UN/SEMI-SUPERVISED LEARNING OF THE SOURCE PROPERTIES IN THE FERMI CATALOGS

Germán Gómez-Vargas
FONDECYT Postdoctoral Fellow at
Pontificia Universidad Católica de Chile

Roberto Muñoz
Metric Computer Vision

DarkMachines

June 19, 2018

DATA: THE FERMI LAT

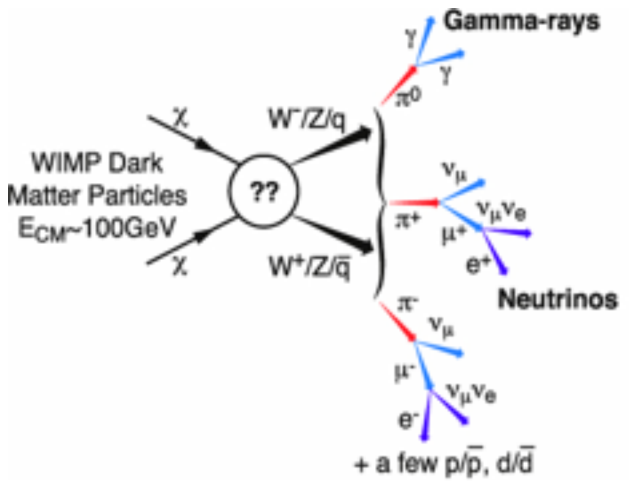
https://fermi.gsfc.nasa.gov/fermi10/

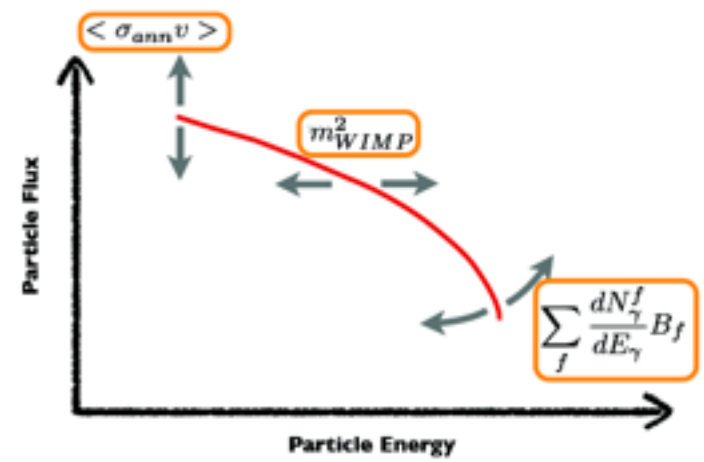
- Onboard the Fermi Gamma-ray Space Telescope. Launched June 11, 2008

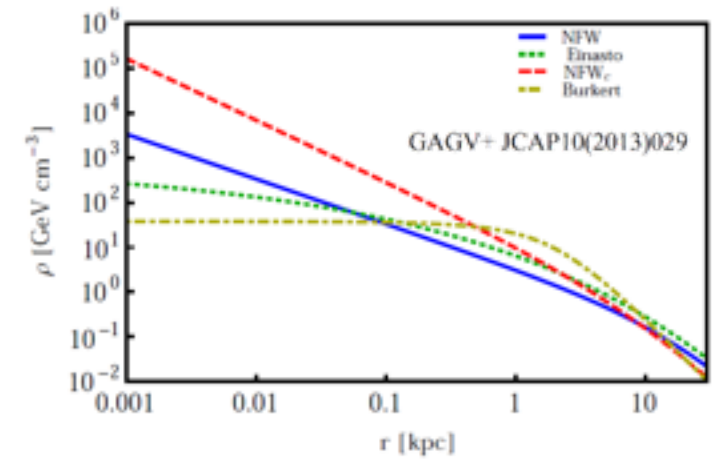
- The Fermi LAT collects high energy gamma rays (~20 MeV to > 300 GeV) with a large effective area (~6200 cm2) and a large field of view (2.4 sr)

CELEBRATING 10 YEARS

GAMMA RAYS FROM DARK MATTER







$$\left(\frac{d\Phi_\gamma}{dE_\gamma}\right)_{prompt} = \boxed{\sum_i \frac{dN^i_\gamma}{dE_\gamma} \frac{\langle \sigma_i v\rangle}{8\pi m^2_{DM}}} \boxed{\bar{J}(\Delta\Omega)\Delta\Omega}$$

Particle Physics
Spectral information

DM Distribution
(J-Factor)
Spatial information

WIMP = Weakly Interacting Massive Particle

– DM candidate (e.g. neutralino)

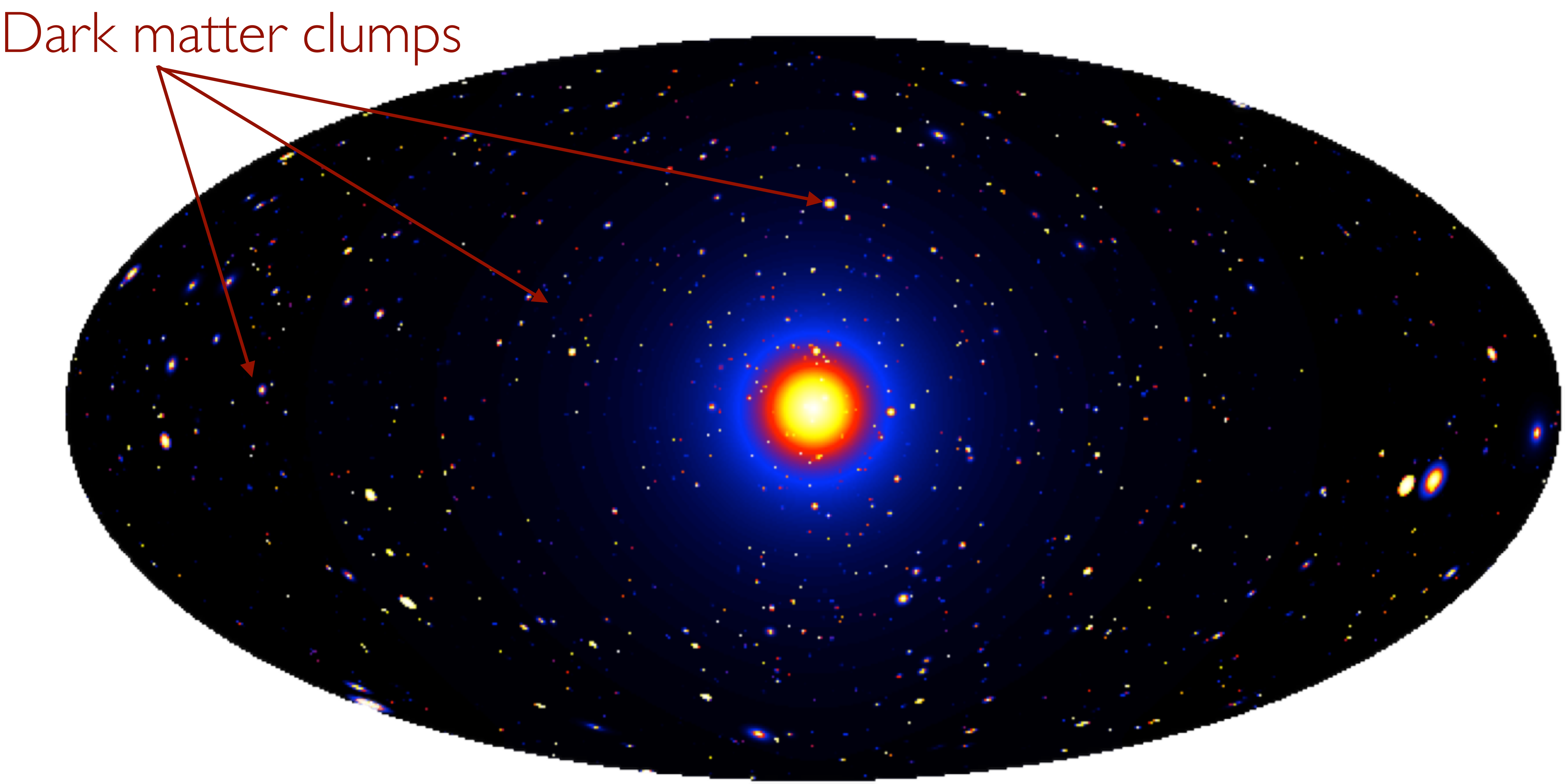
– Believe the Milky Way sits in a large spherical "halo" or cloud of DM

