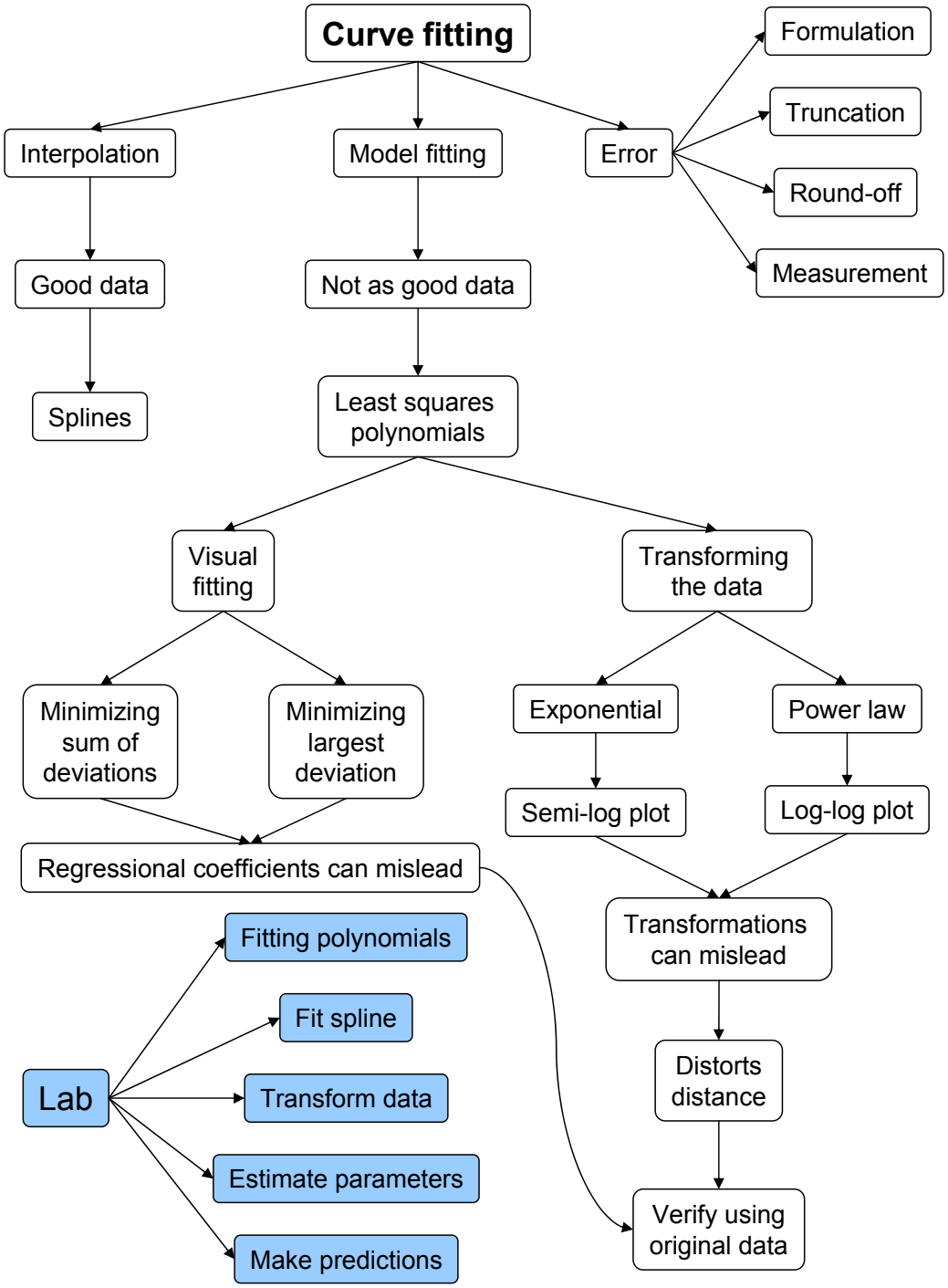
Formulating models

- We can use information from data to formulate mathematical models
- These models rely on assumptions about the data or data not collected
- Different assumptions will lead to different models.


Curve fitting
Interpolation
Model fitting
Error
Formulation
Truncation
Round-off
Measurement
Good data
Not as good data
Splines
Least squares polynomials
Visual fitting
Transforming the data
Minimizing sum of deviations
Minimizing largest deviation
Exponential
Power law
Semi-log plot
Log-log plot
Regressional coefficients can mislead
Transformations can mislead
Distorts distance
Verify using original data
Lab
Fitting polynomials
Fit spline
Transform data
Estimate parameters
Make predictions

An influenza example

Consider an influenza pandemic moving through a population. Assumptions we could make could involve:

- The heterogeneous mixing of the population is proportional to the local population density
- The urban vs. rural environment
- Or we could ignore population heterogeneity altogether.

Model fitting

- Collected data can determine unknown parameters in our model
- We select the curve from each model that "best fits" the data
- We then choose the most appropriate model for our particular situation.

Complexity may be a problem

- A three-dimensional model for the spread of measles might involve partial differential equations for the movement of infectious droplets in three spatial dimensions, plus one temporal dimension
- This will be enormously complicated
- The equations may not even be solvable.

Interpolation

- There is little hope for constructing a master model that can be solved and analysed analytically
- Or there may be so many variables that one would not even attempt to construct an explicit model
- In such cases the data must be used to determine values outside the collected range (interpolation).

Possible tasks for data analysis

1. Fitting a selected model type or types to the data
2. Choosing the most appropriate model from competing types that have been fitted
   - Eg. is the best fitting exponential model a better model than the best-fitting least-squares model?
3. Making predictions from the collected data (interpolation and extrapolation).

Our three tasks

In Task 1, the precise meaning of "best" model must be identified and the resulting mathematical problem resolved

In Task 2, a criterion is needed for computing models of different types

In Task 3 criteria must be established for determining how to make predictions in between the observed data points.

Model fitting → Explain data

When model-fitting,

- We strongly suspect a relationship of a particular type
- We are willing to accept some deviation between the model and the collected data points
- We want a model that satisfactorily *explains* the situation under investigation.

Interpolation → predicting

When interpolating,

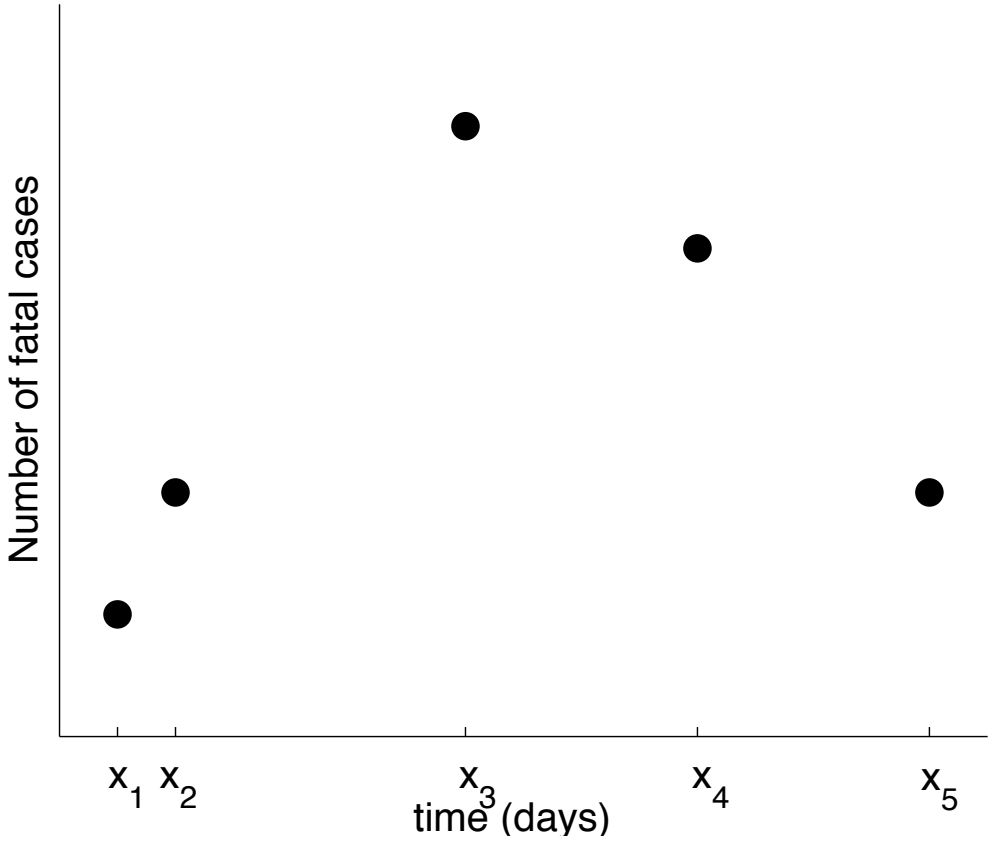
- We are strongly guided by the data that have been carefully collected
- We seek a curve that captures the trend of the data
- We want to *predict* in between the data points.

