LEARNING NEW PHYSICS FROM DATA: A SYMMETRIZED APPROACH

2401.09530

UC Davis HEP Seminar, Apr 2024

INBAR SAVORAY

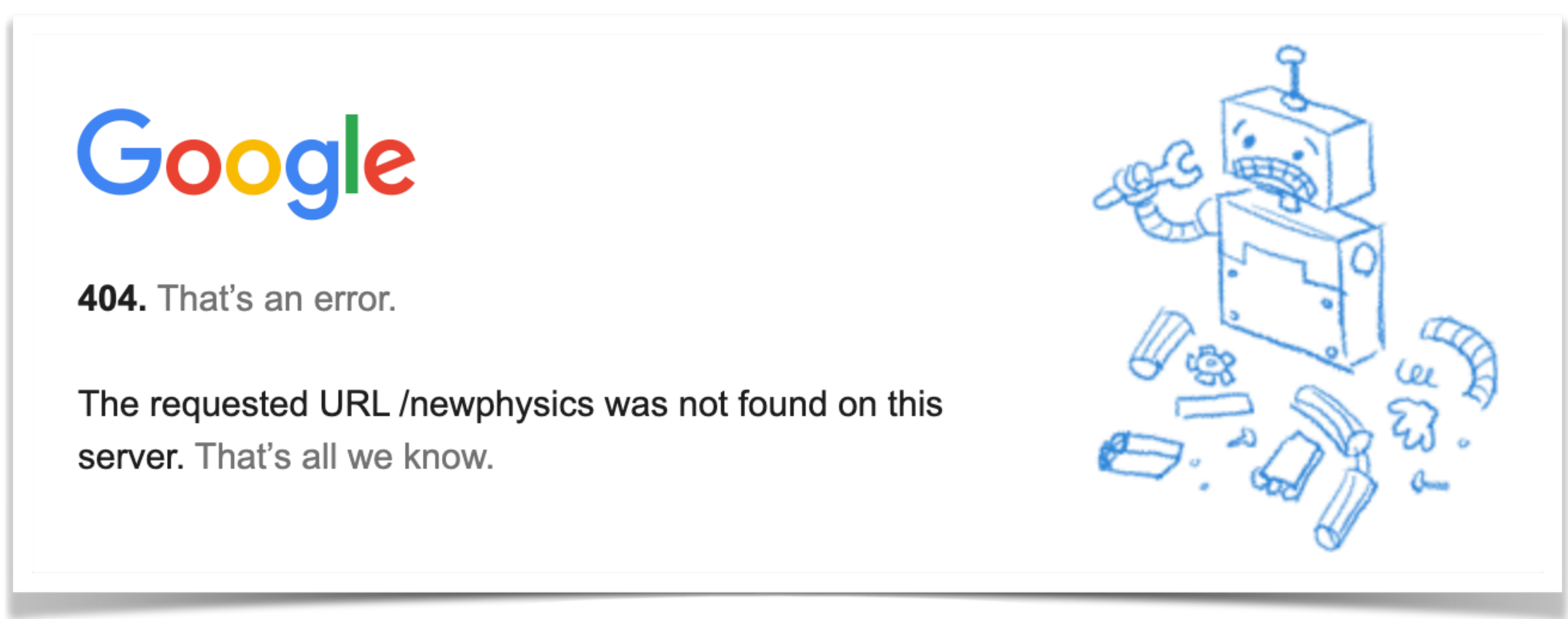
UC Berkeley & Lawrence Berkeley National Lab

IN COLLABORATION WITH: SHIKMA BRESSLER AND YUVAL ZURGIL

Weizmann Institute of Science

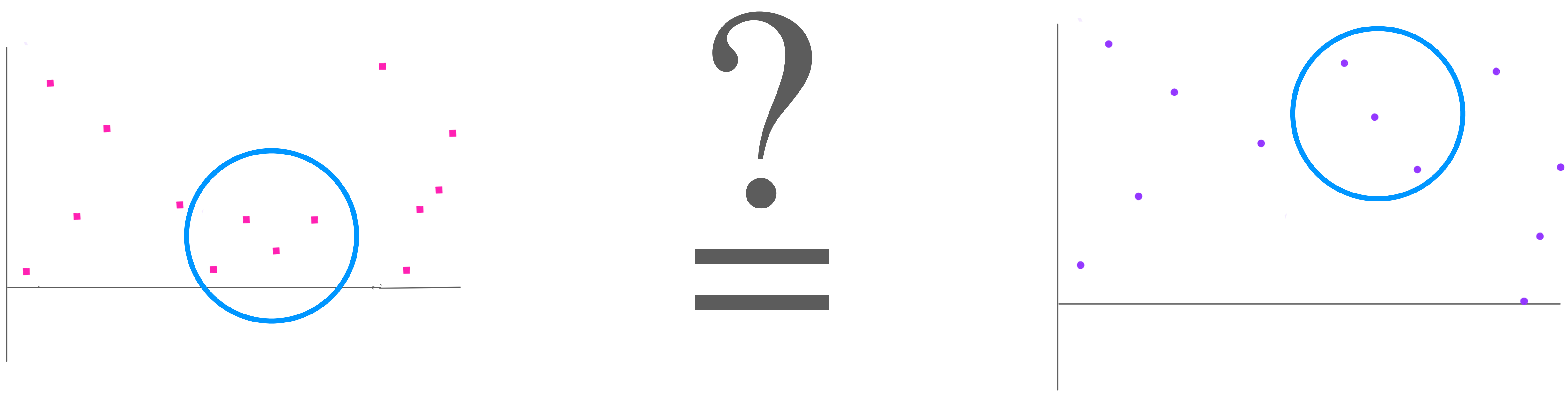
MOTIVATION

- Despite great theoretical and experimental effort, no evidence of New Physics has been found to date.
- Many dedicated searches ruled out a significant portion of the parameter space of theoretically motivated models.
- However, there is still much more to explore:
  - New theoretical models.
  - A lot of data.



MODEL AGNOSTIC SEARCHES

- **Data-directed paradigm (DDP)** for model agnostic searches:
  - Search for deviations from SM properties (what we do know).
  - Scan the data efficiently.
  - Identify anomalous regions for detailed study.

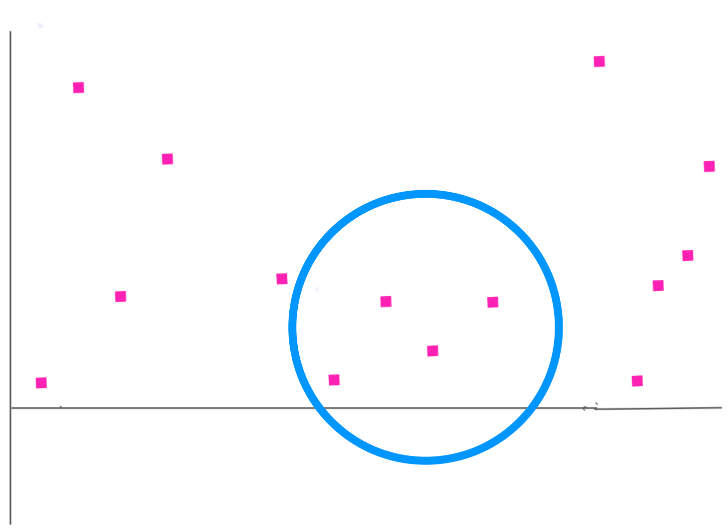


SM *Data*

S. Bressler, A. Dery and A. Efrati, [1405.4545] *M. Birman, B. Nachman, R. Sebbah, G. Sela, O. Turetz, and S. Bressler, [2203.07529]*

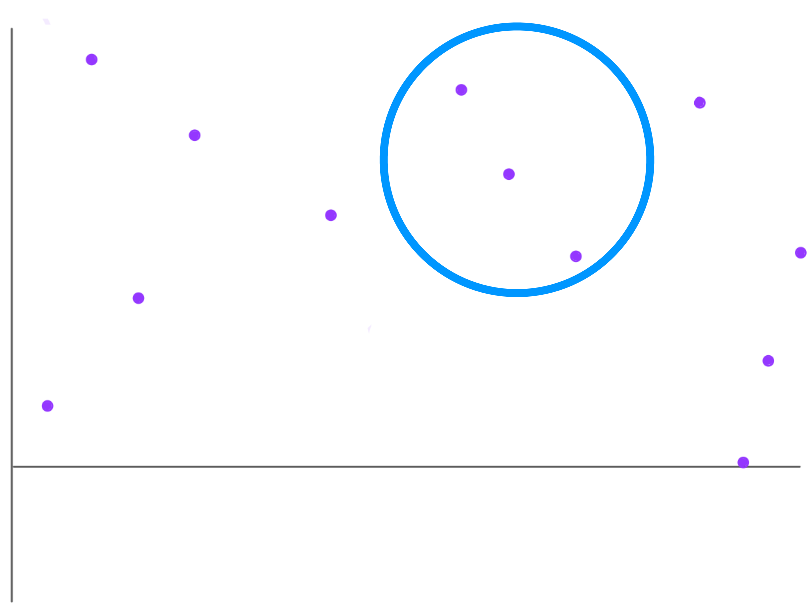
MODEL AGNOSTIC SEARCHES

- **SM symmetries** imply relations between different regions of the data, that if violated could point to NP.
- Example - **lepton flavor universality:** $e/\mu/\tau$ **should be interchangeable** (up to H+phase space).
  - (Hints: neutrino masses + B-anomalies)



e data

$\stackrel{?}{=}$

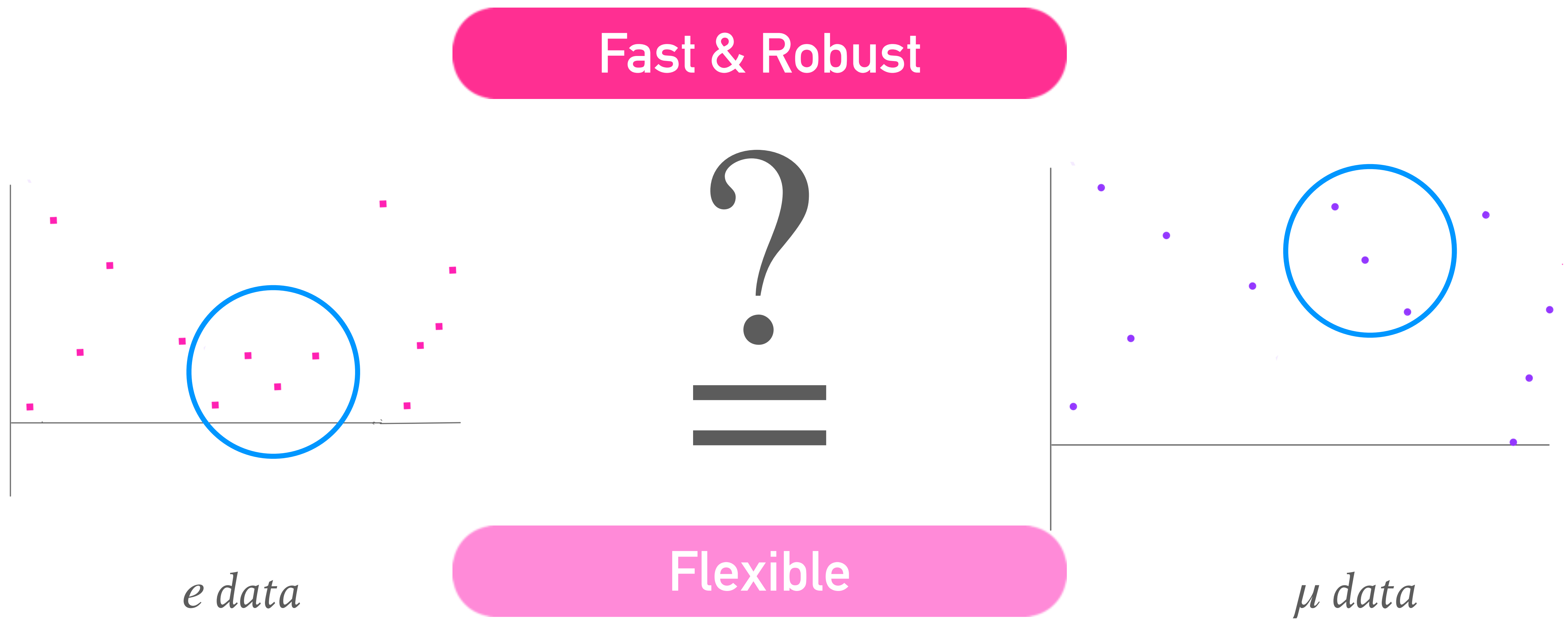


μ data

S. Bressler, A. Dery and A. Efrati, [1405.4545]   *M. Birman, B. Nachman, R. Sebbah, G. Sela, O. Turetz, and S. Bressler, [2203.07529]*

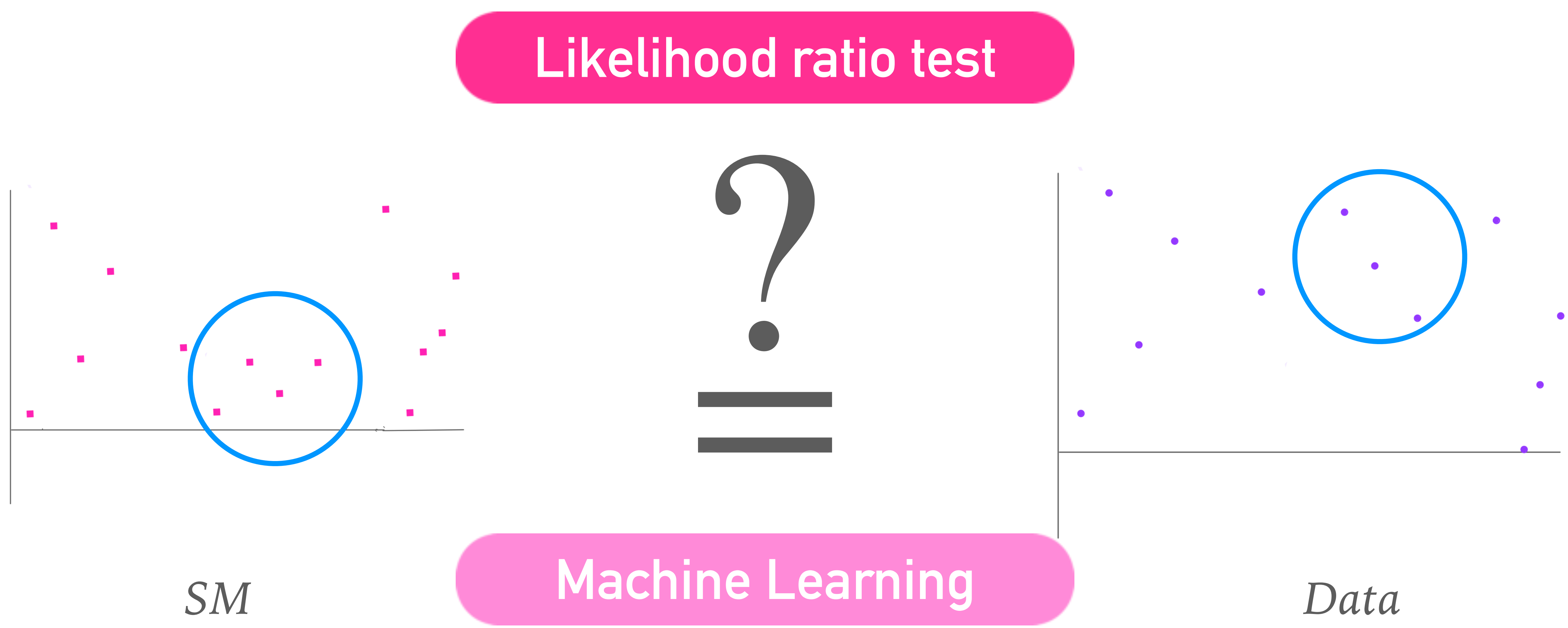
GOAL

- **Efficiently scan data for asymmetries** between samples that should only differ by statistical fluctuations.
- **Model-independent interpretation**: minimal assumptions, no detailed simulations (SM&NP).



METHOD

- Previous proposal - **"Learning NP from a Machine"(NPLM)**
  R. T. D'Agnolo & A. Wulzer, [1806.02350].
  - Testing whether an observed dataset is distributed according to a much larger reference SM sample.



