

Course wrap-up

STAT 4710

December 8, 2022

Where we are

- ✓ **Unit 1:** Intro to modern data mining
- ✓ **Unit 2:** Tuning predictive models
- ✓ **Unit 3:** Regression-based methods
- ✓ **Unit 4:** Tree-based methods
- ✓ **Unit 5:** Deep learning

Where we are

- ✓ **Unit 1:** Intro to modern data mining
- ✓ **Unit 2:** Tuning predictive models
- ✓ **Unit 3:** Regression-based methods
- ✓ **Unit 4:** Tree-based methods
- ✓ **Unit 5:** Deep learning

Today's lecture:

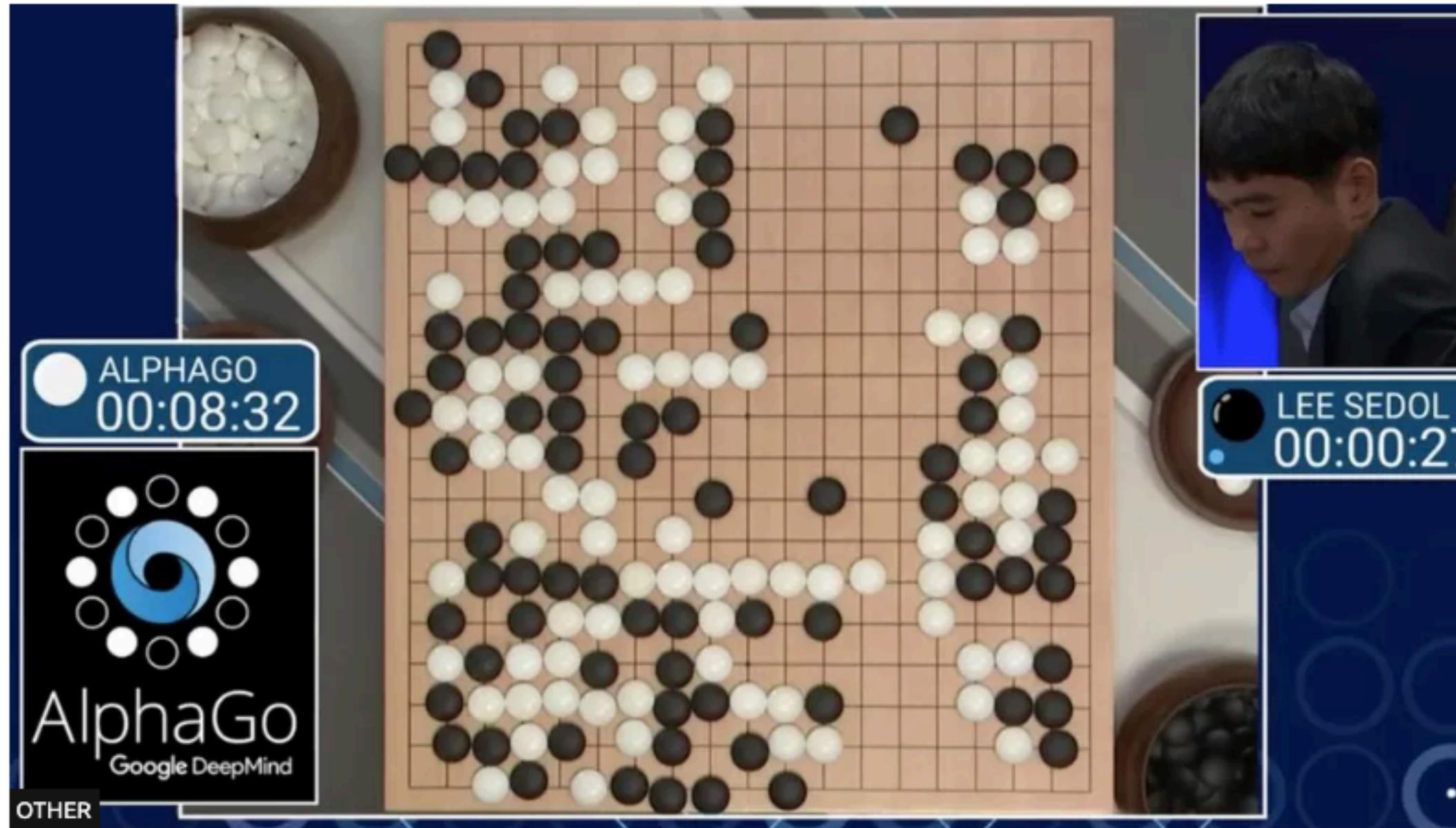
- Deep learning bonus material!
- Looking back at STAT 4710
- Looking beyond STAT 4710

Deep Learning Bonus Material!

AlphaGo (by DeepMind, 2016)

Artificial intelligence: Google's AlphaGo beats Go master Lee Se-dol

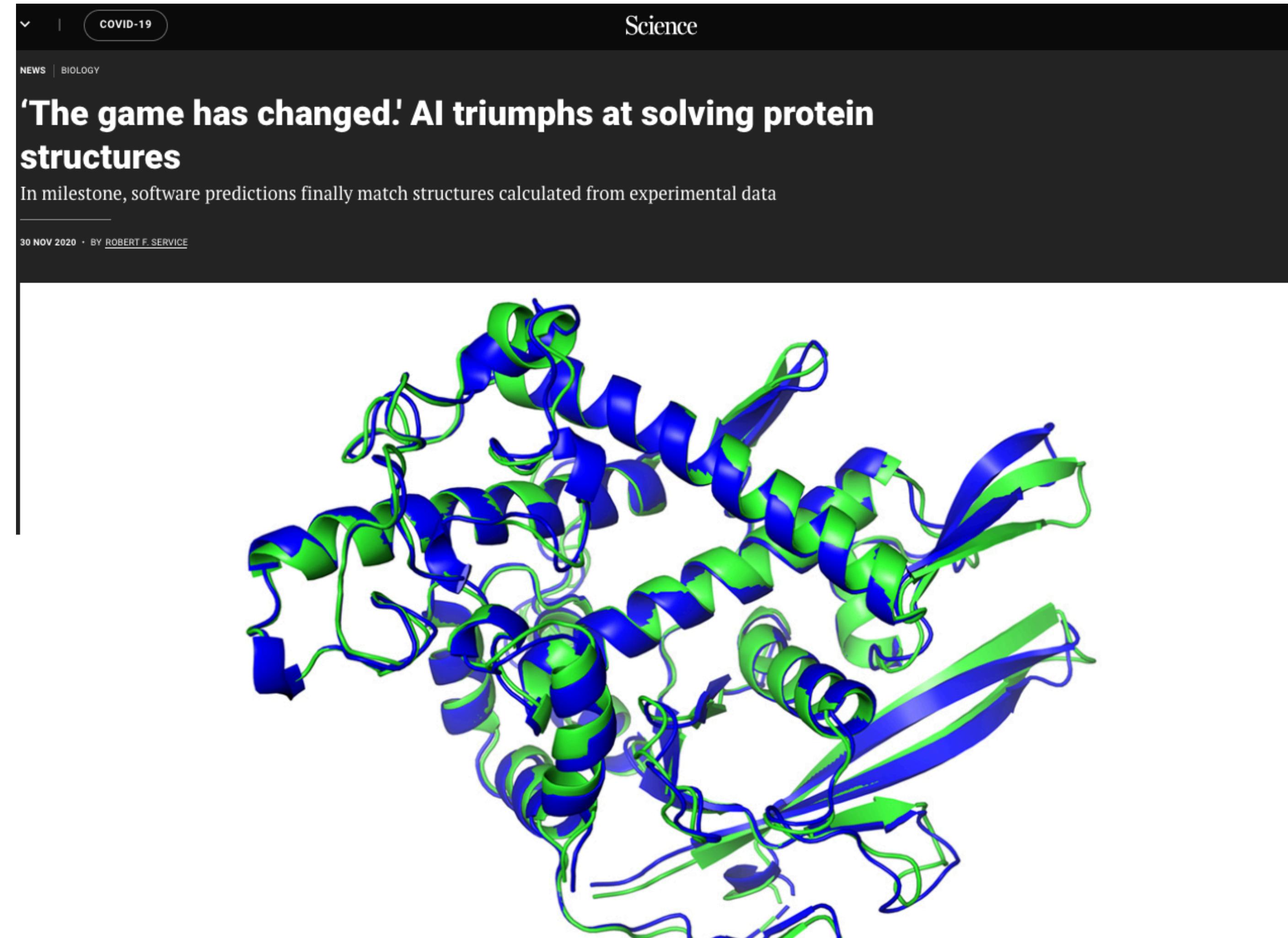
⌚ 12 March 2016



A computer program has beaten a master Go player 3-0 in a best-of-five competition, in what is seen as a landmark moment for artificial intelligence.

Learn more at <https://www.deepmind.com/research/highlighted-research/alphago>

AlphaFold (by DeepMind, 2020)



Learn more at <https://alphafold.ebi.ac.uk/>

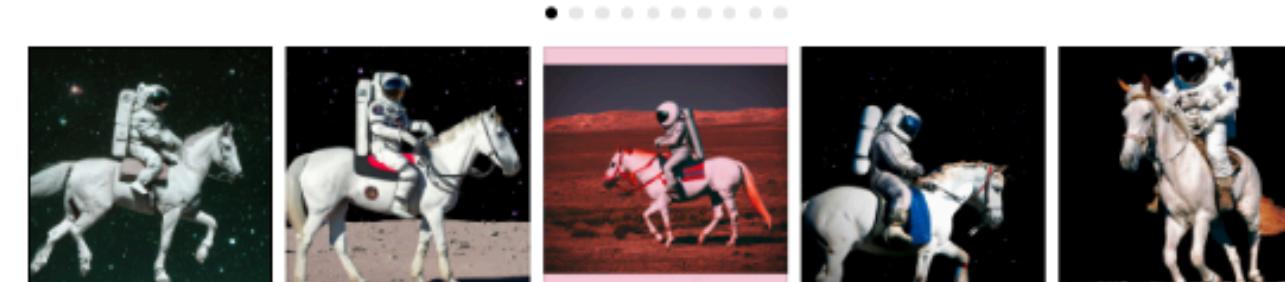
DALL·E 2 (by OpenAI, 2022)

DALL·E 2 can create original, realistic images and art from a text description. It can combine concepts, attributes, and styles.

TEXT DESCRIPTION

An astronaut Teddy bears A bowl of soup
riding a horse lounging in a tropical resort in space playing basketball with cats in space
in a photorealistic style in the style of Andy Warhol as a pencil drawing

DALL·E 2



Learn more at <https://openai.com/dall-e-2/>

ChatGPT (by OpenAI, 2022)

USER this code is not working like i expect — how do i fix it?

```
resultWorkerErr := make(chan error)
defer close(resultWorkerErr)
go func() {
    defer cancel()
    resultWorkerErr <- b.resultWorker(ctx)
}()

err := b.worker(ctx)
cancel()
if err == nil {
    return <-resultWorkerErr
}
return multierror.Append(err, <-resultWorkerErr)
```

CHATGPT It's difficult to say what's wrong with the code without more context. Can you provide more information about what the code is supposed to do and what isn't working as expected? Also, is this the entire code or just a part of it?

USER just part of it — the error never surfaces. I think there is something wrong with the channel

CHATGPT It's difficult to say without more information about what the code is supposed to do and what's happening when it's executed. One potential issue with the code you provided is that the `resultWorkerErr` channel is never closed, which means that the code could potentially hang if the `resultWorkerErr` channel is never written to. This could happen if `b.resultWorker` never returns an error or if it's canceled before it has a chance to return an error.

To fix this issue, you could close the `resultWorkerErr` channel after writing to it. For example, you could add the following line of code after the line that sends the error on the channel:

```
close(resultworkerErr)
```

Learn more at <https://openai.com/blog/chatgpt/>

ChatGPT (by OpenAI, 2022)

USER this code is not working like i expect — how do i fix it?

```
resultWorkerErr := make(chan error)
defer close(resultWorkerErr)
go func() {
    defer cancel()
    resultWorkerErr <- b.resultWorker(ctx)
}()

err := b.worker(ctx)
cancel()
if err == nil {
    return <-resultWorkerErr
}
return multierror.Append(err, <-resultWorkerErr)
```

CHATGPT It's difficult to say what's wrong with the code without more context. Can you provide more information about what the code is supposed to do and what isn't working as expected? Also, is this the entire code or just a part of it?

USER just part of it — the error never surfaces. I think there is something wrong with the channel

CHATGPT It's difficult to say without more information about what the code is supposed to do and what's happening when it's executed. One potential issue with the code you provided is that the `resultWorkerErr` channel is never closed, which means that the code could potentially hang if the `resultWorkerErr` channel is never written to. This could happen if `b.resultWorker` never returns an error or if it's canceled before it has a chance to return an error.

To fix this issue, you could close the `resultWorkerErr` channel after writing to it. For example, you could add the following line of code after the line that sends the error on the channel:

```
close(resultworkerErr)
```

USER help me write a short note to introduce myself to my neighbor

CHATGPT Hi there,

My name is [Your Name] and I just moved in next door. I wanted to reach out and introduce myself. I'm looking forward to getting to know you and the neighborhood. If you ever need anything or just want to chat, please don't hesitate to stop by or give me a call.

Best,
[Your Name]

USER can you make it more formal?

CHATGPT Dear Neighbor,

I am writing to introduce myself. My name is [Your Name] and I have recently moved in next door. I wanted to take a moment to reach out and say hello. I am looking forward to getting to know you and the neighborhood.

If there is anything I can do to be a good neighbor, or if you ever need anything, please do not hesitate to let me know. I am happy to help in any way that I can.

Sincerely,
[Your Name]

USER can you send it for me?

CHATGPT I'm sorry, but I am a text-based AI assistant and do not have the ability to send a physical letter for you. I can only provide suggestions and advice on what to write. It is up to you to take that advice and send the letter yourself.

Learn more at <https://openai.com/blog/chatgpt/>

ChatGPT (by OpenAI, 2022)

The image displays three separate screenshots of the ChatGPT web interface, each showing a different conversation or interaction.

Screenshot 1 (Left): A user asks about code that is not working as expected. The AI provides a detailed explanation of the code's behavior and suggests a fix. It also includes a snippet of Go code and a timestamp.

```
resultWorkerErr := make(chan error)
defer close(resultWorkerErr)
go func() {
    defer cancel()
    resultWorkerErr <- b.resultWorker(ctx)
}()

err := b.worker(ctx)
cancel()
if err == nil {
    return <-resultWorkerErr
}
return multierror.Append(err, <-resultWorkerErr)
```

CHATGPT: It's difficult to say what's wrong with the code without more information about what the code is supposed to do and what you expected? Also, is this the entire code or just a part of it — the error never surfaces. I think the channel is closed.

USER: just part of it — the error never surfaces. I think the channel is closed.

CHATGPT: It's difficult to say without more information about what the code is supposed to do and what's happening when it's executed. One possibility provided is that the `resultWorkerErr` channel is closed, which means the code could potentially hang if the `resultWorker` function never returns an error. This could happen if `b.resultWorker` never returns an error or if there's a chance to return an error.

To fix this issue, you could close the `resultWorkerErr` channel. For example, you could add the following line of code before closing the channel:

```
close(resultworkerErr)
```

5:33 AM · Dec 1, 2022

Screenshot 2 (Middle): A user asks for help writing a short note to introduce themselves to a neighbor. The AI responds with a creative, rhyming response about quantum theory.

USER: help me write a short note to introduce myself to my neighbor

juan @juanbuis

having a particularly bizarre morning thanks to chatgpt

explain quantum theory to a child in the style of snoop dogg

Yo, little homie, let me break it down for you
Quantum theory, it's a crazy thing, it's true
See, the world is made up of tiny, tiny particles
That are always moving, and changing, and actin' wild

These particles, they don't always behave like we think
Sometimes they're waves, and sometimes they're points, it's a trip
They can be in two places at once, and they can be in two states
It's a mind-blowing concept, but that's the quantum fate

So don't be scared, little homie, it's all good
Quantum theory, it's just the way the world works, it's understood
It's a wild ride, but it's a cool one too
Quantum theory, it's the way the world does what it do.

Screenshot 3 (Right): A user asks for help writing a short note to introduce themselves to a neighbor. The AI responds with a template message.

...
I just moved in next door. I wanted to reach out and introduce myself and get to know you and the neighborhood. If you want to chat, please don't hesitate to stop by or give me a call.
My name is [Your Name] and I have recently moved into the neighborhood. I wanted to reach out and say hello. I am looking for a good neighbor, or if you ever need anything, I am here to help. I am happy to help in any way that I can.

