Fifth Workshop on Millimetre Astronomy in Italy

Bologna, June 12-14 2023

UMLAUT: A Data-Driven Machine Learning Approach with Automated Parameter Tuning and Outlier Detection

Ivano Baronchelli

INAF - IRA
ARC – Italian node

The Astrophysical Journal Supplement Series, 249:12 (27pp), 2020 July https://doi.org/10.3847/1538-4365/ab9a3a

CrossMark

Identification of Single Spectral Lines through Supervised Machine Learning in a Large HST Survey (WISP): A Pilot Study for Euclid and WFIRST

I. Baronchelli[1], C. M. Scarlata[2], G. Rodighiero[1], L. Rodríguez-Muñoz[1], M. Bonato[3,4], M. Bagley[5], A. Henry[6], M. Rafelski[6,7], M. Malkan[8], J. Colbert[9], Y. S. Dai (戴昱)[10], H. Dickinson[2,11], C. Mancini[1], V. Mehta[2], L. Morselli[1], and H. I. Teplitz[9]

[1] Dipartimento di Fisica e Astronomia, Università di Padova, vicolo Osservatorio, 3, I-35122 Padova, Italy; ivano.baronchelli@unipd.it
[2] MN Institute for Astrophysics, University of Minnesota, 116 Church Street SE, Minneapolis, MN 55455, USA
[3] INAF-Istituto di Radioastronomia and Italian ALMA Regional Centre, Via Gobetti 101, I-40129, Bologna, Italy
[4] INAF-Osservatorio Astronomico di Padova, Vicolo dell'Osservatorio 5, I-35122, Padova, Italy
[5] College of Natural Sciences, The University of Texas at Austin, 2515 Speedway, Austin, TX 78712, USA
[6] Space Telescope Science Institute, 3700 San Martin Drive, Baltimore, MD 21218, USA
[7] Department of Physics and Astronomy, Johns Hopkins University, Baltimore, MD 21218, USA
[8] Department of Physics and Astronomy, UCLA, Physics and Astronomy Building, 3-714, LA CA 90095-1547, USA
[9] IPAC, Mail Code 314-6, Caltech, 1200 E. California Boulevard, Pasadena, CA 91125, USA
[10] Chinese Academy of Sciences South America Center for Astronomy (CASSACA)/NAOC, 20A Datun Road, Beijing 100101, People's Republic of China
[11] School of Physical Sciences, The Open University, Walton Hall, Milton Keynes, MK7 6AA, UK
Received 2019 November 28; revised 2020 May 25; accepted 2020 June 5; published 2020 July 13

Abstract

Future surveys focusing on understanding the nature of dark energy (e.g., Euclid and WFIRST) will cover large fractions of the extragalactic sky in near-IR slitless spectroscopy. These surveys will detect a large number of galaxies that will have only one emission line in the covered spectral range. In order to maximize the scientific return of these missions, it is imperative that single emission lines are correctly identified. Using a supervised machine-learning approach, we classified a sample of single emission lines extracted from the WFC3 IR Spectroscopic Parallel survey, one of the closest existing analogs to future slitless surveys. Our automatic software integrates a spectral energy distribution (SED)-fitting strategy with additional independent sources of information. We calibrated it and tested it on a "gold" sample of securely identified objects with multiple lines detected. The algorithm correctly classifies real emission lines with an accuracy of 82.6%, whereas the accuracy of the SED-fitting technique alone is low (~50%) due to the limited amount of photometric data available ($\leqslant$6 bands). While not specifically designed for the Euclid and WFIRST surveys, the algorithm represents an important precursor of similar algorithms to be used in these future missions.

The Astrophysical Journal Supplement Series, 249:12 (27pp), 2020 July
https://doi.org/10.3847/1538-4365/ab9a3a

CrossMark

Identification of Single Spectral Lines through Supervised Machine Learning in a Large HST Survey (WISP): A Pilot Study for Euclid and WFIRST

I. Baronchelli[1], C. M. Scarlata[2], G. Rodighiero[1], L. Rodríguez-Muñoz[1], M. Bonato[3,4], M. Bagley[5], A. Henry[6],
M. Rafelski[6,7], M

[1] Dipartimento di
[2] MN
[3] INAF
[5] Co
[8] Departm
[10] Chinese Academy of Scie

Future surveys f
fractions of the
galaxies that wil
return of these
machine-learning
Spectroscopic Pa
integrates a spec
We calibrated it
algorithm correc
fitting technique
not specifically
similar algorithm

The Astrophysical Journal Supplement Series, 257:67 (17pp), 2021 December
https://doi.org/10.3847/1538-4365/ac250c

CrossMark

Identification of Single Spectral Lines in Large Spectroscopic Surveys Using UMLAUT: an Unsupervised Machine-learning Algorithm Based on Unbiased Topology

I. Baronchelli[1,2], C. M. Scarlata[3], L. Rodríguez-Muñoz[2], M. Bonato[1,4], L. Morselli[1,2], M. Vaccari[5,6], R. Carraro[7], L. Barrufet[8], A. Henry[9], V. Mehta[3], G. Rodighiero[2], A. Baruffolo[1], M. Bagley[10], A. Battisti[11], J. Colbert[12], Y. S. Dai (戴昱)[13], M. De Pascale[1], H. Dickinson[14], M. Malkan[15], C. Mancini[2,16], M. Rafelski[9,17], and H. I. Teplitz[12]

[1] INAF-Osservatorio Astronomico di Padova, Vicolo dell'Osservatorio 5, I-35122, Padova, Italy; ivano.baronchelli@inaf.it
[2] Dipartimento di Fisica e Astronomia, Università di Padova, vicolo Osservatorio, 3, I-35122 Padova, Italy
[3] MN Institute for Astrophysics, University of Minnesota, 116 Church Street SE, Minneapolis, MN 55455, USA
[4] INAF-Istituto di Radioastronomia and Italian ALMA Regional Centre, Via Gobetti 101, I-40129, Bologna, Italy
[5] Inter-University Institute for Data Intensive Astronomy (IDIA) - Department of Physics & Astronomy, South Africa
[6] University of the Western Cape, Robert Sobukwe Road, 7535 Bellville, Cape Town, South Africa
[7] Instituto de Física y Astronomía, Universidad de Valparaíso, Gran Bretaña 1111, Playa Ancha, Valparaíso, Chile
[8] Université de Genève, Department of Astronomy, Chemin Pegasi, 51, 1290 Versoix, Switzerland
[9] Space Telescope Science Institute, 3700 San Martin Drive, Baltimore, MD, 21218, USA
[10] College of Natural Sciences, The University of Texas at Austin, 2515 Speedway, Austin, TX 78712, USA
[11] Research School of Astronomy and Astrophysics, Australian National University, Cotter Road, Weston Creek, ACT 2611, Australia
[12] IPAC, Mail Code 314-6, Caltech, 1200 E. California Boulevard, Pasadena, CA 91125, USA
[13] Chinese Academy of Sciences South America Center for Astronomy (CASSACA)/NAOC, 20A Datun Road, Beijing 100101, People's Republic of China
[14] School of Physical Sciences, The Open University, Walton Hall, Milton Keynes, MK7 6AA, UK
[15] Department of Physics and Astronomy, UCLA, Physics and Astronomy Bldg., 3-714, LA CA 90095-1547, USA
[16] INAF-IASF, Via Alfonso Corti 12 I-20133 Milano Italy
[17] Department of Physics and Astronomy, Johns Hopkins University, Baltimore, MD 21218, USA
Received 2020 December 30; revised 2021 August 3; accepted 2021 September 6; published 2021 December 10

Abstract

The identification of an emission line is unambiguous when multiple spectral features are clearly visible in the same spectrum. However, in many cases, only one line is detected, making it difficult to correctly determine the redshift. We developed a freely available unsupervised machine-learning algorithm based on unbiased topology (UMLAUT) that can be used in a very wide variety of contexts, including the identification of single emission lines. To this purpose, the algorithm combines different sources of information, such as the apparent magnitude, size and color of the emitting source, and the equivalent width and wavelength of the detected line. In each specific case, the algorithm automatically identifies the most relevant ones (i.e., those able to minimize the dispersion associated with the output parameter). The outputs can be easily integrated into different algorithms, allowing us to combine supervised and unsupervised techniques and increasing the overall accuracy. We tested our software on WISP (WFC3 IR Spectroscopic Parallel) survey data. WISP represents one of the closest existing analogs to the near-IR spectroscopic surveys that are going to be performed by the future Euclid and Roman missions. These missions will investigate the large-scale structure of the universe by surveying a large portion of the extragalactic sky in near-IR slitless spectroscopy, detecting a relevant fraction of single emission lines. In our tests, UMLAUT correctly identifies real lines in 83.2% of the cases. The accuracy is slightly higher (84.4%) when combining our unsupervised approach with a supervised approach we previously developed.


The Astrophysical Journal Supplement Series, 249:12 (27pp), 2020 July
https://doi.org/10.3847/1538-4365/ab9a3a
© 2020. The American Astronomical Society. All rights reserved.
CrossMark

Identification of Single Spectral Lines through Supervised Machine Learning in a Large HST Survey (WISP): A Pilot Study for Euclid and WFIRST

I. Baronchelli[1], C. M. Scarlata[2], G. Rodighiero[1], L. Rodríguez-Muñoz[1], M. Bonato[3,4], M. Bagley[5], A. Henry[6],
M. Rafelski[6,7], M

[1] Dipartimento di
[2] MN
[3] INAF
[5] Co
[8] Depart
[10] Chinese Academy of Scie

Future surveys f
fractions of the
galaxies that wil
return of these
machine-learning
Spectroscopic P
integrates a spec
We calibrated it
algorithm correc
fitting technique
not specifically
similar algorithm

The Astrophysical Journal Supplement Series, 257:67 (17pp), 2021 December
https://doi.org/10.3847/1538-4365/ac250c
© 2021. The American Astronomical Society. All rights reserved.
CrossMark

Identification of Single Spectral Lines in Large Spectroscopic Surveys Using UMLAUT: an Unsupervised Machine-learning Algorithm Based on Unbiased Topology

I. Baronchelli[1,2], C. M. Scarlata[3], L. Rodríguez-Muñoz[2], M. Bonato[1,4], L. Morselli[1,2], M. Vaccari[5,6], R. Carraro[7],
L. Barrufet[8], A. Henry[9], V. Mehta[3],
Y. S. Dai (戴昱)[13], M. De Pascale[1]

[1] INAF-Osservatorio Astronomico
[2] Dipartimento di Fisica e
[3] MN Institute for Astrophysi
[4] INAF-Istituto di Radioastron
[5] Inter-University Institute for D
[6] University of the Wes
[7] Instituto de Física y Astrono
[8] Université de Genève
[9] Space Telescope
[10] College of Natural Scienc
[11] Research School of Astronomy and A
[12] IPAC, Mail Code
[13] Chinese Academy of Sciences South America Ce
[14] School of Physical S
[15] Department of Physics and
[16]
[17] Department of Phys
Received 2020 December 30;

The identification of an emission l
same spectrum. However, in many
redshift. We developed a freely av
(UMLAUT) that can be used in a
lines. To this purpose, the algorith
size and color of the emitting sourc
case, the algorithm automatically
associated with the output paramete
combine supervised and unsupervis
WISP (WFC3 IR Spectroscopic Pa
near-IR spectroscopic surveys that
missions will investigate the large-
sky in near-IR slitless spectroscopy
correctly identifies real lines in 83.
unsupervised approach with a supe

frontiers in Psychology
ORIGINAL RESEARCH
published: 11 July 2019
doi: 10.3389/fpsyg.2019.01528
Check for updates

Inter- vs. Intra-Speaker Variation in Mixed Heritage Syntax: A Statistical Analysis

Federica Cognola[1], Ivano Baronchelli[2]* and Evelina Molinari[3]*

[1] Dipartimento di Studi Europei, Americani e Interculturali, Sapienza Università di Roma, Rome, Italy, [2] Dipartimento di Fisica e Astronomia, Università di Padova, Padua, Italy, [3] Istituto Culturale Mòcheno, Palù del Fersina, Italy

OPEN ACCESS

Edited by:
Ángel J. Gallego,
Autonomous University of Barcelona, Spain

Reviewed by:
Kleanthes K. Grohmann,

Based on the novel data pertaining to five syntactic phenomena (the position of the finite verb in embedded clauses, in sentences with a modal verb, negative concord, the position of focused light/heavy objects in main clauses with a complex tense and scrambling) in the heritage language Mòcheno collected via original fieldwork, we show that there are two populations – one exhibiting intra-speaker variation between German and Italian word orders, and one lacking it; and these two populations are the result of diatopic variation and, to a lesser extent, of diastratic variation. The results achieved using quantitative statistical analysis are partially convergent with those arrived at via the traditional theoretical syntax for Mòcheno, but our analysis has allowed us to shed new light on a series of phenomena that have been neglected or poorly understood thus far. More specifically, and for the first time, we discovered that there is a micro-variation resulting from diastratic (age) variation within the Roveda variety, which represents the

The Astrophysical Journal Supplement Series, 249:12 (27pp), 2020 July
https://doi.org/10.3847/1538-4365/ab9a3a

CrossMark

Identification of Single Spectral Lines through Supervised Machine Learning in a Large HST Survey (WISP): A Pilot Study for Euclid and WFIRST

I. Baronchelli[1], C. M. Scarlata[2], G. Rodighiero[1], L. Rodríguez-Muñoz[1], M. Bonato[3,4], M. Bagley[5], A. Henry[6], M. Rafelski[6,7], M

[1] Dipartimento di

zenodo

Search

Upload Communities

Log in Sign up

December 10, 2021

Software Open Access

UMLAUT - Little snail

Ivano Baronchelli

UMLAUT 1.0

Unsupervised Machine Learning Algorithm based on Unbiased Topology (IDL VERSION)

DOI: 323649071.svg

https://zenodo.org/badge/323649071.svg

REFERENCE: Baronchelli I. et al. (2021):

https://iopscience.iop.org/article/10.3847/1538-4365/ac250c/meta

https://arxiv.org/abs/2111.01807

UMLAUT is a variant of the KNN (K-closest neighbor) algorithm.

- Given a set of reference data points (training set), for which the value of N+1 parameters is known,

- Given one analysis data point with N parameters known and the (N+1)-th parameter unknown,

--> UMLAUT estimates the value of the (N+1)-th parameter for the analysis data point. To this purpose, UMLAUT finds the closest data-points of the training set, in a ranked N-dimensional space "associated" with the physical parameter space. After finding the closest data points, the unknown parameter is obtained as the combination (ex. from the average) of the values assumed by the closest reference data points along the (N+1)-th dimension.

Available in

GitHub

Indexed in

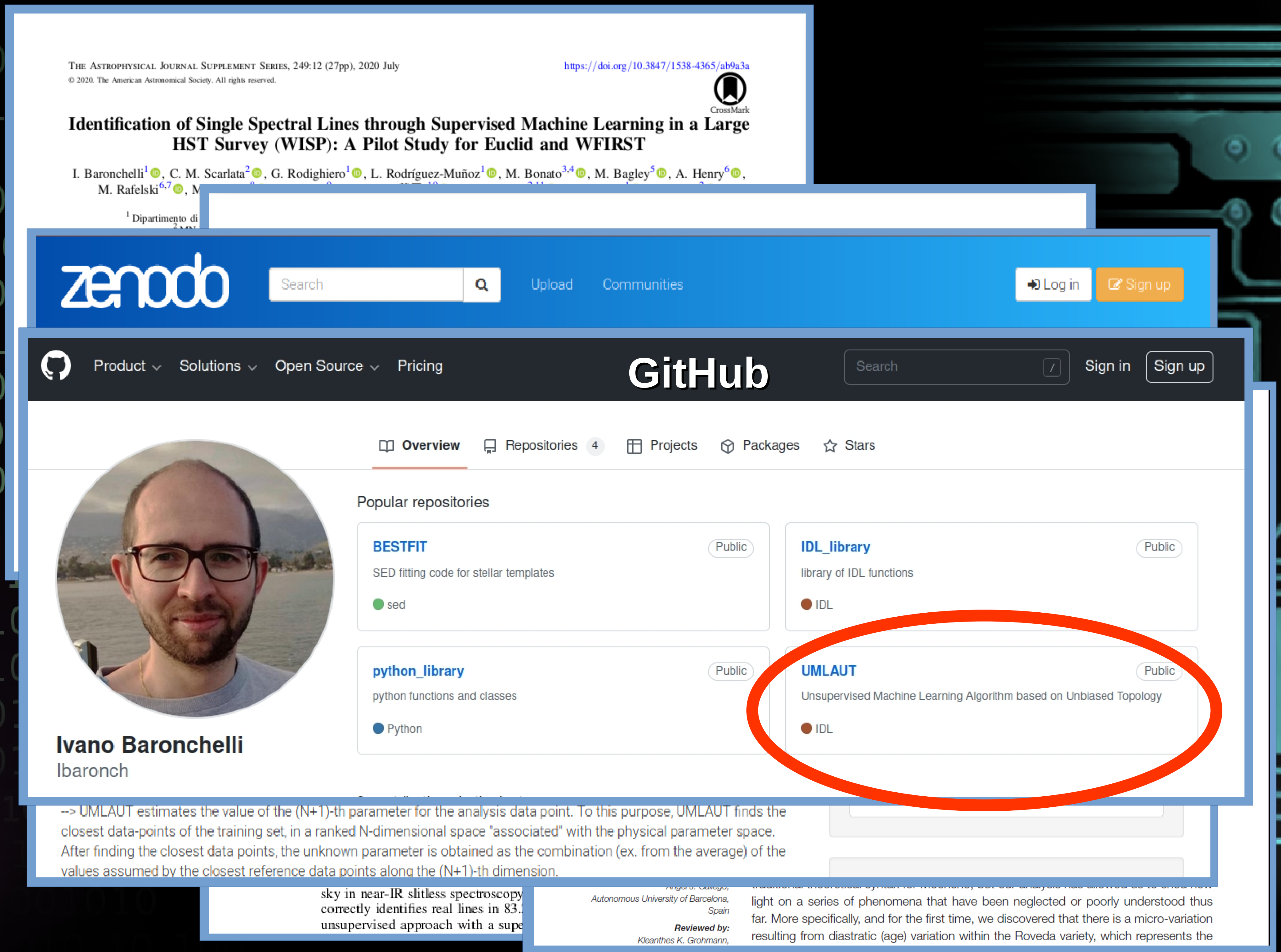
OpenAIRE

sky in near-IR slitless spectroscopy
correctly identifies real lines in 83.
unsupervised approach with a supe

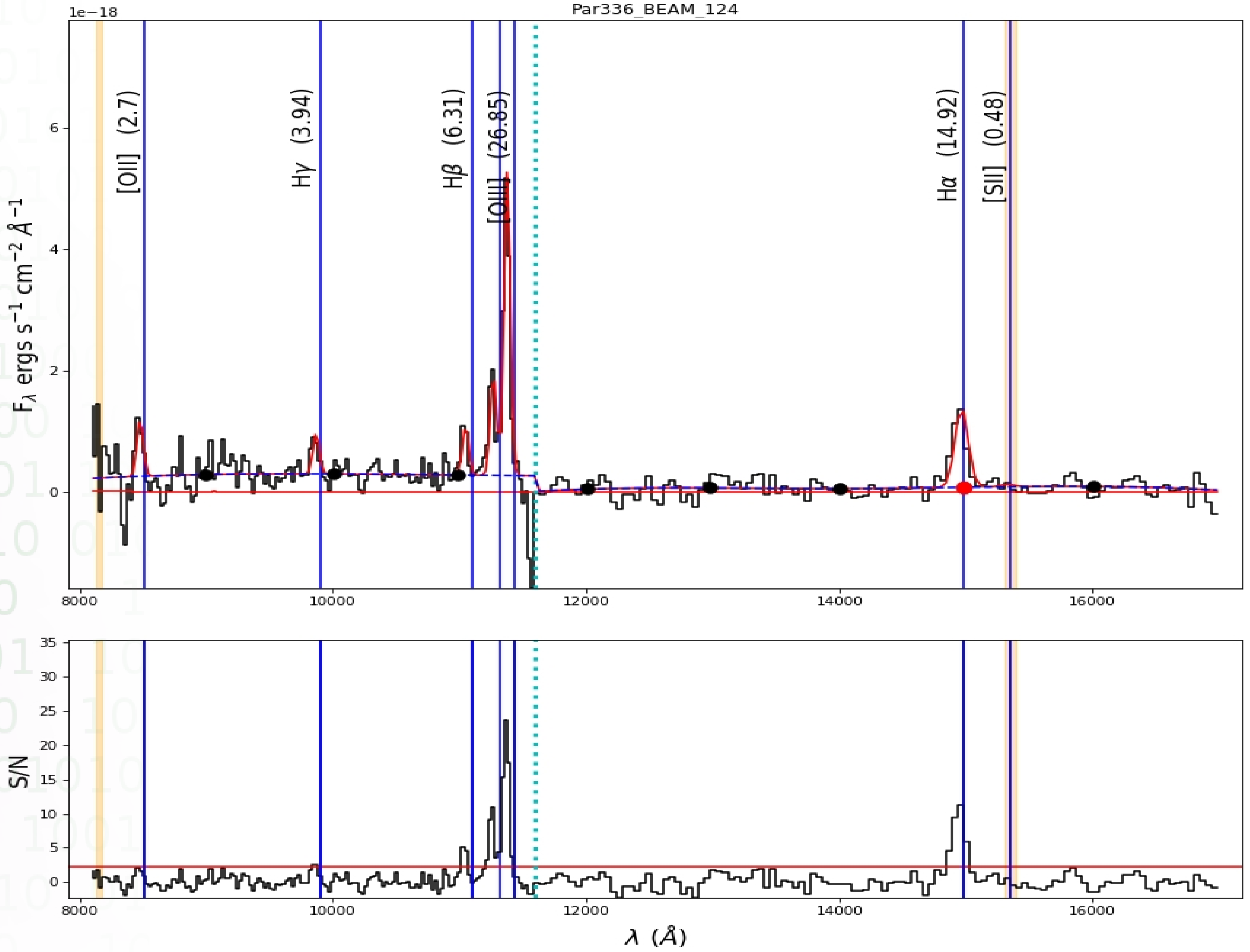
Autonomous University of Barcelona, Spain

Reviewed by:
Kleanthes K. Grohmann,

light on a series of phenomena that have been neglected or poorly understood thus far. More specifically, and for the first time, we discovered that there is a micro-variation resulting from diastratic (age) variation within the Roveda variety, which represents the


The Astrophysical Journal Supplement Series, 249:12 (27pp), 2020 July
https://doi.org/10.3847/1538-4365/ab9a3a
© 2020. The American Astronomical Society. All rights reserved.
CrossMark
Identification of Single Spectral Lines through Supervised Machine Learning in a Large HST Survey (WISP): A Pilot Study for Euclid and WFIRST
I. Baronchelli[1], C. M. Scarlata[2], G. Rodighiero[1], L. Rodríguez-Muñoz[1], M. Bonato[3,4], M. Bagley[5], A. Henry[6],
M. Rafelski[6,7], M
[1] Dipartimento di
zenodo
Search
Upload
Communities
Log in
Sign up
Product
Solutions
Open Source
Pricing
GitHub
Search
Sign in
Sign up
Overview
Repositories 4
Projects
Packages
Stars
Popular repositories
BESTFIT
Public
SED fitting code for stellar templates
sed
IDL_library
Public
library of IDL functions
IDL
python_library
Public
python functions and classes
Python
UMLAUT
Public
Unsupervised Machine Learning Algorithm based on Unbiased Topology
IDL
Ivano Baronchelli
Ibaronch
--> UMLAUT estimates the value of the (N+1)-th parameter for the analysis data point. To this purpose, UMLAUT finds the closest data-points of the training set, in a ranked N-dimensional space "associated" with the physical parameter space. After finding the closest data points, the unknown parameter is obtained as the combination (ex. from the average) of the values assumed by the closest reference data points along the (N+1)-th dimension.
sky in near-IR slitless spectroscopy
correctly identifies real lines in 83.
unsupervised approach with a supe
Autonomous University of Barcelona, Spain
Reviewed by:
Kleanthes K. Grohmann,
light on a series of phenomena that have been neglected or poorly understood thus far. More specifically, and for the first time, we discovered that there is a micro-variation resulting from diastratic (age) variation within the Roveda variety, which represents the

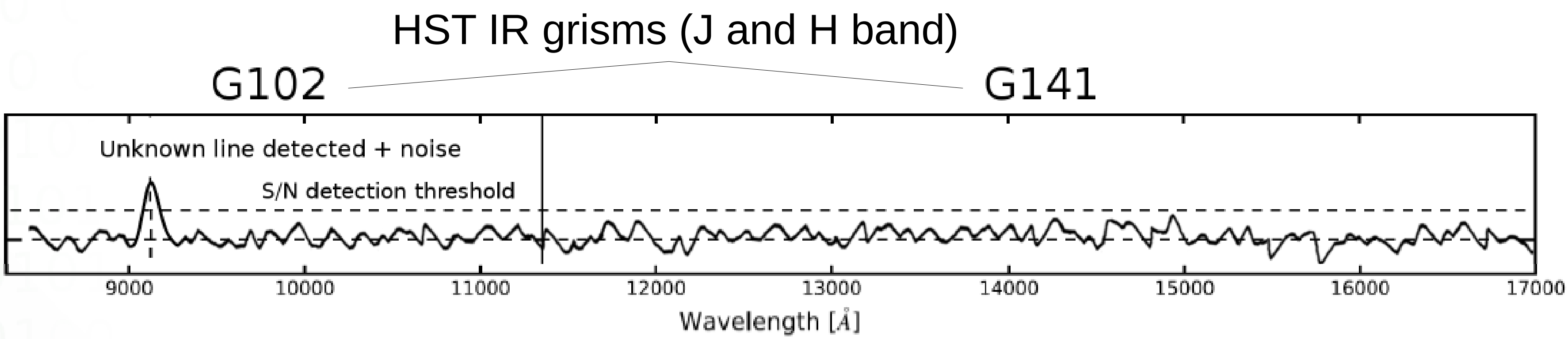
Identifying emission lines is quite easy...
… when multiple lines are detected



WISP data (Atek+2010)

But when only one line is detected...

?

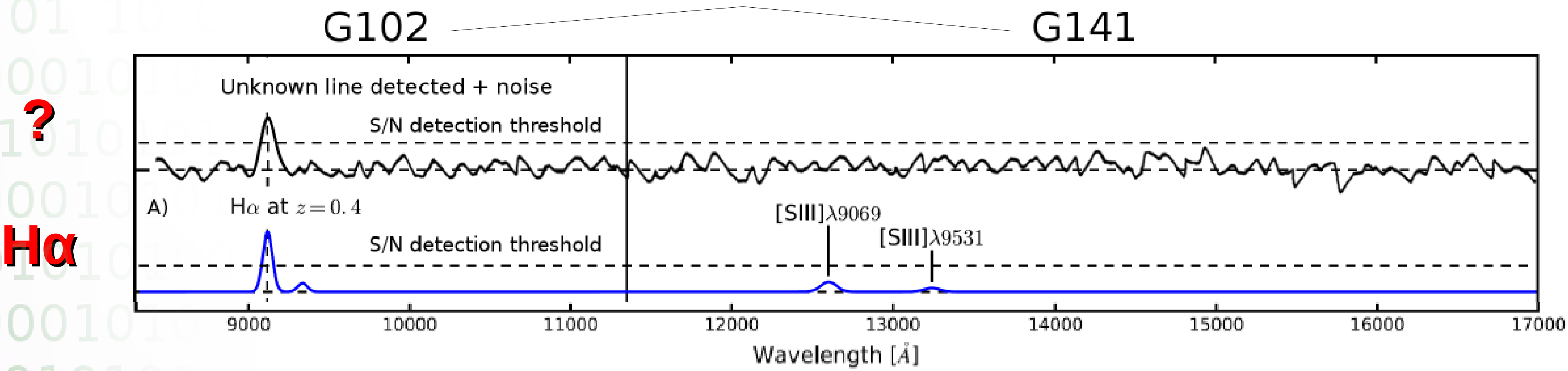


Unknown single line detected

Anything else is beneath the S/N detection threshold

But when only one line is detected...