Model fitting vs. interpolation

- In all situations we may ultimately want to make predictions from the model
- When model fitting, we emphasize the *proposed models* over the data
- When interpolating, we place greater confidence in the *collected data* and attaches less significance to the form of the model.

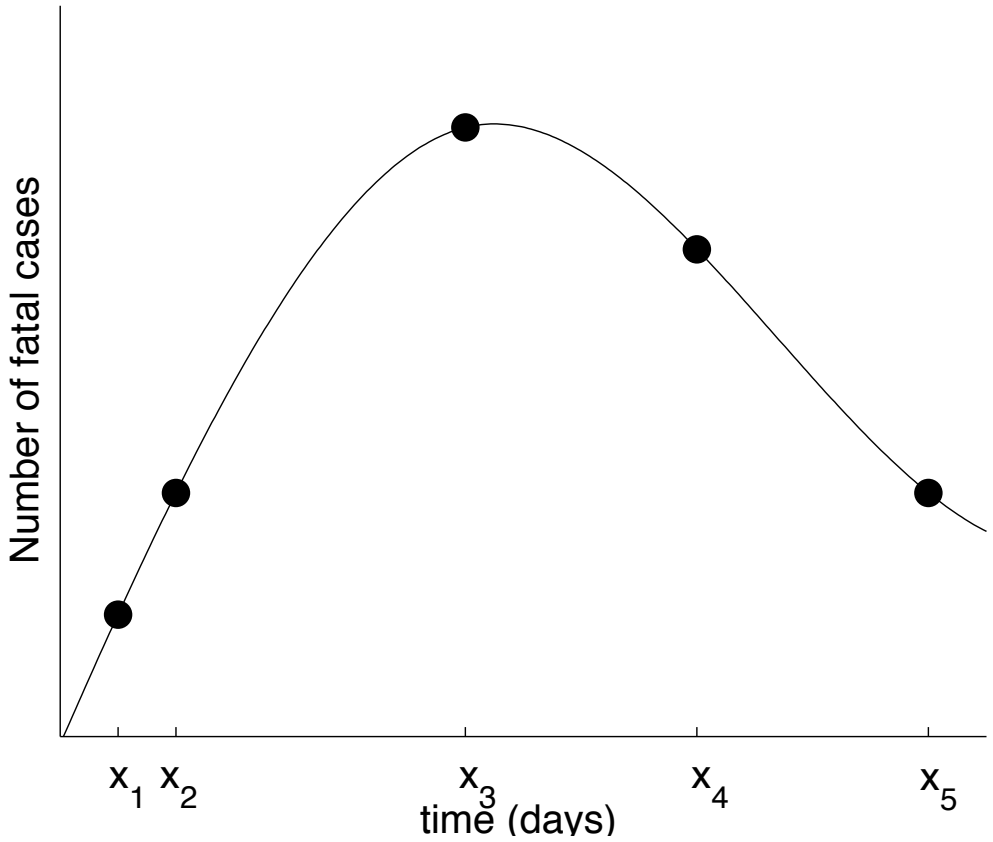
1918 pandemic influenza

- Consider this data of fatal cases of 1918 influenza in Philadelphia over a number of days
- How can we fit curves to this data and make predictions?



If we have faith in the data...

- *Spline interpolation* passes a smooth curve through the points
- Captures the data trend over the observation range
- We'll study this in more detail shortly.

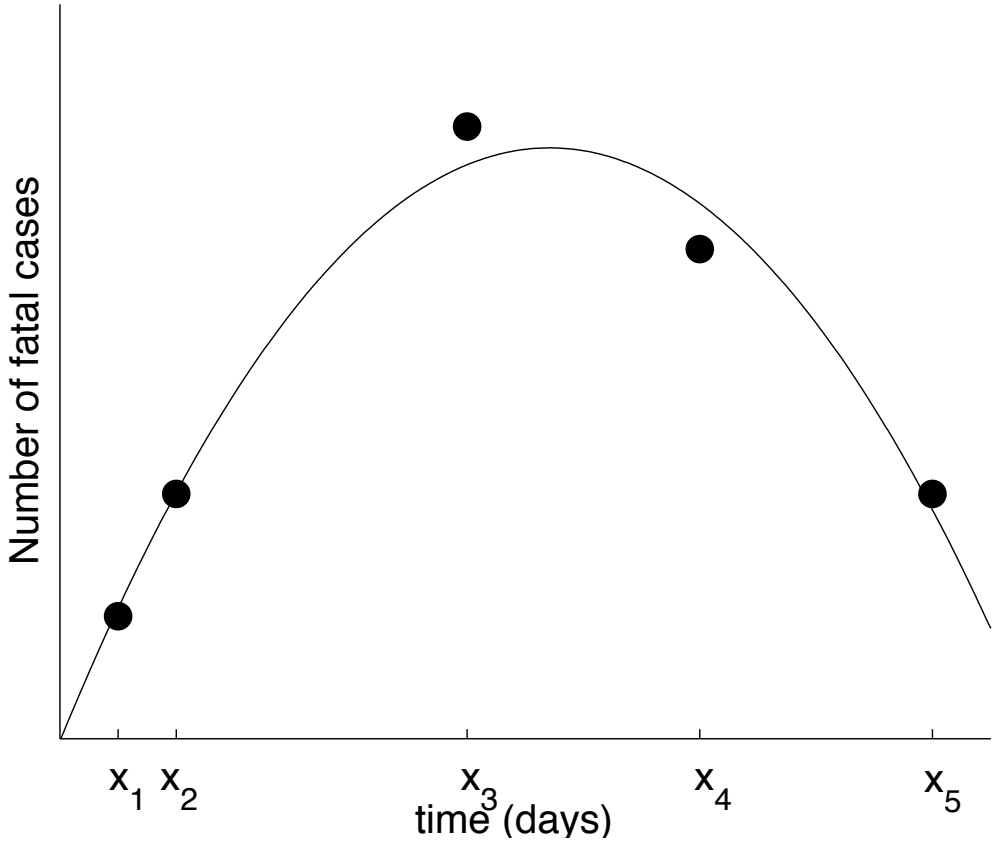


Maybe a parabolic trend...?

