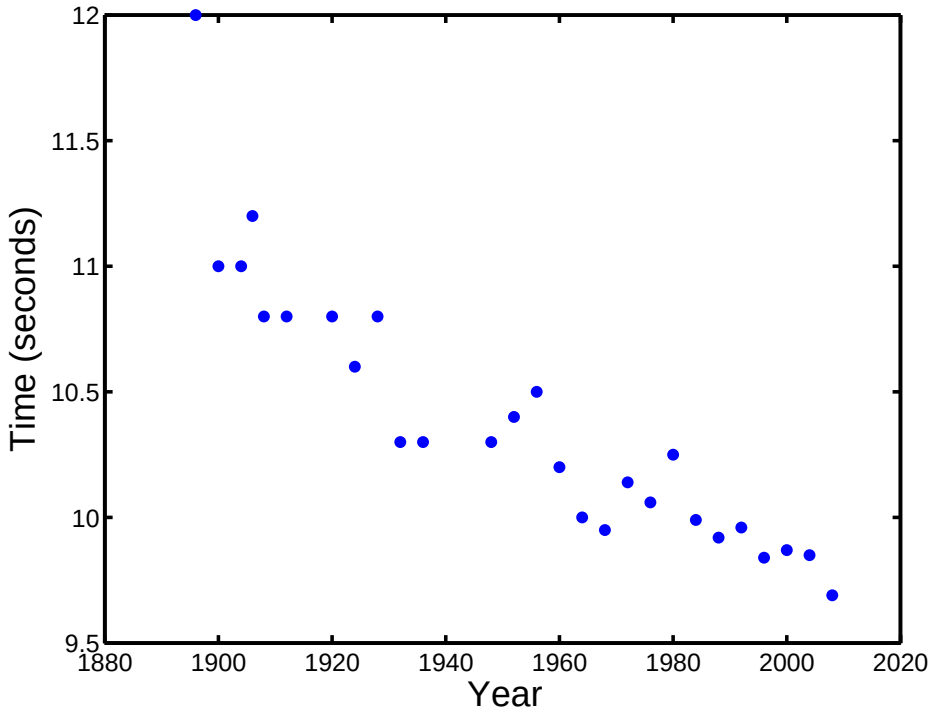
Some data and a problem

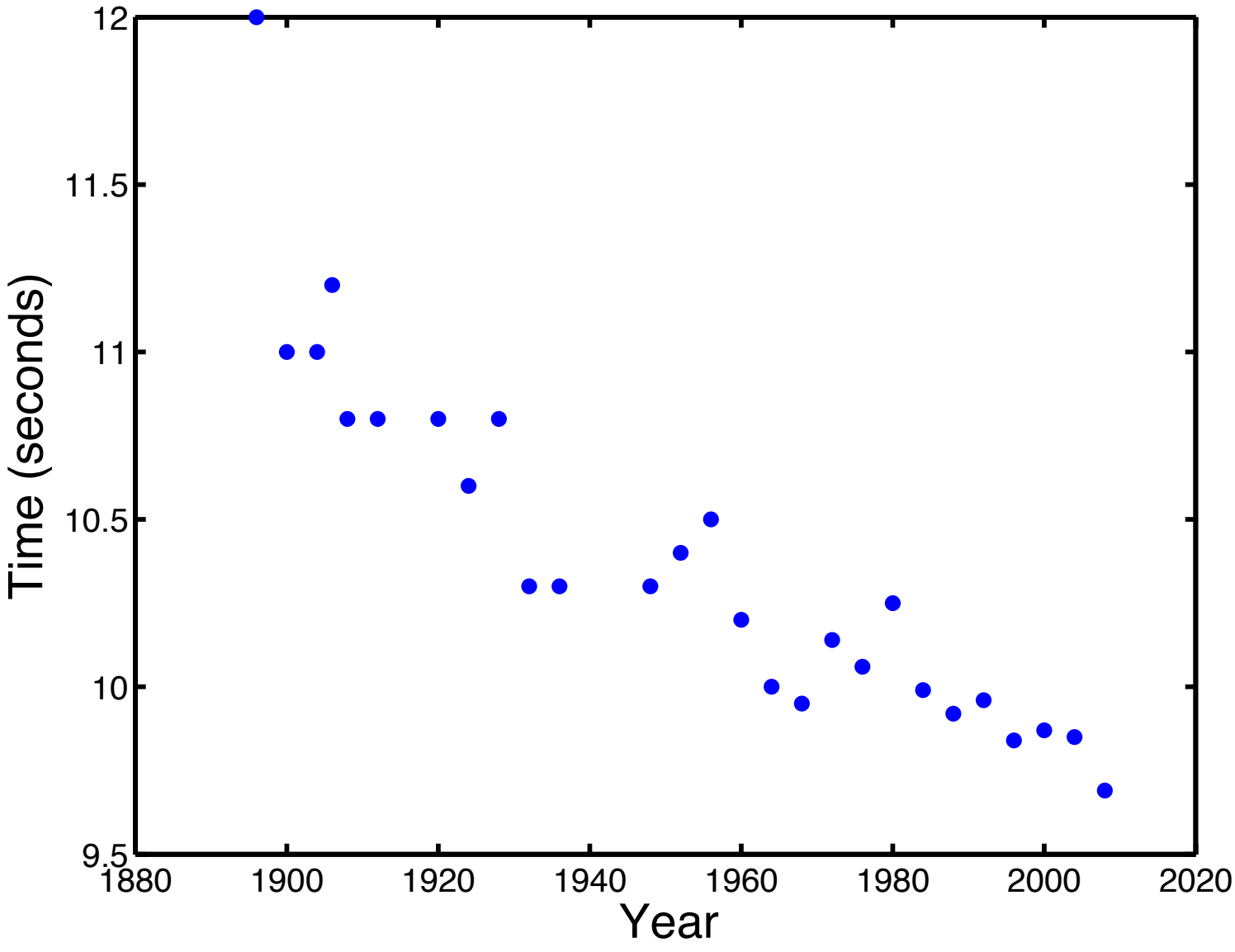


Winning times for the men's Olympic 100m sprint, 1896-2008.

In this lecture, we will use this data to predict the winning time in London 2012

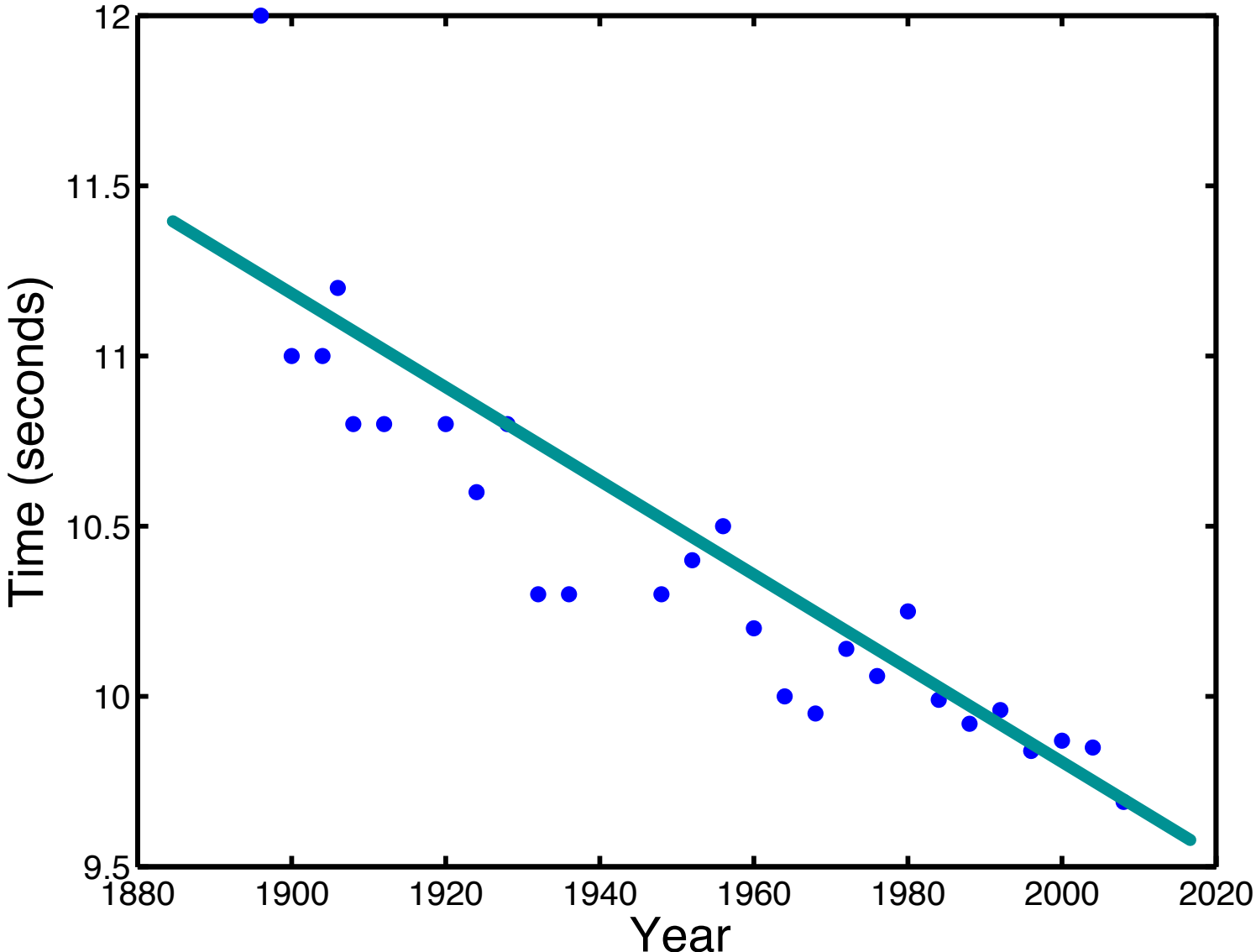
Reading: Section 1.1 of FCML

Draw a line through it!