- Non-relativistic (cold) DM

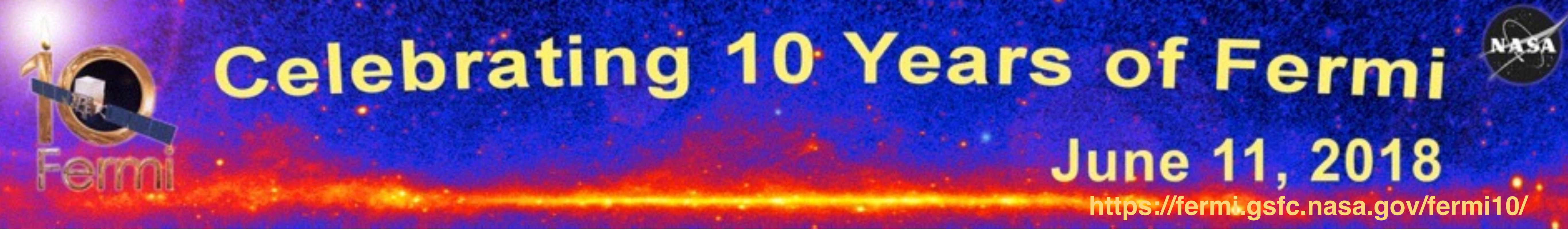
$$\bar{J}(\Delta\Omega) \equiv \frac{1}{\Delta\Omega}\int d\Omega \int_{l.o.s.} \rho^2(r(l,\Psi))\, dl$$

Annihilation

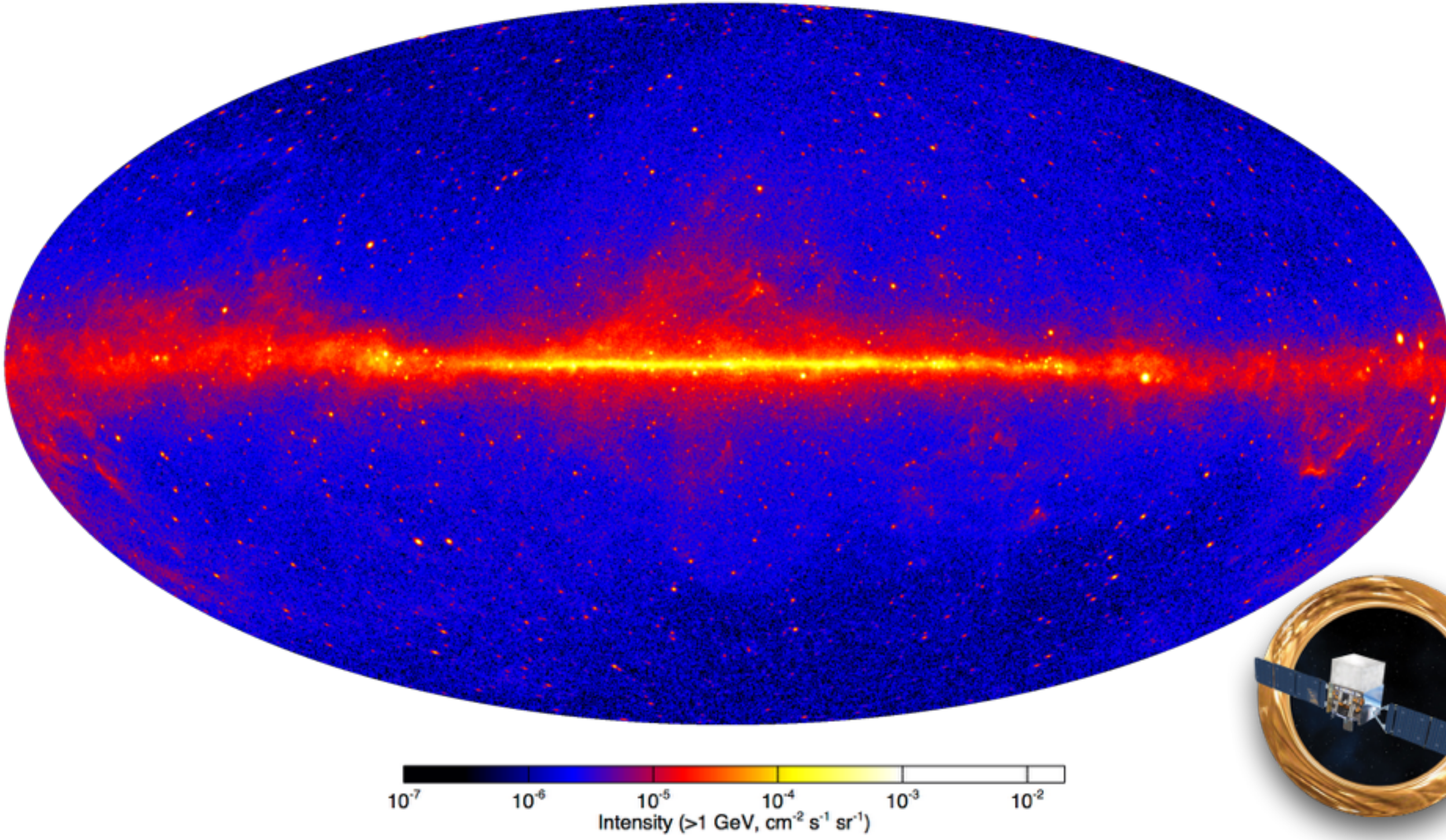
PREDICTION OF ALL-SKY DM-INDUCED GAMMA-RAY EMISSION



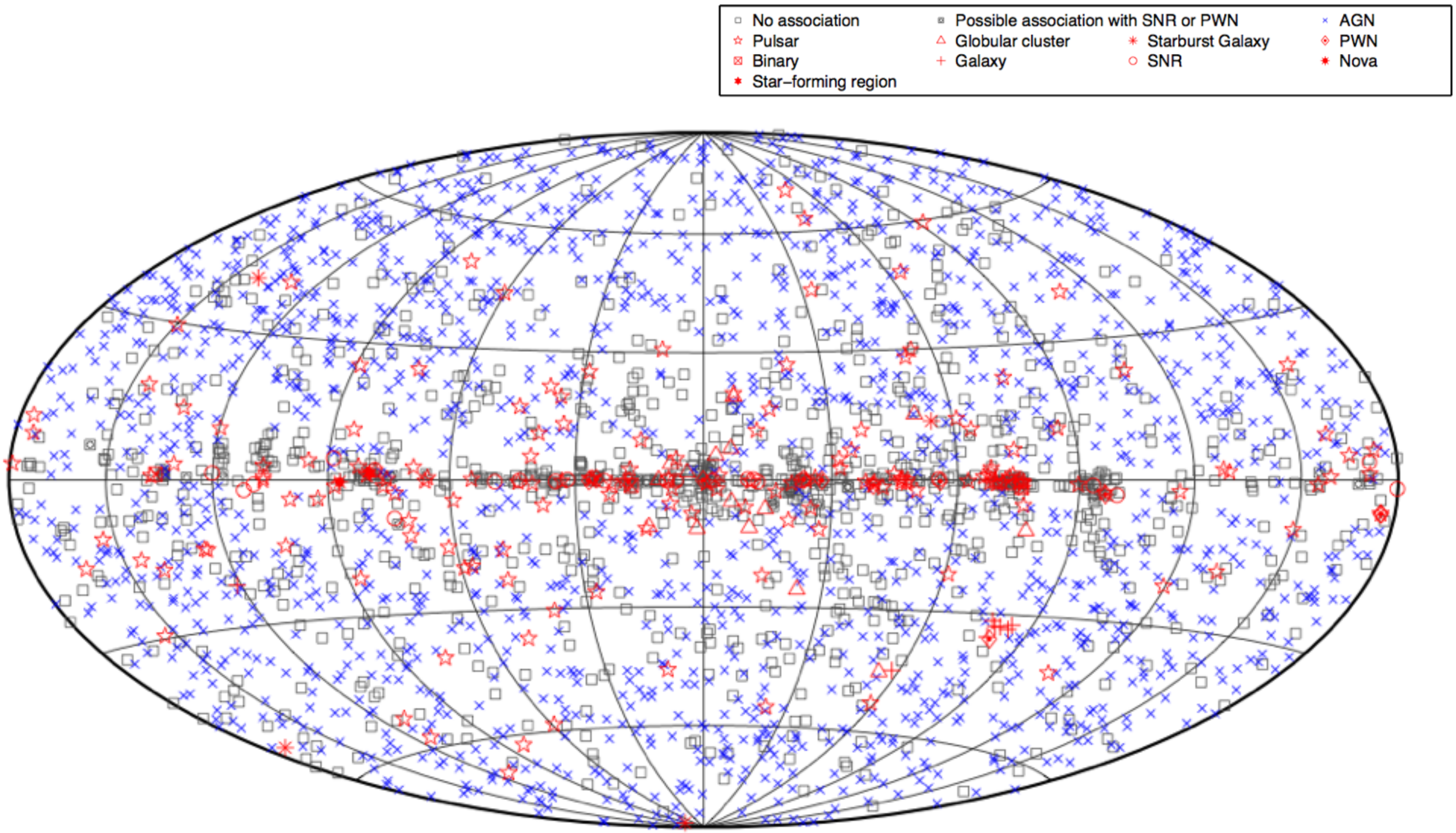
Simulation by Pieri et.al. Phys.Rev. D83 (2011) 023518



