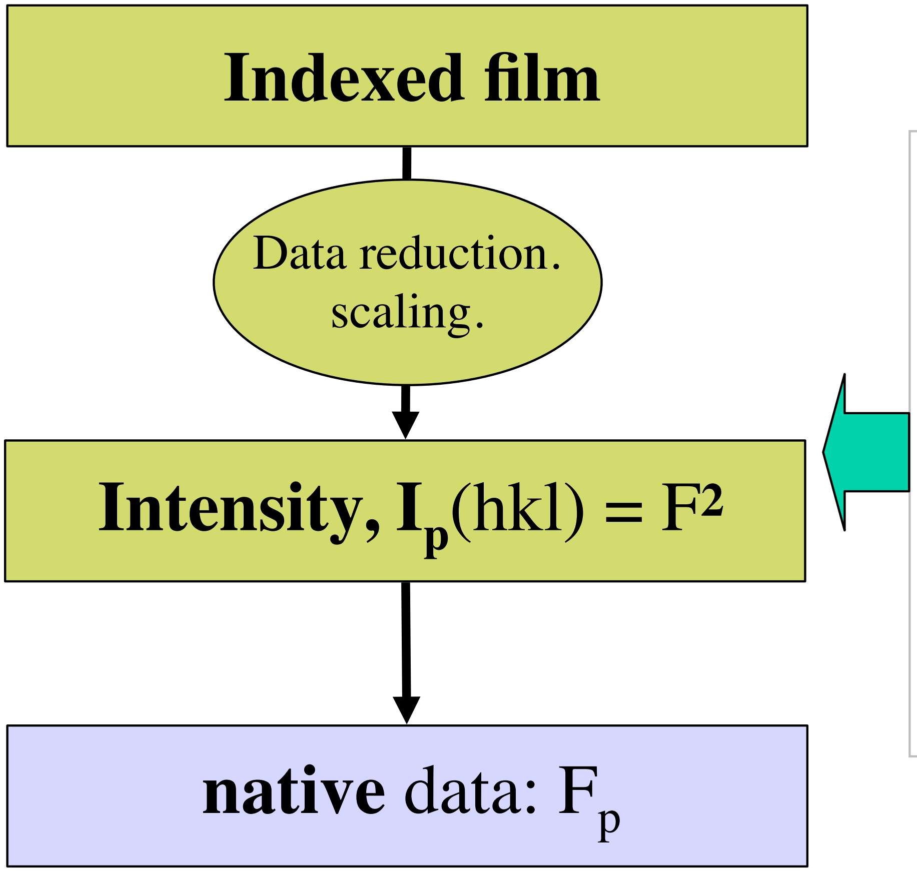
PSD '18 -- Lecture 11

Error sources in Crystallography

Summary flowcharts

From crystal to data



Indexed film

Data reduction. scaling.

