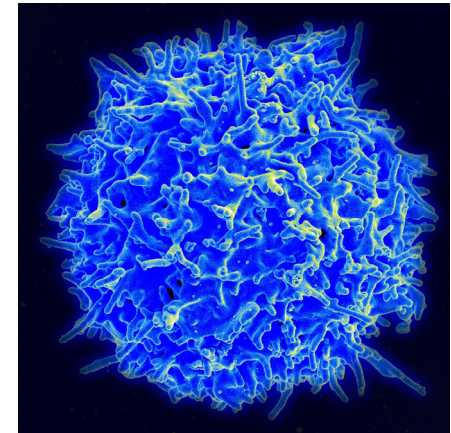


Human T cell expression data

- The matrix contains **47 expression samples** from Lukk et al, Nature Biotechnology 2010
- All samples are **from T cells in different individuals**
- Only the **top 3000 genes** with the largest variability **were used**
- The value is **log2 of gene's expression level** in a given sample as measured by the microarray technology

a T cell



A global map of human gene expression

Margus Lukk, Misha Kapushesky, Janne Nikkilä, Helen Parkinson, Angela Goncalves, Wolfgang Huber, Esko Ukkonen & Alvis Brazma

Affiliations | Corresponding author

Nature Biotechnology **28**, 322–324 (2010) | doi:10.1038/nbt0410-322

Although there is only one human genome sequence, different genes are expressed in many different cell types and tissues, as well as in different developmental stages or diseases. The structure of this 'expression space' is still largely unknown, as most transcriptomics experiments focus on sampling small regions. We have constructed a global gene expression map by integrating microarray data from 5,372 human samples representing 369 different cell and tissue types, disease states and cell lines. These have been compiled in an online resource (<http://www.ebi.ac.uk/gxa/array/U133A>) that allows the user to search for a gene of interest and

“Let’s Make a Deal” show with Monty Hall aired on NBC/ABC 1963-1986





Matlab exercise #1: “Wheel of Fortune”

- Each group gets a pair of genes that are known to be correlated.
- Each group also gets a random pair of genes selected by the “Wheel of Fortune”. They may or may not be correlated
- Download (log-transformed) expression_table.mat
- Run command fitlm(x,y) on assigned and random pairs
- Record β_0 , β_1 , R^2 , P-value of the slope β_1 and write them on the blackboard
- Validate Matlab result for R^2 using your own calculations
- Look up gene names (see gene_description in your workspace) and write down a brief description of biological functions of genes. Does their correlation make biological sense?

**Correlated pairs
plausible biological connection based
on short description**

1, 6 g1=1994; g2=188;

2, g1=2872; g2=1269;

3, g1=1321; g2=10;

4, g1= 886; g2=819;

5, g1=2138; g2=1364;

no obvious biological common function

**g1=1+floor(rand.*3000); g2=1+floor(rand.*3000);
disp([g1, g2])**

Random pairs

```
>> g1=floor(3000.*rand)+1; g2=floor(3000.*rand)+1;  
disp([g1,g2]);
```

```
>> g1=floor(3000.*rand)+1; g2=floor(3000.*rand)+1;  
disp([g1,g2]);
```

```
>> g1=floor(3000.*rand)+1; g2=floor(3000.*rand)+1;  
disp([g1,g2]);
```

```
>> g1=floor(3000.*rand)+1; g2=floor(3000.*rand)+1;  
disp([g1,g2]);
```

Matlab code

- load expression_table.mat
- g1=2907; g2=288;
- x=exp_t(g1,:)' ; y=exp_t(g2,:)' ;
- figure; plot(x,y,'ko');
- lm=fitlm(x,y)
- y_fit=lm.Fitted;
- hold on; plot(x,lm.Fitted,'r-');
- SST=sum((y-mean(y)).^2);
- SSR=sum((y_fit-mean(y)).^2);
- SSE=sum((y-y_fit).^2);
- R2=SSR./SST
- disp([gene_names(g1), gene_names(g2)]);
- disp(gene_description(g1)); disp(gene_description (g2));

Credit: XKCD
comics

WHY ARE THERE SLAVES IN THE BIBLE

WHY DO TWINS HAVE DIFFERENT FINGERPRINTS
WHY ARE AMERICANS AFRAID OF DRAGONS

QUESTIONS

FOUND IN GOOGLE AUTOCOMPLETE

WHY IS HTTPS CROSSED OUT IN RED
WHY IS THERE A LINE THROUGH HTTPS
WHY IS THERE A RED LINE THROUGH HTTPS ON FACEBOOK
WHY IS HTTPS IMPORTANT

