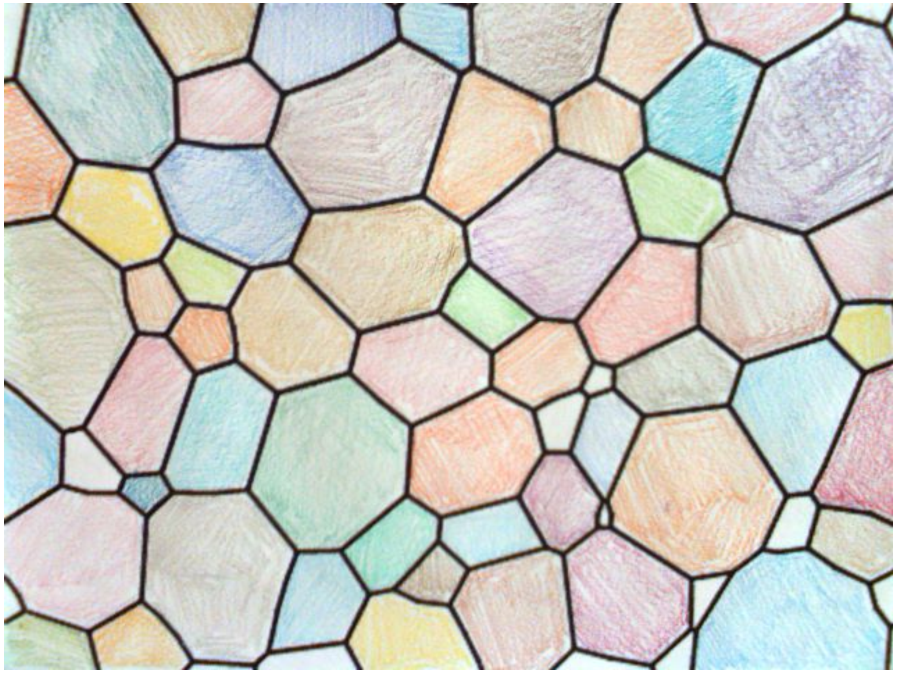
Edge Detecting New Physics the Voronoi Way



Dipsikha Debnath

Pheno 2016

Work with Jamie Gainer, Doojin Kim, and Konstantin Matchev

arXiv:1506.04141[hep-ph]

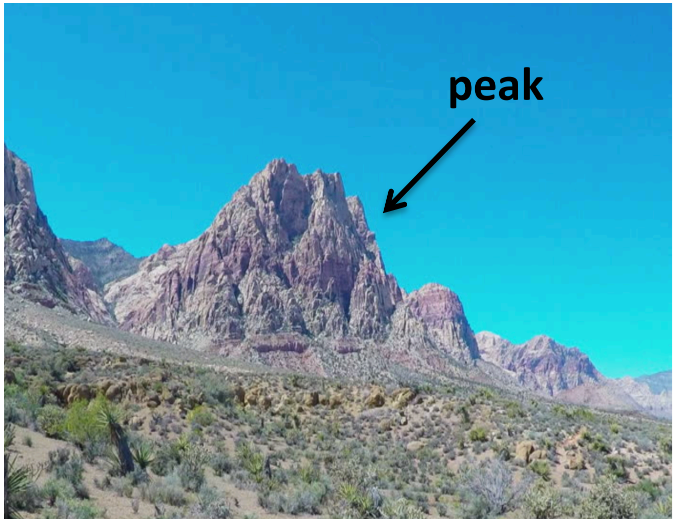
Motivation

- Run 2 Goal: Discovery of new physics, ideal search strategies should be-
    - **Model independent**: BSM may show up somewhere, not expected from theoretical viewpoint
    - **Maximally Sensitive**: need every last bit of information to have enough statistical significance for discovery

- A powerful approach for new physics searches: Identify structural **"features"** in data (i.e, odd features in the distribution of some variable)

  Example:

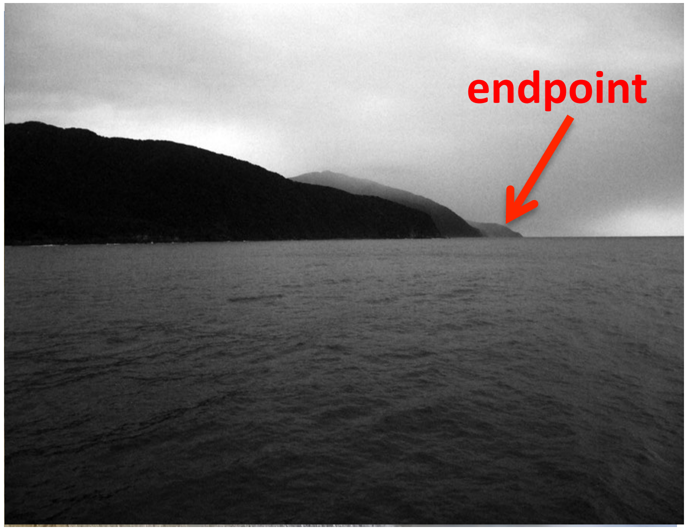
  ✓ Resonance peak



Motivation

- Run 2 Goal: Discovery of new physics, ideal search strategies should be-
  - **Model independent**: BSM may show up somewhere, not expected from theoretical viewpoint
  - **Maximally Sensitive**: need every last bit of information to have enough statistical significance for discovery

- A powerful approach for new physics searches: Identify structural **"features"** in data (i.e, odd features in the distribution of some variable)



Example:

✓ Resonance peak

✓ Kinematic endpoint

Motivation

- Run 2 Goal: Discovery of new physics, ideal search strategies should be-
  - ❑ **Model independent**: BSM may show up somewhere, not expected from theoretical viewpoint
  - ❑ **Maximally Sensitive**: need every last bit of information to have enough statistical significance for discovery

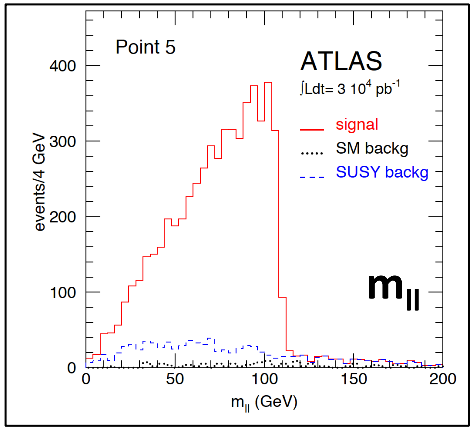
- A powerful approach for new physics searches: Identify structural **"features"** in data (i.e, odd features in the distribution of some variable)

Example:

✓ Resonance peak
✓ Kinematic endpoint
✓ Kinematic edge

- **Well established technique for SUSY search (1-dim)**: visible decay products have edges and endpoints in their invariant mass distribution

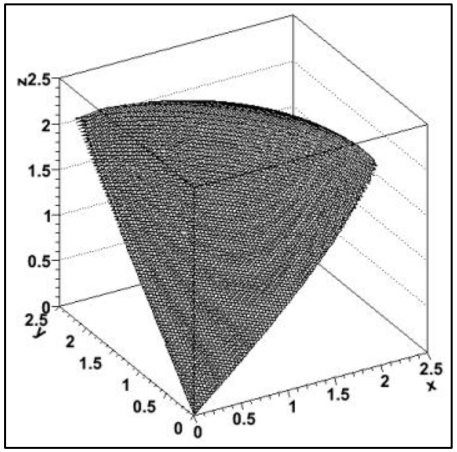
Example: $\tilde{\chi}_0^2 \to \tilde{l}^{\pm} l^{\mp} \to l^+ l^- \tilde{\chi}_0^1$



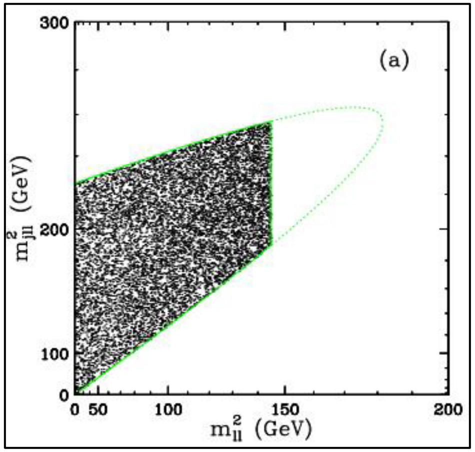
[ATLAS 1999]

- For SUSY discovery we can use **additional variables** (e.g. if more than two visible final state particles, consider invariant masses of different pairs of particles).

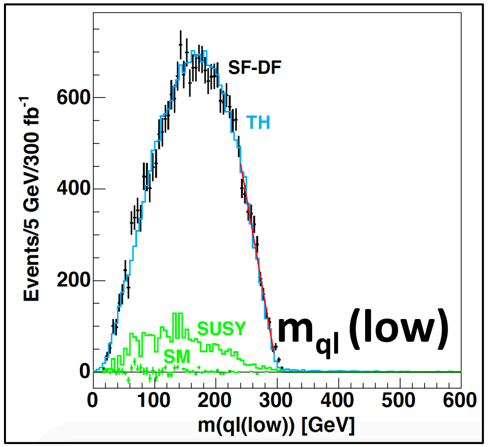
Example: $\tilde{q} \to q\tilde{\chi}_0^2 \to q\tilde{l}^{\pm} l^{\mp} \to q l^+ l^- \tilde{\chi}_0^1$



[Constanzo, Tovey (2009)]



[Burns, Matchev, Park (2009)]



[Gjelsten, Miller, Osland(2005)]

Challenges of edge detection

Then the goal is to find edges in more than 1 dimension. Edge detection in HEP data is non trivial

- ❑ Especially with relatively sparse data (as opposed to edge detection in image)

- ❑ We may not know analytically the class of distributions describing the data.

- ❑ The data may be in more than two dimensions.

