Modern Analysis Techniques for Large Data Sets

Miguel F. Morales
Bryna Hazelton

Graduate elective

- You get out what you put in
  - Homework (first ~7 weeks; 40%; 100%; + options)
  - Final project & presentation based on your research
  - No exams
- Fill in quiz describing what you'd like to get out of the course

Key topics

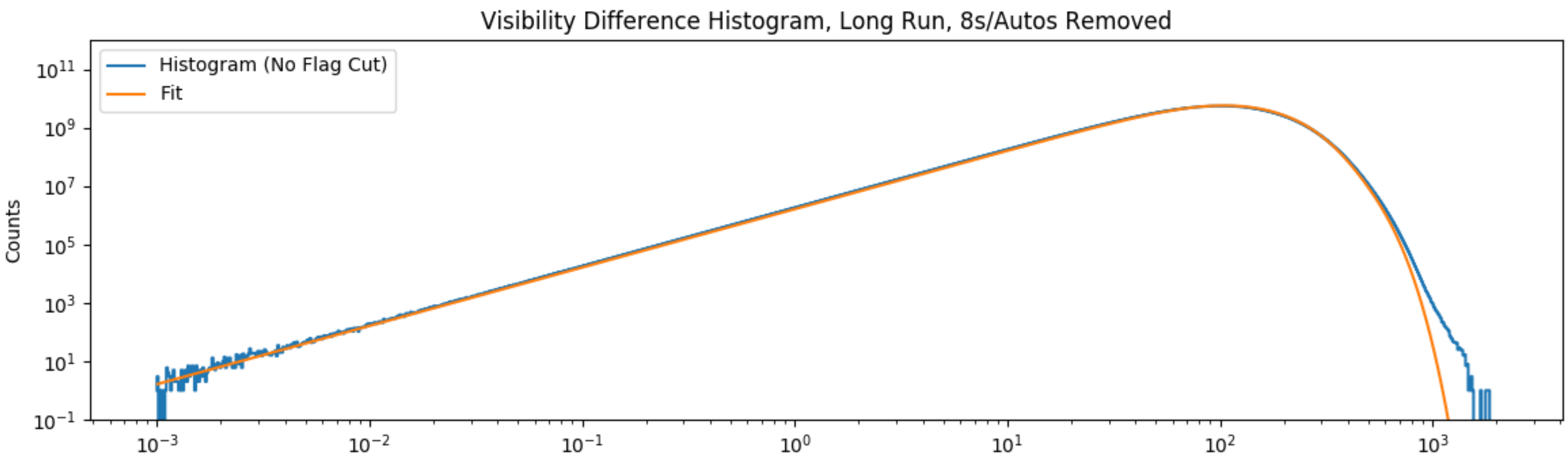
- Practical statistics
  - Turning research questions into statistical math
  - Systematics, non-Gaussianity, covariance, avoiding common mistakes
- Data visualization
  - High quality visualizations (data density, perception, accessibility)
  - Types of visualizations (research, monitoring, presentation)
- Collaborative data analysis
  - Using GitHub to your advantage (branching, provenance)
  - Collaborative development
  - Peer reviewed code
- Advanced practices
  - Statistically rigorous visualization
  - Jackknife tests, testing thickets, testing below the noise
  - Question driven data rampages

Course structure

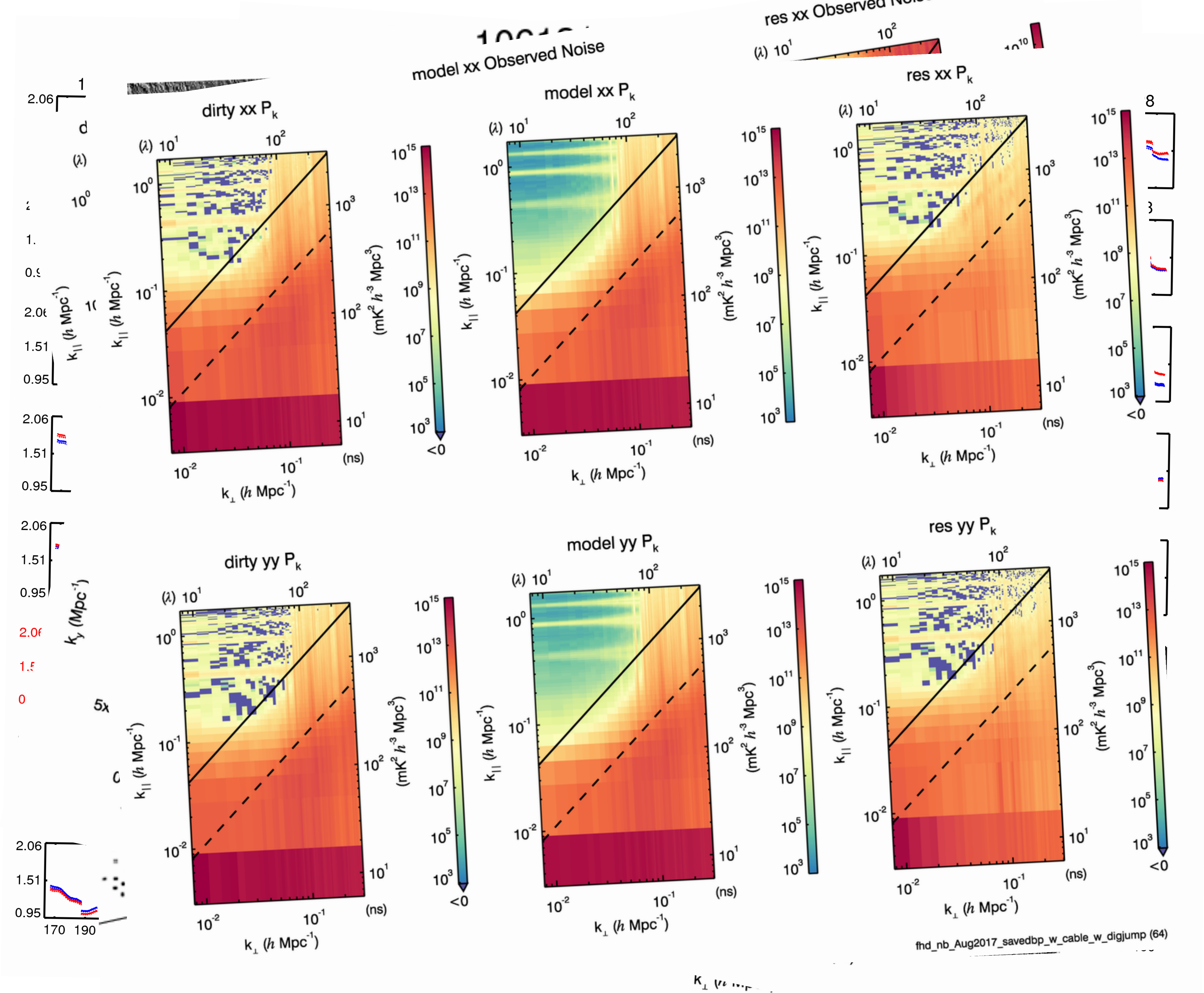
- Building blocks (stats; visualization; git & collaborative code)