Intensity, I_p(hkl) = F^2

native data: F_p

<u>I is always a relative term</u>

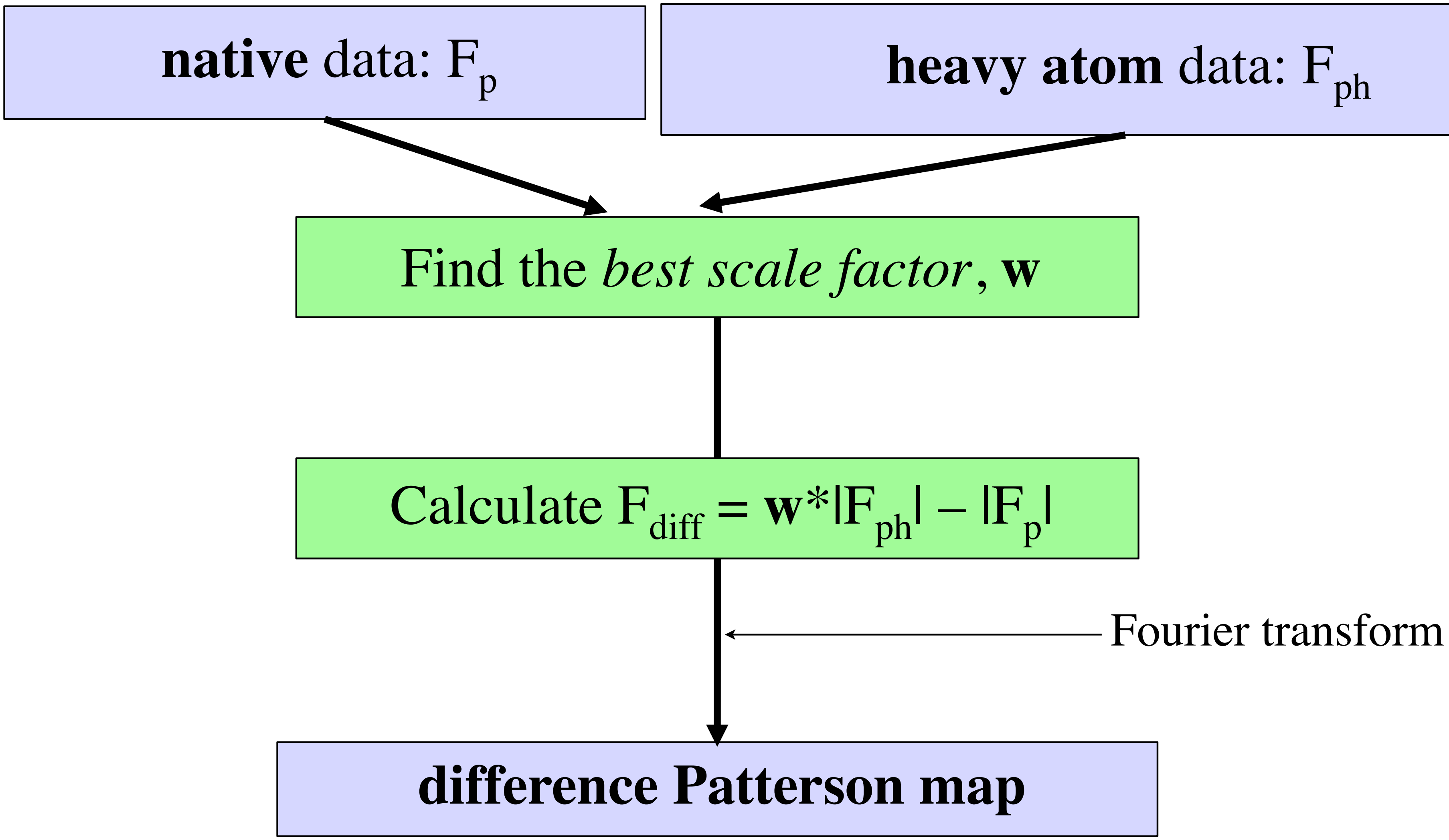
Bigger crystal, higher I

Better crystal, higher I

Longer exposure, higher I

More intense Xrays, higher I

From data to Patterson map



native data: F_p

heavy atom data: F_{ph}

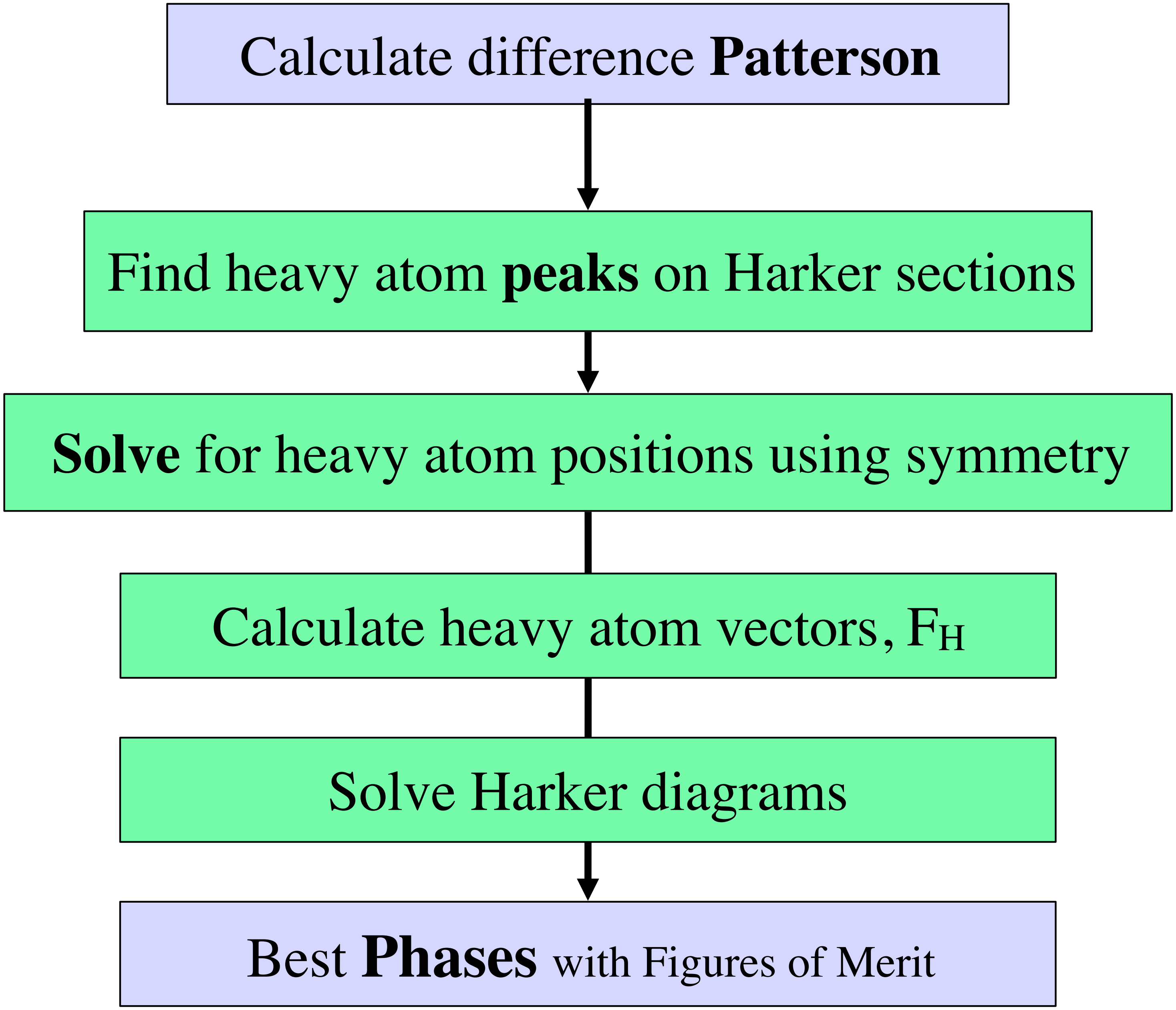
Find the *best scale factor*, **w**

Calculate $F_{diff} = \mathbf{w}^*|F_{ph}| - |F_p|$

