Lecture 08

Model Evaluation Part 1: Introduction to Overfitting and Underfitting

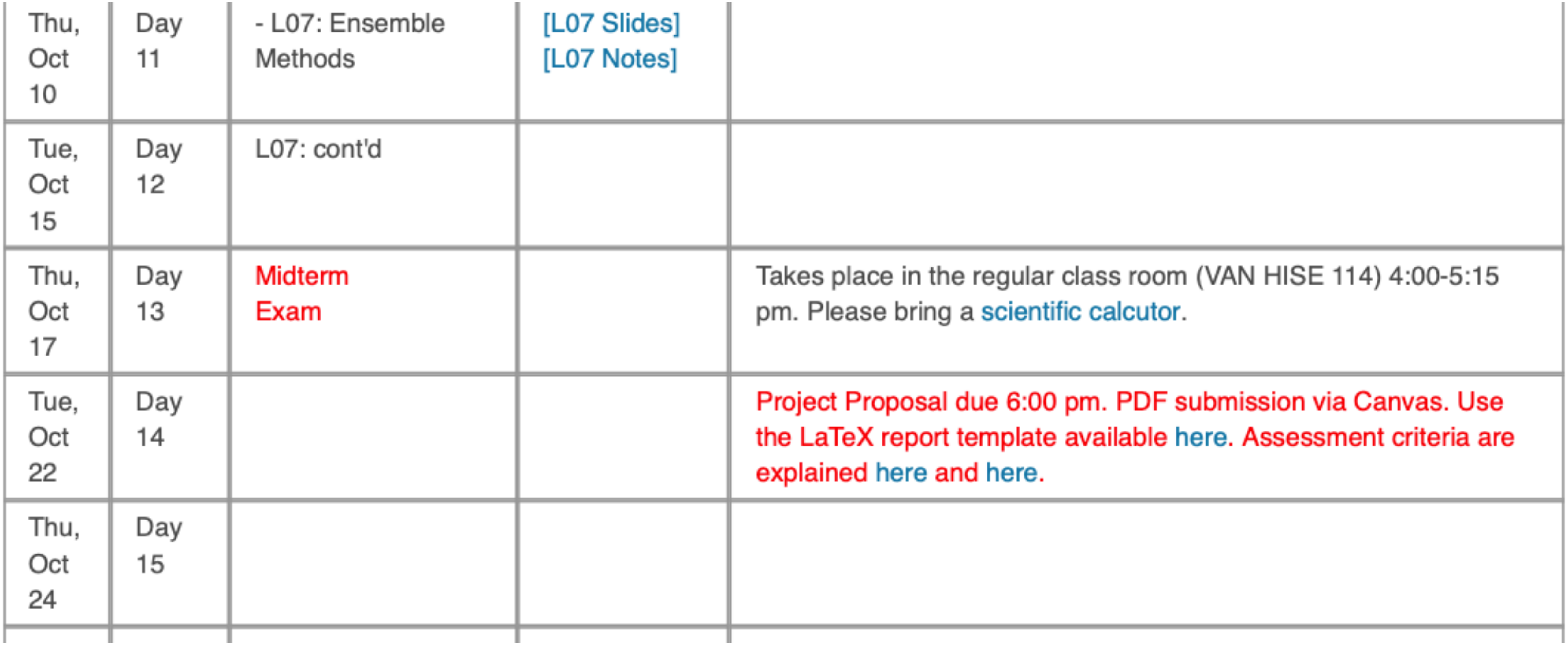
STAT 479: Machine Learning, Fall 2019

Sebastian Raschka

http://stat.wisc.edu/~sraschka/teaching/stat479-fs2019/

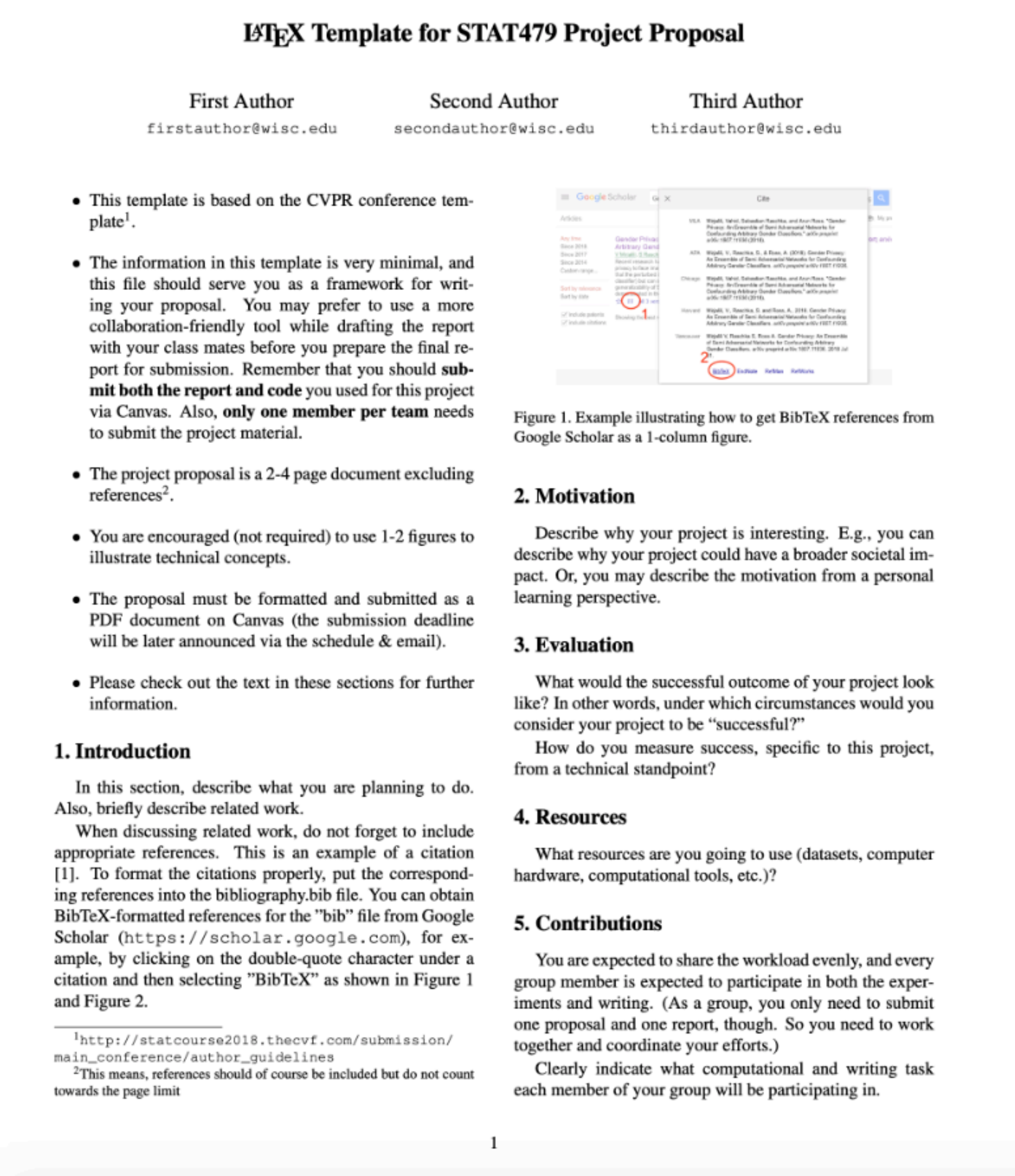
Announcements First

Source: https://pittnews.com/article/149994/news/starship-food-delivery-robots/



Thu, Oct 10	Day 11	- L07: Ensemble Methods	[L07 Slides] [L07 Notes]	
Tue, Oct 15	Day 12	L07: cont'd		
Thu, Oct 17	Day 13	Midterm Exam		Takes place in the regular class room (VAN HISE 114) 4:00-5:15 pm. Please bring a scientific calcutor.
Tue, Oct 22	Day 14			Project Proposal due 6:00 pm. PDF submission via Canvas. Use the LaTeX report template available here. Assessment criteria are explained here and here.
Thu, Oct 24	Day 15			

http://pages.stat.wisc.edu/~sraschka/teaching/stat479-fs2019/#calendar



LaTeX Template for STAT479 Project Proposal

First Author
firstauthor@wisc.edu

Second Author
secondauthor@wisc.edu

Third Author
thirdauthor@wisc.edu

- This template is based on the CVPR conference template[1].
- The information in this template is very minimal, and this file should serve you as a framework for writing your proposal. You may prefer to use a more collaboration-friendly tool while drafting the report with your class mates before you prepare the final report for submission. Remember that you should **submit both the report and code** you used for this project via Canvas. Also, **only one member per team** needs to submit the project material.
- The project proposal is a 2-4 page document excluding references[2].
- You are encouraged (not required) to use 1-2 figures to illustrate technical concepts.
- The proposal must be formatted and submitted as a PDF document on Canvas (the submission deadline will be later announced via the schedule & email).
- Please check out the text in these sections for further information.

1. Introduction

In this section, describe what you are planning to do. Also, briefly describe related work.

When discussing related work, do not forget to include appropriate references. This is an example of a citation [1]. To format the citations properly, put the corresponding references into the bibliography.bib file. You can obtain BibTeX-formatted references for the "bib" file from Google Scholar (https://scholar.google.com), for example, by clicking on the double-quote character under a citation and then selecting "BibTeX" as shown in Figure 1 and Figure 2.

Figure 1. Example illustrating how to get BibTeX references from Google Scholar as a 1-column figure.

2. Motivation

Describe why your project is interesting. E.g., you can describe why your project could have a broader societal impact. Or, you may describe the motivation from a personal learning perspective.

3. Evaluation

What would the successful outcome of your project look like? In other words, under which circumstances would you consider your project to be "successful?"

How do you measure success, specific to this project, from a technical standpoint?

4. Resources

What resources are you going to use (datasets, computer hardware, computational tools, etc.)?

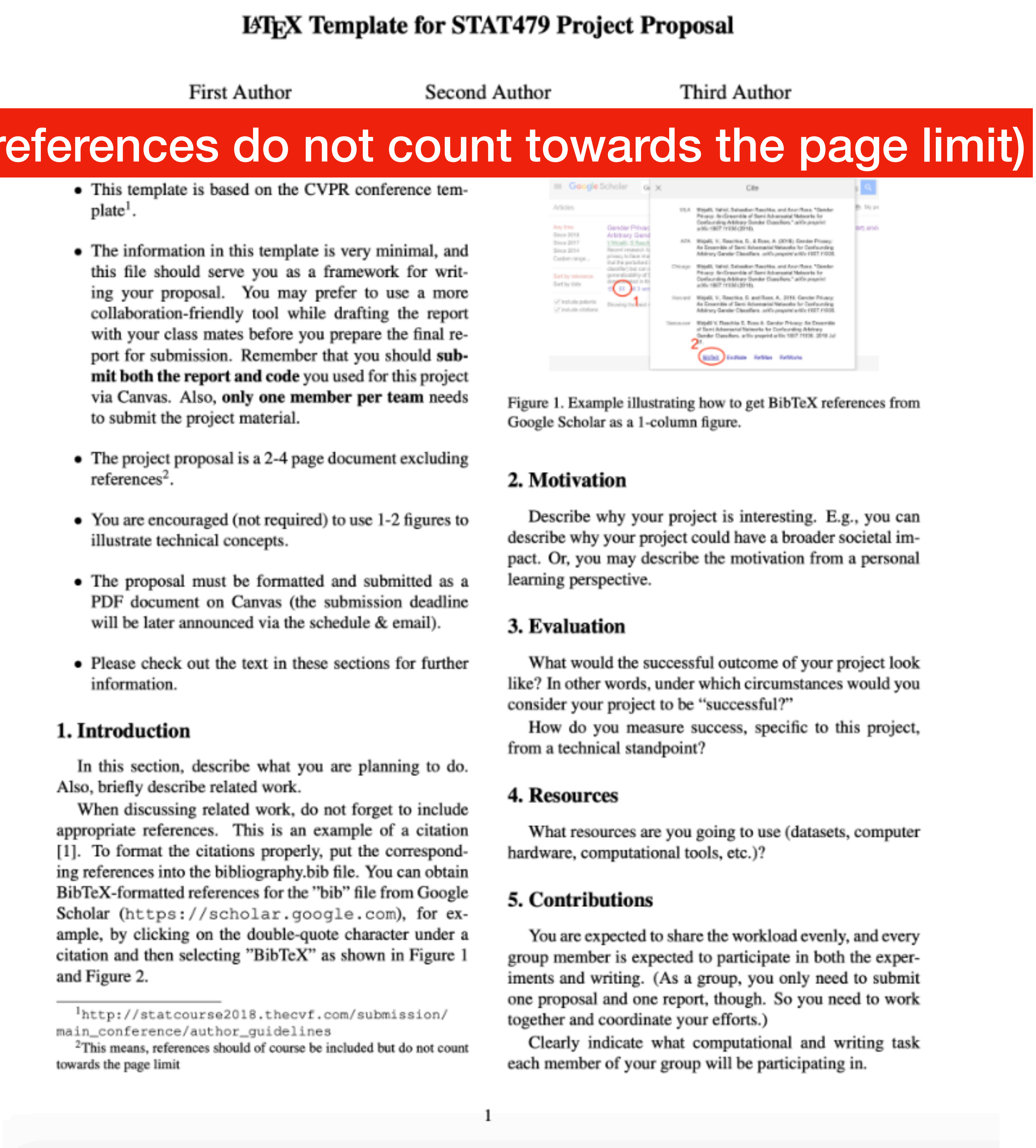
5. Contributions

You are expected to share the workload evenly, and every group member is expected to participate in both the experiments and writing. (As a group, you only need to submit one proposal and one report, though. So you need to work together and coordinate your efforts.)

Clearly indicate what computational and writing task each member of your group will be participating in.

[1] http://statcourse2018.thecvf.com/submission/main_conference/author_guidelines

[2] This means, references should of course be included but do not count towards the page limit



LaTeX Template for STAT479 Project Proposal

First Author Second Author Third Author

2-4 pages (references do not count towards the page limit)

- This template is based on the CVPR conference template[1].

- The information in this template is very minimal, and this file should serve you as a framework for writing your proposal. You may prefer to use a more collaboration-friendly tool while drafting the report with your class mates before you prepare the final report for submission. Remember that you should **submit both the report and code** you used for this project via Canvas. Also, **only one member per team** needs to submit the project material.

- The project proposal is a 2-4 page document excluding references[2].

- You are encouraged (not required) to use 1-2 figures to illustrate technical concepts.

- The proposal must be formatted and submitted as a PDF document on Canvas (the submission deadline will be later announced via the schedule & email).

- Please check out the text in these sections for further information.

1. Introduction

In this section, describe what you are planning to do. Also, briefly describe related work.

When discussing related work, do not forget to include appropriate references. This is an example of a citation [1]. To format the citations properly, put the corresponding references into the bibliography.bib file. You can obtain BibTeX-formatted references for the "bib" file from Google Scholar (`https://scholar.google.com`), for example, by clicking on the double-quote character under a citation and then selecting "BibTeX" as shown in Figure 1 and Figure 2.

Figure 1. Example illustrating how to get BibTeX references from Google Scholar as a 1-column figure.

2. Motivation

Describe why your project is interesting. E.g., you can describe why your project could have a broader societal impact. Or, you may describe the motivation from a personal learning perspective.

3. Evaluation

What would the successful outcome of your project look like? In other words, under which circumstances would you consider your project to be "successful?"

How do you measure success, specific to this project, from a technical standpoint?

4. Resources

What resources are you going to use (datasets, computer hardware, computational tools, etc.)?

5. Contributions

You are expected to share the workload evenly, and every group member is expected to participate in both the experiments and writing. (As a group, you only need to submit one proposal and one report, though. So you need to work together and coordinate your efforts.)

Clearly indicate what computational and writing task each member of your group will be participating in.

[1] `http://statcourse2018.thecvf.com/submission/main_conference/author_guidelines`

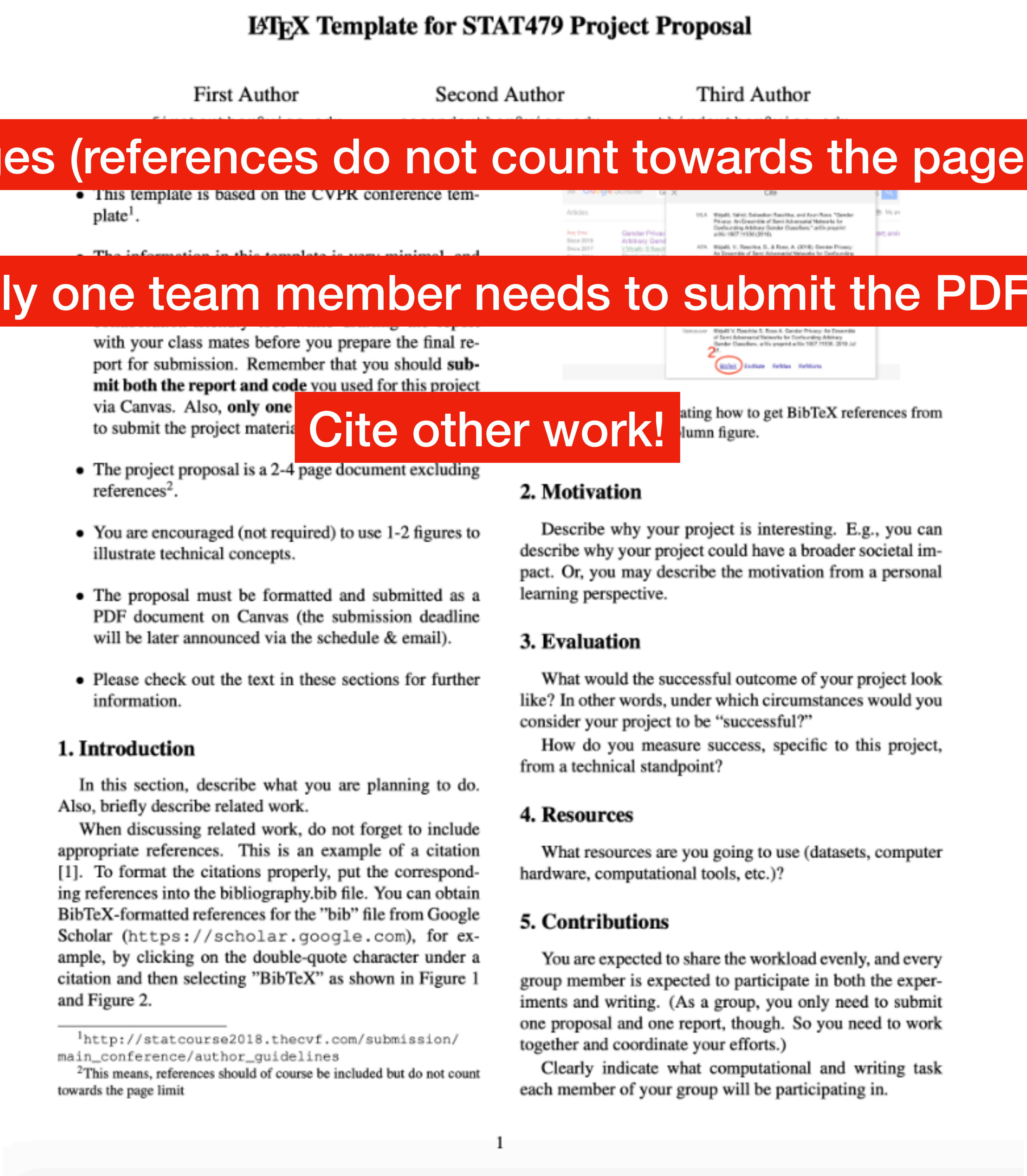
[2] This means, references should of course be included but do not count towards the page limit

1

2-4 pages (references do not count towards the page limit)

only one team member needs to submit the PDF

Cite other work!



LaTeX Template for STAT479 Project Proposal

First Author   Second Author   Third Author

- This template is based on the CVPR conference template[1].
- The information in this template is very minimal, and

with your class mates before you prepare the final report for submission. Remember that you should **submit both the report and code** you used for this project via Canvas. Also, **only one** to submit the project materia

- The project proposal is a 2-4 page document excluding references[2].
- You are encouraged (not required) to use 1-2 figures to illustrate technical concepts.
- The proposal must be formatted and submitted as a PDF document on Canvas (the submission deadline will be later announced via the schedule & email).
- Please check out the text in these sections for further information.

1. Introduction

In this section, describe what you are planning to do. Also, briefly describe related work.

When discussing related work, do not forget to include appropriate references. This is an example of a citation [1]. To format the citations properly, put the corresponding references into the bibliography.bib file. You can obtain BibTeX-formatted references for the "bib" file from Google Scholar (https://scholar.google.com), for example, by clicking on the double-quote character under a citation and then selecting "BibTeX" as shown in Figure 1 and Figure 2.

ating how to get BibTeX references from lumn figure.