Challenges of edge detection

Then the goal is to find edges in more than 1 dimension. Edge detection in HEP data is non trivial

- ❑ Especially with relatively sparse data (as opposed to edge detection in image) ✔
- ❑ We may not know analytically the class of distributions describing the data. ✔
- ❑ The data may be in more than two dimensions. ✔

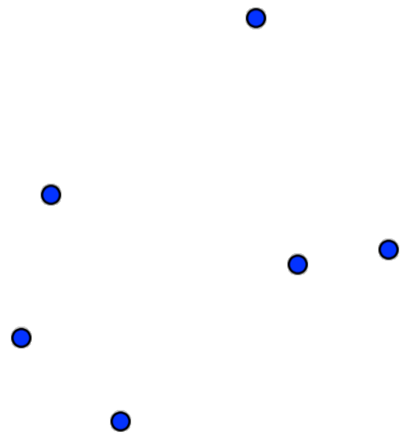
We propose method of edge detection using **geometric properties of Voronoi tessellations**

What are Voronoi tessellations?

- Tessellation: breaking up space into regions

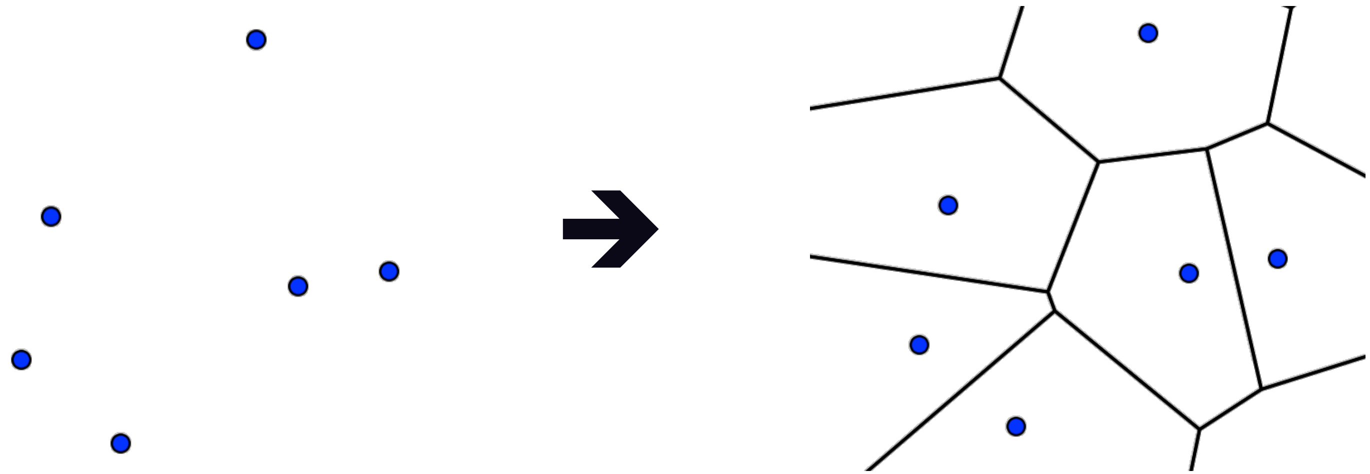
What are Voronoi tessellations?

- Tessellation: breaking up space into regions
- Voronoi tessellation method:
  - Take a set of seed points in space



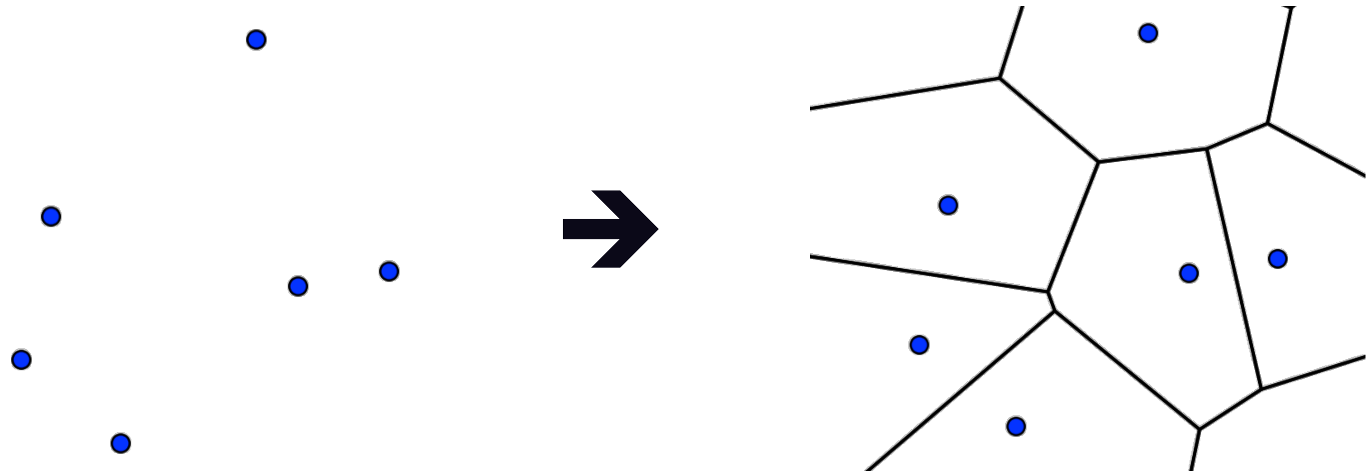
What are Voronoi tessellations?

- Tessellation: breaking up space into regions
- Voronoi tessellation method:
  - Take a set of seed points in space
  - Divide space into regions where a given data point is the closest data point
  - Region = "Voronoi cell"



What are Voronoi tessellations?

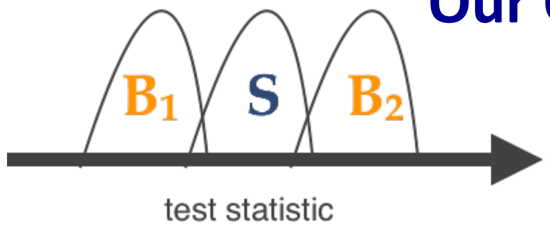
- Tessellation: breaking up space into regions
- Voronoi tessellation method:
  - Take a set of seed points in space
  - Divide space into regions where a given data point is the closest data point
  - Region = “Voronoi cell”



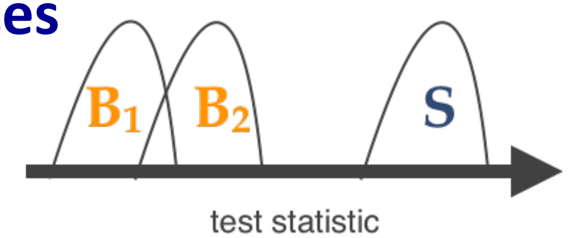
Voronoi tessellations have been widely applied in Mathematics, Condensed matter physics, Astrophysics, and occasionally **in particle physics** (SLEUTH [hep-ex/ 006011], FastJet[Cacciari, Salam, Soyez:1111.6097])

Method of Edge Detection

Our Choices



vs.



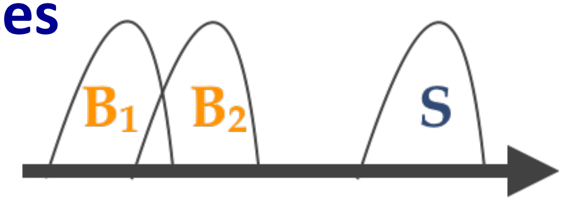
Signal (edge) is **in-between** background (non-edge).

Signal (edge) is **well-separated** from background (non-edge)

Method of Edge Detection

Our Choices

vs.



Signal (edge) is **in-between** background (non-edge).

Signal (edge) is **well-separated** from background (non-edge)

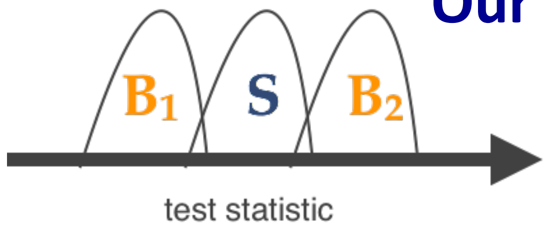
- Choices: **Scaled standard deviation**

$$\bar{\sigma}_i \equiv \frac{1}{\bar{a}} \sqrt{\sum_{j \in N_i} \frac{(a_j - \bar{a})^2}{|N_i| - 1}}$$

For a given cell, consider the neighboring cells and their areas. **An edge cell will have a big spread in neighboring areas.**

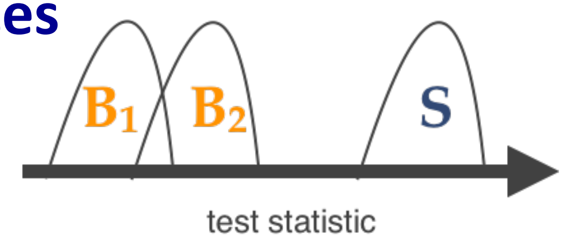
Method of Edge Detection

Our Choices



Signal (edge) is **in-between** background (non-edge).

vs.

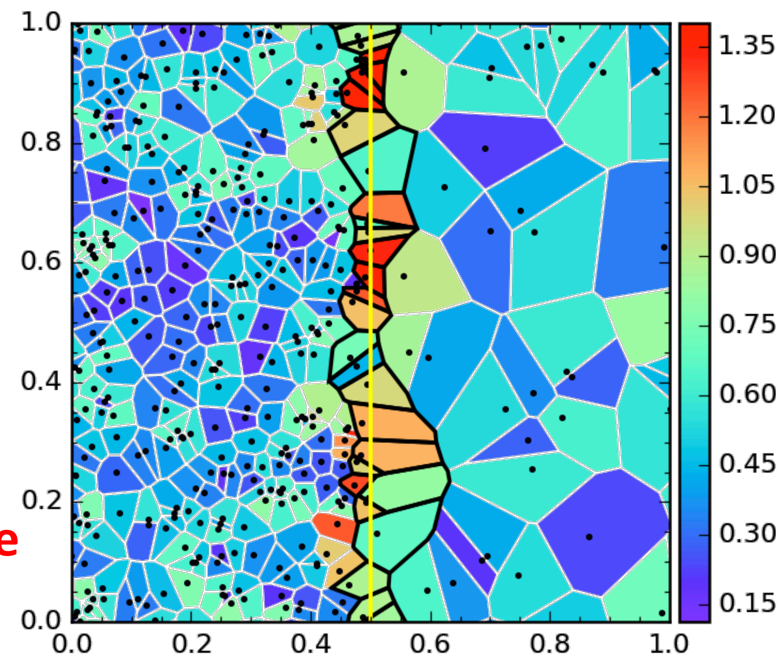


Signal (edge) is **well-separated** from background (non-edge)

