九州大学 海洋システム工学専攻

システム設計特論 （木村）

# (4) 教師なし学習（クラスタリング・自己組織化）

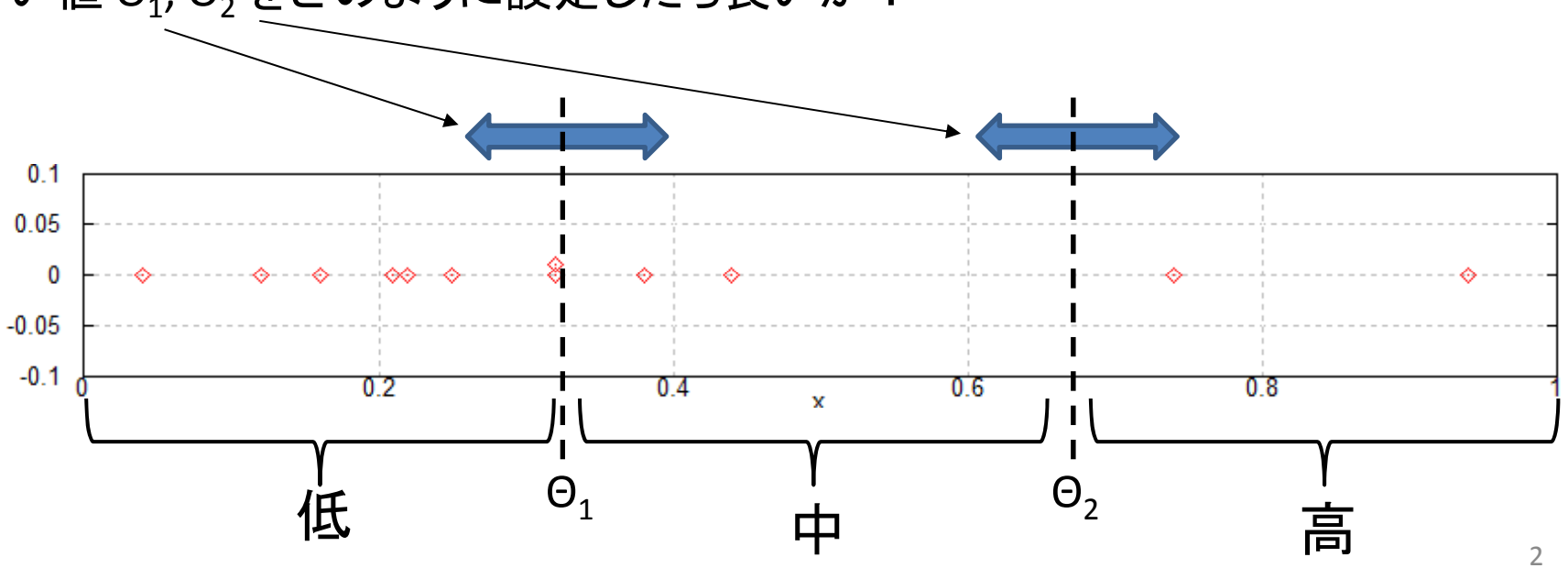
授業の資料等は

http://sysplan.nams.kyushu-u.ac.jp/gen/index.html

# 【復習】データの量子化（クラスタリング・教師無し学習）

| 試行 | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Xの観測値 | 0.21 | 0.74 | 0.32 | 0.25 | 0.12 | 0.38 | 0.32 | 0.22 | 0.04 | 0.44 | 0.94 | 0.16 |

【例】
表のように観測される確率事象において，確率変数 x の観測値は $0 \leqq x < 1$ の連続値だが，あるしきい値 $\Theta_1, \Theta_2$を設定して、$0 \leqq x < \Theta_1$ のとき「低」，
$\Theta_1 \leqq x < \Theta_2$ のとき「中」,$\Theta_2 \leqq x < 1$ のとき「高」, という表示を行うことを考える.
このとき，「低」「中」「高」による表示の**平均情報量を最大化**するには，
しきい値 $\Theta_1, \Theta_2$ をどのように設定したら良いか？



# 【復習】データの量子化（クラスタリング・教師無し学習）

「低」に分類される事象の確率を　$P_1$
「中」に分類される事象の確率を　$P_2$
「高」に分類される事象の確率を　$P_3$　と表すと、全表示の平均情報量Ｉは

$$I = - P_1 \log_2 P_1 - P_2 \log_2 P_2 - P_3 \log_2 P_3$$

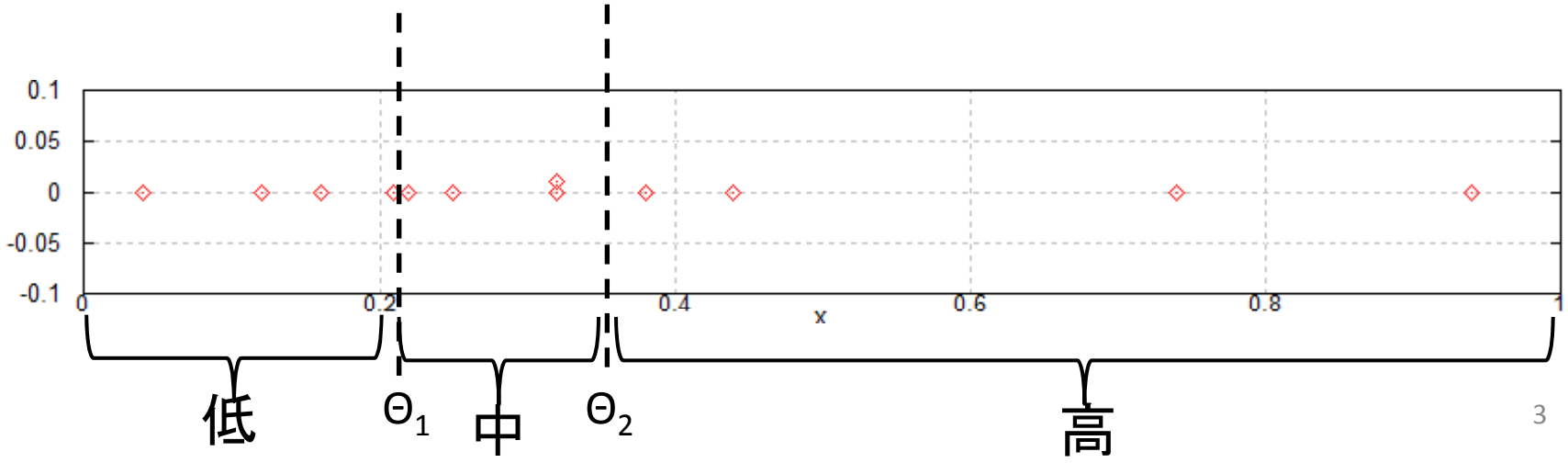
ただし　$P_1 + P_2 + P_3 = 1$ である。
よって等式制約条件下での極値問題になるので、
ラグランジュの未定乗数法によってＩが最大になる $P_1, P_2, P_3$ を求めると

$P_1 = P_2 = P_3 = 1／3$　である。

空間がn次元の場合は?

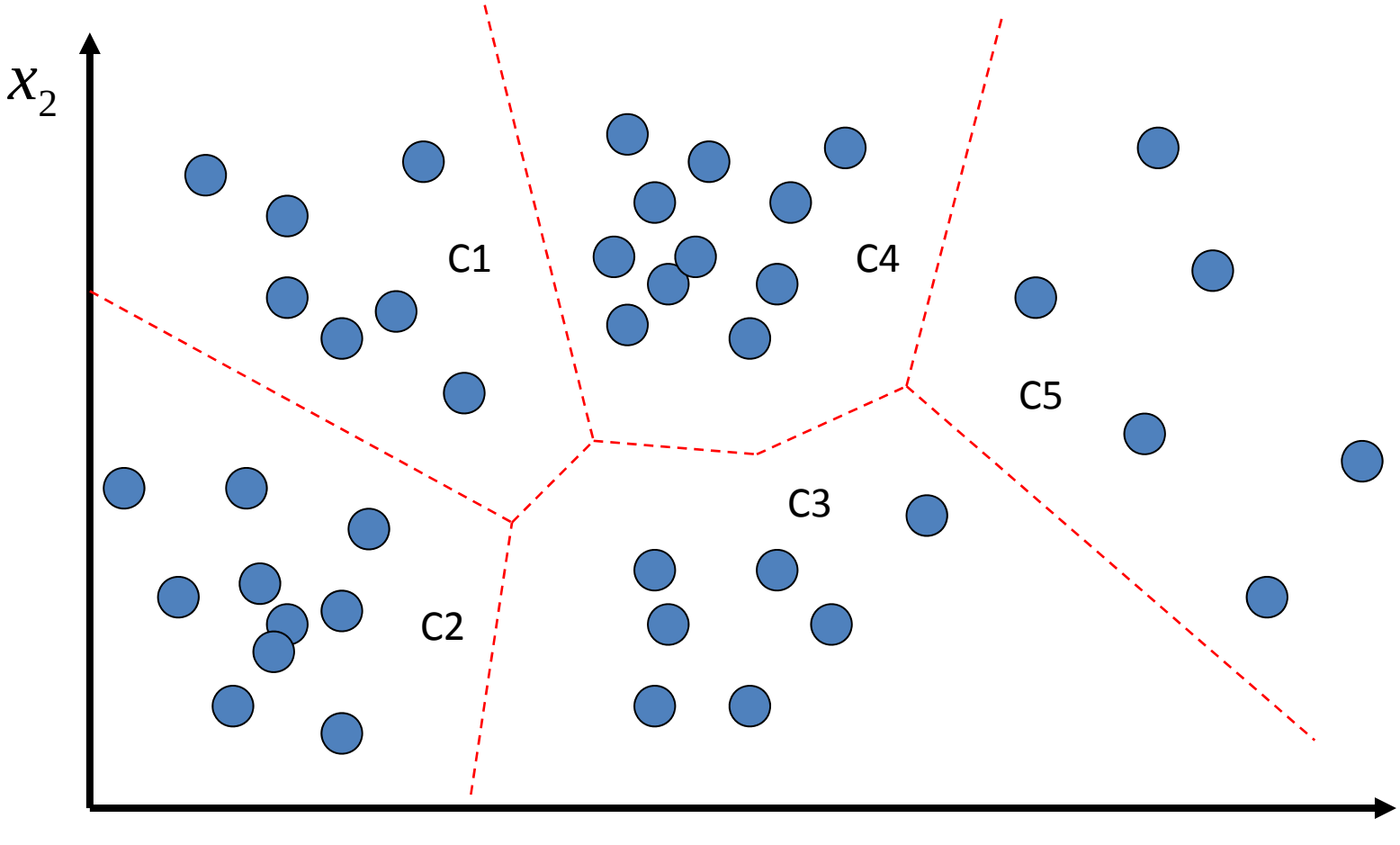
データは12個なので、これらを

ちょうど4個ずつ3等分するようにしきい値$\Theta_1, \Theta_2$ を設定すれば良い。



# 教師なし学習（ベクトル量子化・自己組織化）

**正解ラベルの付いていない観測データを、その分布に従っていくつかのクラスタに分類**
（ただし、分割すべきクラスタの個数は与える）
→ **クラスタリング**（clustering），**自己組織化**（self-organization）
正解ラベルすなわち教師信号がない → **教師なし学習**（unsupervised learning）
連続な信号空間の離散化 → **ベクトル量子化**（vector quantization）