WHY AREN'T MY
ARMS GROWING



WHY ARE THERE WEEKS
WHY DO I FEEL DIZZY

WHY AREN'T ECONOMISTS RICH

WHY DO AMERICANS CALL IT SOCCER

WHY ARE MY EARS RINGING

WHY ARE THERE SO MANY AVENGERS

WHY ARE THE AVENGERS FIGHTING THE X MEN

WHY IS WOLVERINE NOT IN THE AVENGERS

WHY ARE THERE ANTS IN MY LAPTOP

WHY IS EARTH TILTED

WHY IS SPACE BLACK

WHY IS OUTER SPACE SO COLD

WHY ARE THERE PYRAMIDS ON THE MOON

WHY IS NASA SHUTTING DOWN

WHY ARE THERE MALE AND FEMALE BIKES

WHY ARE THERE TINY SPIDERS IN MY HOUSE

WHY DO SPIDERS COME INSIDE

WHY ARE THERE HUGE SPIDERS IN MY HOUSE

WHY ARE THERE LOTS OF SPIDERS IN MY HOUSE

WHY ARE THERE SPIDERS IN MY ROOM

WHY ARE THERE SO MANY SPIDERS IN MY ROOM

WHY DO SPIDER BITES ITCH

WHY IS DYING SO SCARY

WHY IS THERE NO GPS IN LAPTOPS

WHY DO KNEES CLICK

WHY AREN'T THERE E GRADES

WHY IS ISOLATION BAD

WHY DO BOYS LIKE ME

WHY DON'T BOYS LIKE ME

WHY IS THERE ALWAYS A JAVA UPDATE

WHY ARE THERE RED DOTS ON MY THIGHS

WHY IS LYING GOOD

WHY ARE THERE
GHOSTS



WHY IS THERE AN OWL IN MY BACKYARD

WHY IS THERE AN OWL OUTSIDE MY WINDOW

WHY IS THERE AN OWL ON THE DOLLAR BILL

WHY DO OWLS ATTACK PEOPLE

WHY ARE AK 47s SO EXPENSIVE

WHY ARE THERE HELICOPTERS CIRCLING MY HOUSE

WHY ARE THERE GODS

WHY ARE THERE TWO SPOCKS

WHY IS MT VESUVIUS THERE

WHY DO THEY SAY T MINUS

WHY ARE THERE OBELISKS

WHY ARE WRESTLERS ALWAYS WET

WHY ARE OCEANS BECOMING MORE ACIDIC

WHY IS ARWEN DYING

WHY AREN'T MY QUAIL LAYING EGGS

WHY AREN'T MY QUAIL EGGS HATCHING

WHY AREN'T THERE ANY FOREIGN MILITARY BASES IN AMERICA

WHY ARE CIGARETTES LEGAL

WHY ARE THERE DUCKS IN MY POOL

WHY IS JESUS WHITE

WHY IS THERE LIQUID IN MY EAR

WHY DO Q TIPS FEEL GOOD

WHY DO GOOD PEOPLE DIE

WHY AREN'T
THERE GUNS IN
HARRY POTTER



WHY ARE ULTRASOUNDS IMPORTANT

WHY ARE ULTRASOUND MACHINES EXPENSIVE

WHY IS STEALING WRONG

WHY ARE THERE FIREWORKS

WHY ARE THERE DOGS AFRAID OF FIREWORKS
WHY IS THERE NO KING IN ENGLAND

WHY DO WHALES JUMP
WHY ARE WITCHES GREEN
WHY ARE THERE MIRRORS ABOVE BEDS

WHY DO I SAY UH

WHY IS SEA SALT BETTER

WHY ARE THERE TREES IN THE MIDDLE OF FIELDS

WHY IS THERE NOT A POKEMON MMO

WHY IS THERE LAUGHING IN TV SHOWS

WHY ARE THERE DOORS ON THE FREEWAY

WHY ARE THERE SO MANY SVCHOST.EXE RUNNING

WHY AREN'T THERE ANY COUNTRIES IN ANTARCTICA

WHY ARE THERE SCARY SOUNDS IN MINECRAFT

WHY IS THERE KICKING IN MY STOMACH

WHY ARE THERE TWO SLASHES AFTER HTTP

WHY ARE THERE CELEBRITIES

WHY DO SNAKES EXIST

WHY DO OYSTERS HAVE PEARLS

WHY ARE DUCKS CALLED DUCKS

WHY DO THEY CALL IT THE CLAP

WHY ARE KYLE AND CARTMAN FRIENDS

WHY IS THERE AN ARROW ON AANG'S HEAD

WHY ARE TEXT MESSAGES BLUE

WHY ARE THERE MUSTACHES ON CLOTHES

WHY ARE THERE MUSTACHES ON CARS

WHY ARE THERE MUSTACHES EVERYWHERE

WHY ARE THERE SO MANY BIRDS IN OHIO

WHY IS THERE SO MUCH RAIN IN OHIO

WHY IS OHIO WEATHER SO WEIRD

WHY ARE THERE BRIDESMAIDS

WHY DO DYING PEOPLE REACH UP

WHY AREN'T THERE VARICOSE ARTERIES

WHY ARE OLD KINGDOMS DIFFERENT

WHY ARE THERE SQUIRRELS

WHY IS PROGRAMMING SO HARD
WHY IS THERE A 0 OHM RESISTOR
WHY DO AMERICANS HATE SOCCER
WHY DO RHYMES SOUND GOOD
WHY DO TREES DIE
WHY IS THERE NO SOUND ON CNN
WHY AREN'T POKEMON REAL
WHY AREN'T BULLETS SHARP
WHY DO DREAMS SEEM SO REAL



WHY IS SEX
SO IMPORTANT



Multiple Linear Regression

(Chapters 12-13 in
Montgomery, Runger)

12-1: Multiple Linear Regression Model

12-1.1 Introduction

- Many applications of regression analysis involve situations in which there are more than one regressor variable X_k used to predict Y .
- A regression model then is called a **multiple regression model**.

Multiple Linear Regression Model

$$Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \dots + \beta_k x_k + \varepsilon$$

One can also use powers and products of other variables or even non-linear functions like $\exp(x_i)$ or $\log(x_i)$

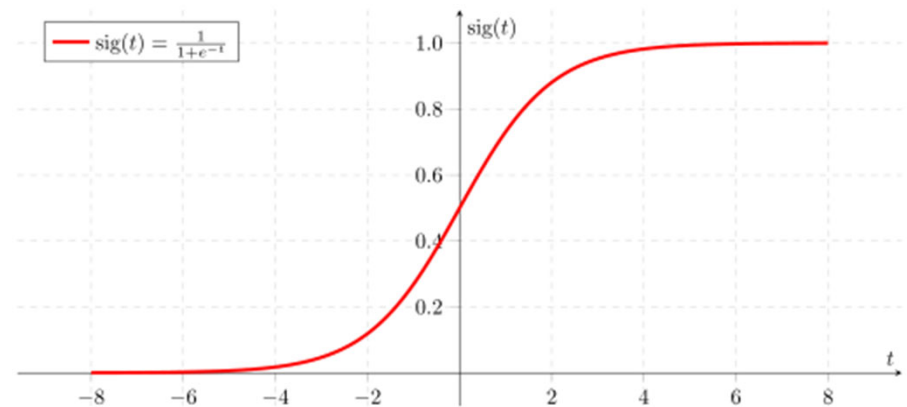
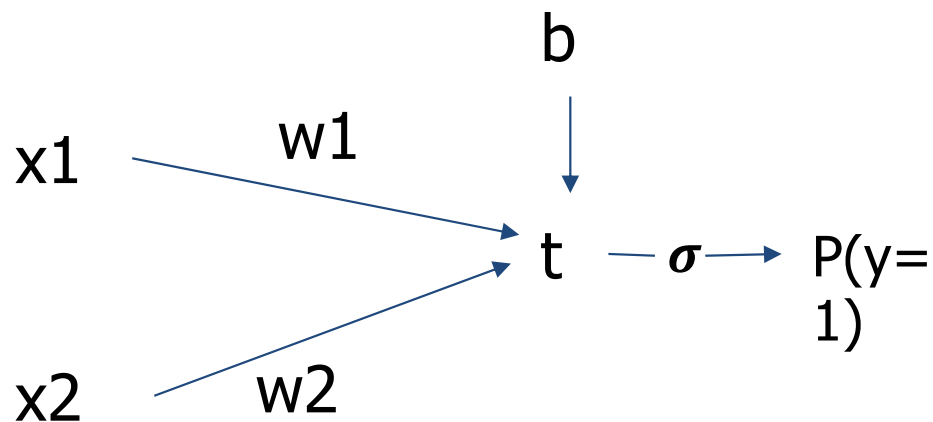
instead of x_3, \dots, x_k .

Example: the general two-variable quadratic regression has 6 constants:

$$Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 (x_1)^2 + \beta_4 (x_2)^2 + \beta_5 (x_1 x_2) + \varepsilon$$