- A parabolic model would be of the form

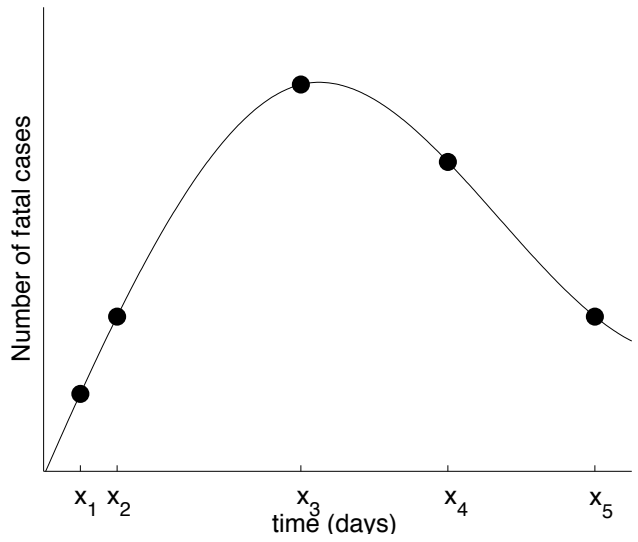
  $y=C_1x^2+C_2x+C_3$

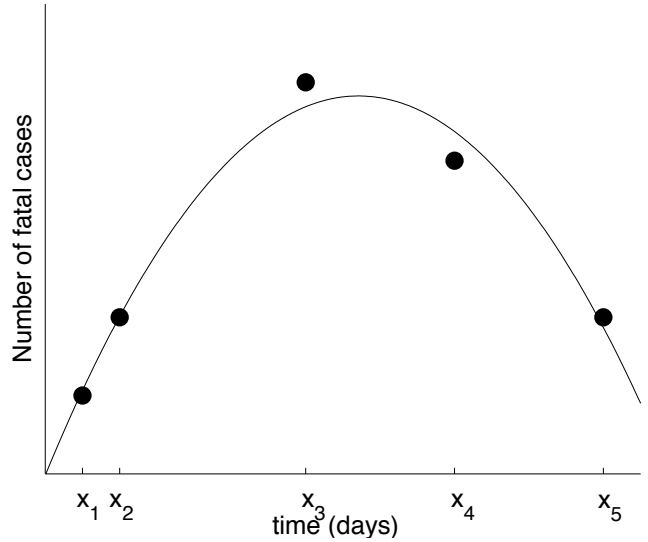
- The data would be used to determine C_1, C_2 and C_3
- In such a way that selects the "best" parabola.



Splines vs. parabolas

- We don't care if the parabola deviates from some or all data points
- Outside the range of data points, the curves may vary significantly
- Beyond x_5, the predictions will be quite different.





Sources of error

- We need to be informed about sources of possible error
- Otherwise undue confidence may be placed in intermediate results
- This will cause faulty decisions in subsequent steps.

Classification of errors

1. Formulation error
2. Truncation error
3. Round-off error
4. Measurement error.

Formulation errors

- Come from assuming certain variables are negligible
- Or from simplifying relationships among variables
- E.g. ignoring spatial heterogeneity → we may be neglecting important relationships among individuals that facilitate disease transmission

Formulation errors are present in even the best models.

Truncation errors

- Come from the numerical method used to solve a mathematical problem
- E.g. truncating polynomial approximations

$$e^x = 1 + x + \frac{x^2}{2} + \frac{x^3}{6} + \frac{x^4}{24} + \cdots$$

$$\approx 1 + x + \frac{x^2}{2} + \frac{x^3}{6} + \frac{x^4}{24}$$

Computers and calculators do this all the time.

Round-off errors

- Come from using a finite digit machine for computation
- Eg. the number 1/3 is represented by 0.33333333 in 8-digit arithmetic
- 3 x 1/3 =0.99999999, rather than 1
- The error of 10^{-8} is due to round-off

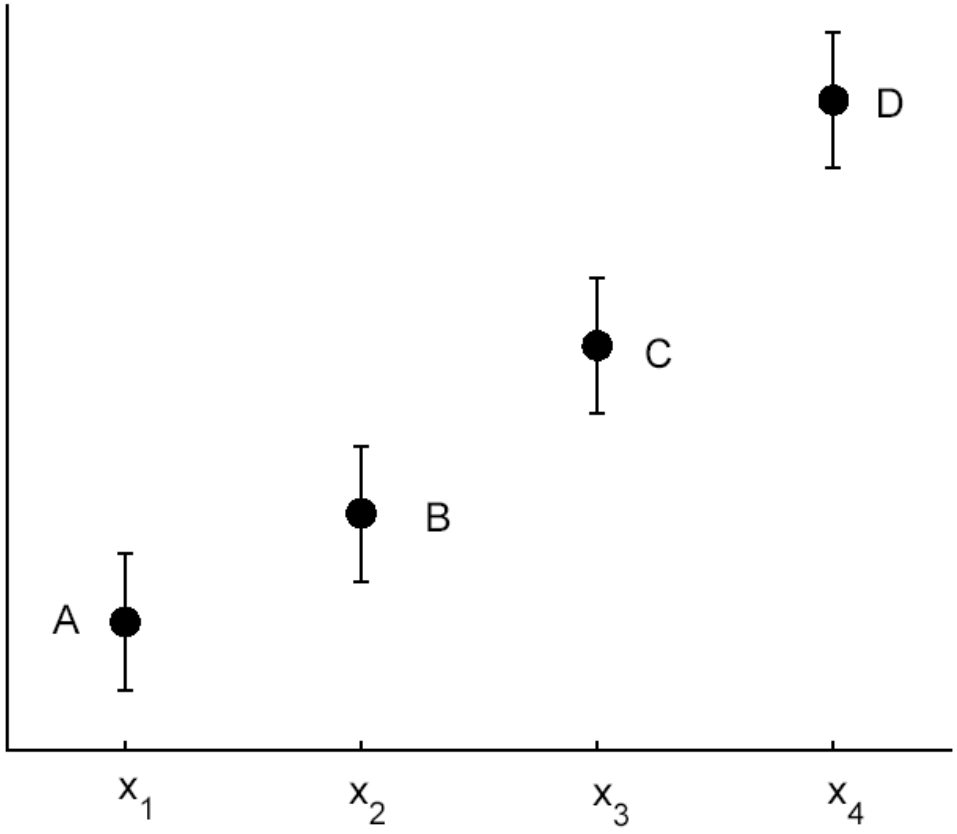
We must always expect round-off errors to be present.

Measurement errors

- Come from imprecision in the data collection
- May include human errors in recording or reporting the data, or the actual physical limitations of the laboratory equipment.

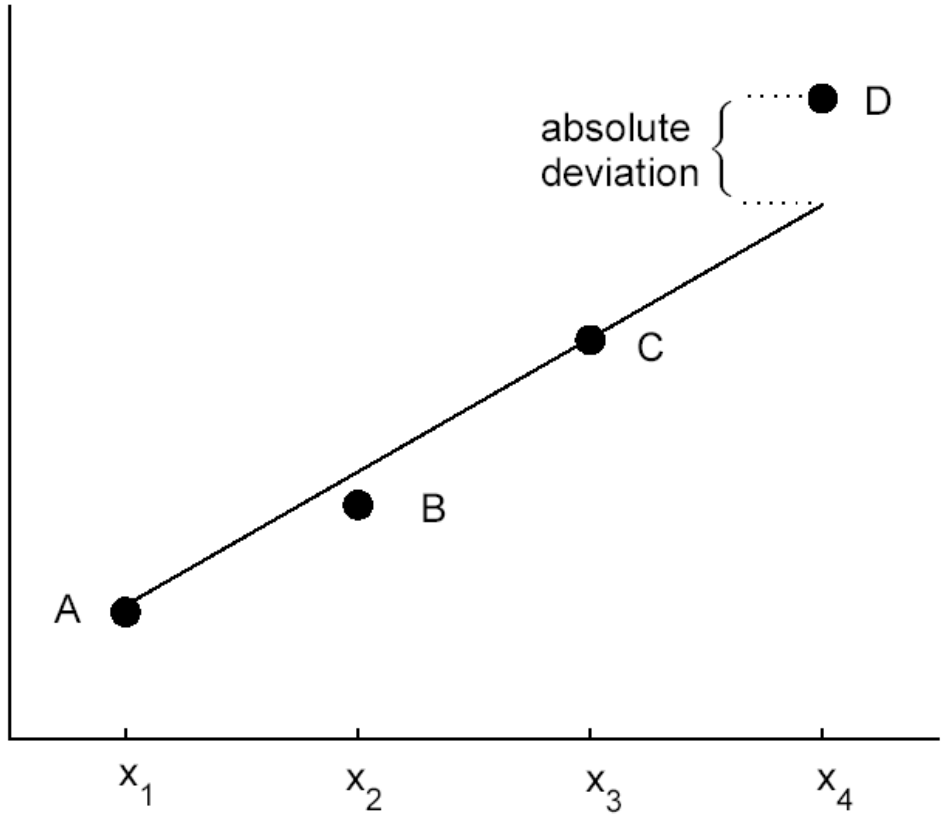
Confidence in the data

- Data points can be thought of an an interval of confidence
- To "best fit" a line through these points, we might want to minimize the sum of these deviations.