METHOD

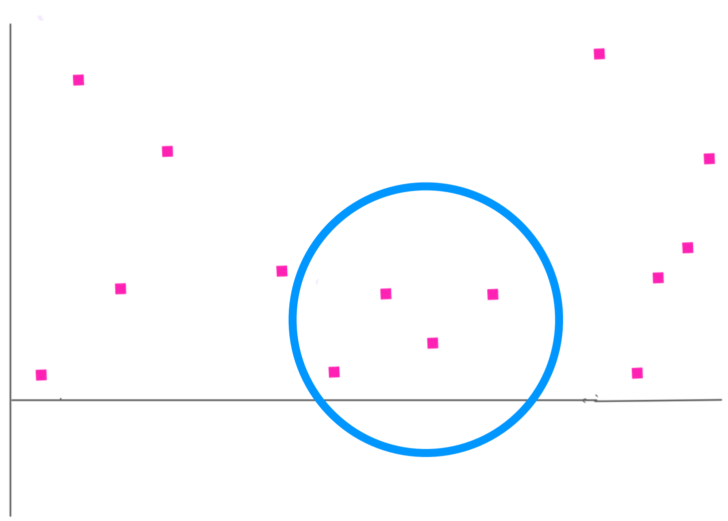
- Previous proposal - **"Learning NP from a Machine" (NPLM)**

  R. T. D'Agnolo & A. Wulzer, [1806.02350].

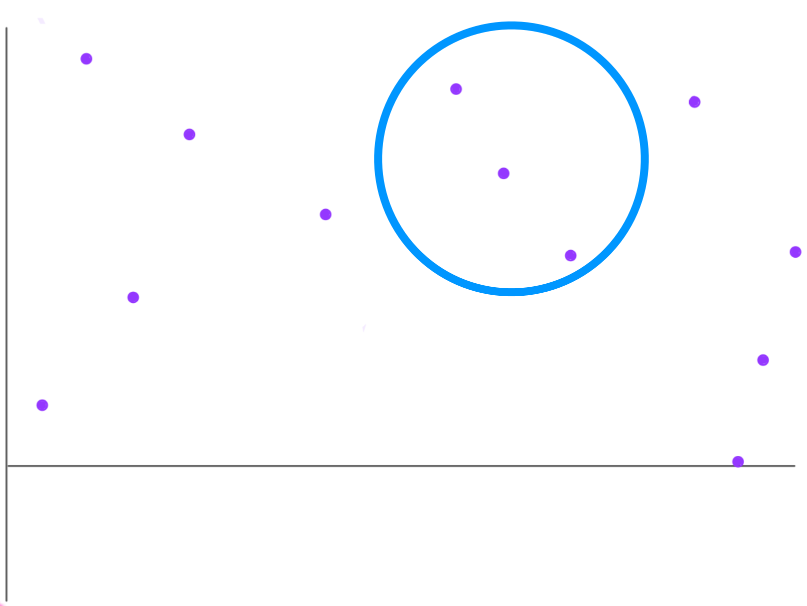
  - Testing whether an observed dataset is distributed according to a much larger reference SM sample.

Can it be implemented for small asymmetry searches?

Likelihood ratio test



$\stackrel{?}{=}$



e data

Machine Learning

μ data

LIKELIHOOD 101

- Likelihood - probability of obtaining result x had θ been true:

$$\mathscr{L}(\theta \mid x) = p(x \mid \theta)$$

  - Model fitting
  - Hypothesis testing

LIKELIHOOD 101 - MODEL FITTING

- Likelihood - probability of obtaining result x had θ been true:

$$\mathscr{L}(\theta \mid x) = p(x \mid \theta)$$

- The model in which the probability of obtaining the observed is the highest is the most likely (MLE)

$$\text{MLE:} \quad \hat{\theta} = \text{argmax}\left(\mathscr{L}\left(\theta \mid x_{\text{obs}}\right)\right)$$

  - Likelihood always maximal if prediction=observed.
  - If something occurred, it cannot have zero probability.

LIKELIHOOD 101 - MODEL FITTING

- Likelihood - probability of obtaining result x had θ been true:

$$\mathscr{L}(\theta|x) = p(x|\theta)$$

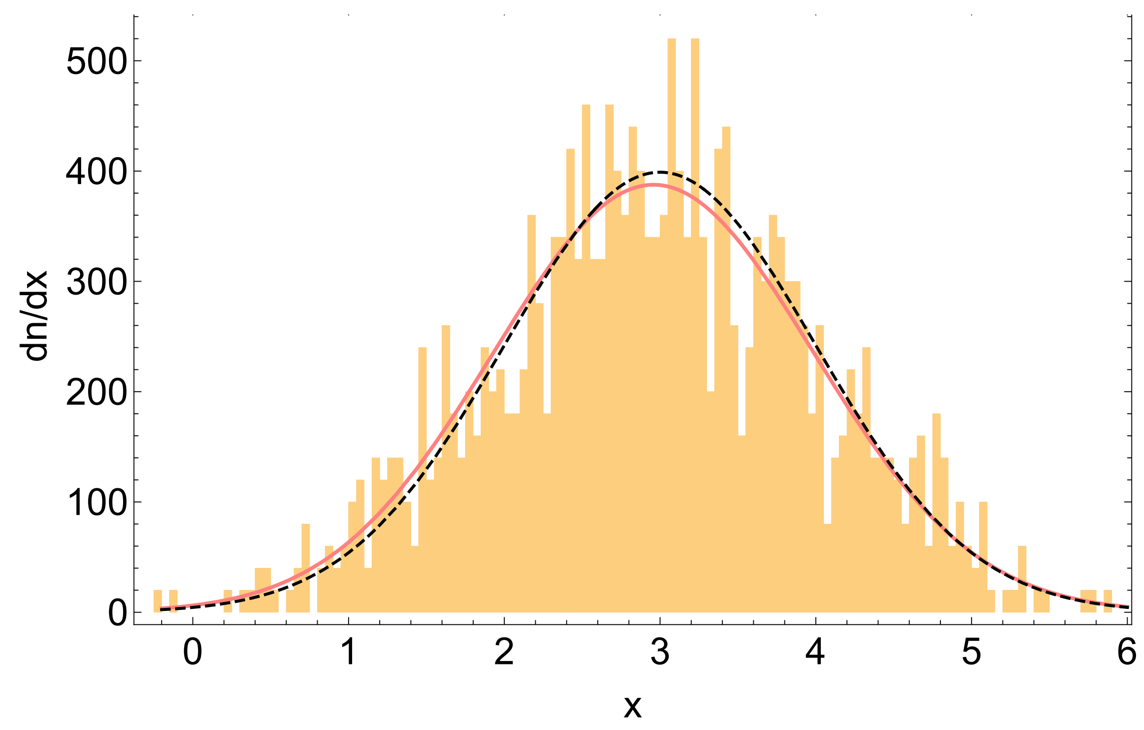
- The model in which the probability of obtaining the observed is the highest is the most likely (MLE)

$$\text{MLE:} \quad \hat{\theta} = \text{argmax}\left(\mathscr{L}\left(\theta|x_{\text{obs}}\right)\right)$$

- Example: Gaussian PDF $\{x_0, \sigma\} = \theta$

  - $$\mathscr{L}\left(x_0, \sigma|x\right) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{\sum_i (x_i - x_0)^2}{2\sigma^2}}$$

  - $$\text{MLE: } \hat{x}_0 = \bar{x}, \hat{\sigma} = \sqrt{\frac{1}{N}\left(x_i - \bar{x}\right)^2}.$$



LIKELIHOOD 101 - HYPOTHESES TESTING

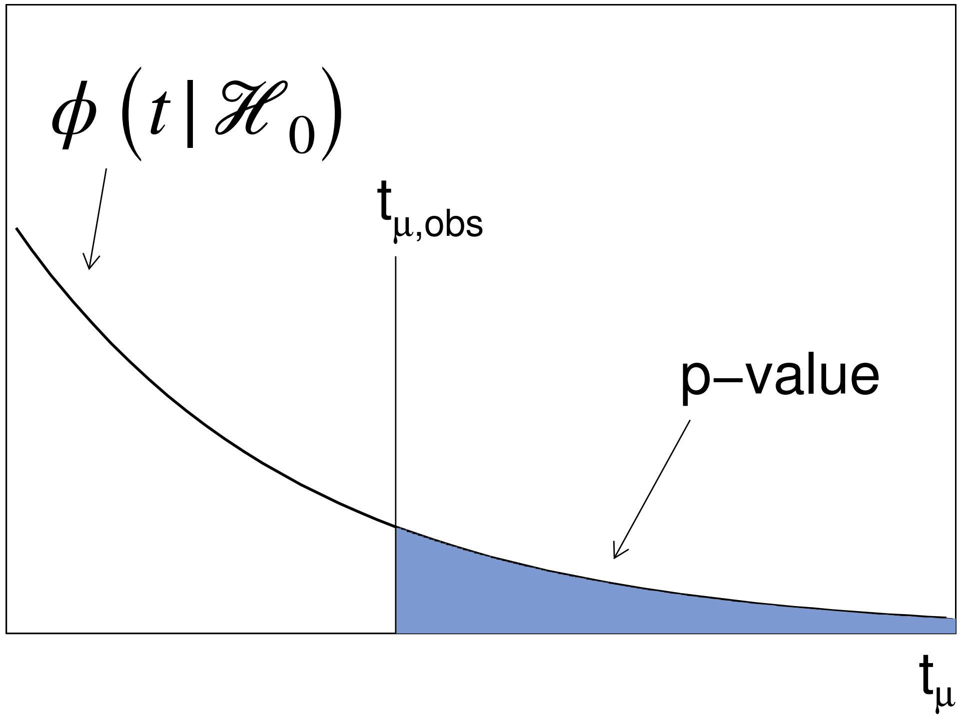
- Maximum profile likelihood test of $\mathscr{H}_0(\mu_0, \nu)$ vs. $\mathscr{H}_1(\mu, \nu)$

$$t_{\text{obs}} = 2\log\left(\frac{\max_{\mu,\nu}\left(\mathscr{L}\left(\mathscr{H}_1 | x_{\text{obs}}\right)\right)}{\max_{\nu}\left(\mathscr{L}\left(\mathscr{H}_0 | x_{\text{obs}}\right)\right)}\right)$$

SM+NP

SM

- **Optimal according to the Neyman-Pearson Lemma.**
- Generate toy datasets $\{x_{\text{toy}}\}$ from $\mathscr{H}_0$
- Find the distribution of t
- Calculate p-value for rejecting $\mathscr{H}_0$



S. S. Wilks, Annals Math. Statist. 9 (1938) 60.

G. Cowan, K. Cranmer, E. Gross & O. Vitells, Eur. Phys. J. C (2011) 71: 1554, [1007.1727]

LIKELIHOOD 101 - HYPOTHESES TESTING

- Maximum profile likelihood test of $\mathscr{H}_0(\mu_0, \nu)$ vs. $\mathscr{H}_1(\mu, \nu)$

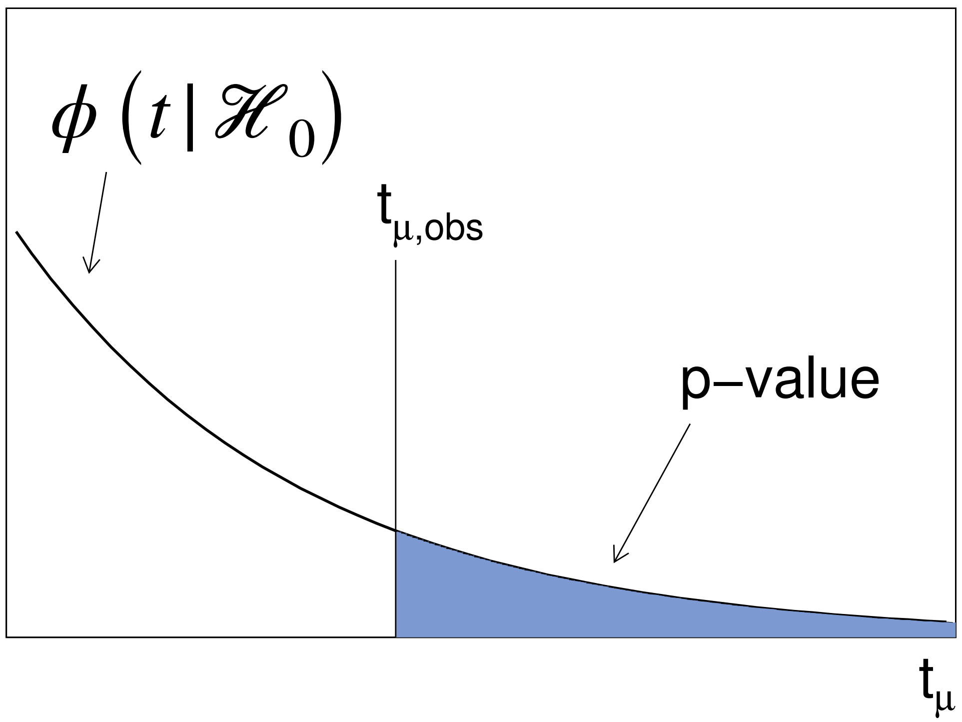
$$t_{\text{obs}} = 2\log\left(\frac{\max_{\mu,\nu}\left(\mathscr{L}\left(\mathscr{H}_1 \mid x_{\text{obs}}\right)\right)}{\max_{\nu}\left(\mathscr{L}\left(\mathscr{H}_0 \mid x_{\text{obs}}\right)\right)}\right)$$

SM+NP

SM

- **Optimal according to the Neyman-Pearson Lemma.**
- **Asymptotic null-distribution - for high enough statistics known regardless of underlying model:**
  $\phi\left(t \mid \mathscr{H}_0\right) \to \chi^2_n$,
  n=#dof NP

Fast & Robust



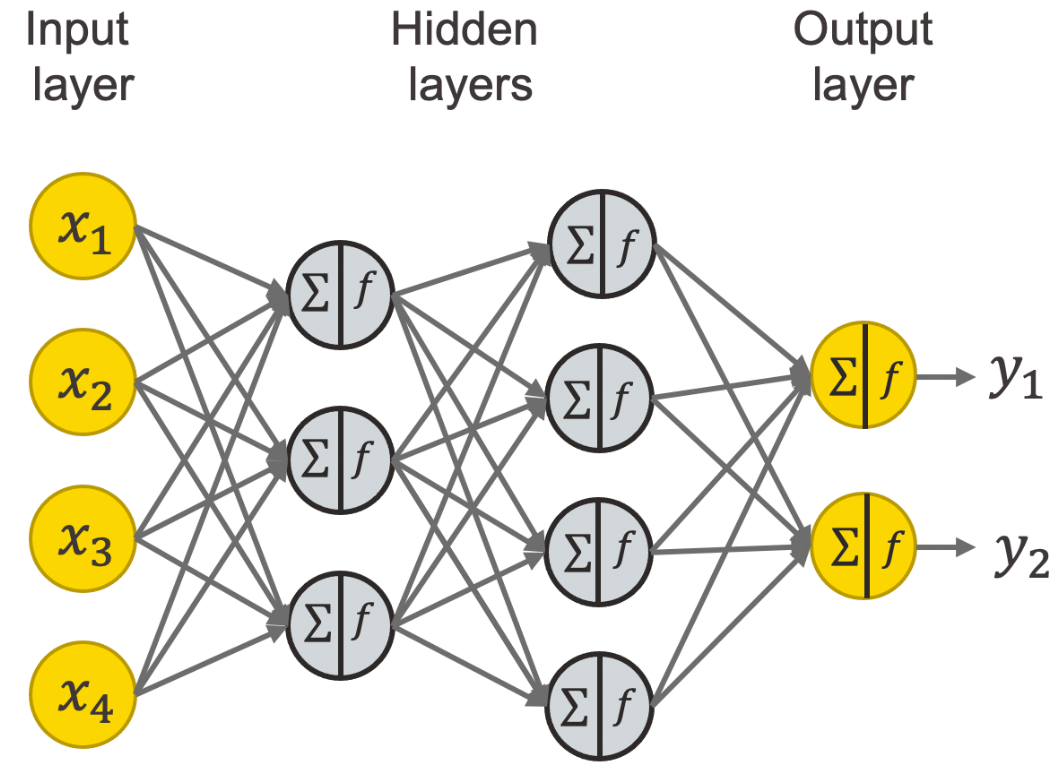
S. S. Wilks, Annals Math. Statist. 9 (1938) 60.

G. Cowan, K. Cranmer, E. Gross & O. Vitells, Eur. Phys. J. C (2011) 71: 1554, [1007.1727]

MACHINE LEARNING 101

- A family of functions - expressive, universal approximators
- Neural Network (NN) - a specific family of functions.
- Training - NN parameters θ found by minimizing some "loss".

