Lecture 10

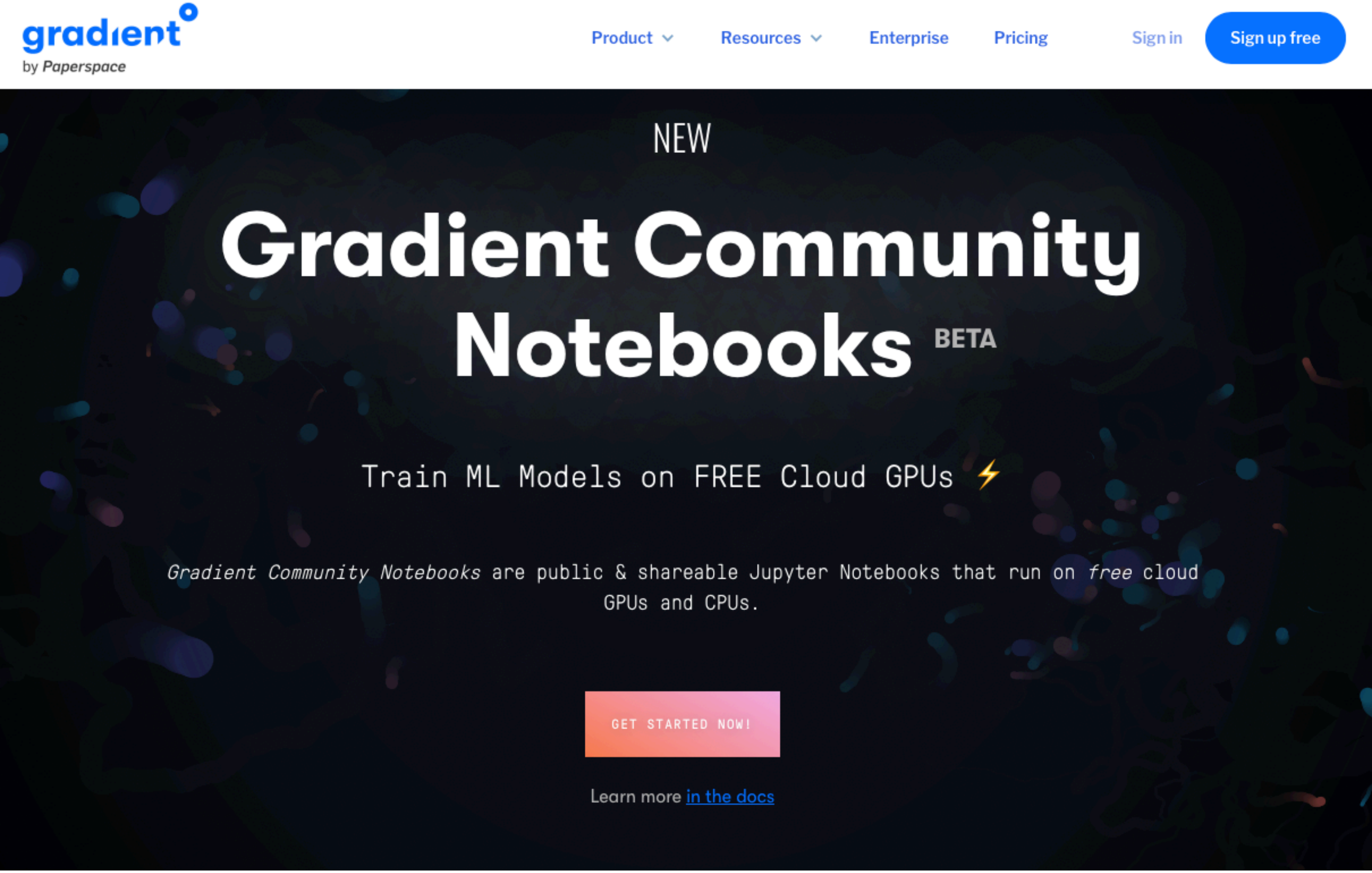
Model Evaluation 3:
Cross Validation

STAT 479: Machine Learning, Fall 2018

Sebastian Raschka

http://stat.wisc.edu/~sraschka/teaching/stat479-fs2018/

Some News



https://gradient.paperspace.com/free-gpu

Happy Halloween!

deeplearning.ai

THE BATCH

October 30, 2019 Essential news for deep learners

Subscribe Tips

https://www.deeplearning.ai/thebatch/

Happy Halloween!

Visions of Machine Learning AI gone wrong

& Malevolent superintelligences

1) AI Goes Rogue

- Code awakens into sentience, will be able to think better than humans (btw. have you heard of Movarek's paradox?)

- Computers already “think” much faster than humans; they remember more information than humans, too

- Artificial intelligence already manages crucial systems in fields like finance, security, and communications

- It will enslave or exterminate our species (https://www.nytimes.com/2009/07/26/science/26robot.html)

1) AI Goes Rogue

- Code awakens into sentience, will be able to think better than humans (btw. have you heard of Movarek's paradox?)

How scared should we be?

- Artificial intelligence already manages crucial systems in fields like finance, security, and communications

- It will enslave or exterminate our species (https://www.nytimes.com/2009/07/26/science/26robot.html)

2) Deepfakes Wreak Havoc

- Deepfakes of celebrities (https://www.engadget.com/2019/10/11/deepfake-celebrity-impresonations);

- GPT-2 language model's ability to churn out faux articles that convince readers they're from the New York Times (https://www.foreignaffairs.com/articles/2019-08-02/not-your-fathers-bots)

2) Deepfakes Wreak Havoc

- Deepfakes of celebrities (https://www.engadget.com/2019/10/11/deepfake-celebrity-impresonations);

How scared should we be?

www.foreignaffairs.com/articles/2019-08-02/not-your-fathers-bots)

2) Deepfakes Wreak Havoc

- Deepfakes of celebrities (https://www.engadget.com/2019/10/11/deepfake-celebrity-impresonations);
- Fake videos of politicians (https://www.foreignaffairs.com/articles/2019-08-02/not-your-fathers-bots)

How scared should we be?

https://www.biometricupdate.com/201905/researchers-develop-digital-watermarks-to-detect-deepfakes

3) No Escape from Surveillance

- Artificial intelligence will boost the power of surveillance, effectively making privacy obsolete

- Smartphone applications track your location, browsing history, and even mine your contact data, thanks to the permissions you give them in exchange for free apps.

- More than half of all US companies monitor their employees — including email monitoring and biometric tracking (https://www.gartner.com/smarterwithgartner/the-future-of-employee-monitoring)

- AI surveillance is used by local, state, or national governments in over 40 percent of the world's countries (https://carnegieendowment.org/2019/09/17/global-expansion-of-ai-surveillance-pub-79847)

3) No Escape from Surveillance

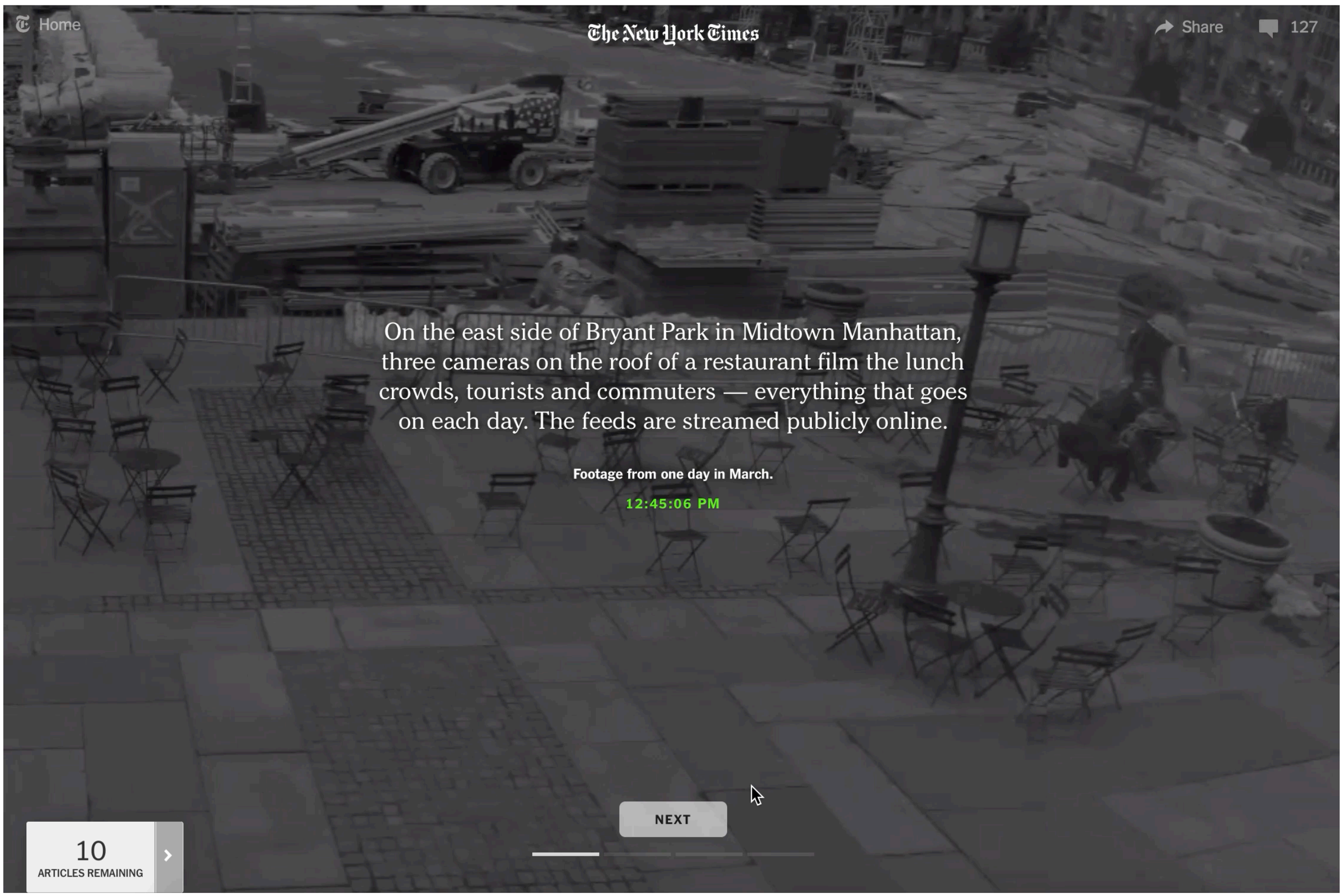
- Artificial intelligence will boost the power of surveillance, effectively making privacy obsolete

- Smartphone applications track your location, browsing history, and even mine your contact data, thanks to the permissions

How scared should we be?

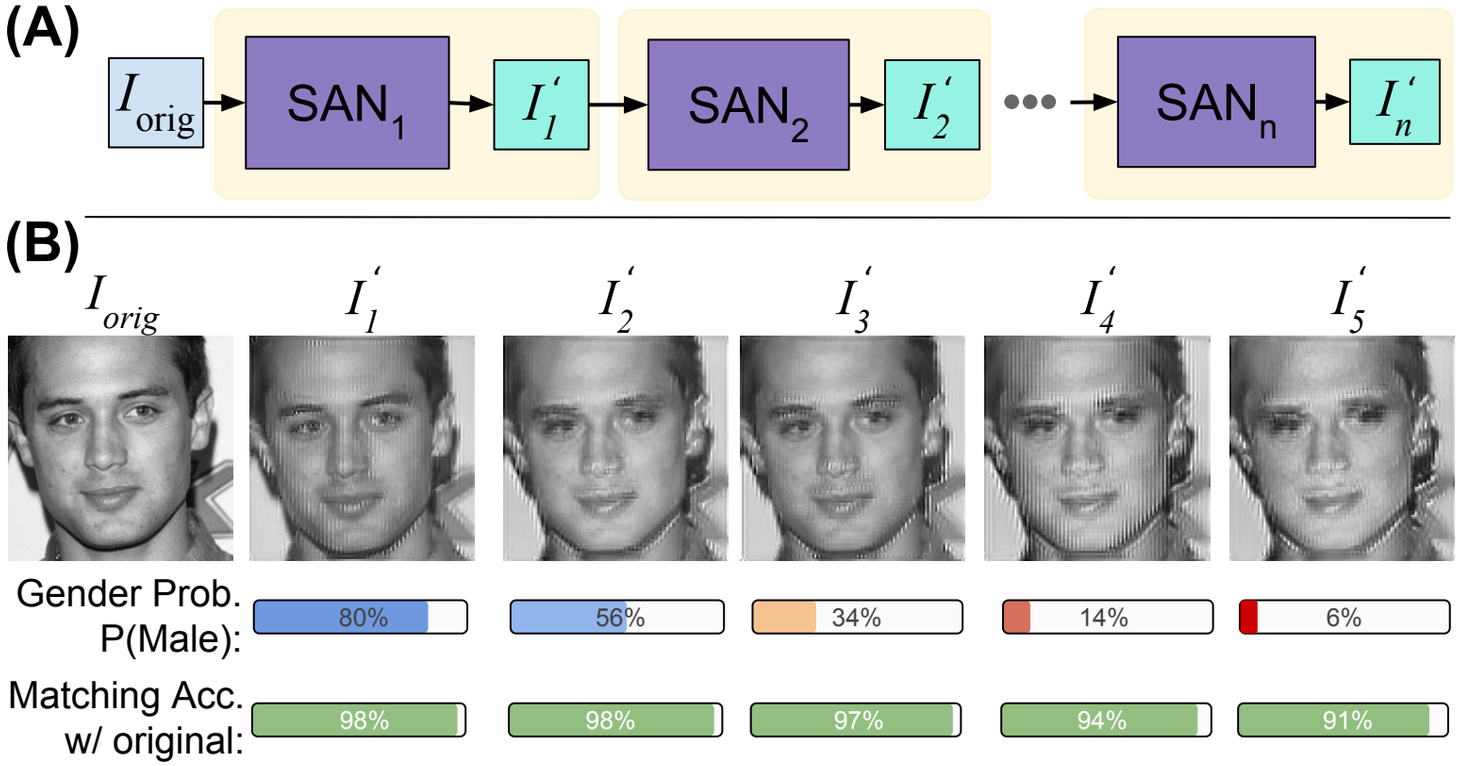
- More than half of all US companies monitor their employees including email monitoring and biometric tracking (https://www.gartner.com/smarterwithgartner/the-future-of-employee-monitoring)