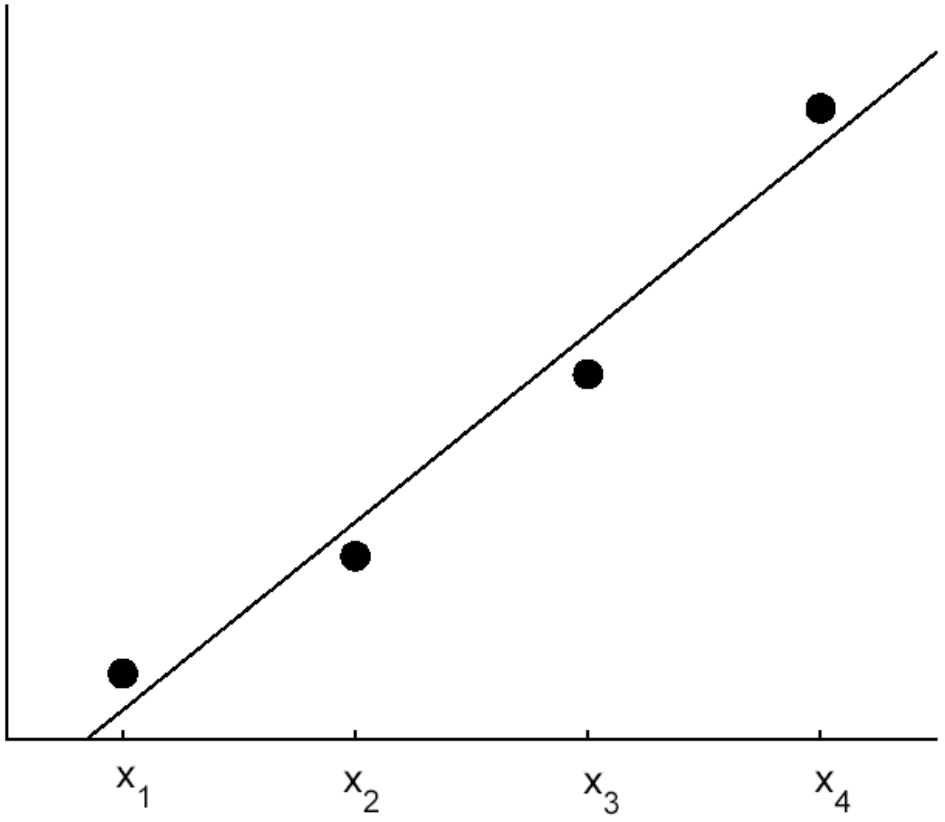
Minimizing the sum of deviations

- A, B and C are quite close to the line
- But D is very far
- If we have confidence in D's accuracy, we should be concerned about predictions made from this model near D.



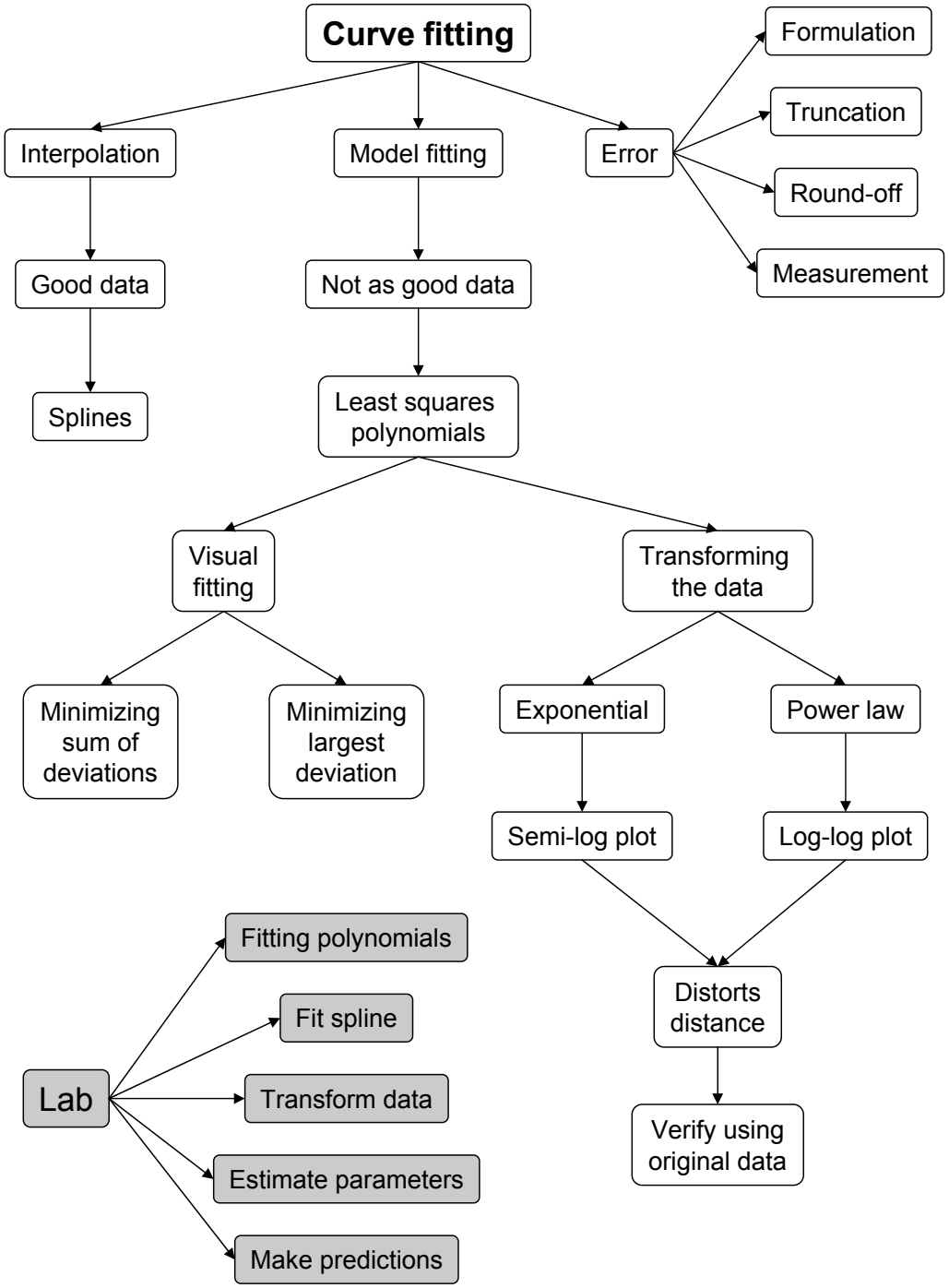
Minimizing the largest deviation

- In this case no point is exactly on the line
- But no point is too far from it either
- A visual inspection suggests this is a pretty good fit.



Simple models may be sufficient

- These visual methods for fitting a line to data points may appear imprecise
- However, the methods are often quite compatible with the accuracy of the modelling process itself
- The grossness of the assumptions and the imprecision involved in the data collection may not warrant a more sophisticated analysis.