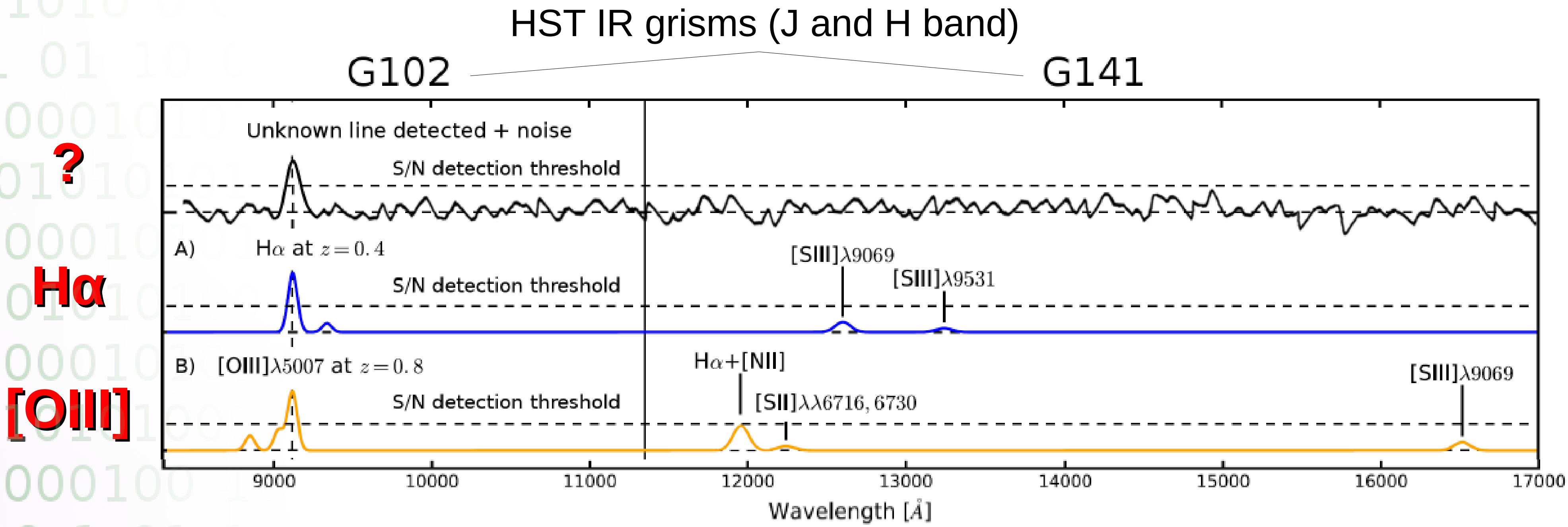
HST IR grisms (J and H band)



Unknown single line detected

Anything else is beneath the S/N detection threshold

But when only one line is detected...

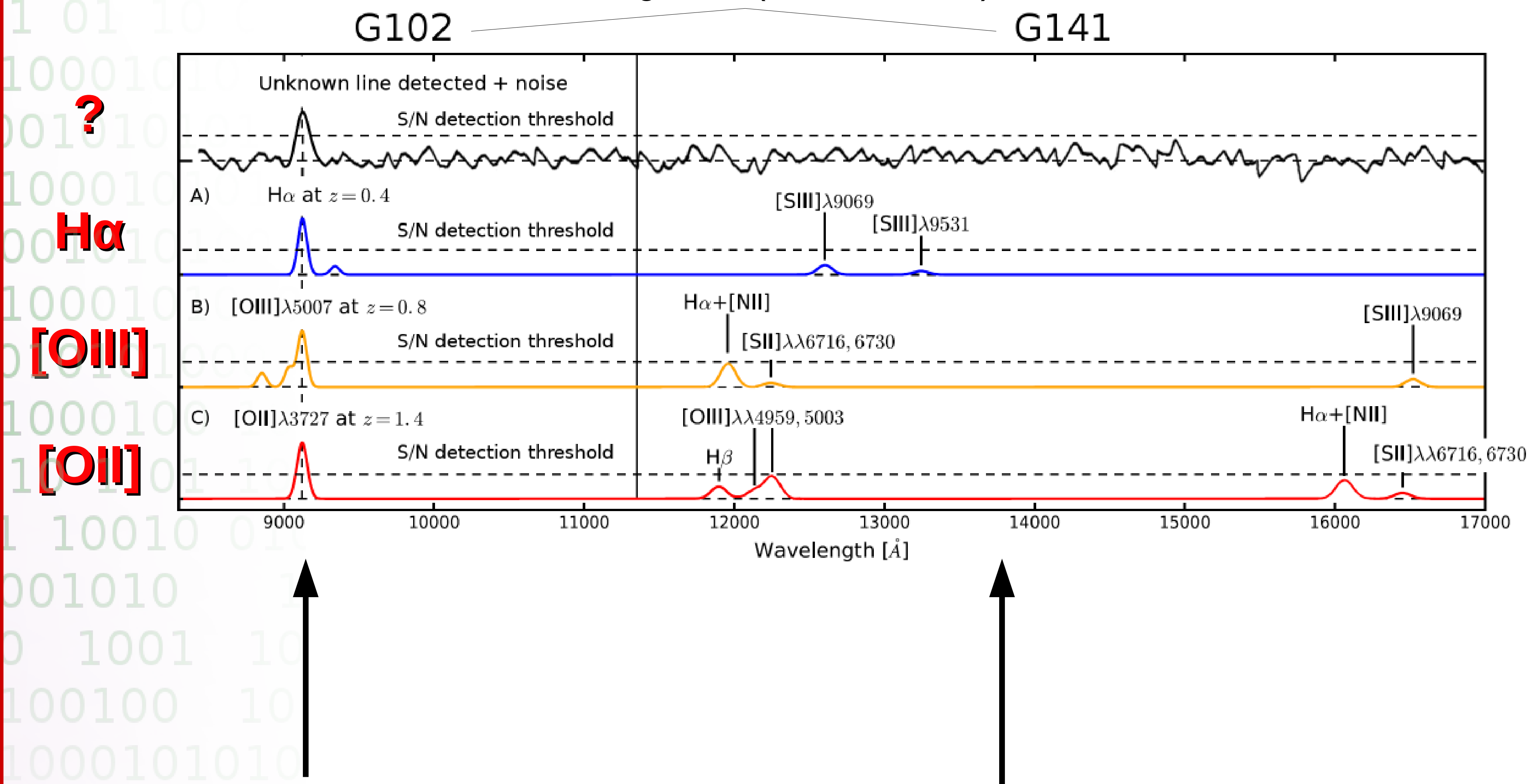


Unknown single line detected

Anything else is beneath the S/N detection threshold

But when only one line is detected...

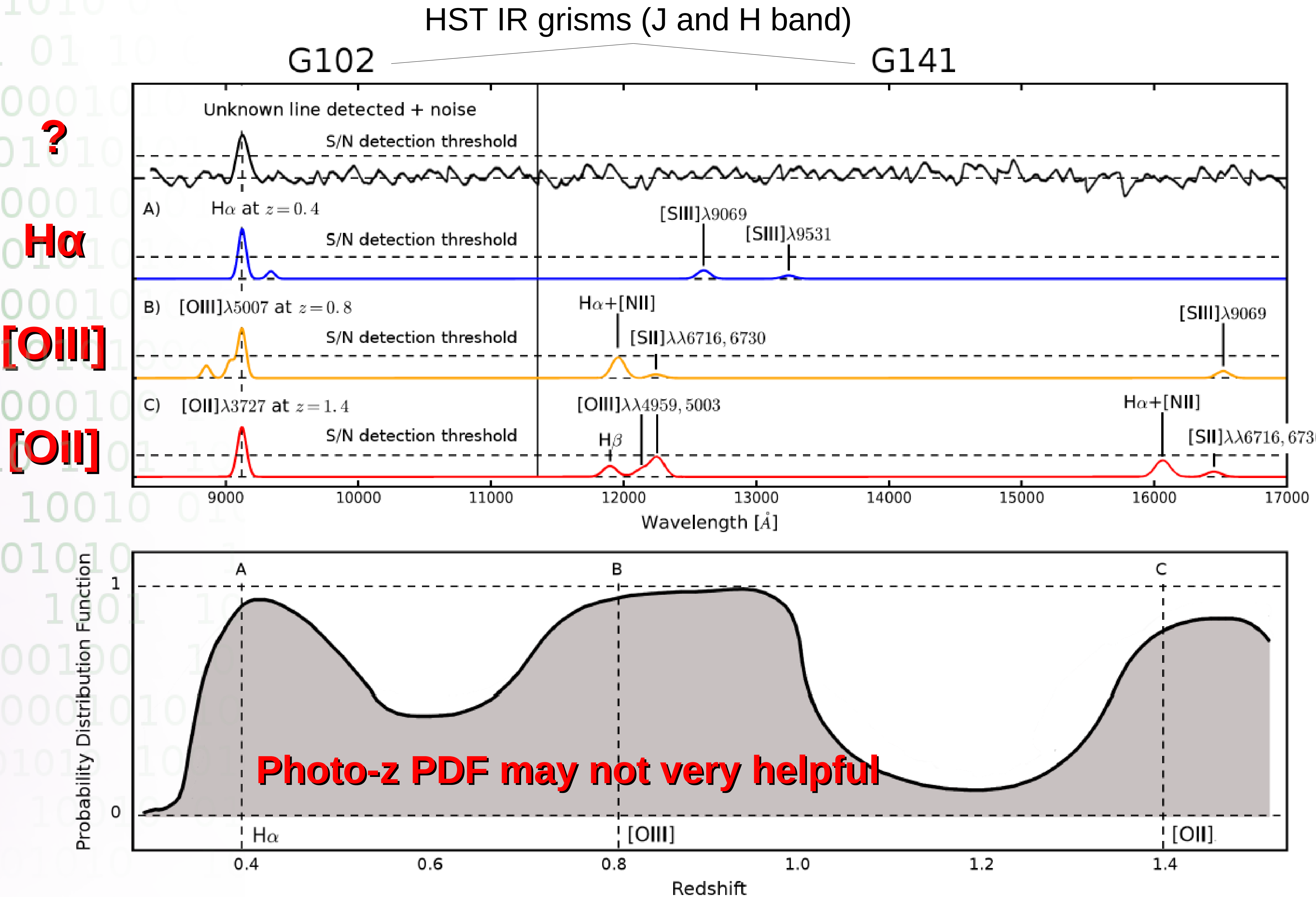
HST IR grisms (J and H band)



Unknown single line detected

Anything else is beneath the S/N detection threshold

But when only one line is detected...



WISP

PI: M. A. Malkan (University of California - Los Angeles)

The WFC3 Infrared Spectroscopic Parallel survey

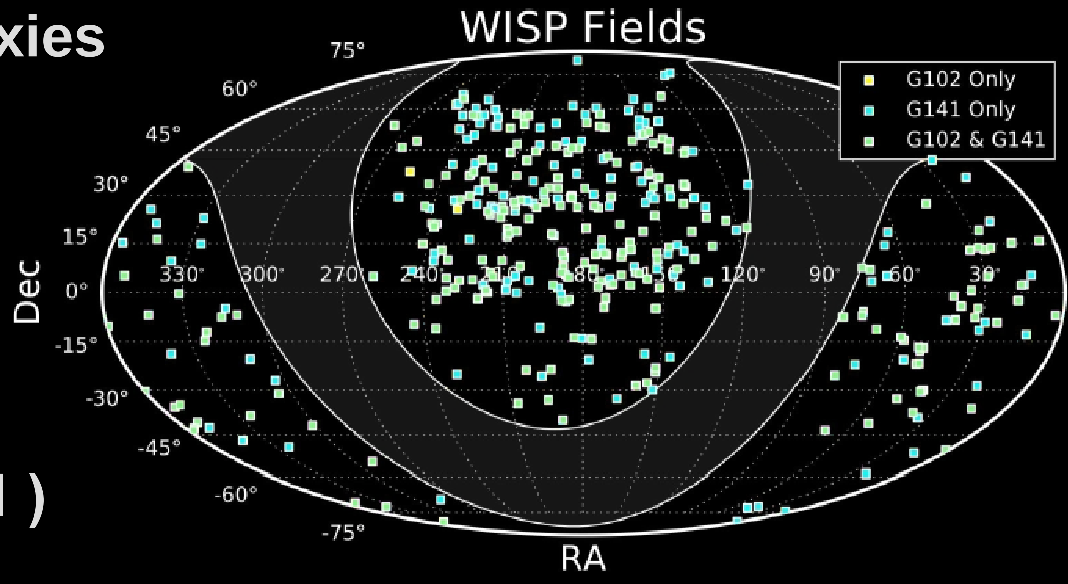
H. Atek et al. (2010)

HST pure parallel large program (~2000 orbits, >1500 arcmin^2)

- slitless spectroscopy { G102: 0.8 – 1.1 μm (R~210) / G141: 1.07 – 1.7 μm (R~130) }
- direct imaging { J band (F110W), / H band (F140W/F160W) / Ancillary UVIS + IRAC + ground based optical }

- One of the most importan proxies For the future surveys of <u>Euclid</u> and <u>Roman</u>

- Now collecting data with <u>JWST</u> (PASSAGE survey, P.I. M. Malkan, proposal ID: GO-1571 - Cycle 1)



An artificial intelligence is not an artificial GOD

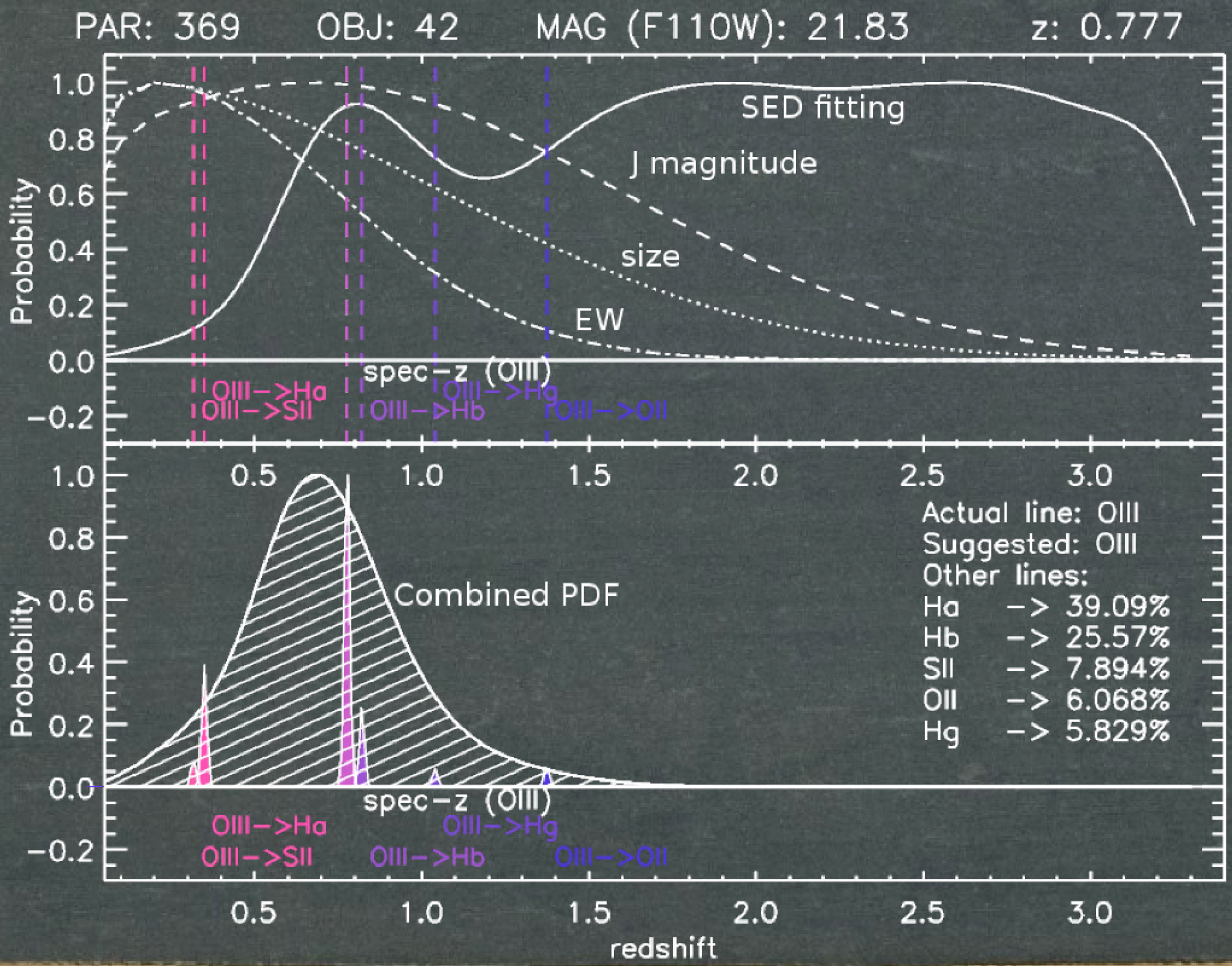
A learner can help exploiting hidden information, but it does not *create* information from nothing.

Additional information that can help identifying a spectral line:

- Apparent Magnitude (F110W);
- Apparent Size (F110W);
- Color index (J-H);
- Line equivalent width;

... etc ...

ALGORITHM 1 (Naïve Bayes approach)



UMLAUT basics

Unbiased Machine Learning Algoirthm Using Topology

Empirical principle:

Objects that are close to each other in N dimensions tend to be close also in an indipendent N+1 dimension, when N>>

UMLAUT basics

Unbiased Machine Learning Algoirthm Using Topology

Empirical principle:

Objects that are close to each other in N dimensions tend to be close also in an indipendent N+1 dimension, when N>>

Toucans that are similar to each other in 2 dimensions, tend to be similar also in 3 dimensions...

UMLAUT basics

Unbiased Machine Learning Algoirthm Using Topology

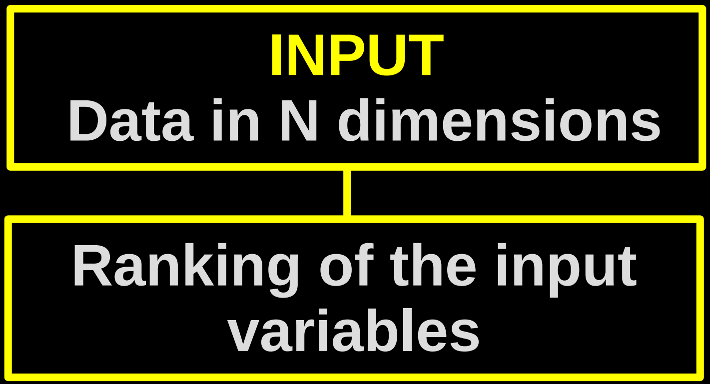
Empirical principle:

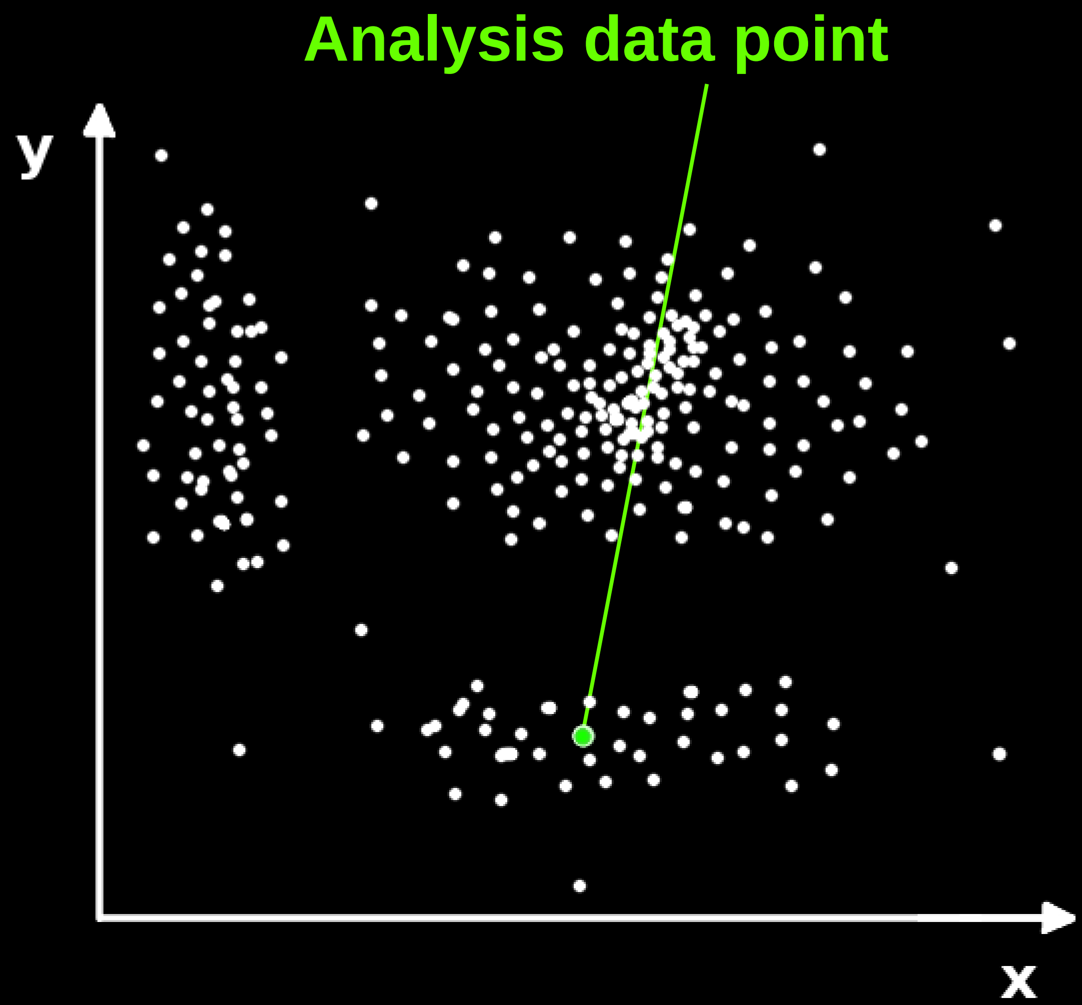
Objects that are close to each other in N dimensions tend to be close also in an indipendent N+1 dimension, when N>>

Toucans that are similar to each other in 2 dimensions, tend to be similar also in 3 dimensions...

...unless one of the pictures is depicting the picture of a toucan.

How it works in practice:



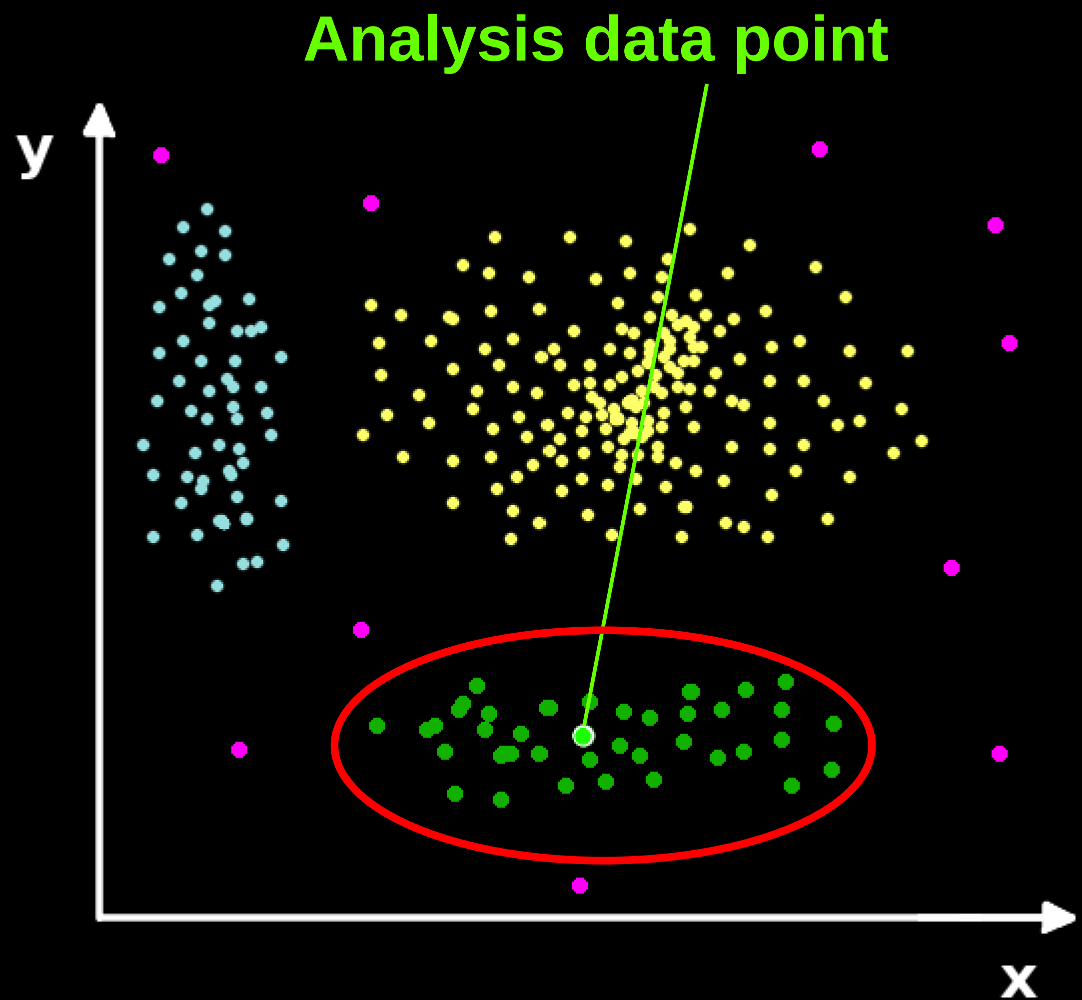


How it works in practice:

INPUT
Data in N dimensions

Ranking of the input variables

Groups identification (DBSCAN-like)



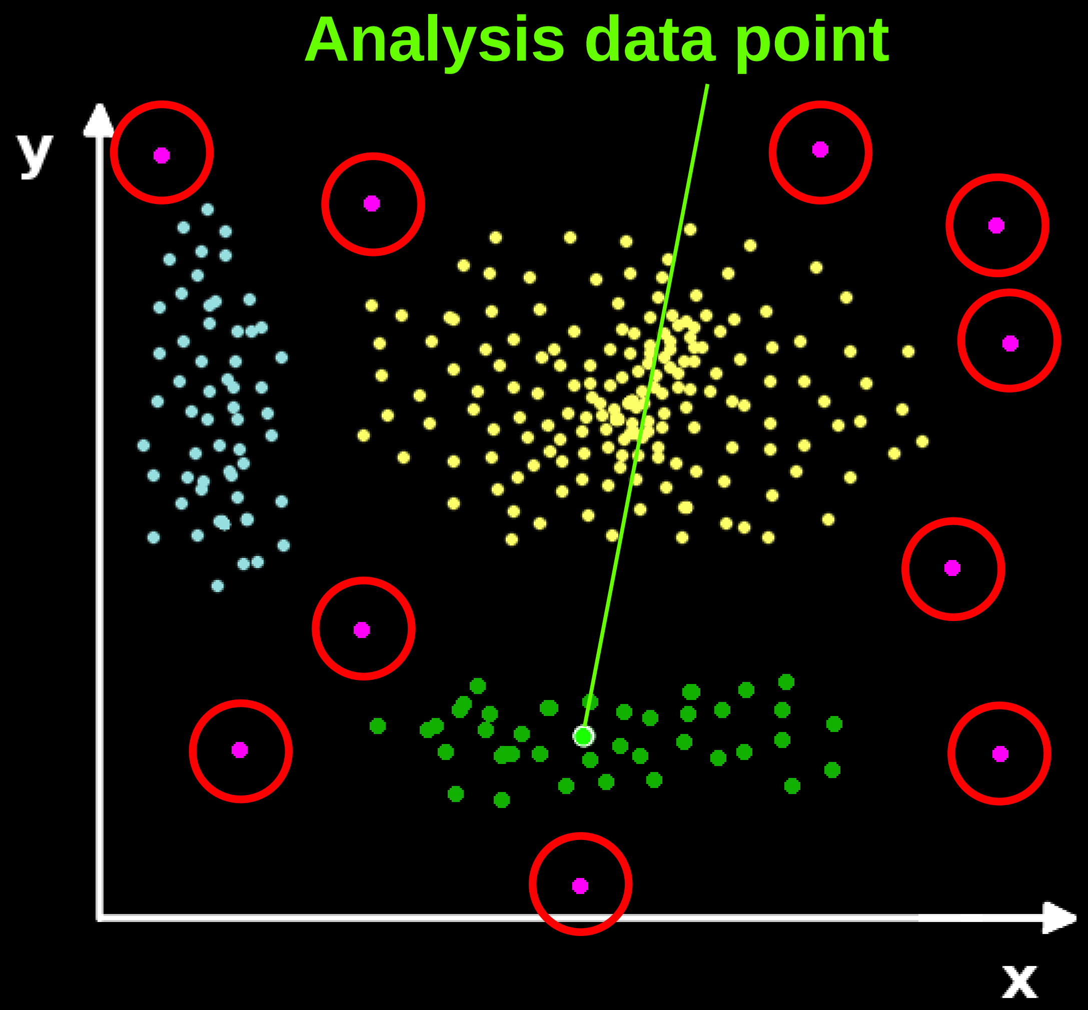
How it works in practice:

INPUT
Data in N dimensions

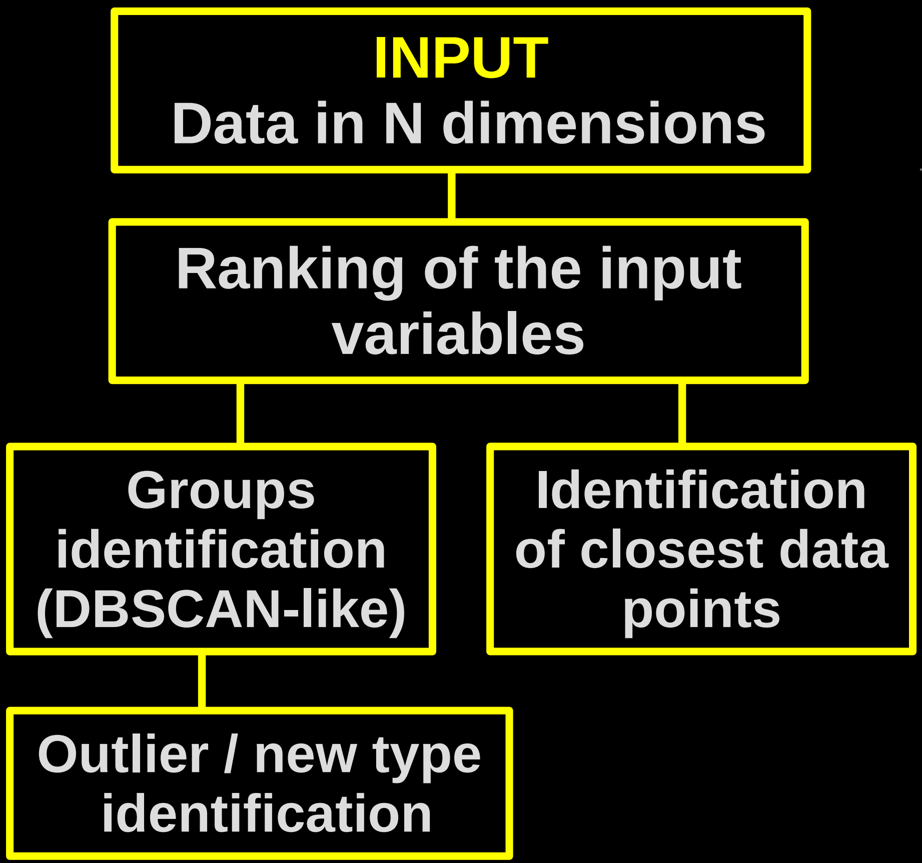
Ranking of the input variables

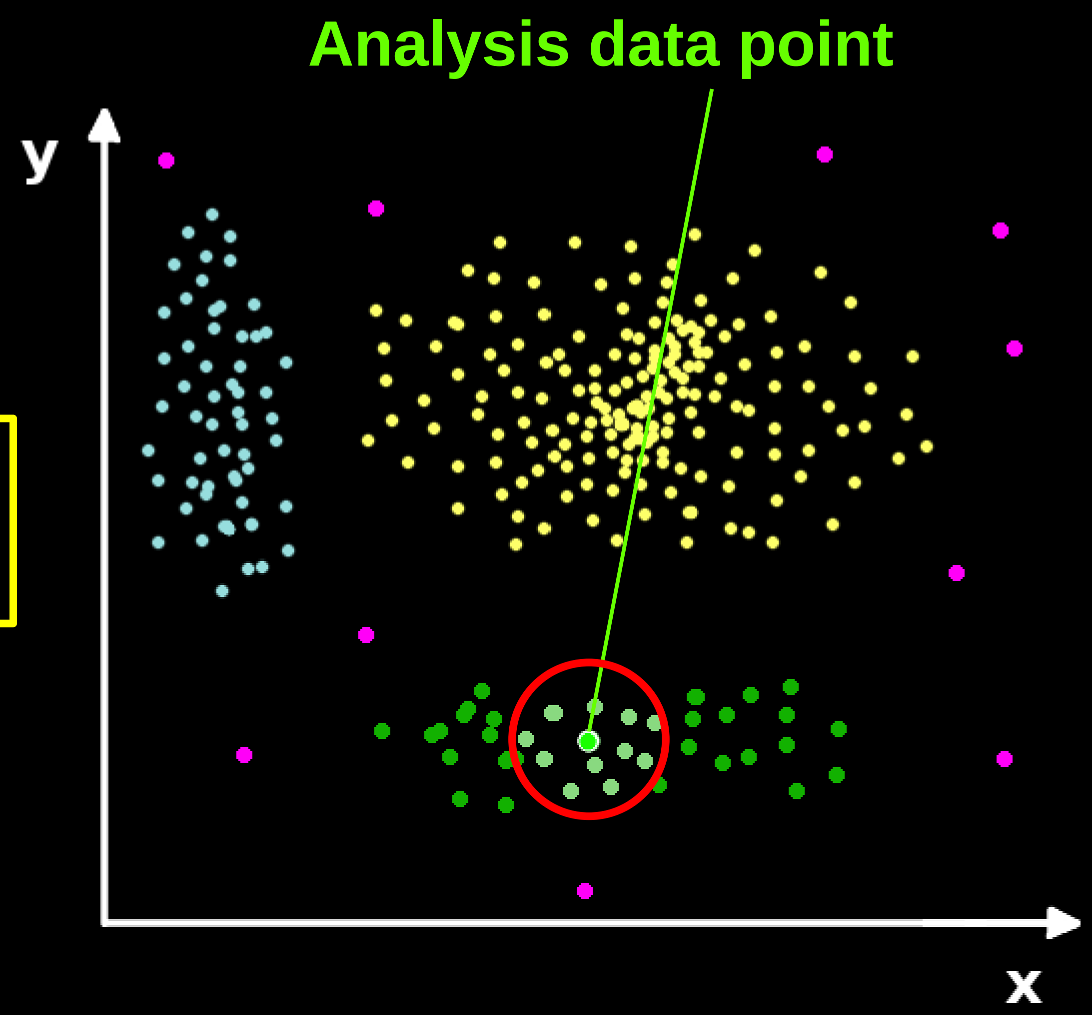
Groups identification (DBSCAN-like)

Outlier / new type identification



How it works in practice:





How it works in practice:

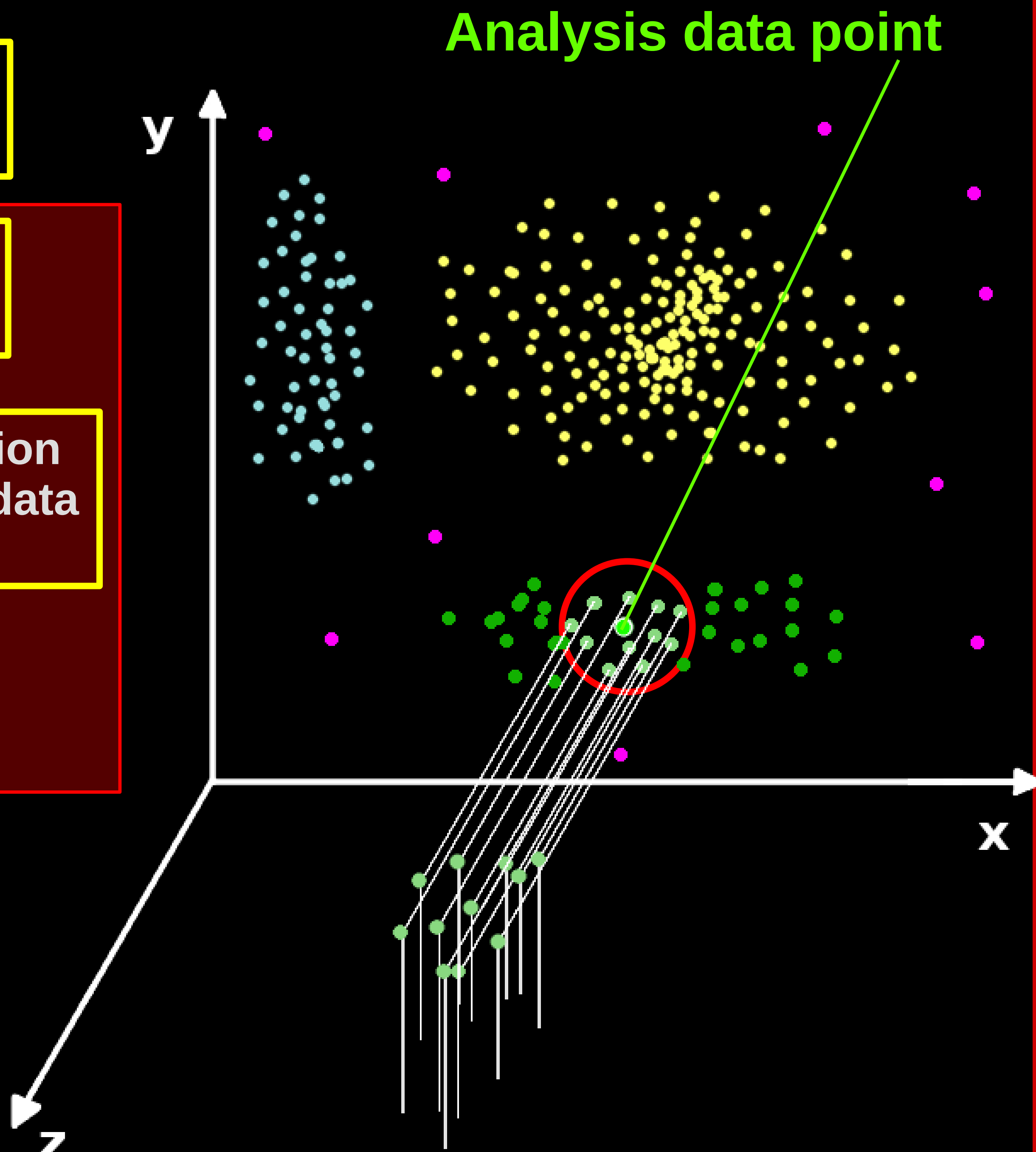
INPUT
Data in N dimensions

Ranking of the input variables

Groups identification (DBSCAN-like)

Identification of closest data points

Outlier / new type identification



How it works in practice:

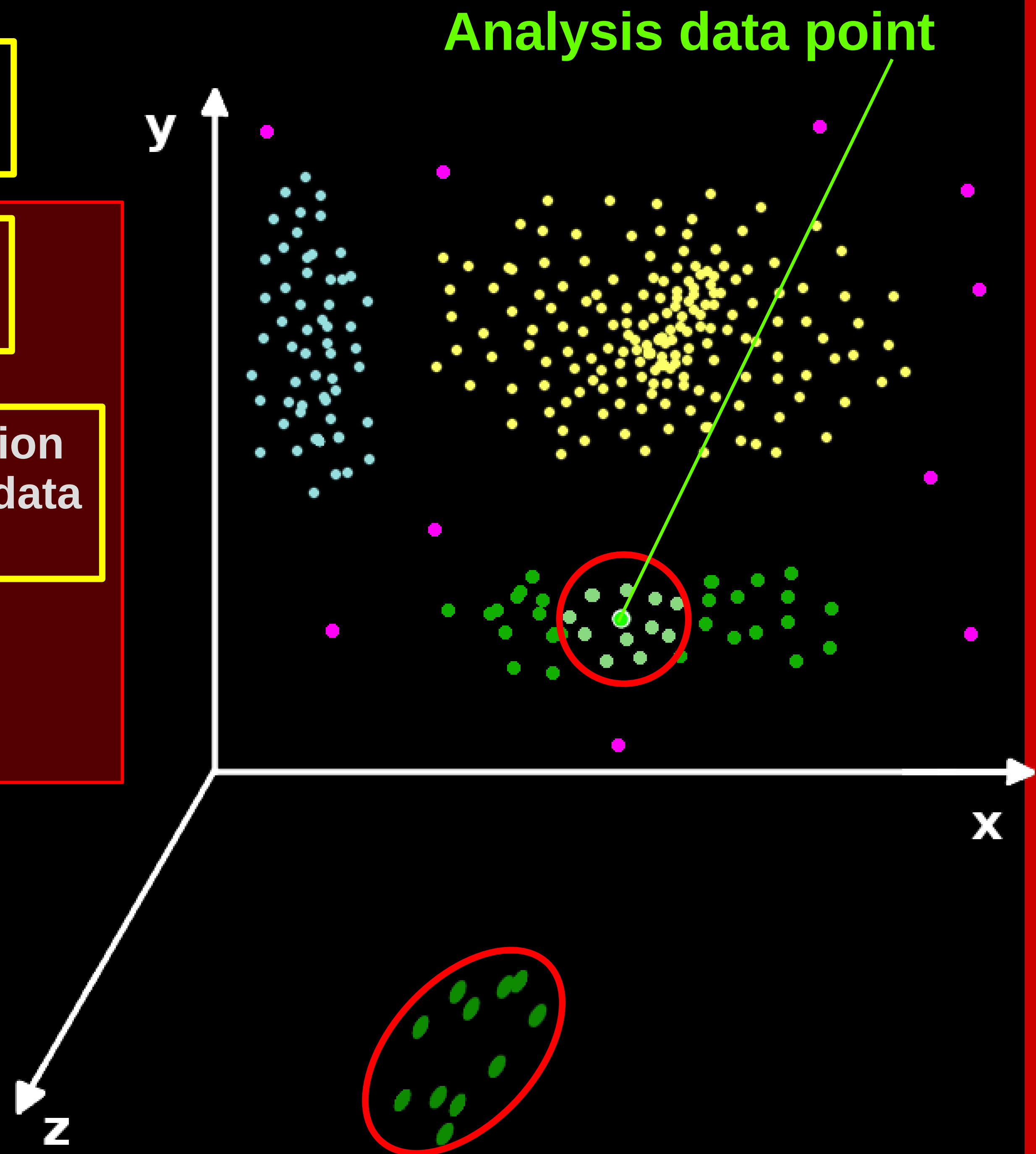
INPUT
Data in N dimensions

Ranking of the input variables

Groups identification (DBSCAN-like)

Identification of closest data points

Outlier / new type identification



How it works in practice:

INPUT
Data in N dimensions

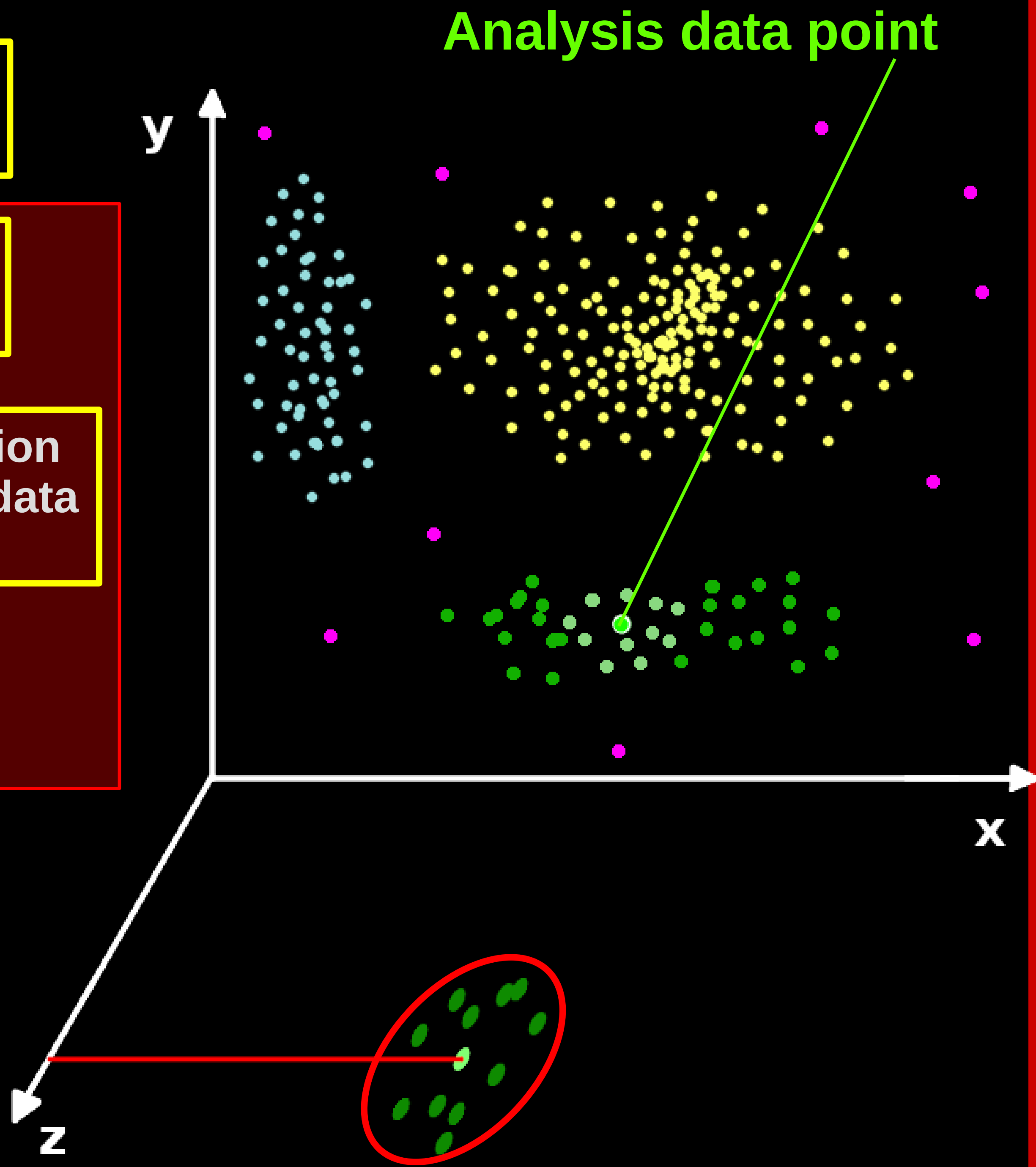
Ranking of the input variables

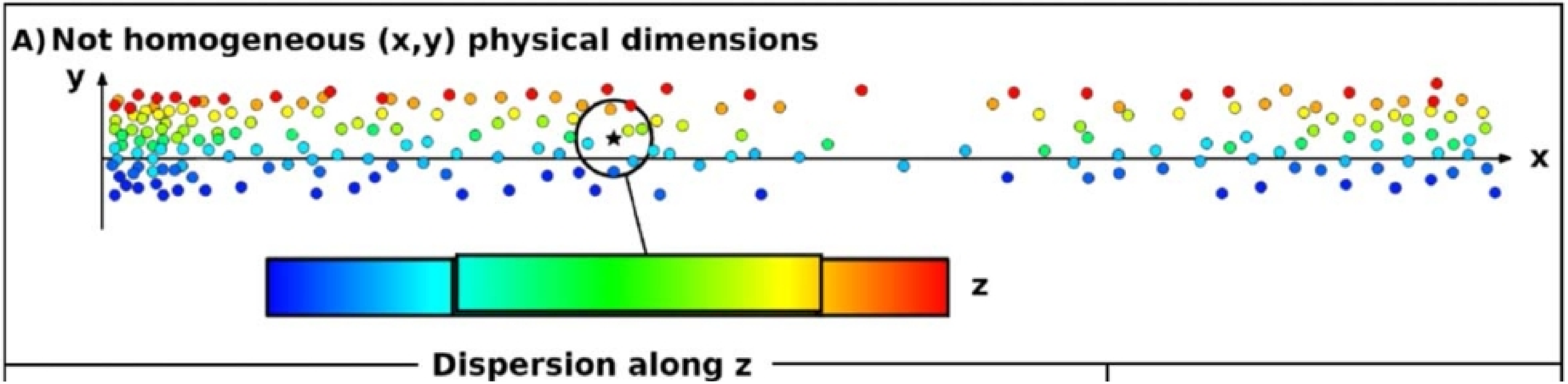
Groups identification (DBSCAN-like)

Identification of closest data points

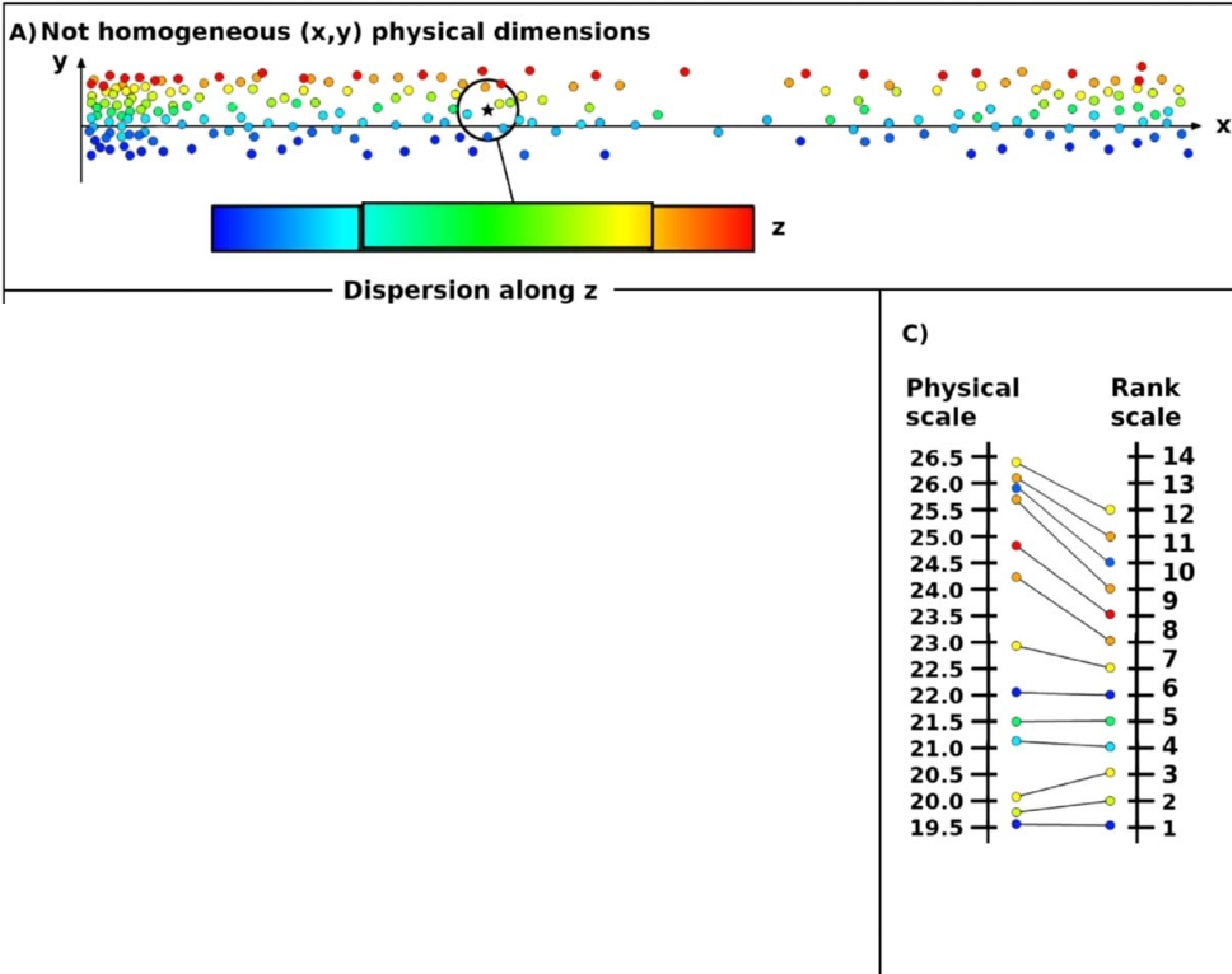
Outlier / new type identification

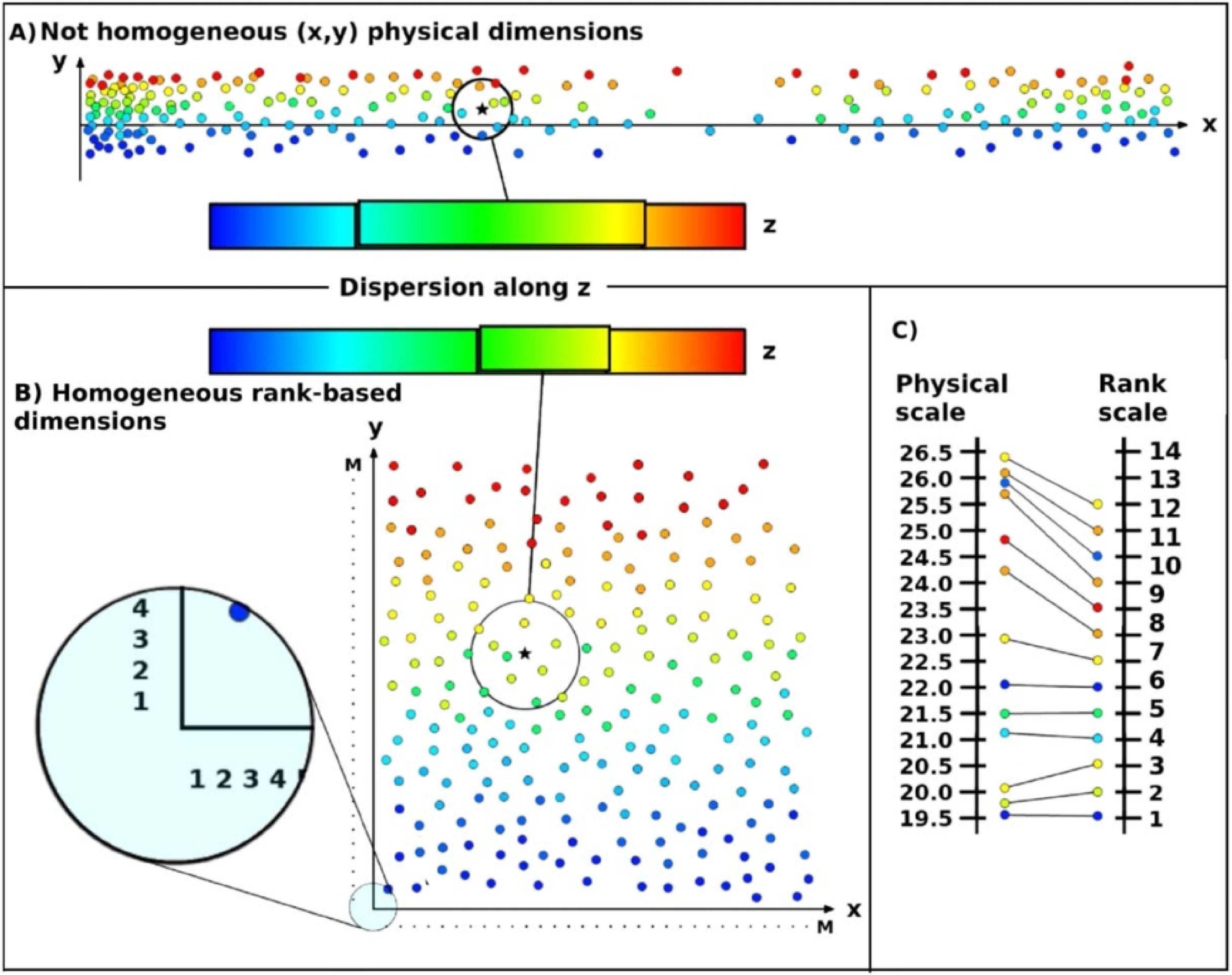
OUTPUT
Average value / fit / PDF along the N+1 dimension, for the closest data-points