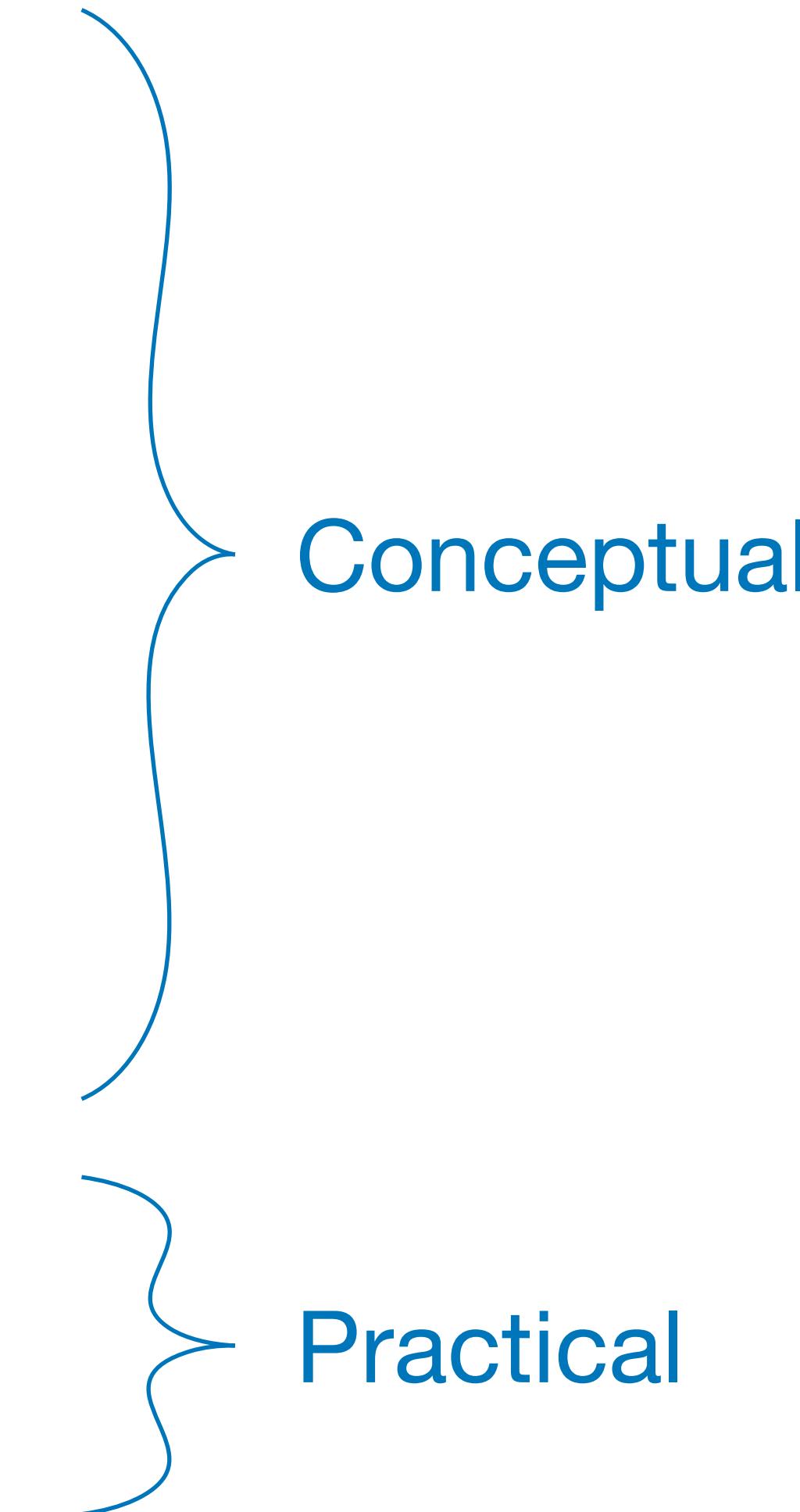
AI assistant and do not have the ability to send a message. You can provide suggestions and advice on what to write. It is up to you to take that advice and send the letter yourself.

Learn more at <https://openai.com/blog/chatgpt/>

Looking back at STAT 4710

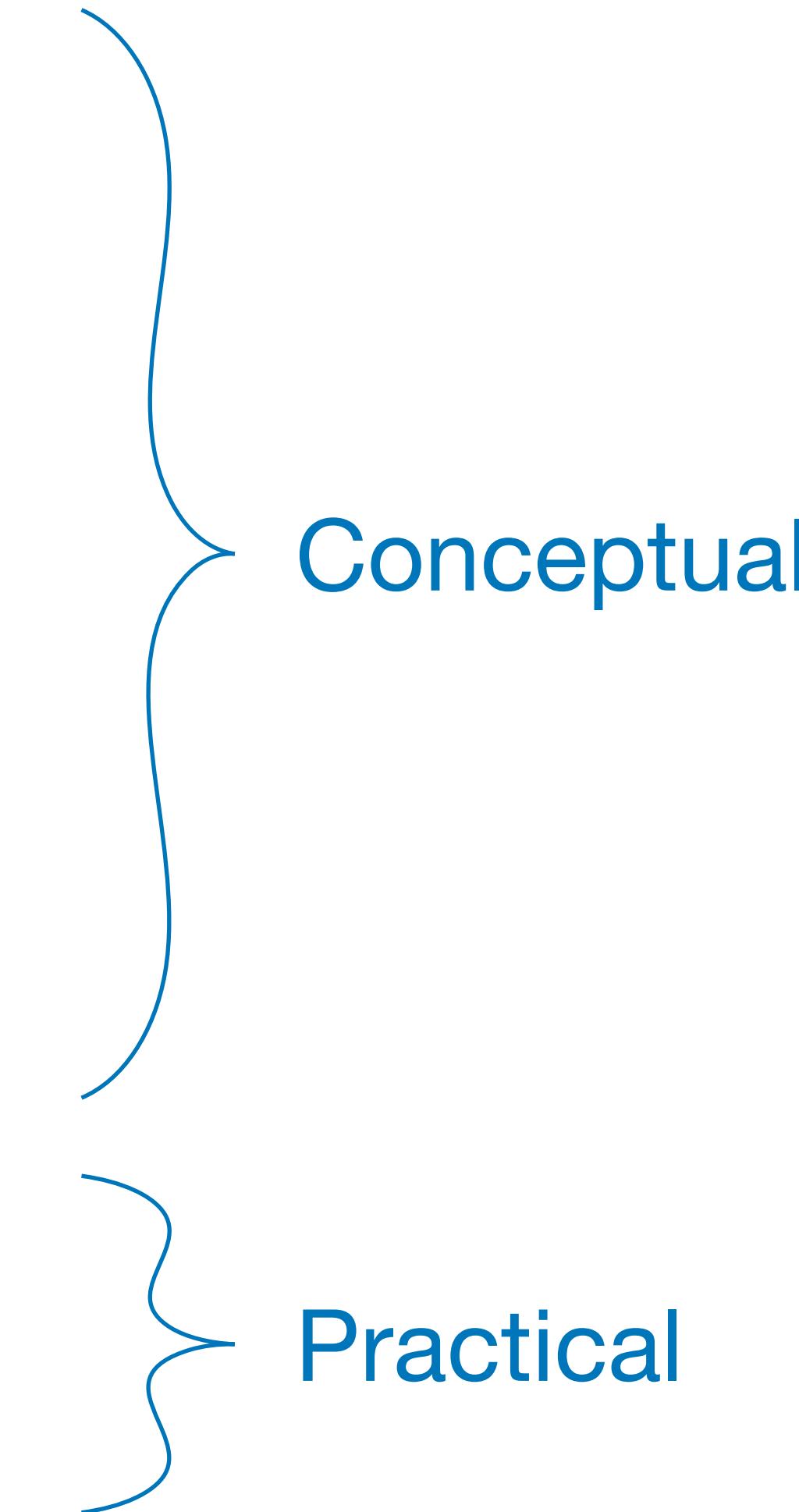
Themes of the class, and a lingering question

1. Regression and classification
2. Training predictive models
3. Model complexity
4. Bias-variance trade-off
5. Model selection and model assessment
6. Interpretability of predictive models
7. R programming tools
8. Working with data



Themes of the class, and a lingering question

1. Regression and classification
2. Training predictive models
3. Model complexity
4. Bias-variance trade-off
5. Model selection and model assessment
6. Interpretability of predictive models
7. R programming tools
8. Working with data



Lingering question: What is the best prediction method?

Theme: Regression and classification

Theme: Regression and classification

Prediction methods vary based on the **response type**:

- Regression: continuous responses
- Classification: discrete responses (binary or multi-class)

Theme: Regression and classification

Prediction methods vary based on the **response type**:

- Regression: continuous responses
- Classification: discrete responses (binary or multi-class)

Most methods have versions for regression and classification, e.g. linear regression and logistic regression.

Theme: Regression and classification

Prediction methods vary based on the **response type**:

- Regression: continuous responses
- Classification: discrete responses (binary or multi-class)

Most methods have versions for regression and classification, e.g. linear regression and logistic regression.

Classification methods are indirect in the sense that they predict probabilities of each class. They are also a little more fussy; need to make sure probabilities are between 0 and 1, class imbalance, misclassification error versus Gini index,...

Theme: Regression and classification

Prediction methods vary based on the **response type**:

- Regression: continuous responses
- Classification: discrete responses (binary or multi-class)

Most methods have versions for regression and classification, e.g. linear regression and logistic regression.

Classification methods are indirect in the sense that they predict probabilities of each class. They are also a little more fussy; need to make sure probabilities are between 0 and 1, class imbalance, misclassification error versus Gini index,...

Many of the same intuitions apply for regression and classification.

Theme: Training predictive models

Define **class of predictive models** $f_\beta(X)$ indexed by some parameter vector β .

Find member of this class that best fits the training data, as measured by the **loss function** L of predictions given true responses, possibly regularized:

$$\hat{\beta} = \arg \min_{\beta} \frac{1}{n} \sum_{i=1}^n L(Y_i, f_\beta(X_i)) + \lambda \cdot \text{penalty}(\beta).$$

Theme: Training predictive models

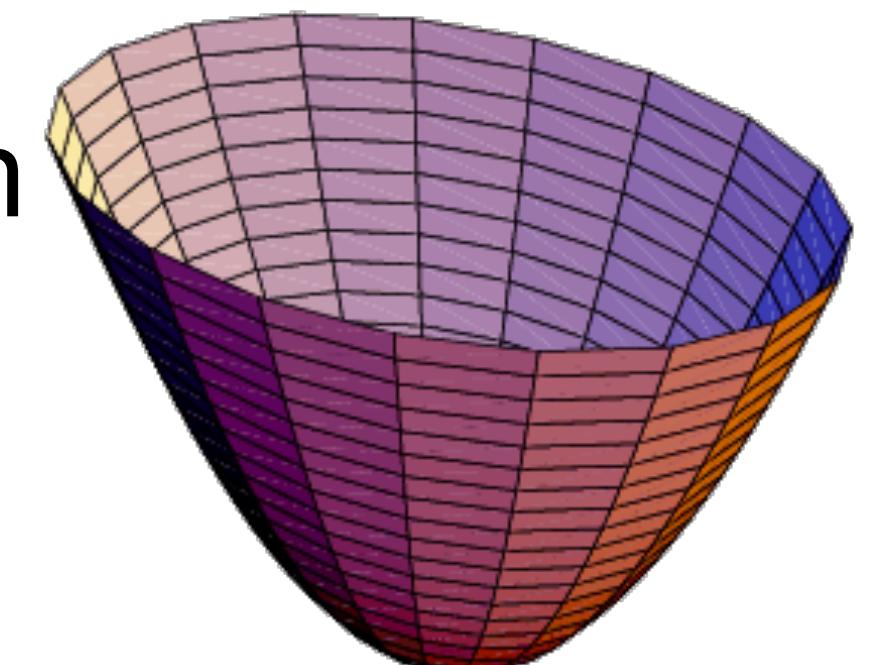
Define **class of predictive models** $f_{\beta}(X)$ indexed by some parameter vector β .

Find member of this class that best fits the training data, as measured by the **loss function** L of predictions given true responses, possibly regularized:

$$\hat{\beta} = \arg \min_{\beta} \frac{1}{n} \sum_{i=1}^n L(Y_i, f_{\beta}(X_i)) + \lambda \cdot \text{penalty}(\beta).$$

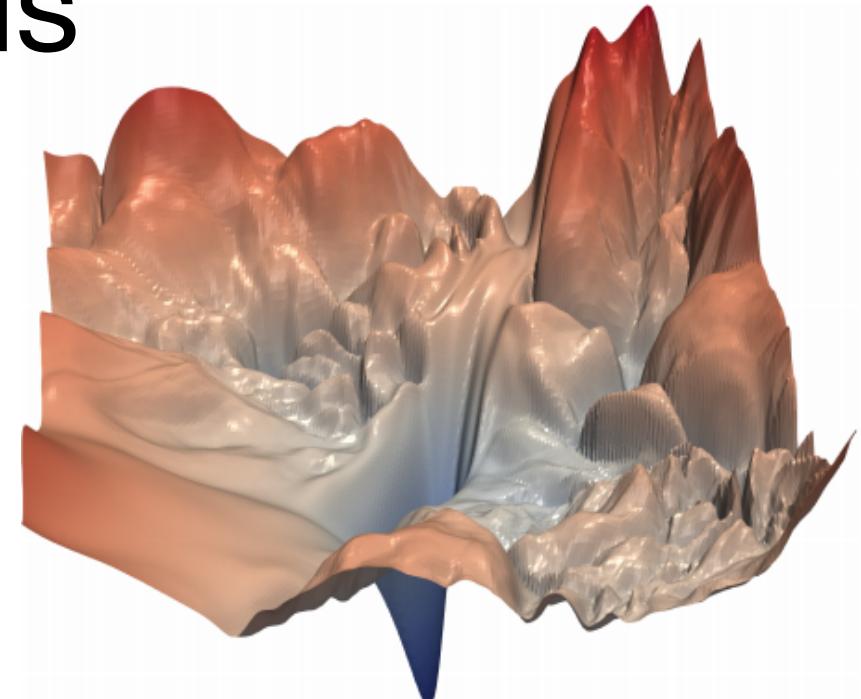
Convex (optimization is easy)

- Linear and logistic regression
- Linear and logistic regression with ridge or lasso penalties



Not convex (optimization is hard)

- Tree-based methods
- Neural networks



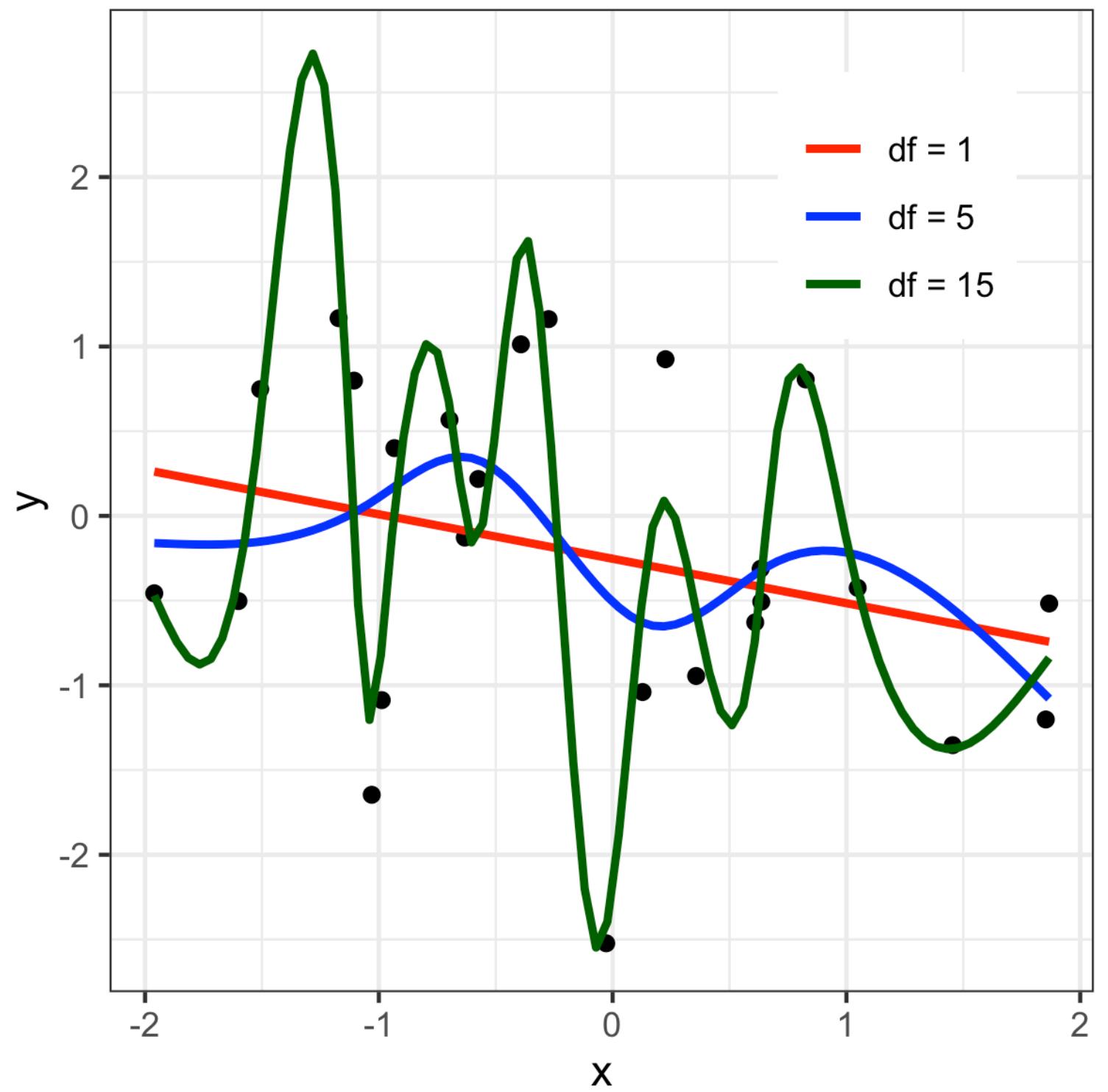
Theme: Model complexity

Theme: Model complexity

One of the key characteristics of a predictive model is its complexity: how flexibly does it fit the training data?

