理解発見データマイニング
－ AI はなんでもしてくれるわけじゃない－

宇野 毅明　（国立情報学研究所
＆総合研究大学院大学）

http://research.nii.ac.jp/~uno/index-j.html
e-mail: uno@nii.ac.jp

2018年8月24日 市民講座

データマイニングとは

データマイニングとは、データの中から、

＋ おもしろいもの
＋ 特徴的なもの
＋ 共通性の高いもの
＋ めずらしいもの

こういったものを見つける、計算技術なのですが、

説明の前に、ちょっと **AI** の話をします

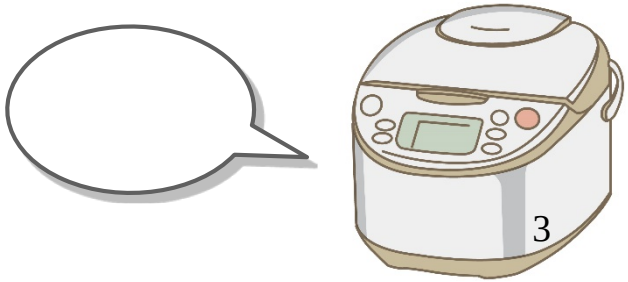
「暑いですね」

- 暑い日に、**「暑いですね」**と言ってもらえたら少しうれしいので、そう言ってくれるAIを作ろう、と思ったとします

  きっと、**家電製品**とかに取り付けるんです

- さて、どういう仕組みにしましょうか

  人工知能っぽいですよね



気温で判定

- 暑い日は、気温が高いので、ある温度を超えたら**「暑いですね」**と言いましょう

30度超えたらでいいですかね
でも、どのくらいが暑いかって、人によって違いますよね

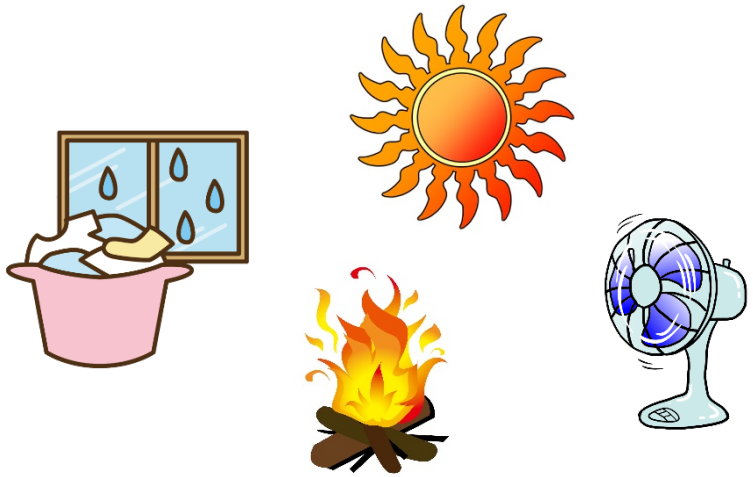
- じゃあ、**「汗かいてる人」**がいたら、にしましょうか
  お化粧してると、だめですね…

これ、**「AI」**でしょうか？

様々な要因

- 人が「暑い」と思う要因は、いろいろあります

温度、湿度、
風、直射日光、
熱源、、、



いろんな条件を考え、判定条件を作れば良さそうですね

複雑な条件だったら、「AI」と言っていいでしょうか

その 人 の観察も

- 同じ気温、湿度でも、人によって状況が違います

＋ 運動しているか
＋ 体が冷えているか
＋ 風があたる位置にいるか
＋ 涼しい服装をしているか

人の観察をすると、もっと良くなるようですね

機械的じゃなく、これくらい人を見れば「AI」でしょうか

暑い、って、どういうこと？

- そもそも、暑い、って、言葉で説明できるでしょうか？

  + 気温が高い　← 何度以上？
  + むしむしする　← 湿度？どれくらい？
  + 毎日暑い　←　いつから？ 途中に寒い日あっていい？
  + 汗をかいている　← カレー食べたかも

実は、そもそも、「言葉で言い表す」ことさえ難しいんです
　なら、（数字で）条件づけるのは、もっと難しい

犬と猫の違いを言葉で言えますか？

データ型 の AI

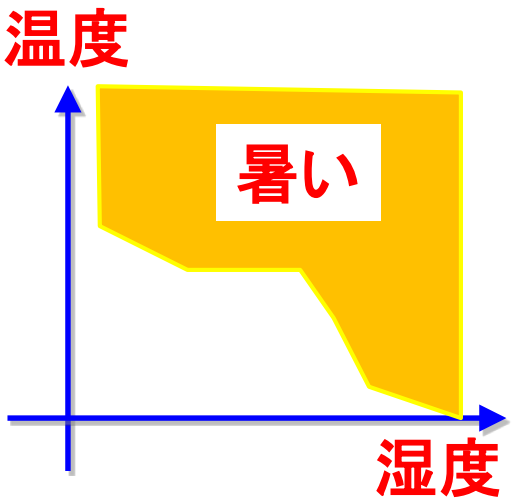
- 気温、湿度、風、日光、こういうものなら数字で扱える
  しかし、これらの数字がそれぞれどうなると「暑い」と感じるかは、たぶん複雑

  ならば**データで調べよう**

- 気温、湿度、風、日光、これらの条件いろいろ変わる
  それぞれで、暑いかどうか、みんなに答えてもらえば、
  「この辺りが暑いかそうでないかの境目だろう」
  という**線引き**がわかる

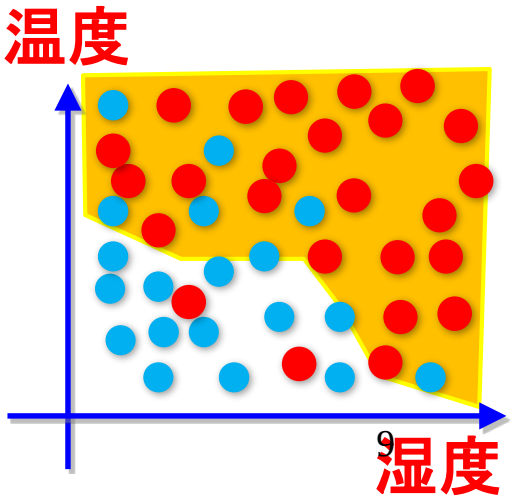
複雑な条件を作る

- 温度だけ、湿度だけでは説明できないのだから
  多数の数値がからんでいるだろう



他が同じ条件で、温度だけが上がったときに、寒く感じるってことはないだろうから、一応きれいに分かれるだろう

人によって、暑いか寒いか、感じ方が違うけど、多数の人からデータをとれば「普通はこれくらい」がわかるだろう



最近の AI の基本

- 温度や湿度などの数値を、単純に使うだけでは難しいことを、データをたくさん集めてできるようにする、のが最近の AI のブームの意味

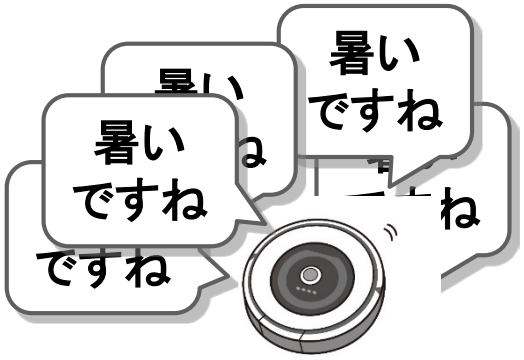
- 自動翻訳も、手書き文字認識も、音声認識も、将棋も、顔認証も、基本は同じ

- IT技術の発達で、データがたくさん、簡単に安く手に入るようになったのが、そもそも

「できた」として

- さて、適切なときに「暑い」と言えたとしましょう
  どうなるでしょうか

- 温度や湿度がある点を超えると「暑い」と言います
  ずっとしゃべり続けます



  ← 何分したらもう一回言う？
  ← 暑くなくなり、もう一度暑くなったら？

- いろんな家電がしゃべります
  うるさいです

では、何をしてほしい？

- 気分や、体調、周りの人などの状況によって、**「暑いですね」**と言ってもらってうれしいでしょうか

これも、やっぱり**「言葉で説明できない」**ものですね

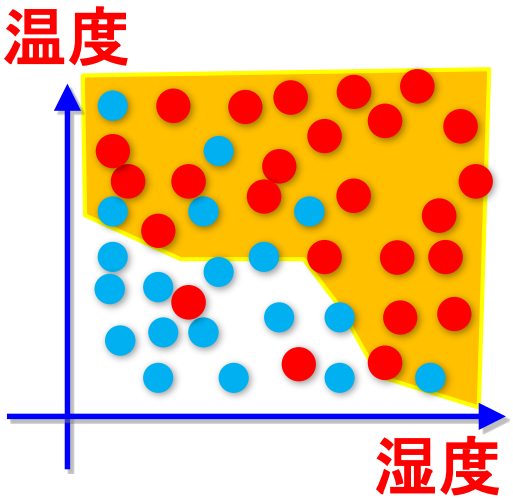
- でも、**「暑いかどうか」**とは違って、気温や湿度などの数値から判断できるものと違い、**「状況」**は、時と場合でみなければいけないものが違います

さらに、それを**「理解する」**のも難しい。データから導くのは難しいです

データは意味を教えない

- 判断基準が数値（画像や音声も）であるなら、データを集めれば何かわかる

　判断基準が状況や感情などは、そもそも（客観的なデータの）何を基準に判断すればいいか、わからない



無理に「見えるデータ」を使うと、推薦・広告のようになる

意味を知りたいと思ったら

- では、状況把握や意味の理解、主観的な意思、そういうものがないと困るものは、どうしましょうか

やっぱり、**「人間が考えるしかない」**

＋ 新人の採用、結婚相手
＋ 引っ越し先、趣味、職業選択
＋ 新商品開発、科学・技術
＋ 経営戦略、経営判断
＋ 裁判、調停、監査

…

人間が考えるなら

やっぱり、**「人間が考えるしかない」**

そうなったら、コンピュータのできることは、

「人間の思考を助ける」 こと

この方向で、特徴的なものを見つける、となると

はっきりわからないものを見つける ので、

「それっぽいものを全部見つけ出す」となります

「あなたにとってこれが面白いですよ」

「このデータの特徴がこれです」

など、自動的に判断して見つけてくれるのではないです

データマイニングの理念

（意味理解を伴わないで） データの特徴を見つける

どんなものを見つけるか

＋ 多くのデータ、多くの場合に共通するもの
　　たくさん現れるもの、）

＋ （データの中で）まとまっているもの
　　似ているものが多いもの、集中して現れるもの

数字で表現できるものに取り組みます
　そうでないと、計算できないからです

データマイニングの技術
1．パターンマイニング

たくさん現れるものを見つける

- データの中に、何か同じものがたくさん現れてたら、きっとそれはデータの特徴だろう

（こういうのを **パターン** とよびます）

- たくさん現れるパターンは面白い(?)ので、それらを全部見つけよう

（という問題が **パターンマイニング** です）

買い物データで

- 右の表、1行が、
一人の買い物かごの中身
と思ってください

- さて、どういうものが
たくさん現れているでしょう？

品目？ 数量？

簡単に調べられるものは、
調べればいいので、
「品物の組合せ」に注目

お茶、おにぎり
お茶、弁当
弁当
ポテチ、雑誌、せんべい
ポテチ、せんべい、ジュース
お茶、せんべい
お茶、おにぎり
お茶、おにぎり、雑誌
弁当
お茶、雑誌、ポテチ、せんべい
ポテチ、せんべい
弁当、ポテチ、せんべい
弁当、雑誌、ポテチ、せんべい

たくさん現れる組合せ

- 例えば、

  お茶とおにぎり
  ポテチと雑誌とせんべい

等がたくさん現れる

意味を解釈せずにやると、
どの品目がえらいとかわからないので、一定回数以上現れたら、よしとする

お茶、おにぎり
お茶、弁当
弁当
ポテチ、雑誌、せんべい
ポテチ、せんべい、ジュース
お茶、せんべい
お茶、おにぎり
お茶、おにぎり、雑誌
弁当
お茶、雑誌、ポテチ、せんべい
ポテチ、せんべい
弁当、ポテチ、せんべい
弁当、雑誌、ポテチ、せんべい

仮に2回以上とすると

- 2回以上現れる商品の組合せを全部見つけると

お茶、おにぎり
お茶、せんべい
お茶、弁当
お茶、雑誌
せんべい、ポテチ
せんべい、雑誌
せんべい、弁当
雑誌、ポテチ、せんべい
…

意外とたくさんある

お茶、おにぎり
お茶、弁当
弁当
ポテチ、雑誌、せんべい
ポテチ、せんべい、ジュース
お茶、せんべい
お茶、おにぎり
お茶、おにぎり、雑誌
弁当
お茶、雑誌、ポテチ、せんべい
ポテチ、せんべい
お茶、弁当、ポテチ、せんべい
弁当、雑誌、ポテチ、せんべい

一般には

- データの中こういった、売れ筋の「組合せ」は、販売戦略にヒントを与える

＋ 店舗レイアウトの設計
＋ クーポンの発行、広告
＋ 顧客層の理解

実際は、1億件とかのデータから、大量のパターンを見つけることになる

… 商品の組合せ、たくさんあるから計算が大変。

効率的な探索

- うまく工夫すると、必要なところだけ上手に探索できる

鍵は、**「良く現れる組合せ」** を簡単にしても、現れる回数は減らない、という事実

「おにぎり、お茶、弁当」の組合せが3回現れる
→**「おにぎり、お茶」**は少なくとも3回

おにぎり、お茶、弁当、雑誌
おにぎり、お茶、弁当
おにぎり、お茶、弁当、せんべい
…

効率的な探索

「おにぎり、お茶、弁当」の組合せが3回現れる
→「おにぎり、お茶」は少なくとも3回

10回以上現れる組合せ
を見つけたいとして

おにぎり、お茶、弁当、雑誌
おにぎり、お茶、弁当
おにぎり、お茶、弁当、せんべい
…

＋「おにぎり」を買った人が10人以上いる
　→「おにぎり＋〇〇」を調べにかかる
＋「おにぎり＋弁当」を買った人は10人未満
　→「おにぎり＋弁当＋〇〇」は調べに行かない

