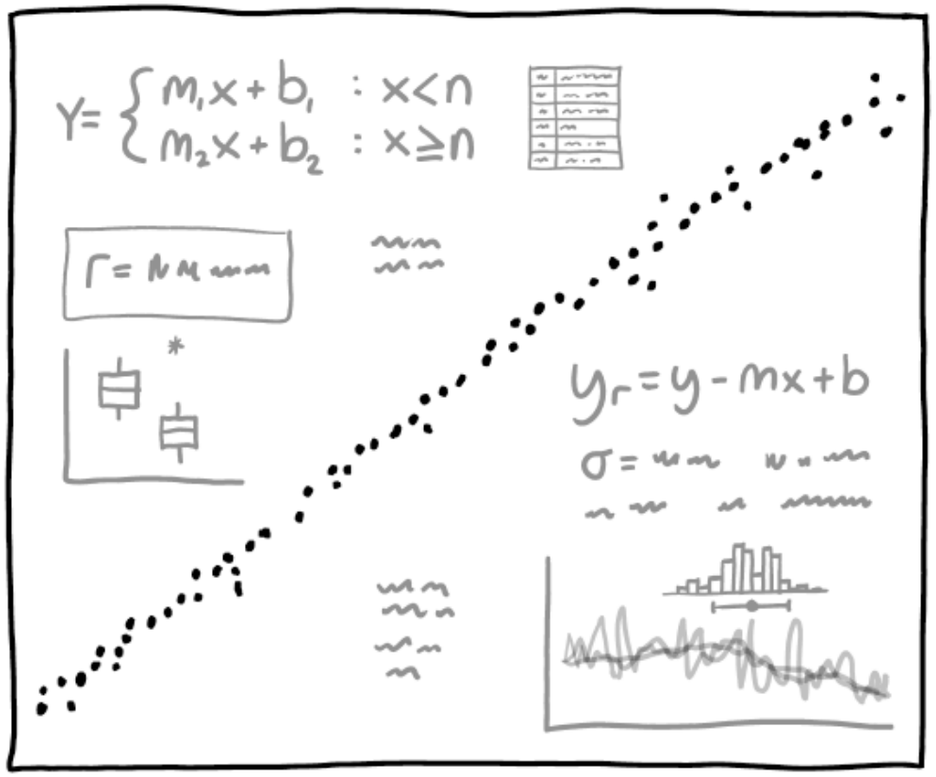
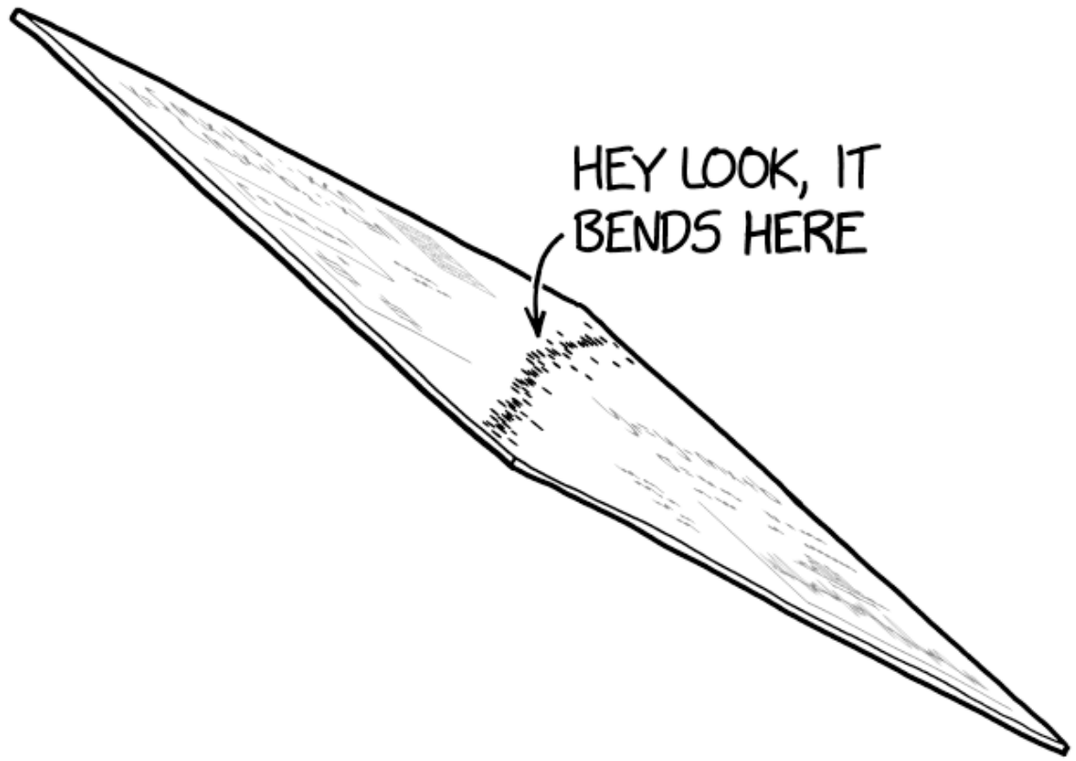
HOW TO DETECT A CHANGE IN THE SLOPE OF YOUR DATA

NOVICE METHOD:



DO A BUNCH OF STATISTICS

EXPERT METHOD:



TIP THE GRAPH SIDEWAYS

https://xkcd.com/2701/

Class 19: Blind & semi-blind analyses, Killing Trees for ***Science!***

Miguel F. Morales
Bryna Hazelton

Blind & semi-blind analyses

What is selection bias?

1) Try some data cuts

2) Look to see if the answer is ‘better’

   a) If yes, keep cuts

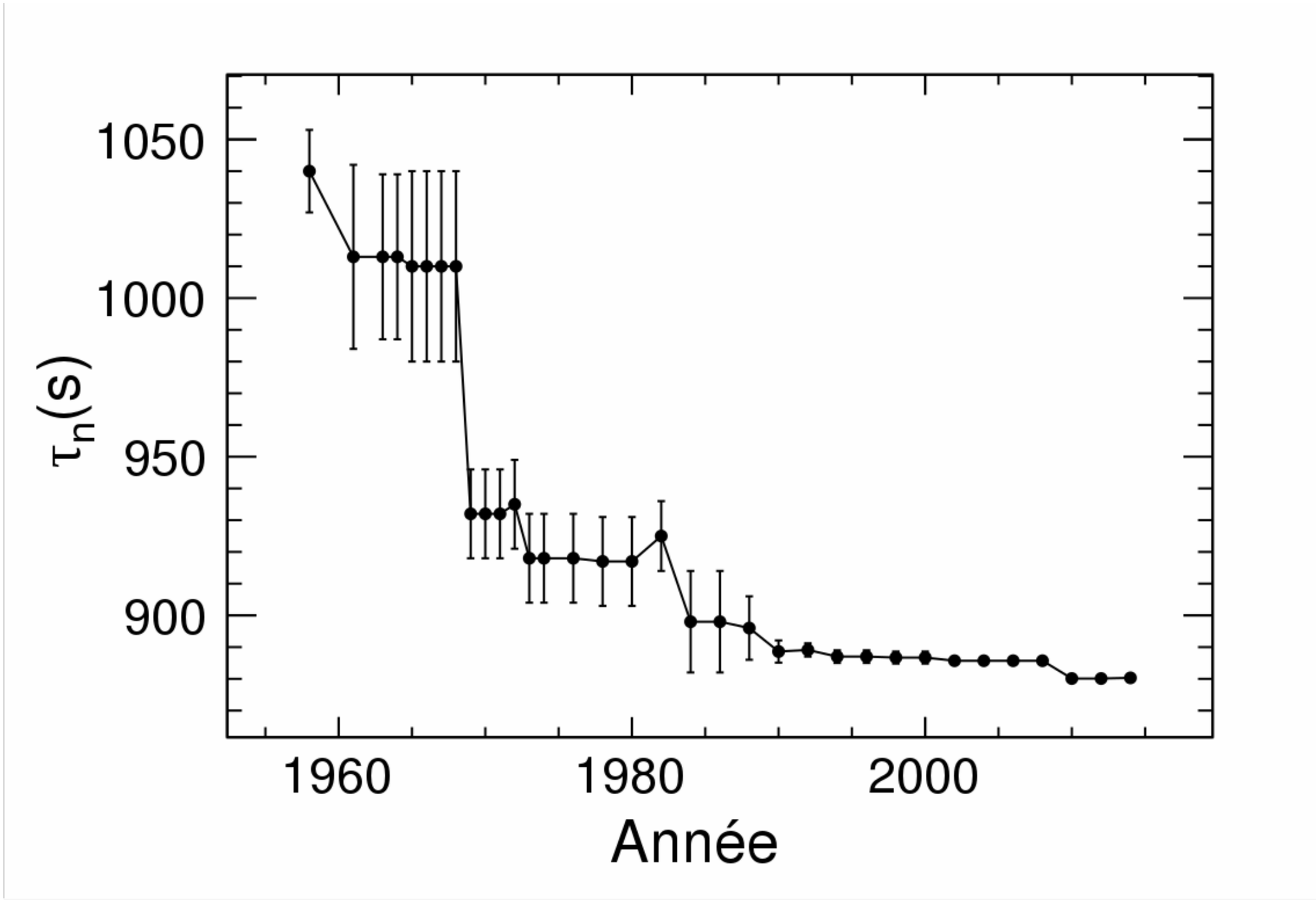
   b) If no, don’t use cuts

3) Return to step 1

Selection bias

- Higher significance (signal boosting)
- Getting a ‘better’ answer (value steering)