Theme: Model complexity

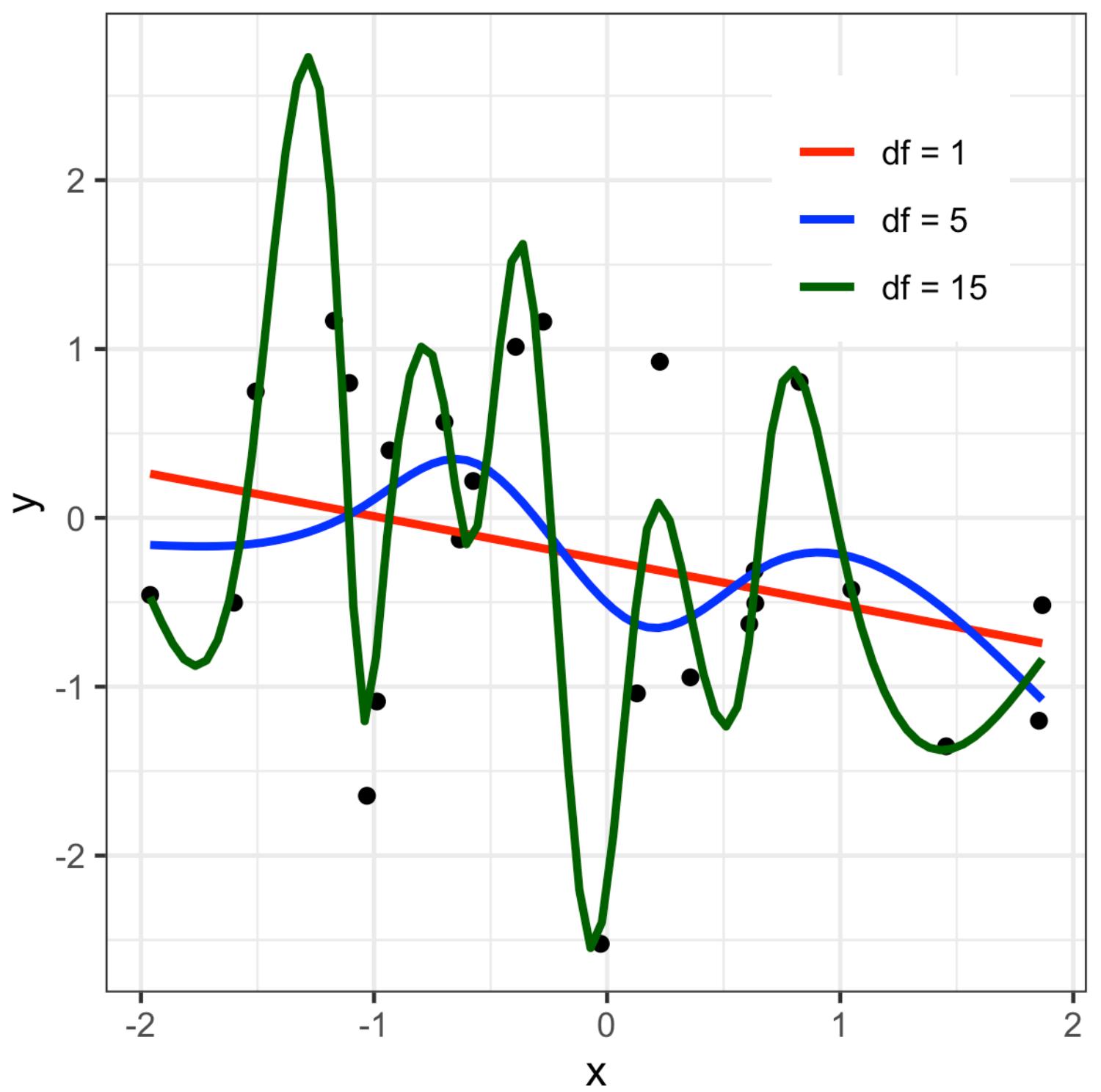
One of the key characteristics of a predictive model is its complexity: how flexibly does it fit the training data?



Theme: Model complexity

One of the key characteristics of a predictive model is its complexity: how flexibly does it fit the training data?

How is model complexity defined? Depends on model:

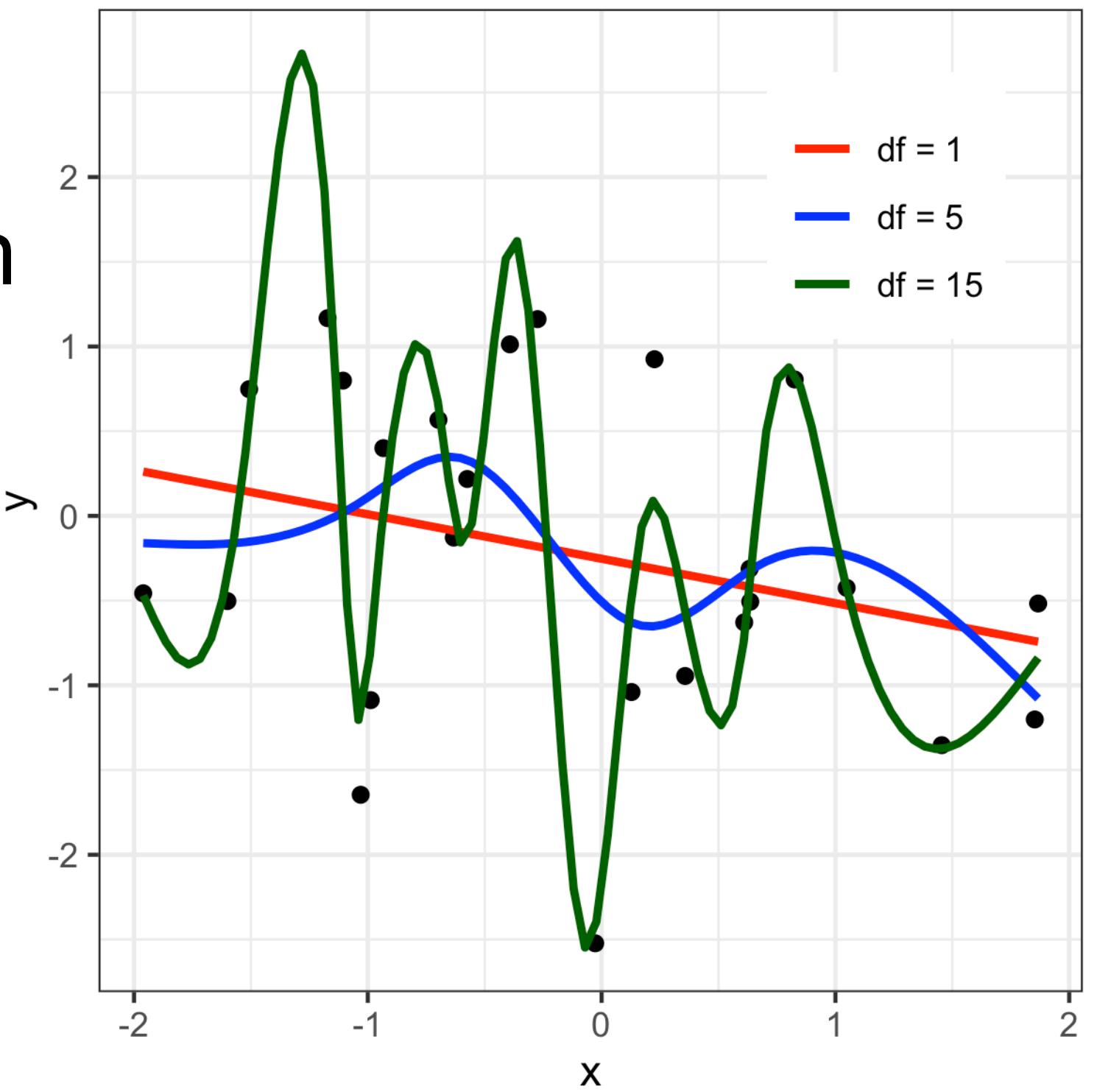


Theme: Model complexity

One of the key characteristics of a predictive model is its complexity: how flexibly does it fit the training data?

How is model complexity defined? Depends on model:

- Number of parameters in linear or logistic regression

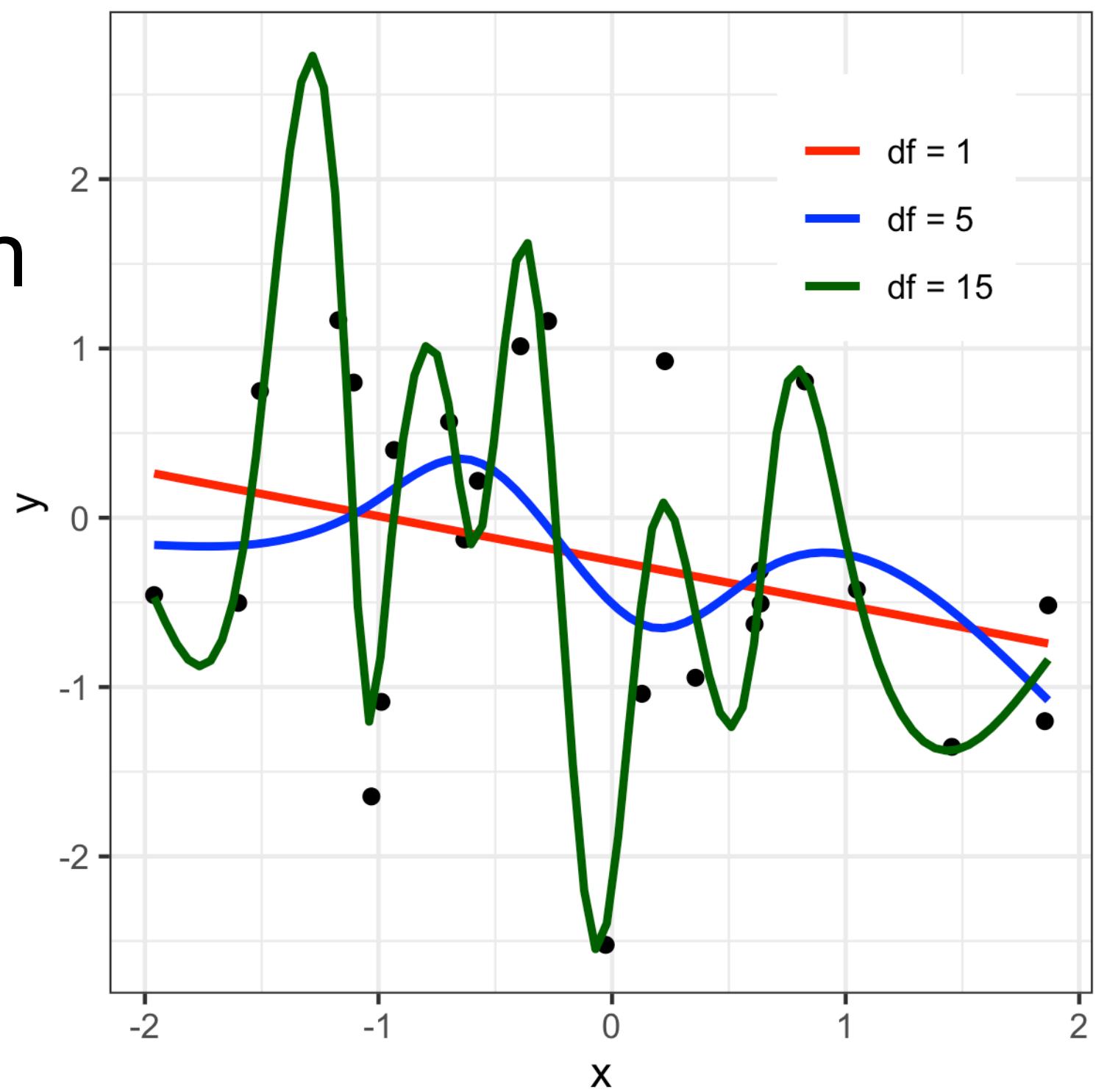


Theme: Model complexity

One of the key characteristics of a predictive model is its complexity: how flexibly does it fit the training data?

How is model complexity defined? Depends on model:

- Number of parameters in linear or logistic regression
- Depth of a decision tree

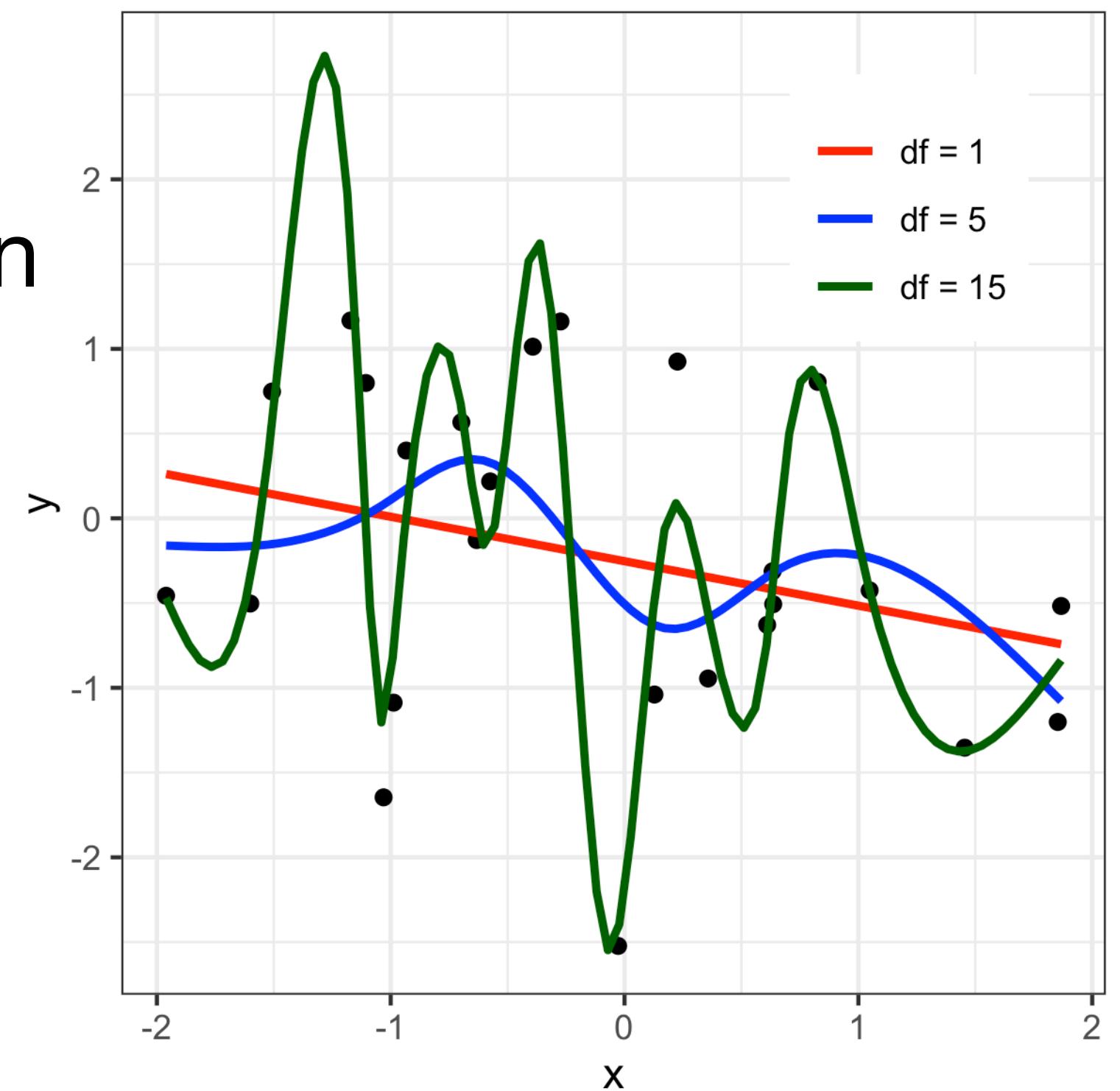


Theme: Model complexity

One of the key characteristics of a predictive model is its complexity: how flexibly does it fit the training data?

How is model complexity defined? Depends on model:

- Number of parameters in linear or logistic regression
- Depth of a decision tree
- Number of neighbors used in K-nearest-neighbors



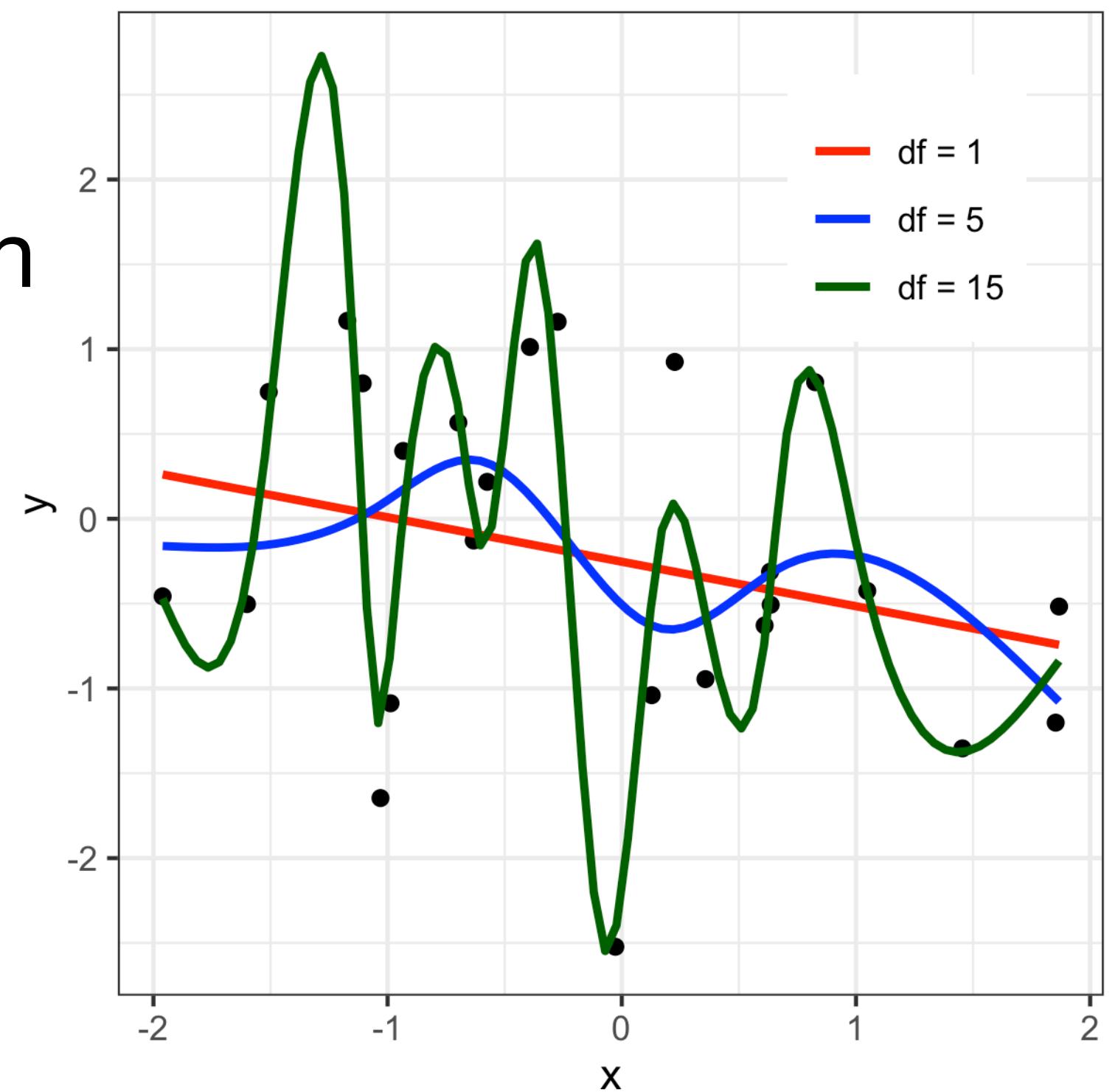
Theme: Model complexity

One of the key characteristics of a predictive model is its complexity: how flexibly does it fit the training data?