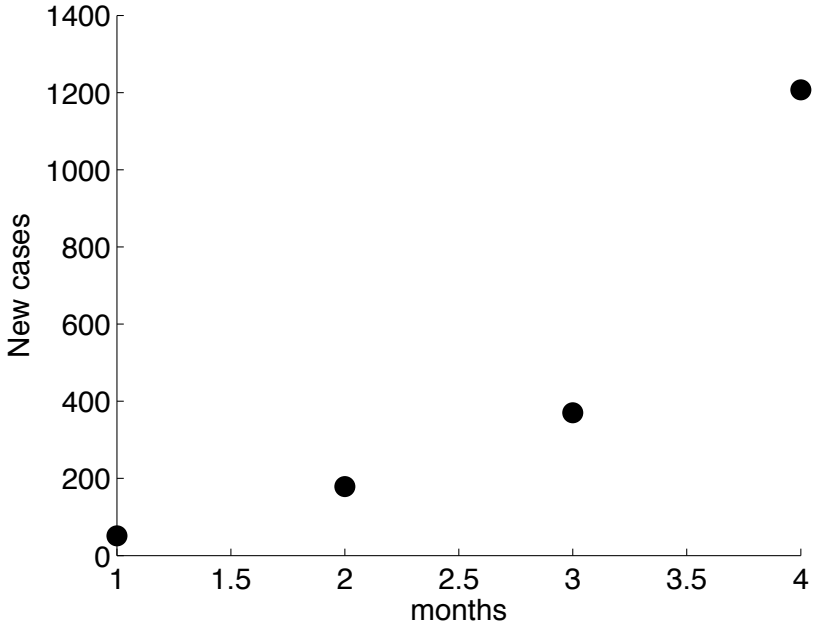

Curve fitting
Interpolation
Model fitting
Error
Formulation
Truncation
Round-off
Measurement
Good data
Not as good data
Splines
Least squares polynomials
Visual fitting
Transforming the data
Minimizing sum of deviations
Minimizing largest deviation
Exponential
Power law
Semi-log plot
Log-log plot
Distorts distance
Verify using original data
Lab
Fitting polynomials
Fit spline
Transform data
Estimate parameters
Make predictions

Transforming the data

- To fit curves other than lines, we have to transform the data
- Consider the number of new cases of HIV infections detected in 1981:

Month	Jan	Feb	Mar	Apr
New cases	51	179	370	1207

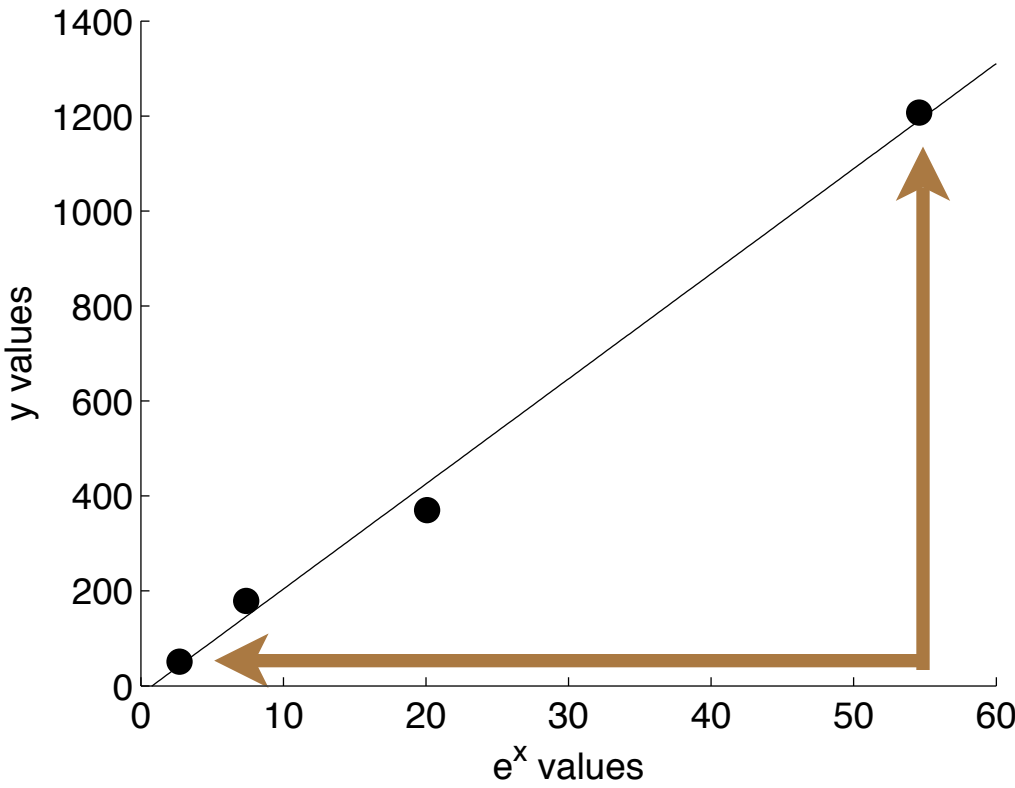
- What sort of curve should we fit?



Maybe exponential?

- Let's fit models of the form $y = Ce^x$
- Plot y vs. e^x
- The slope of the line is

$$C = \frac{1207 - 51}{54.6 - 2.7} \approx 22.3\,.$$



x	1	2	3	4
e^x	2.7	7.4	20.1	54.6

Or transform the data first

- Take logs of both sides

$$y = Ce^x$$

$$\ln y = \ln[Ce^x]$$

$$\ln y = \ln C + \ln[e^x]$$

$$\ln y = \ln C + x$$

$\ln(ab) = \ln(a)+\ln(b)$

$\ln$ *and* e *are inverses*

- Thus if we plot ln *y* vs. *x*, the intercept should be ln *C*.

Plotting transformed points

- The transformed data is

x	1	2	3	4
ln$y$	3.932	5.167	5.914	7.096

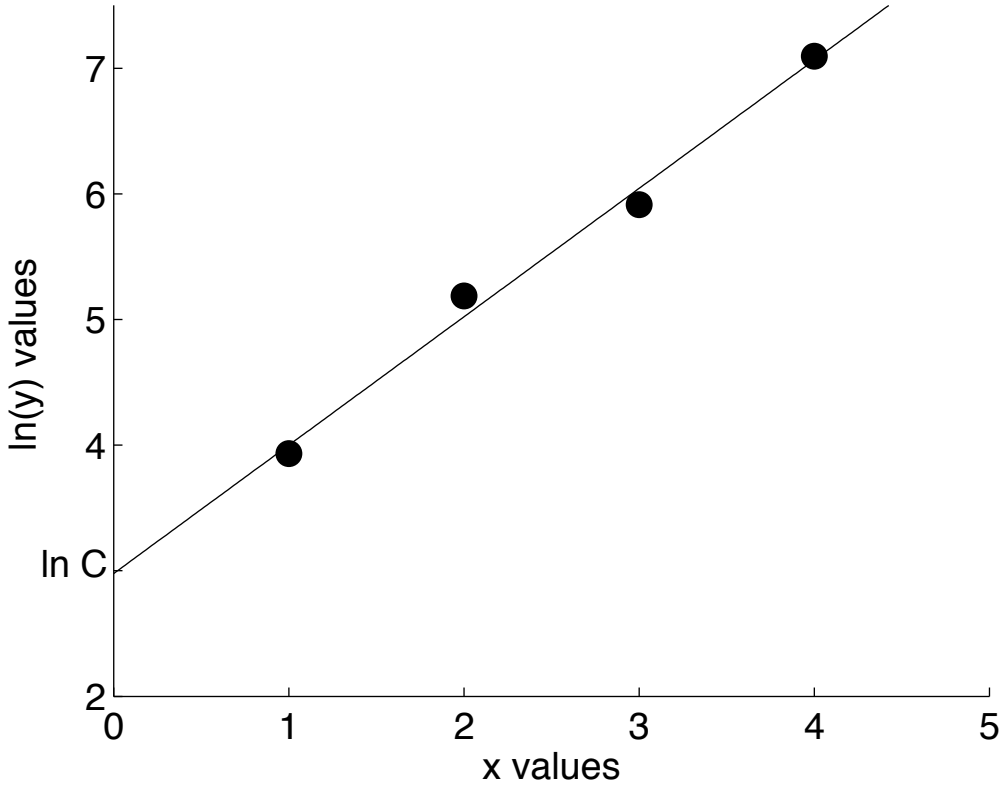
- From the intercept,

$$\ln C \approx 2.9776$$

$$C \approx e^{2.9776}$$

$$\approx 19.6$$

- Which is more accurate?



Other transformations

- Whenever we have unknowns in the power, use a logarithm
- E.g. the power law

$$y = x^a$$

$$\ln y = \ln(x^a)$$

$$\ln y = a \ln x$$

$\ln(b^c) = c \ln(b)$

- A linear relationship between ln *y* and ln *x*
- The slope is *a*.

