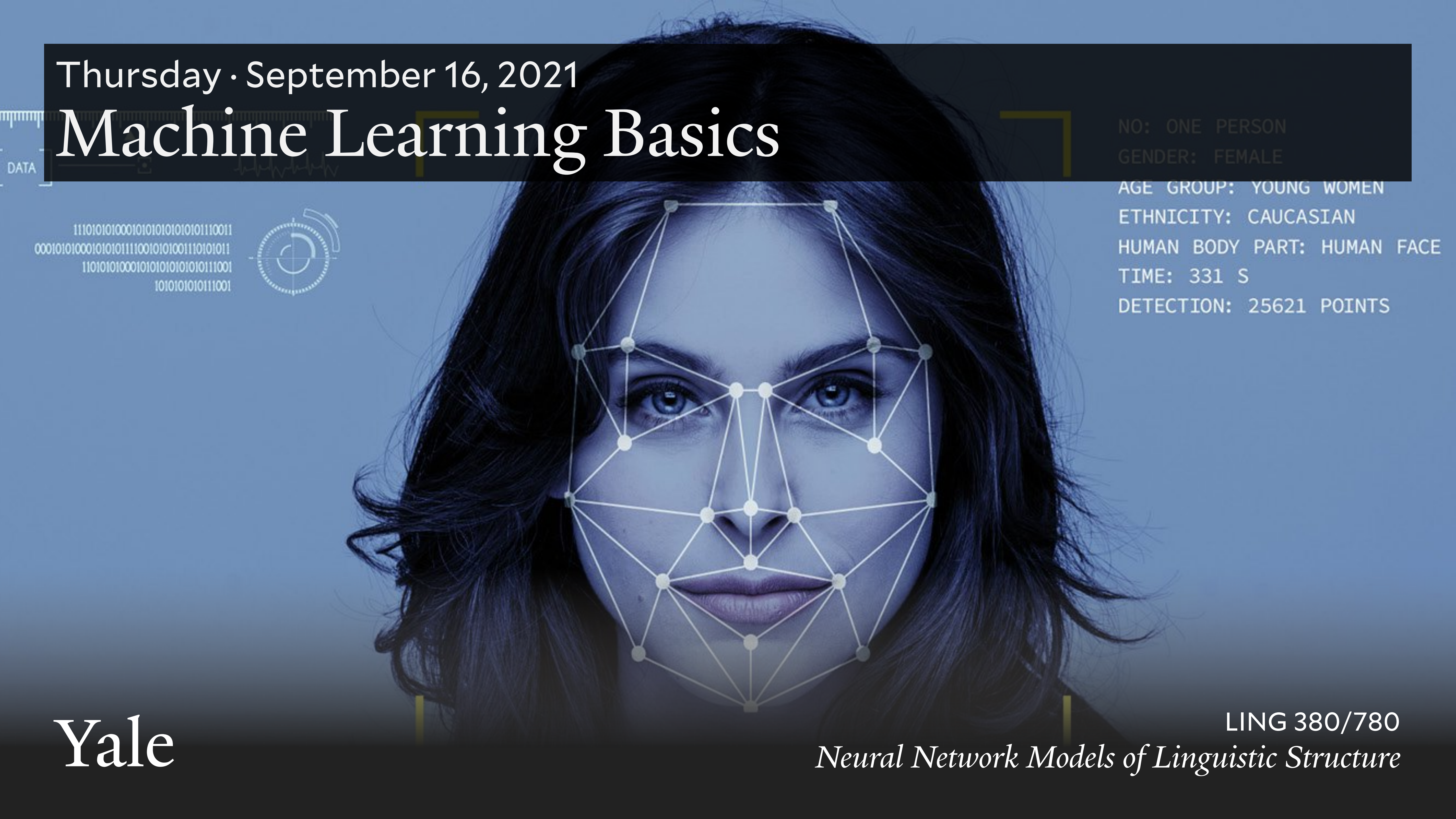
Thursday · September 16, 2021

Machine Learning Basics



Yale

LING 380/780

Neural Network Models of Linguistic Structure

What is Learning?

