Sleuth

A quasi-model-independent search strategy for new physics

Introduced by Prof. Bruce Knuteson in 2000 –
arXiv:hep-ex/0311059 or hep-ex/0006011

Ran Feldesh Ben Gurion Uni. Ehud Duchovni, Weizmann Inst.

Sleuth

- Introduction of the Problem
- Solution suggested by Sleuth
- Comparison with other search methods in Particle Physics
- Implementations for SUSY searches
- Current Status and Future Plans of Wiezmann-Sleuth

The Problem:
Identify if a **set of events** fits to **a given distribution**

In Our Particle-Physics realm:

- The ***set of events*** is the final states driven out of the raw detector's data
- The ***given distribution*** is the prediction of the Standard Model

The Question Asked:
What is the chance that the point at x=61 comes out of new physics? Is it due to uncertainties in the measurements or flag for new physics?
After all the chance of seeing *exactly* this event is zero . We need to define a *notion for regions*

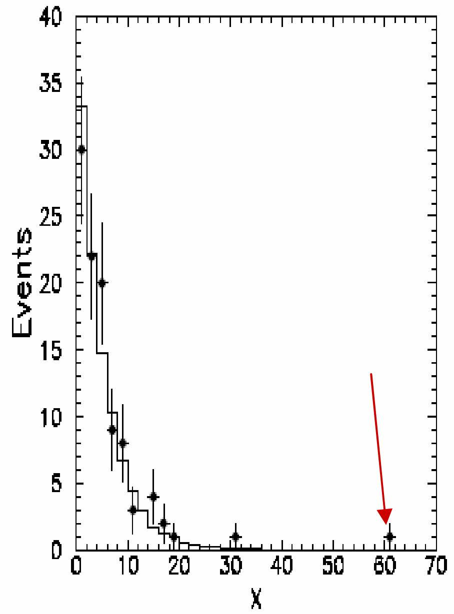
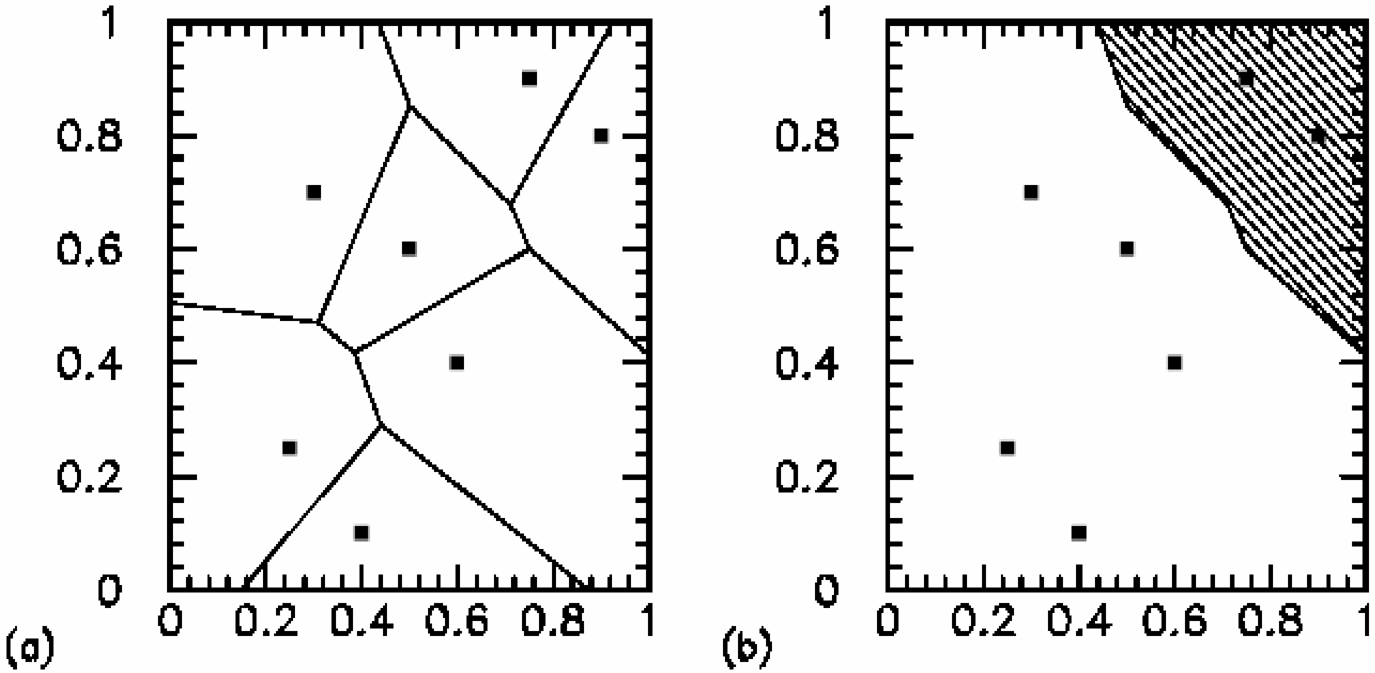