Fourier transform

difference Patterson map

From Patterson map to phases



Calculate difference **Patterson**

Find heavy atom **peaks** on Harker sections

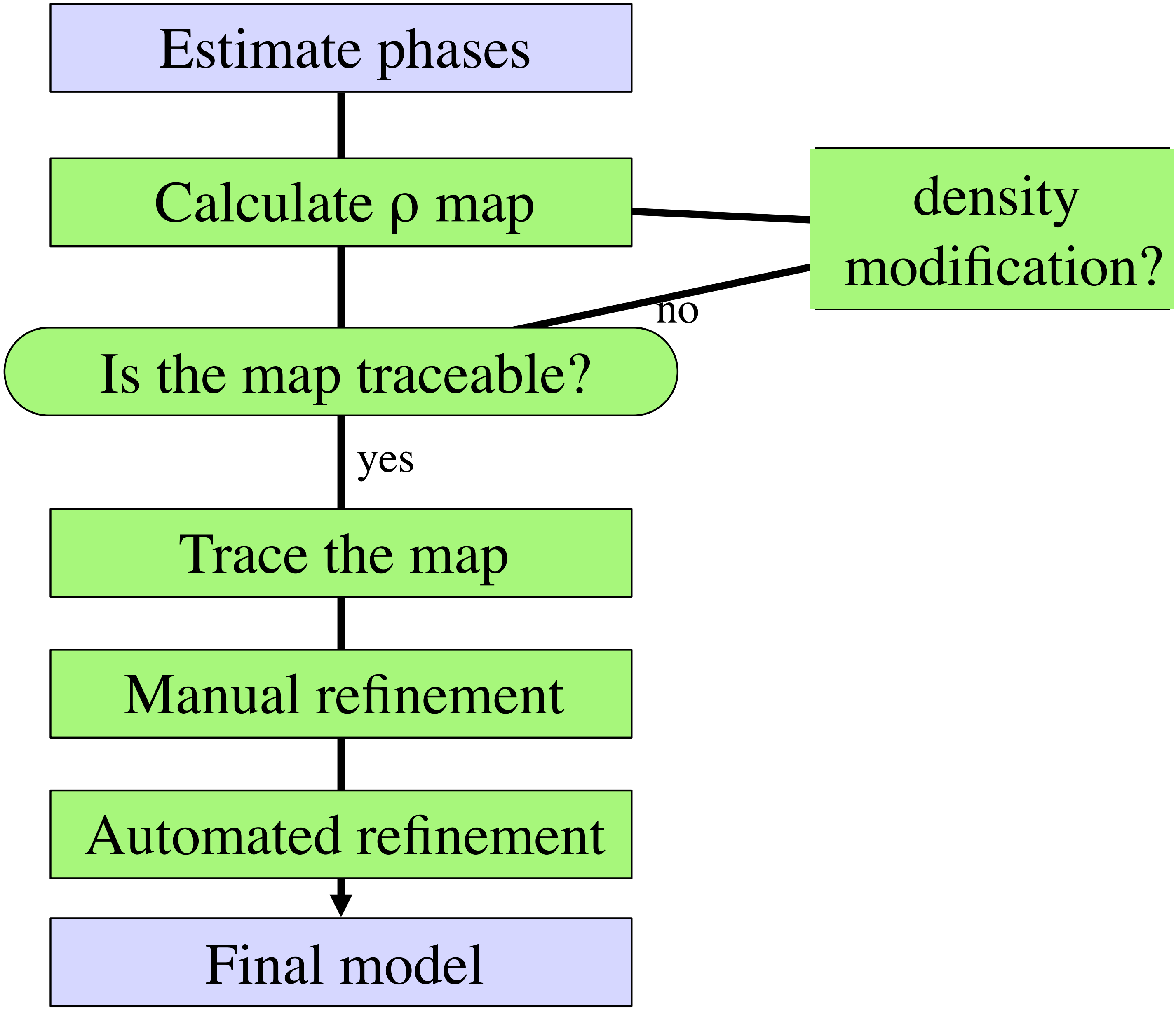
Solve for heavy atom positions using symmetry

Calculate heavy atom vectors, F_H

Solve Harker diagrams

Best **Phases** with Figures of Merit

From Phases to Model



Estimate phases

Calculate ρ map

density modification?

no

Is the map traceable?