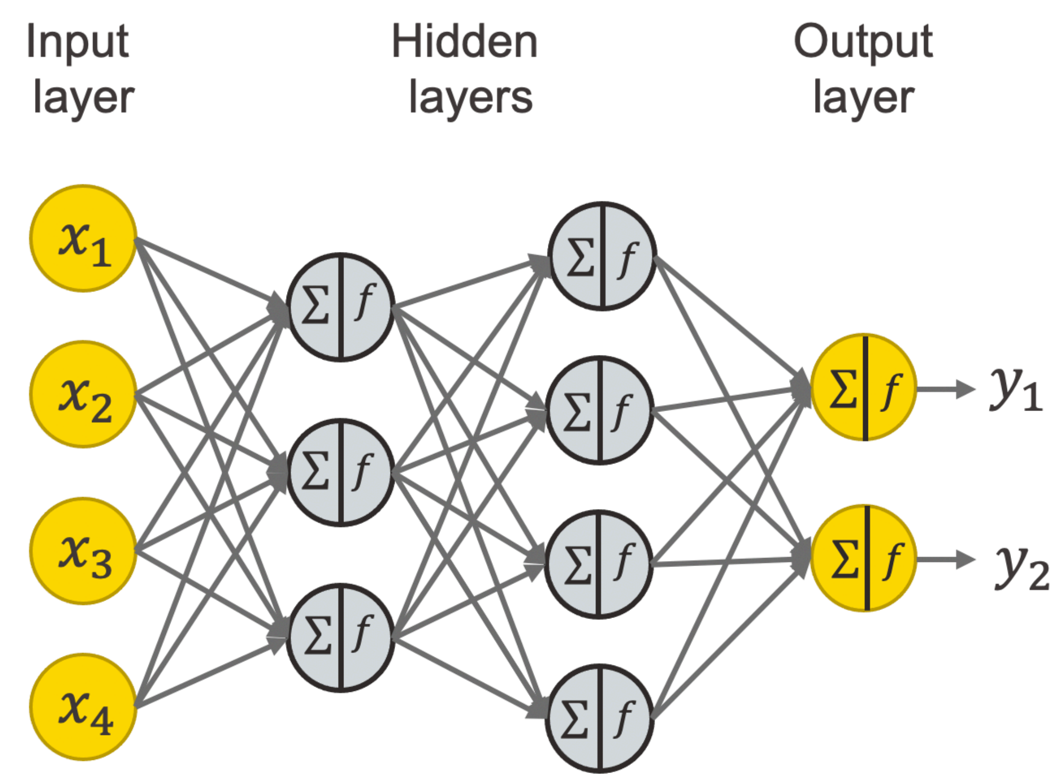
Flexible



MACHINE LEARNING 101

- A family of functions - expressive, universal approximators
- Neural Network (NN) - a specific family of functions.
- Training - NN parameters θ found by minimizing some "loss".
  - $p(x\,|\,\theta) = \mathscr{L}\left(\mathscr{H}(\theta)\,|\,x\right)$ given by the output of a NN.
  - NN loss $= \; -\mathscr{L}\left(\mathscr{H}(\theta)\,|\,x_{\text{obs}}\right)$.
  - $$t_{\text{obs}} = 2\log\left(\frac{\max_{\mu,\nu}\left(\mathscr{L}\left(\mathscr{H}_1\,|\,x_{\text{obs}}\right)\right)}{\max_{\nu}\left(\mathscr{L}\left(\mathscr{H}_0\,|\,x_{\text{obs}}\right)\right)}\right)$$

Flexible



NPLM

- Determine if sample **A** is drawn from SM or SM+NP distribution.

$$\mathscr{H}_0 : n_{\mathbf{A}}(x) = n_{\mathbf{A}}^{\rm SM}(x)\,, \qquad \mathscr{H}_1 : n_{\mathbf{A}}(x) = n_{\mathbf{A}}^{\rm SM+NP}(x)\,,$$

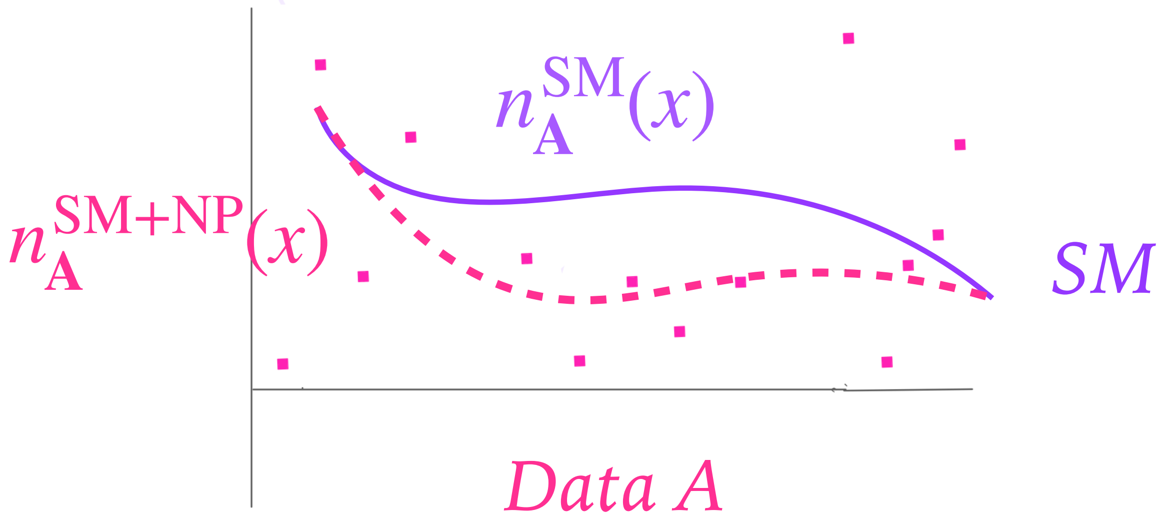
- Profile likelihood test

$$n(x) = N\,p(x)$$

$$t = 2\log\left(\frac{\max\left(\mathscr{L}\left(\mathscr{H}_1 | \mathbf{A}\right)\right)}{\max\left(\mathscr{L}\left(\mathscr{H}_0 | \mathbf{A}\right)\right)}\right),$$

- Poisson likelihood:

$$\mathscr{L}\left(\mathscr{H} | \mathbf{A}\right) = \frac{e^{-N_{\rm A}(\mathscr{H})}}{\tilde{N}_{\mathbf{A}}!}\prod_{x\in\mathbf{A}} n_{\mathbf{A}}(x | \mathscr{H})\,.$$



R. T. D'Agnolo & A. Wulzer, [1806.02350].

NPLM

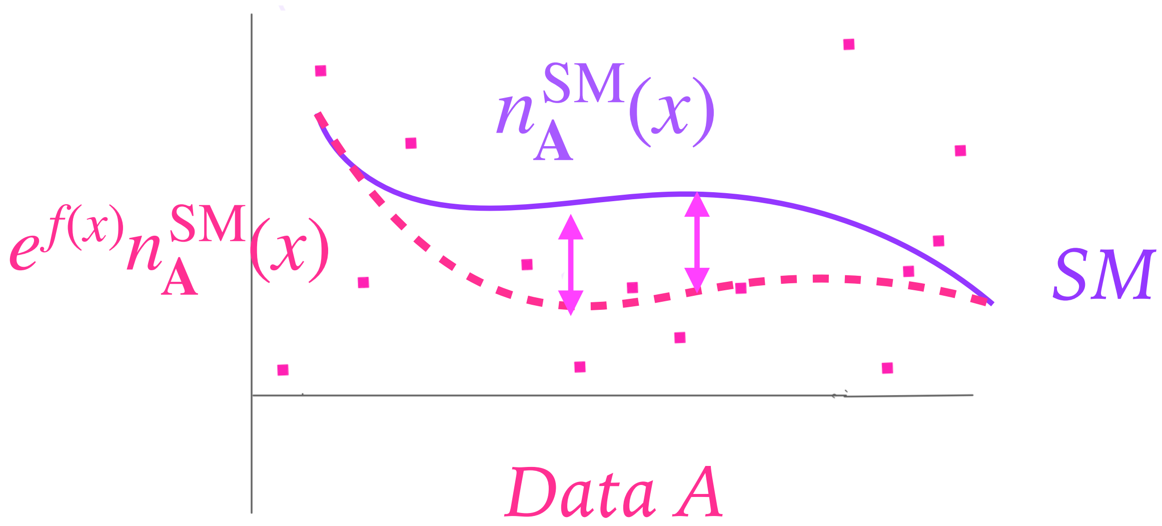
- Determine if sample **A** is drawn from SM or SM+NP distribution.

$$\mathcal{H}_0 : n_{\mathbf{A}}(x) = n_{\mathbf{A}}^{\rm SM}(x)\ , \qquad \mathcal{H}_1 : n_{\mathbf{A}}(x) = e^{f(x)} n_{\mathbf{A}}^{\rm SM}(x)\ ,$$

- Profile likelihood test

$$t = 2\left(-\int \left(e^{\hat{f}(x)} - 1\right) \hat{n}_{\mathbf{A}}^{\rm SM}(x)dx + \sum_{x\in\mathbf{A}} \hat{f}(x)\right)$$

NP: $f(x)$ is an output of a NN maximizing t



R. T. D'Agnolo & A. Wulzer, [1806.02350].

NPLM

- Determine if sample **A** is drawn from SM or SM+NP distribution.

$$\mathcal{H}_0 : n_{\mathbf{A}}(x) = n_{\mathbf{A}}^{\mathrm{SM}}(x) \,, \qquad \mathcal{H}_1 : n_{\mathbf{A}}(x) = e^{f(x)} n_{\mathbf{A}}^{\mathrm{SM}}(x) \,,$$

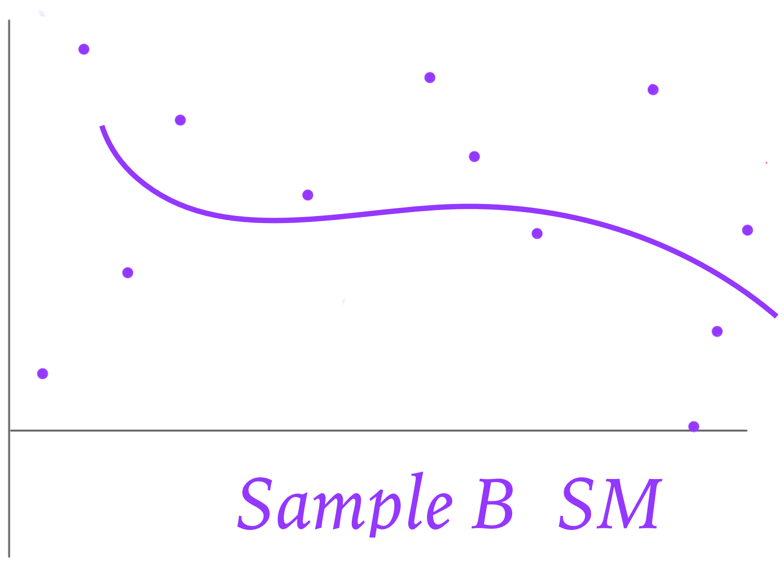
- Profile likelihood test

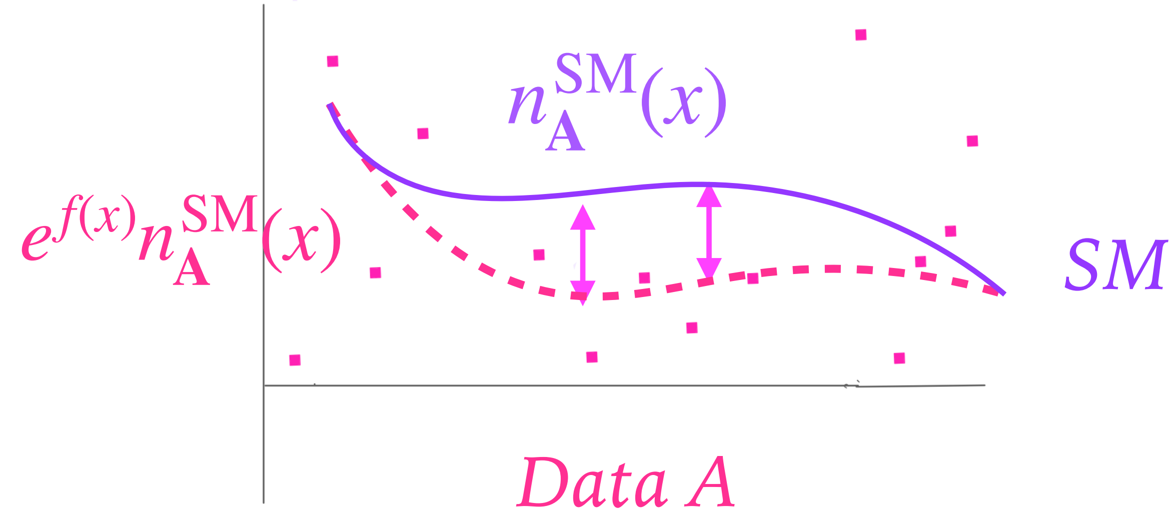
*SM: fit from control sample **B***

$$t = 2\left(-\int \left(e^{\hat{f}(x)} - 1 \right) \hat{n}_{\mathbf{A}}^{\mathrm{SM}}(x) dx + \sum_{x \in \mathbf{A}} \hat{f}(x) \right)$$

NP: f(x) is an output of a NN maximizing t

- SM dist. given by large sample **B** drawn from it, $\tilde{N}_{\mathbf{B}} \gg \tilde{N}_{\mathbf{A}}$





R. T. D'Agnolo & A. Wulzer, [1806.02350].

NPLM

- Determine if sample **A** is drawn from SM or SM+NP distribution.

$$\mathscr{H}_0 : n_{\mathbf{A}}(x) = N_{\mathbf{A}}\tilde{p}_{\mathbf{B}}(x)\,, \qquad \mathscr{H}_1 : n_{\mathbf{A}}(x) = e^{f(x)}N_{\mathbf{A}}\tilde{p}_{\mathbf{B}}(x)\,,$$

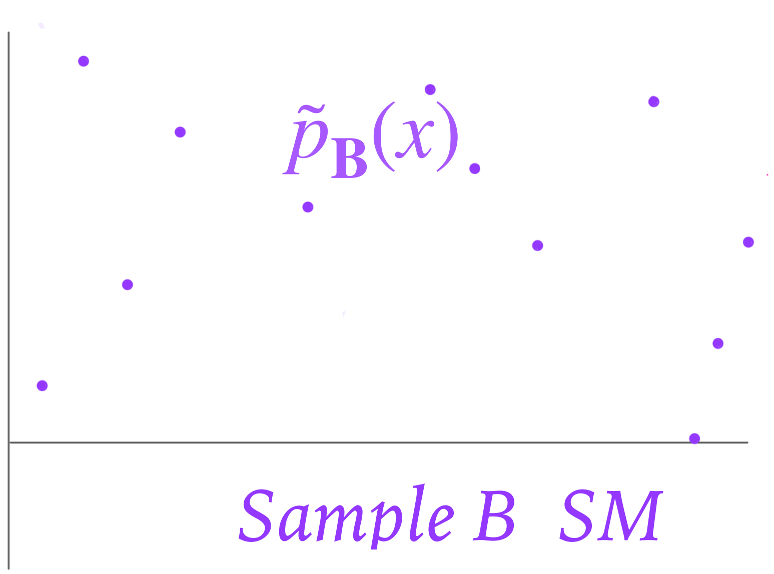
- Profile likelihood test

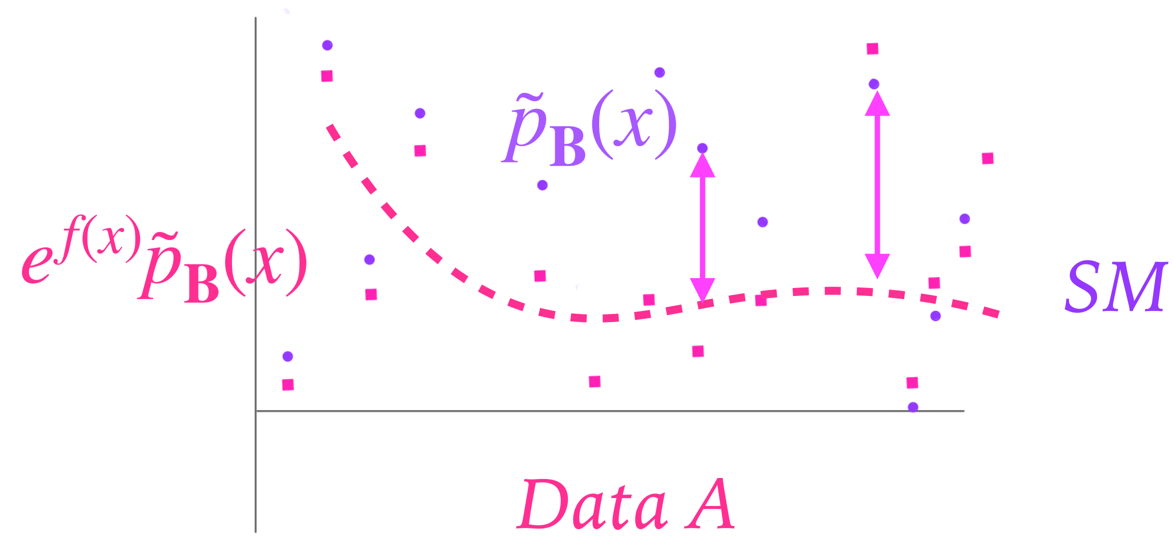
SM: empiric observation B

$$t = 2\left(-\frac{N_{\mathbf{A}}}{\tilde{N}_{\mathbf{B}}}\sum_{x\in\mathbf{B}}\left(e^{\hat{f}(x)}-1\right)+\sum_{x\in\mathbf{A}}\hat{f}(x)\right)$$

NP: $f(x)$ is an output of a NN maximizing t

- SM dist. **represented** by large sample **B** drawn from it, $\tilde{N}_{\mathbf{B}} \gg \tilde{N}_{\mathbf{A}}$