- AI surveillance is used by local, state, or national governments in over 40 percent of the world's countries (https://carnegieendowment.org/2019/09/17/global-expansion-of-ai-surveillance-pub-79847)



https://www.nytimes.com/interactive/2019/04/16/opinion/facial-recognition-new-york-city.html

FlowSAN: Privacy-enhancing Semi-Adversarial Networks to Confound Arbitrary Face-based Gender Classifiers



Improvements to better control the perturbations and enhance the removal of soft-biometric information

Vahid Mirjalili, Sebastian Raschka, Arun Ross (2019)
FlowSAN: Privacy-enhancing Semi-Adversarial Networks to Confound Arbitrary Face-based Gender Classifiers
IEEE Access 2019, 10.1109/ACCESS.2019.2924619

4) Biased Data Trains Oppressive AI

- AI learns from data to reach its own conclusions

- But training datasets are often gathered from and curated by humans who have social biases.

- The risk that AI will reinforce existing social biases is rising as the technology increasingly governs education, employment, loan applications, legal representation, and press coverage.

Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification

[edit]

Joy Buolamwini, Timnit Gebru ; Proceedings of the 1st Conference on Fairness, Accountability and Transparency, PMLR 81:77-91, 2018.

Abstract

Recent studies demonstrate that machine learning algorithms can discriminate based on classes like race and gender. In this work, we present an approach to evaluate bias present in automated facial analysis algorithms and datasets with respect to phenotypic subgroups. Using the dermatologist approved Fitzpatrick Skin Type classification system, we characterize the gender and skin type distribution of two facial analysis benchmarks, IJB-A and Adience. We find that these datasets are overwhelmingly composed of lighter-skinned subjects (79.6% for IJB-A and 86.2% for Adience) and introduce a new facial analysis dataset which is balanced by gender and skin type. We evaluate 3 commercial gender classification systems using our dataset and show that darker-skinned females are the most misclassified group (with error rates of up to 34.7%). The maximum error rate for lighter-skinned males is 0.8%. The substantial disparities in the accuracy of classifying darker females, lighter females, darker males, and lighter males in gender classification systems require urgent attention if commercial companies are to build genuinely fair, transparent and accountable facial analysis algorithms.

Related Material

http://proceedings.mlr.press/v81/buolamwini18a.html

4) Biased Data Trains Oppressive AI

- AI learns from data to reach its own conclusions
- AI learns from data that is labeled by humans who have social biases.

How scared should we be?

- The risk that AI will reinforce existing social biases is rising as the technology increasingly governs education, employment, loan applications, legal representation, and press coverage.

Gender Privacy: An Ensemble of Semi Adversarial Networks for Confounding Arbitrary Gender Classifiers

Improvements to construct a more diverse set of SAN models for better generalizability via ensembling

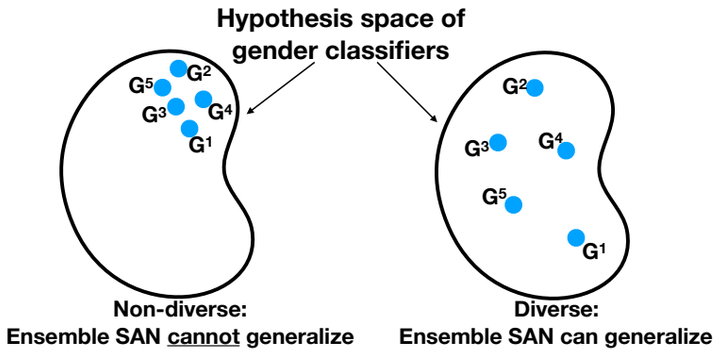


Figure 1: Diversity in an ensemble SAN can be enhanced through its auxiliary gender classifiers (see Figure 2). When the auxiliary gender classifiers lack diversity, ensemble SAN cannot generalize well to arbitrary gender classifiers.

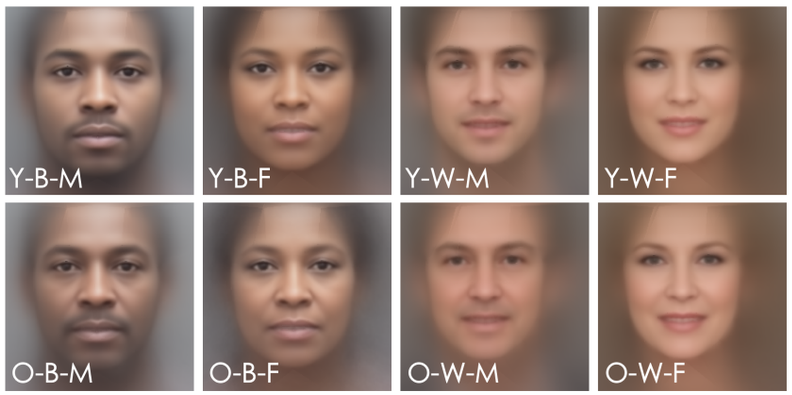


Figure 4: Face prototypes computed for each group of attribute labels. The abbreviations at the bottom of each image refer to the prototype attribute-classes, where Y=young, O=old, M=male, F=female, W=white, B=black.

Vahid Mirjalili, Sebastian Raschka, and Arun Ross (2018) *Gender Privacy: An Ensemble of Semi Adversarial Networks for Confounding Arbitrary Gender Classifiers.* 9th IEEE International Conference on Biometrics: Theory, Applications, and Systems (BTAS 2018)

5) Machines Take Everyone's Jobs

- AI will exceed human performance at a wide range of activities. Huge populations will become jobless

- Historically, technology created more jobs than it destroyed. What makes AI different is it threatens to outsource the one thing humans have always relied on for employment: their brains.

- Massive unemployment in the past have brought severe social disruption. The U.S. Great Depression in the 1930s saw jobless rates above 34 percent.

5) Machines Take Everyone's Jobs

- AI will exceed human performance at a wide range of activities. Huge populations will become jobless

- thing humans have always relied on for employment: their brains.

- Massive unemployment in the past have brought severe social disruption. The U.S. Great Depression in the 1930s saw jobless rates above 34 percent.

How scared should we be?

5) Machines Take Everyone's Jobs

- AI will exceed human performance at a wide range of activities. Huge populations will become jobless
- thing humans have always relied on for employment: their brains

How scared should we be?

Lifelong learning is a front-line defense (and a rewarding pursuit!). Education can help you stay ahead of partial automation in your current profession or change lanes if your profession is being automated away.

6) AI Winter Sets In

- Could the flood of hype for artificial intelligence lead to a catastrophic collapse in funding?

- AI will fail to deliver on promises inflated by businesses and researchers

- Funding will dry up, research will sputter, and progress will stall

6) AI Winter Sets In

- Could the flood of hype for artificial intelligence lead to a

How scared should we be?

- AI will fail to deliver on promises inflated by businesses and researchers

- Funding will dry up, research will sputter, and progress will stall

6) AI Winter Sets In

How scared should we be?

- Could the flood of hype for artificial intelligence lead to a catastrophic collapse in funding?

- **As AI practitioners, we should strive to present our work honestly, criticize one another fairly and openly, and promote projects that demonstrate clear value. Genuine progress in improving peoples' lives is the best way to ensure that AI enjoys perpetual springtime.**

deeplearning.ai

THE BATCH

October 30, 2019 Essential news for deep learners

Subscribe Tips

https://www.deeplearning.ai/thebatch/

Lecture 10

Model Evaluation 3:
Cross Validation

STAT 479: Machine Learning, Fall 2018

Sebastian Raschka

http://stat.wisc.edu/~sraschka/teaching/stat479-fs2018/

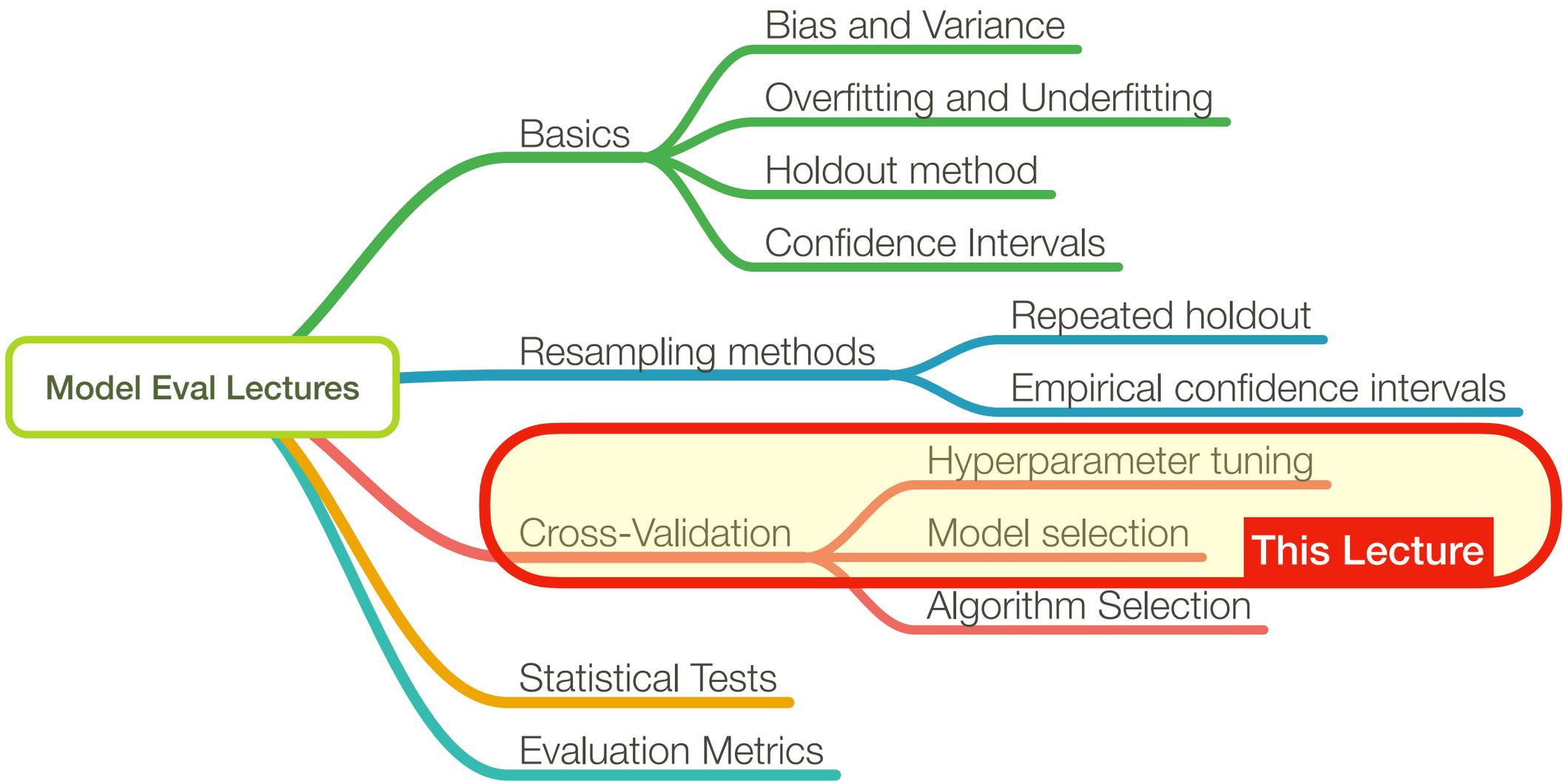
- **Performance estimation**
  - Large dataset
    - 2-way holdout method (train/test split)
    - Confidence interval via normal approximation
  - Small dataset
    - (Repeated) k-fold cross-validation without independent test set
    - Leave-one-out cross-validation without independent test set
    - Confidence interval via 0.632(+) bootstrap
- **Model selection (hyperparameter optimization) and performance estimation**
  - Large dataset
    - 3-way holdout method (train/validation/test split)
  - Small dataset
    - (Repeated) k-fold cross-validation with independent test set
    - Leave-one-out cross-validation with independent test set
- **Model & algorithm comparison**
  - Large dataset
    - Multiple independent training sets + test sets (algorithm comparison, AC)
    - McNemar test (model comparison, MC)
    - Cochran's Q + McNemar test (MC)
  - Small dataset
    - Combined 5x2cv *F* test (AC)
    - Nested cross-validation (AC)

These are my personal recommendations;

MC = model comparison,
AC = algorithm comparison;

Terms that are still unfamiliar (McNemar's test, 5x2cv F test, etc.) will be covered next lecture.