- Putting it all together

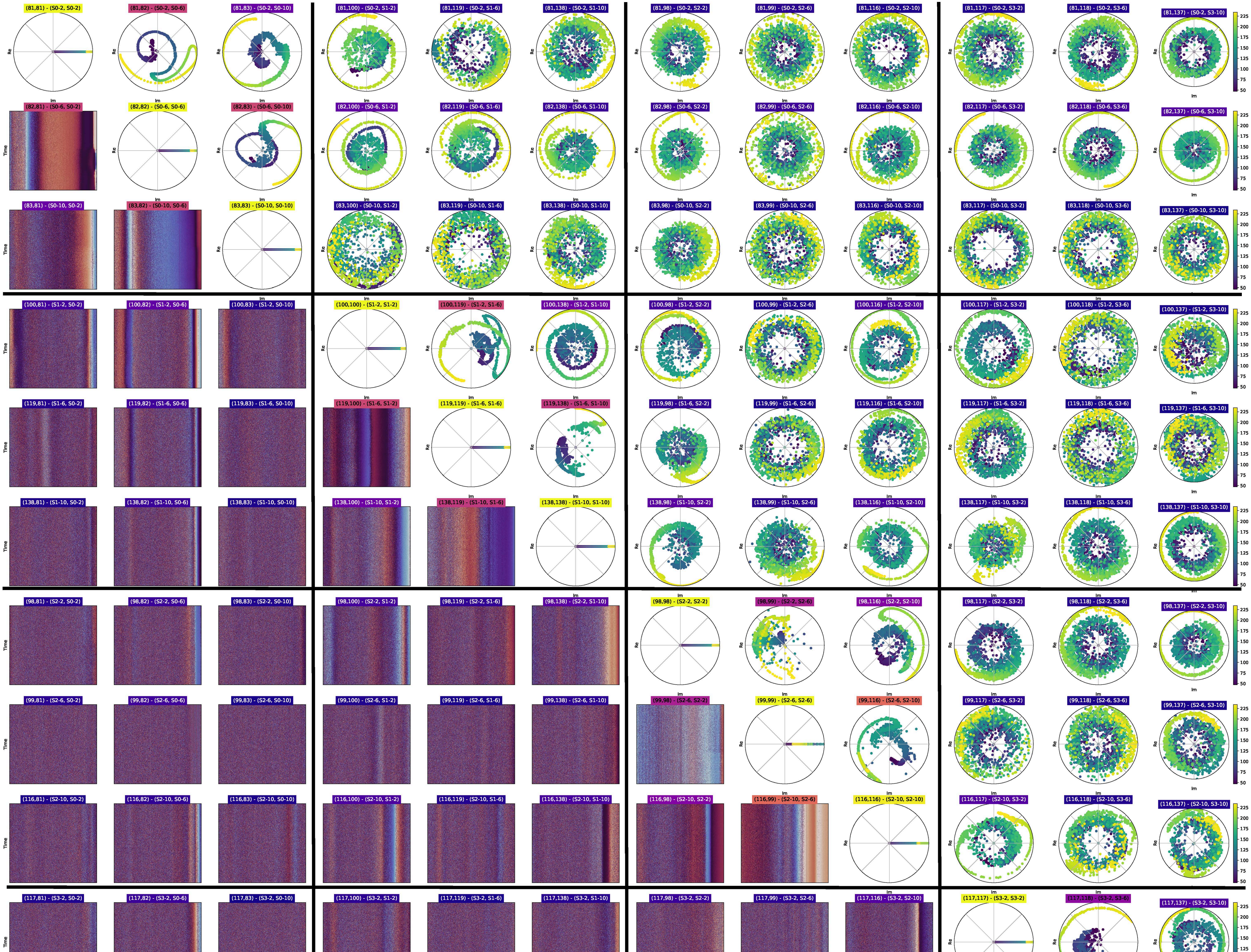
Statistics



100 billion data points



model xx Observed Noise
res xx Observed Noise
dirty xx P$_k$
model xx P$_k$
res xx P$_k$
dirty yy P$_k$
model yy P$_k$
res yy P$_k$
$k_{||}$ (h Mpc^{-1})
$k_\perp$ (h Mpc^{-1})
(mK^2 h^{-3} Mpc^3)
(λ)
(ns)
<0
k_y (Mpc^{-1})
fhd_nb_Aug2017_savedbp_w_cable_w_digjump (64)



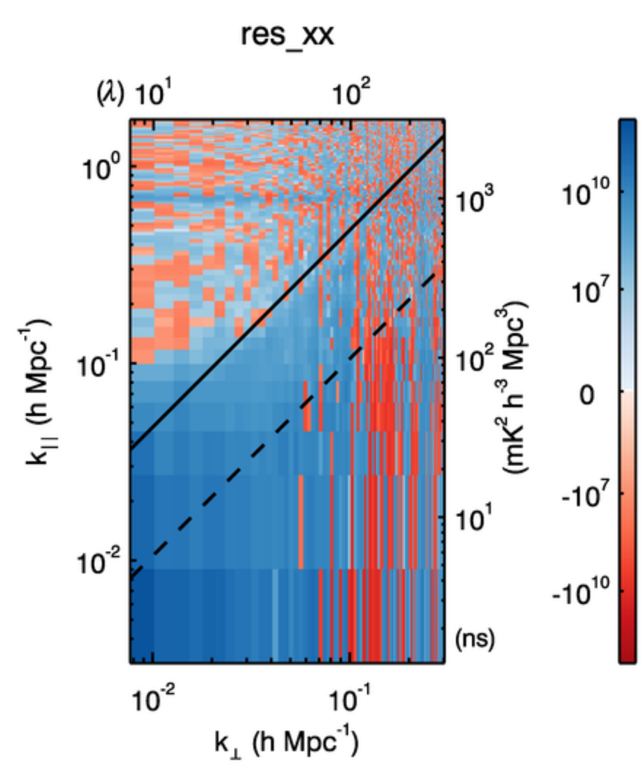
(81,81) - (S0-2, S0-2)
(81,82) - (S0-2, S0-6)
(81,83) - (S0-2, S0-10)
(81,100) - (S0-2, S1-2)
(81,119) - (S0-2, S1-6)
(81,138) - (S0-2, S1-10)
(81,98) - (S0-2, S2-2)
(81,99) - (S0-2, S2-6)
(81,116) - (S0-2, S2-10)
(81,117) - (S0-2, S3-2)
(81,118) - (S0-2, S3-6)
(81,137) - (S0-2, S3-10)
(82,81) - (S0-6, S0-2)
(82,82) - (S0-6, S0-6)
(82,83) - (S0-6, S0-10)
(82,100) - (S0-6, S1-2)
(82,119) - (S0-6, S1-6)
(82,138) - (S0-6, S1-10)
(82,98) - (S0-6, S2-2)
(82,99) - (S0-6, S2-6)
(82,116) - (S0-6, S2-10)
(82,117) - (S0-6, S3-2)
(82,118) - (S0-6, S3-6)
(82,137) - (S0-6, S3-10)
(83,81) - (S0-10, S0-2)
(83,82) - (S0-10, S0-6)
(83,83) - (S0-10, S0-10)
(83,100) - (S0-10, S1-2)
(83,119) - (S0-10, S1-6)
(83,138) - (S0-10, S1-10)
(83,98) - (S0-10, S2-2)
(83,99) - (S0-10, S2-6)
(83,116) - (S0-10, S2-10)
(83,117) - (S0-10, S3-2)
(83,118) - (S0-10, S3-6)
(83,137) - (S0-10, S3-10)
(100,81) - (S1-2, S0-2)
(100,82) - (S1-2, S0-6)
(100,83) - (S1-2, S0-10)
(100,100) - (S1-2, S1-2)
(100,119) - (S1-2, S1-6)
(100,138) - (S1-2, S1-10)
(100,98) - (S1-2, S2-2)
(100,99) - (S1-2, S2-6)
(100,116) - (S1-2, S2-10)
(100,117) - (S1-2, S3-2)
(100,118) - (S1-2, S3-6)
(100,137) - (S1-2, S3-10)
(119,81) - (S1-6, S0-2)
(119,82) - (S1-6, S0-6)
(119,83) - (S1-6, S0-10)
(119,100) - (S1-6, S1-2)
(119,119) - (S1-6, S1-6)
(119,138) - (S1-6, S1-10)
(119,98) - (S1-6, S2-2)
(119,99) - (S1-6, S2-6)
(119,116) - (S1-6, S2-10)
(119,117) - (S1-6, S3-2)
(119,118) - (S1-6, S3-6)
(119,137) - (S1-6, S3-10)
(138,81) - (S1-10, S0-2)
(138,82) - (S1-10, S0-6)
(138,83) - (S1-10, S0-10)
(138,100) - (S1-10, S1-2)
(138,119) - (S1-10, S1-6)
(138,138) - (S1-10, S1-10)
(138,98) - (S1-10, S2-2)
(138,99) - (S1-10, S2-6)
(138,116) - (S1-10, S2-10)
(138,117) - (S1-10, S3-2)
(138,118) - (S1-10, S3-6)
(138,137) - (S1-10, S3-10)
(98,81) - (S2-2, S0-2)
(98,82) - (S2-2, S0-6)
(98,83) - (S2-2, S0-10)
(98,100) - (S2-2, S1-2)
(98,119) - (S2-2, S1-6)
(98,138) - (S2-2, S1-10)
(98,98) - (S2-2, S2-2)
(98,99) - (S2-2, S2-6)
(98,116) - (S2-2, S2-10)
(98,117) - (S2-2, S3-2)
(98,118) - (S2-2, S3-6)
(98,137) - (S2-2, S3-10)
(99,81) - (S2-6, S0-2)
(99,82) - (S2-6, S0-6)
(99,83) - (S2-6, S0-10)
(99,100) - (S2-6, S1-2)
(99,119) - (S2-6, S1-6)
(99,138) - (S2-6, S1-10)
(99,98) - (S2-6, S2-2)
(99,99) - (S2-6, S2-6)
(99,116) - (S2-6, S2-10)
(99,117) - (S2-6, S3-2)
(99,118) - (S2-6, S3-6)
(99,137) - (S2-6, S3-10)
(116,81) - (S2-10, S0-2)
(116,82) - (S2-10, S0-6)
(116,83) - (S2-10, S0-10)
(116,100) - (S2-10, S1-2)
(116,119) - (S2-10, S1-6)
(116,138) - (S2-10, S1-10)
(116,98) - (S2-10, S2-2)
(116,99) - (S2-10, S2-6)
(116,116) - (S2-10, S2-10)
(116,117) - (S2-10, S3-2)
(116,118) - (S2-10, S3-6)
(116,137) - (S2-10, S3-10)
(117,81) - (S3-2, S0-2)
(117,82) - (S3-2, S0-6)
(117,83) - (S3-2, S0-10)
(117,100) - (S3-2, S1-2)
(117,119) - (S3-2, S1-6)
(117,138) - (S3-2, S1-10)
(117,98) - (S3-2, S2-2)
(117,99) - (S3-2, S2-6)
(117,116) - (S3-2, S2-10)
(117,117) - (S3-2, S3-2)
(117,118) - (S3-2, S3-6)
(117,137) - (S3-2, S3-10)
Re
Im
Time
225
200
175
150
125
100
75
50