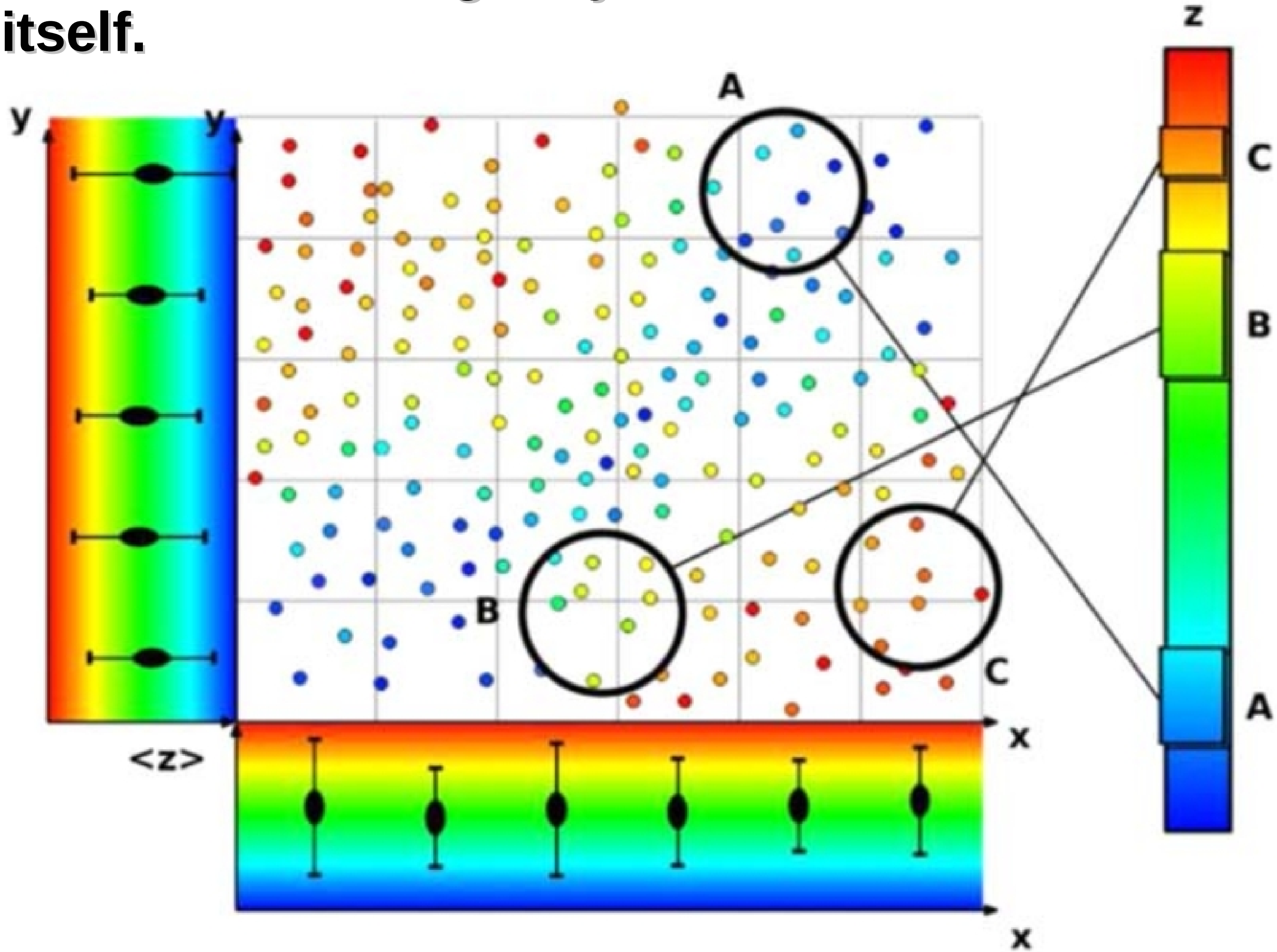
A) Not homogeneous (x,y) physical dimensions
y
x
z
Dispersion along z


A) Not homogeneous (x,y) physical dimensions
y
x
z
Dispersion along z
C)
Physical scale
Rank scale
26.5
26.0
25.5
25.0
24.5
24.0
23.5
23.0
22.5
22.0
21.5
21.0
20.5
20.0
19.5
14
13
12
11
10
9
8
7
6
5
4
3
2
1


A) Not homogeneous (x,y) physical dimensions
y
x
z
Dispersion along z
z
B) Homogeneous rank-based dimensions
y
M
4
3
2
1
1 2 3 4
x
M
C)
Physical scale
Rank scale
26.5
26.0
25.5
25.0
24.5
24.0
23.5
23.0
22.5
22.0
21.5
21.0
20.5
20.0
19.5
14
13
12
11
10
9
8
7
6
5
4
3
2
1

Among the UMLAUT Advantages:

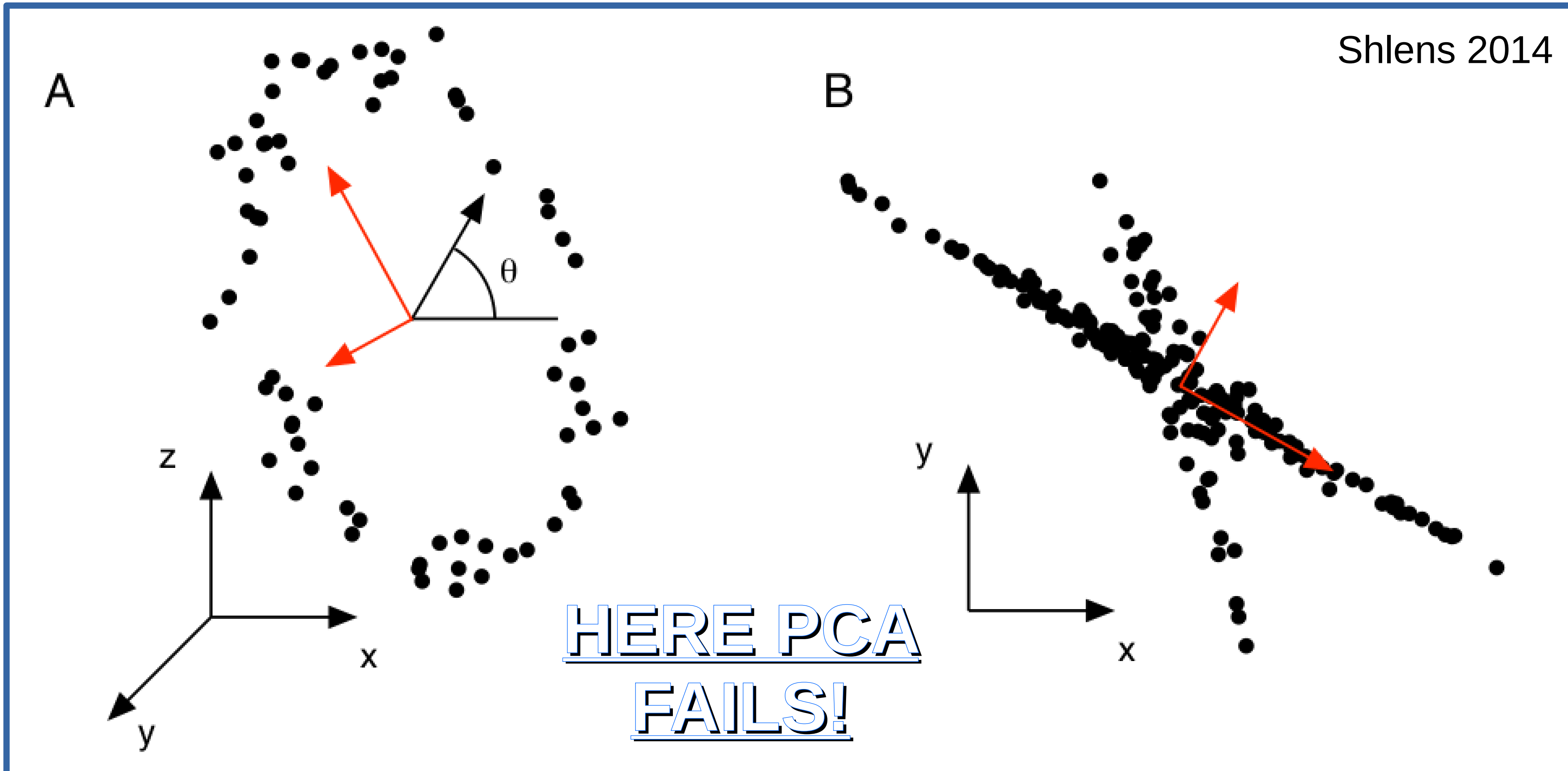
Single dimensions could retain useful information about the output without showing any correlations at all with the output itself.



→ UMLAUT takes into account all the input dimensions at the same time!

Among the UMLAUT Advantages:

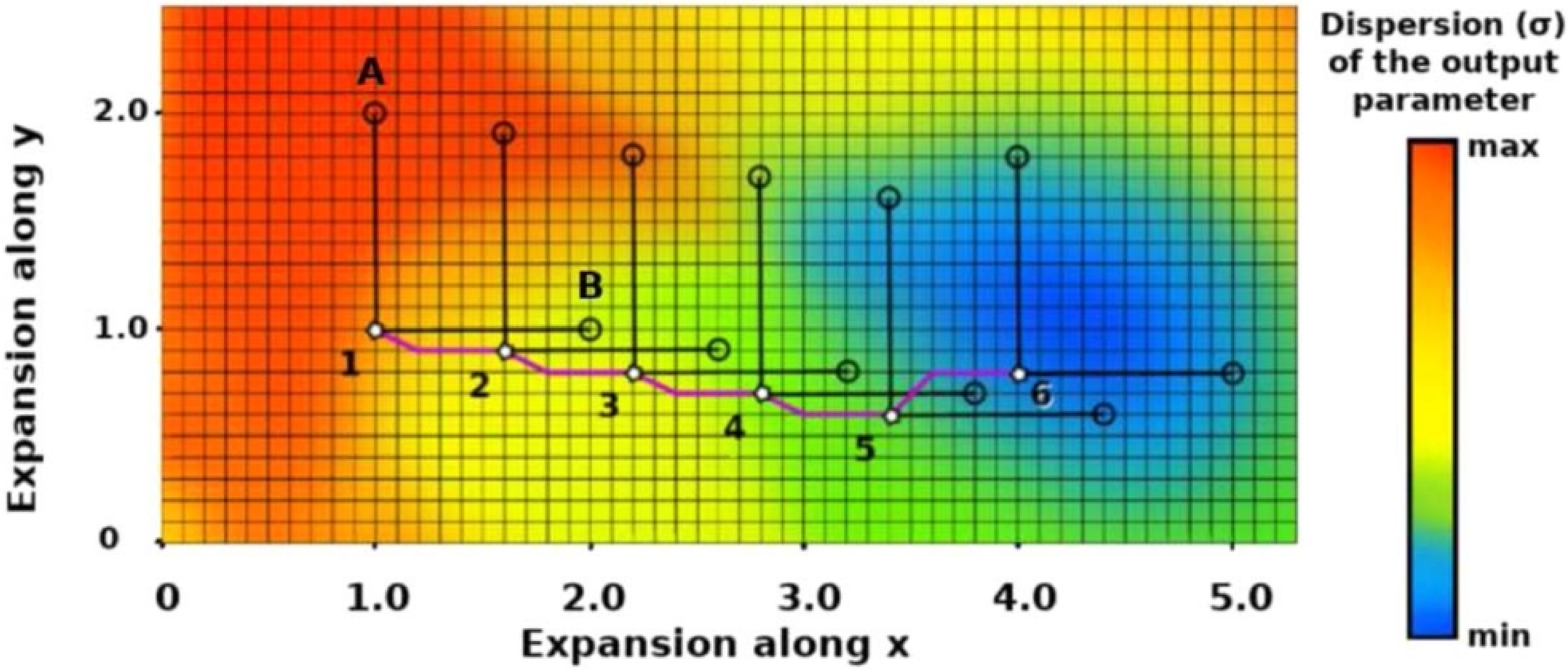
Dimensionality reduction strategies (such as PCA) often assume orthogonal native axes and/or Gaussian distributions



→ UMLAUT does not require/assume orthogonal native axes or assumptions on the data distribution.

Dimensionality reduction strategy

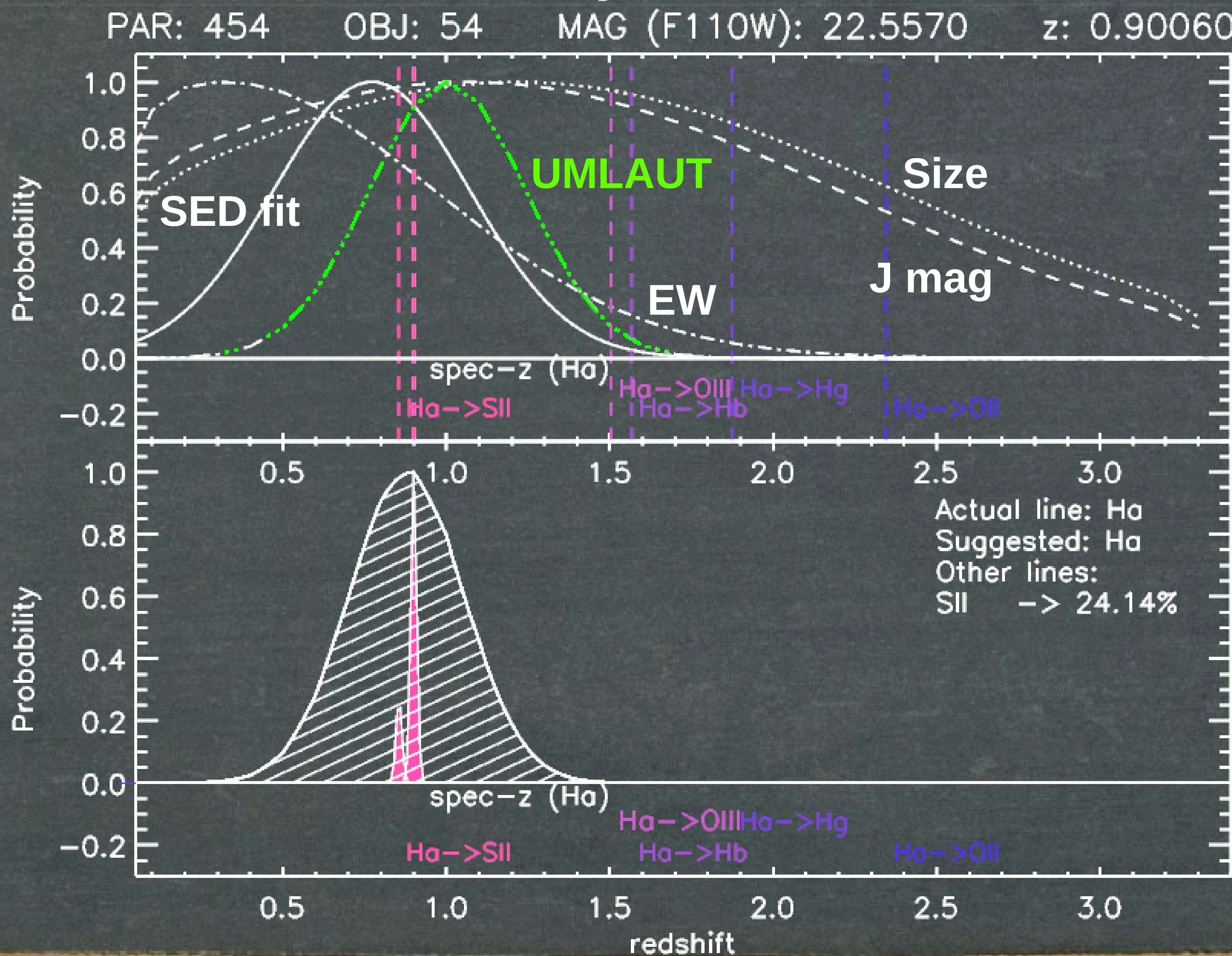
N-dimensional space of expansion of the N input parameters (N=2 in this example)



Strategy (simplified):

1) Measure the dispersion on the output parameter in positions 1, A and B
2) Move into the direction of decreasing gradient (position 2)
3) Iterate until a local minimum is found

Naïve Bayes + UMLAUT



Naïve Bayes alone

	Completeness	Contamination	Purity
Ha	88.3 %	15.2 %	84.8 %
[OIII]	77.0 %	17.9 %	82.1 %
[OII]	50.0 %	73.3 %	26.7 %

UMLAUT alone

	Completeness	Contamination	Purity
Ha	89.6 %	16.0 %	84.0 %
[OIII]	77.8 %	18.0 %	82.0 %
[OII]	12.5 %	33.3 %	66.7 %

Naïve Bayes + UMLAUT

	Completeness	Contamination	Purity
Ha	89.3 %	12.8 %	87.2 %
[OIII]	79.6 %	16.2 %	83.8 %
[OII]	66.7 %	70.9 %	29.1 %

ALMA Band 3 number counts

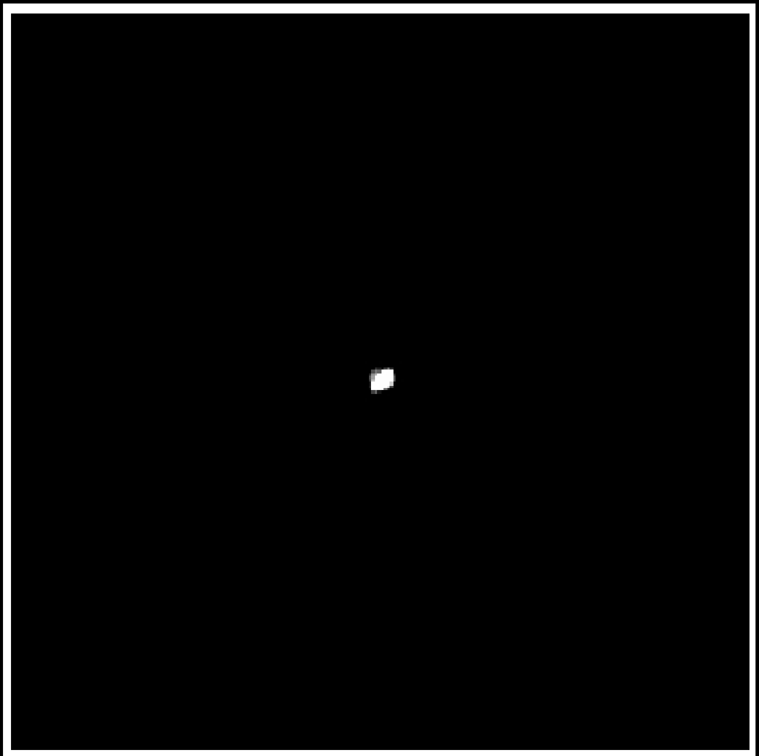


Image of a band 3 calibrator

- Are there any other serendipitous detections around?

Compute number counts !

ALMA Band 3 number counts

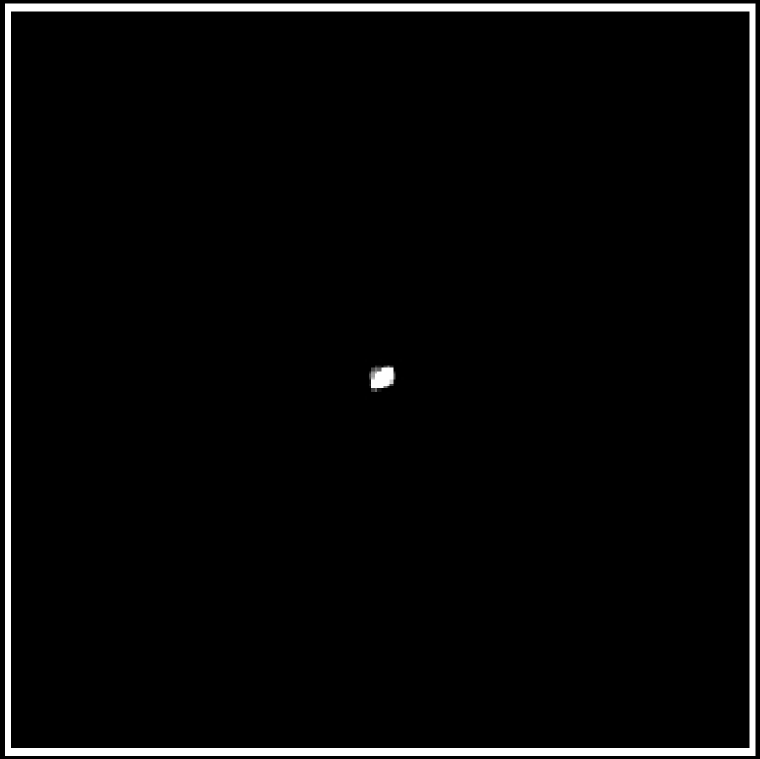
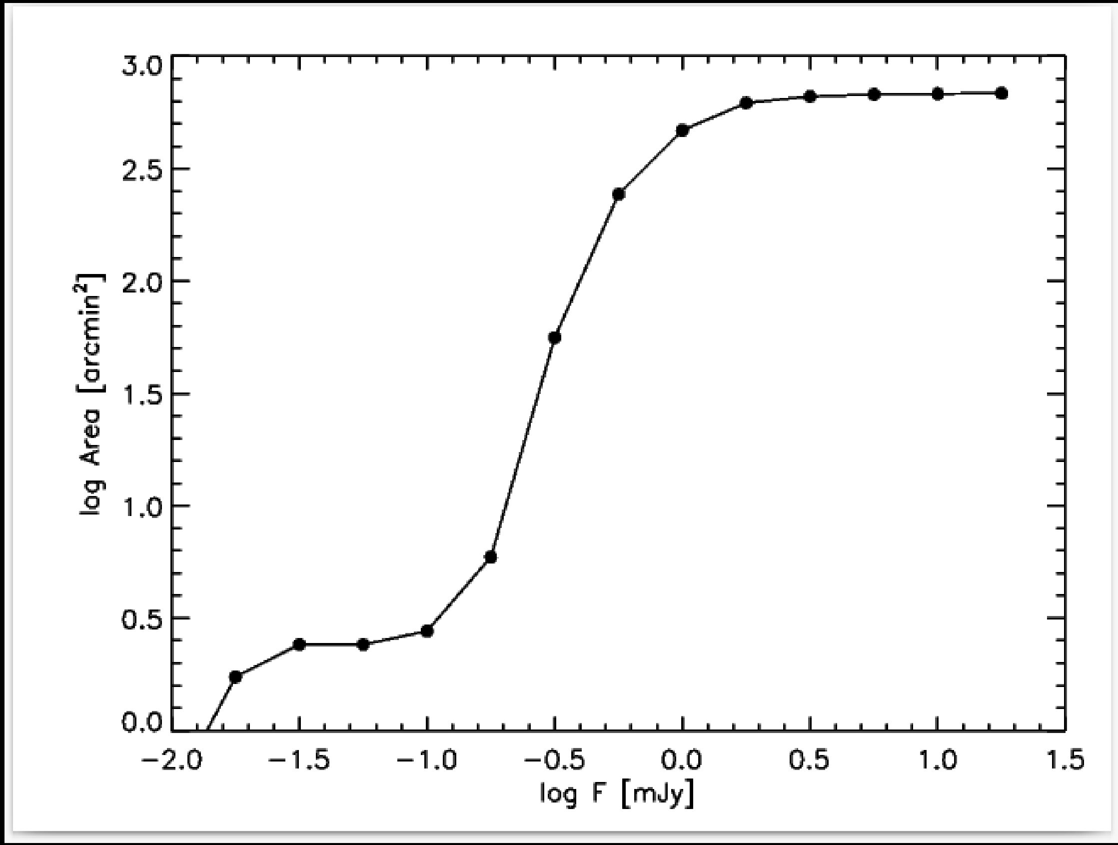


Image of a band 3 calibrator



~ 500 valid calibrator images
~ 700 arcmin^2 (at 1mJy)

- Are there any other serendipitous detections around?

Compute number counts !

Contribution of each single detection

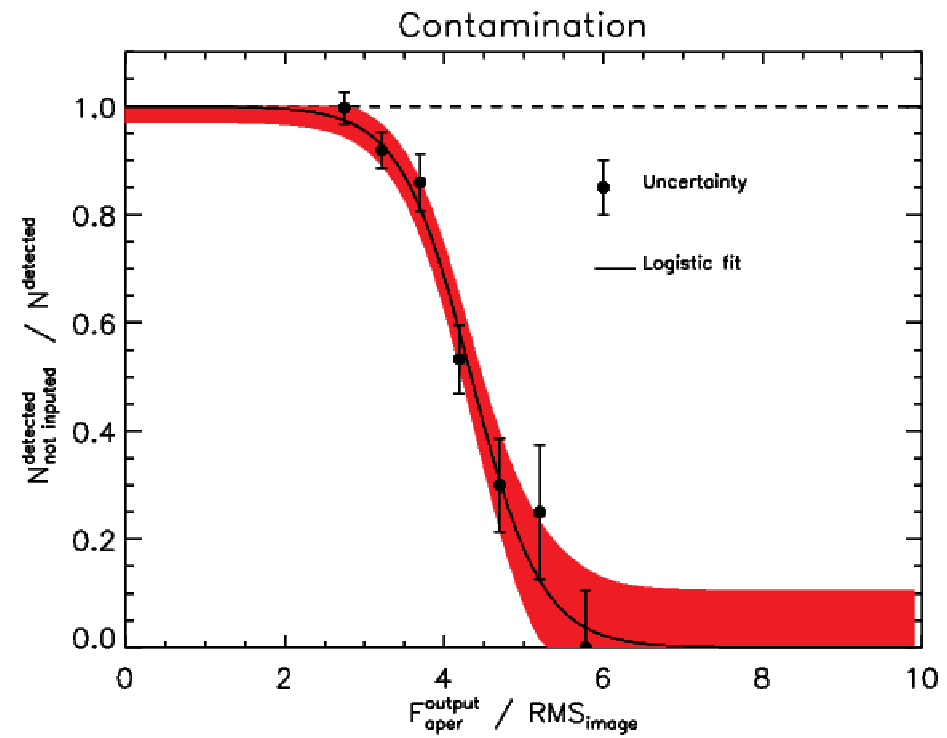
$$N_i(S_i) = \frac{1 - f_{sp(S_i)}}{C(S_i) \cdot A(S_i)},$$

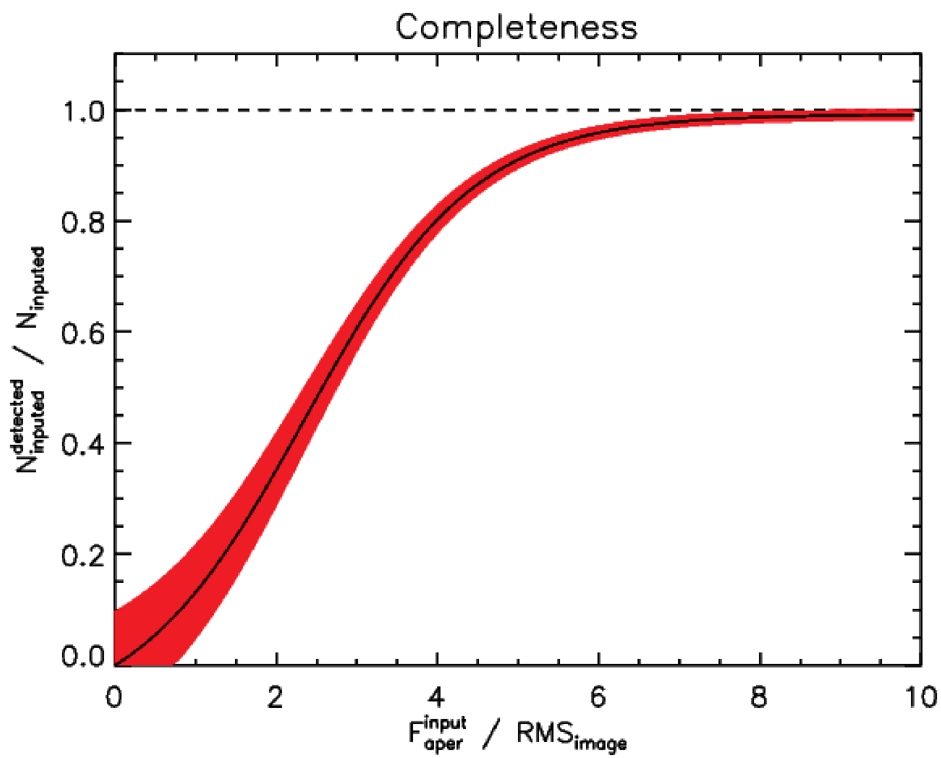
S_i is the flux density of the detection;