Fermi-LAT, P8data, 9 years, energies > 1 GeV

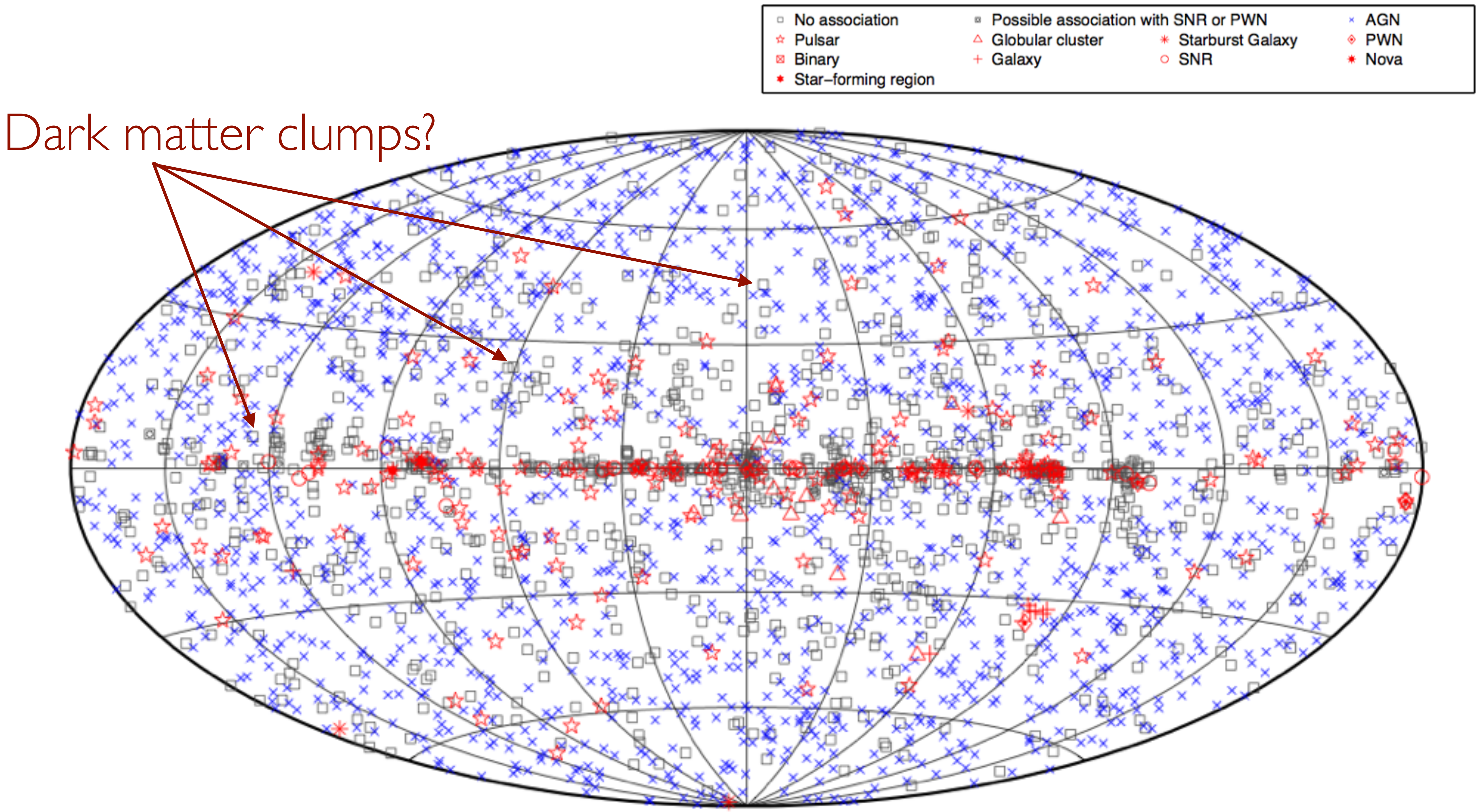


3FGL catalog of point sources



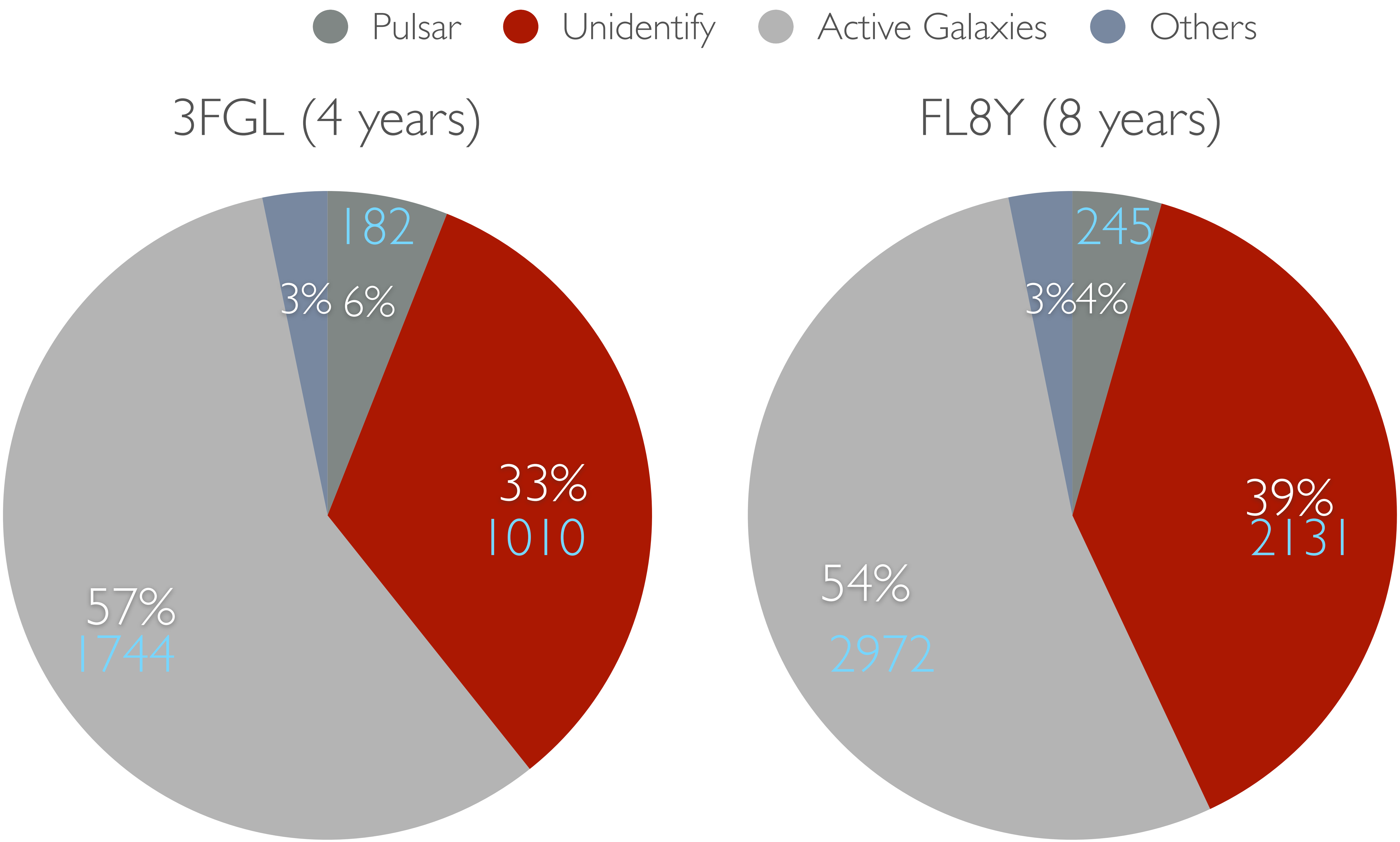
Fermi-LAT 4 years Source Catalog extracted from arXiv:1501.02003