Figure 4.1: Example of a data set with a potentially anomalous point. The solid histogram is the expected distribution, and the points with error bars are the data. The bulk of the data is well described by the background prediction, but the point located at $x = 61$ appears out of place.

A notion of a region: Voronoi diagrams

An *N-region* (about a cluster of N data points) is the set of all values of x closer to a data point in that cluster than to any other data point in the sample.



seven 1-regions

one 2-region

Voronoi diagrams

Voronoi regions in the 1D example

This is a sketch of the division of the x axes to segments with the Voronoi method.

Each point in a segment has its closest event inside the segment it's in.

This way each event has its "region of influence" and a notion of an event-region is established.

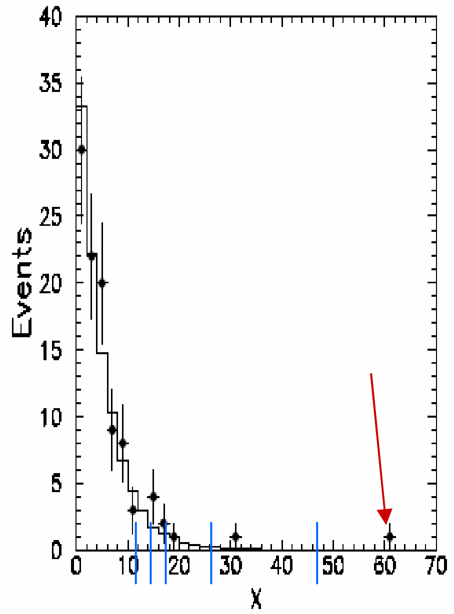


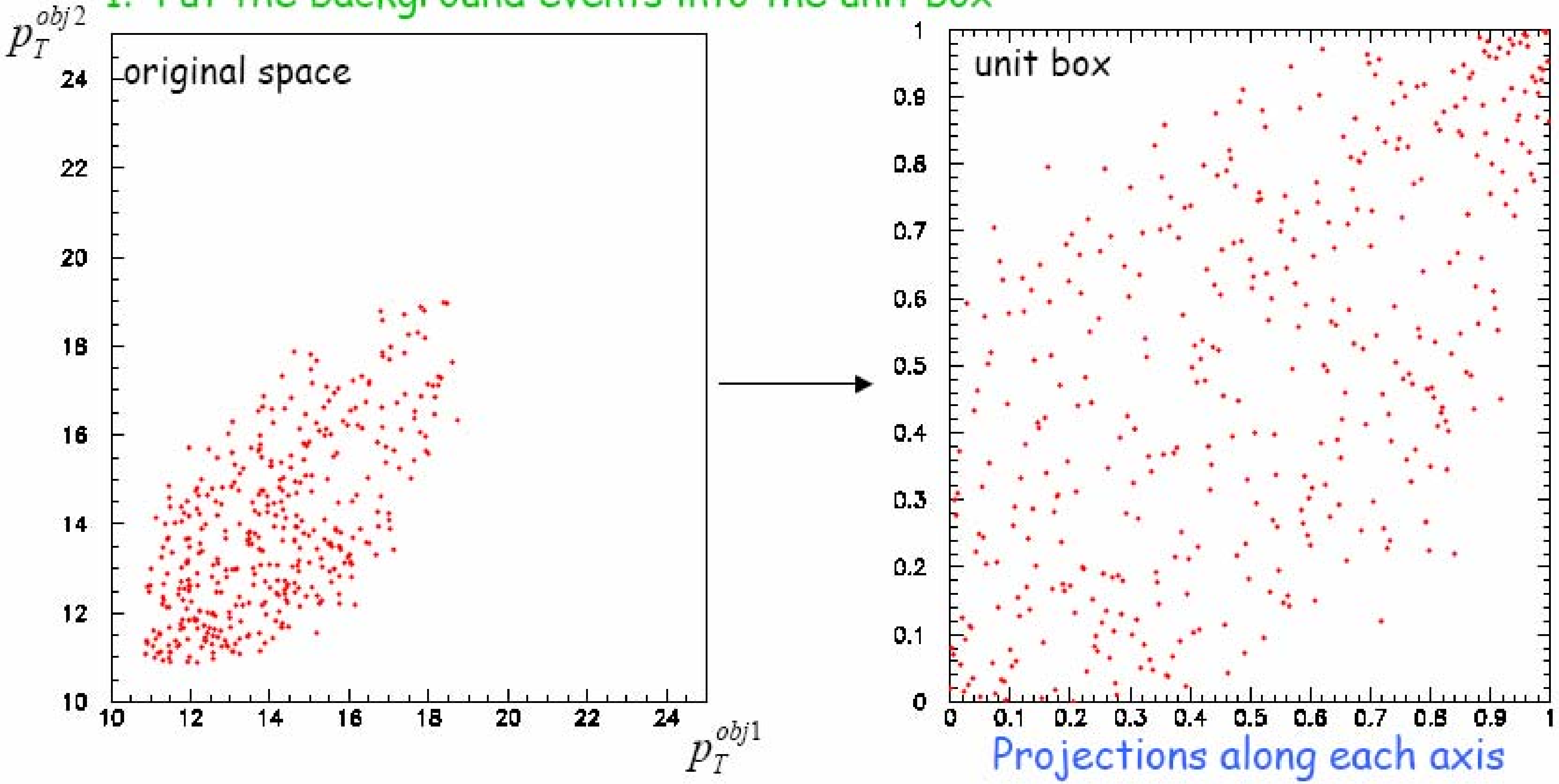
Figure 4.1: Example of a data set with a potentially anomalous point. The solid histogram is the expected distribution, and the points with error bars are the data. The bulk of the data is well described by the background prediction, but the point located at $x = 61$ appears out of place.

But before implementing this definition – a transformation of the variable set will make the rest of the algorithm simpler and better understood

Transformation of the Variable Space: normalizing the distribution

We begin by applying a variable transformation that makes the background distribution uniform in the "unit box" — $[0,1]^d$