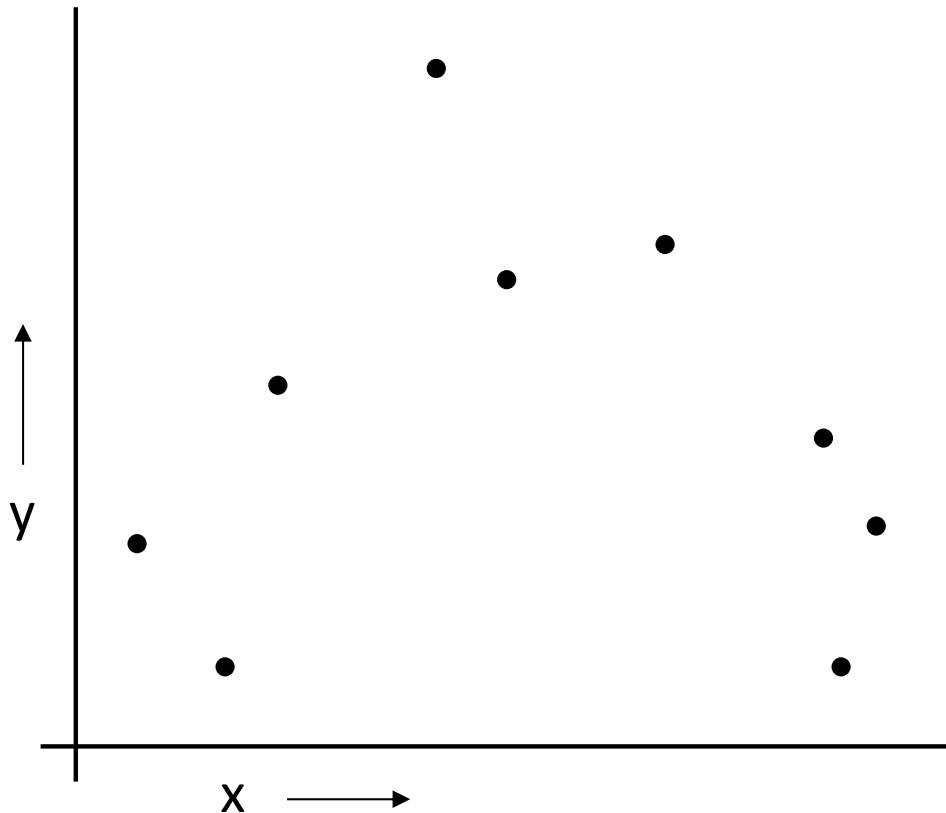
Logistic Regression

$$P(y=1) = \sigma(x_1 * w_1 + x_2 * w_2 + b)$$



How to know where to stop
adding new variables or
powers of old variables?

A Regression Problem

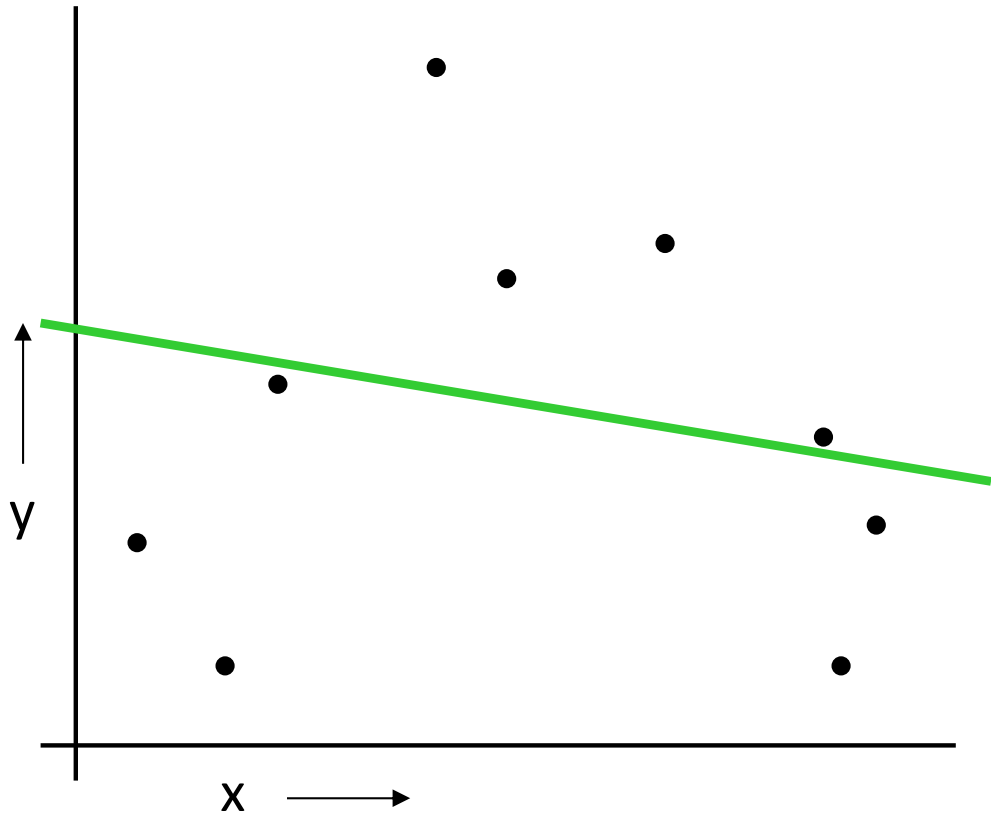


$$y = f(x) + \text{noise}$$

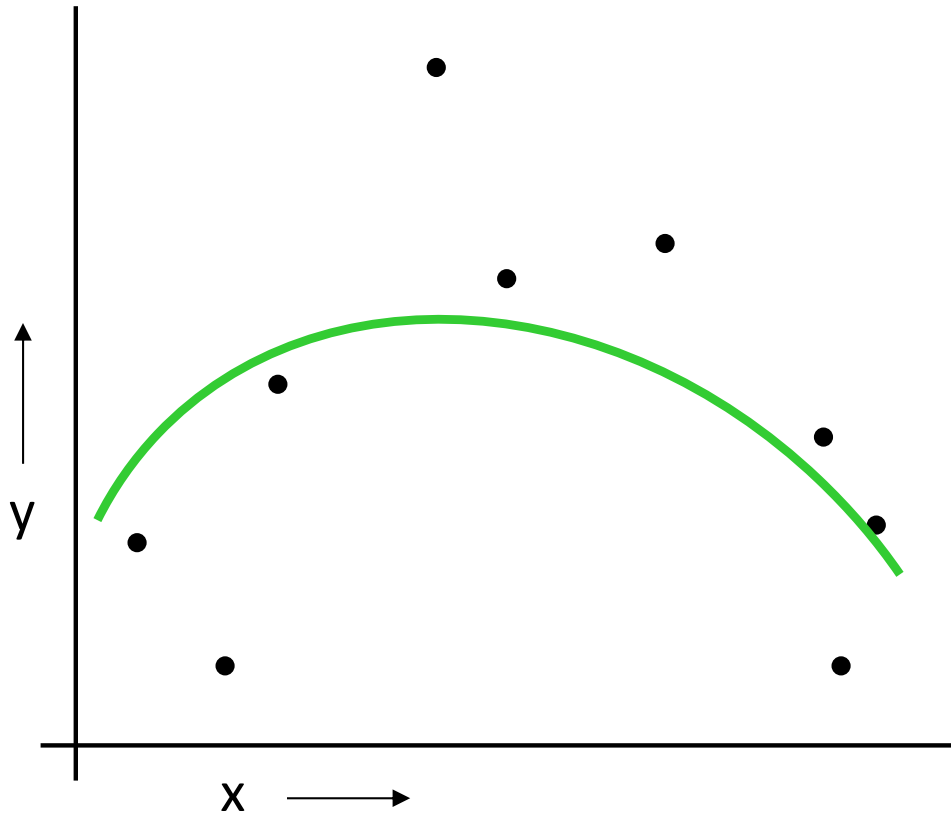
Can we learn f from this data?

Let's consider three methods...