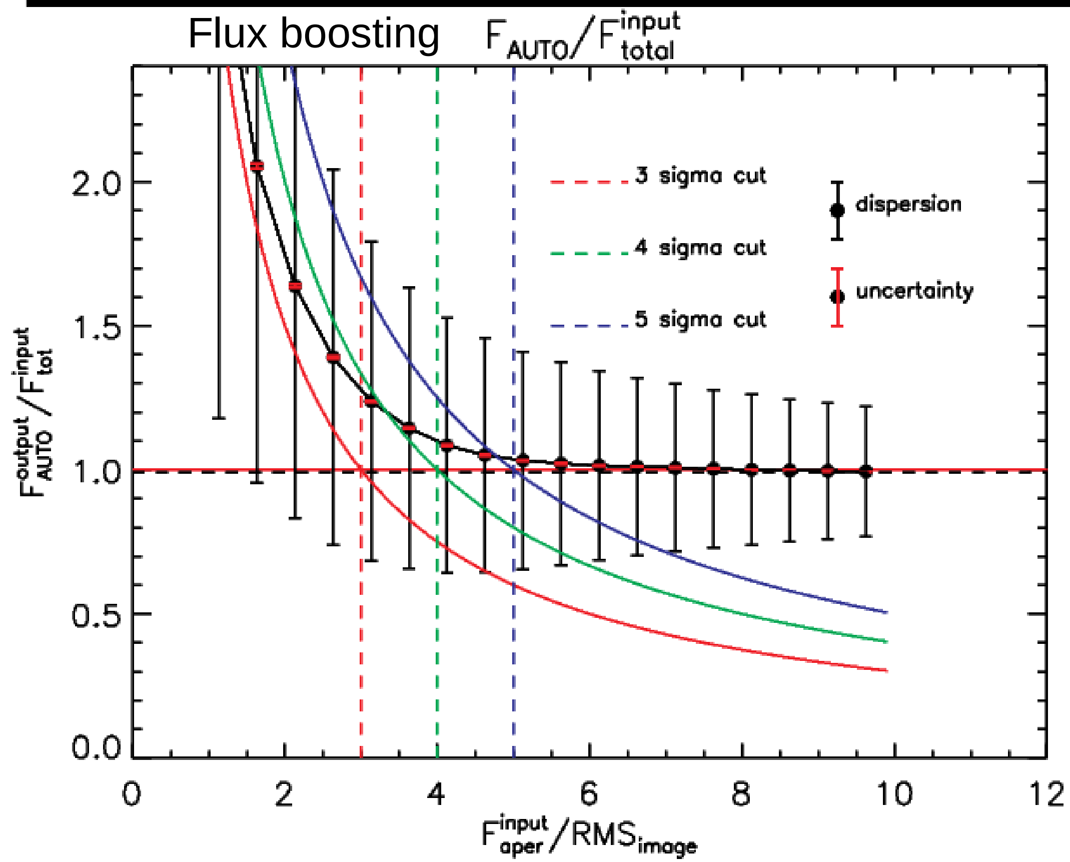
$f_{sp}(S_i)$ is the fraction of spurious sources at S_i,

$A(S_i)$ is the effective area at S_i,

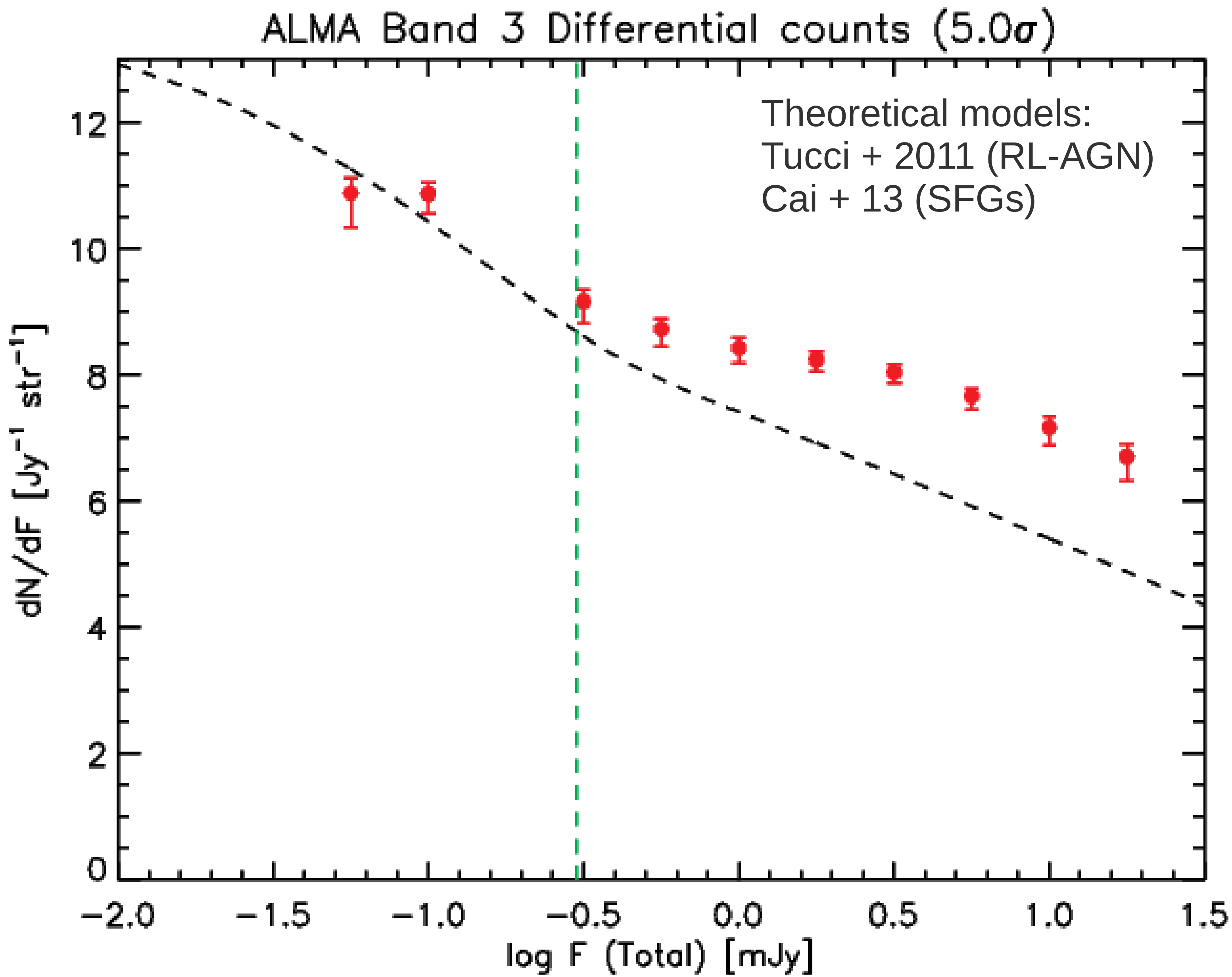
$C(S_i)$ is the completeness at S_i,



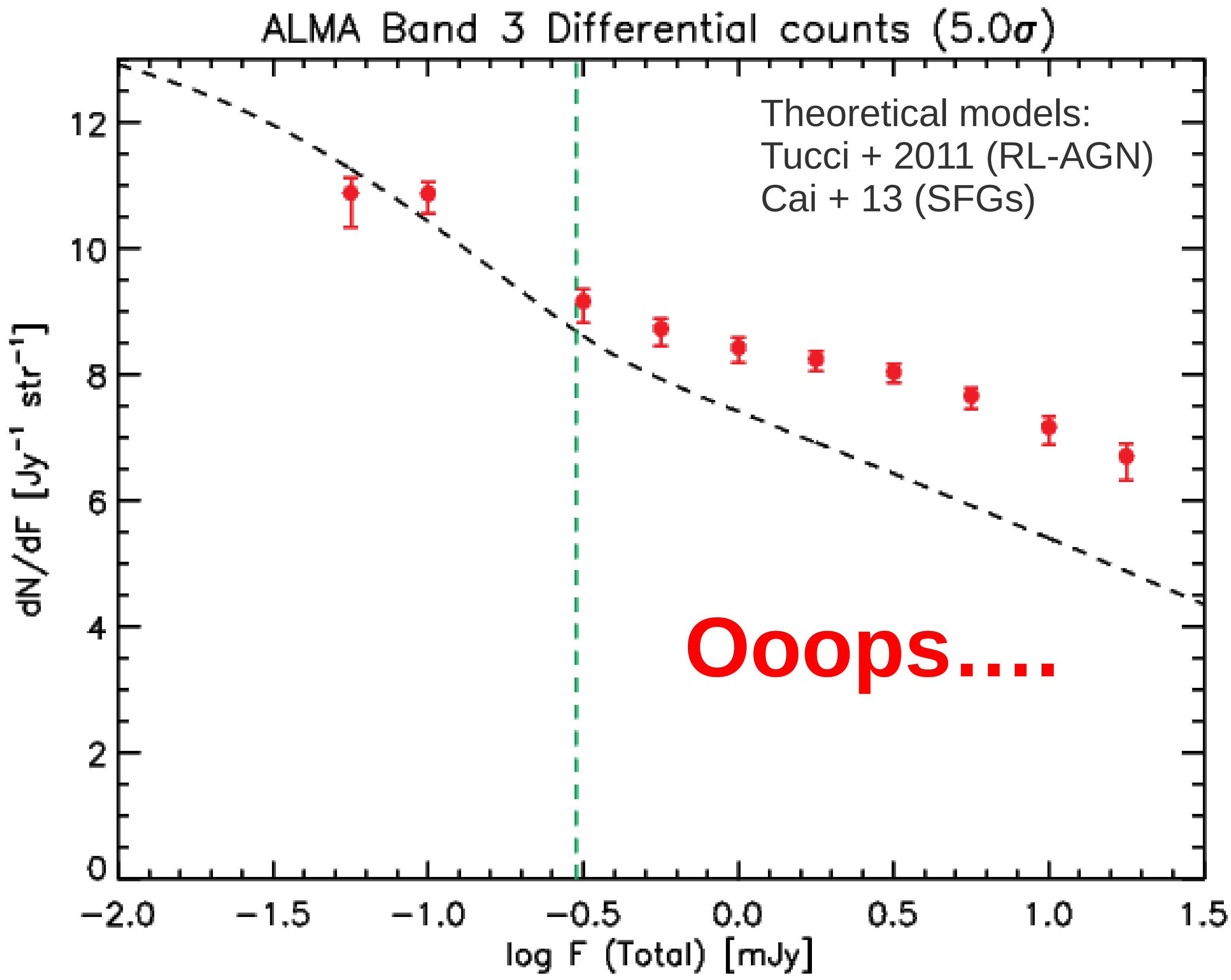




ALMA Band 3 number counts



ALMA Band 3 number counts



ALMA Band 3 number counts

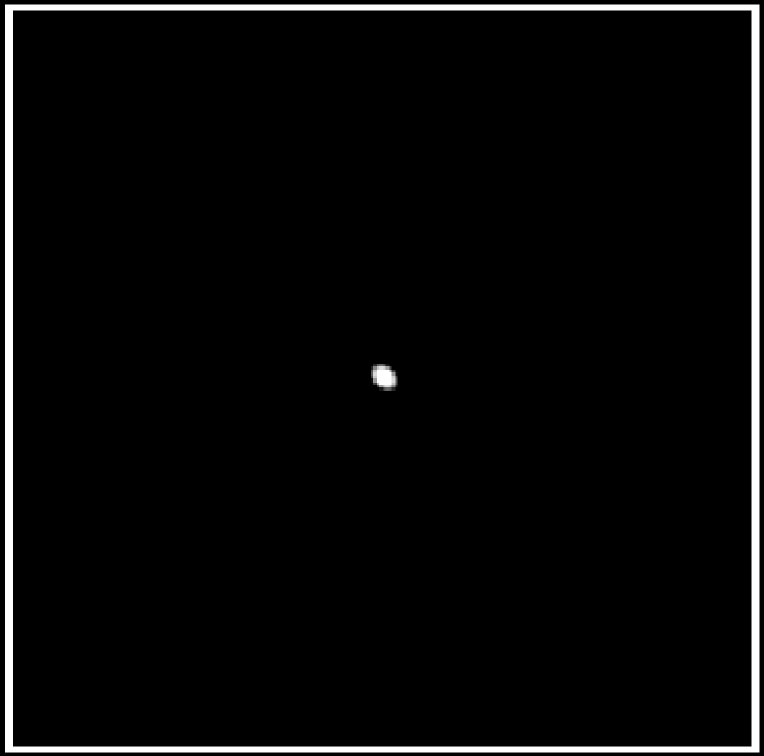


Image of a band 3 calibrator

- Are there any other serendipitous detections around?

Compute number counts !

ALMA Band 3 number counts

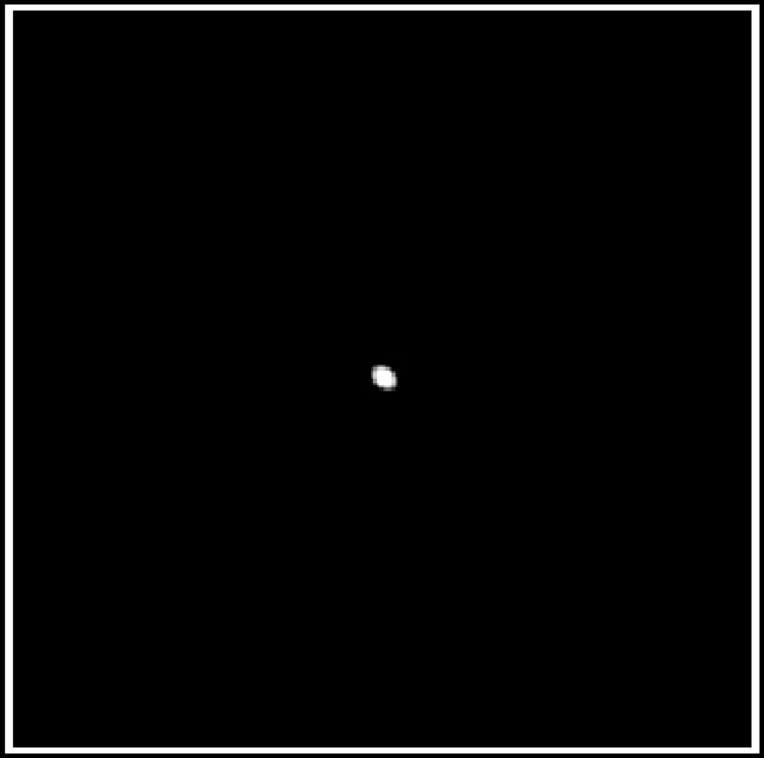
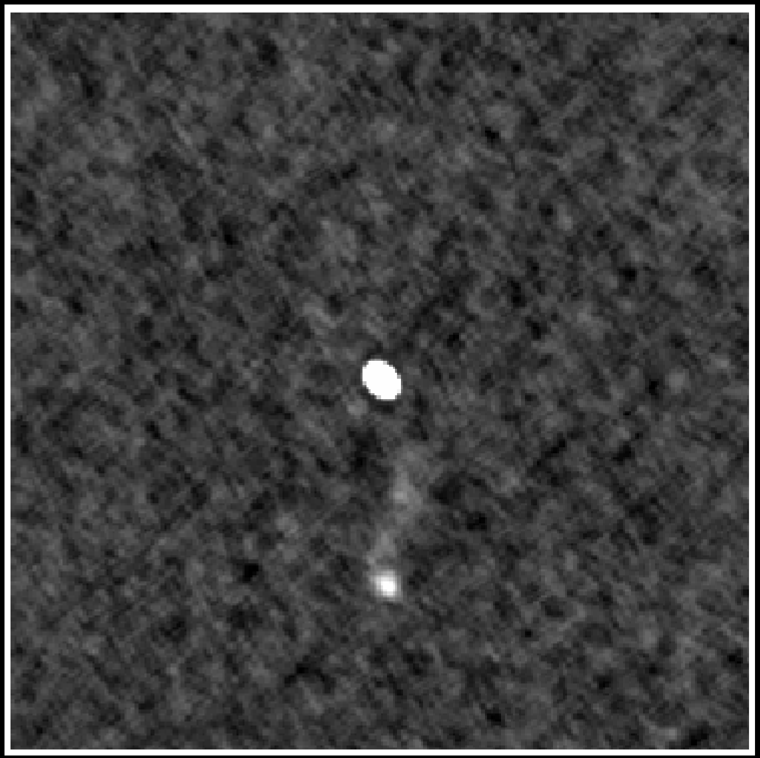


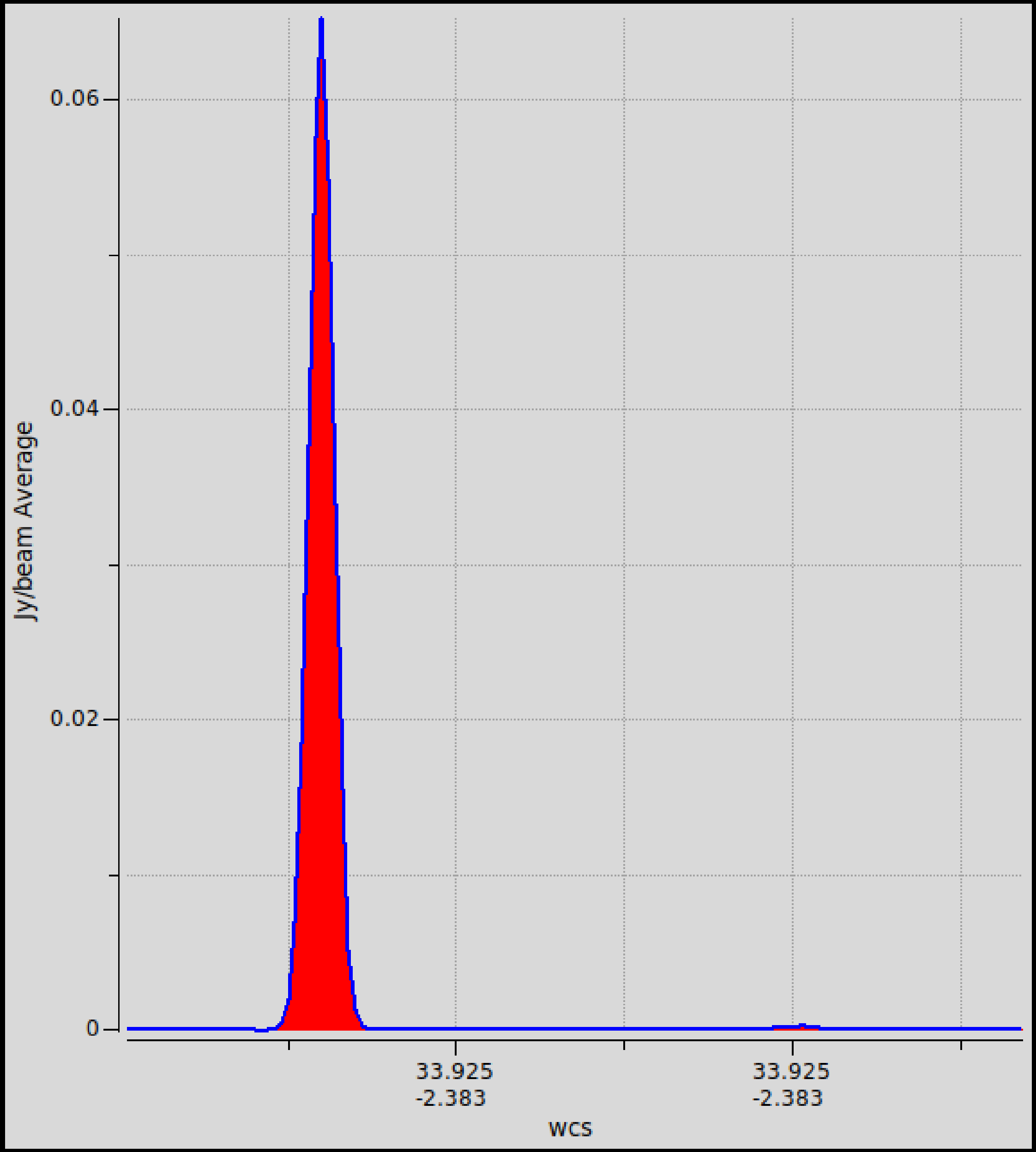
Image of a band 3 calibrator

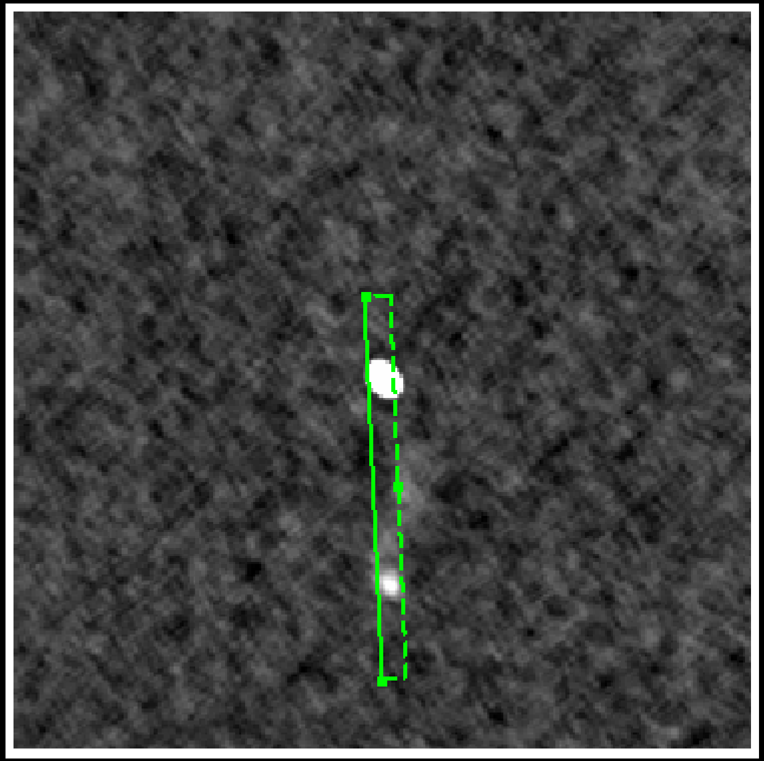


Same calibrator (after changing scale)

Extensions of the central calibrator are counted as (multiple) serendipitous sources

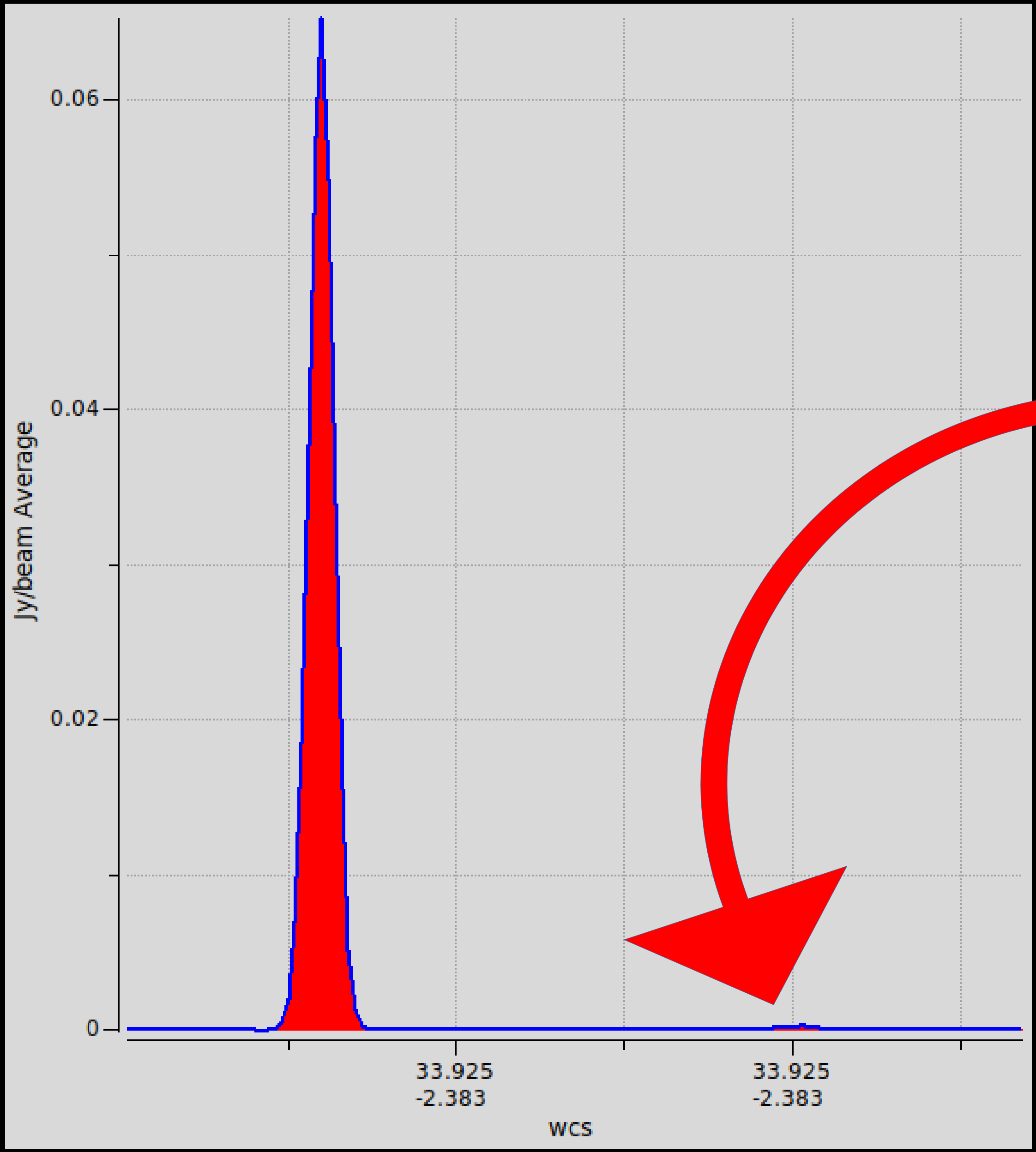
ALMA Band 3 number counts

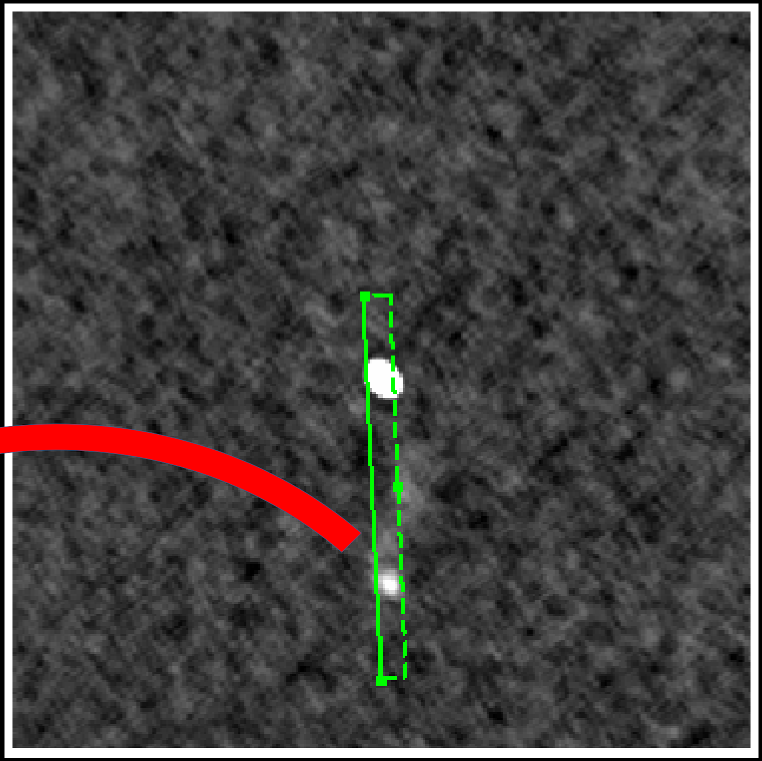




Same calibrator (after changing scale)

ALMA Band 3 number counts





Same calibrator (after changing scale)

ALMA Band 3 number counts

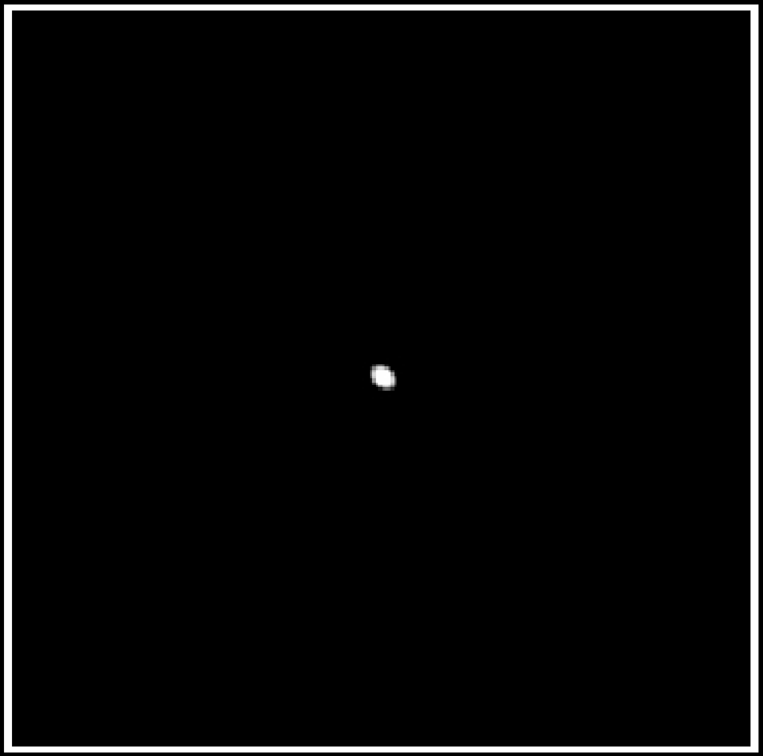
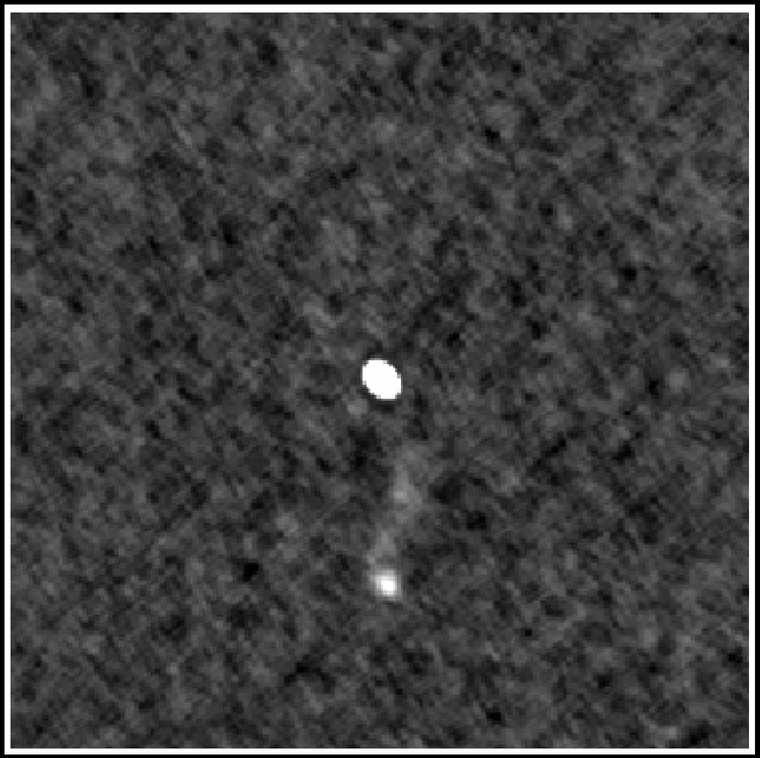


