Clustering

Andrew Askew

Clustering:

- What is a cluster?

Clustering:

- What is a cluster?
- No, seriously, that’s a real question.
- Once upon a time I was writing a paper that mentioned calorimetry in context and it was decided collectively that “cluster” did not require a definition because “everyone knows what it means”.

Clustering:

- What is a cluster?
- No, seriously, that's a real question.
- I sat down to write out a definition once that I found did the trick: A cluster of energy in the calorimeter is a local group of calorimeter elements that are spatially consistent with a shower.
- Originally I wrote that out for EM objects, which meant I really implied electromagnetic shower.
- Whatever, what this really means is that you are making a logical connection between calorimeter elements in order to make further hypotheses about the origin of the energy deposition.

Clusters and Tracks:

- You'll note we HAVE tracks, but we're not DOING tracking.
  - PFlow starts with the tracks as inputs, but performs the clustering with its own algorithm.
  - That algorithm, with different settings, is used for both ECAL and HCAL.
  - This algorithm is a two step process: first creating "topological" clusters, and second, attempting to find subclusters within the topological clusters.

Step 1:

- Topological clusters are pretty easy. There are only a couple of parameters:
  - Seed threshold: at least one element in the cluster must have an energy (or optionally ET) above a threshold.
  - Gathering threshold: other elements associated with this cluster must have some threshold energy.
- Topological clusters are contiguous regions of energy with at least one seed. This is accomplished via a nearest-neighbor algorithm.

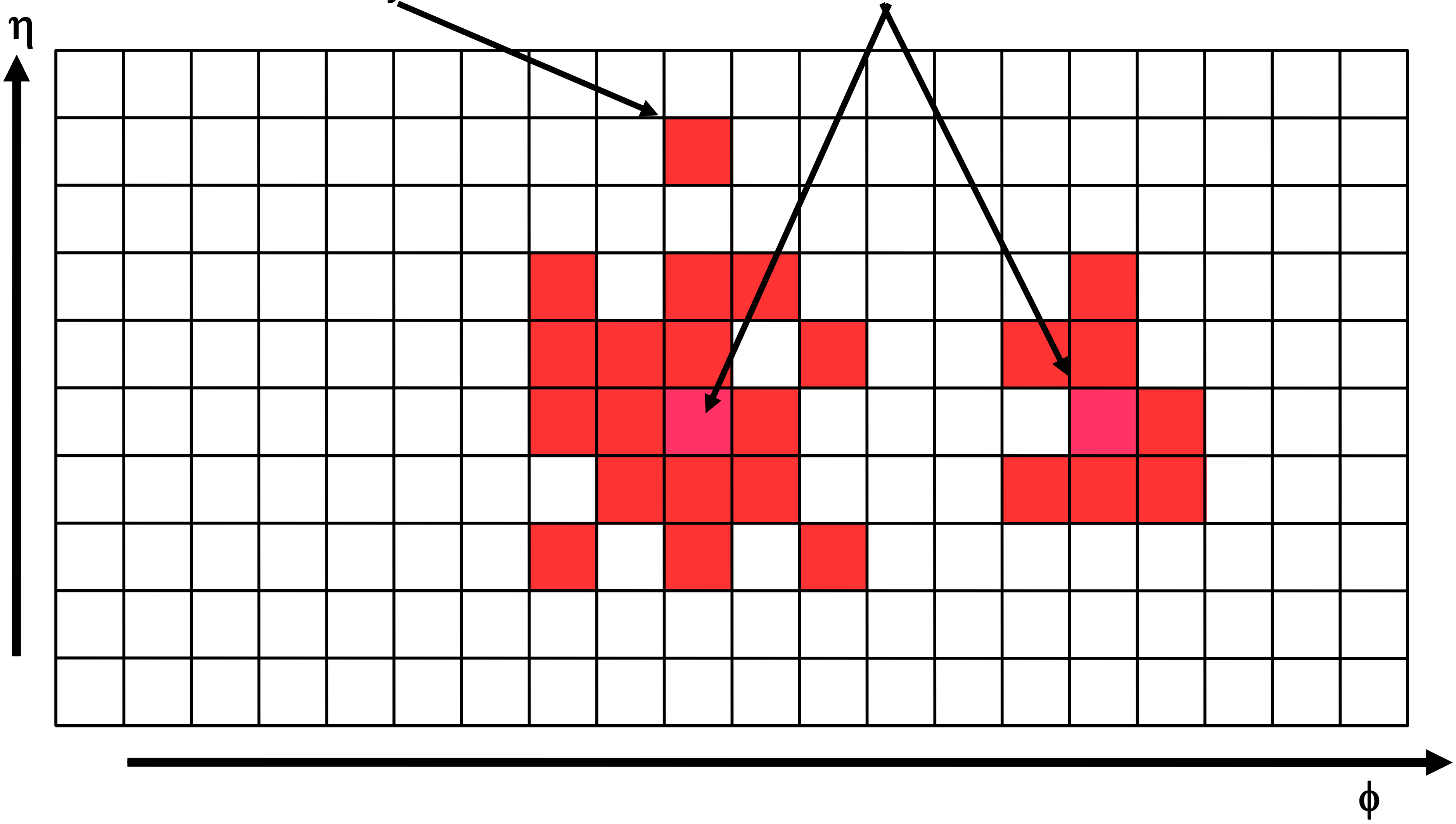
Neighbors?

- Except for the edges, crystals within the barrel ECAL (or towers within the barrel HCAL) are surrounded by eight other elements (forming a three-by-three array).
- For selecting seeds, particle flow can be configured to use only four (the ones that share faces) or all eight elements. A seed MUST be the highest energy of its neighbors. Our standard configuration is neighbors-4 for HCAL and neighbors-8 for ECAL.

The Particle Flow algorithm

- If you recognize the jargon, the process by which crystals are added to a topological cluster is a recursive nearest-neighbor algorithm.
  - E.g. I take a step in a given direction and check my neighbor…
    - Then I take a step from the neighbor, to the next neighbor…
      - Then I take a step from the neighbor's neighbor, to the next neighbor…
        - And so on and so forth, until I run out of neighbors (and proceed to the next step, working back up the chain) or go insane…
- The illustration will make this slightly easier, but the entire contiguous area of energy will be added to the topological cluster, terminating where the crystals drop below threshold. In this case, the crystals that are white are either negative energy, or below the threshold for adding them.
  - The threshold for a seed in the ECAL is 0.23.
  - The threshold for adding crystals in the ECAL is 0.08.

PFlow Clustering:



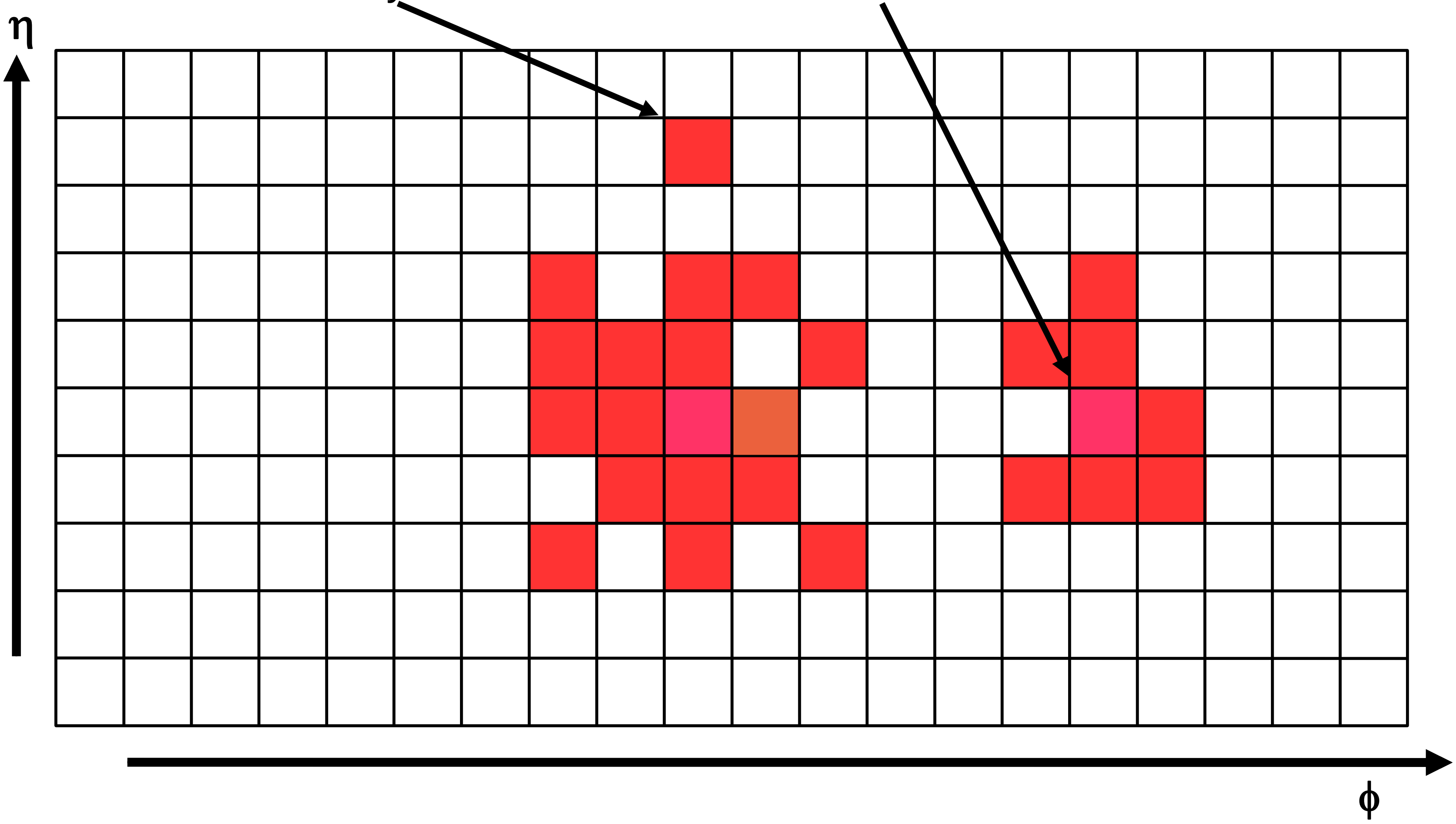
Unclustered crystal.

Seed cells, to start clustering.

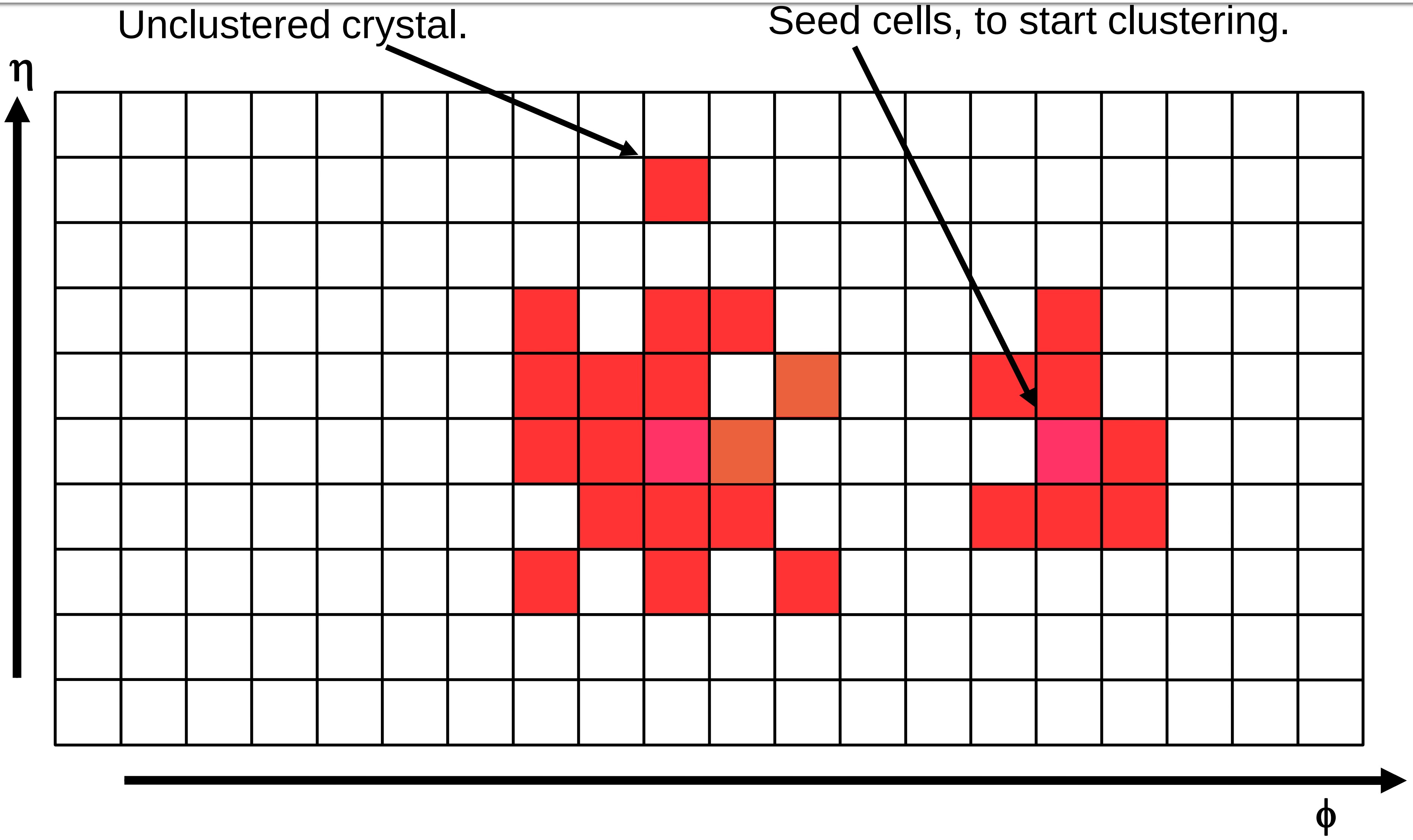
PFlow Clustering:

Unclustered crystal.