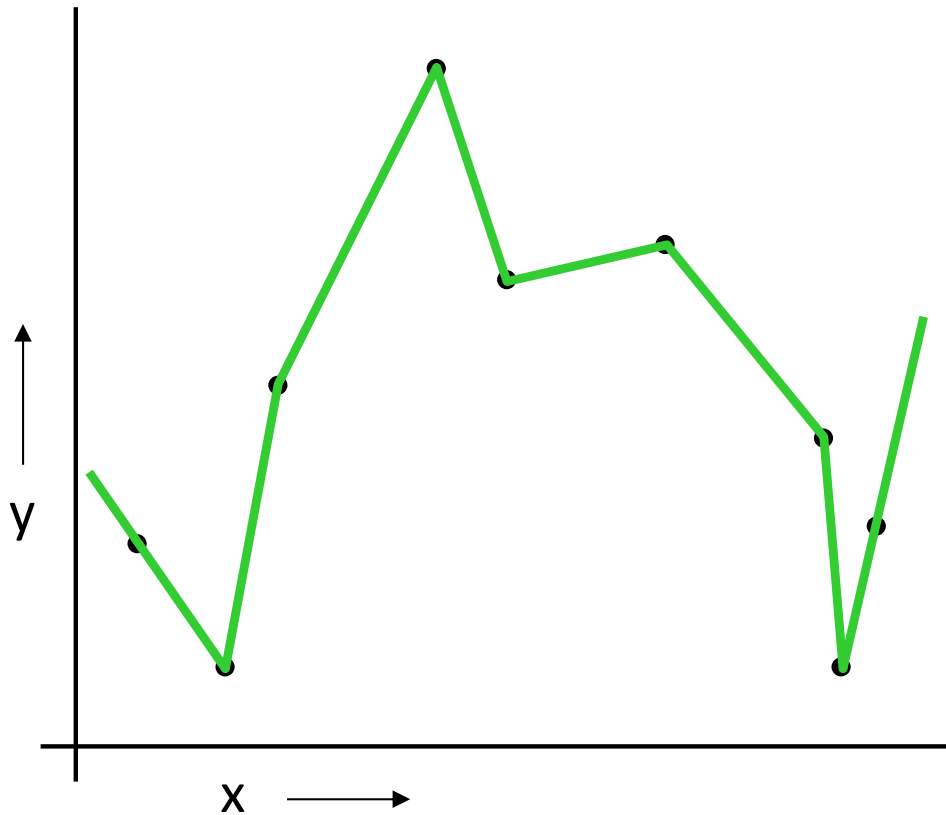
Linear Regression



Quadratic Regression

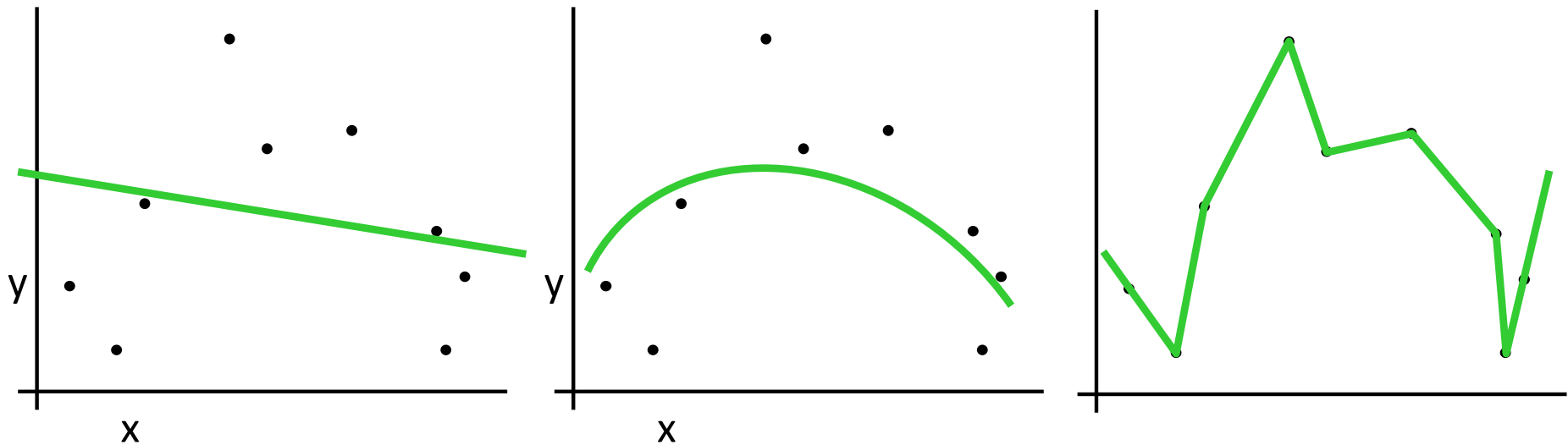


Join-the-dots



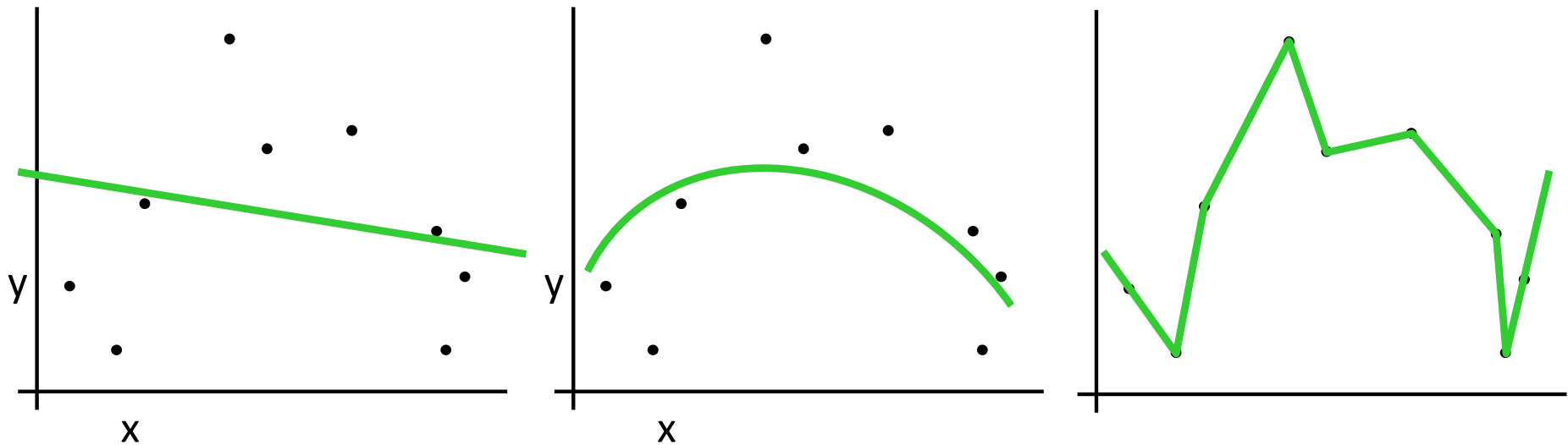
Also known as **piecewise linear nonparametric regression** if that makes you feel better

Which is best?



Why not choose the method with the best fit to the data?