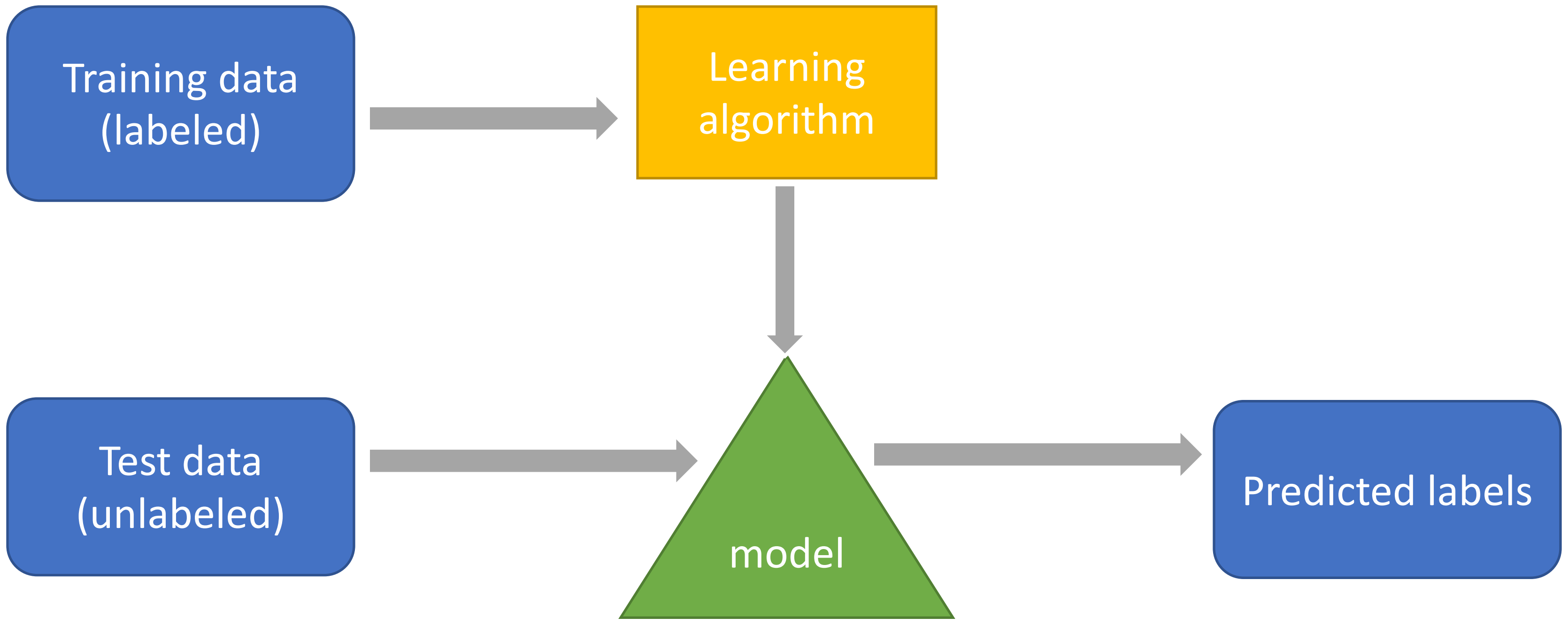
- Use the past to predict the future

	Inside Out	Good Will Hunting	Mean Girls	Terminator	Titanic	Warrior
Tina Fey	3	1	5	1	?	1
Helen Mirren	2	?	?	2	5	1
Sylvester Stallone	1	3	1	4	2	5
Tom Hanks	?	3	1	?	4	3
George Clooney	2	2	1	3	1	4

Types of learning problems

- **Regression:** predict a numerical value
- **Binary classification:** predict a yes-no response
- **Multiclass classification:** predict membership into one of a number of classes
- **Ranking:** order a set of objects with respect to relevance

Framework for learning



Patterns, Learning, and Inductive Reasoning

- A learner needs to find patterns in the world.
- But the learner has an **inductive bias** that tells them what patterns are possible.
- The learner's task: to find the **best possible description** of the world around them, within the constraints of the learner's inductive bias.

Ingredients of Machine Learning

- **Architecture**: The learner's inductive bias.
- **Loss Function:** A measure of how bad a model is.
- **Optimization Algorithm:** An algorithm that tries to minimize how bad the model is.

Binary Classification

Input	Label	Input	Label
1, 2, 4	True	4, 8, 16	???
2, 4, 8	True	16, 8, 4	???
16, 32, 64	True	3, 6, 12	???
2, 1, 4	False	1, 2, 3	???
3, 2, 1	False	0, 0, 0	???

Binary Classification

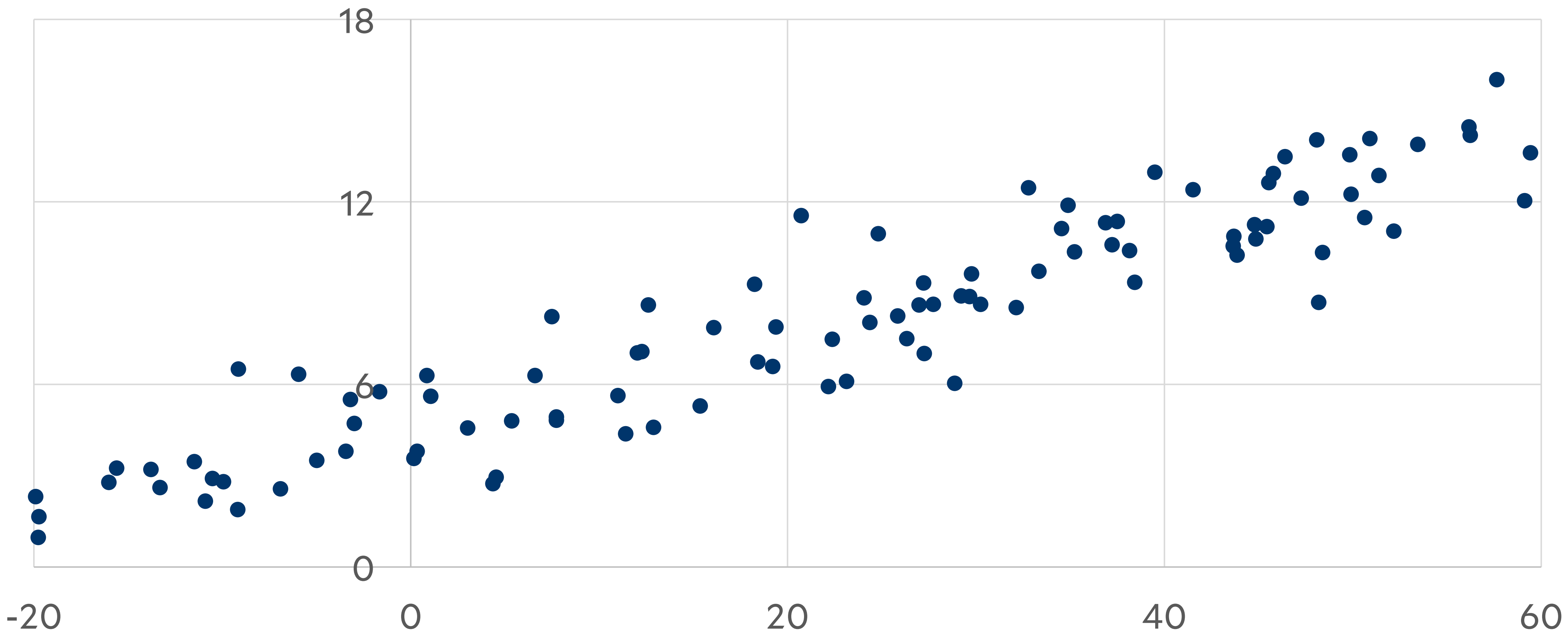
Input	Label	Input	Label
1, 2, 4	True	4, 8, 16	True
2, 4, 8	True	16, 8, 4	False
16, 32, 64	True	3, 6, 12	True
2, 1, 4	False	1, 2, 3	True
3, 2, 1	False	0, 0, 0	False
5, 6, 7	???		
5, 10, 20	???		
3, 6, 9	???		