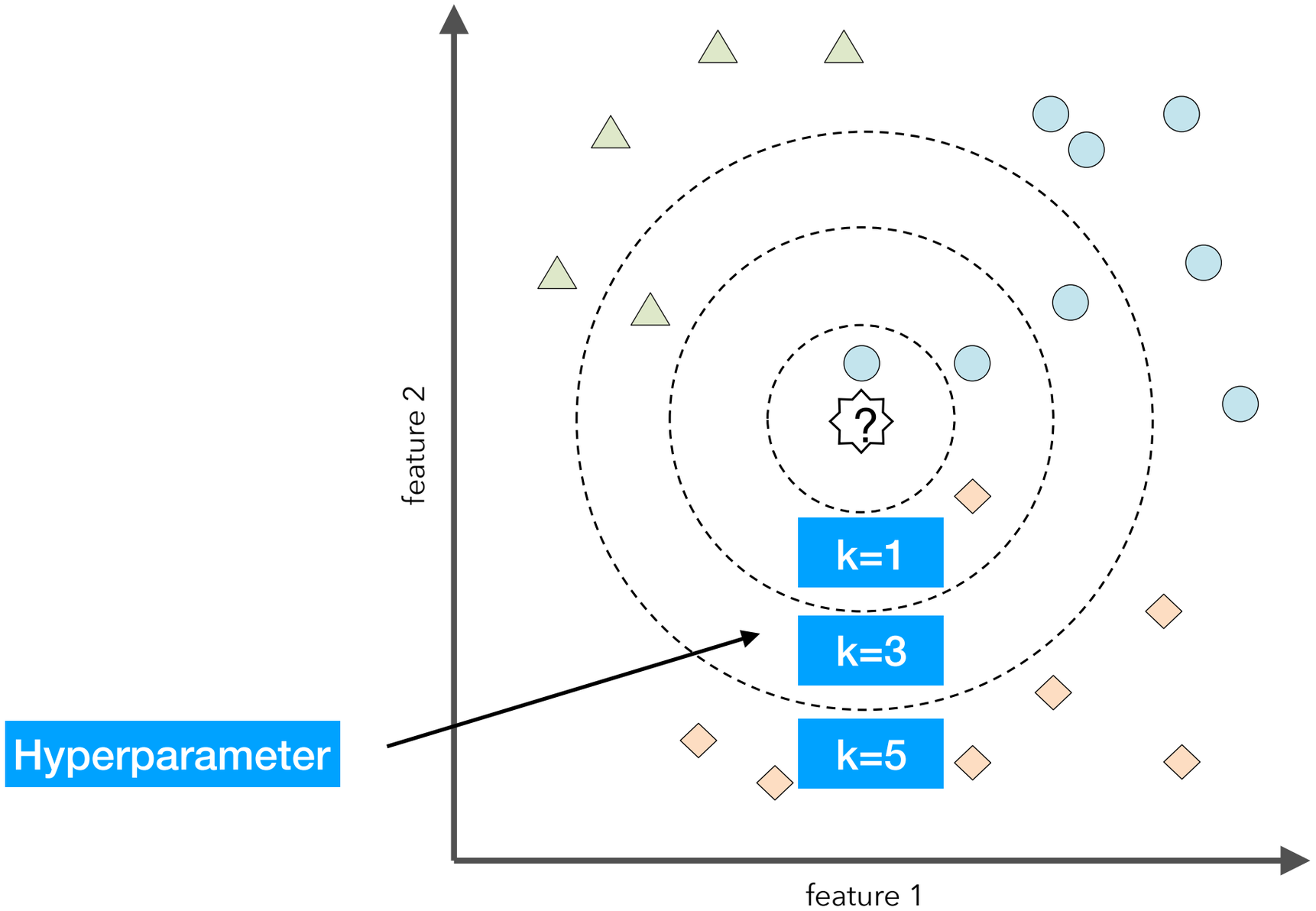
Overview


Model Eval Lectures
Basics
Bias and Variance
Overfitting and Underfitting
Holdout method
Confidence Intervals
Resampling methods
Repeated holdout
Empirical confidence intervals
Cross-Validation
Hyperparameter tuning
Model selection
Algorithm Selection
This Lecture
Statistical Tests
Evaluation Metrics

What are Hyperparameters?

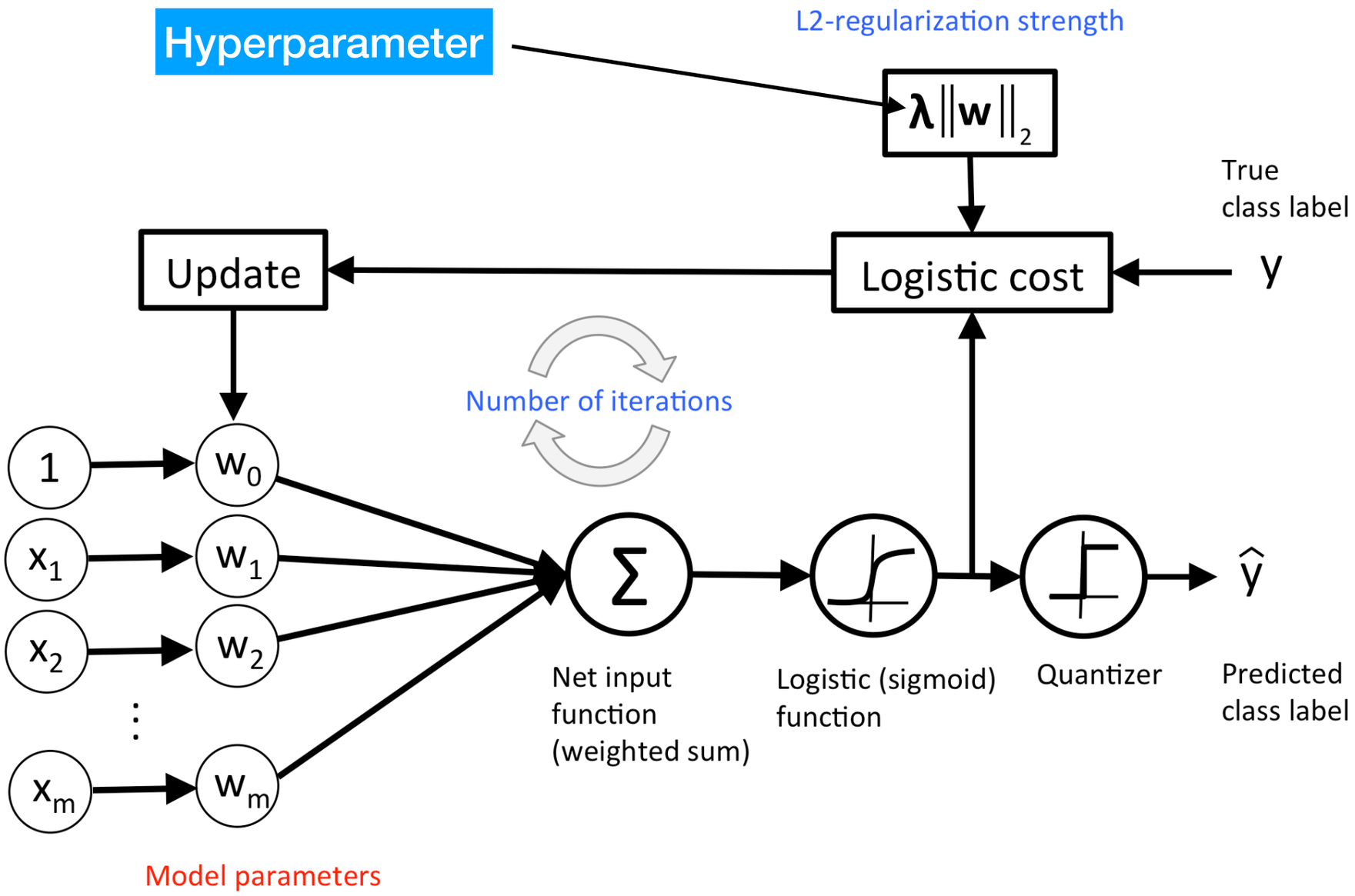
Hyperparameters

nonparametric model: k-nearest neighbors



Hyperparameters

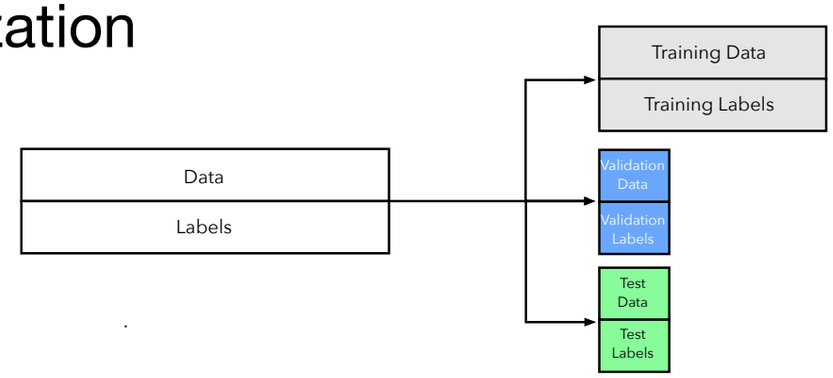
parametric model: logistic regression



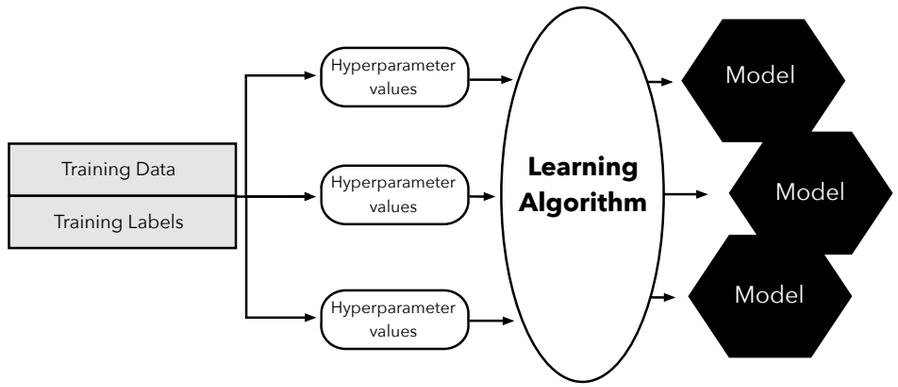
3-Way Holdout

instead of "regular" holdout to avoid "data leakage" during hyperparameter optimization