広い広い組合せ空間の中の、必要な場所だけ調べられる

計算の工夫

- 「お茶」を買った人の数
を調べるには、データベース
全部調べないといけない

では、「お茶とおにぎり」は？

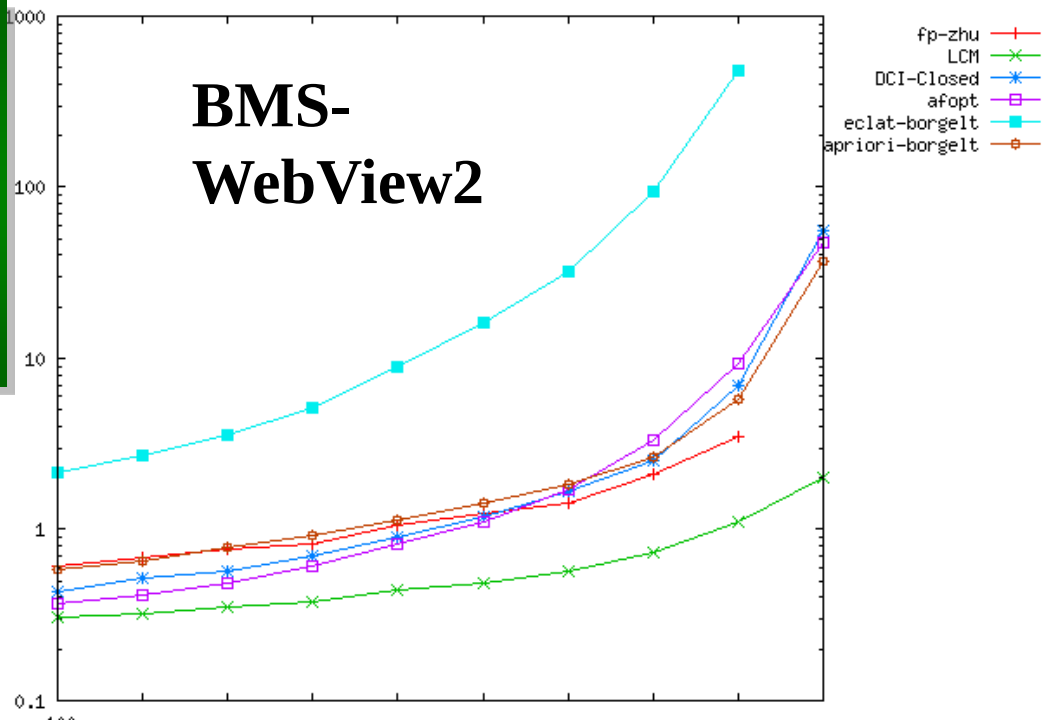
実は、お茶を買った人だけ
調べれば十分！

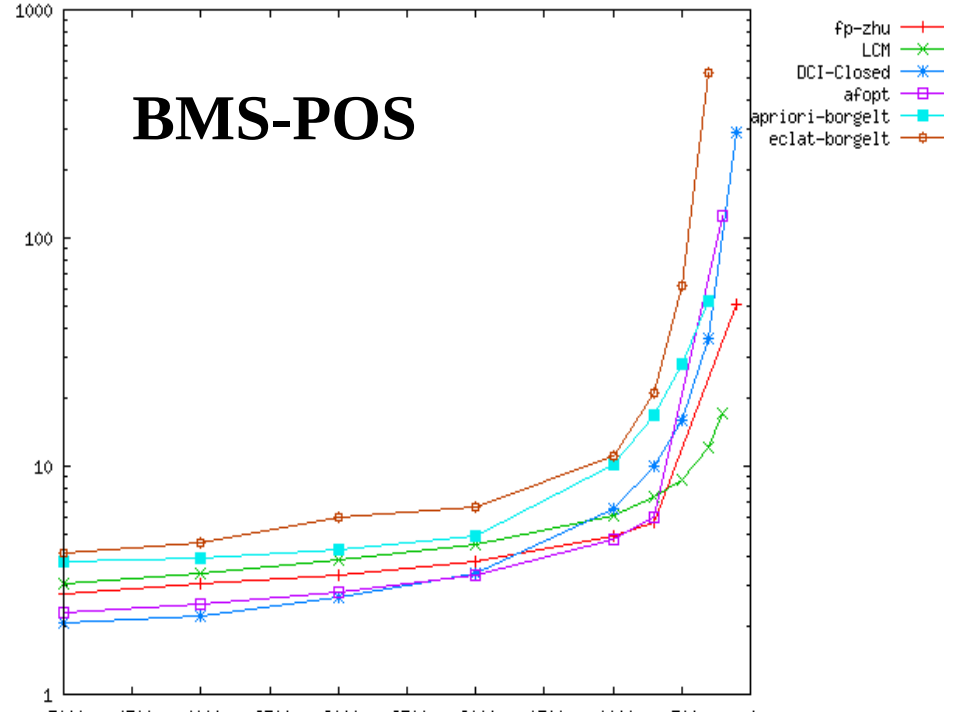
探索途中は、けっこう楽
ができるんです

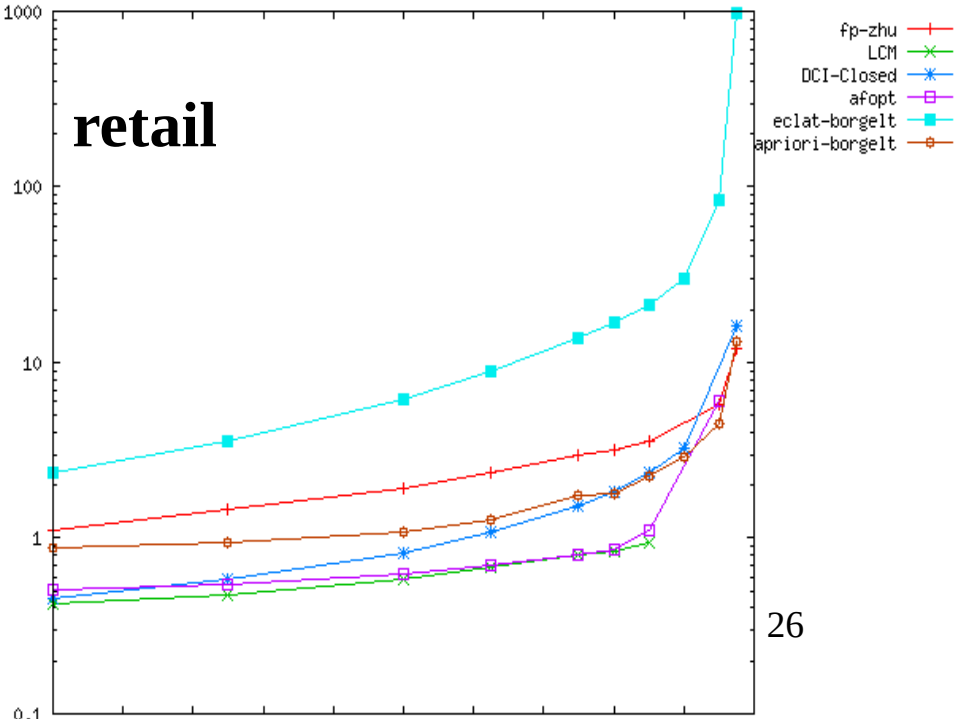
いろいろやると、100万倍くらい
速くなります

お茶、おにぎり
お茶、弁当
弁当
ポテチ、雑誌、せんべい
ポテチ、せんべい、ジュース
お茶、せんべい
お茶、おにぎり
お茶、おにぎり、雑誌
弁当
お茶、雑誌、ポテチ、せんべい
ポテチ、せんべい
弁当、ポテチ、せんべい
弁当、雑誌、ポテチ、せんべい

実データ（すかすか）平均の大きさ5-10







ルールマイニング

- 組合せの代わりに「ルール」を見つけるのが ルールマイニング

おにぎり を買った人は
お茶 を買う

お茶 を買っても
おにぎり を買うとは限らない

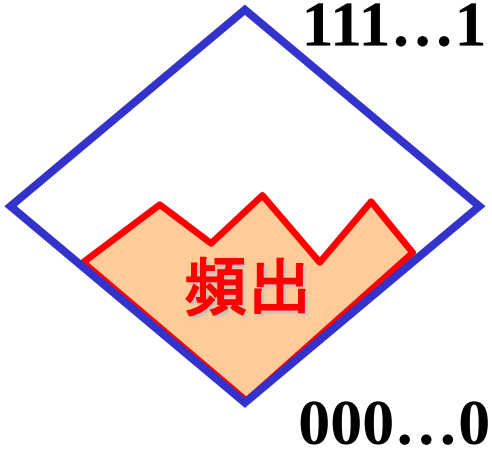
組合せがたくさん買われていればいい、というわけではない

計算は同じ方法で高速にできる

お茶、おにぎり
お茶、弁当
弁当
ポテチ、雑誌、せんべい
ポテチ、せんべい、ジュース
お茶、せんべい
お茶、おにぎり
お茶、おにぎり、雑誌
弁当
お茶、雑誌、ポテチ、せんべい
ポテチ、せんべい
弁当、ポテチ、せんべい
弁当、雑誌、ポテチ、せんべい

マイニングの技術の歴史

- いくつかのブレイクスルーがある



第1世代

apriori（幅優先）: ディスク上のデータ用

第2世代

深さ優先探索: データをメモリに保持

第3世代

データベースの圧縮: trie(FP-tree) などで再帰的に圧縮

第4世代？

基数ソート、振り分け、および ppc拡張による飽和集合列挙

第3世代：再帰的データ圧縮（FP-tree）

- データベースの縮約により、再帰の深いレベルでの高速化する

（1）前回追加したアイテムより小さいアイテムは消す

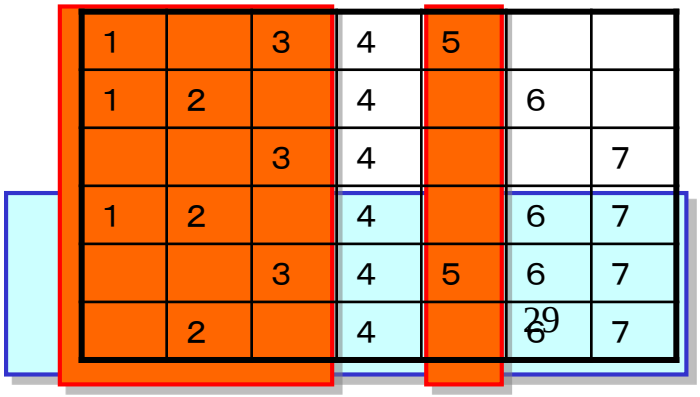
（2）現在の出現集合からできるデータベースの中で、頻出になっていないアイテムは消去する

（再帰呼び出しの中で加えられることが無いから）

（3）まったく同一のトランザクションは、1つにまとめる

- 実データだと、下のほうのレベルではだいたい大きさが定数になる

- FP-tree（trie）を使うと、共有する prefix も圧縮されるが、オーバーヘッドも大きい

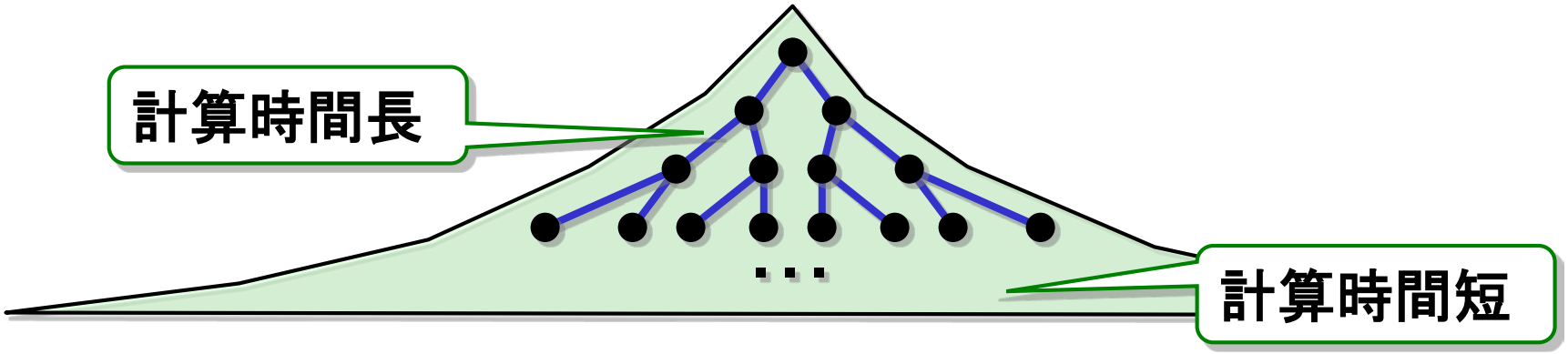


Σ が小さいときと速度の大きな差はない

末広がり性

- 再帰の末端以外はさほど高速化されていないのに、なぜ実際には大幅に高速化されるのか？

- バックトラックは、各反復で複数の再帰呼び出しをする
  - ➔ 計算木は、下に行くほど大きくなる
  - ➔ 反復の平均計算時間を支配するのは一番下の数レベル