Gavagai!
Gavagai

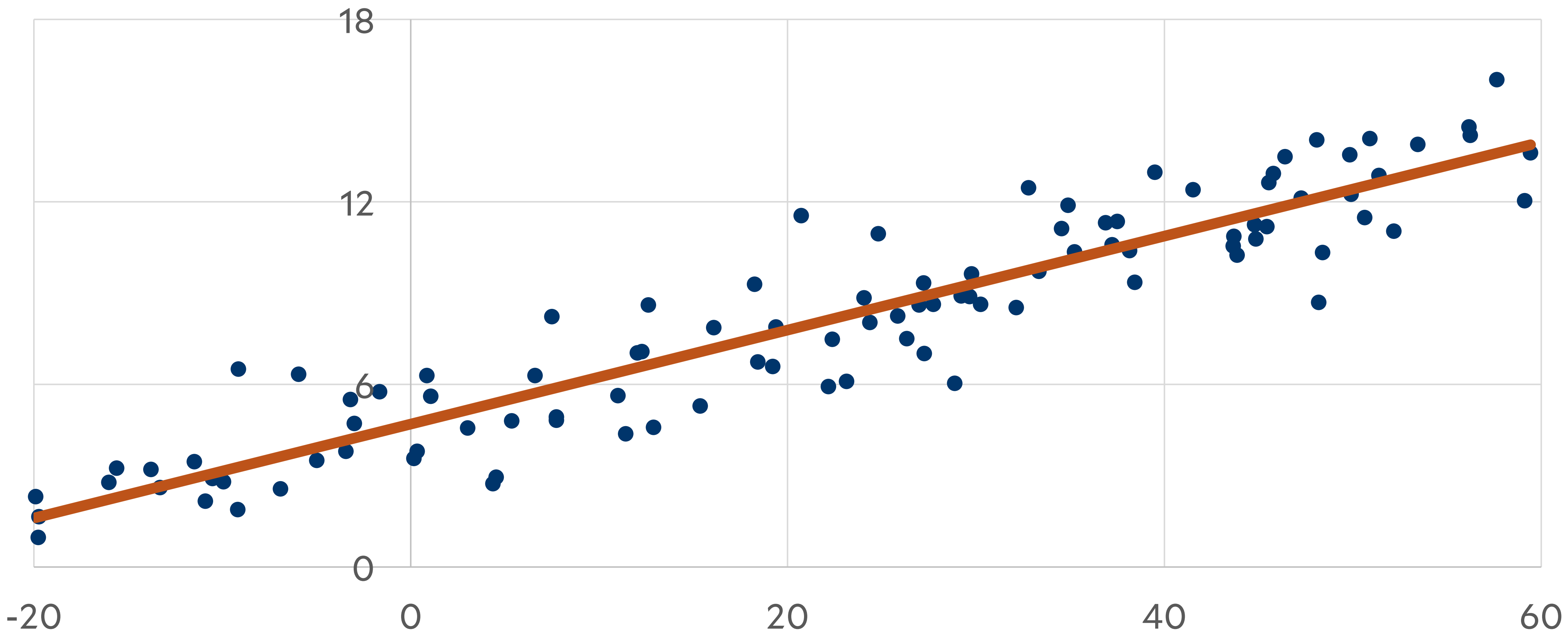

Gavagai?
Gavagai!


Gavagai!
Gavagai?