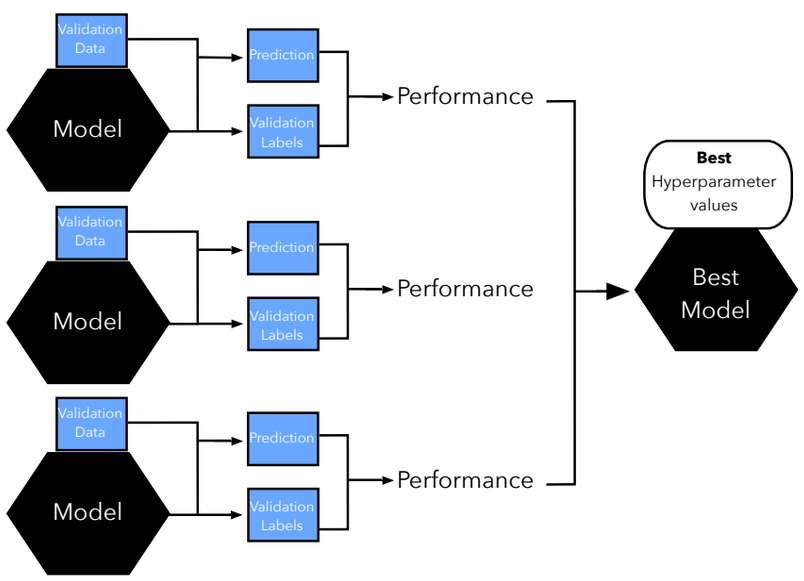
1



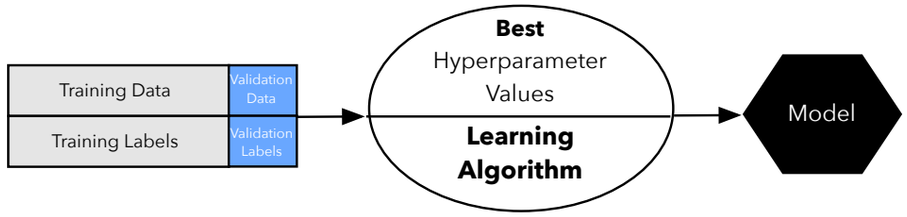
2



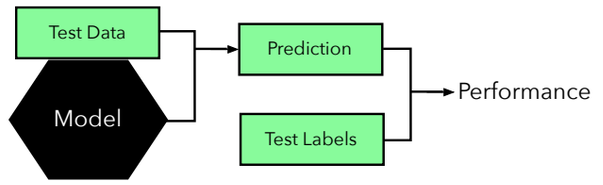
3



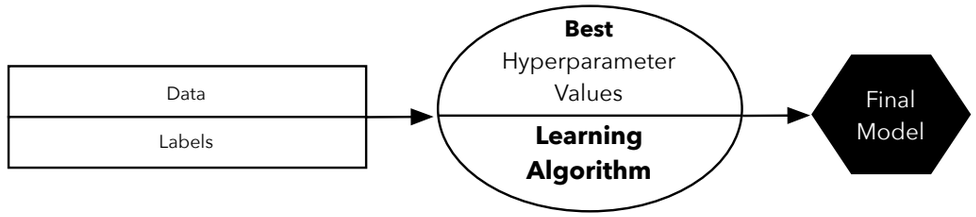
4



5



6



Main points why we evaluate the predictive performance of a model:

1. Want to estimate the generalization performance, the predictive performance of our model on future (unseen) data.

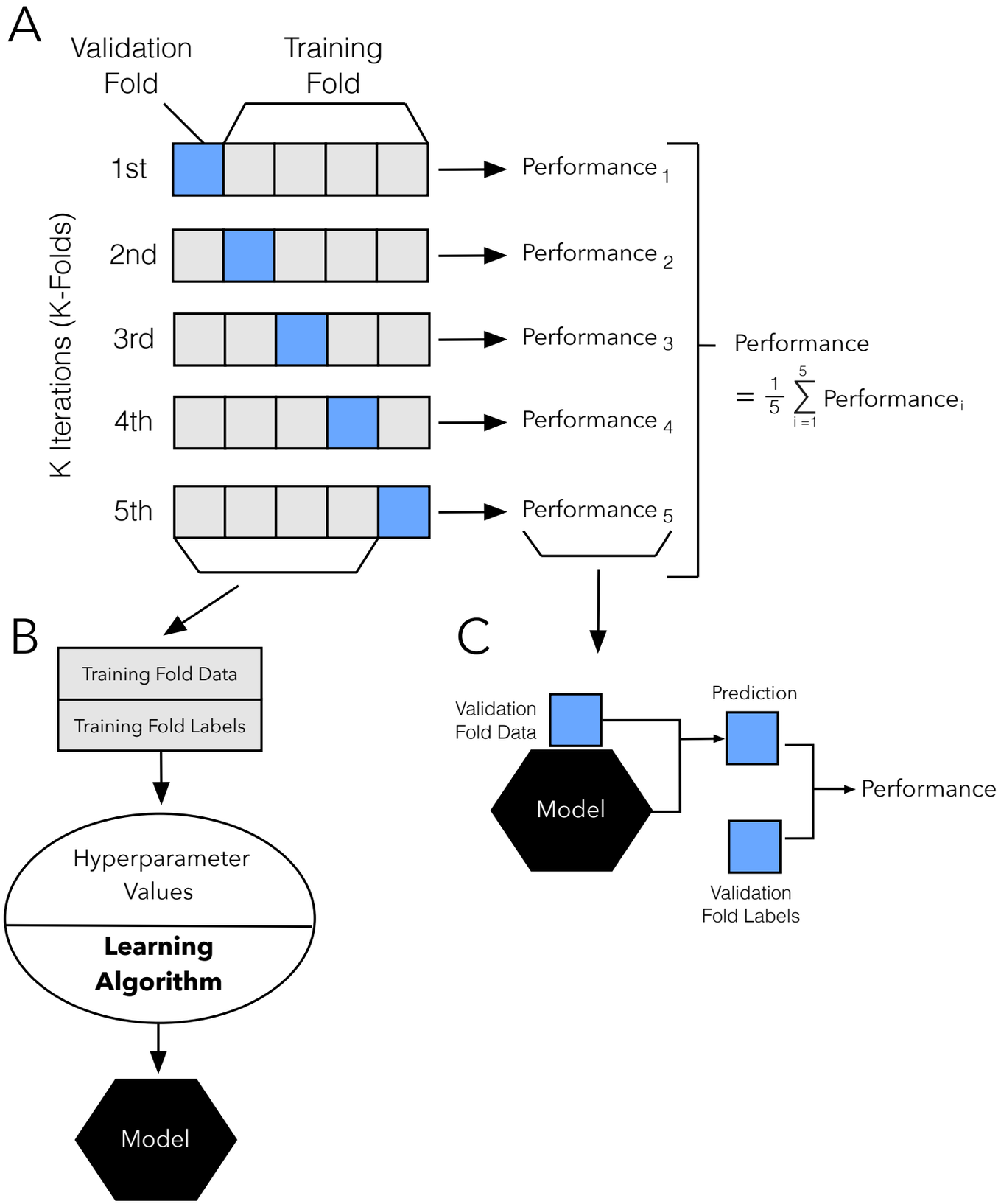
2. Want to increase the predictive performance by tweaking the learning algorithm and selecting the best performing model from a given hypothesis space.

3. Want to identify the ML algorithm that is best-suited for the problem at hand; thus, we want to compare different algorithms, selecting the best-performing one as well as the best performing model from the algorithm's hypothesis space.

k-Fold Cross-Validation Part 1

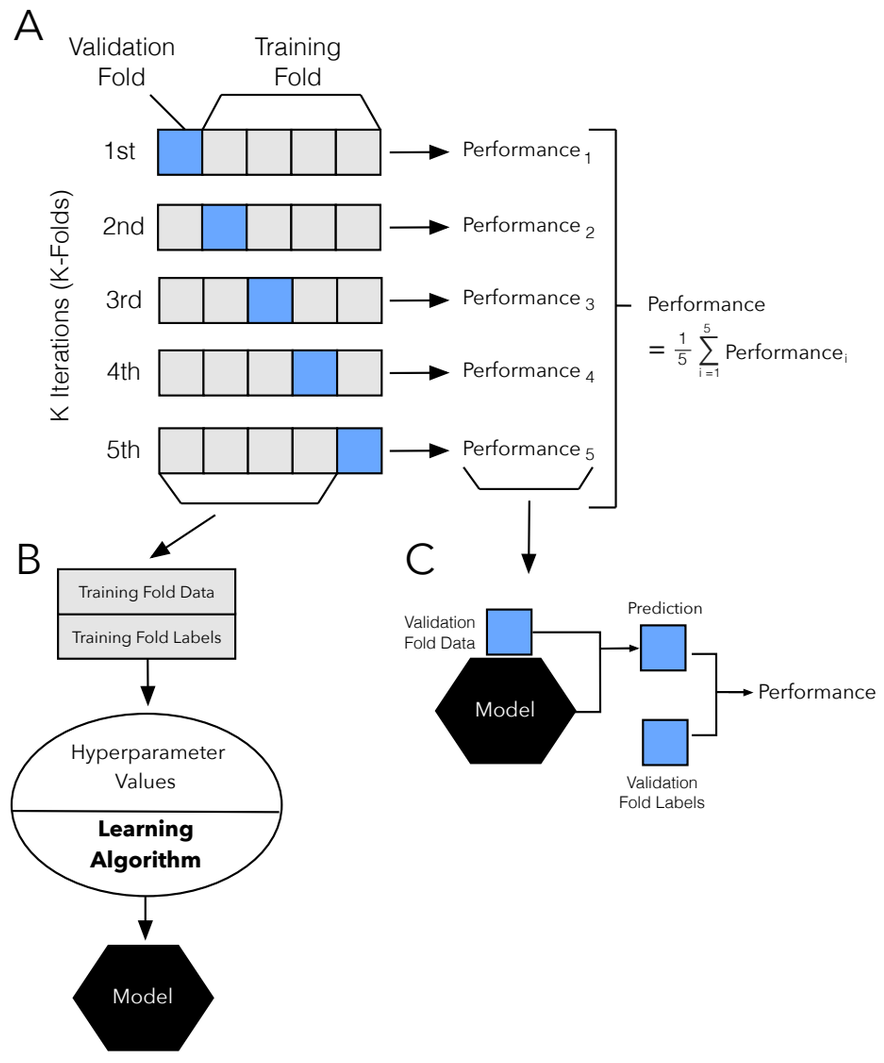
Model Evaluation

k-Fold Cross-Validation

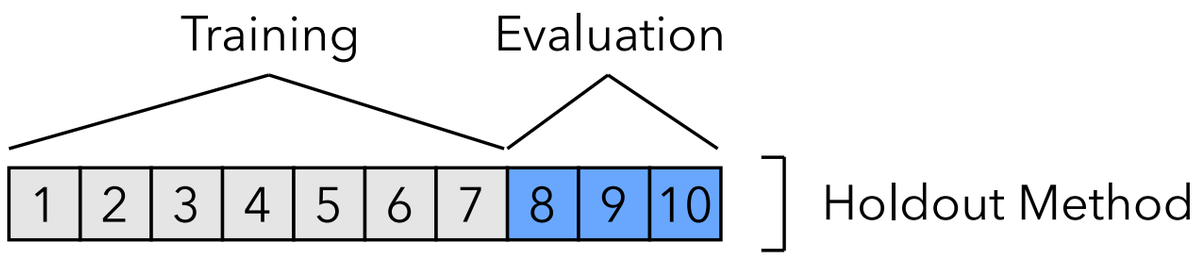


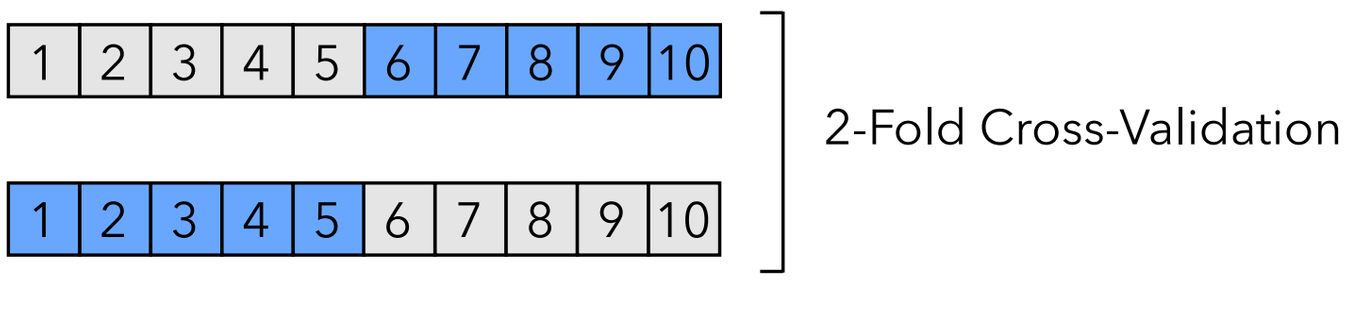
k-Fold Cross-Validation

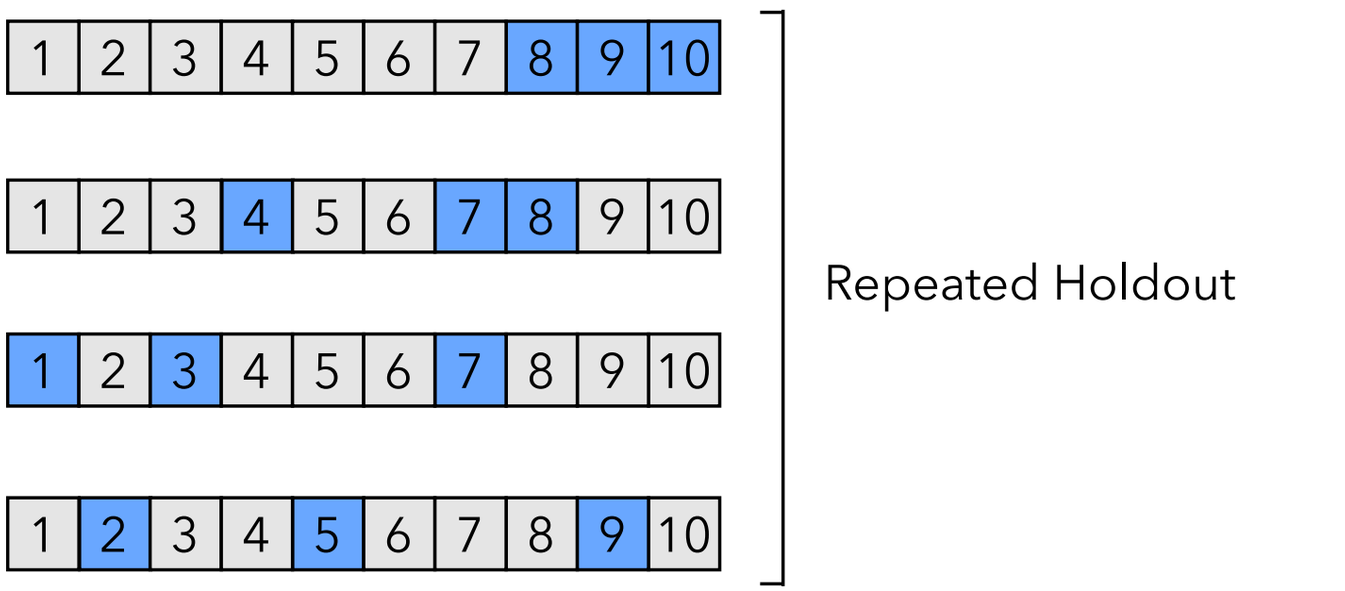
- non-overlapping test folds; utilizes all data for testing
- overlapping training folds
- some variance estimate from different training sets, (but no unbiased estimate)
- more pessimistic for small k because we withhold data from fitting