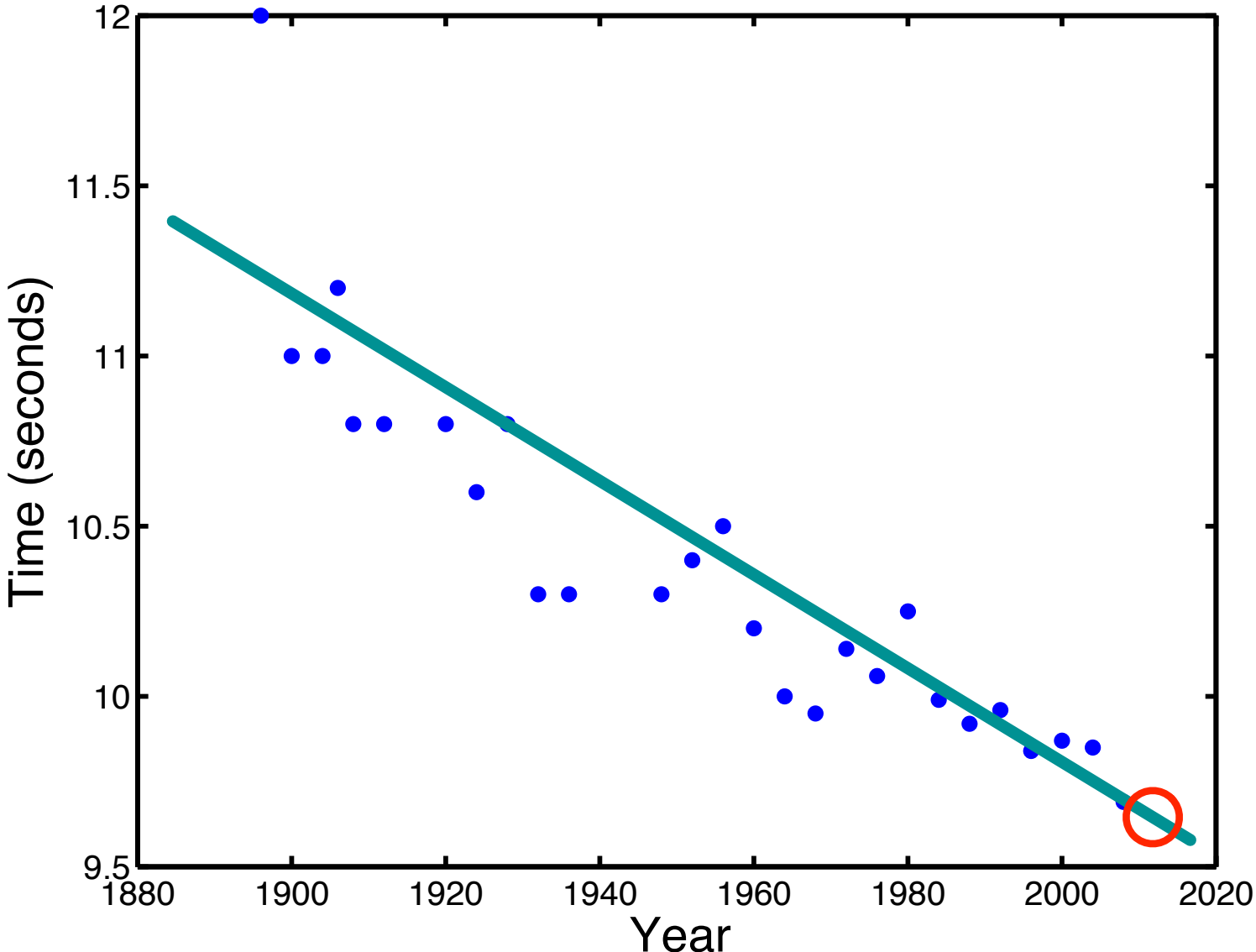
Draw a line through it!

Draw a line through it!



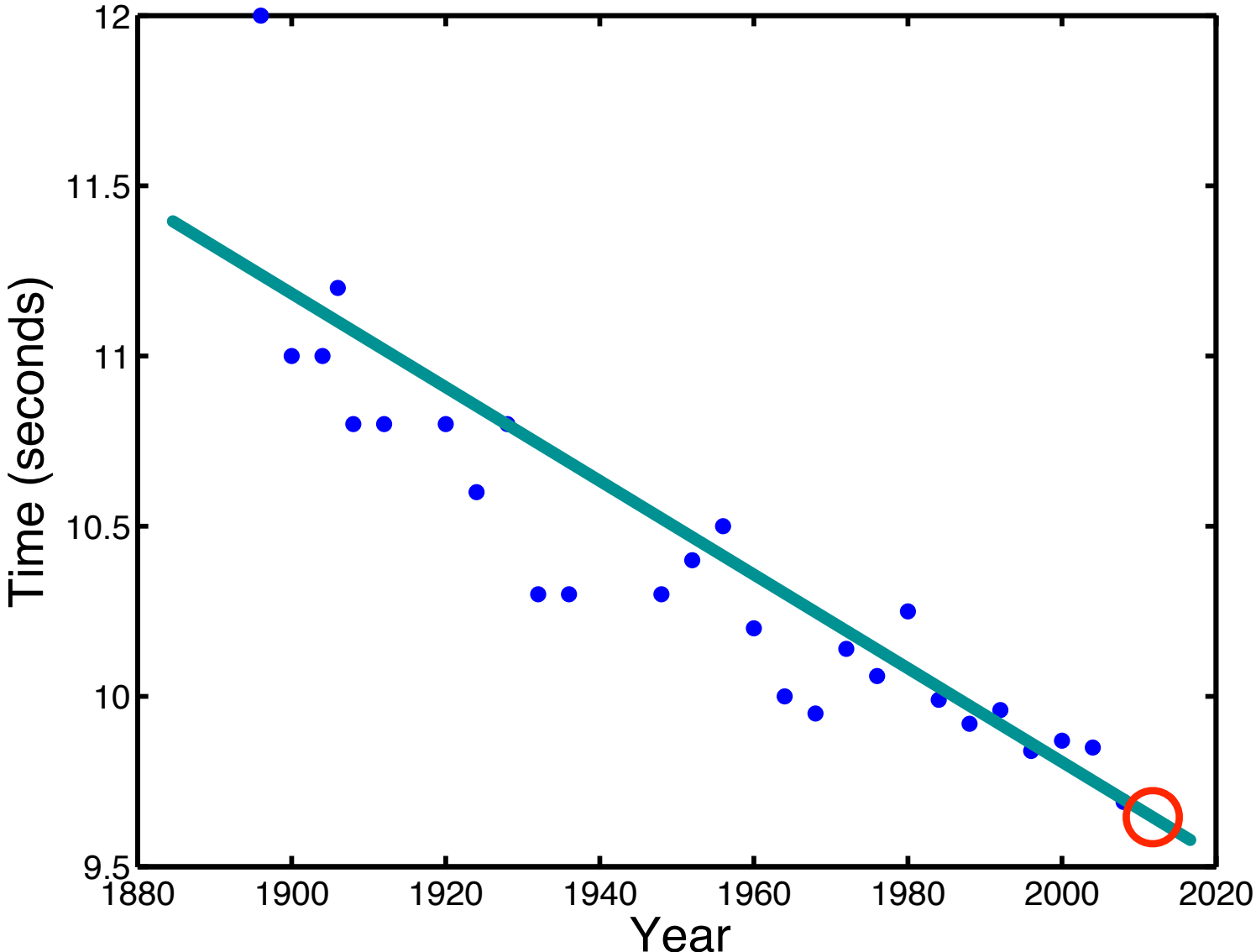
Draw a line through it!

Draw a line through it!



Draw a line through it!

Draw a line through it!



Overview

Linear models

- Introduce the idea of **building models**.
- Talk about **assumptions**.
- Use a **linear** model.
- What constitutes a **good** model?
- Find the **best** linear model.
- Use it to **predict** the winning time in 2012.

What did we do?

Basically:

- **Decided to draw a line through our data.**
- Chose a straight line.
- Drew a good straight line.
- Extended the line to 2012.
- Read off the winning time for 2012.

Technically

- Decided we needed a model.
- Chose a linear model.
- Fitted a linear model.
- Evaluated the model at 2012.
- Used this as our prediction.

What did we do?

Basically:

- Decided to draw a line through our data.
- **Chose a straight line.**
- Drew a good straight line.
- Extended the line to 2012.
- Read off the winning time for 2012.

Technically

- Decided we needed a model.
- Chose a linear model.
- Fitted a linear model.
- Evaluated the model at 2012.
- Used this as our prediction.

What did we do?

Basically:

- Decided to draw a line through our data.
- Chose a straight line.
- **Drew a good straight line.**
- Extended the line to 2012.
- Read off the winning time for 2012.

Technically

- Decided we needed a model.
- Chose a linear model.
- Fitted a linear model.
- Evaluated the model at 2012.
- Used this as our prediction.

What did we do?

Basically:

- Decided to draw a line through our data.
- Chose a straight line.
- Drew a good straight line.
- **Extended the line to 2012.**
- Read off the winning time for 2012.

Technically

- Decided we needed a model.
- Chose a linear model.
- Fitted a linear model.
- Evaluated the model at 2012.
- Used this as our prediction.

What did we do?

Basically:

- Decided to draw a line through our data.
- Chose a straight line.
- Drew a good straight line.
- Extended the line to 2012.
- **Read off the winning time for 2012.**

Technically

- Decided we needed a model.
- Chose a linear model.
- Fitted a linear model.
- Evaluated the model at 2012.
- Used this as our prediction.

Assumptions

Assumptions

Our Assumptions

1. That there exists a relationship between Olympic year and winning time.
2. **That this relationship is linear (i.e. a straight line).**

Assumptions

Our Assumptions

1. That there exists a relationship between Olympic year and winning time.
2. That this relationship is linear (i.e. a straight line).
3. **That this relationship will continue into the future.**

Assumptions

Our Assumptions

1. That there exists a relationship between Olympic year and winning time.
2. That this relationship is linear (i.e. a straight line).
3. That this relationship will continue into the future.

Are they any good?

Summary

So far..

- We have some data.
- We want to predict the winning time in 2012.
- We've made some assumptions.

Summary

So far..

- We have some data.
- We want to predict the winning time in 2012.
- We've made some assumptions.

The rest of the lecture...

- Mathematically define the model.
- Fit the model – there are many possible lines!
- Make our prediction.
- Some discussion about the assumptions.

Definitions

Attributes and targets

Typically in Supervised Machine Learning, we have a set of attributes and corresponding targets:

Definitions

Attributes and targets

Typically in Supervised Machine Learning, we have a set of **attributes** and corresponding targets:

- **Attributes:** Olympic year.

Definitions

Attributes and targets

Typically in Supervised Machine Learning, we have a set of attributes and corresponding **targets**:

- **Attributes:** Olympic year.
- **Targets:** Winning time.

Definitions

Attributes and targets

Typically in Supervised Machine Learning, we have a set of attributes and corresponding targets:

- **Attributes:** Olympic year.
- **Targets:** Winning time.

Variables

Mathematically, each is described by a variable:

Definitions

Attributes and targets

Typically in Supervised Machine Learning, we have a set of attributes and corresponding targets:

- **Attributes:** Olympic year.
- **Targets:** Winning time.

Variables

Mathematically, each is described by a variable:

- Olympic year: x.

Definitions

Attributes and targets

Typically in Supervised Machine Learning, we have a set of attributes and corresponding targets:

- **Attributes:** Olympic year.
- **Targets:** Winning time.

Variables

Mathematically, each is described by a variable:

- Olympic year: x.
- Winning time: t.

Definitions

Model

Our goal is to create a model.

- This is a function that can relate x to t.

$$t = f(x)$$

- Hence, we can work out t when $x = 2012$.

Definitions

Model

Our goal is to create a model.

- This is a function that can relate x to t.

$$t = f(x)$$

- Hence, we can work out t when $x = 2012$.

Data

We're going to create the model from data:

- N attribute-response pairs, (x_n, t_n)
- e.g. $(1896, 12s), (1900, 11s), \ldots, (2008, 9.69s)$
- $x_1 = 1896,\ t_1 = 12$, etc

Definitions

Model

Our goal is to create a model.

- This is a function that can relate x to t.

$$t = f(x)$$

- Hence, we can work out t when $x = 2012$.

Data

We're going to create the model from data:

- N attribute-response pairs, (x_n, t_n)
- e.g. $(1896, 12s), (1900, 11s), \ldots, (2008, 9.69s)$
- $x_1 = 1896,\ t_1 = 12$, etc

Often called **training** data

A linear model

$$t = f(x)$$

A linear model

$$t = f(x) = w_0 + w_1 x$$

A linear model

$$t = f(x) = w_0 + w_1 x = f(x; w_0, w_1)$$

- w_0 and w_1 are *parameters* of the model.