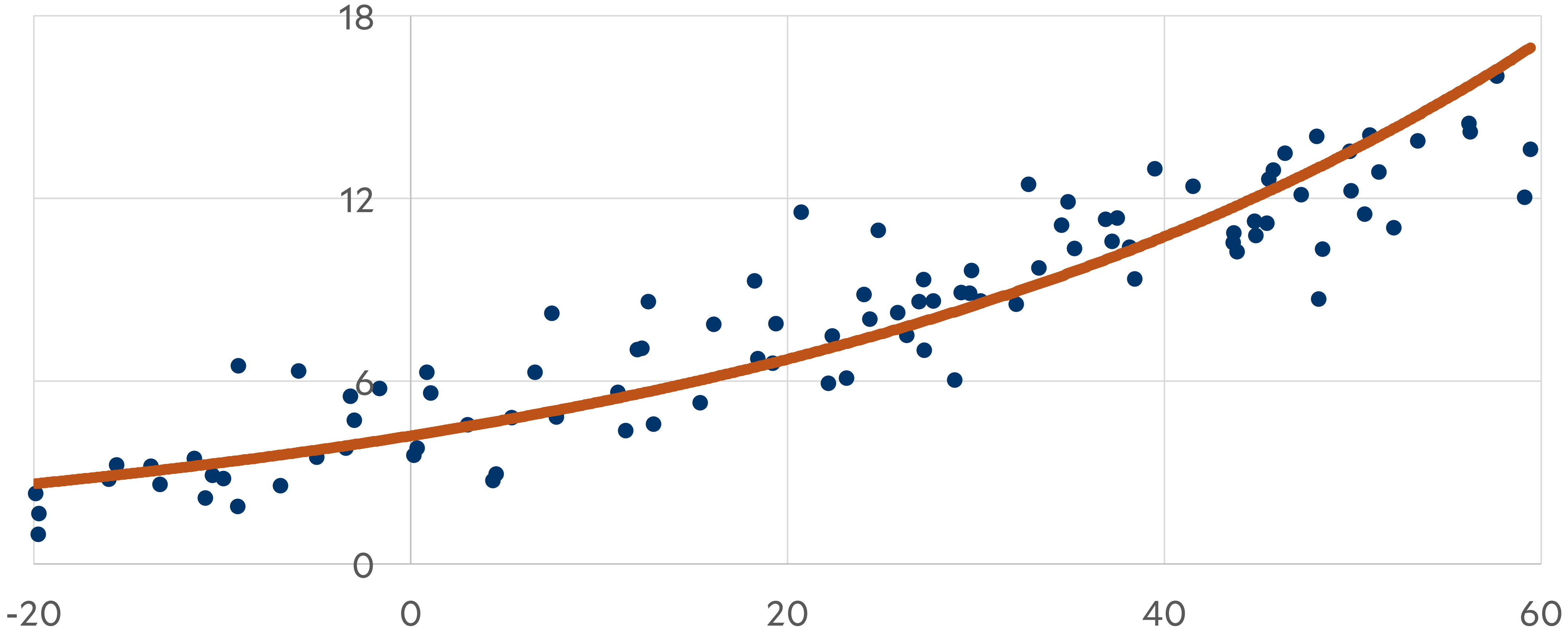
What's the Pattern?



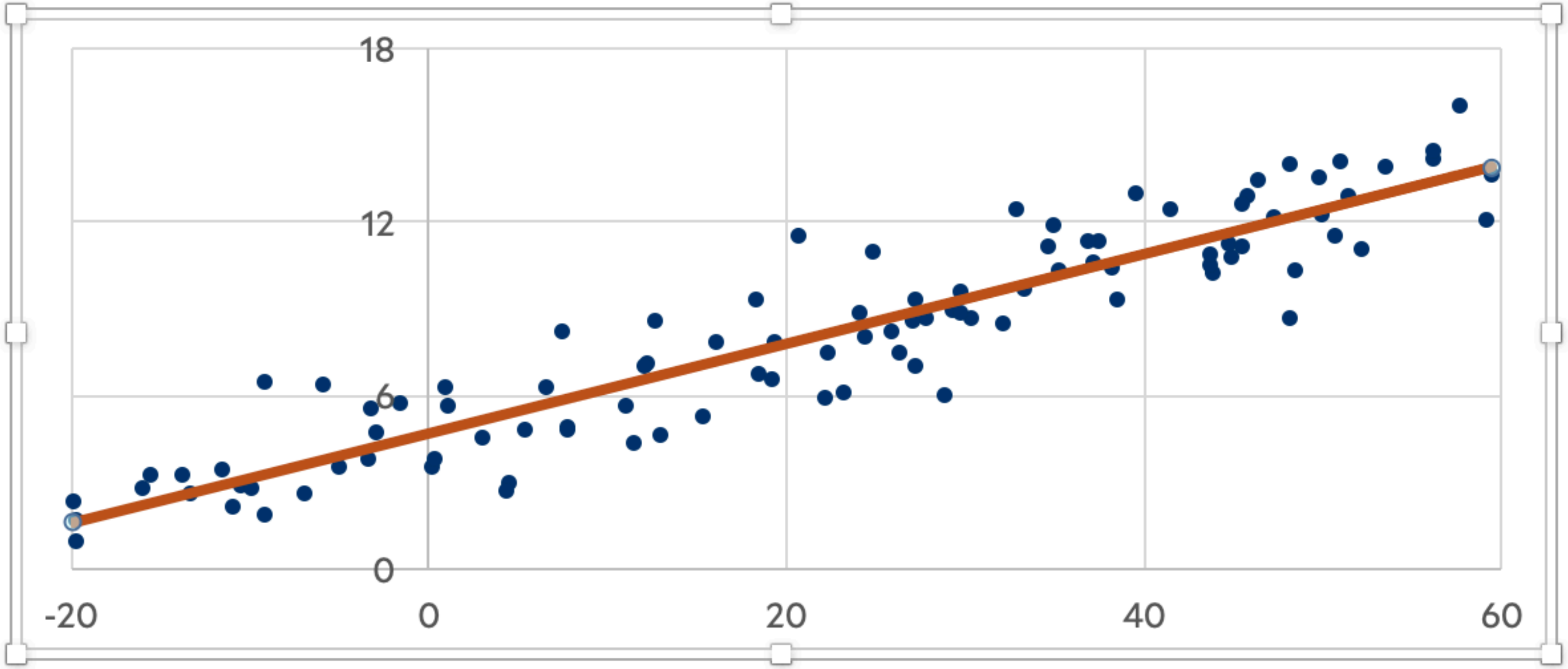
What's the Pattern?