2. Motivation

Describe why your project is interesting. E.g., you can describe why your project could have a broader societal impact. Or, you may describe the motivation from a personal learning perspective.

3. Evaluation

What would the successful outcome of your project look like? In other words, under which circumstances would you consider your project to be "successful?"

How do you measure success, specific to this project, from a technical standpoint?

4. Resources

What resources are you going to use (datasets, computer hardware, computational tools, etc.)?

5. Contributions

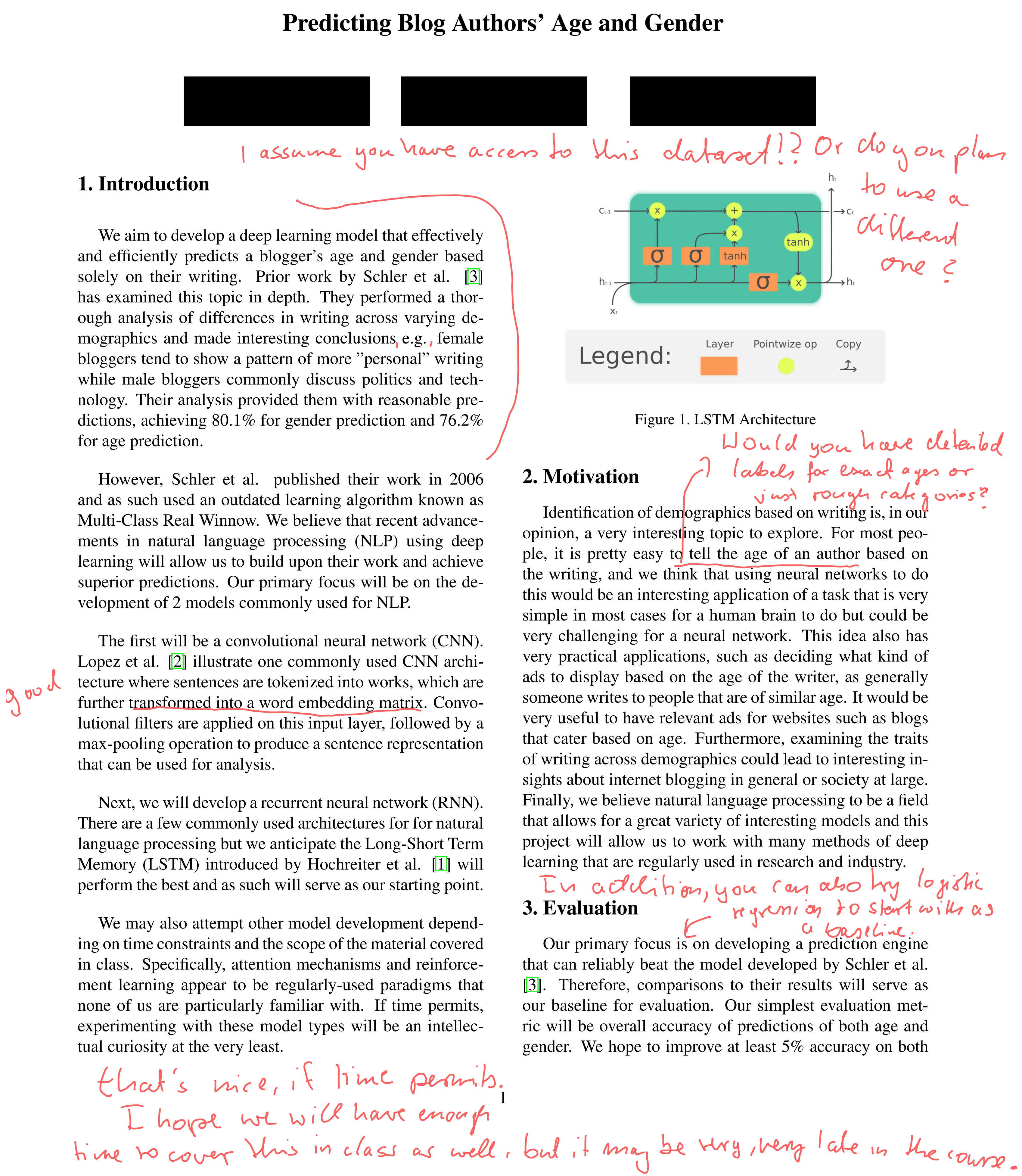
You are expected to share the workload evenly, and every group member is expected to participate in both the experiments and writing. (As a group, you only need to submit one proposal and one report, though. So you need to work together and coordinate your efforts.)

Clearly indicate what computational and writing task each member of your group will be participating in.

[1] http://statcourse2018.thecvf.com/submission/main_conference/author_guidelines

[2] This means, references should of course be included but do not count towards the page limit

Example Proposal



c_{t-1}
h_{t-1}
x_t
h_t
c_t
σ
σ
tanh
σ
tanh
x
+
Legend:
Layer
Pointwize op
Copy

Predicting Blog Authors' Age and Gender

[redacted] [redacted] [redacted]

I assume you have access to this dataset!? Or do you plan to use a different one?

1. Introduction

We aim to develop a deep learning model that effectively and efficiently predicts a blogger's age and gender based solely on their writing. Prior work by Schler et al. [3] has examined this topic in depth. They performed a thorough analysis of differences in writing across varying demographics and made interesting conclusions, e.g., female bloggers tend to show a pattern of more ”personal” writing while male bloggers commonly discuss politics and technology. Their analysis provided them with reasonable predictions, achieving 80.1% for gender prediction and 76.2% for age prediction.

However, Schler et al. published their work in 2006 and as such used an outdated learning algorithm known as Multi-Class Real Winnow. We believe that recent advancements in natural language processing (NLP) using deep learning will allow us to build upon their work and achieve superior predictions. Our primary focus will be on the development of 2 models commonly used for NLP.

The first will be a convolutional neural network (CNN). Lopez et al. [2] illustrate one commonly used CNN architecture where sentences are tokenized into works, which are further transformed into a word embedding matrix. Convolutional filters are applied on this input layer, followed by a max-pooling operation to produce a sentence representation that can be used for analysis.

good

Next, we will develop a recurrent neural network (RNN). There are a few commonly used architectures for for natural language processing but we anticipate the Long-Short Term Memory (LSTM) introduced by Hochreiter et al. [1] will perform the best and as such will serve as our starting point.

We may also attempt other model development depending on time constraints and the scope of the material covered in class. Specifically, attention mechanisms and reinforcement learning appear to be regularly-used paradigms that none of us are particularly familiar with. If time permits, experimenting with these model types will be an intellectual curiosity at the very least.

that's nice, if time permits. I hope we will have enough time to cover this in class as well, but it may be very, very late in the course.

Figure 1. LSTM Architecture

Would you have detailed labels for exact ages or just rough categories?

2. Motivation

Identification of demographics based on writing is, in our opinion, a very interesting topic to explore. For most people, it is pretty easy to tell the age of an author based on the writing, and we think that using neural networks to do this would be an interesting application of a task that is very simple in most cases for a human brain to do but could be very challenging for a neural network. This idea also has very practical applications, such as deciding what kind of ads to display based on the age of the writer, as generally someone writes to people that are of similar age. It would be very useful to have relevant ads for websites such as blogs that cater based on age. Furthermore, examining the traits of writing across demographics could lead to interesting insights about internet blogging in general or society at large. Finally, we believe natural language processing to be a field that allows for a great variety of interesting models and this project will allow us to work with many methods of deep learning that are regularly used in research and industry.

In addition, you can also try logistic regression to start with as a baseline.

3. Evaluation

Our primary focus is on developing a prediction engine that can reliably beat the model developed by Schler et al. [3]. Therefore, comparisons to their results will serve as our baseline for evaluation. Our simplest evaluation metric will be overall accuracy of predictions of both age and gender. We hope to improve at least 5% accuracy on both

1

Example Proposal

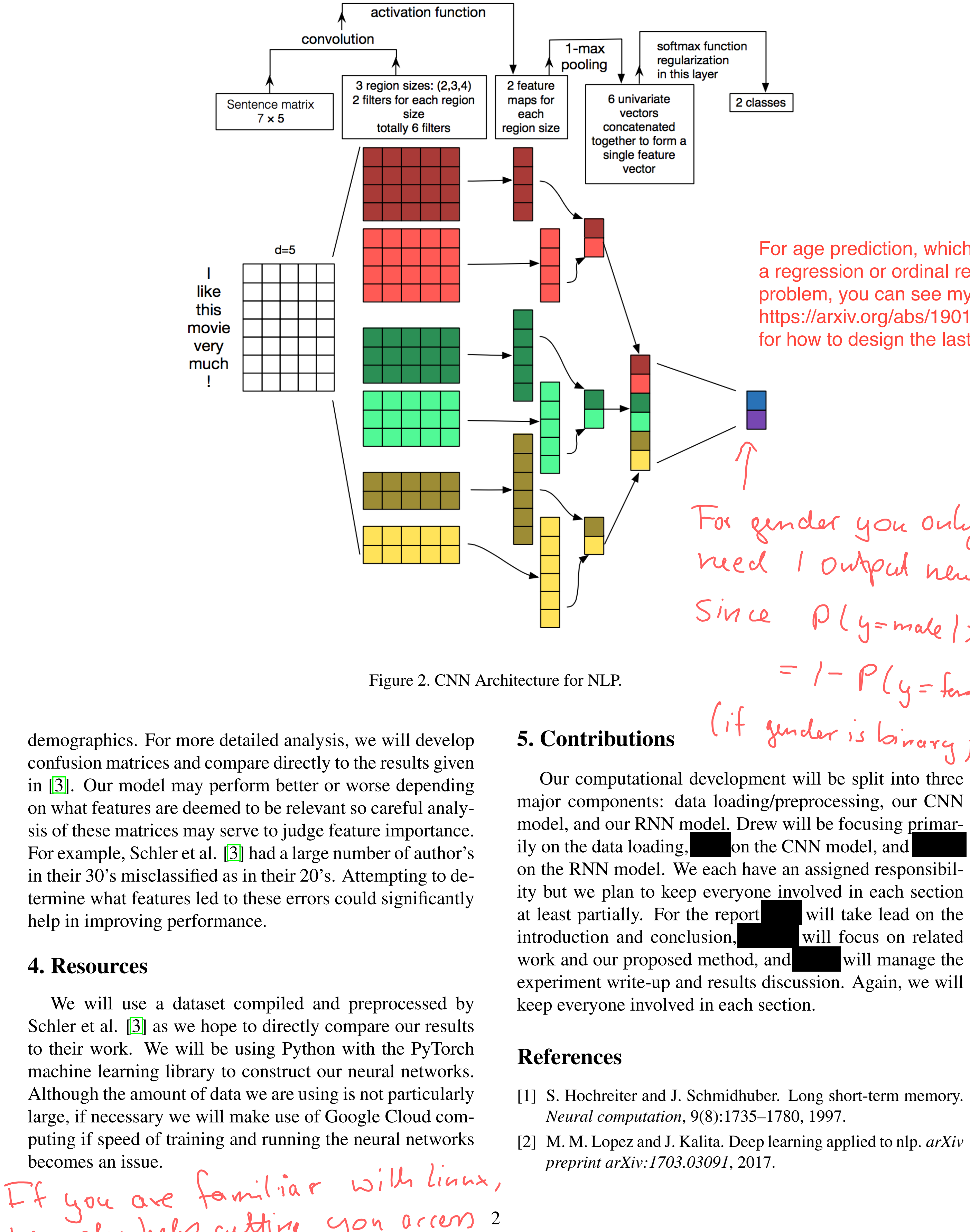

activation function
convolution
1-max pooling
softmax function regularization in this layer
Sentence matrix 7 × 5
3 region sizes: (2,3,4) 2 filters for each region size totally 6 filters
2 feature maps for each region size
6 univariate vectors concatenated together to form a single feature vector
2 classes
d=5
I
like
this
movie
very
much
!

Figure 2. CNN Architecture for NLP.

demographics. For more detailed analysis, we will develop confusion matrices and compare directly to the results given in [3]. Our model may perform better or worse depending on what features are deemed to be relevant so careful analysis of these matrices may serve to judge feature importance. For example, Schler et al. [3] had a large number of author's in their 30's misclassified as in their 20's. Attempting to determine what features led to these errors could significantly help in improving performance.

4. Resources

We will use a dataset compiled and preprocessed by Schler et al. [3] as we hope to directly compare our results to their work. We will be using Python with the PyTorch machine learning library to construct our neural networks. Although the amount of data we are using is not particularly large, if necessary we will make use of Google Cloud computing if speed of training and running the neural networks becomes an issue.

5. Contributions

Our computational development will be split into three major components: data loading/preprocessing, our CNN model, and our RNN model. Drew will be focusing primarily on the data loading, [redacted] on the CNN model, and [redacted] on the RNN model. We each have an assigned responsibility but we plan to keep everyone involved in each section at least partially. For the report [redacted] will take lead on the introduction and conclusion, [redacted] will focus on related work and our proposed method, and [redacted] will manage the experiment write-up and results discussion. Again, we will keep everyone involved in each section.

References

[1] S. Hochreiter and J. Schmidhuber. Long short-term memory. *Neural computation*, 9(8):1735–1780, 1997.

[2] M. M. Lopez and J. Kalita. Deep learning applied to nlp. *arXiv preprint arXiv:1703.03091*, 2017.

2

For age prediction, which is either a regression or ordinal regression problem, you can see my preprint https://arxiv.org/abs/1901.07884 for how to design the last layer

For gender you only need 1 output neuron, since $p(y=male|x) = 1 - P(y=female|x)$ (if gender is binary)

If you are familiar with linux, I can also help getting you access to the department cluster.

Citing Verbatim

Some text. "Methods based on voting can reduce variance" [1].
Some text.

References

[1] Dietterich, T. G., & Kong, E. B. (1995). Machine learning bias, statistical bias, and statistical variance of decision tree algorithms. Technical report, Department of Computer Science, Oregon State University.

Paraphrasing

Some text. It has been shown that the variance of a learning algorithm can be lowered by combining multiple models via voting [1]. Some text.

References

[1] Dietterich, T. G., & Kong, E. B. (1995). Machine learning bias, statistical bias, and statistical variance of decision tree algorithms. Technical report, Department of Computer Science, Oregon State University.

"Citing" Images

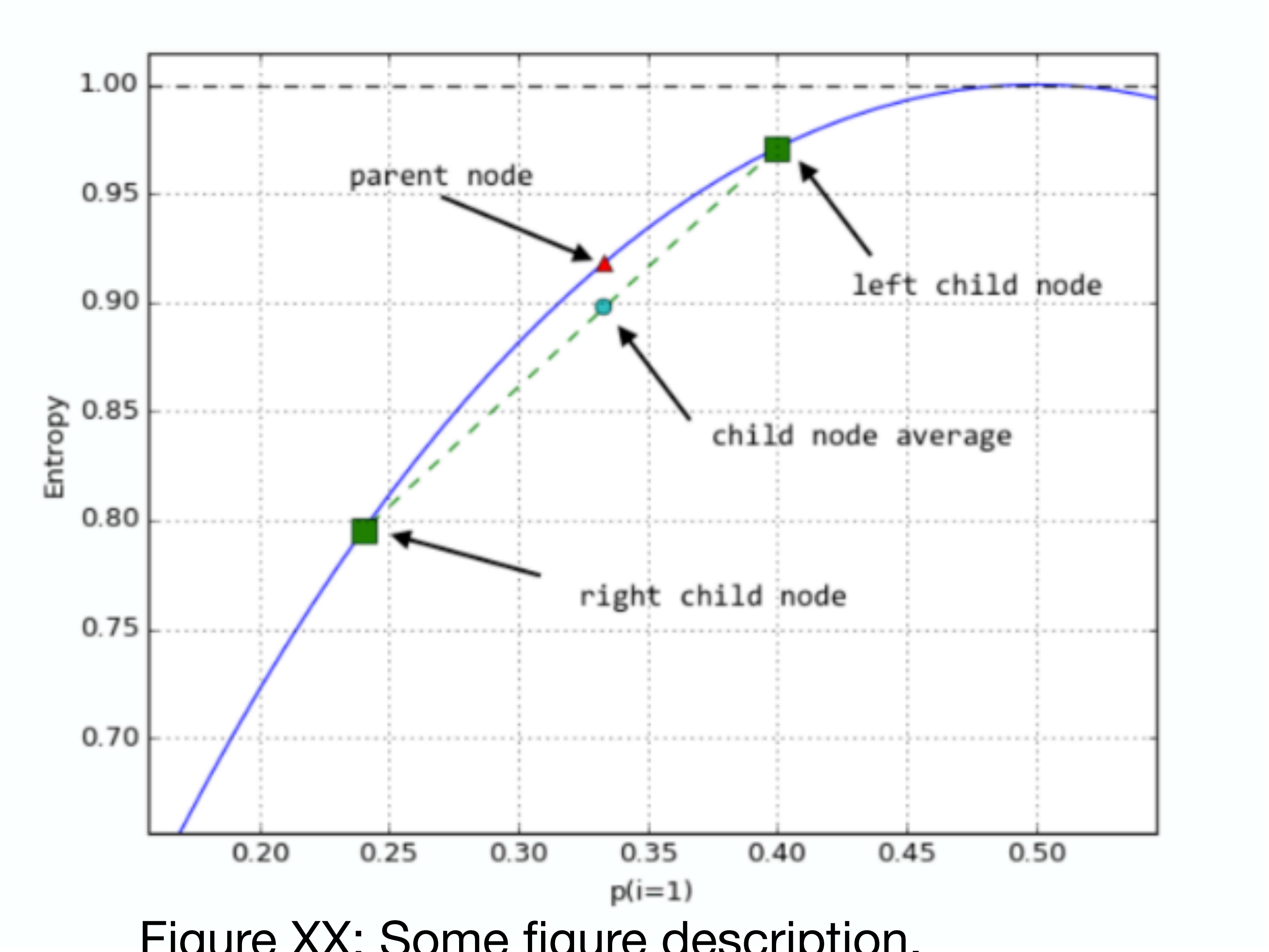
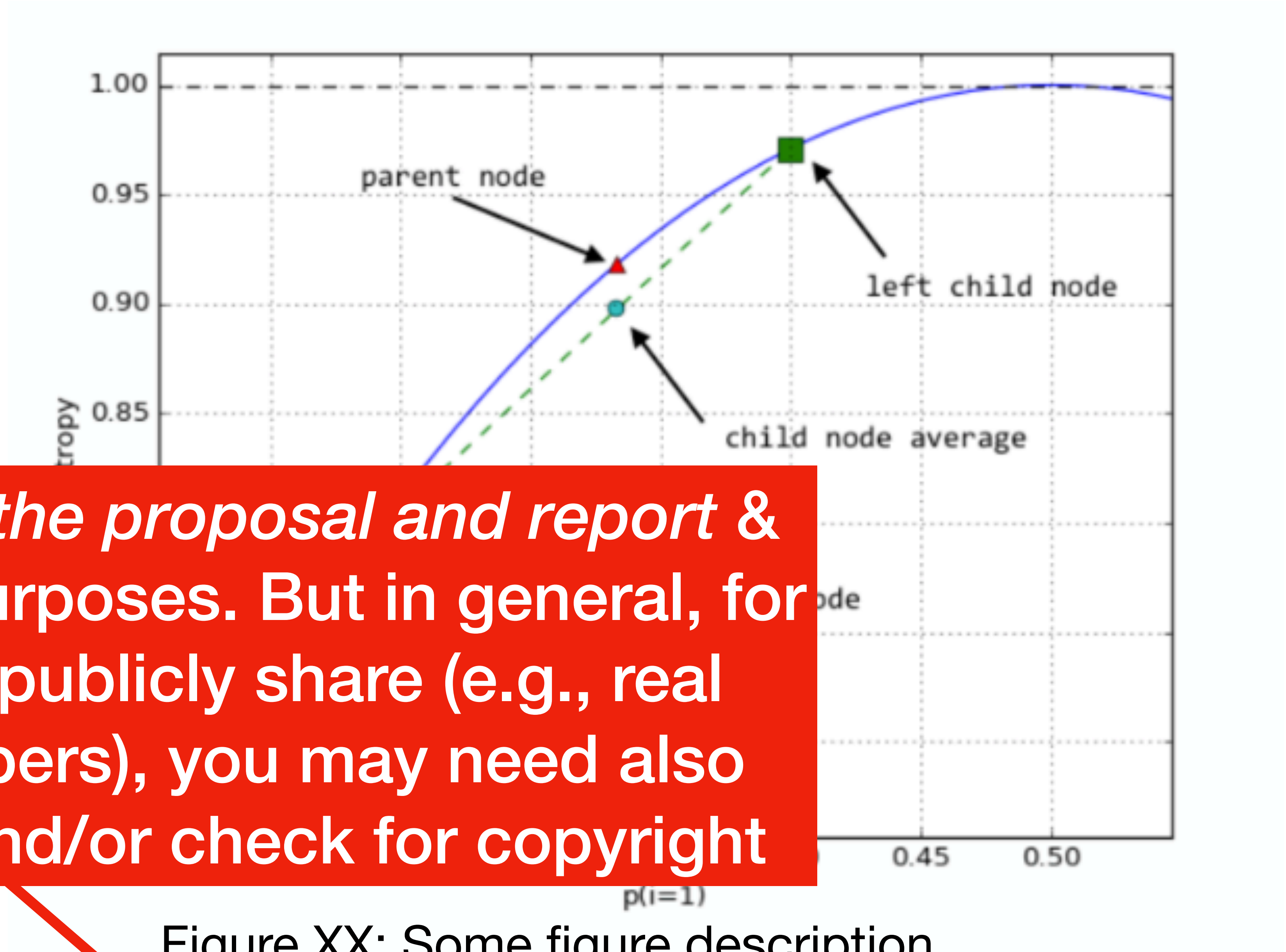