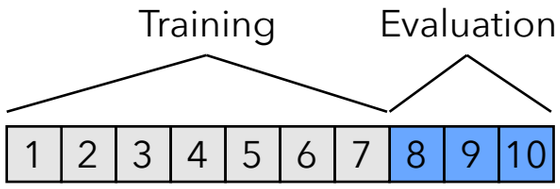
k-Fold CV special cases: k=2 & k=n



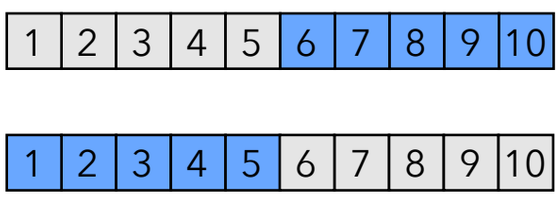




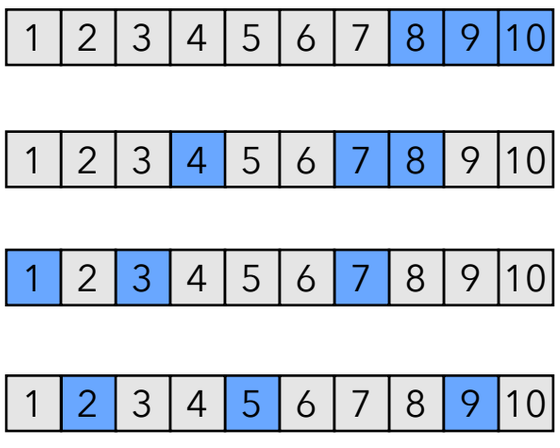
k-Fold CV special cases: k=2 & k=n



Holdout Method



2-Fold Cross-Validation



Repeated Holdout

LOOCV

k-Fold Cross-Validation

"[...] where available sample sizes are modest, holding back compounds for model testing is ill-advised. This fragmentation of the sample harms the calibration and does not give a trustworthy assessment of fit anyway. It is better to use all data for the calibration step and check the fit by cross-validation, making sure that the cross-validation is carried out correctly. [...] The only motivation to rely on the holdout sample rather than cross-validation would be if there was reason to think the cross-validation not trustworthy -- biased or highly variable. But neither theoretical results nor the empiric results sketched here give any reason to disbelieve the cross-validation results." [1]

1. Hawkins, D. M., Basak, S. C., & Mills, D. (2003). Assessing model fit by cross-validation. *Journal of chemical information and computer sciences*, *43*(2), 579-586.

LOOCV vs Holdout

Experiment	**Mean**	**Standard deviation**
True R^2 — q^2	0.010	0.149
True R^2 — hold 50	0.028	0.184
True R^2 — hold 20	0.055	0.305
True R^2 — hold 10	0.123	0.504

1. Hawkins, D. M., Basak, S. C., & Mills, D. (2003). Assessing model fit by cross-validation. *Journal of chemical information and computer sciences*, *43*(2), 579-586.

The reported "mean" refers to the averaged difference between the true coefficients of determination (R^2) and the coefficients obtained via LOOCV (here called q^2) after repeating this procedure on multiple, different 100-example training sets

(why not changing the random seed in LOOCV?)

LOOCV vs Holdout

Experiment	**Mean**	**Standard deviation**
True R^2 — q^2	0.010	0.149
True R^2 — hold 50	0.028	0.184
True R^2 — hold 20	0.055	0.305
True R^2 — hold 10	0.123	0.504

1. Hawkins, D. M., Basak, S. C., & Mills, D. (2003). Assessing model fit by cross-validation. *Journal of chemical information and computer sciences*, *43*(2), 579-586.

The reported "mean" refers to the averaged difference between the true coefficients of determination (R2) and the coefficients obtained via LOOCV (here called q^2) after repeating this procedure on different 100-example training

In rows 2-4, the researchers used the holdout method for fitting models to the 100-example training sets, and they evaluated the performances on holdout sets of sizes 10, 20, and 50 samples. Each experiment was repeated 75 times, and the mean column shows the average difference between the estimated R^2 and the true R^2 values.

Problems with LOOCV for Classification

- While LOOCV is almost unbiased, one downside of using LOOCV over k-fold cross-validation with $k < n$ is the large variance of the LOOCV estimate.

- LOOCV is "defect" when using a discontinuous loss-function such as the 0-1 loss in classification or even in continuous loss functions such as the mean-squared-error.

- LOOCV has high variance because the test set only contains one example.

Problems with LOOCV for Classification

"With *k=n*, the cross-validation estimator is approximately unbiased for the true (expected) prediction error, but can have high variance because the n "training sets" are so similar to one another." [1]

[1] Friedman, J., Hastie, T., & Tibshirani, R. (2001). *The elements of statistical learning* (Vol. 1, No. 10). New York, NY, USA: Springer series in statistics.

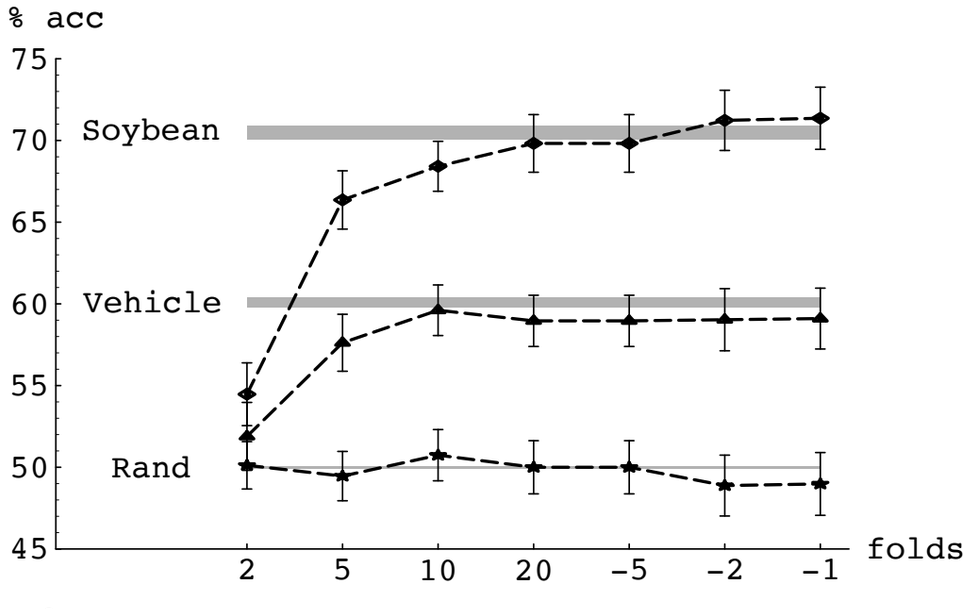
Problems with LOOCV for Classification

$$\text{Var}\left(\sum_{i=1}^{n} X_i\right) = \sum_{i=1}^{n}\sum_{j=1}^{n} \text{Cov}(X_i, X_j) = \sum_{i=1}^{n} \text{Var}(X_i) + 2 \sum_{1 \leq i < j \leq n} \text{Cov}(X_i, X_j)$$

For correlated variables, the variance of their sum is the sum of their covariances

Or in other words, we can attribute the high variance to the fact that the mean of highly correlated variables has a higher variance than the mean of variables that are not highly correlated?

Empirical Study and Recommendation



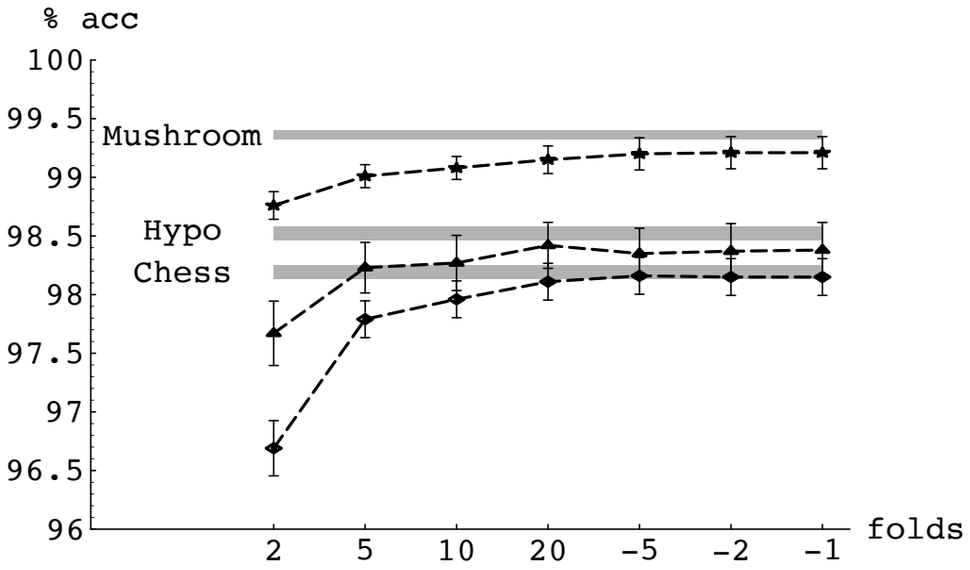
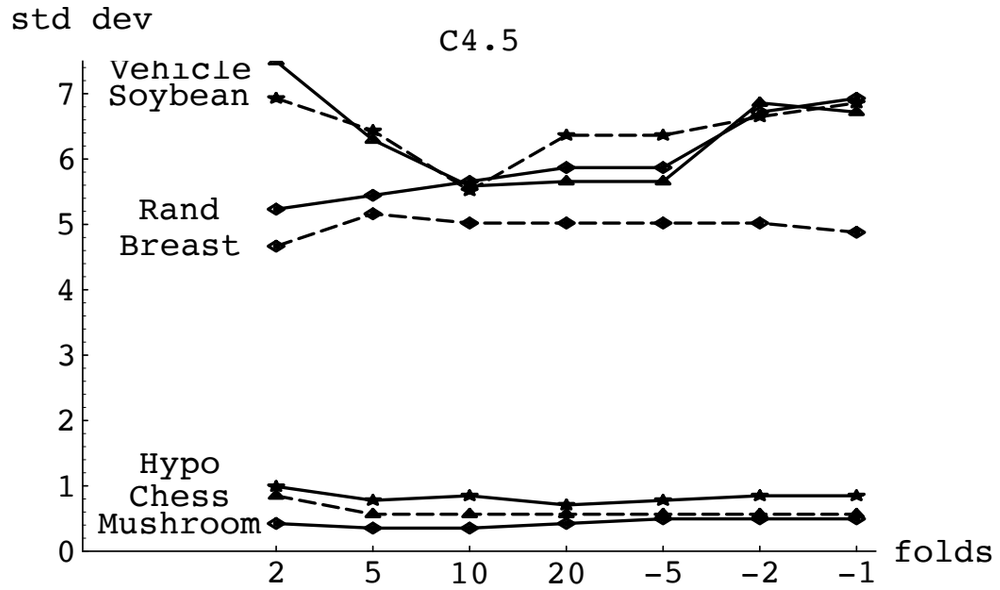


Figure 1: C4.5: The bias of cross-validation with varying folds. A negative k folds stands for leave-k-out. Error bars are 95% confidence intervals for the mean. The gray regions indicate 95% confidence intervals for the true accuracies. Note the different ranges for the accuracy axis.