Avian influenza

- Consider the number of birds needing to be culled per positive case of bird flu:

x	3	7	20	148
y	8	65	549	36300

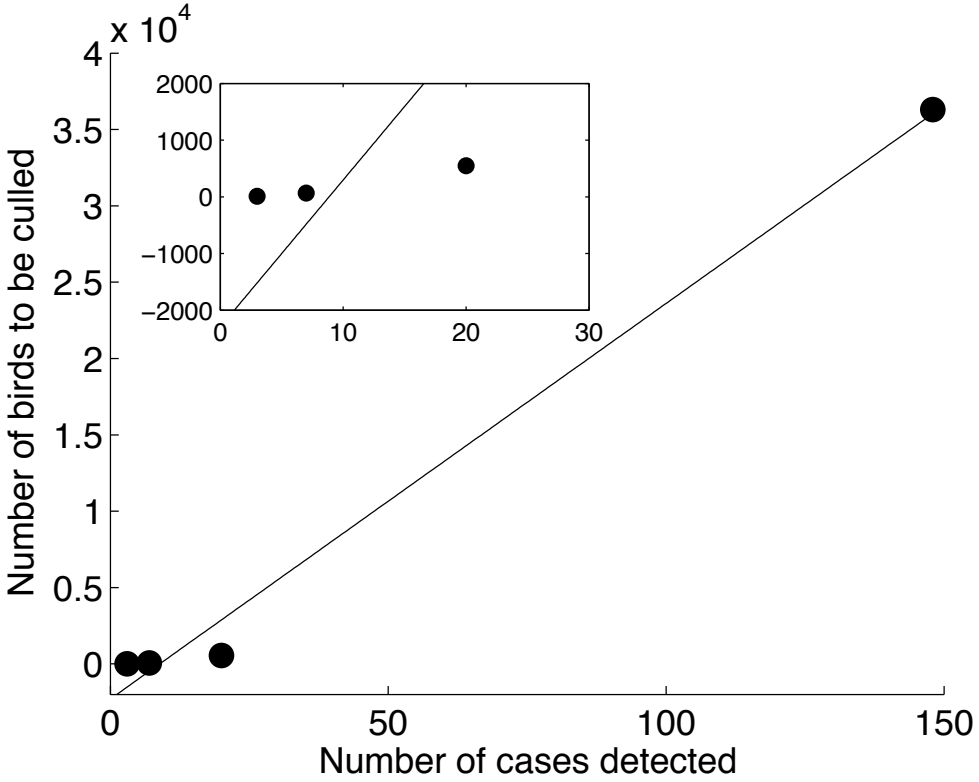
- Fitting a line of best fit using linear regression (eg your calculator) gives

$$r=0.9956$$

- Seems pretty good.

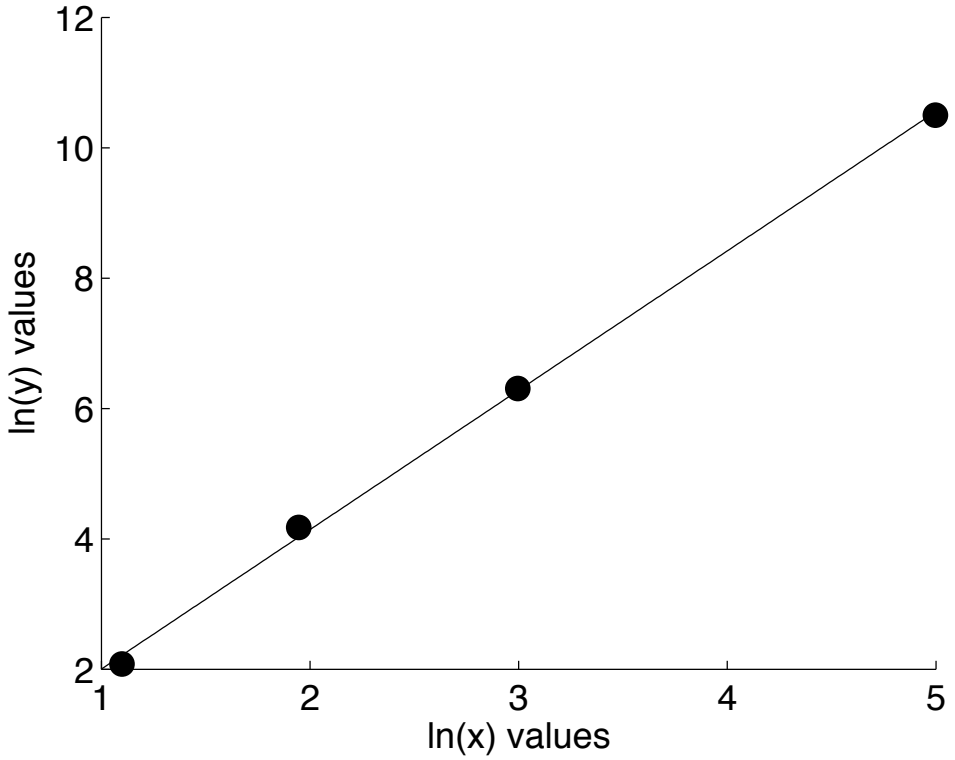
A very poor fit

- But plotting these data against the line of best fit tells a different story
- Look at the inset: this is a terrible fit.



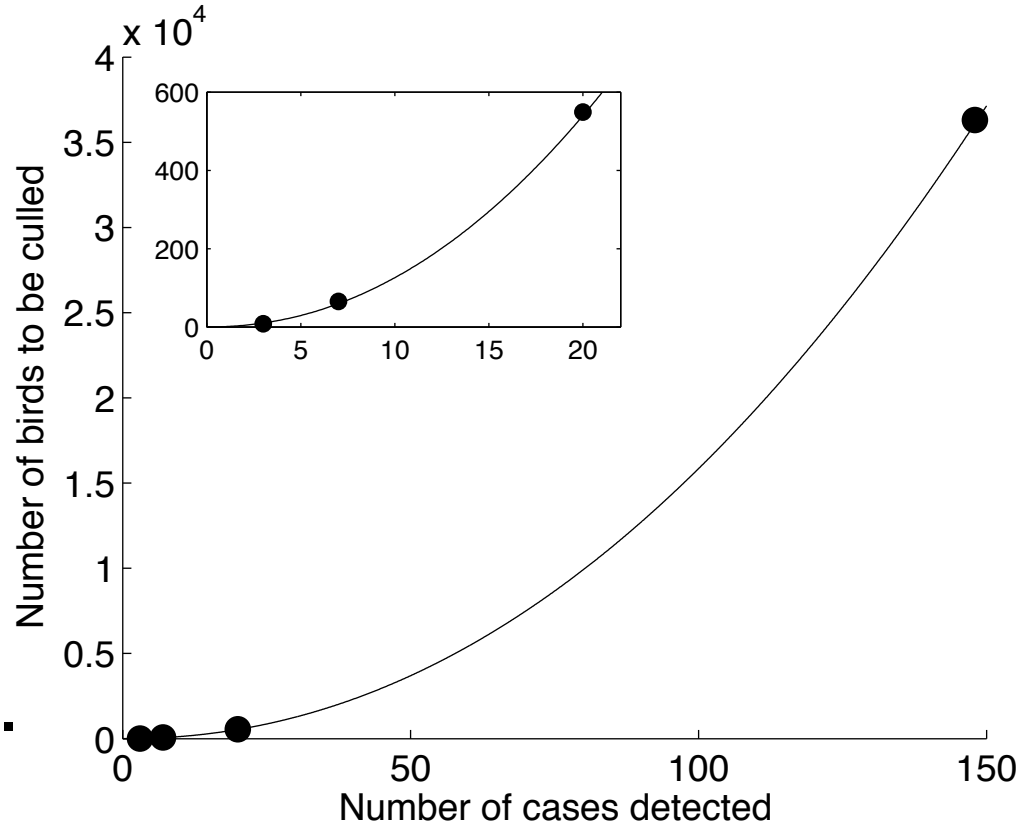
A different relationship

- Instead, let's try $y=x^a$
- When we plot ln y vs ln x, we find r=0.9996 and the slope is 2.1496
- Thus a $\approx$ 2.1.