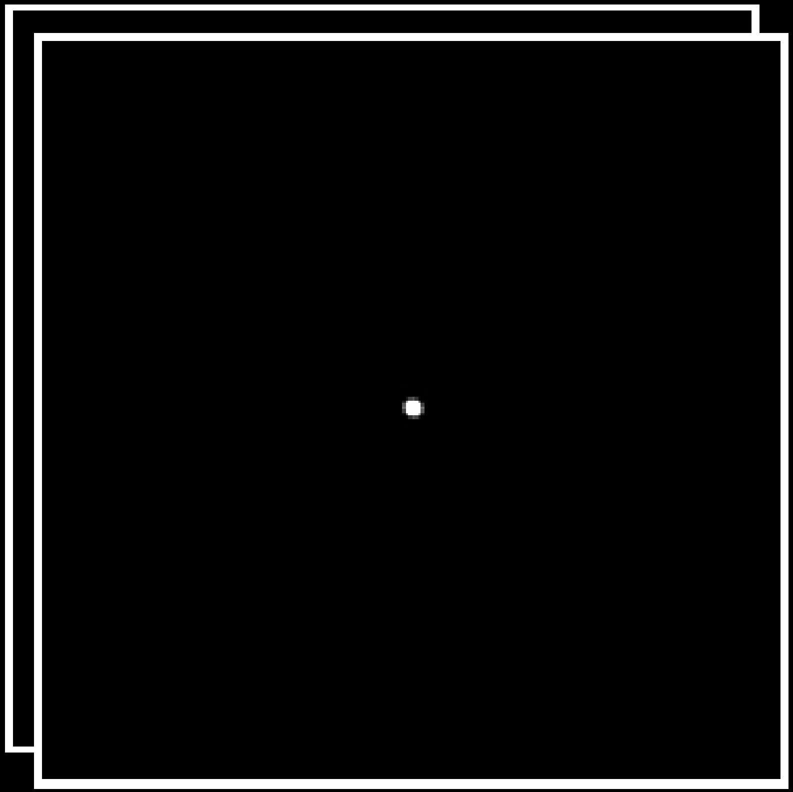
Image of a band 3 calibrator



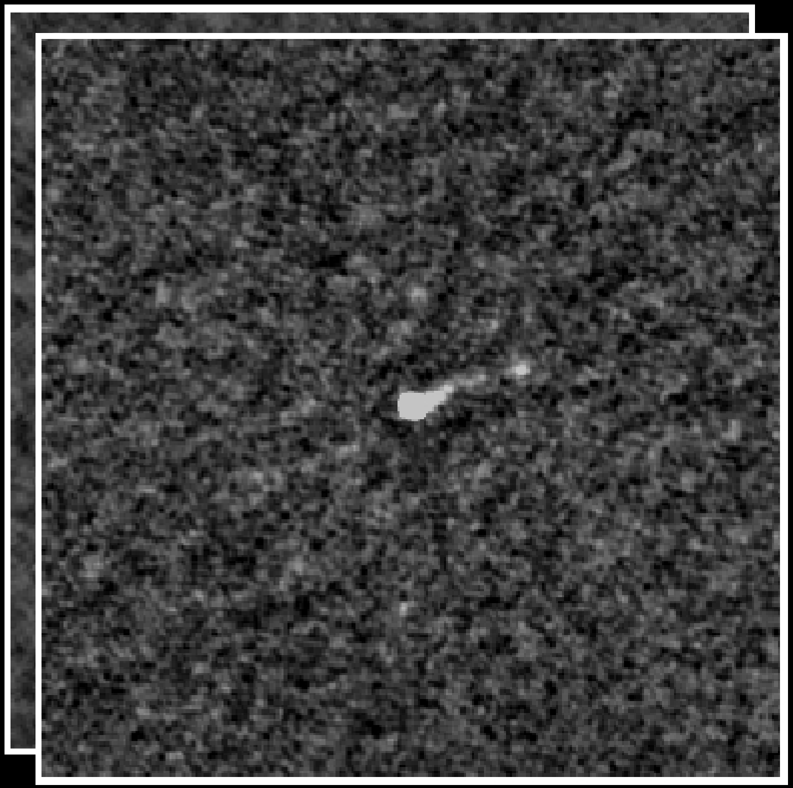
Same calibrator (after changing scale)

Extensions of the central calibrator are counted as (multiple) serendipitous sources

ALMA Band 3 number counts



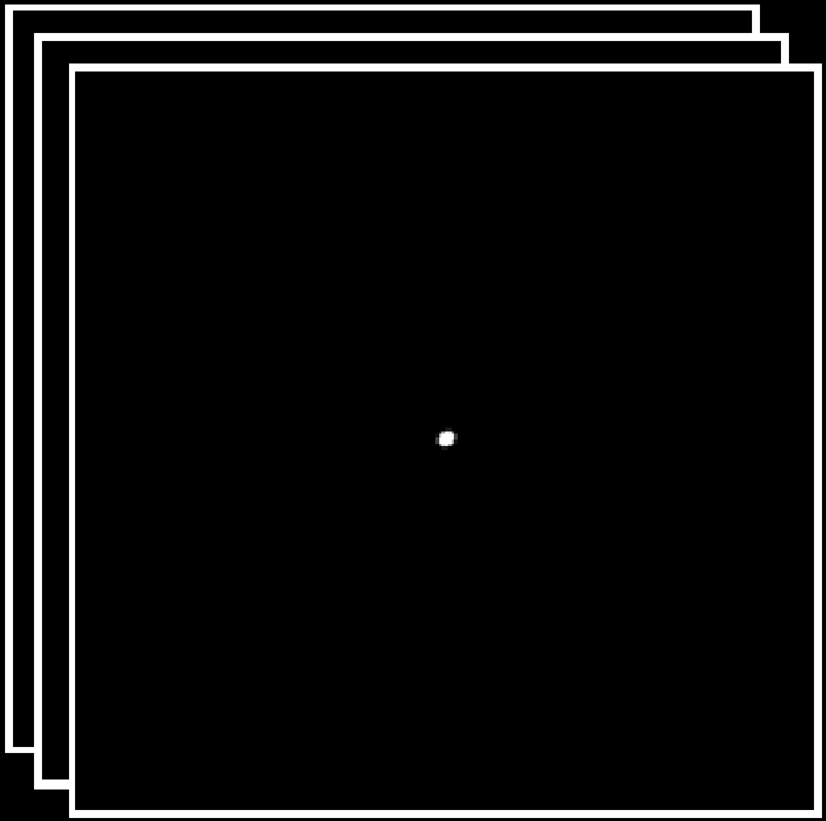
Another calibrator



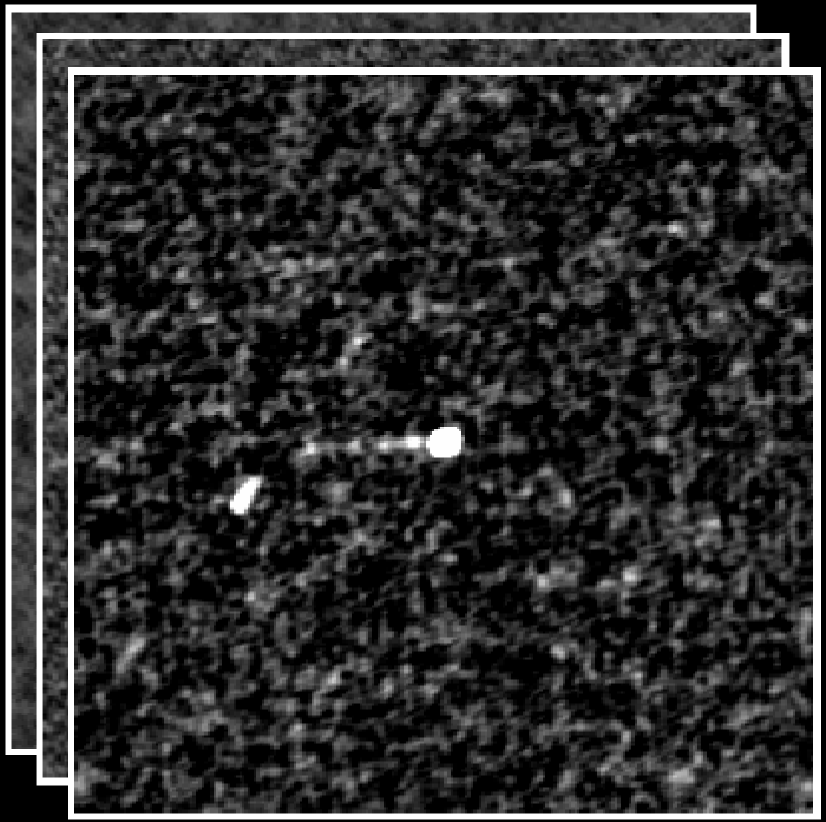
Same calibrator (after changing scale)

Extensions of the central calibrator are counted as (multiple) serendipitous sources

ALMA Band 3 number counts



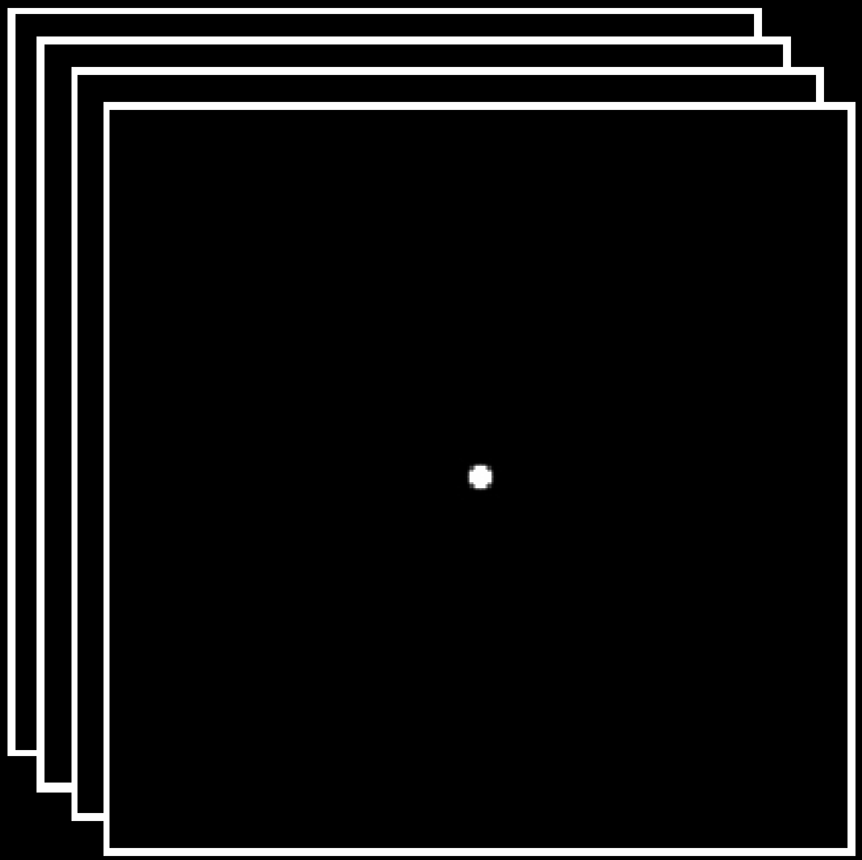
Another calibrator



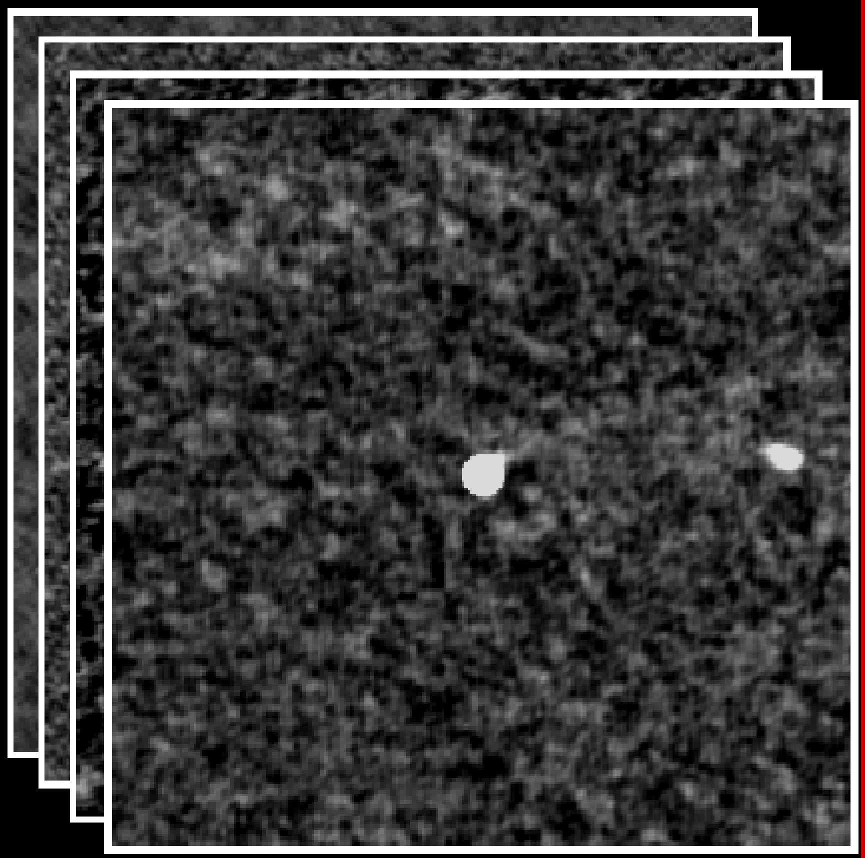
Same calibrator (after changing scale)

Extensions of the central calibrator are counted as (multiple) serendipitous sources

ALMA Band 3 number counts



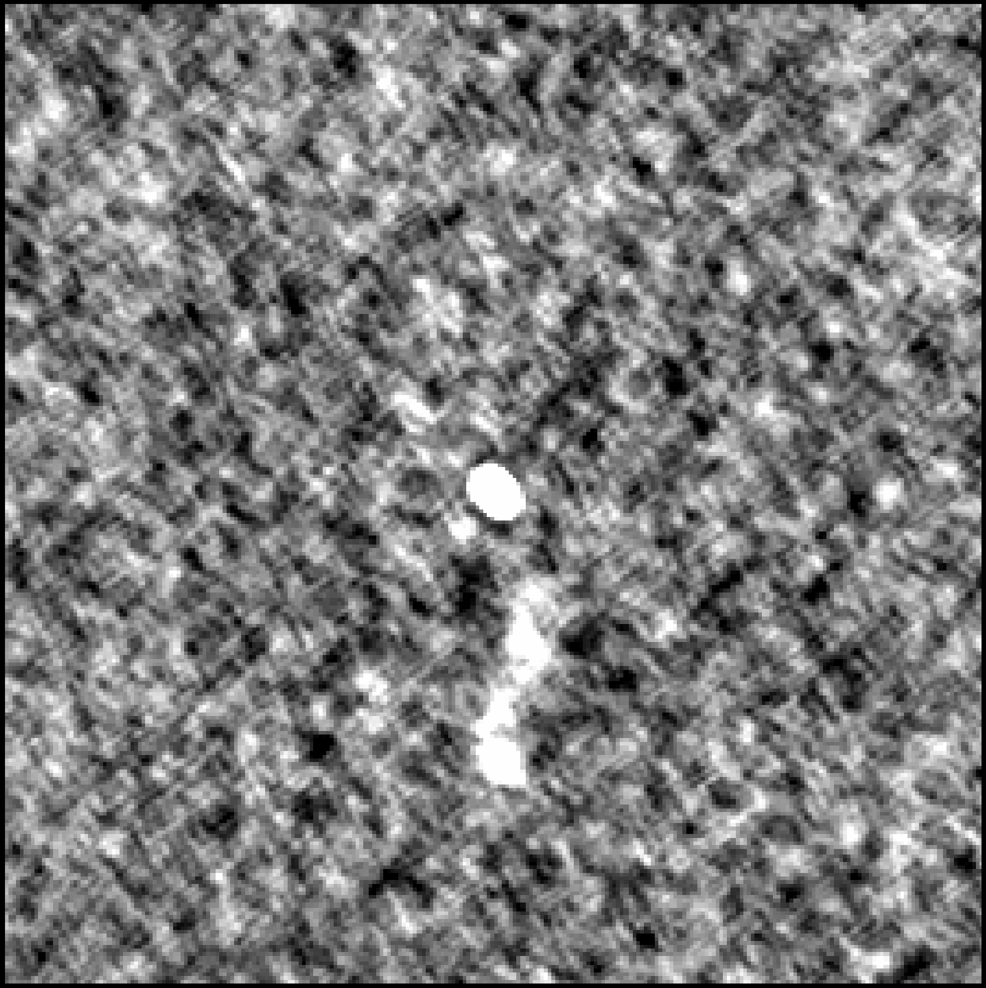
Another calibrator



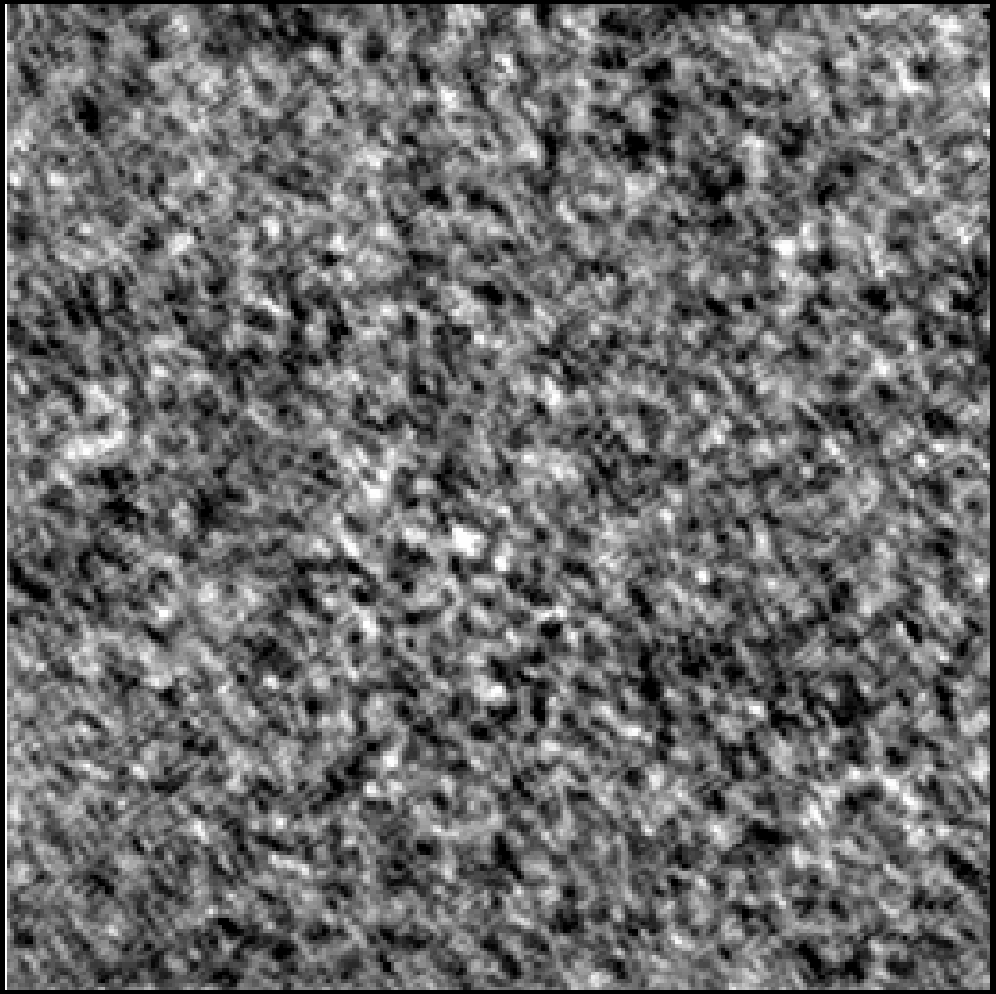
Same calibrator (after changing scale)

Extensions of the central calibrator are counted as (multiple) serendipitous sources → Down to a few sigma, it becomes more and more difficult to identify these contaminants by eye