Neutron lifetime



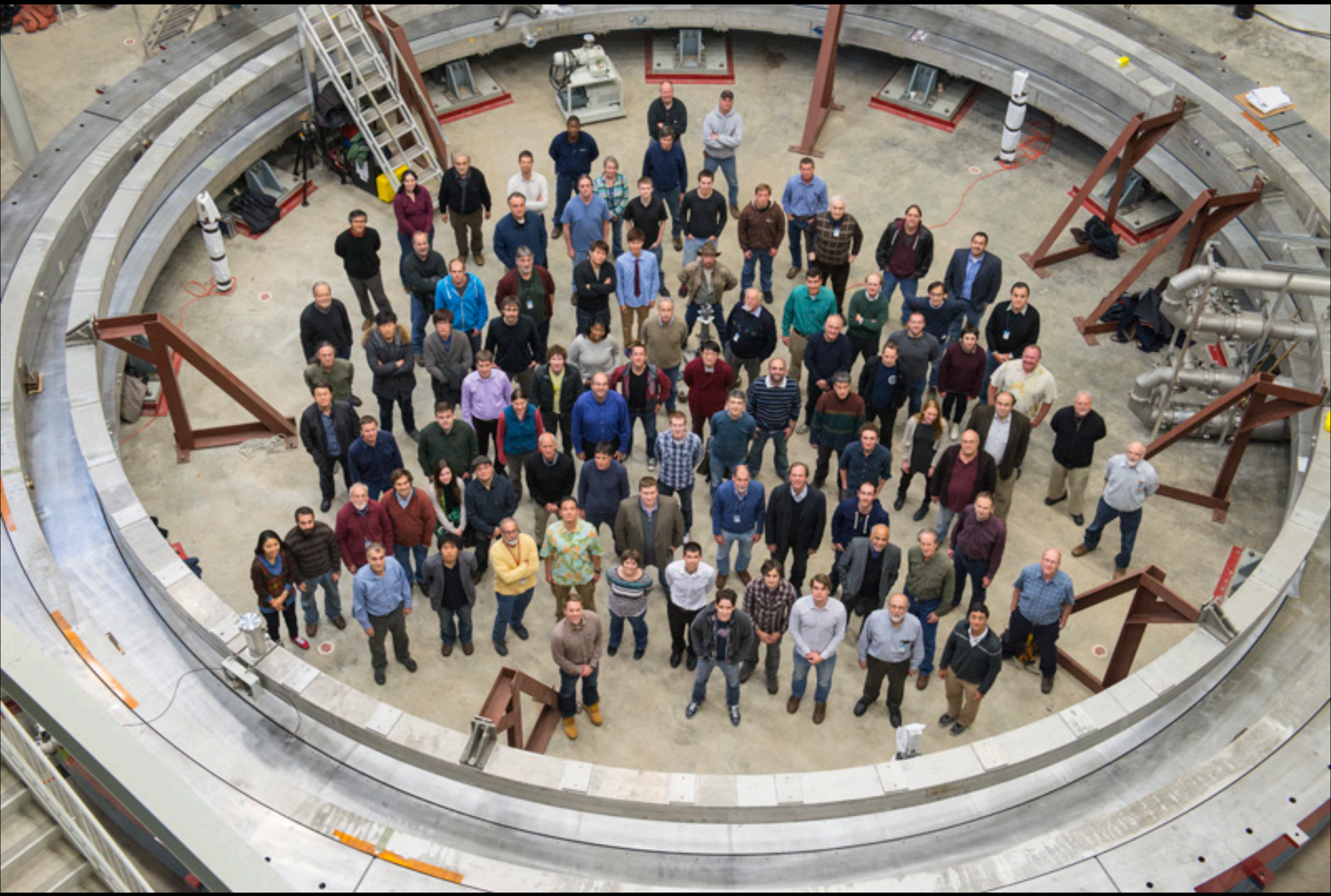
So easy to do

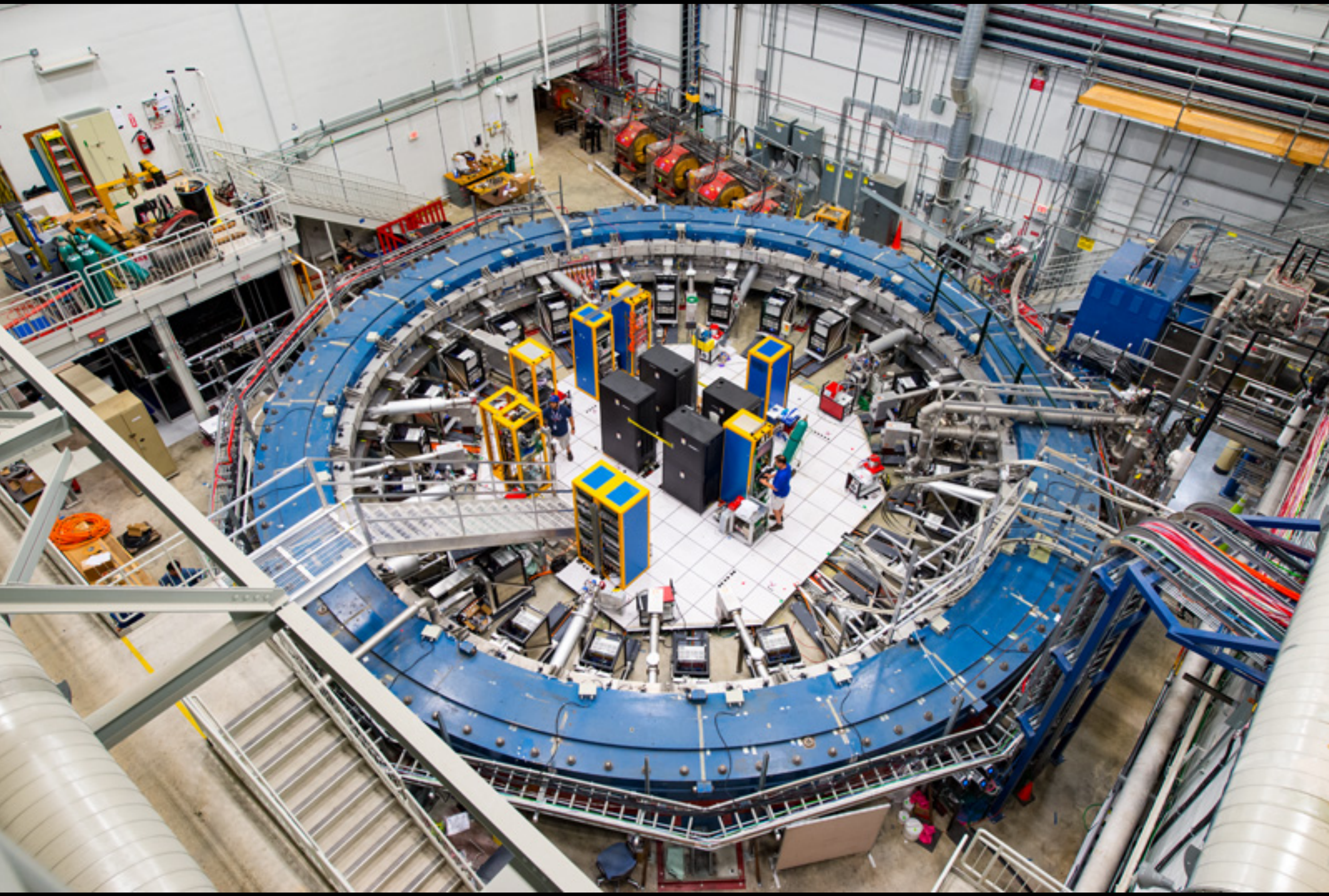
- Nothing malicious
- You have to filter your data
- Natural human response, science is done by people

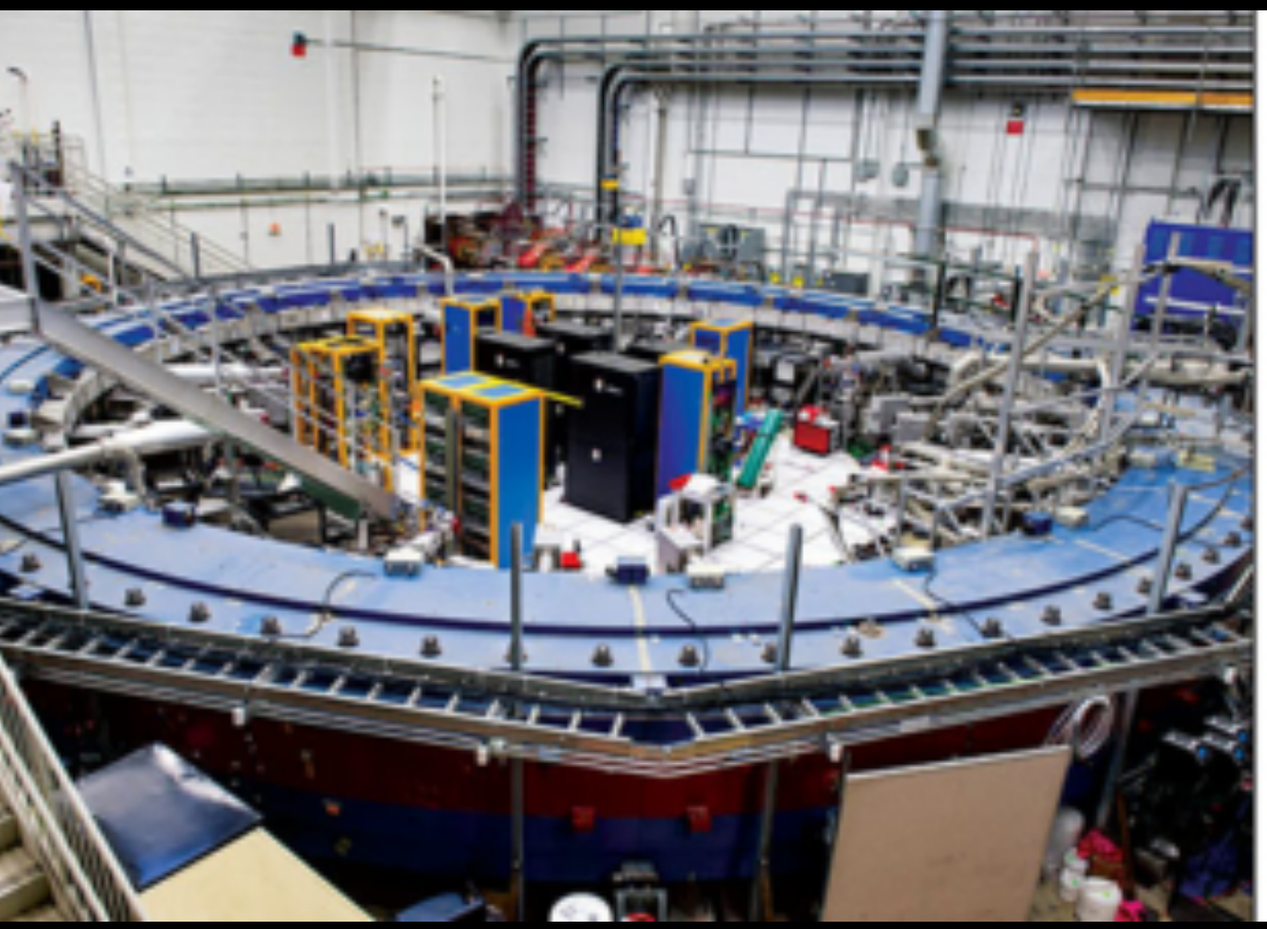
Goal is to not make decisions based
on how it affects the answer