R. T. D'Agnolo & A. Wulzer, [1806.02350].

NPLM

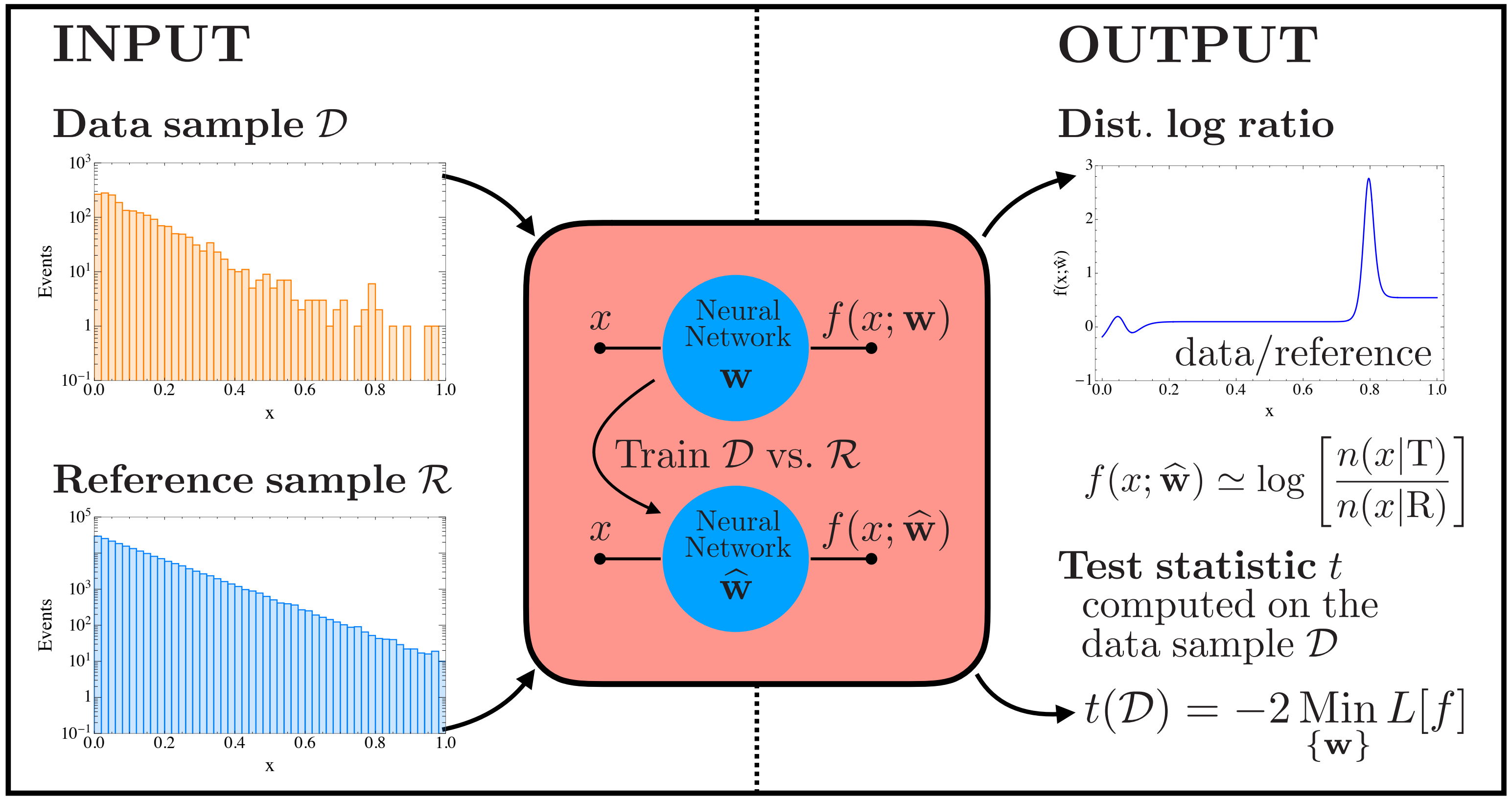
- Profile likelihood test loss

SM: empiric observation B

$$L = -\left(-\frac{N_{\mathbf{A}}}{\tilde{N}_{\mathbf{B}}}\sum_{x\in\mathbf{B}}\left(e^{\hat{f}(x)}-1\right)+\sum_{x\in\mathbf{A}}\hat{f}(x)\right)$$

NP: f(x) is an output of a NN minimizing L

- Interpretation:
  - **NP**: deviations of $f(x)$ from zero
  - **Significance**: p-value of obtained $t = -2L$



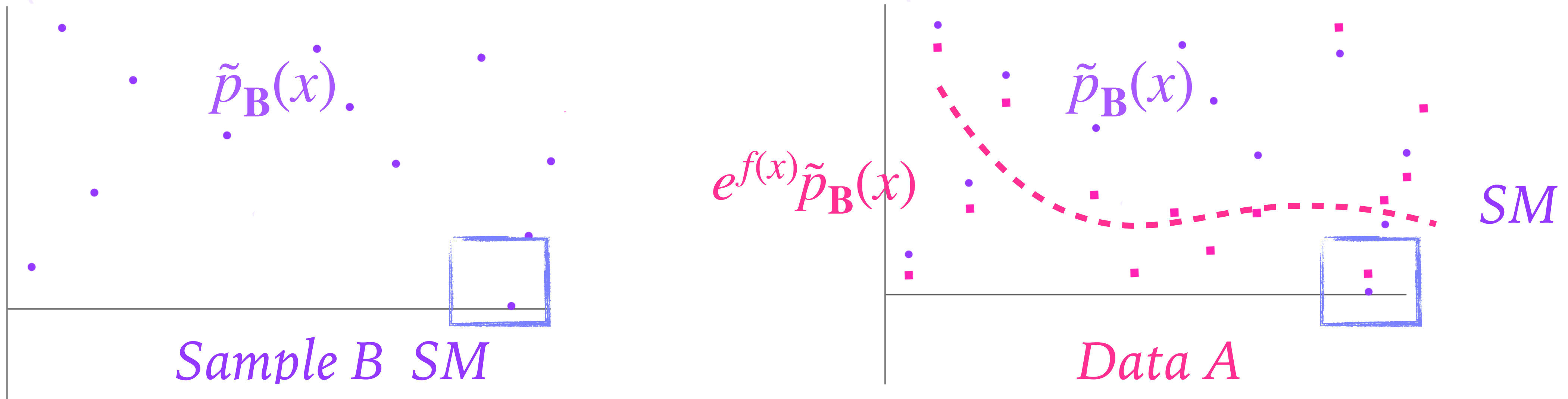
R. T. D'Agnolo & A. Wulzer, [1806.02350].

NPLM CHALLENGES: SEVERE WEIGHT-CLIPPING

- Unbounded loss

$$L = -\left(-\frac{N_{\mathbf{A}}}{\tilde{N}_{\mathbf{B}}} \sum_{x \in \mathbf{B}} \left(e^{\hat{f}(x)} - 1 \right) + \sum_{x \in \mathbf{A}} \hat{f}(x) \right)$$

- For $x_\star \in (A - A \cap B)$, if $f(x_\star) \to \infty$ then $L \to -\infty$.



NPLM CHALLENGES: SEVERE WEIGHT-CLIPPING

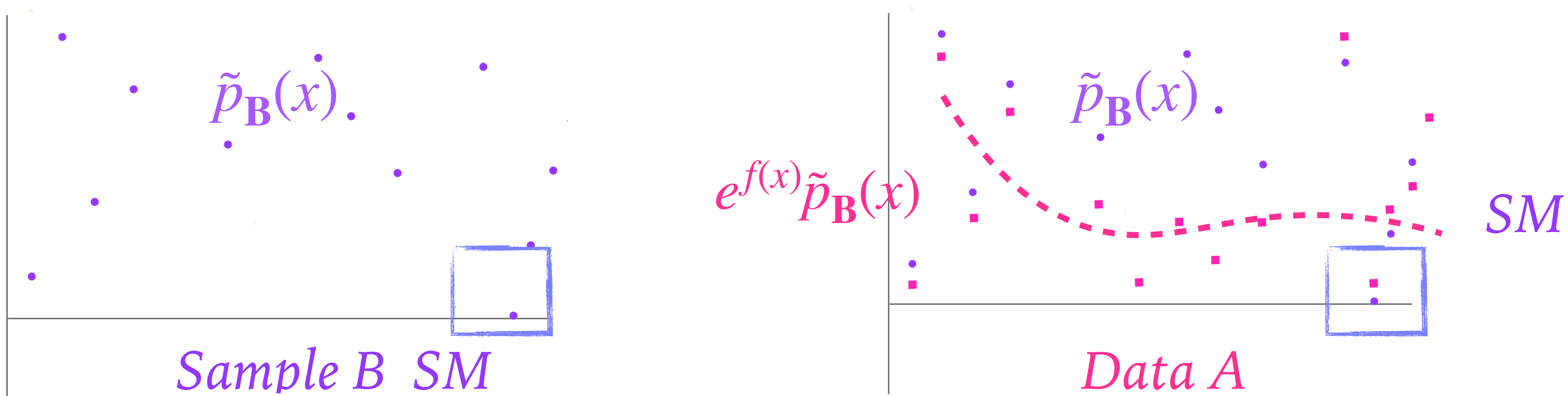
- Unbounded loss

$$L = -\left(-\frac{N_{\mathbf{A}}}{\tilde{N}_{\mathbf{B}}} \sum_{x \in \mathbf{B}} \left(e^{\hat{f}(x)} - 1 \right) + \sum_{x \in \mathbf{A}} \hat{f}(x) \right)$$

- For $x_\star \in (A - A \cap B)$, if $f(x_\star) \to \infty$ then $L \to -\infty$.
- **This is a result of a false null-hypothesis.**

$$\mathscr{H}_1 : \quad n_{\mathbf{A}}(x_\star) = e^{f(x_\star)} N_{\mathbf{A}} \tilde{p}_{\mathbf{B}}(x_\star), \qquad t = 2\log\left(\frac{max\left(\mathscr{L}\left(\mathscr{H}_1 | \mathbf{A}\right)\right)}{max\left(\mathscr{L}\left(\mathscr{H}_0 | \mathbf{A}\right)\right) = 0} \right)$$

$$\mathscr{H}_0 : \quad n_{\mathbf{A}}(x_\star) = N_{\mathbf{A}} \tilde{p}_{\mathbf{B}}(x_\star) = 0,$$

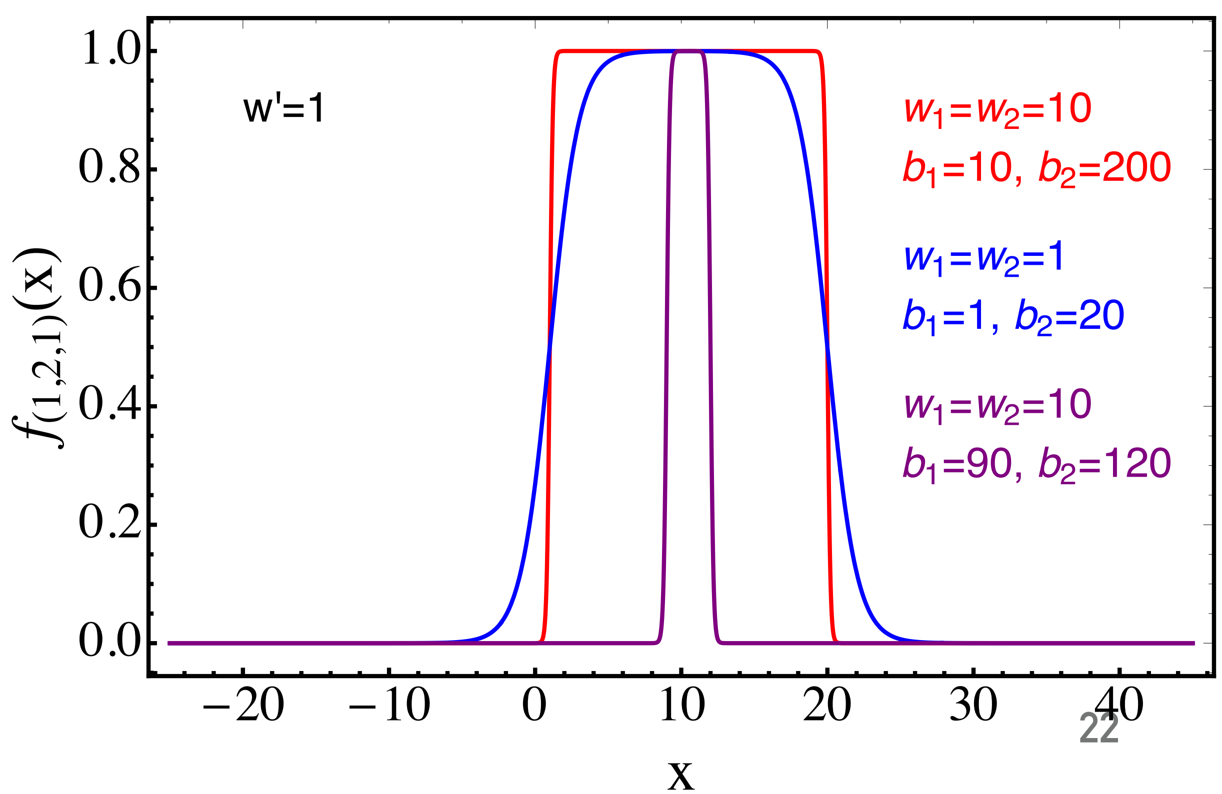


NPLM CHALLENGES: SEVERE WEIGHT-CLIPPING

- Unbounded loss

$$L = -\left(-\frac{N_{\mathbf{A}}}{\tilde{N}_{\mathbf{B}}} \sum_{x \in \mathbf{B}} \left(e^{\hat{f}(x)} - 1 \right) + \sum_{x \in \mathbf{A}} \hat{f}(x) \right)$$

- For $x_\star \in (A - A \cap B)$, if $f(x_\star) \to \infty$ then $L \to -\infty$.
- Weight-clipping - setting a max for NN weights (~gradients).
  - Not allow divergences on small scales $\Delta x \leq 1/W$.
  - Smooth out local fluctuations.

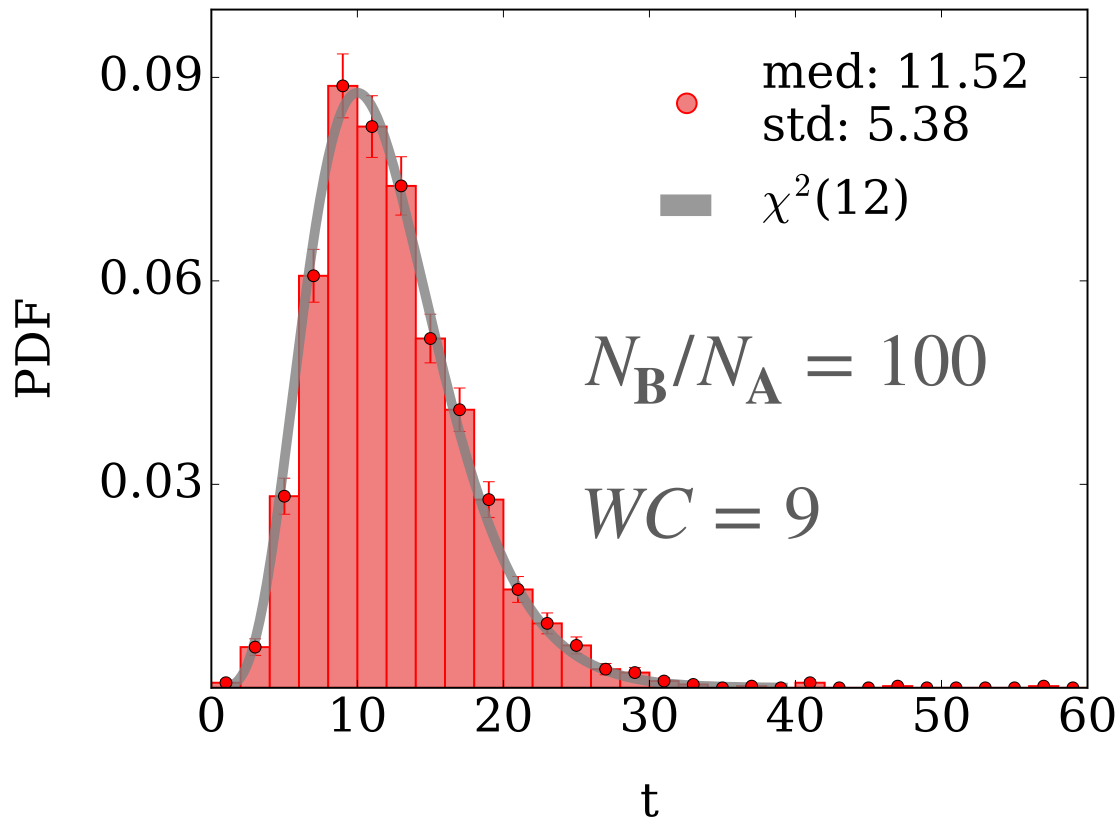


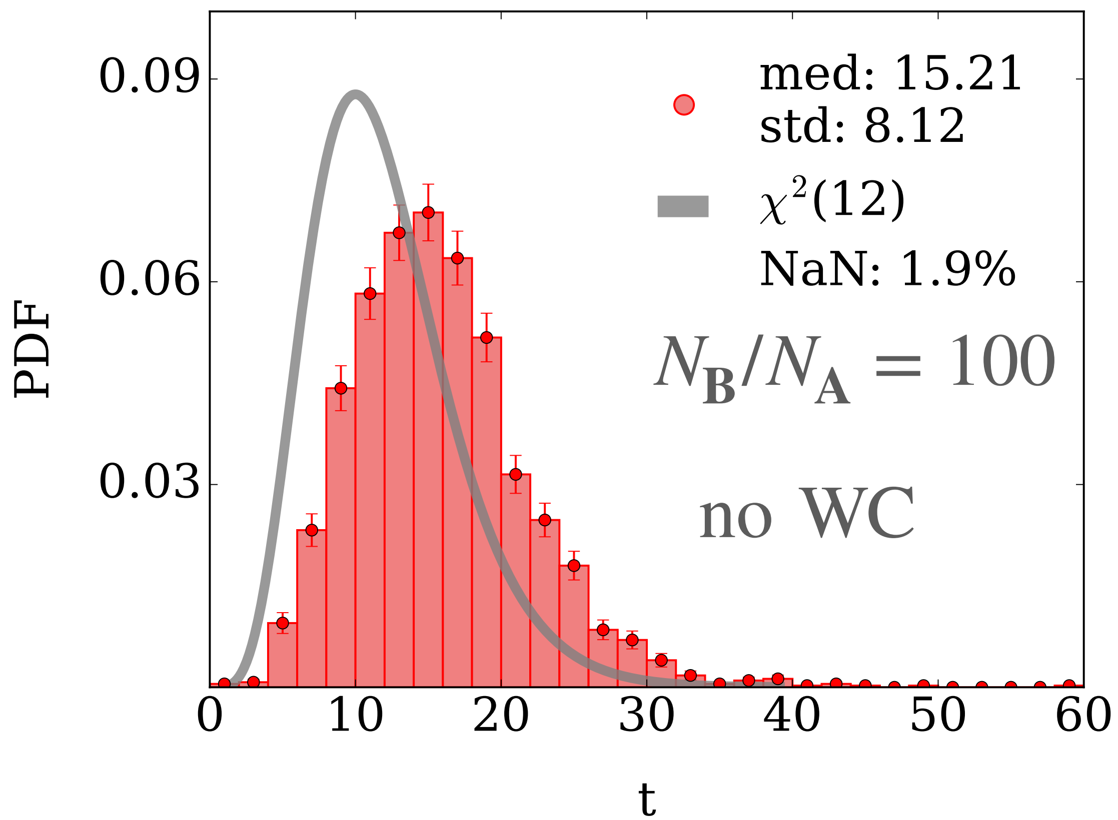
R. T. D'Agnolo & A. Wulzer, [1806.02350].