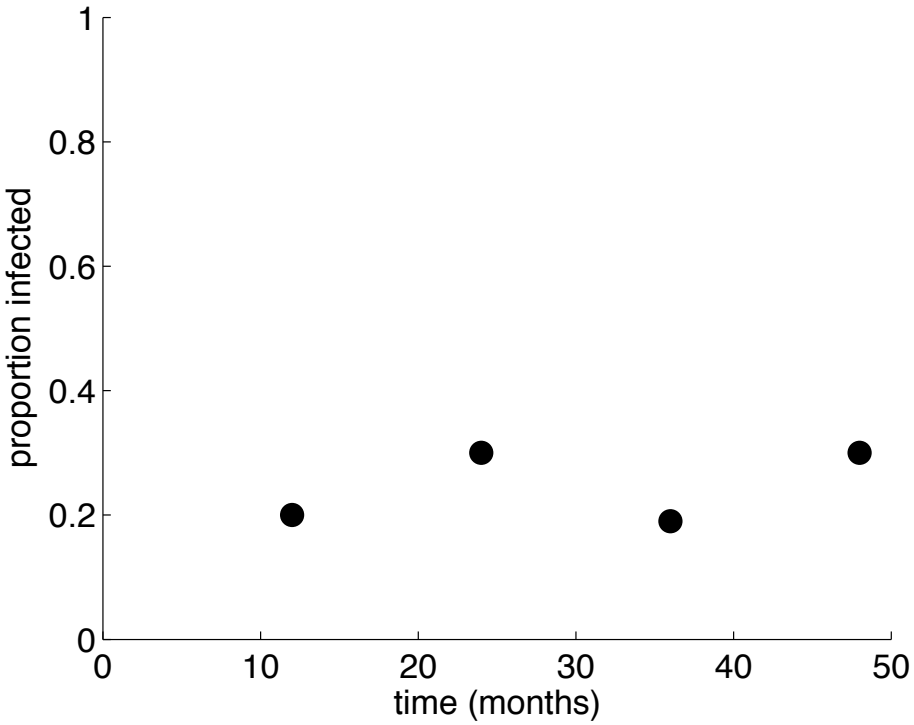
A much better fit

- Now plot $y=x^{2.1}$
- This is a really good fit
- Fitting the curve to the *original* data is when we make decisions about which curve is best.



Biannual influenza outbreaks

- Consider this seasonal influenza data
- Suppose we suspect the data fit a model of the form $y = Ce^{1/x}$
- We want to find the "best" C.



Variations are condensed

- Transform the data as before

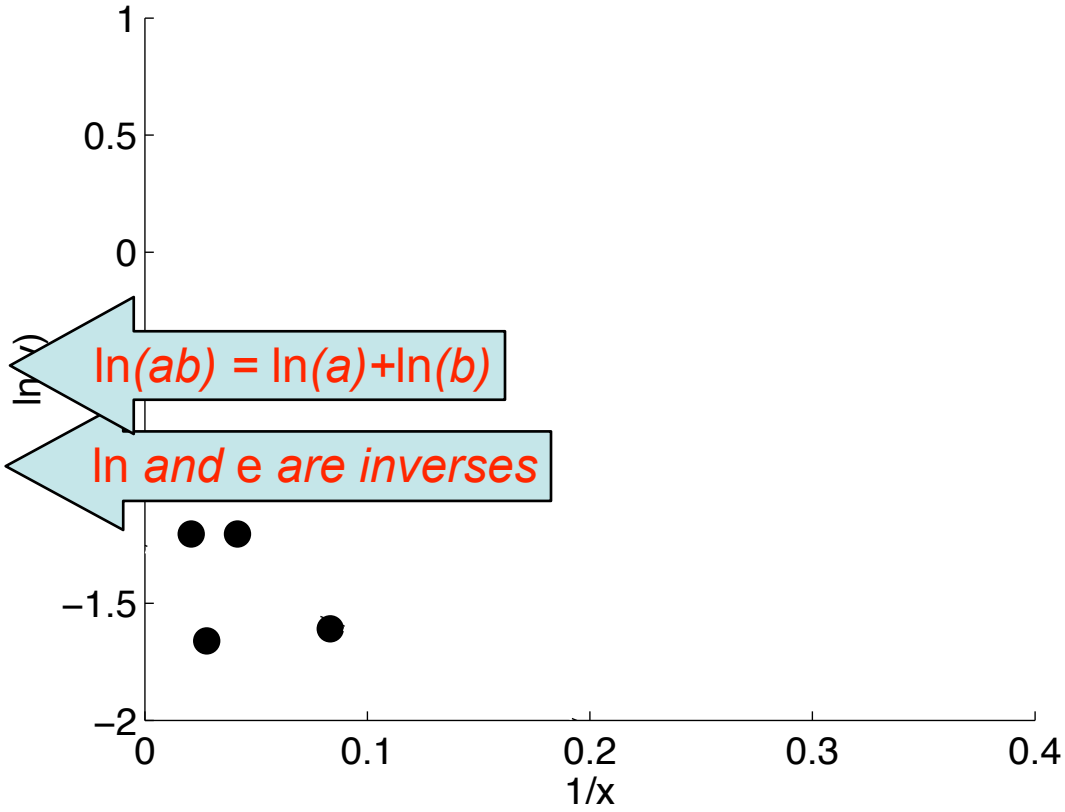
$$y = Ce^{1/x}$$

$$\ln y = \ln(Ce^{1/x})$$

$$\ln y = \ln e^{1/x} + \ln C$$

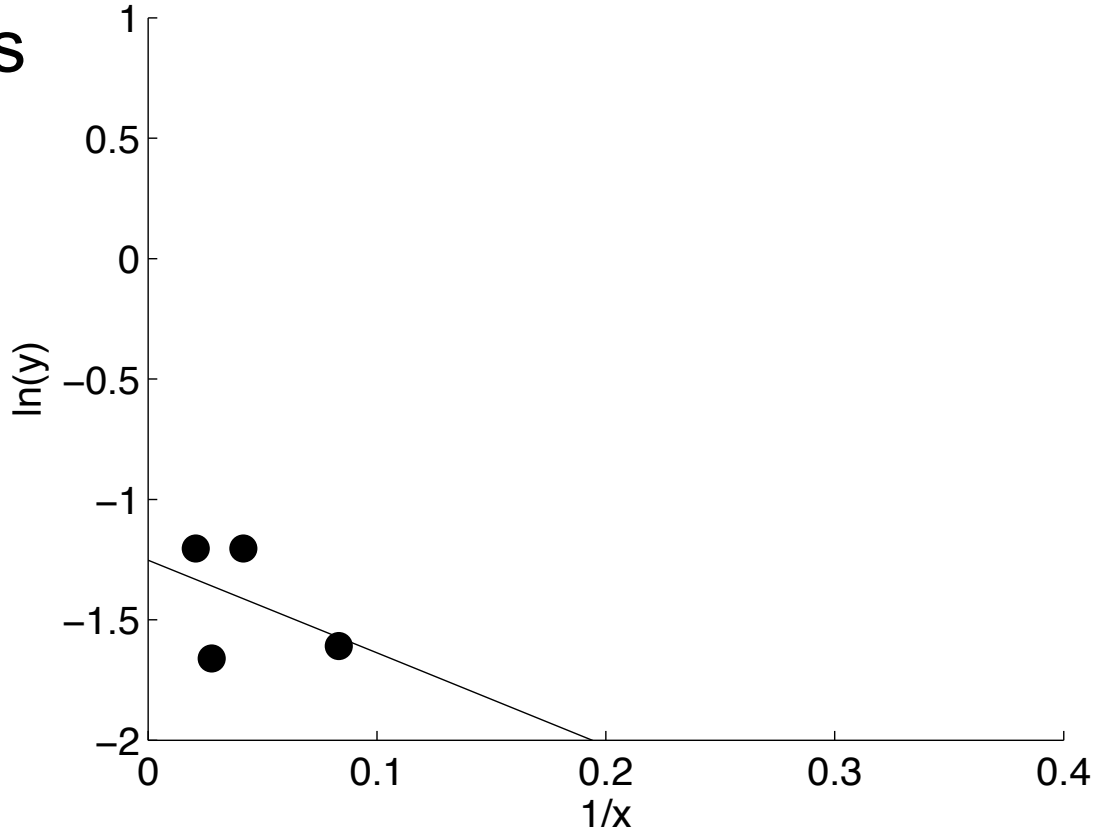
$$\ln y = 1/x + \ln C$$

- Plot ln *y* vs. 1/*x*
- See how distance is distorted?



