

ML for fundamental physics:

Linking Particle Physics & Astro

My background: LHC $\rightarrow$ Astro (Gaia, PTA, ...)

Motivations

- Astro data fascinating window into fundamental physics: dark matter, dark energy, early univ. cosmology, gravitational waves, ...

- Open data! Can motivate creative new idea, demo proof-of-concept, and apply to real data, all "in-house"

- Much simpler than LHC data in many respects

- if new physics can be produced @ LHC
- great! But no guarantee. Meanwhile astro for some probe of new physics (DM, DE, baryogenesis, ...) but less controlled environment than LHC

	particle physics	Astro
data	mostly closed iid events	mostly open stars, galaxies, <u>time series</u> - only one Universe, one MW galaxy, ...
	very high dim'l triggering	lower dim'l selection biases, also triggering

controlled int
state

uncontrolled

Simulations high quality
cheap

crude, expensive

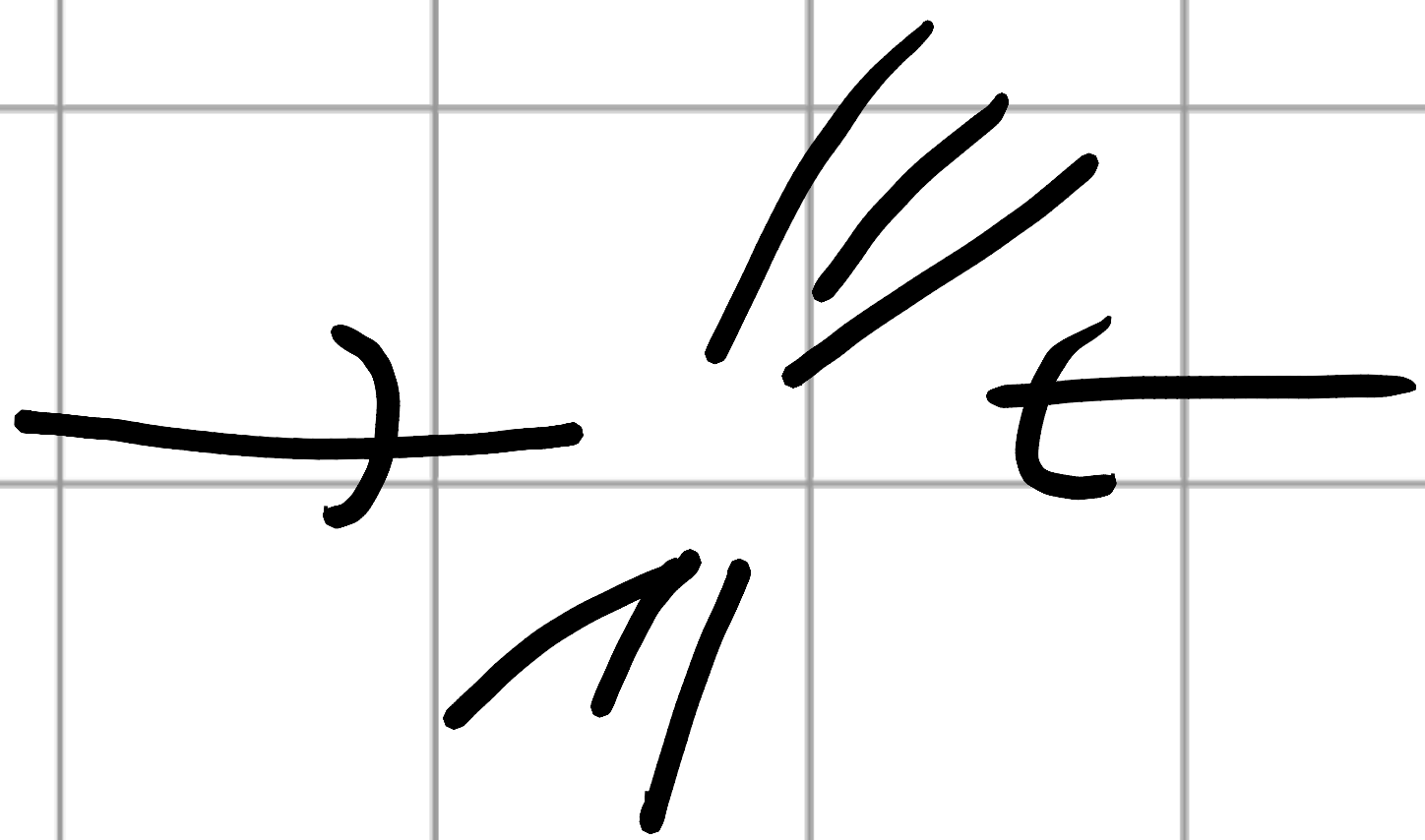
standards very high
(> 50 for
discovery)

much lower ("hype")
but no guarantee
anyone will believe
you

stats Frequentist
 $L(x_{data} | \theta)$
 $= \prod L(event | \theta)$

Bayesian
 $L(\theta | Data)$
since Data often
not iid

Particle Physics Data at a glance



collisions (pp @ LHC)

bunch x-y rate 40 MHz

→ collisions rate 6 Hz

→ trigger L1: $\sim 10^4$ kHz

L2: ~ 1 kHz

each "event"
recorded is an
"interesting" collision

Huge reduction!

$\sim 10^{10}$ events recorded
by LHC to date
expts

RAW: each event 10^8 channels (hits, calo pixels, ...)
→ $\mathcal{O}(\text{MB})$ → $\mathcal{O}(\text{PB})/\text{year!}$

compress/reduce (reconstruction of tracks, particles, jets, ...)

$\mathcal{O}(100)$ particles/event

$\lesssim 10$ features/ptcl
(p_x, p_y, p_z , charge, displacement, ...)