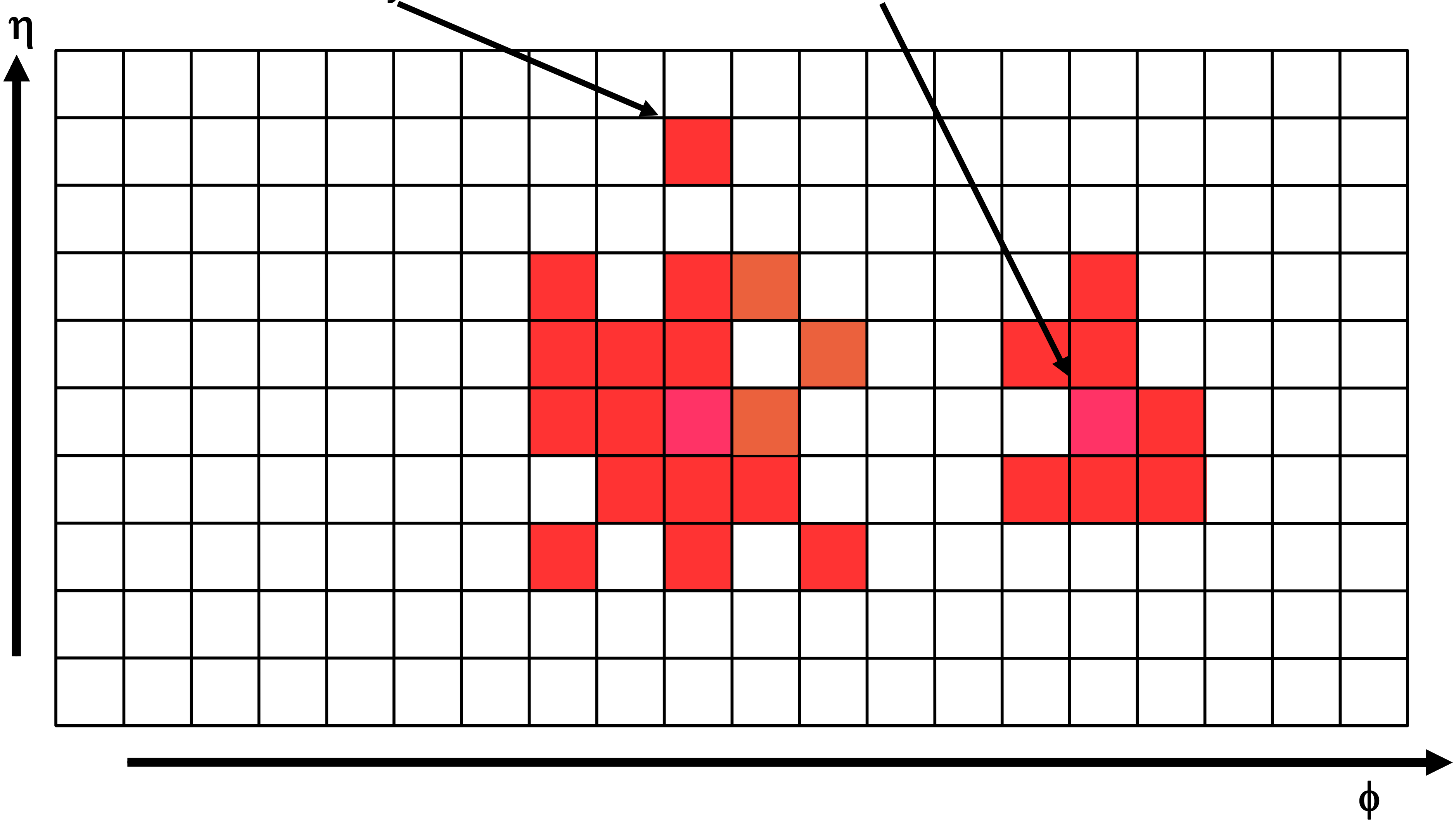
Seed cells, to start clustering.



PFlow Clustering:



PFlow Clustering:



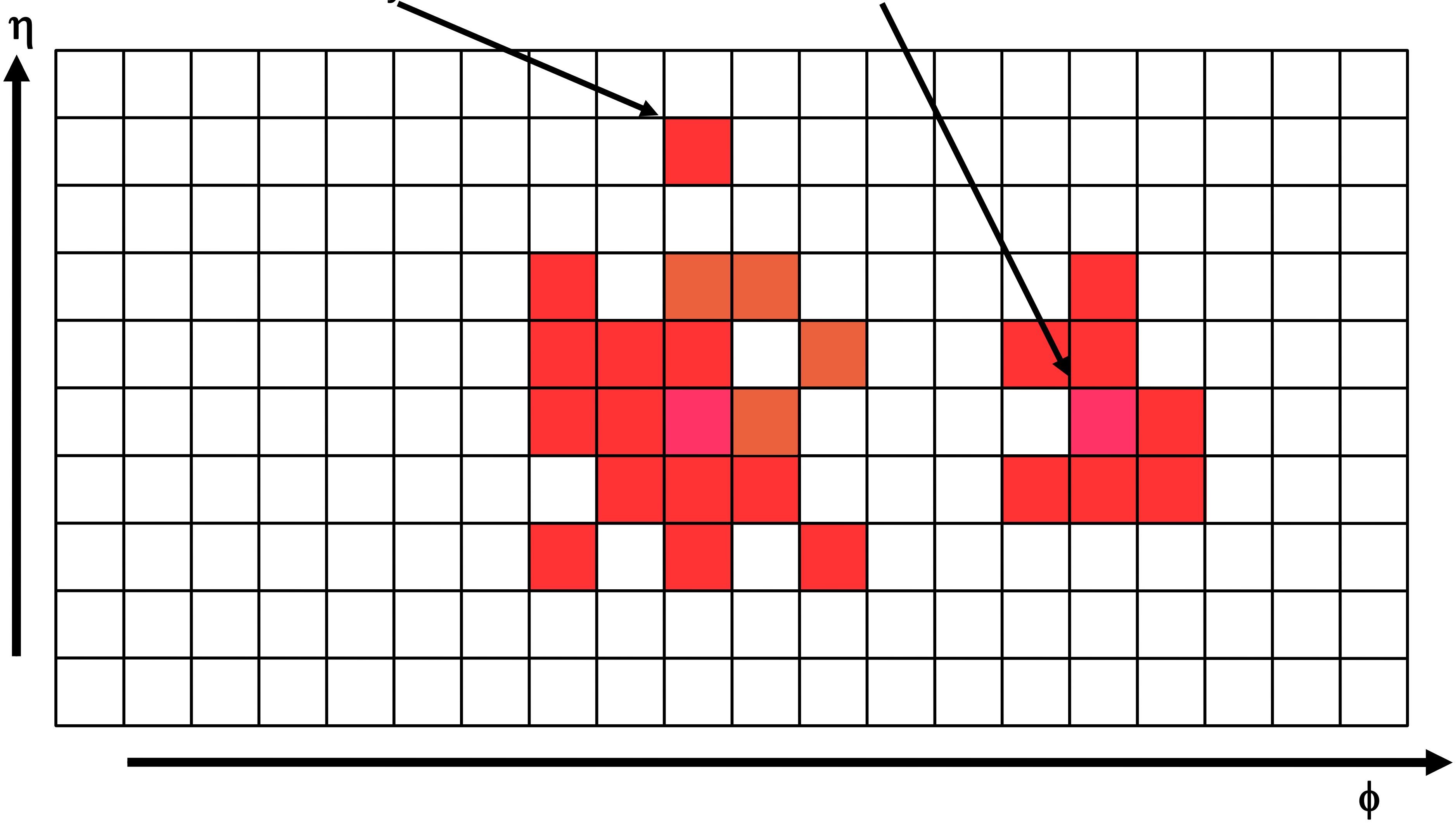
Unclustered crystal.

Seed cells, to start clustering.

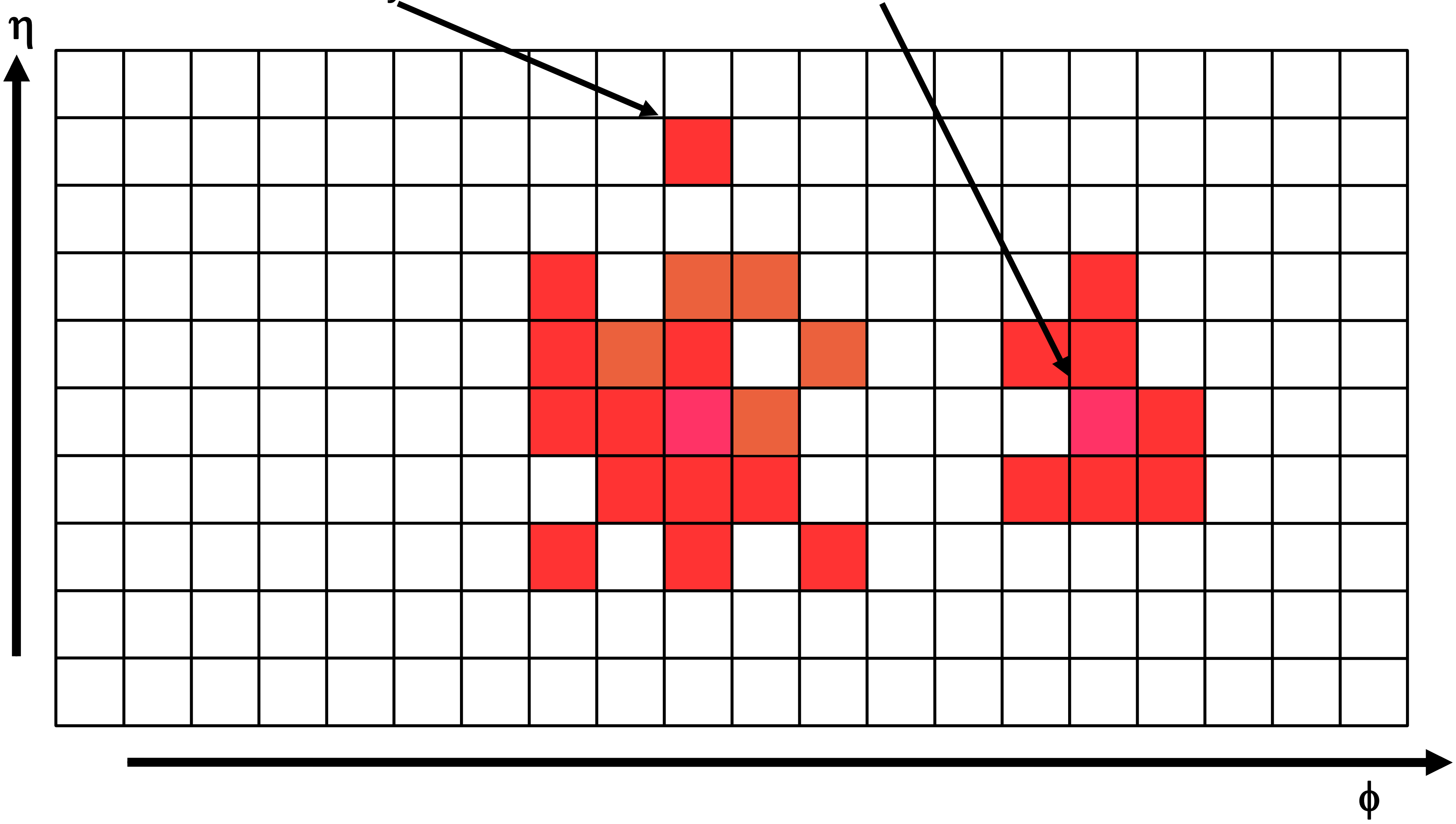
PFlow Clustering:

Unclustered crystal.

Seed cells, to start clustering.