yes

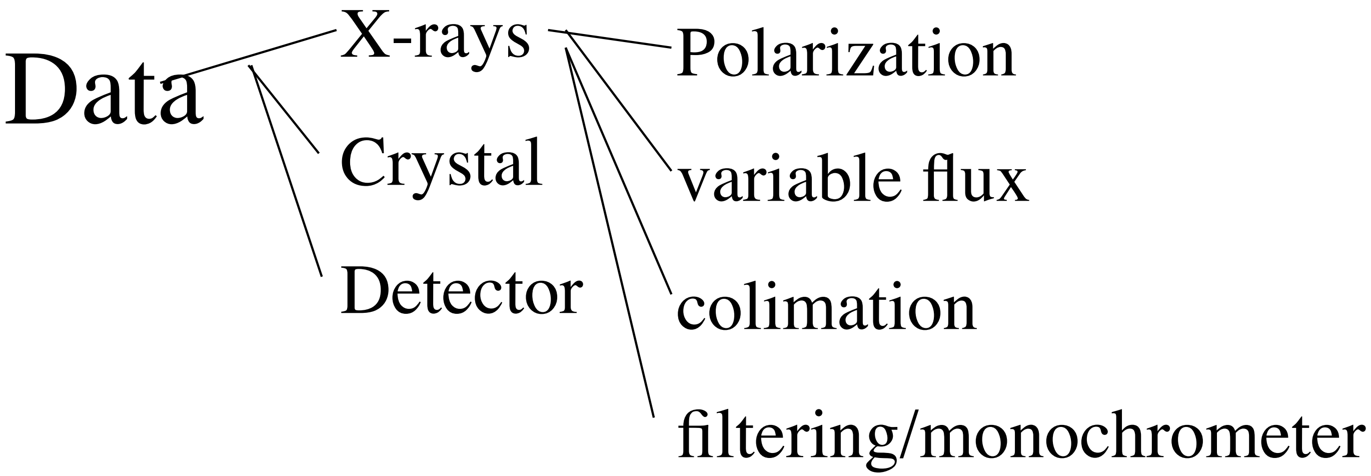
Trace the map

Manual refinement

Automated refinement

Final model

Sources of error in crystal structures



Data

X-rays

Crystal

Detector

Polarization

variable flux

colimation

filtering/monochrometer

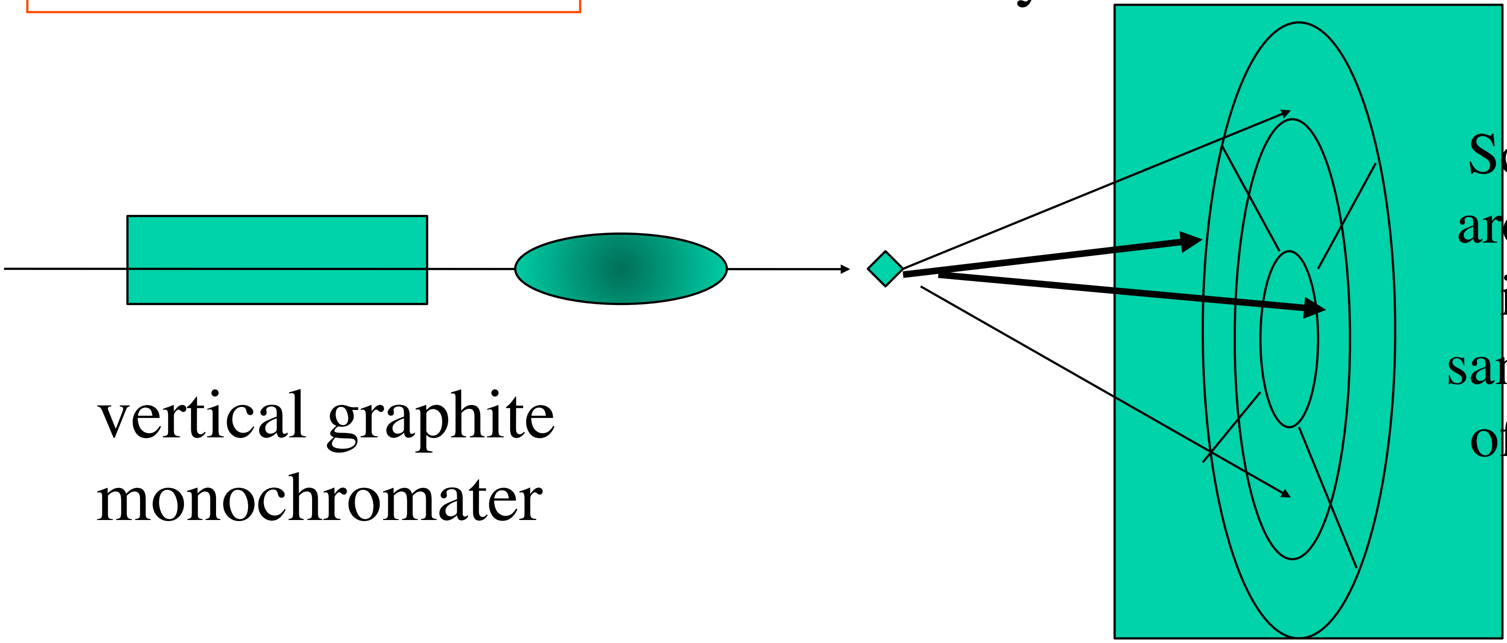
Model

Experimental sources of error

Polarization

weaker scatter
vertically

Solution:
zonal scaling.



vertical graphite
monochromater

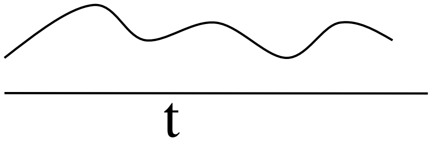
Scale factors
are calculated
in evenly-
sampled zones
of reciprocal
space.

horizontally polarized X-rays

Experimental sources of error

variable flux

A problem for *synchrotron X-rays*. Solution: Use an external flux meter + scaling.



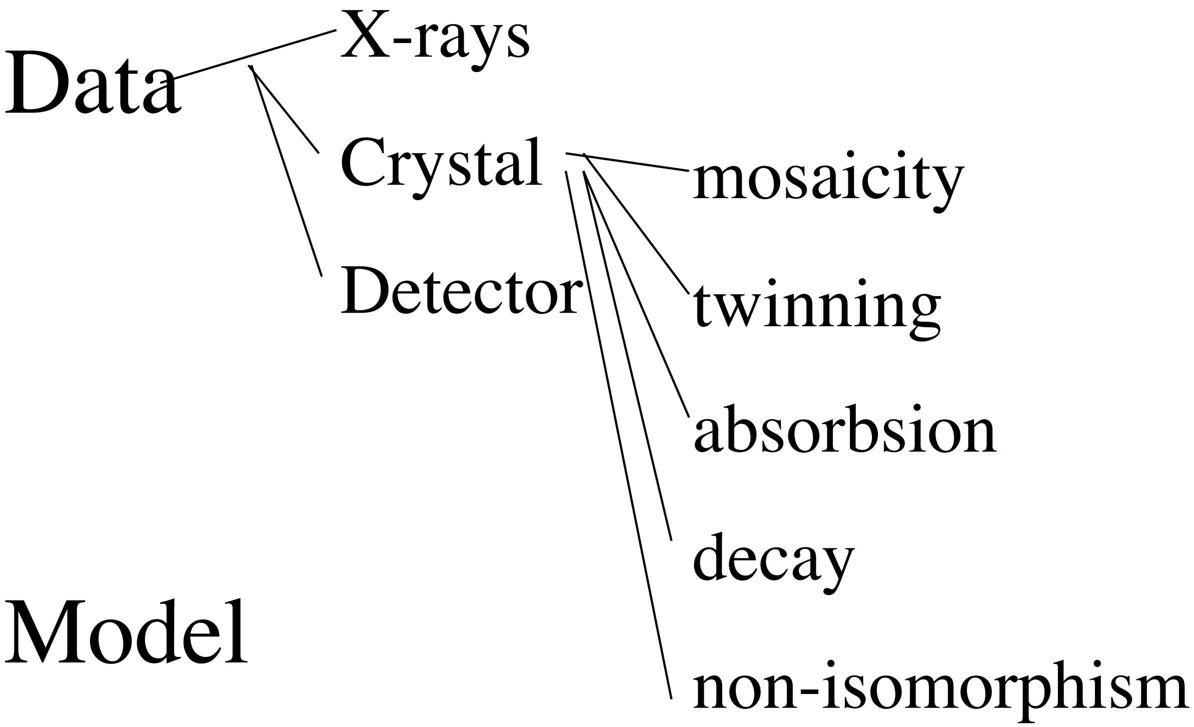
colimation

Wide beam means high background, large spots, spot overlap. Narrow beam means longer exposures, uneven exposure of crystal

variable wavelength

Spots may be radially smeared. Solution: Use *monochromator*.