- Choices: **Scaled standard deviation**

$$\bar{\sigma}_i \equiv \frac{1}{\bar{a}} \sqrt{\sum_{j \in N_i} \frac{(a_j - \bar{a})^2}{|N_i| - 1}}$$

For a given cell, consider the neighboring cells and their areas. **An edge cell will have a big spread in neighboring areas.**



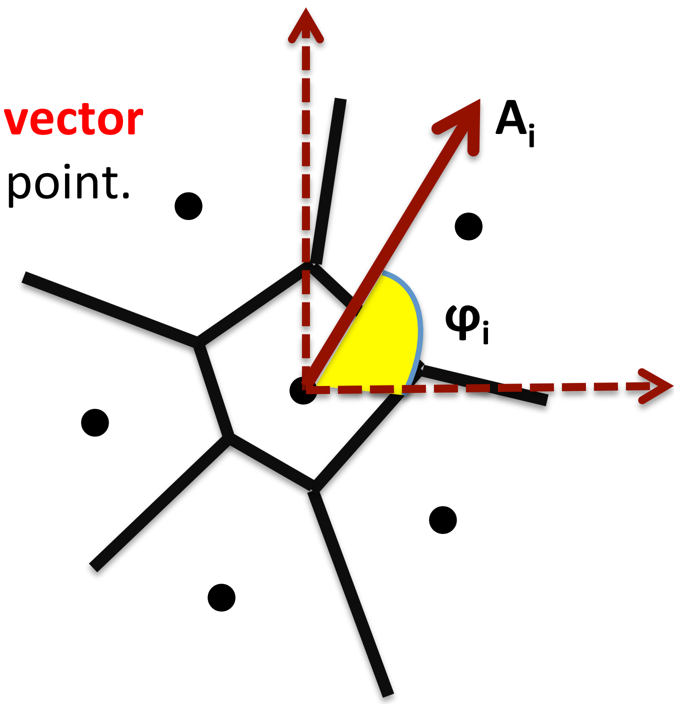
350 randomly generated data points with density of points from left to right region as 6.

Method of Edge Detection

Our choices

❖ Choices: **Amplitude and phase angle of gradient vector**
Compute the gradient vector (A_i, φ_i) at each data point.

$$(\nabla f)_i \equiv (A_i \cos\varphi_i, A_i \sin\varphi_i) \equiv \vec{\boldsymbol{A}}_i$$



Method

- The directional derivative for the i-th Voronoi cell toward the j-th neighbor.

$$(\nabla_{\hat{n}_{ij}} f)_i = (a_i a_j)^{\frac{3}{4}} \frac{f(\vec{\boldsymbol{x}}_j) - f(\vec{\boldsymbol{x}}_i)}{|\vec{r}_j - \vec{r}_i|}$$

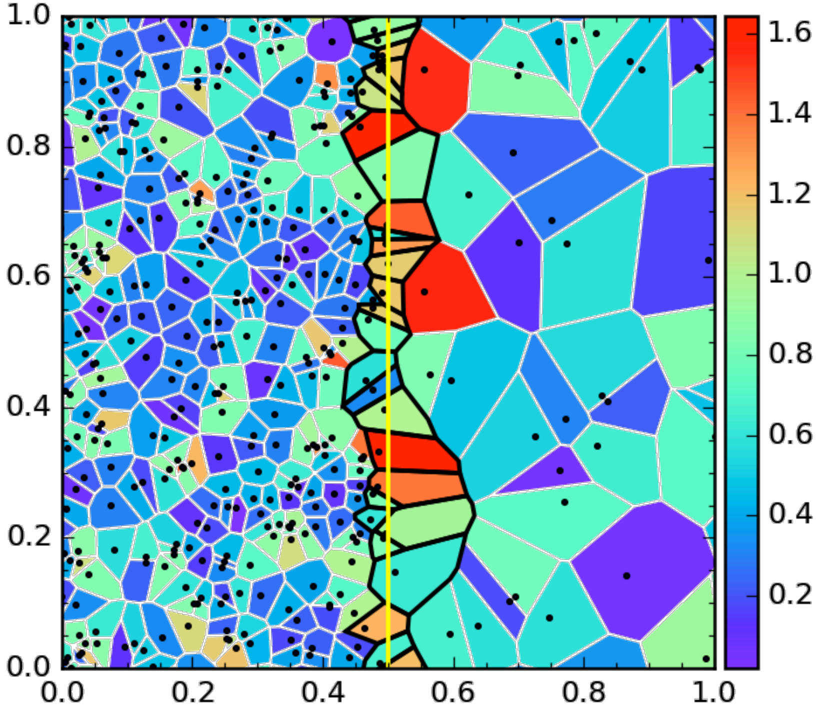
where $f(\vec{\boldsymbol{x}}_i) \simeq \frac{1}{N a_i}$, $\hat{n}_{ij} = \frac{\vec{r}_j - \vec{r}_i}{|\vec{r}_j - \vec{r}_i|} \equiv (\cos\varphi_{ij}, \sin\varphi_{ij})$

- Finally, extract the amplitude and phase of the gradient vector by fitting the directional derivative to

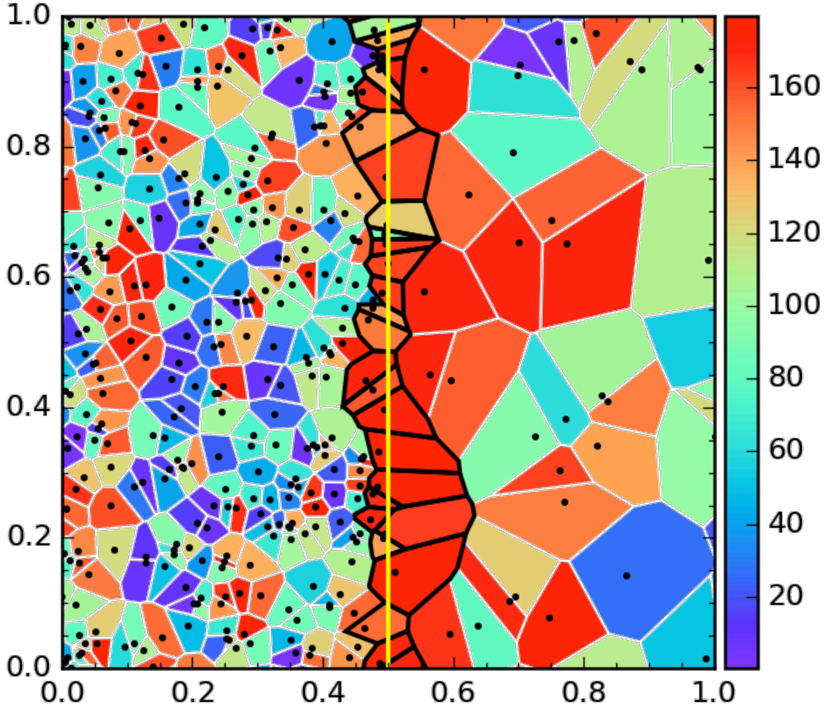
$$(\nabla_{\hat{n}_{ij}} f)_i \equiv (\nabla f)_i \cdot \hat{n}_{ij} = \overset{\checkmark}{A_i} \cos(\overset{\checkmark}{\varphi_i} - \varphi_{ij})$$

Method of Edge Detection

Amplitude



Phase angle (in degree)



- **Edge cells are characterized with relatively large gradient magnitudes.**
- The directions of their gradients are correlated.

350 randomly generated data points within unit square with density of data points from left to right region as 6.

Method of Edge Detection

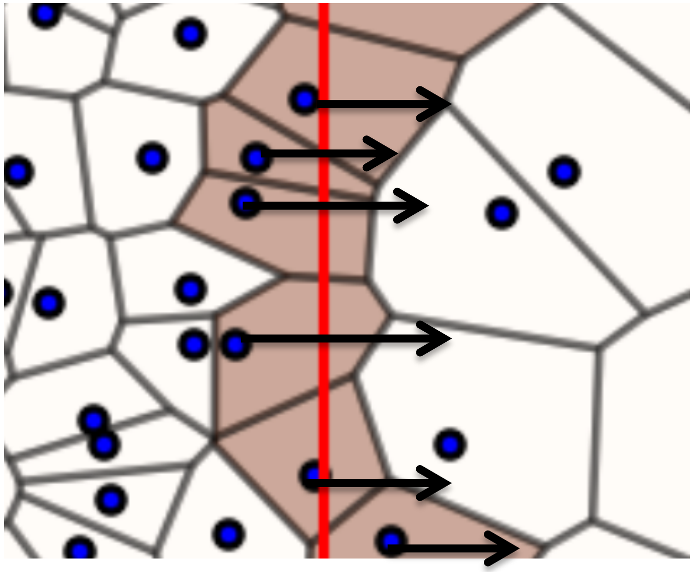
Our choices

- Choices: **Average scalar product of the gradient vectors**

  The average scalar product of the gradient vectors for a given cell.

$$\bar{s}_i \equiv \frac{1}{|N_i|} \sum_{j \in N_i} \vec{\boldsymbol{A}}_i \cdot \vec{\boldsymbol{A}}_j$$