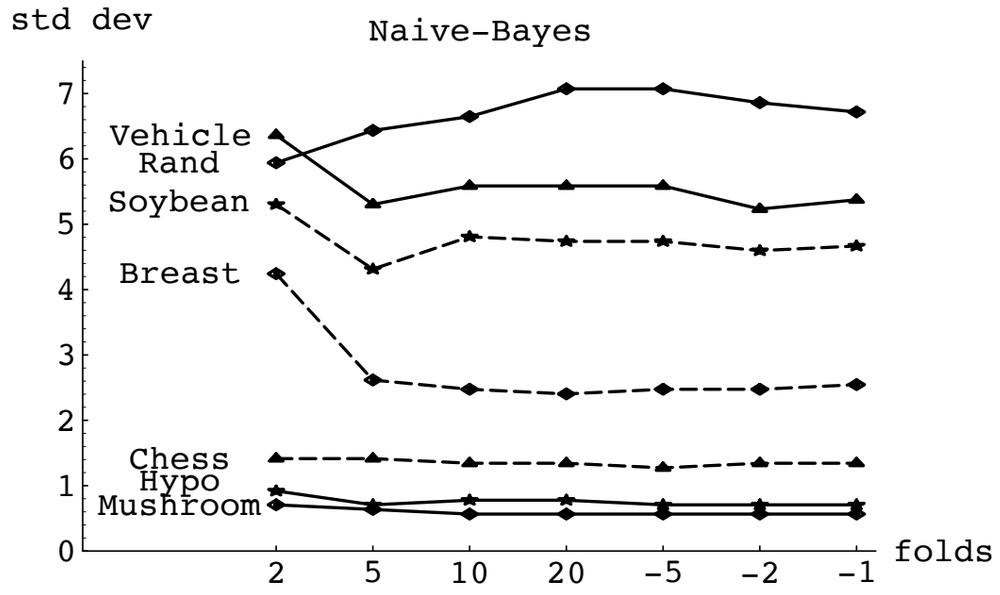


Figure 3: Cross-validation: standard deviation of accuracy (population). Different line styles are used to help differentiate between curves.

Kohavi, R. (1995, August). A study of cross-validation and bootstrap for accuracy estimation and model selection. In *Ijcai* (Vol. 14, No. 2, pp. 1137-1145).

Summarizing k-Fold CV for Model Evaluation

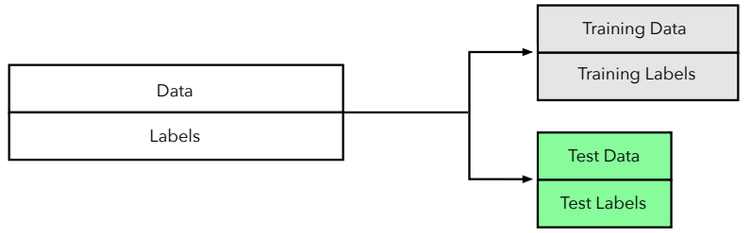
What happens if we increase *k*?

- The bias of the performance estimator ____creases (more accurate / more variable?)
- The variance of the performance estimators ____creases (more accurate / more variable?)
- The computational cost ____creases (more iterations, larger training sets during fitting)
- Exception: decreasing the value of k in k-fold cross-validation to small values (for example, 2 or 3) also ____creases the variance on small datasets due to random sampling effects

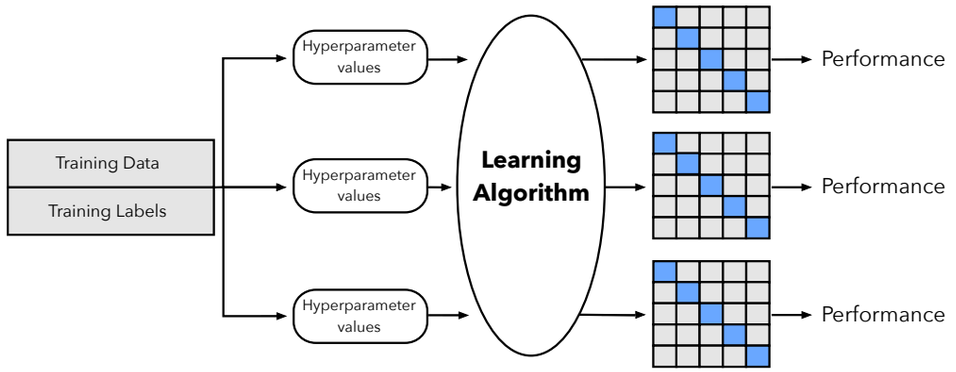
k-Fold Cross-Validation Part 2

Model Selection

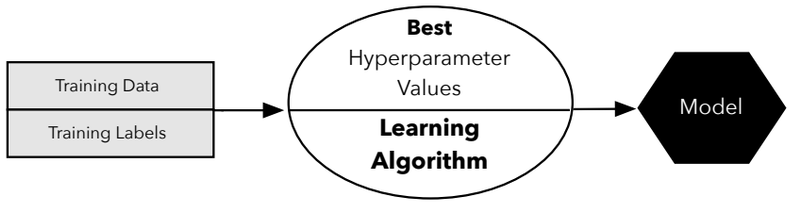
1


Data
Labels
Training Data
Training Labels
Test Data
Test Labels

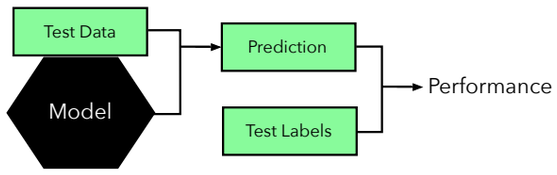
2


Training Data
Training Labels
Hyperparameter values
Hyperparameter values
Hyperparameter values
Learning Algorithm
Performance
Performance
Performance

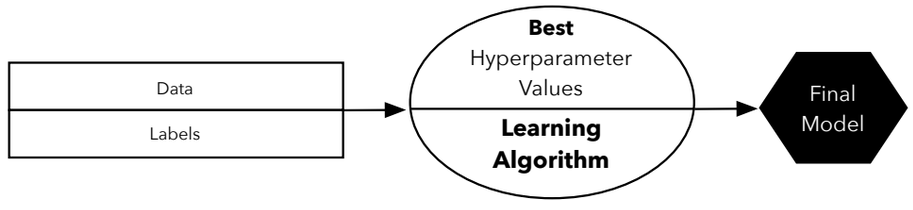
3


Training Data
Training Labels
Best
Hyperparameter
Values
Learning
Algorithm
Model

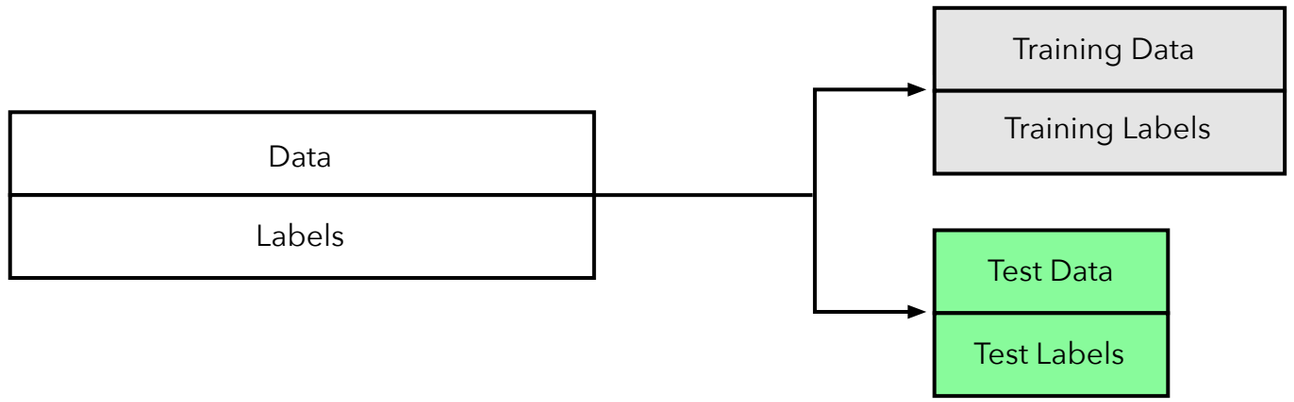
4


Test Data
Model
Prediction
Test Labels
Performance

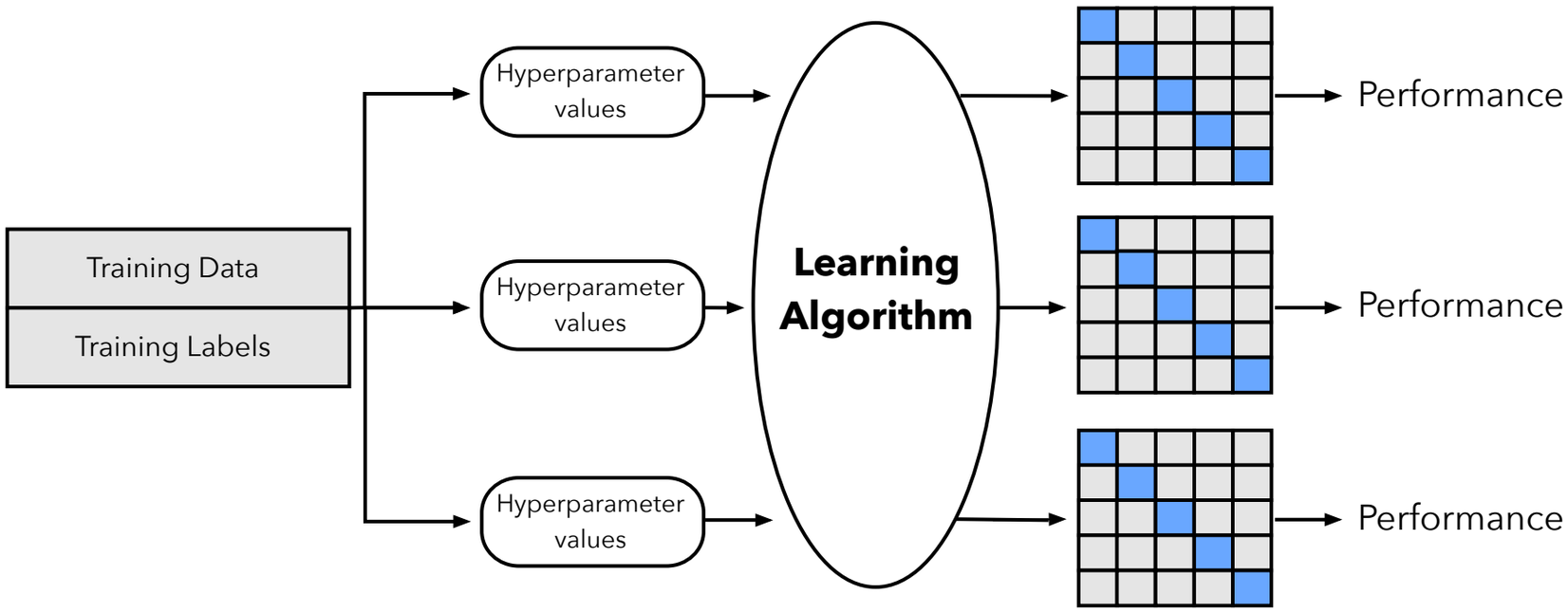
5


Data
Labels
Best
Hyperparameter
Values
Learning
Algorithm
Final
Model

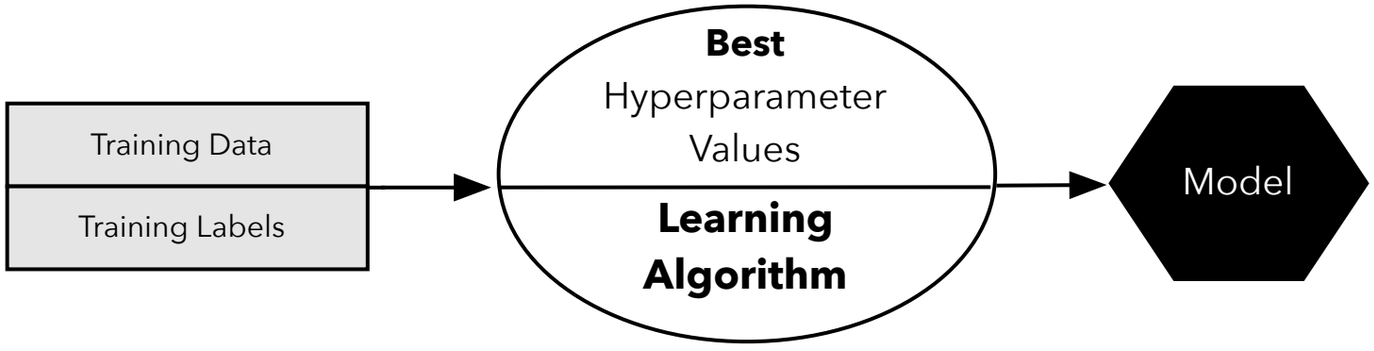
1


Data
Labels
Training Data
Training Labels
Test Data
Test Labels

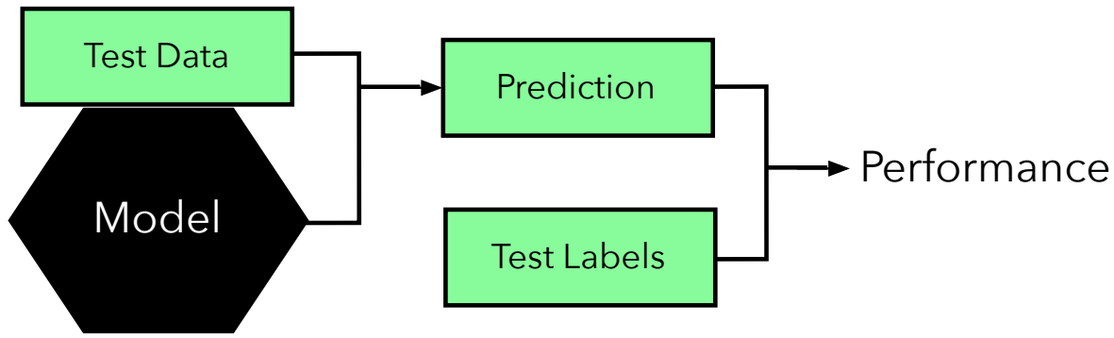
2


Training Data
Training Labels
Hyperparameter values
Hyperparameter values
Hyperparameter values
Learning Algorithm
Performance
Performance
Performance