ALMA Band 3 number counts - simulations

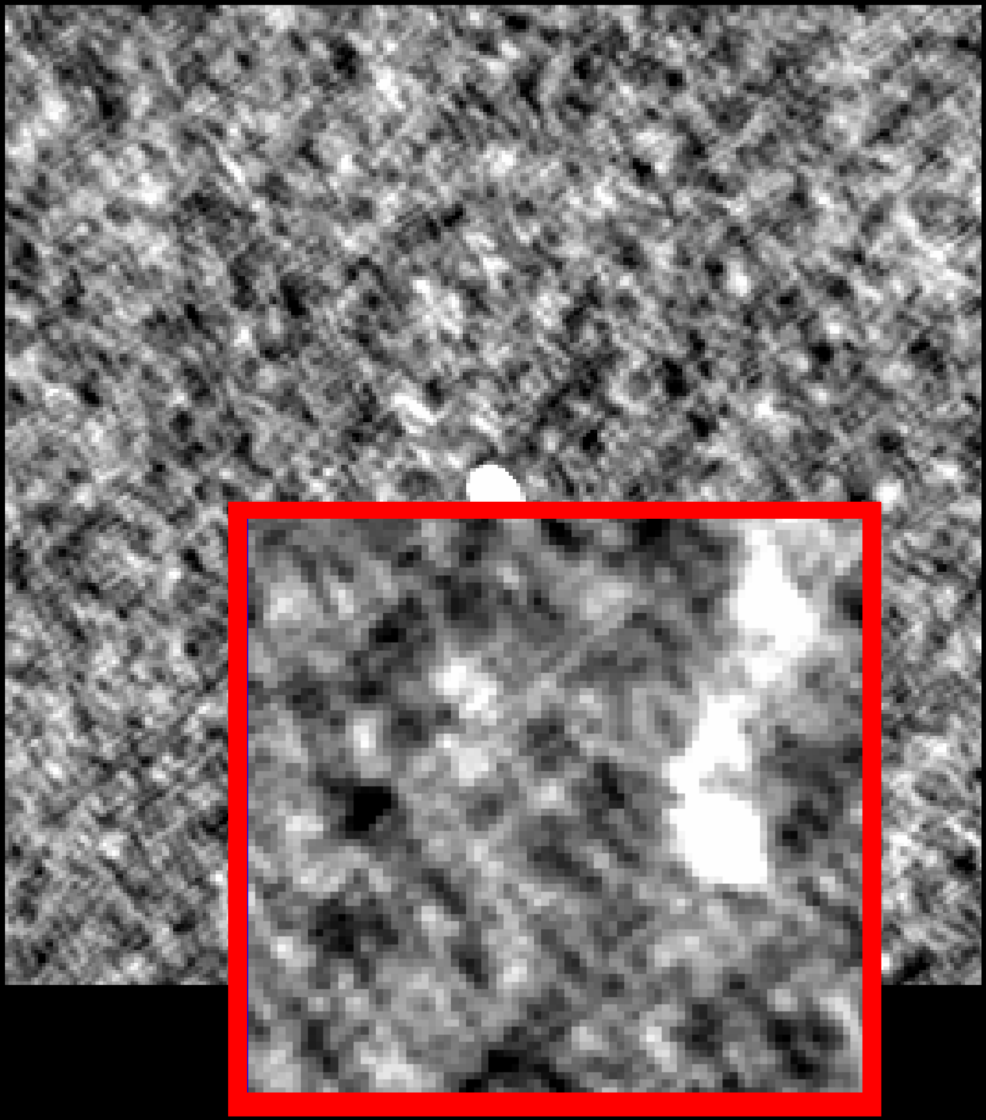


Original image

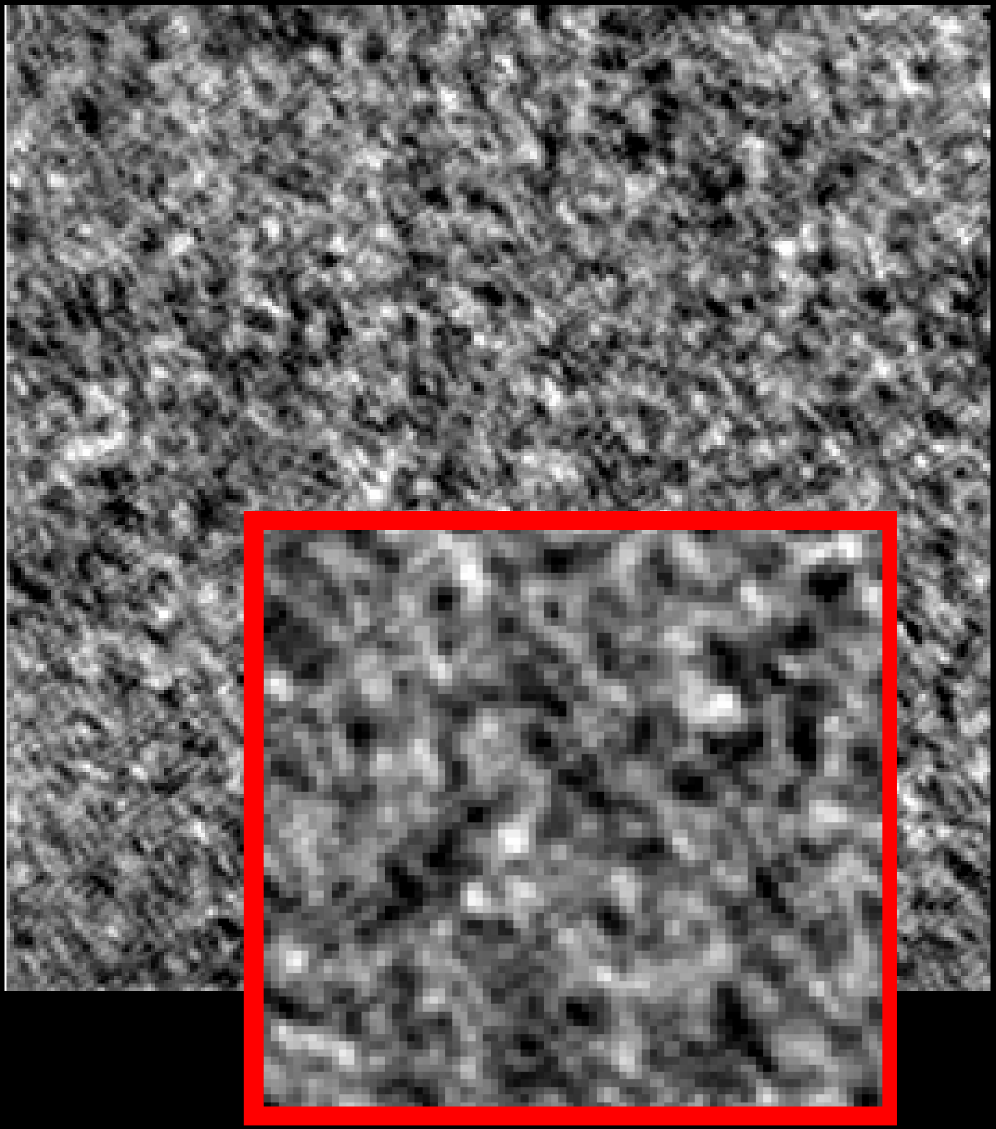


Simulated image

ALMA Band 3 number counts - simulations

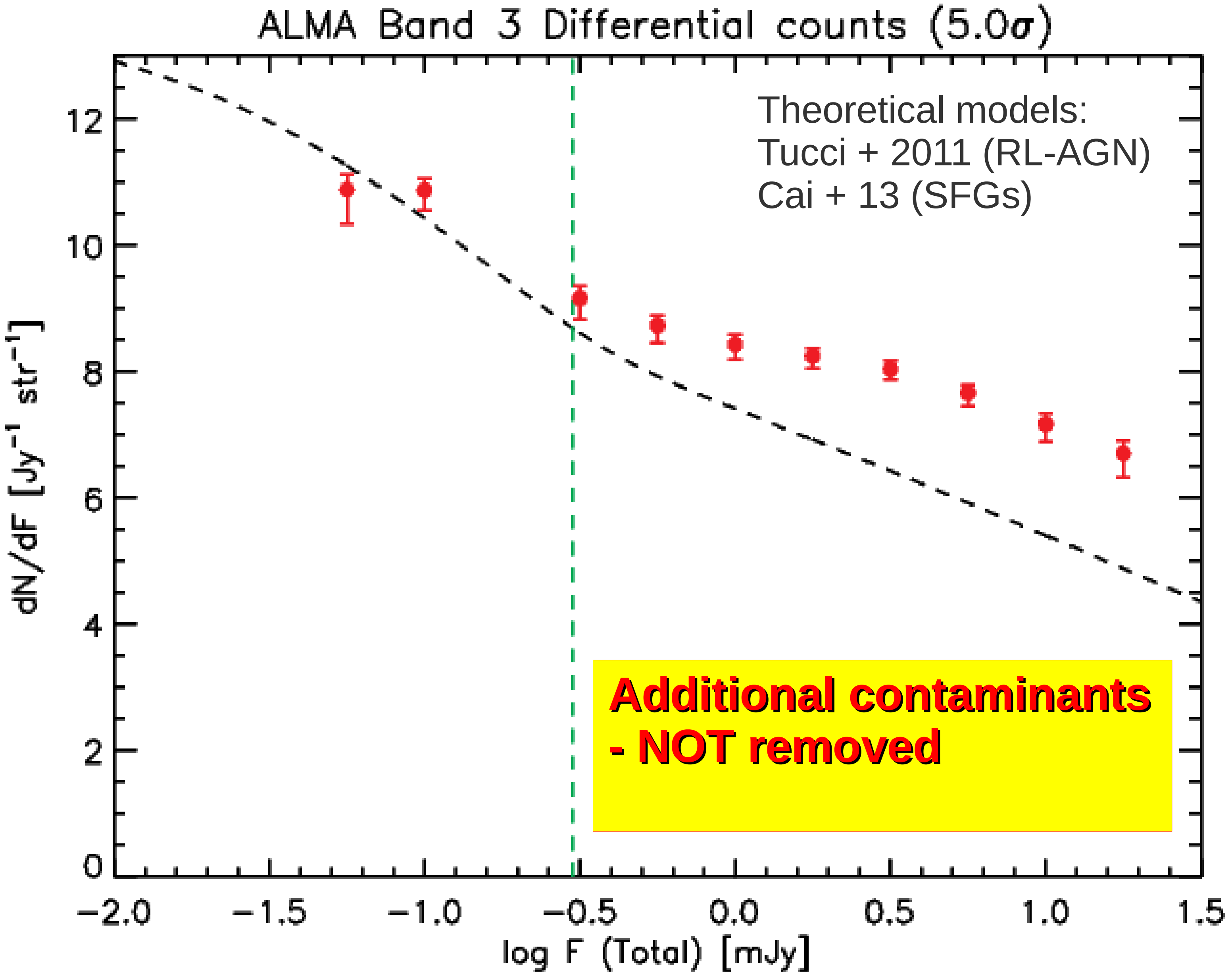


Original image

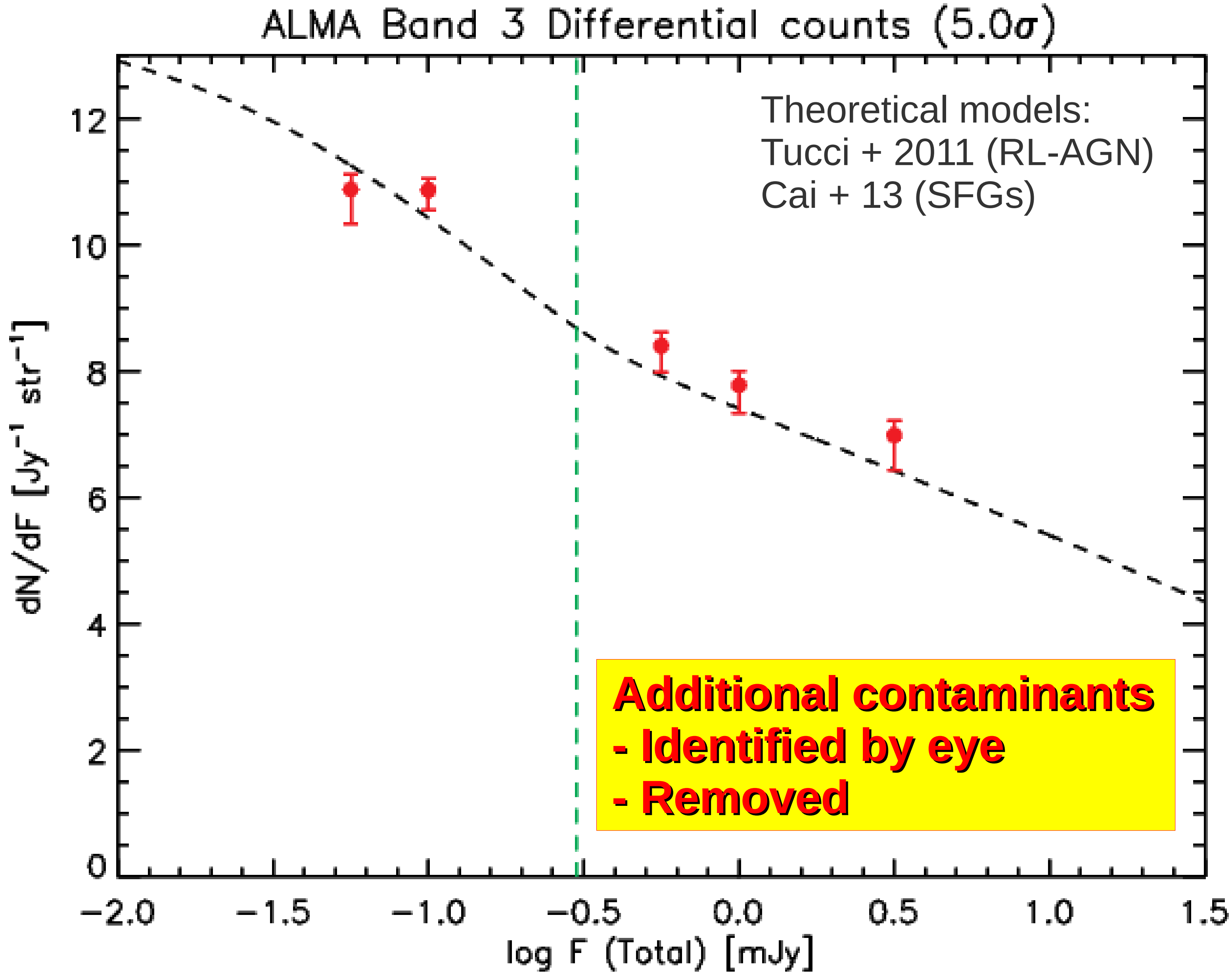


Simulated image

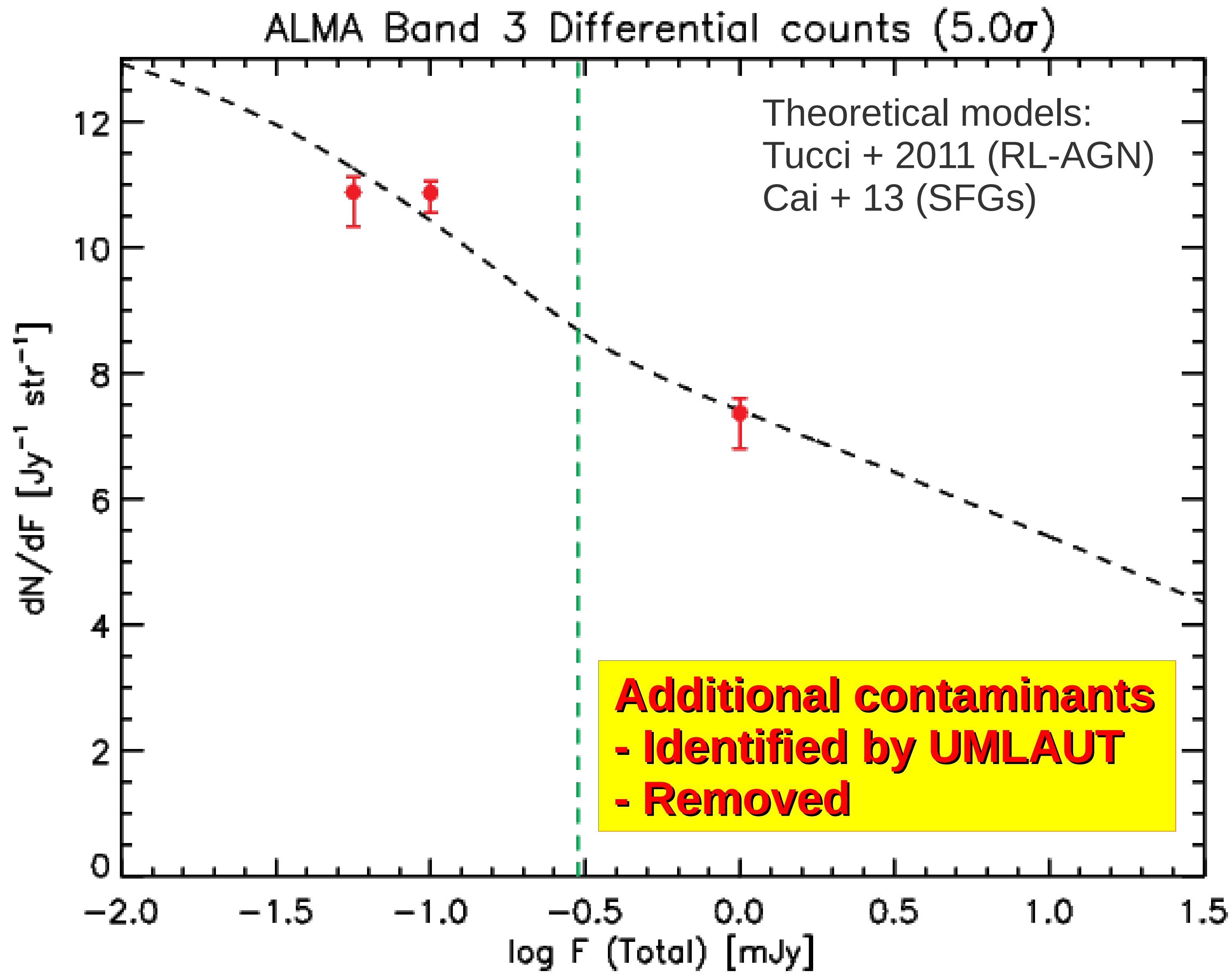
ALMA Band 3 number counts (5σ)



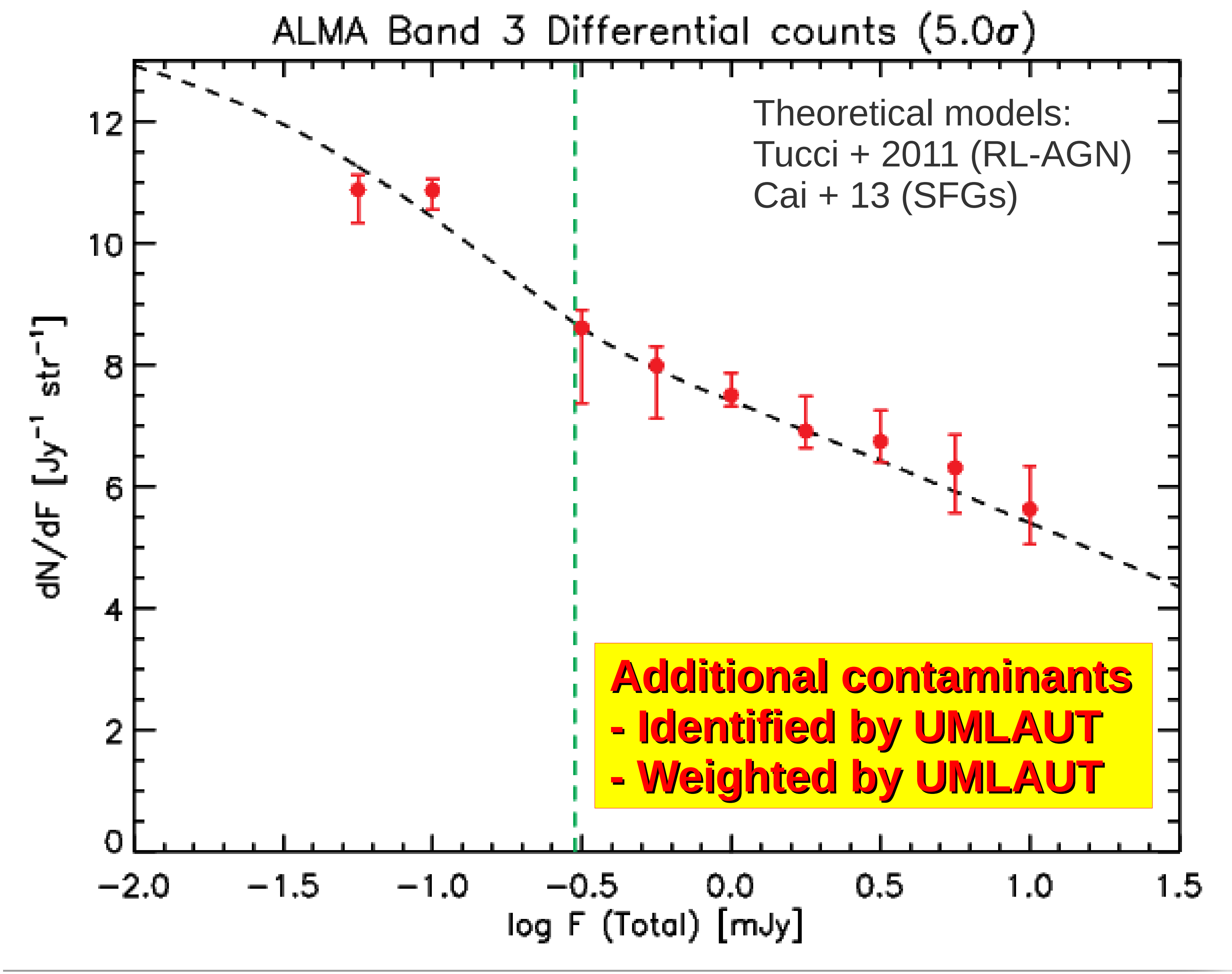
ALMA Band 3 number counts (5σ)



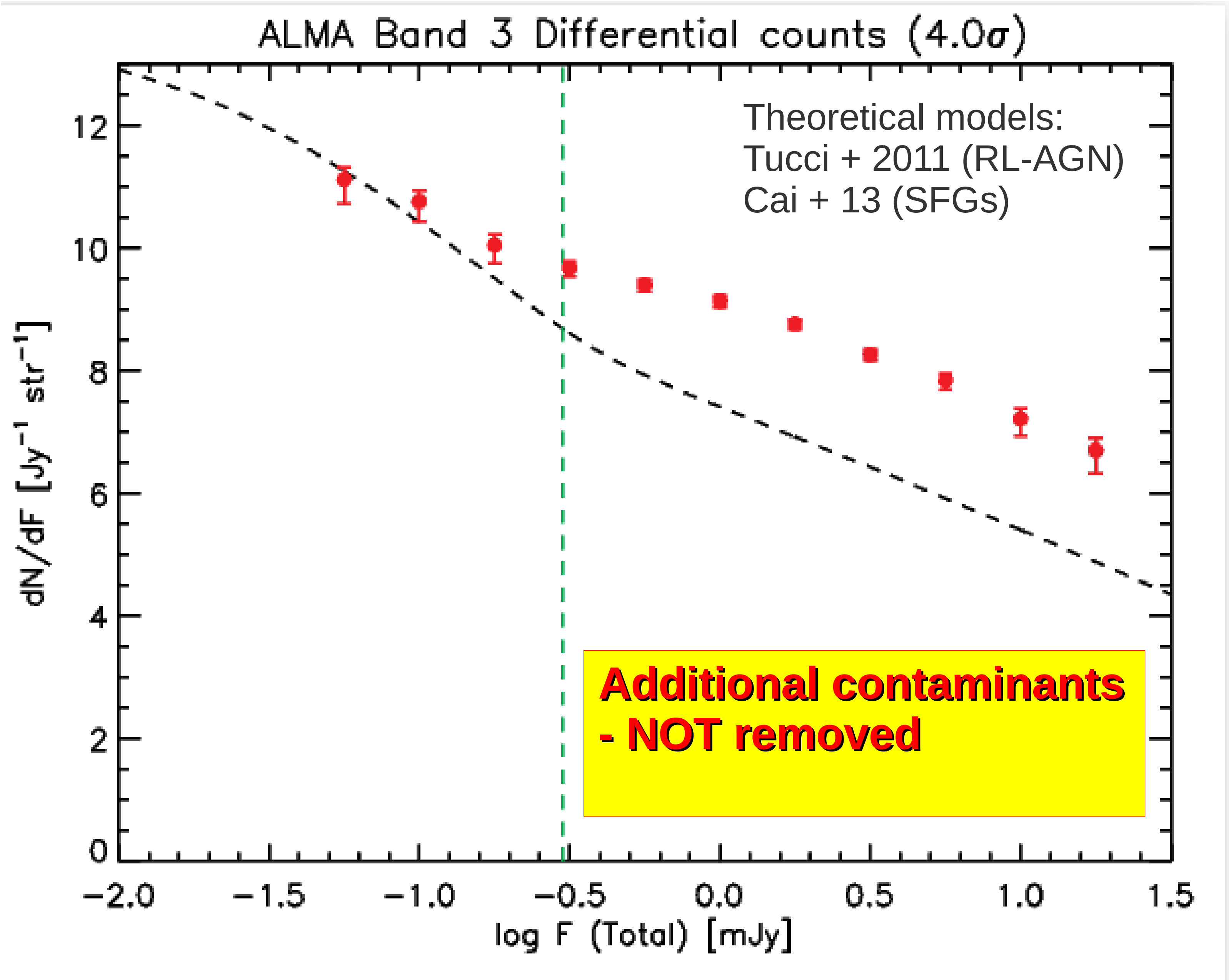
ALMA Band 3 number counts (5σ)



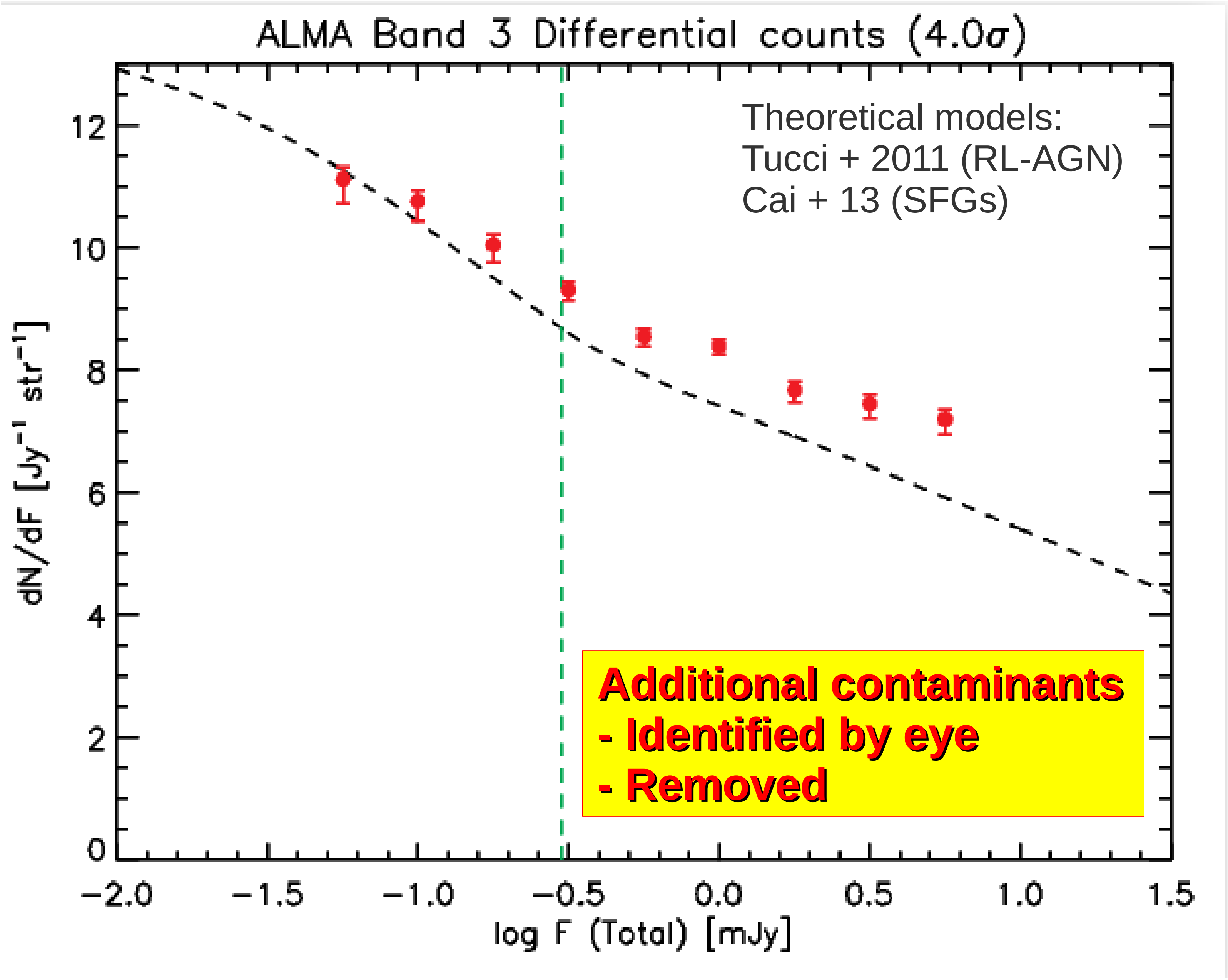
ALMA Band 3 number counts (5σ)



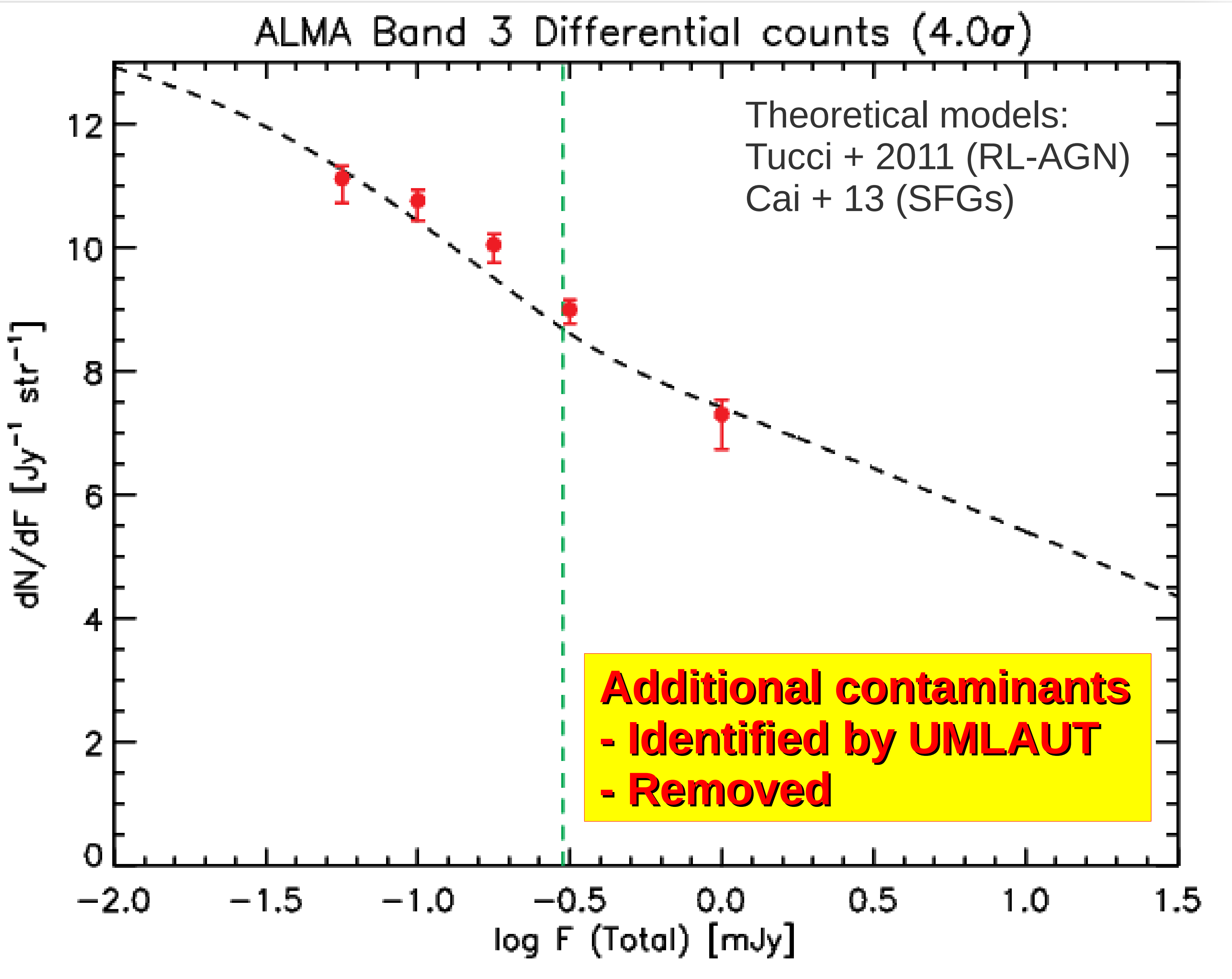
ALMA Band 3 number counts (4σ)



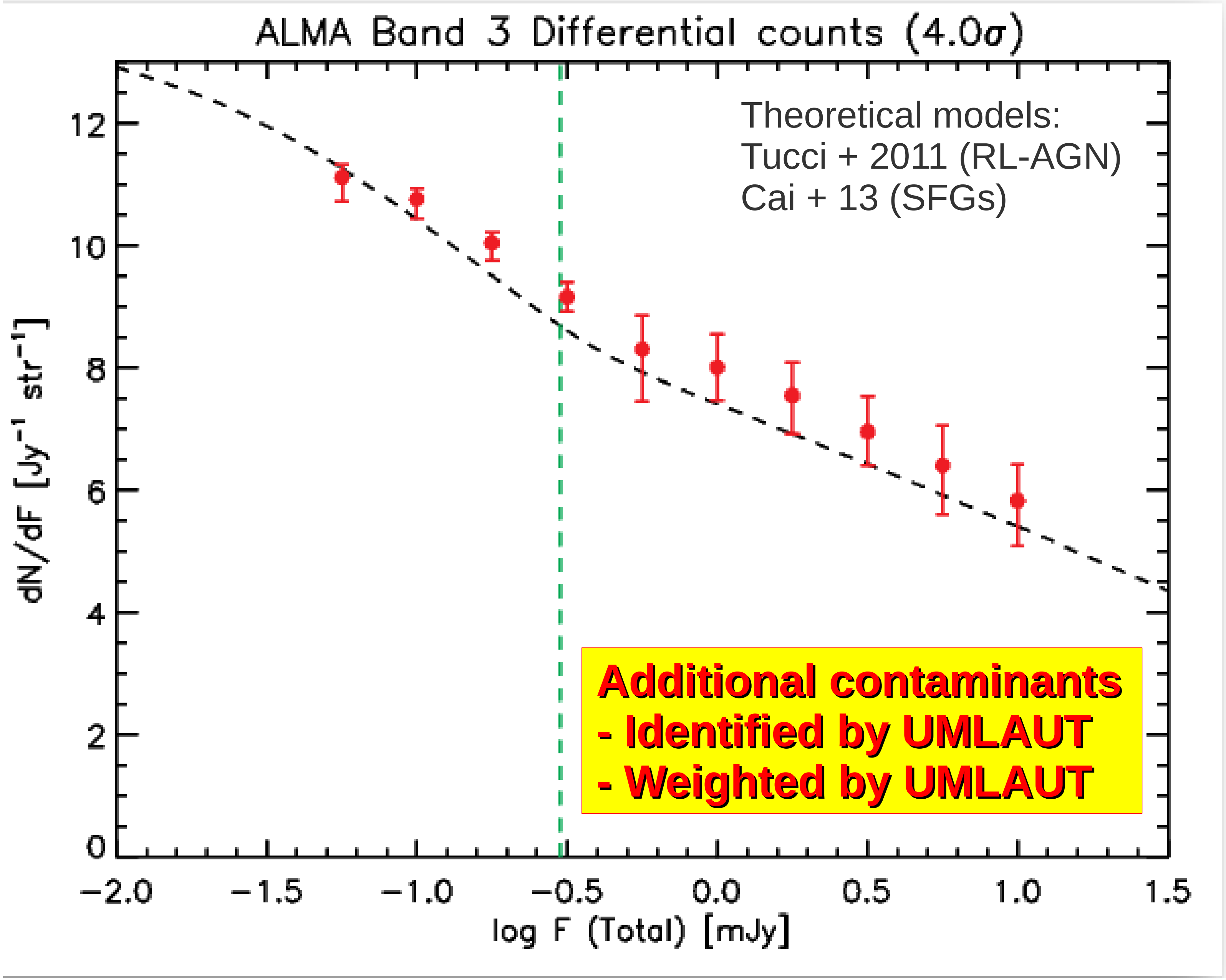
ALMA Band 3 number counts (4σ)