- No canonical ordering of ptcls — "point cloud"

$\dots$
 $\dots (p_x, p_y, p_z, \dots)$
 $\dots$

graph NNS

+ transformers

generally best arch.
for LHC events

Astro data

- surveys

- targeted instruments (e.g. HST, JWST)

- images

- timeseries

- catalogs — tabular data

each object (star, galaxy, ...)
represented by column of features

- pos on sky

- proper
motion

- distance

- radial velocity

- spectra

- ...

ground based

↓ vs space based

partial sky
coverage

"footprint"

↓
all sky
coverage

Examples: Gaia $\sim 10^9$ stars in MW

- star catalog
- space based
- optical
- pos sky
- proper mot.
- photometry (3 bands)
- RV & distance (subset)

DESI

- ongoing galaxy survey
- ground based
- $\sim 30\text{M}$ galaxies
- spectrometer
- pos sky
- redshift
- spectra

LSST/Rubin

ground based

$\sim 20\text{B}$ galaxies (over next 10 years)

- imaging
- also time domain
- camera
- optical

Nanograv
PTA

ground based, radio

$\sim 10^2$ pulsars

time series $\sim 2.5\text{wk}$ cadence
12.5 years
 ~ 50 time residuals each

Common ML tools btw PP & astro

- classifier/regression

time series
images

- RNNs, CNNs

← Astro

point clouds - GNNs, transformers, ... ← LHC

symmetries: rotations, translations

Lorentz, permutation

example applications

PP	Ast
jet tagging	galaxy/star classif
event classif	transient ident (eg supernovae, exoplanets, ...)
energy calibrati	galaxy/star property predict
particle ID	
mass measurment	
:	

- generative modeling (GANs, VAEs, flows, diffusion, ...)

example apps

	PP	Ast
surrogate modeling	fast calo sim pion showers phase space int. superres calo sim	cosmological sims baryon in painting mock star catalogs
SBI	LHC moments EFT fits ⋮	cosmology w/ galaxy surveys LIGO PTAs ⋮
Anomaly detection	resonant AD outlier detection w/ VAEs	stellar stream searches outlier detection

- Topics / Outline for rest of lectures

- SBI

- Resonant Anomaly Detection

- Surrogate Modeling

Simulation Based Inference

- Conventional Data Analysis:

$$D \xrightarrow{\text{data}} \vec{S} = (S_1, S_2, \dots) \text{ summary statistics}$$

$$L(S_1, S_2, \dots | \theta)$$

likelihood given parameters θ

$$\chi^2 = \vec{S}^T C(\theta)^{-1} \vec{S} \quad (\chi^2)$$

$$\text{or } \prod_{i=1}^{N_{\text{bins}}} \text{Pois}(S_i | \mu_i(\theta)) \quad \text{histogram}$$

- Then max L wrt $\theta \rightarrow$ get best fit value & uncertainty on θ .
- Idea of SBI: learn full likelihood $L(D|\theta)$ using ML - more optimal analysis!
- "simulation based"
need samples from forward model $\theta \rightarrow D$ to train ML on

Given samples from: $\theta \sim p(\theta)$ "prior"
 $D \sim p(D|\theta)$ "forward model / simulator"
 $(\theta, D) \sim p(\theta, D) \equiv p(\theta)p(D|\theta)$

Can attempt to learn

- $p(D|\theta)$ "Neural Likelihood Estimator"
(NFs, diffusion/CFM, ...)

↓
prior independent!

can be difficult if D is very high dim'l

- $p(\theta|D)$ "Neural Posterior Estimator"

prior dependent

high dim'l D can be mitigated w/
summary network

$$p(\theta|D) \leftarrow p(\theta|F(D))$$

- Also: can learn w/ classifier trick
alternatively

From samples (θ, D)

make random pairs $\rightarrow p(\theta)p(D)$

Binary classifier learns

$$\frac{p(\theta, D)}{p(\theta)p(D)} = \frac{p(\theta|D)}{p(\theta)}$$

$$\frac{p(D|\theta)}{p(D)}$$

plug in actual data $\hat{p}$ just an unnormalized

$\rightarrow$ get either posterior or likelihood!

"Neural density ratio estimation"

SBI only as good as the simulator!

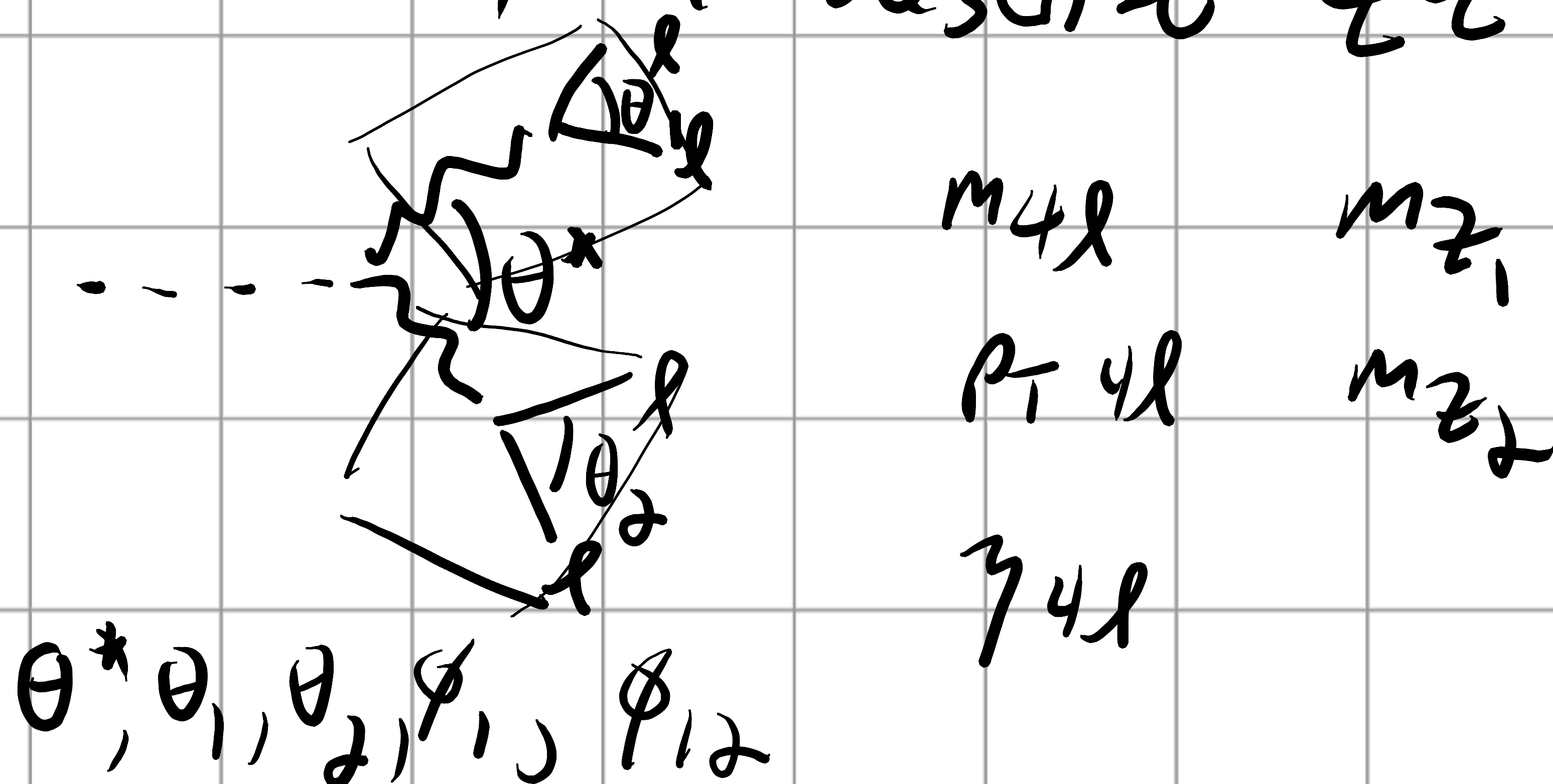
if sim-data diff large, can lead to biased results