A linear model

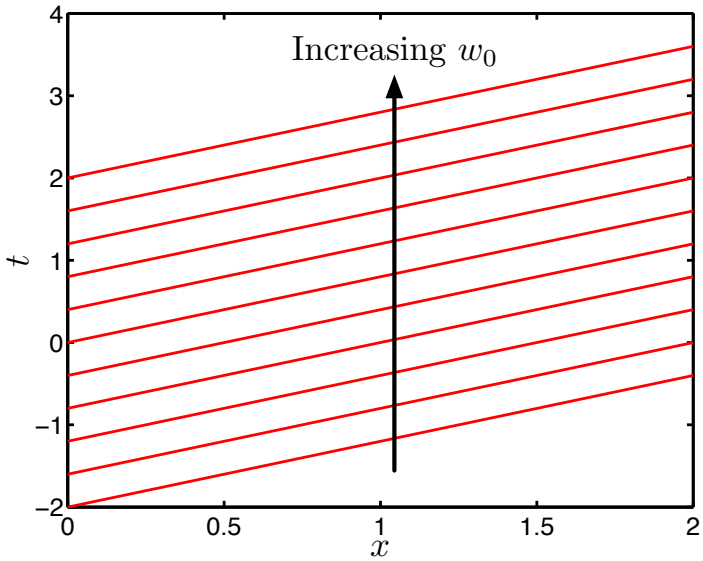
$$t = f(x) = w_0 + w_1 x = f(x; w_0, w_1)$$

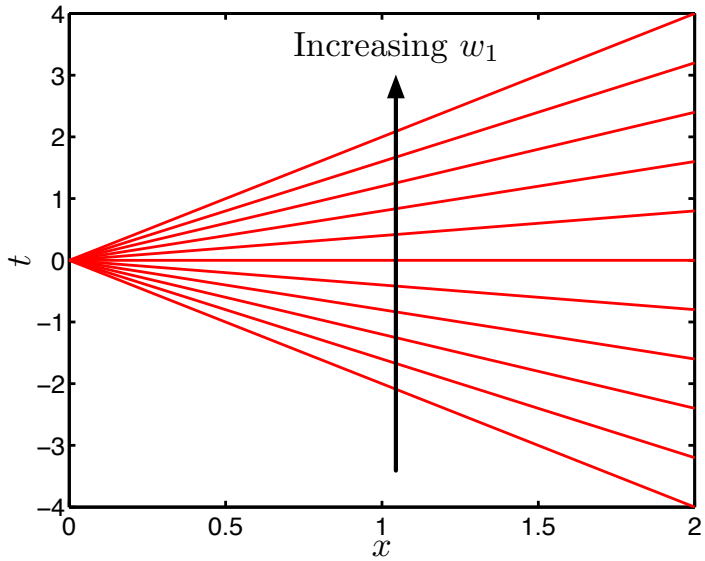
- w_0 and w_1 are *parameters* of the model.
- They determine the properties of the line.

A linear model

$$t = f(x) = w_0 + w_1 x = f(x; w_0, w_1)$$

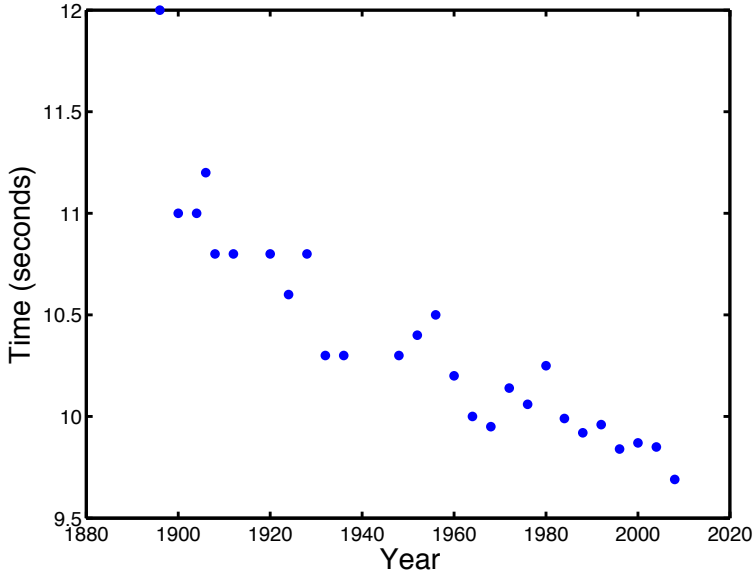
- w_0 and w_1 are *parameters* of the model.
- They determine the properties of the line.



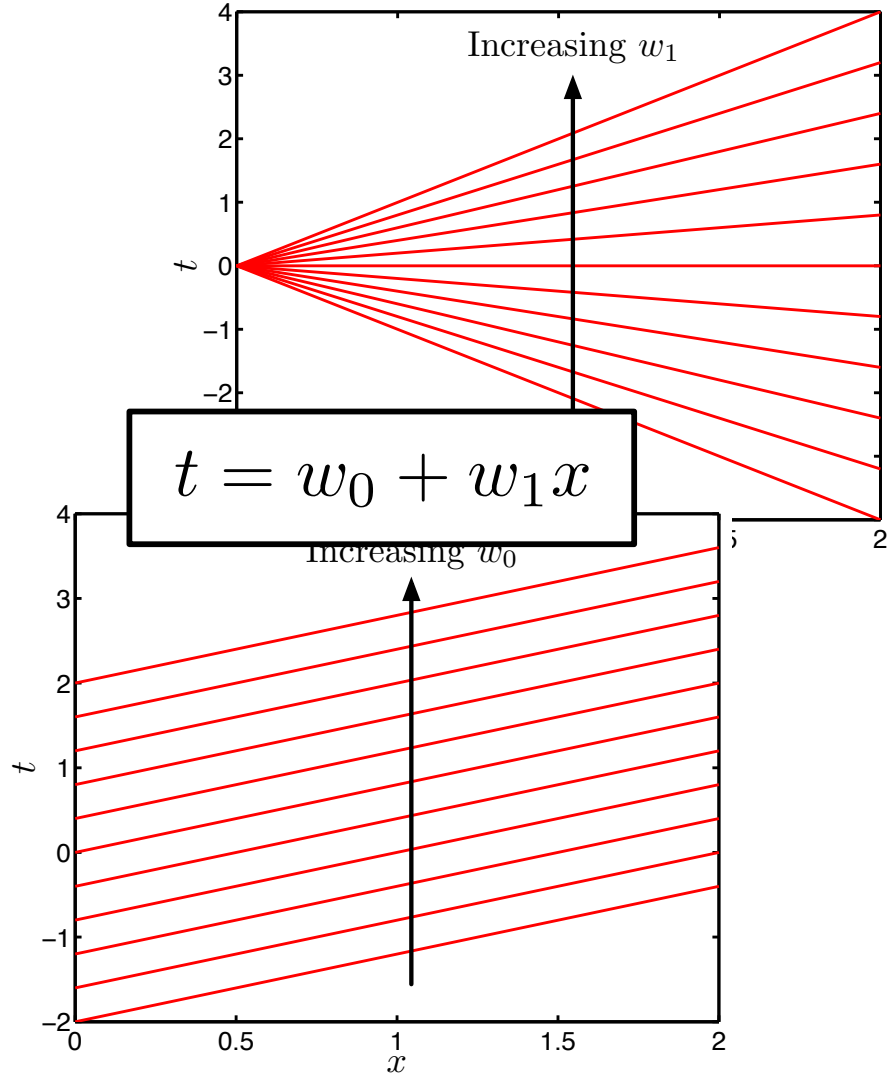


What next?

We have data and a family of models:

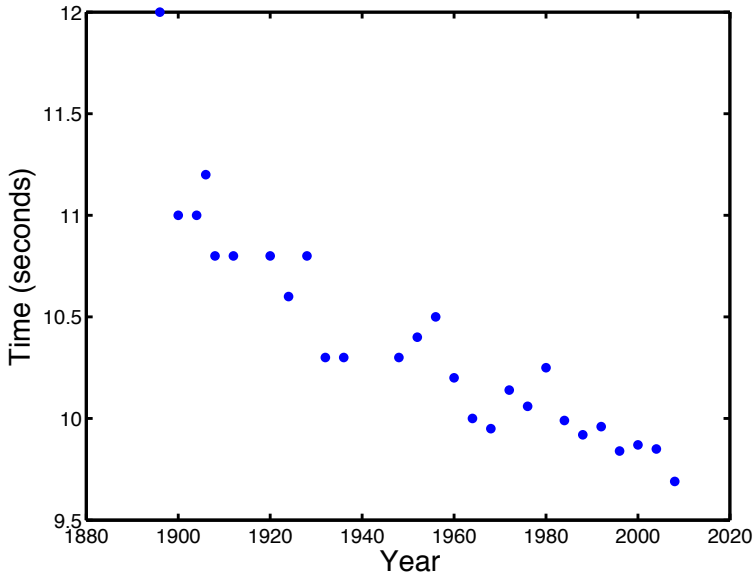


?

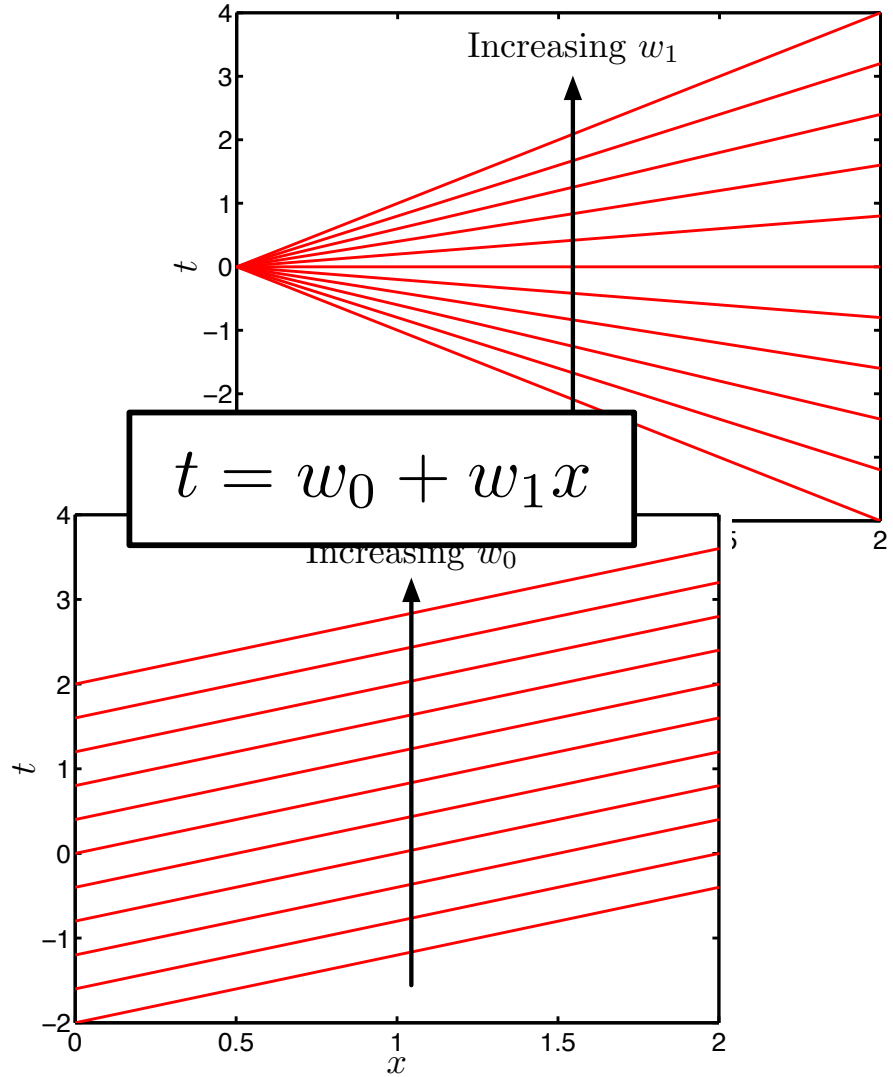


What next?