Figure XX: Some figure description.
Source: https://github.com/rasbt/stat479-machine-learning-fs19/blob/master/06_trees/06-trees__slides.pdf

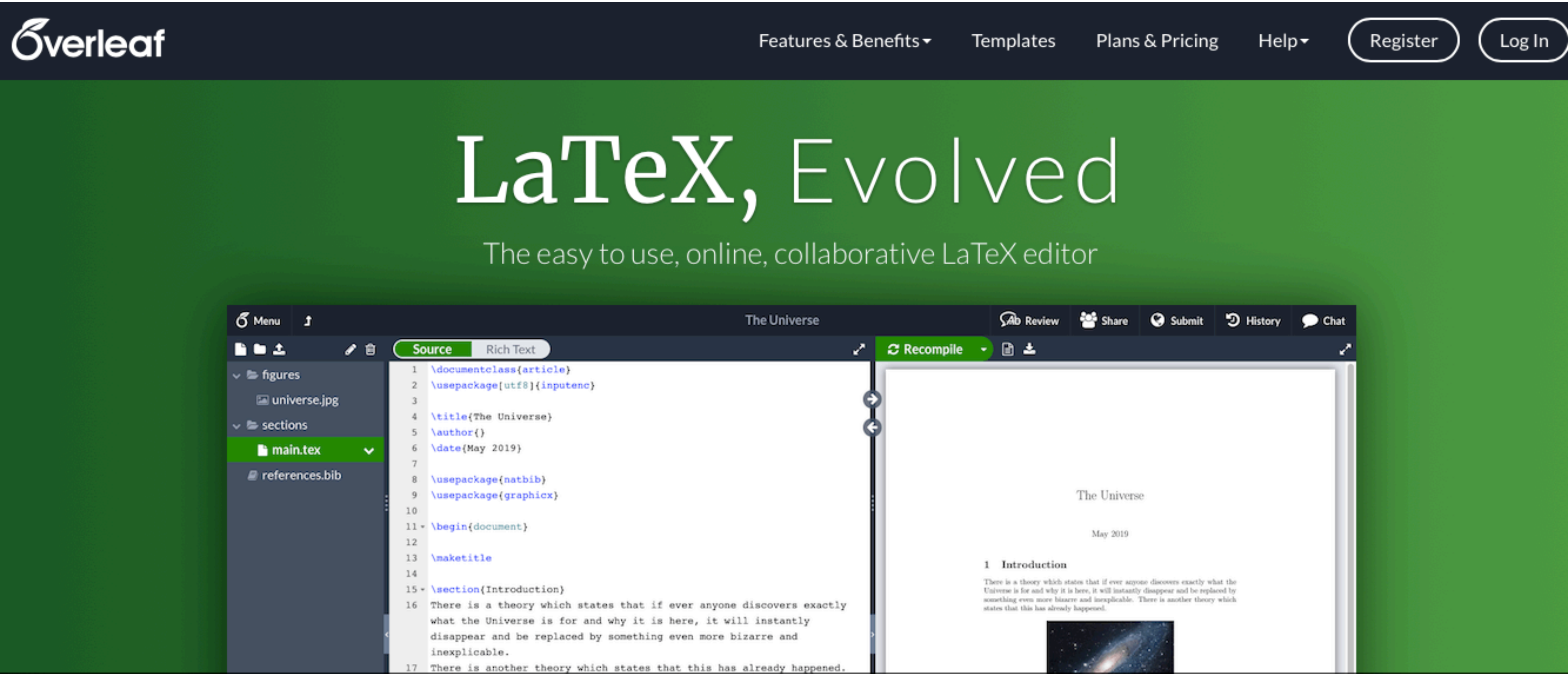
"Citing" Images



Sufficient for *the proposal and report* & educational purposes. But in general, for papers you publicly share (e.g., real research papers), you may need also permission and/or check for copyright

Figure XX: Some figure description.
Source: https://github.com/rasbt/stat479-machine-learning-fs19/blob/master/06_trees/06-trees__slides.pdf

Optional Tool/Platform for collaborative writing in LaTeX



LaTeX, Evolved

The easy to use, online, collaborative LaTeX editor

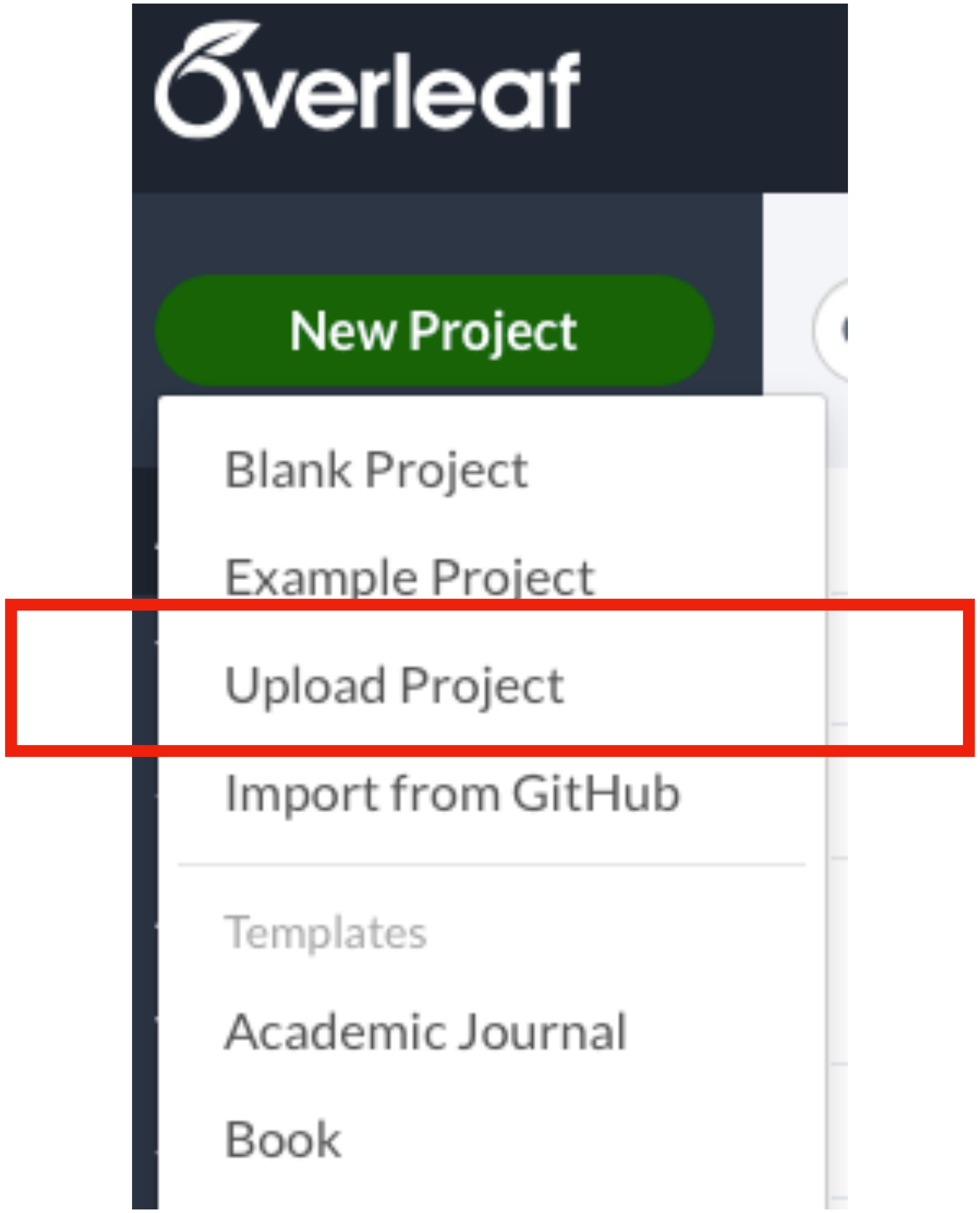
```
\documentclass{article}
\usepackage[utf8]{inputenc}

\title{The Universe}
\author{}
\date{May 2019}

\usepackage{natbib}
\usepackage{graphicx}

\begin{document}

\maketitle

\section{Introduction}
There is a theory which states that if ever anyone discovers exactly
what the Universe is for and why it is here, it will instantly
disappear and be replaced by something even more bizarre and
inexplicable.
There is another theory which states that this has already happened.
```

The Universe

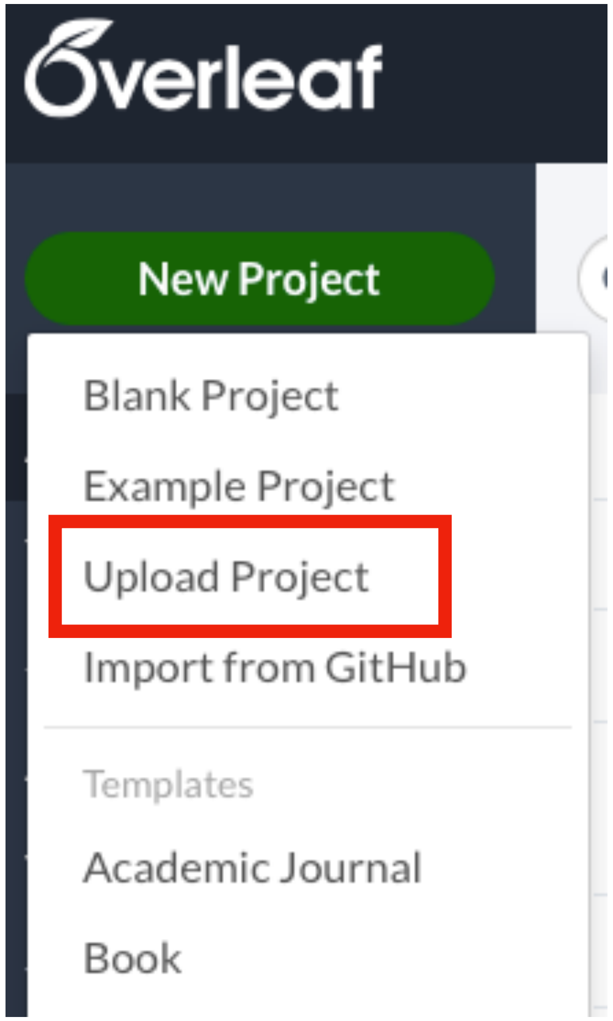
May 2019

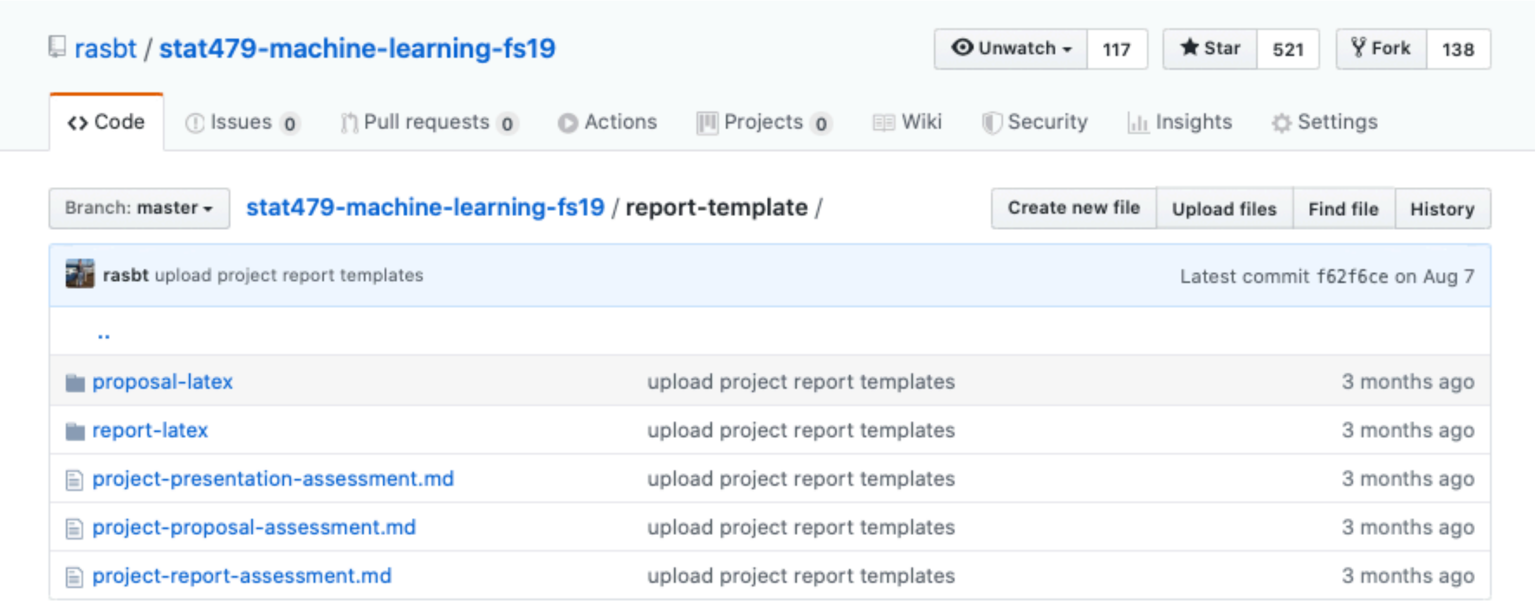
1 Introduction

There is a theory which states that if ever anyone discovers exactly what the Universe is for and why it is here, it will instantly disappear and be replaced by something even more bizarre and inexplicable. There is another theory which states that this has already happened.



Overleaf
New Project
Blank Project
Example Project
Upload Project
Import from GitHub
Templates
Academic Journal
Book



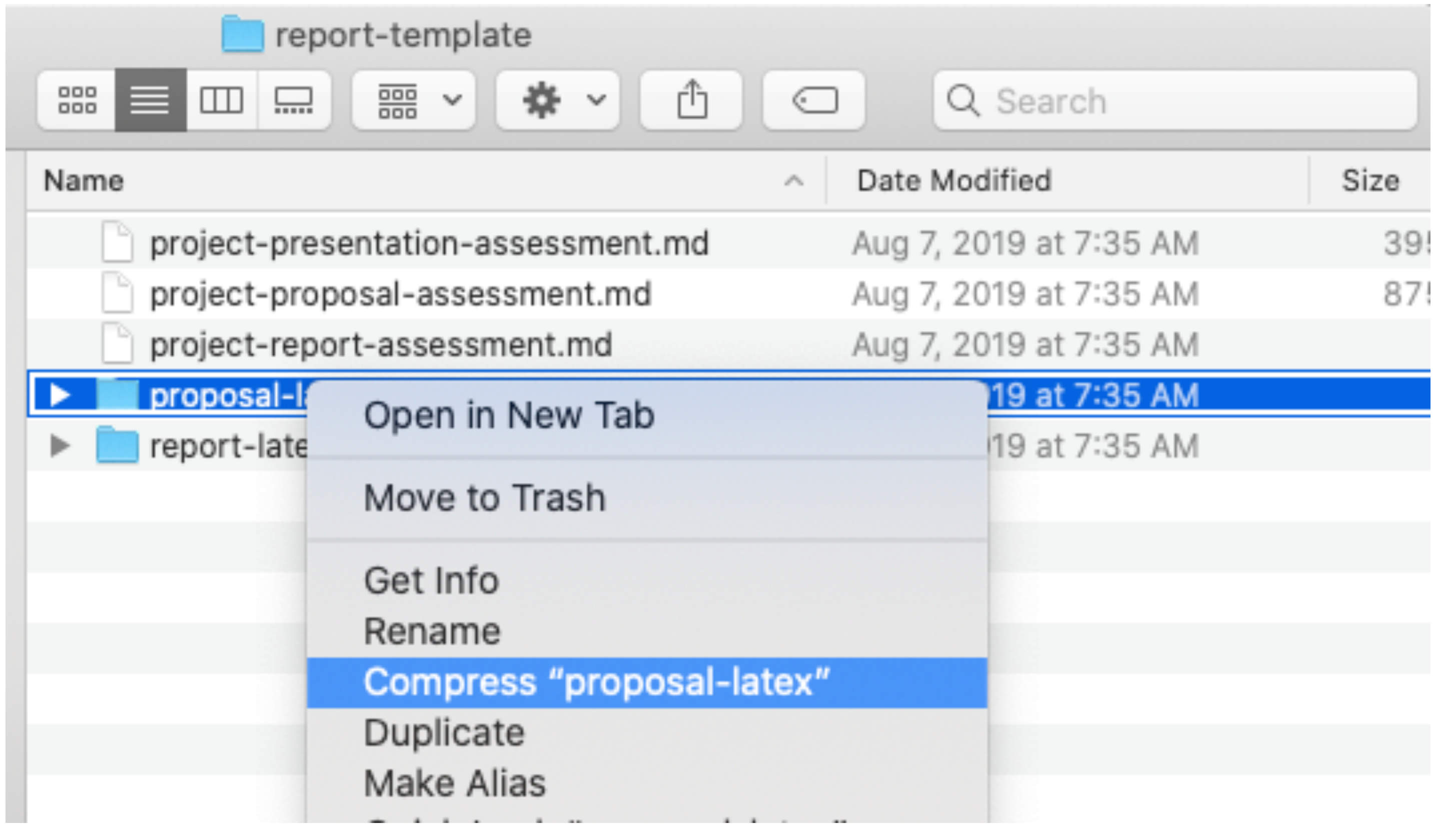


stat479-machine-learning-fs19 / report-template /

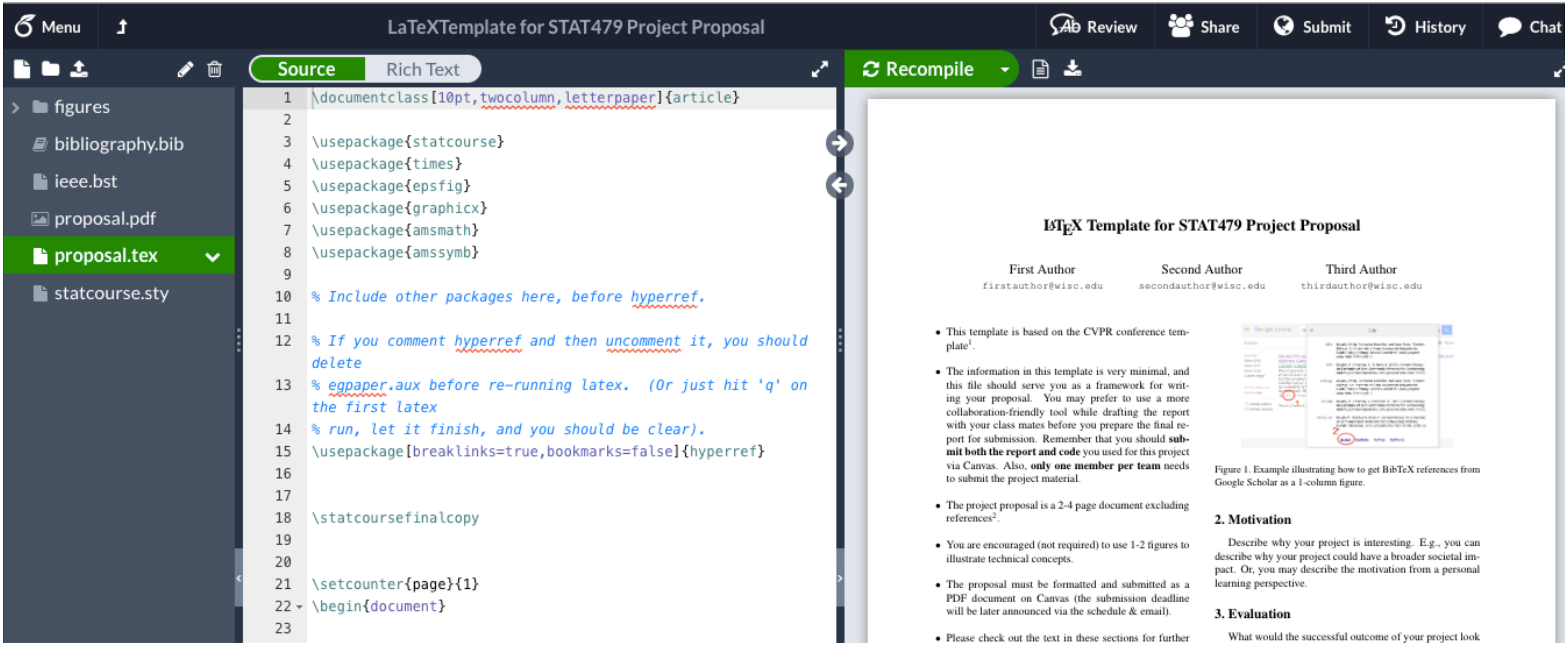
rasbt upload project report templates — Latest commit f62f6ce on Aug 7

..		
proposal-latex	upload project report templates	3 months ago
report-latex	upload project report templates	3 months ago
project-presentation-assessment.md	upload project report templates	3 months ago
project-proposal-assessment.md	upload project report templates	3 months ago
project-report-assessment.md	upload project report templates	3 months ago

https://github.com/rasbt/stat479-machine-learning-fs19/tree/master/report-template



Name	Date Modified	Size
project-presentation-assessment.md	Aug 7, 2019 at 7:35 AM	39
project-proposal-assessment.md	Aug 7, 2019 at 7:35 AM	87
project-report-assessment.md	Aug 7, 2019 at 7:35 AM	
▶ proposal-l	19 at 7:35 AM	
▶ report-late	19 at 7:35 AM	



```
\documentclass[10pt,twocolumn,letterpaper]{article}

\usepackage{statcourse}
\usepackage{times}
\usepackage{epsfig}
\usepackage{graphicx}
\usepackage{amsmath}
\usepackage{amssymb}

% Include other packages here, before hyperref.

% If you comment hyperref and then uncomment it, you should delete
% egpaper.aux before re-running latex.  (Or just hit 'q' on the first latex
% run, let it finish, and you should be clear).
\usepackage[breaklinks=true,bookmarks=false]{hyperref}


\statcoursefinalcopy


\setcounter{page}{1}
\begin{document}

```

LaTeX Template for STAT479 Project Proposal

First Author
firstauthor@wisc.edu

Second Author
secondauthor@wisc.edu

Third Author
thirdauthor@wisc.edu

- This template is based on the CVPR conference template[1].
- The information in this template is very minimal, and this file should serve you as a framework for writing your proposal. You may prefer to use a more collaboration-friendly tool while drafting the report with your class mates before you prepare the final report for submission. Remember that you should **submit both the report and code** you used for this project via Canvas. Also, **only one member per team** needs to submit the project material.
- The project proposal is a 2-4 page document excluding references[2].
- You are encouraged (not required) to use 1-2 figures to illustrate technical concepts.
- The proposal must be formatted and submitted as a PDF document on Canvas (the submission deadline will be later announced via the schedule & email).
- Please check out the text in these sections for further

Figure 1. Example illustrating how to get BibTeX references from Google Scholar as a 1-column figure.