How is model complexity defined? Depends on model:

- Number of parameters in linear or logistic regression
- Depth of a decision tree
- Number of neighbors used in K-nearest-neighbors

How is model complexity controlled (regularization)?



Theme: Model complexity

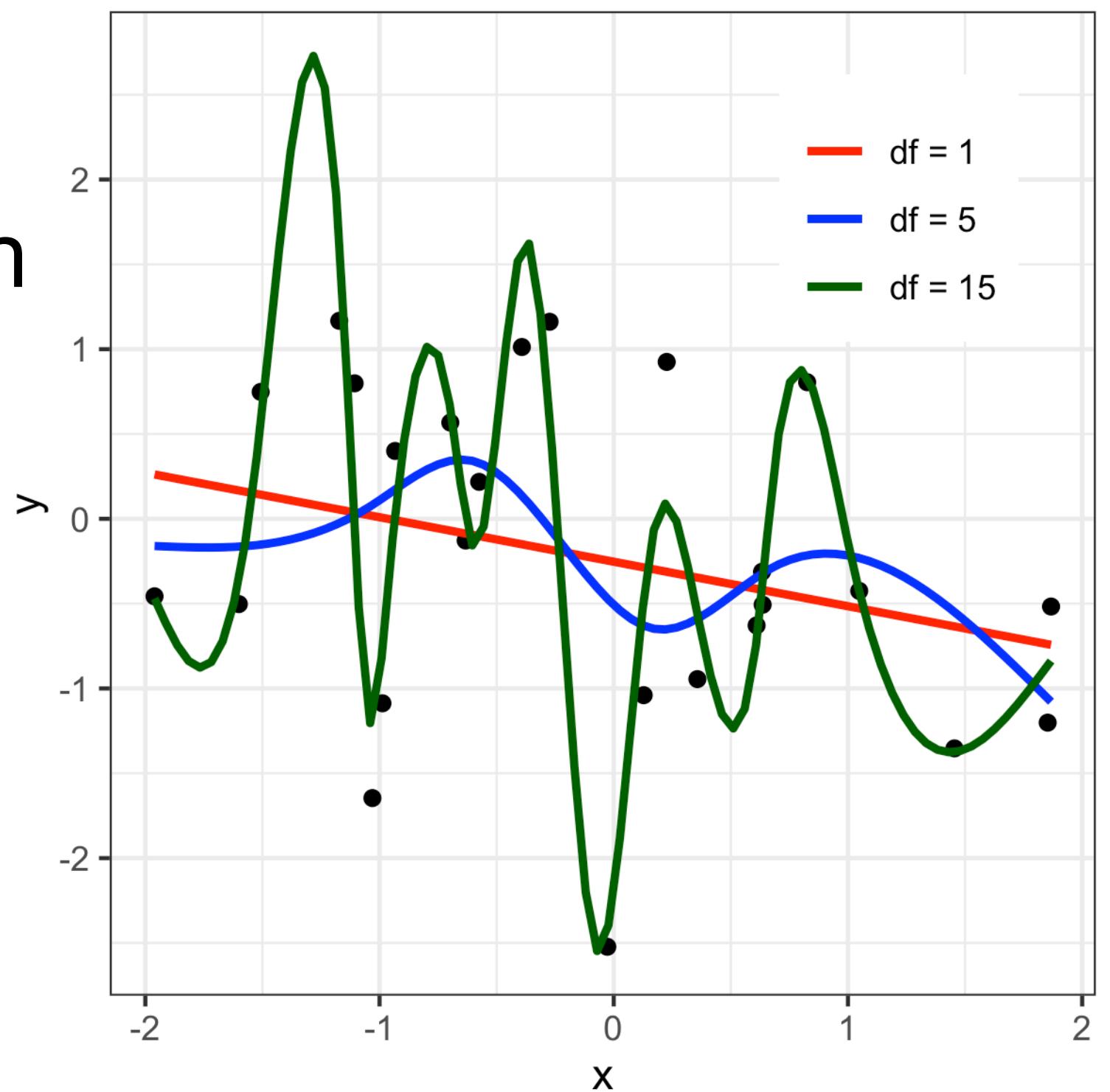
One of the key characteristics of a predictive model is its complexity: how flexibly does it fit the training data?

How is model complexity defined? Depends on model:

- Number of parameters in linear or logistic regression
- Depth of a decision tree
- Number of neighbors used in K-nearest-neighbors

How is model complexity controlled (regularization)?

- Explicit regularization via penalization (lasso, ridge)



Theme: Model complexity

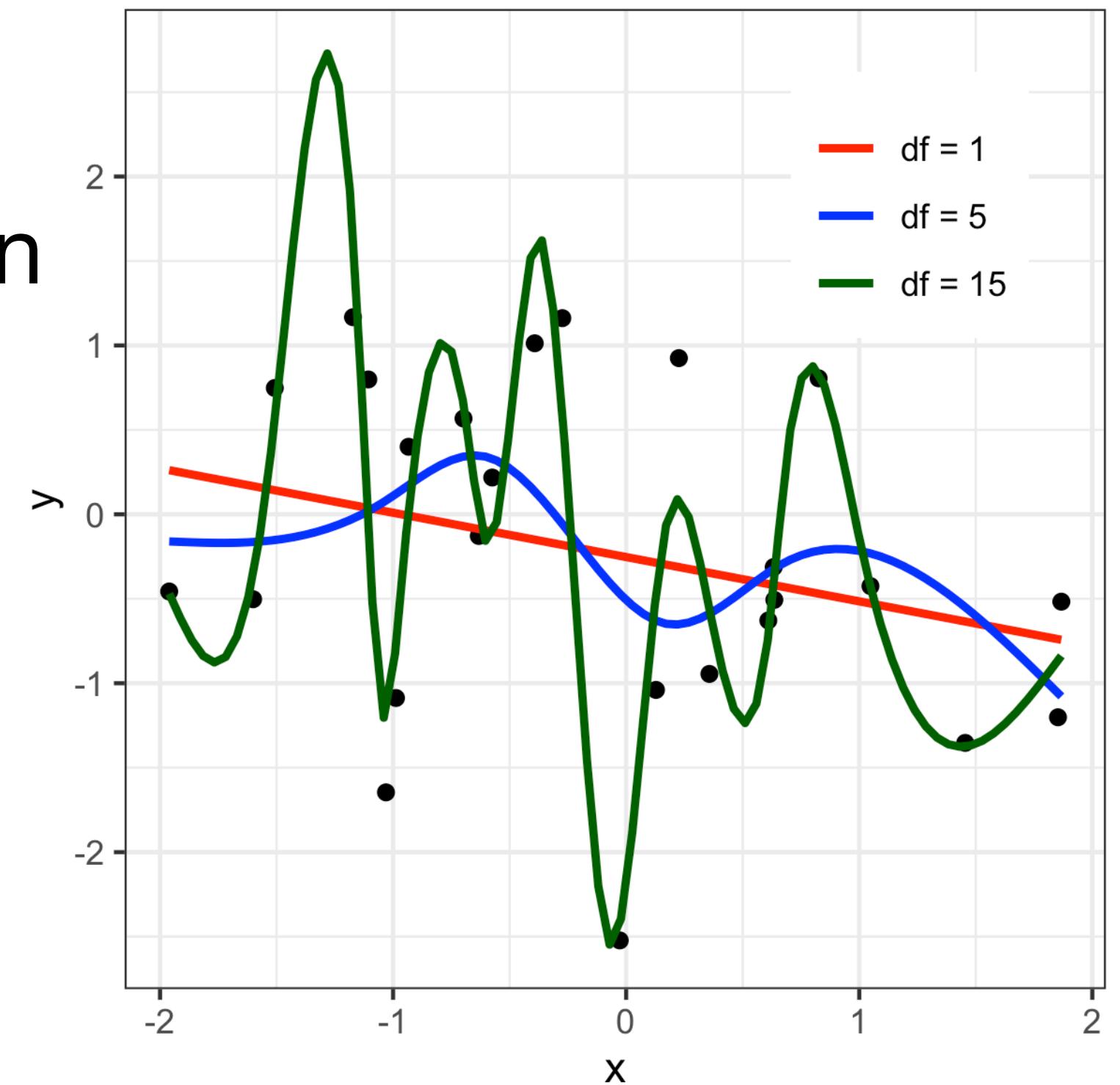
One of the key characteristics of a predictive model is its complexity: how flexibly does it fit the training data?

How is model complexity defined? Depends on model:

- Number of parameters in linear or logistic regression
- Depth of a decision tree
- Number of neighbors used in K-nearest-neighbors

How is model complexity controlled (regularization)?

- Explicit regularization via penalization (lasso, ridge)
- Implicit regularization, e.g. sub-sampling features during random forest model training

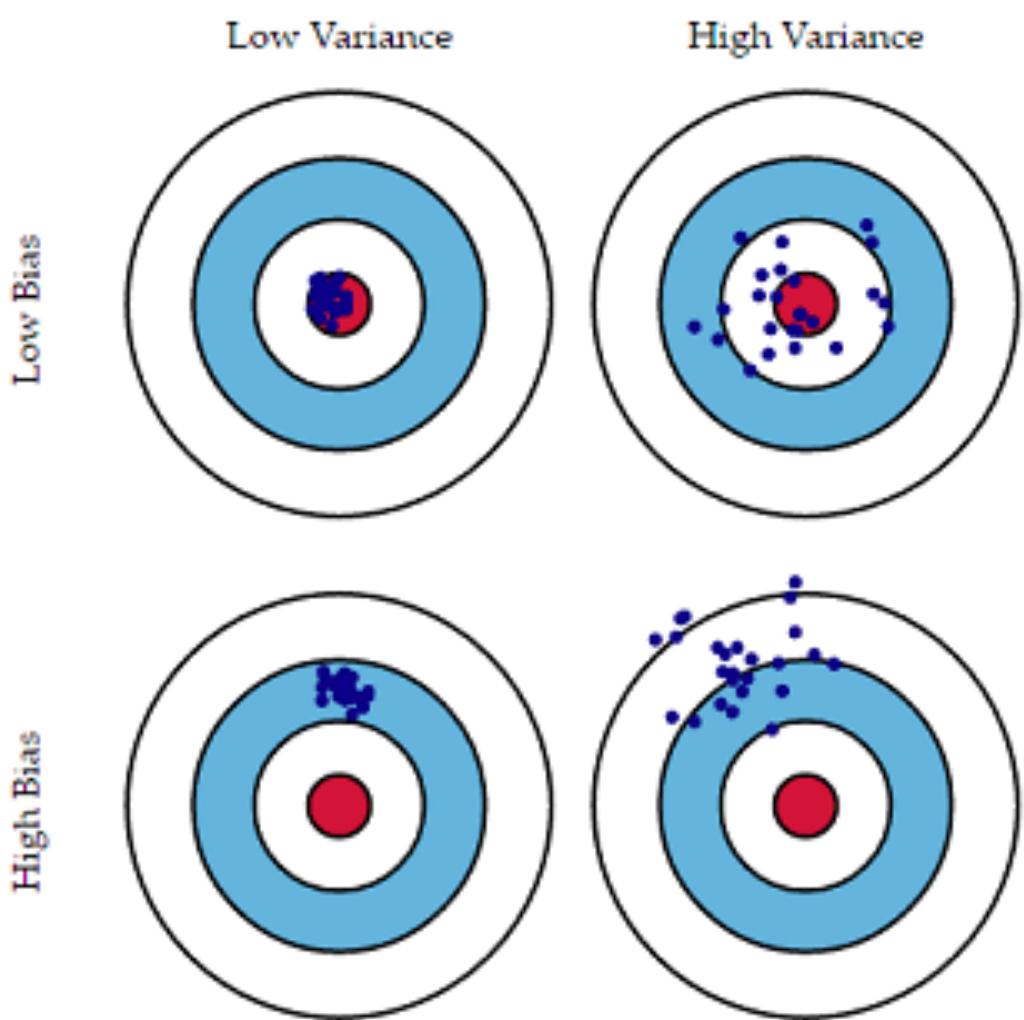


Theme: Bias-variance trade-off

Theme: Bias-variance trade-off

Consider sampling many different training sets.

- Bias: How far off are predictions on average?
- Variance: How much do the predictions wobble around?



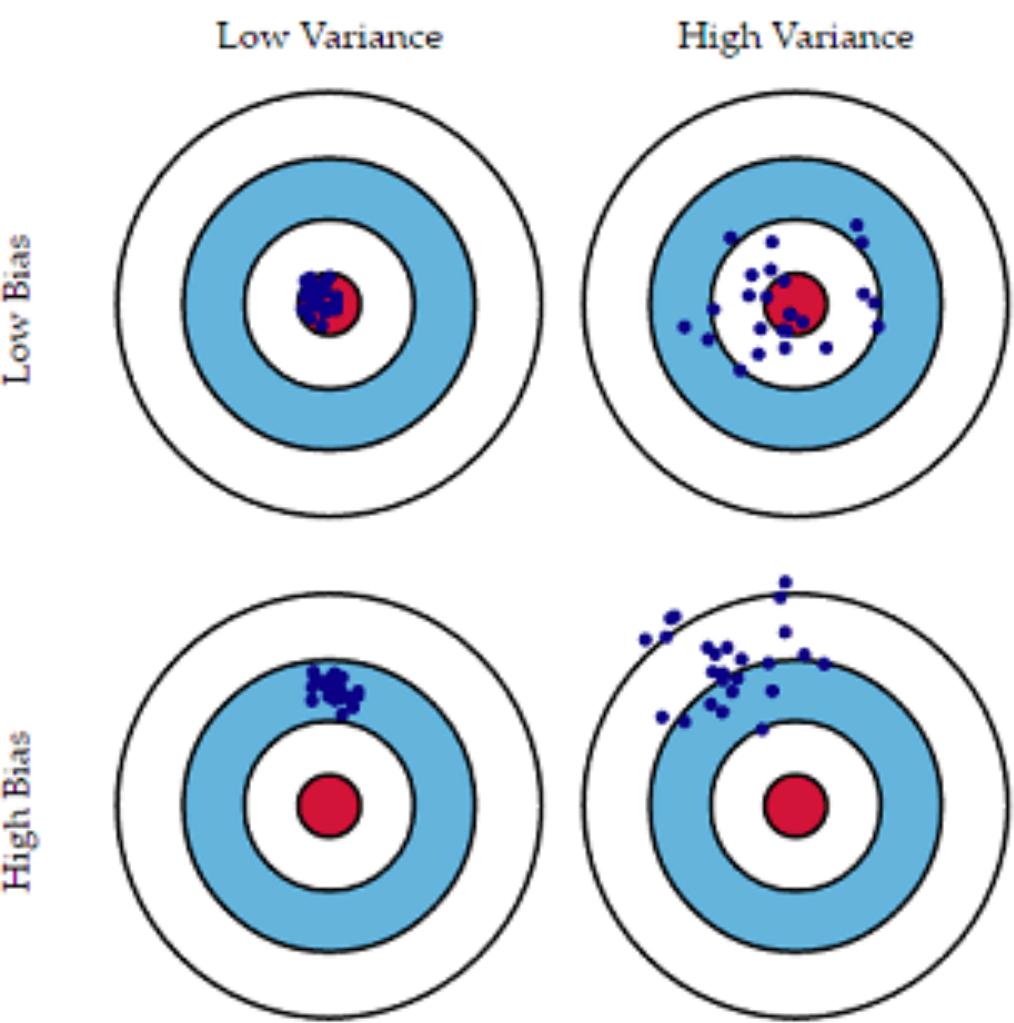
<https://www.listendata.com/2017/02/bias-variance-tradeoff.html>

Theme: Bias-variance trade-off

Consider sampling many different training sets.

- Bias: How far off are predictions on average?
- Variance: How much do the predictions wobble around?

Higher model complexity leads to less bias but more variance.



<https://www.listendata.com/2017/02/bias-variance-tradeoff.html>

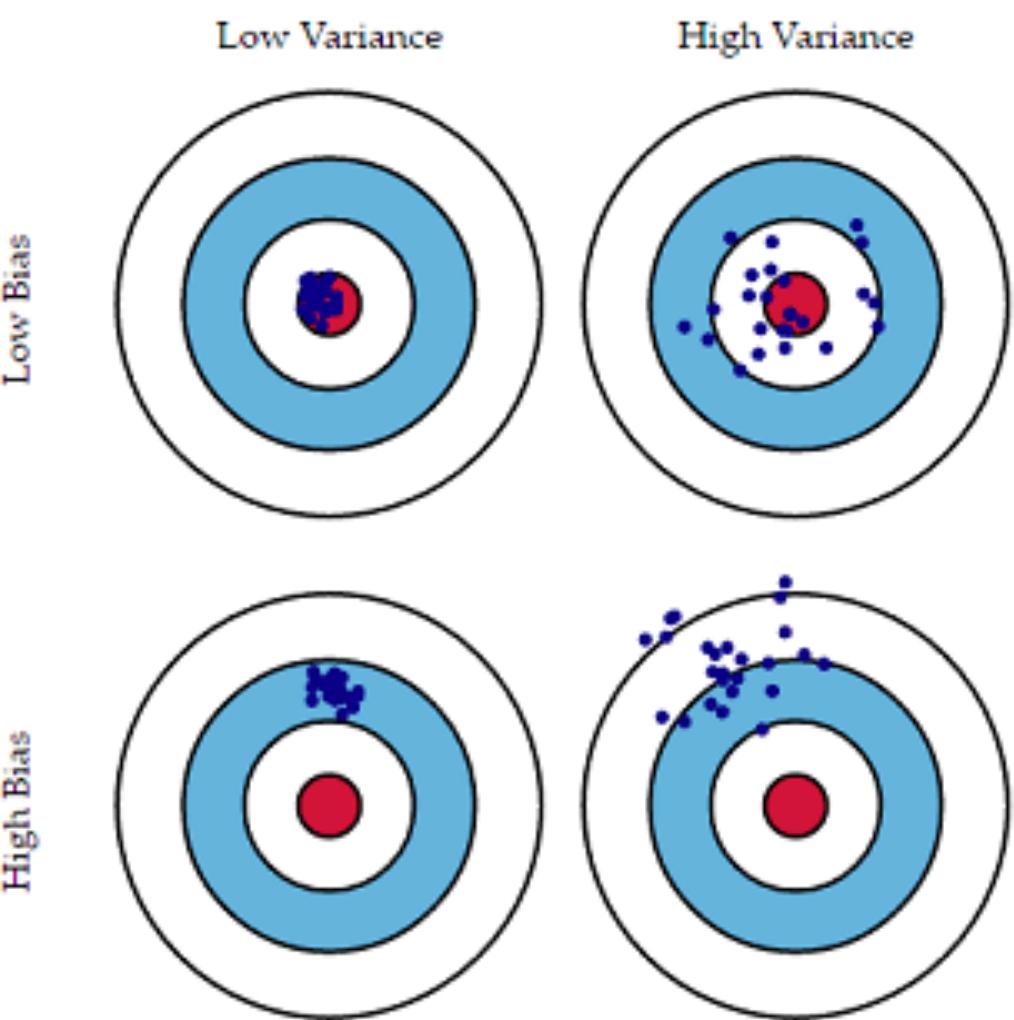
Theme: Bias-variance trade-off

Consider sampling many different training sets.