# クラスタリングの目的：　何のためにデータを分割するのか？

【情報理論】 符号化・量子化　＝　**平均情報量の最大化**

【**関数近似**】 データ分布に重みづけされた近似誤差の最小化
ボロノイ分割とも関連

【データからの**知識抽出**】 決定木と関連

# クラスタリングの目的：　何のためにデータを分割するのか？

1）膨大なデータを、なるべくそのデータの持つ「情報」を
損なわないように圧縮・単純化する

2）（連続値で構成された）データを離散化する

【情報理論】 符号化・量子化　＝　平均情報量の最大化

【関数近似】 データ分布に重みづけされた近似誤差の最小化
ボロノイ分割とも関連

【データからの知識抽出】 決定木と関連

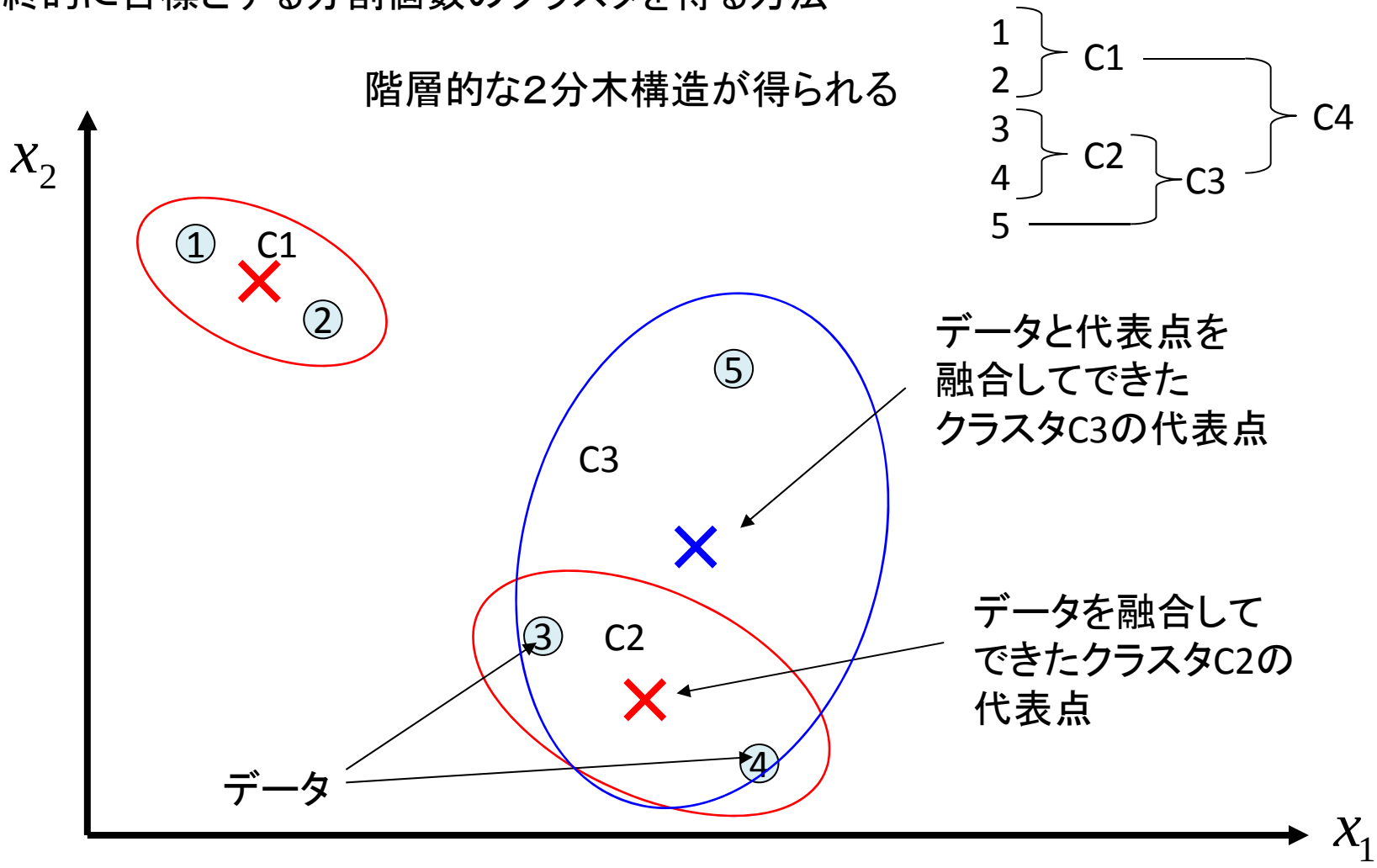
# クラスタリングの代表的手法：ボトムアップクラスタリング法

（ヒューリスティクス：最適解の保障なし）

「距離尺度」が重要！

Nコのデータ点をすべて別のクラスとみなす状態から出発して、
あらかじめ定義されたクラス間の**距離**に従って**最も近いデータ点同士を融合**してゆき、
最終的に目標とする分割個数のクラスタを得る方法

階層的な2分木構造が得られる



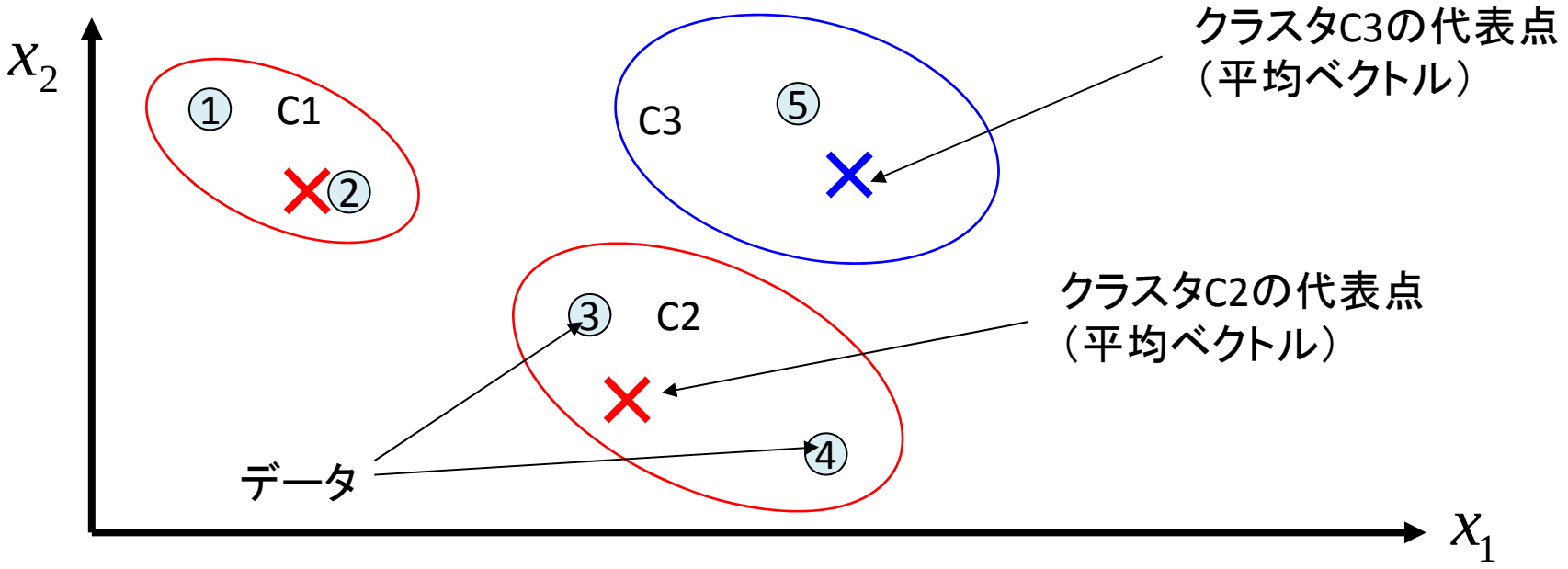

データと代表点を
融合してできた
クラスタC3の代表点

データを融合して
できたクラスタC2の
代表点

# クラスタリングの代表的手法：K一平均法（k-means clustering method）

（ヒューリスティクス：最適解の保障なし）
クラスの数Kを既知として、以下の繰り返し計算により各クラスの平均ベクトルを求める：

1） Kコの平均ベクトルの初期値を適当に選ぶ （学習データからランダムにKコ選ぶ）。
2） 現在の平均ベクトルを用いて、Nコのデータを最も**距離の近い**平均ベクトルのクラスへ分類する。
3） 全データの分類が終わったら、改めて各クラスの平均ベクトルを計算し直す。
4） 全ての平均ベクトルが変化しなくなるまで手順2・3を繰り返す




**クラスタリングの結果は、平均ベクトルの初期値に大きく依存する**

# 【参考】

## クラスタリングとして一般的ではないが、「勢力範囲」の距離を与えてクラスタを生成し、その個数で分布を分析・評価する場合もある

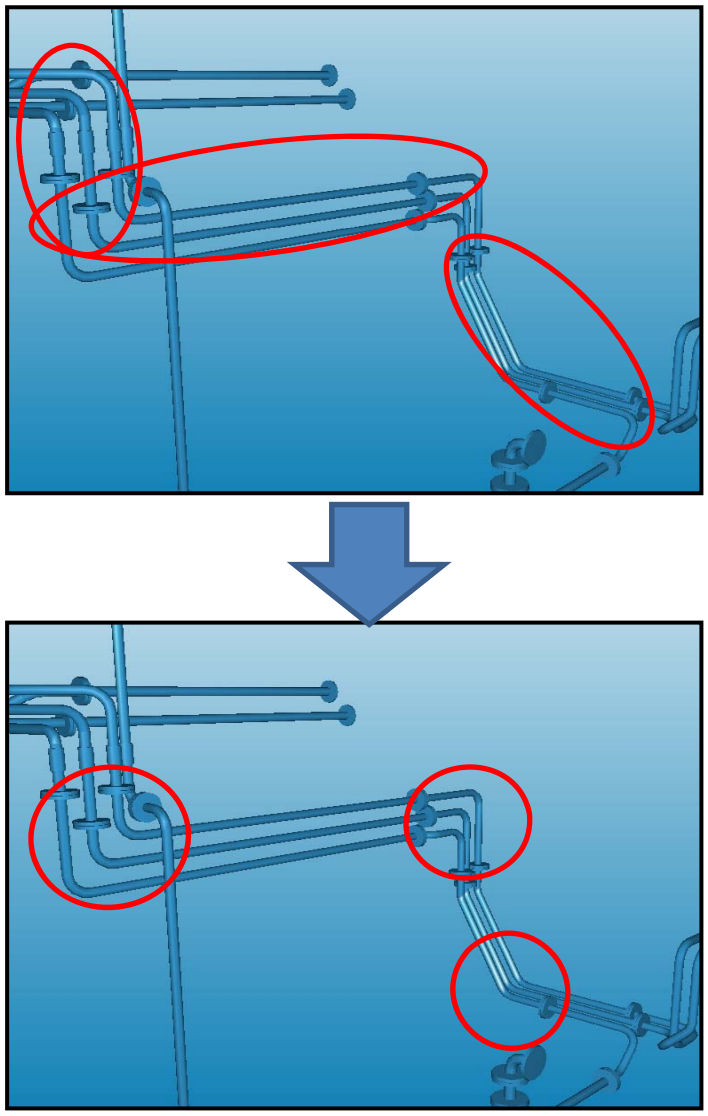
- パイプの「グループ化」の必要性
  - ●メンテナンスのしやすさ
  - ●スペースの確保
  - ●見栄え
- 集合性をどう表現するか？

↓ 着眼点を変える

- ●エルボが密集している。
- ●エルボからパイプの伸びる方向が一致している
- ●壁付近に配置されている。

↓

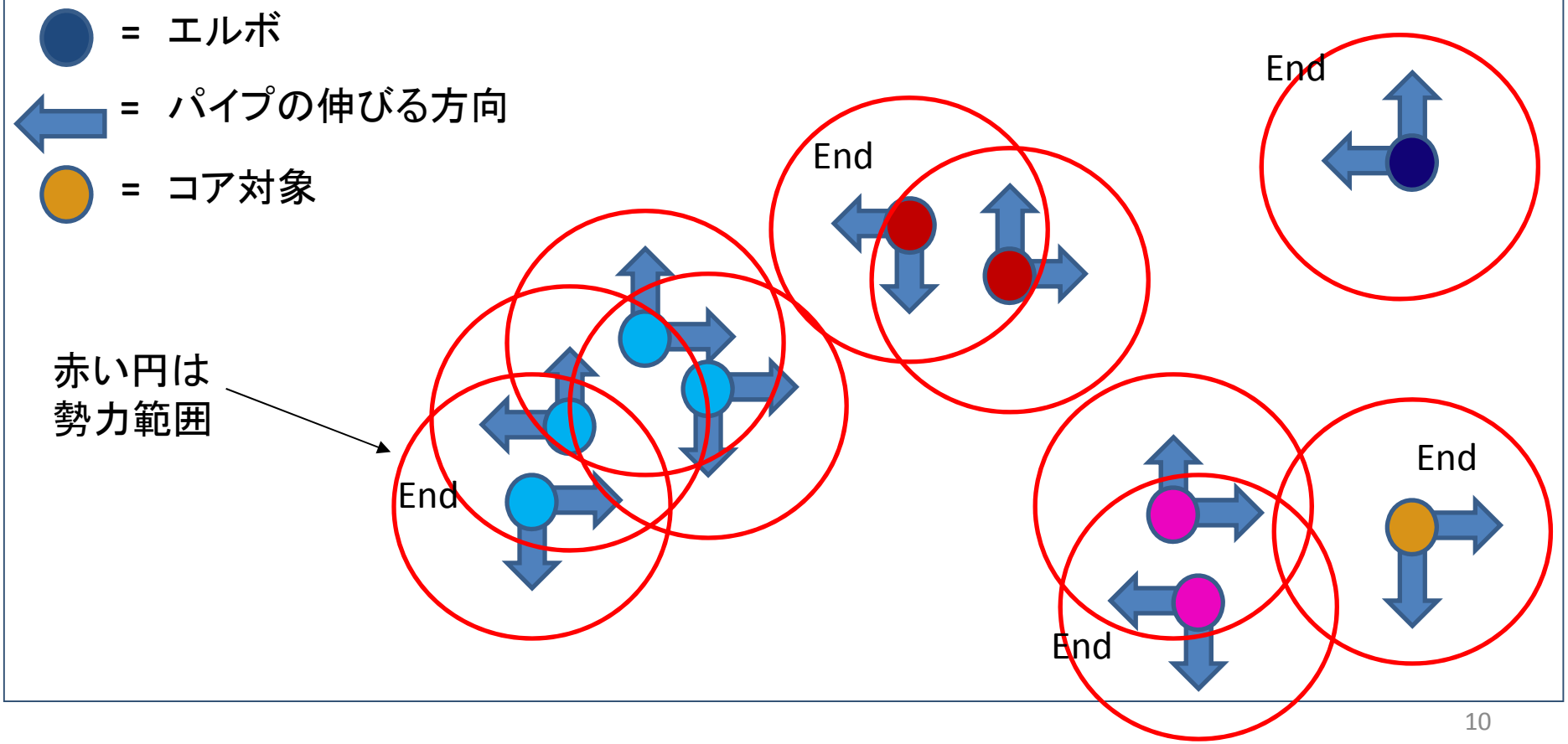
**エルボの分布のクラスタリング**により
パイプの「グループ化」度合いを数値化