Examples: ATLAS measurement of
offshell Higgs prod w/ NSBI
2412.01548

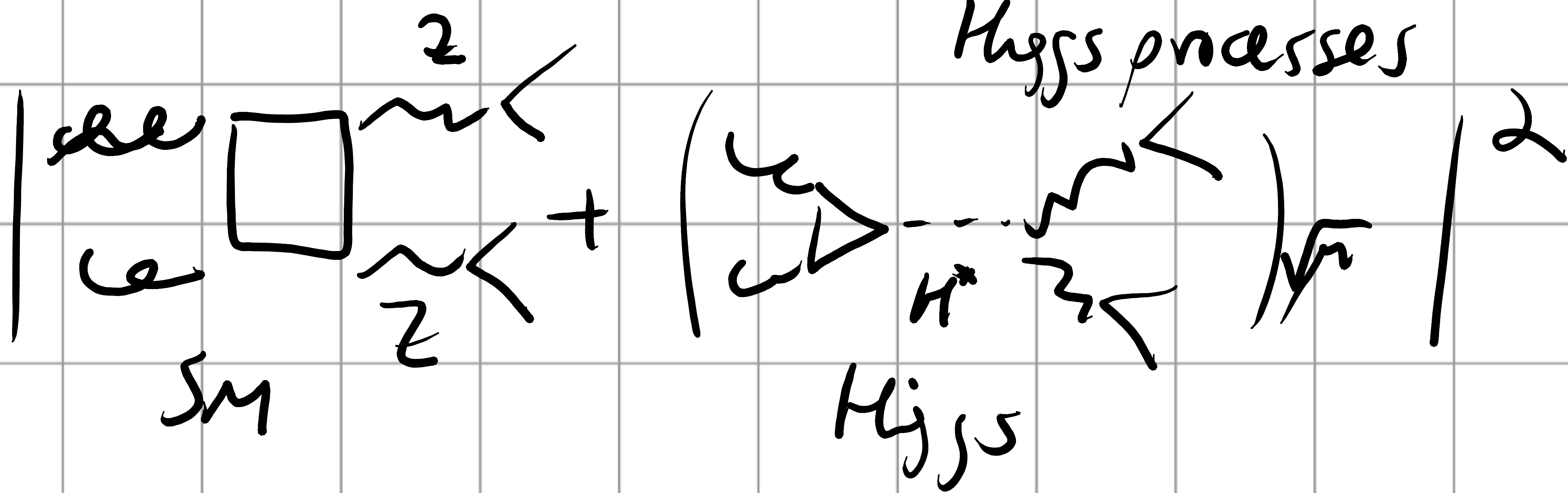
improved expected precision of measurement
by factor of 2.6 compared to previous
analyses on same data that used traditional
techniques

$\left\{ \begin{array}{l} gg \rightarrow H^* \rightarrow ZZ \rightarrow 4l \\ \text{main SM b.s. } gg \rightarrow ZZ \rightarrow 4l \end{array} \right.$
has quantum interference!
→ well-described by simulations
→ good target for NSBI

Features: 10 kinematic variables
that describe $^{(H^\pm)} ZZ \rightarrow 4l @ LHC$



Main parameter: $\mu \equiv$ signal strength
Higgs processes



Clever trick: don't even need to sample
from $(\mu, \mathcal{D})$!

$$p(\mathcal{D}|\mu) = \mu p_S(\mathcal{D}) + \sqrt{\mu} f_{\pm}(\mathcal{D}) + p_B(\mathcal{D})$$

densities

not a density

But! $f_{\pm}(D) = p_{\mu=1}(D) - p_S(D) - p_B(D)$

5. ! $p(D|\mu) = (\mu - \sqrt{\mu}) p_S^{(D)} + \sqrt{\mu} p_{\mu=1}(D) + (1 - \sqrt{\mu}) p_B(D)$

Can generate ^{3 sets of} samples from sig only p_S
 $\downarrow$ b_S only p_B
 $\mu=1$ $p_{\mu=1}$

perf. NRE for each
 set general $p(D|\mu)$ for any μ .

Lots of validations, closure tests, control regions
 $\downarrow$ systematics $\downarrow$ VBF proc.

full analysis includes nuisances α
 $p(D|\mu, \alpha) \rightarrow$ need to be profiled
 over in final analysis

final result: $\mu = 0.87^{+0.75}_{-0.54} (1.00^{+1.04}_{-0.95})$

↓
factor of 2.6
better than
prev. analysis!