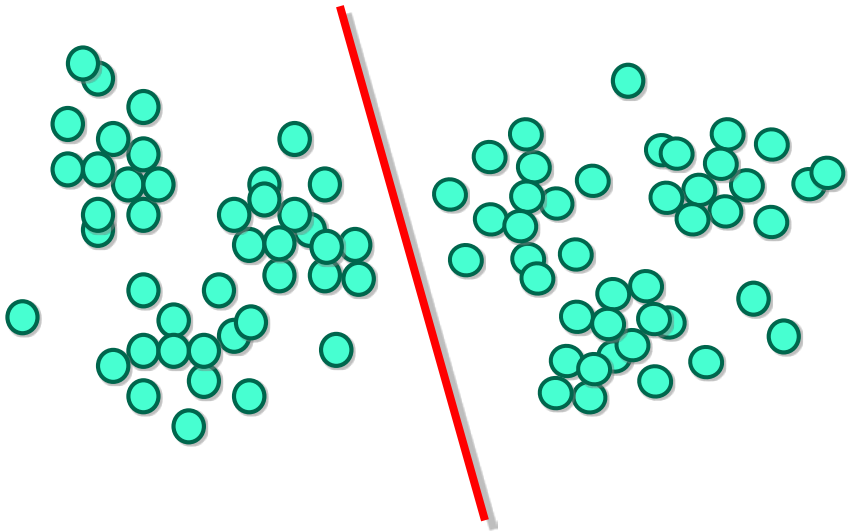
ダウンプロジェクトや圧縮による末端の高速化が全体を高速化している

データマイニングの技術
2. クラスタリング

クラスタリング

クラスタリング： データをいくつかの集まりに分ける問題

「性別」「年齢」のような、分ける要素がわかってないけれど、それでも分けてみたい、分かれ目があるのか知りたい、というときに使う　（集まりを **クラスタ** という）