NPLM CHALLENGES: SEVERE WEIGHT-CLIPPING

- Weight-clipping - setting a max for NN weights (~gradients).
- Determined to reach the asymptotic distribution and avoid divergences.
  - **Costly optimization process.**
  - **The stricter the WC, the less flexible the NN.**

χ^2_n predicted for likelihood test

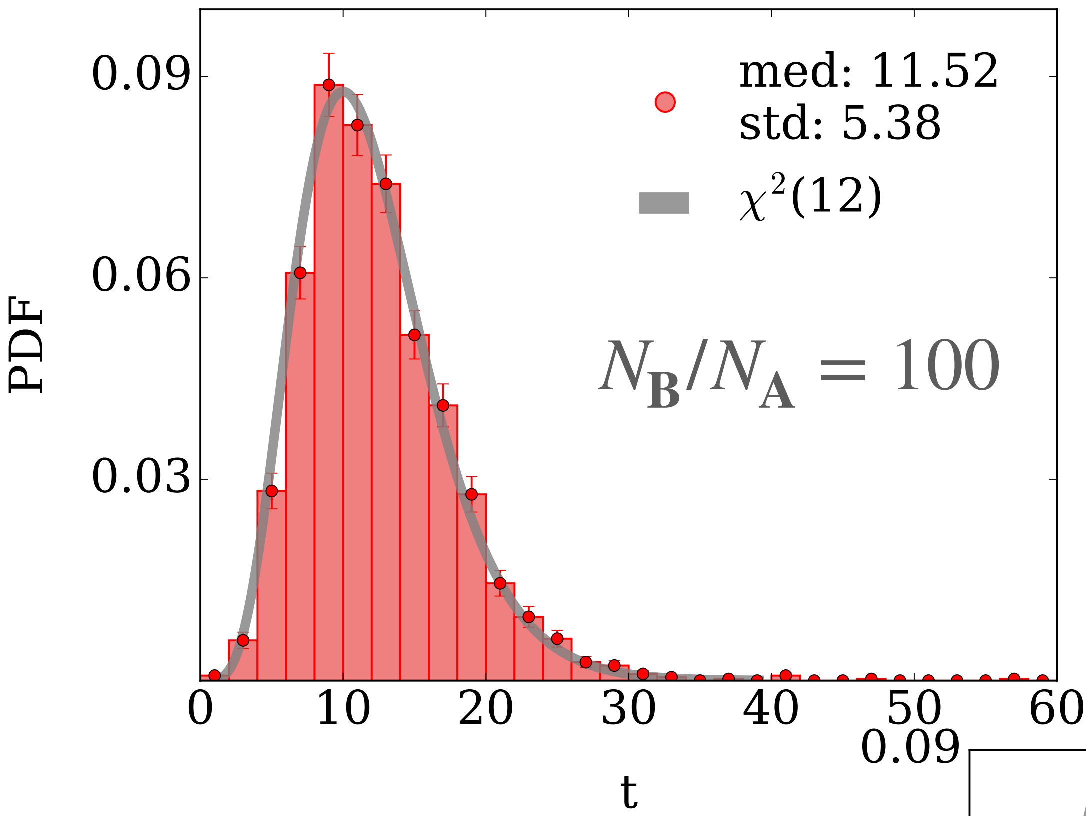




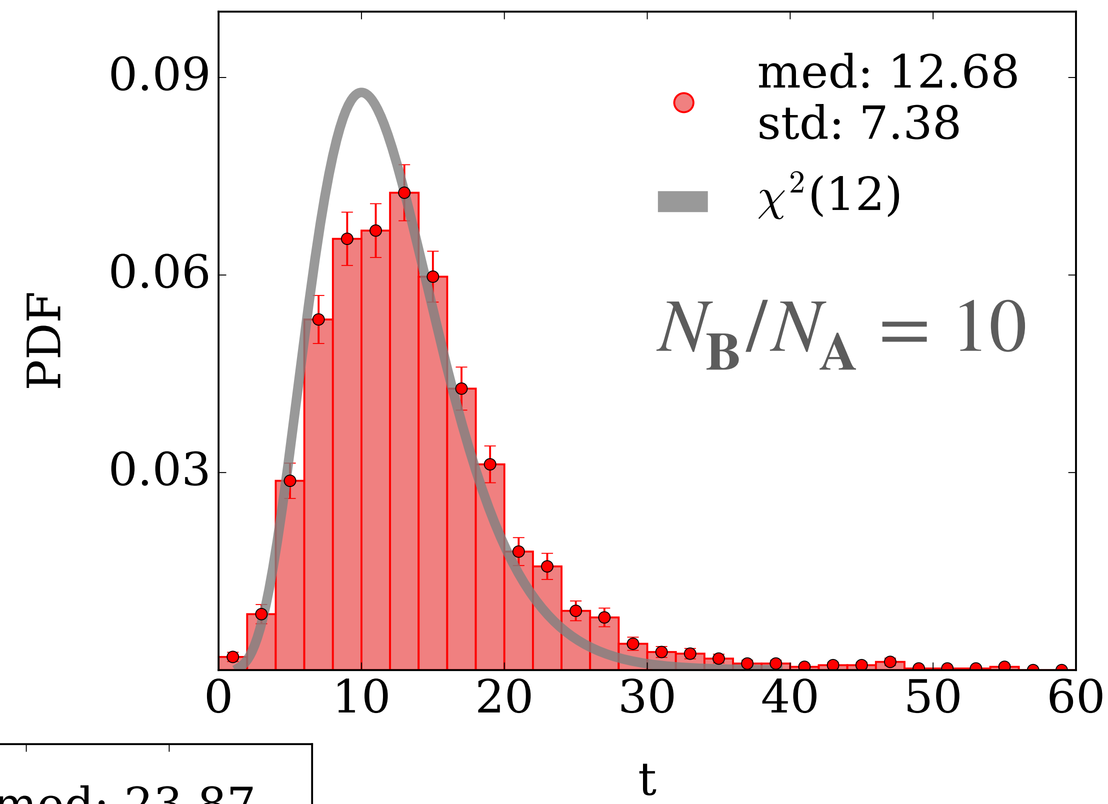
R. T. D'Agnolo & A. Wulzer, [1806.02350].

NPLM CHALLENGES: IMBALANCED SAMPLES

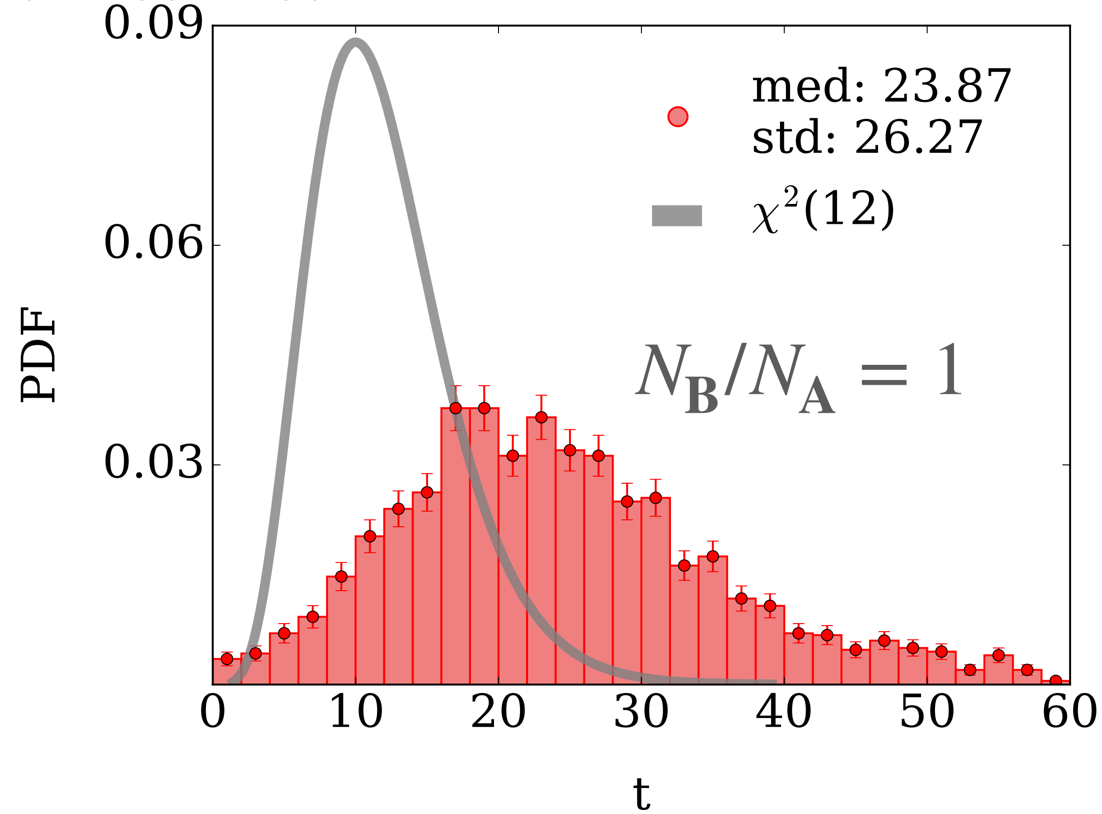
- Requires a large ratio between sample sizes $\tilde{N}_\mathbf{B} \gg \tilde{N}_\mathbf{A}$



WC = 9



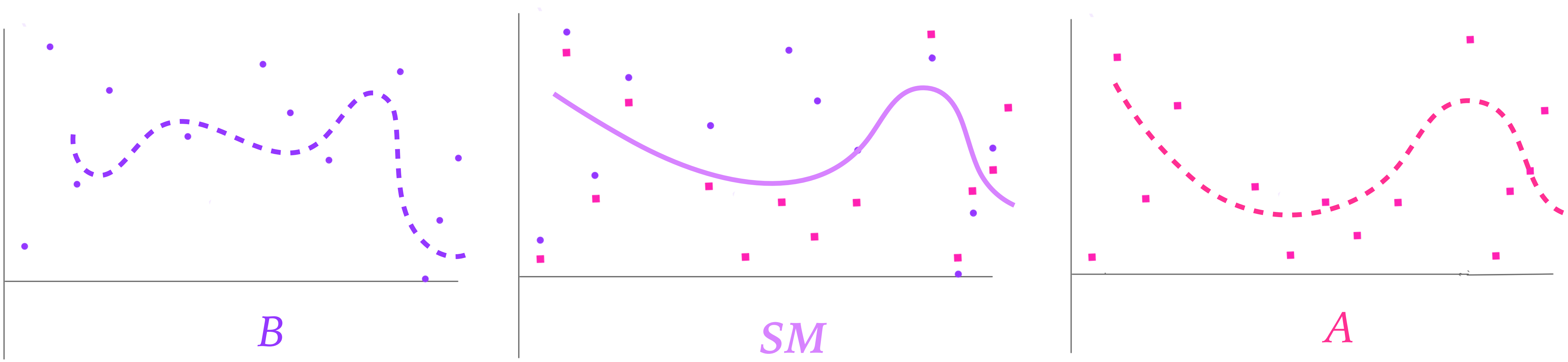
χ^2_n predicted for likelihood test



t distribution for toy data A and B generated from the same PDF

THE SYMMETRIZED FORMALISM

- **Symmetric question:** instead of asking if sample A comes from the distribution of sample B, we ask if A and B come from the same distribution.
- **Symmetric (democratic) modeling:** account for fluctuations in both samples.
  - **Improved sensitivity for any sample sizes ratio N_A/N_B.**
  - **Avoid artificial singularities.**



THE SYMMETRIZED FORMALISM - SYMMETRIC TEST

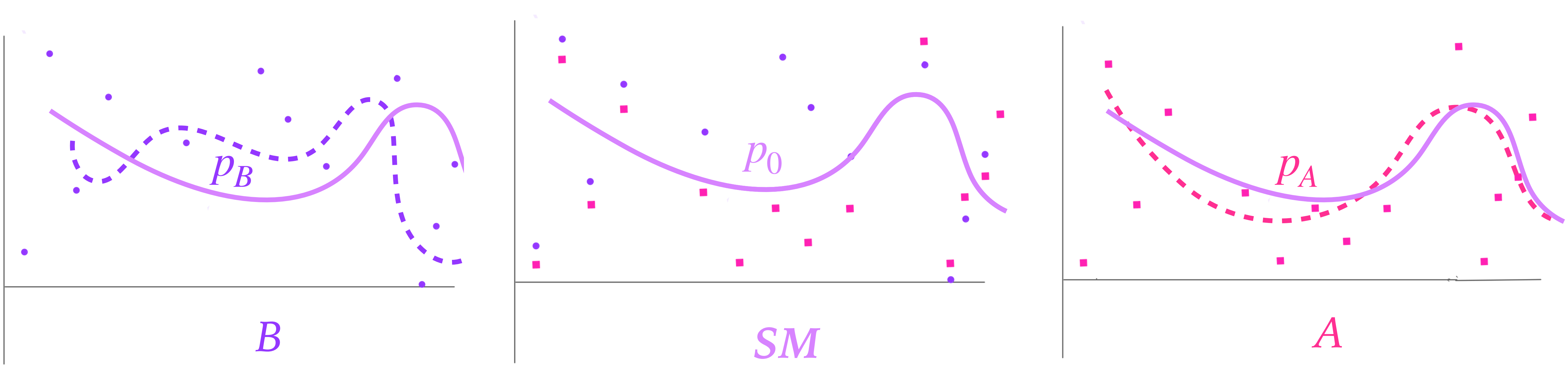
- Determine if samples **A** and **B** are drawn from the same distribution.

$$\mathscr{H}_0: \quad n_{\mathbf{A}}(x) = N_{\mathbf{A}} p_0(x) \ , \qquad n_{\mathbf{B}}(x) = N_{\mathbf{B}} p_0(x)$$

$$\mathscr{H}_1: \quad n_{\mathbf{A}}(x) = N_{\mathbf{A}} p_{\mathbf{A}}(x) \ , \qquad n_{\mathbf{B}}(x) = N_{\mathbf{B}} p_{\mathbf{B}}(x) \ ,$$

- Symmetric test - learn common PDF from both samples, test on both

$$t = 2\log\left(\frac{\max_{p_{\mathbf{A}},p_{\mathbf{B}}}\left(\mathscr{L}\left(N_{\mathbf{A}}, p_{\mathbf{A}}(x)\,|\,\mathbf{A}\right)\mathscr{L}\left(N_{\mathbf{B}}, p_{\mathbf{B}}(x)\,|\,\mathbf{B}\right)\right)}{\max_{p_0}\left(\mathscr{L}\left(N_{\mathbf{A}}, p_0(x)\,|\,\mathbf{A}\right)\mathscr{L}\left(N_{\mathbf{B}}, p_0(x)\,|\,\mathbf{B}\right)\right)}\right)$$



THE SYMMETRIZED FORMALISM - SYMMETRIC TEST

- **Determine if samples A and B are drawn from the same distribution.**

$$\mathscr{H}_0: \quad n_{\mathbf{A}}(x) = N_{\mathbf{A}} p_0(x) \,, \quad n_{\mathbf{B}}(x) = N_{\mathbf{B}} p_0(x)$$
$$\mathscr{H}_1: \quad n_{\mathbf{A}}(x) = N_{\mathbf{A}} p_{\mathbf{A}}(x) \,, \quad n_{\mathbf{B}}(x) = N_{\mathbf{B}} p_{\mathbf{B}}(x) \,,$$