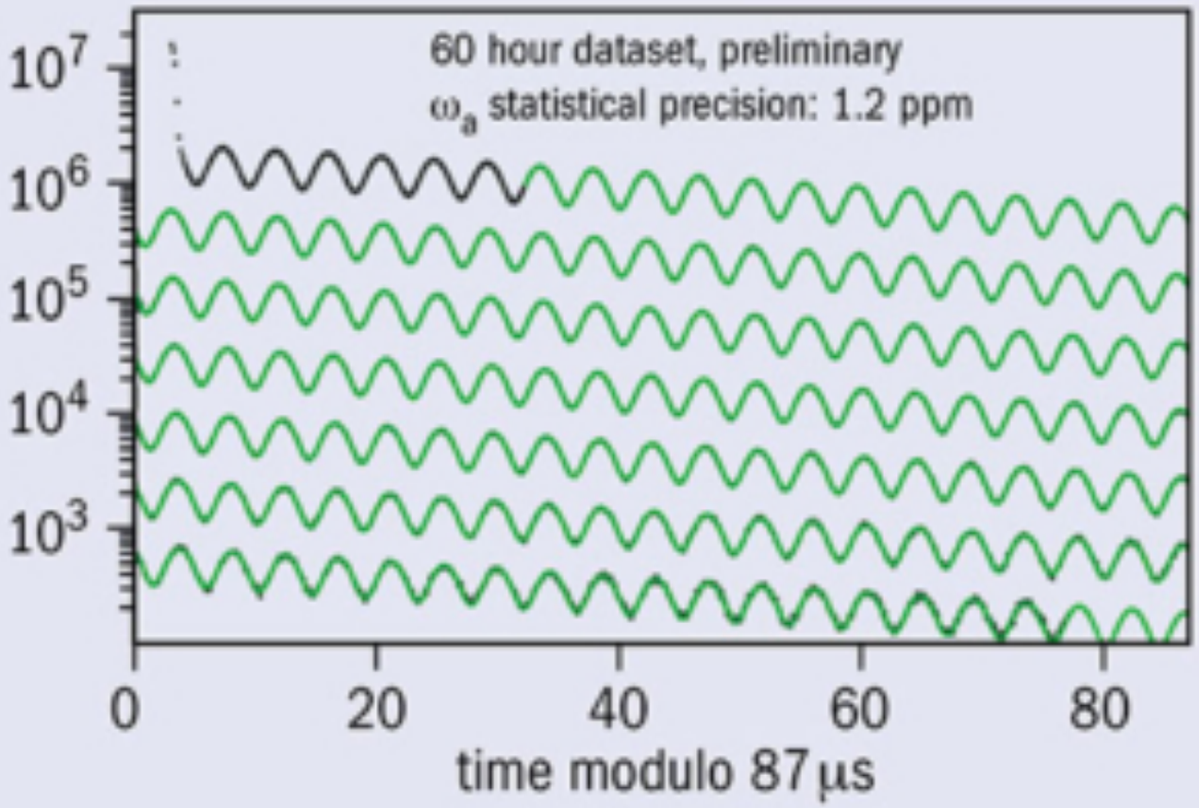
Blind analysis

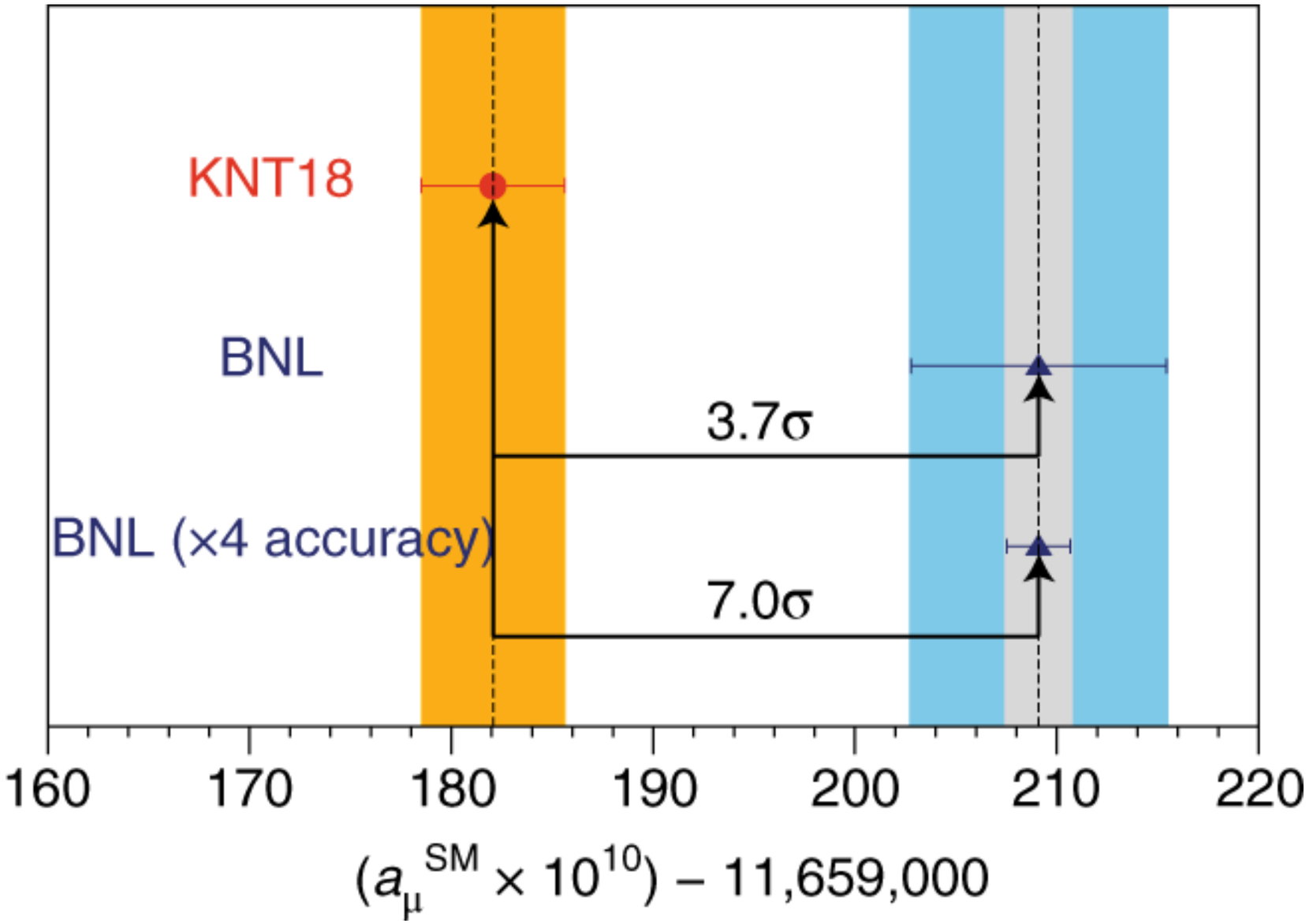
- You don’t know what the answer is until the end
- Option 1: develop all cuts on simulation or small data sample (LHC)
- Option 2: fake signal is injected, but you don’t know which are real (LIGO)
- Option 3: data is obfuscated, only corrected after analysis is done (unblinding)








60 hour dataset, preliminary
ω_a statistical precision: 1.2 ppm
10^7
10^6
10^5
10^4
10^3
0
20
40
60
80
time modulo 87 μs