Edge cells



Method of Edge Detection

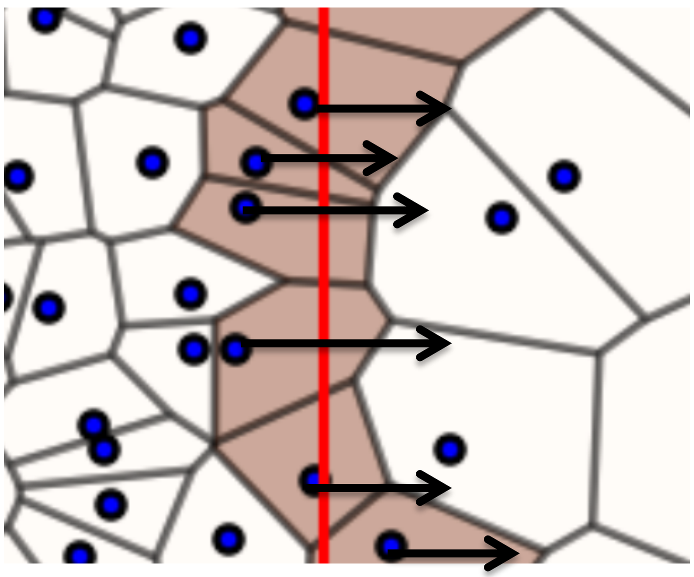
Our choices

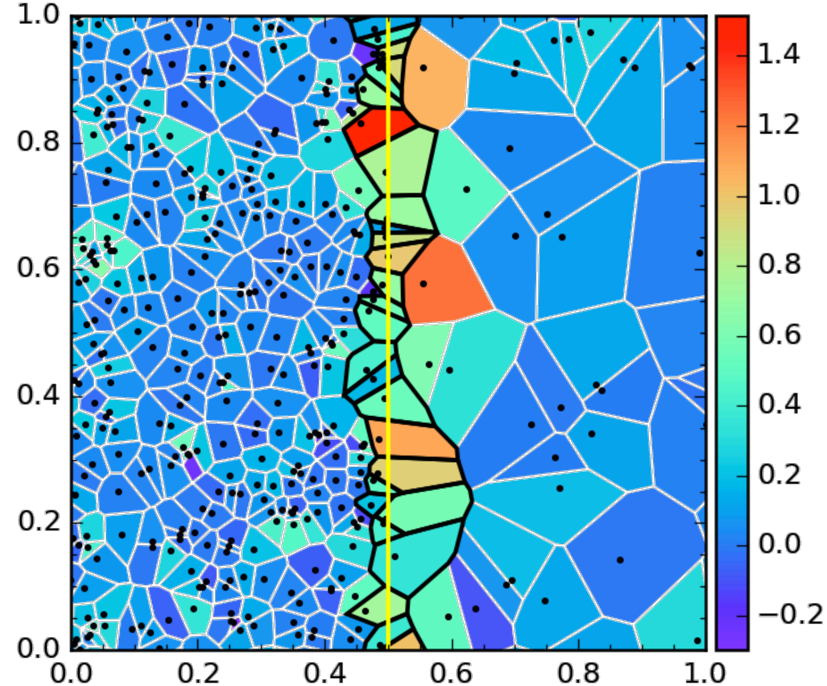
- Choices: **Average scalar product of the gradient vectors**

The average scalar product of the gradient vectors for a given cell.

$$\bar{s}_i \equiv \frac{1}{|N_i|} \sum_{j \in N_i} \vec{A}_i \cdot \vec{A}_j$$

Edge cells



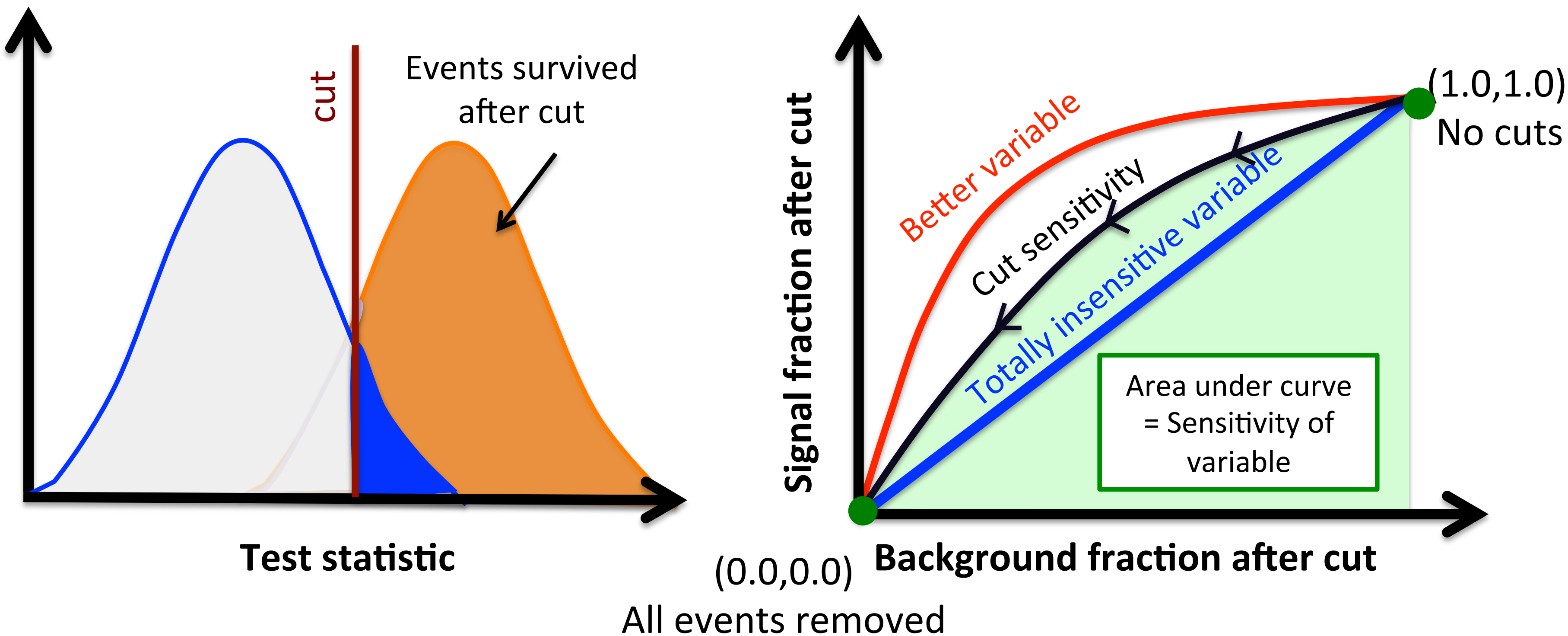


The average scalar product of the gradient vectors should be higher for edge cells as they have large gradient magnitudes.

Quantify Sensitivity of Variables

ROC curves

Assumption: edge cells ⇔ signal, non-edge cells ⇔ background



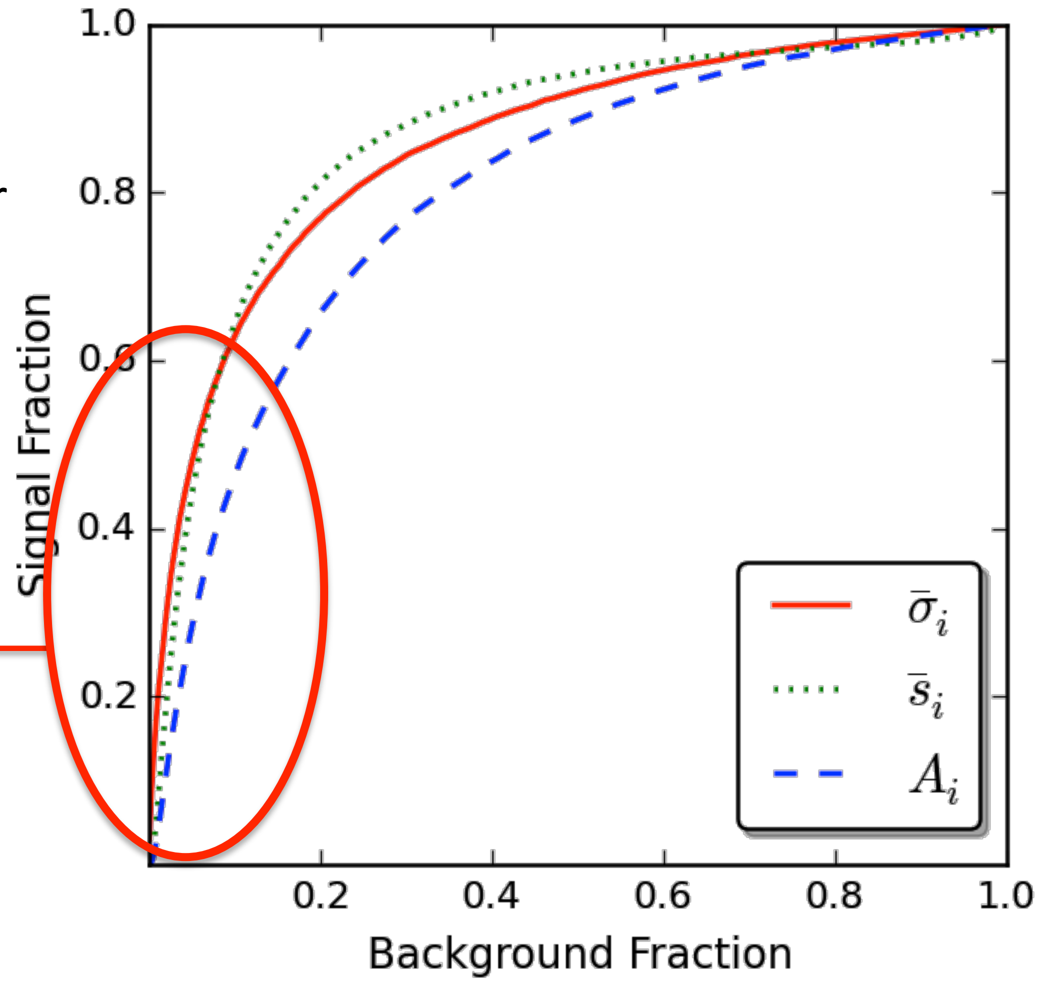
ROC curve with greater areas are more sensitive variable

Quantify Sensitivity of Variables

- We notice that three quantities scaled standard deviation, amplitude, average scalar product of the gradient vectors are quite successful in identifying edge cells.