- Symmetric test - learn common PDF from both samples, test on both

$$t = 2 \log \left(\frac{\max_{p_{\mathbf{A}}, p_{\mathbf{B}}} \left(\mathscr{L}\left(N_{\mathbf{A}}, p_{\mathbf{A}}(x) \,|\, \mathbf{A}\right) \mathscr{L}\left(N_{\mathbf{B}}, p_{\mathbf{B}}(x) \,|\, \mathbf{B}\right) \right)}{\max_{p_0} \left(\mathscr{L}\left(N_{\mathbf{A}}, p_0(x) \,|\, \mathbf{A}\right) \mathscr{L}\left(N_{\mathbf{B}}, p_0(x) \,|\, \mathbf{B}\right) \right)} \right)$$

- **NPLM**: if $\tilde{N}_{\mathbf{B}} \gg \tilde{N}_{\mathbf{A}}$, learn common PDF from **B** - $\hat{p}_0 \approx \hat{p}_{\mathbf{B}}$, test on **A**

$$t_{N_{\mathbf{B}} \gg N_{\mathbf{A}}} \to 2 \log \left(\frac{\max_{p_{\mathbf{A}}} \left(\mathscr{L}\left(N_{\mathbf{A}}, p_{\mathbf{A}}(x) \,|\, \mathbf{A}\right) \right)}{\mathscr{L}\left(N_{\mathbf{A}}, \hat{p}_B(x) \,|\, \mathbf{A}\right)} \right)$$

THE SYMMETRIZED FORMALISM

- Determine if observed samples **A** and **B** are drawn from the same distribution.

$$\mathscr{H}_0: \quad n_{\mathbf{A}}(x) = N_{\mathbf{A}} e^{f_0} p_0(x)\ , \qquad n_{\mathbf{B}}(x) = N_{\mathbf{B}} e^{g_0} p_0(x)$$

$$\mathscr{H}_1: \quad n_{\mathbf{A}}(x) = N_{\mathbf{A}} e^{f(x)} p_0(x)\ , \qquad n_{\mathbf{B}}(x) = N_{\mathbf{B}} e^{g(x)} p_0(x)\ ,$$

- **The symmetric null distribution -**

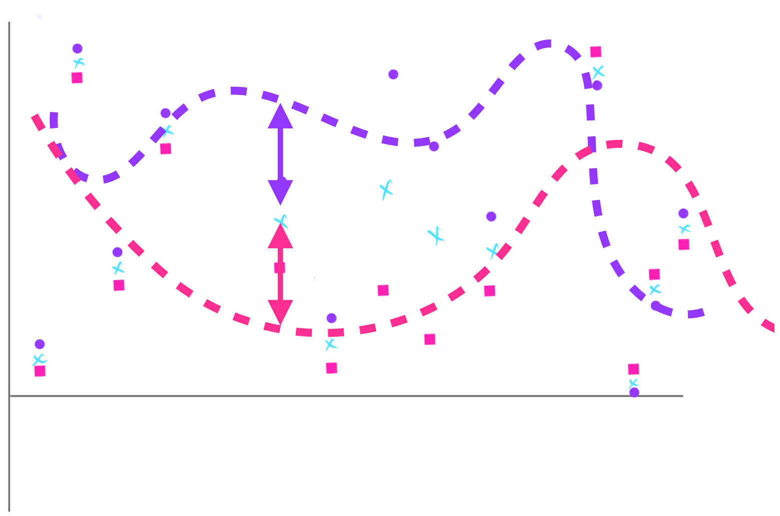
NPLM

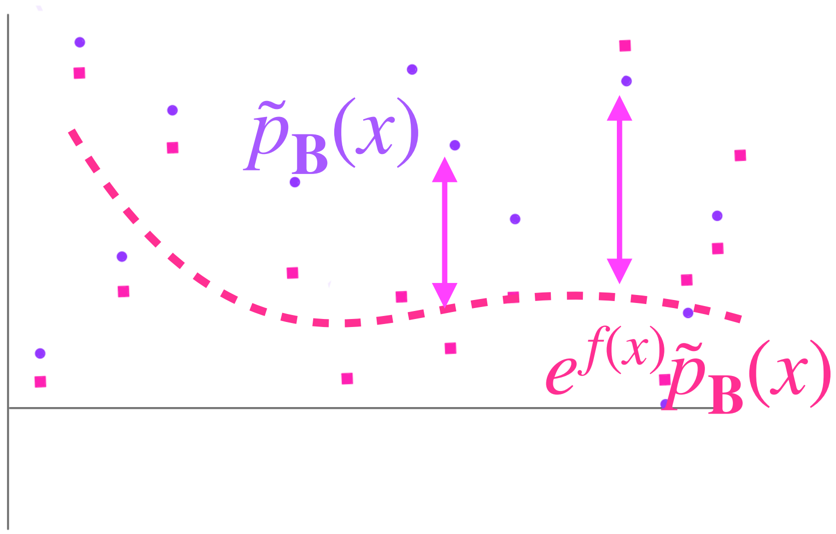
True global MLE $\mathscr{H}_0$

$$p_0(x) = \frac{\tilde{n}_{\mathbf{A}}(x) + \tilde{n}_{\mathbf{B}}(x)}{\tilde{N}_{\mathbf{A}} + \tilde{N}_{\mathbf{B}}}$$

$$p_0(x) = \frac{\tilde{n}_{\mathbf{B}}(x)}{\tilde{N}_{\mathbf{B}}}$$

Approx. global MLE $\mathscr{H}_0$





THE SYMMETRIZED FORMALISM

- Determine if observed samples **A** and **B** are drawn from the same distribution.

$$\mathcal{H}_0: \quad n_{\mathbf{A}}(x) = N_{\mathbf{A}} e^{f_0} p_0(x) , \qquad n_{\mathbf{B}}(x) = N_{\mathbf{B}} e^{g_0} p_0(x)$$

$$\mathcal{H}_1: \quad n_{\mathbf{A}}(x) = N_{\mathbf{A}} e^{f(x)} p_0(x) , \qquad n_{\mathbf{B}}(x) = N_{\mathbf{B}} e^{g(x)} p_0(x) ,$$

- The symmetric null distribution -

True global MLE $\mathcal{H}_0$

$$p_0(x) = \frac{\tilde{n}_{\mathbf{A}}(x) + \tilde{n}_{\mathbf{B}}(x)}{\tilde{N}_{\mathbf{A}} + \tilde{N}_{\mathbf{B}}}$$

NPLM

$$p_0(x) = \frac{\tilde{n}_{\mathbf{B}}(x)}{\tilde{N}_{\mathbf{B}}}$$

- The symmetric test statistic -

Approx. global MLE $\mathcal{H}_0$

$$t_{\mathbf{A+B}}(\mathbf{A}) = -2\min\left[-\frac{1}{\tilde{N}_{\mathbf{A}} + \tilde{N}_{\mathbf{B}}} \sum_{x \in \mathbf{A,B}} \tilde{N}_{\mathbf{A}} \left(e^{f(x)} - 1\right) + \sum_{x \in \mathbf{A}} f(x)\right]$$

$$t_{\mathbf{A+B}}(\mathbf{B}) = -2\min\left[-\frac{1}{\tilde{N}_{\mathbf{A}} + \tilde{N}_{\mathbf{B}}} \sum_{x \in \mathbf{A,B}} \tilde{N}_{\mathbf{B}} \left(e^{g(x)} - 1\right) + \sum_{x \in \mathbf{B}} g(x)\right]$$

$$t_{\mathbf{B}}(\mathbf{A}) = 2\left(-\frac{N_{\mathbf{A}}}{\tilde{N}_{\mathbf{B}}} \sum_{x \in \mathbf{B}} \left(e^{\hat{f}(x)} - 1\right) + \sum_{x \in \mathbf{A}} \hat{f}(x)\right)$$

No divergences

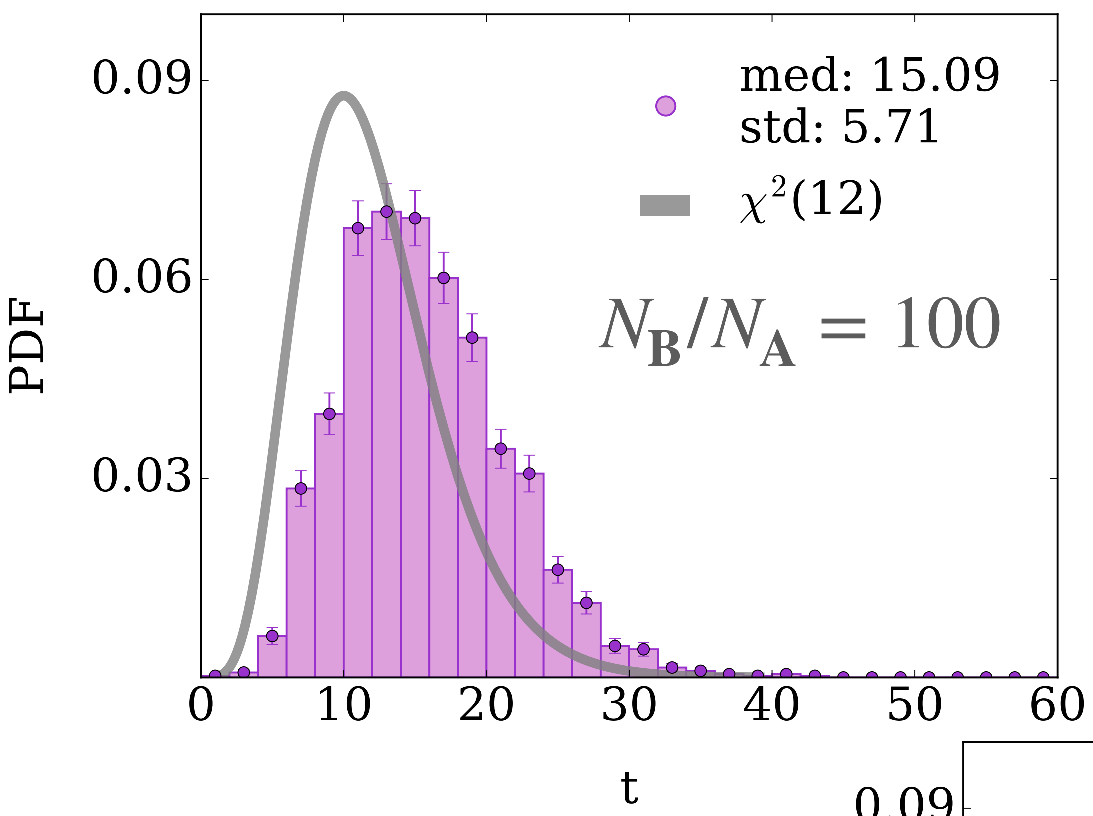
Unbounded

RESULTS

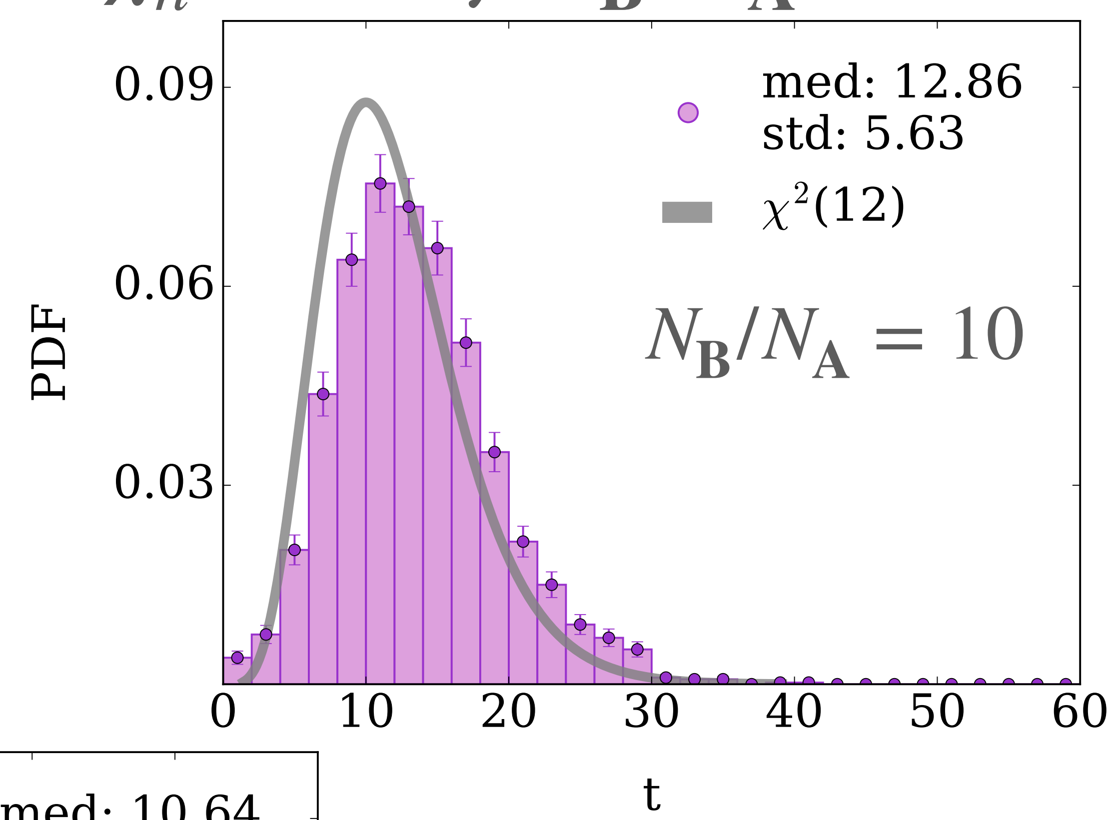
- Toy LFV - $e^{\pm}\mu^{\mp}$ samples with $\sim 2.2 \times 10^5$ events.
- 1-d variable: $x = \dfrac{m_{coll}}{100\,\mathrm{GeV}}$
- Hyper-parameters: 500k epochs, 1 hidden layer of 4 neurons
- Symmetric - $\mathbf{A}$ and $\mathbf{B}$ randomly drawn from the $e\mu$ sample
- Asymmetric - $H \to \tau e, \tau \to \mu + X$ added to $\mathbf{A}$.

RESULTS - THE SYMMETRIC CASE

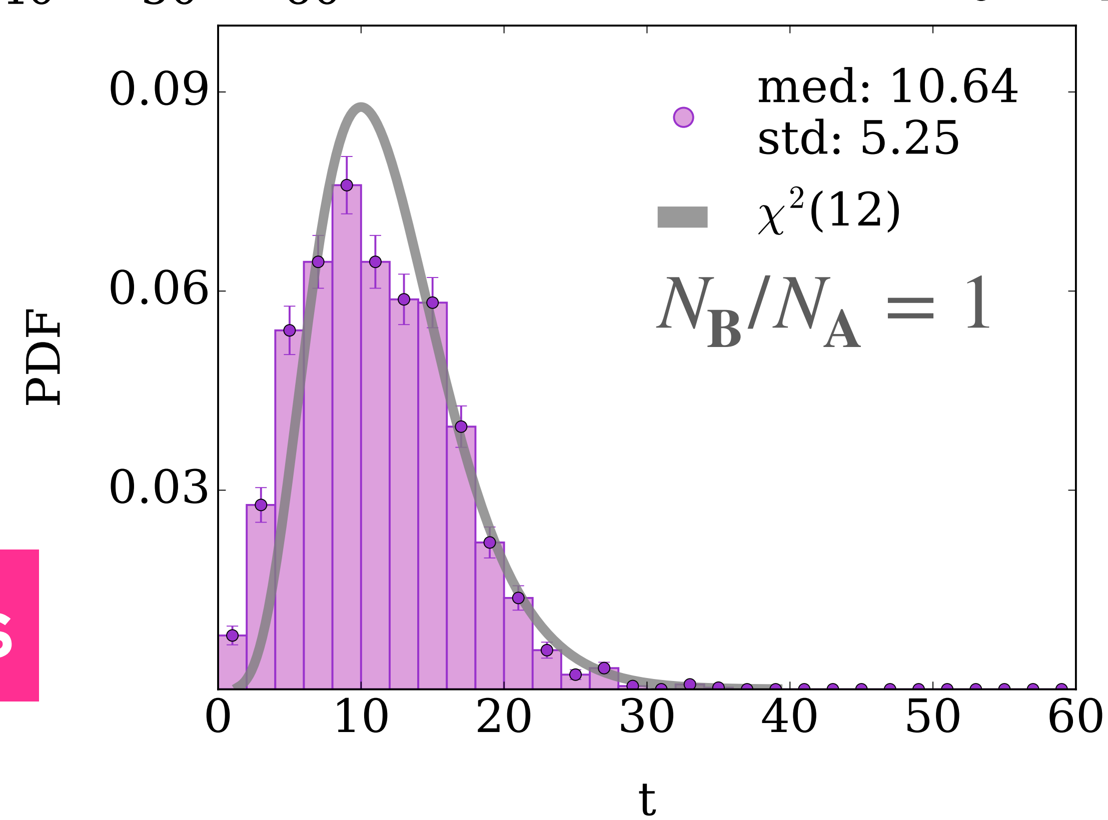
- Good agreement with asymptotic χ^2_n for any $\tilde{N}_{\mathrm{B}}/\tilde{N}_{\mathrm{A}}$



No WC!