2. Motivation

Describe why your project is interesting. E.g., you can describe why your project could have a broader societal impact. Or, you may describe the motivation from a personal learning perspective.

3. Evaluation

What would the successful outcome of your project look

Lecture 08

Model Evaluation Part 1: Introduction to Overfitting and Underfitting

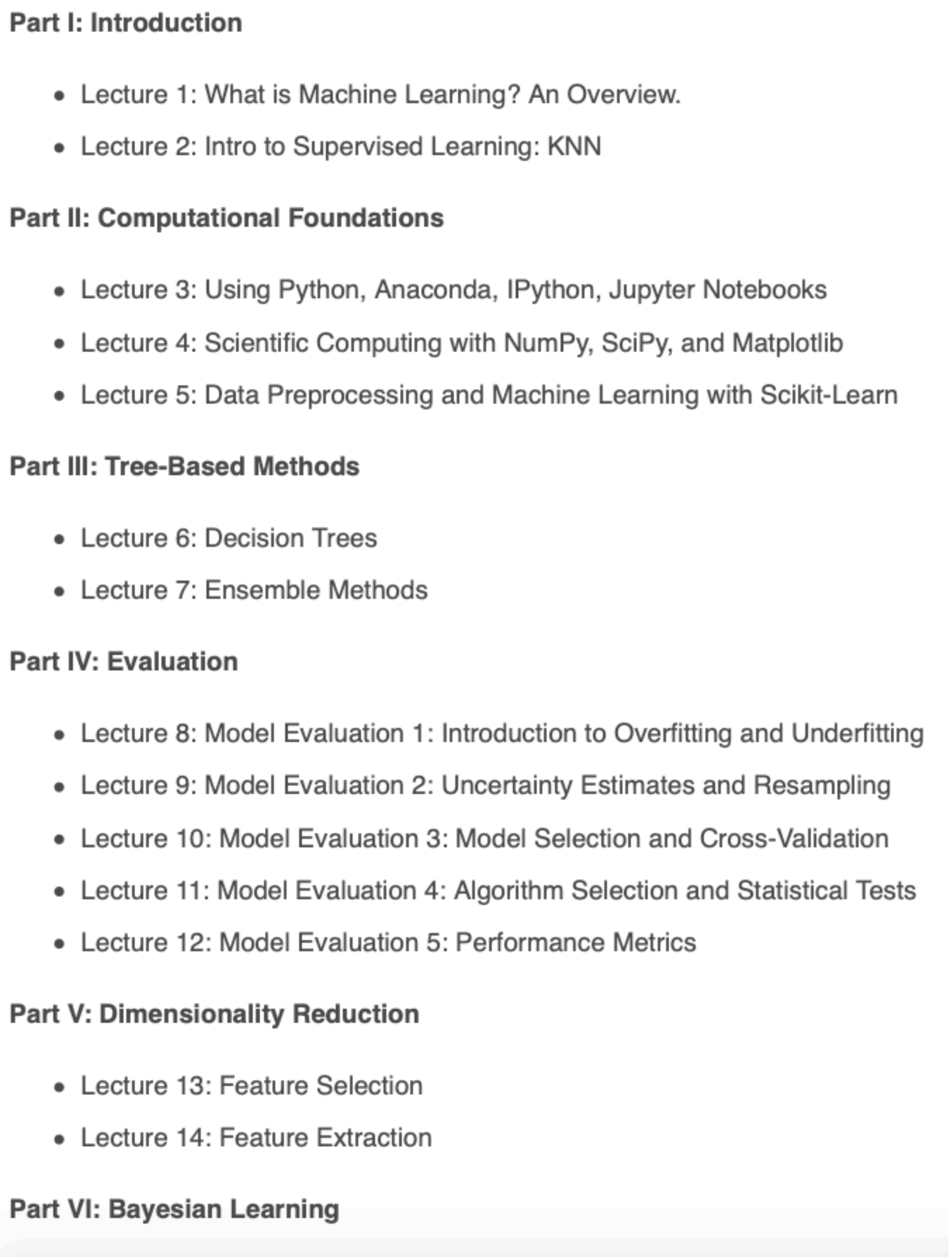
STAT 479: Machine Learning, Fall 2019

Sebastian Raschka

http://stat.wisc.edu/~sraschka/teaching/stat479-fs2019/

Where we are in this course

http://pages.stat.wisc.edu/~sraschka/teaching/stat479-fs2019/#topics-summary-planned



Part I: Introduction

- Lecture 1: What is Machine Learning? An Overview.
- Lecture 2: Intro to Supervised Learning: KNN

Part II: Computational Foundations

- Lecture 3: Using Python, Anaconda, IPython, Jupyter Notebooks
- Lecture 4: Scientific Computing with NumPy, SciPy, and Matplotlib
- Lecture 5: Data Preprocessing and Machine Learning with Scikit-Learn

Part III: Tree-Based Methods

- Lecture 6: Decision Trees
- Lecture 7: Ensemble Methods

Part IV: Evaluation

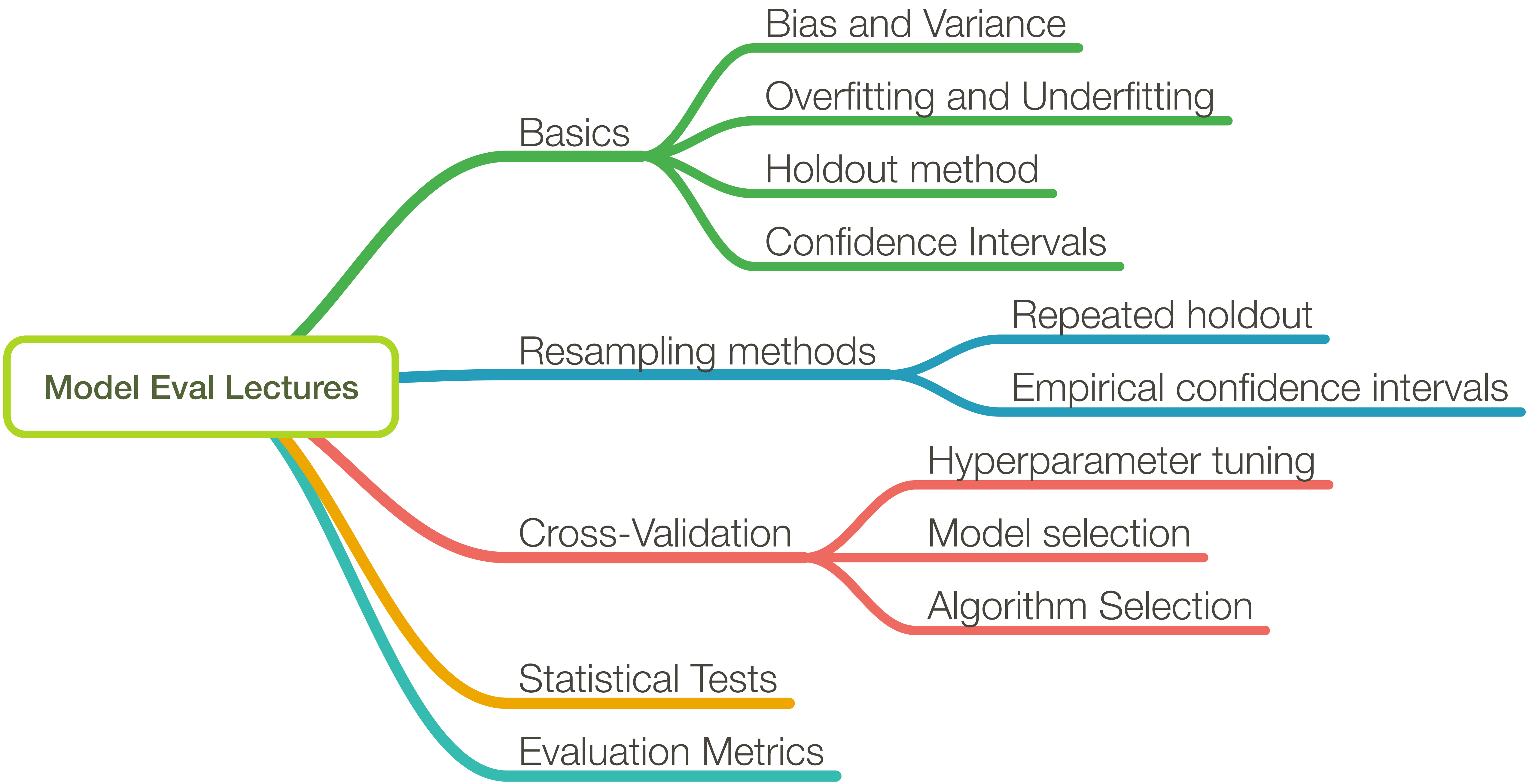
- Lecture 8: Model Evaluation 1: Introduction to Overfitting and Underfitting
- Lecture 9: Model Evaluation 2: Uncertainty Estimates and Resampling
- Lecture 10: Model Evaluation 3: Model Selection and Cross-Validation
- Lecture 11: Model Evaluation 4: Algorithm Selection and Statistical Tests
- Lecture 12: Model Evaluation 5: Performance Metrics

Part V: Dimensionality Reduction

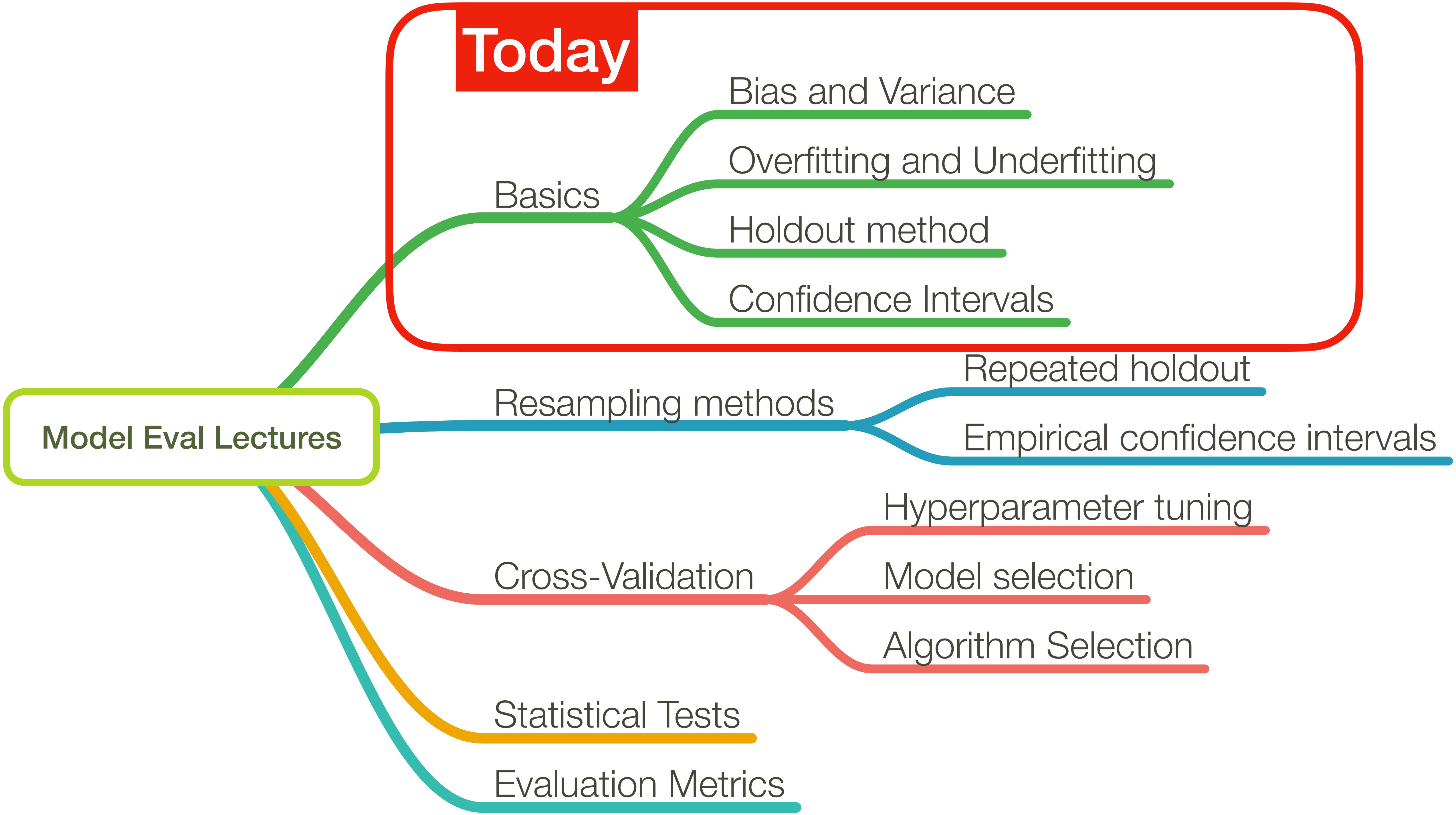
- Lecture 13: Feature Selection
- Lecture 14: Feature Extraction

Part VI: Bayesian Learning

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
Evaluation Metrics

Overview


Today
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
Evaluation Metrics

Overfitting and Underfitting

Overfitting and Underfitting

Generalization Performance

Want a model to "generalize" well to _________ data

(Want "high generalization accuracy" or "low generalization error")

Overfitting and Underfitting

Assumptions

- i.i.d. assumption: training and test examples are independent and identically distributed (drawn from the same joint probability distribution, P(**X**, y))

- For some random model that has not been fitted to the training set,
  we expect **the training error is ______ _____ the test error**

- The training error or accuracy provides
  an _______imistically biased estimate of the generalization performance

Overfitting and Underfitting

Model Capacity

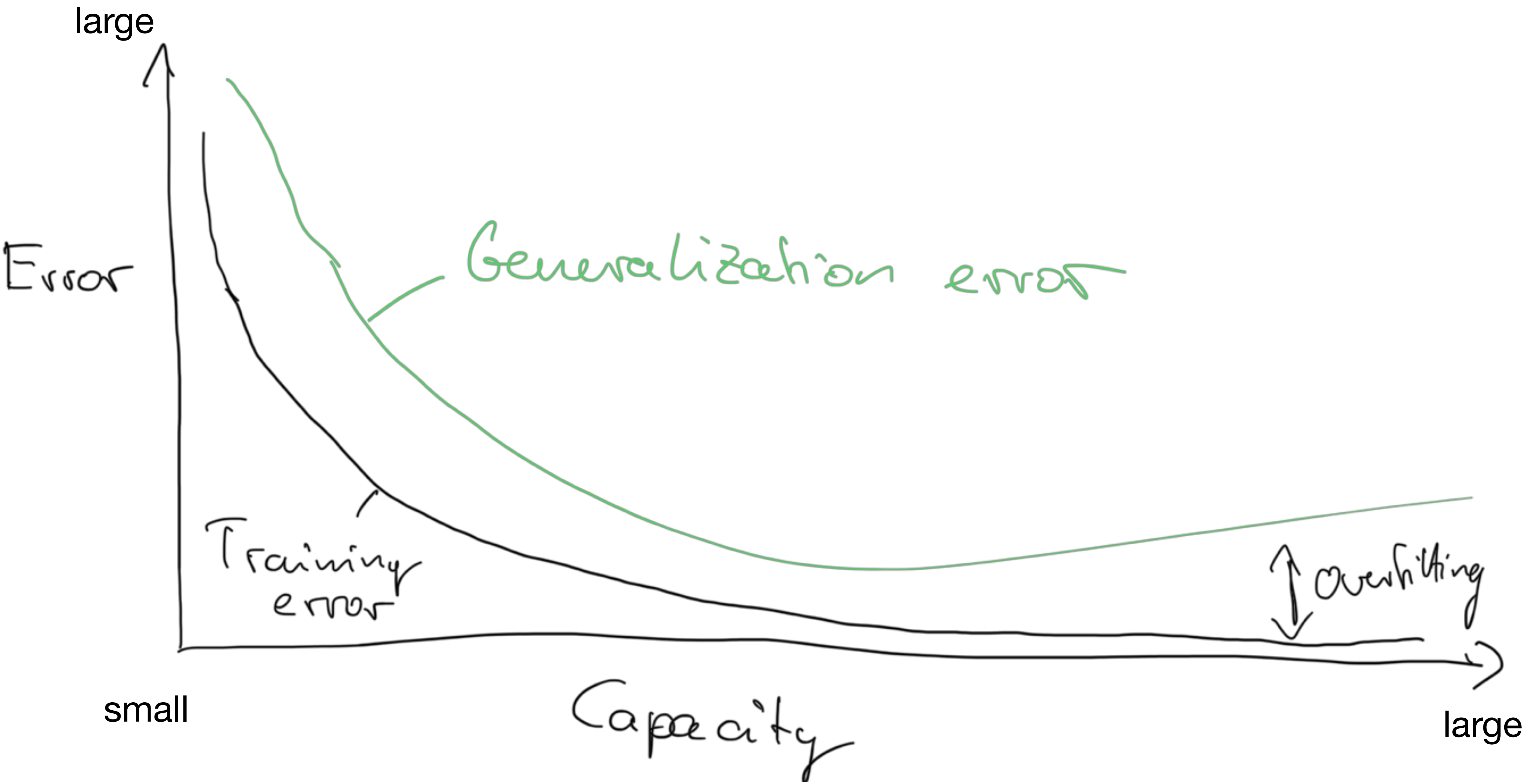
- Underfitting: both the training and test error are _________