Sources of error in crystal structures



Data

X-rays

Crystal

Detector

mosaicity

twinning

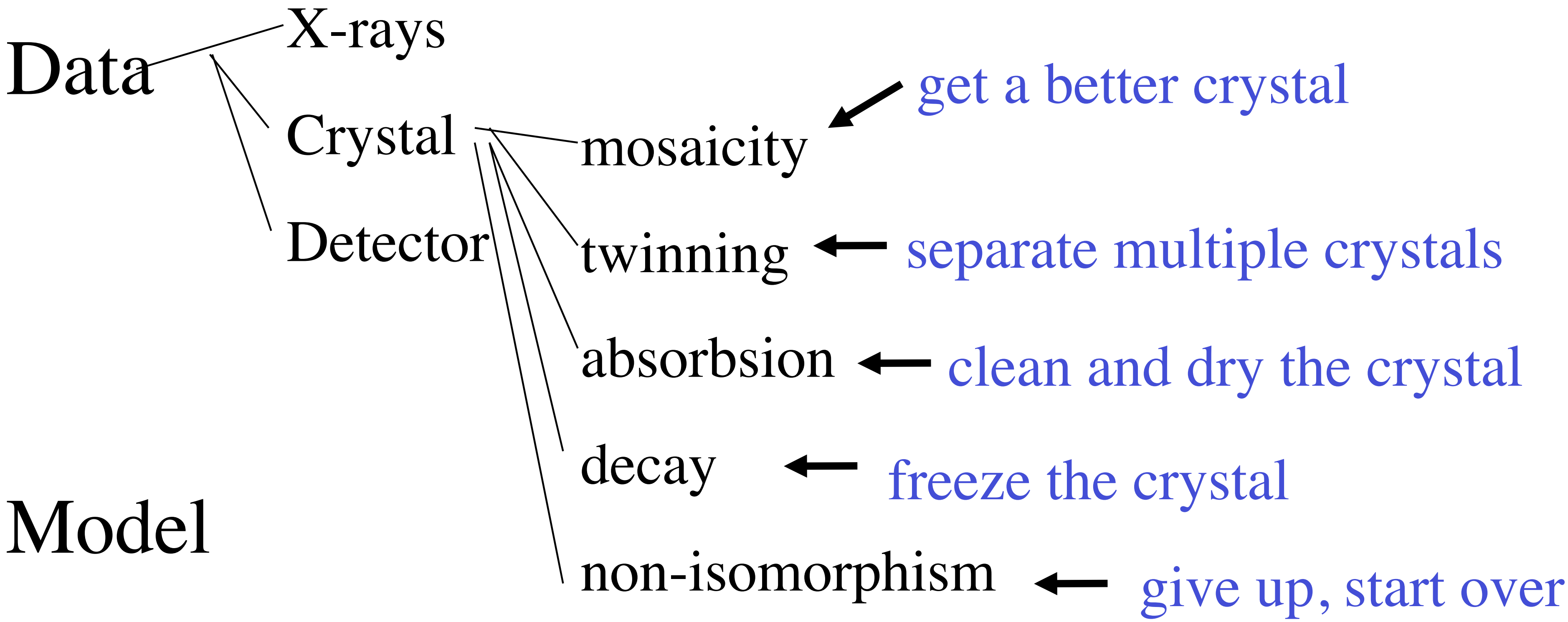
absorbsion

decay

non-isomorphism

Model

Sources of error in crystal structures



Data

X-rays

Crystal

Detector

mosaicity ← get a better crystal

twinning ← separate multiple crystals

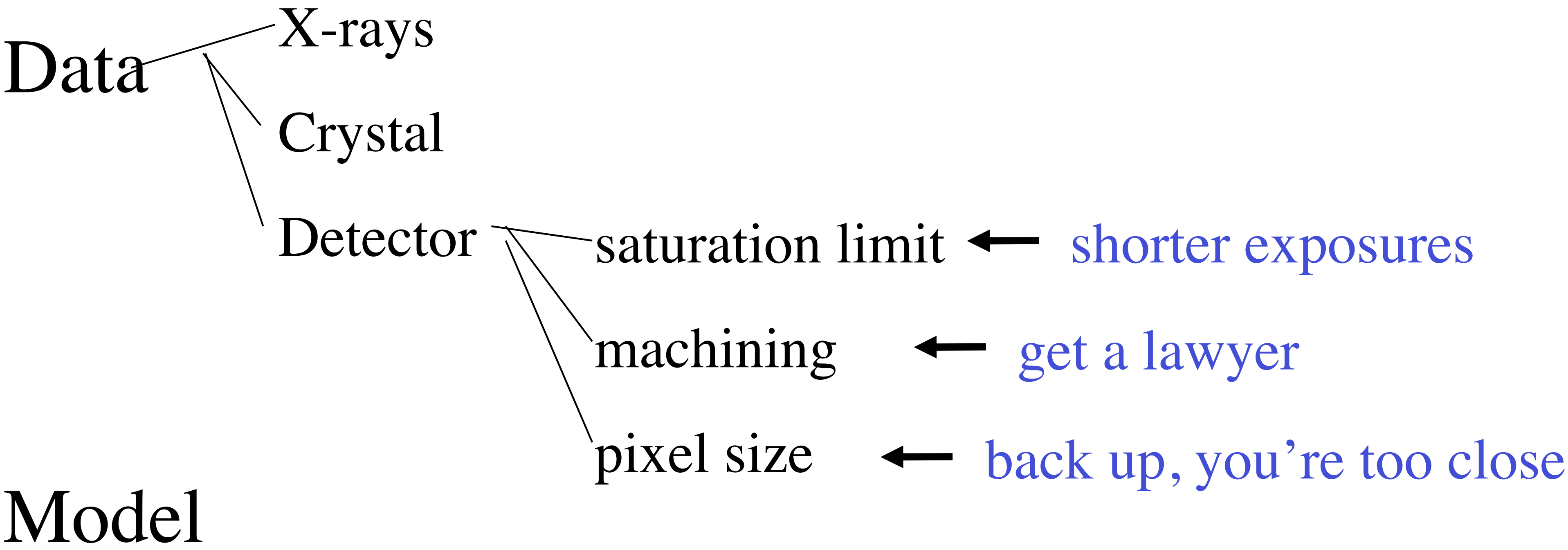
absorbsion ← clean and dry the crystal

decay ← freeze the crystal

non-isomorphism ← give up, start over

Model

Sources of error in crystal structures



Data

X-rays

Crystal

Detector

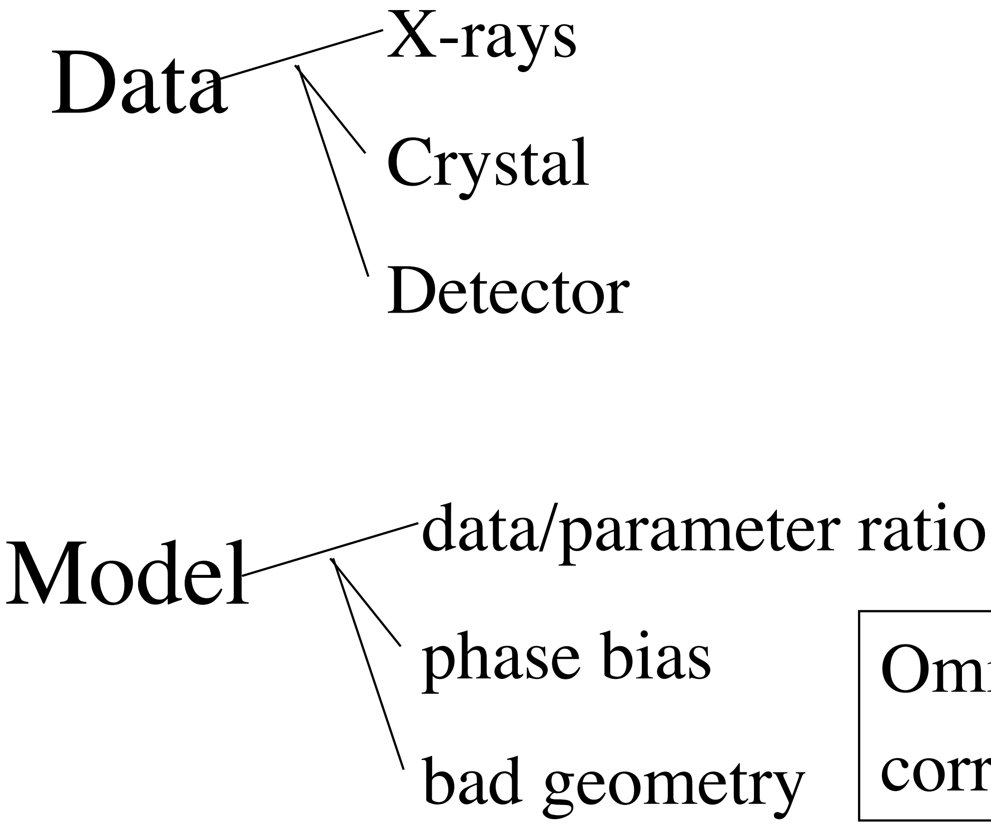
saturation limit ← shorter exposures

machining ← get a lawyer

pixel size ← back up, you’re too close

Model

Computational Sources of error



Data

X-rays

Crystal

Detector

Model