We have data and a family of models:



?



Need to find w_0, w_1 from $(x_1, t_1), (x_2, t_2), \ldots, (x_N, t_N)$

How good is a particular w_0, w_1?

- How good is a particular line (w_0, w_1)?

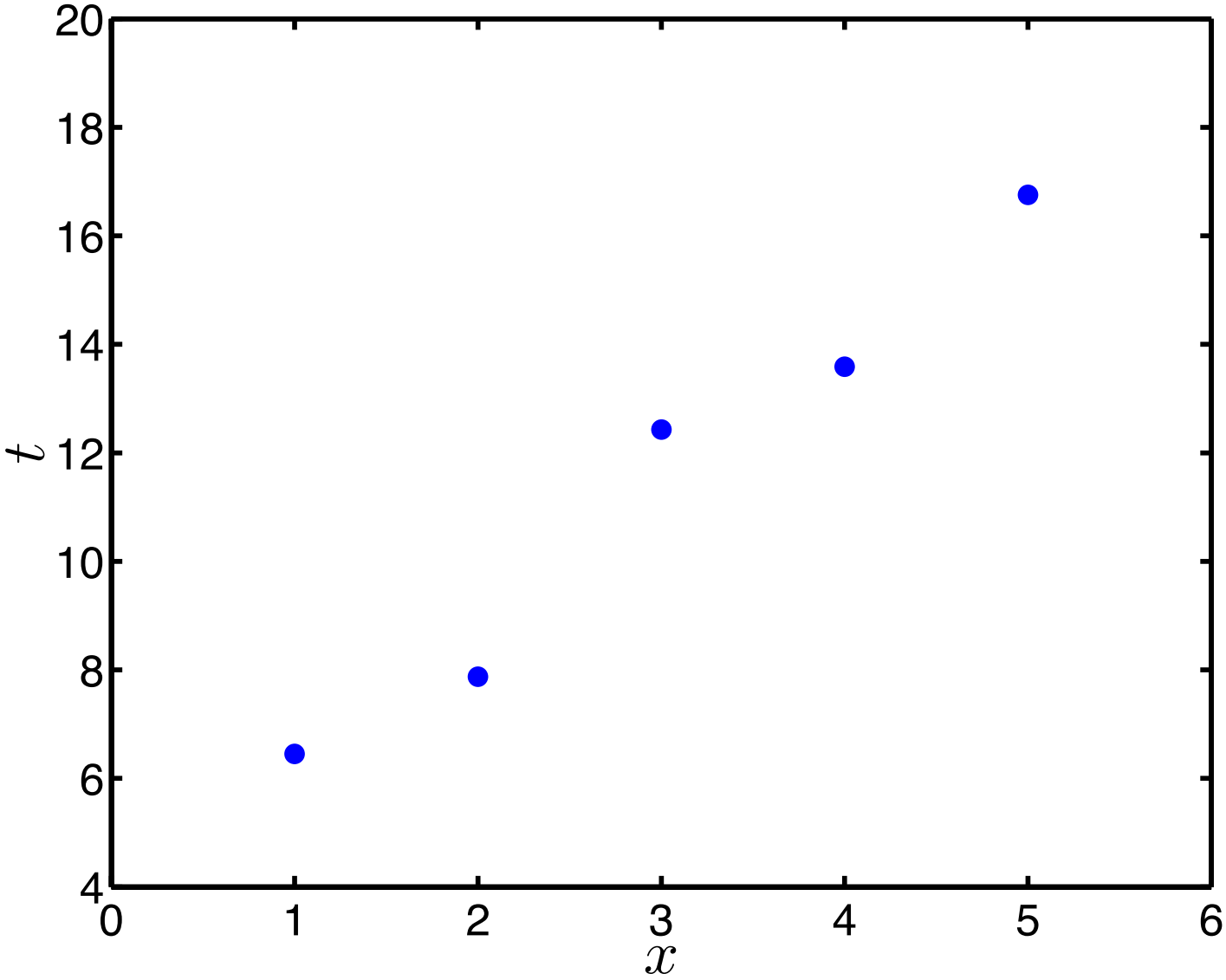
How good is a particular w_0, w_1?

- How good is a particular line (w_0, w_1)?
- We need to be able to provide a numerical value of goodness for any w_0, w_1.
  - How good is $w_0 = 5, w_1 = 0.1$?
  - Is $w_0 = 5, w_1 = -0.1$ better or worse?

How good is a particular w_0, w_1?

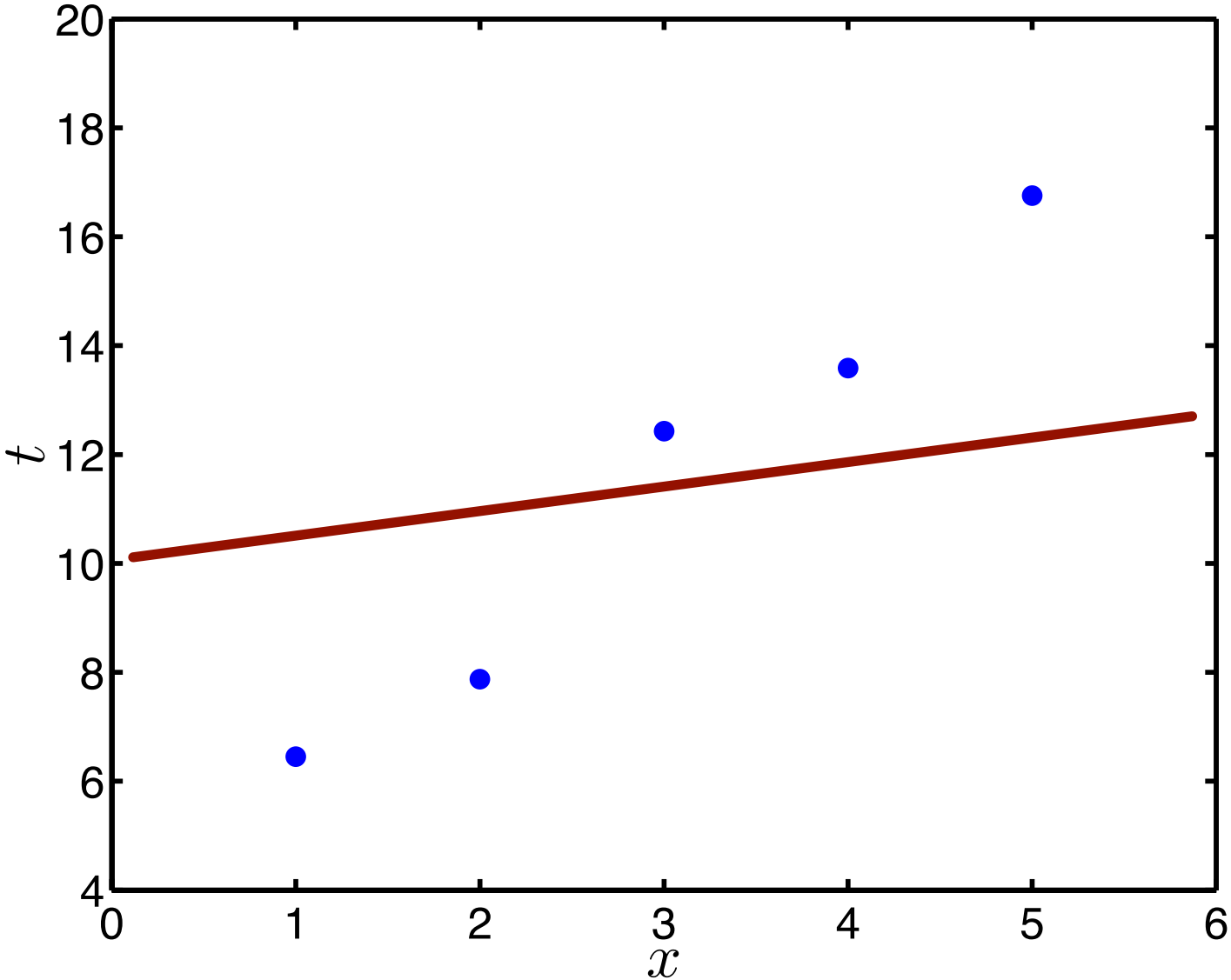
- How good is a particular line (w_0, w_1)?
- We need to be able to provide a numerical value of goodness for any w_0, w_1.
  - How good is $w_0 = 5, w_1 = 0.1$?
  - Is $w_0 = 5, w_1 = -0.1$ better or worse?
- Once we can answer these questions, we can search for the best w_0, w_1 pair.

