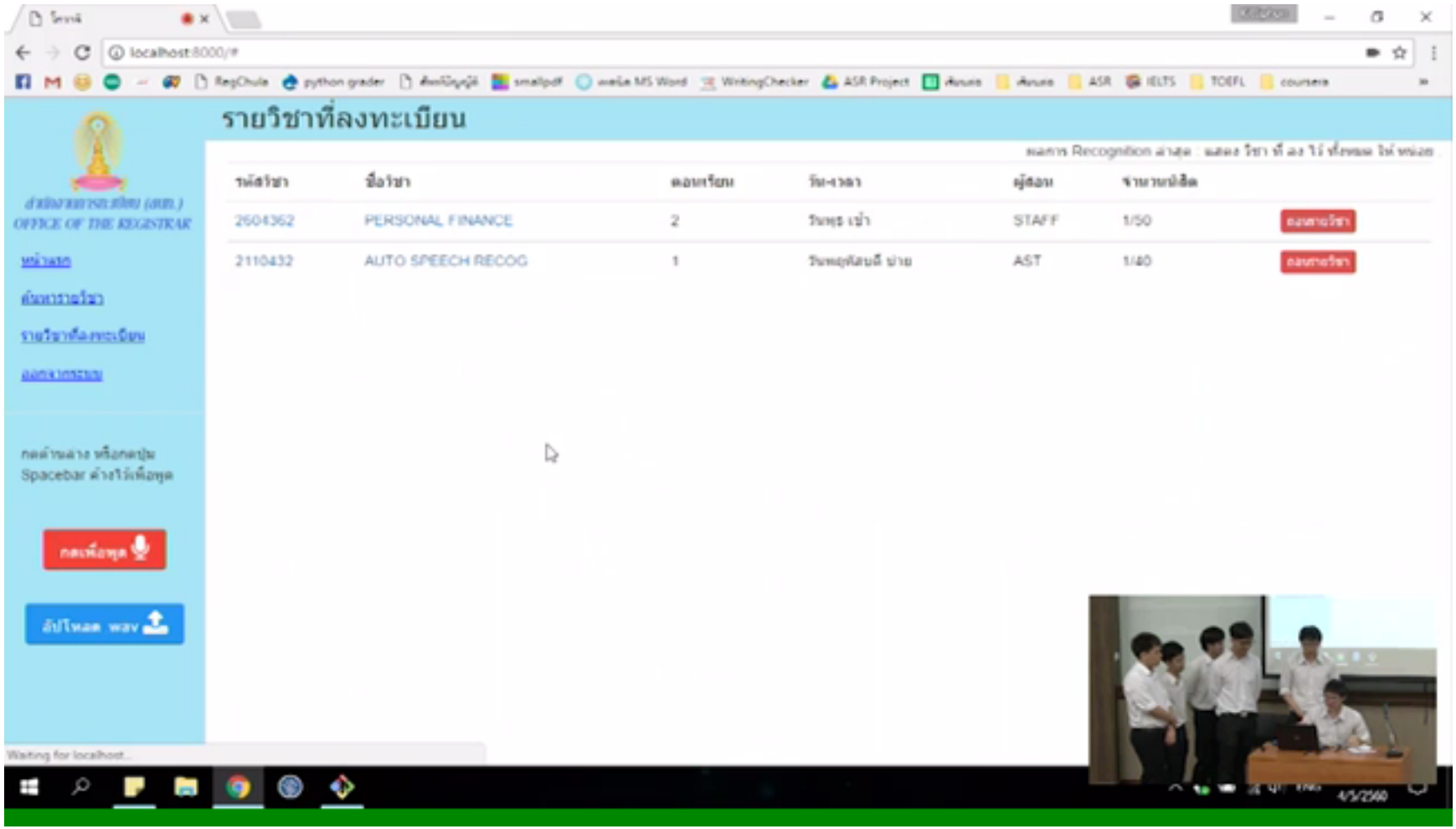
Beyond this course

- 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


localhost:8000/#
RegChula
python grader
smallpdf
WritingChecker
ASR Project
ASR
IELTS
TOEFL
coursera
รายวิชาที่ลงทะเบียน
OFFICE OF THE REGISTRAR
หน้าแรก
ค้นหารายวิชา
รายวิชาที่ลงทะเบียน
ออกจากระบบ
กดค้างเอาง หรือกดปุ่ม
Spacebar ค้างไว้เพื่อพูด
กดเพื่อพูด
อัปโหลด wav
รหัสวิชา
ชื่อวิชา
หน่วยกิต
วัน-เวลา
ผู้สอน
จำนวนที่นั่ง
2504362
PERSONAL FINANCE
2
STAFF
1/50
2110432
AUTO SPEECH RECOG
1
AST
1/40
Waiting for localhost...
4/5/2560

Natural language processing

- Conditional Random Fields
- Recursive Neural Networks
- Basics of language understanding
- Chatbots

- Take both for best experience!

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