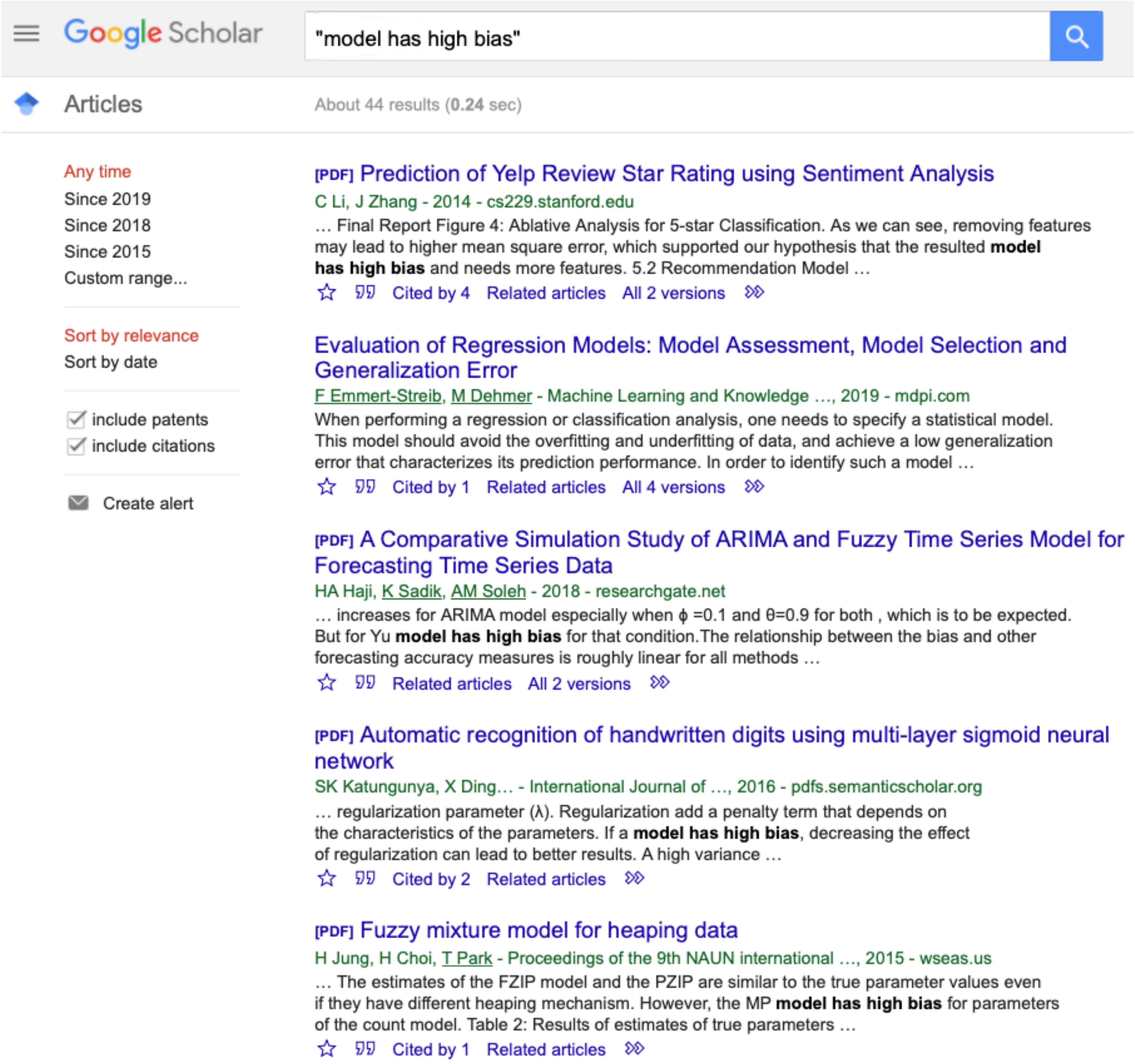
- Overfitting: gap between training and test error (where test error is larger)

- Large hypothesis space being searched by a learning algorithm

-> high tendency to ______fit

Overfitting and Underfitting





"model has high bias"

About 44 results (**0.24** sec)

[PDF] Prediction of Yelp Review Star Rating using Sentiment Analysis

C Li, J Zhang - 2014 - cs229.stanford.edu

… Final Report Figure 4: Ablative Analysis for 5-star Classification. As we can see, removing features may lead to higher mean square error, which supported our hypothesis that the resulted **model has high bias** and needs more features. 5.2 Recommendation Model …

☆ ❞ Cited by 4 Related articles All 2 versions »

Evaluation of Regression Models: Model Assessment, Model Selection and Generalization Error

F Emmert-Streib, M Dehmer - Machine Learning and Knowledge …, 2019 - mdpi.com

When performing a regression or classification analysis, one needs to specify a statistical model. This model should avoid the overfitting and underfitting of data, and achieve a low generalization error that characterizes its prediction performance. In order to identify such a model …

☆ ❞ Cited by 1 Related articles All 4 versions »

[PDF] A Comparative Simulation Study of ARIMA and Fuzzy Time Series Model for Forecasting Time Series Data

HA Haji, K Sadik, AM Soleh - 2018 - researchgate.net

… increases for ARIMA model especially when ϕ =0.1 and θ=0.9 for both , which is to be expected. But for Yu **model has high bias** for that condition.The relationship between the bias and other forecasting accuracy measures is roughly linear for all methods …

☆ ❞ Related articles All 2 versions »

[PDF] Automatic recognition of handwritten digits using multi-layer sigmoid neural network

SK Katungunya, X Ding… - International Journal of …, 2016 - pdfs.semanticscholar.org

… regularization parameter (λ). Regularization add a penalty term that depends on the characteristics of the parameters. If a **model has high bias**, decreasing the effect of regularization can lead to better results. A high variance …

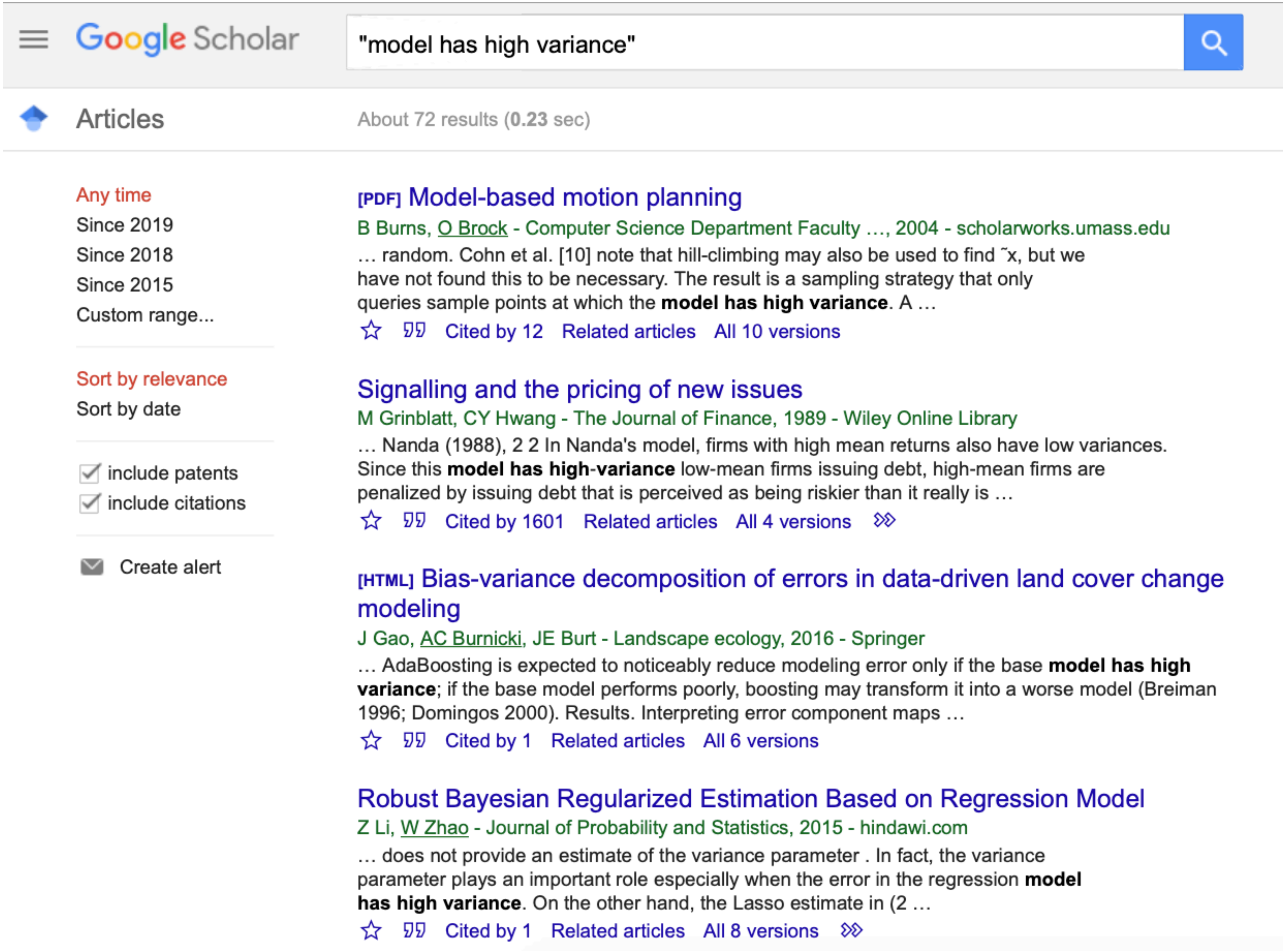
☆ ❞ Cited by 2 Related articles »

[PDF] Fuzzy mixture model for heaping data

H Jung, H Choi, T Park - Proceedings of the 9th NAUN international …, 2015 - wseas.us

… The estimates of the FZIP model and the PZIP are similar to the true parameter values even if they have different heaping mechanism. However, the MP **model has high bias** for parameters of the count model. Table 2: Results of estimates of true parameters …

☆ ❞ Cited by 1 Related articles »



"model has high variance"

About 72 results (**0.23** sec)

[PDF] Model-based motion planning

B Burns, O Brock - Computer Science Department Faculty …, 2004 - scholarworks.umass.edu

… random. Cohn et al. [10] note that hill-climbing may also be used to find ˜x, but we have not found this to be necessary. The result is a sampling strategy that only queries sample points at which the **model has high variance**. A …

☆ 99 Cited by 12 Related articles All 10 versions

Signalling and the pricing of new issues

M Grinblatt, CY Hwang - The Journal of Finance, 1989 - Wiley Online Library

… Nanda (1988), 2 2 In Nanda's model, firms with high mean returns also have low variances. Since this **model has high-variance** low-mean firms issuing debt, high-mean firms are penalized by issuing debt that is perceived as being riskier than it really is …

☆ 99 Cited by 1601 Related articles All 4 versions »

[HTML] Bias-variance decomposition of errors in data-driven land cover change modeling

J Gao, AC Burnicki, JE Burt - Landscape ecology, 2016 - Springer

… AdaBoosting is expected to noticeably reduce modeling error only if the base **model has high variance**; if the base model performs poorly, boosting may transform it into a worse model (Breiman 1996; Domingos 2000). Results. Interpreting error component maps …

☆ 99 Cited by 1 Related articles All 6 versions

Robust Bayesian Regularized Estimation Based on Regression Model

Z Li, W Zhao - Journal of Probability and Statistics, 2015 - hindawi.com

… does not provide an estimate of the variance parameter . In fact, the variance parameter plays an important role especially when the error in the regression **model has high variance**. On the other hand, the Lasso estimate in (2 …

☆ 99 Cited by 1 Related articles All 8 versions »

"[...] model has high bias/variance"
-- What does that mean?

Bias-Variance Decomposition and Bias-Variance Trade-off

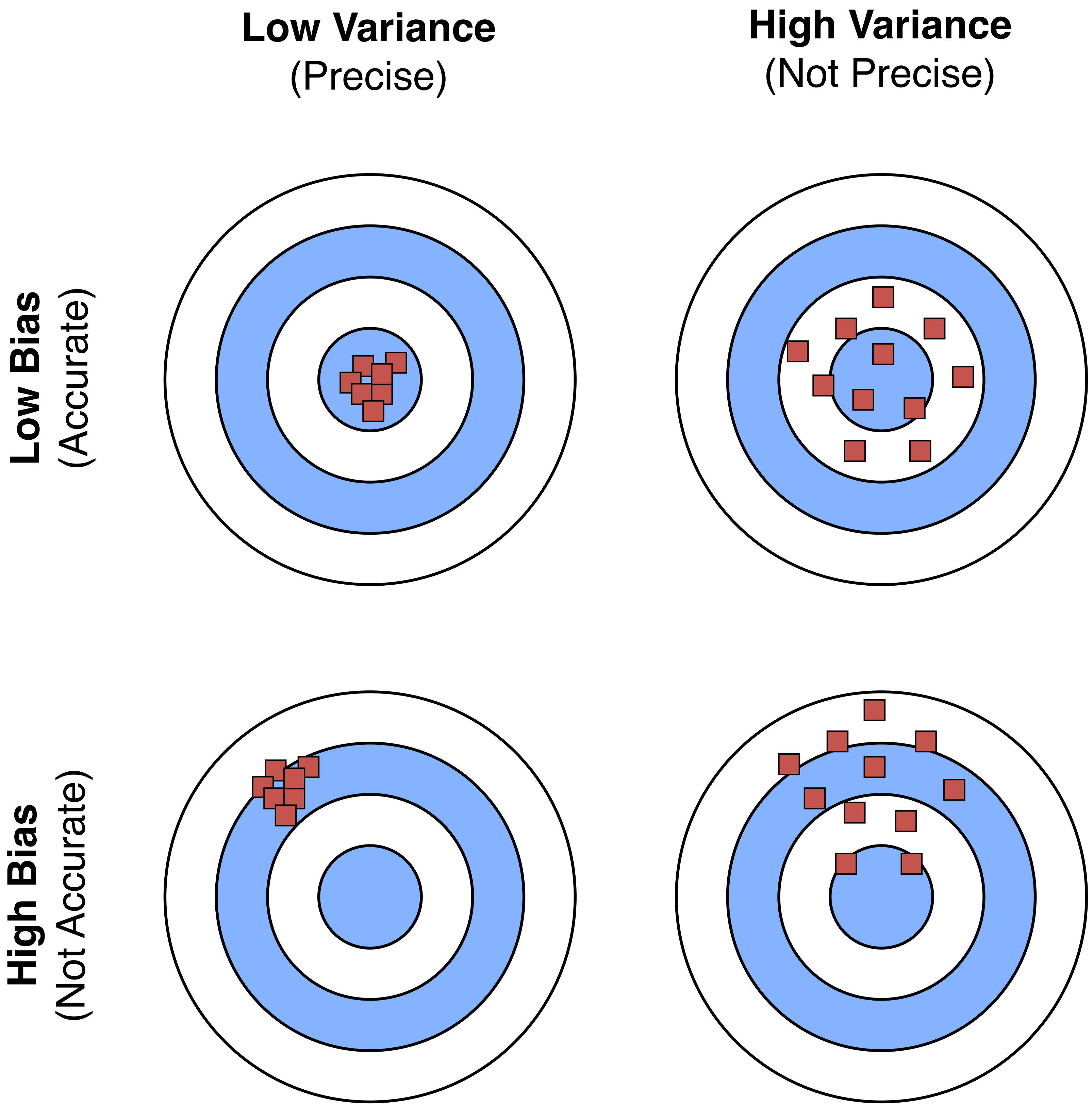
(and how it related to overfitting and underfitting)

Bias-Variance Decomposition

- Decomposition of the loss into bias and variance help us understand learning algorithms, concepts are related to underfitting and overfitting

- Helps explain why ensemble methods (last lecture) might perform better than single models

Bias-Variance Intuition



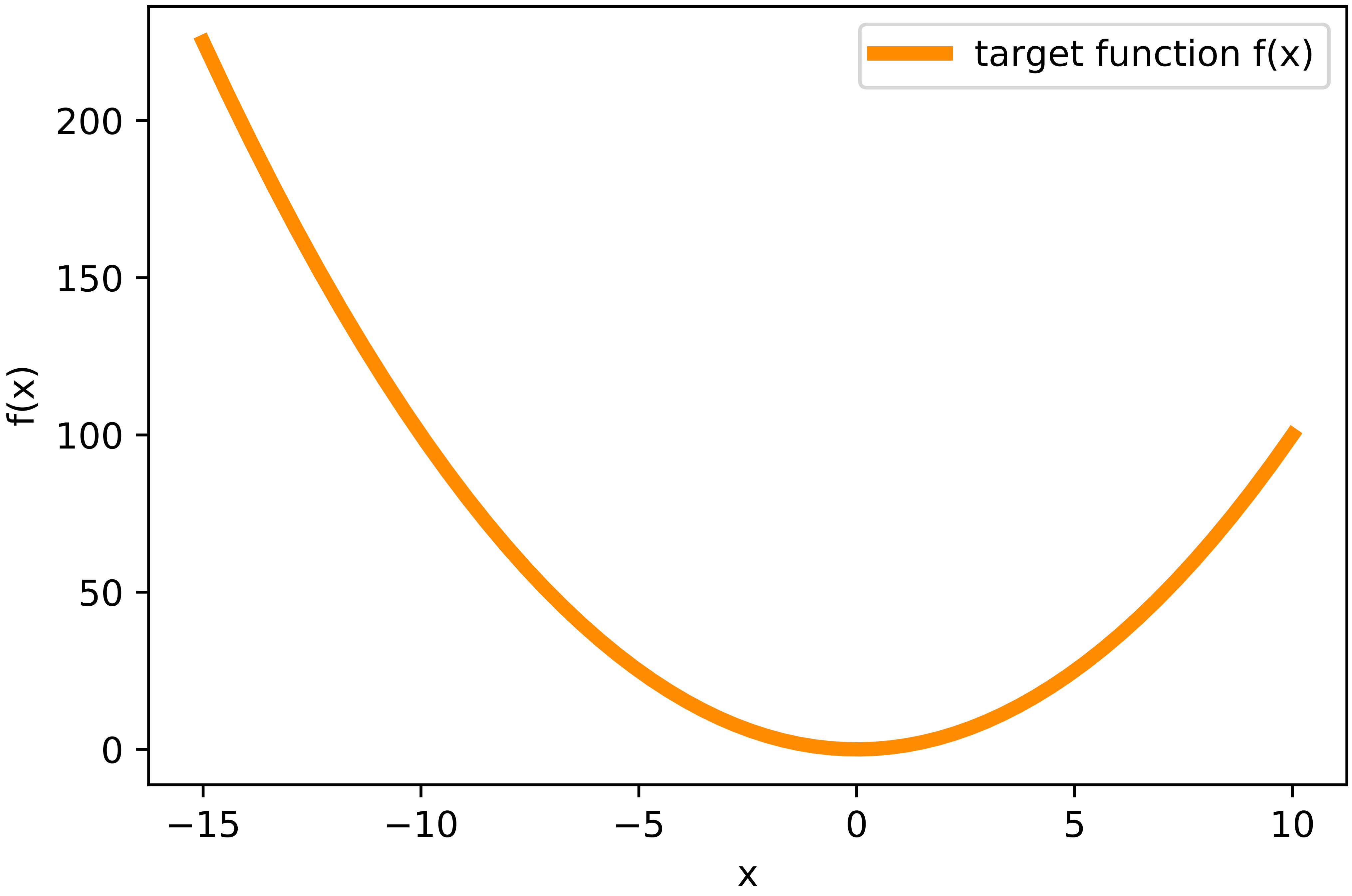
Low Variance
(Precise)

High Variance
(Not Precise)