# エルボのクラスタリングアルゴリズム

- パイプのグループ性（集合性）を数値化するアルゴリズム

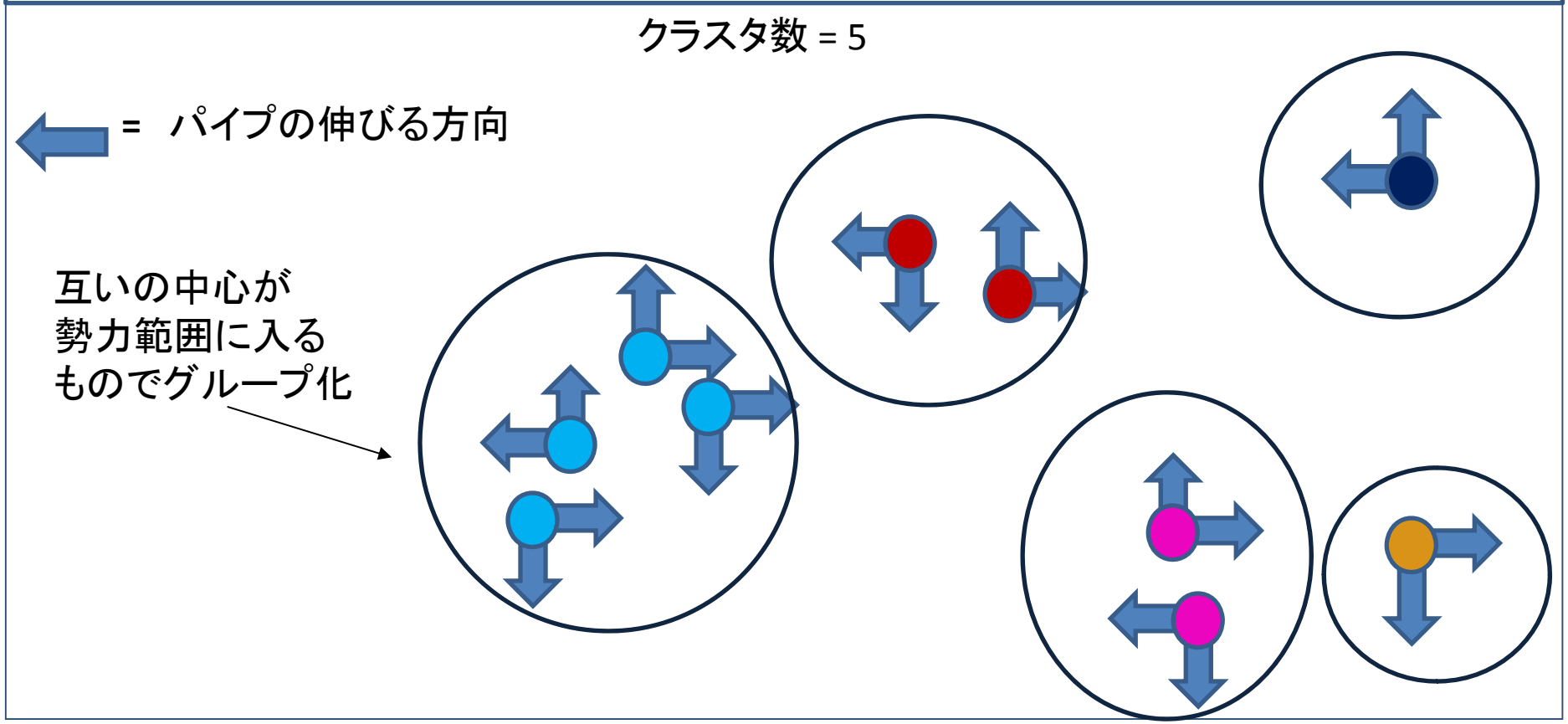
5. 1～4を繰りかえす。全てのエルボがコア対象となれば終了。



# エルボのクラスタリングアルゴリズム

- パイプのグループ性（集合性）を数値化するアルゴリズム

6. クラスタの個数を計算する

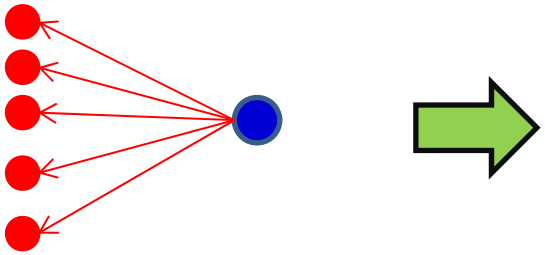



操作性の良いバルブ配置の評価アルゴリズムにも使用できる

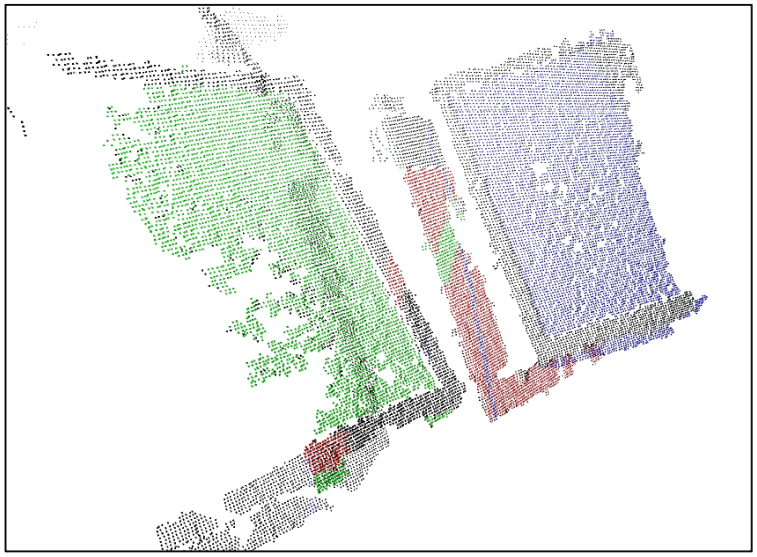
# 【参考2】 計測機器Kinect


## 深度計測機器Kinect

KinectはMicrosoftが発売したXbox360の付属機器であり、縦480×横640pixelそれぞれの深度データを取得することが出来る。



対象との距離をスキャン



深度マップ生成

# Kinectの深度データからの3DCADデータ生成

## (1)クラスタリング

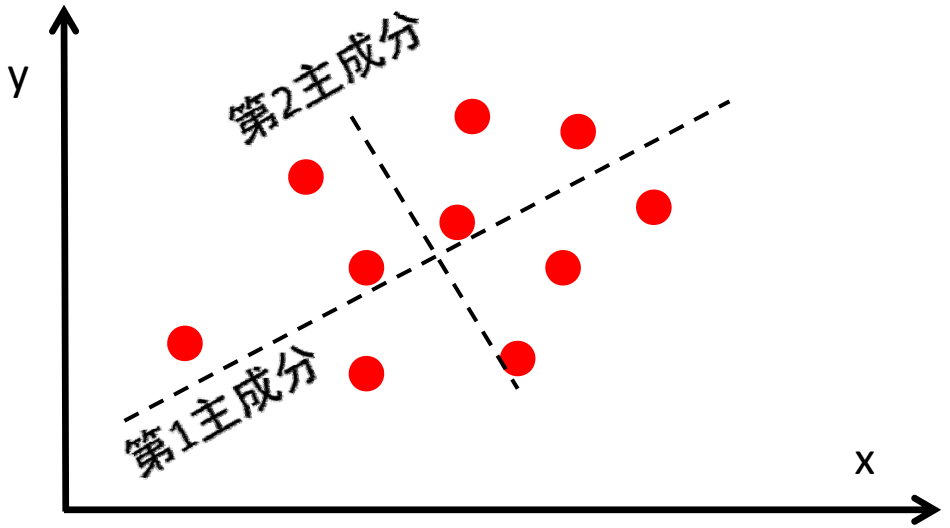
計測データである点集合から平面を検出し、距離を考慮した上で複数のクラスタに振り分ける。

本研究ではクラスタに振り分ける条件の一つ、**平面の判定**には主成分分析を用いる。主成分分析によって得られる三つの固有値に対し、一つの固有値が他の二つの固有値に比べて極端に小さい時、平面であると判定する。

## 【主成分分析とは】

データを無相関な(少数の)特性値に縮約する統計処理

データが3次元空間中の点列の場合、第3主成分の分散の大きさが第1・第2主成分の分散に比べて極端に小さければ点列は平面に分布している

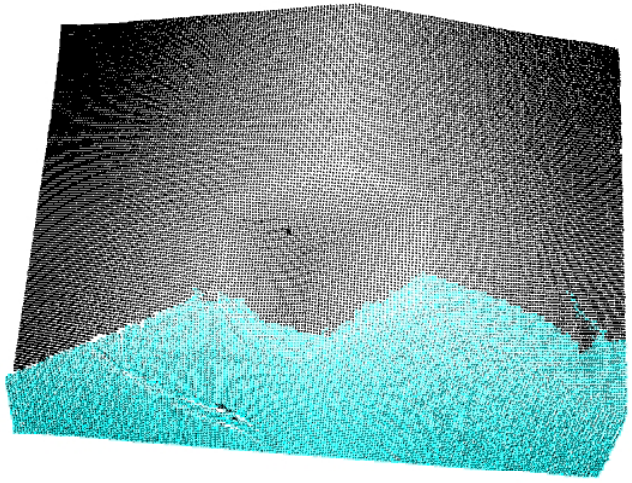



点列が平面かどうかの判定に利用

# Kinectの深度データからの3DCADデータ生成

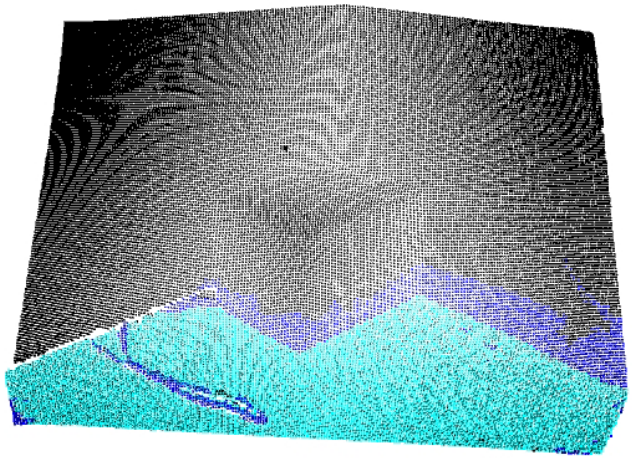
## （1）クラスタリング

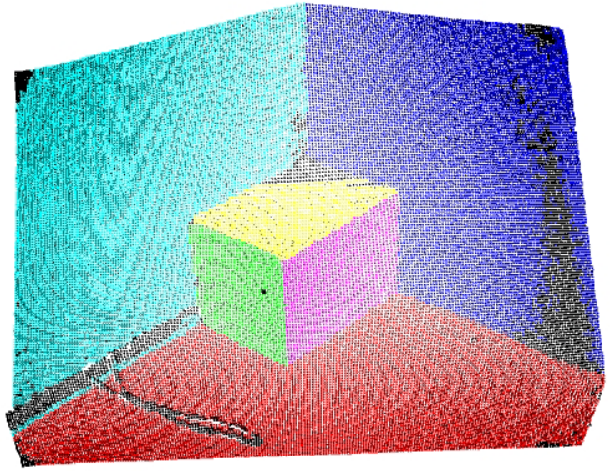
まずデータ点列からランダムに1点を選び、その近傍20点について主成分分析により平面上に分布しているかどうか判定する。平面に分布していたらそれらの点列を1つのクラスタとする。