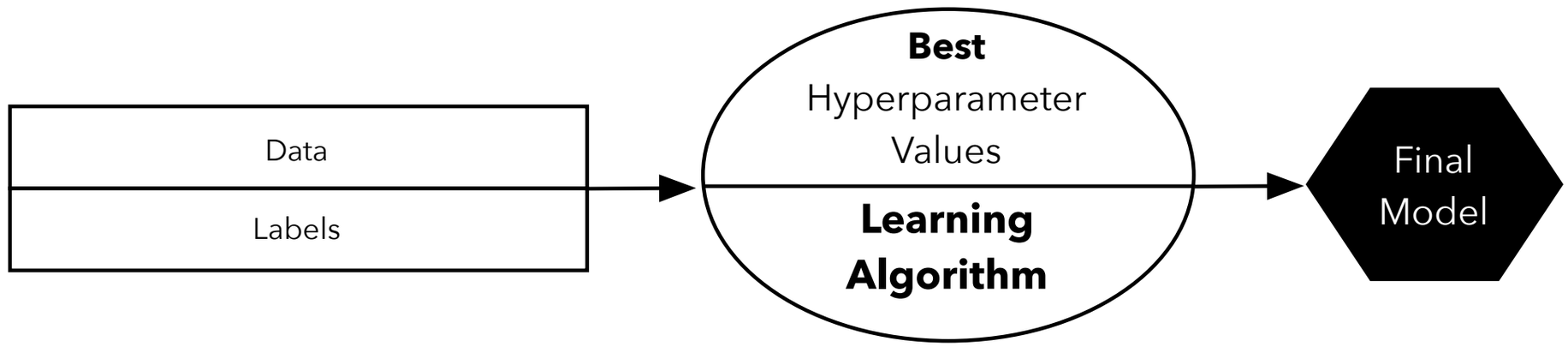
3


Training Data
Training Labels
Best
Hyperparameter
Values
Learning
Algorithm
Model

4


Test Data
Model
Prediction
Test Labels
Performance

5


Data
Labels
Best
Hyperparameter
Values
Learning
Algorithm
Final
Model

The Law of Parsimony

Occam's Razor: "Among competing hypotheses, the one with the fewest assumptions should be selected."

https://en.wikipedia.org/wiki/Occam%27s_razor

The Law of Parsimony

"Simpler models are more accurate. This belief is sometimes equated with Occam's razor, but the razor only says that simpler explanations are preferable, not why. They're preferable because they're easier to understand, remember, and reason with. Sometimes the simplest hypothesis consistent with the data is less accurate for prediction than a more complicated one. Some of the most powerful learning algorithms output models that seem gratuitously elaborate -- sometimes even continuing to add to them after they've perfectly fit the data -- but that's how they beat the less powerful ones."

Pedro Domingos: "Then Myths about Machine Learning"
https://medium.com/@pedromdd/ten-myths-about-machine-learning-d888b48334a3

The 1-standard error method

"... However, if two models perform equally well, the simpler one seems more likely (among other advantages)"

Pedro Domingos: "Then Myths about Machine Learning"
https://medium.com/@pedromdd/ten-myths-about-machine-learning-d888b48334a3

The 1-standard error method

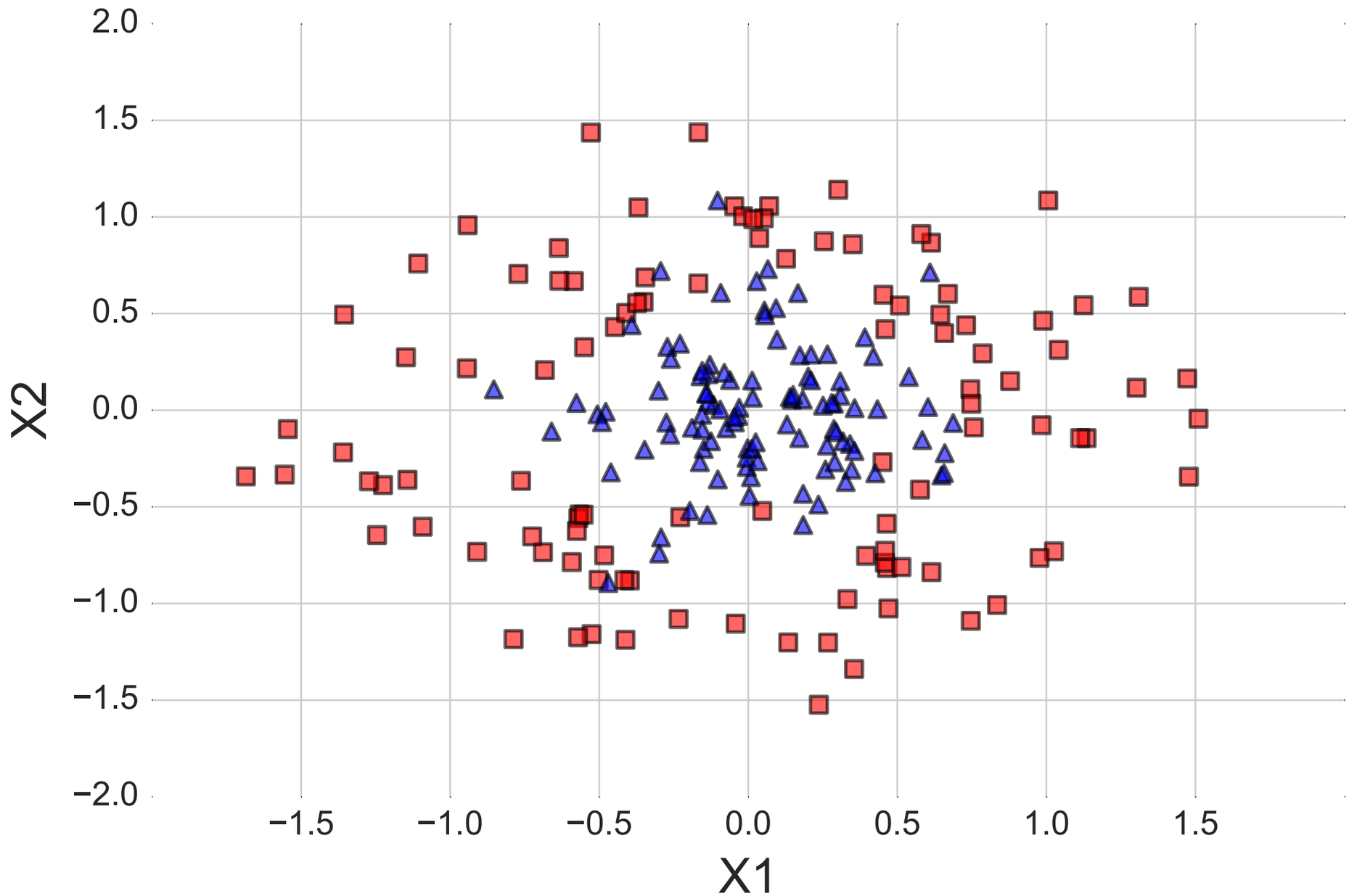
"... However, if two models perform equally well, the simpler one seems more likely (among other advantages)"

Pedro Domingos: "Then Myths about Machine Learning"

1. Consider the numerically optimal estimate and its standard error.

2. Select the model whose performance is within one standard error of the value obtained in step 1.

The 1-standard error method



(Some toy data I generated via scikit-learn)

The 1-standard error method

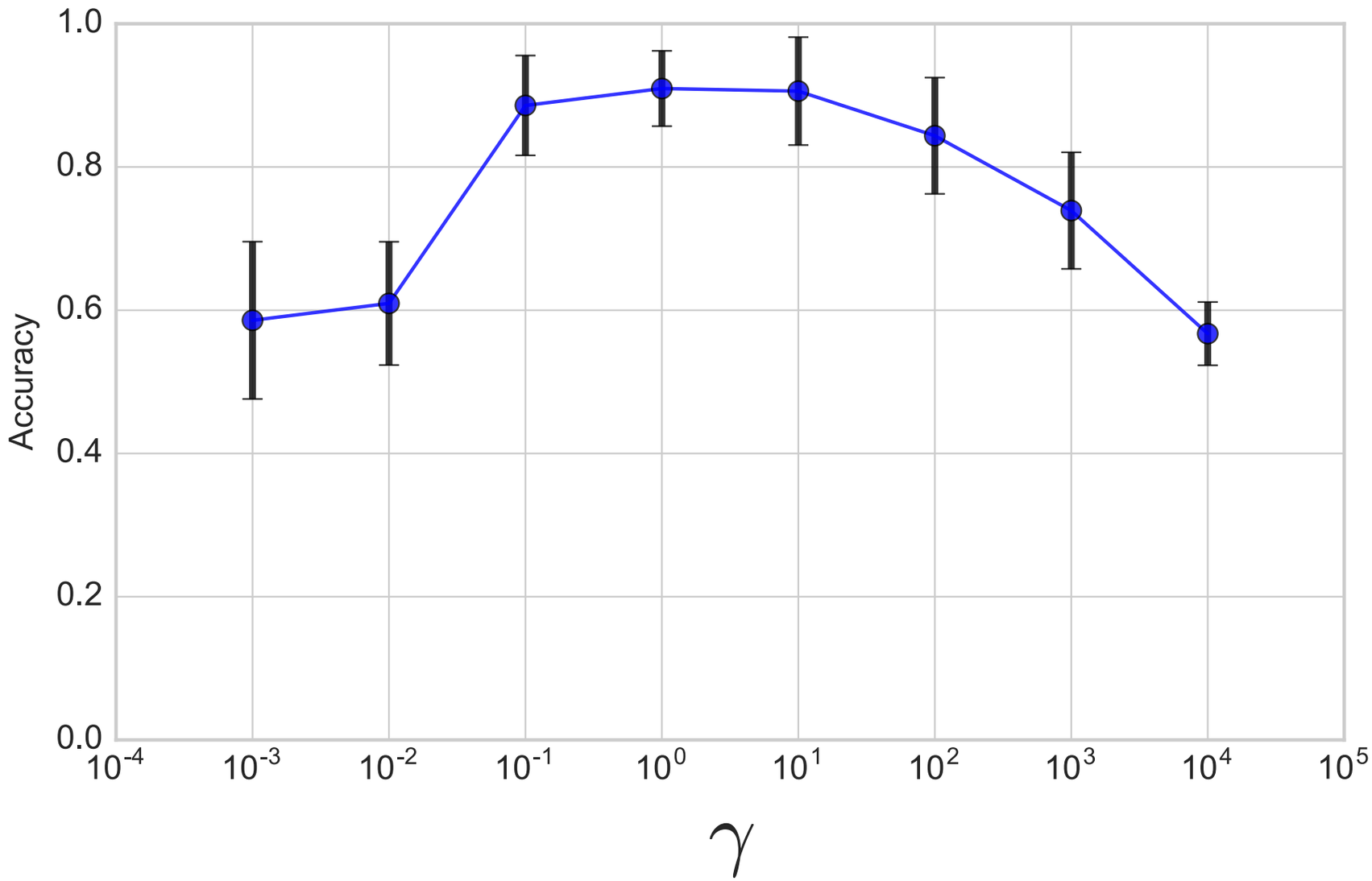
Consider a RBF-kernel SVM, where gamma controls the influence of the training points