- Bias: How far off are predictions on average?
- Variance: How much do the predictions wobble around?

Higher model complexity leads to less bias but more variance.

$$\text{Prediction error} = \text{Bias}^2 + \text{Variance}.$$



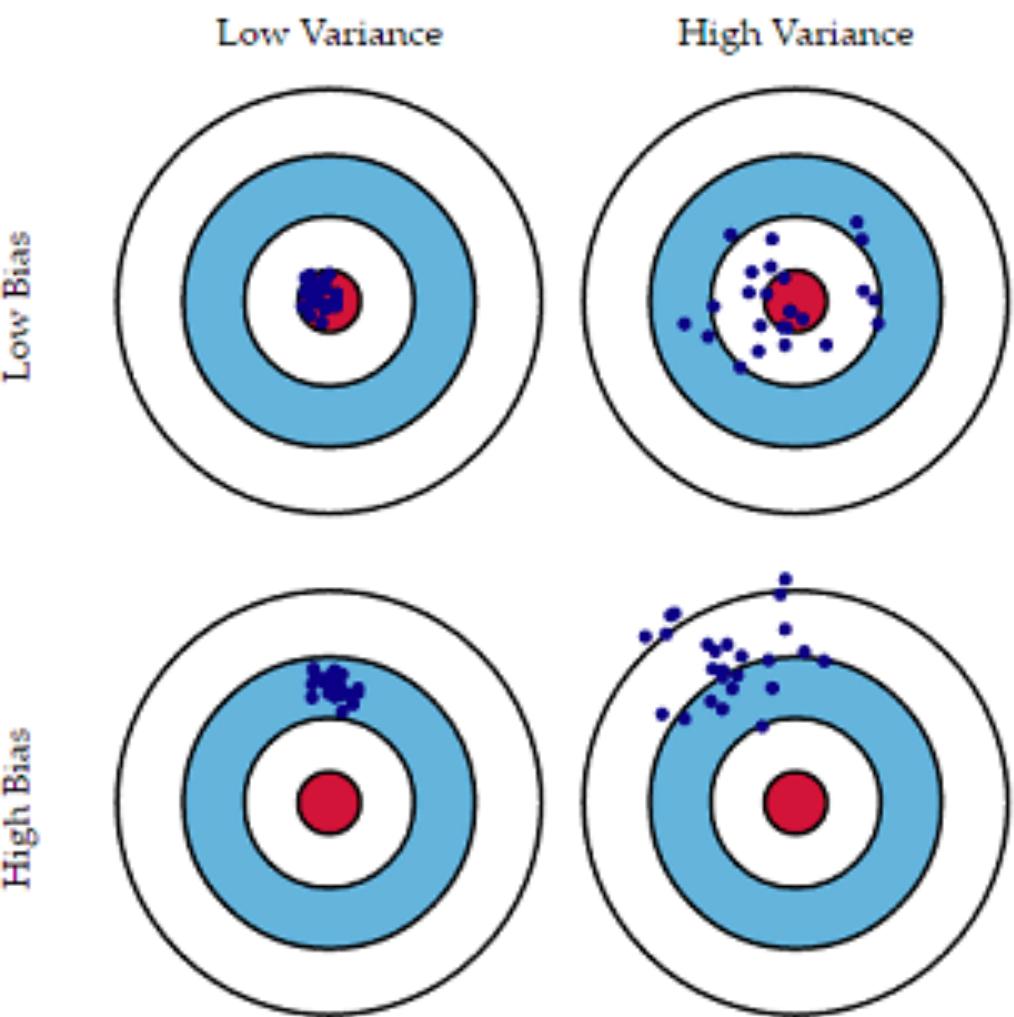
<https://www.listendata.com/2017/02/bias-variance-tradeoff.html>

Theme: Bias-variance trade-off

Consider sampling many different training sets.

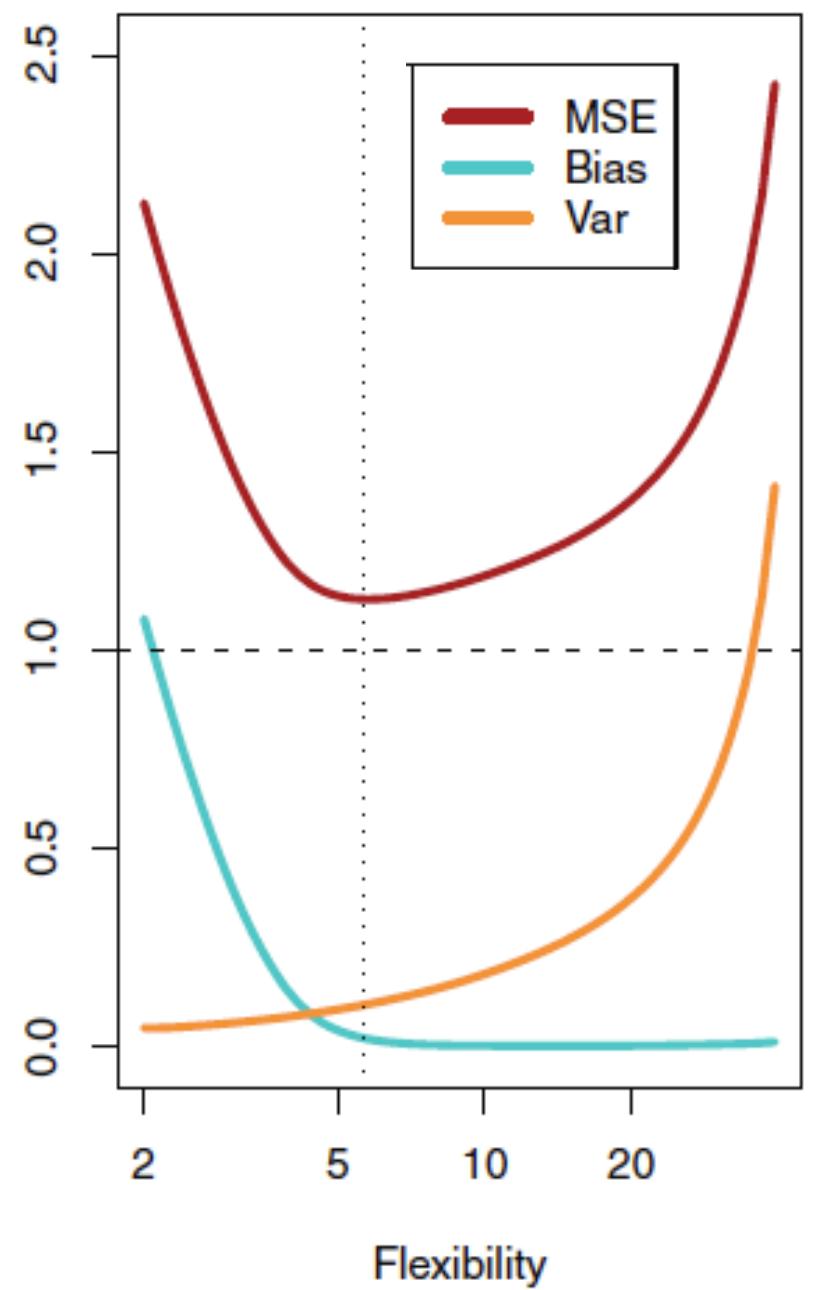
- Bias: How far off are predictions on average?
- Variance: How much do the predictions wobble around?

Higher model complexity leads to less bias but more variance.



<https://www.listendata.com/2017/02/bias-variance-tradeoff.html>

Prediction error = Bias²+Variance.

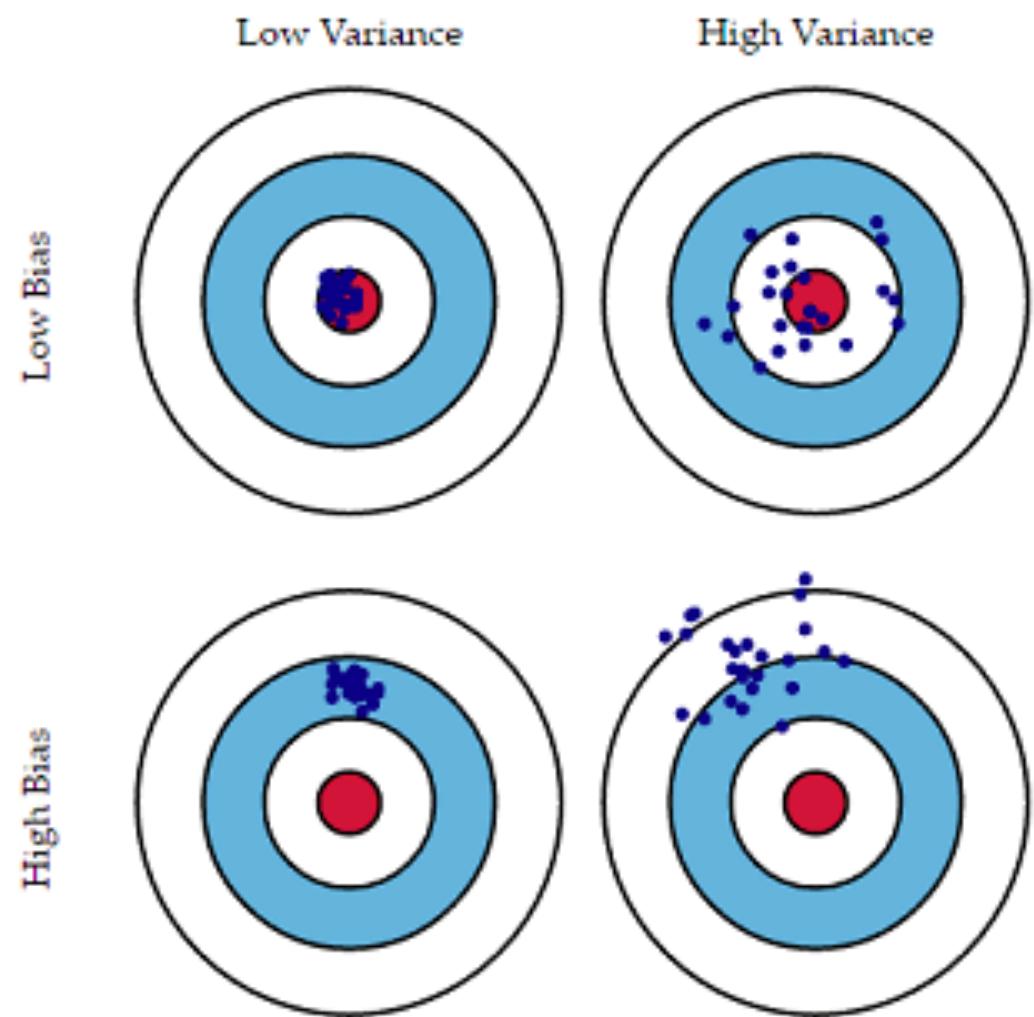


Theme: Bias-variance trade-off

Consider sampling many different training sets.

- Bias: How far off are predictions on average?
- Variance: How much do the predictions wobble around?

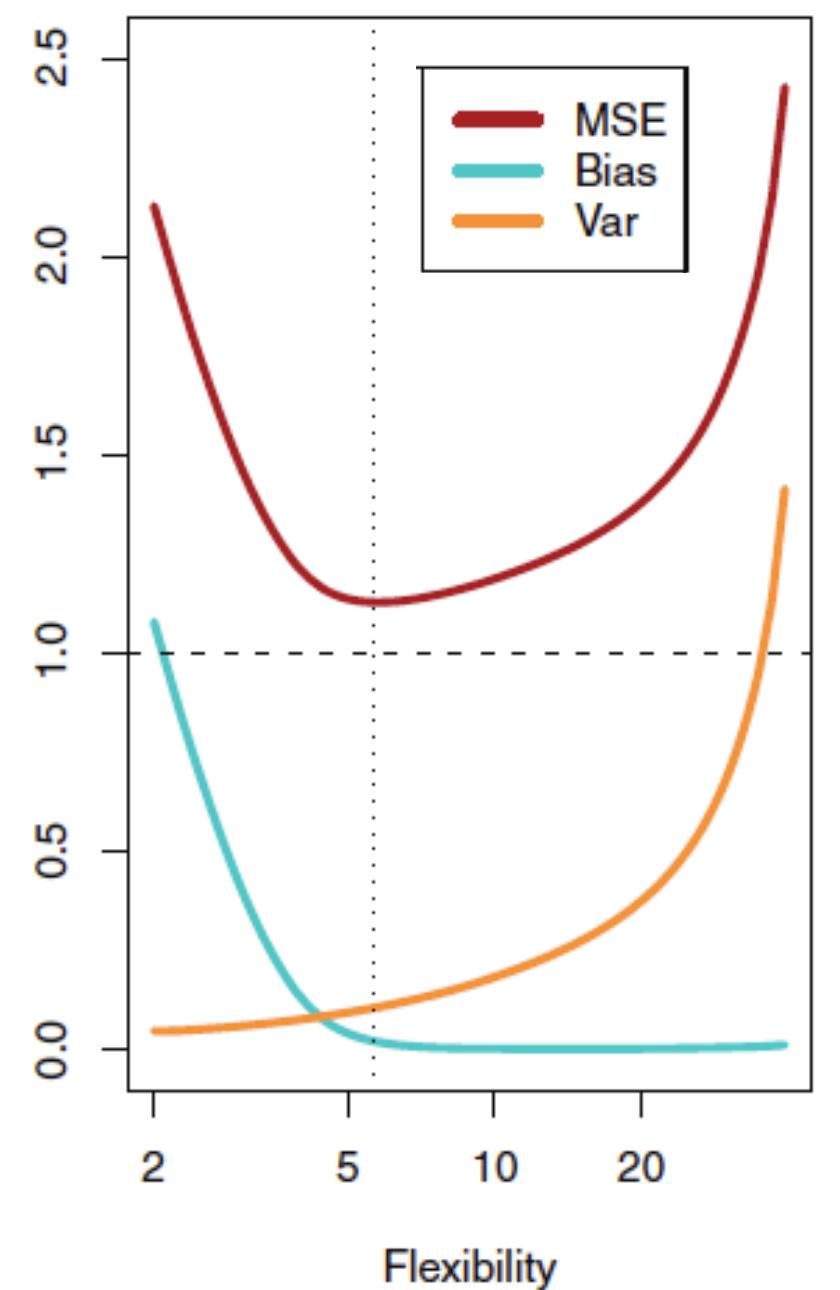
Higher model complexity leads to less bias but more variance.



<https://www.listendata.com/2017/02/bias-variance-tradeoff.html>

$$\text{Prediction error} = \text{Bias}^2 + \text{Variance}.$$

Overfitting: complex models' sensitivity to noise in the training data (high variance) → low training error but high test error.

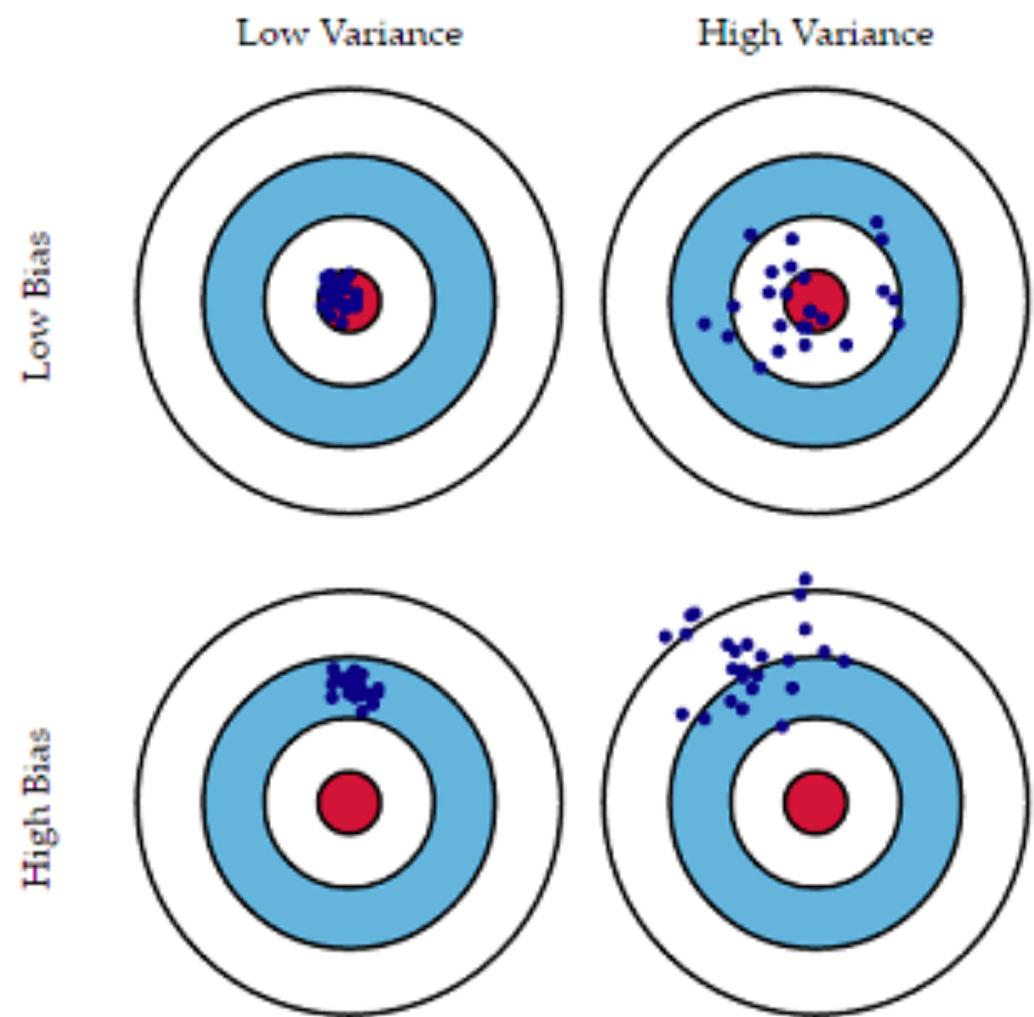


Theme: Bias-variance trade-off

Consider sampling many different training sets.