3FGL catalog of point sources



Fermi-LAT 4 years Source Catalog extracted from arXiv:1501.02003

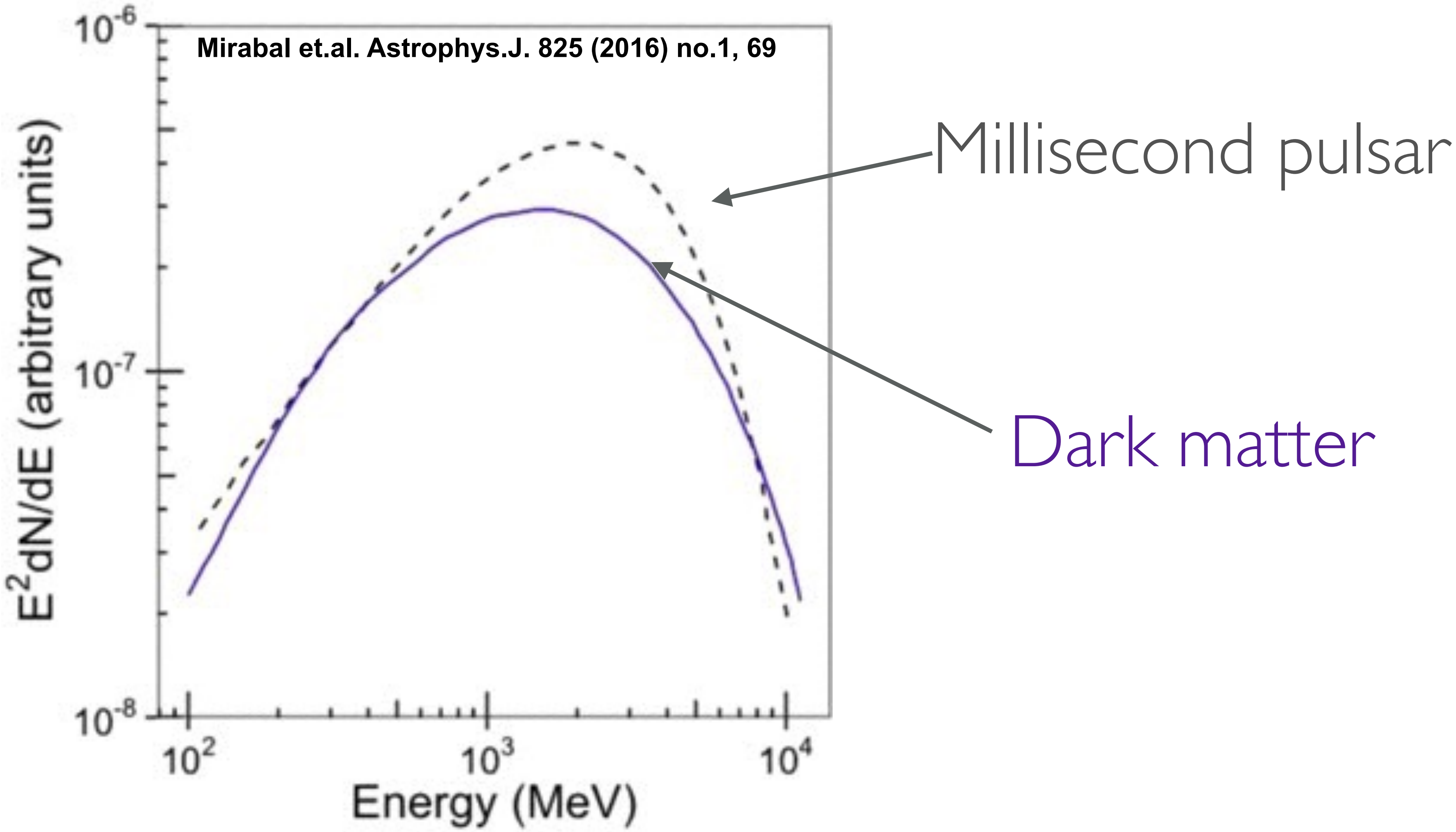
Unidentify sources



https://fermi.gsfc.nasa.gov/ssc/data/access/lat/4yr_catalog/

https://fermi.gsfc.nasa.gov/ssc/data/access/lat/fl8y/

Dark matter vs. pulsars



- Spectral shape from dark matter annihilation (purple line) of a 30 GeV particle into bottom quark pairs (Fornengo et al. 2004). Other Standard Model annihilation channels are expected to produce similar spectra. Also shown is a representative millisecond pulsar spectrum (black dashed line) from the second Fermi pulsar catalog (Abdo et al. 2013).

Methods to match DM clumps with UND sources I

N-body simulation

↓

Fermi-LAT performance

↓

Predict detectability/distribution of DM clumps

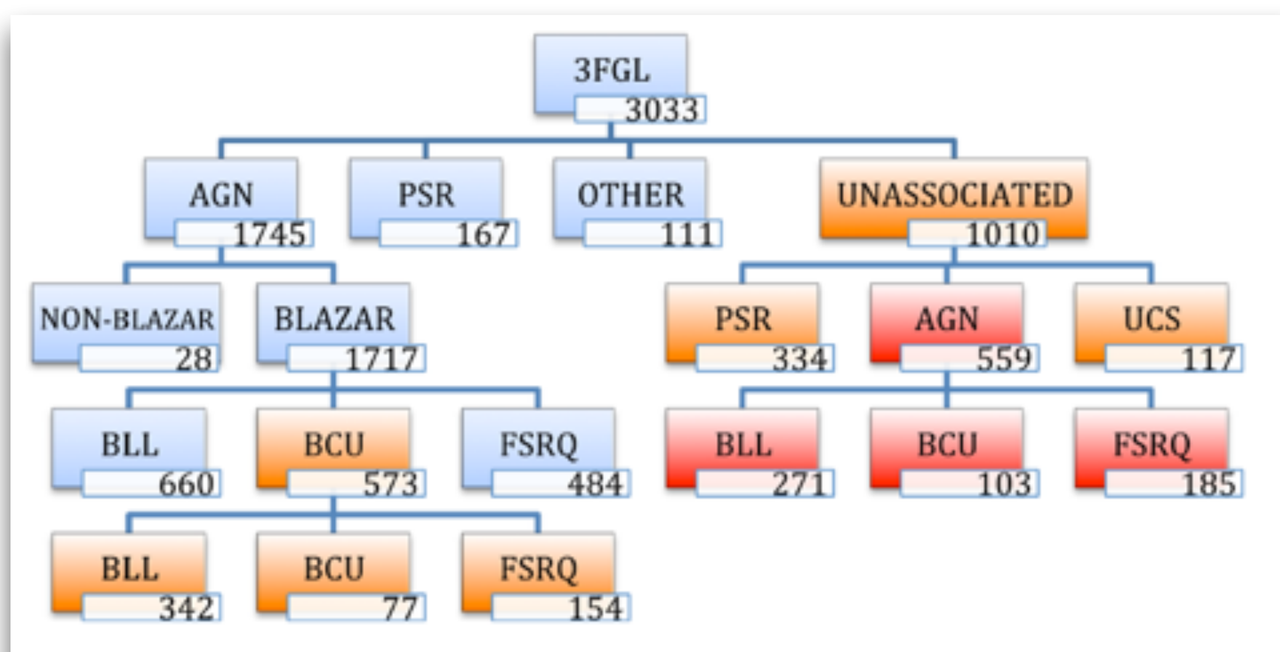
↓

Compare to Fermi-LAT catalogs
How many unidentify sources match my DM predictions?