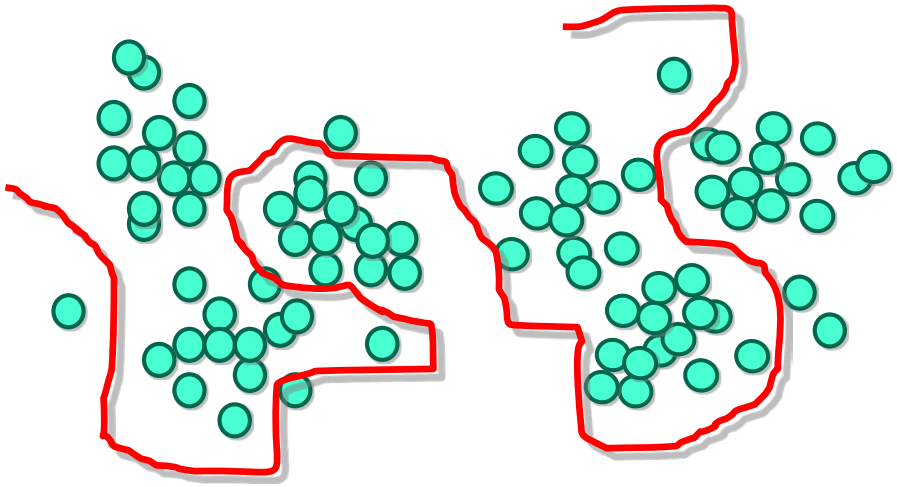


難しさ

- 分けようと思ったら、どんな分け方でもできます

じゃあ、どんな分け方だったら「それっぽい」でしょうか？

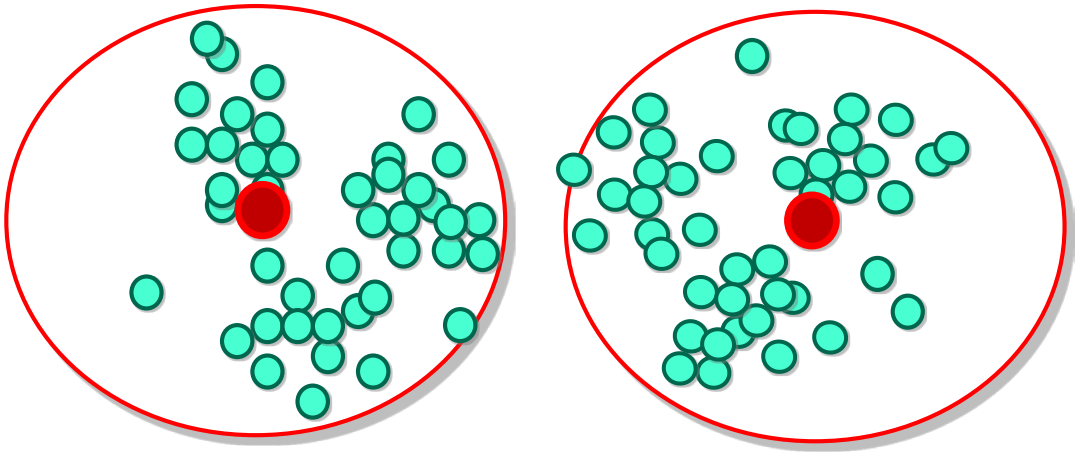
＋ 分け方がシンプル
＋ 分け目のところがはっきりしている



中心を与える

- こんな分け方を考えましょう

＋ 中心っぽい点が2つある
＋ 各点を、近い方の中心点のグループに入れる

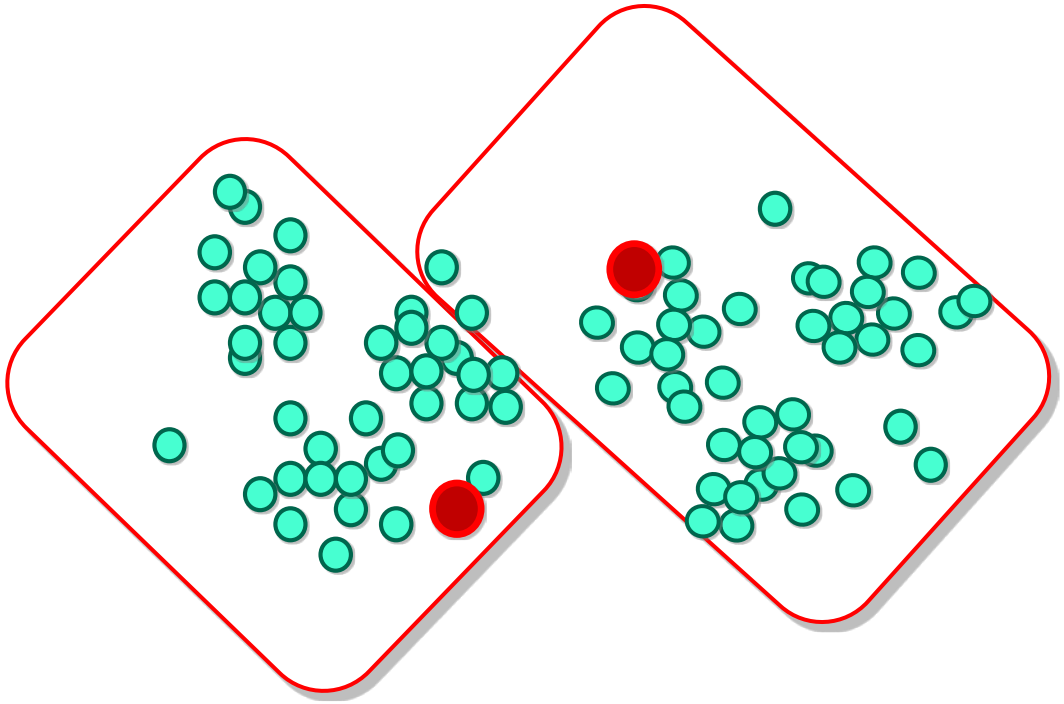


なんとなく、いい感じ？

じゃあ、これをどうやって計算するか考えましょう

だんだんと調整する

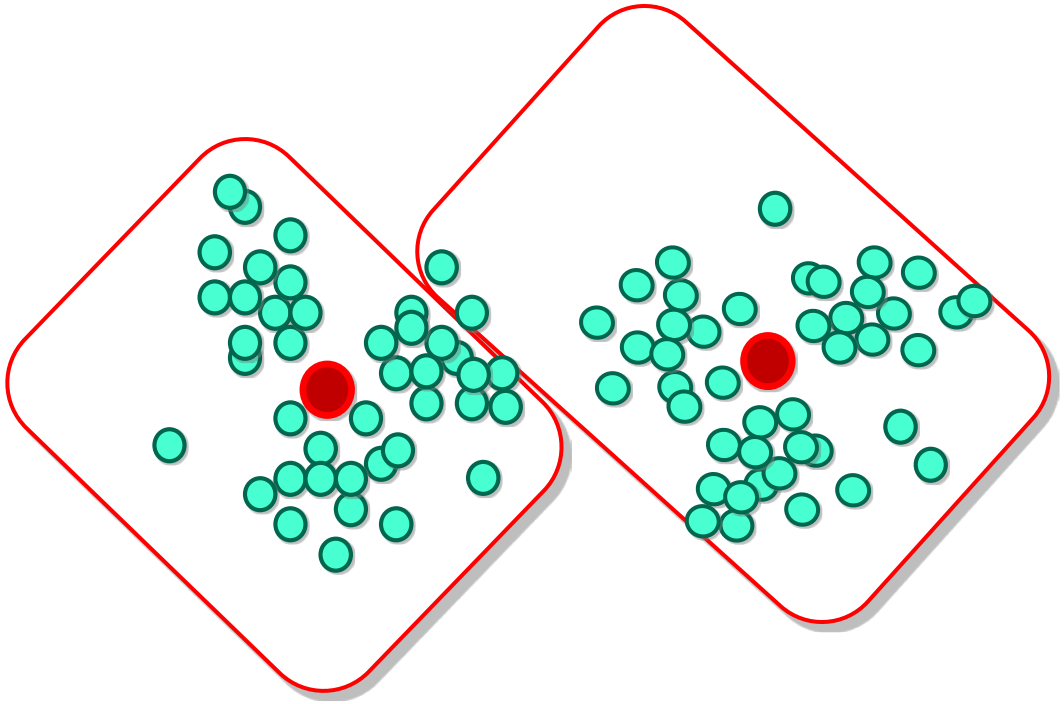
- 最初に適当に2つ点をおく
  これを元にグループ分けする



やっぱり、適当に中心を置くと、バランスが悪いですよね

だんだんと調整する

- 最初に適当に2つ点をおく
  これを元にグループ分けする

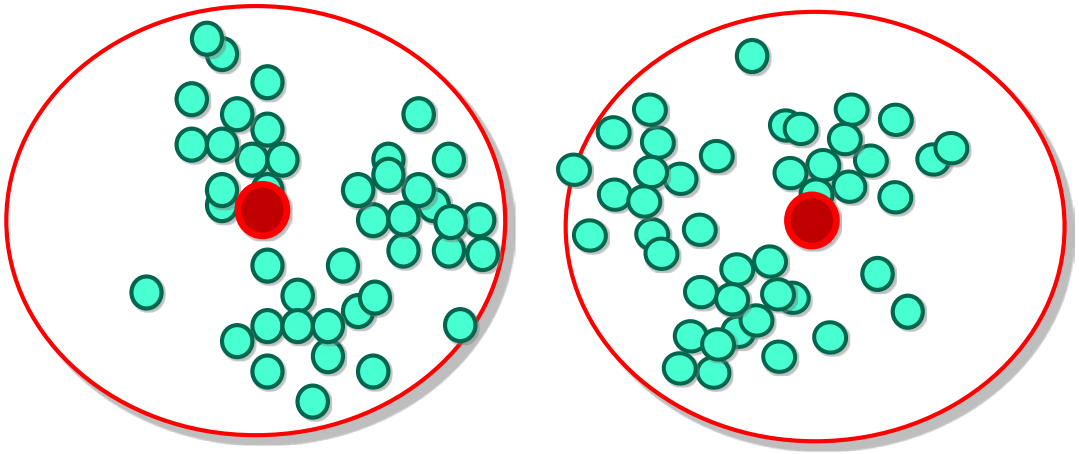


やっぱり、適当に中心を置くと、バランスが悪いですよね
じゃあ、点点の中心（正確には重心）に、移動しましょう

最後には

そして、またグループ分けをして、と繰り返す

・・・というようなことを繰り返していくと、最後には落ち着く
そこを「クラスタリングの結果」として出します



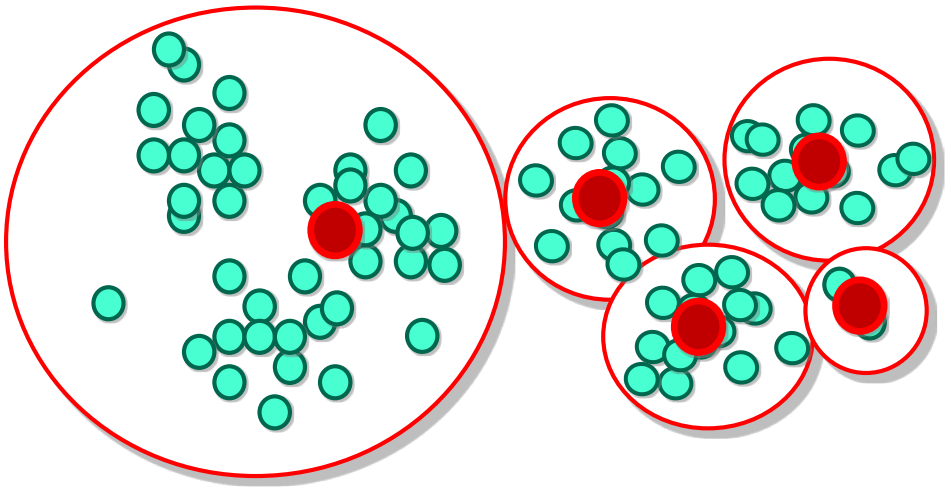
K-means

- この計算方法（アルゴリズム）、**k-means** っていいます

中心の点、を **k** 個にして同じことをすると、**k** 個のクラスタができます

最初に「ランダムに」点を置くので、毎回結果が違う

いい感じにわけてくれるとは限らない

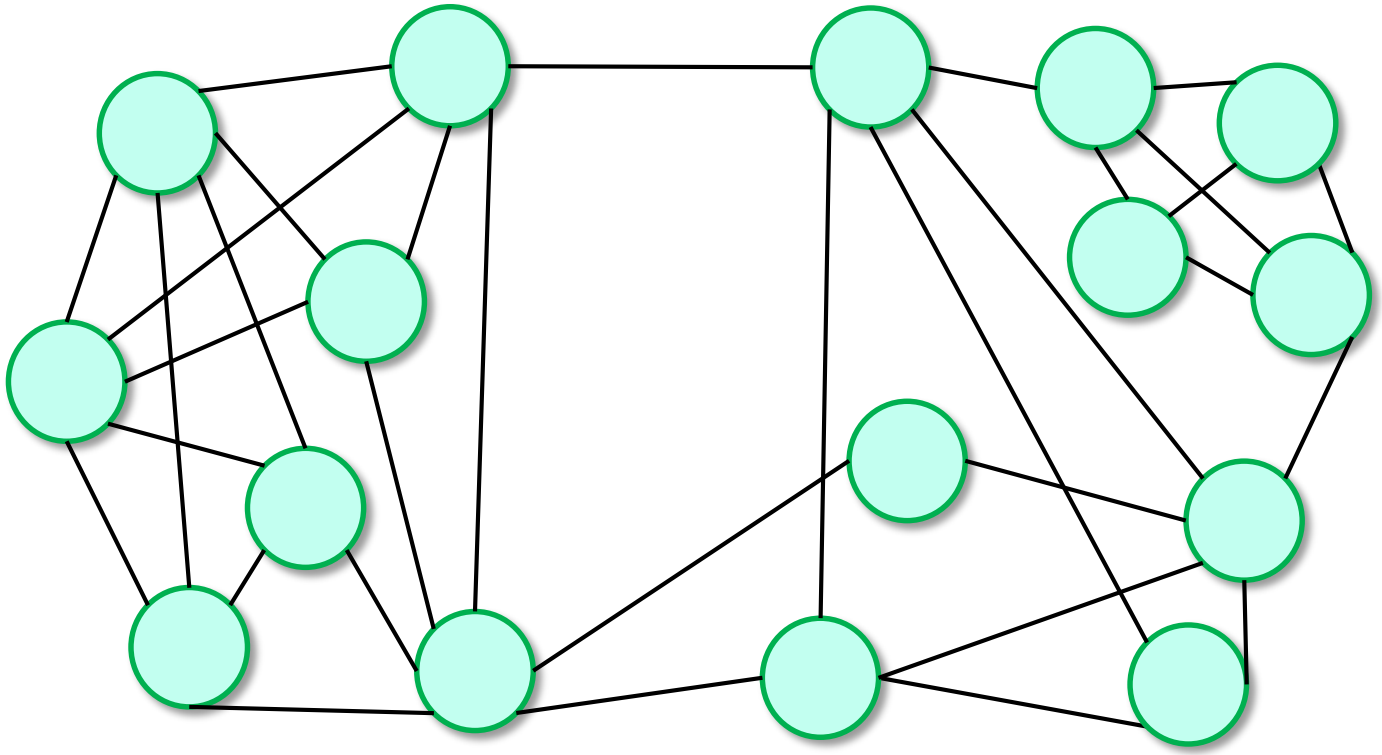


ネットワークの場合

ネットワークを、つながりが薄い部分を境目にして分割したい

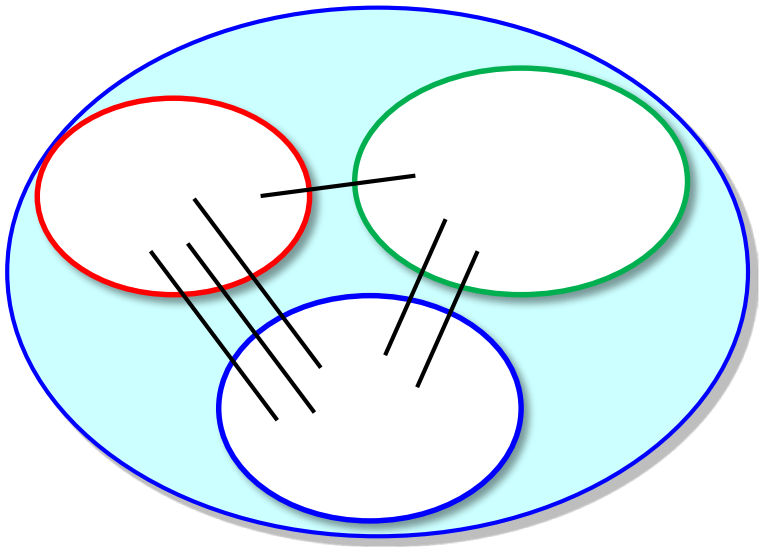
あるいは、濃くつながっているところを見つけたい

（距離が近い物、似ている物を結んだネットワークでも良い）



切り目最適化（グラフカット）