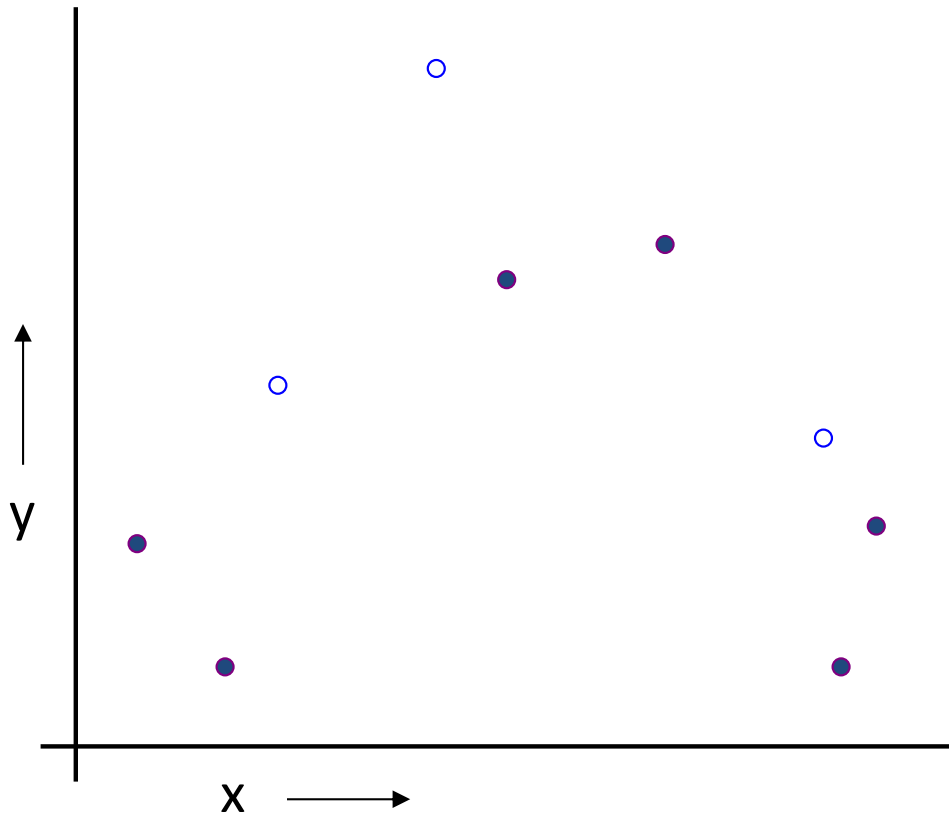
What do we really want?



Why not choose the method with the best fit to the data?

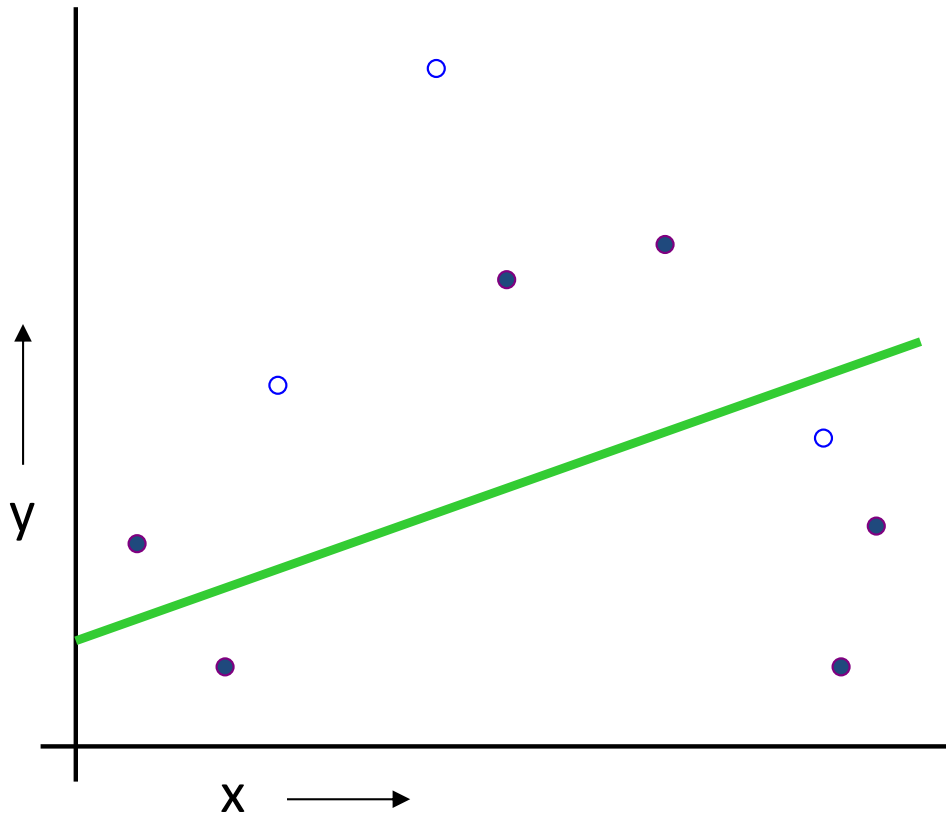
“How well are you going to predict future data drawn from the same distribution?”

The test set method



1. Randomly choose 30% of the data to be in a **test set**
2. The remainder is a **training set**

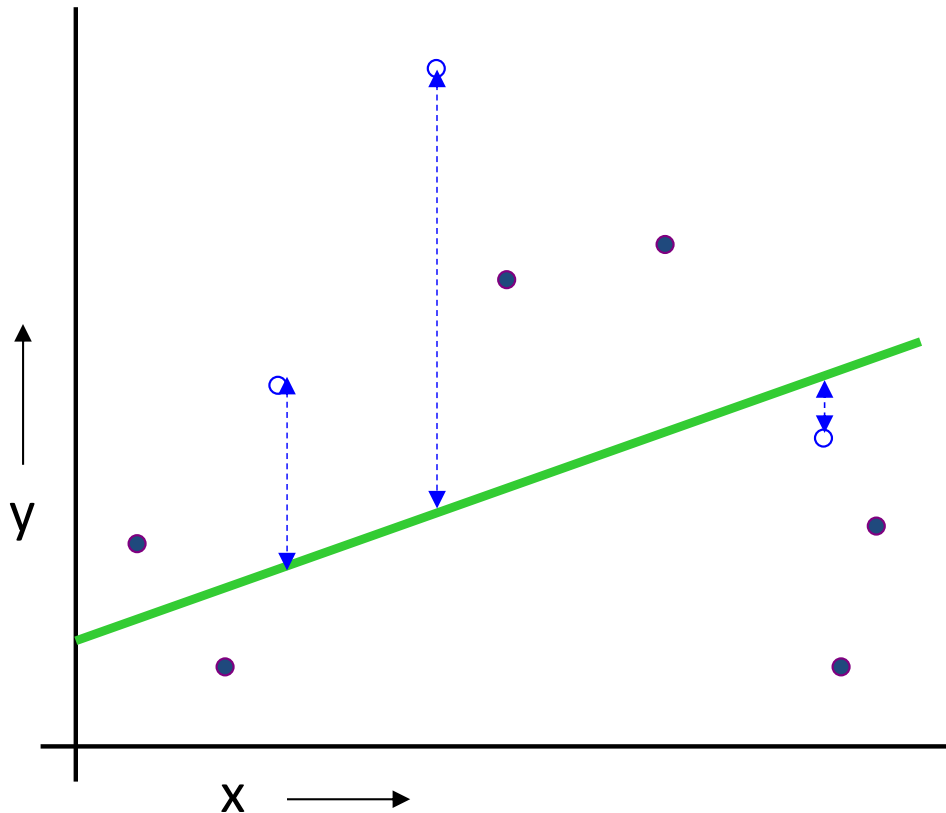
The test set method



(Linear regression example)