Kinectの精度を考慮し、主成分分析での判定は比較的ゆるくなっているため、主成分分析のみでクラスタリングを行った場合、図のように平面からやや離れている点も含んでしまう。

# Kinectの深度データからの3DCADデータ生成





そのため、一度クラスタから平面の方程式を導く。この時に最小メジアン法を使用することで平面から離れた点に影響されず平面の方程式を導くことが出来る。この平面の方程式とクラスタの点の距離を計算することにより、一定以上離れている点を除外することが出来る。

**誤差が許容範囲内の点を追加していく＝RANSAC法**

追加出来る点が出来なくなった場合、クラスタに含まれる点の数が一定以上の場合は次のクラスタに移る。点の数が足りない場合はやり直す。

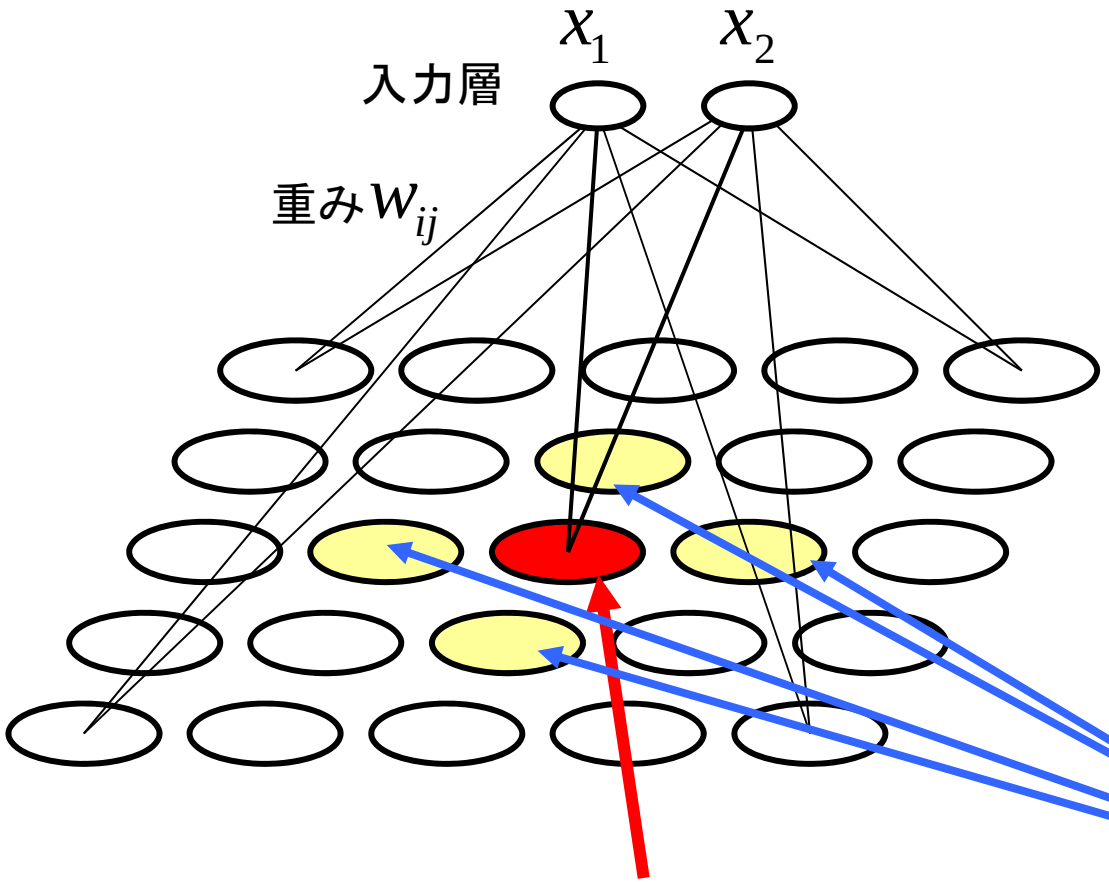
最終的に一定数のクラスタを作成するか、クラスタが作成出来なくなった時点でクラスタリングの処理は終了する。

**問題に応じて様々な「距離尺度」が用いられる！**

詳しい説明は「河内 基樹，木村 元：深度計測デバイスによるデータからの3D 図面自動生成に関する研究，日本船舶海洋工学会講演会論文集 Vol.17, pp.489--492 」
http://sysplan.nams.kyushu-u.ac.jp/gen/papers/papersj.html

# Kohonen（1982）の自己組織化マップ（self-organization map: SOM）

感覚器上で近い部分は、大脳皮質でも近くの部位へ刺激が投射される　→　特徴地図




出力層：重みが入力ベクトルxに
最も近い素子1コだけが反応

→　素子1コが1クラスを分担

$i^*$　：反応した素子

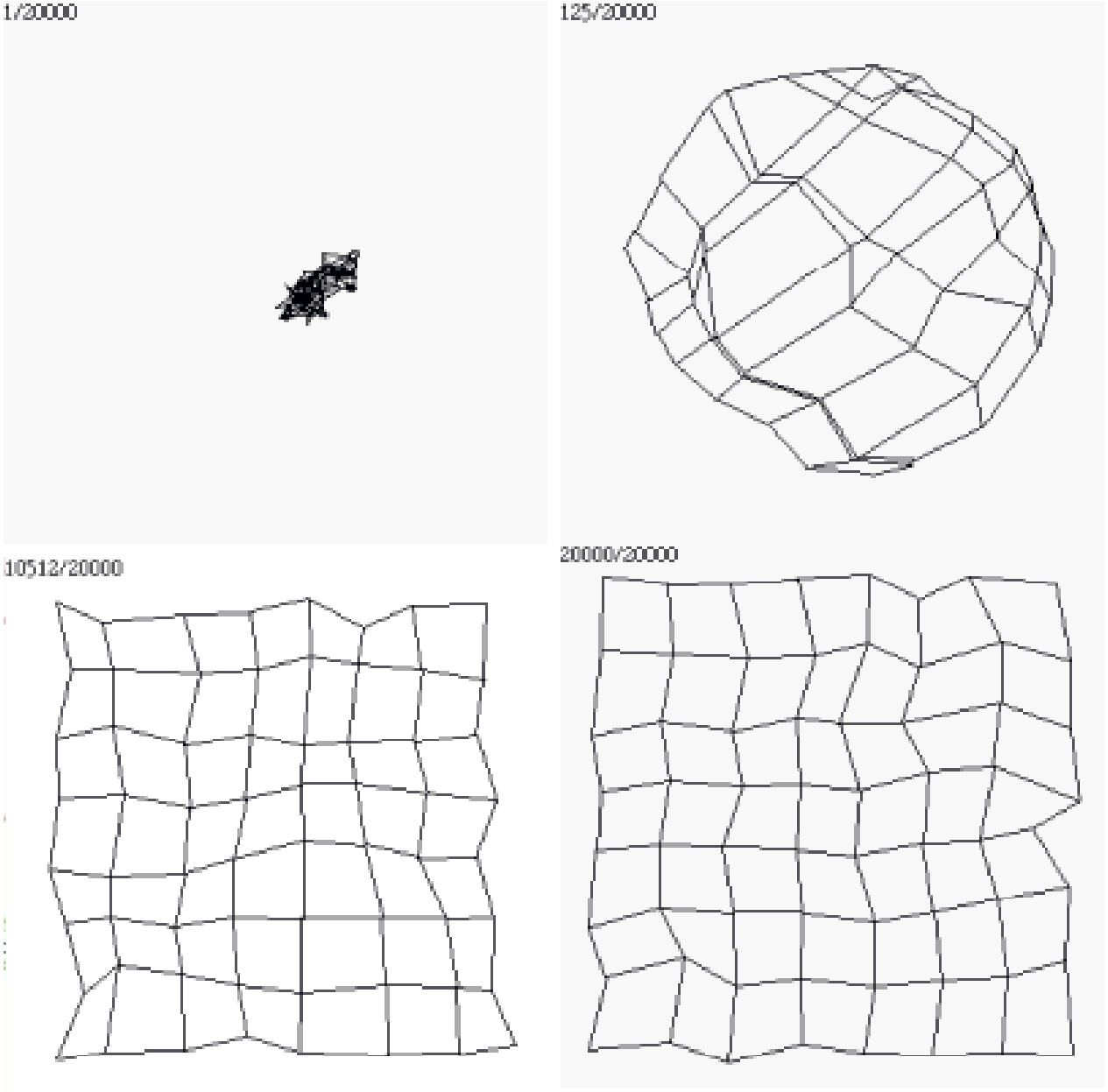
$\Lambda(i,i^*)$：近傍関数

**入力ベクトルxに最も近い重みを持つ素子**、および**その近傍の素子**の
重み変数を、入力ベクトルxに近づける方向へ少しずつ更新していく
（競合学習：勝者総取り）

$$w_{ij} \leftarrow w_{ij} + \alpha\Lambda(i,i^*)(x_i - w_{ij})$$

**データの分布密度を反映したクラスタリング**　→　空間の充填、データ圧縮符号化

# SOMの学習の様子

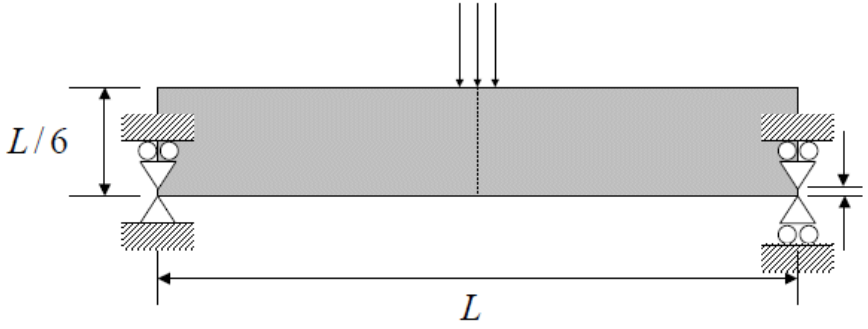


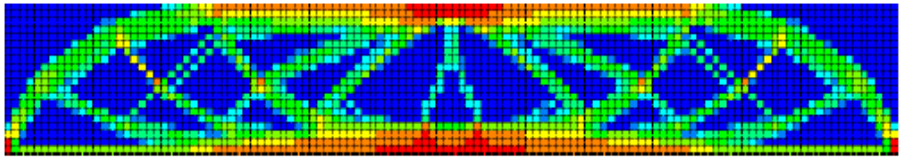
# そもそも自己組織化とは？

『 **要素間の相互作用のみ**によって自発的に進行する組織化 』

他の自己組織化の例： **骨のリモデリング**

骨は、応力が繰り返しかかる部分では
骨細胞が骨を強化していき、応力が
あまりかからない部分では消えていく
→ **トポロジー最適化**



畔上秀幸：形状最適化問題の解法，計算工学Vol.2 No.4(1997)

(a) $\bar{m}_S = 0.3,\ \bar{m} = 0.30,\ \bar{G} = 0.9,\ G = 0.89,\ C/C_0 = 1.49$

(b) $\bar{m}_S = 0.4,\ \bar{m} = 0.40,\ \bar{G} = 0.80,\ G = 0.86,\ C/C_0 = 1.10$