- ネットワークを、クラスタ間を結ぶ枝がなるべく少なくなるように **k** 個に分割する （トライ＆エラー、逐次改良、など）

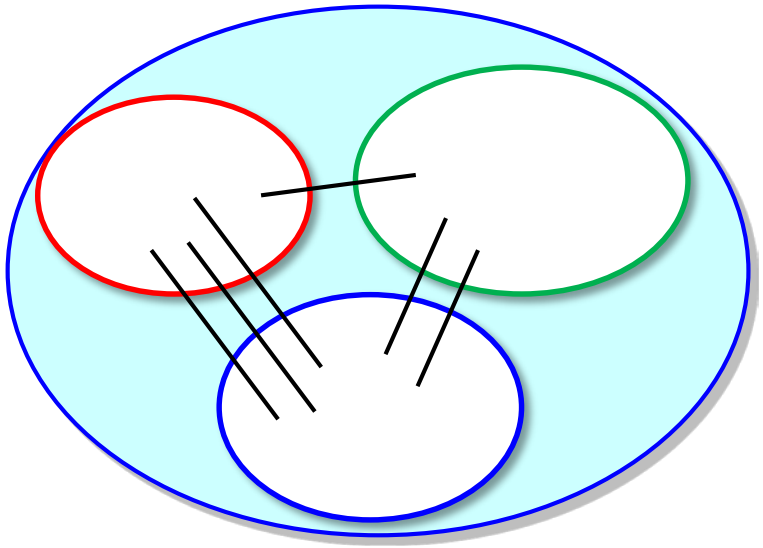


モジュラリティ

- クラスタの良さをはかる指標
  （内部の枝密度）／（外につながる部分の枝密度）

- Girvan-Newman 法は、すべてのクラスタのモジュラリティの総和を大きくしている

分ける個数を指定しない
でいいのが楽ちん

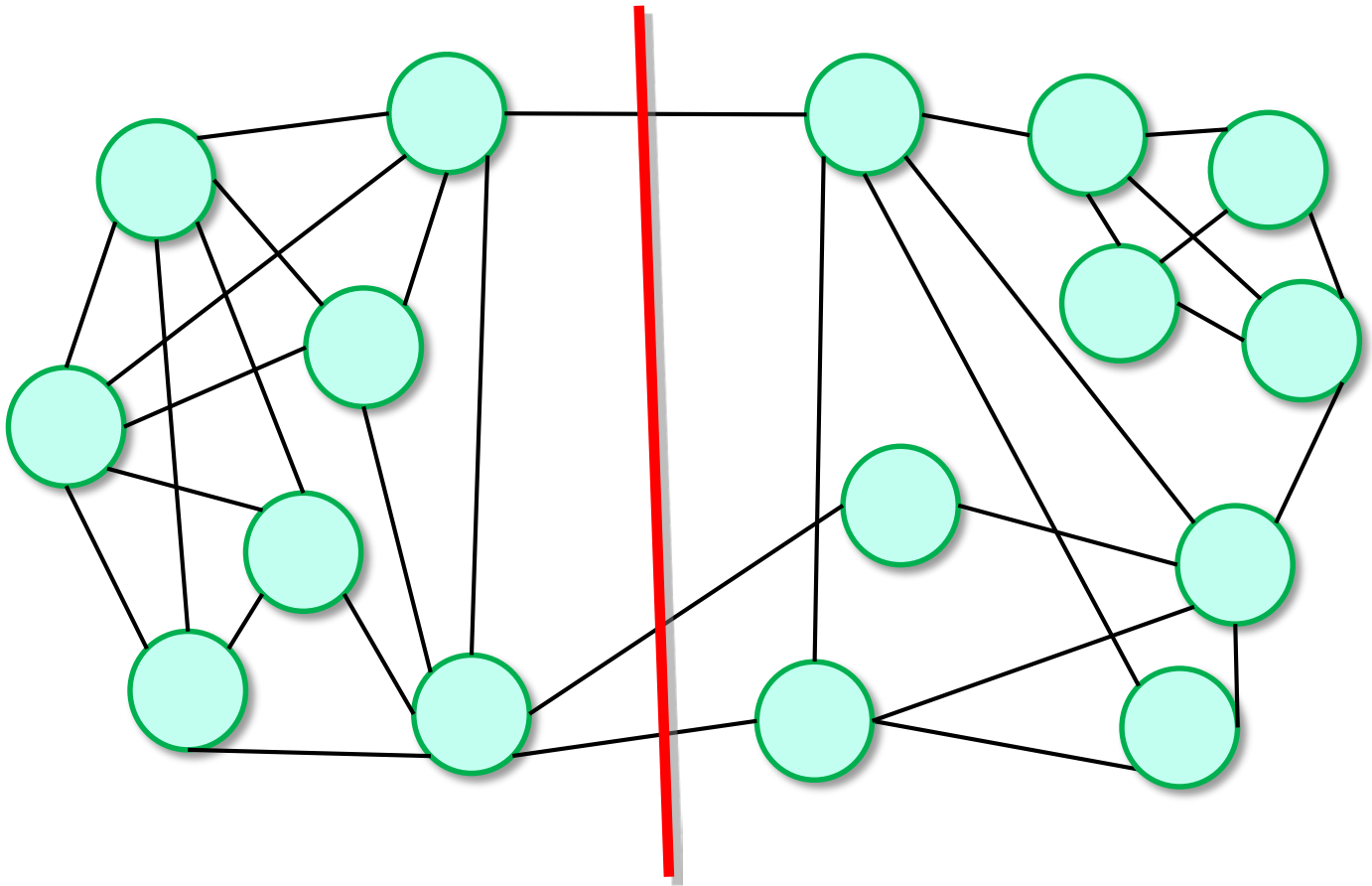


ネットワークの場合

ネットワークを、つながりが薄い部分を境目にして分割したい

あるいは、濃くつながっているところを見つけたい

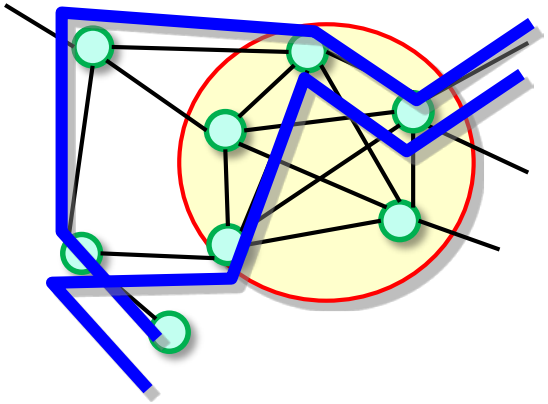
（距離が近い物、似ている物を結んだネットワークでも良い）



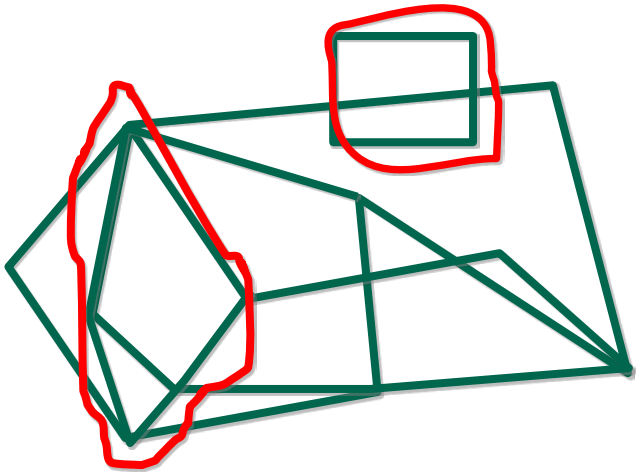
データマイニングの技術
3. 構造マイニング

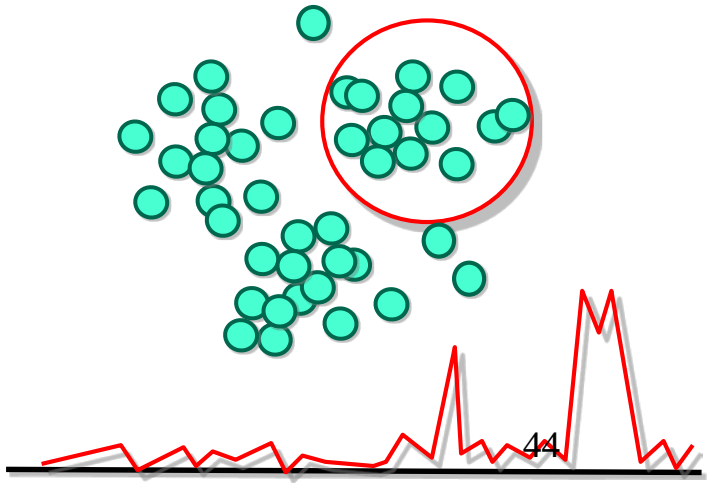
構造マイニング

データの中で、**「特定の形」**をしてる部分を見つける問題



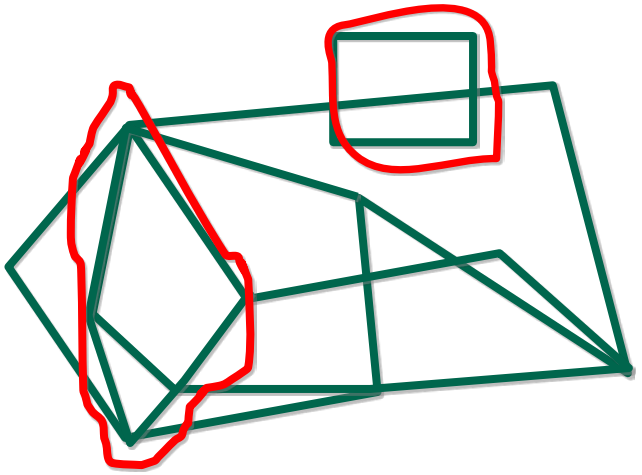
＋ ネットワークのパス、クリーク
＋ 図形データの三角、四角
＋ 点々や時系列データの
濃いところ、集まってるところ

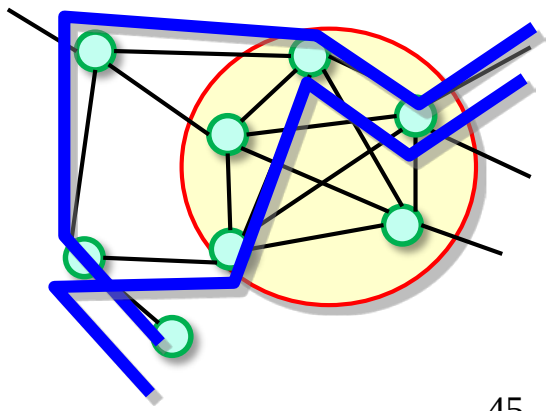




構造マイニングの特徴

- 部分的な特徴を見つけられる
  （通り道、友達グループ、異常…）
- どういうものを見つけるか考える必要あり
- 見つけるのはけっこう簡単 （場合分けを繰り返す）
  でも全部となるとすごい量　（1億とか）





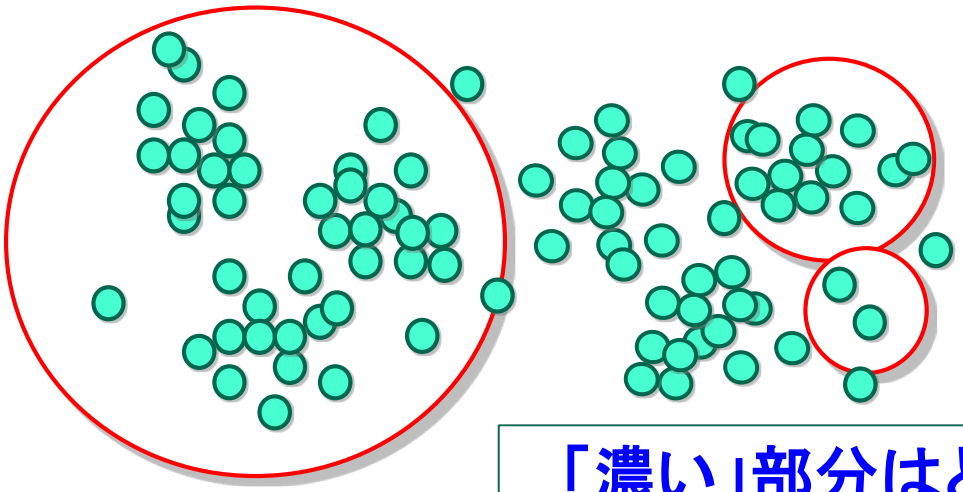
データマイニングの弱点

見つけたいものの「ゆらぎ」

- 数理的に「こういうもの！」と決まってるわけではないので、何が「正解」かわかりにくい

どっちつかずの場所がたくさんある

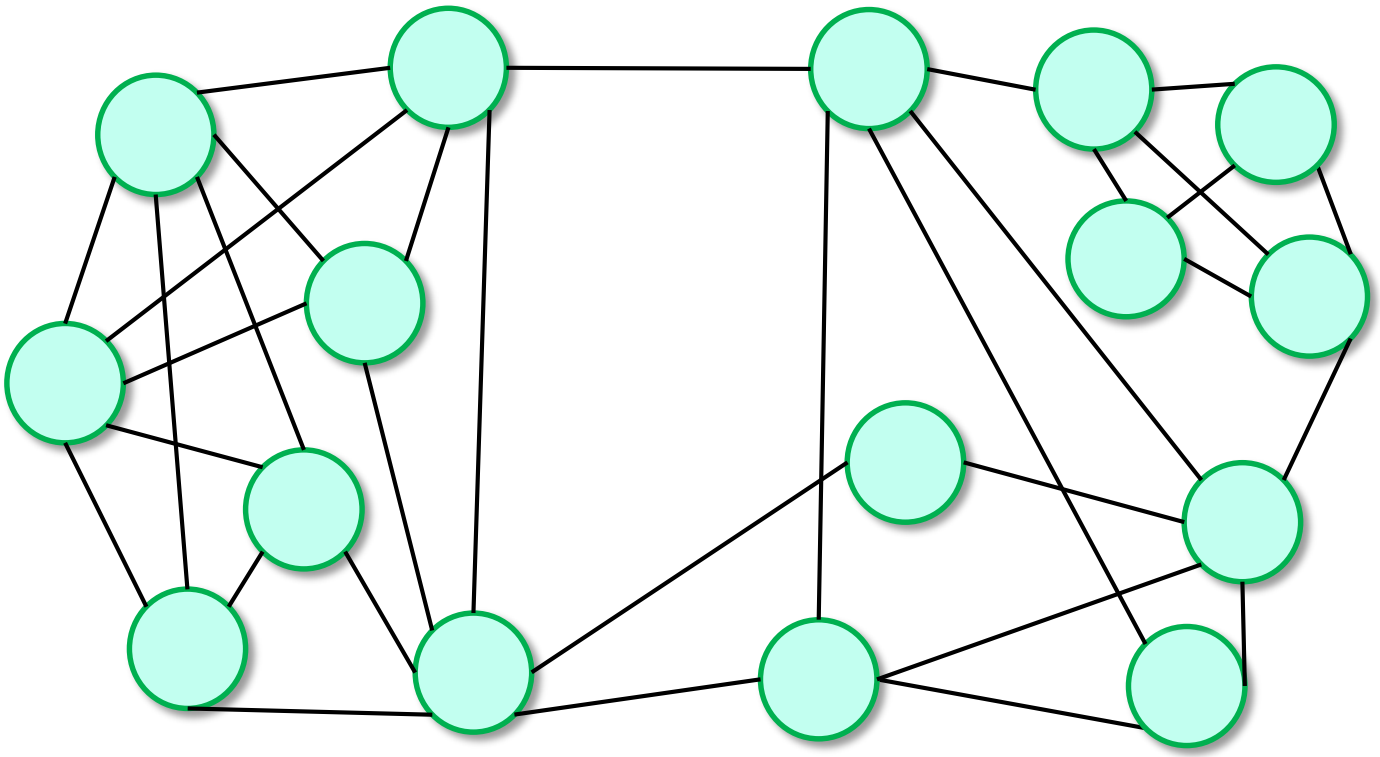
　大きいものなら、ぶれても困らないが、細かいものは影響が大きいし、場合を尽くすと大変なことになる