1. Randomly choose 30% of the data to be in a **test set**
2. The remainder is a **training set**
3. Perform your regression on the training set

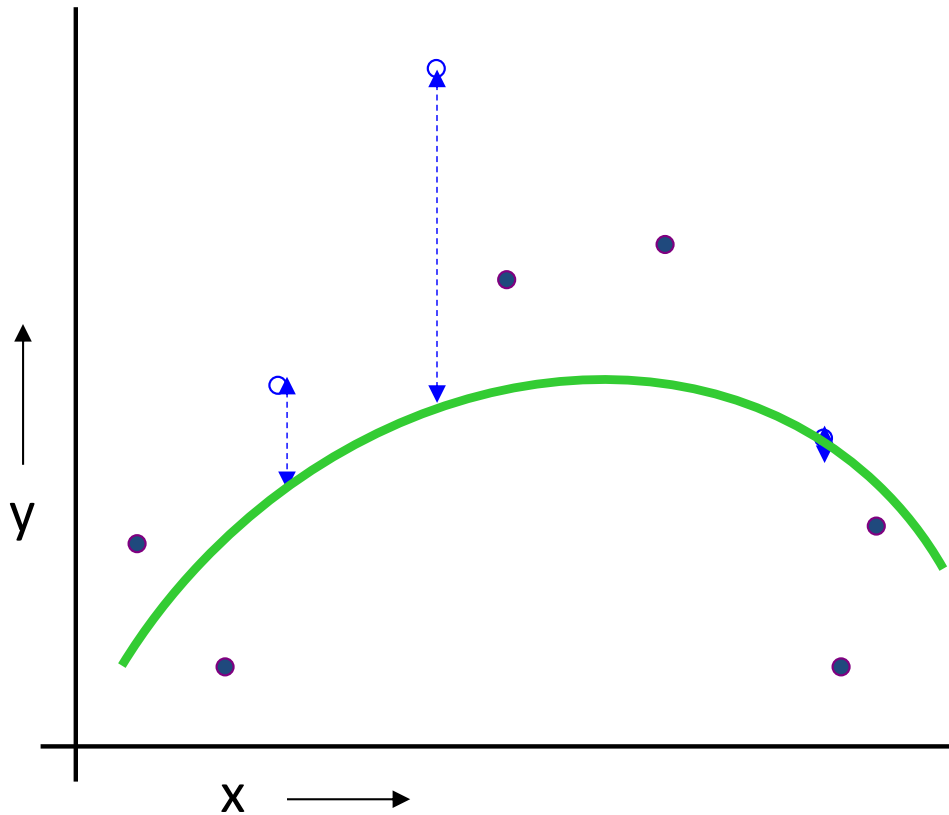
The test set method



(Linear regression example)
Mean Squared Error = 2.4

1. Randomly choose 30% of the data to be in a **test set**
2. The remainder is a **training set**
3. Perform your regression on the training set
4. Estimate your future performance with the test set

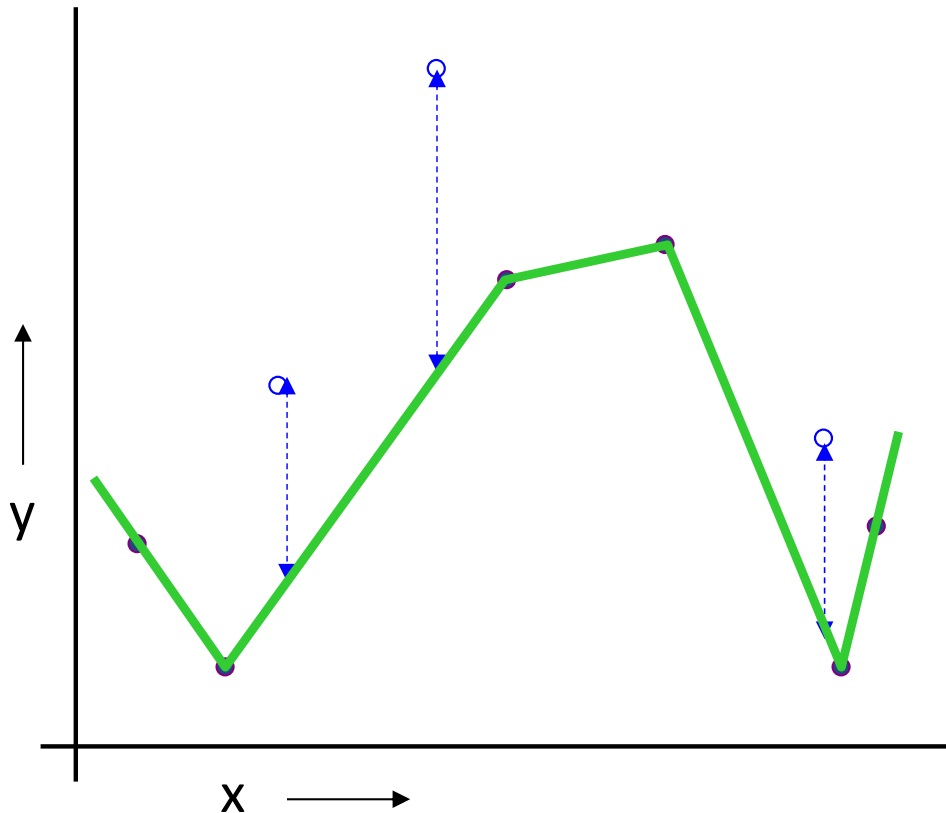
The test set method



(Quadratic regression example)
Mean Squared Error = 0.9

1. Randomly choose 30% of the data to be in a **test set**
2. The remainder is a **training set**
3. Perform your regression on the training set
4. Estimate your future performance with the test set