1. Put the background events into the unit box

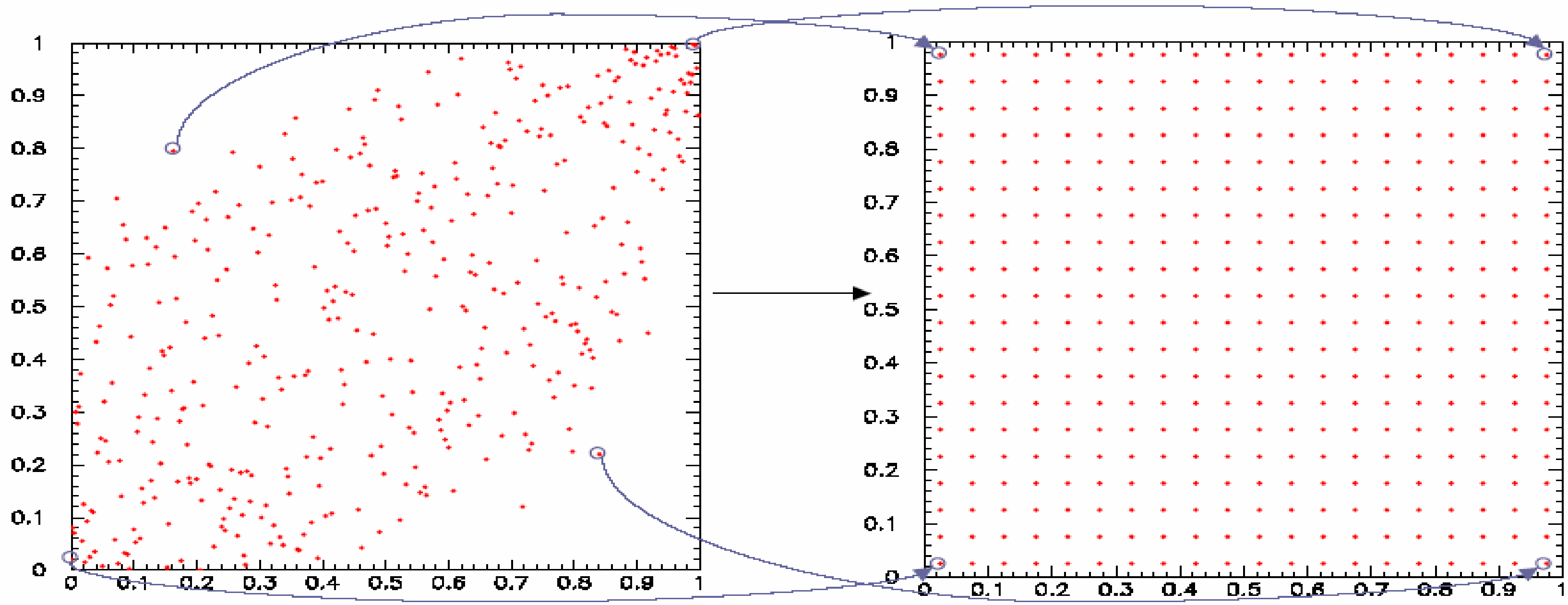


Projections along each axis are now uniform

Flattening the distribution

2. Map the background events onto a uniform grid

[Iteratively switch pairings to minimize the maximum distance moved]



How is it done in practice?

- assign a grid point to every event (randomly or otherwise)
- calculate the distance of each such a couple.