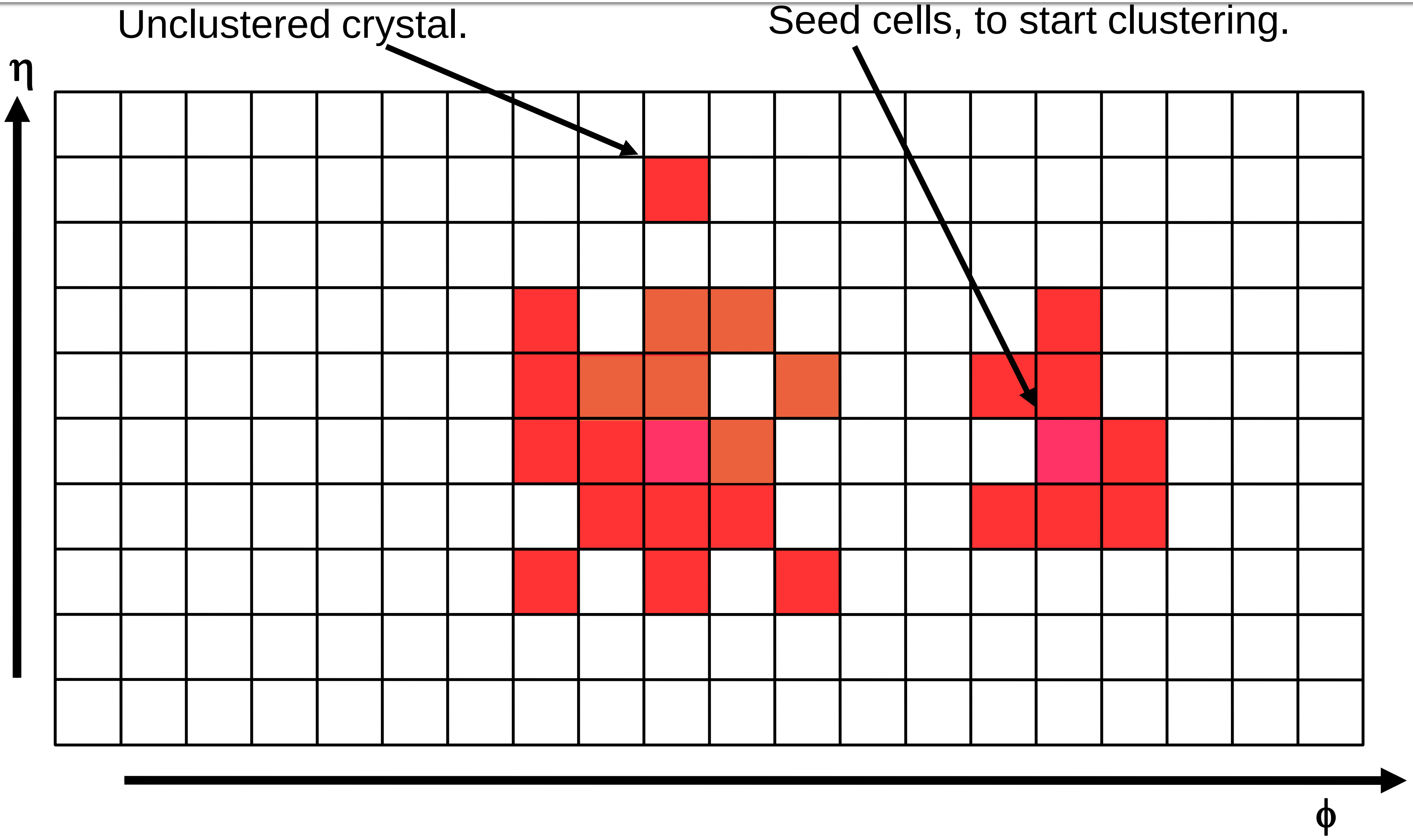
PFlow Clustering:



Unclustered crystal.

Seed cells, to start clustering.

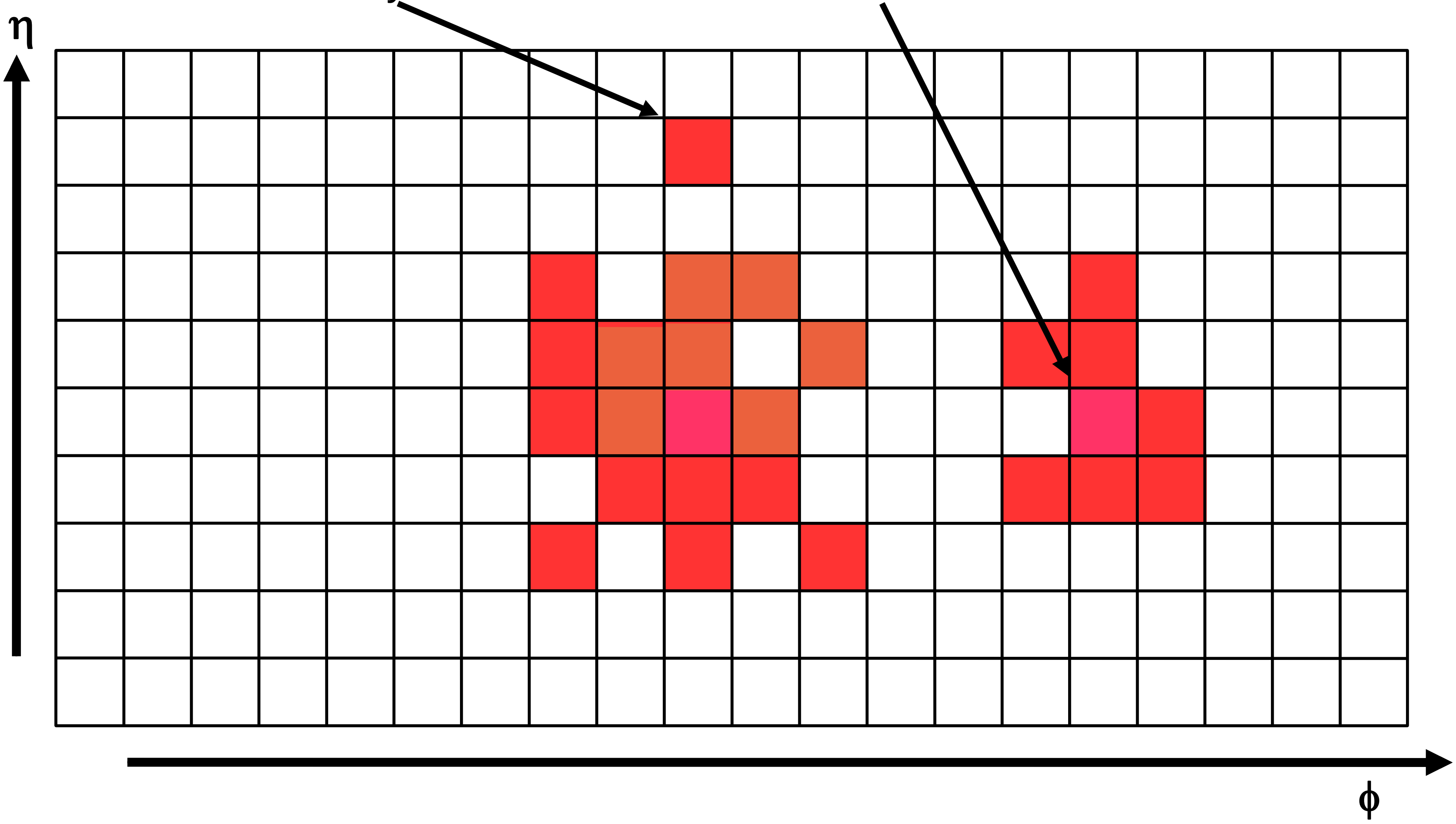
PFlow Clustering:



Unclustered crystal.

Seed cells, to start clustering.

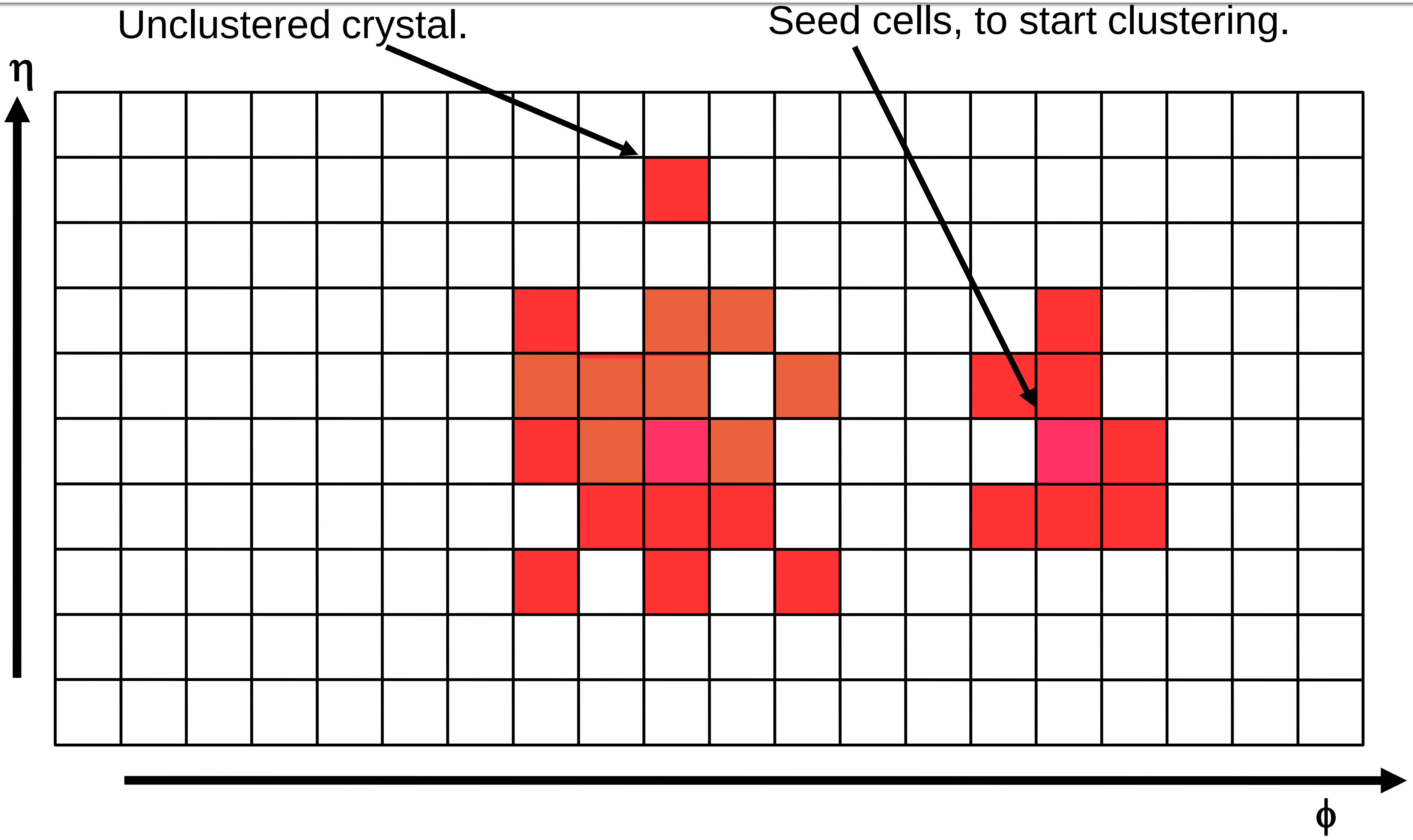
PFlow Clustering:



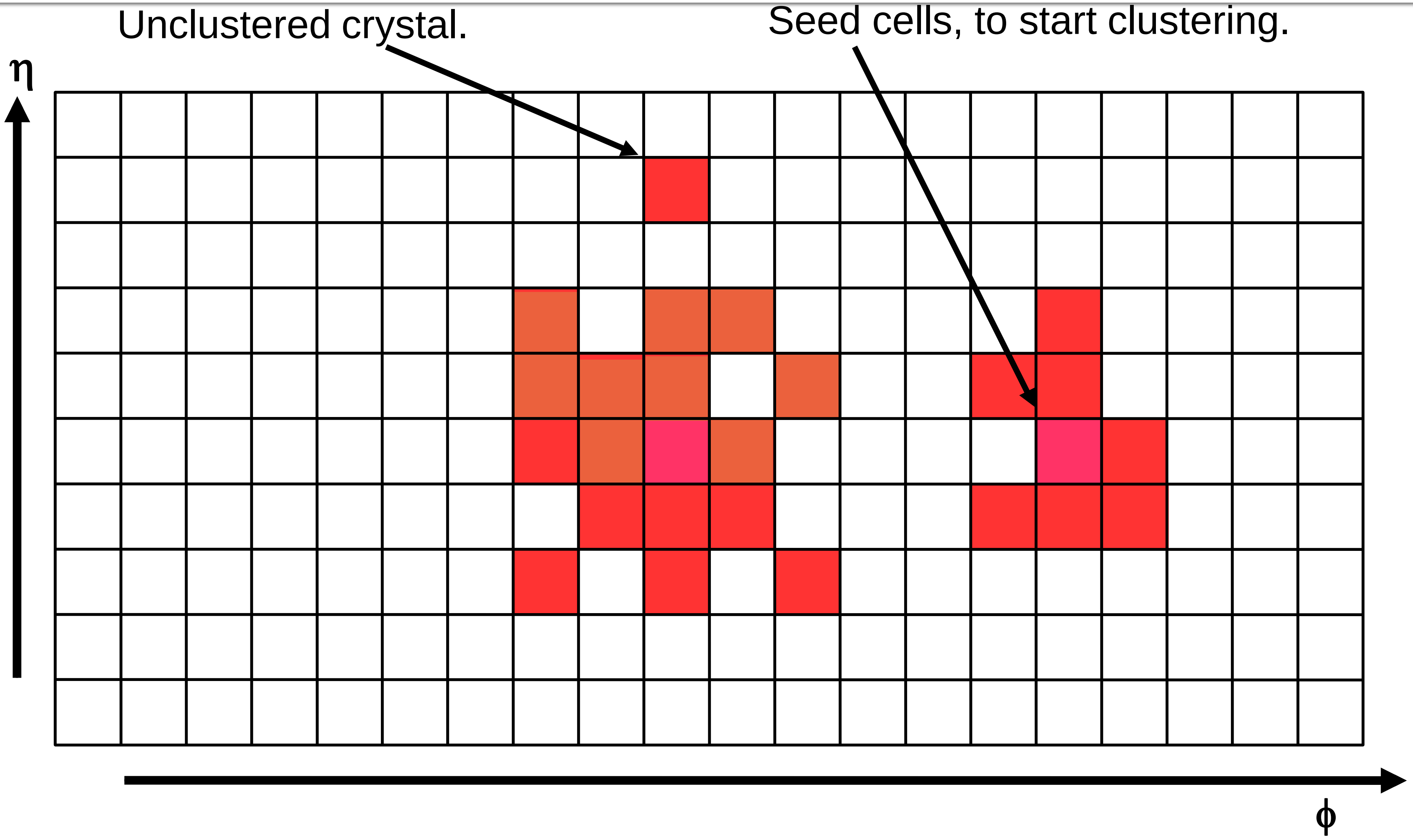
Unclustered crystal.

Seed cells, to start clustering.

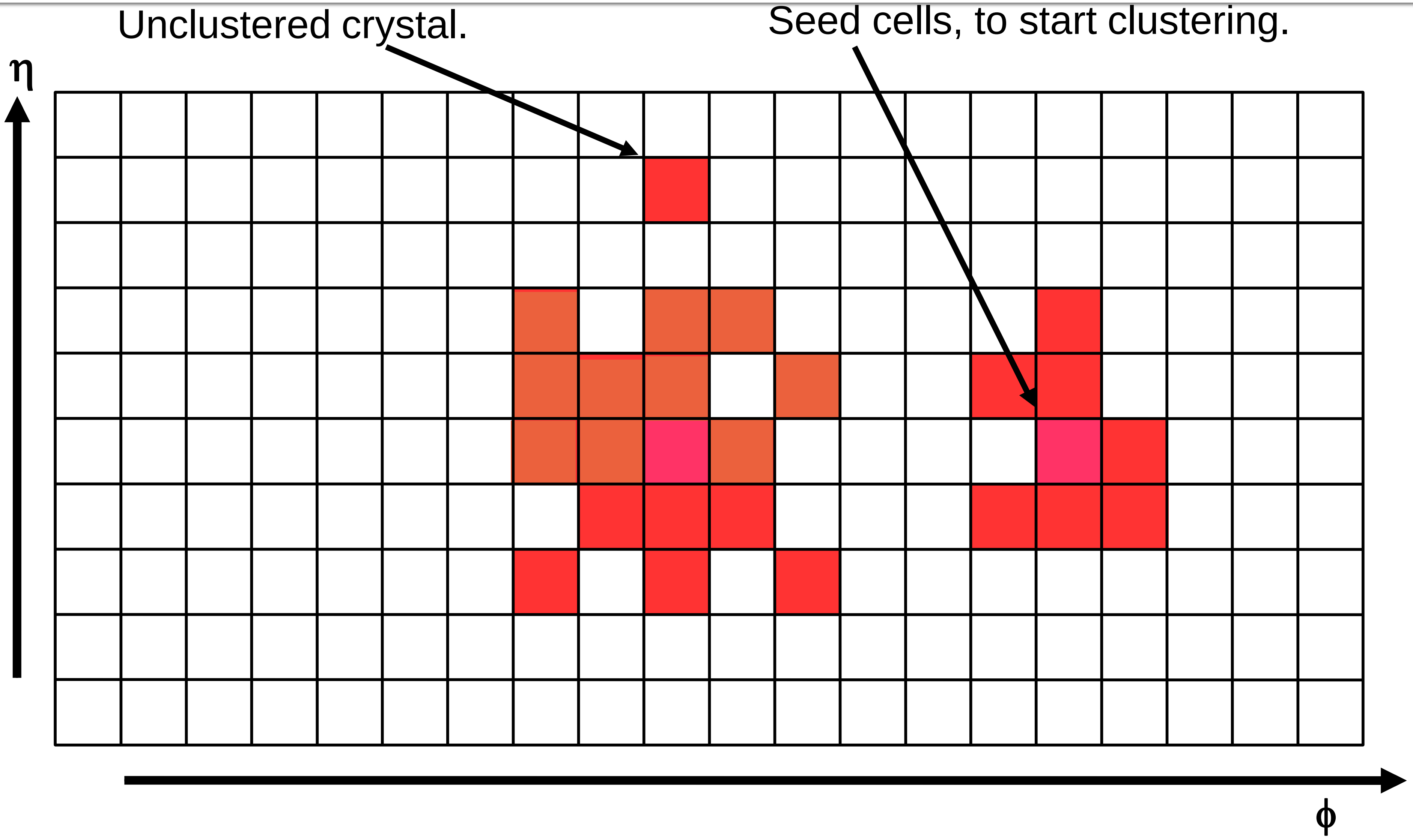
PFlow Clustering:



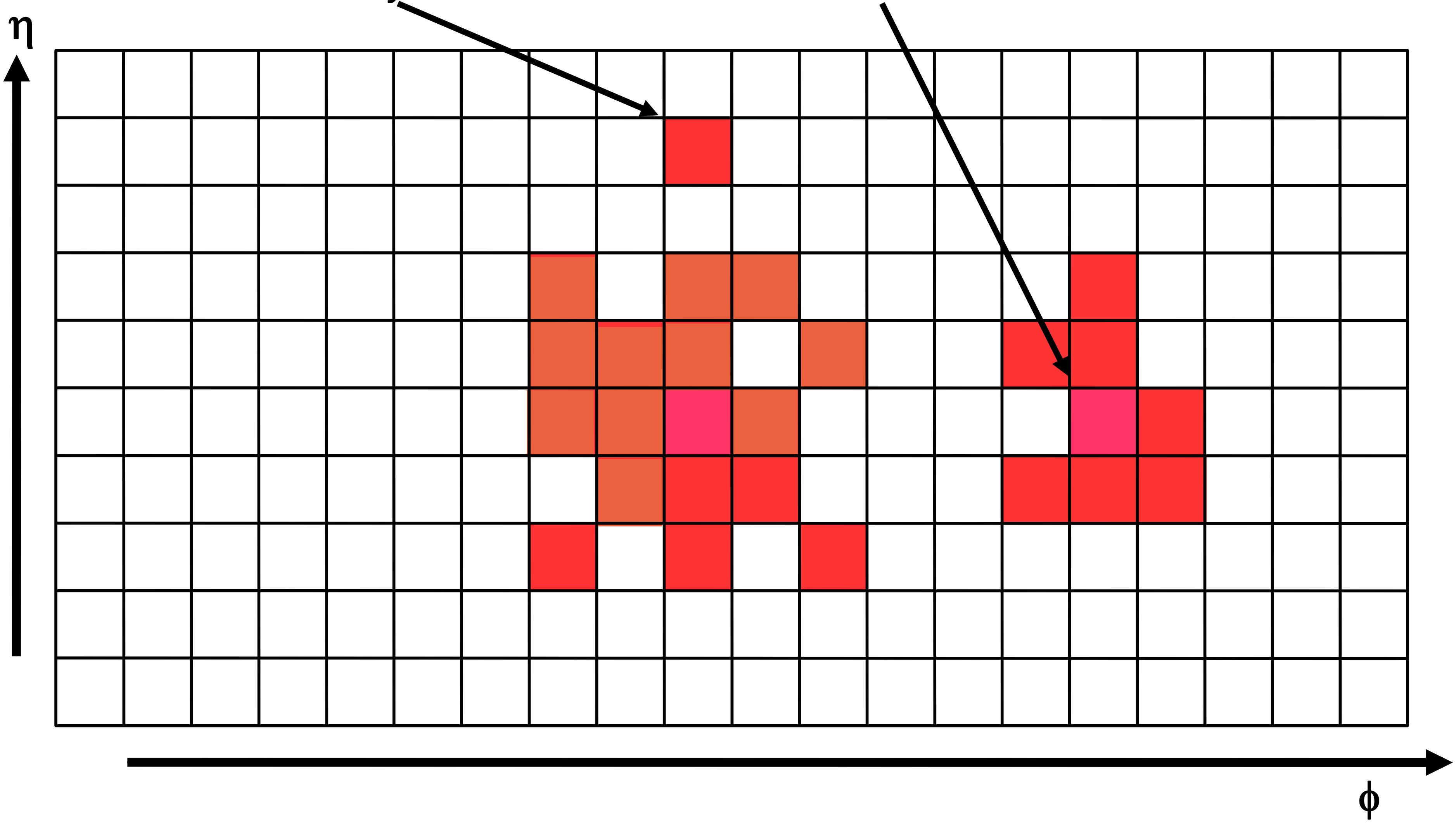
PFlow Clustering:



PFlow Clustering:

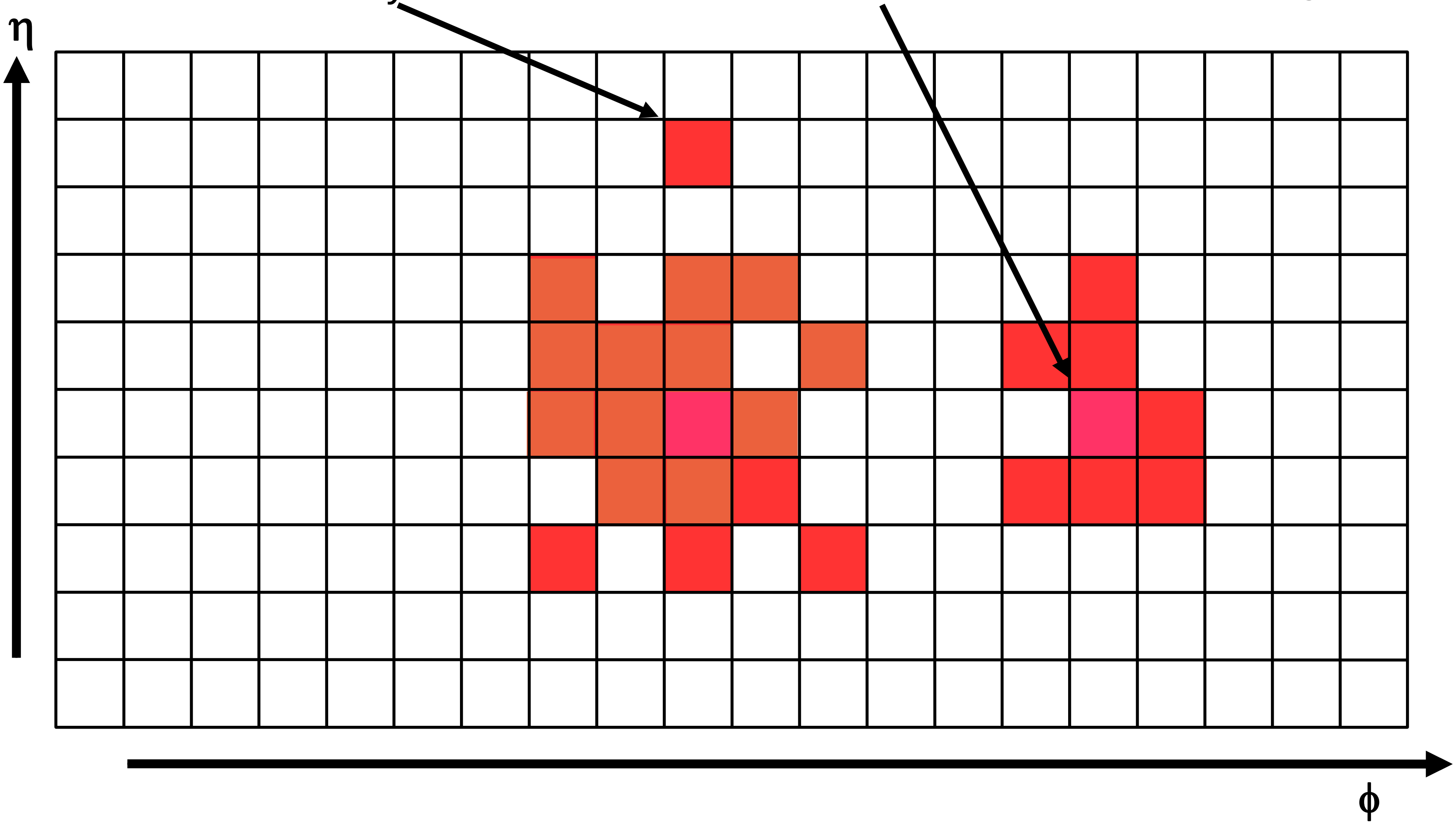


PFlow Clustering:


Unclustered crystal.