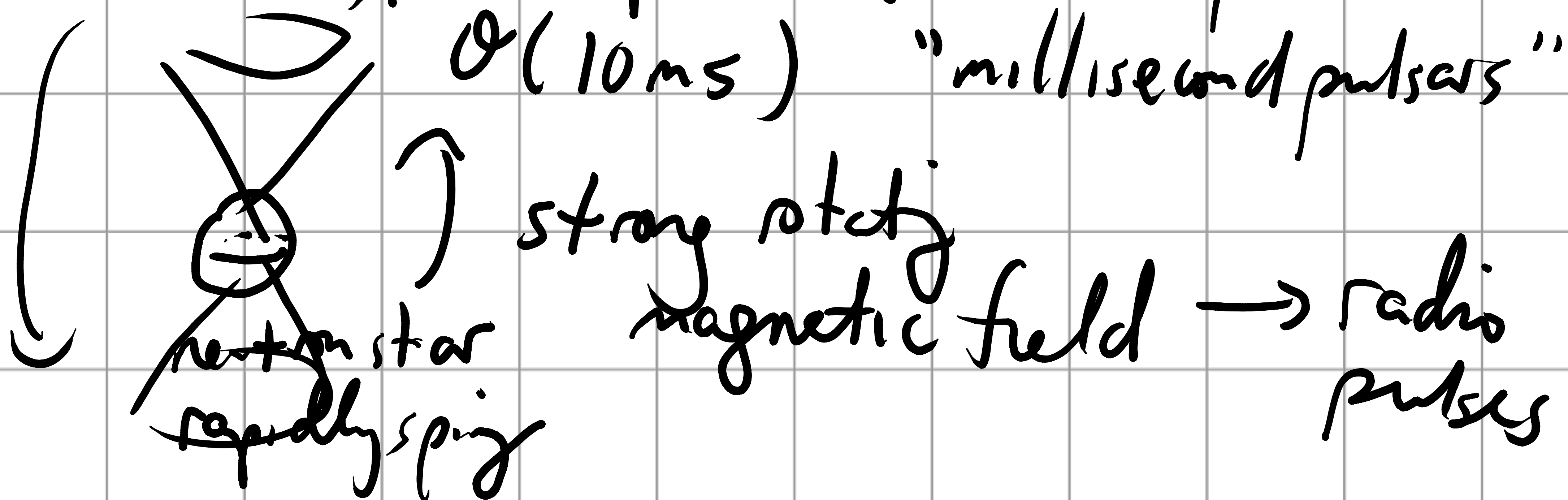
2nd example: OS, Freytsis, Taylor, Oror, Smyth
PRL 133 (2024) 1, 2310.12209

First application of SBI to PTAs

Nanograw PTA: ~50 pulsars observed

extremely regular for ~15 years every ~2.5 weeks
See radio pulses from each pulsar

$\mathcal{O}(10\text{ms})$ "millisecond pulsars"



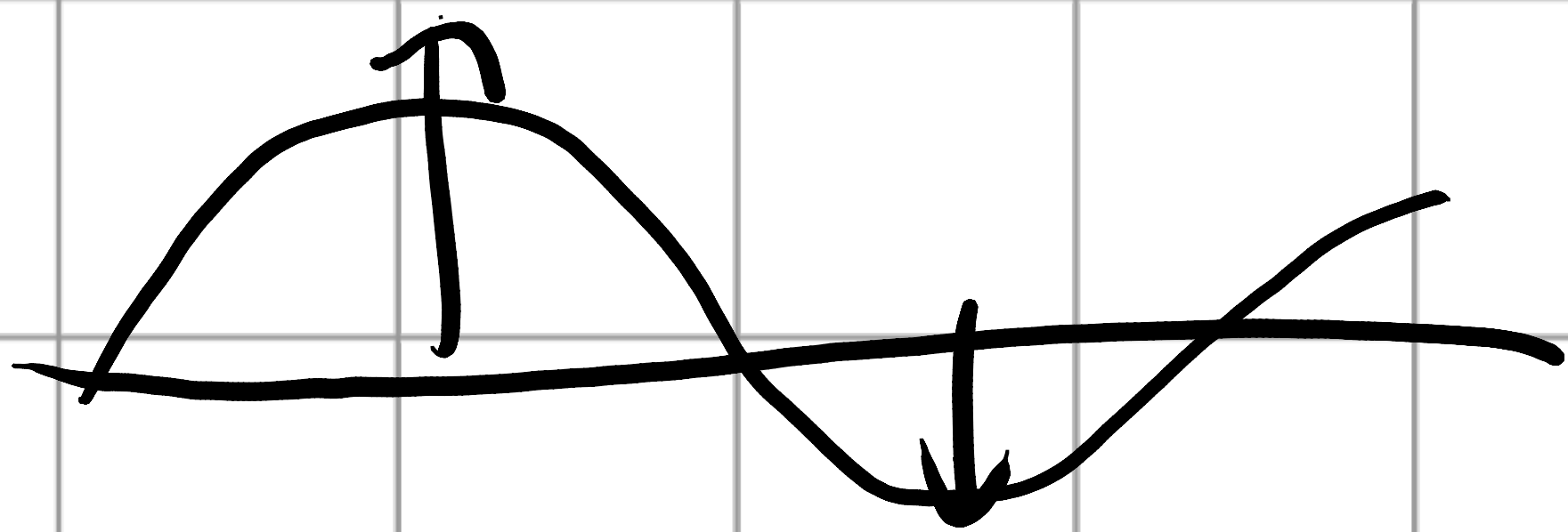
- ToAs very predictable "astronomical clocks"
for best pulsars can get ~ 10 ns precision
on ToAs (10^{-6})

$$\delta t = t_{\text{ToA}} - t_{\text{pred}} \quad \text{"timing residuals"}$$

GW

use array of well char.
ms pulsars
can be sensitive to
gravitational waves!