- Some recent papers using this method: Schoonenberg et al JCAP05(2016)028, Hooper and Witte JCAP04(2017)018, Calore et.al. Phys. Rev. D 96, 063009 (2017)

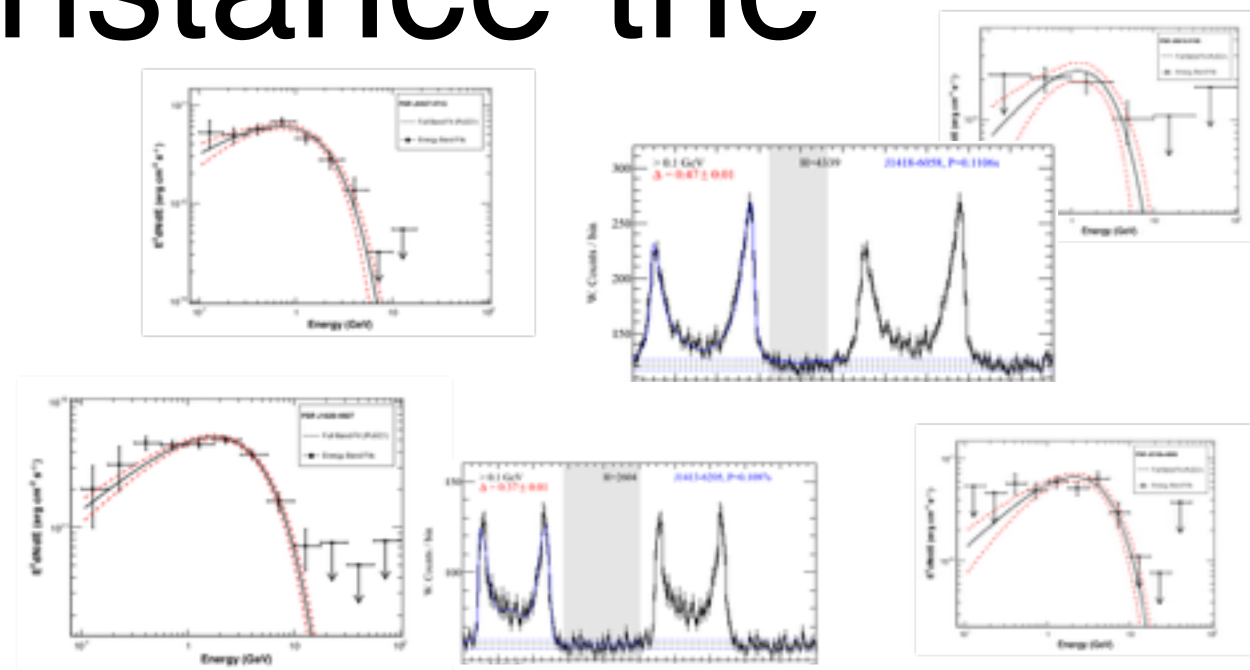
Methods to match DM clumps with UND sources II

- To use **supervised** machine learning methods to label unidentify sources in the catalog.
- Some recent papers: Mirabal et.al. Astrophys.J. 825 (2016) no.1, 69, Salveti et.al, MNRAS 470 (2017) no.2, 1291-1297, SazParkison et.al. Astrophys.J. 820 (2016) no.1, 8, Chiaro et.al. MNRAS 462 (2016) no.3, 3180-3195.
- The only one dealing with dark matter vs pulsars is Mirabal et.al 2016. The other works focus on classify the unidentify sources.
- They use "classical methods" including Decision Trees, Support Vector Machines, a Logistic Regression (LR) model, various modified versions of LR (e.g., Boosted LR, logistic decision trees) , RF, as well as some combination of methods (e.g., a two-step method involving decision trees followed by LR). In multi-wavelength observations of tagged UNS sources as pulsars some have been confirmed Salveti et.al. MNRAS 470 (2017) no.1, 466-480

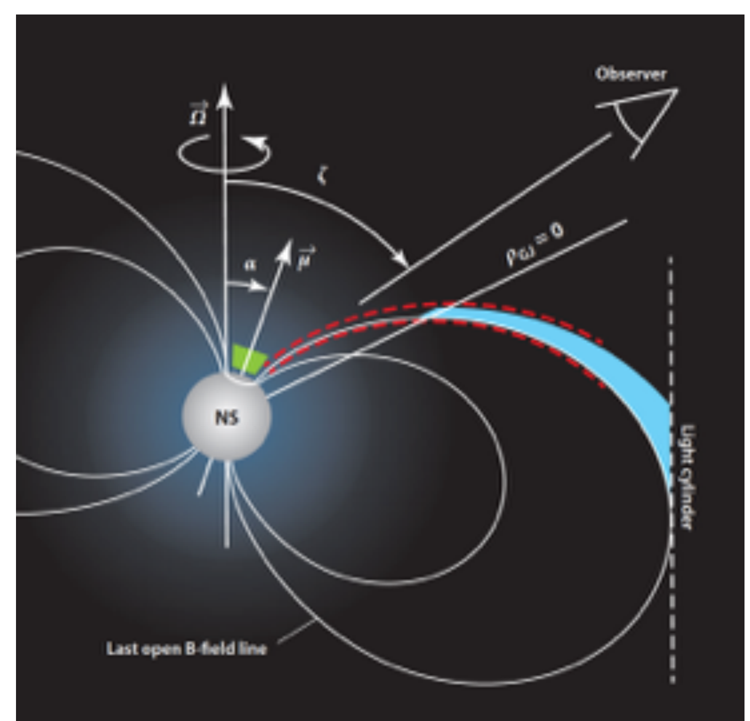


Orange: source classification provided by Chiaro et al. (2016) and Saz Parkinson et al. (2016).
Red: Salveti et.al. (2017) classification of unassociated sources classified as likely AGN.
Here UCS are unassociated sources that are not classified
as PSR or AGN candidates.
3FGLzoo from Salveti et.al. 2017

Human bias, for instance the pulsar spectra



- Pulsars detected in gamma rays, counterpart in radio -> timing information.
- Light curves and spectral shapes of a bunch of pulsars
- Pulsar mission models tweaked to reproduce data
- Some UND sources may be pulsars without energy cutoff



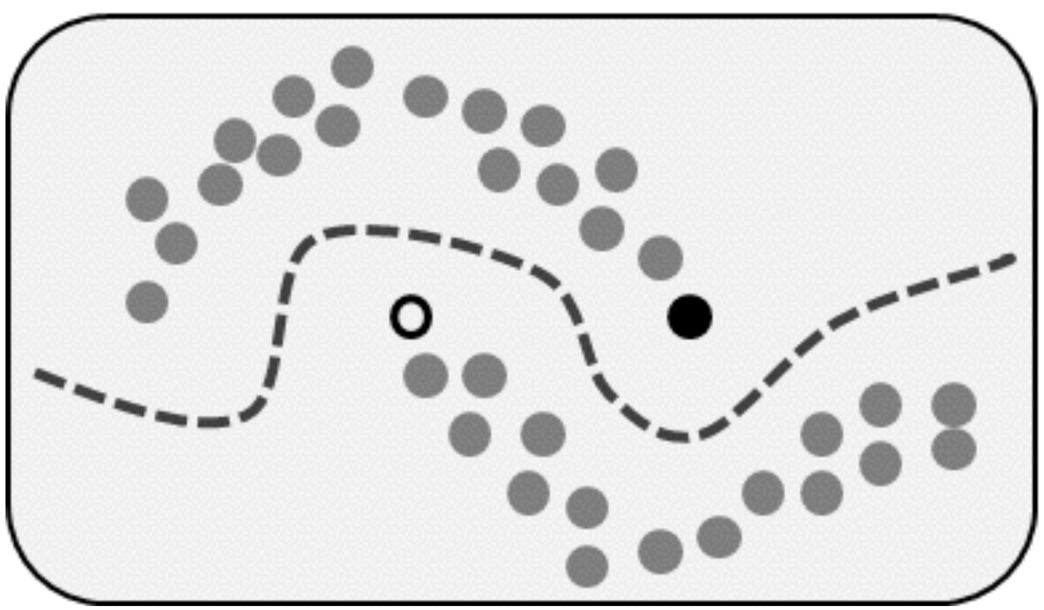
Scheme of a neutron star magnetosphere with the internal emission regions highlighted: polar cap model is in green, outer gap model in dark blue, and slot gap model in red. Figure from Caraveo, 2014