(don't need to know the details, yet)

Gaussian/RBF-kernel: $K(x_i, x_j) = exp(-\gamma ||x_i - x_j||^2), \gamma > 0.$

The 1-standard error method

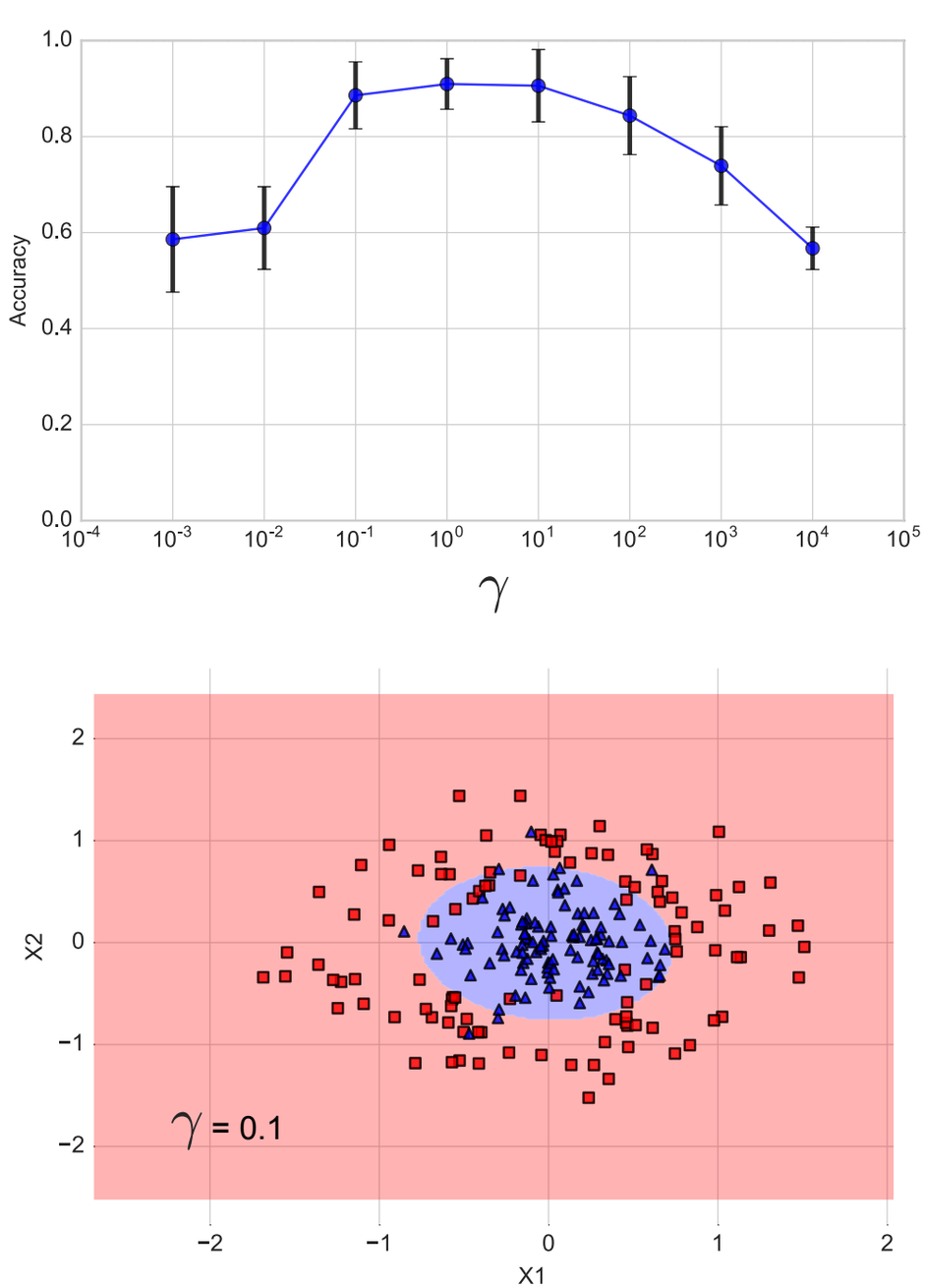
Which parameter would you select?

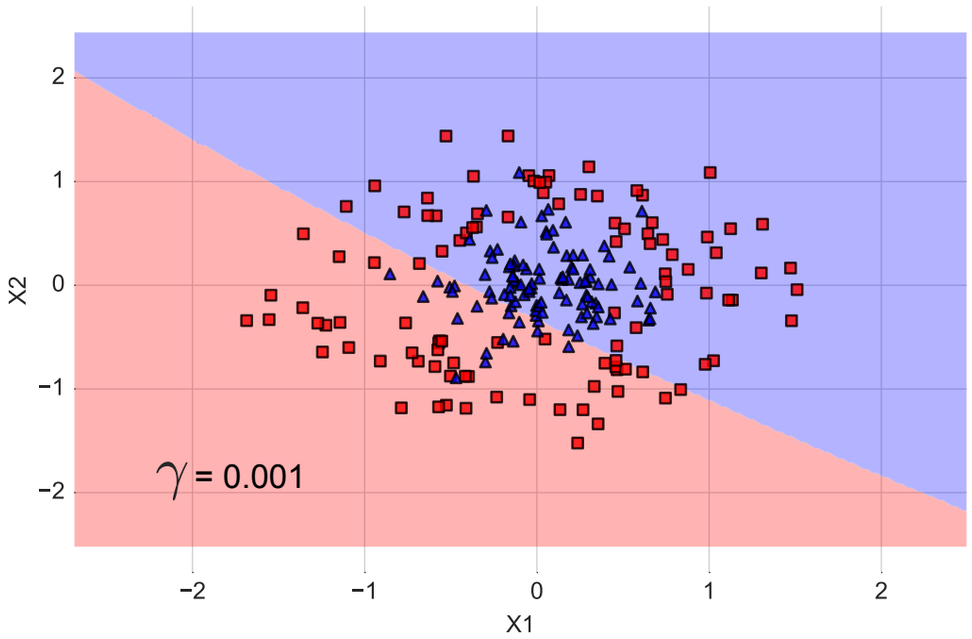


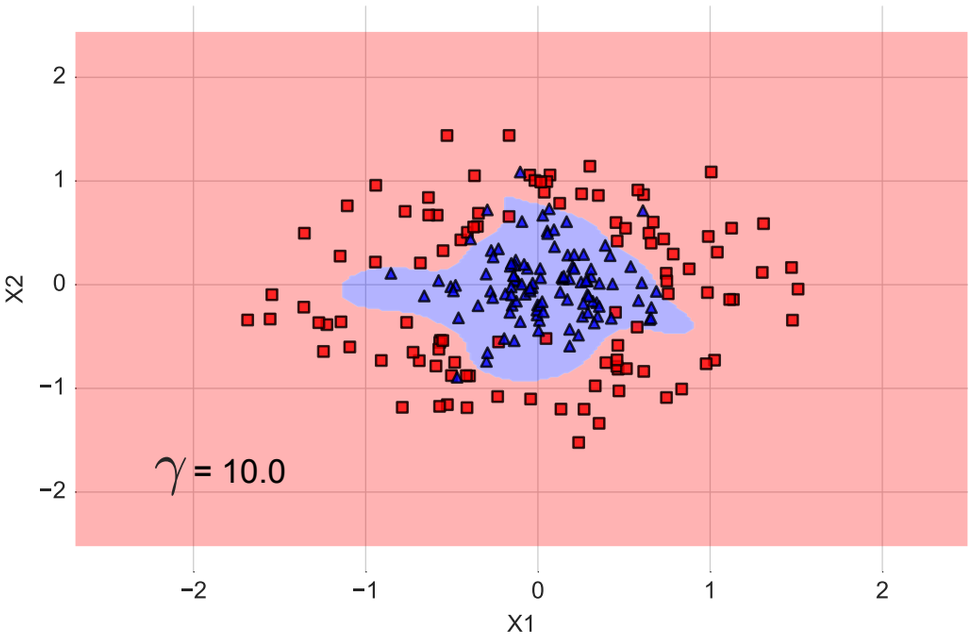
(note: here I used 10-fold CV)

The 1-standard error method

Which parameter would you select?







(note: here I used 10-fold CV)

Code Examples

https://github.com/rasbt/stat479-machine-learning-fs19/blob/master/10_eval3-cv/code/10_eval3-cv__code.ipynb

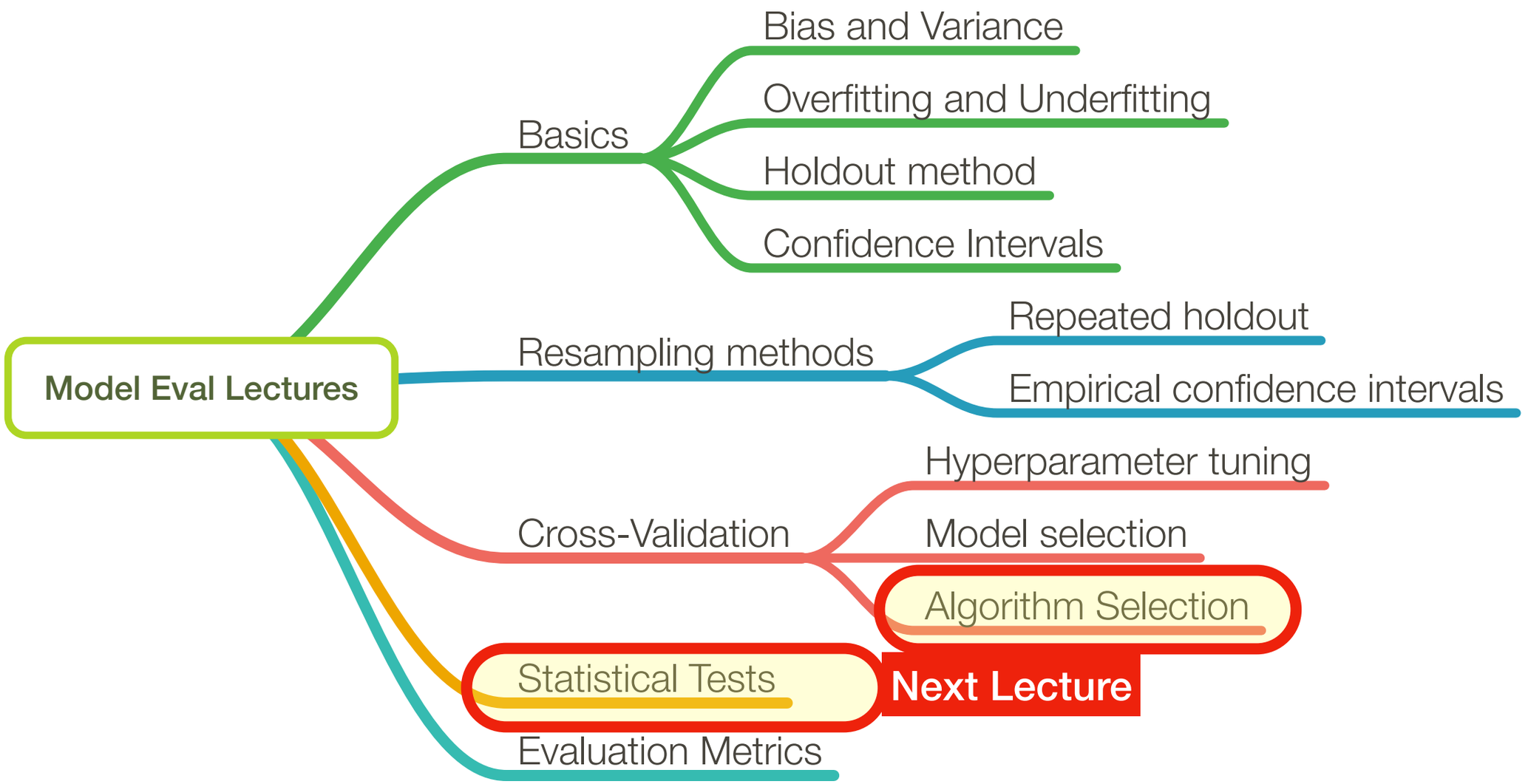
Reading Assignment

L10 Lecture Notes:

https://github.com/rasbt/stat479-machine-learning-fs19/blob/master/10_eval3-cv/10-eval3-cv__notes.pdf

Kohavi, Ron. "A study of cross-validation and bootstrap for accuracy estimation and model selection." *IJCAI*. Vol. 14. No. 2. 1995. https://ai.stanford.edu/~ronnyk/accEst.pdf

Overview


Model Eval Lectures
Basics
Bias and Variance
Overfitting and Underfitting
Holdout method
Confidence Intervals
Resampling methods
Repeated holdout
Empirical confidence intervals
Cross-Validation
Hyperparameter tuning
Model selection
Algorithm Selection
Statistical Tests
Next Lecture
Evaluation Metrics