ALMA Band 3 number counts (4σ)



ALMA Band 3 number counts (4σ)



UMLAUT

Easily adaptable to a wide variety of problems

frontiers in Psychology

ORIGINAL RESEARCH
published: 11 July 2019
doi: 10.3389/fpsyg.2019.01528

Check for updates

Inter- vs. Intra-Speaker Variation in Mixed Heritage Syntax: A Statistical Analysis

Federica Cognola[1], Ivano Baronchelli[2]* and Evelina Molinari[3]*

[1] Dipartimento di Studi Europei, Americani e Interculturali, Sapienza Università di Roma, Rome, Italy, [2] Dipartimento di Fisica e Astronomia, Università di Padova, Padua, Italy, [3] Istituto Culturale Mòcheno, Palù del Fersina, Italy

OPEN ACCESS

Edited by:
Ángel J. Gallego, Autonomous University of Barcelona, Spain

Reviewed by:
Kleanthes K. Grohmann,

Based on the novel data pertaining to five syntactic phenomena (the position of the finite verb in embedded clauses, in sentences with a modal verb, negative concord, the position of focused light/heavy objects in main clauses with a complex tense and scrambling) in the heritage language Mòcheno collected via original fieldwork, we show that there are two populations – one exhibiting intra-speaker variation between German and Italian word orders, and one lacking it; and these two populations are the result of diatopic variation and, to a lesser extent, of diastratic variation. The results achieved using quantitative statistical analysis are partially convergent with those arrived at via the traditional theoretical syntax for Mòcheno, but our analysis has allowed us to shed new light on a series of phenomena that have been neglected or poorly understood thus far. More specifically, and for the first time, we discovered that there is a micro-variation resulting from diastratic (age) variation within the Roveda variety, which represents the

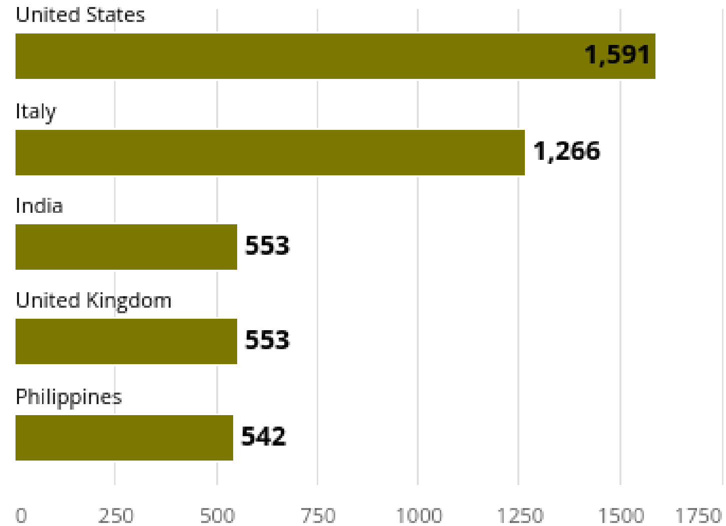
Views

8,866

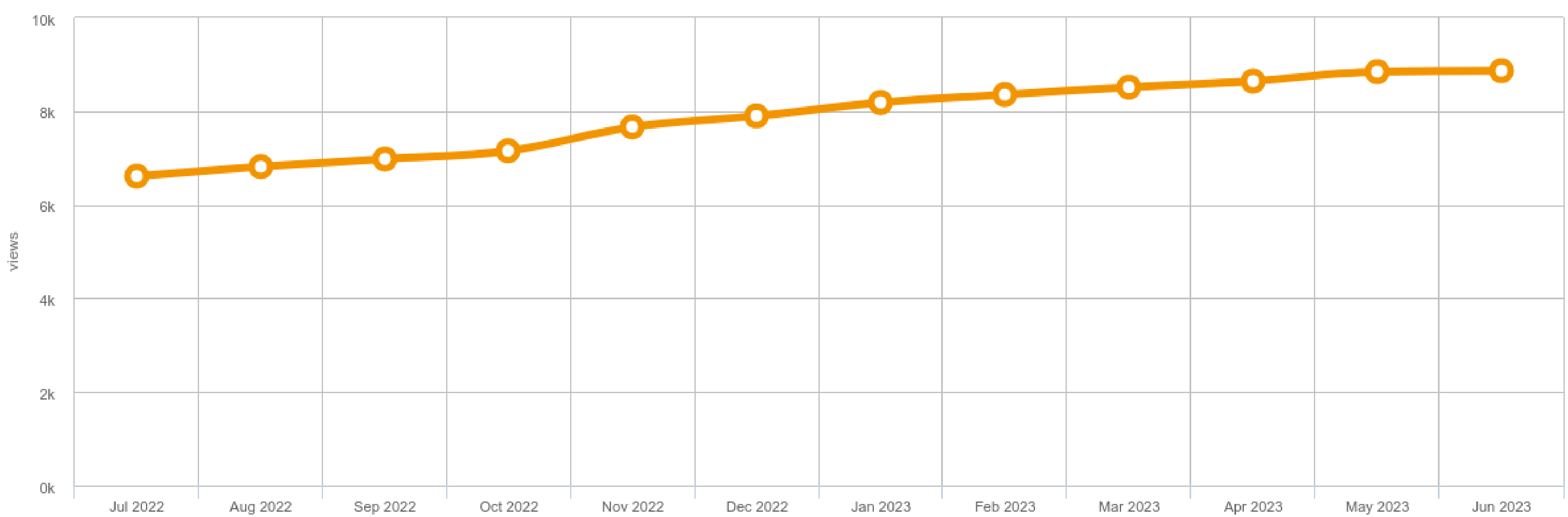
Frontiers

This article has more views than **93%** of all **Frontiers** articles.

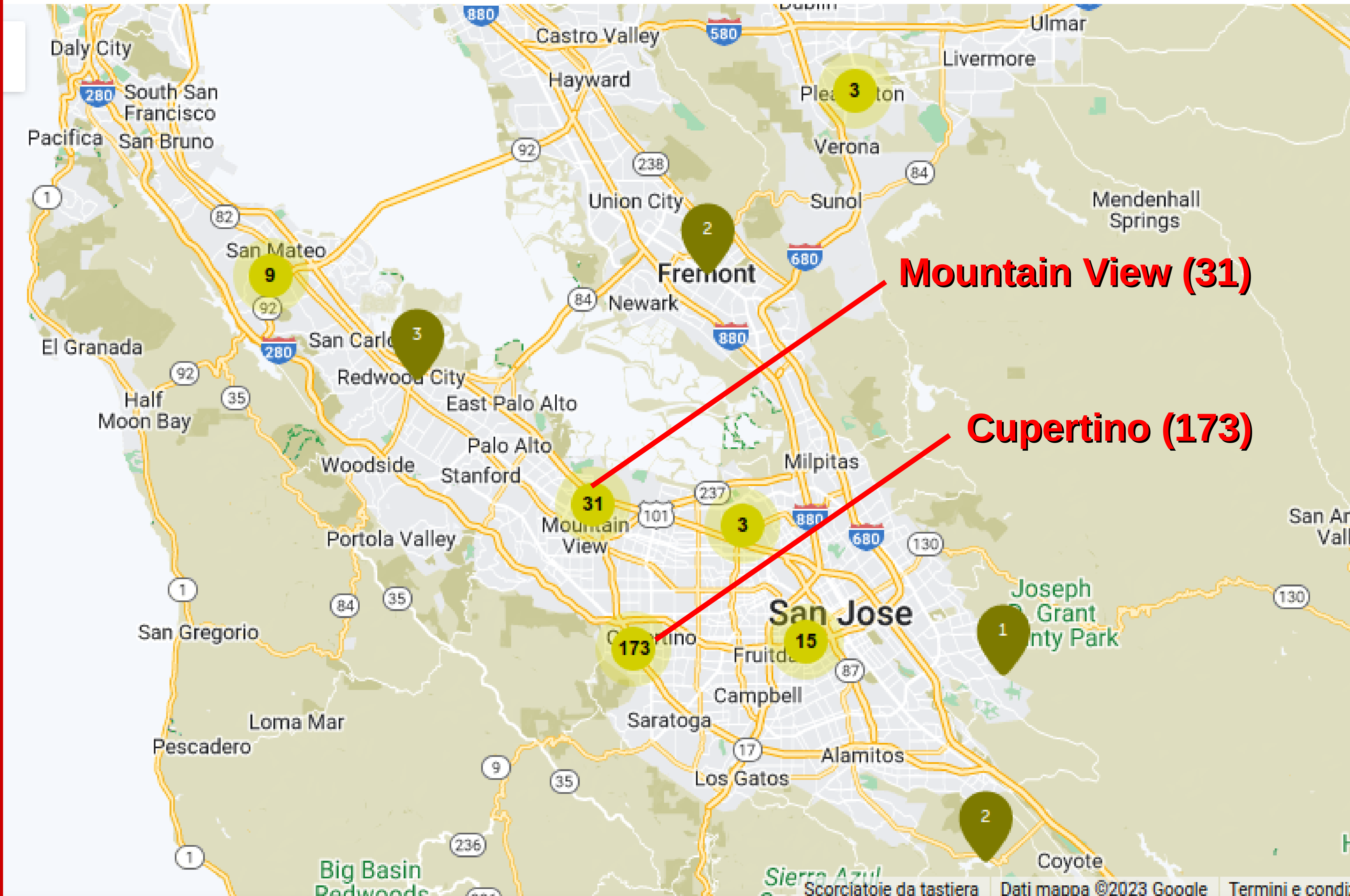
Top countries

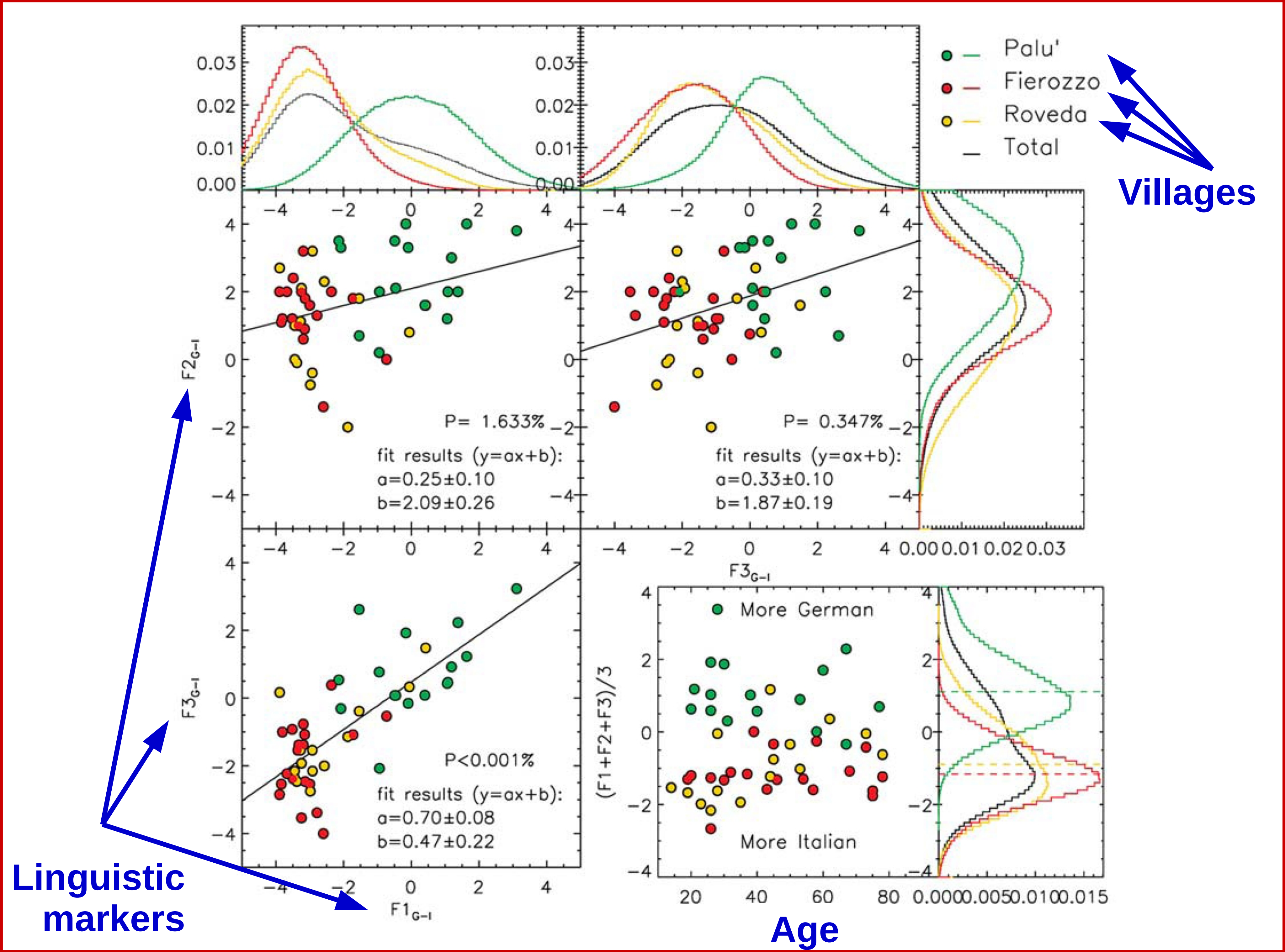


Last 12 months

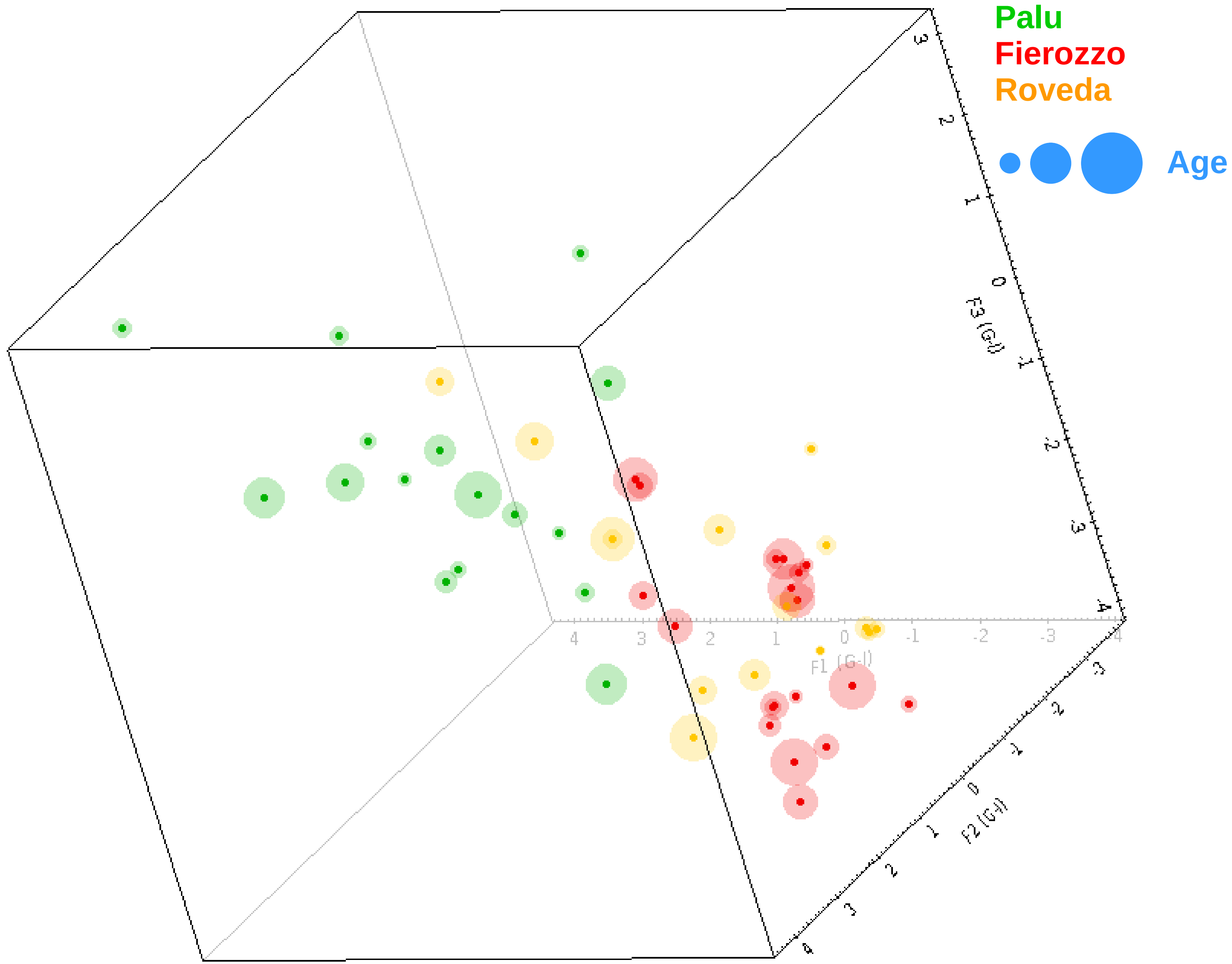


Demographics



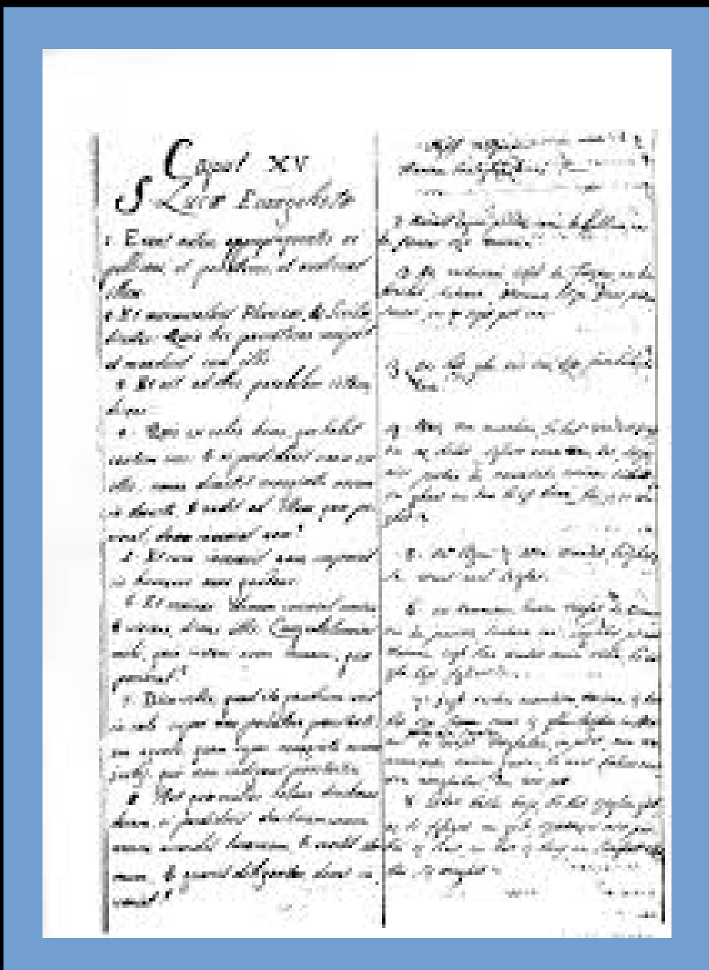


Palu'
Fierozzo
Roveda
Total
Villages
P= 1.633%
fit results (y=ax+b):
a=0.25±0.10
b=2.09±0.26
P= 0.347%
fit results (y=ax+b):
a=0.33±0.10
b=1.87±0.19
P<0.001%
fit results (y=ax+b):
a=0.70±0.08
b=0.47±0.22
$F2_{G-I}$
$F3_{G-I}$
$F1_{G-I}$
More German
More Italian
(F1+F2+F3)/3
Linguistic markers
Age


Palu
Fierozzo
Roveda
Age
F1 (G-I)
F2 (G-I)
F3 (G-I)

1811 – Napoleon wants a census of all the dialects spoken in the Italian kindom

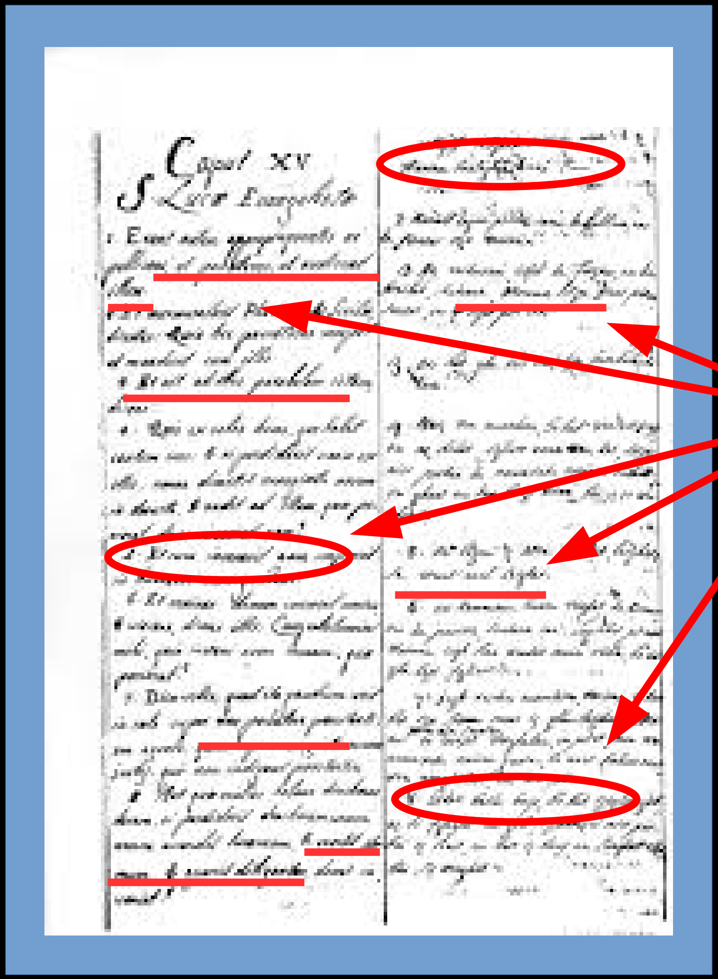
→ **First important official document written in Mocheno language**



Who wrote it?

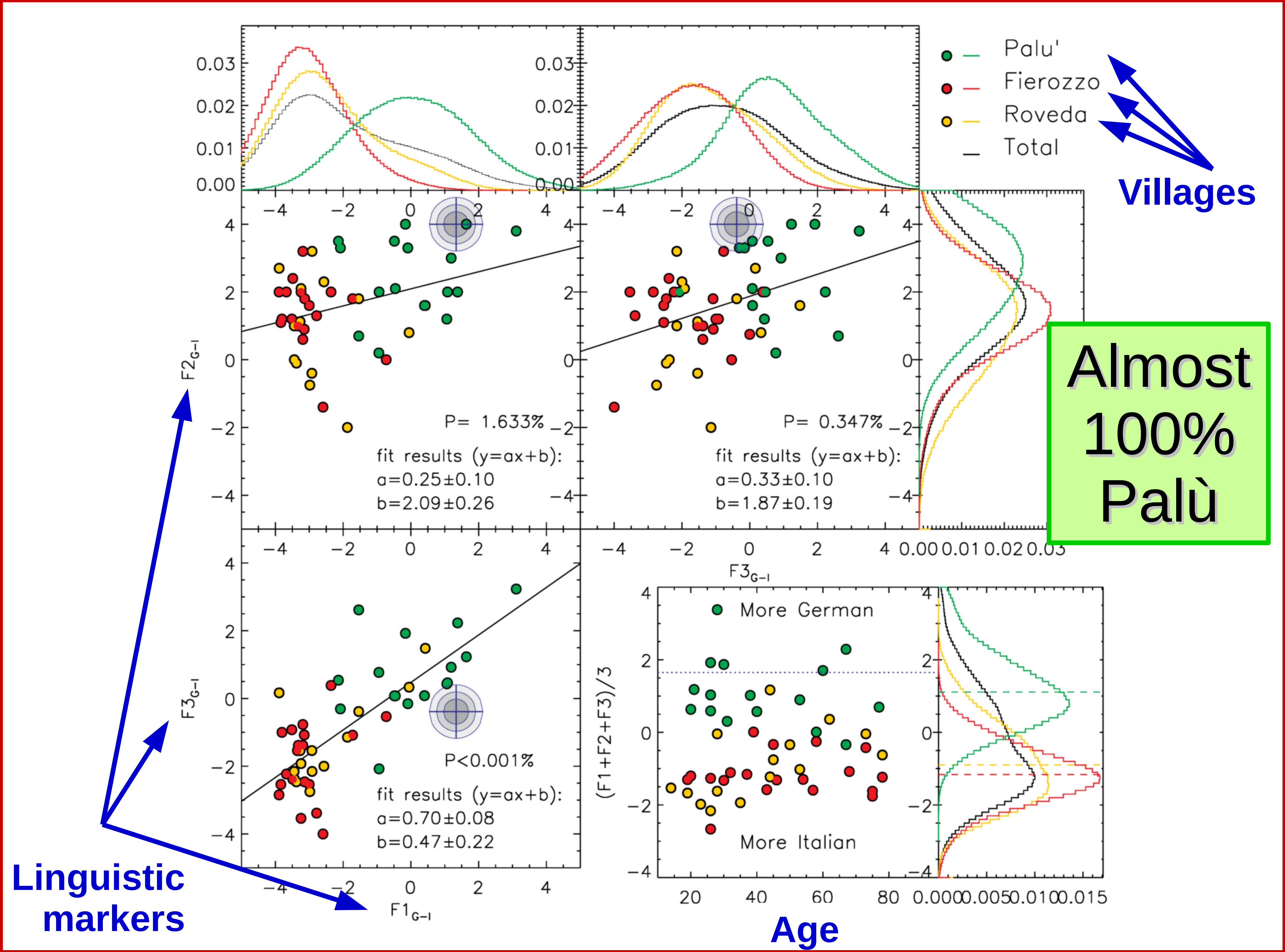
1811 – Napoleon wants a census of all the dialects spoken in the Italian kindom

→ First important official document written in Mocheno language



Who wrote it?

Linguistic markers



Palu'
Fierozzo
Roveda
Total
Villages
Almost 100% Palù
P= 1.633%
fit results (y=ax+b):
a=0.25±0.10
b=2.09±0.26
P= 0.347%
fit results (y=ax+b):
a=0.33±0.10
b=1.87±0.19
P<0.001%
fit results (y=ax+b):
a=0.70±0.08
b=0.47±0.22
F2G-I
F3G-I
F1G-I
(F1+F2+F3)/3
More German
More Italian
Linguistic markers
Age

Ongoing development:

- Code optimization (same algorithm - faster run);

- Translate UMLAUT in different languages (Python)

- Parallelization of the code (NVIDIA-CUDA);

- Improvement of the current dimensionality reduction strategy
  - <u>example1:</u> using a PCA-like strategy
  - <u>example2:</u> by replacing each of the input dimensions with random ranks and comparing the output variance.

Conclusions:

- New facilities producing unprecedented amount of data are going to be available in the next few years: <u>every aspect of the treatment of such data must be automated;</u>

UMLAUT is:

- a ML algorithm based on the analogy paradigm;
- extremely (and easily) adaptable to very different contexts;
- freely available online

3 applications were presented:

- Identification of single spectral lines
- Recognition of ancient linguistic variants
- Identification of contaminants in sub-mm number counts