「濃い」部分はどこまで？

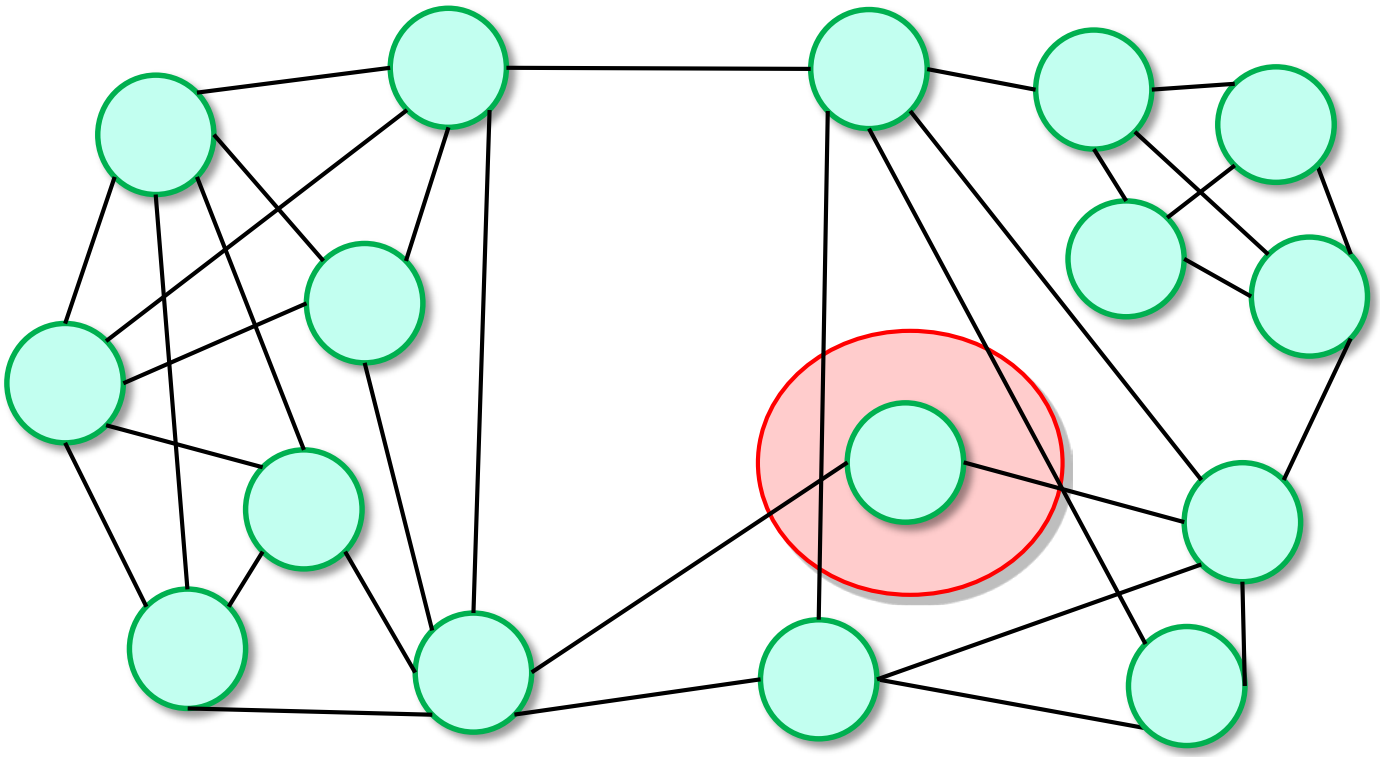
グラフカットの例

以下のグラフを2個に分割しろ



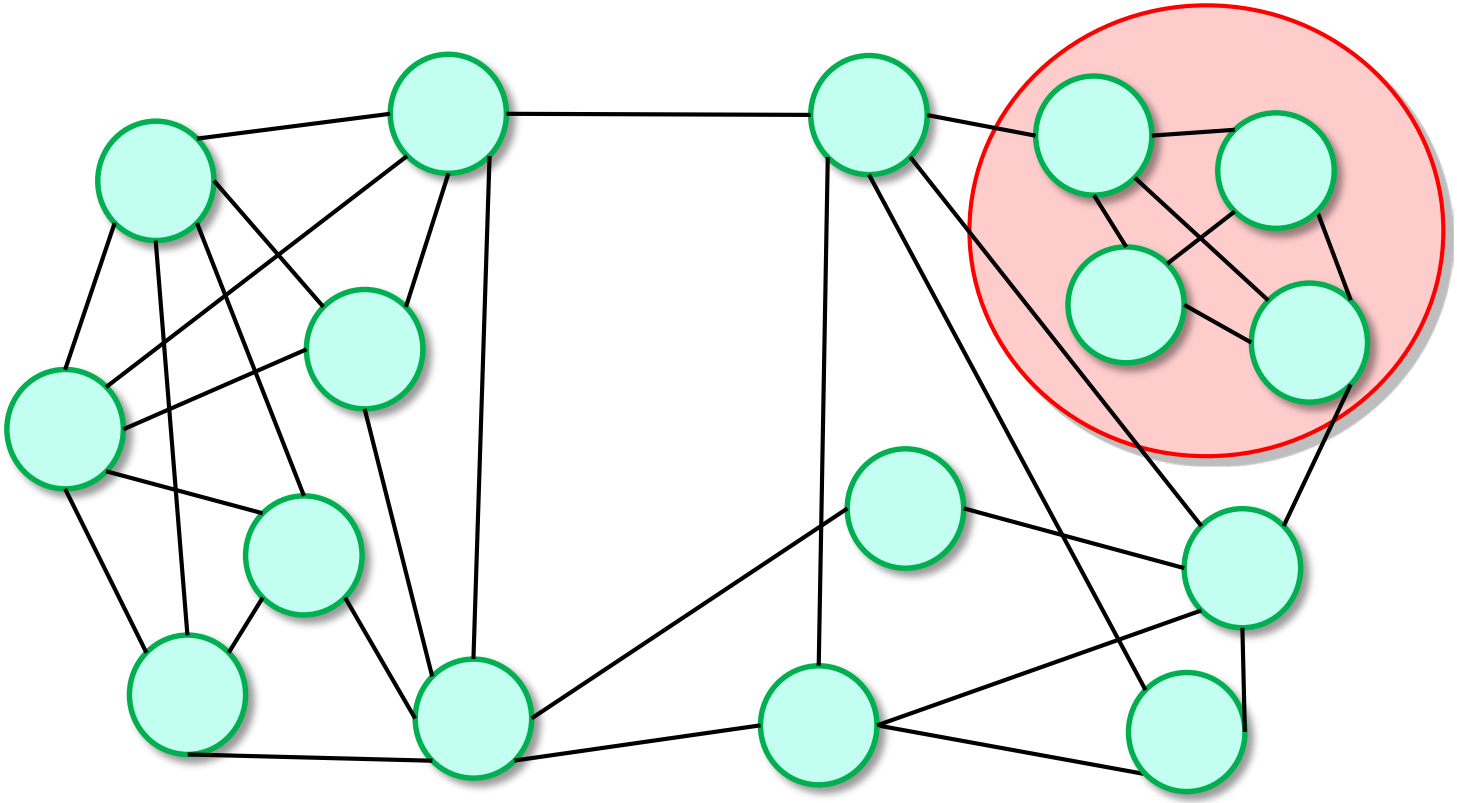
グラフカットの例

以下のグラフを2個に分割しろ



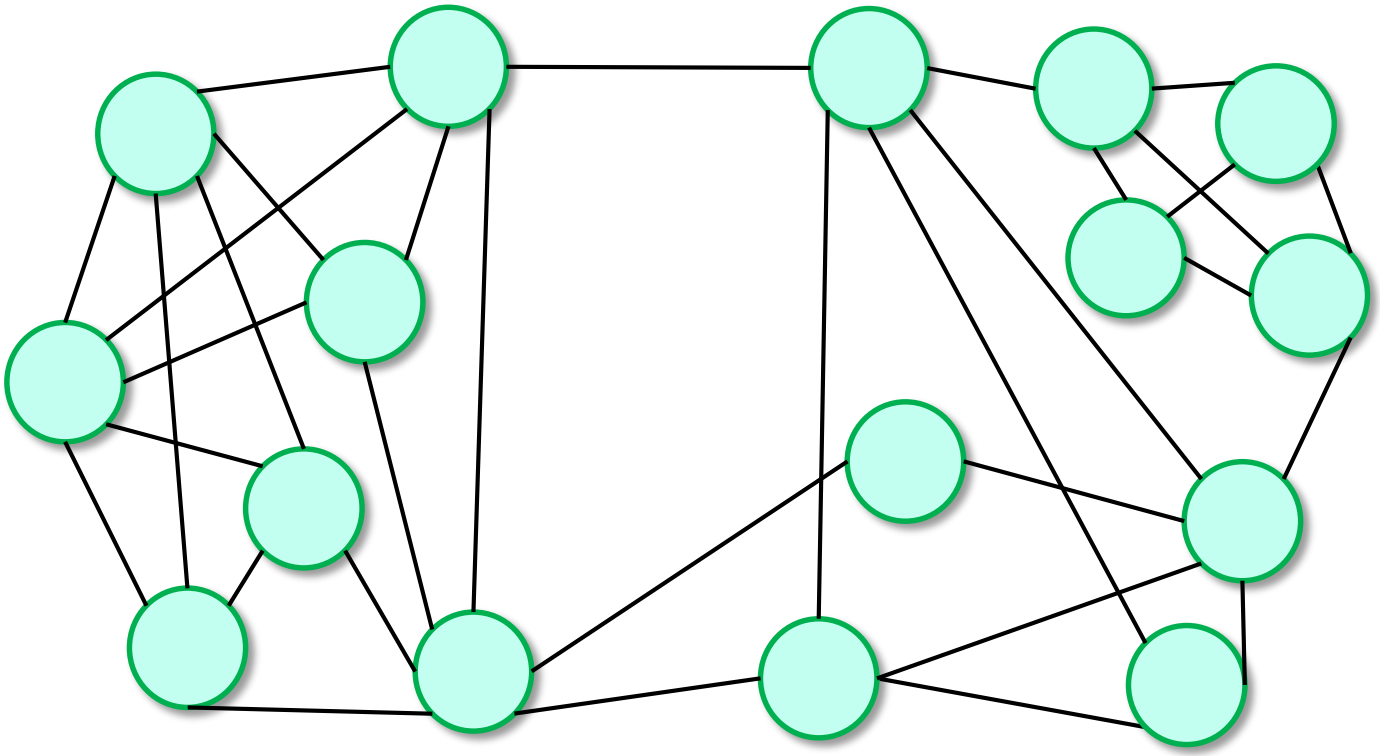
グラフカットの例

以下のグラフを2個に分割しろ



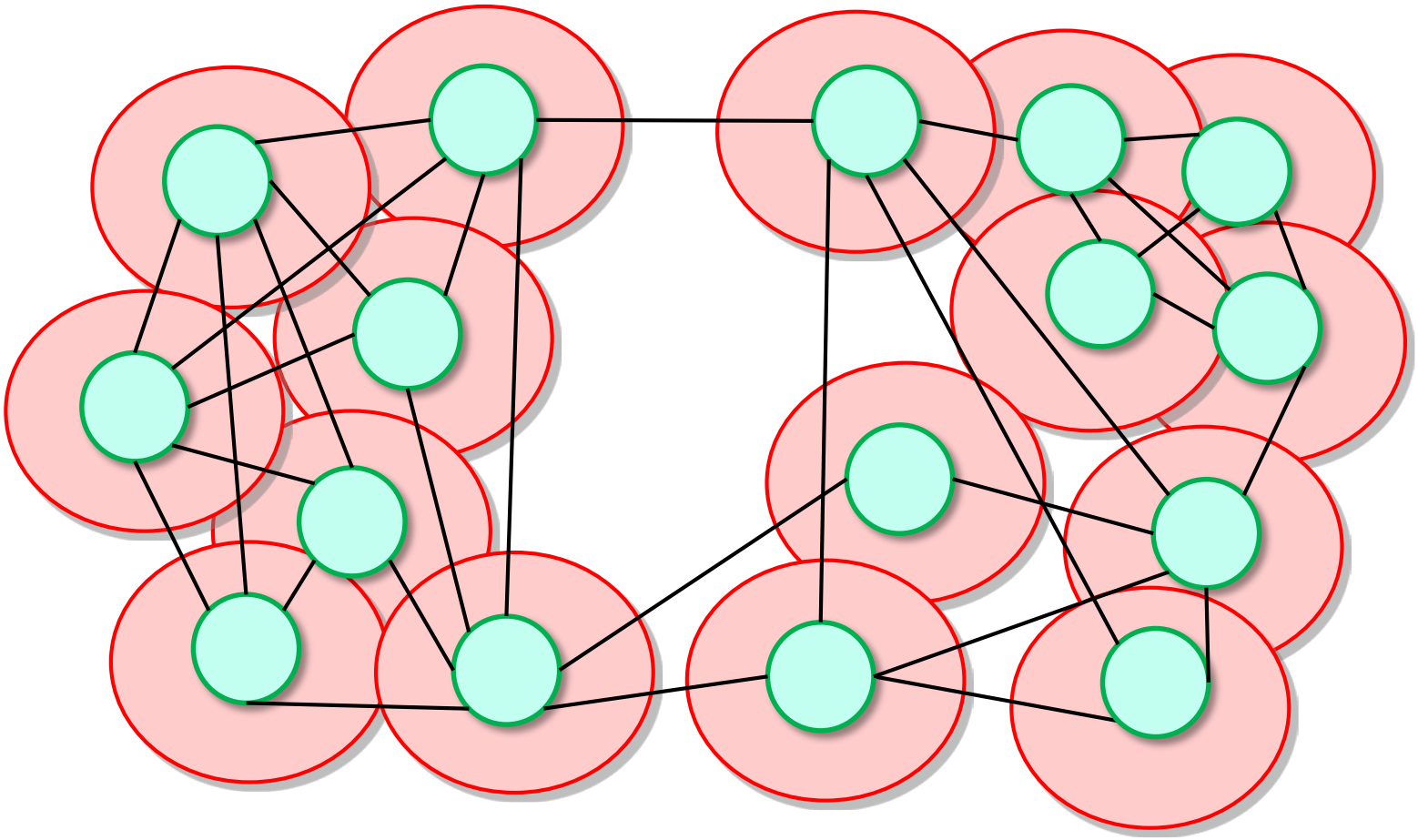
モジュラリティ最大化の例

以下のグラフをモジュラリティが最大になるように分割しろ



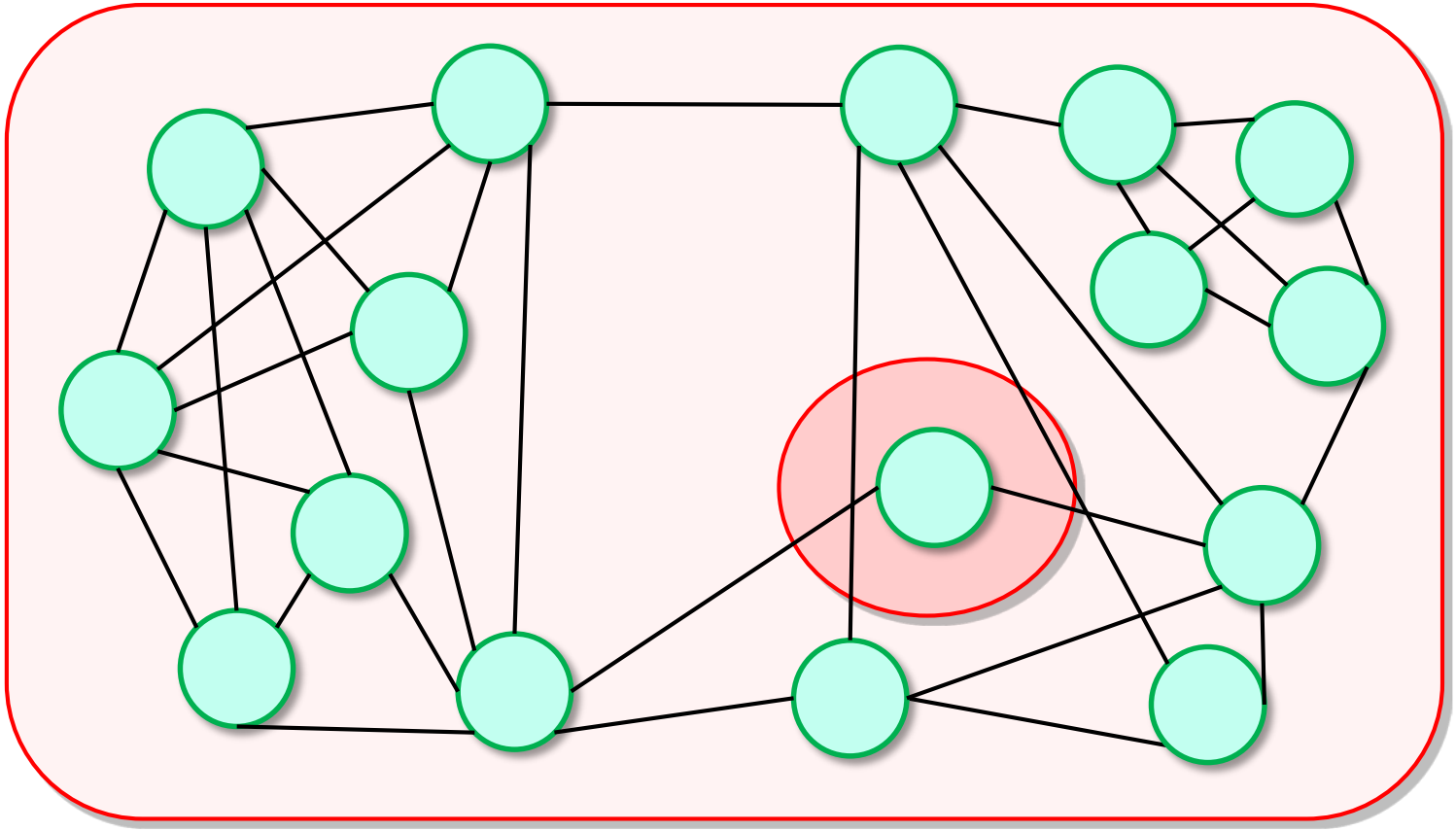
モジュラリティ最大化の例

以下のグラフをモジュラリティが最大になるように分割しろ



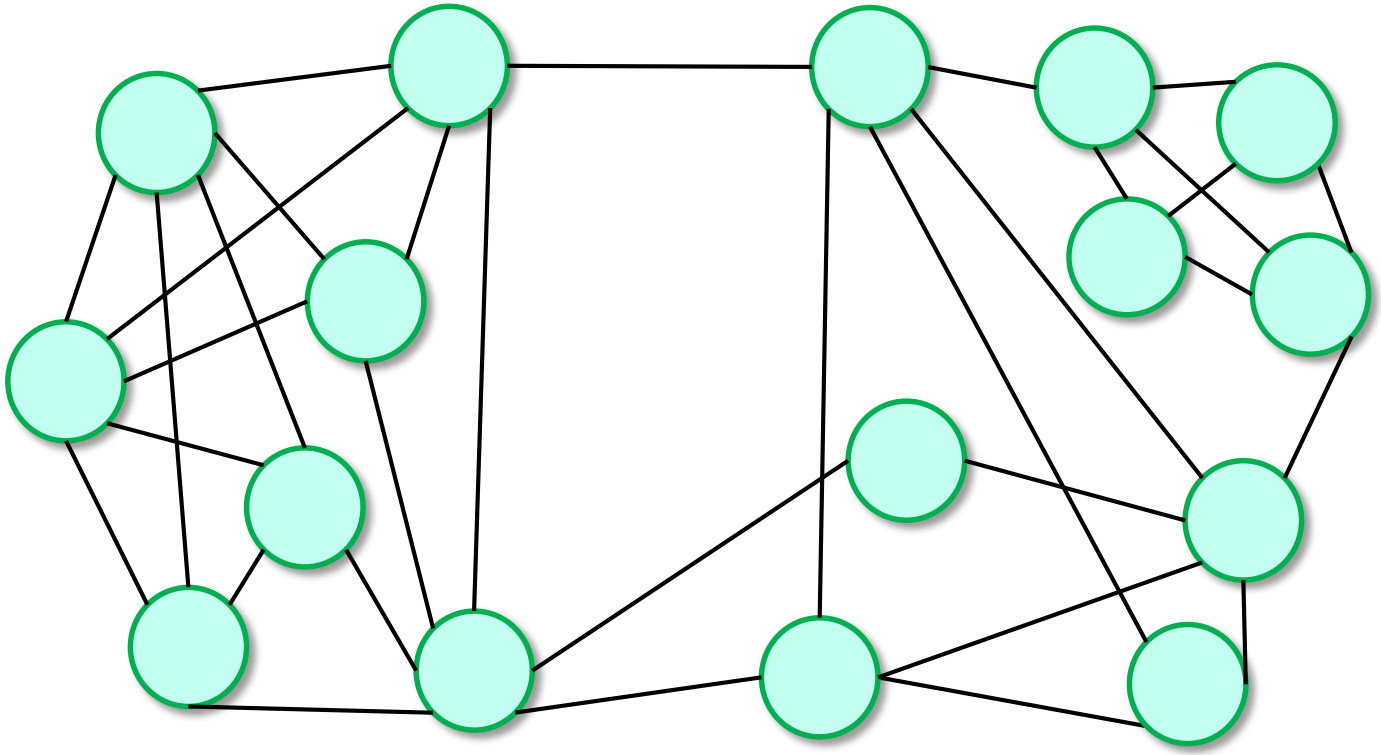
モジュラリティ最大化の例

以下のグラフをモジュラリティが最大になるように分割しろ



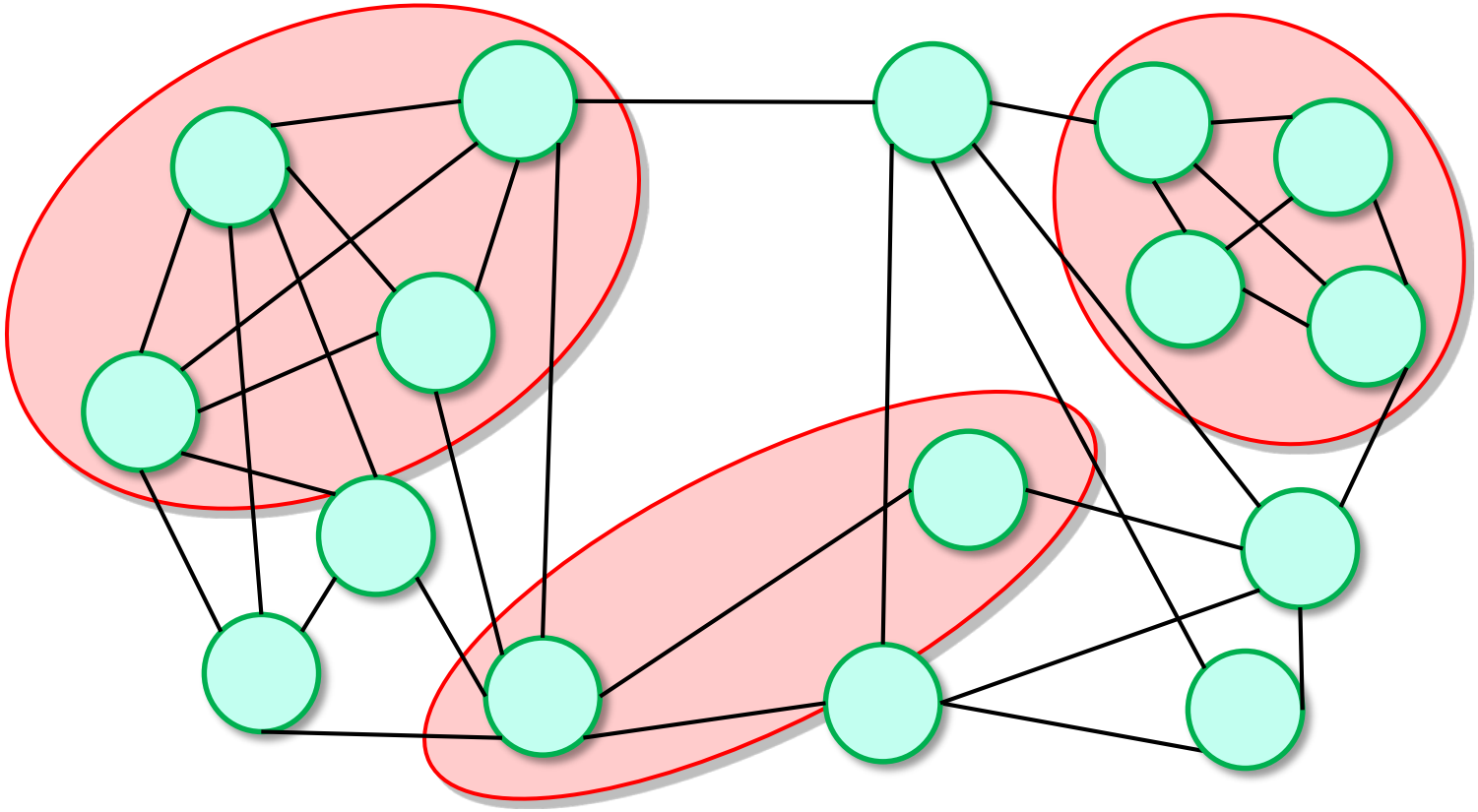
極大クリーク列挙

完全グラフである部分グラフをクリークという