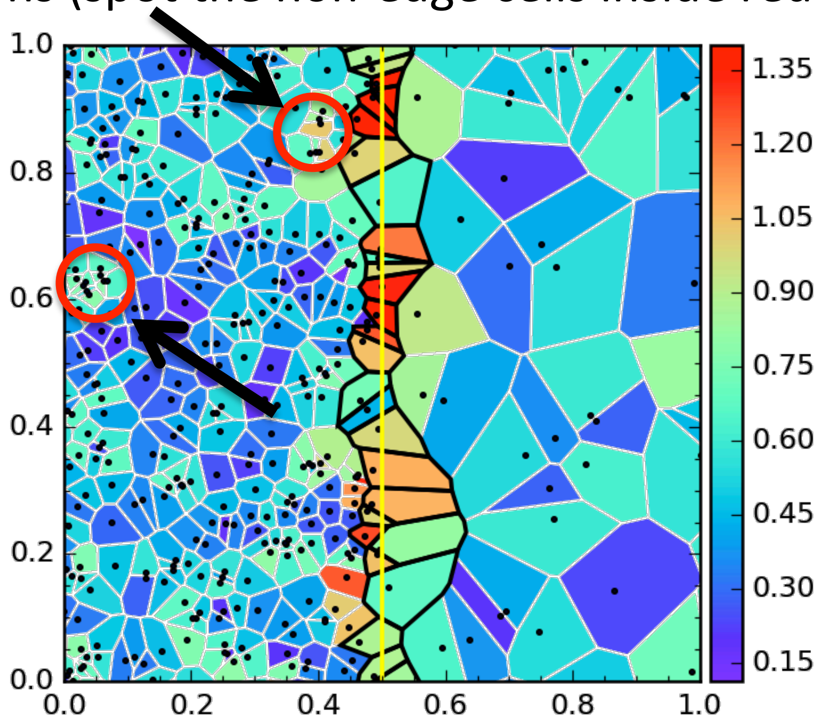
- Plot signal selection efficiency vs. the background efficiency for different values of cut on these three variables.

The scaled standard deviation does best in the relevant range of very low background fraction



Problem:

One issue in distinguishing edge cells from non-edge cells is the presence of statistical fluctuations (spot the non-edge cells inside red circles)



Color: scaled standard deviation

Is there a way to filter out statistical fluctuations but preserve the underlying features?

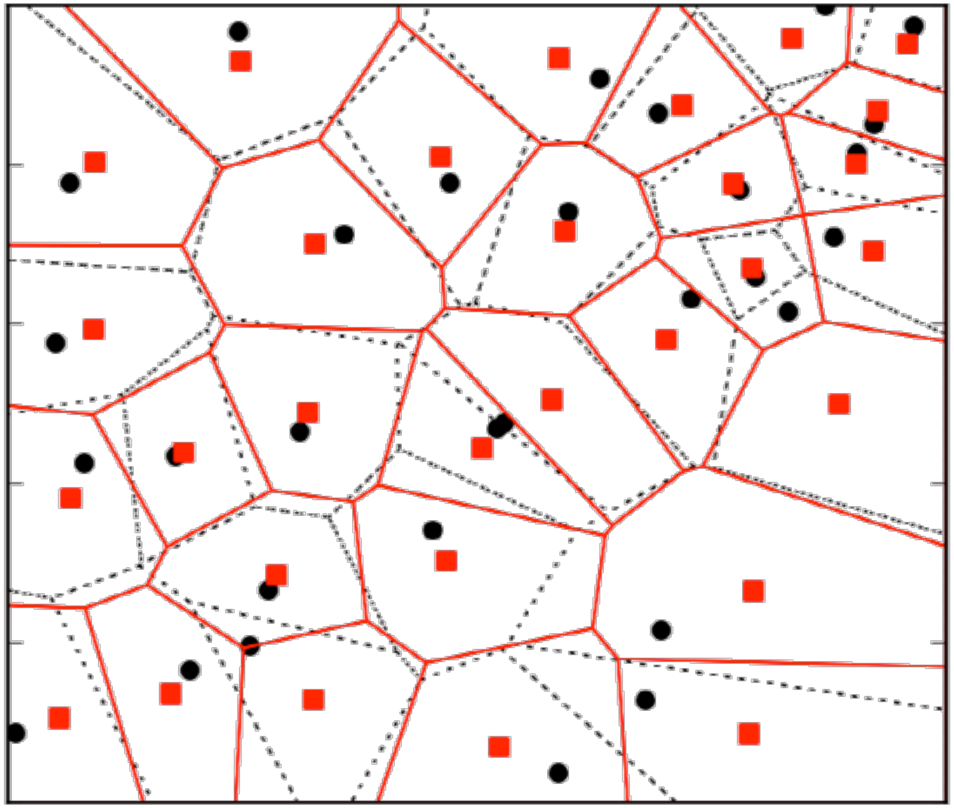
Improved Edge Finding

Lloyd's algorithm

- We can smooth out statistical fluctuations in the data by replacing each point (black dot) with the centroid (red square) of its cell

- Points (black) are not necessarily located at the center of mass / centroid (red) of cell

- **Black**: Voronoi tessellations of randomly generated points



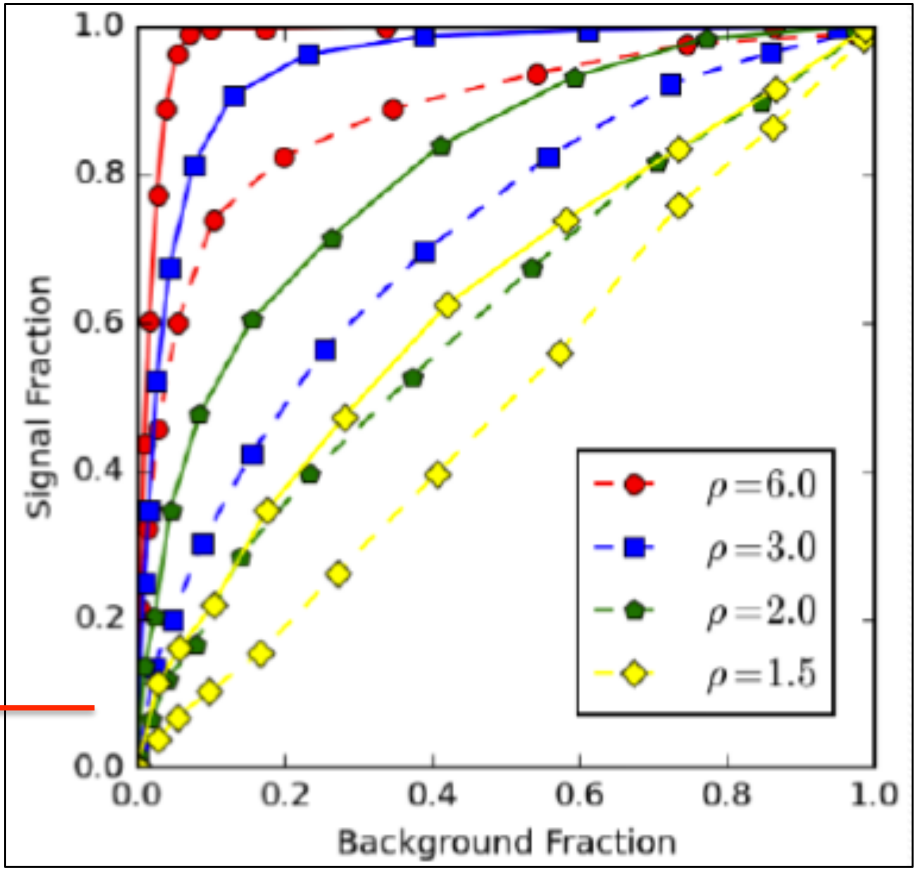
Red: Voronoi tessellations after 1 Lloyd iteration

Increased sensitivity from Lloyd's algorithms

Signal selection efficiency vs. the background efficiency for different values of cut on scaled standard deviation.

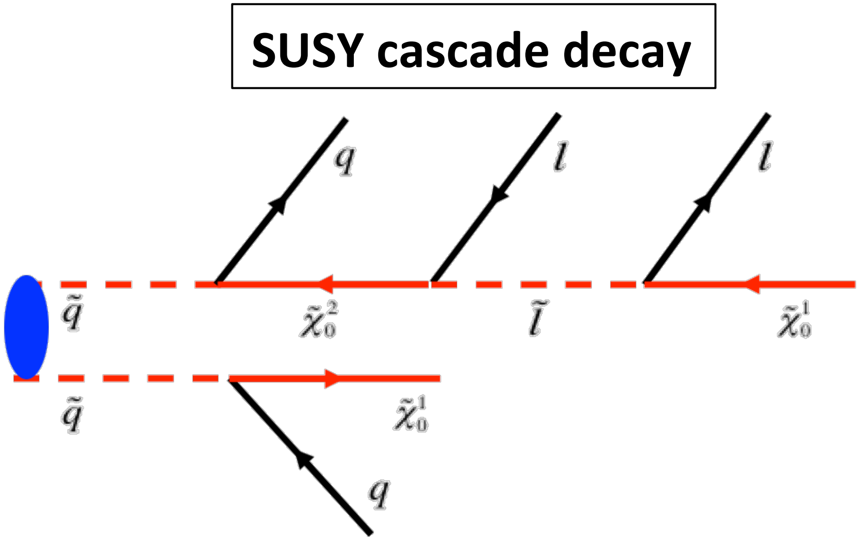
- Solid (dashed) lines show ROC curves after (before) using Lloyd's Algorithm.
- Tested with different density ratios.