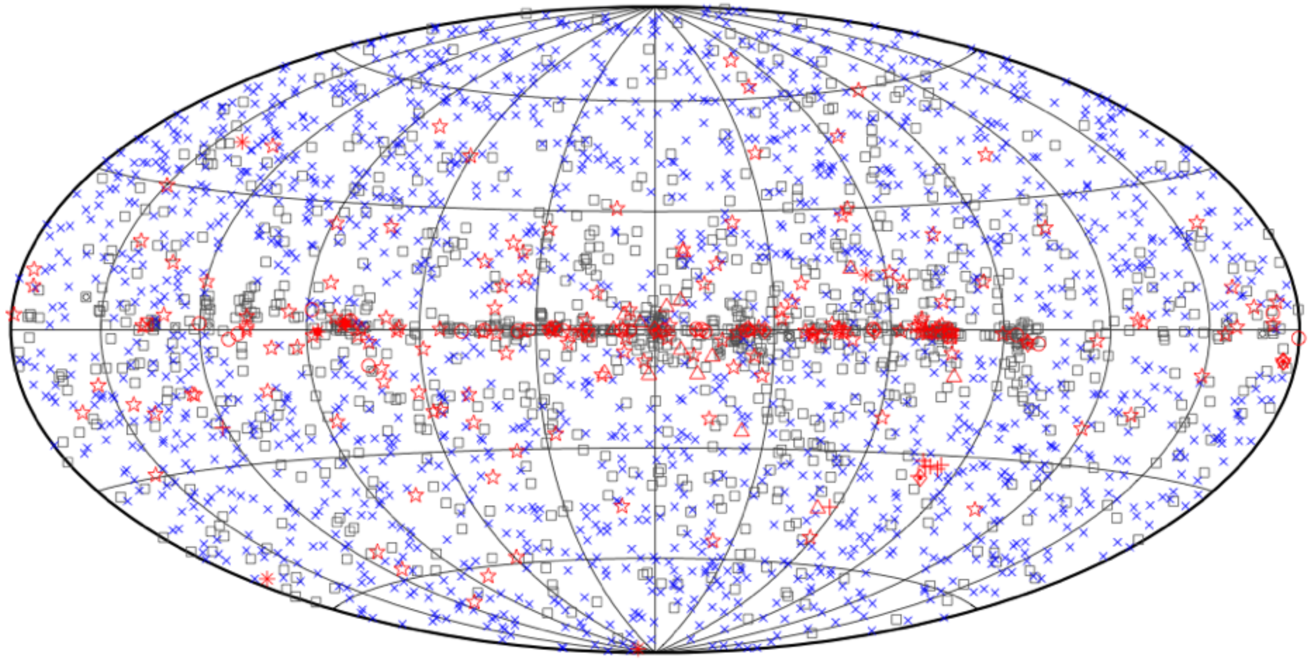
Data Challenge

- The way we put labels to objects is based on modelization of nature that can be wrong.
- Let’s think a machine taking the data and creating categories, then we scientist working on making sense of the different categories.

Objective of the data challenge

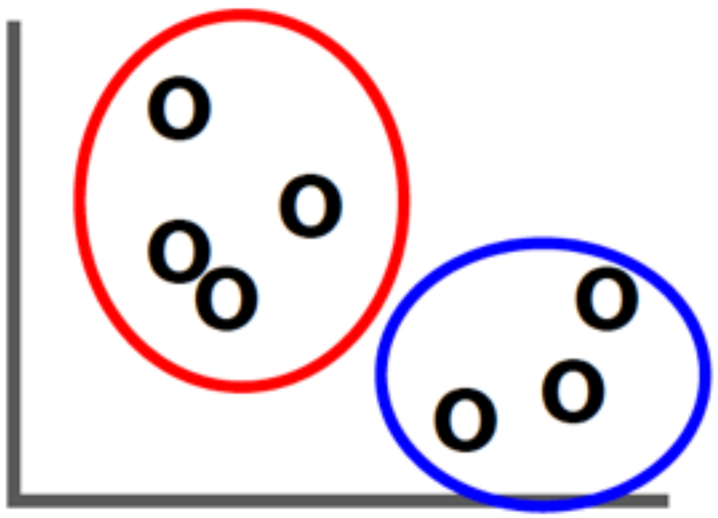
- Identify different groups of sources in the Fermi catalogs of point sources using semi-supervised or supervised learning



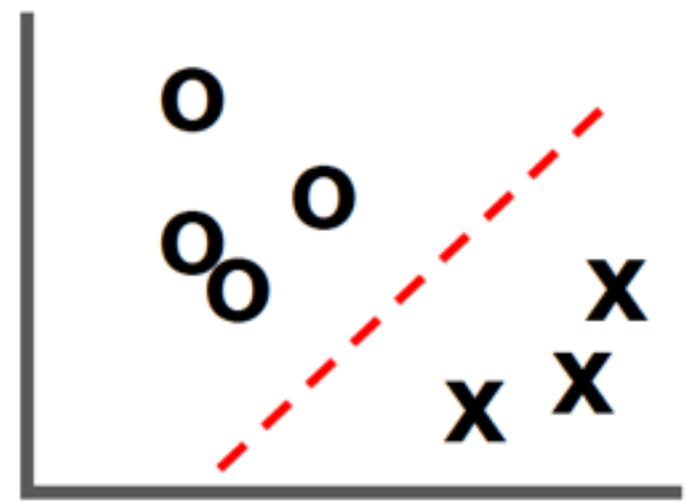


Semi-supervised Learning

Traditionally in ML, we have



Unsupervised Learning:
Learn structure from the **data**, e.g. density estimation, clustering



Supervised Learning:
Learn a mapping between the **data** and a **target**, e.g. regression, classification

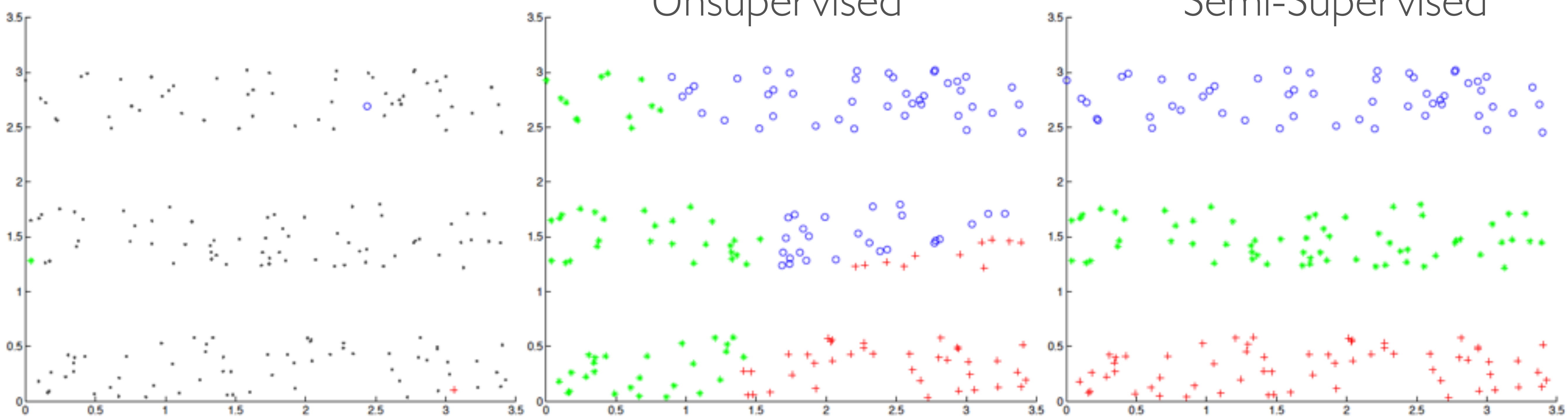
Semi-supervised Learning

- The goal is to use both labelled and unlabelled data to build better learners, than using each one alone.
- Is motivated by real world scenarios:
    - Abundant unlabelled data
    - High labelling costs