Seed cells, to start clustering.

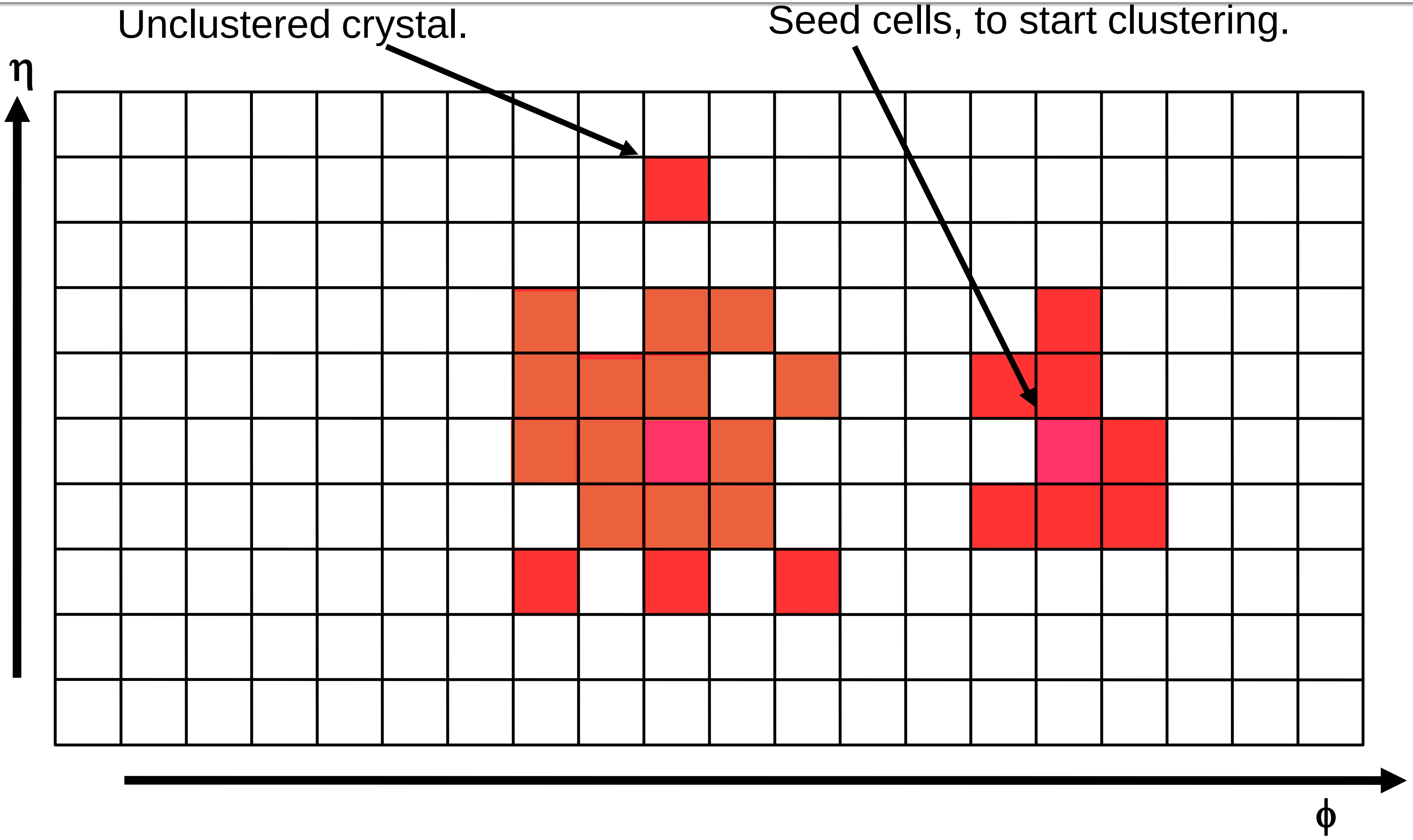
PFlow Clustering:



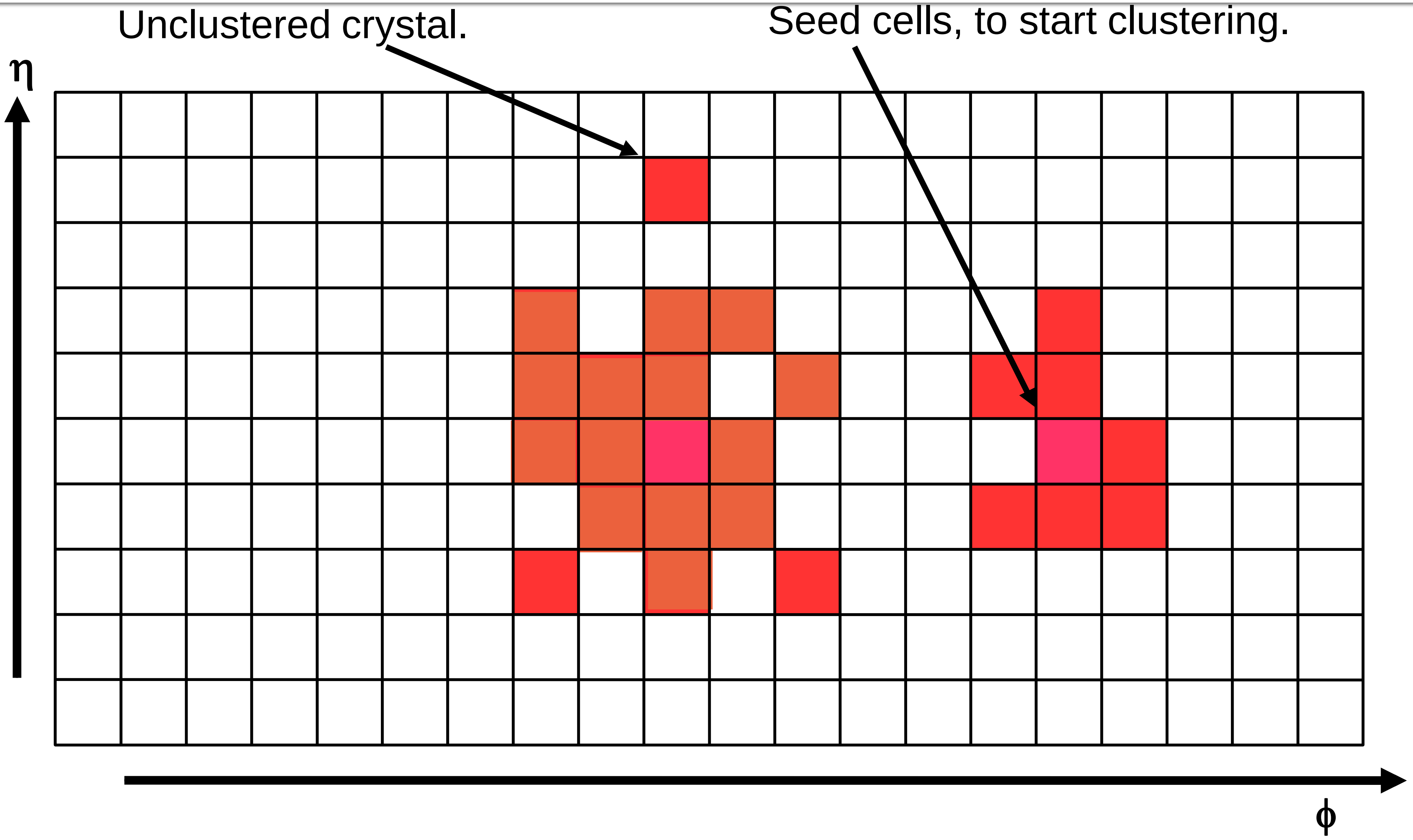
Unclustered crystal.

Seed cells, to start clustering.

PFlow Clustering:



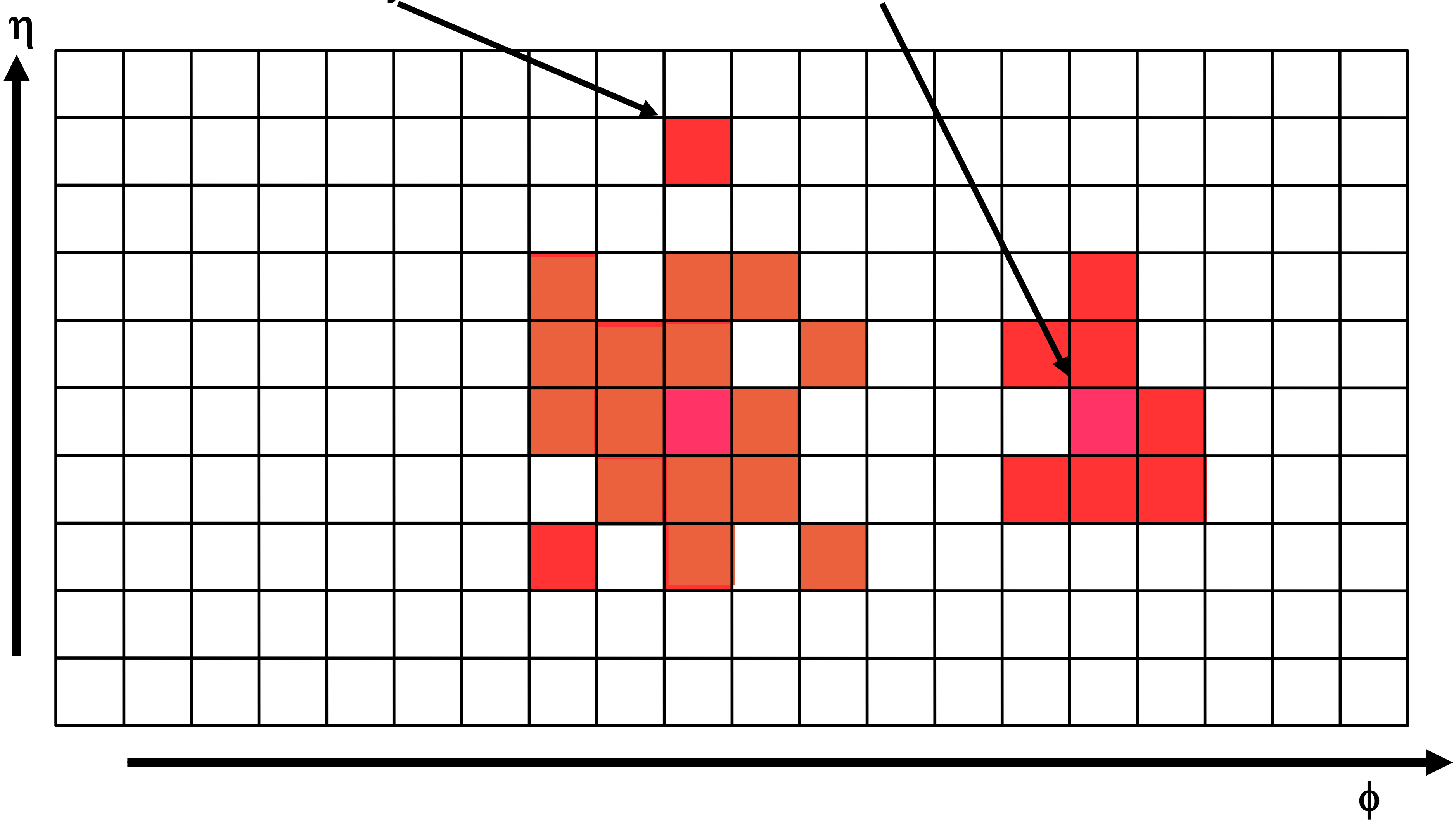
PFlow Clustering:



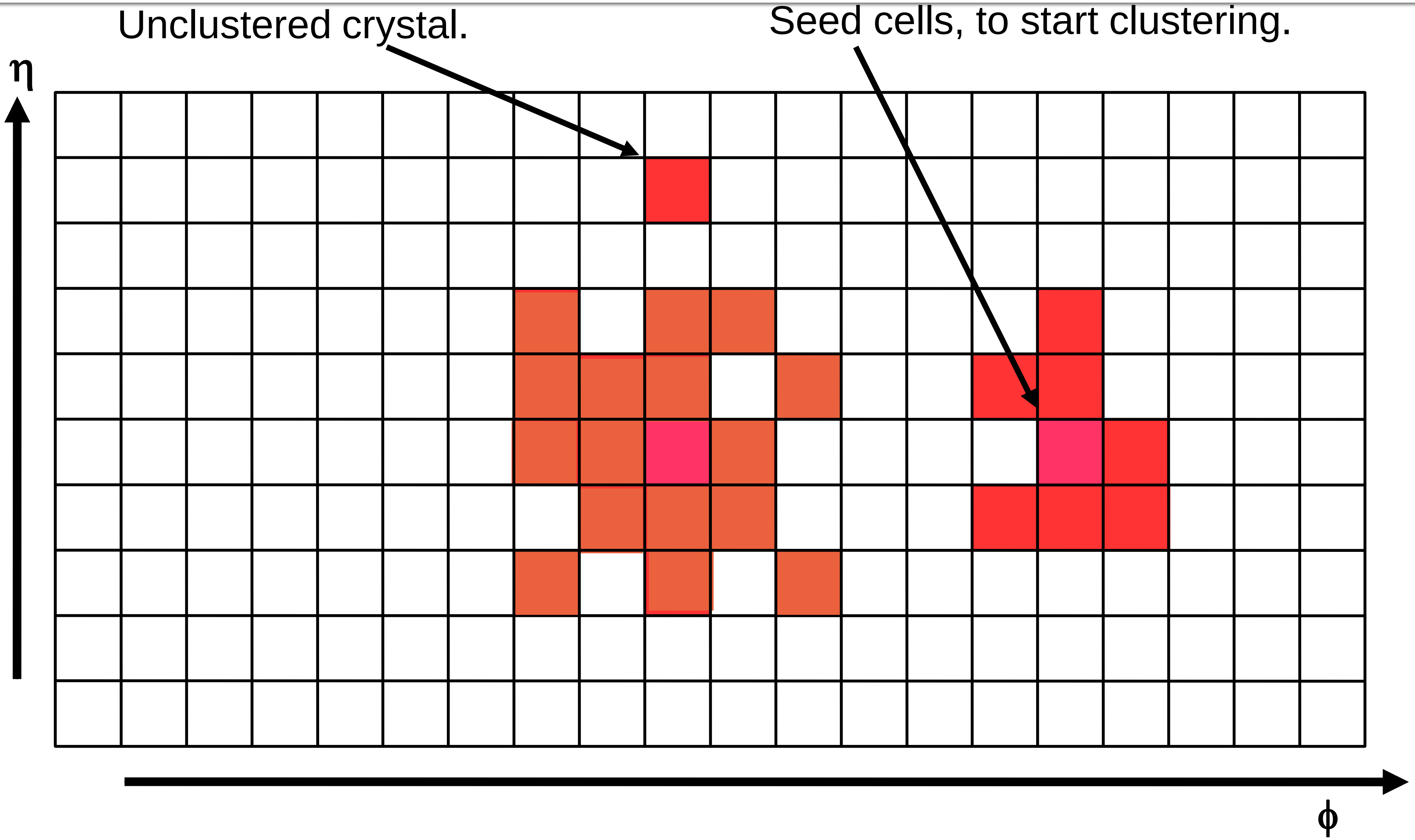
Unclustered crystal.

Seed cells, to start clustering.

PFlow Clustering:

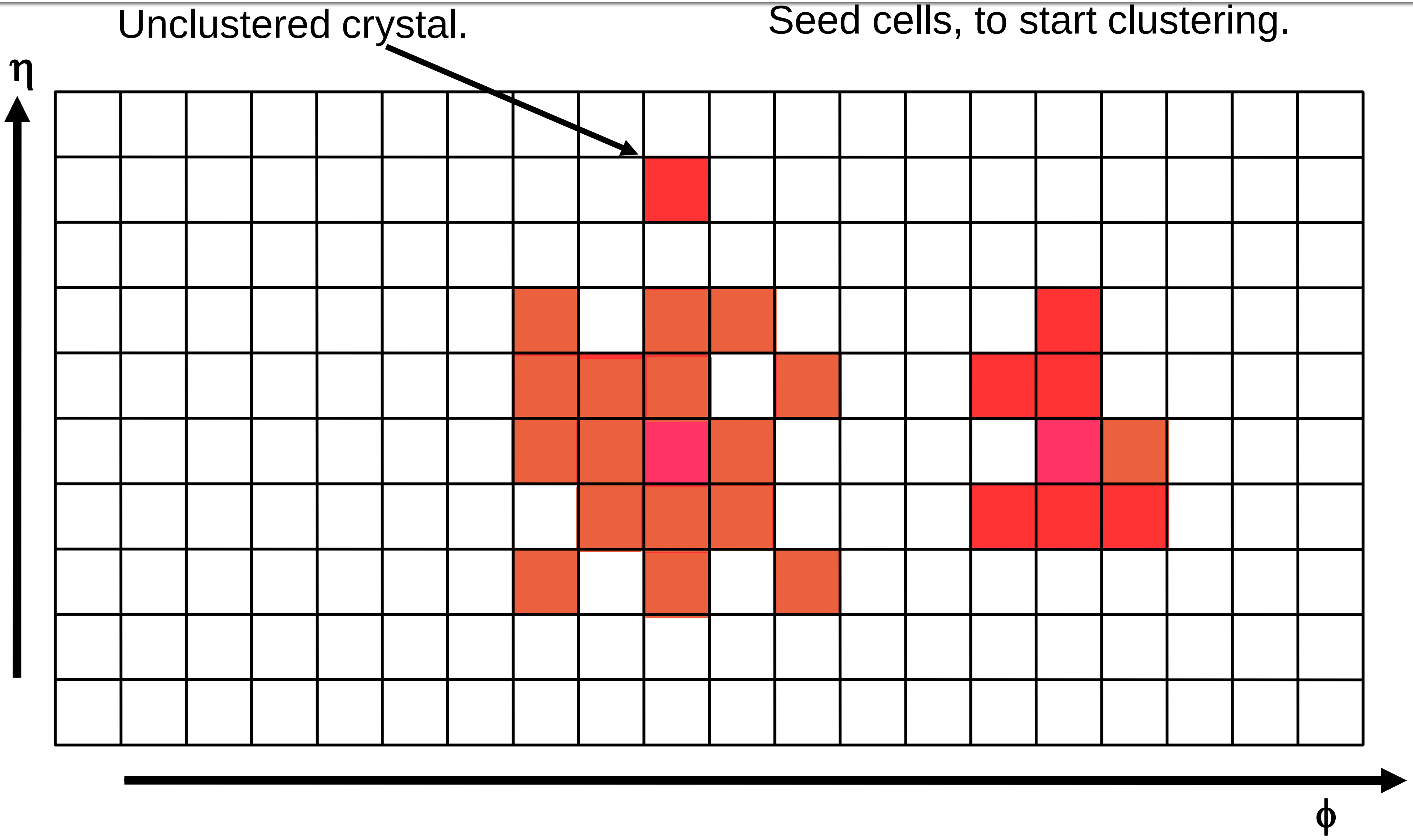


PFlow Clustering:

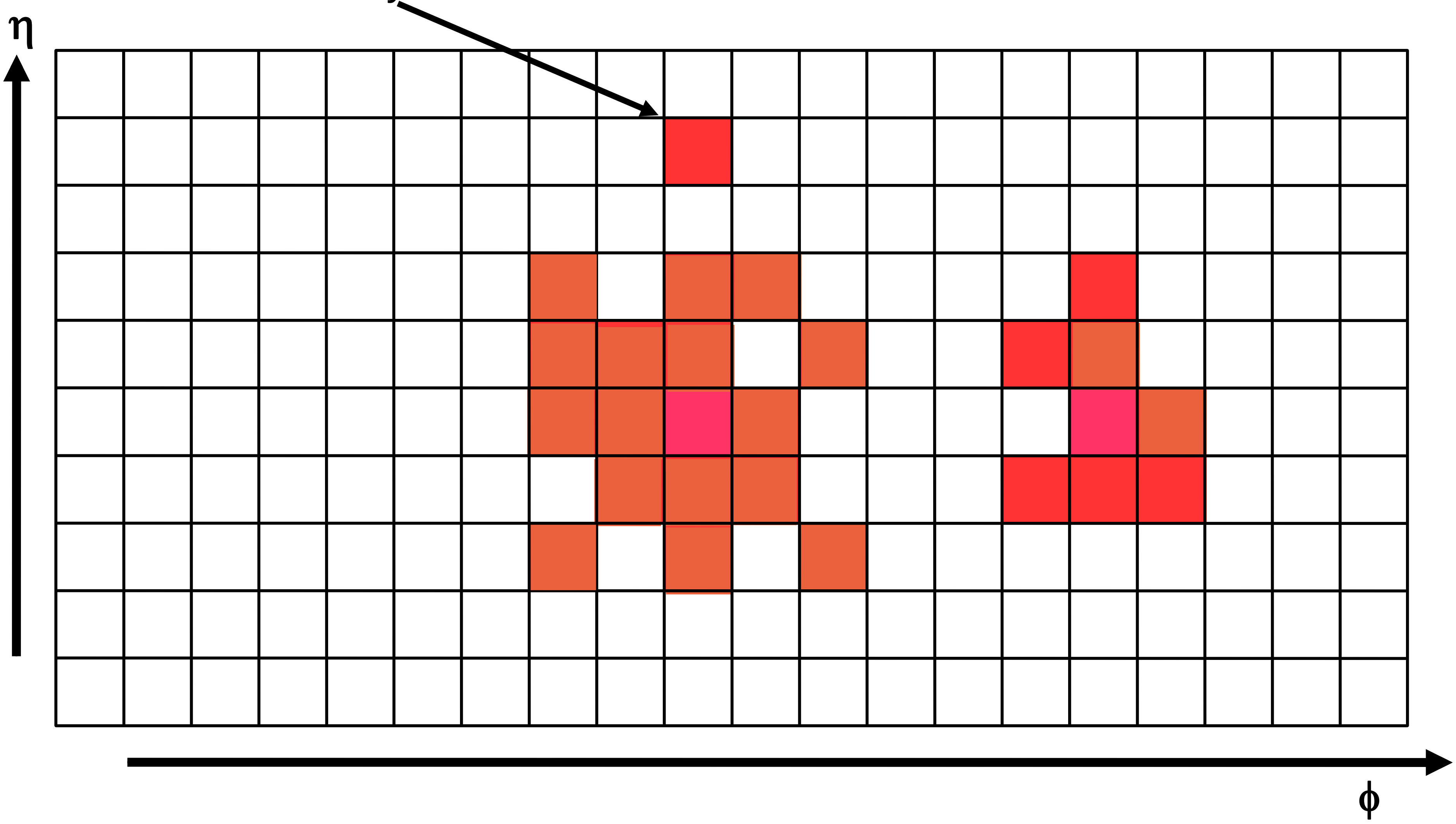

Unclustered crystal.

Seed cells, to start clustering.

PFlow Clustering:



PFlow Clustering:



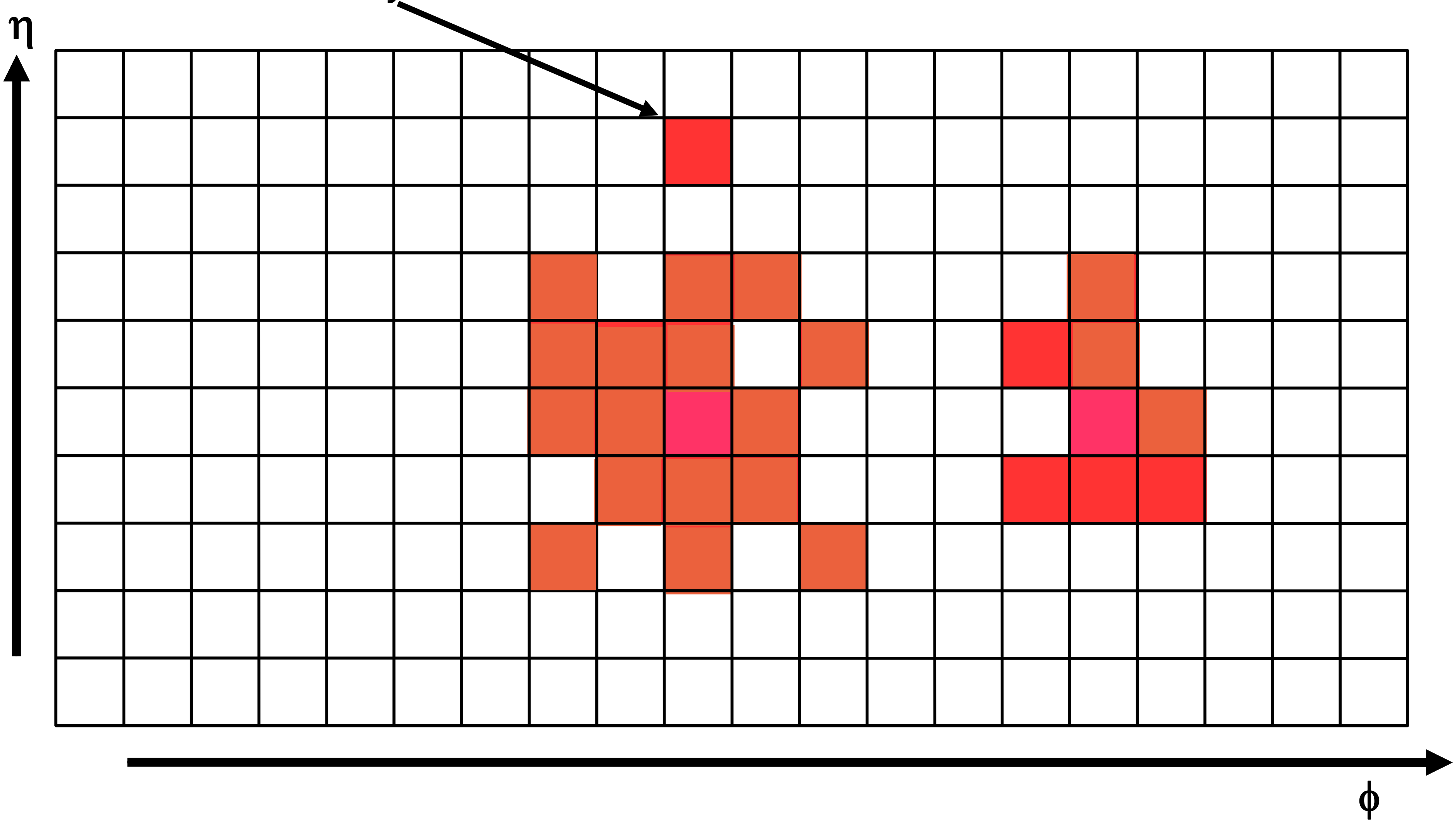
Unclustered crystal.