git & analysis traceability

Data unit tests

2 fhd_core/fhd_struct_init_antenna.pro View

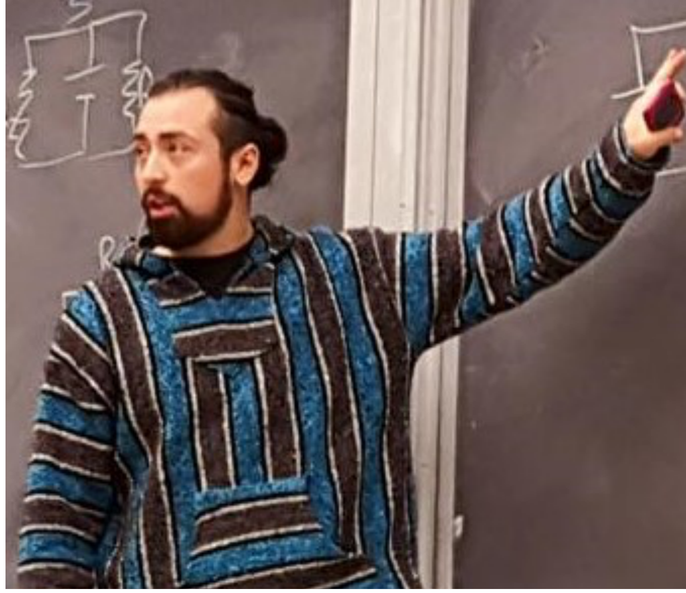
```
@@ -86,7 +86,7 @@ dec_use=dec_arr[valid_i]

 ;NOTE: Eq2Hor REQUIRES Jdate to have the same number of elements as RA and Dec for precession!!
 ;;NOTE: The NEW Eq2Hor REQUIRES Jdate to be a scalar! They created a new bug when they fixed the old one
-Eq2Hor,ra_use,dec_use,Jdate,alt_arr1,az_arr1,lat=obs.lat,lon=obs.lon,alt=obs.alt,precess=1
+Eq2Hor,ra_use,dec_use,Jdate,alt_arr1,az_arr1,lat=obs.lat,lon=obs.lon,alt=obs.alt,precess=1,/nutate
 za_arr=fltarr(psf_image_dim,psf_image_dim)+90. & za_arr[valid_i]=90.-alt_arr1
 az_arr=fltarr(psf_image_dim,psf_image_dim) & az_arr[valid_i]=az_arr1

```



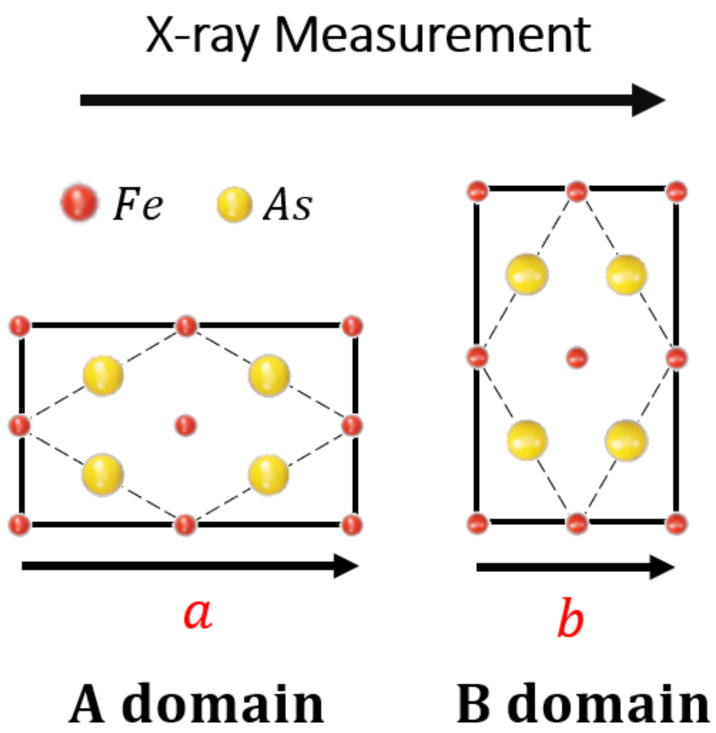

NERD

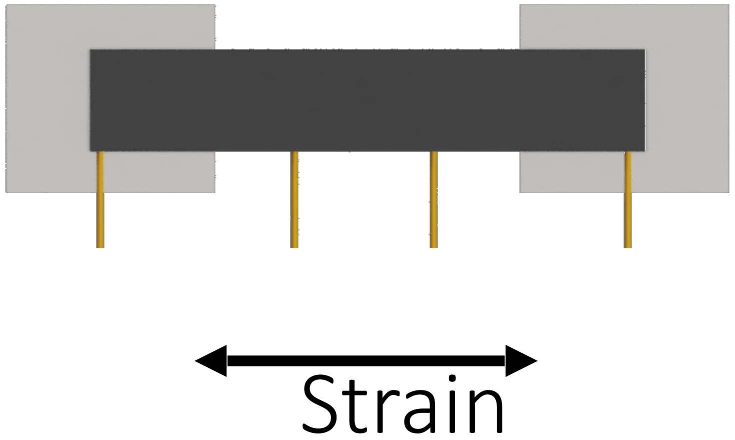
$BaFe_2As_2$ – Iron Based High Temperature Superconductor