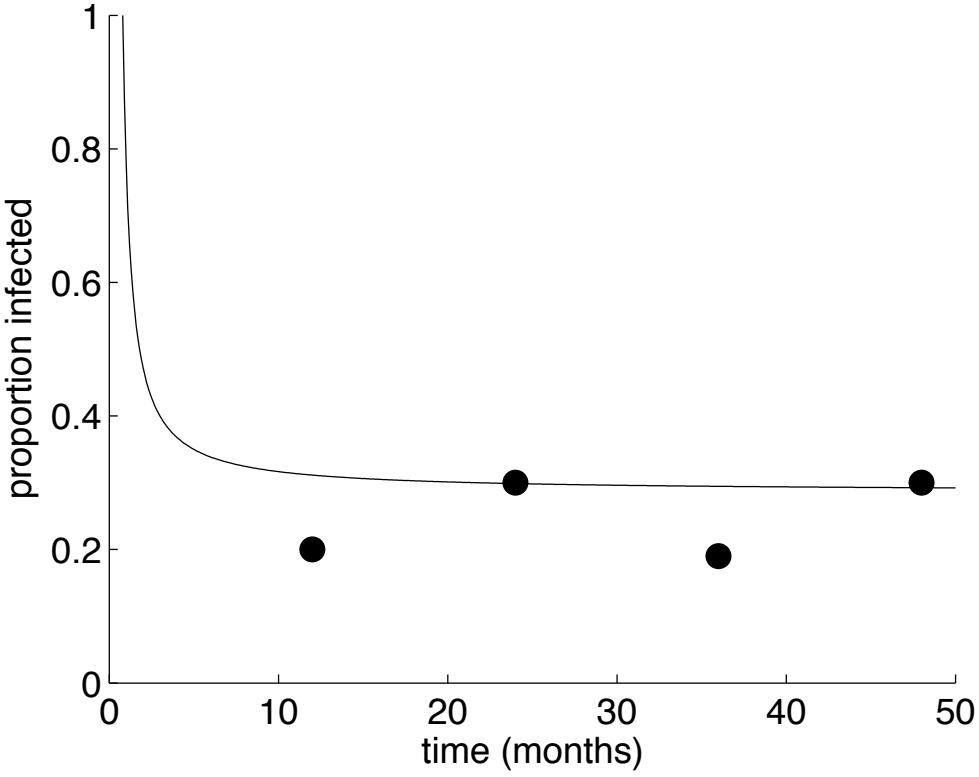
A line fits this quite well

- Fitting a line to the transformed data is a pretty good fit
- The deviations are small
- The best fit gives ln $C \approx$ -1.25.



A poor fit in the original data

- But in the original data, the fitted model $y = Ce^{1/x}$ fits the model poorly
- No seasonal peaks
- Heads up to ∞ near $x = 0$.



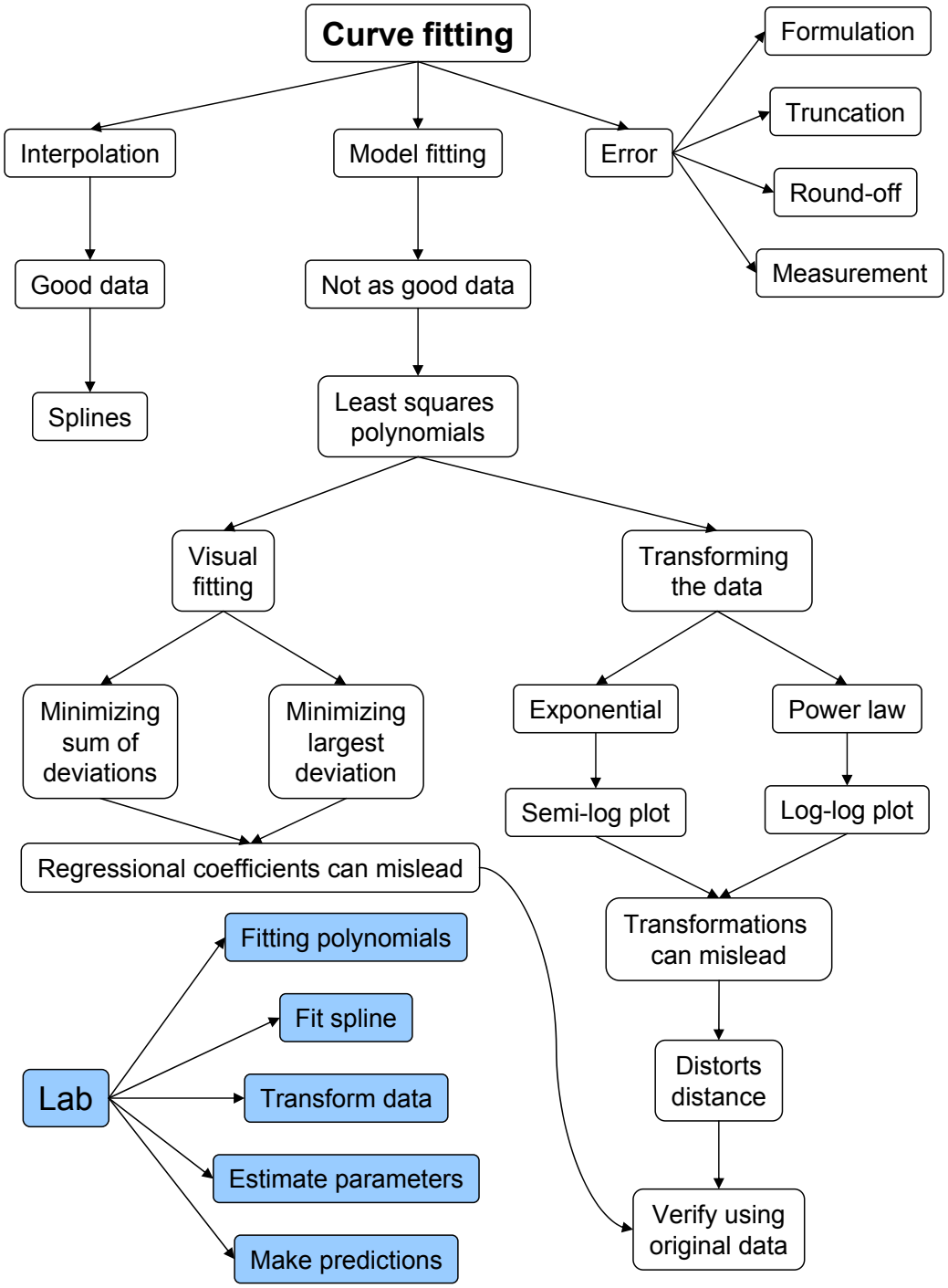
What's gone wrong?

- Answer: the data were never supposed to fit a model of the form $y = Ce^{1/x}$
- But we wouldn't know that from the transformation, which is a pretty good fit
- Many computer codes fit models by first making a transformation

Be careful: always verify your model using the original data.

Lab work

- In the lab we'll fit both polynomials and splines to data
- We'll also transform our data and use that to estimate parameters
- We'll also use our data and models to make predictions.


Curve fitting
Interpolation
Model fitting
Error
Formulation
Truncation
Round-off
Measurement
Good data
Not as good data
Splines
Least squares polynomials
Visual fitting
Transforming the data
Minimizing sum of deviations
Minimizing largest deviation
Exponential
Power law
Semi-log plot
Log-log plot
Regressional coefficients can mislead
Transformations can mislead
Distorts distance
Verify using original data
Lab
Fitting polynomials
Fit spline
Transform data
Estimate parameters
Make predictions