- Bias: How far off are predictions on average?
- Variance: How much do the predictions wobble around?

Higher model complexity leads to less bias but more variance.

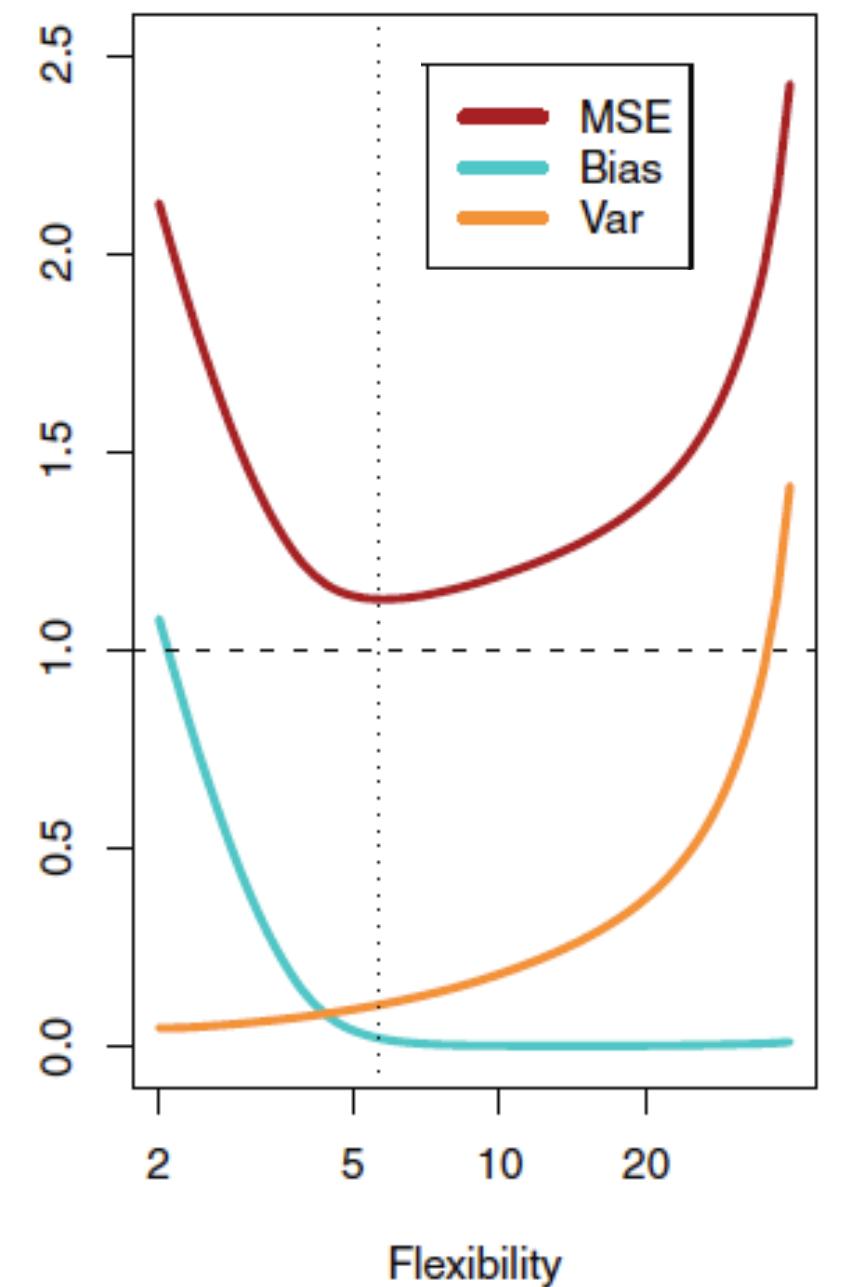


<https://www.listendata.com/2017/02/bias-variance-tradeoff.html>

$$\text{Prediction error} = \text{Bias}^2 + \text{Variance}.$$

Overfitting: complex models' sensitivity to noise in the training data (high variance) → low training error but high test error.

Variance increases as noise variance increases, model flexibility increases, training sample size decreases.

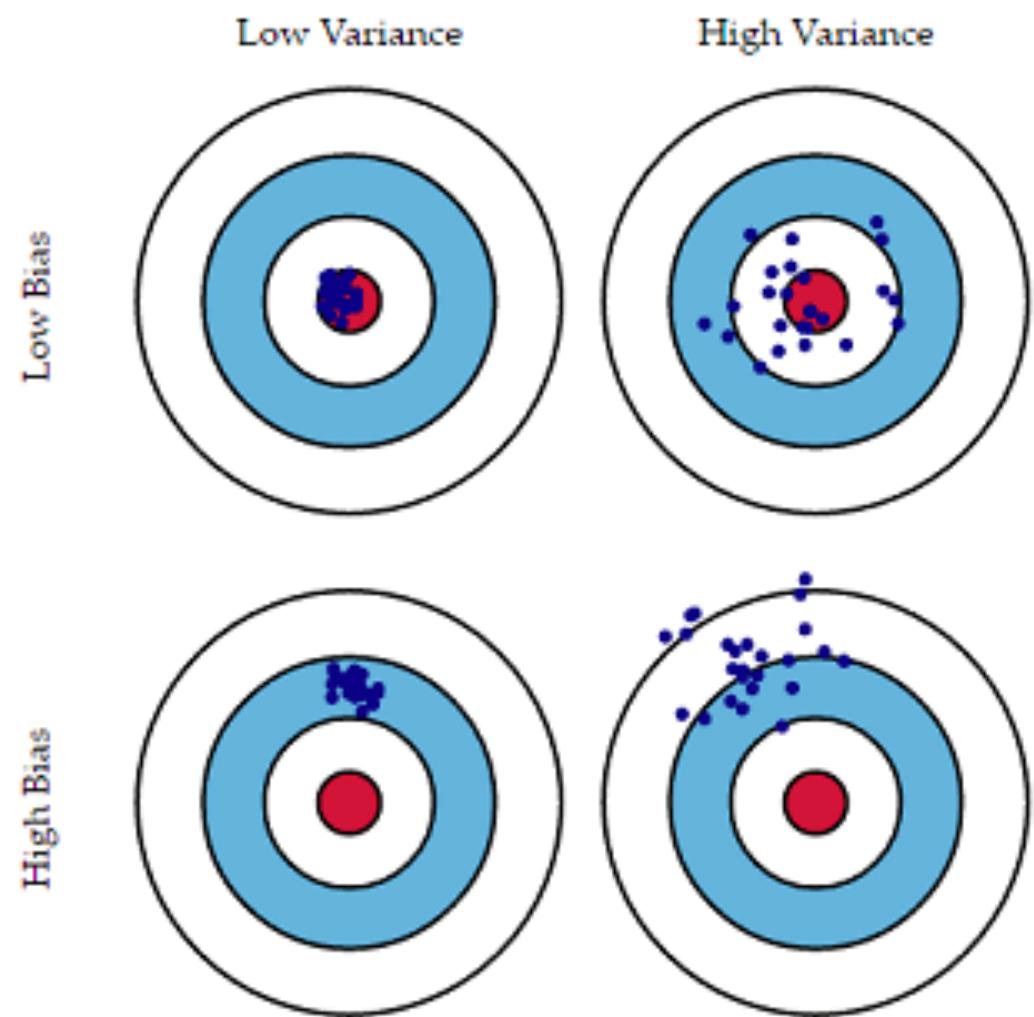


Theme: Bias-variance trade-off

Consider sampling many different training sets.

- Bias: How far off are predictions on average?
- Variance: How much do the predictions wobble around?

Higher model complexity leads to less bias but more variance.



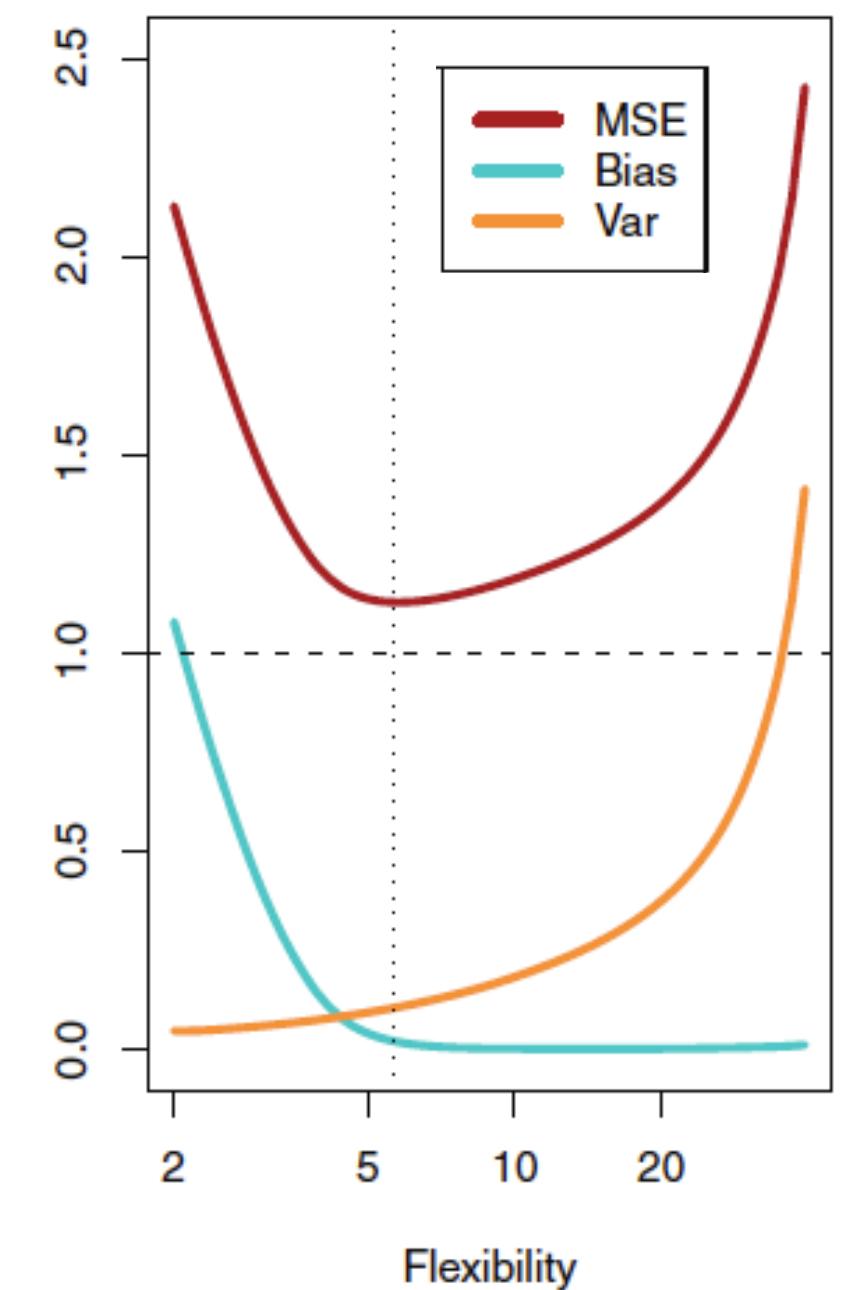
<https://www.listendata.com/2017/02/bias-variance-tradeoff.html>

$$\text{Prediction error} = \text{Bias}^2 + \text{Variance}.$$

Overfitting: complex models' sensitivity to noise in the training data (high variance) → low training error but high test error.

Variance increases as noise variance increases, model flexibility increases, training sample size decreases.

Lower noise or larger sample size means you can afford more complex model (think of deep learning).



Theme: Model selection and model assessment

Theme: Model selection and model assessment

Model assessment

Theme: Model selection and model assessment

Model assessment

- Predictive models are assessed based on test data that is separate from the data used to train these models.

Theme: Model selection and model assessment

Model assessment

- Predictive models are assessed based on test data that is separate from the data used to train these models.
- Different criteria are used to quantify the accuracy of predictions, like RMSE, misclassification error, F-score, and confusion matrix.

Theme: Model selection and model assessment

Model assessment

- Predictive models are assessed based on test data that is separate from the data used to train these models.
- Different criteria are used to quantify the accuracy of predictions, like RMSE, misclassification error, F-score, and confusion matrix.
- Model assessment can be subtle for classification problems.

Theme: Model selection and model assessment

Model assessment

- Predictive models are assessed based on test data that is separate from the data used to train these models.
- Different criteria are used to quantify the accuracy of predictions, like RMSE, misclassification error, F-score, and confusion matrix.
- Model assessment can be subtle for classification problems.

Model selection

Theme: Model selection and model assessment

Model assessment

- Predictive models are assessed based on test data that is separate from the data used to train these models.
- Different criteria are used to quantify the accuracy of predictions, like RMSE, misclassification error, F-score, and confusion matrix.
- Model assessment can be subtle for classification problems.

Model selection

- Main tool: Cross-validation, which mimics the train-test split using folds.

Theme: Model selection and model assessment

Model assessment

- Predictive models are assessed based on test data that is separate from the data used to train these models.
- Different criteria are used to quantify the accuracy of predictions, like RMSE, misclassification error, F-score, and confusion matrix.
- Model assessment can be subtle for classification problems.

Model selection

- Main tool: Cross-validation, which mimics the train-test split using folds.
- Other schemes for model assessment: validation set approach and out-of-bag error, the latter for random forests.

Theme: Model selection and model assessment

Model assessment

- Predictive models are assessed based on test data that is separate from the data used to train these models.
- Different criteria are used to quantify the accuracy of predictions, like RMSE, misclassification error, F-score, and confusion matrix.
- Model assessment can be subtle for classification problems.

Model selection

- Main tool: Cross-validation, which mimics the train-test split using folds.
- Other schemes for model assessment: validation set approach and out-of-bag error, the latter for random forests.
- One-standard-error rule reflects preference for simpler models.

Theme: Interpretability of predictive models

We want to understand how our predictive model is arriving at its conclusions.

Theme: Interpretability of predictive models

We want to understand how our predictive model is arriving at its conclusions.



Theme: Interpretability of predictive models

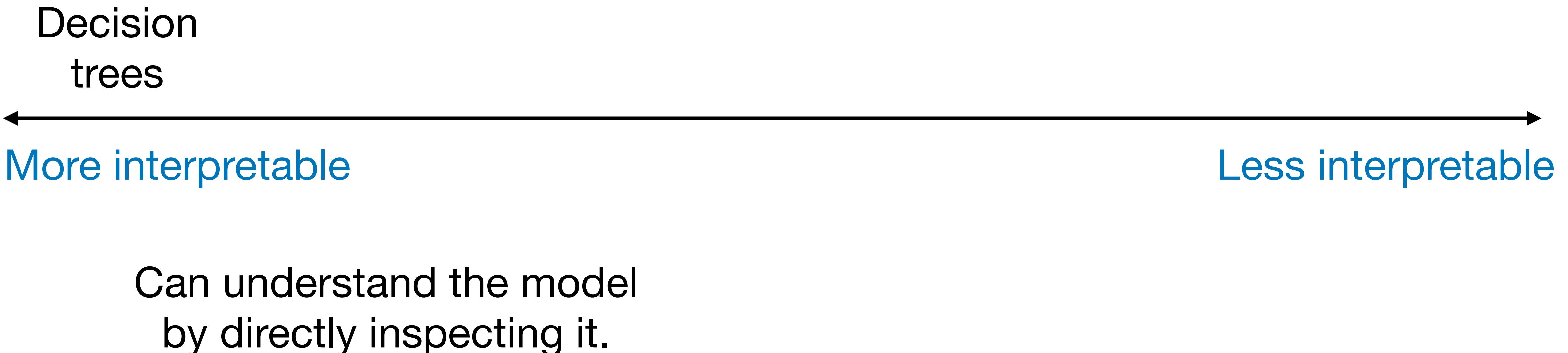
We want to understand how our predictive model is arriving at its conclusions.



Can understand the model
by directly inspecting it.