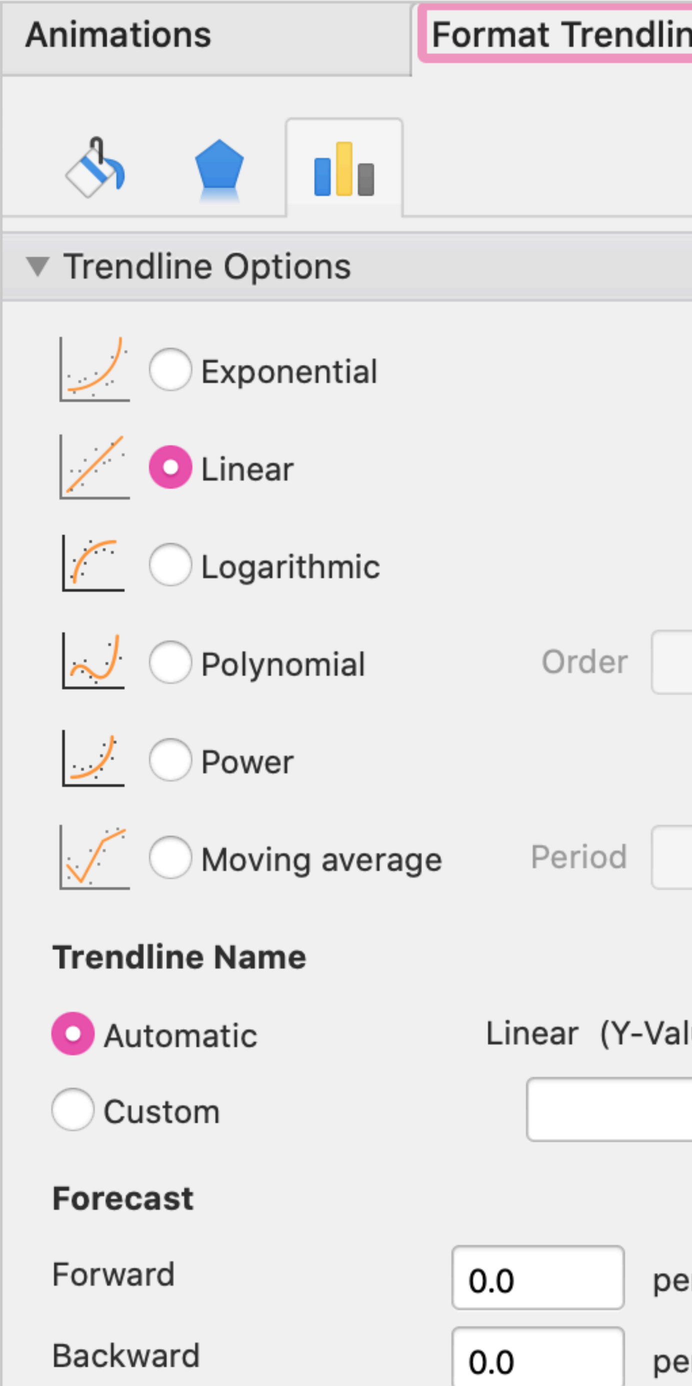


What's the Pattern?



What's the Pattern?





Model Architectures

A model architecture is a family of parameterized functions of the form

$$\hat{\boldsymbol{y}} = \hat{f}(\boldsymbol{x}; \boldsymbol{\theta})$$

where $\boldsymbol{\theta}$ is a vector of parameters.

Example: Social Science

- Hamermesh and Parker (2004): Do good-looking instructors get better course evaluations?
- Create a model that predicts course evaluation scores from course feature vectors.
- The model learns from UT Austin course evaluations.

Example: Social Science

- Feature vectors for courses: $\boldsymbol{x} \in \mathbb{R}^7$, where
  - x_1: “beauty score” from 0 (ugly) to 1 (beautiful)
  - x_2: 1 if instructor is female, 0 if male
  - x_3: 1 if instructor is non-white, 0 if white
  - x_4: 1 if instructor is a native English speaker, 0 otherwise
  - x_5: 1 if instructor is tenure-track, 0 otherwise
  - x_6: 1 if the course is 100/200, 0 if 300/400
  - x_7: 1 if course is only one credit, 0 otherwise

Example: Social Science

Linear model architecture:

$$\hat{y} = \hat{f}(\boldsymbol{x}; \boldsymbol{\theta}) = \boldsymbol{w}^\top \boldsymbol{x} + b$$

where

- $\boldsymbol{x} \in \mathbb{R}^7$ is the feature vector for a course
- $\hat{y} \in [0, 1]$ is the predicted course evaluation
- parameters are $\boldsymbol{\theta} = \begin{bmatrix} \boldsymbol{w} \\ b \end{bmatrix}$.

Example: Social Science

- Learned model parameters:
  - $w_1 = 0.275$ (beautiful?)
  - $w_2 = -0.239$ (female?)
  - $w_3 = -0.249$ (non-white?)
  - $w_4 = -0.253$ (native English speaker?)
  - $w_5 = -0.136$ (tenure-track?)
  - $w_6 = -0.045$ (100/200?)
  - $w_7 = 0.687$ (one-credit?)