Loss



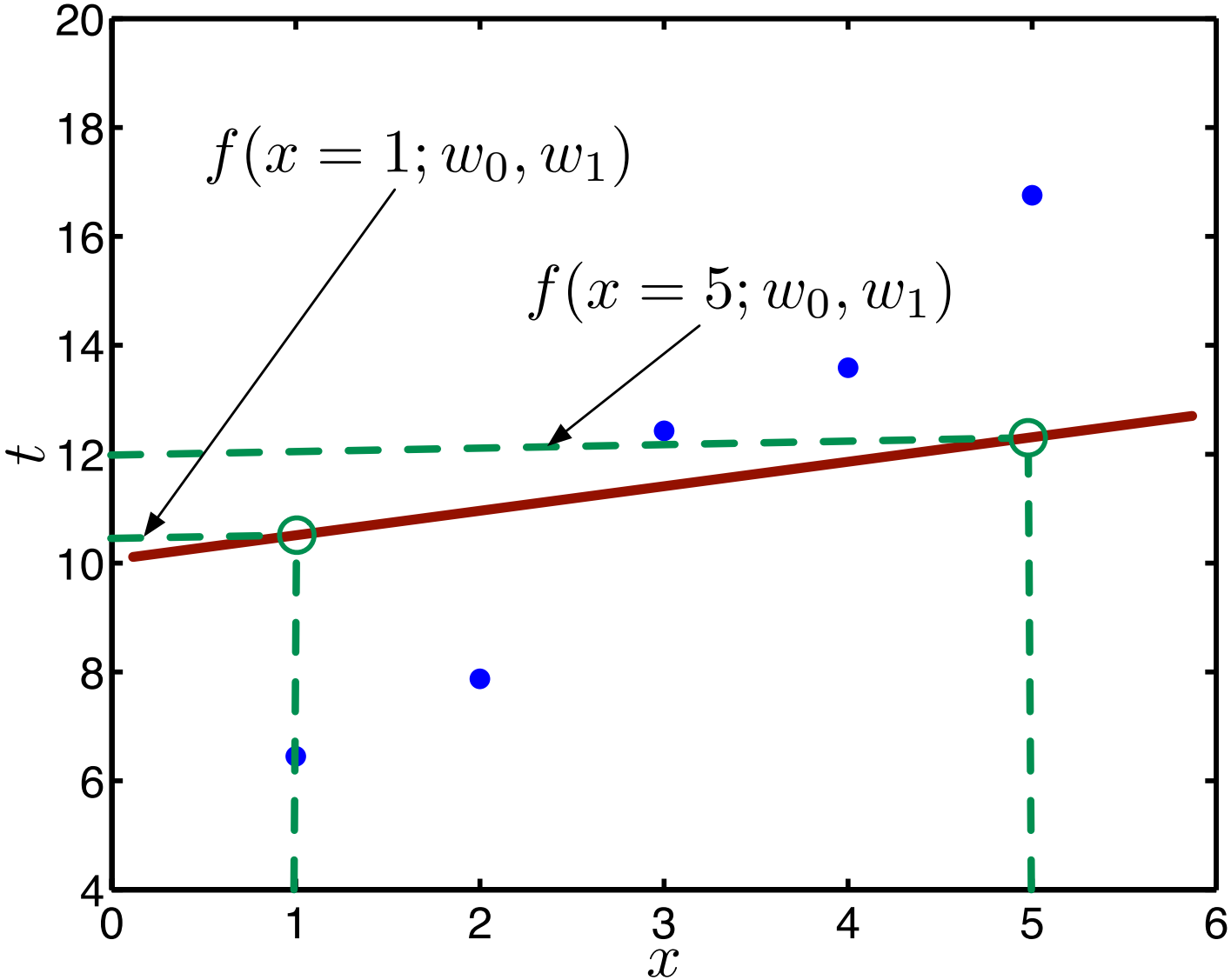
Some different data

Loss



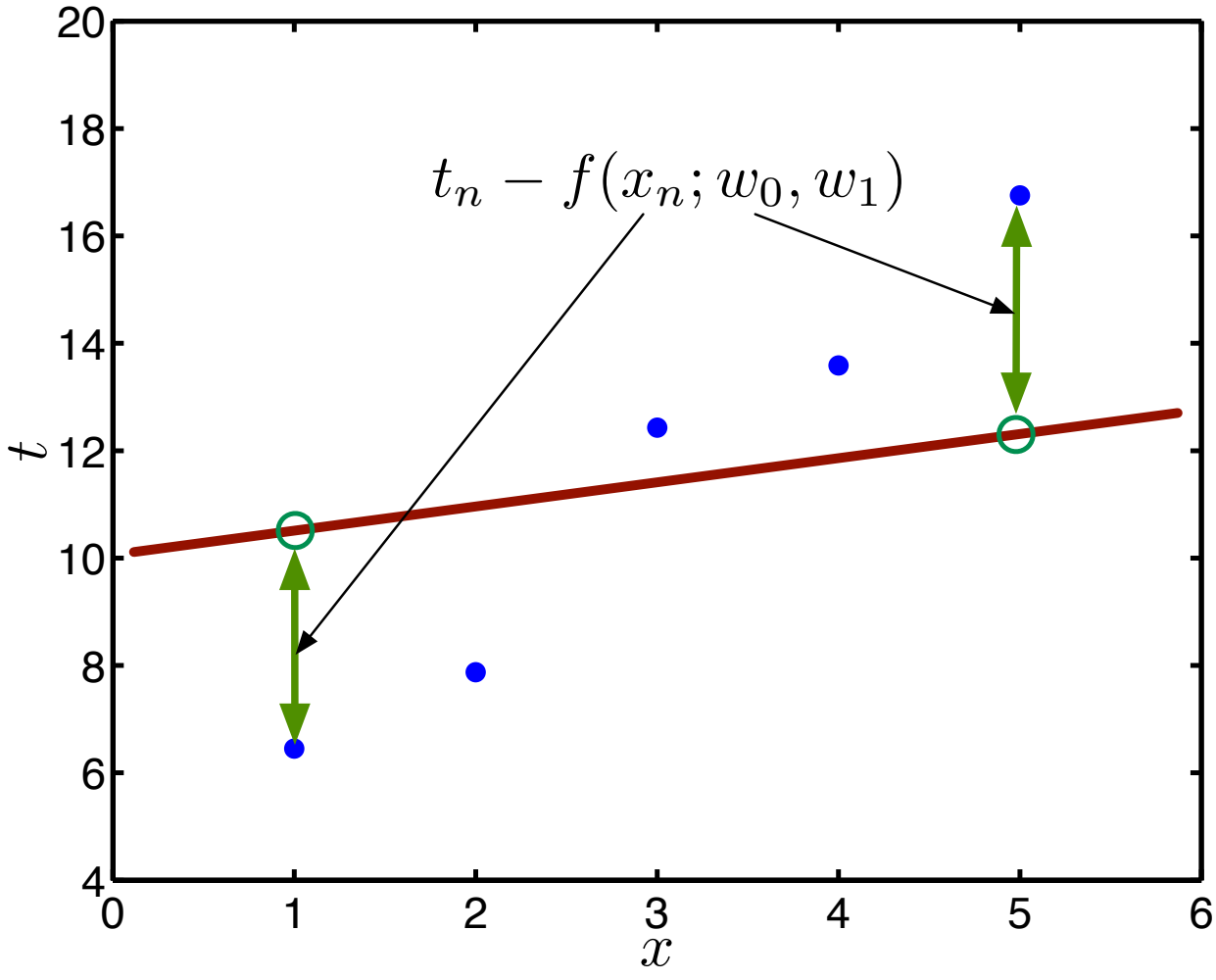
Given w_0 and w_1 you can draw a line

Loss



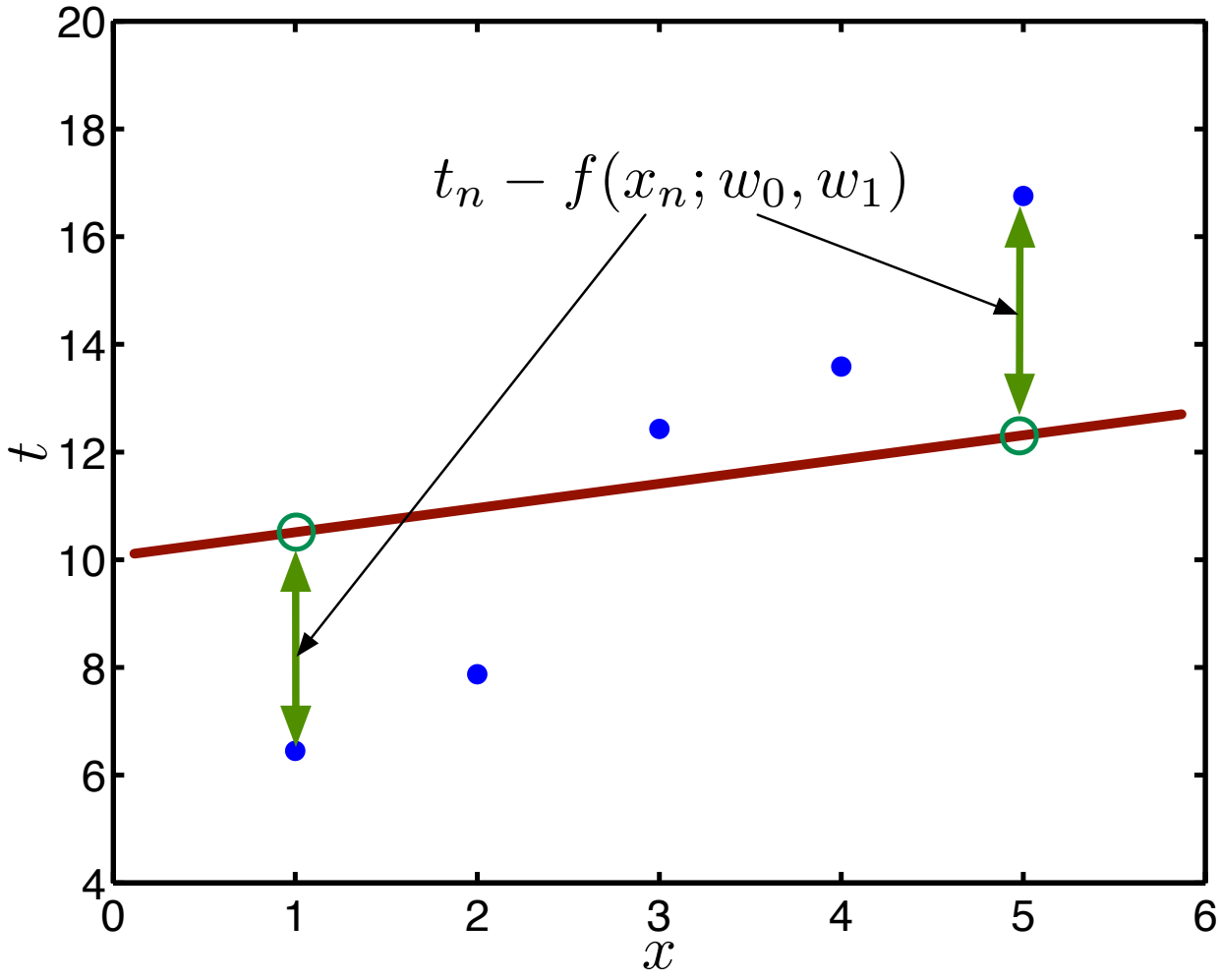
This means that we can compute $f(x_n; w_0, w_1)$ for each x_n

Loss



$f(x_n; w_0, w_1)$ can be compared with the truth, t_n

Loss



$f(x_n; w_0, w_1)$ can be compared with the truth, t_n

$|t_n - f(x_n; w_0, w_1)|$ tells us how *badly* we model (x_n, t_n)

Squared loss

- The *Squared loss* of training point n is defined as:

$$\mathcal{L}_n = (t_n - f(x_n; w_0; w_1))^2$$

Squared loss

- The *Squared loss* of training point n is defined as:

$$\mathcal{L}_n = (t_n - f(x_n; w_0; w_1))^2$$

- It is the squared difference between the true response (winning time), t_n when the input is x_n and the response predicted by the model, $f(x_n; w_0, w_1) = w_0 + w_1 x_n$

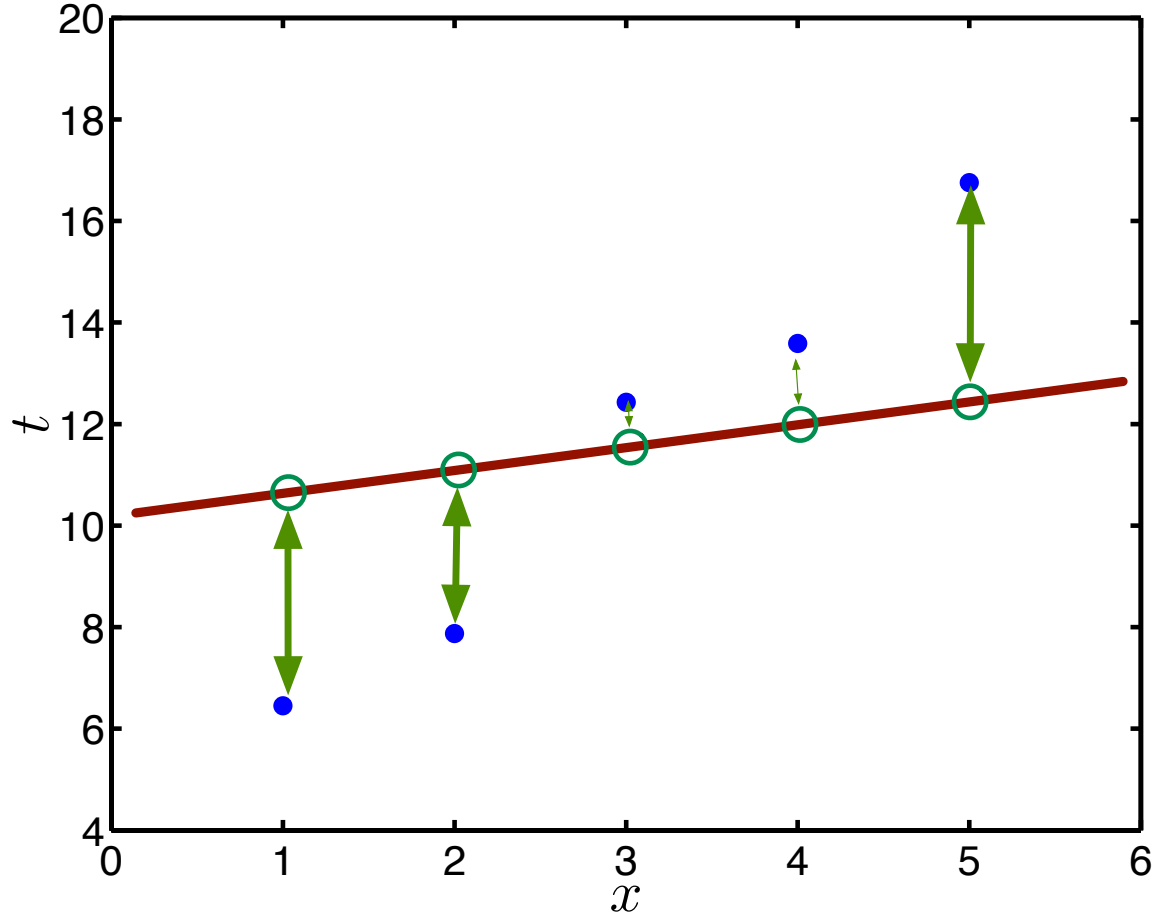
Squared loss

- The *Squared loss* of training point n is defined as:

$$\mathcal{L}_n = (t_n - f(x_n; w_0; w_1))^2$$

- It is the squared difference between the true response (winning time), t_n when the input is x_n and the response predicted by the model, $f(x_n; w_0, w_1) = w_0 + w_1 x_n$
- The lower $\mathcal{L}_n$, the closer the line at x_n passes to t_n