χ^2_n predicted for likelihood test

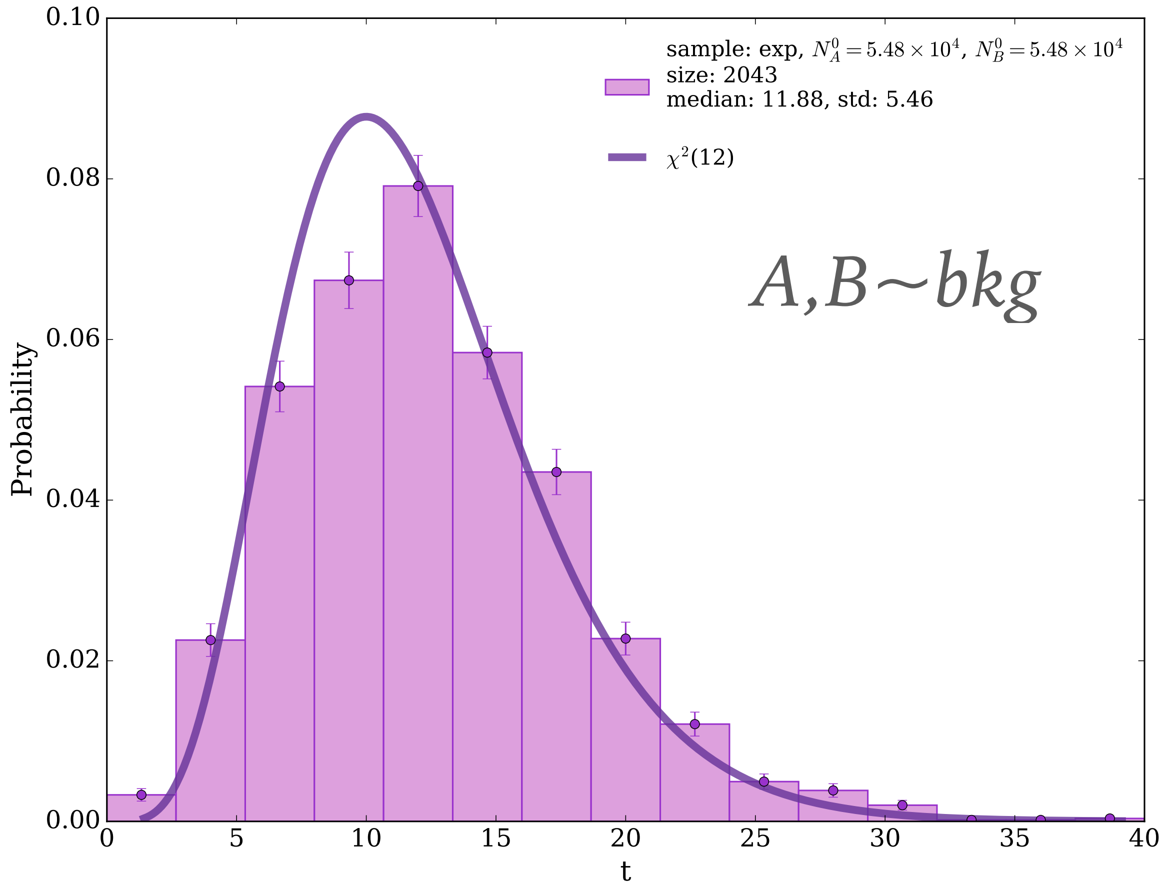


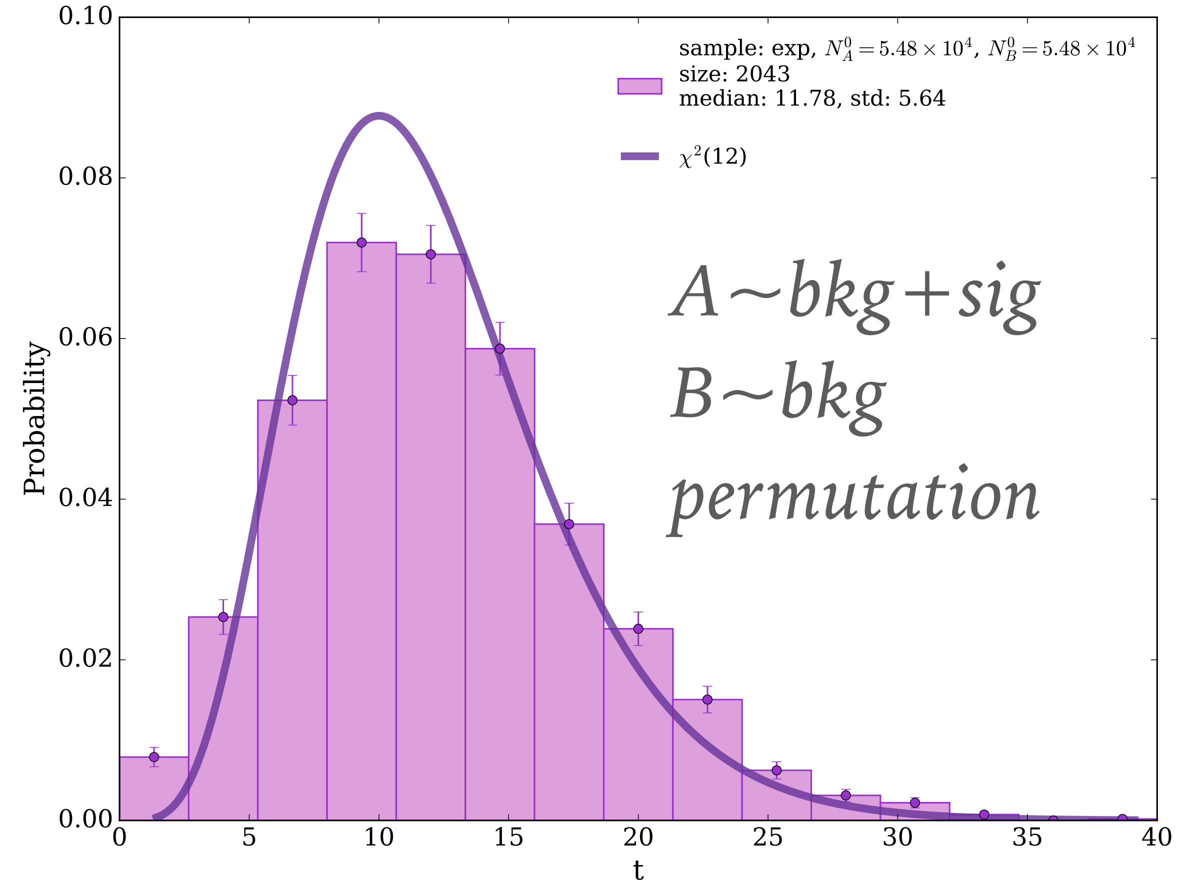
t distribution for toy data A and B generated from the same PDF

No divergences

RESULTS - EMPIRIC SYMMETRIC DISTRIBUTION

- Narrower and <u>predictable</u> background-only distribution.
  - Good agreement with asymptotic χ^2_n
  - Can generate empiric distribution from permutations of **<u>observed</u> A** and **B**



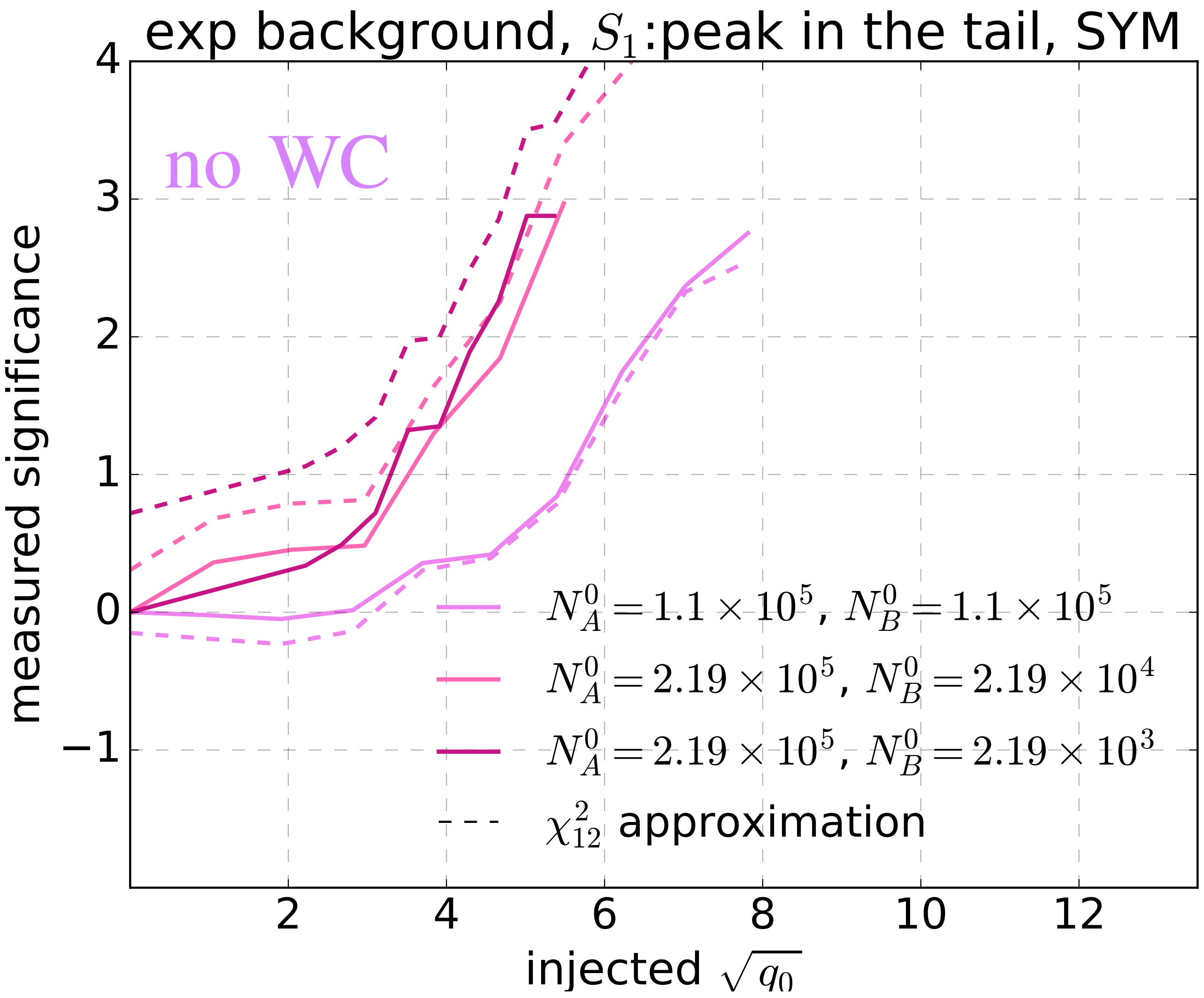


Symmetry assumptions only, no need for detailed SM simulation!

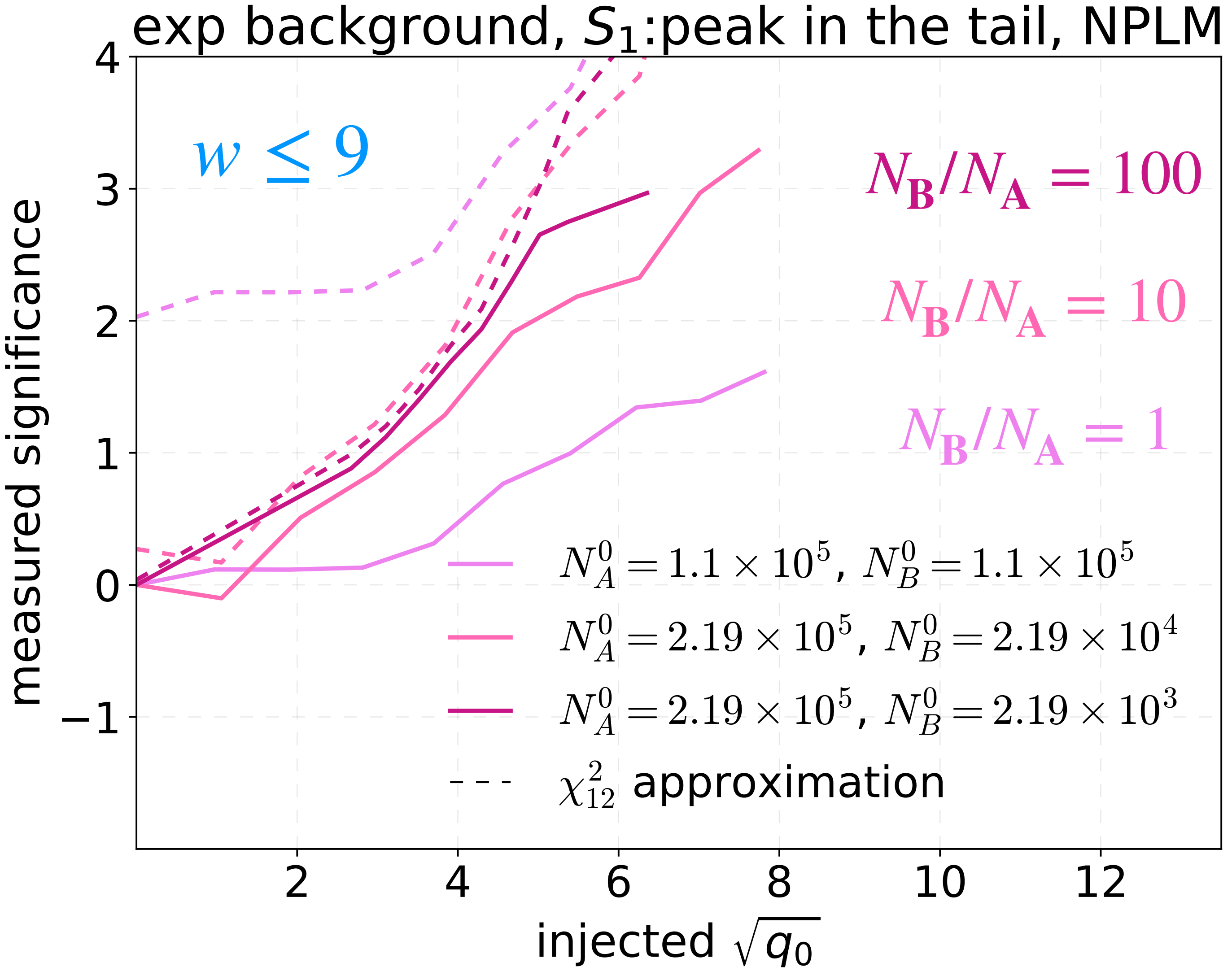
RESULTS - THE ASYMMETRIC CASE

➤ **Better sensitivity due to narrower background-only distribution and relaxed weight-clipping.**

$t_{\mathbf{A+B}}(\mathbf{A}) + t_{\mathbf{A+B}}(\mathbf{B})$



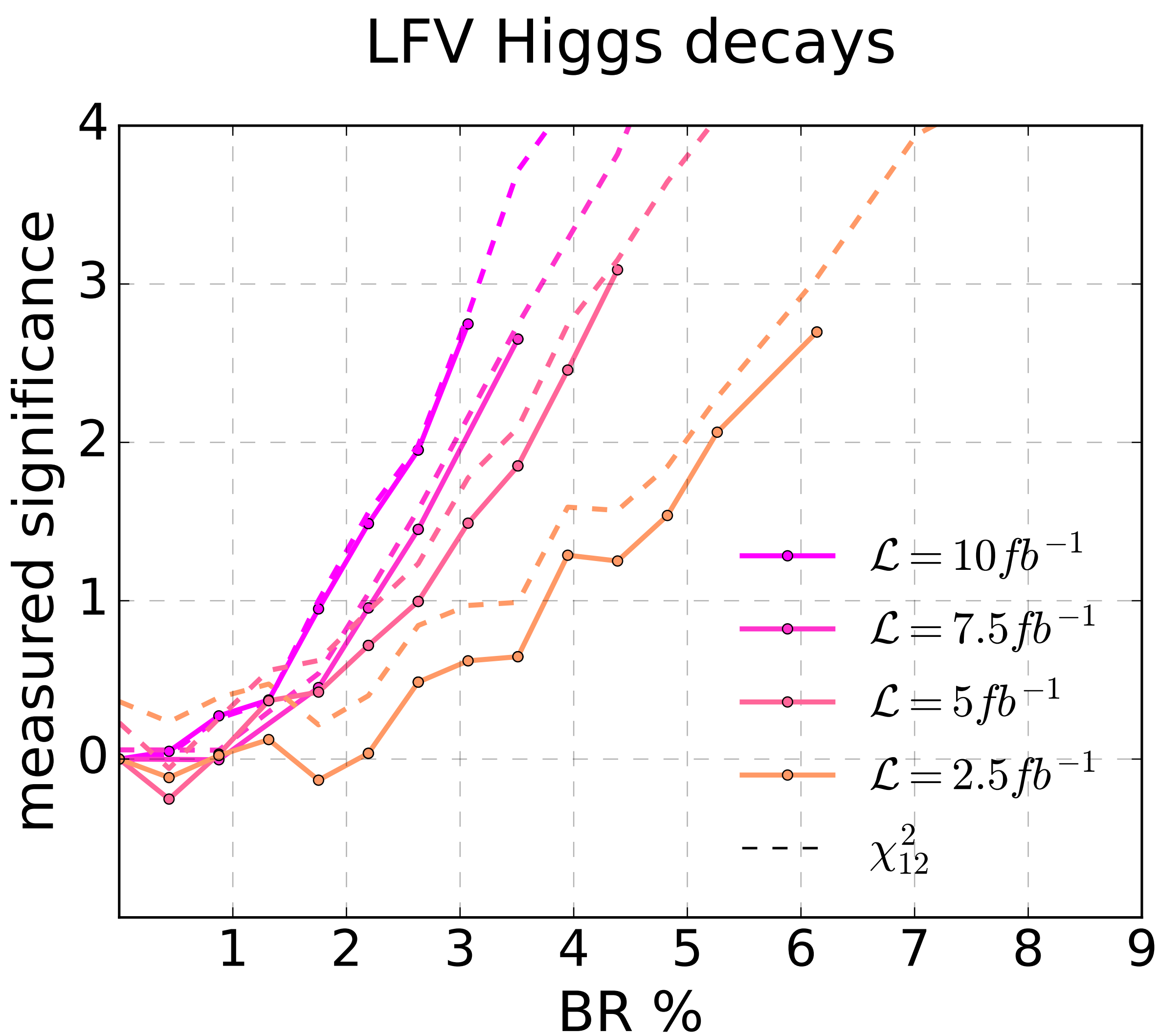
$t_{\mathbf{B}}(\mathbf{A})$



RESULTS - THE ASYMMETRIC CASE

- **Sensitivity to** $BR\,(H \to \tau e) \sim 2.6\,\%$ **at** $L = 10\,\mathrm{fb}^{-1}$
- Current bound: $BR\,(H \to \tau e) \sim 0.2\,\%$ at $L = 138\,\mathrm{fb}^{-1}$

ATLAS, [2302.05225], CMS , [2105.03007]



CONCLUSIONS

- SM symmetries can be exploited for model-agnostic NP searches that are fully data-based.
- NPLM: ML+likelihood-loss test for deviations of observed data from much larger reference dataset.
- **The symmetrized formalism** -
  - **Symmetric statistical test** to account for fluctuations in both samples.
  - **Symmetric reference distribution** - assigning non-zero probability to all observed events.
- **Allows for searches for differences between samples of arbitrary sizes, and relaxing the tuning of the model parameters.**

OPEN QUESTIONS AND FUTURE WORK

- Realistic analysis -
  - Including more kinematical variables.