Seed cells, to start clustering.

PFlow Clustering:

Unclustered crystal.

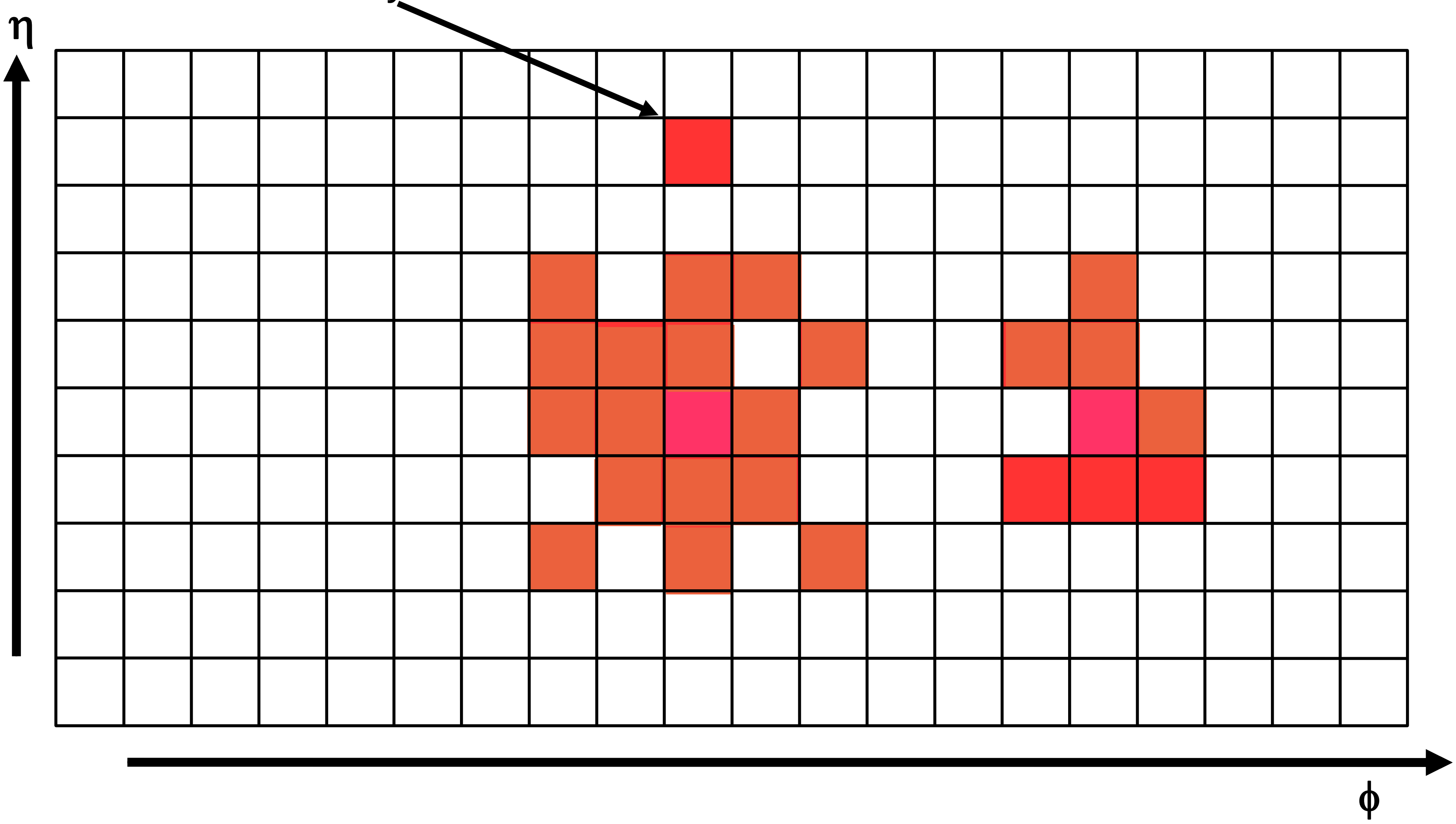
Seed cells, to start clustering.



PFlow Clustering:

Unclustered crystal.

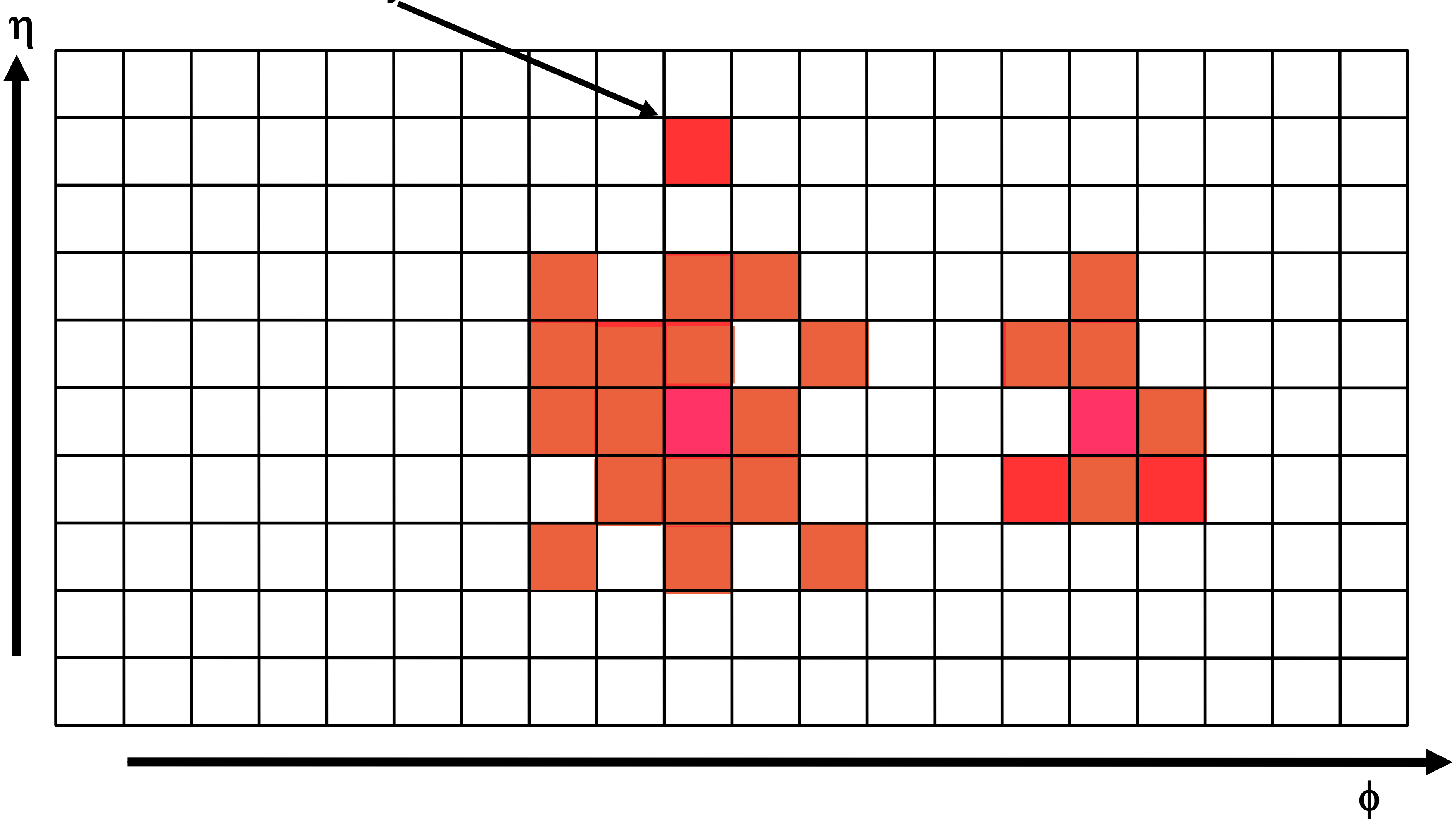
Seed cells, to start clustering.



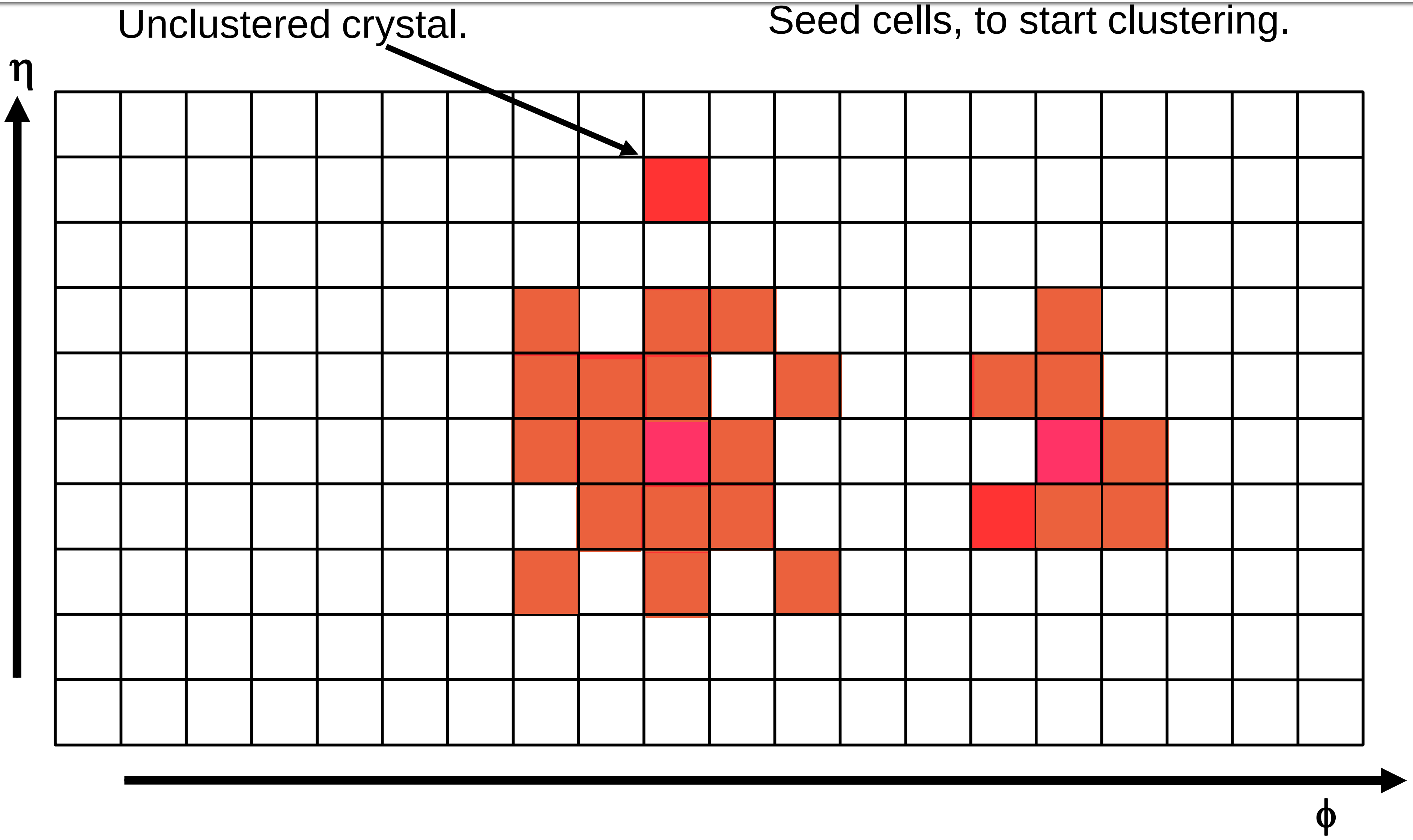
PFlow Clustering:

Unclustered crystal.