Shua Sanchez, Prof. Jiun-Haw Chu's Quantum Materials group



- Project goal: apply strains and x-ray to precisely detwin a sample and strain-tune superconductivity.
- This crystal has 2 structural domains (A and B) where iron atoms form rectangular lattice.
- Under zero stress, the A and B domains have the same total volume (domain population). Applying tension detwins the crystal to turn B domains into A, and compression to A to B.
- We combined x-rays to measure the a and b lattice constants directly while applying strain.
- The video shows 162 strain states sequentially and the intensity of the x-ray diffraction on the area detector

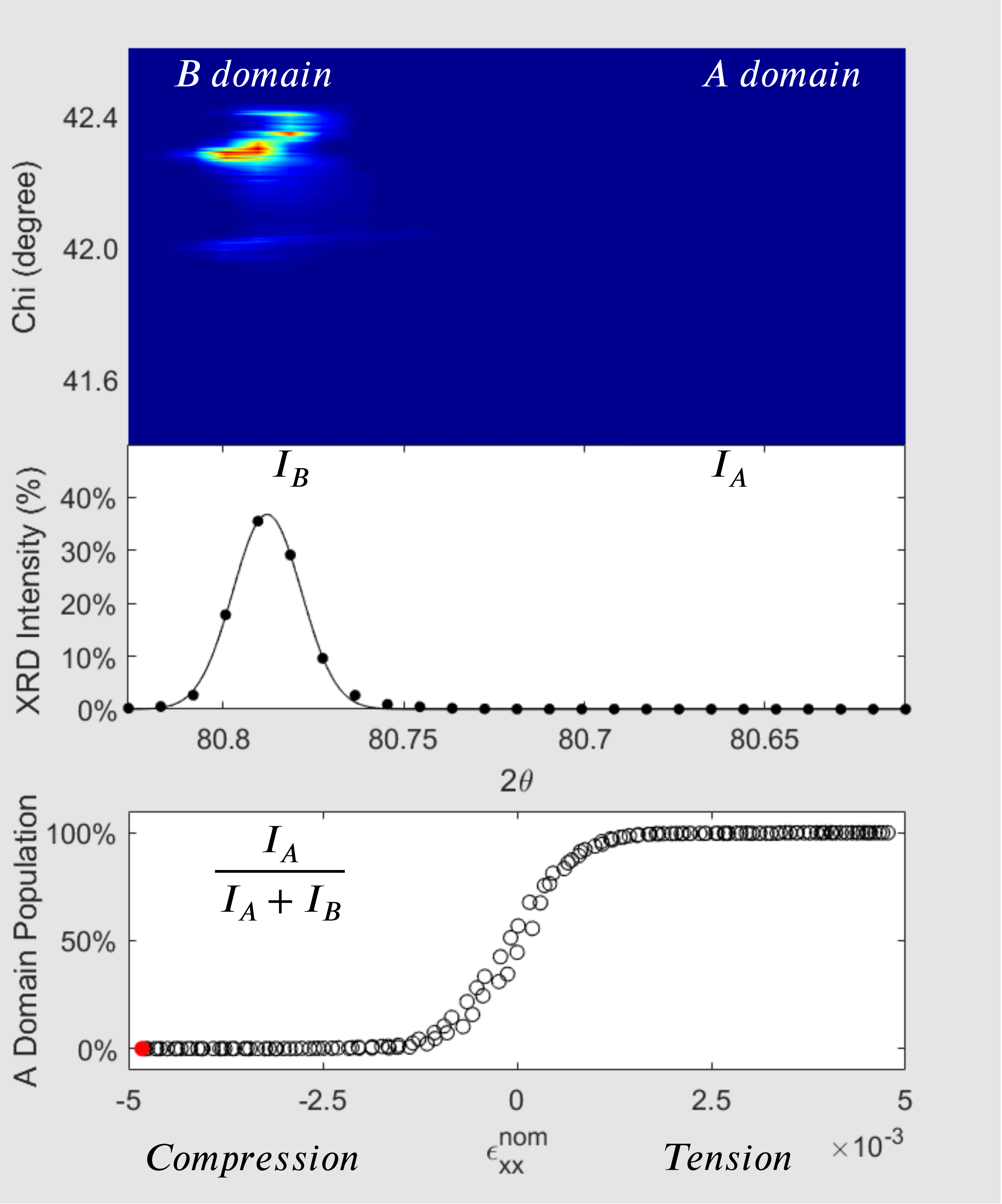




- (Top plot) the intensity position on the detector gives the length values of a and b which change with strain

- (Middle plot) the intensity is summed vertically and fit to 2 Gaussians

- (Bottom plot) The relative intensity $\dfrac{I_A}{I_A + I_B}$ gives the relative A domain population which change vs strain.



Interesting result!
Lattice constants freeze in place during detwinning!

Implies that the domain pinning is much softer than the crystal lattice

Can smoothly detwin the sample from B to A and back

Introductions

Thinking about statistics

What is significance anyway?

What is the question?