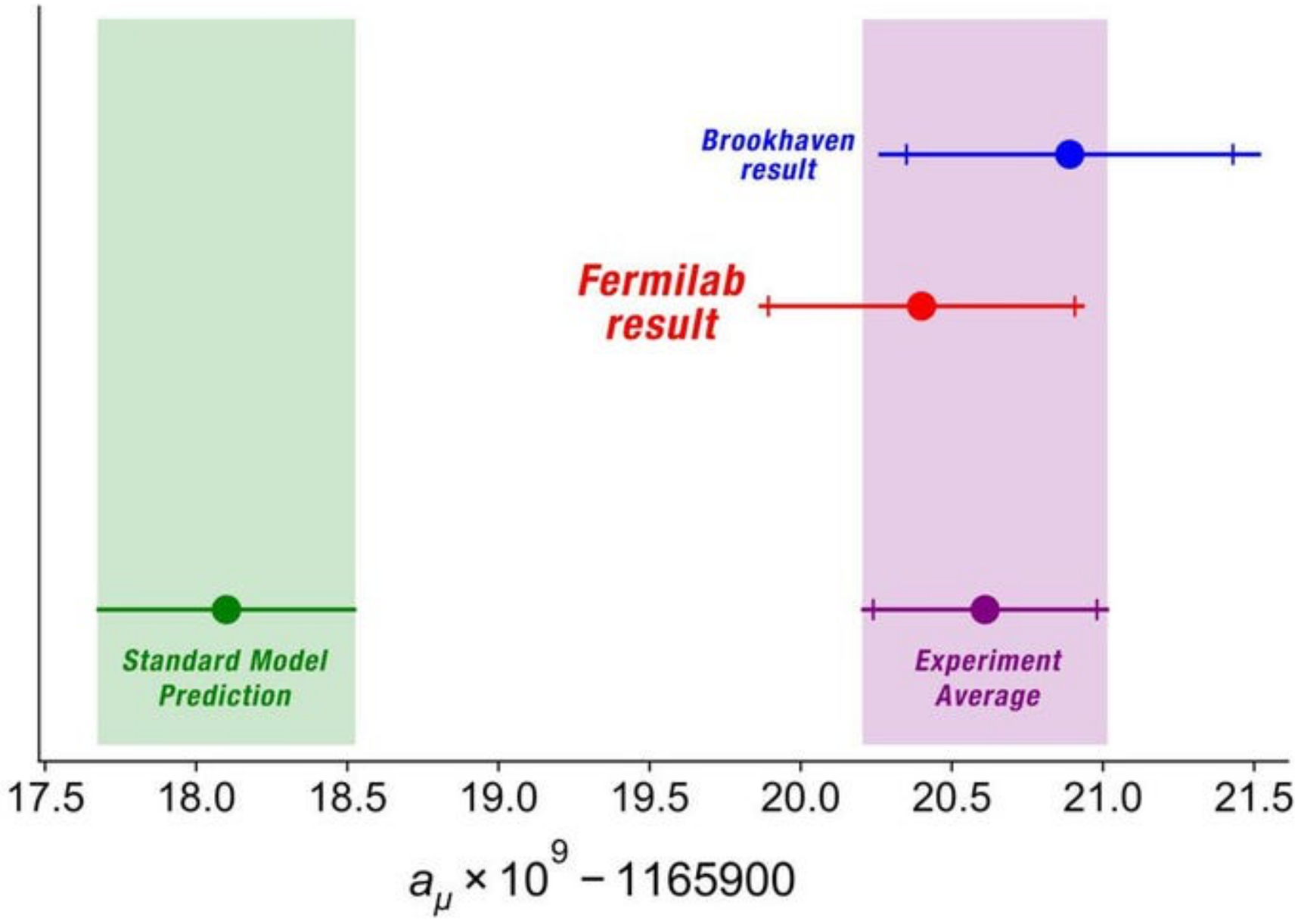

KNT18
BNL
3.7σ
BNL (×4 accuracy)
7.0σ
160
170
180
190
200
210
220
$(a_\mu^{SM} \times 10^{10}) - 11{,}659{,}000$

g-2 tiered unblinding

- Entire experiment is run off a mis-set clock. Only two people, both outside collaboration, know what clock is actually set to.

- Stages in data analysis where data is off by different amounts to allow relative unblinding on smaller subsets.

5:49

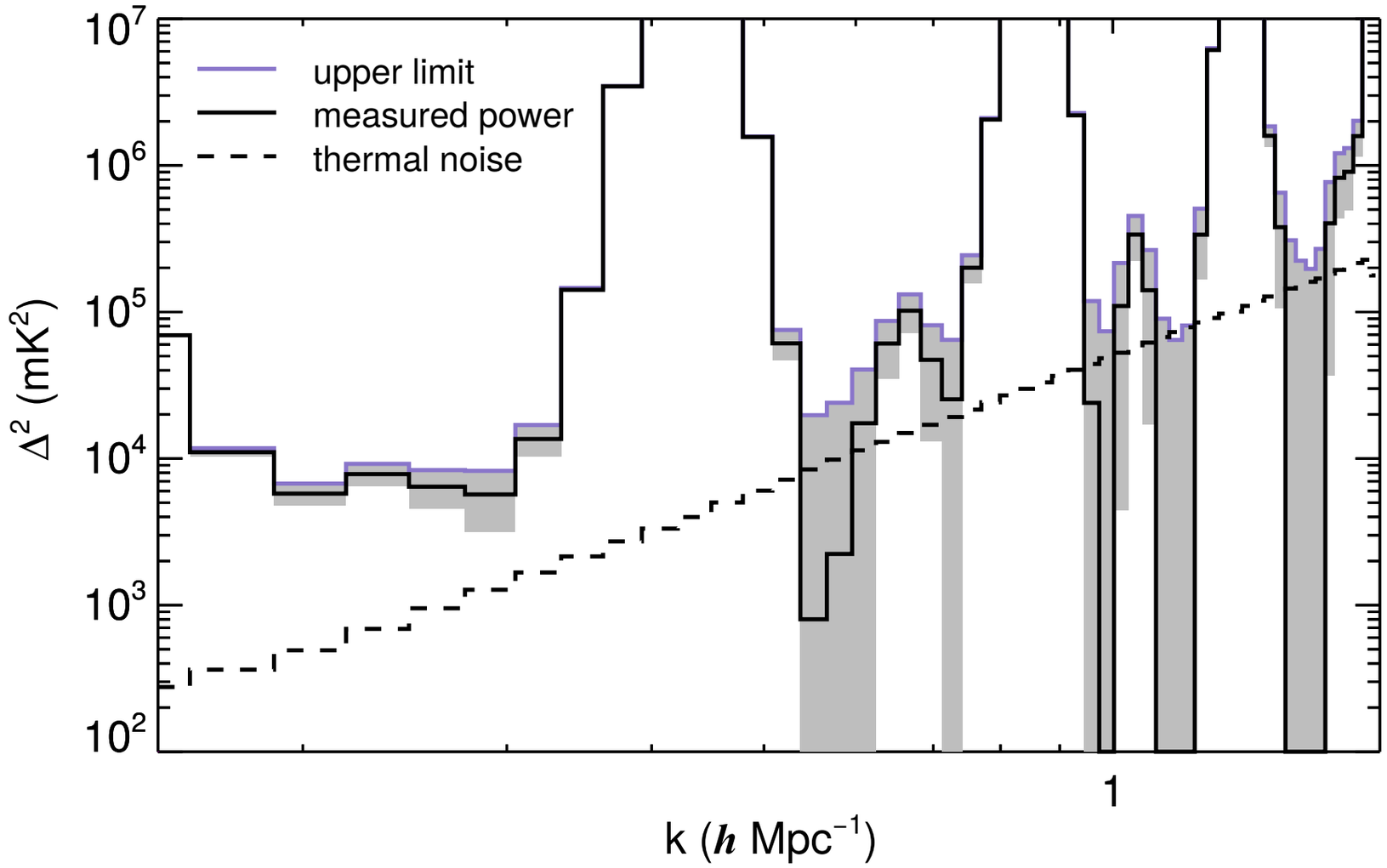

Brookhaven result
Fermilab result
Standard Model Prediction
Experiment Average
17.5
18.0
18.5
19.0
19.5
20.0
20.5
21.0
21.5
$a_\mu \times 10^9 - 1165900$