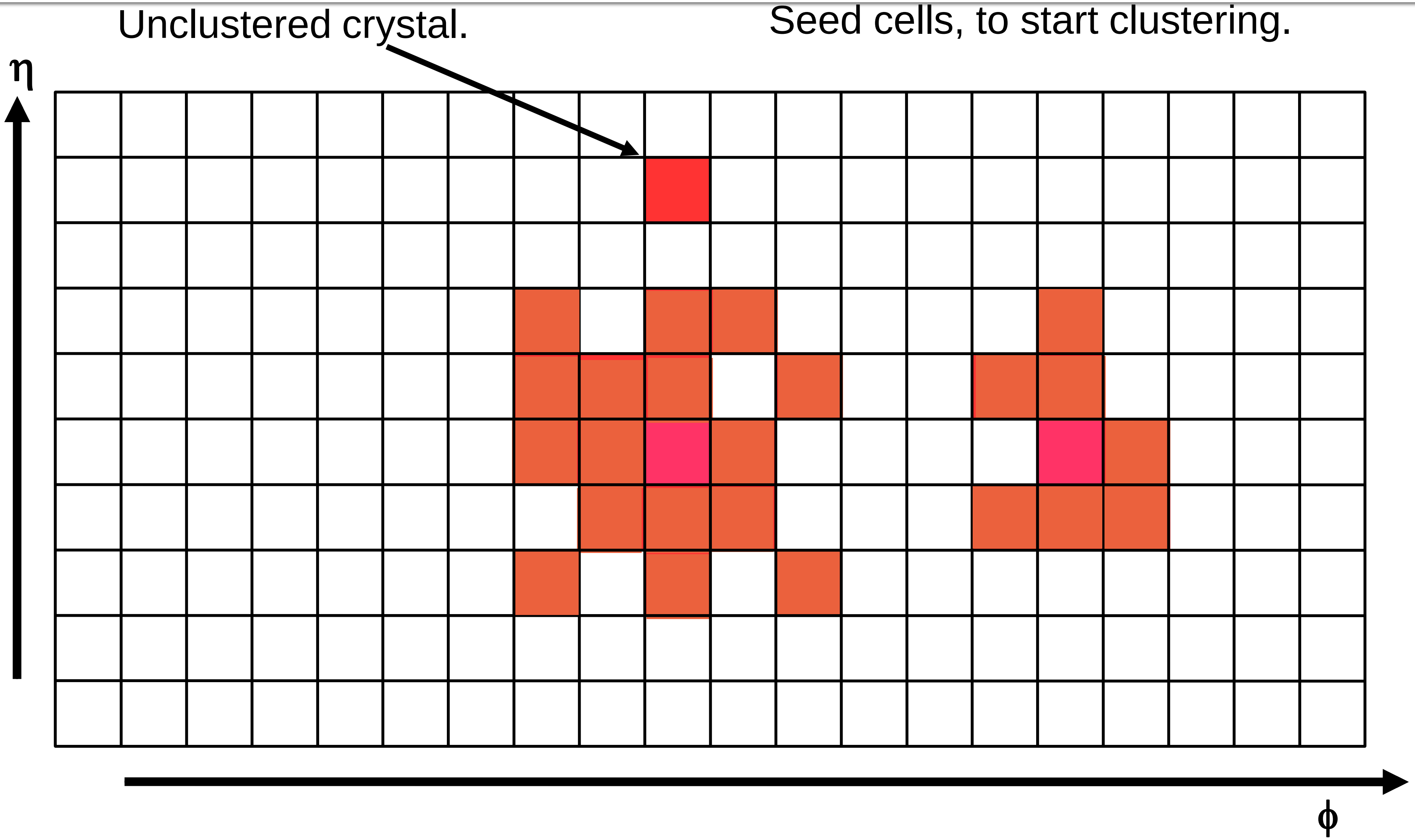
Seed cells, to start clustering.



PFlow Clustering:



PFlow Clustering:


Unclustered crystal.

Seed cells, to start clustering.

HCAL

- The HCAL is slightly easier, since the granularity is larger.
- The thresholds are different too:
  - Seed threshold in HCAL is 0.8 GeV.
  - The gathering threshold is also 0.8 GeV.
- Just to reiterate, HCAL also uses neighbors-4, not all eight.

Next step:

- If there is only one seed in the topological cluster, then good news: you're done!
    - This happens a lot with isolated EM-showers, like from electrons and photons.
- If there is more than one seed in the topological cluster, particle flow allows the sharing of the energy between the two maxima, under the assumption that the individual showers are Gaussian.
- The sharing is decided iteratively.

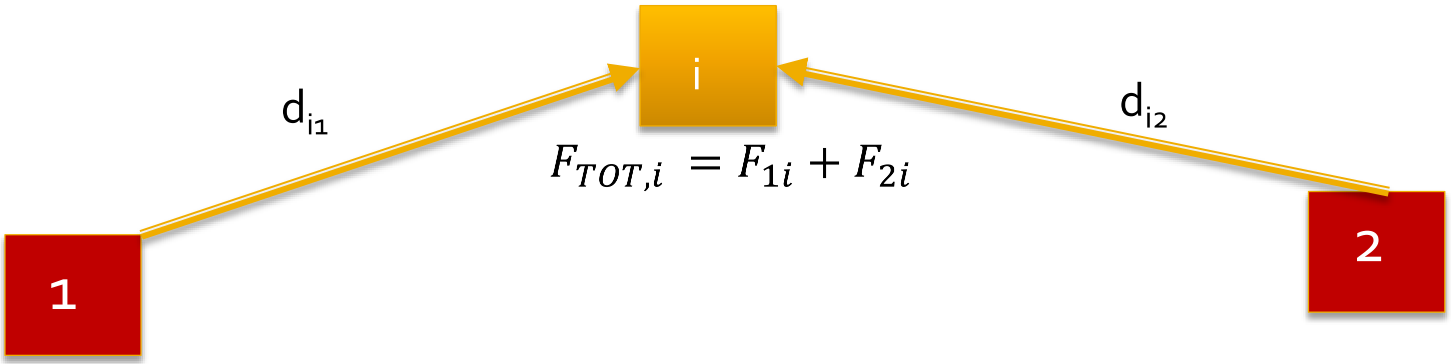
Sharing energy:

- Each iteration of the algorithm is a two-step process:
  - Start from the position of each of the clusters.
    - BTW—Seeds of clusters always award 100% of their energy to their own clusters, that’s the only energy that CANNOT be shared.
  - Calculate the fraction of energy of each element separately according to:
  - $$F = \frac{E_{CLUS}}{E_{NORM}} * \frac{e^{\frac{-\left(\overrightarrow{\mathrm{R}_{\mathrm{crys}}} - \overrightarrow{R_{clus}}\right)^{2}}{2\sigma^{2}}}}{F_{TOT}}$$
  - A lot going on there.

Definitions:

- $F = \frac{E_{CLUS}}{E_{NORM}} * \frac{e^{\frac{-\left(\overrightarrow{\mathrm{R}_{\mathrm{crys}}}-\overrightarrow{R_{clus}}\right)^{2}}{2\sigma^{2}}}}{F_{TOT}}$
- E_{CLUS} is the cluster energy, E_{NORM} is a normalization parameter (usually the gathering threshold), the numerator in the exponent is the magnitude of the distance between the cluster position and the hit position, and σ is the expected Gaussian width of the cluster (10cm for HCAL, 1.5cm for ECAL) .
- F_{TOT} is the sum of all of the fractions calculated from all the seeds (ensures unitarity).

Example:



$$F_{TOT,i} = F_{1i} + F_{2i}$$

$$F_{1i} = \frac{E_1}{E_{norm}} * e^{\left(\frac{-d_{i1}^2}{2\sigma^2}\right)}$$

$$F_{2i} = \frac{E_2}{E_{norm}} * e^{\left(\frac{-d_{i2}^2}{2\sigma^2}\right)}$$

Cluster 1 awarded fraction: $F = \frac{F_{i1}}{F_{TOT}}$

Cluster 2 awarded fraction: $F = \frac{F_{i2}}{F_{TOT}}$

- So here I have simply replaced the distances from cluster 1 to crystal i with the Cartesian distance, and done the same with cluster 2. Sigma is a constant related to the Moliere radius, and ultimately E_{NORM}, doesn't matter.

Step 2:

- Once you have assigned all of the energy, you recalculate the cluster energy and cluster position.
- The cluster energy is just the sum of the hits times the fraction assigned.
- The position is calculated for ECAL according to:
  - $x = \frac{\Sigma w_i x_i}{\Sigma w_i}, w_i = \max(0, w_0 + \log\left(\frac{E_i}{E_{TOT}}\right))$, where w_0=4.2, and E_i is the energy of the hit times the fraction assigned and E_{TOT} is the cluster energy.
- The position is calculated for HCAL according to:
  - $x = \frac{\Sigma w_i x_i}{\Sigma w_i}, w_i = \max(0, \log\left(\frac{E_i}{E_{norm}}\right))$, where Ei is the energy of the hit times the fraction assigned, and Enorm is typically the gathering threshold.

How do you start?

- I swear this is less complicated than it sounds.
- It’s iterative, so you have to have a starting point. The starting point is that each cluster has the energy of only the seed and the position of the seed.
  - You then calculate the fractions of each non-seed hit
  - Calculate the new energy of the cluster
  - Calculate the new position of the cluster
  - Repeat

How do you stop?

- You stop either when you hit the maximum number of iterations OR when the energies/positions have converged.
  - For both ECAL and HCAL, you have 50 iterations maximum, and a stopping tolerance 1e-8.
- It sounds like every cluster owns every element, but that isn't strictly true, the minimum fraction is 1e-7 (any less and you're excluded from the final cluster).

Linking and Blocks

Andrew Askew

So now we have the “atoms” of PF

- With the calorimeter information formed into logical groups (under the hypothesis of individual showers), the next step is to determine how to combine information across detectors.
- This step is known as “linking”. This is actually technically simpler than what we just did, probably to your relief.

Definitions:

- Two “linked” items are associated by their physical proximity.
- Groups of linked items are known as “blocks”.
- Each block will go on to become one or more particle candidates.
- The simplest example of this is a track which is not in proximity of any clusters in the calorimeter. This becomes a block of one item (the track), which then almost certainly becomes a charged hadron.

A slight aside:

- Clearly we shouldn't test all objects against all other objects (would take forever).
- Technically what is done in the code is to form a KDTree of all the elements and then traverse the tree only testing the closest elements against each other.
- You as humans can skip this step (hopefully you aren't considering linking things that are back to back in phi, for example).

What is meant by proximity (tracks)?

- Tracks are extrapolated to the calorimeter (for our purpose, the position has been given at the ECAL). A link is established if:
  - The position of the track is within the area defined by all the elements in the cluster
  - Further, if within one extra element in any direction, a link can be established (accounts for gaps, cracks, etc.).
- If more than one HCAL cluster can be linked to a track, OR if more than one track can be linked to an ECAL cluster, only the shortest distance link is kept (when making particle hypotheses).

What is meant by proximity (clusters)?

- For linking ECAL clusters to HCAL clusters, a link is formed when the ECAL cluster position is within the cluster envelope (as previously defined).
- When multiple HCAL clusters are linked to a single ECAL cluster, only the shortest link is kept (when making particle hypotheses).

In Plainer English:

- Final linking is decided by size and proximity.
- “A big thing can have multiple little things, but a little thing cannot have multiple big things”—Julie Hogan. A very neat summation that I choose to steal.
- So while an HCAL cluster (the biggest thing) can have multiple ECAL clusters, an ECAL cluster cannot have multiple HCAL clusters. An HCAL cluster can have multiple tracks, but a track may not have multiple HCAL clusters.

Combinations:

- In real life, we link every unique combination we can.
- Blocks include objects which are linked, either directly, or by indirectly linking to objects linked to other objects.
    - That sounded confusing. Consider a track linked to an ECAL cluster, and that ECAL cluster is linked to an HCAL cluster, but the HCAL cluster is not directly linked to the track. They all end up in the block.
- There's no TECHNICAL limit to the number of objects that can fall in a block, but the rules tend to limit the combinations (nearest objects, plus linking multiplicity).

Examples:

- A block could contain:
  - A single track
  - A single ECAL cluster
  - A single HCAL cluster
  - A track AND an ECAL cluster
  - A track AND an HCAL cluster
  - An ECAL cluster and an HCAL cluster
  - A track AND an ECAL cluster AND an HCAL cluster
  - Etc.
- Don't forget: these are NOT particle hypotheses YET. These are associations between objects.