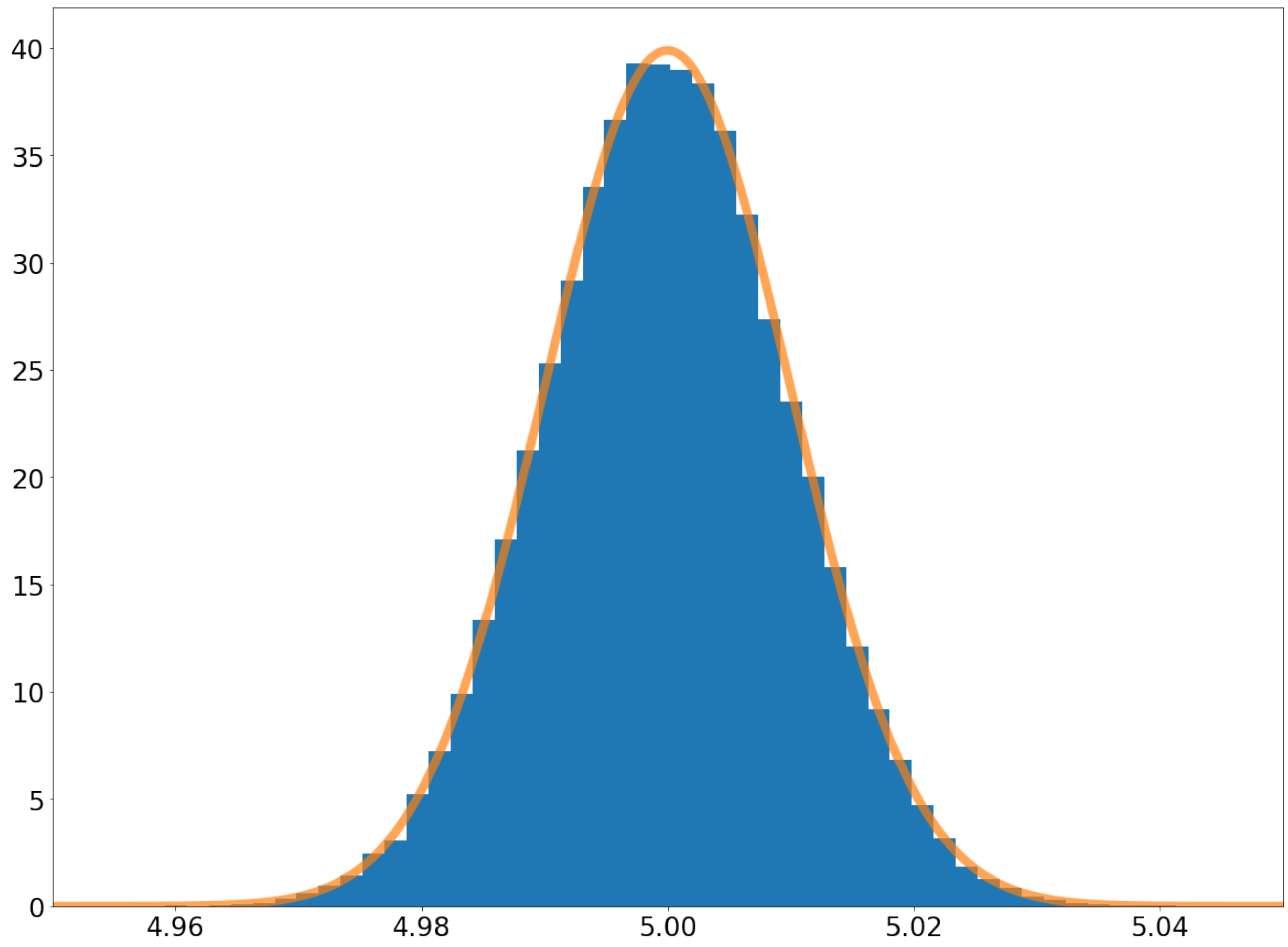
- Must clearly & precisely state the question

Example (null hypothesis):

If there is no signal; what is the probability that the background produces a signal that is equally or more signal-like than what I observed?

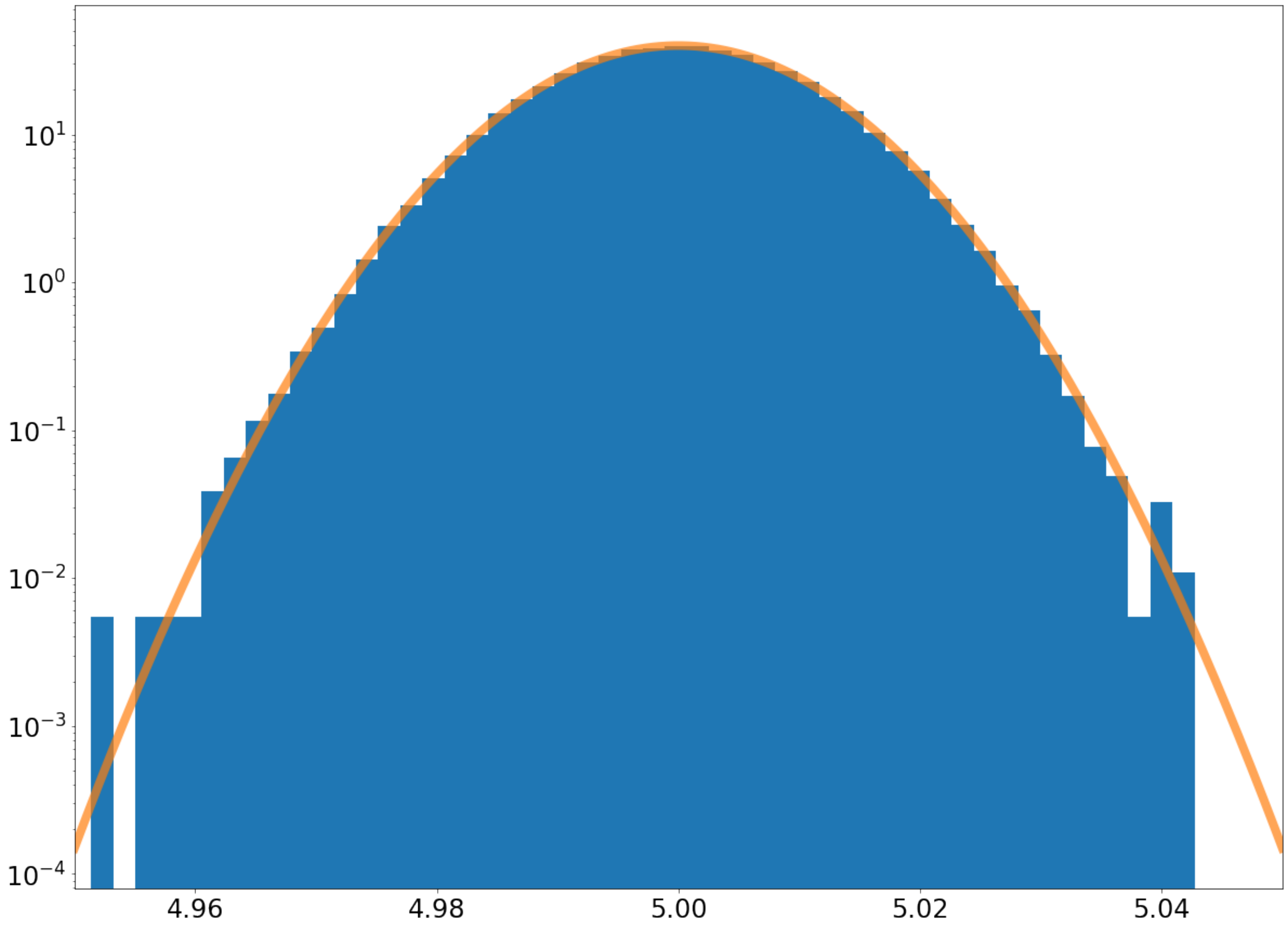
$$d = \frac{1}{\sigma\sqrt{2\pi}} e^{\frac{-(x-\mu)^2}{2\sigma^2}}$$

```
data = stats.norm.rvs(loc = 5., scale = 0.01, size = 100000)
```