Example: SGNS

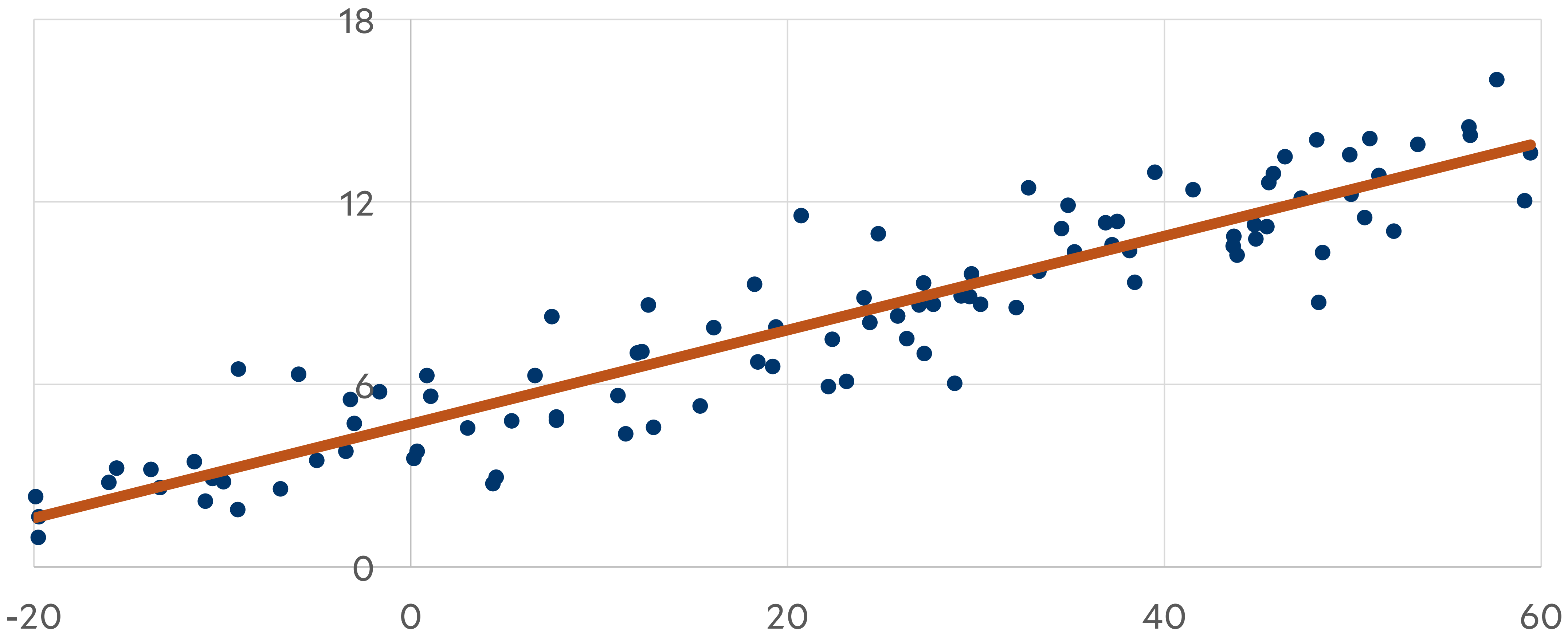
What is the architecture for SGNS?

$$\hat{y} = \hat{f}(w, c; \boldsymbol{\theta}) = \sigma(\langle c \rangle^\top [\![w]\!])$$