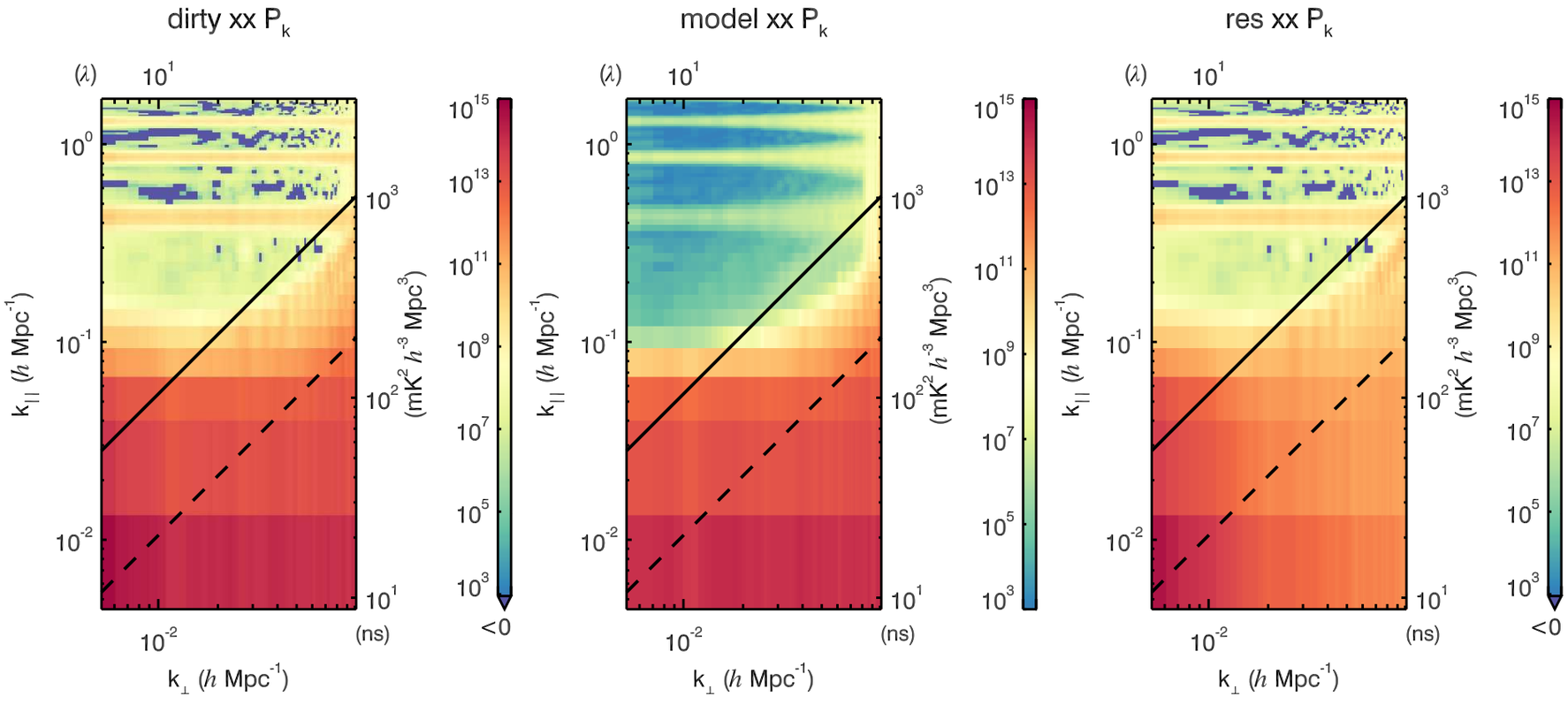
Pros & cons of blind analysis

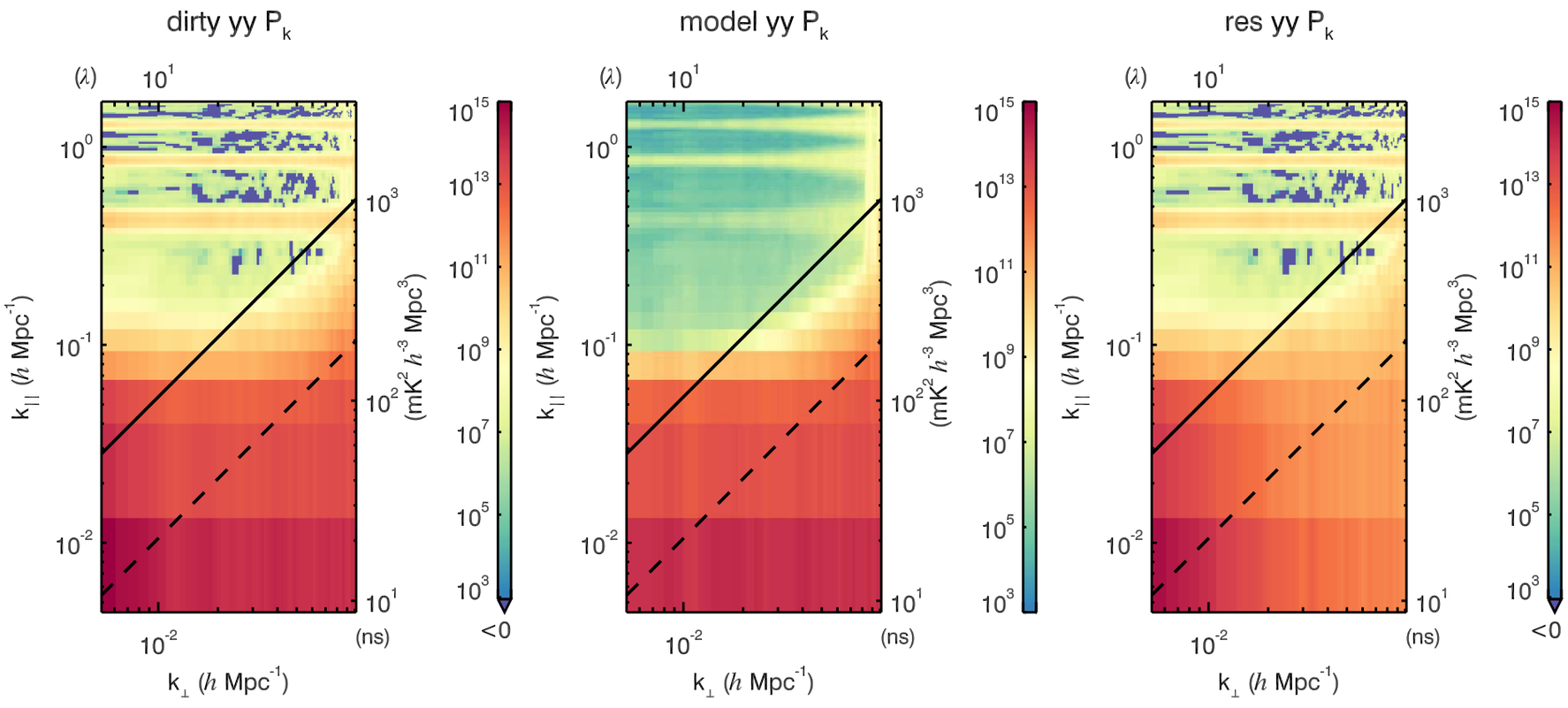
- Really works (for selection bias)
- Very slow (uncertainty of unblinding)
- Hampers data exploration and hunting for new systematics (slow learning)

What if blind doesn’t work?

Try not to look at the answer (intermediate diagnostics)



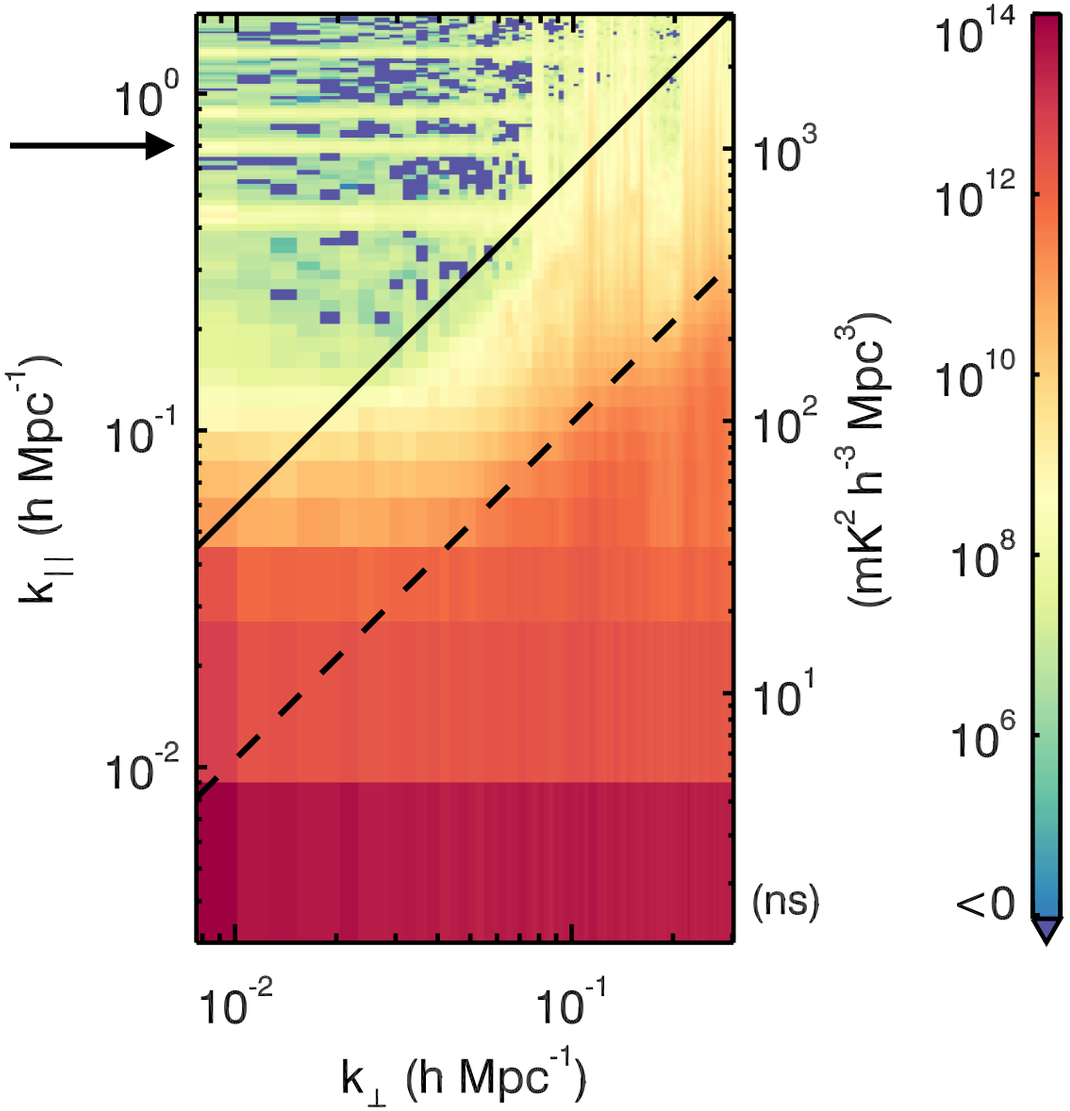

dirty xx P_k
model xx P_k
res xx P_k
(λ)
10^1
10^0
10^-1
10^-2
10^3
10^2
10^1
k_|| (h Mpc^-1)
(mK^2 h^-3 Mpc^3)
10^15
10^13
10^11
10^9
10^7
10^5
10^3
<0
10^-2
(ns)
k_⊥ (h Mpc^-1)


dirty yy P_k
model yy P_k
res yy P_k
(λ)
10^1
10^0
10^-1
10^-2
10^3
10^2
10^1
k_|| (h Mpc^-1)
(mK^2 h^-3 Mpc^3)
10^15
10^13
10^11
10^9
10^7
10^5
10^3
<0
10^-2
(ns)
k_⊥ (h Mpc^-1)