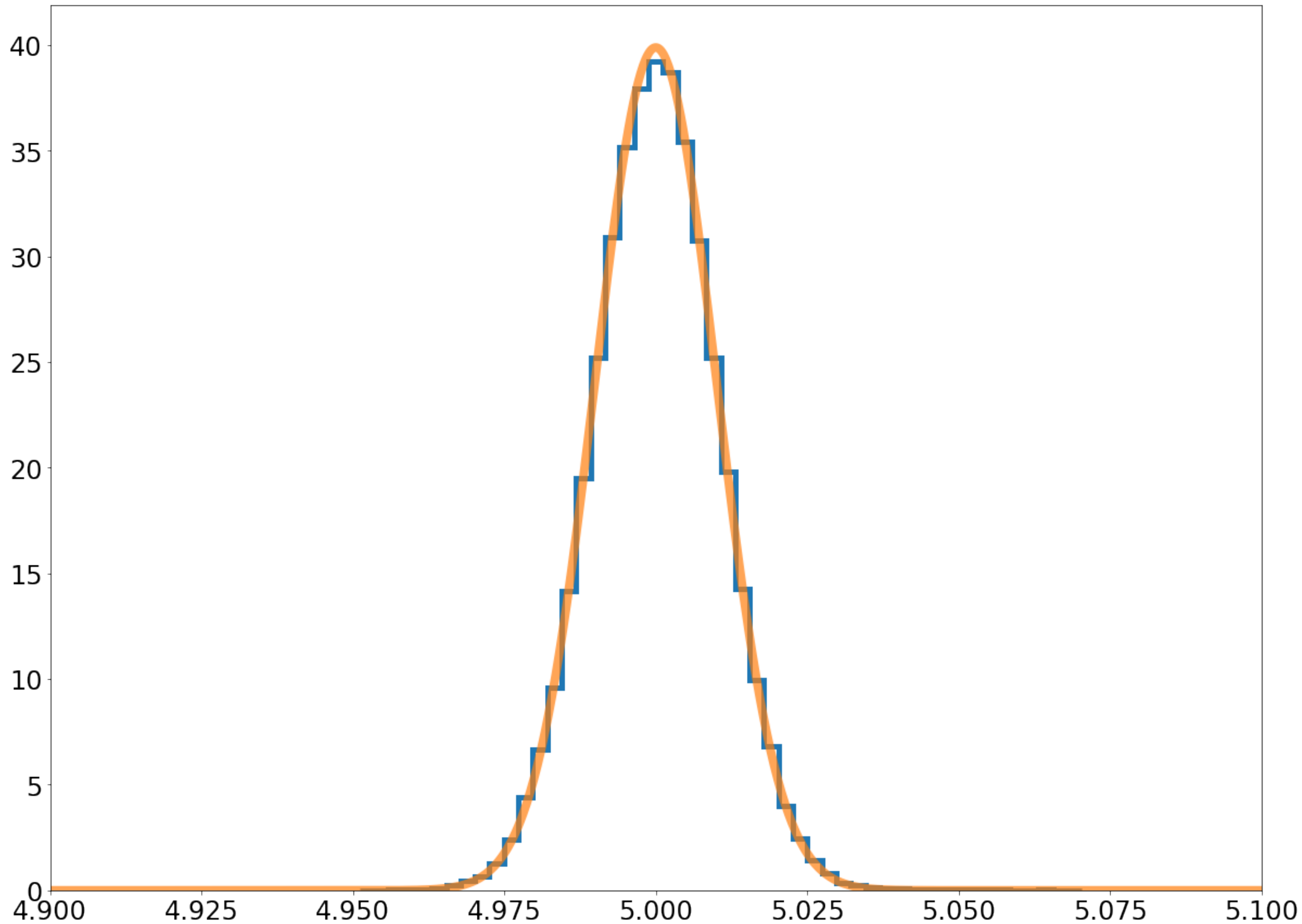


```
fig, ax = plt.subplots(1, 1)
ax.hist(data,50, density=True)
plt.yscale('log')
plt.tick_params(labelsize = 24)
plt.xlim([4.95,5.05])
x = np.linspace(4.95,5.05,1000)
ax.plot(x,stats.norm.pdf(x,loc = 5., scale = 0.01), linewidth = 8, alpha = 0.7)
plt.show()
```

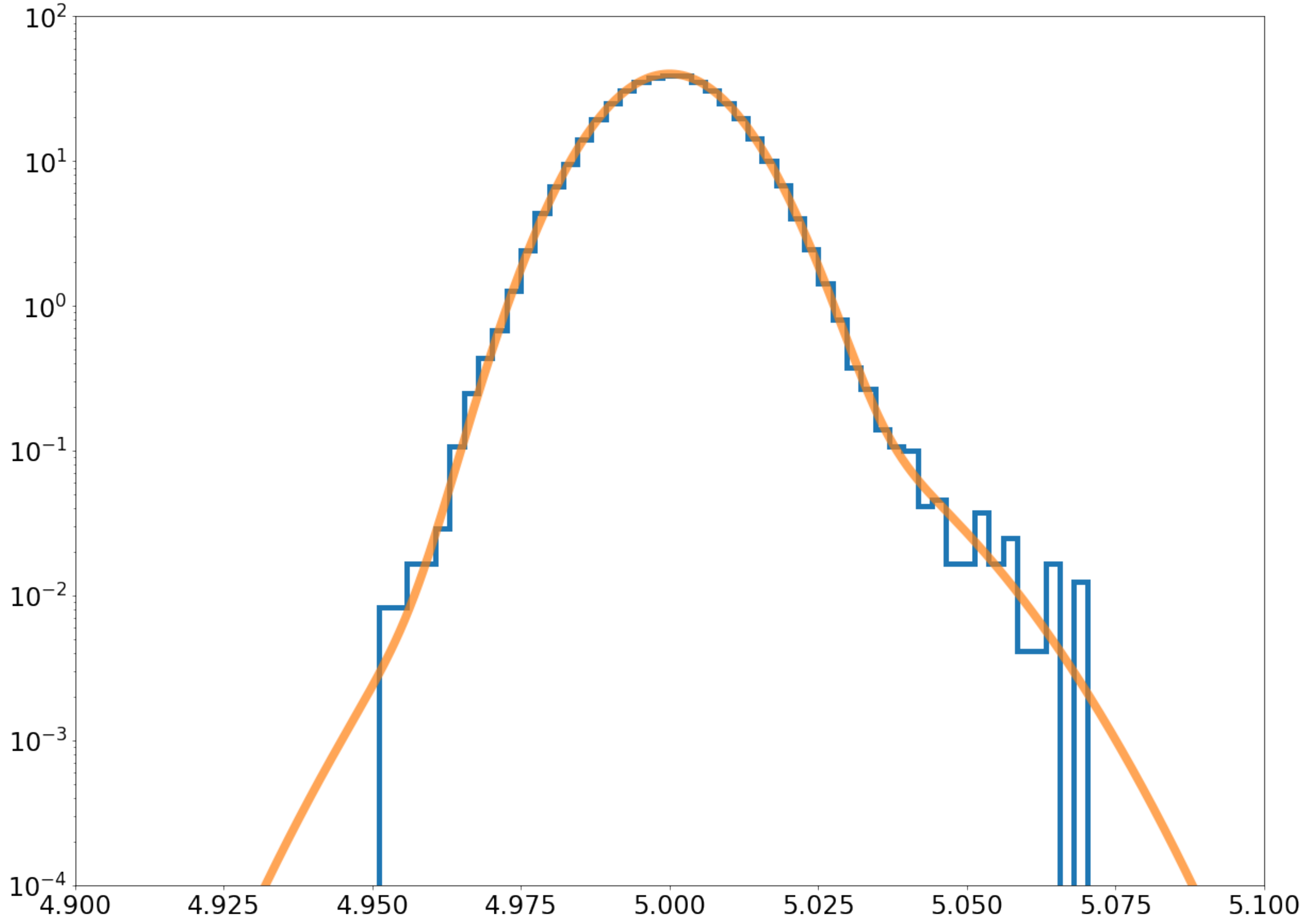
$$\log d = \frac{-(x-\mu)^2}{2\sigma^2} + c$$