GW will distort δt by $\mathcal{O}(10 \text{ ns})$ across
a period



max period set by
length of obs
 $\mathcal{O}(10 \text{ yr}) \rightarrow \text{nano Hz}$
(Hence, Nanograv)

So PTAs are pot'ly sensitive
to GWs in nHz regime!
(by contrast: LIGO Hz $\rightarrow$ kHz)

Sources of GWs in nHz range:

- Stochastic GWB - incoherent superpos
of GWs from supermassive
BH binary systems
 10^4 rows
- individual SMBH binaries

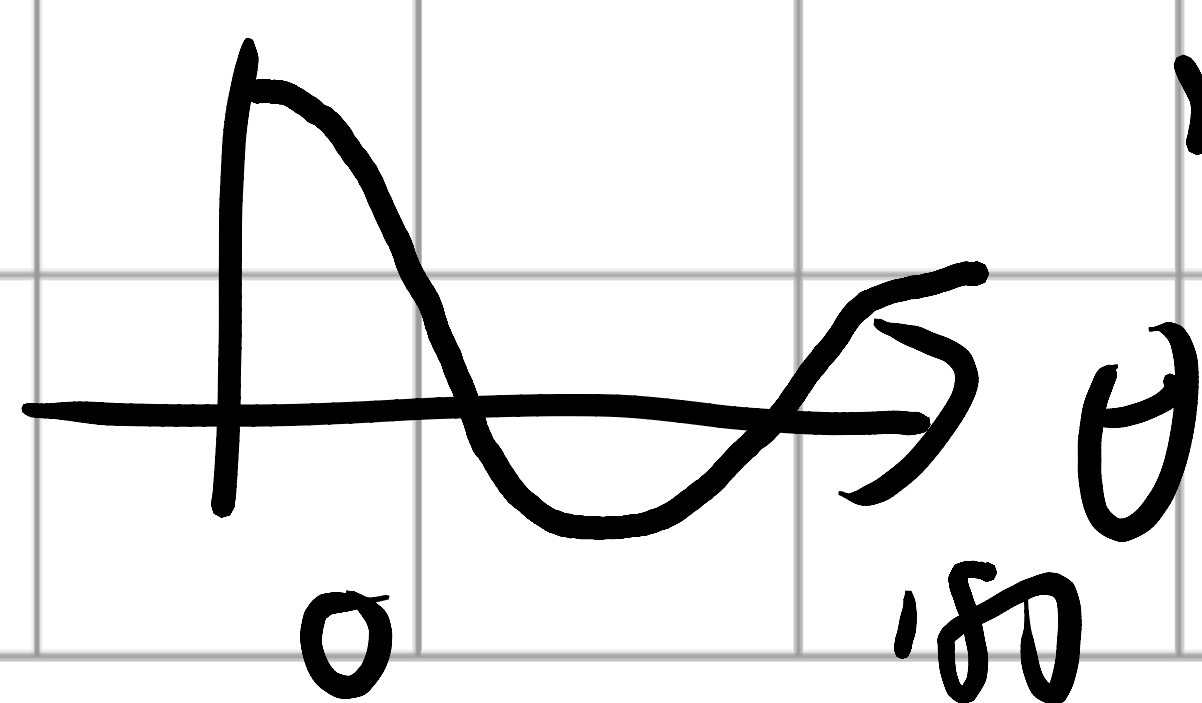
described by power law spectrum in freq

I: pulsar
index

$$\delta_I(t) = \int_{-\infty}^{\infty} df a_I(f) \cos 2\pi f t$$

$$\langle a_I(f) a_J(f') \rangle = A_{GW}^2 f^{-\gamma_{GW}} \chi(\cos \theta_{IJ})$$

$$\left\{ \begin{array}{l} \gamma_{GW} = 13/3 \\ \chi(\cos \theta_{IJ}) \end{array} \right\} \text{predicted} \quad \cdot \delta(f-f')$$



"Hellings-Downs Curve"

Also: noise sources

$$\langle a_I(f) a_J(f') \rangle = \delta_{IJ} A_{rI}^2 f^{-\delta_I} \delta(f-f')$$

"red noise"

$$\langle a_I(f) a_J(f') \rangle = \delta_{IJ} A_w^2 \delta(f-f')$$

"white noise"

uncorr. Gauss noise
added to each residual

Add. +
together
FT

$$\langle \delta_I(t_i) \delta_J(t_j) \rangle = C_{IJ}(t_i - t_j; A_{\text{GWB}}, A_r, A_w)$$

"Gaussian process"

Think of as big mtx w/ $N_{\text{pulsar}} \times N_{\text{res}} \sim 10^4$ rows

Likelihood is simple $(\delta_{Ii})^T (C^{-1})_{Ii, Ji} (\delta_{Jj})$

- But slow to eval - need to invert $10^4 \times 10^4$ mtx!
- To estimate posterior $p(A_{\text{GWB}}, A_r, A_w | D)$
need to run MCMC over $\sim 10^6$ samples
— can take ~ 1 wk for single plot!

"But likelihood fast to sample! diagonal in freq
FT is fast

- So NPE is feasible, takes ≤ 1 day
we trained NF (MAF) w/ RQS transt.
to learn $p(\theta | \vec{T})$

A_{GUB}, A_r, A_w

↓ too high dim'l
first send to ~~summary~~
network $F(\vec{T})$

used LSTM
but transt would have
been better

- Can also sample from NF $\rightarrow$ get posterior
from samplg in seconds!
- coverage, bias not guaranteed - have to
check after training

II. Surrogate Models

The basic idea: given slow but accurate sim

$$X_i \sim p_{\text{sim}}(x)$$

suppose p_{sim} intractable but can
(eventually) sample suff. from it
to train a gen model

$$X_i \sim p_{\text{gen}}(x) \approx p_{\text{sim}}(x)$$

now fast! hopefully p_{gen} good approx to p_{sim}
and interpolates through training
data well.

p_{gen} is a "surrogate" for p_{sim}

Metrics:

- Classifier Test
- Histograms

- ...

- Multi model Classifier Test

"log posterior score" (OS+Linet 2011.1765)

• Suppose have multiple models, maybe none are great ($AUC=1$ on clf test) or maybe binary clf training unreliable for some models

• Instead: train multi-class clf btw samples from each model

get $p(i|x)$ $i=1, \dots, N_{\text{model}}$
 x data

$$LP(i) = \frac{1}{N_{\text{ref}}} \sum_{x \in \text{ref}} \log p(i|x)$$

model w/ highest LP
is closest to ref
data!

To slides of surrogate model analysis

III. Anomaly Detection

Anomaly Detection key task in
PP & Astro (also: Data quality monitoring)
↓ new physics ↓ new astrophysical objects / phenomena

Two types of anomalies

data
"outliers"

data
"overdensities"

if data is like this, don't need ML!