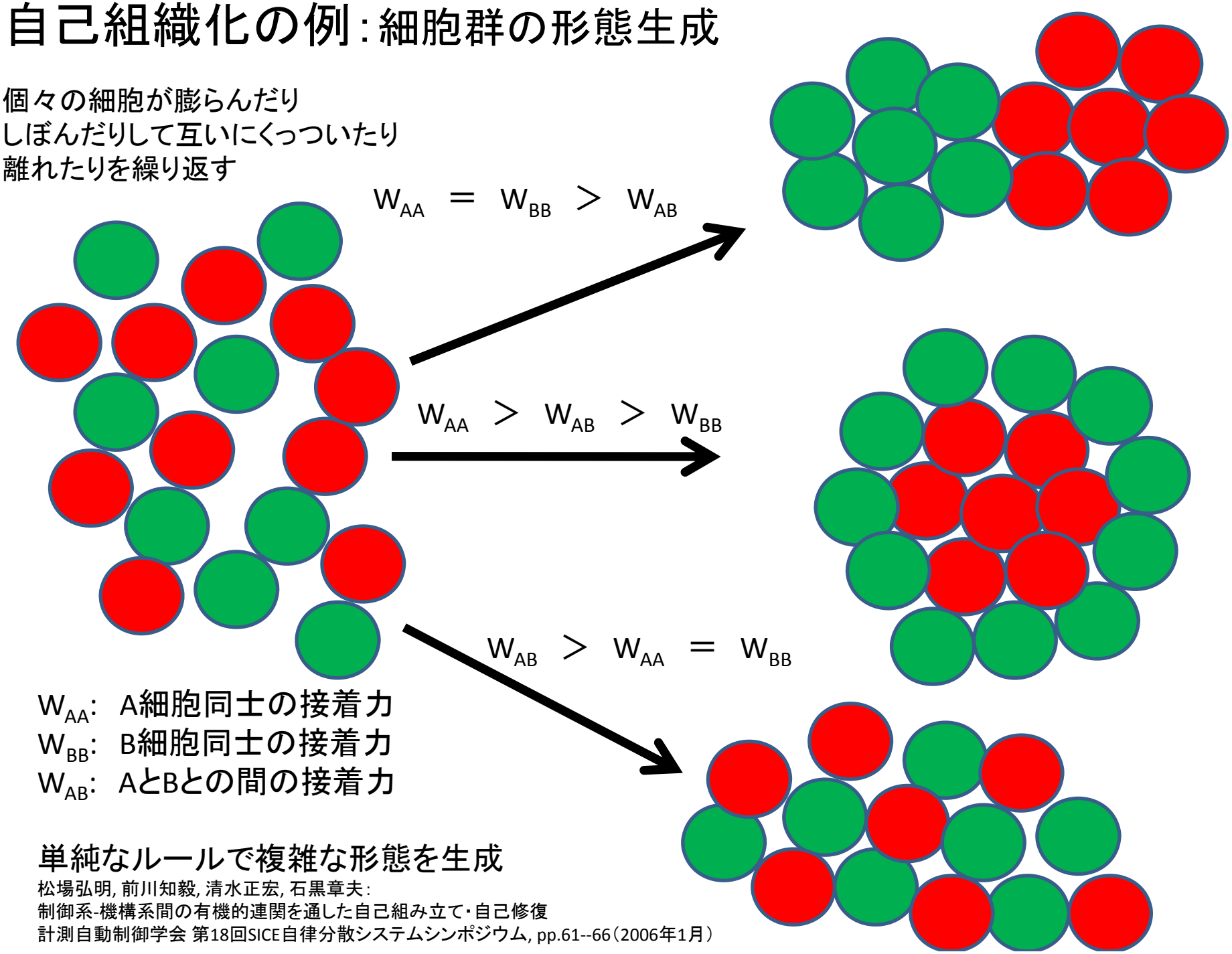
(c) $\bar{m}_S = 0.5,\ \bar{m} = 0.50,\ \bar{G} = 0.75,\ G = 0.84,\ C/C_0 = 1.04$

(d) $\bar{m}_S = 0.6,\ \bar{m} = 0.60,\ \bar{G} = 0.75,\ G = 0.84,\ C/C_0 = 1.02$

東京大学　藤井大地教授のPPT資料より引用

# 自己組織化の例：細胞群の形態生成

個々の細胞が膨らんだり
しぼんだりして互いにくっついたり
離れたりを繰り返す




$W_{AA}$: A細胞同士の接着力
$W_{BB}$: B細胞同士の接着力
$W_{AB}$: AとBとの間の接着力

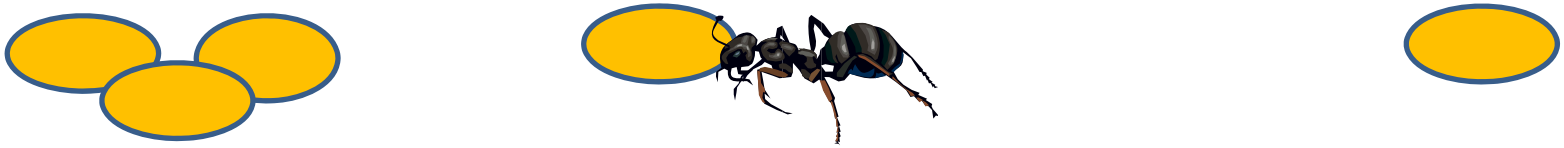
単純なルールで複雑な形態を生成

松場弘明, 前川知毅, 清水正宏, 石黒章夫:
制御系-機構系間の有機的連関を通した自己組み立て・自己修復
計測自動制御学会 第18回SICE自律分散システムシンポジウム, pp.61--66（2006年1月）

# 自己組織化の例： アリの行動

## アリが巣穴の中で物をきれいに（1か所に）格納する仕組み

1） 何も持っていない状態で物にぶつかったら、その物をつかむ
2） 物をつかんだ状態で他の物にぶつかったら、持っている物を放す
3） 上記以外の場合、ランダムに動き回る

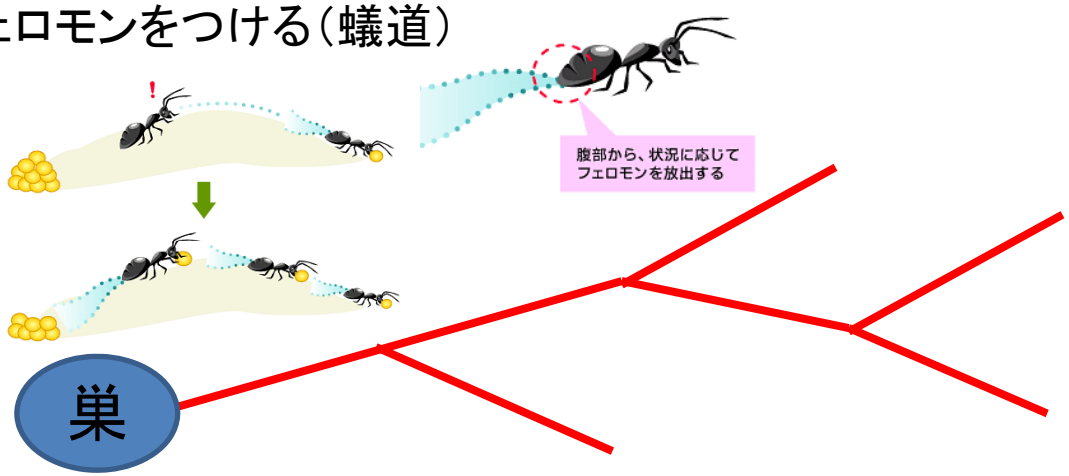


## エサを持ったアリがエサ場から一直線に巣へ戻れる仕組み

1） エサを見つけたアリは地面にフェロモンをつける（蟻道）

2） アリはフェロモンをたどって歩く

3） 分岐点では、
必ず鈍角に曲がっている
ほうへ曲がる

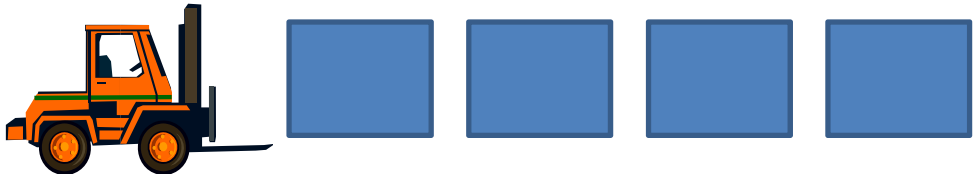



このような単純な制御ルールをいくつか組み合わせるだけで、複雑な動きが可能で、
しかも故障などのトラブルに強い**頑健（ロバストと呼ばれる）なシステム**を構築することが
可能な場合がある。トヨタのカンバン方式はこれに類似。
このような分散システムのパフォーマンスは一般に「最適」では無いが、時に非常に強力である

# 自己組織化を用いたシステム

・倉庫内の物品の配置の最適化
大きな倉庫から必要な物品を短時間で取り出せるようにしたい

→　物品が使用される毎に、その品の置場を一定距離だけ
入口近くへ移動

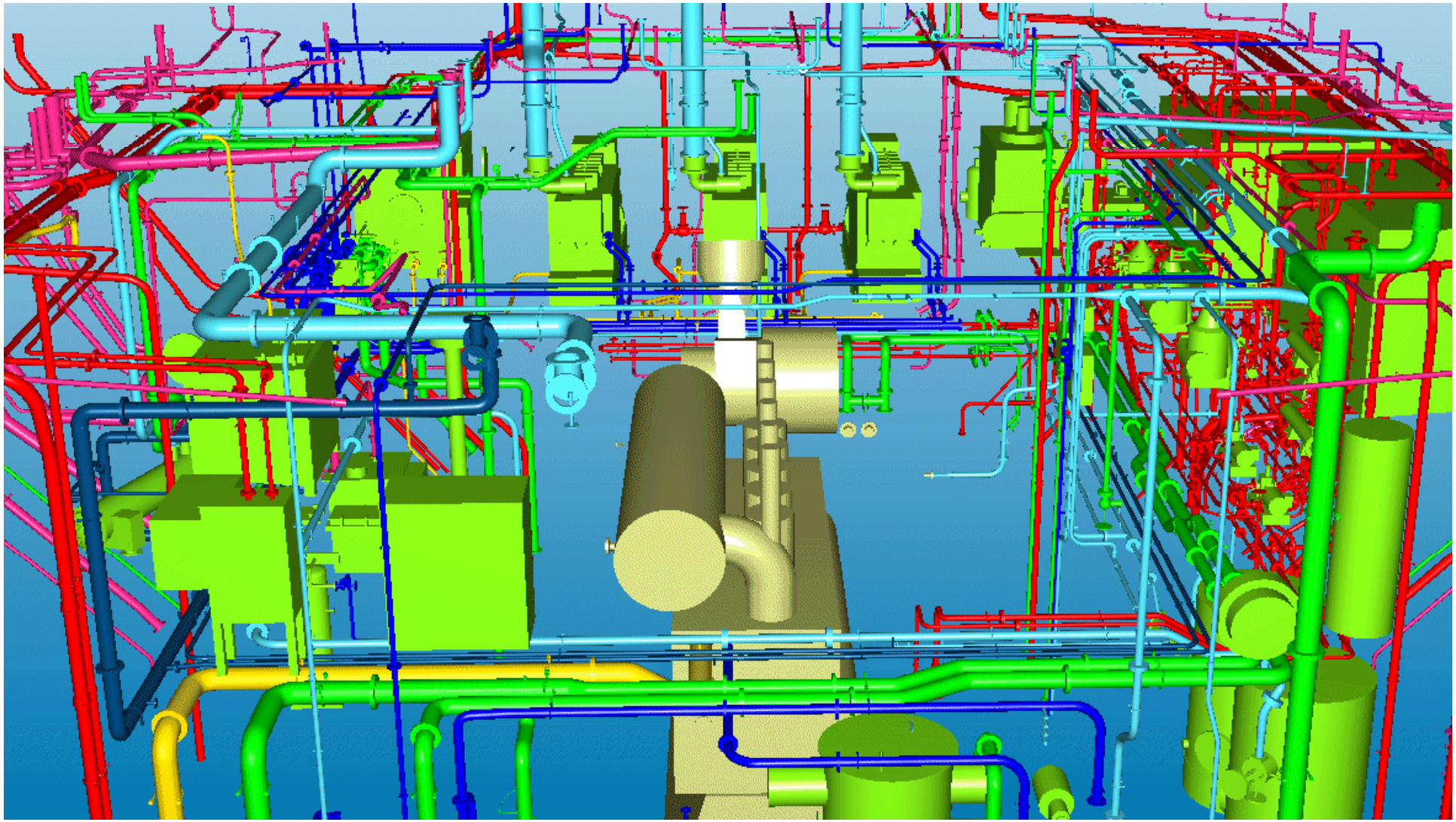


この操作を繰り返すと、使用頻度の高い順に物品がソートされ、
使用頻度の高い物品ほど入口近く（取出し時間が短い場所）へ

これはデータベースにおいて検索時間を減らすようなデータの
格納方法としても使用される

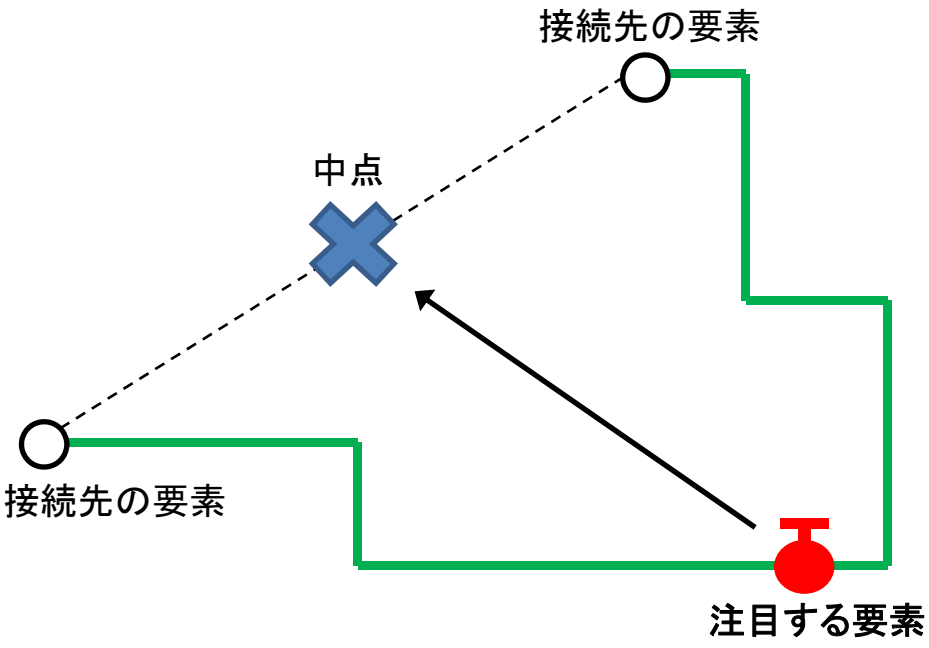
類似例：不要書類の整理法（超整理法）
新しい書類が発生する度にファイルして時間順に並べていき、使用したら、また先頭に戻すという方法。デメリットはめったに使わないファイルはひたすら探すしかない点