Xiaojin Zhu Semi-Supervised Learning Tutorial ICML 2007

Unsupervised

Semi-Supervised



Data

learning w/o labelled

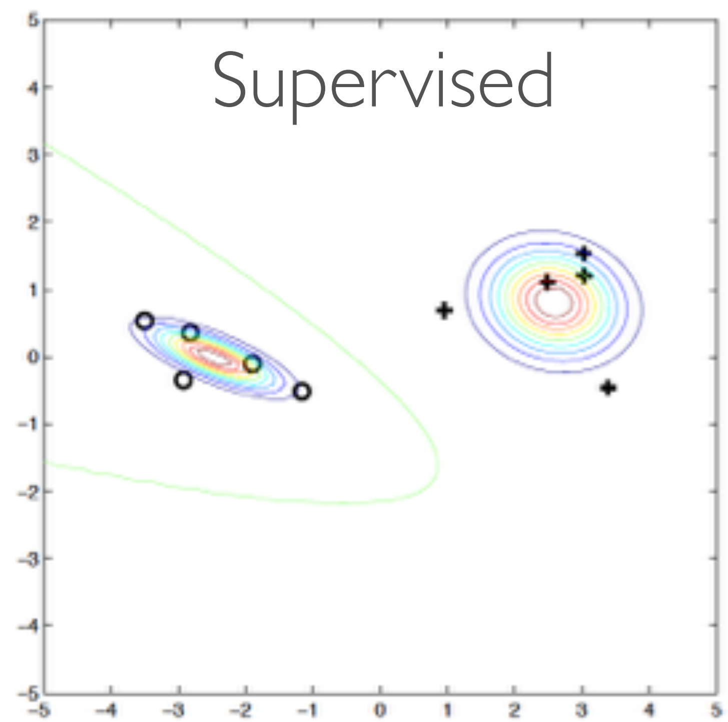
learning w labelled

Semi-supervised Learning

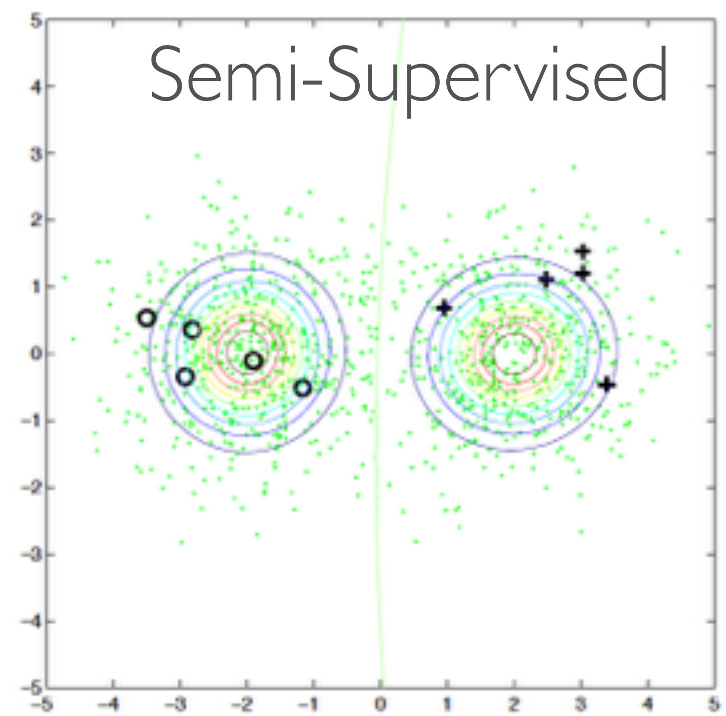
- The goal is to use both labelled and unlabelled data to build better learners, than using each one alone.
- Is motivated by real world scenarios:
  - Abundant unlabelled data
  - High labelling costs