But if data is
- higher dim
- anomaly features not obvious

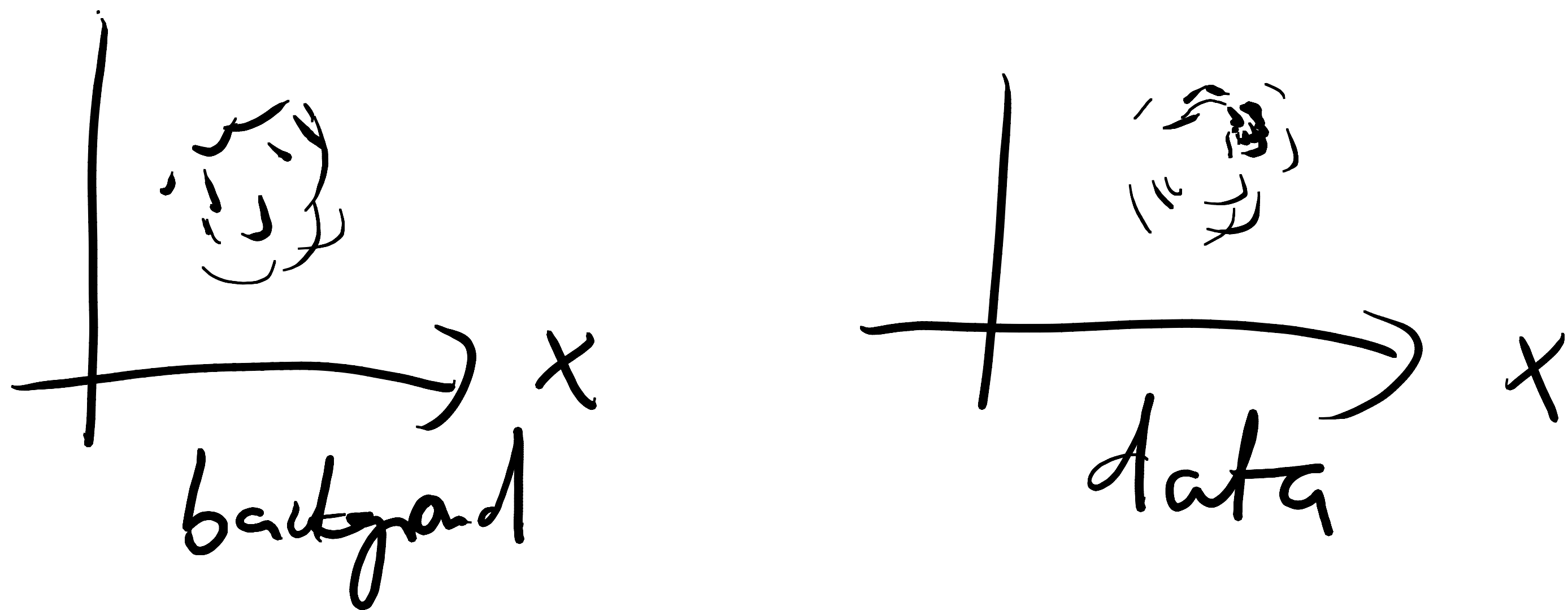
Then ML can be powerful tool.

outliers $\rightarrow$ rare events on tails
(esp @ LMC unlikely to have
truly $p_{\mathcal{D}_j^{(x)}} = 0$)

$\downarrow$
autoencoders, density estimation
 $\rightarrow$ see Thea's talk

Here focus on overdensities

5/1/25 Overdensity Detection



suppose $\text{data} \sim \text{bg} + \text{little bit of signal}$
can we select signal preferentially from data?

$R(x)$ anomaly score

How about

$$R(x) = \frac{p_{\text{data}}(x)}{p_{\text{bg}}(x)}$$

$R(x) > R_c > 1 \rightarrow$ selects locations in
data w/ overdensities.

In fact, $R(x)$ is optimal!

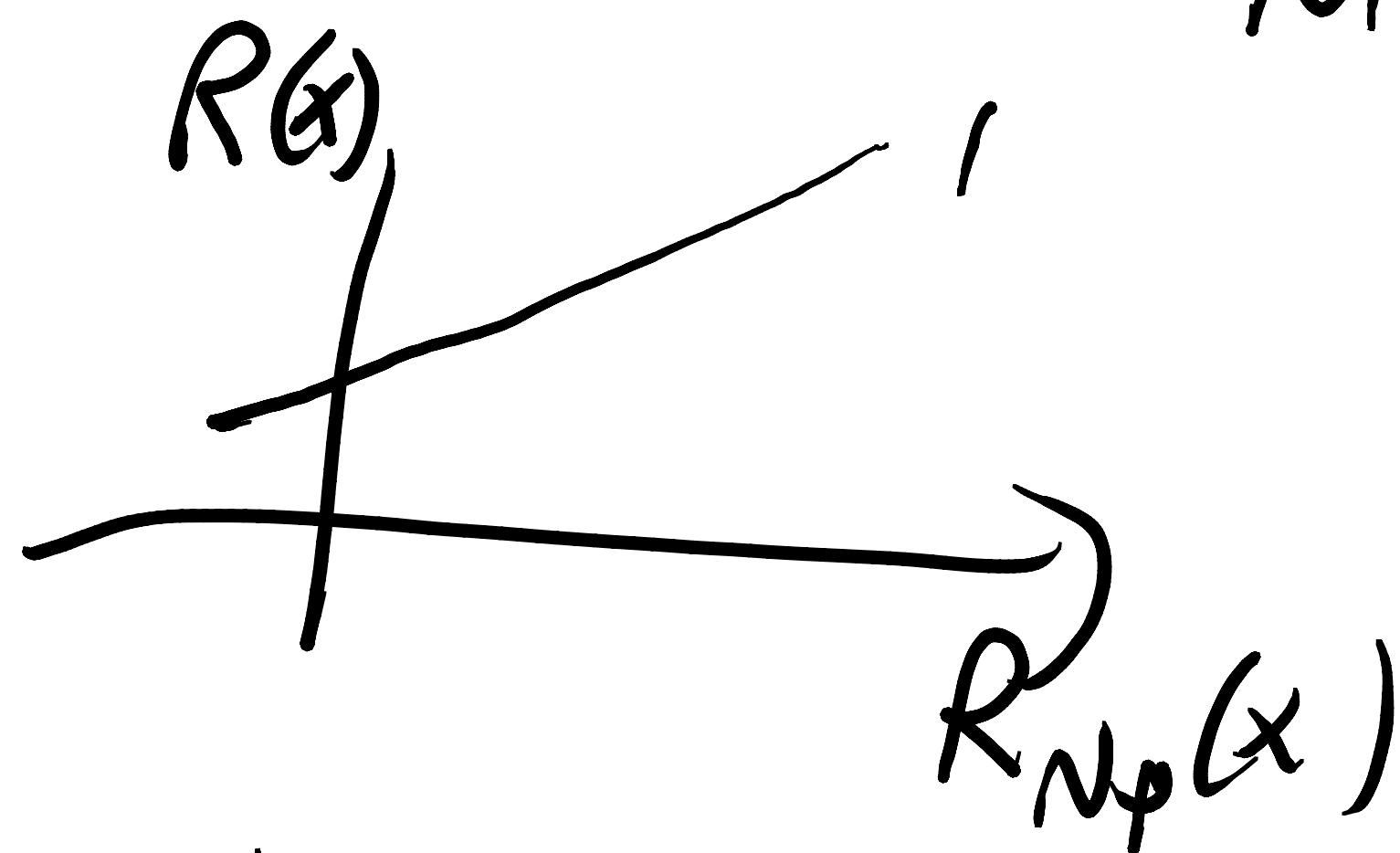
$$p_{\text{data}}(x) = (1 - \epsilon) p_{\text{bg}}(x) + \epsilon p_{\text{sig}}(x)$$