Better signal sensitivity (edge) after Lloyd iterations



(Debnath, Gainer, Kim, Matchev, 2015)

Application

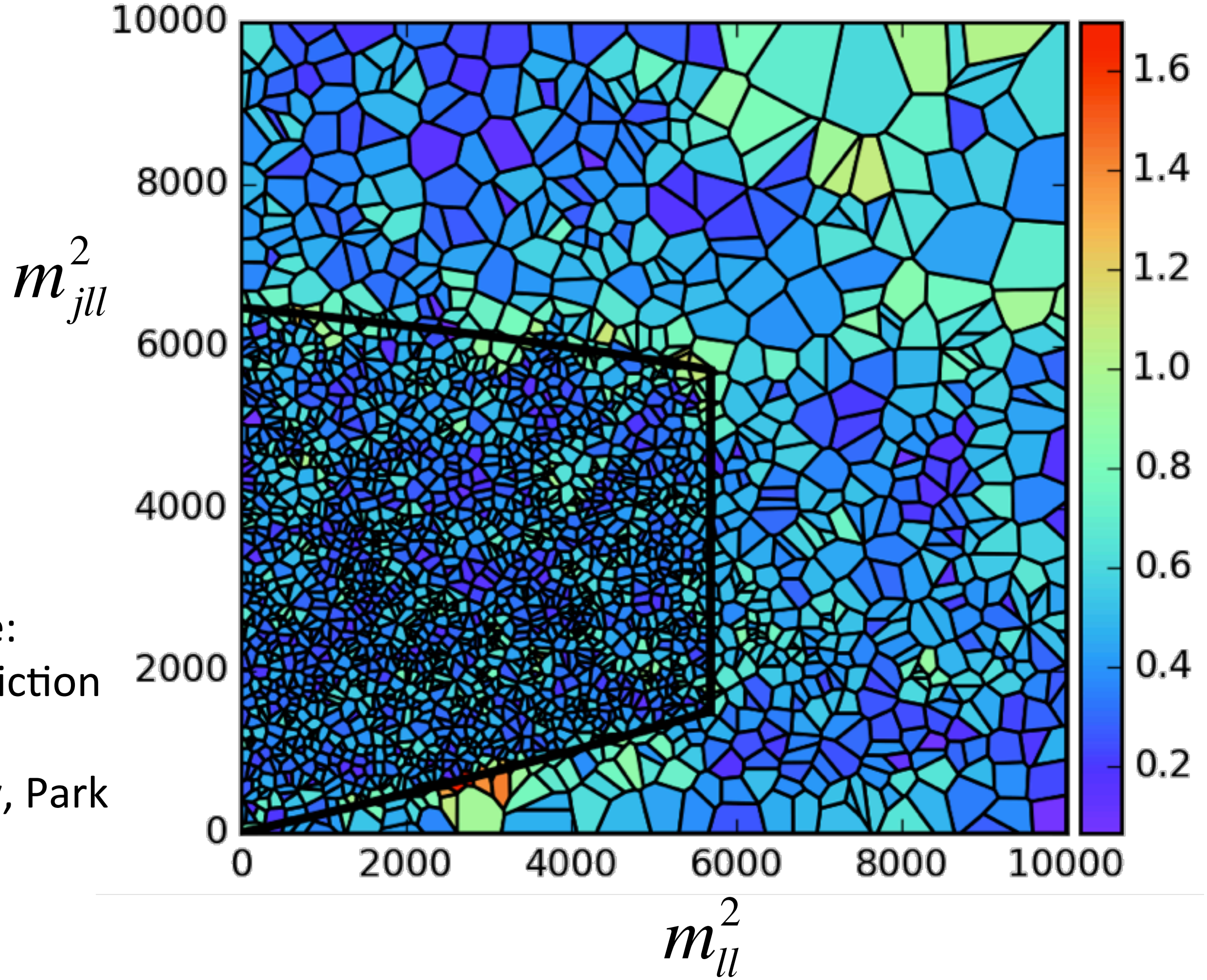


- Signal: Squark pair production; asymmetric topology. Signature 2 jets and 2 leptons.

- Mass spectrum: 400 (Squark), 300 (2nd neutralino), 280(slepton), 200 (LSP) GeV.

- For a given event, we consider m_{ll}^2 and m_{jll}^2
  1. Combinatorial background arising from the choice of jets is considered.
  2. We also include background from top pair productions.

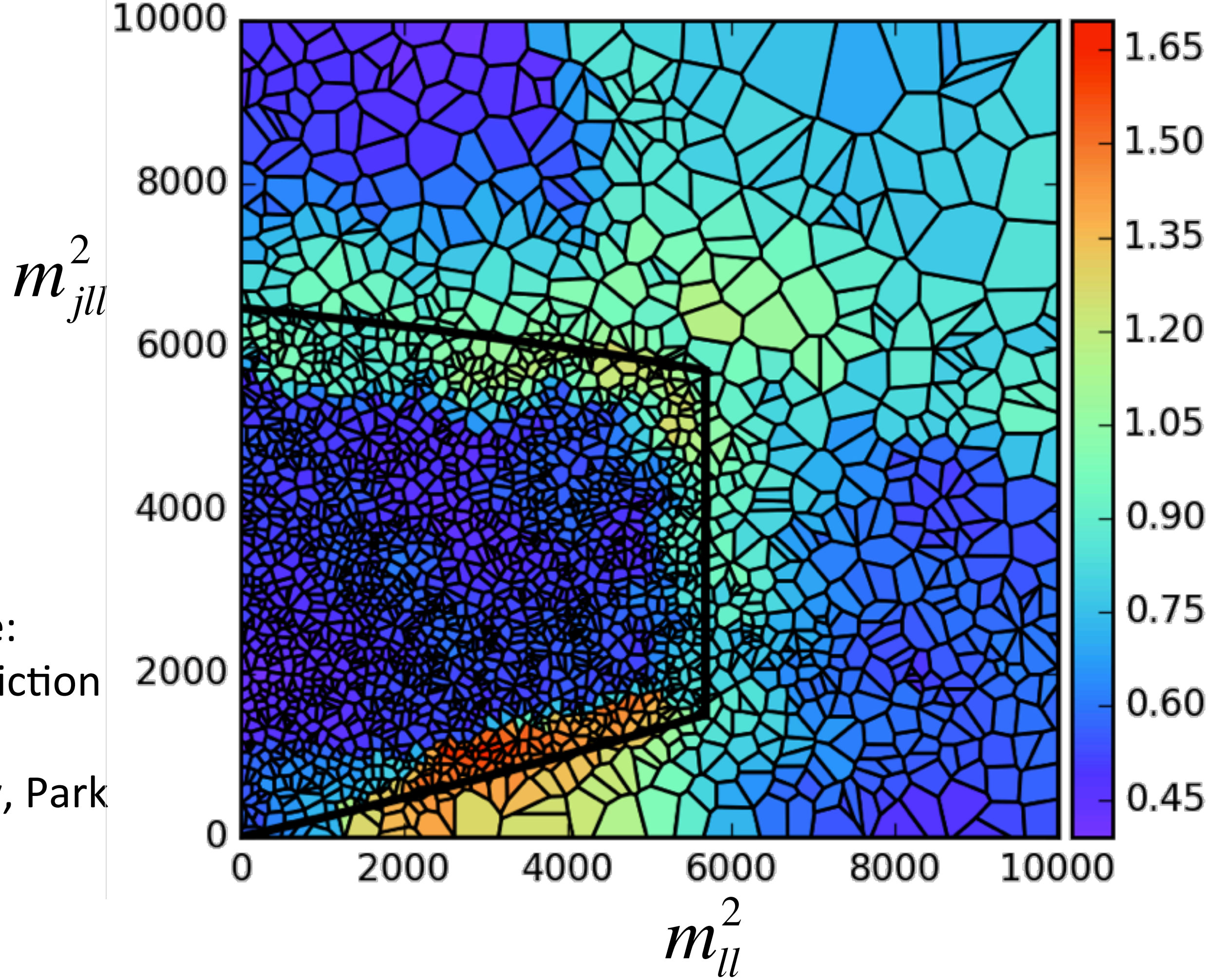
Result: Scaled standard deviation for just the Voronoi cells



Black solid curve: theoretical prediction for edge [Burns, Matchev, Park (2009)]

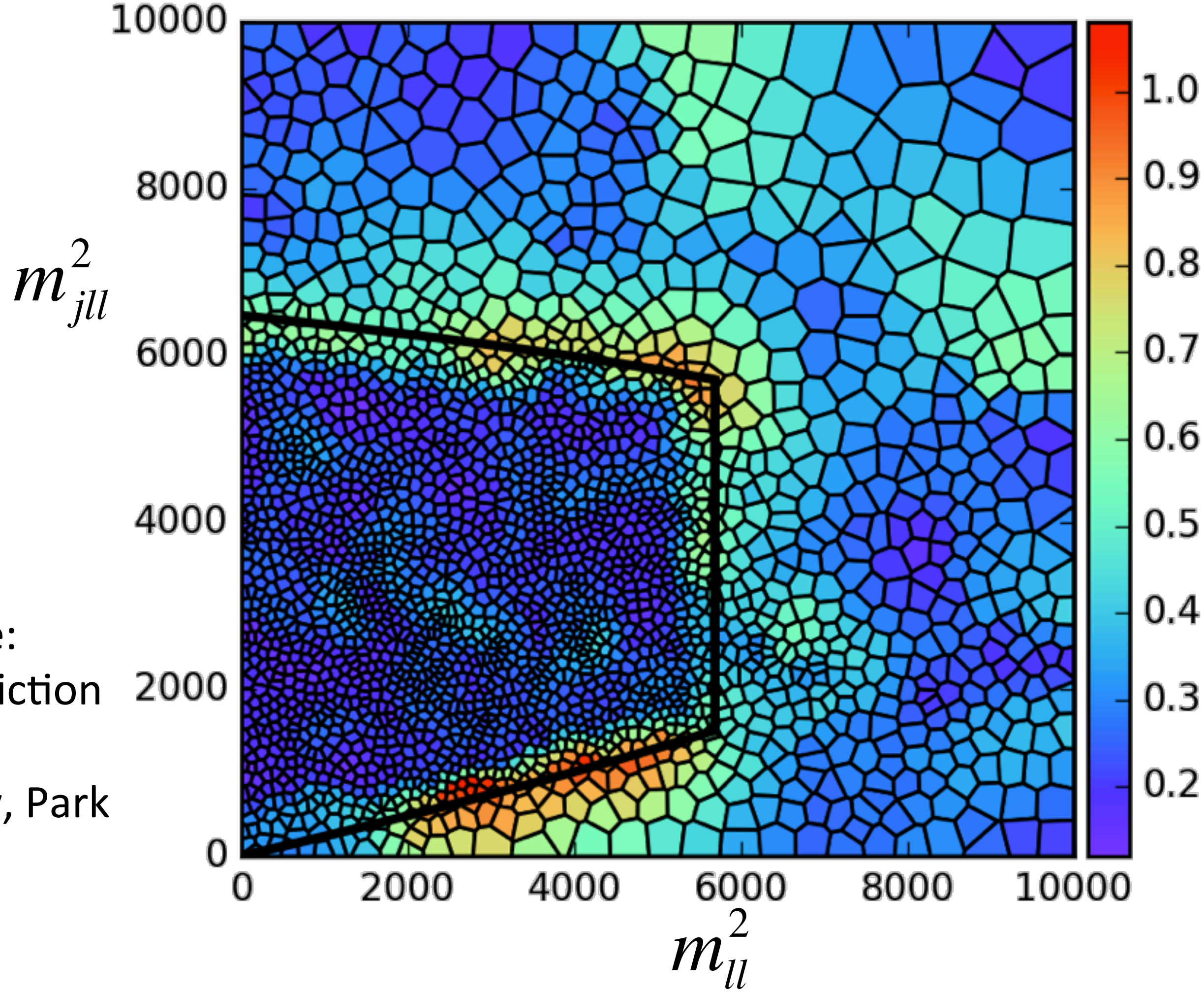
Application

Result: Scaled standard deviation after 5 Lloyd steps



Black solid curve: theoretical prediction for edge [Burns, Matchev, Park (2009)]

Result: Scaled standard deviation including 5 th nearest neighbor



Black solid curve: theoretical prediction for edge [Burns, Matchev, Park (2009)]

Conclusion

- I have discussed-
  - finding kinematic “features” in collider physics data is an essential step toward the discovery of BSM physics and Voronoi tessellations have greater role in achieving it.
  - in general, Voronoi-based analyses are qualitatively different than standard analyses: the value of a variable calculated for an event depends on “neighboring” events in phase space
  - our proposed methods have been tested with 2-dimensional HEP data.
- Recently, we are involved in finding edge in 3-D and mass measurement of new physics particle using Voronoi tessellation.
- Apart from above aspects, a wide range of applications are possible with Voronoi technique. Stay tuned!