data/parameter ratio

phase bias

bad geometry

Luzatti or Σ_A plot will estimate errors. Real-space R will locate errors.

Omit maps, $2F_o$-F_c maps correct for phase bias.

PROCHECK, Molprobity find bad contacts, bad rotamers, etc.

avoiding phase bias errors…

Cross-validation: The free R-factor

The R-factor measures the *residual difference between observed and calculated amplitudes*.

Free R is summed on a "test set". **Test set data was not used for refinement.**

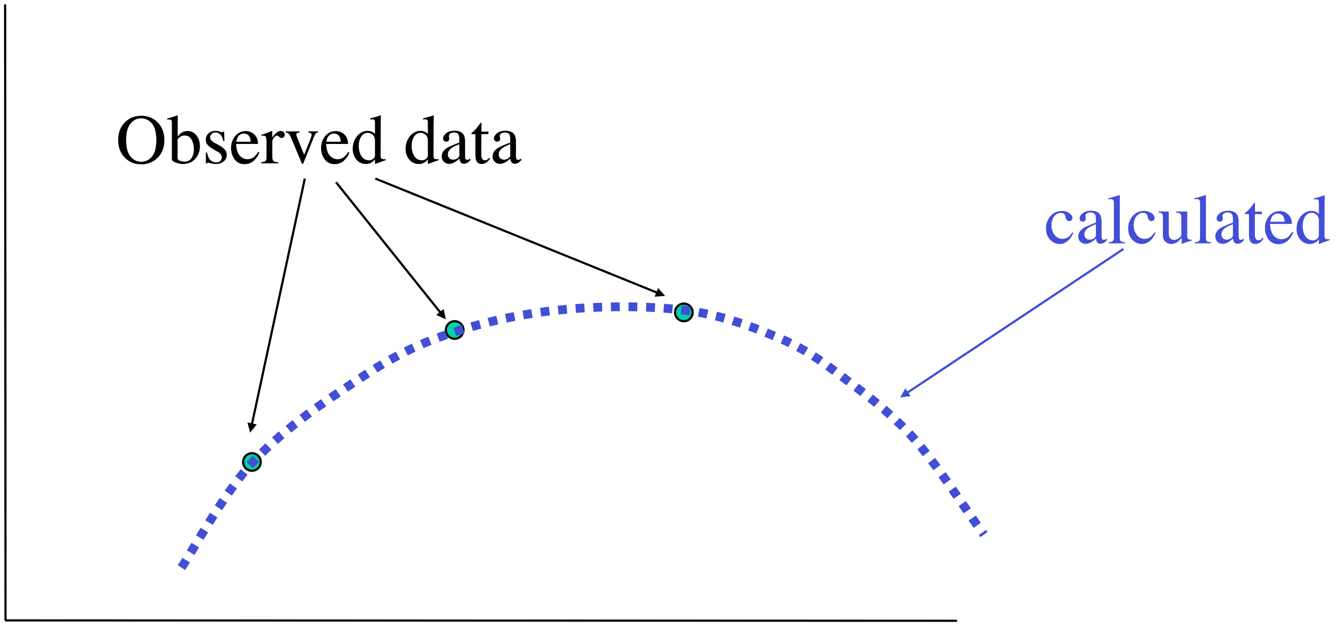
Free R ask: *"How well does your model predict the data it hasn't been fit to?"*

$$R_{free} = \frac{\sum_{h \in T} \left\| F_{obs}(h) \right| - k \left| F_{calc}(h) \right\|}{\sum_{h \in T} \left| F_{obs}(h) \right|}$$

Note: T = independent test set of *F's*.