So

$$R(x) = \frac{p_{\text{data}}(x)}{p_{b_s}(x)} = 1 - \varepsilon + \varepsilon \frac{p_{\text{sig}}(x)}{p_{b_s}(x)}$$

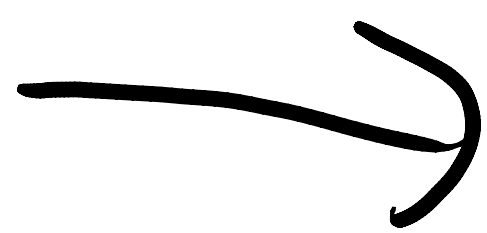
$R(x)$ is monotonically w/ $\frac{p_{\text{sig}}}{p_{b_s}}$ LR!

$$R(x) = 1 - \varepsilon + \varepsilon R_{NP}(x)$$

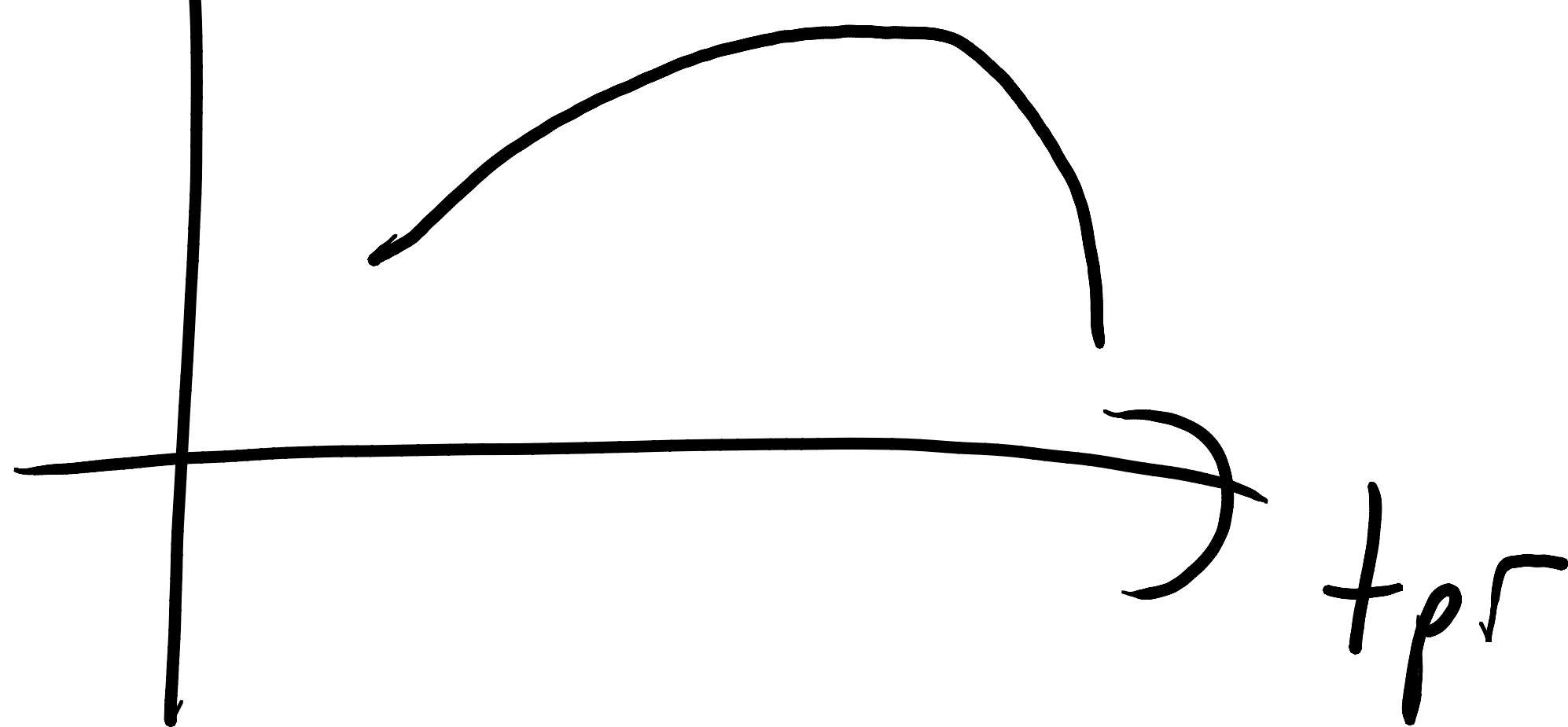


So $R(x) > R_c \Leftrightarrow R_{NP}(x) > R'_c$
 $\Rightarrow R(x)$ is optimal sig/bg classifier
for any unknown signal!!
 $\rightarrow$ very powerful anomaly detection
method!