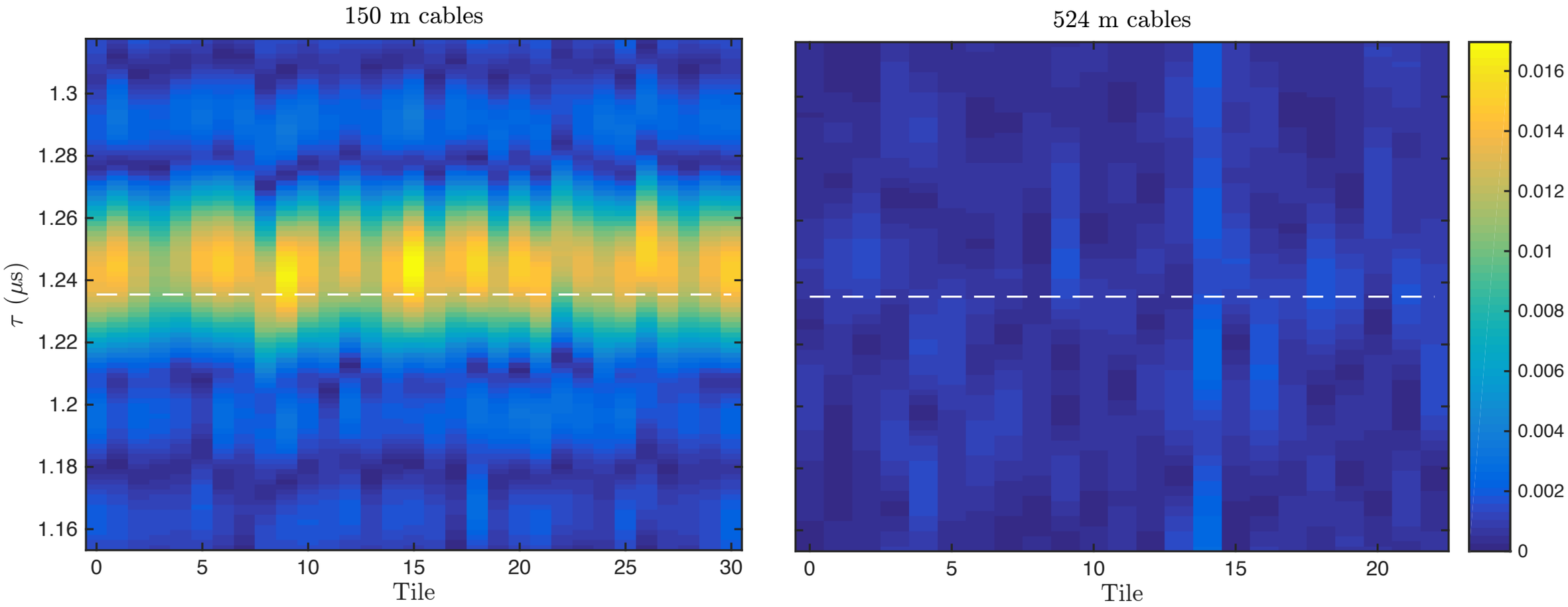
Don’t make decisions based on the science result

- Identify an effect (often is science associated intermediate plot)
- Hypothesize what might be causing it (instrumental systematics)
- Create tests to see if instrumental hypotheses are correct (jackknifes usually)
- Correct identified systematics *even* if the intermediate plot gets worse

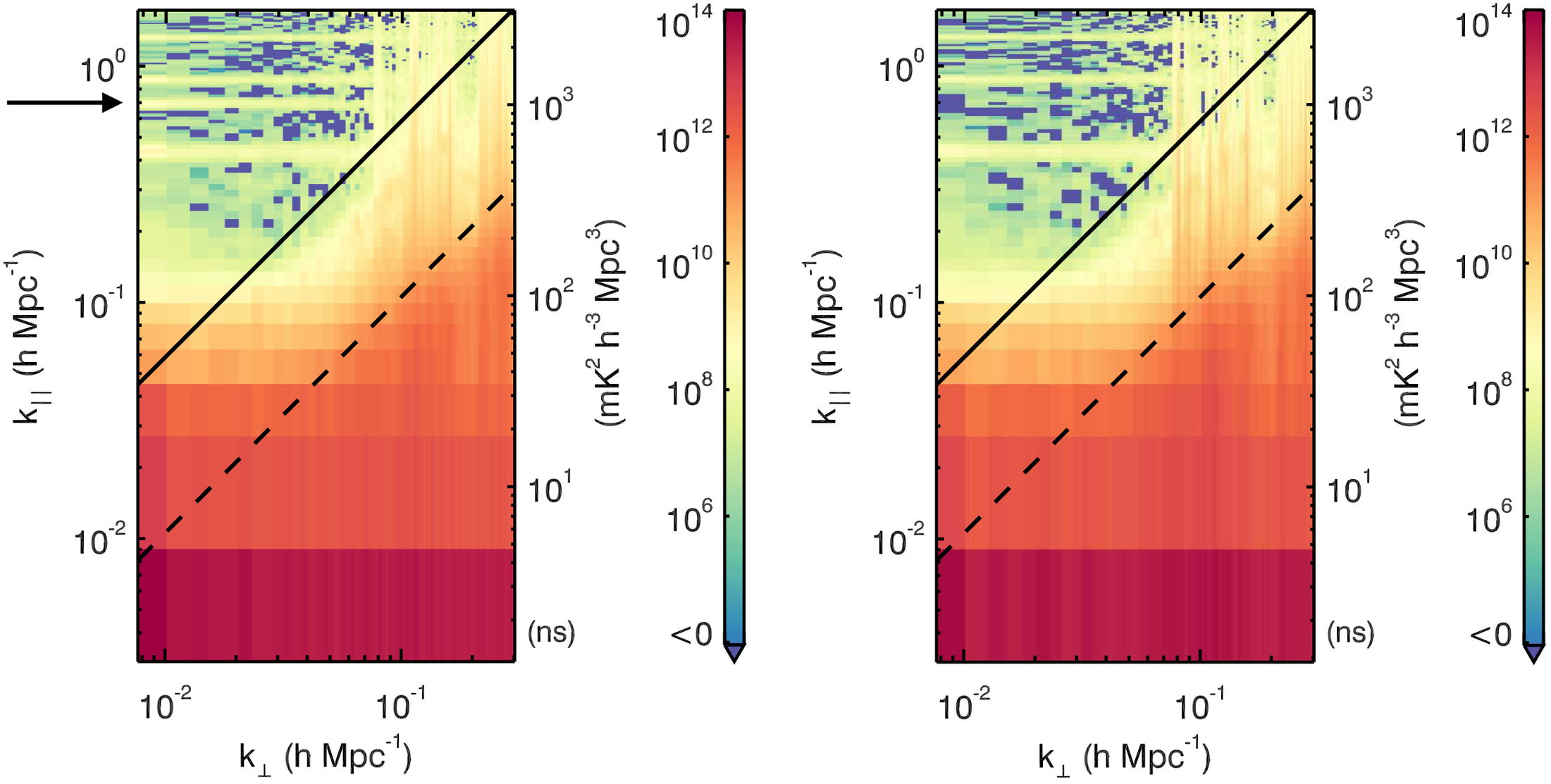
Example


10^0
10^{-1}
10^{-2}
$k_{||}$ (h Mpc^{-1})
10^3
10^2
10^1
(ns)
10^{-2}
10^{-1}
$k_\perp$ (h Mpc^{-1})
10^{14}
10^{12}
10^{10}
10^8
10^6
<0
(mK^2 h^{-3} Mpc^3)

Instrument test



Apply


10^0
10^{-1}
10^{-2}
$k_{||}$ (h Mpc^{-1})
10^3
10^2
10^1
(ns)
10^{-2}
10^{-1}
$k_\perp$ (h Mpc^{-1})
10^{14}
10^{12}
10^{10}
10^8
10^6
<0
(mK^2 h^{-3} Mpc^3)

Pros & cons of semi-blind

- Can explore in a way that is often hard in blind analyses (learn fast)
- Selection bias still present at some level
- Takes self discipline and faith

Blind: gold standard, only sometimes useful

- Beautiful simulations
- Well understood systematics
- Long experimental heritage

Selection bias: protecting you from you

- Blind means you don’t know what the answer is while tuning the analysis
- Semi-blind means you don’t use the answer to make decisions
- For you, not the audience

Data Rampage!