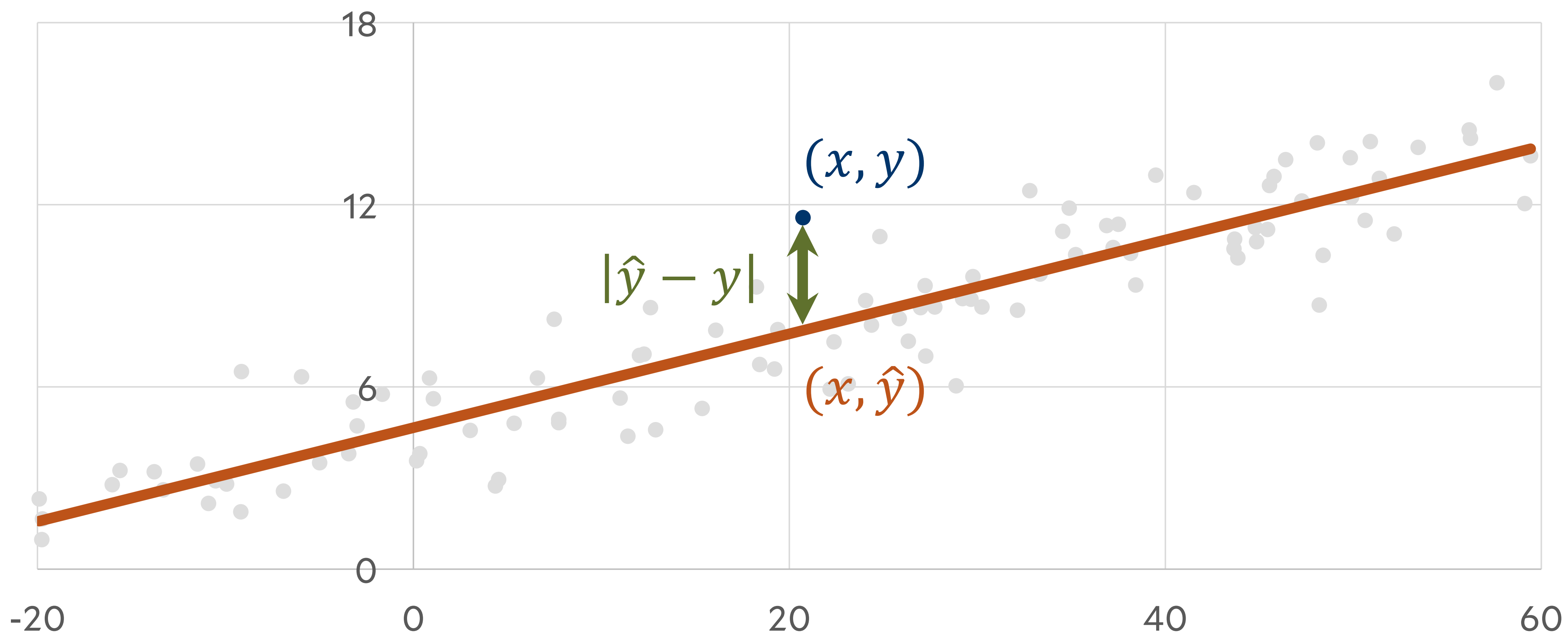
where

- $w \in \mathbb{V}$ is a target word and $c \in \mathbb{V}$ is a context
- $\hat{y} \in (0, 1)$ is the probability that w and c occur together
- $\boldsymbol{\theta}^\top = [\langle c_1 \rangle^\top \quad \langle c_2 \rangle^\top \quad \cdots \quad \langle c_n \rangle^\top \quad [\![w_1]\!]^\top \quad [\![w_2]\!]^\top \quad \cdots \quad [\![w_n]\!]^\top]$

How Bad Is My Model?



How Bad Is My Model?



Loss Functions

Let $\hat{f}(\cdot; \boldsymbol{\theta}): \mathbb{A} \to \mathbb{B}$ be an architecture that predicts $\widehat{\boldsymbol{y}} = \hat{f}(\boldsymbol{x}; \boldsymbol{\theta}) \in \mathbb{B}$ from input $\boldsymbol{x} \in \mathbb{A}$.

A loss function is a function $L: \mathbb{B} \times \mathbb{B} \to \mathbb{R}$ such that $L(\widehat{\boldsymbol{y}}, \boldsymbol{y})$ measures how bad the prediction $\widehat{\boldsymbol{y}}$ is for the true value $\boldsymbol{y}$.

Loss Functions

Mean Squared Error Loss Function (Linear Regression)

$$L_{\mathrm{MSE}}(\hat{y}, y) = (\hat{y} - y)^2$$

Binary Cross-Entropy Loss Function (SGNS)

$$L_{\mathrm{CE}}(\hat{y}, y) = -y \ln(\hat{y}) - (1 - y)\ln(1 - \hat{y})$$

Loss Minimization

Let $\hat{f}(\cdot, \boldsymbol{\theta}): \mathbb{A} \to \mathbb{B}$ be a model architecture. We **train** $\hat{f}(\cdot, \boldsymbol{\theta})$ on a dataset $\mathbb{D} \subseteq \mathbb{A} \times \mathbb{B}$ by finding the parameters $\boldsymbol{\theta}^*$ that minimize average loss:

$$\boldsymbol{\theta}^* = \underset{\boldsymbol{\theta}}{\operatorname{argmin}} \overbrace{\sum_{(\boldsymbol{x},\boldsymbol{y})\in\mathbb{D}} L(\hat{f}(\boldsymbol{x}; \boldsymbol{\theta}), \boldsymbol{y})}^{\text{Objective } (\mathcal{L})}$$

Minimizing average loss is the same as minimizing total loss!

Example: Linear Regression

Linear Regression Model

$$\hat{f}(\boldsymbol{x}; \boldsymbol{w}, b) = \boldsymbol{w}^\top \boldsymbol{x} + b$$

Linear Regression Objective

$$\mathcal{L} = \sum_{(\boldsymbol{x}, y) \in \mathbb{D}} L_{\text{MSE}}(\hat{f}(\boldsymbol{x}; \boldsymbol{w}, b), y)$$

Example: Linear Regression

Linear Regression Model

$$\hat{f}(\boldsymbol{x}; \boldsymbol{w}, b) = \boldsymbol{w}^\top \boldsymbol{x} + b$$

Linear Regression Objective

$$\mathcal{L} = \sum_{(\boldsymbol{x}, y) \in \mathbb{D}} L_{\mathrm{MSE}}(\boldsymbol{w}^\top \boldsymbol{x} + b\,, y)$$

Example: Linear Regression

Linear Regression Model

$$\hat{f}(\boldsymbol{x}; \boldsymbol{w}, b) = \boldsymbol{w}^\top \boldsymbol{x} + b$$

Linear Regression Objective

$$\mathcal{L} = \sum_{(\boldsymbol{x}, y) \in \mathbb{D}} (\boldsymbol{w}^\top \boldsymbol{x} + b - y)^2$$

Example: Linear Regression

Linear Regression Model

$$\hat{f}(\boldsymbol{x}; \boldsymbol{w}, b) = \boldsymbol{w}^\top \boldsymbol{x} + b$$

Linear Regression Minimization Problem

$$\boldsymbol{w}^*, b^* = \underset{\boldsymbol{w},b}{\operatorname{argmin}}\, \mathcal{L} = \underset{\boldsymbol{w},b}{\operatorname{argmin}} \sum_{(\boldsymbol{x},y)\in\mathbb{D}} (\boldsymbol{w}^\top \boldsymbol{x} + b - y)^2$$

Example: SGNS

SGNS Model

$$\hat{f}(w, c; \langle\cdot\rangle, [\![\cdot]\!]) = \sigma(\langle c\rangle^\top [\![w]\!])$$

SGNS Objective

$$\mathcal{L} = \sum_{(w,c,y)\in\mathbb{D}} L_{\mathrm{CE}}(\sigma(\langle c\rangle^\top [\![w]\!])\,, y)$$