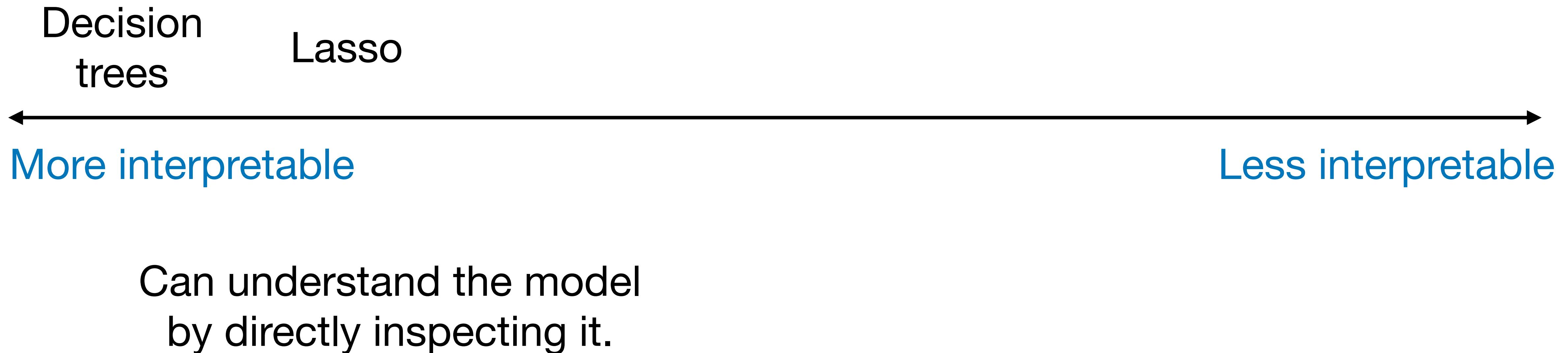
Theme: Interpretability of predictive models

We want to understand how our predictive model is arriving at its conclusions.



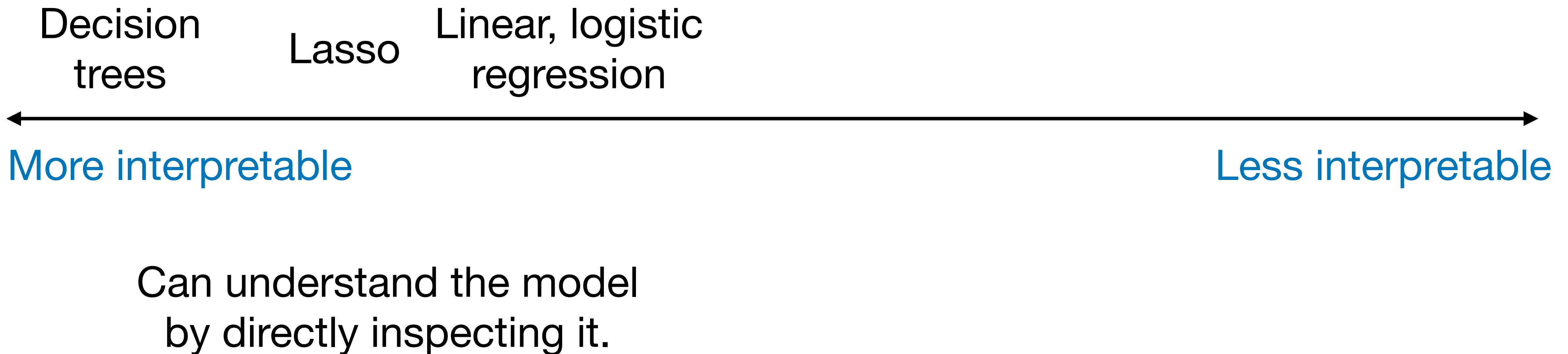
Theme: Interpretability of predictive models

We want to understand how our predictive model is arriving at its conclusions.



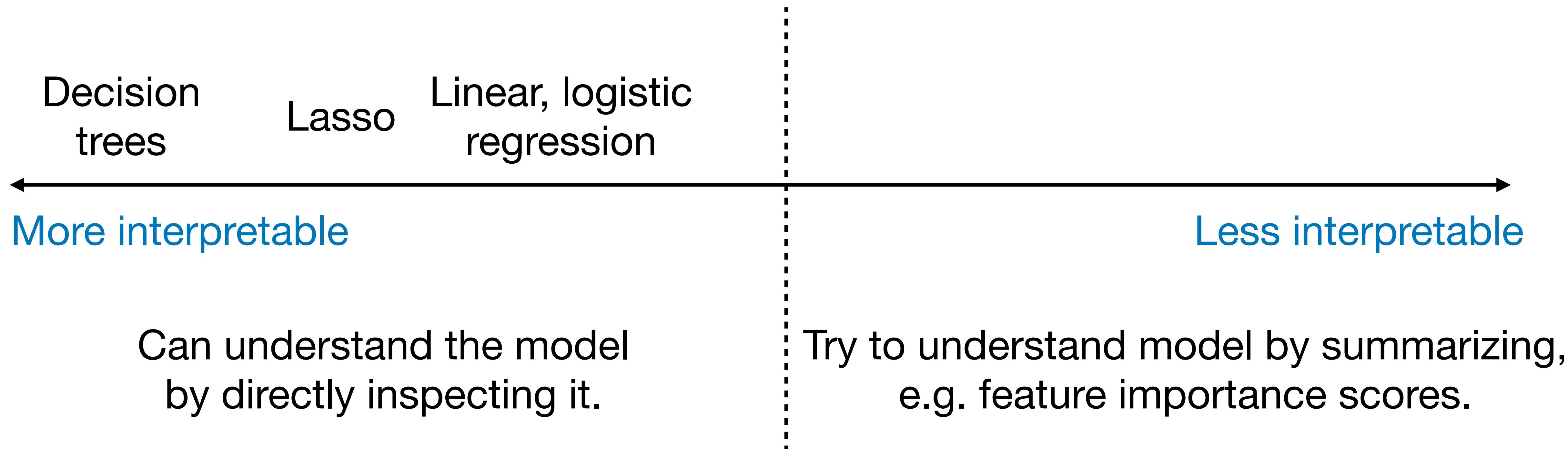
Theme: Interpretability of predictive models

We want to understand how our predictive model is arriving at its conclusions.



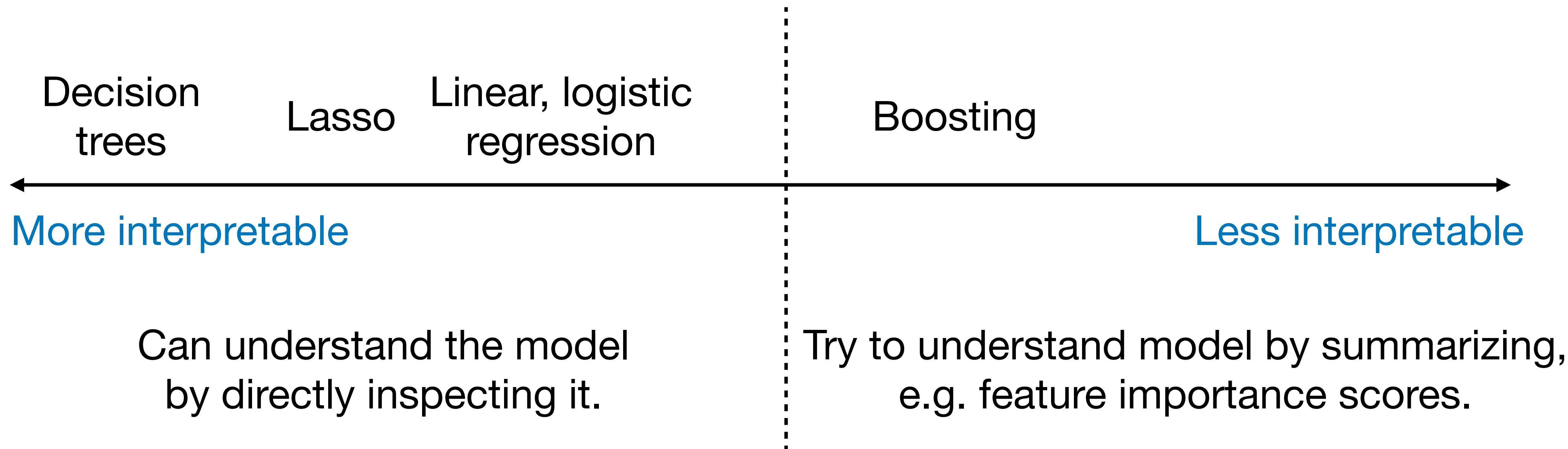
Theme: Interpretability of predictive models

We want to understand how our predictive model is arriving at its conclusions.



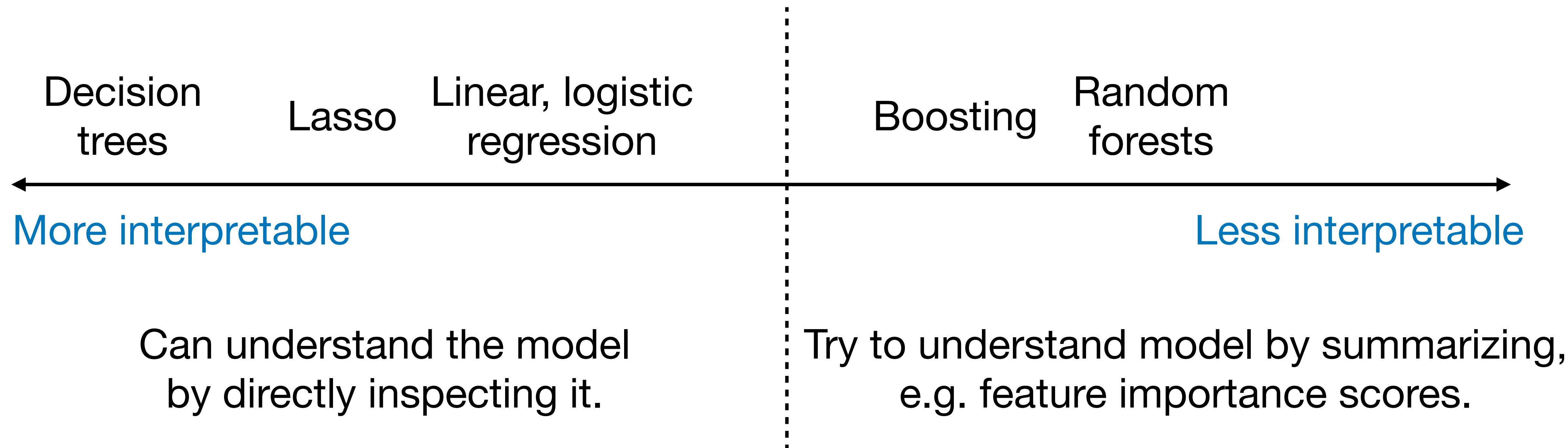
Theme: Interpretability of predictive models

We want to understand how our predictive model is arriving at its conclusions.



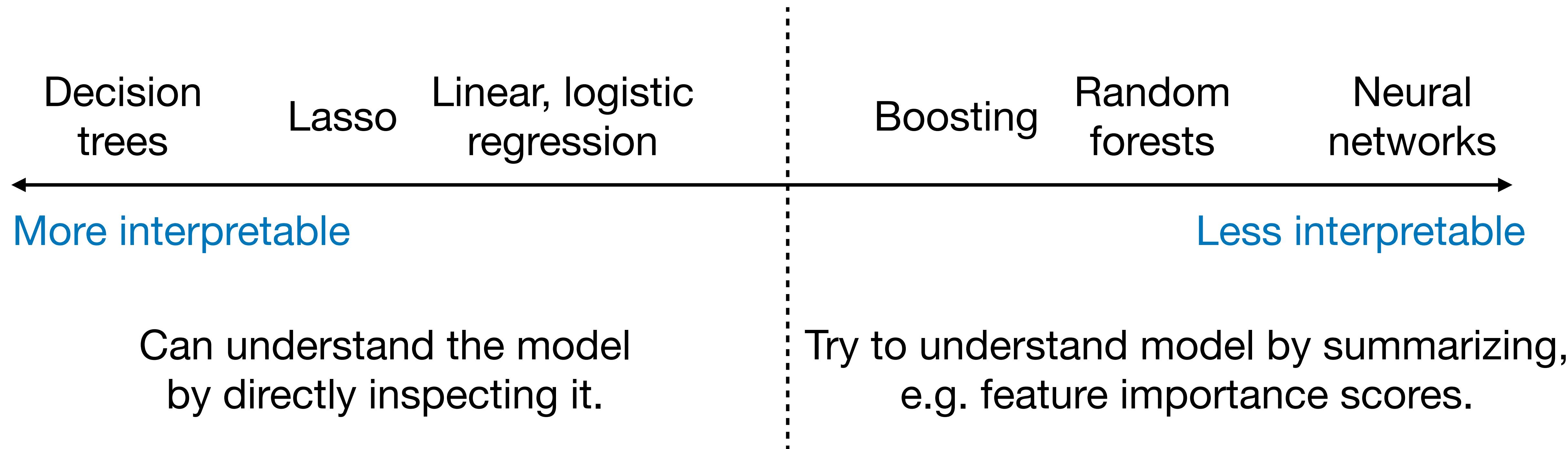
Theme: Interpretability of predictive models

We want to understand how our predictive model is arriving at its conclusions.



Theme: Interpretability of predictive models

We want to understand how our predictive model is arriving at its conclusions.



Theme: R programming tools

Theme: R programming tools

Tools:

Theme: R programming tools

Tools:

- `tidyverse` is a nice way to import, clean, transform, and visualize data.

Theme: R programming tools

Tools:

- `tidyverse` is a nice way to import, clean, transform, and visualize data.
- `R` is well suited for data science; lots of packages available to analyze data.

Theme: R programming tools

Tools:

- **tidyverse** is a nice way to import, clean, transform, and visualize data.
- **R** is well suited for data science; lots of packages available to analyze data.
- **R Markdown** is a nice way to integrate text, code, and output.

Theme: R programming tools

Tools:

- **tidyverse** is a nice way to import, clean, transform, and visualize data.
- **R** is well suited for data science; lots of packages available to analyze data.
- **R Markdown** is a nice way to integrate text, code, and output.
- **RStudio Cloud** gives a ready-to-go R programming environment.

Theme: R programming tools

Tools:

- **tidyverse** is a nice way to import, clean, transform, and visualize data.
- **R** is well suited for data science; lots of packages available to analyze data.
- **R Markdown** is a nice way to integrate text, code, and output.
- **RStudio Cloud** gives a ready-to-go R programming environment.

Lessons learned:

Theme: R programming tools

Tools:

- **tidyverse** is a nice way to import, clean, transform, and visualize data.
- **R** is well suited for data science; lots of packages available to analyze data.
- **R Markdown** is a nice way to integrate text, code, and output.
- **RStudio Cloud** gives a ready-to-go R programming environment.

Lessons learned:

- Programming takes patience, attention to detail, and lots of Googling.

Theme: R programming tools

Tools:

- **tidyverse** is a nice way to import, clean, transform, and visualize data.
- **R** is well suited for data science; lots of packages available to analyze data.
- **R Markdown** is a nice way to integrate text, code, and output.
- **RStudio Cloud** gives a ready-to-go R programming environment.

Lessons learned:

- Programming takes patience, attention to detail, and lots of Googling.
- Each R package and each software tool has its own quirks and limitations.

Theme: R programming tools

Tools:

- [tidyverse](#) is a nice way to import, clean, transform, and visualize data.
- [R](#) is well suited for data science; lots of packages available to analyze data.
- [R Markdown](#) is a nice way to integrate text, code, and output.
- [RStudio Cloud](#) gives a ready-to-go R programming environment.

Lessons learned:

- Programming takes patience, attention to detail, and lots of Googling.
- Each R package and each software tool has its own quirks and limitations.
- With practice, these programming and software tools can be very powerful.

Theme: Working with data

Theme: Working with data

- Though increasingly abundant, data are still a precious resource, more of which gives better predictions.

Theme: Working with data

- Though increasingly abundant, data are still a precious resource, more of which gives better predictions.
- Especially in the real world, data are messy and require cleaning.

Theme: Working with data

- Though increasingly abundant, data are still a precious resource, more of which gives better predictions.
- Especially in the real world, data are messy and require cleaning.
- Exploratory data analysis and visualization can reveal a lot about a dataset.

Theme: Working with data

- Though increasingly abundant, data are still a precious resource, more of which gives better predictions.
- Especially in the real world, data are messy and require cleaning.
- Exploratory data analysis and visualization can reveal a lot about a dataset.
- The most successful analyses couple statistical intuition and data intuition.

Theme: Working with data

- Though increasingly abundant, data are still a precious resource, more of which gives better predictions.
- Especially in the real world, data are messy and require cleaning.
- Exploratory data analysis and visualization can reveal a lot about a dataset.
- The most successful analyses couple statistical intuition and data intuition.
- Ultimate goal of data science is to create knowledge and/or make decisions; we must make conclusions relevant to the underlying real-world problem.

A lingering question: What is the best method?

A lingering question: What is the best method?