EXIT
EXIT
EXIT

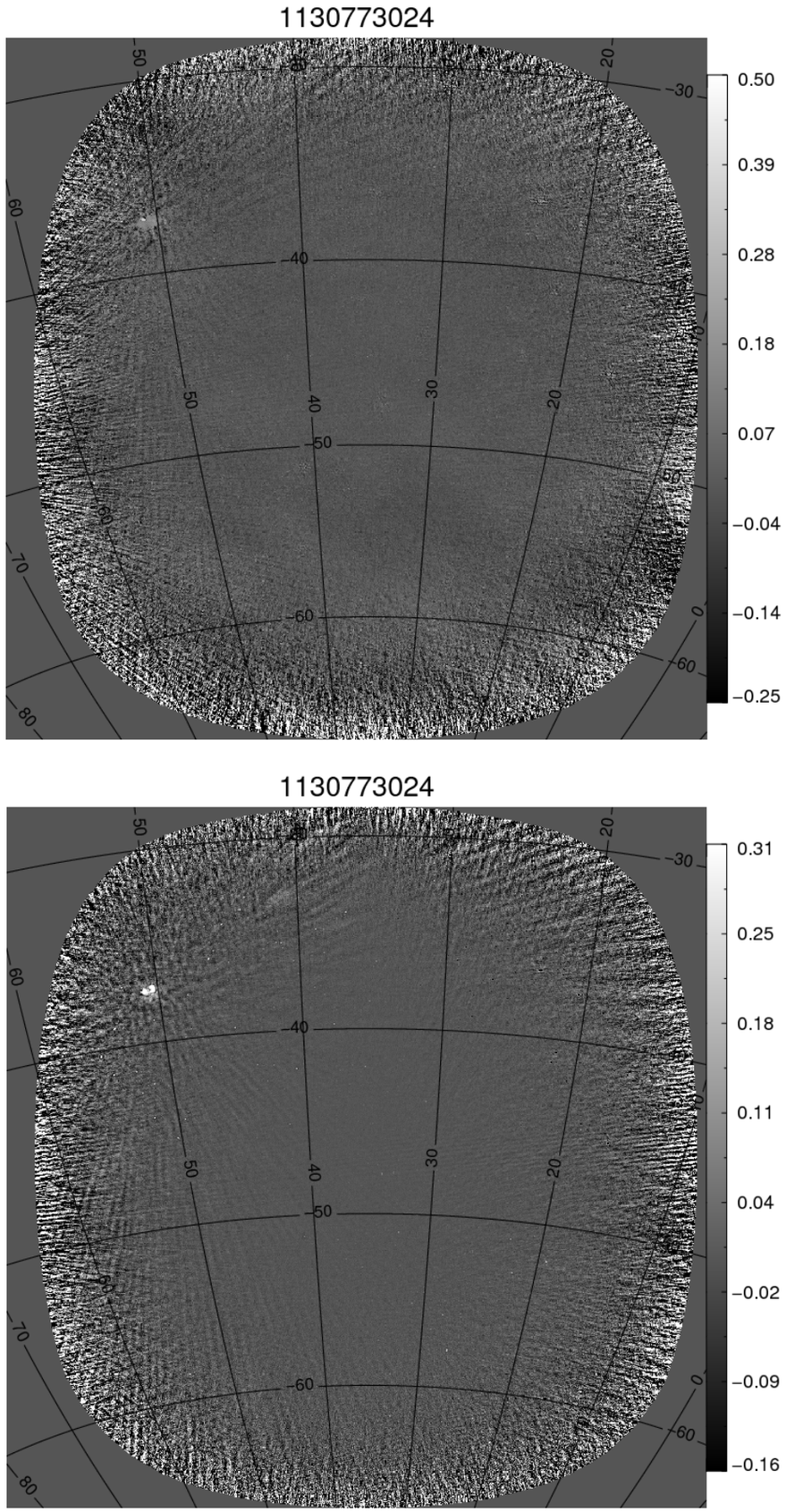
Collaborative visual brainstorming

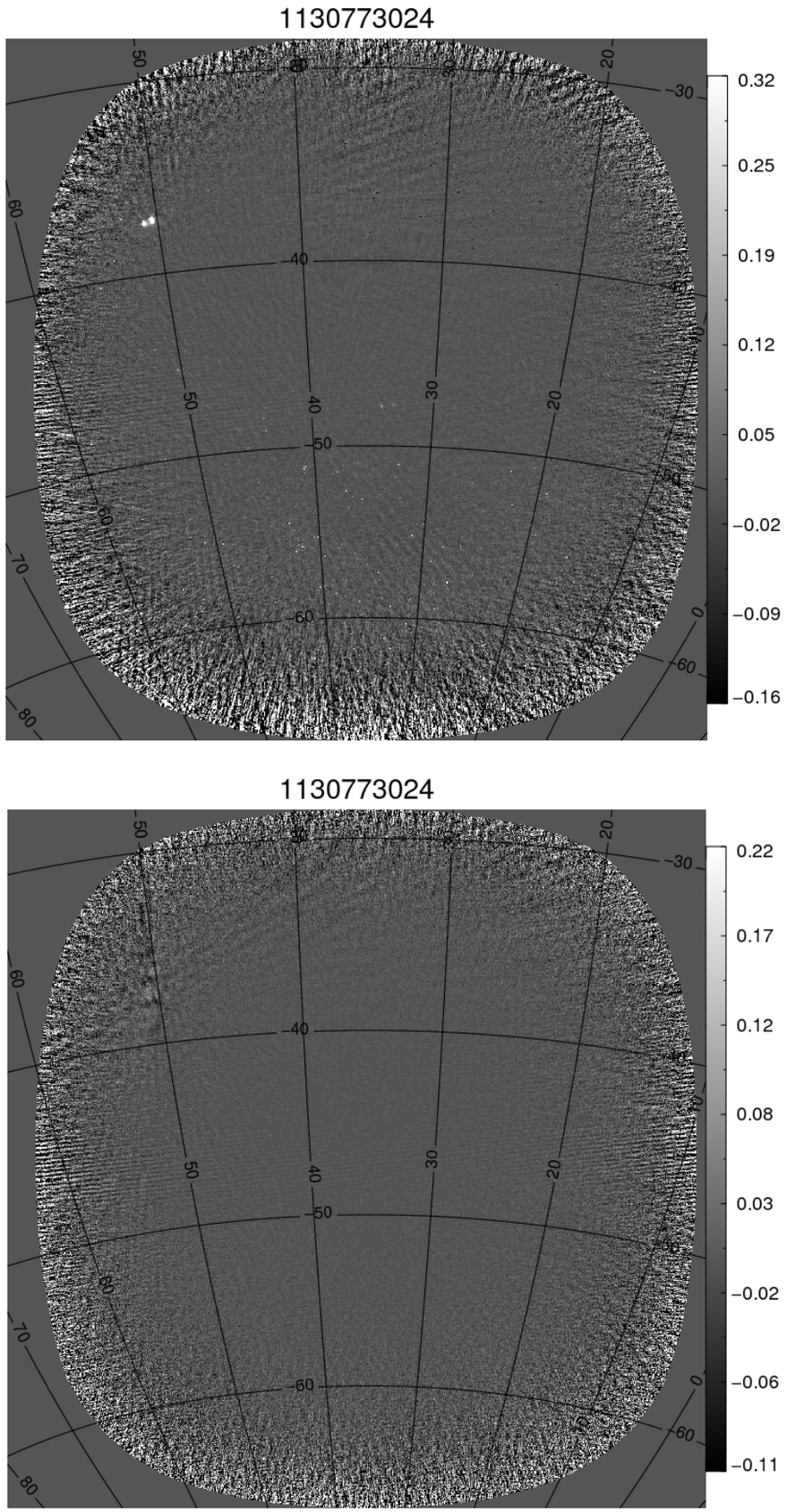
- Sense some pattern is hiding in the data, *or*
- What is the next problems/systematics/worries to attack?

Structured way to explore the data

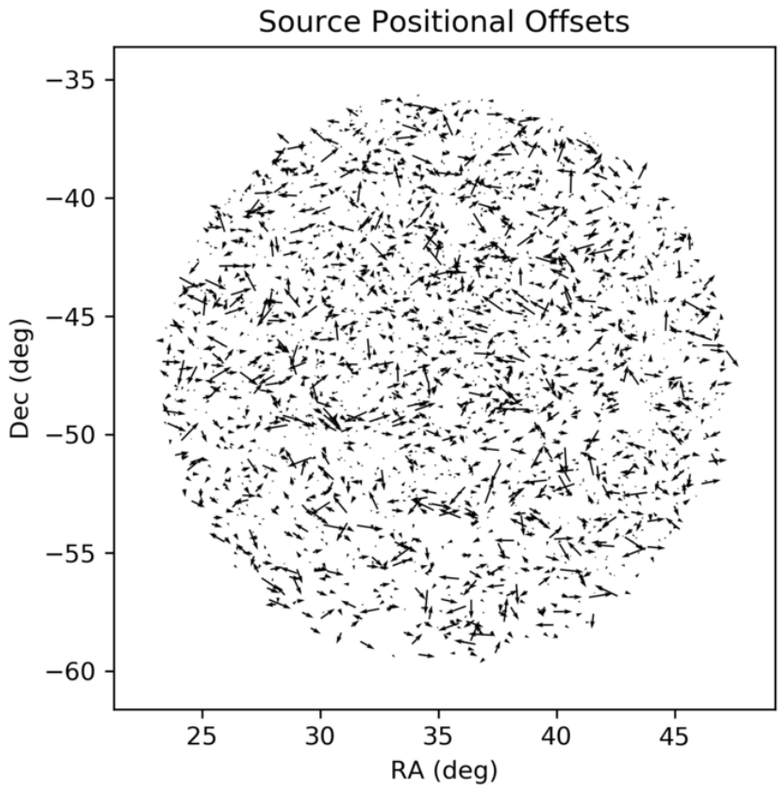
- Concentrate on one aspect of the data, with enough information to put that aspect into context
- Make a set of plots that all go together (complimentary views)
- Key metadata printed on the plot
- A meaningful initial layout

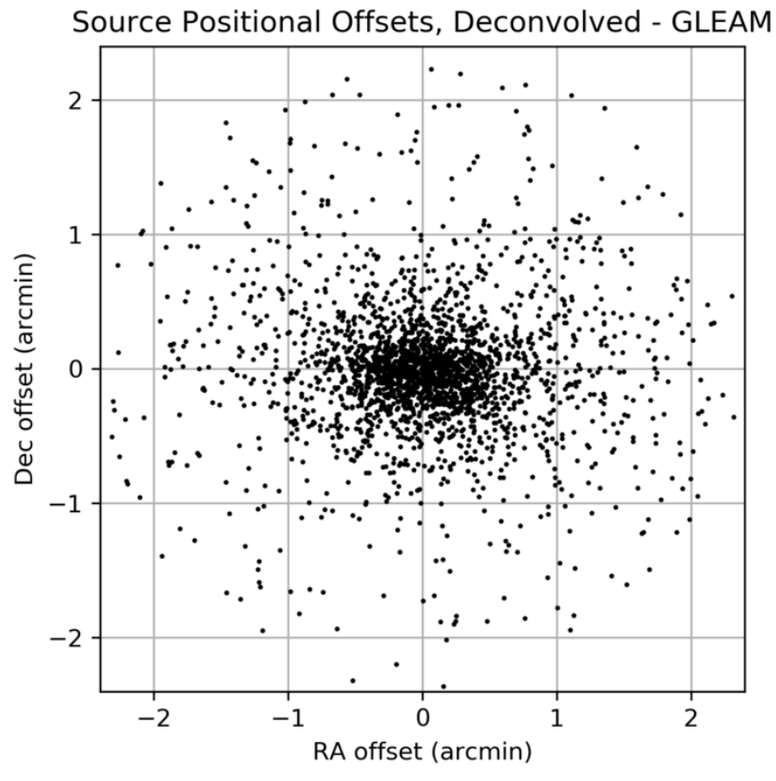
Plot sets

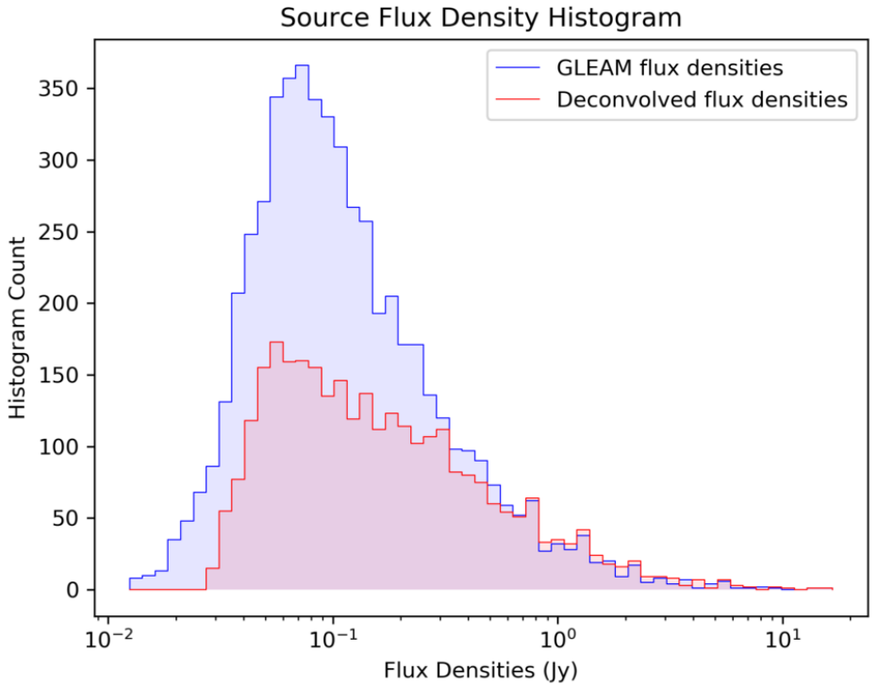

1130773024
0.50
0.39
0.28
0.18
0.07
−0.04
−0.14
−0.25
1130773024
0.31
0.25
0.18
0.11
0.04
−0.02
−0.09
−0.16