機器の配置は、それらを繋ぐパイプのコストを大きく左右

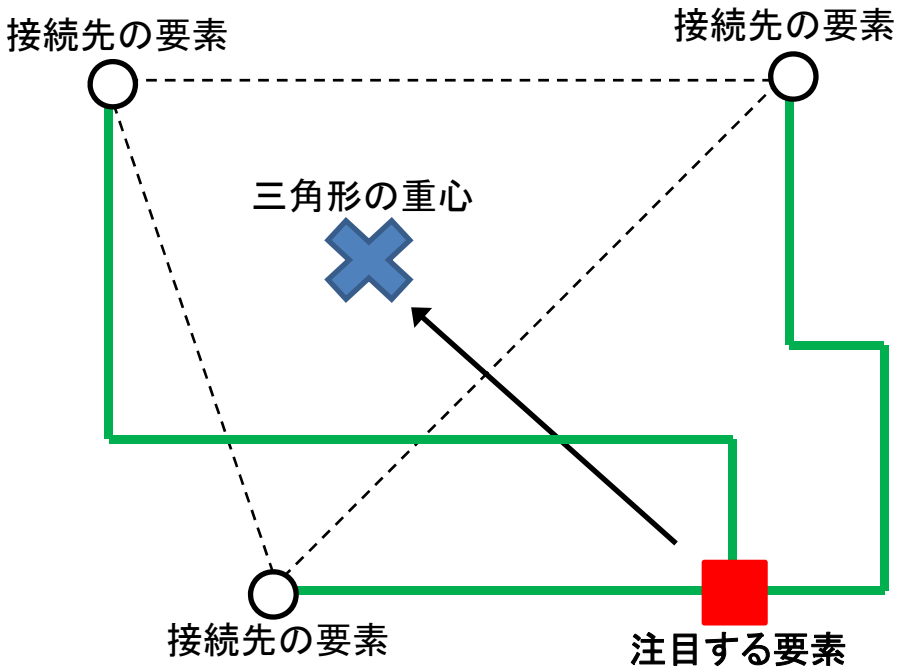


機器配置問題に「**自己組織化**」を利用できるか？

# 自己組織化を利用した解決法




(a) 注目する要素の接続先が2箇所の場合



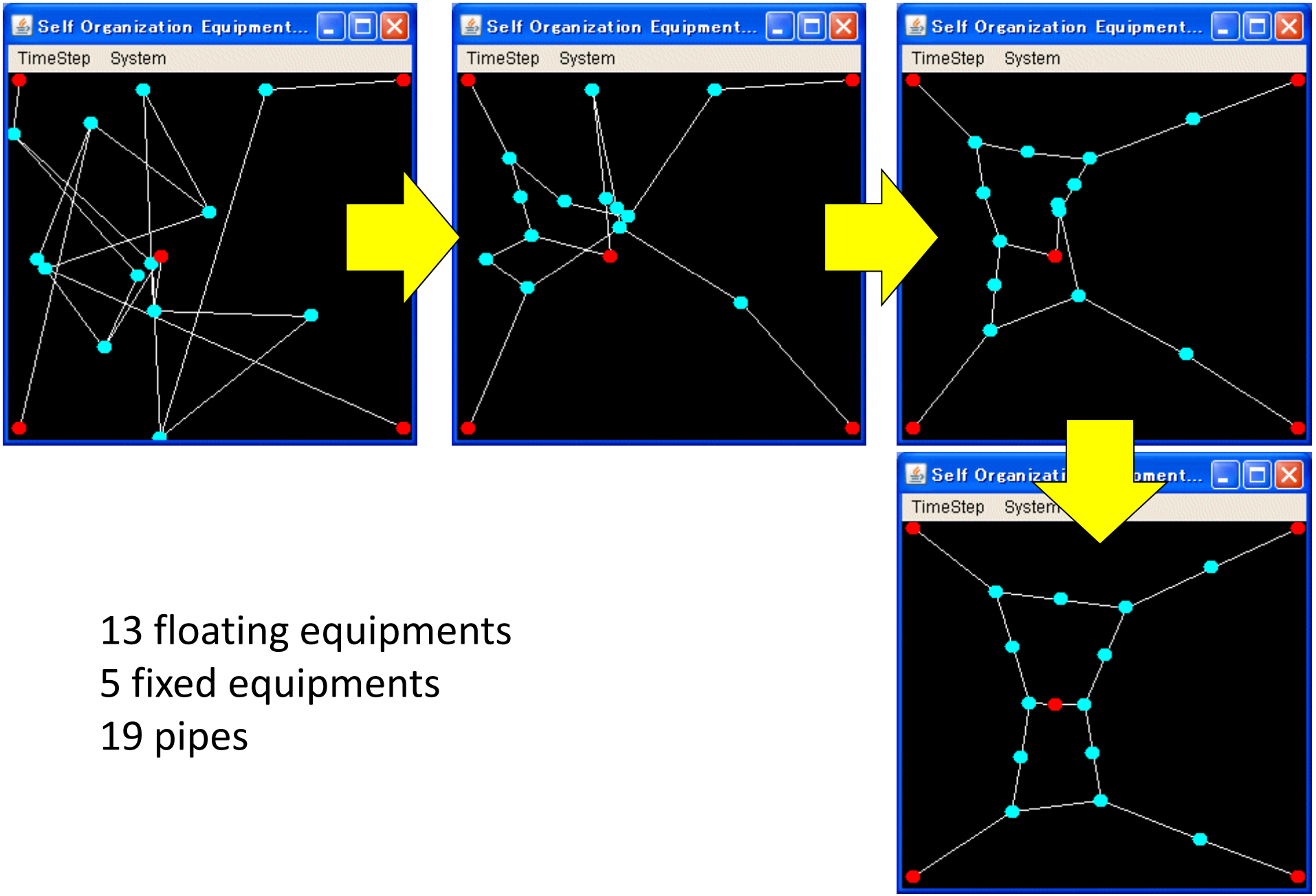

(b) 注目する要素の接続先が3箇所の場合

上記の操作を全ての要素についてランダムな順序で繰り返す

→ こんがらがっていた要素が整然と並ぶ

自己組織化

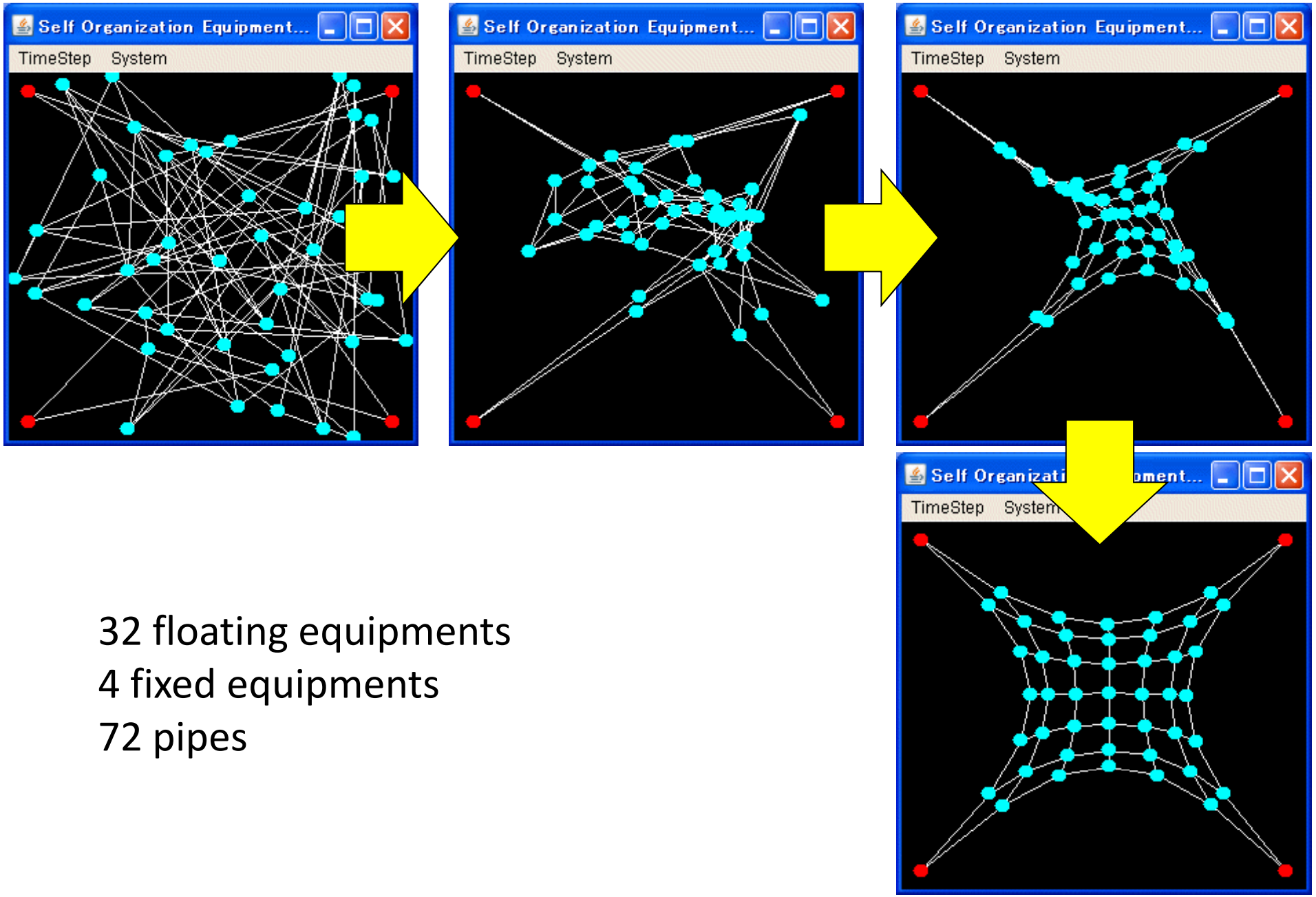
# Demo



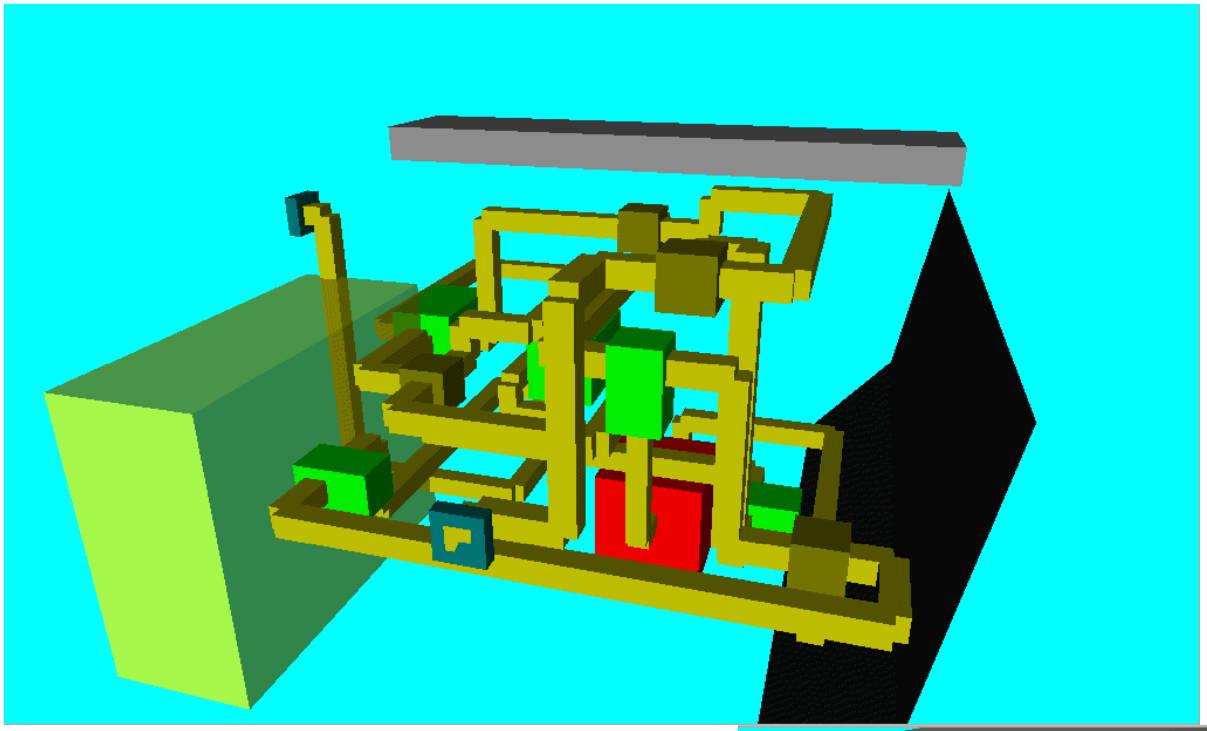

13 floating equipments
5 fixed equipments
19 pipes

自己組織化

# Demo

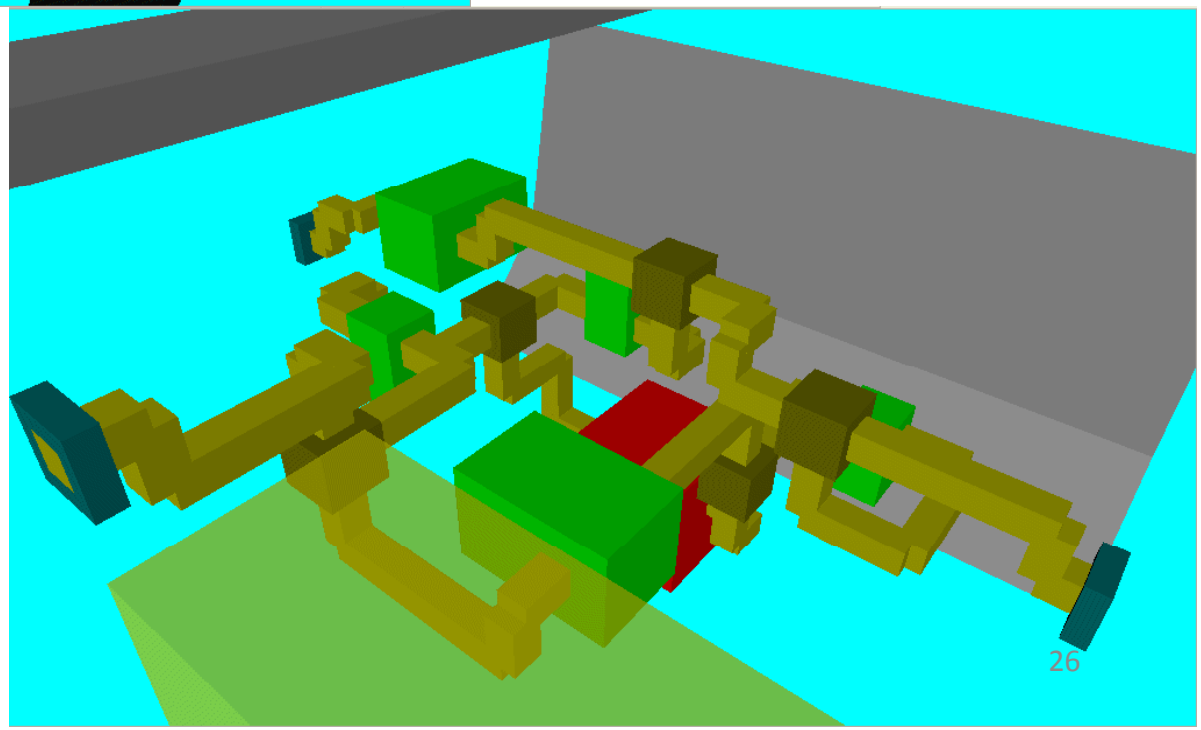



32 floating equipments
4 fixed equipments
72 pipes

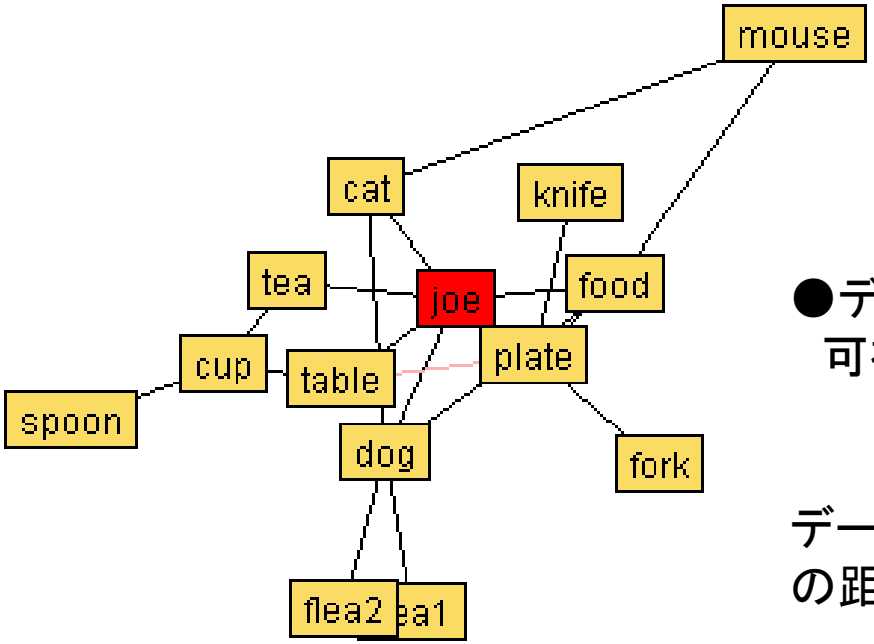


ランダム配置

自己組織化
オペレーション
適用後



# グラフ構造可視化： Eadesのスプリングモデル




●データ間の接続関係を利用してグラフ構造を可視化するグラフ自動描画アルゴリズムの一種