What is over-fitting?

If you have three points, you can fit them to a quadratic equation (3 parameters) with *zero residual*, but is it right?

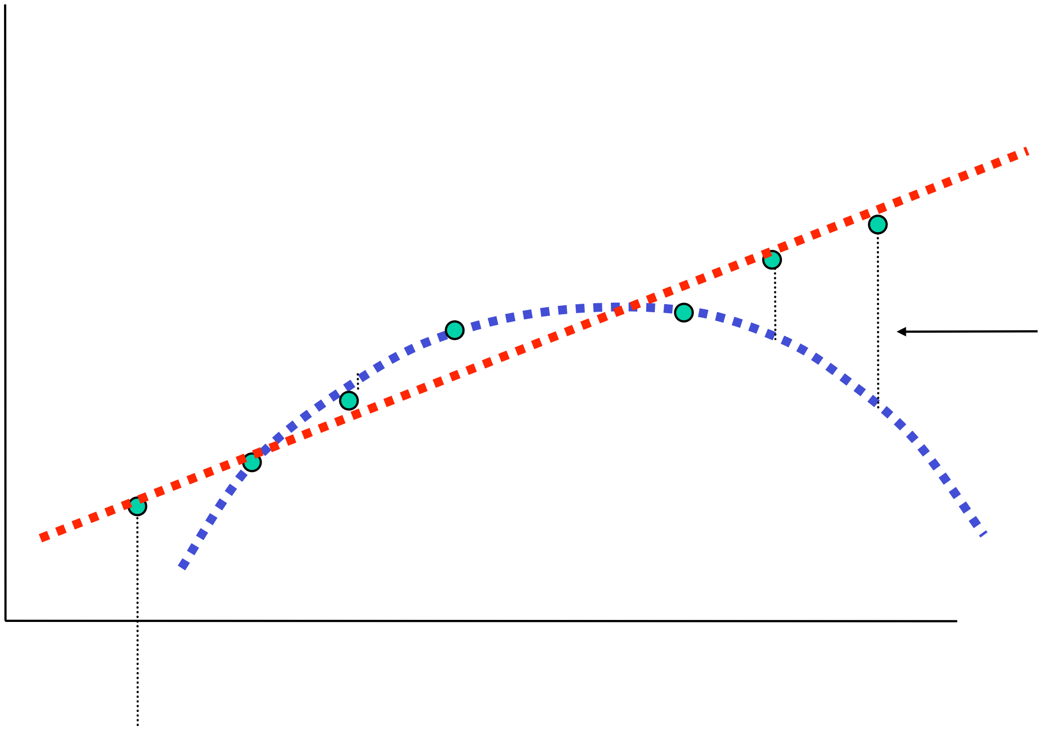


R-factor = 0.000!!

Fitting unseen data, as a test

Fit is correct if *additional data*, not used in fitting the curve, fall <u>on the curve</u>.

Low residual in the “test set” *validates* the fit.



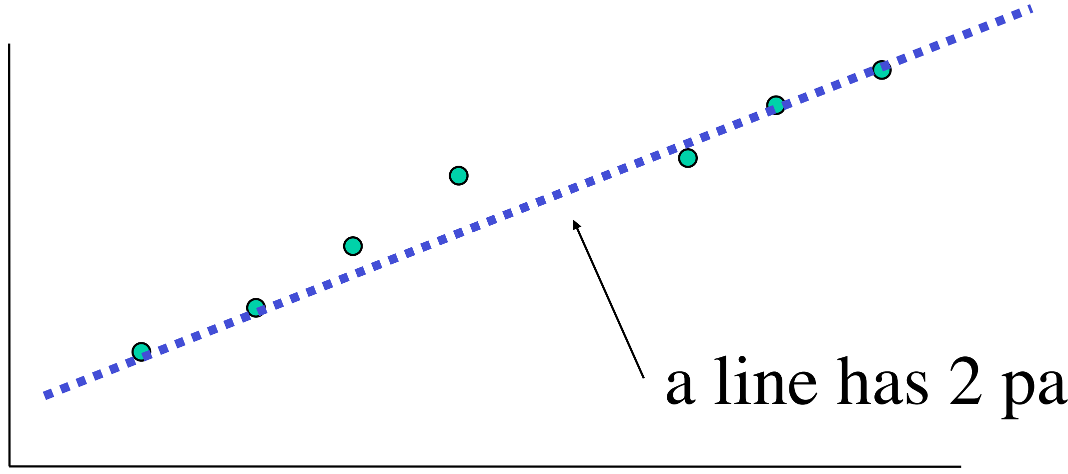
residual≠0

cross-validation

Means: measuring the residual on data (a "test set") that were not used to refine (or fit) the model.

The residual on test data is likely to be small if

$$\frac{\textit{data}}{\textit{parameters}}$$ is large.



a line has 2 parameters

Parameters versus Data

Example :

Papain crystal structure has 25,000 reflections.