    NPLM: R. T. D'Agnolo, G. Grosso, M. Pierini, A. Wulzer & M. Zanetti [1912.12155].
  - Including systematic uncertainties.

    NPLM: R. T. D'Agnolo, G. Grosso, M. Pierini, A. Wulzer & M. Zanetti [2111.13633].
  - Imperfect symmetries - correct for detector effects that violate LFU.

    M. Birman, S. Bressler et al, ATLAS [1604.07730].
- Hyper-parameters choice -
  - Improve robustness by preventing overfitting.

THANK YOU!

OPEN QUESTIONS AND FUTURE WORK

OPEN QUESTIONS AND FUTURE WORK

- Symmetry corrections - electrons and muons are not identical
  - Different efficiencies
  - Different fake rates

$$n_{e,\mu} = n^{\text{truth}}_{e,\mu}\varepsilon_{e,\mu} + f_{e,\mu}$$

- Previous works on restoring the symmetry by re-weighting

$$R^{\epsilon} = \frac{N_{\mu}}{N_e}$$

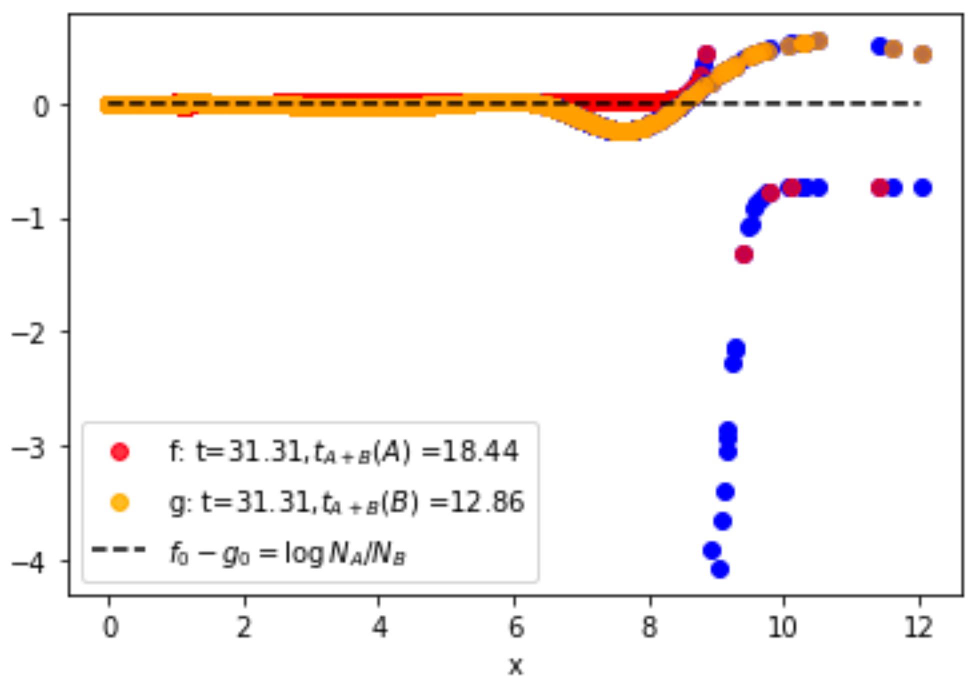
M. Birman, S.Bressler et al, ATLAS[1604.07730]

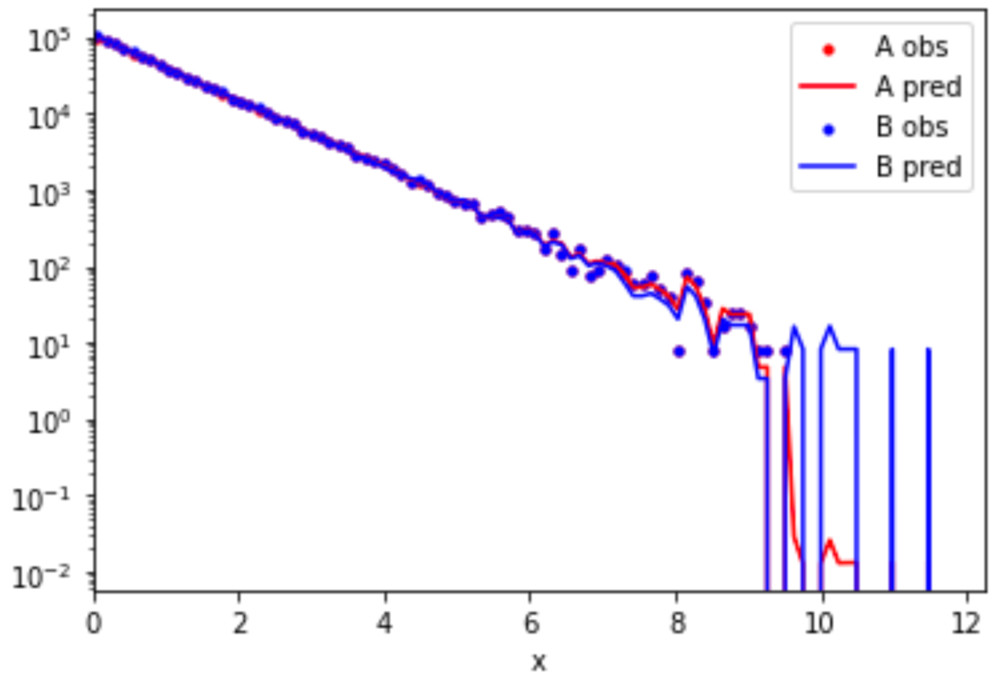
- Can use a similar approach to the symmetrized test?

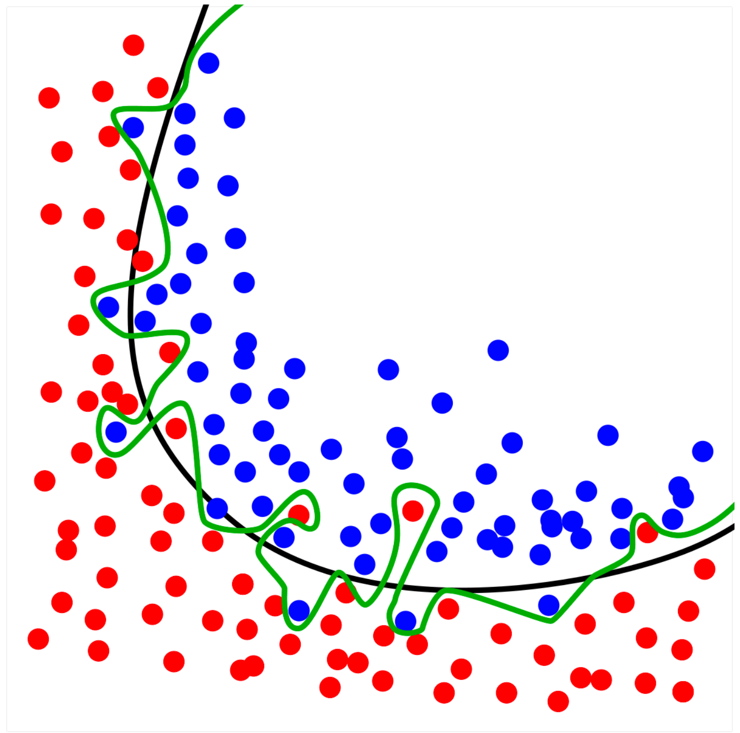
$$t = 2\log\left(\frac{\max_{p_e,p_\mu,p_{SM},\varepsilon_e,\varepsilon_\mu,f_e,f_\mu}\left(\mathscr{L}\left(n_e\varepsilon_e + f_e \,|\, \mathbf{e}_{\text{obs}}\right)\mathscr{L}\left(n_\mu\varepsilon_\mu + f_\mu \,|\, \mu_{\text{obs}}\right)\mathscr{L}\left(n^{SM}\varepsilon_e + f_e \,|\, \mathbf{e}_{\text{sim}}\right)\mathscr{L}\left(n^{SM}\varepsilon_\mu + f_\mu \,|\, \mu_{\text{sim}}\right)\right)}{\max_{p_0,\varepsilon_e,\varepsilon_\mu,f_e,f_\mu}\left(\mathscr{L}\left(n^{SM}\varepsilon_e + f_e \,|\, \mathbf{e}_{\text{obs}}\right)\mathscr{L}\left(n^{SM}\varepsilon_\mu + f_\mu \,|\, \mu_{\text{obs}}\right)\mathscr{L}\left(n^{SM}\varepsilon_e + f_e \,|\, \mathbf{e}_{\text{sim}}\right)\mathscr{L}\left(n^{SM}\varepsilon_\mu + f_\mu \,|\, \mu_{\text{sim}}\right)\right)}\right)$$

OPEN QUESTIONS AND FUTURE WORK

- Overfitting - too flexible functions are also able to perfectly fit statistical fluctuations.

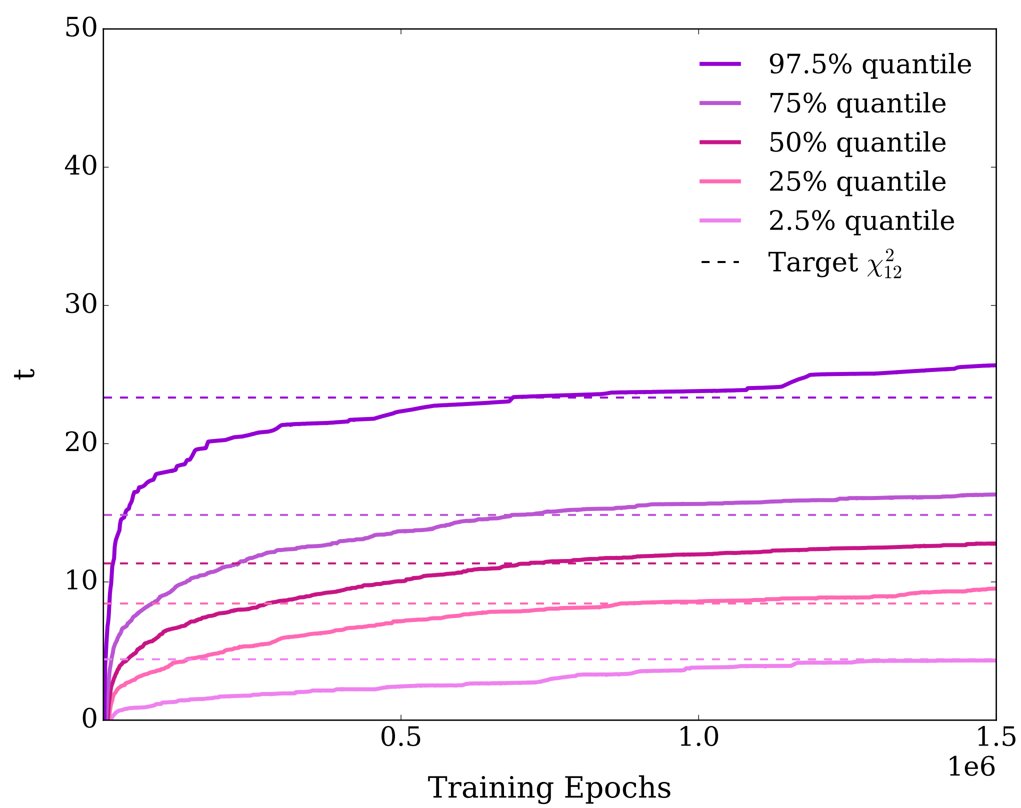


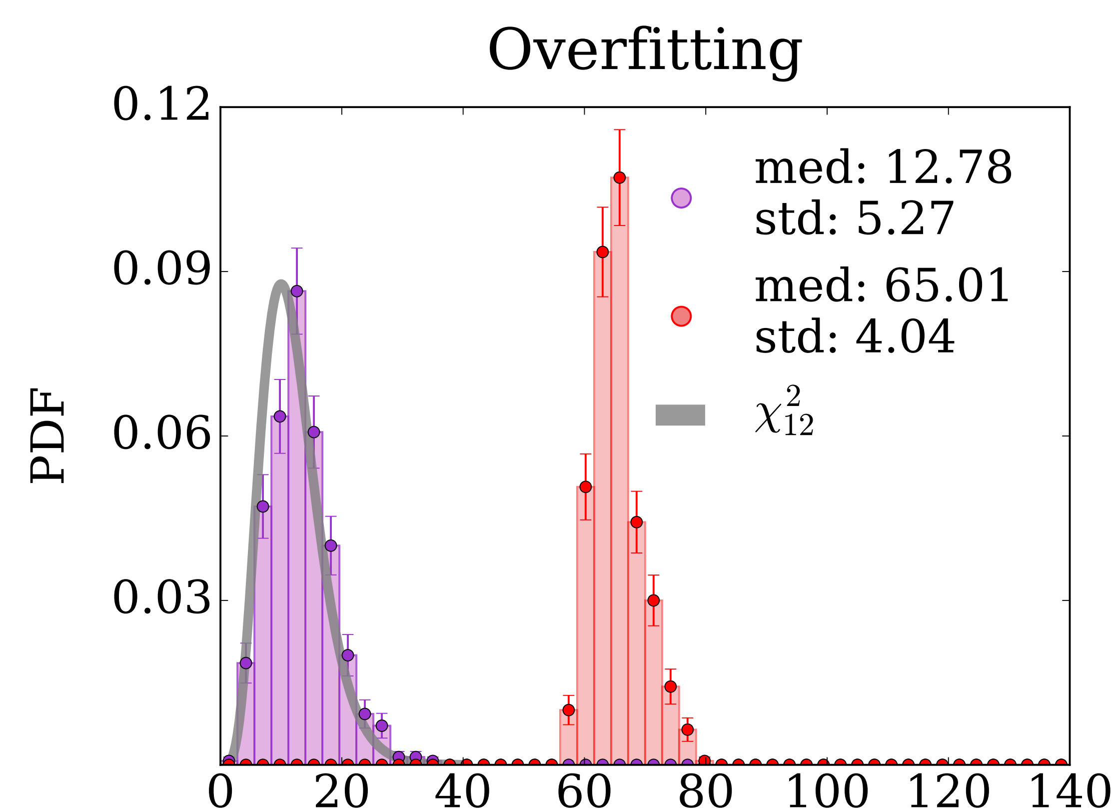




OPEN QUESTIONS AND FUTURE WORK

- Overfitting - too flexible functions are also able to perfectly fit statistical fluctuations.
- Distribution drifts away from the asymptotic χ^2 for a large number of epochs
- "Slightly" overfit solutions could be severe - locating longest runs





OPEN QUESTIONS AND FUTURE WORK

- Overfitting - potential solutions:
- Different fitting schemes -
  - Smooth functions + averaging.
  - Fit symmetric and asymmetric components instead of A and B.
- Standard ML regularization techniques -
  - Validation set - should understand resulting distribution.
  - Adding a cost term to the loss penalizing high weights/ complex models.
  - Understand relation between overfitting and a normal distribution of the parameters under the null hypothesis.