データ同士に接続関係がある場合、2次元平面上の距離を $d_{ij}$ とすると

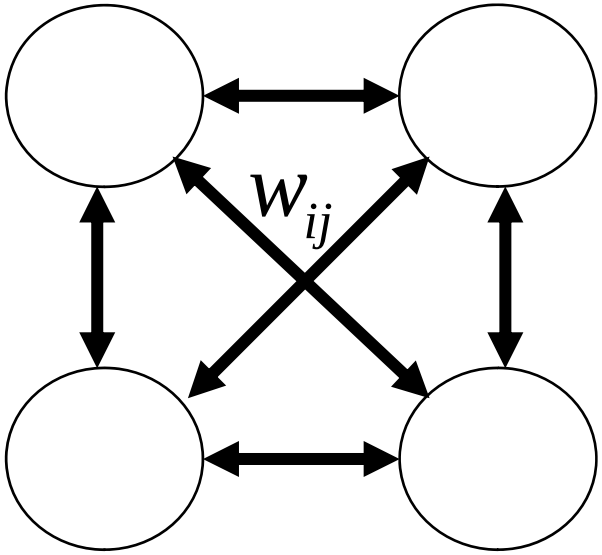
$$fa_{ij} = \alpha \log \frac{d_{ij}}{\bar{d}}$$ の引力で引き合う

またデータ同士の接続に関係なく、データ間は

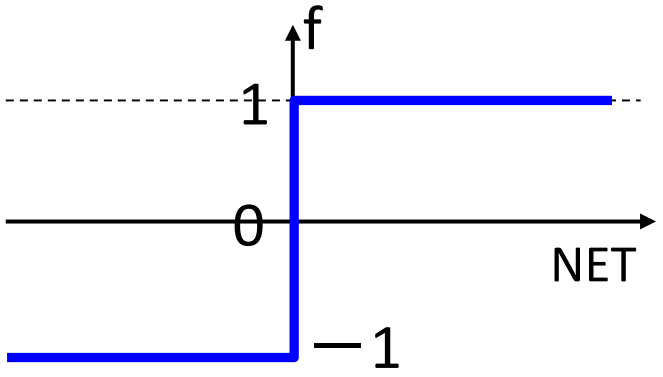
$$fr_{ij} = \beta \frac{1}{d_{ij}^2}$$ の斥力によって反発する。

ただし $\bar{d}$ はデータ間の理想長、$\alpha, \beta$ は定数

# 相互結合型ニューラルネット：ホップフィールドネットワーク




**ホップフィールドネットワーク**
**のニューロン素子：　１ or －１の２値**




●自分自身への結合重みはゼロ　$w_{ii} = 0$
●結合重みは対称　$w_{ij} = w_{ji}$

●ネットワークへの入力：　全てのニューロンユニットに初期値「１」または「－１」を与える
●**ネットワークによるパターンの記憶**：　以下の計算により重み$w_{ij}$ を決める

$$w_{ij} = \begin{cases} \sum_{t=1}^{T} x_i^t x_j^t & i \neq j \\ 0 & i = j \end{cases}$$