Mean squared loss



Average the loss at each training point to give single figure for all data:

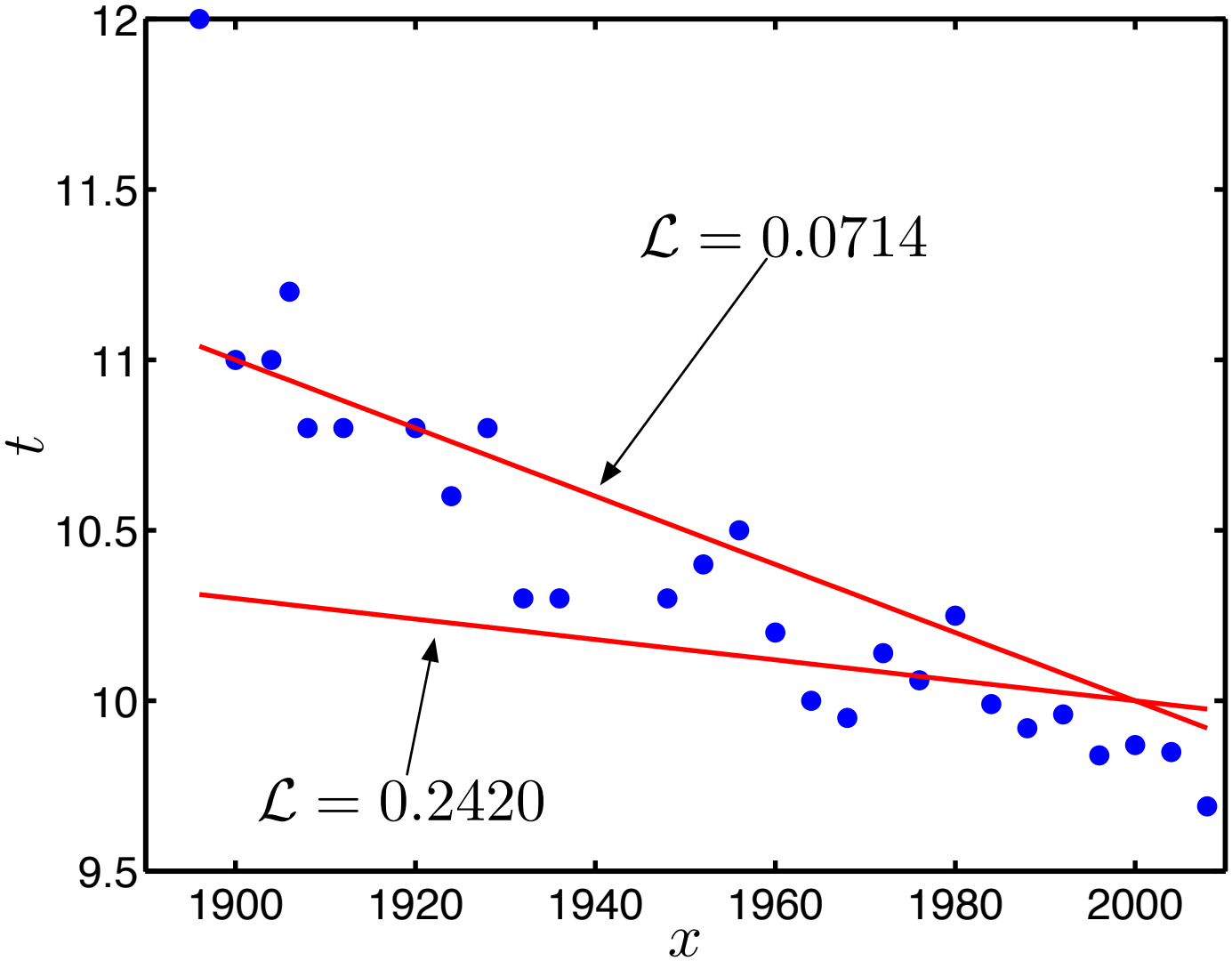
$$\mathcal{L} = \frac{1}{N}\sum_{n=1}^{N}(t_n - f(x_n; w_0, w_1))^2$$

- The average loss:

$$\mathcal{L} = \frac{1}{N} \sum_{n=1}^{N} (t_n - f(x_n; w_0, w_1))^2$$

- $\mathcal{L}$ tells us how good the model is as a function of w_0 and w_1.
  - Remember that lower is better!

Example


12
11.5
11
10.5
10
9.5
1900
1920
1940
1960
1980
2000
t
x
$\mathcal{L} = 0.0714$
$\mathcal{L} = 0.2420$

An optimisation problem

- We've derived an expression for how good the model is for any w_0 and w_1.

$$\mathcal{L} = \frac{1}{N} \sum_{n=1}^{N} (t_n - f(x_n; w_0, w_1))^2$$

An optimisation problem

- We've derived an expression for how good the model is for any w_0 and w_1.

$$\mathcal{L} = \frac{1}{N}\sum_{n=1}^{N}(t_n - f(x_n; w_0, w_1))^2$$

- Want...

$$\underset{w_0, w_1}{\text{argmin}}\ \mathcal{L} = \underset{w_0, w_1}{\text{argmin}}\ \frac{1}{N}\sum_{n=1}^{N}(t_n - f(x_n; w_0, w_1))^2$$

An optimisation problem

- We've derived an expression for how good the model is for any w_0 and w_1.

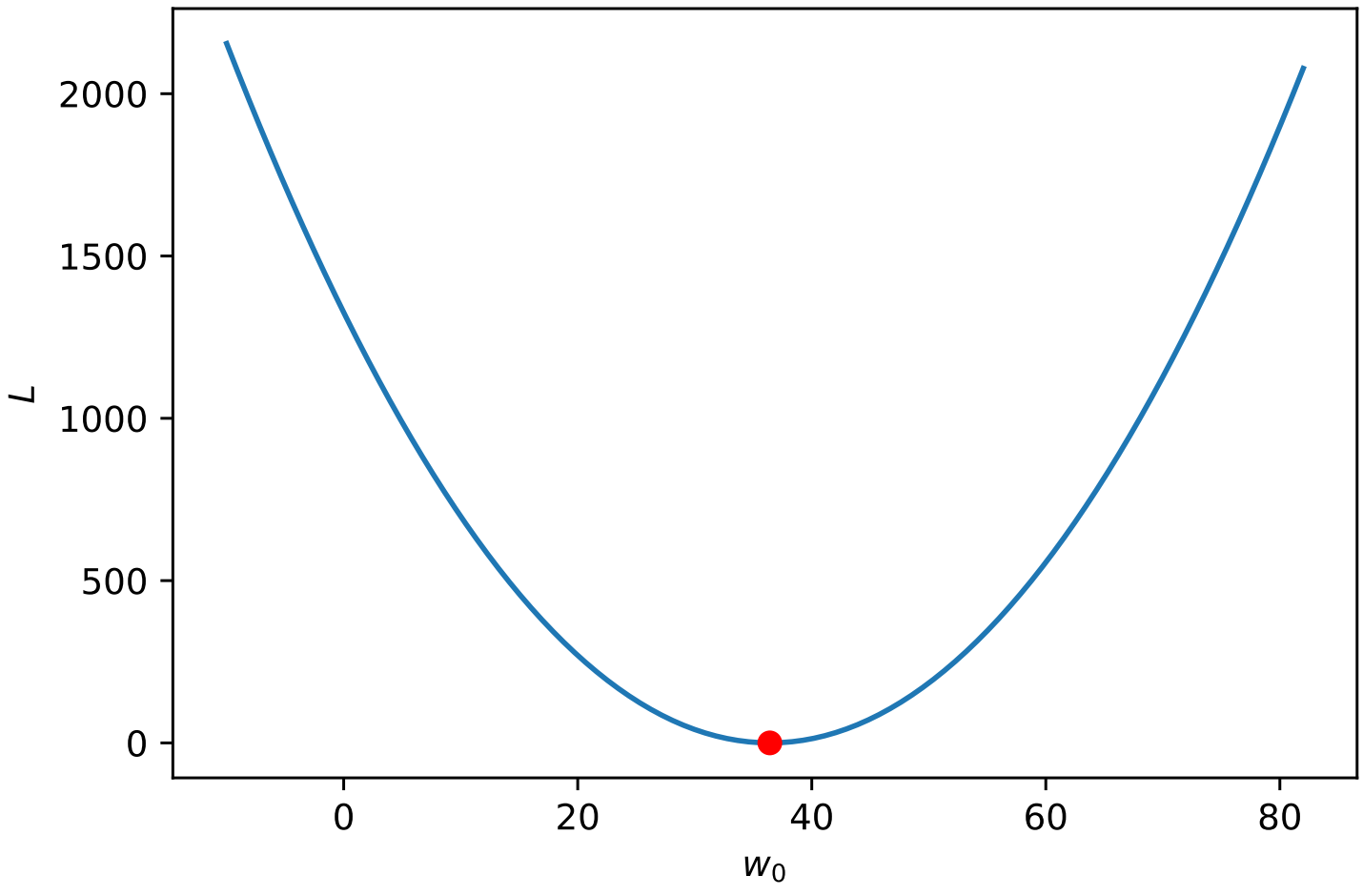
$$\mathcal{L} = \frac{1}{N}\sum_{n=1}^{N}(t_n - f(x_n; w_0, w_1))^2$$

- Want...