他のクリークに含まれないクリークが極大クリーク
極大クリークはコミュニティやクラスタの基本的なモデル



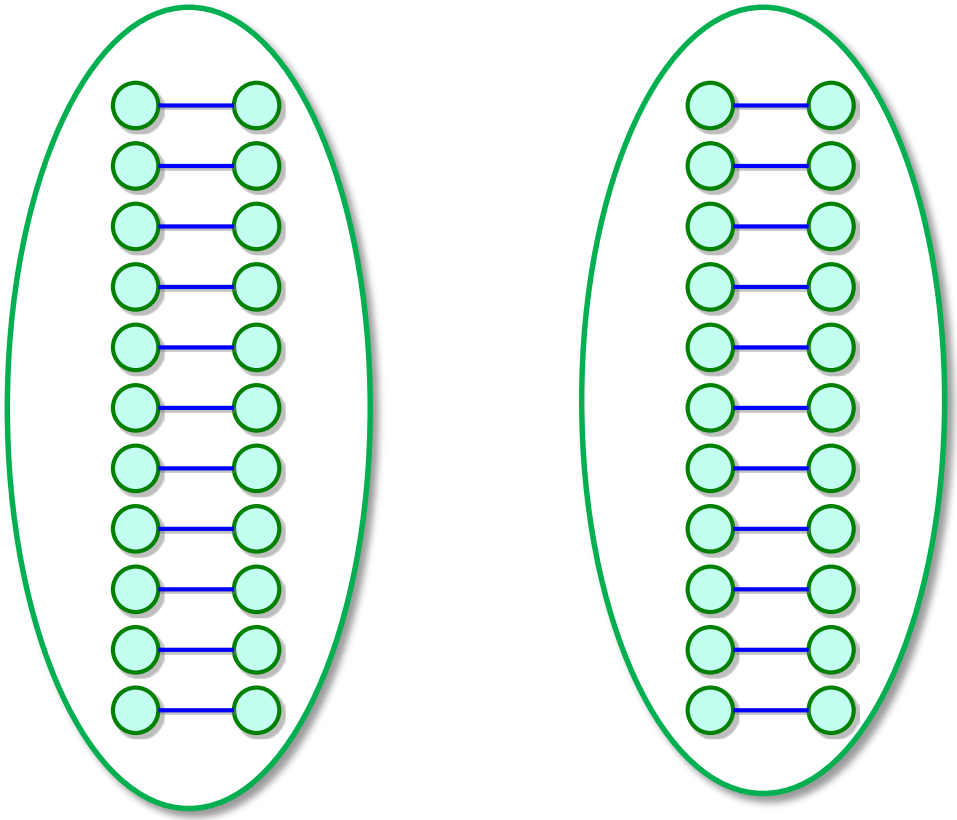
極大クリーク列挙

完全グラフである部分グラフをクリークという
他のクリークに含まれないクリークが極大クリーク
極大クリークはコミュニティやクラスタの基本的なモデル



極大クリーク列挙の例

以下のグラフ、極大クリークはどこにいくつあるか



$2^{11} + 2^{11} = 4096$

密度 80%以上の極大グラフは、少なくとも
$20(2^{11} + 2^{11}) = 81920$

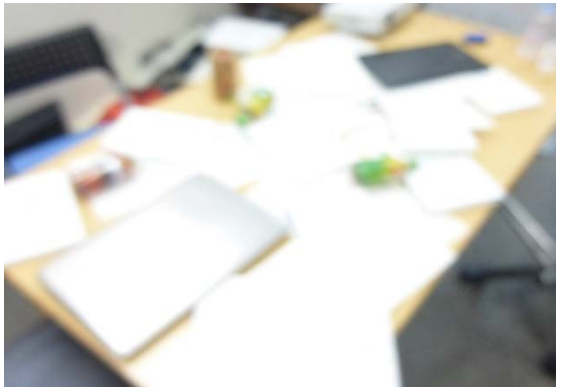
自分のほしいモノを数理で表現するのは非常に難しい

今まで研究では、変なものを無理矢理ユーザに押しつけていた

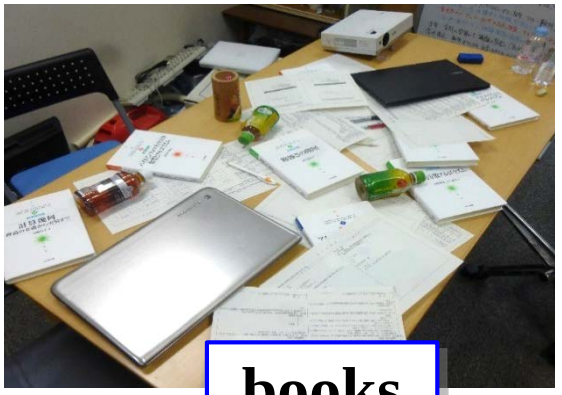
データ研磨による「明確化」

データを見えるようにする

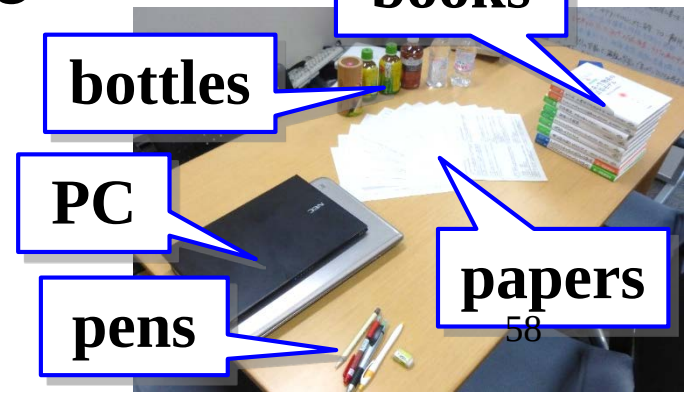
- **密な部分**の境目があいまいなので、解が大量／計算が大変／精度が出ない
  - → 境目がクッキリしていればいいのに、、、