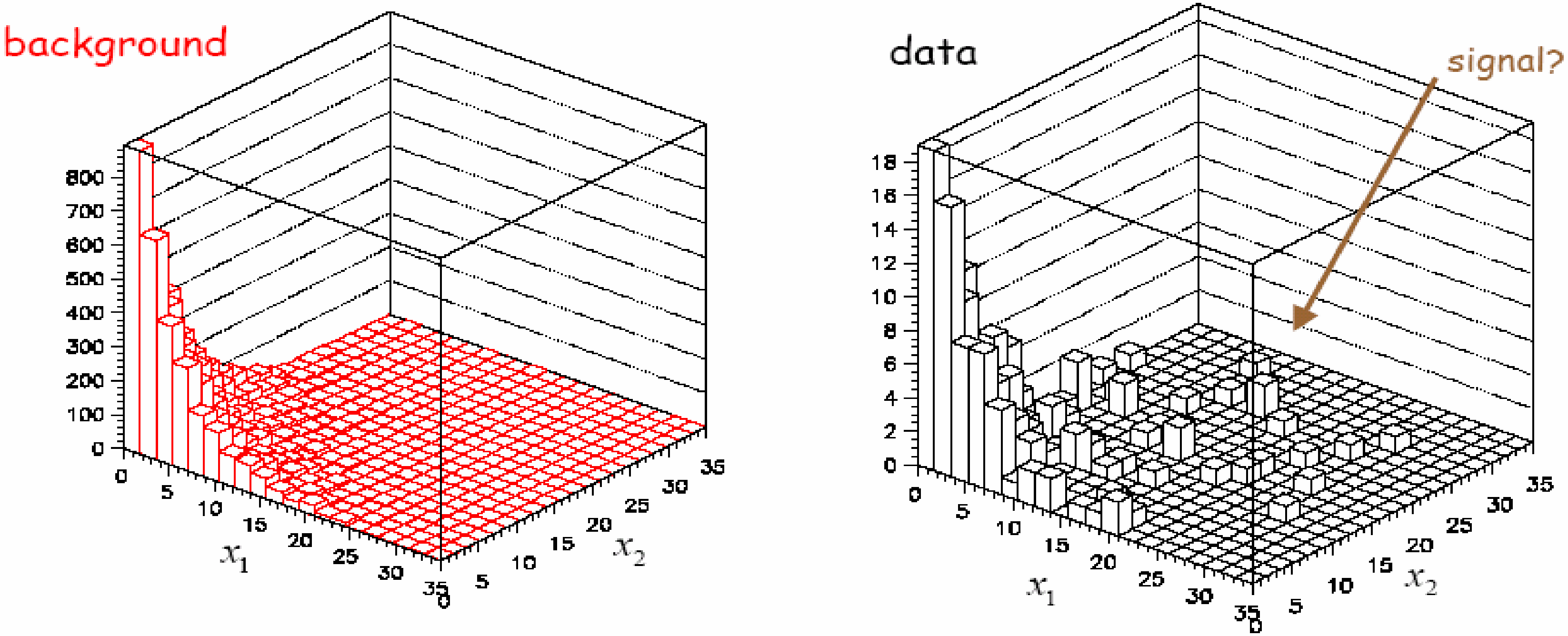
Then start an iterative process: check two couples and try switching their association.
if by this switch the maximum of the two distances is smaller
than the previous maximum than make the switch.