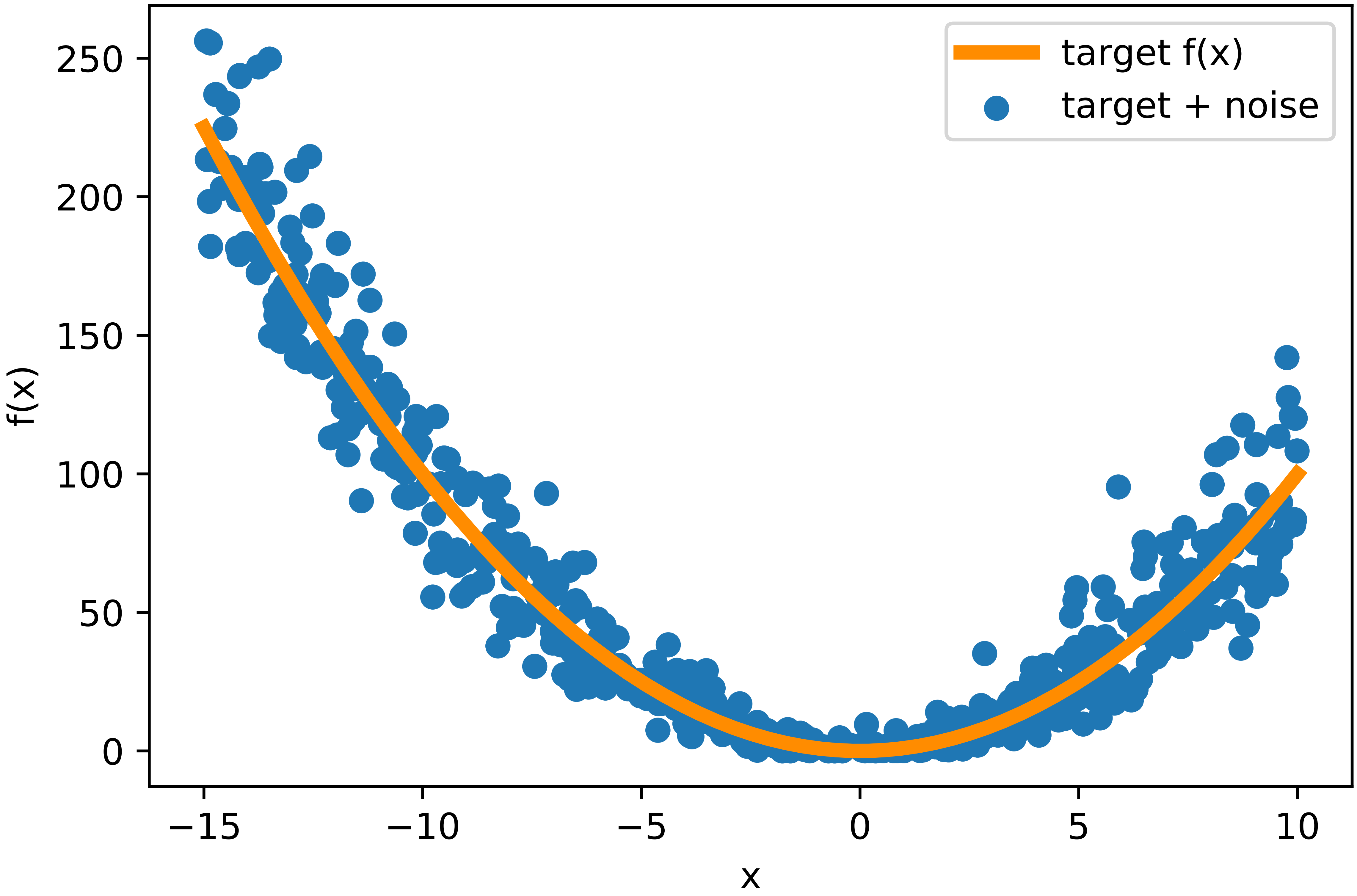
Low Bias
(Accurate)

High Bias
(Not Accurate)

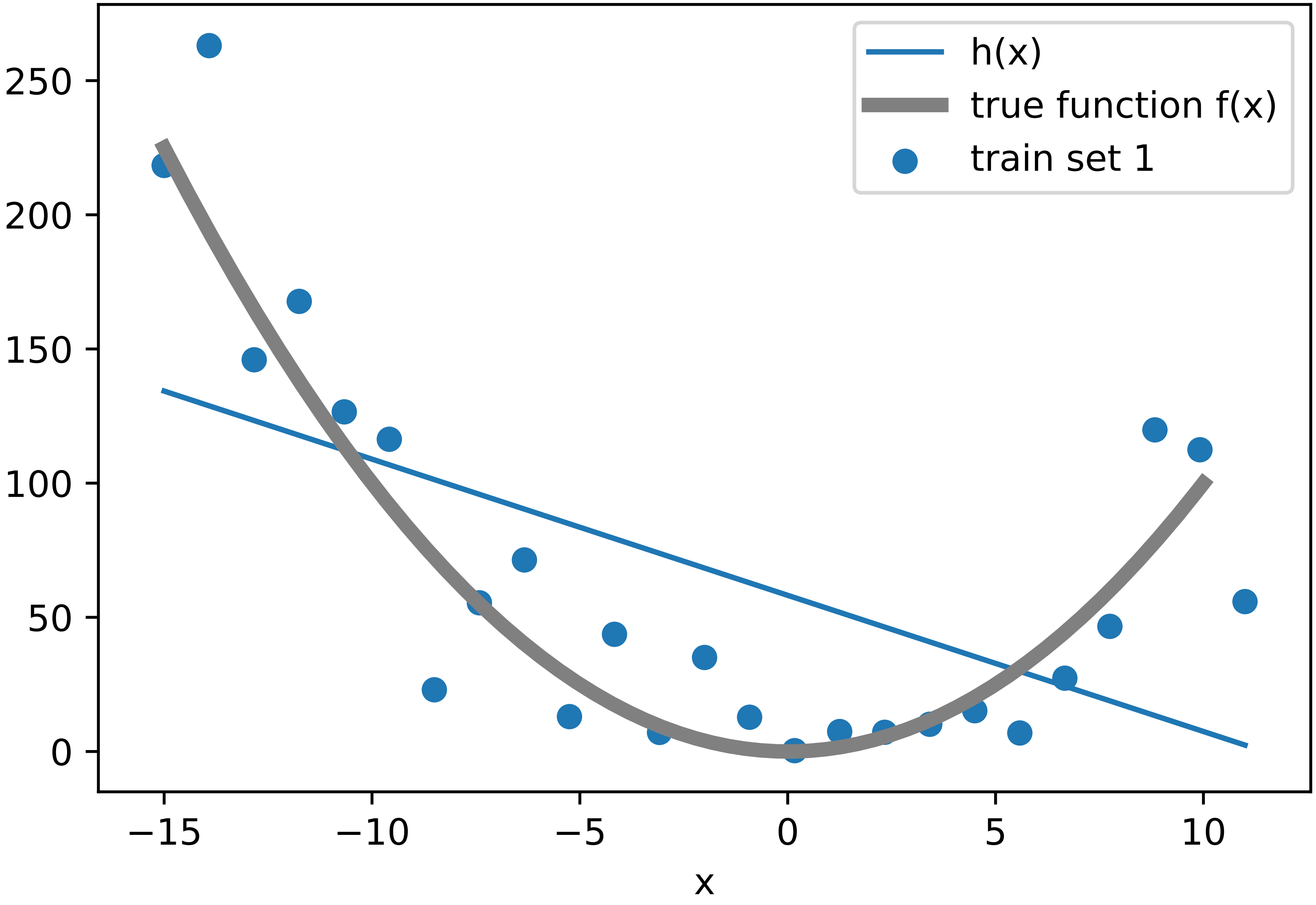
Bias-Variance Intuition



Bias-Variance Intuition

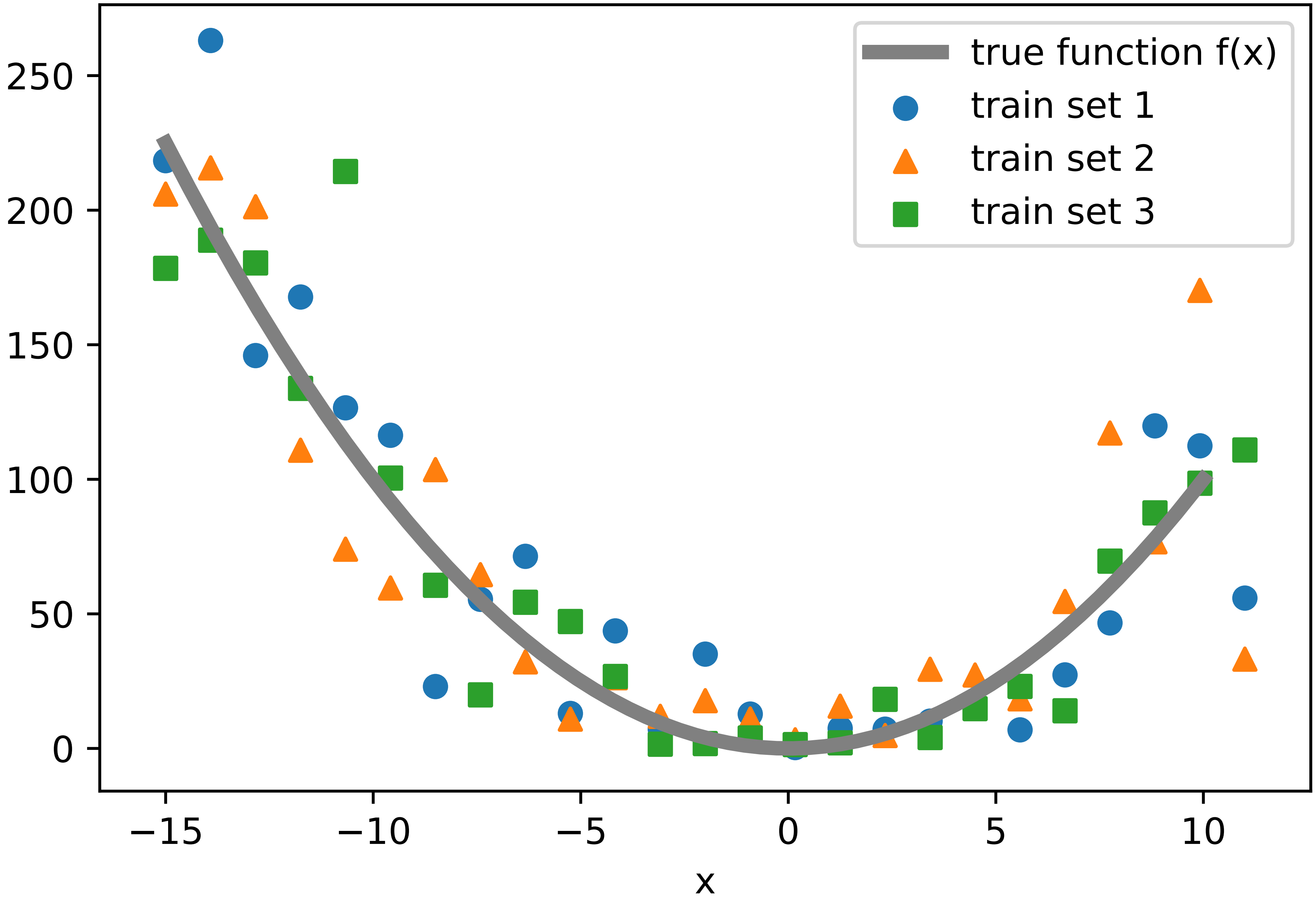


Bias-Variance Intuition

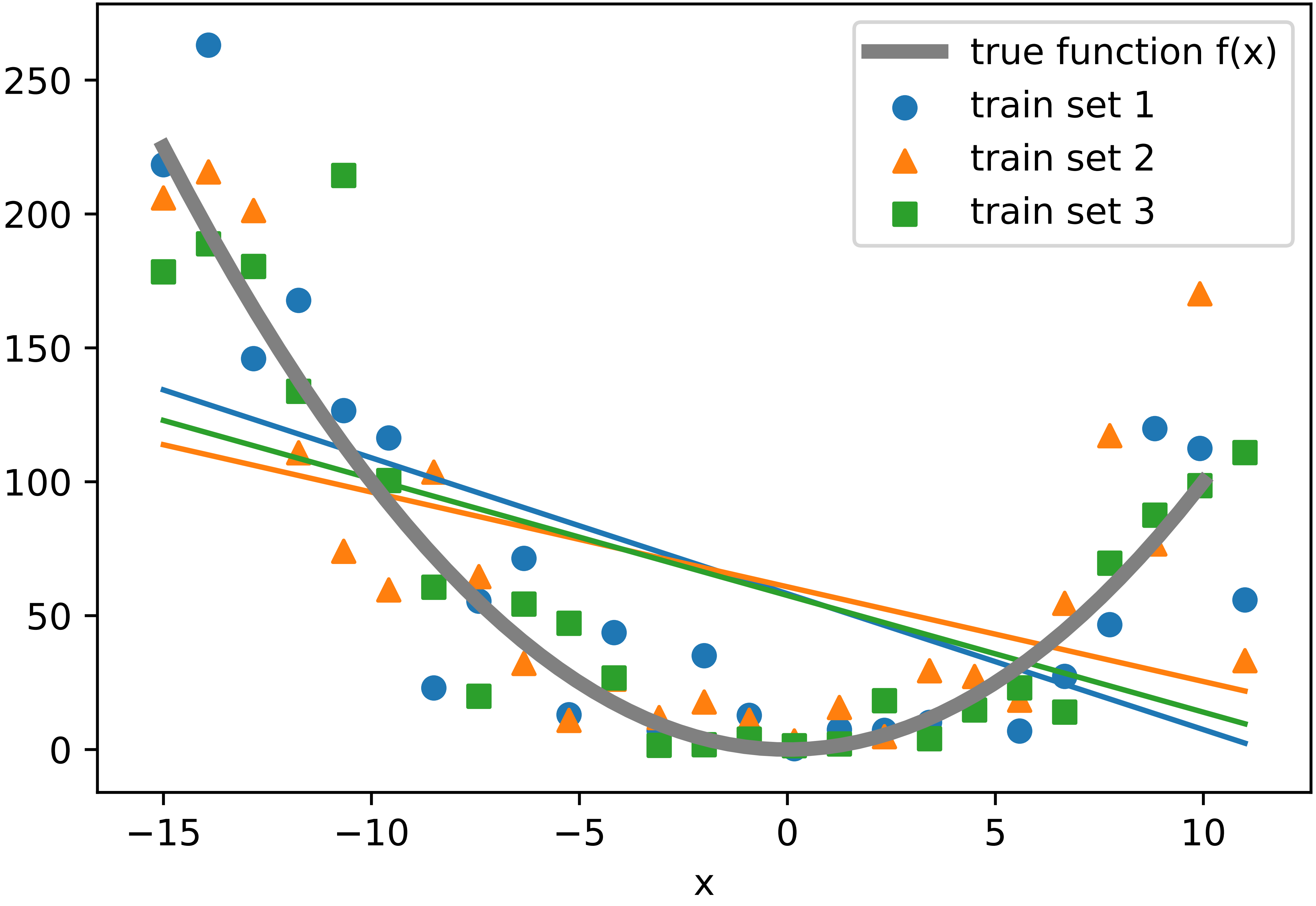


Bias-Variance Intuition

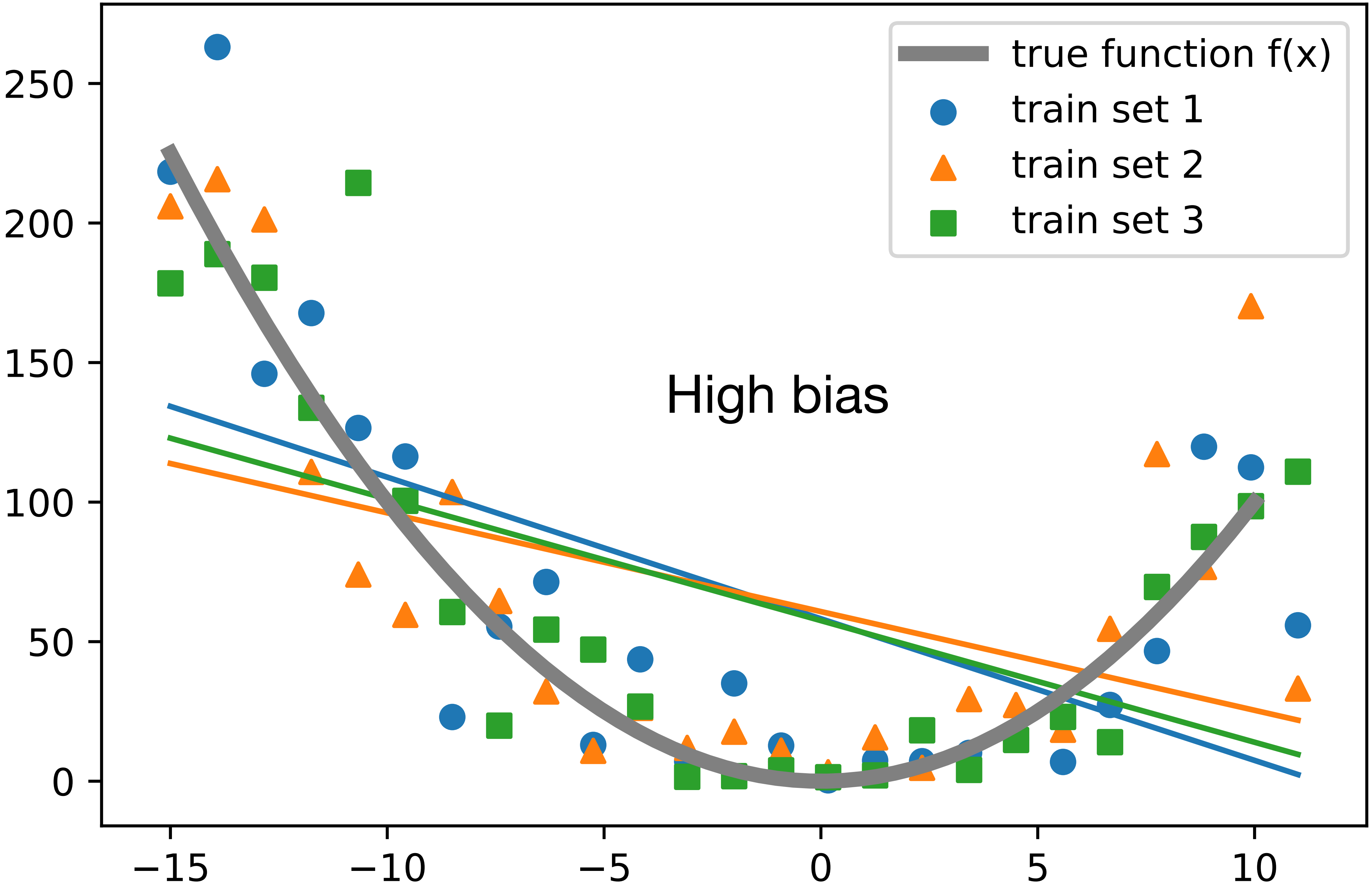
suppose we have multiple training sets



Bias-Variance Intuition



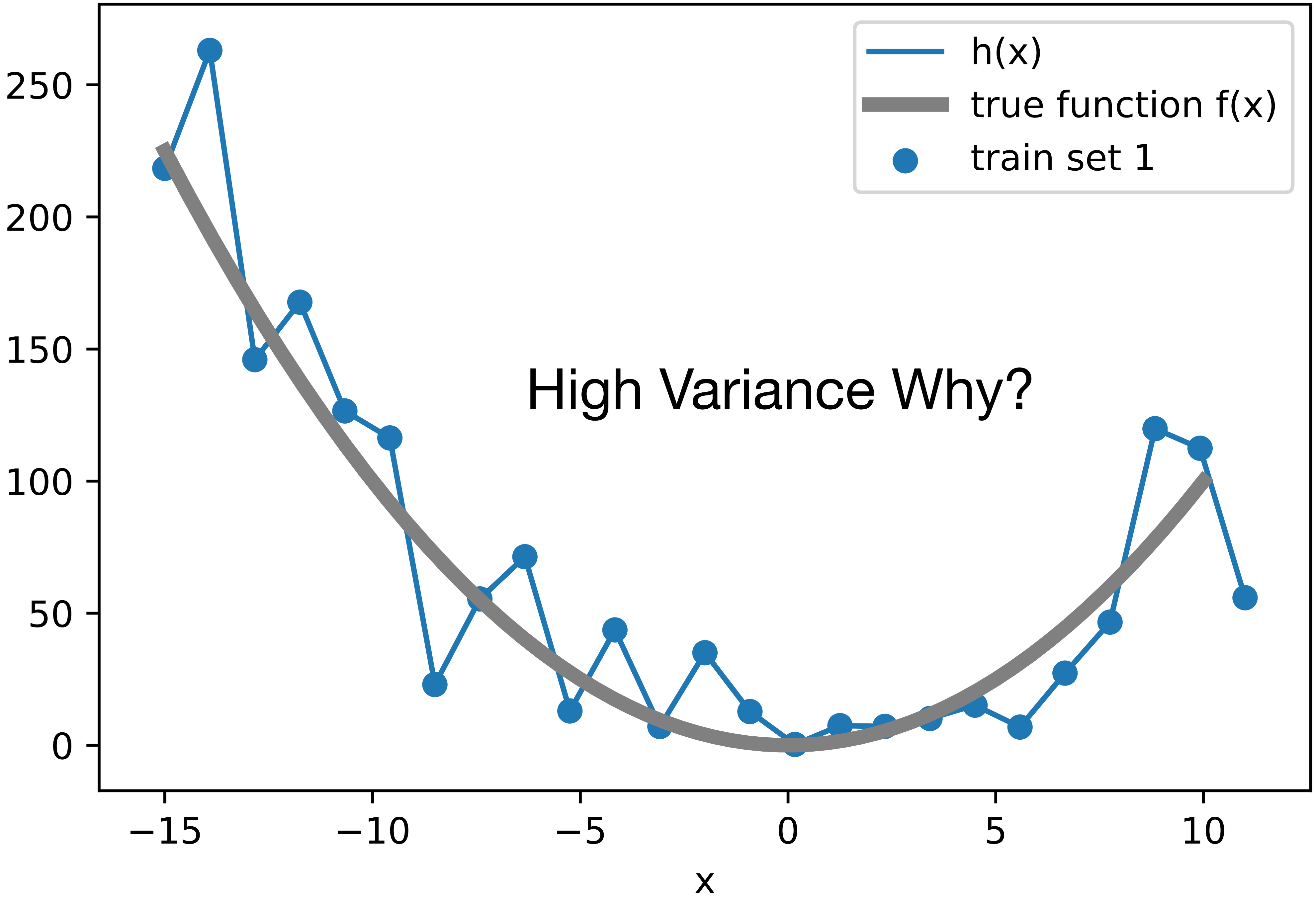
Bias-Variance Intuition



(There are points where the bias is zero ...)