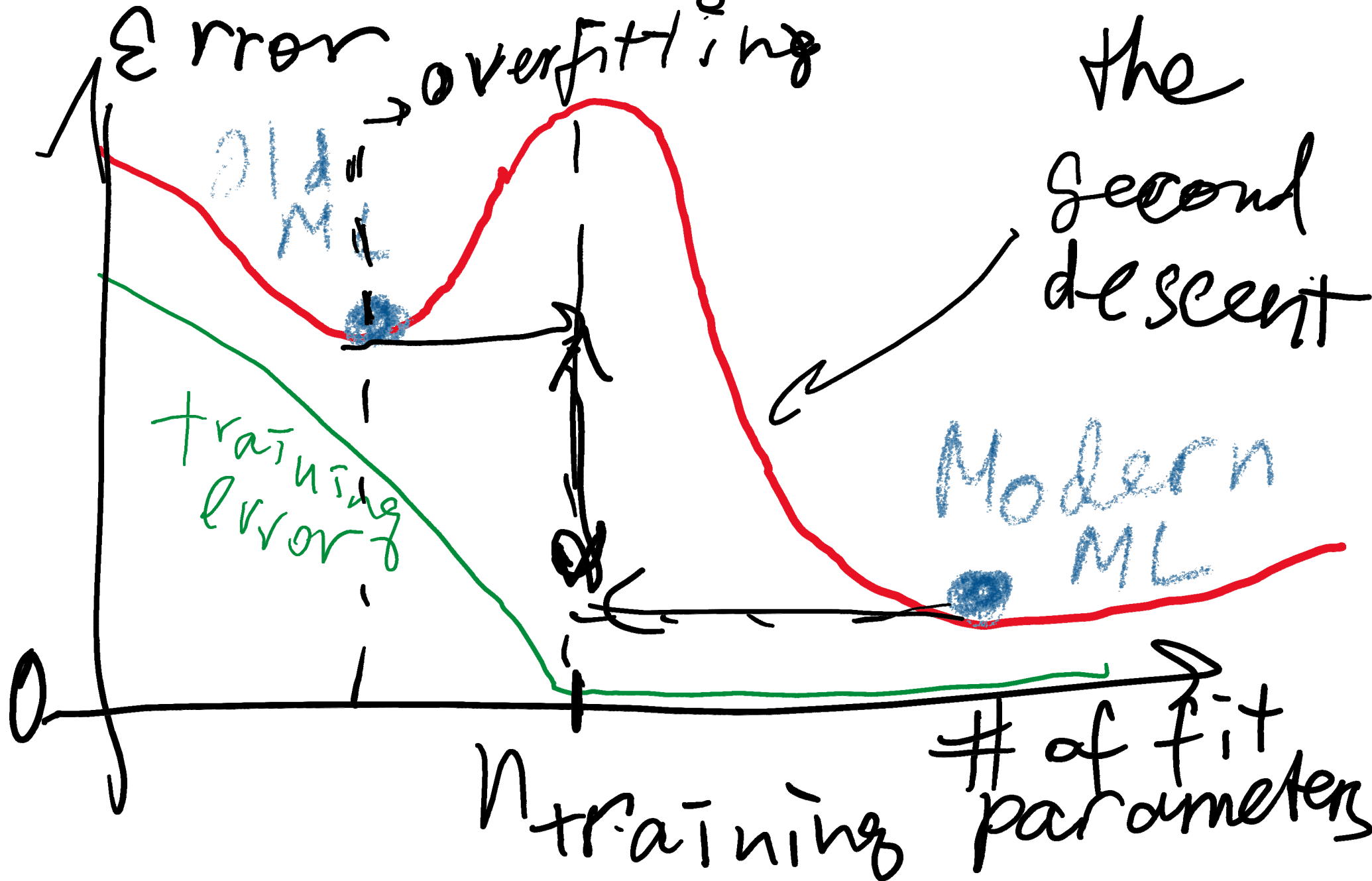
The test set method



(Join the dots example)
Mean Squared Error = 2.2

1. Randomly choose 30% of the data to be in a **test set**
2. The remainder is a **training set**
3. Perform your regression on the training set
4. Estimate your future performance with the test set

Double descend- the main reason modern
Machine Learning works so well



12-1: Multiple Linear Regression Model

12-1.3 Matrix Approach to Multiple Linear Regression

Suppose the model relating the regressors to the response is

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_k x_{ik} + \varepsilon_i \quad i = 1, 2, \dots, n$$

In matrix notation this model can be written as

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon} \quad (12-6)$$

12-1: Multiple Linear Regression Model

12-1.3 Matrix Approach to Multiple Linear Regression

where

$$\mathbf{y} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix} \quad \mathbf{X} = \begin{bmatrix} 1 & x_{11} & x_{12} & \cdots & x_{1k} \\ 1 & x_{21} & x_{22} & \cdots & x_{2k} \\ \vdots & \vdots & \vdots & & \vdots \\ 1 & x_{n1} & x_{n2} & \cdots & x_{nk} \end{bmatrix} \quad \boldsymbol{\beta} = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_k \end{bmatrix} \quad \text{and} \quad \boldsymbol{\varepsilon} = \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{bmatrix}$$

12-1.3 Matrix Approach to Multiple Linear Regression

We wish to find the vector $\hat{\beta}$ that minimizes the sum of squares of error terms:

$$L = \sum_{i=1}^n \varepsilon_i^2 = \varepsilon' \varepsilon = (\mathbf{y} - \mathbf{X}\beta)' (\mathbf{y} - \mathbf{X}\beta)$$

$$0 = \frac{\partial L}{\partial \beta} = -\mathbf{X}' (\mathbf{y} - \mathbf{X}\beta) = -\mathbf{X}' \mathbf{y} + (\mathbf{X}' \mathbf{X}) \beta$$

The resulting least squares estimate is

$$\hat{\beta} = (\mathbf{X}' \mathbf{X})^{-1} \mathbf{X}' \mathbf{y} \quad (12-7)$$

Analog of $\frac{1}{\text{Var}(x)}$

Analog of $\text{Cov}(x, y)$

Multiple Linear Regression Model

$$\hat{\beta} = (X'X)^{-1} X'y$$

*H is an
idempotent
matrix*

$$\hat{y} = X\hat{\beta} = X(X'X)^{-1}X'y,$$

$$\hat{y} = Hy, \quad \text{and} \quad e = (I - H)y.$$



$$H = H'; \quad H^2 = X \underbrace{(X'X)^{-1} X' X}_{I} (X'X)^{-1} X' = X(X'X)^{-1} X' = H$$

Vectors \hat{y} & e are orthogonal since

$$\hat{y}'e = y'H(I-H)y = 0 \quad \text{since}$$

$$H(I - H) = H - H^2 = H - H = 0.$$

12-1: Multiple Linear Regression Models

12-1.4 Properties of the Least Squares Estimators

Unbiased estimators:

$$\begin{aligned} E(\hat{\beta}) &= E[(X'X)^{-1}X'Y] \\ &= E[(X'X)^{-1}X'(X\beta + \epsilon)] \\ &= E[(X'X)^{-1}X'X\beta + (X'X)^{-1}X'\epsilon] \\ &= \beta \end{aligned}$$

Covariance Matrix of Estimators:

$$\mathbf{C} = (X'X)^{-1} = \begin{bmatrix} C_{00} & C_{01} & C_{02} \\ C_{10} & C_{11} & C_{12} \\ C_{20} & C_{21} & C_{22} \end{bmatrix}$$

12-1: Multiple Linear Regression Models

12-1.4 Properties of the Least Squares Estimators

Individual variances and covariances:

$$V(\hat{\beta}_j) = \sigma^2 C_{jj}, \quad j = 0, 1, 2$$
$$\text{cov}(\hat{\beta}_i, \hat{\beta}_j) = \sigma^2 C_{ij}, \quad i \neq j$$

In general,

$$\text{cov}(\hat{\beta}) = \sigma^2 (\mathbf{X}'\mathbf{X})^{-1} = \sigma^2 \mathbf{C}$$

12-1: Multiple Linear Regression Models

Estimating error variance σ_ε^2

An unbiased estimator of error variance σ_ε^2 is

$$\hat{\sigma}_\varepsilon^2 = \frac{\sum_{i=1}^n e_i^2}{n - p} = \frac{SS_E}{n - p} \quad (12-16)$$

Here $p=k+1$ for k -variable multiple linear regression

R^2 and Adjusted R^2

The **coefficient of multiple determination R^2**

$$R^2 = \frac{SS_R}{SS_T} = 1 - \frac{SS_E}{SS_T}$$

The **adjusted R^2** is

$$R^2_{\text{adj}} = 1 - \frac{SS_E/(n - p)}{SS_T/(n - 1)} \approx 1 - \frac{\overset{\uparrow}{2} \varepsilon}{\sum y^2} \quad (12-23)$$

- The adjusted R^2 statistic penalizes **adding terms** to the MLR model.
- It can help guard against **overfitting** (including regressors that are not really useful)

How to know where to stop adding variables?