After cutting on $R(x)$



$$SIC = \frac{\epsilon_5}{\sqrt{\epsilon_B}}$$



$SIC > 1$: enhancing
significance
of anomaly

How to learn $R(x)$?

1. Direct Density Estimation

NF or something $\rightarrow p_{data}(x)$
 $\searrow p_{sig}(x)$

2. Classifier-based methods

binary clf data vs by "likelihood ratio
trick"

approx. $R_{\text{data}} = \frac{p_{\text{data}}(x)}{p_{\text{bg}}(x)}$ "weak supervision"

data = $\begin{cases} \text{correlated} \\ \text{w/ synd vs } b_j \\ \text{but not perfect, noisy} \end{cases}$

Supervise w/ noisy labels

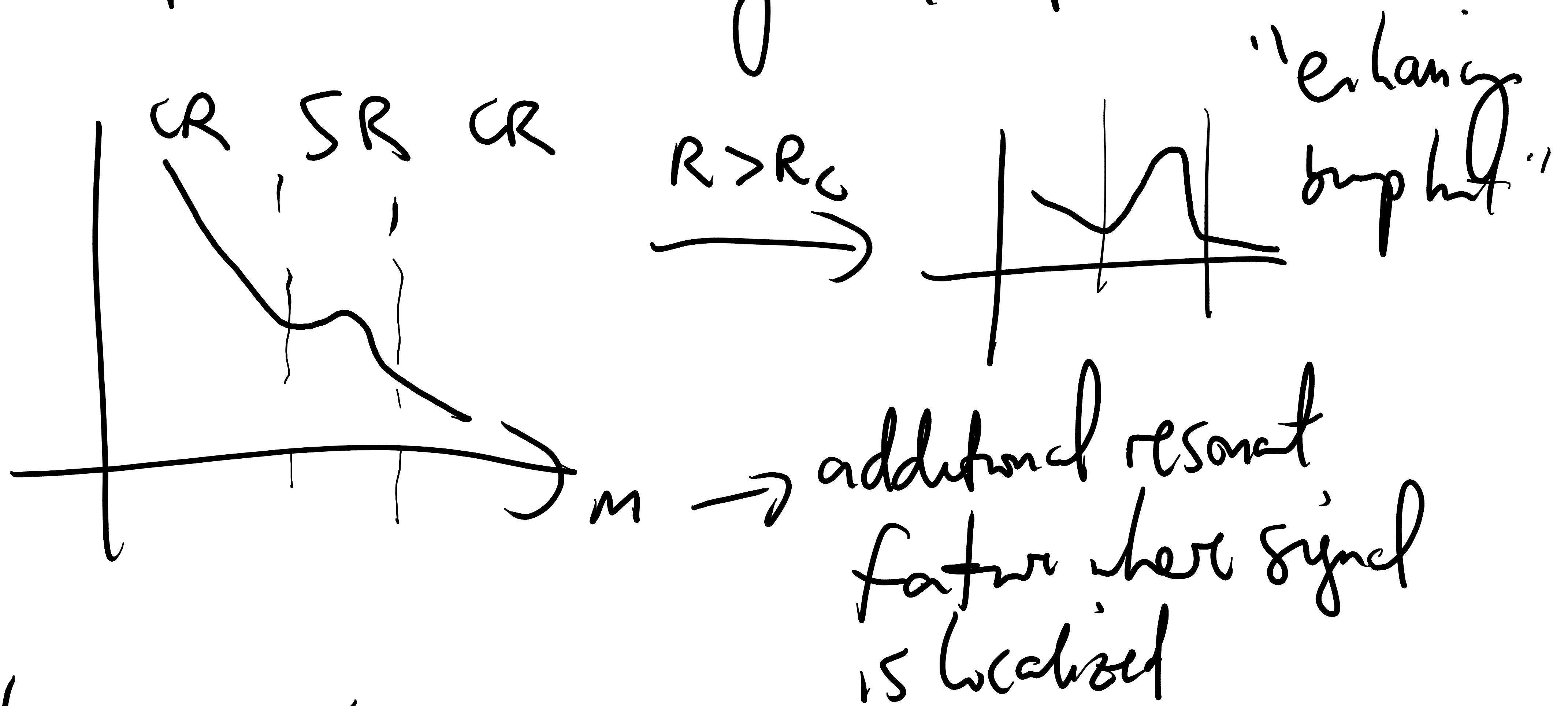
Show 2nd gaussian example

Now: how to obtain background sample??

1. Simulations? generally not good enough
Sim vs data $\rightarrow$ just learn all places
Sim is wrong, no
sensitivity to anomalies

2. data-driven
— by control regions

One well studied example:
resonant anomaly detection



- learn $p_{\mathcal{Y}}(x)$ ($p_{\mathcal{Y}}(x|m)$)
for $m \in CR$ (w/NF) eg $H \rightarrow \delta\delta$
 $n = n_H = n_{\text{in}}(\delta\delta)$
- interpolate into SR
to get $p_{\mathcal{Y}}(x)$ in SR
- learn $p_{data}(x)$ $\rightarrow$ direct density method
for $m \in SR$ (w/NF) "Anomaly Detected through Density Estimation"

Nachter & DS 2001.04990

- classify events in CR vs events in SR
 $p_{bs}(x|m \in CR)$ $p_{data}(x|m \in SR)$

get $R(x)$ if $\underline{\underline{p_{bs}(x|m) = p_{bs}(x)}}$ (x & m
 indep
 for bs)

"C.W. La Plante" Collins, Howe
 Nachman
 1805.02664, 1902.02634

- if x & m not indep ...

learn $p_{bs}(x|m \in CR)$ w/ NF

↓
 $p_{bs}(x|m \in SR)$

- Sample from $p_{bs}(x|m \in SR)$ (since NF
 it's possible!)

- classify data vs by samples

Classifying Anomalies Through Outlier
 Density Estimate (CAWODE)

OS + Hall et al 2109.00546