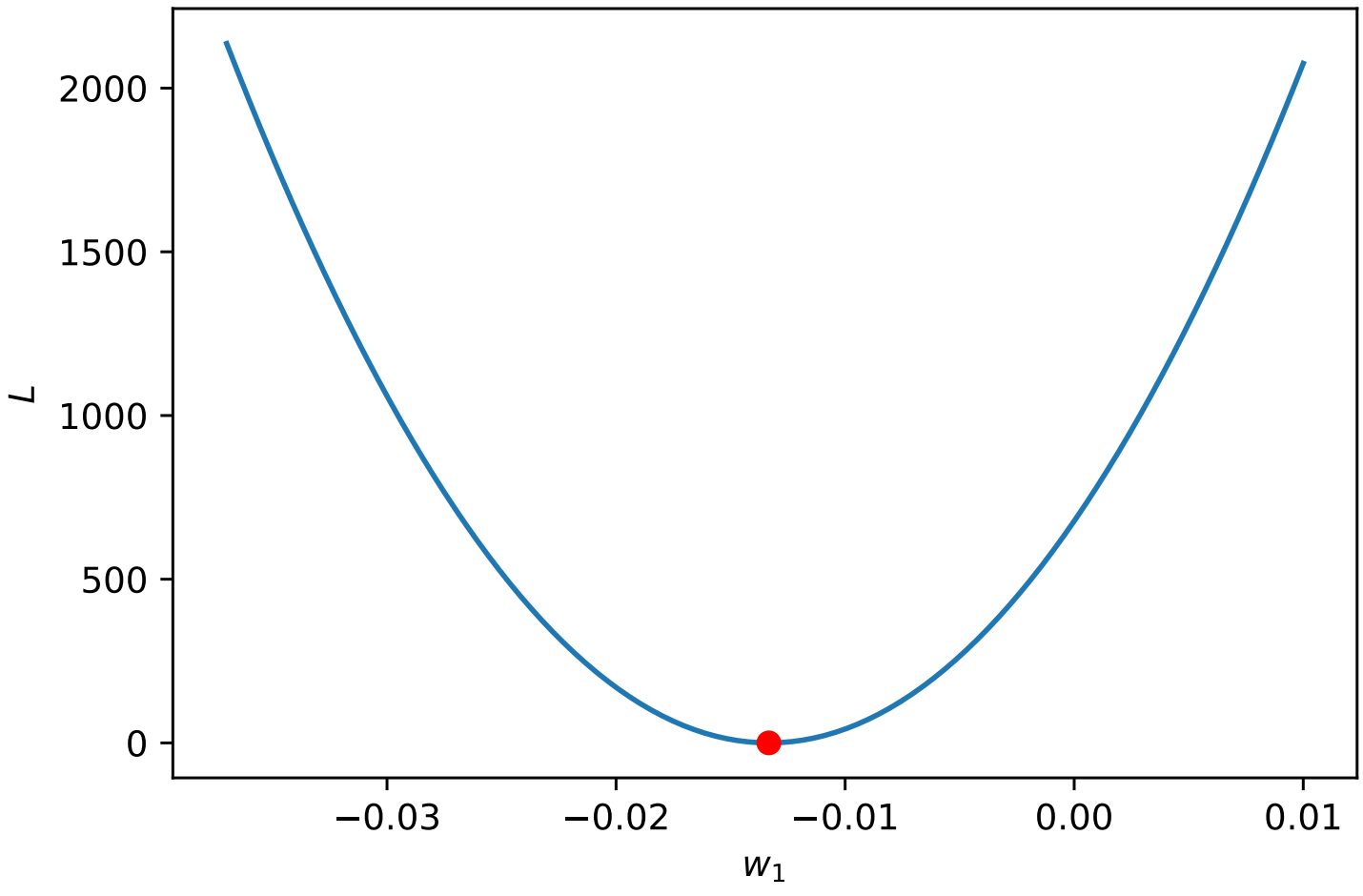
$$\underset{w_0, w_1}{\text{argmin}}\ \mathcal{L} = \underset{w_0, w_1}{\text{argmin}}\ \frac{1}{N}\sum_{n=1}^{N}(t_n - f(x_n; w_0, w_1))^2$$

- Could use trial and error to find a good w_0, w_1 combination.
- Can we be **more efficient**?

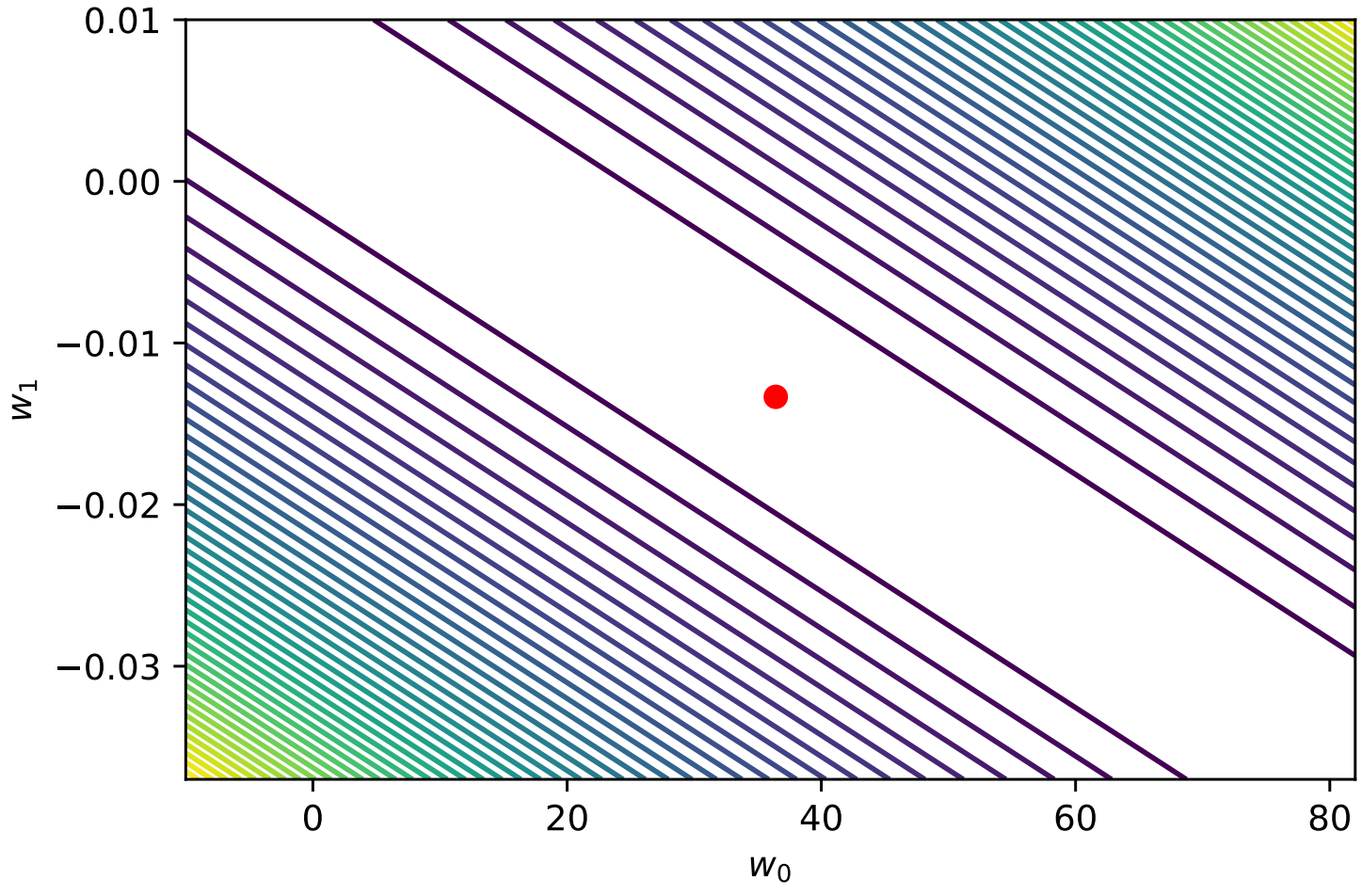
Let's look at the average squared loss in 1D



Let's look at the average squared loss in 1D



Let's look at the average squared loss in 2D



Maths Revision: Gradients

- if $f(x_1, x_2, \ldots)$ is a scalar-valued function of $(x_1, x_2, \ldots)$
- ...then gradient $\nabla f = \begin{pmatrix} \frac{\partial f}{\partial x_1} \\ \frac{\partial f}{\partial x_2} \\ \vdots \end{pmatrix}$

  i.e. vector of partial derivatives

Maths Revision: Gradients

- if $f(x_1, x_2, \ldots)$ is a scalar-valued function of $(x_1, x_2, \ldots)$
- ...then gradient $\nabla f = \begin{pmatrix} \frac{\partial f}{\partial x_1} \\ \frac{\partial f}{\partial x_2} \\ \vdots \end{pmatrix}$

  i.e. vector of partial derivatives
- always points in direction of **steepest ascent**

Gradient Descent

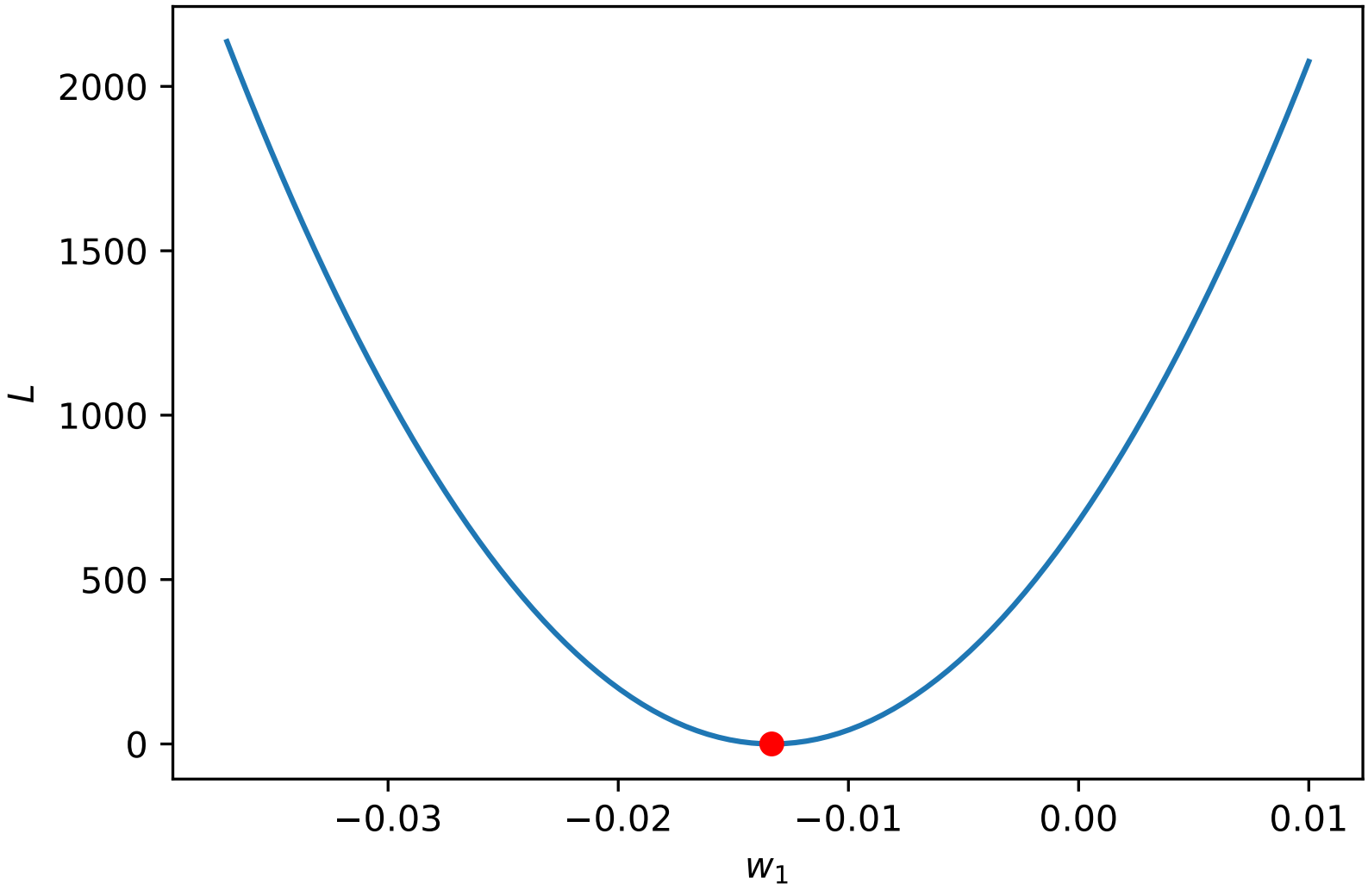
Initial guess w_0 w_1;
Set α;
while *Not converged* **do**
  for *i=0,1* **do**
    $g_i = \frac{\partial L(w_0, w_1)}{\partial w_i}$
  end
  for *i=0,1* **do**
    $w_i = w_i - \alpha g_i$
  end
end