- Adding new variables x_i to MLR
watch the adjusted R^2
- Once the adjusted R^2
no longer increases = stop.
Now you did the best you can.

Credit: XKCD
comics

WHY ARE THERE SLAVES IN THE BIBLE

WHY DO TWINS HAVE DIFFERENT FINGERPRINTS
WHY ARE AMERICANS AFRAID OF DRAGONS

QUESTIONS

FOUND IN GOOGLE AUTOCOMPLETE

WHY IS HTTPS CROSSED OUT IN RED
WHY IS THERE A LINE THROUGH HTTPS
WHY IS THERE A RED LINE THROUGH HTTPS ON FACEBOOK
WHY IS HTTPS IMPORTANT

WHY AREN'T MY
ARMS GROWING



WHY ARE THERE WEEKS
WHY DO I FEEL DIZZY

WHY AREN'T ECONOMISTS RICH

WHY DO AMERICANS CALL IT SOCCER

WHY ARE MY EARS RINGING

WHY ARE THERE SO MANY AVENGERS

WHY ARE THE AVENGERS FIGHTING THE X MEN

WHY IS WOLVERINE NOT IN THE AVENGERS

WHY ARE THERE ANTS IN MY LAPTOP

WHY IS EARTH TILTED

WHY IS SPACE BLACK

WHY IS OUTER SPACE SO COLD

WHY ARE THERE PYRAMIDS ON THE MOON

WHY IS NASA SHUTTING DOWN

WHY ARE THERE MALE AND FEMALE BIKES

WHY ARE THERE TINY SPIDERS IN MY HOUSE

WHY DO SPIDERS COME INSIDE

WHY ARE THERE HUGE SPIDERS IN MY HOUSE

WHY ARE THERE LOTS OF SPIDERS IN MY HOUSE

WHY ARE THERE SPIDERS IN MY ROOM

WHY ARE THERE SO MANY SPIDERS IN MY ROOM

WHY DO SPIDER BITES ITCH

WHY IS DYING SO SCARY

WHY IS THERE NO GPS IN LAPTOPS

WHY DO KNEES CLICK

WHY AREN'T THERE E GRADES

WHY IS ISOLATION BAD

WHY DO BOYS LIKE ME

WHY DON'T BOYS LIKE ME

WHY IS THERE ALWAYS A JAVA UPDATE

WHY ARE THERE RED DOTS ON MY THIGHS

WHY IS LYING GOOD

WHY ARE THERE
GHOSTS



WHY IS THERE AN OWL IN MY BACKYARD

WHY IS THERE AN OWL OUTSIDE MY WINDOW

WHY IS THERE AN OWL ON THE DOLLAR BILL

WHY DO OWLS ATTACK PEOPLE

WHY ARE AK 47s SO EXPENSIVE

WHY ARE THERE HELICOPTERS CIRCLING MY HOUSE

WHY ARE THERE GODS

WHY ARE THERE TWO SPOCKS

WHY IS MT VESUVIUS THERE

WHY DO THEY SAY T MINUS

WHY ARE THERE OBELISKS

WHY ARE WRESTLERS ALWAYS WET

WHY ARE OCEANS BECOMING MORE ACIDIC

WHY IS ARWEN DYING

WHY AREN'T MY QUAIL LAYING EGGS

WHY AREN'T MY QUAIL EGGS HATCHING

WHY AREN'T THERE ANY FOREIGN MILITARY BASES IN AMERICA

WHY ARE CIGARETTES LEGAL

WHY ARE THERE DUCKS IN MY POOL

WHY IS JESUS WHITE

WHY IS THERE LIQUID IN MY EAR

WHY DO Q TIPS FEEL GOOD

WHY DO GOOD PEOPLE DIE

WHY AREN'T
THERE GUNS IN
HARRY POTTER



WHY ARE ULTRASOUNDS IMPORTANT

WHY ARE ULTRASOUND MACHINES EXPENSIVE

WHY IS STEALING WRONG

WHY ARE THERE FIREWORKS

WHY ARE THERE DOGS AFRAID OF FIREWORKS
WHY IS THERE NO KING IN ENGLAND

WHY DO WHALES JUMP
WHY ARE WITCHES GREEN
WHY ARE THERE MIRRORS ABOVE BEDS
WHY DO I SAY UH
WHY IS SEA SALT BETTER
WHY ARE THERE TREES IN THE MIDDLE OF FIELDS
WHY IS THERE NOT A POKEMON MMO
WHY IS THERE LAUGHING IN TV SHOWS
WHY ARE THERE DOORS ON THE FREEWAY
WHY ARE THERE SO MANY SVCHOST.EXE RUNNING
WHY AREN'T THERE ANY COUNTRIES IN ANTARCTICA
WHY ARE THERE SCARY SOUNDS IN MINECRAFT
WHY IS THERE KICKING IN MY STOMACH
WHY ARE THERE TWO SLASHES AFTER HTTP
WHY ARE THERE CELEBRITIES
WHY DO SNAKES EXIST
WHY DO OYSTERS HAVE PEARLS
WHY ARE DUCKS CALLED DUCKS
WHY DO THEY CALL IT THE CLAP
WHY ARE KYLE AND CARTMAN FRIENDS
WHY IS THERE AN ARROW ON AANG'S HEAD
WHY ARE TEXT MESSAGES BLUE
WHY ARE THERE MUSTACHES ON CLOTHES
WHY ARE THERE MUSTACHES ON CARS
WHY ARE THERE MUSTACHES EVERYWHERE
WHY ARE THERE SO MANY BIRDS IN OHIO
WHY IS THERE SO MUCH RAIN IN OHIO
WHY IS OHIO WEATHER SO WEIRD
WHY ARE THERE BRIDESMAIDS
WHY DO DYING PEOPLE REACH UP
WHY AREN'T THERE VARICOSE ARTERIES
WHY ARE OLD KINGDOMS DIFFERENT

WHY ARE THERE
SQUIRRELS



WHY IS PROGRAMMING SO HARD
WHY IS THERE A 0 OHM RESISTOR
WHY DO AMERICANS HATE SOCCER
WHY DO RHYMES SOUND GOOD
WHY DO TREES DIE
WHY IS THERE NO SOUND ON CNN
WHY AREN'T POKEMON REAL
WHY AREN'T BULLETS SHARP
WHY DO DREAMS SEEM SO REAL

WHY IS SEX
SO IMPORTANT