Properties of the procedure:

- Does not uniquely defined - different initial status will yield different final status.
- Adjacent points stay adjutants - a necessary condition for the flattening to be meaningful.

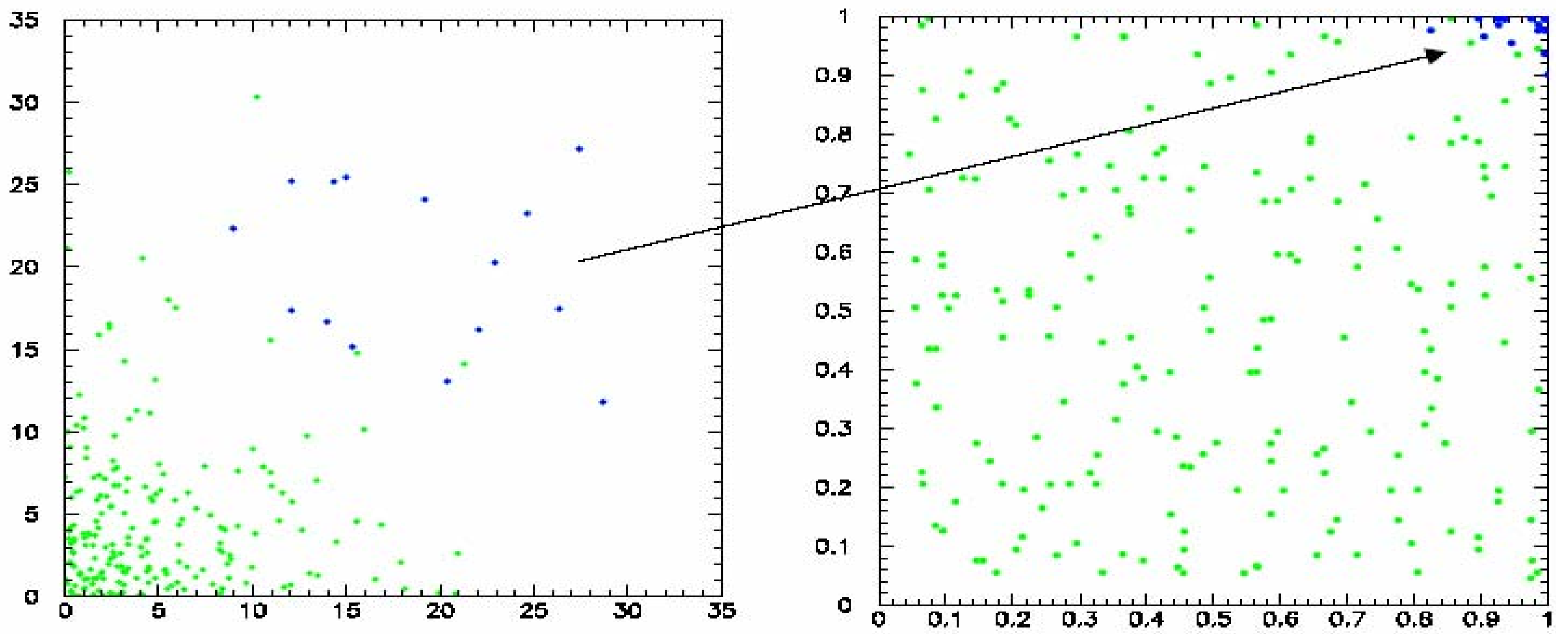
Example:

A quick example of how this might look for data:



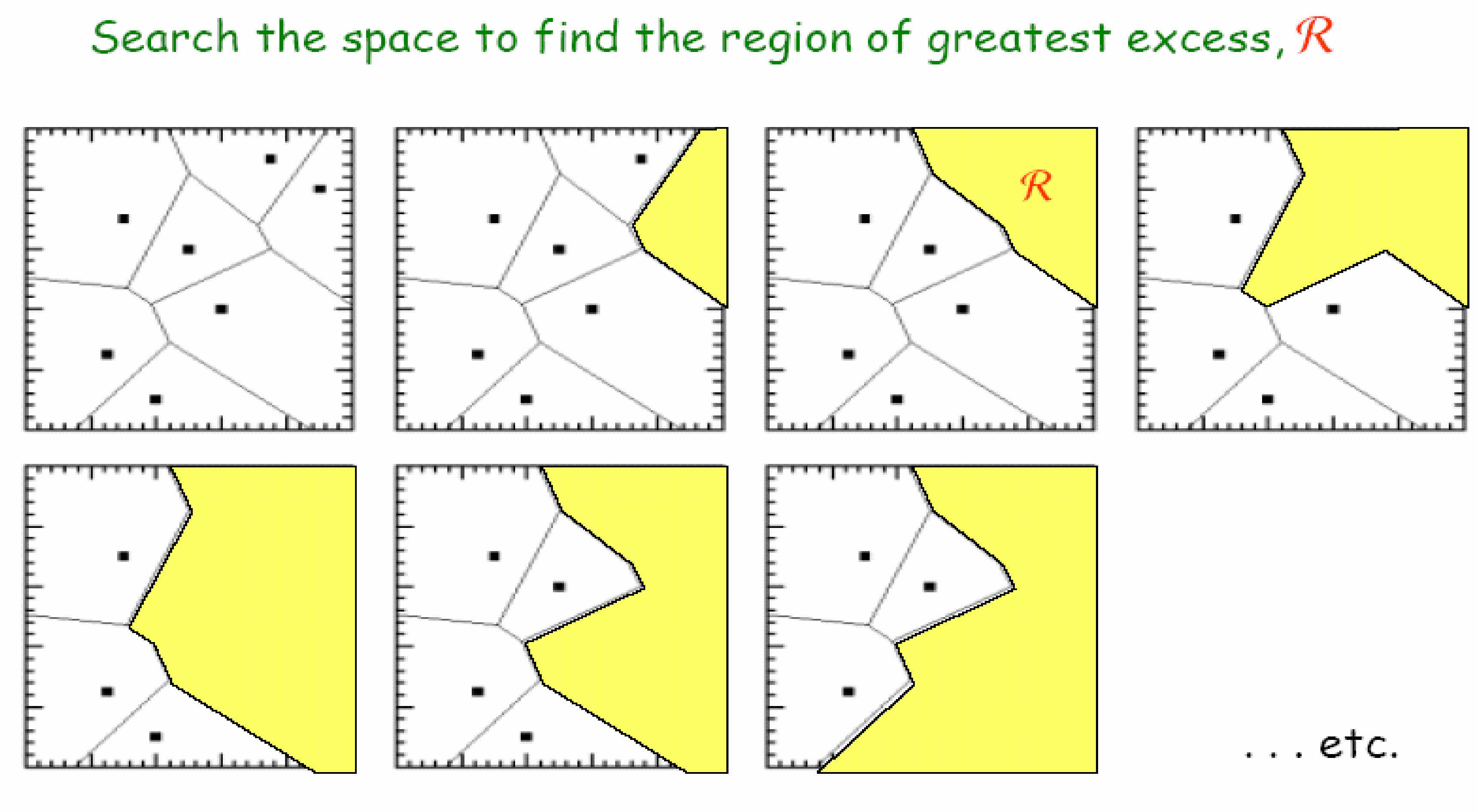
Example:

The transformation maps the signal region into the upper right-hand corner of the unit box



The background data events are uniformly distributed, as desired, and the signal cluster is "obvious"

Enlargement of the Top-Right corner:



There is a number of N 1-sized Regions and N*(N-1) of 2-sized events.
From combinatorics we know that the total number of possible combinations of an N-set is 2^N. This number of region of interest becomes too big for establishing a thorough check.
Therefore criterions can be imposed on the regions.
Possible candidates can be: considering adjacent events only; setting of all-close-by events as a one undivided region and checking portion of the whole volume which fits a given definition (such as AntiCornerSphere).

Quantify the Measure of Interest

Two Ways:

Plain:

For each n-sized region: find the set of events with the smallest volume.
Divide this volume with the expected volume for such a n-sized region.
This ratio could naively represent the measure of interest.

Sophisticated (and better motivated):

General Theme: Make "hypothetical similar experiments" (hse's) – produce events out of the given distribution (considering its uncertainties).
The measure of interest would be the fraction of hse's which fluctuates above the data set in a selected region

Current status of Wiezmann-Sleuth

- Built using Mathematica, contains all mentioned stages
- Simplicity was implemented throughout in order to make results useful for people other than the programmer…

Results

A first run was done with 3 variables

(total P_t, P_t of 1^{st} and missing p_t) at a point which is at the 5 sigma limit of LSL i.e. discovery is not trivial.

The 1-event region and 2-event region which were chosen as most interesting were indeed containing signal but the 3-event to the 6th regions were not pure-signal.

A run with 4 variables (circularity added) is now being conducted and expected to perform better (more information)

Comparison with other methods

**Although Dedicated searches are bound to be more sensitive than this method,
It has it's points of strength:**

- **Well motivated**
  (built upon founded notions – region definition, transformation of variable space…)
- **Emphasize understanding of the data**
  (rather than what the data has to say about a certain model)
- **Inclusive and Generic**
  (one run can suffice for many models. in particle physics this property is highlighted since the number of candidate models is very large)
- **Systematic**
  (Leaves no stone unturned and restricted by its sensitivity alone)

- **Allow for surprises….**

We would like to pinpoint that although this algorithm was originally built for use in the particle physics field, it is not restricted to it, and it looks like that it can be used for comparing and analyzing different data sets.

Future Goals

- Strength of the code: add more criterion options, discuss and refine different aspects of it in order to make it sound as it can be
- Ease of use: build an interface with simple input/output/computation options
- Keep the code simple and transparent for the user
- As of now we consider Sleuth to be a generic method for data analysis which could be used in many fields. We would like to check its sensitivity for different types of data sets.

Thank You !

Thanks for Listening!

Any Questions? ☺

All the useful graphs were obtained from Knuteson's presentations.
His homepage: http://mit1.fnal.gov/~knuteson/