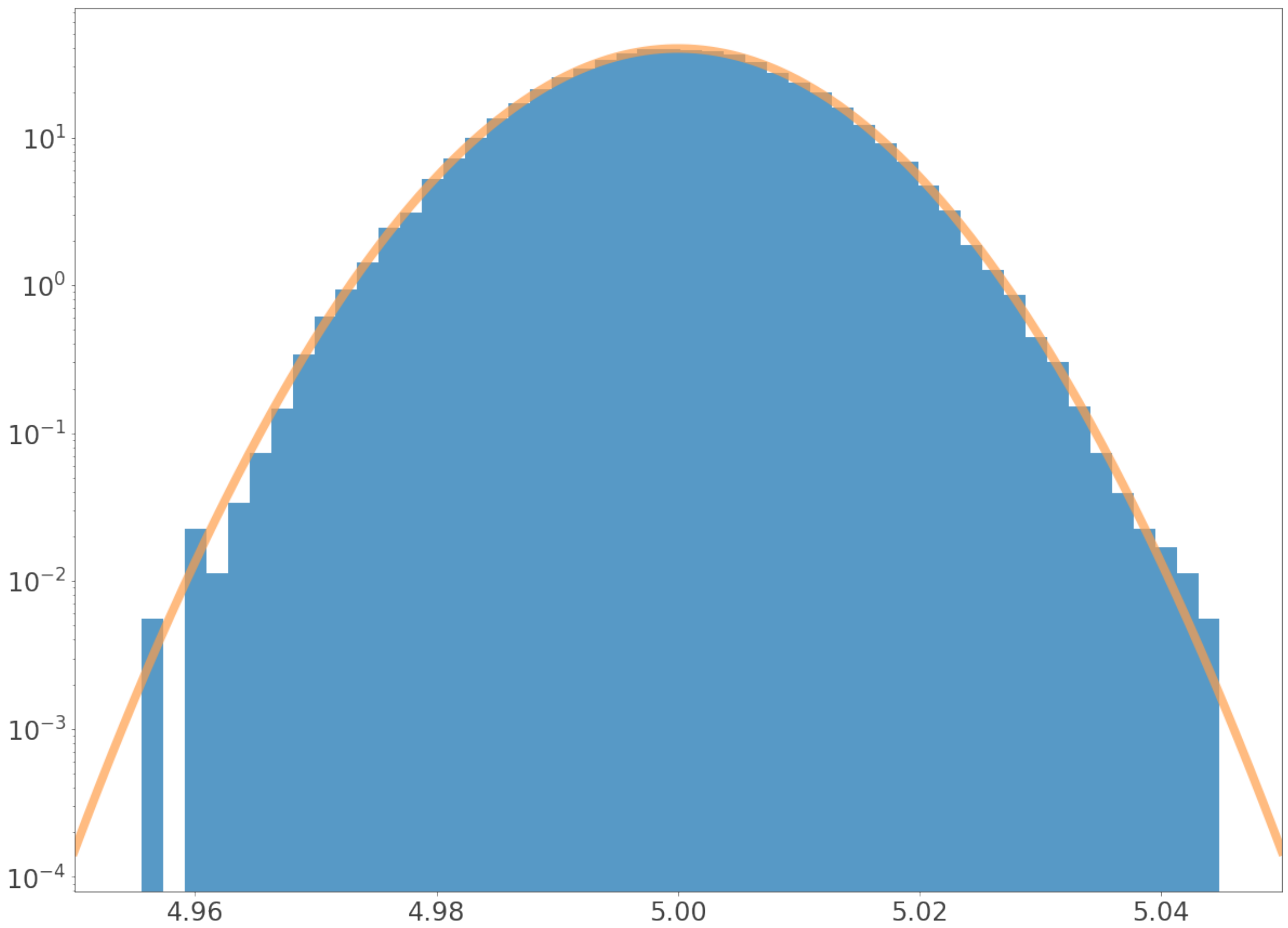
Never assume Gaussian statistics



Never assume Gaussian statistics



If there is no signal; what is the probability that the background produces a signal that is equally or more signal-like than what I observed?



$$\text{'Probability'} = \int_a^\infty \mathrm{pdf(x)dx}$$

in this case!

In physics, $X\sigma$ is shorthand for a probability

- 5σ means: the probability of signal-free data giving a measurement that is equally or more signal like than your observation is less than 2.87×10^{-7} (or 1 in 3.5 million)

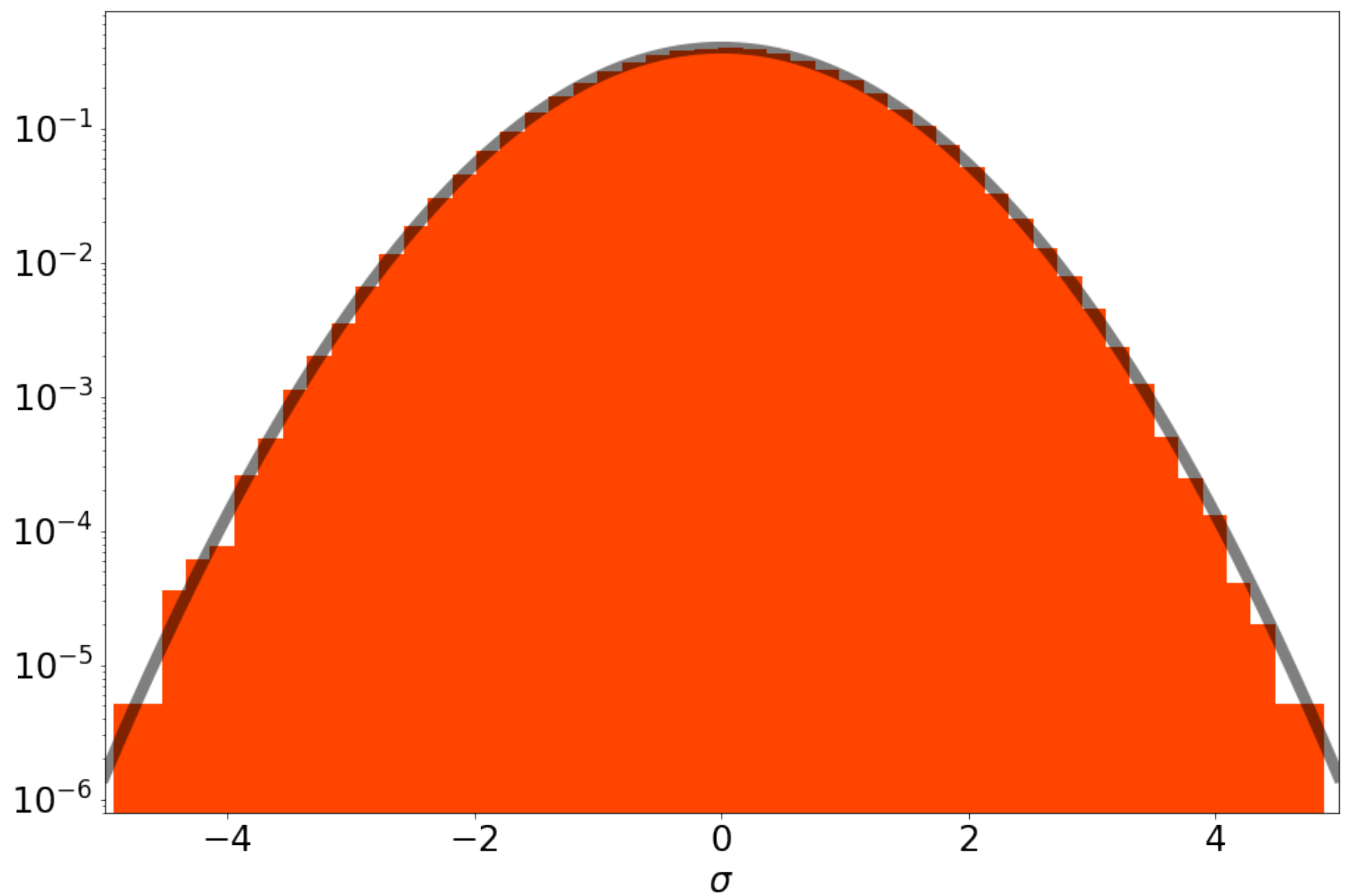
Common mistakes

- $X\sigma$ does not imply Gaussian distributed data
- $X\sigma$ is not $X\sigma$ away from the mean
- $X\sigma$ does not mean your question is one-sided