Example: SGNS

$$\mathcal{L} = \sum_{(w,c,y)\in\mathbb{D}} L_{\mathrm{CE}}(\sigma(\langle c\rangle^\top [\![w]\!]), y)$$

$$= \sum_{(w,c,y)\in\mathbb{D}} -\cancel{y}\ln(\sigma(\langle c\rangle^\top [\![w]\!])) - \cancel{(1-y)\ln(1-\sigma(\langle c\rangle^\top [\![w]\!]))}$$

$y = 1$

$$= \left(\sum_{(w,c,1)\in\mathbb{D}} -\ln(\sigma(\langle c\rangle^\top [\![w]\!]))\right) + \left(\sum_{(w,c,0)\in\mathbb{D}} -\ln(1-\sigma(\langle c\rangle^\top [\![w]\!]))\right)$$

Example: SGNS

$$\mathcal{L} = \sum_{(w,c,y)\in\mathbb{D}} L_{\mathrm{CE}}(\sigma(\langle c\rangle^\top [\![w]\!]), y)$$

$$= \sum_{(w,c,y)\in\mathbb{D}} -\cancel{y\ln(\sigma(\langle c\rangle^\top [\![w]\!]))} - \cancel{(1-y)}\ln(1-\sigma(\langle c\rangle^\top [\![w]\!]))$$

$y = 0$

$$= \left(\sum_{(w,c,1)\in\mathbb{D}} -\ln(\sigma(\langle c\rangle^\top [\![w]\!]))\right) + \left(\sum_{(w,c,0)\in\mathbb{D}} -\ln(1-\sigma(\langle c\rangle^\top [\![w]\!]))\right)$$

Example: SGNS

$$\mathcal{L} = \sum_{(w,c,y)\in\mathbb{D}} L_{\mathrm{CE}}(\sigma(\langle c\rangle^\top [\![w]\!]), y)$$

$$= \sum_{(w,c,y)\in\mathbb{D}} -y\ln(\sigma(\langle c\rangle^\top [\![w]\!])) - (1-y)\ln(1-\sigma(\langle c\rangle^\top [\![w]\!]))$$

$$= \left(\sum_{(w,c,1)\in\mathbb{D}} -\ln(\sigma(\langle c\rangle^\top [\![w]\!]))\right) + \left(\sum_{(w,c,0)\in\mathbb{D}} -\ln(1-\sigma(\langle c\rangle^\top [\![w]\!]))\right)$$

Optimization Algorithm

An optimization algorithm is any algorithm that can minimize the objective.

Given input $\mathbb{D}$ and model architecture $\hat{f}(\cdot; \boldsymbol{\theta})$, return:

$$\underset{\boldsymbol{\theta}}{\operatorname{argmin}} \sum_{(\boldsymbol{x},\boldsymbol{y}) \in \mathbb{D}} L(\hat{f}(\boldsymbol{x}; \boldsymbol{\theta}), \boldsymbol{y})$$

Optimization for Linear Regression

Use first derivative test:

$$\nabla_{\boldsymbol{\theta}} \mathcal{L} = \mathbf{0}$$

where $\nabla_{\boldsymbol{\theta}} \mathcal{L}$ is the **gradient of $\mathcal{L}$**:

$$\nabla_{\theta} \mathcal{L} = \begin{bmatrix} \partial \mathcal{L} / \partial \theta_1 \\ \partial \mathcal{L} / \partial \theta_2 \\ \vdots \\ \partial \mathcal{L} / \partial \theta_n \end{bmatrix}$$

Optimization for Linear Regression

For all i,

$$\frac{\partial \mathcal{L}}{\partial w_i} = 0$$

$$\frac{\partial \mathcal{L}}{\partial b} = 0$$

Solve for w_i and b!

From the first equation we get

$$
\begin{aligned}
0 &= 2 \sum_{(x,y)\in\mathbb{D}} x(ax + b - y) \\
&= 2 \sum_{(x,y)\in\mathbb{D}} \left(ax^2 + \left(\frac{x}{|\mathbb{D}|} \sum_{(x',y')\in\mathbb{D}} (y' - ax') \right) - yx \right) \\
&= 2a \left(\sum_{(x,y)\in\mathbb{D}} x^2 - \frac{x}{|\mathbb{D}|} \sum_{(x',y')\in\mathbb{D}} x' \right) + \sum_{(x,y)\in\mathbb{D}} \left(-yx + \frac{x}{|\mathbb{D}|} \sum_{(x',y')\in\mathbb{D}} y' \right),
\end{aligned}
$$

hence

$$
a = \frac{\sum_{(x,y)\in\mathbb{D}} \left(yx - \frac{x}{|\mathbb{D}|} \sum_{(x',y')\in\mathbb{D}} y' \right)}{2\left(\sum_{(x,y)\in\mathbb{D}} x^2 - \frac{x}{|\mathbb{D}|} \sum_{(x',y')\in\mathbb{D}} x' \right)}
$$

and

$$
b = \frac{1}{|\mathbb{D}|} \sum_{(x,y)\in\mathbb{D}} \left(y - x \frac{\sum_{(x'',y'')\in\mathbb{D}} \left(y''x'' - \frac{x''}{|\mathbb{D}|} \sum_{(x',y')\in\mathbb{D}} y' \right)}{2\left(\sum_{(x'',y'')\in\mathbb{D}} x''^2 - \frac{x''}{|\mathbb{D}|} \sum_{(x',y')\in\mathbb{D}} x' \right)} \right).
$$