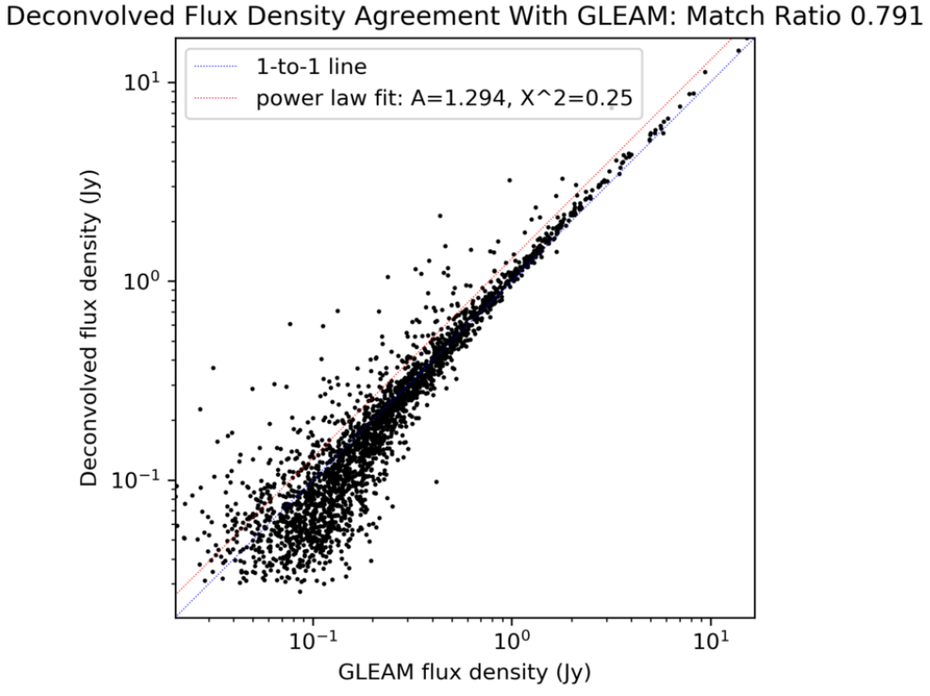

1130773024
0.32
0.25
0.19
0.12
0.05
−0.02
−0.09
−0.16
1130773024
0.22
0.17
0.12
0.08
0.03
−0.02
−0.06
−0.11

Plot Set pt 2


Source Positional Offsets
Dec (deg)
−35
−40
−45
−50
−55
−60
25
30
35
40
45
RA (deg)


Source Positional Offsets, Deconvolved - GLEAM
Dec offset (arcmin)
2
1
0
−1
−2
−2
−1
0
1
2
RA offset (arcmin)


Source Flux Density Histogram
GLEAM flux densities
Deconvolved flux densities
Histogram Count
350
300
250
200
150
100
50
0
10^{-2}
10^{-1}
10^{0}
10^{1}
Flux Densities (Jy)


Deconvolved Flux Density Agreement With GLEAM: Match Ratio 0.791
1-to-1 line
power law fit: A=1.294, X^2=0.25
Deconvolved flux density (Jy)
10^{1}
10^{0}
10^{-1}
10^{-1}
10^{0}
10^{1}
GLEAM flux density (Jy)

A meaningful (initial) grid

A structured approach

1. Look through all the data without thinking too hard, try and let it speak to you.

2. Start picking out distinctive sets of similar data. Label them.

3. Look for patterns. Let your grid help. Rearrange your grid as needed (using metadata)

4. As patterns arise, write them on the whiteboard

5. Distill into a set of tests and actions

A structured approach

1. Look through all the data without thinking too hard, try and let it speak to you.

2. Start picking out distinctive sets of similar data. Label them.

3. Look for patterns. Let your grid help. Rearrange your grid as needed (using metadata)

4. As patterns arise, write them on the whiteboard

5. Distill into a set of tests and actions


No RFI
RF
30
13
3
7
7
4 short baseline prob.
2 bad cal/high res.
Bad cal very high

A structured approach

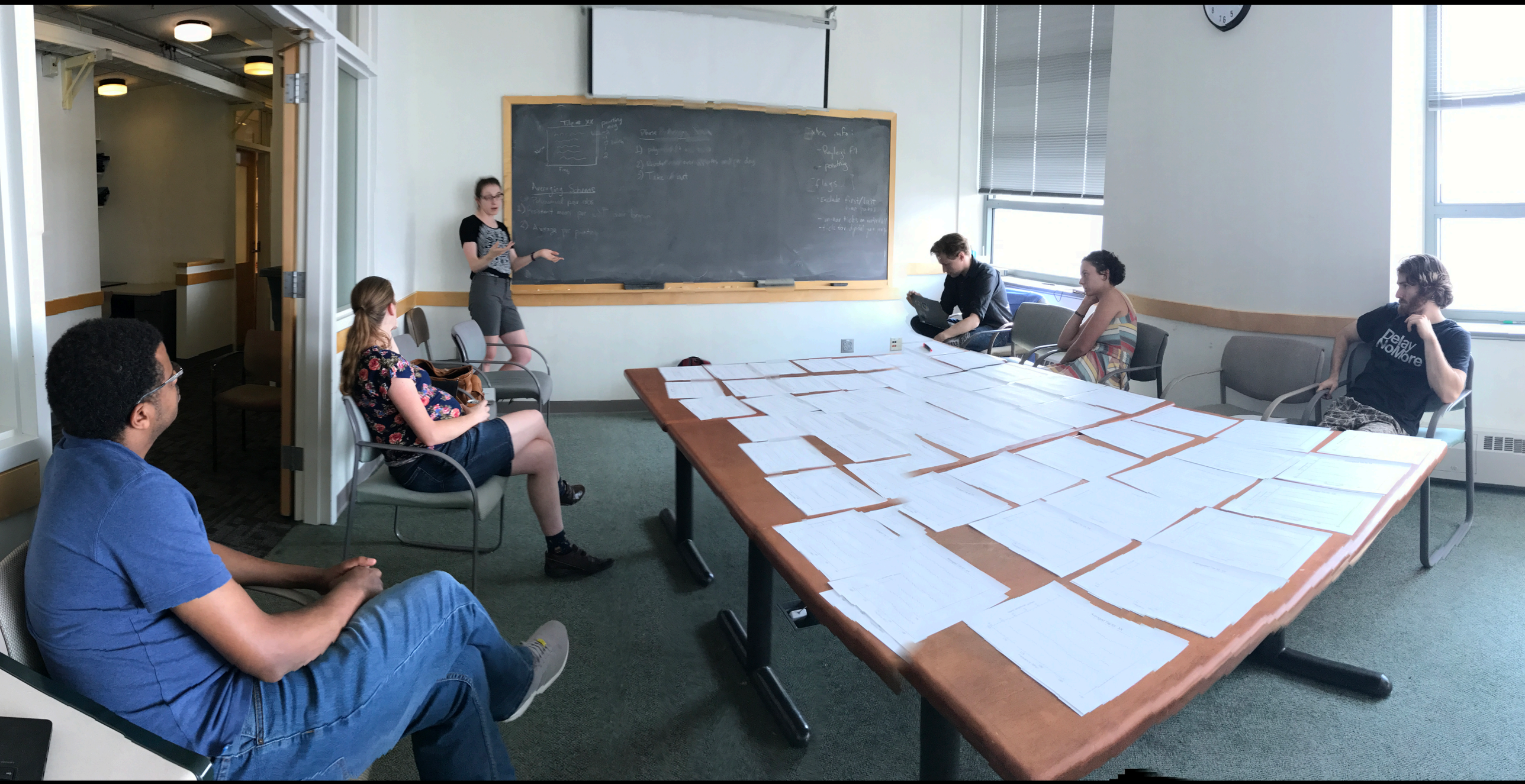
1. Look through all the data without thinking too hard, try and let it speak to you.

2. Start picking out distinctive sets of similar data. Label them.

3. Look for patterns. Let your grid help. Rearrange your grid as needed (using metadata)

4. As patterns arise, write them on the whiteboard

5. Distill into a set of tests and actions


Averaging Schemes
Extra info:
- Replicate #
- pointing
Flags:
- Exclude first/last
Delay NoMore

A structured approach

1. Look through all the data without thinking too hard, try and let it speak to you.

2. Start picking out distinctive sets of similar data. Label them.

3. Look for patterns. Let your grid help. Rearrange your grid as needed (using metadata)

4. As patterns arise, write them on the whiteboard

5. Distill into a set of tests and actions

Why would you kill so many trees?

- 100 million data points
- A way for a group to look for systematic patterns in a lot of data
- Very free form and natural
- Way to find the next questions
- Can often set the direction of the next 3-6 months of work