Thank You!

EXTRA

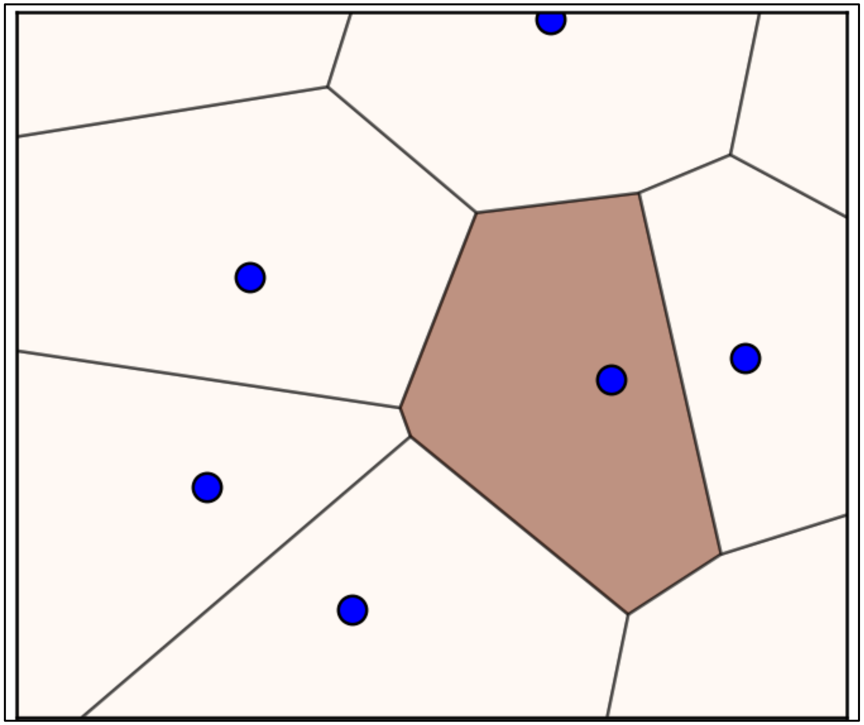
Method of Edge Detection

- We aim to find edge (cells) in 2D using geometric properties of Voronoi cells.
- Geometrical properties:

  Cell area

  Cell perimeter

  Number of sides



Method of Edge Detection

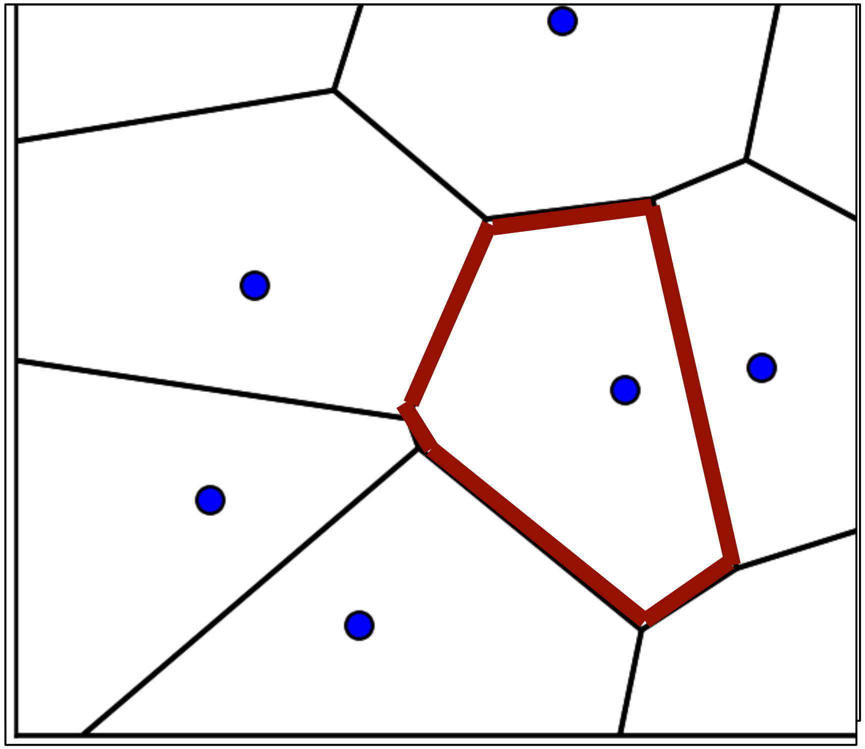
- We aim to find edge (cells) in 2D using geometric properties of Voronoi cells.

- Geometrical properties:

  Cell area

  Cell perimeter

  Number of sides



Method of Edge Detection

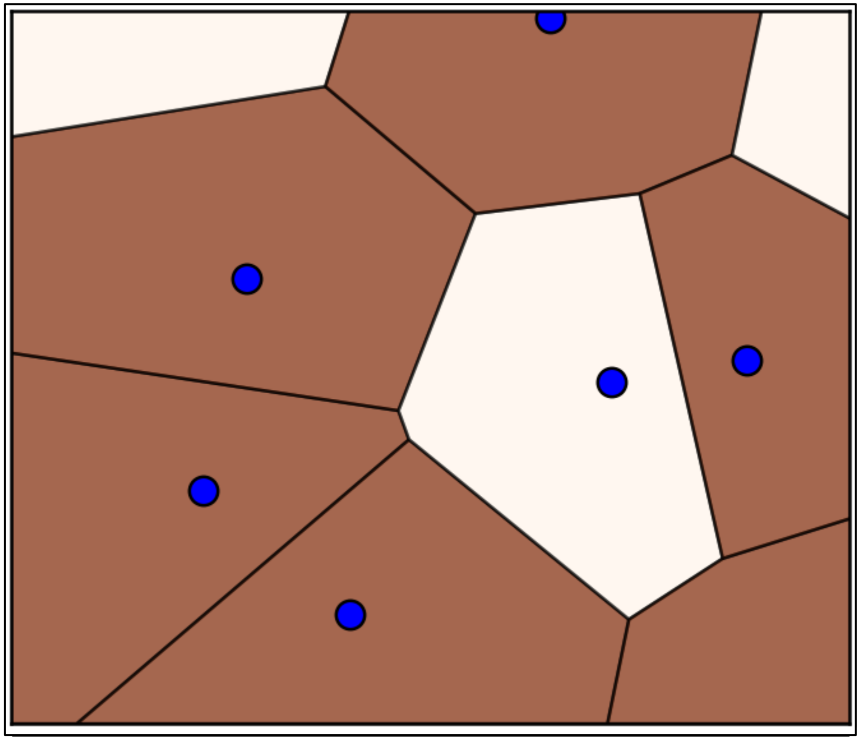
- We aim to find edge (cells) in 2D using geometric properties of Voronoi cells.

- Geometrical properties:

  Cell area

  Cell perimeter

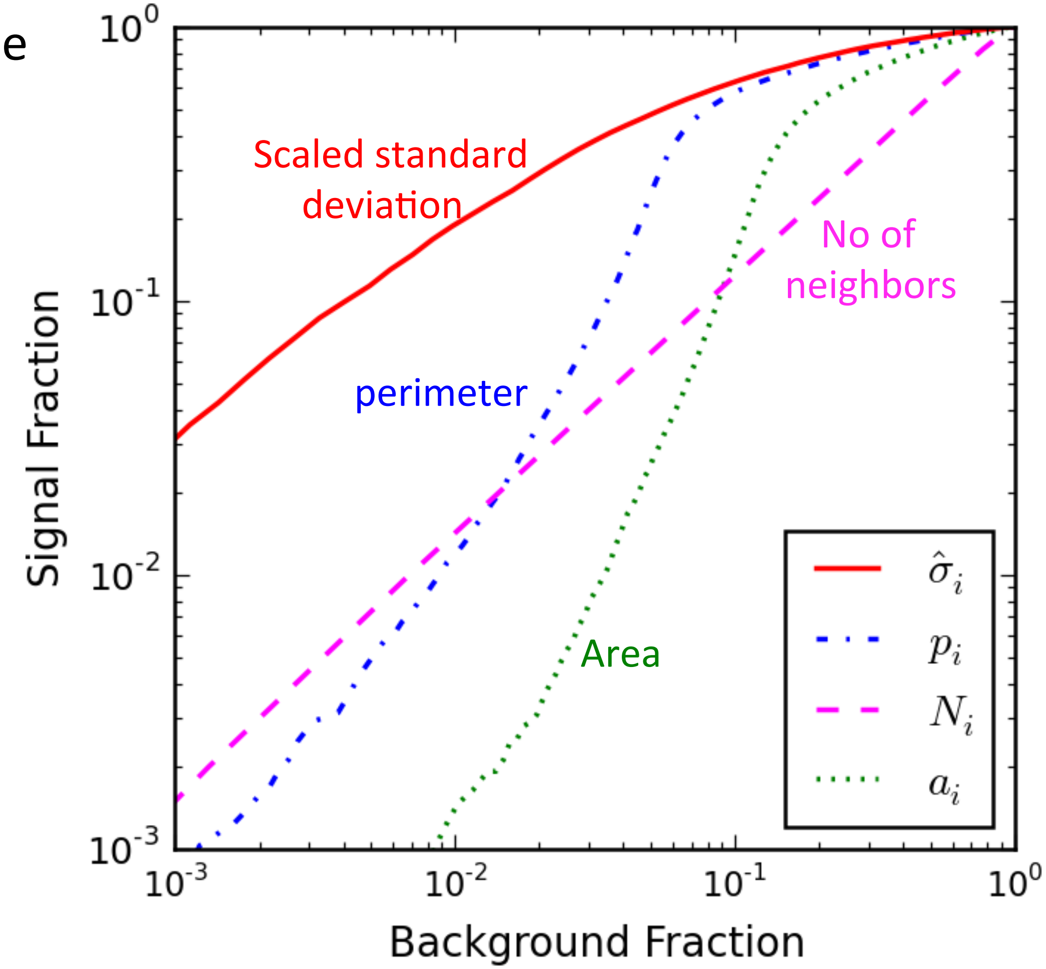
  Number of sides



Method of Edge detection

ROC curves

- Signal selection efficiency vs. the background efficiency for different values of cut on area, perimeter, no of neighbor, and scaled standard deviation.
- **ROC curve for Scaled standard deviation is well-separated from the ROC curves of other variables in low background fraction region**.
- Scaled standard deviation is quite successful in identifying edge cells.