Best interpretation of $X\sigma$ (null hypothesis case)

- The probability of the background giving me a data point that looks as or more signal-like than the reading I have is the same probability as if my data was Gaussian and I was $X\sigma$ away from the mean

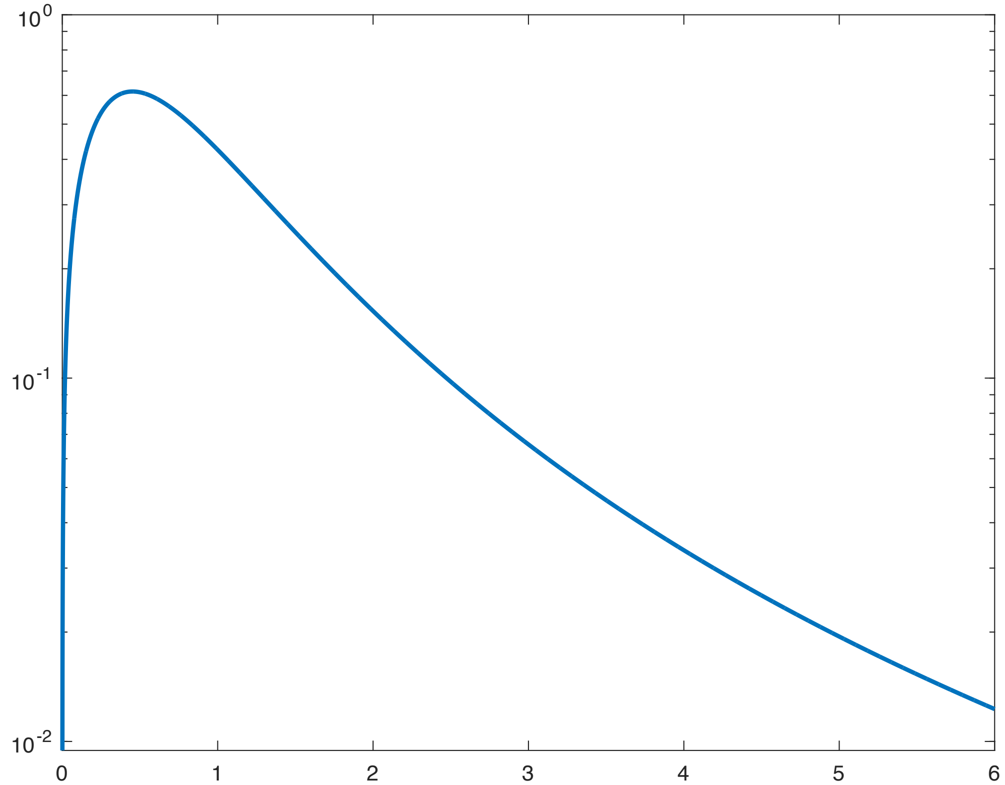
If there is no signal; what is the probability that the background produces a signal that is equally or more signal-like than what I observed?

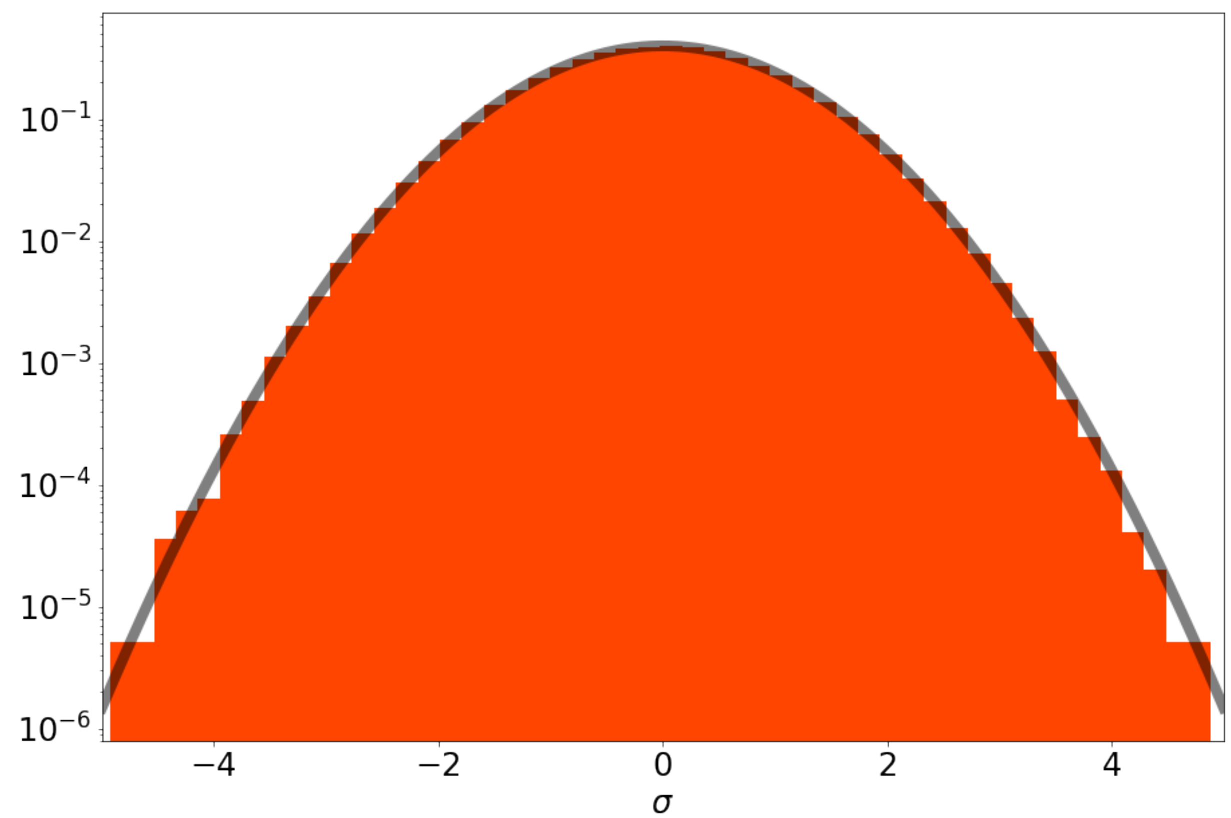


$$\text{Probability } X\sigma = \int_{X\sigma}^{\infty} \frac{1}{\sigma\sqrt{2\pi}} e^{\frac{-(x-\mu)^2}{2\sigma^2}} dx = \frac{1}{2}\text{erfc}\left(\frac{X}{\sqrt{2}}\right)$$

Best interpretation of $X\sigma$ (null hypothesis case)

- The probability of the background giving me a data point that looks as or more signal-like than the reading I have is the same probability as if my data was Gaussian and I was $X\sigma$ away from the mean





Key statistical steps

- Clearly state the question (& turn into math)
- Determine the background distribution
- Integrate background to find probability
- Convert probability into equivalent sigma

$$\text{'Probability'} = \int_a^\infty \mathrm{pdf(x)dx}$$