Gradient Descent

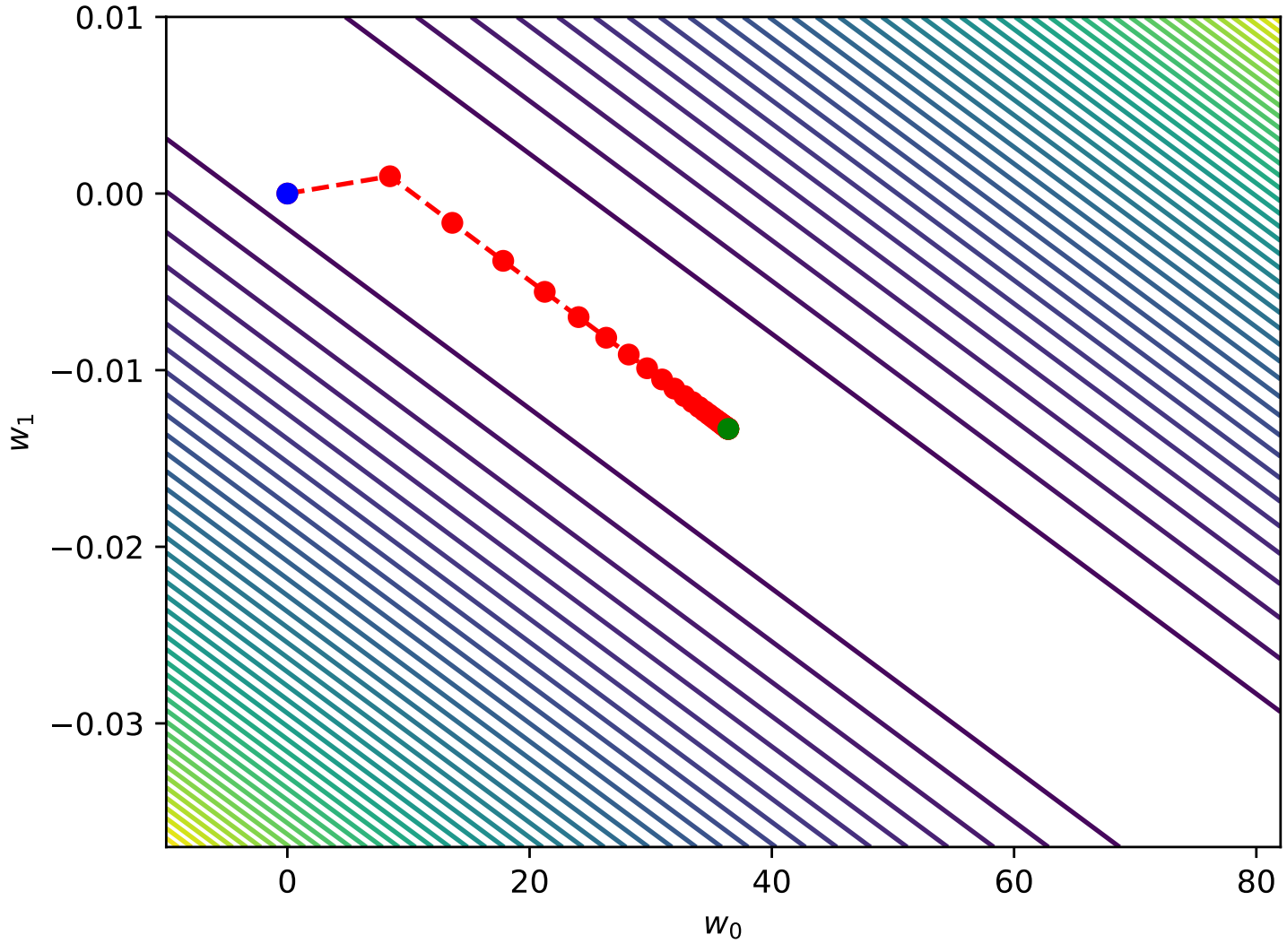
Initial guess w_0 w_1;
Set α;
while *Not converged* **do**
  for *i=0,1* **do**
    $g_i = \frac{\partial L(w_0, w_1)}{\partial w_i}$
  end
  for *i=0,1* **do**
    $w_i = w_i - \alpha g_i$
  end
end

while *Not converged* **do**
  $\mathbf{w} = \mathbf{w} - \alpha \nabla L(w_0, w_1)$
end

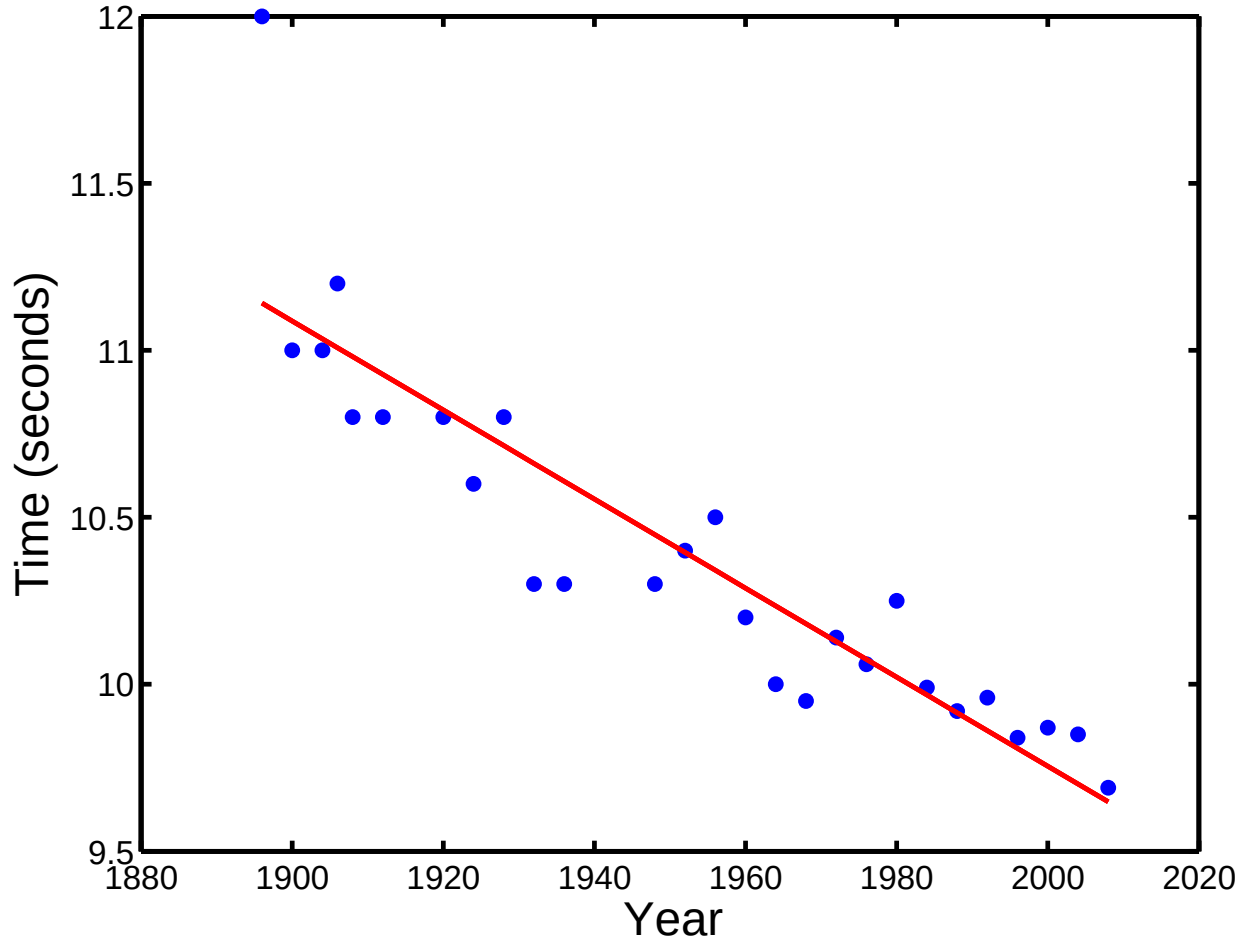
Gradient Descent



Gradient Descent



The model



$$t = 36.416 - 0.0133x$$

Our prediction

- We want to predict the winning time at London 2012.
- Substitute $x = 2012$ into our model.

$$\begin{aligned} t &= 36.416 - 0.0133x \\ t_{2012} &= 36.416 - 0.0133 \times 2012 \\ t_{2012} &= 9.5947\ s \end{aligned}$$

- Based on our modelling assumptions and the previous data, we predict a winning time of 9.5947 seconds.

Summary

- Introduced some ideas about modelling.
- Found some data.
- Derived a way of saying how good a model is.
- Found an expression for the best model.
- Used this to fit a model to the Olympic data.
- Made a prediction for the winning time in 2012.

Assumptions

Our Assumptions

1. **That there exists a relationship between Olympic year and winning time.**

Are they any good?

Assumptions

Our Assumptions

1. That there exists a relationship between Olympic year and winning time.

Are they any good?

1. Is the relationship really between Olympic year and time?

Assumptions

Our Assumptions

1. That there exists a relationship between Olympic year and winning time.
2. **That this relationship is linear (i.e. a straight line).**

Are they any good?

1. Is the relationship really between Olympic year and time?

Assumptions

Our Assumptions

1. That there exists a relationship between Olympic year and winning time.
2. That this relationship is linear (i.e. a straight line).

Are they any good?

1. Is the relationship really between Olympic year and time?
2. Seems a bit simple? Does the line go through all of the points?

Assumptions

Our Assumptions

1. That there exists a relationship between Olympic year and winning time.
2. That this relationship is linear (i.e. a straight line).
3. **This this relationship will continue into the future.**

Are they any good?

1. Is the relationship really between Olympic year and time?
2. Seems a bit simple? Does the line go through all of the points?

Assumptions

Our Assumptions

1. That there exists a relationship between Olympic year and winning time.
2. That this relationship is linear (i.e. a straight line).
3. This this relationship will continue into the future.

Are they any good?

1. Is the relationship really between Olympic year and time?
2. Seems a bit simple? Does the line go through all of the points?
3. Forever? Negative winning times?

Assumptions

Our Assumptions

1. That there exists a relationship between Olympic year and winning time.
2. That this relationship is linear (i.e. a straight line).
3. This this relationship will continue into the future.

Are they any good?

1. Is the relationship really between Olympic year and time?
2. Seems a bit simple? Does the line go through all of the points?
3. Forever? Negative winning times?

The model is 'wrong' but it might still be useful! How useful depends on the questions we wish to answer.