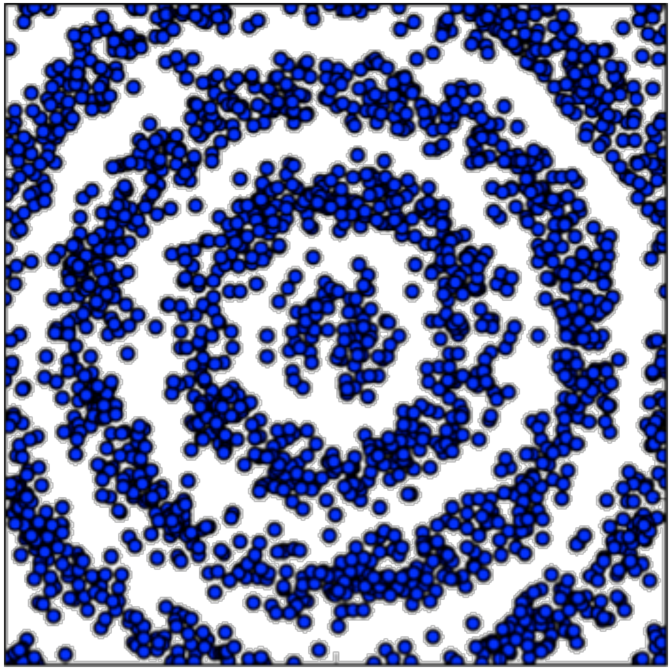


Why Voronoi approach?

- Voronoi tessellations have nice properties for use in HEP data:
  - **Automatic** binning
  - **Preserves maximum spatial resolution**: Each cell has its own bin with shape determined by tessellation
  - Applicable in **many dimensions**

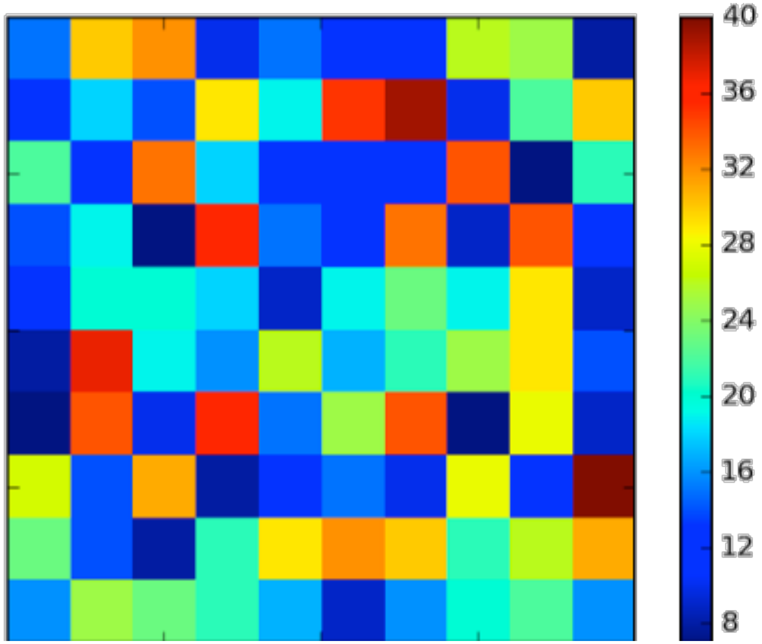
Example: We generate 2000 random data points using pdf

$$f(x,y) = 1 + \sin\left(6\pi\sqrt{x^2 + y^2}\right)$$



Binning vs Voronoi tessellations

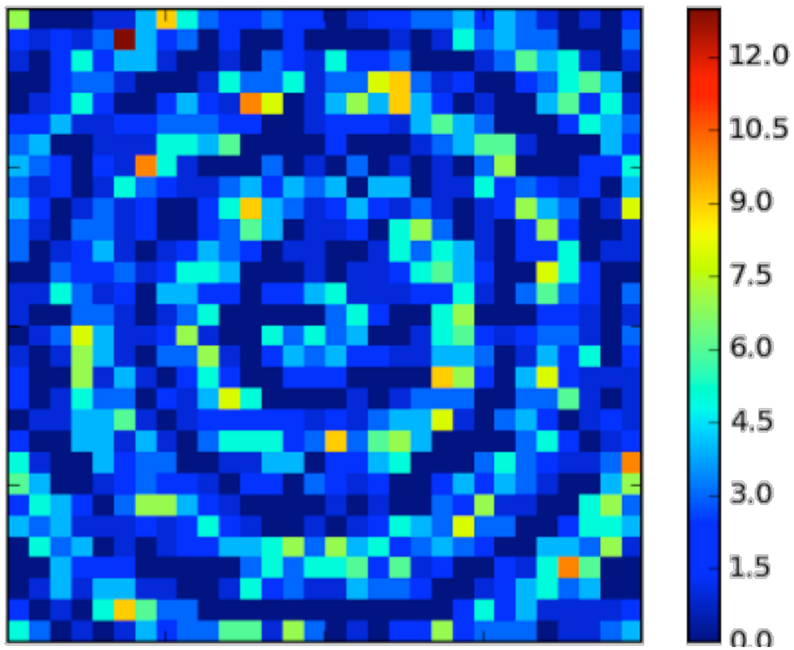
Color: bin height



Why Voronoi approach?

Binning vs Voronoi tessellations

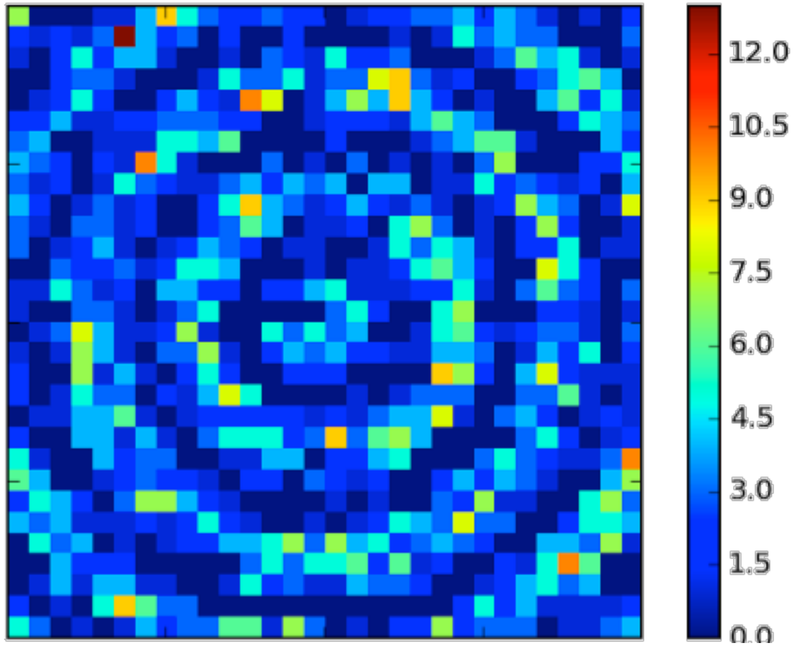
Color: bin height



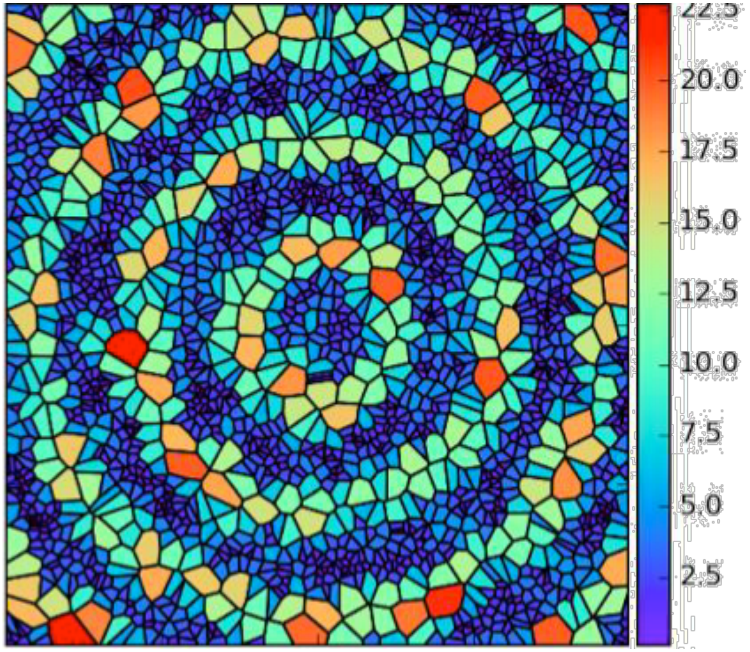
Why Voronoi approach?

Binning vs Voronoi tessellations

Color: bin height



Color: Cell area



(Debnath, Gainer, Kim, Matchev, 2015)

Advantage of Voronoi method: Each cell contains one data point **prevents poor choices of binning** from obscuring structure in the data.