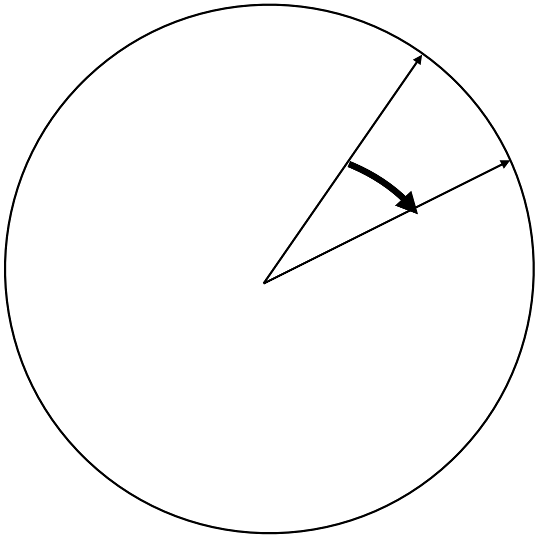
Papain has 2000 non-H atoms

times 4 parameters each (x, y, z, B)

equals 8000 parameters

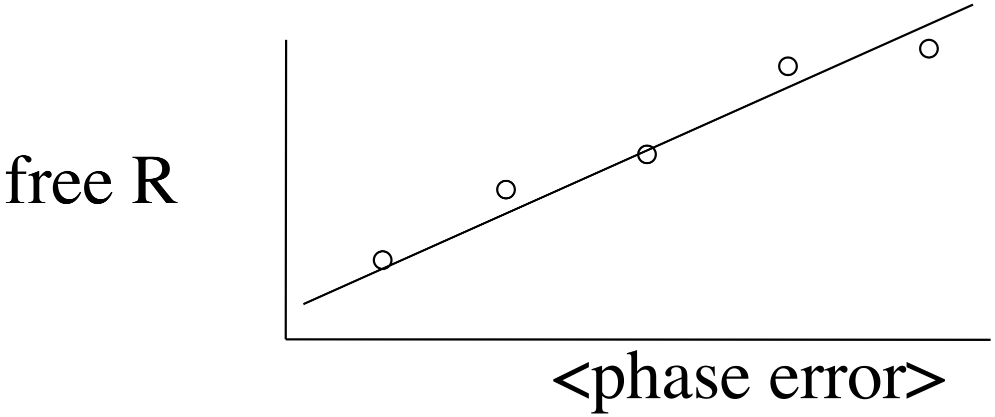
data/parameters = 25,000/8000 ≈ 3 <-- *this is too small!*

Phase error



Every reflection has a phase error, which is the difference of the calculated phase from the true phase (unknown).

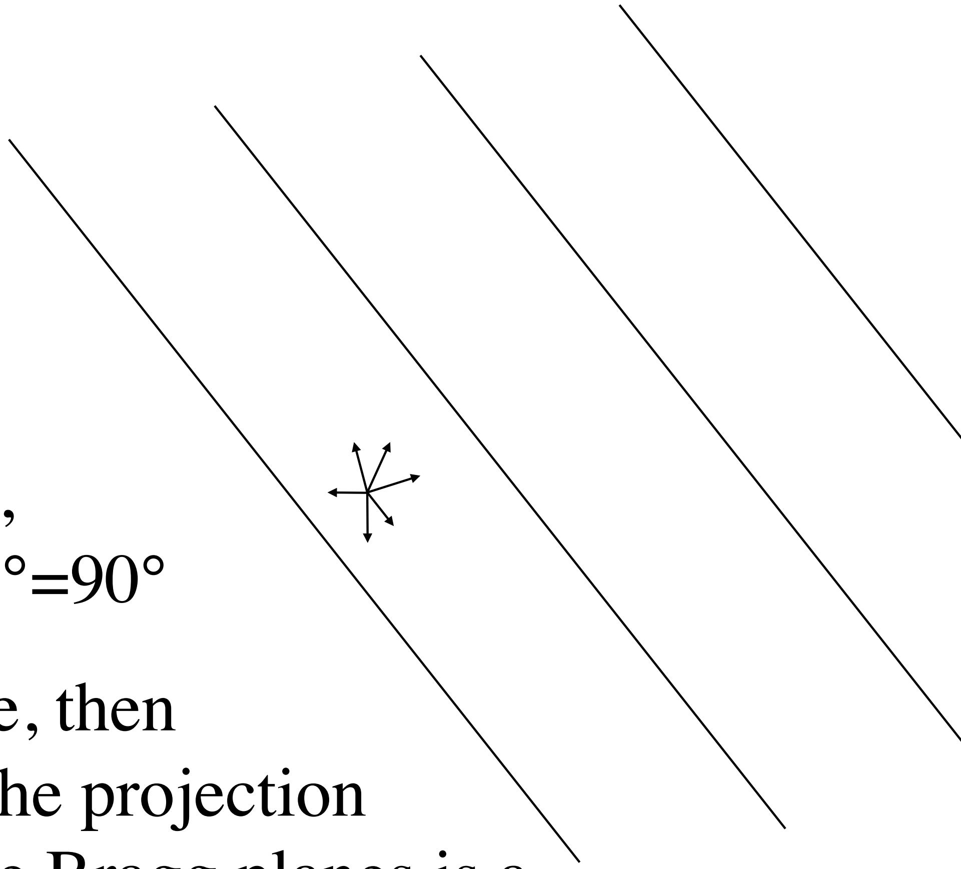
Free R-factor *correlates* with phase error in molecular replacement studiess



Thought experiment

What is the phase error for 4Å resolution reflections if the average coordinate error is 1Å?

Coordinate error causes phase error



If the error in atomic position is 1Å,
and the Bragg plane separation is 4Å,
then the error in phase is ≤ (1/4)*360°=90°

If the error is a Gaussian in real space, then the phase error is also a Gaussian. (The projection of a 3D Gaussian on the normal to the Bragg planes is a 1D Gaussian)

Exercise 8 -- Reading a crystallography paper

Due mon Nov 26. Send email, or write on paper.

Download the PDF linked to "Ex 8 paper"
Read the section labeled "Structure Determination"
Explain: " All non-hydrogen atoms were refined anisotropically."
Read Table 1.
Answer: (1) What is the meaning of the second resolution range (in parentheses)?
(2) Why are there values in parentheses for the other items, such a R_{sym}?
(3) What is Wilson B?