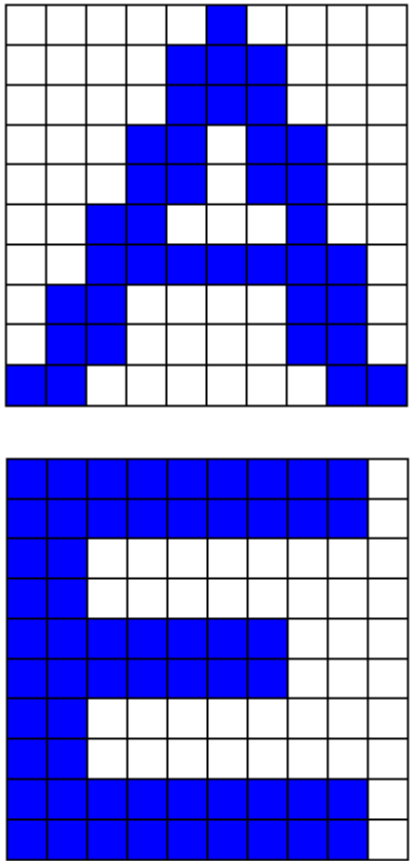
ただし $x_i^t$ は t 番目のパターンの i 番目のユニット

記憶させるパターン数Tは、ユニットの個数の１５%程度

**同じ値のユニット同士を結ぶリンクを強化＝Hebb則**

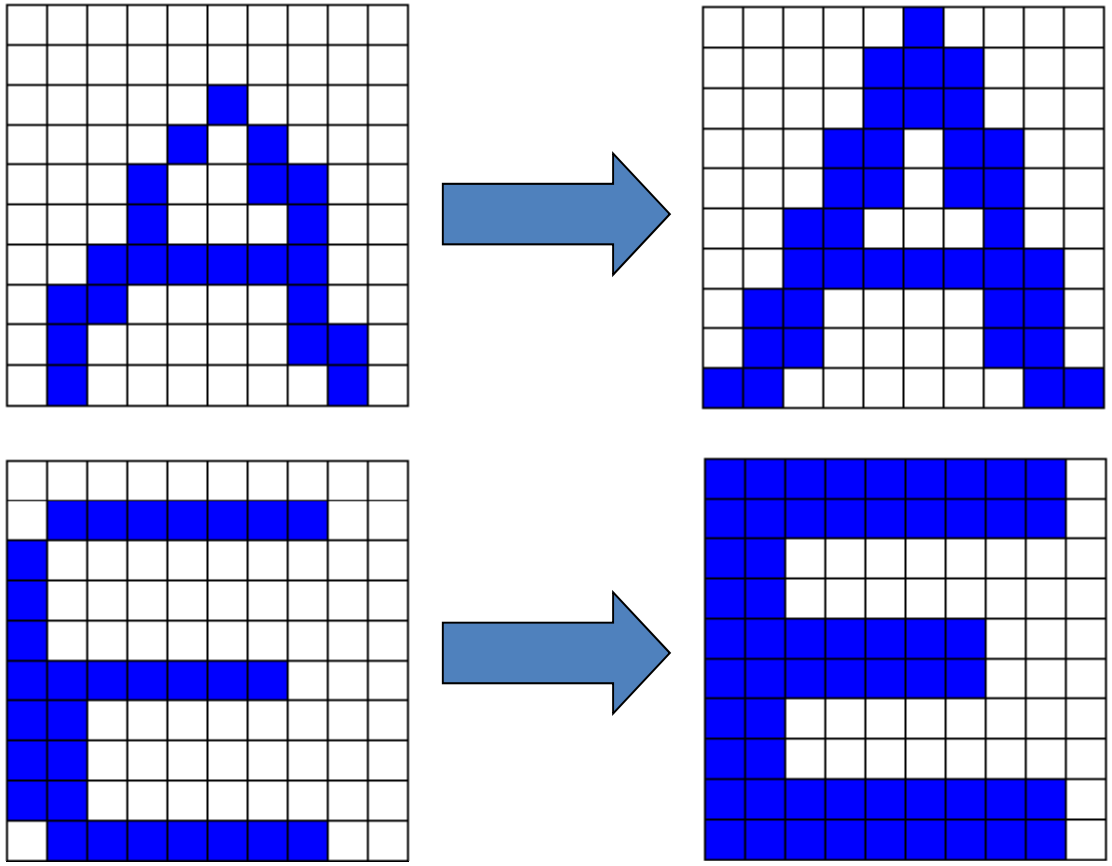
●ネットワークによる計算：　ネットワークに未知パターンを入力し、
ユニットを適当に選んで計算。状態が安定するまで続けると記憶したパターンへ収束

記憶パターン



入力パターン

想起されたパターン

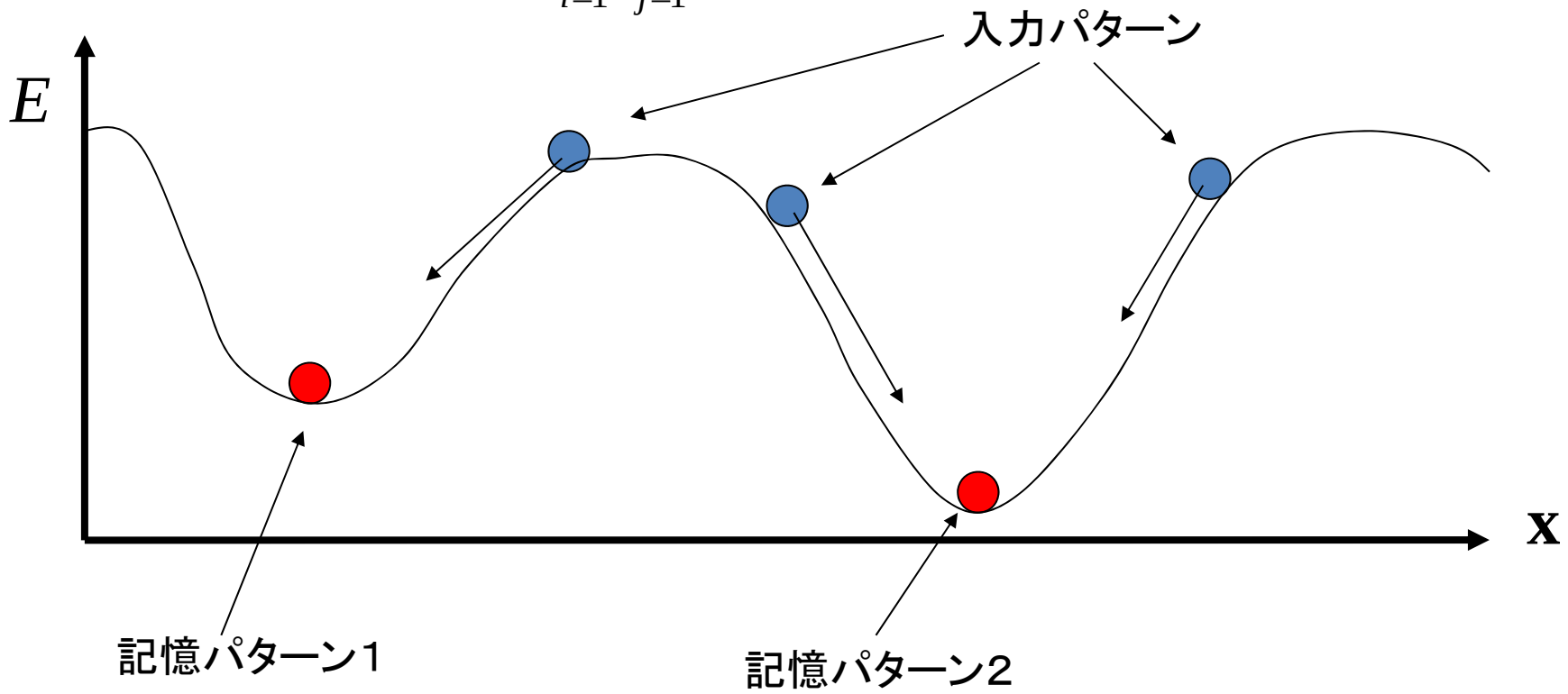


入力パターンが多少歪んでいたりノイズを含んでいても修正

あまり変なパターンを入れると安定しなかったり、反転したり変なパターンへ収束したりする

# ホップフィールドネットワークのエネルギー関数と最適化

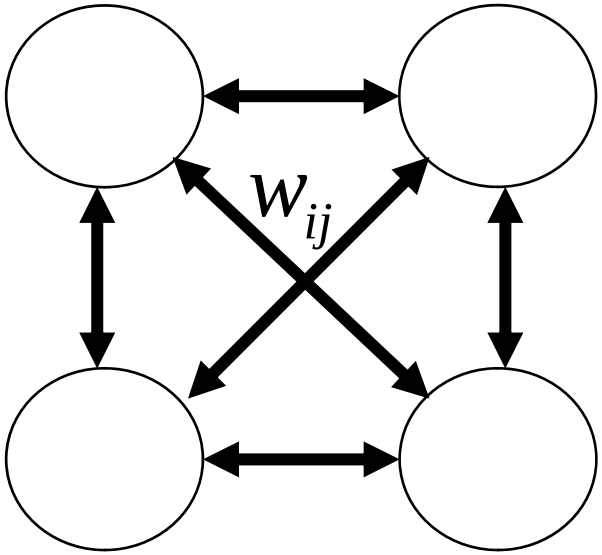
エネルギー関数 $E = -\frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n} w_{ij} x_i x_j$



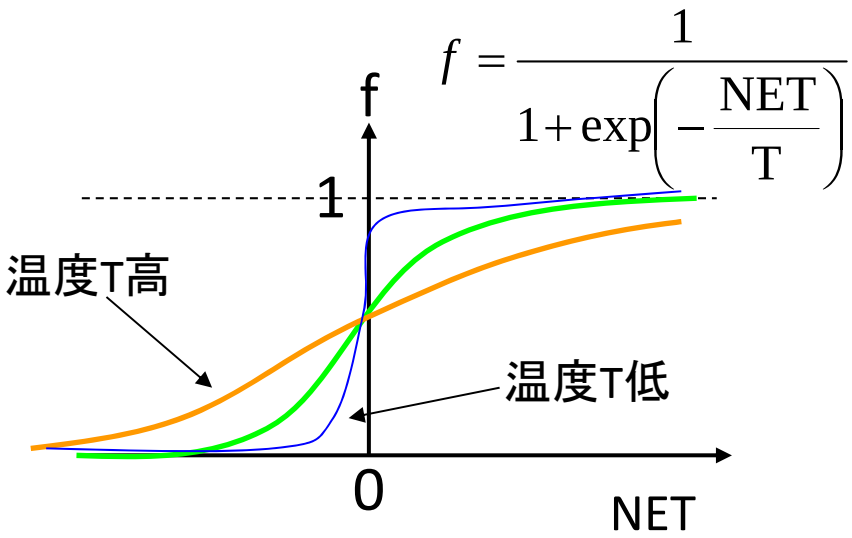

「想起」プロセスによって、エネルギー関数値が
極小値になる方向へユニットの値Xが遷移

$w_{ij} = x_i x_j$ に正比例するとき、Eは常に負の値でエネルギー関数を最小にする

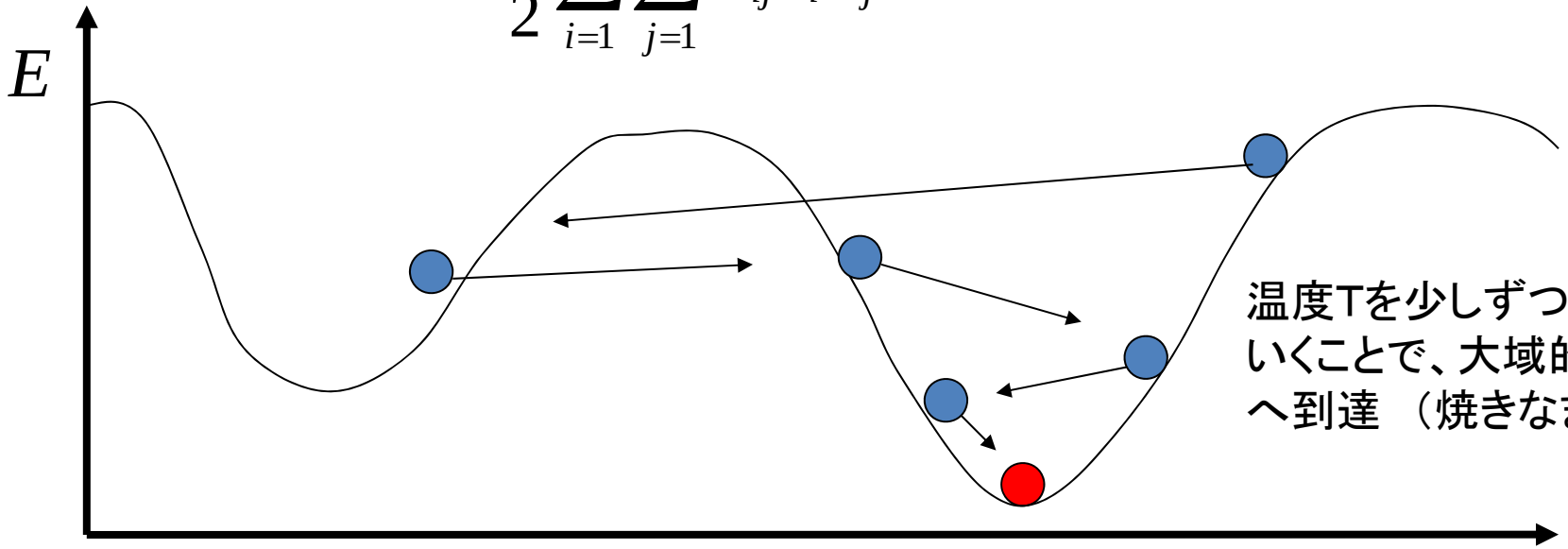
# ボルツマンマシン




**ボルツマンマシンのニューロン素子：**
**シグモイド関数を確率分布とした**
**0または1の離散的な値**




エネルギー関数 $E = -\frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n} w_{ij} x_i x_j$




温度Tを少しずつ下げて
いくことで、大域的な最小値
へ到達 （焼きなまし法）

# まとめ

（1） クラスタリング＝データ圧縮：平均情報量をなるべく大きく

（2） クラスタリング法：

「距離尺度」が重要！

- ボトムアップクラスタリング法
- k平均法
- Kohonenの自己組織化マップ

最適化が保障されないヒューリスティクスだが、
少ない計算量で準最適解を得る

（3） 自己組織化とは？
骨のリモデリング・アリ・配置法

要素間の単純な相互作用で
自発的に進行する組織化

頑健で、場合によっては
非常に効果的

（4） 相互結合型ニューラルネット
パターンの想起＝エネルギー関数の最小化

---

【参考文献】
・麻生英樹、津田宏治、村田昇： パターン認識と学習の統計学、岩波書店（2003年）
・Stuart Russell and Peter Norvig: Artificial Intelligence – A Modern Approach,
Pearson Education Inc. (2003)