Bias-Variance Intuition

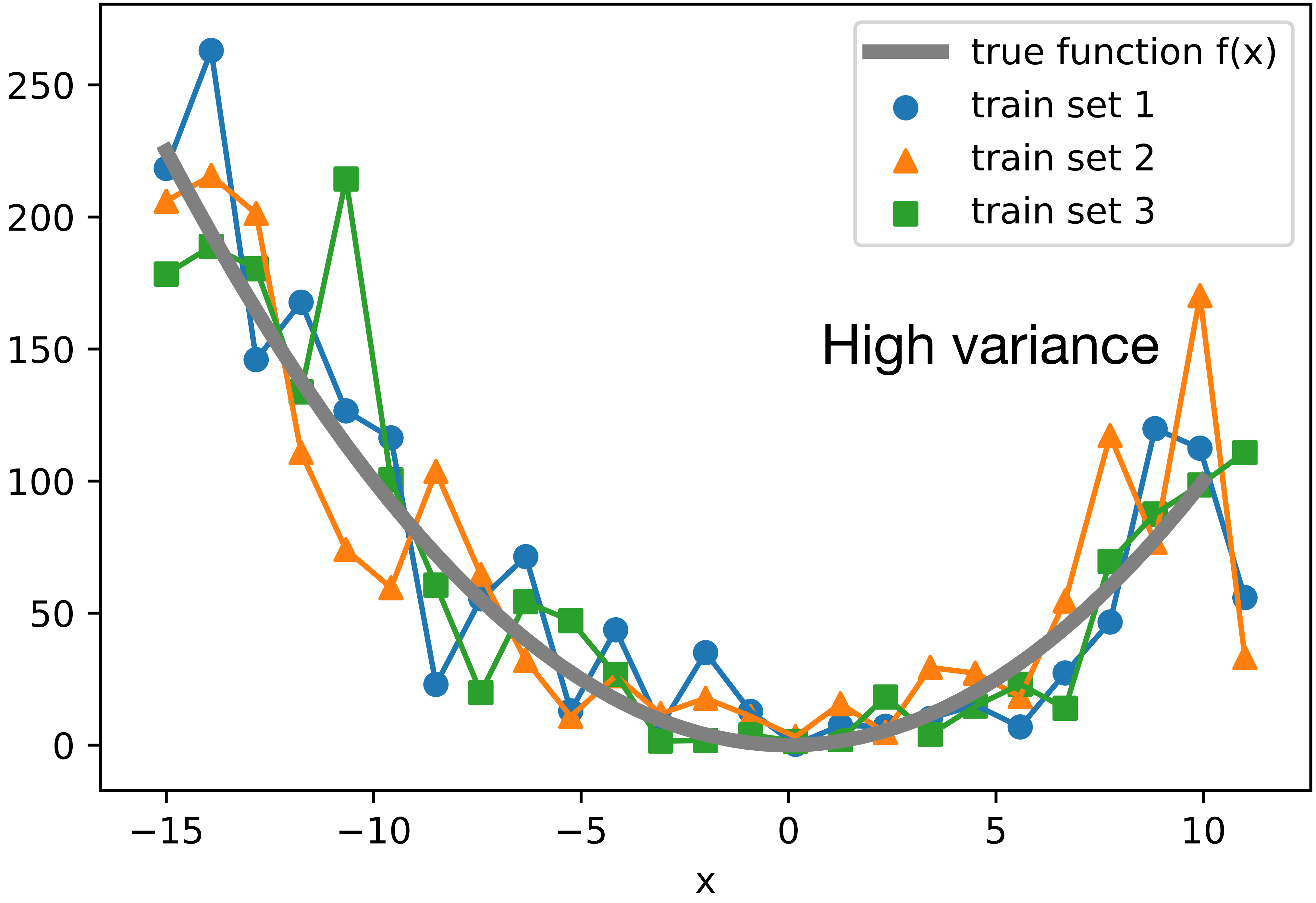
(here, I fit an unpruned decision tree)



High Variance Why?

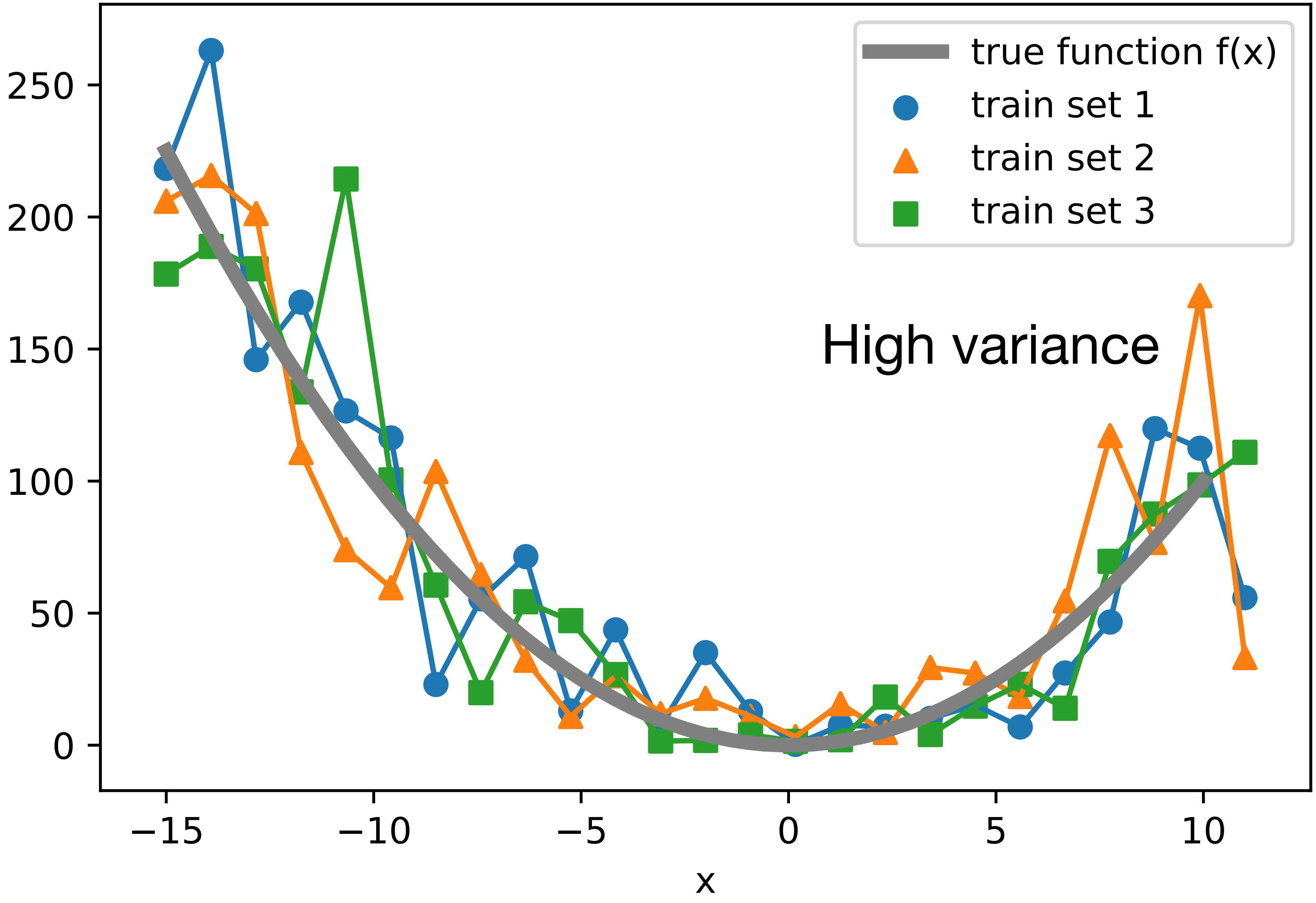
Bias-Variance Intuition

suppose we have multiple training sets



Bias-Variance Intuition

What happens if we take the average?
Does this remind you of something?



Terminology

Point estimator $\hat{\theta}$ of some parameter θ

(could also be a function, e.g., the hypothesis is an estimator of some target function)

Terminology

Point estimator $\hat{\theta}$ of some parameter θ

(could also be a function, e.g., the hypothesis is an estimator of some target function)

Bias $= E[\hat{\theta}] - \theta$

Terminology

General Definition

$$\text{Bias}[\hat{\theta}] = E[\hat{\theta}] - \theta$$

$$\text{Var}[\hat{\theta}] = E[\hat{\theta}^2] - (E[\hat{\theta}])^2$$

$$\text{Var}[\hat{\theta}] = E\left[(E[\hat{\theta}] - \hat{\theta})^2\right]$$

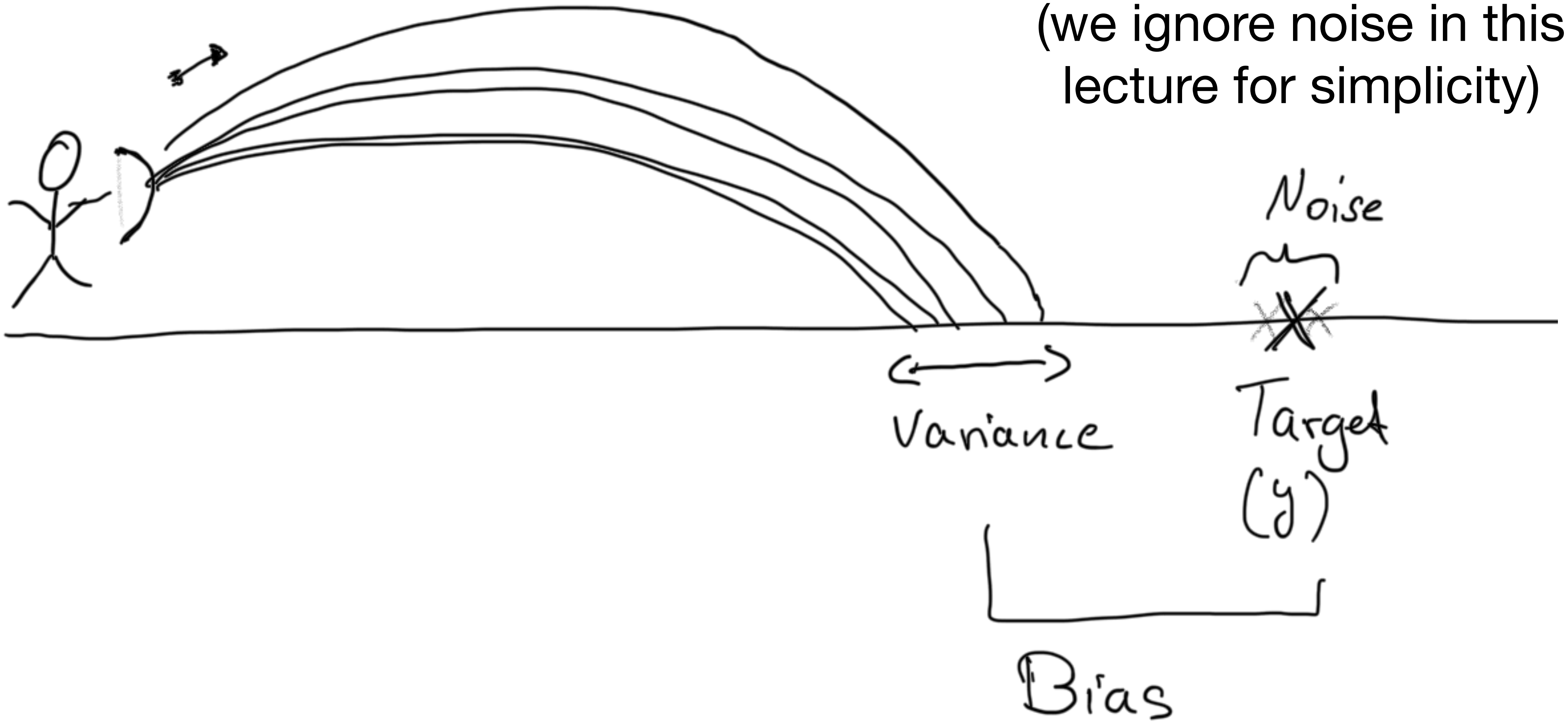
Terminology

Bias$[\hat{\theta}] = E[\hat{\theta}] - \theta$

Var$[\hat{\theta}] = E\left[(E[\hat{\theta}] - \hat{\theta})^2\right]$

Intuition

(we ignore noise in this lecture for simplicity)



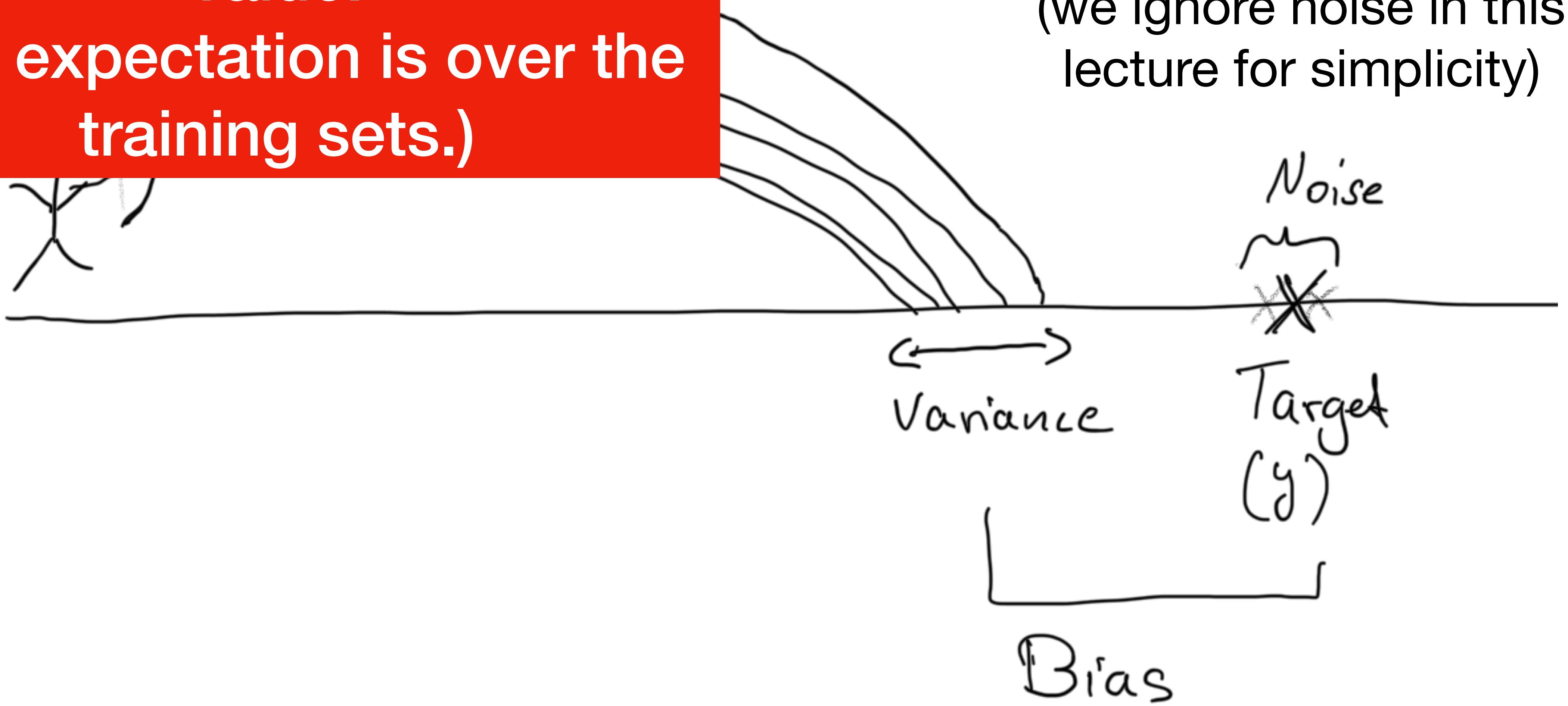
Terminology

$$\text{Bias}[\hat{\theta}] = E[\hat{\theta}] - \theta$$

$$\text{Var}[\hat{\theta}] = E\left[(E[\hat{\theta}] - \hat{\theta})^2\right]$$

Bias is the difference between the average estimator from different training samples and the true value.
(The expectation is over the training sets.)

(we ignore noise in this lecture for simplicity)



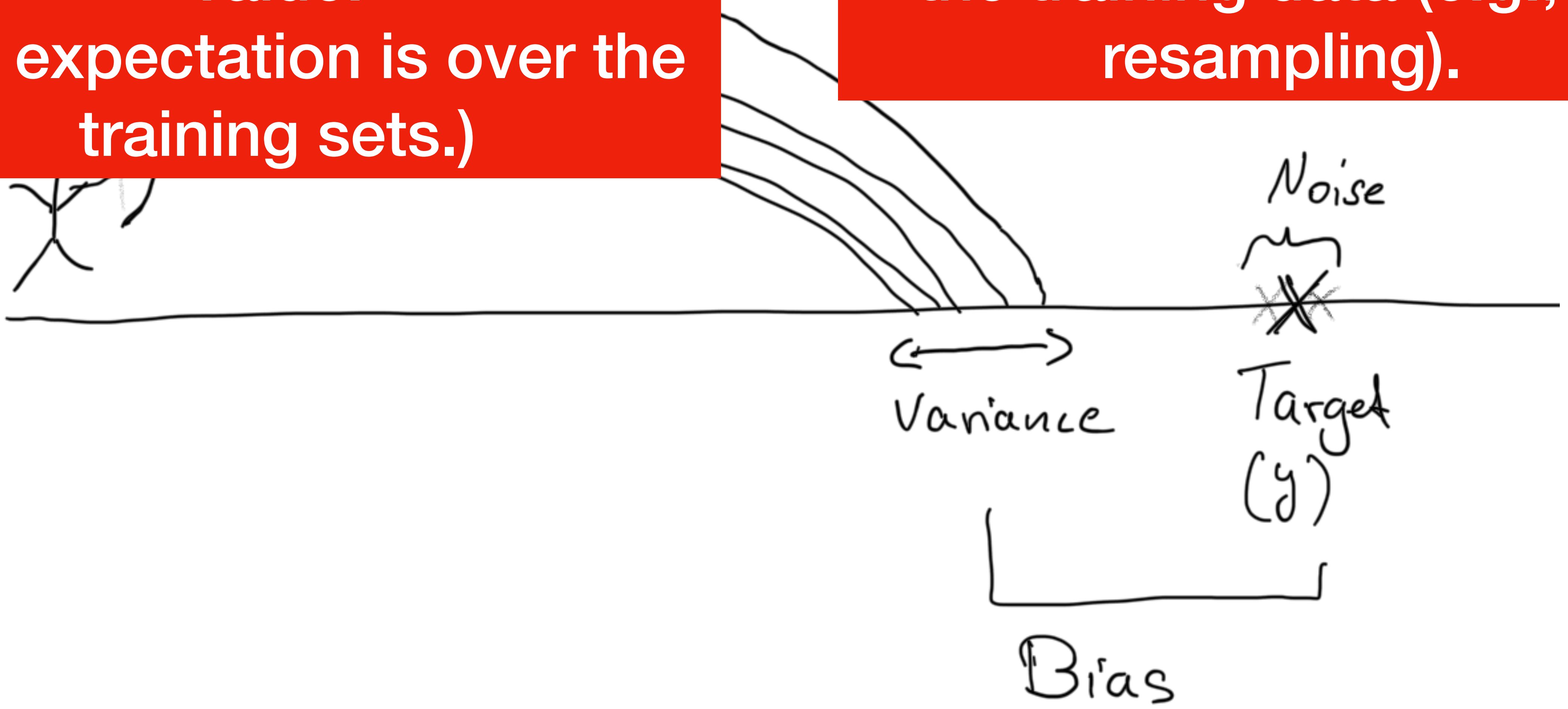
Terminology

$$\text{Bias}[\hat{\theta}] = E[\hat{\theta}] - \theta$$

Bias is the difference between the average estimator from different training samples and the true value.
(The expectation is over the training sets.)

$$\text{Var}[\hat{\theta}] = E\left[(E[\hat{\theta}] - \hat{\theta})^2\right]$$

The variance provides an estimate of how much the estimate varies as we vary the training data (e.g., by resampling).