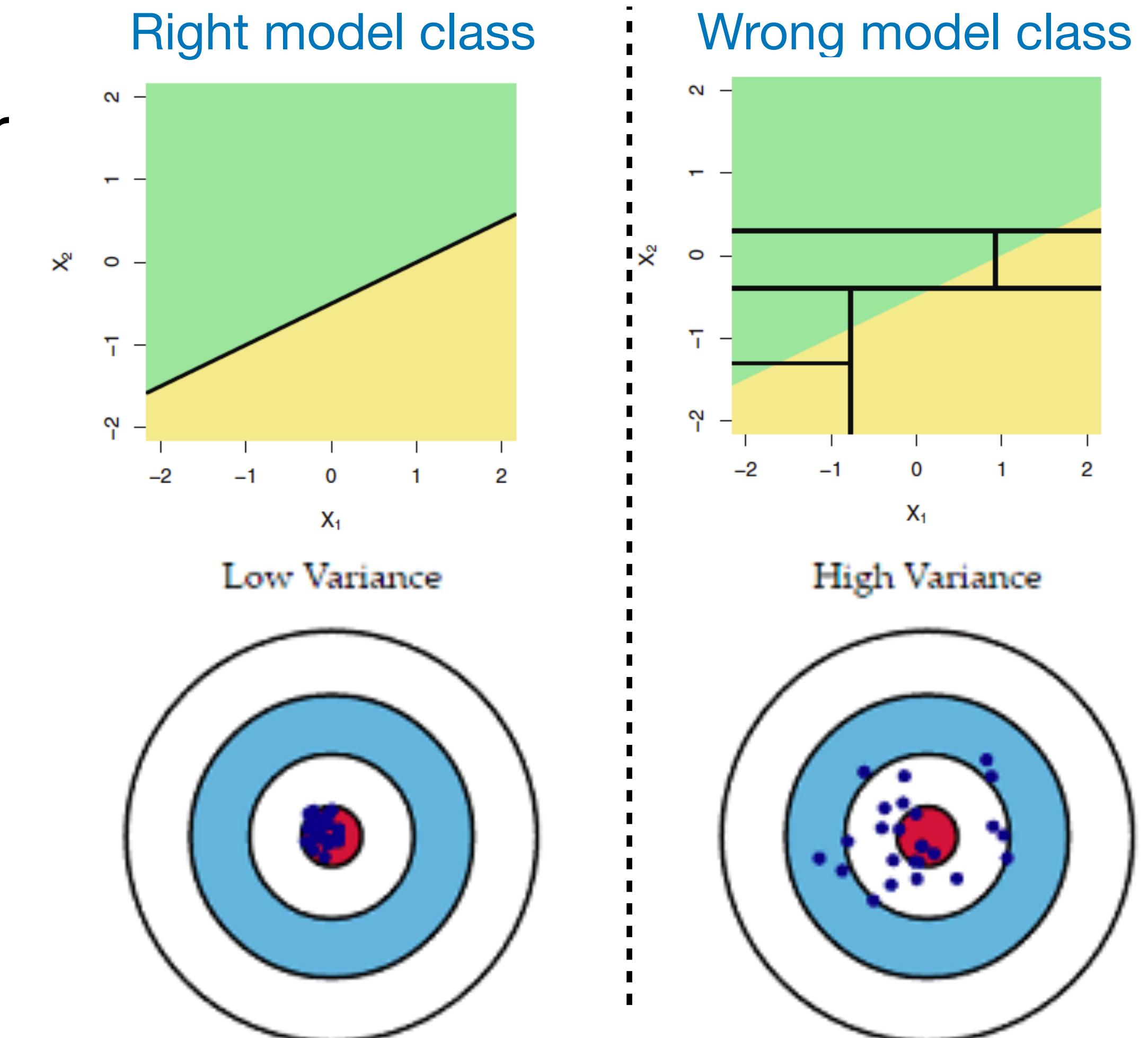
There is no best method. Different methods will work better in different situations.

A lingering question: What is the best method?

There is no best method. Different methods will work better in different situations.

Each prediction method “has in mind” an underlying model class, e.g. linear models for linear regression versus piece-wise constant models for trees.

If true feature-response relationship matches model class our method “has in mind,” will take fewer parameters (less variance) to fit underlying trend (less bias).



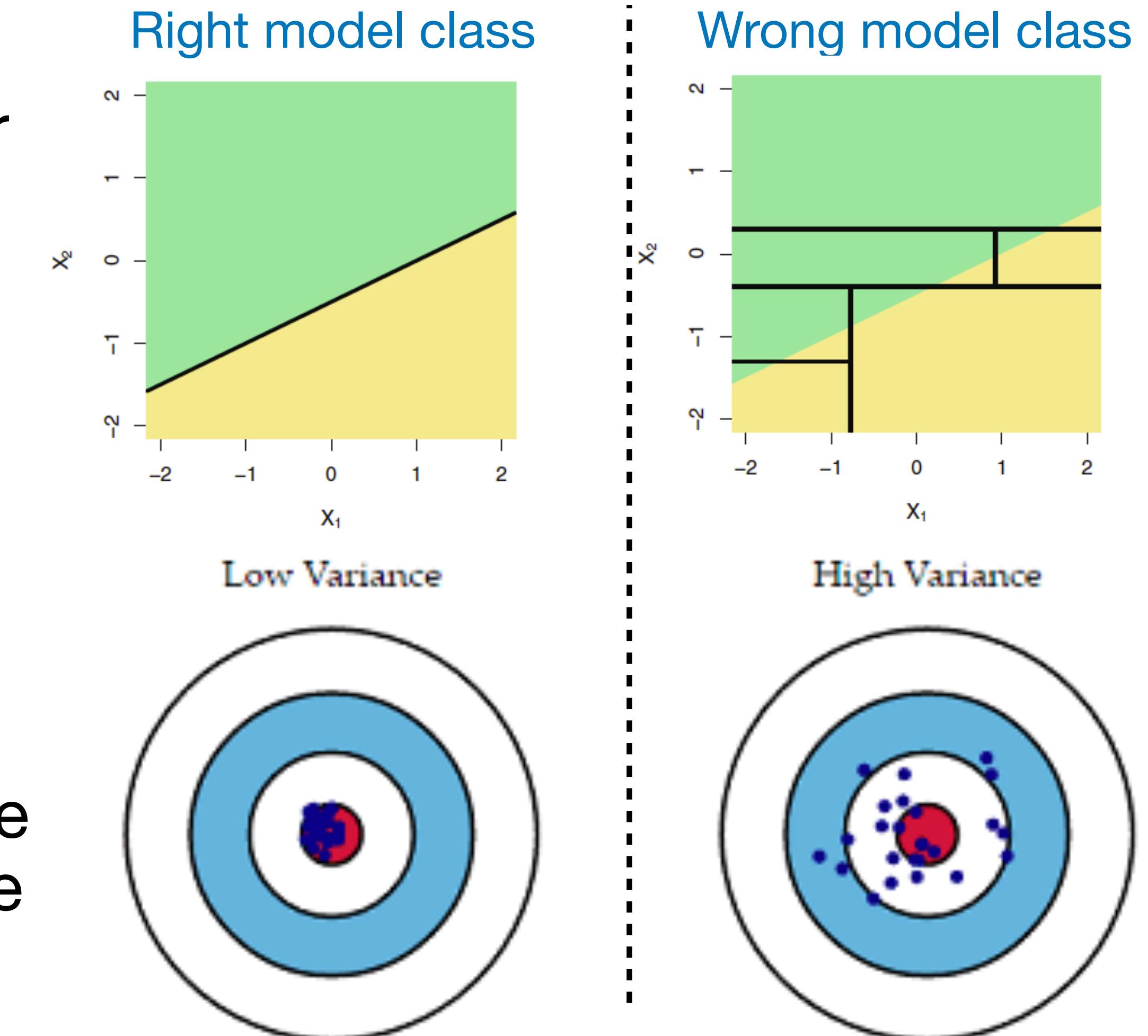
A lingering question: What is the best method?

There is no best method. Different methods will work better in different situations.

Each prediction method “has in mind” an underlying model class, e.g. linear models for linear regression versus piece-wise constant models for trees.

If true feature-response relationship matches model class our method “has in mind,” will take fewer parameters (less variance) to fit underlying trend (less bias).

Moral of the story: It’s good to know several prediction methods. Seek out the ones whose underlying model class you think matches the true feature-response relationship.



Looking beyond STAT 4710

What comes next?

What comes next?

- Where are data mining and machine learning going in the future?

What comes next?

- Where are data mining and machine learning going in the future?
- How do I learn more about data mining and machine learning?

What comes next?

- Where are data mining and machine learning going in the future?
- How do I learn more about data mining and machine learning?
- What other topics are relevant to data science beyond predictive modeling?

What comes next?

- Where are data mining and machine learning going in the future?
- How do I learn more about data mining and machine learning?
- What other topics are relevant to data science beyond predictive modeling?
- What jobs out there value the skills I learned in STAT 4710? How do I get those jobs?

Where are data mining and ML heading?

Where are data mining and ML heading?

- Businesses increasingly data driven; data mining and ML will continue becoming increasingly common in marketing, finance, e-commerce, etc.

Where are data mining and ML heading?

- Businesses increasingly data driven; data mining and ML will continue becoming increasingly common in marketing, finance, e-commerce, etc.
- **Regression and tree-based methods** will continue to be used for general-purpose data mining; **deep learning** for images and natural language.

Where are data mining and ML heading?

- Businesses increasingly data driven; data mining and ML will continue becoming increasingly common in marketing, finance, e-commerce, etc.
- **Regression and tree-based methods** will continue to be used for general-purpose data mining; **deep learning** for images and natural language.
- Datasets are becoming increasingly bigger, so more emphasis will be placed on large-scale/cloud **computation** and parallelization.

Where are data mining and ML heading?

- Businesses increasingly data driven; data mining and ML will continue becoming increasingly common in marketing, finance, e-commerce, etc.
- **Regression and tree-based methods** will continue to be used for general-purpose data mining; **deep learning** for images and natural language.
- Datasets are becoming increasingly bigger, so more emphasis will be placed on large-scale/cloud **computation** and parallelization.
- As deep learning matures, more emphasis will be placed on understanding and interpretation, making it more safe, robust, and fair, developing **theory**.

Learning more about data mining and ML

Computation	Theory	Deep learning
<ul style="list-style-type: none">Python programming (CIS1920, STAT4770)	<ul style="list-style-type: none">Probability (STAT4300, ESE3010)Linear algebra (MATH2400, MATH3120)Calculus (MATH1140, MATH1150)	
<ul style="list-style-type: none">Large-scale computing (NETS2120, CIS5450)Databases (CIS4500)	<ul style="list-style-type: none">Mathy machine learning (CIS5200, ESE5450)Optimization (STAT 4810, CIS5150, ESE5040, ESE6050)	<ul style="list-style-type: none">Deep learning (CIS5220)Computer vision (CIS6800)NLP (CIS5300)

Offered next semester!

Beyond predictive modeling

Beyond predictive modeling

Not every problem in the world can be solved by machine learning!



Beyond predictive modeling

Not every problem in the world can be solved by machine learning!



Statistical inference versus prediction: understanding the world is different from predicting it. E.g., science is driven by understanding rather than prediction.

Beyond predictive modeling

Not every problem in the world can be solved by machine learning!



Statistical inference versus prediction: understanding the world is different from predicting it. E.g., science is driven by understanding rather than prediction.

Causal inference: Cause and effect is the ultimate question in science and policy. Decision makers (e.g. the FDA) need to estimate the effects of different policies.

Beyond predictive modeling

Not every problem in the world can be solved by machine learning!



Statistical inference versus prediction: understanding the world is different from predicting it. E.g., science is driven by understanding rather than prediction.

Causal inference: Cause and effect is the ultimate question in science and policy. Decision makers (e.g. the FDA) need to estimate the effects of different policies.

Quantifying uncertainty: If a neural network predicts there is no pedestrian ahead with probability 0.99, what does this mean? Can we make a statistical guarantee?

Beyond predictive modeling

Not every problem in the world can be solved by machine learning!



Statistical inference versus prediction: understanding the world is different from predicting it. E.g., science is driven by understanding rather than prediction.

Causal inference: Cause and effect is the ultimate question in science and policy. Decision makers (e.g. the FDA) need to estimate the effects of different policies.

Quantifying uncertainty: If a neural network predicts there is no pedestrian ahead with probability 0.99, what does this mean? Can we make a statistical guarantee?

Robustness and safety: Will the self-driving car recognize a pedestrian if she is holding an umbrella? In what sense can we ensure the robustness of an algorithm?

Beyond predictive modeling

Not every problem in the world can be solved by machine learning!



Statistical inference versus prediction: understanding the world is different from predicting it. E.g., science is driven by understanding rather than prediction.

Causal inference: Cause and effect is the ultimate question in science and policy. Decision makers (e.g. the FDA) need to estimate the effects of different policies.

Quantifying uncertainty: If a neural network predicts there is no pedestrian ahead with probability 0.99, what does this mean? Can we make a statistical guarantee?

Robustness and safety: Will the self-driving car recognize a pedestrian if she is holding an umbrella? In what sense can we ensure the robustness of an algorithm?

Fairness: In what sense can a prediction rule be considered fair? How can we assure that our predictive rules live up to this standard?

Data science jobs



Data Scientist: The Sexiest Job of the 21st Century

Meet the people who can coax treasure out of messy, unstructured data. by Thomas H. Davenport and D.J. Patil

From the Magazine (October 2012)



glassdoor

50 Best Jobs in America for 2022

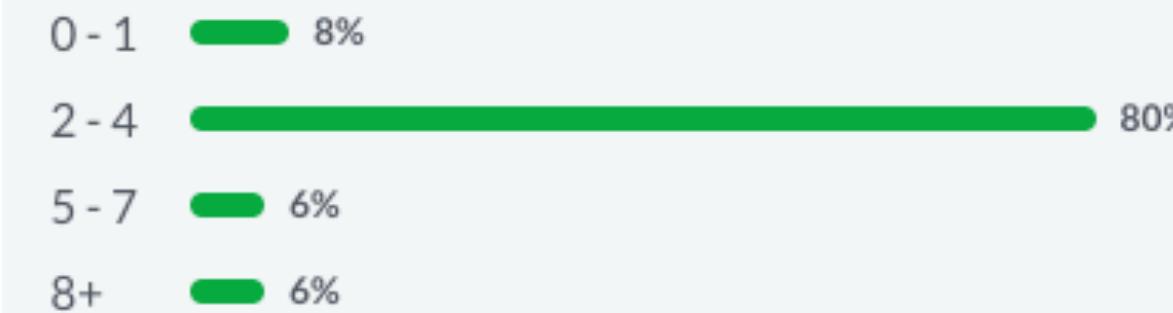
Job Title	Median Base Salary	Job Satisfaction	Job Openings
#1 Enterprise Architect	\$144,997	4.1/5	14,021
#2 Full Stack Engineer	\$101,794	4.3/5	11,252
#3 Data Scientist	\$120,000	4.1/5	10,071
#4 Devops Engineer	\$120,095	4.2/5	8,548
#5 Strategy Manager	\$140,000	4.2/5	6,977

What does a Data Scientist do?

Data scientists utilize their analytical, statistical, and programming skills to collect, analyze, and interpret large data sets. They then use this information to develop data-driven solutions to difficult business challenges. Data scientists commonly have a bachelor's degree in statistics, math, computer science, or economics. Data scientists have a wide range of technical...

[Read More](#)

Average Years of Experience



Common Skill Sets

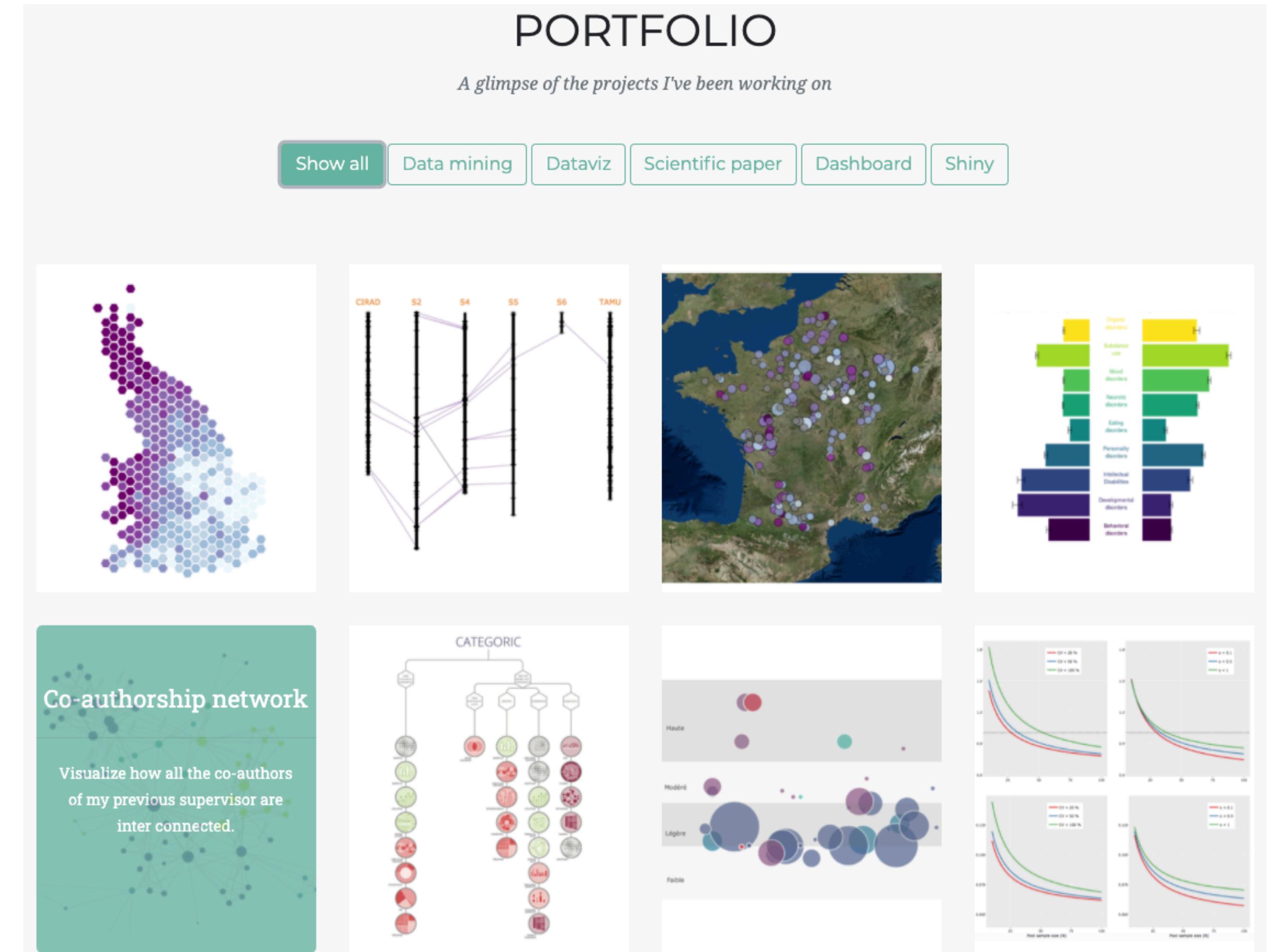
- Machine Learning Statistics
- Python Natural Language Processing
- Hadoop SPARK Algorithms
- SQL Programming Languages

How to get a data science job?

- Learn the skills through classes or on your own.
- Build your skills through data science projects.
- Share your work by posting code on Github and making a portfolio of your projects.
- Apply to internships to gain data science experience.

How to get a data science job?

- Learn the skills through classes or on your own.
- Build your skills through data science projects.
- Share your work by posting code on Github and making a portfolio of your projects.
- Apply to internships to gain data science experience.



<https://www.yan-holtz.com/>