Labelled data O and +

Unlabelled data •



learning w/o unlabelled

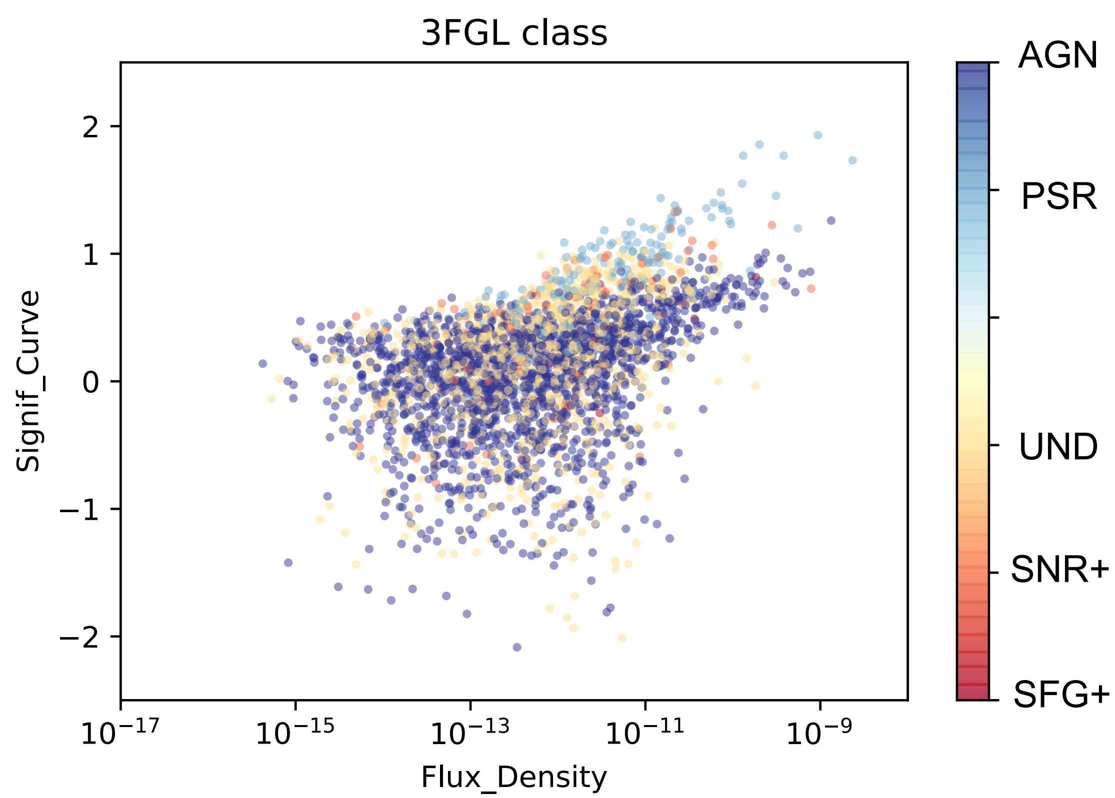


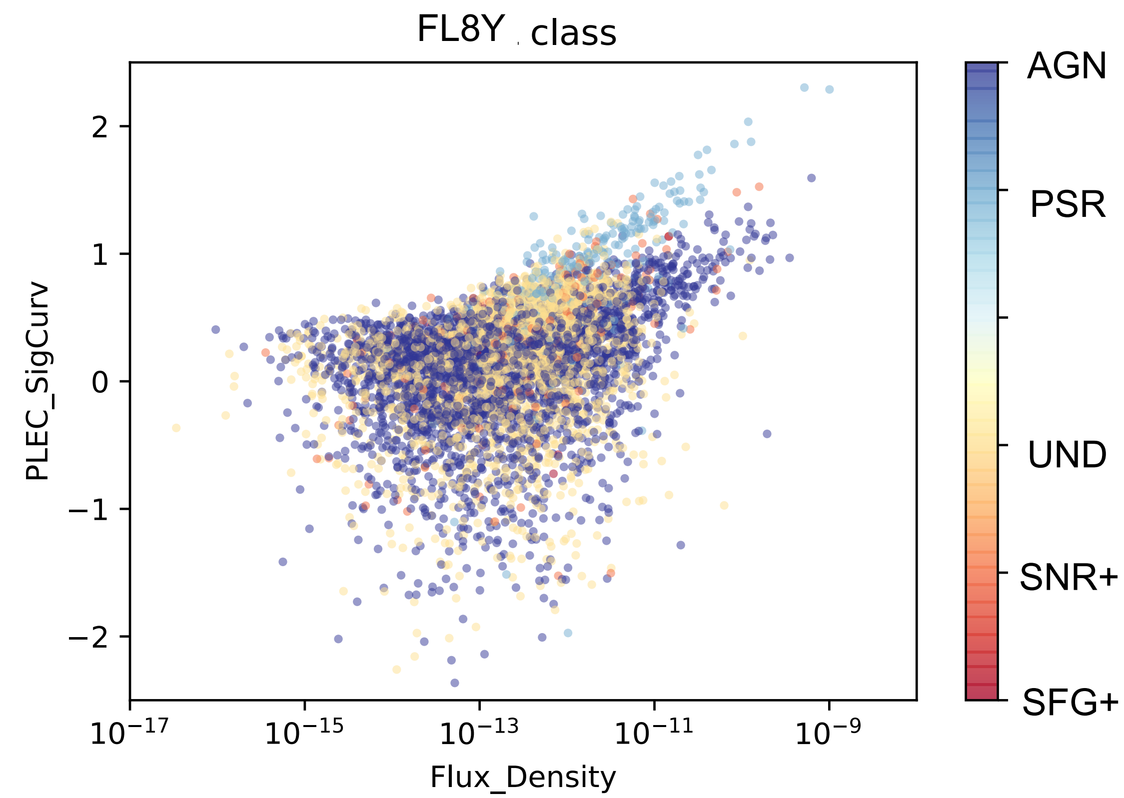
learning w unlabelled

Xiaojin Zhu Semi-Supervised Learning Tutorial ICML 2007

Semi-supervised learning on the 3FGL and FL8Y

http://scikit-learn.org/stable/modules/label_propagation.html

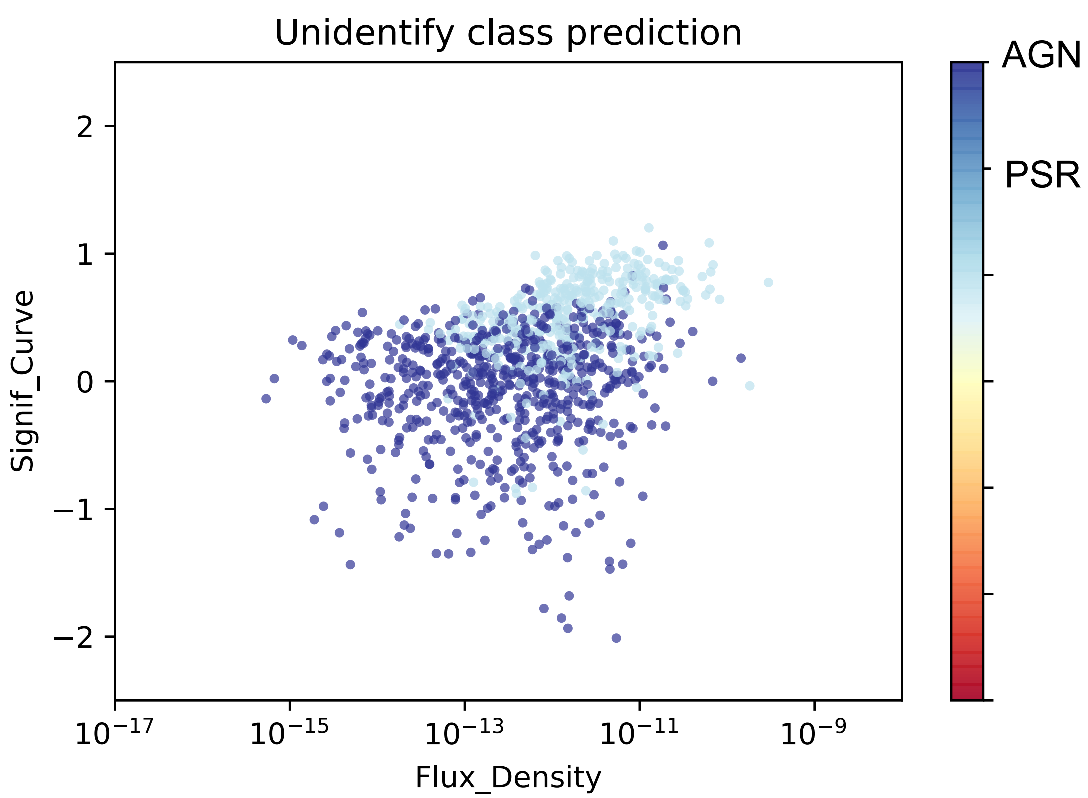


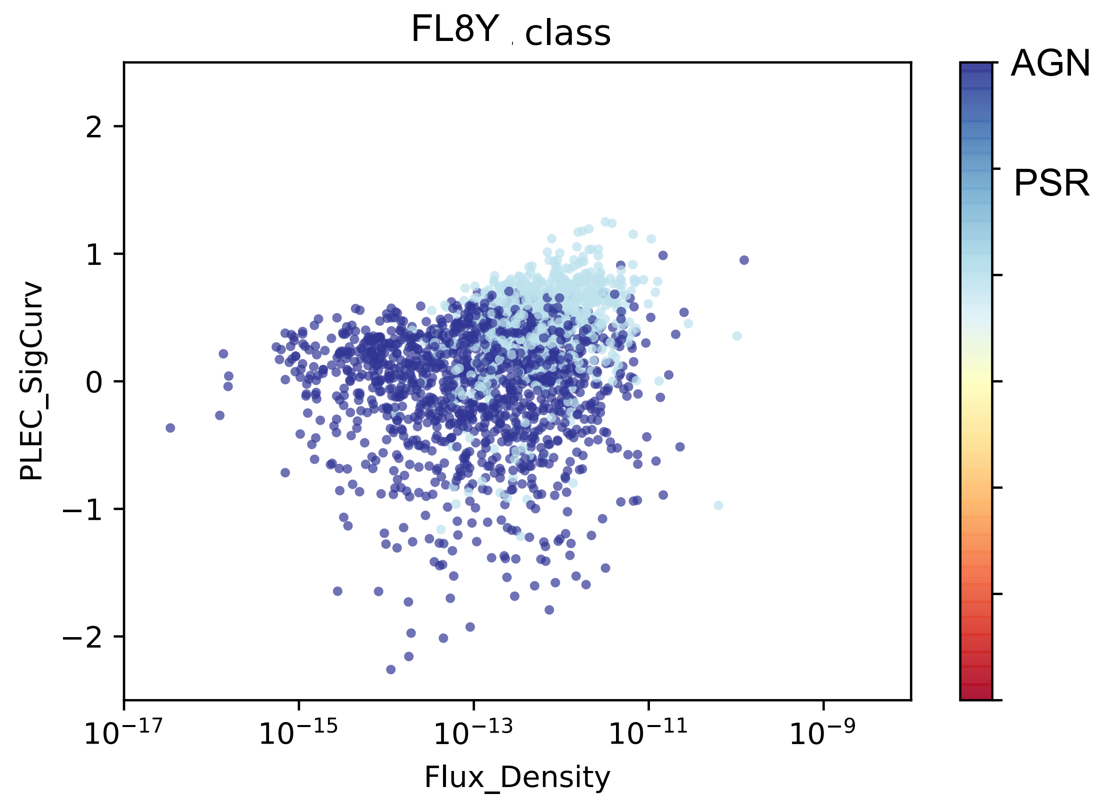


- We apply a method called “Label propagation” to the Fermi Catalog using many of the columns (features) to propagate PSR and AGN labels to the UND sources.
- For 3FGL same features as in Saz Parkinson 2016.
- For FL8Y all features in the catalog.

Semi-supervised learning on the 3FGL and FL8Y

http://scikit-learn.org/stable/modules/label_propagation.html





	Pulsar as pulsar	Overall accuracy
3FGL	95%	95.0%
FL8Y	99.5%	97.7%
3FGL Saz Parkinson 2016	96% (RF) 98% (LR)	96.7% (RF) 94.7% (LR)

Details data challenge

- Catalogs available here https://fermi.gsfc.nasa.gov/ssc/data/access/lat/
- When does the challenge/project start?
  - Starts in July
- How can I participate / who do I need to contact ?
  - Contact us: German and Roberto
- Are we working via a slack page?
  - Yes, we will set a slack page