Bias-Variance Decomposition

Loss = Bias + Variance + Noise

Bias-Variance of the Squared Error

$\text{Bias}[\hat{\theta}] = E[\hat{\theta}] - \theta$

$\text{Var}[\hat{\theta}] = E[\hat{\theta}^2] - (E[\hat{\theta}])^2$

$\text{Var}[\hat{\theta}] = E\left[(E[\hat{\theta}] - \hat{\theta})^2\right]$

"ML Notation" for Squared Error Loss

$y = f(x)$ target

$\hat{y} = \hat{f}(x) = h(x)$ prediction

$S = (y - \hat{y})^2$ squared error

for simplicity, we ignore the noise term

(Next slides: the expectation is over the training data, i.e, the average estimator from different training samples)

Bias-Variance of the Squared Error

"ML Notation" for Squared Error Loss

$y = f(x)$ target

$\hat{y} = \hat{f}(x) = h(x)$ prediction

$S = (y - \hat{y})^2$ squared error

$$S = (y - \hat{y})^2$$

$$\begin{aligned}(y - \hat{y})^2 &= (y - E[\hat{y}] + E[\hat{y}] - \hat{y})^2 \\ &= (y - E[\hat{y}])^2 + (E[\hat{y}] - \hat{y})^2 + 2(y - E[\hat{y}])(E[\hat{y}] - \hat{y})\end{aligned}$$

Bias-Variance of the Squared Error

$$S = (y - \hat{y})^2$$

$$(y - \hat{y})^2 = (y - E[\hat{y}] + E[\hat{y}] - \hat{y})^2$$
$$= (y - E[\hat{y}])^2 + (E[\hat{y}] - \hat{y})^2 + 2(y - E[\hat{y}])(E[\hat{y}] - \hat{y})$$

$$E[S] = E\left[(y - \hat{y})^2\right]$$
$$E\left[(y - \hat{y})^2\right] = (y - E[\hat{y}])^2 + E\left[(E[\hat{y}] - \hat{y})^2\right]$$

$$= \text{Bias}^2 + \text{Var}$$

Bias-Variance of the Squared Error

$$S = (y - \hat{y})^2$$

$$\begin{aligned}(y - \hat{y})^2 &= (y - E[\hat{y}] + E[\hat{y}] - \hat{y})^2 \\ &= (y - E[\hat{y}])^2 + (E[\hat{y}] - \hat{y})^2 \overset{???}{\boxed{+ 2(y - E[\hat{y}])(E[\hat{y}] - \hat{y})}}\end{aligned}$$

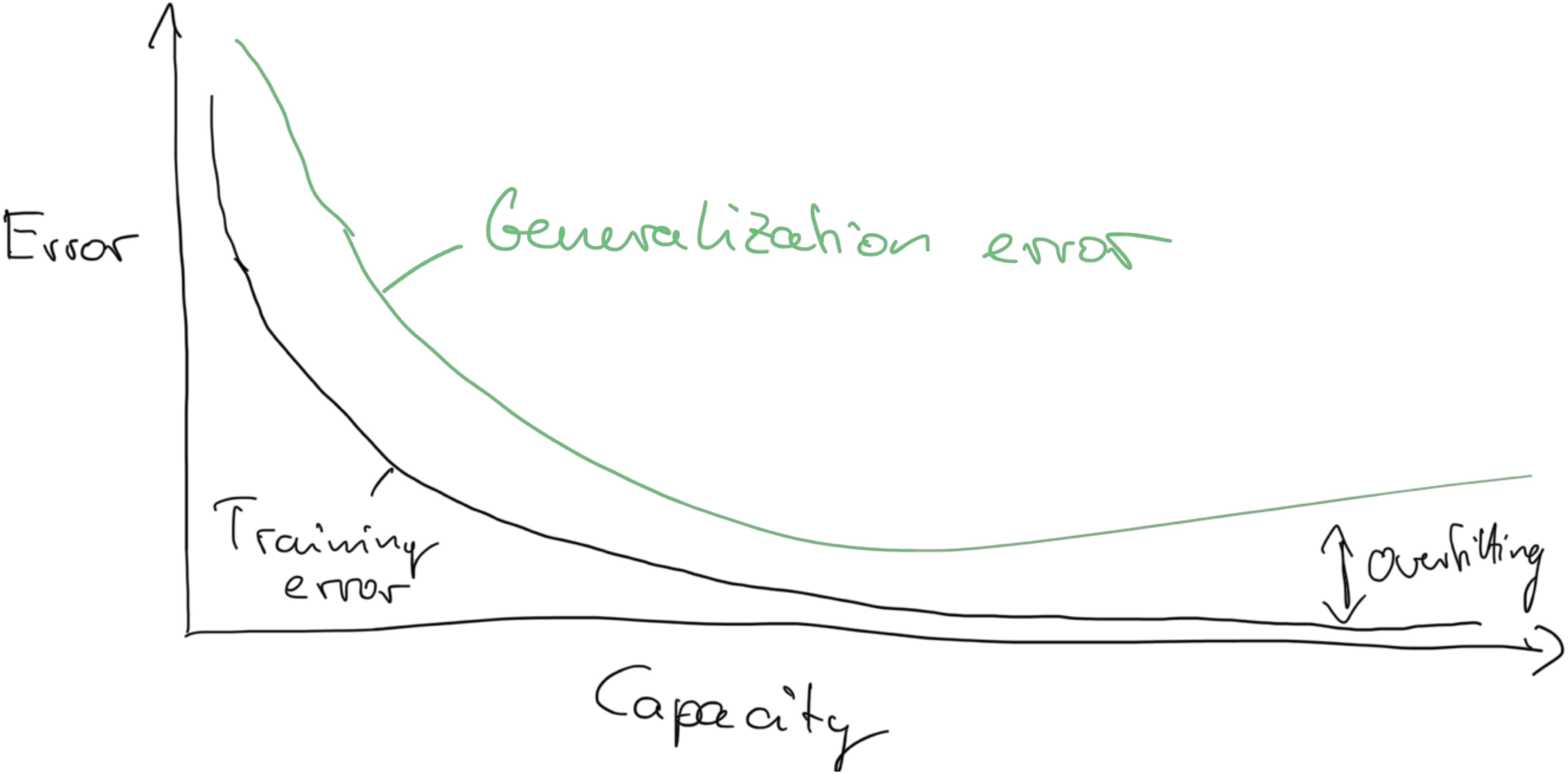
Bias-Variance of the Squared Error

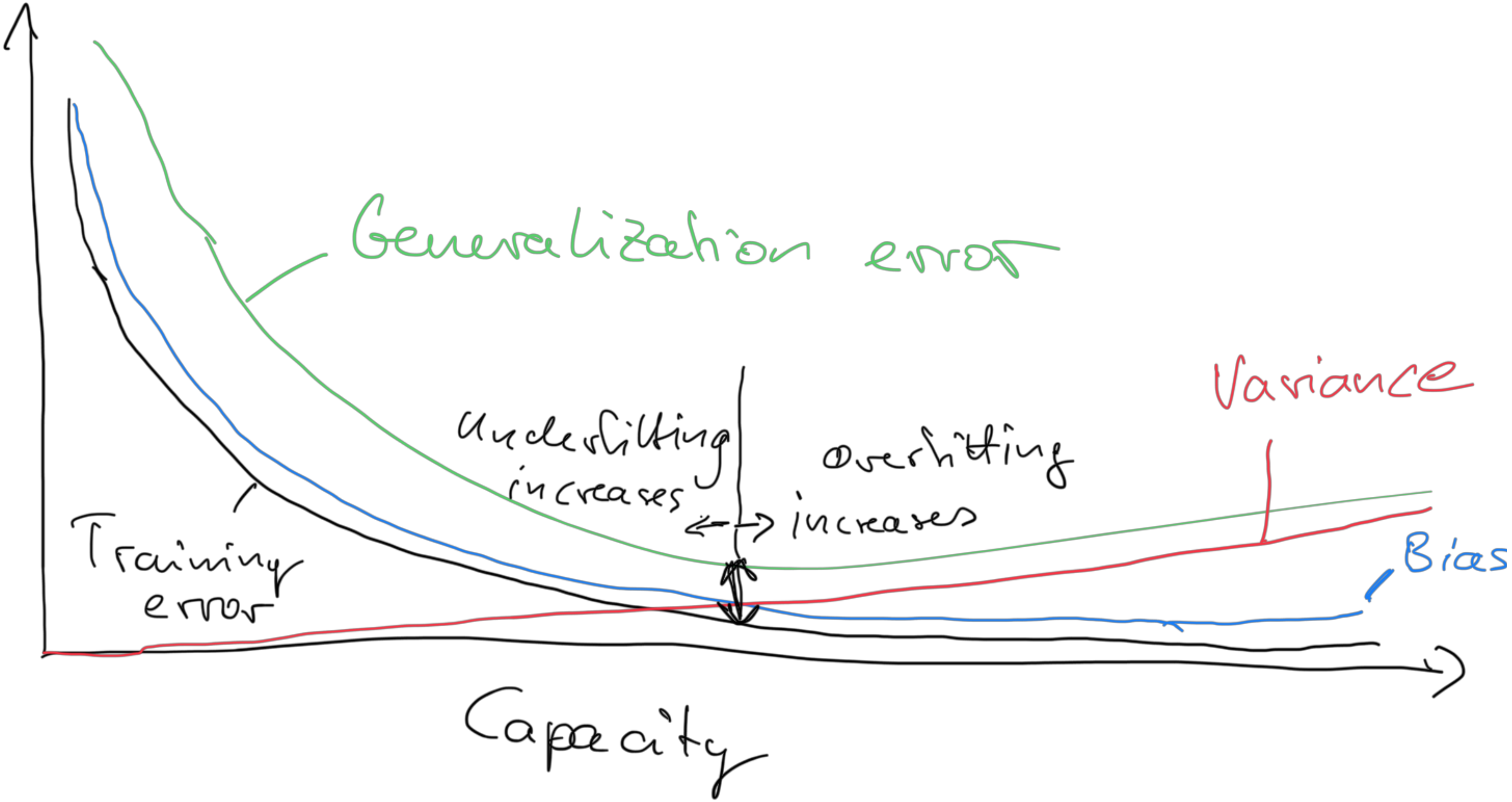
$$S = (y - \hat{y})^2$$

$$\begin{aligned}
(y - \hat{y})^2 &= (y - E[\hat{y}] + E[\hat{y}] - \hat{y})^2 \\
&= (y - E[\hat{y}])^2 + (E[\hat{y}] - \hat{y})^2 \overset{???}{\boxed{+ 2(y - E[\hat{y}])(E[\hat{y}] - \hat{y})}}
\end{aligned}$$

$$\begin{aligned}
E[2(y - E[\hat{y}])(E[\hat{y}] - \hat{y})] &= 2E[(y - E[\hat{y}])(E[\hat{y}] - \hat{y})] \\
&= 2(y - E[\hat{y}])E[(E[\hat{y}] - \hat{y})] \\
&= 2(y - E[\hat{y}])(E[E[\hat{y}]] - E[\hat{y}]) \\
&= 2(y - E[\hat{y}])(E[\hat{y}] - E[\hat{y}]) \\
&= 0
\end{aligned}$$

Now, how is this related to overfitting and underfitting?


Error
Generalization error
Training error
overfitting
Capacity


Generalization error
Variance
underfitting increases
overfitting increases
Bias
Training error
Capacity

How can we think of the bias-variance decomposition in the context of the classification error (0/1 loss)?

Domingos, P. (2000). A unified bias-variance decomposition.
In Proceedings of 17th International Conference on Machine Learning
(pp. 231-238).

"several authors have proposed bias-variance decompositions related to zero-one loss (Kong & Dietterich, 1995; Breiman, 1996b; Kohavi & Wolpert, 1996; Tibshirani, 1996; Friedman, 1997). However, each of these decompositions has significant shortcomings."

Domingos, P. (2000). A unified bias-variance decomposition.
In Proceedings of 17th International Conference on Machine Learning
(pp. 231-238).

"several authors have proposed bias-variance decompositions related to zero-one loss (Kong & Dietterich, 1995; Breiman, 1996b; Kohavi & Wolpert, 1996; Tibshirani, 1996; Friedman, 1997). However, each of these decompositions has significant shortcomings."

Detailed explanation of the 0/1 loss decomposition in the lecture notes

Other "Biases"

Statistical Bias vs "Machine Learning Bias"

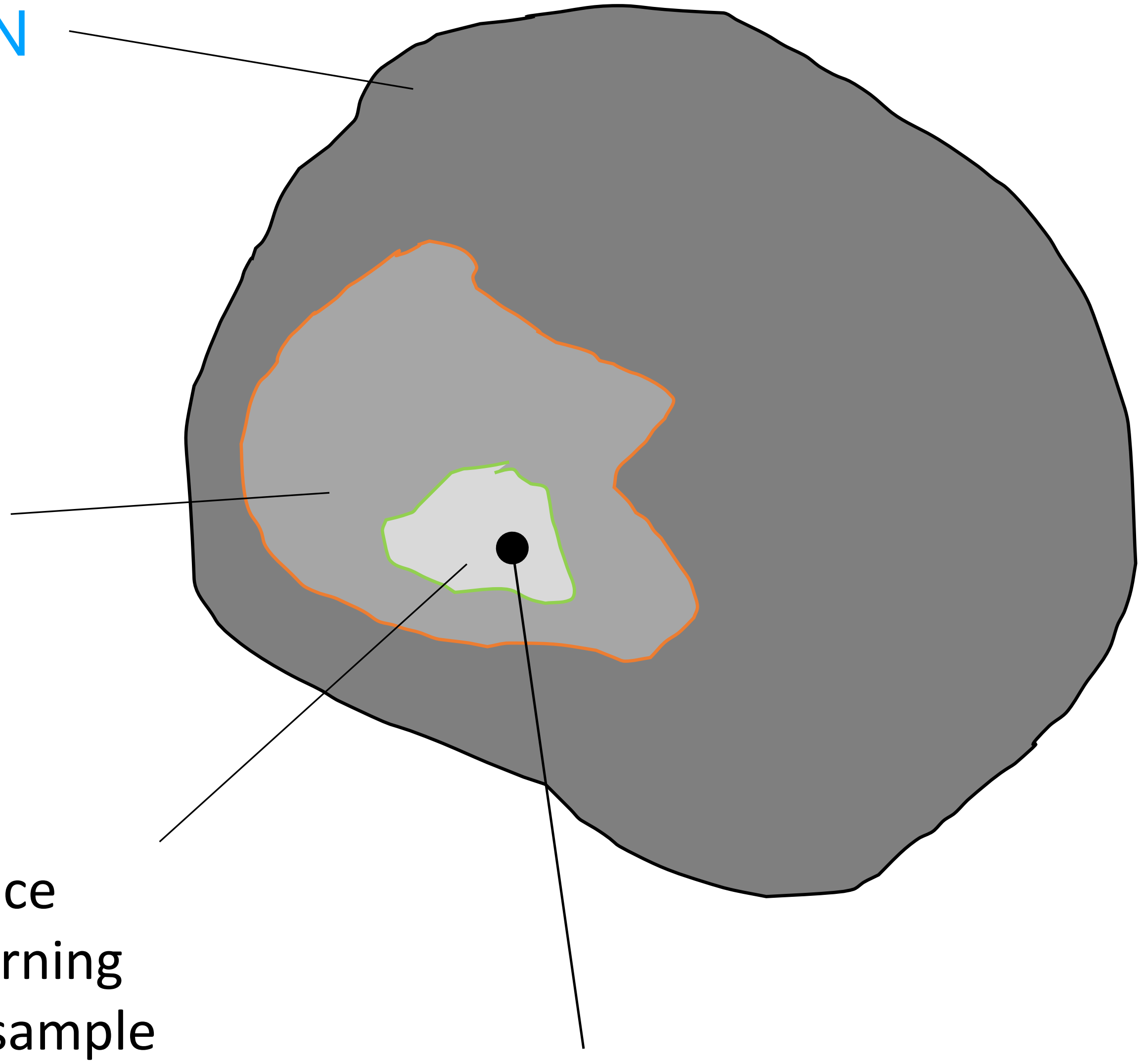
"Machine learning bias" sometimes also called "inductive bias"

e.g., decision tree algorithms consider small trees before they consider large trees

(if training data can be classified by small tree, large trees are not considered)

Hypothesis Space

(From Lecture 1)



Entire hypothesis space

e.g,. decision tree + KNN

Hypothesis space
a particular learning
algorithm category
has access to

e.g,. decision tree

Hypothesis space
a particular learning
algorithm can sample

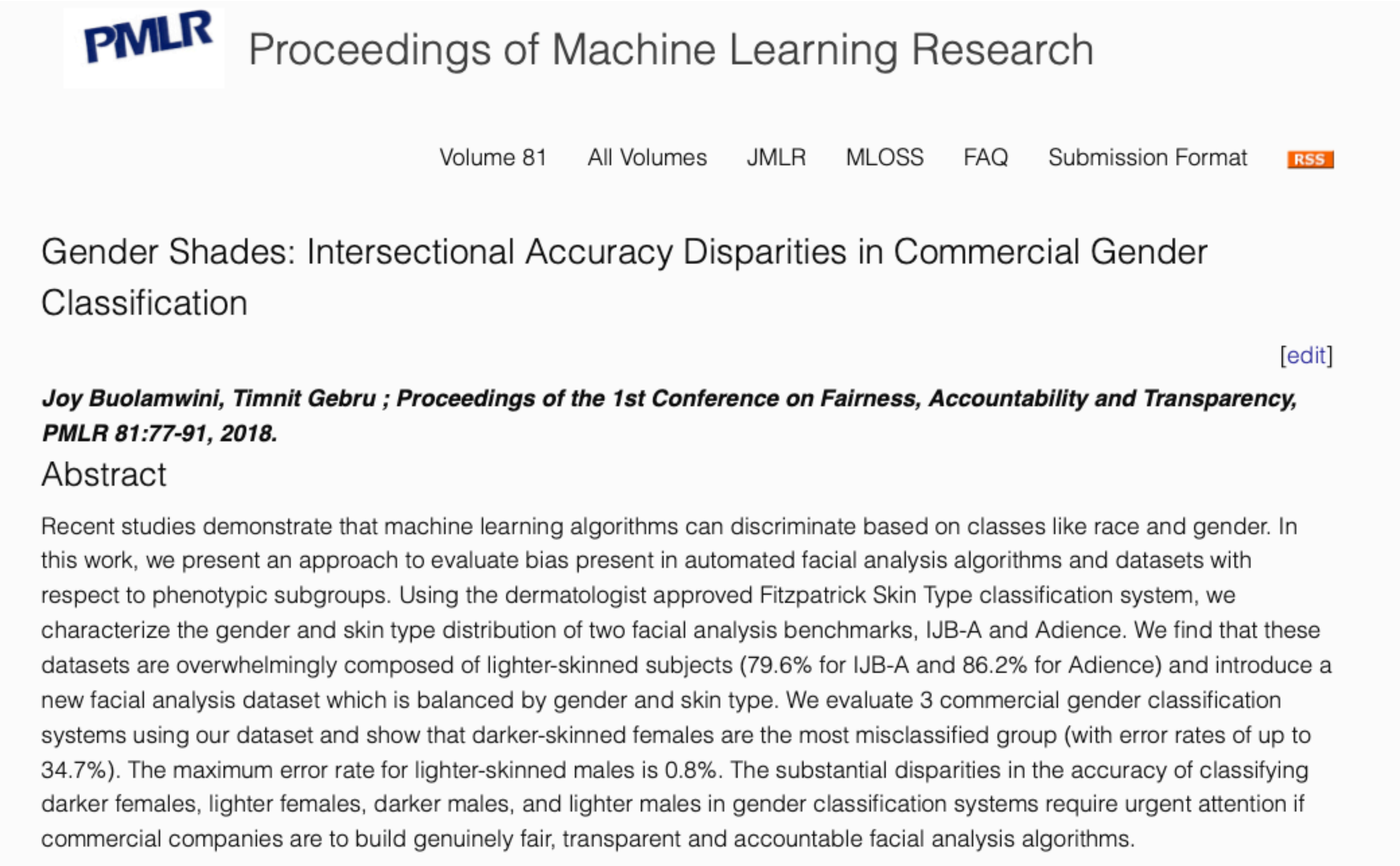
e.g,. ID3

Particular hypothesis
(i.e., a model/classifier)

"Fairness" Bias

"The term bias is often used to refer to demographic disparities in algorithmic systems that are objectionable for societal reasons. "

Barocas, S., Hardt, M., & Narayanan, A. Fairness and Machine Learning.
https://fairmlbook.org/introduction.html



Proceedings of Machine Learning Research

Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification

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.

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

Reading Assignments

Lecture notes

https://github.com/rasbt/stat479-machine-learning-fs19/tree/master/08_model-eval-1