- 画像処理だったら、
  - ＋ コントラストを強めてエッジを強調する
  - ＋ ぶつぶつのノイズを消してフラットにする



- 中身がわかりやすくなり、認識精度も上がる

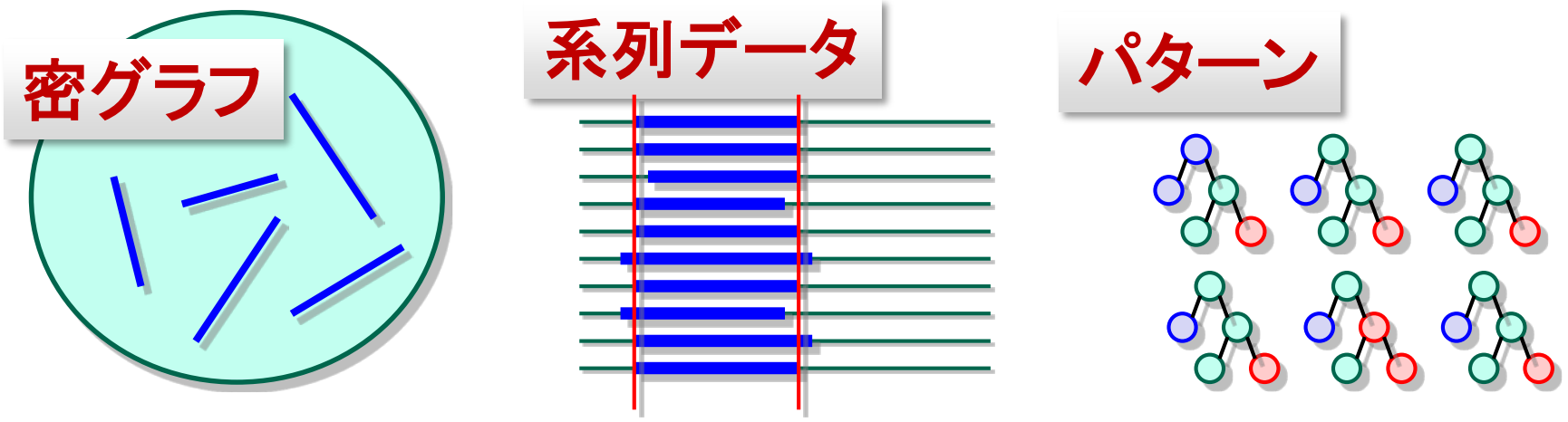


データでも同じことをしよう！

新しいアプローチ： データ研磨

データ研磨

確実な根拠に基づき、データの「揺らぎ」を消す



- 「明らかにこうだろう」を変更。損失なく、網羅性も担保
- 揺らぎが消え、大量の類似解はまとまる

グラフ研磨：周辺情報を利用

2つの頂点が同じクラスタに属するかどうか、周辺情報で判断

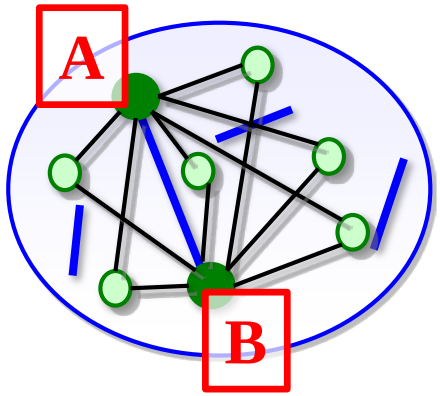
実行可能仮説

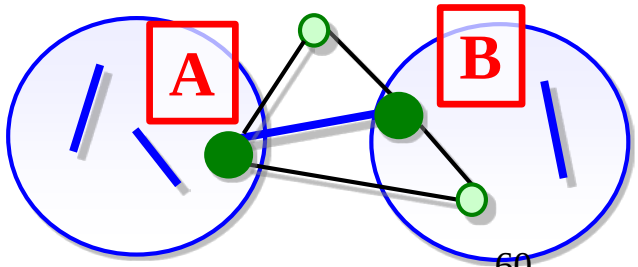
A と B が同じクラスタにいる ⇔ A と B は共通隣人を多く持つ

Aの隣人 ∩ Bの隣人 ≧ k → AB 間に枝をはる

Aの隣人 ∩ Bの隣人 ＜ k → AB 間の枝を切る

+ これを一度に全部行い新しいグラフを作る
+ 変化しなくなるまで繰り返す





- 友達集合の類似性を使うと、
  クラスタの粒度がそろう

なぜ今までは無かったか

- アイディア自体は簡単。でも、無かった。

★「データを変えてしまっていいのか」という疑念

→ 中規模の構造は保存するようにしている
（誤字やピンぼけの修正のようなもの）

★ 計算の難しさ （密部分を網羅的に見つける）

→ 新しいモデル化で対応

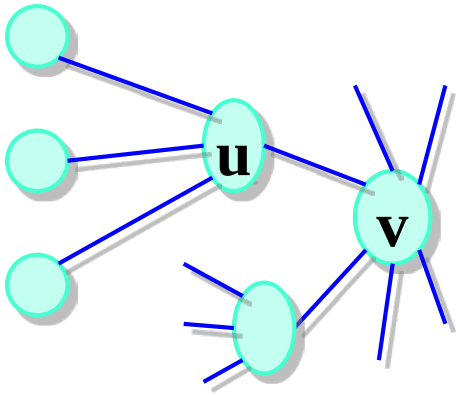
→ 最新の高速アルゴリズムを利用

友達の友達

- 「v」さんと**友達**を共有している ➔ 「v」の**友達の友達**

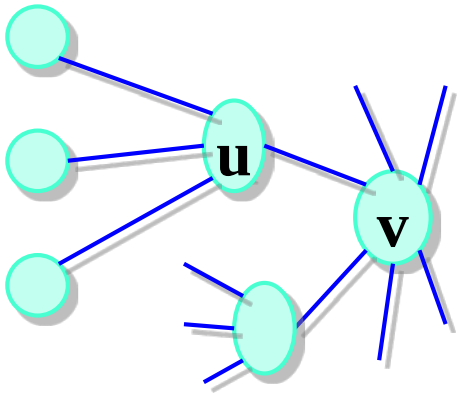
- 「v」さんは、友達の友達とだけ比べればいい
  ➔ 比較対象がものすごく少なくなる
  （が100万点でも、友達の友達は100-1000人くらい）

- 「友達ごとに、その友達の友達の
カウントを上げる」で線形時間に



- 「次数分布がべき乗則」なら全体でも線形時間

べき乗則



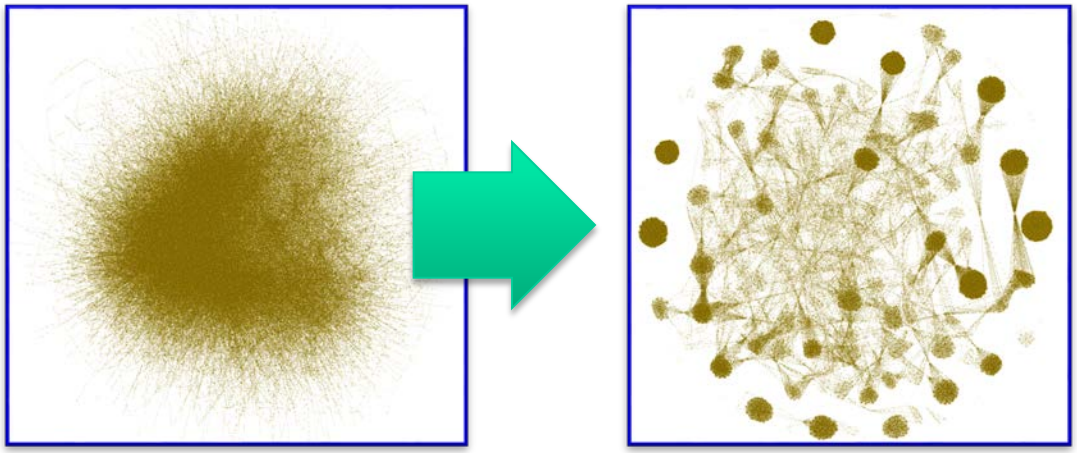
- $N(v)$ を v の隣接頂点の集合とする
- v に関わる計算時間は $\Sigma_{u \in N(v)} | N(u) |$
- すべての v について足すと $\Sigma_v \Sigma_{u \in N(v)} | N(u) |$
- 各 $|N(u)|$ は u の各隣人から足されるので、$|N(u)|$ 回足される
  - ➔ $\Sigma_u |N(u)|^2$
- $|N(u)|$ がジップ則に従い、次数順で i 番目の次数が $\alpha / i^k + \beta$ で押さえられるとしよう（α はグラフの最大次数に対応）
  - ➔ $\Sigma_u |N(u)|^2 \leqq \Sigma_i (\alpha / i^k + \beta)^2 = O(|E| + \alpha^2)$ （$k>1/2$ なら）

データ研磨の威力

- 帝国データバンク様、企業間取引データ

  点：企業、線：取引ある、似てる：共通友達 PMI ≧0.6

	研磨前	研磨後
頂点数	3,282	3,282
枝数	35,168	73,132
クリーク数	32,953	343



買い物データを用いた、顧客が健康志向かどうかの予測精度

	クリーク	Newman	グラフカット
研磨前	60.60%	59.70%	60.03%
研磨後	**71.36%**	**62.76%**	**67.78%**

人工生成したベンチマークデータにおける、意味的構造の検出率

	研磨	Newman	グラフカット
ノイズ10%	**100.00%**	68.74%	76.10%
ノイズ40%	**99.69%**	7.91%	77.03%

結果：データ研磨

- 数が妥当。大きさの分布も良い。中身も妥当
  境目は明確になったようだ

データ研磨によるクラスタの一例

トヨタ14年3月期営業利益は1兆8000億円へ、市場予測下回る
ホンダの今期連結営業利益は前年比＋43.2%の見通し、市場予測下回る
NTTドコモの今期営業利益は0.3%増益を予想、市場予測をやや上回る
ファーストリテ、12年9―13年2月期営業利益は前年比＋5.3%
オリンパス、今期営業利益予想を前年比-1.5%の350億円に下方修正
日産自、12年4―12月期営業益は18.4%減の3491億円
ソニー、今期当期損益予想を200億円の黒字で変更せず
ニコン、13年3月期営業利益予想を前年比-40.1%に下方修正
三菱重が13年3月期営業益予想を上方修正、予測上回る

…

サイト分類

- Web訪問履歴データ
  各項目は各ユーザが訪問したサイトの列
  (10万ユーザ20万サイト)

user 1	site A	site B	site C	…
user 2	site C	site F	site A	…
…				

- サイト間の類似グラフを、訪れたユーザの類似度で作成

site A	user 1	user 2	user 4	…
…				

- データ研磨により、およそ1000個のクラスタを発見
  しかも、各クラスタは密に関係しているものばかり

+ サーフィン情報 / 海洋天気予報 / サーフィン商品…
+ オークション（複数サーバ） / オンラインバンク
+ 地震情報 / 地震雲 / 地理センサーデータのサイト

顧客分類

- マクロミル社の購買履歴データを利用 (6500人、1年間)
  モニターさんが、購入品目と購入した店を記録
  各モニターさんは、ライフスタイルについてのアンケートを記入

- 顧客の嗜好分析をしたい　→ 例として、健康志向をとりあげる
  アンケート項目に記入されいている、健康志向的な項目に〇をつけた人を健康志向とみなす

- 顧客の購買行動から、健康志向かどうかを予測してみる

顧客分類

- 学習モデルに、ロジスティック回帰を使用
  - ×モニター数に比べ、変数の数が多すぎる
  - ×各説明変数を含むモニター数が非常に小さく、スパース
    （店と商品の組合せなので）

- 説明変数を、その変数を含む顧客の集合、の類似度でクラスタリングする

店 **A** で商品 **B** を買った人が　（**X**、**Y**、**Z**）
店 **C** で商品 **D** を買った人が　（**W**、**X**、**Y**、**U**）
} **似ている**

- クラスタリングの質を上げるために、データ研磨を用いる

顧客分類

- グラフそのままでクラスタリングすると、多くのクラスタが出てくる (2000個ほど) 研磨すると、きれいになるので、大幅に減る

- 研磨は、パラメータによって、出てくるクラスタの粒度が変わるので、値を刻み、すべてのパラメータについてクラスタを求め、すべてを使う (合計1500ほど)

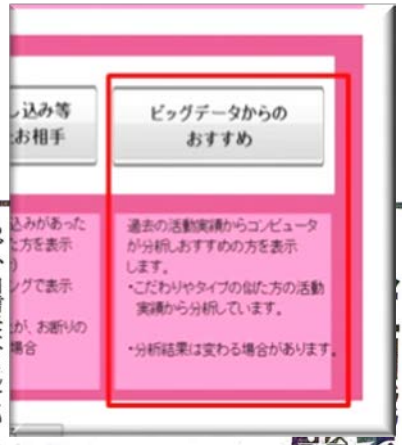
結果、研磨前と後で比較して、予測精度が3-10%ほど向上した

	クリーク	Newman	グラフカット
研磨前	60.60%	59.70%	60.03%
研磨後	**71.36%**	**62.76%**	**67.78%**

婚活におけるお勧め（愛媛結婚支援センター）

- プロフィール検索では、出会い人の範囲がせまくなりすぎる
  → 「好みが似ている人は人格も似ている」仮説から

閲覧した異性の類似によるユーザクラスタを使って推薦



お見合いを受ける率が２．２倍に上昇！

愛媛新聞ONLINE

Ads by Google 愛媛県 婚活の 松山市

愛媛のニュース

愛媛の婚活方式　特別賞、総務省表彰

総務省は25日までに、ICT（情
賞に、県のえひめ結婚支援センタ
同連合会や四国総合通信局に
たシステムを基に、お薦めの異性
を促した結果、お見合いに至る引
地域情報化大賞は全国から85件の応募があり、大賞や部門賞など12件を選出した。3月9日に東京で表彰する。

総務省 地域情報化大賞 特別賞

5年間活動してきました。「ビッグデータのお勧め」で、初めて、結婚してもいいと思える人と出会えました！

インターネット広告のターゲッティング精度向上

- ユーザのWeb閲覧履歴を分析、広告に興味ありそうな人を絞込むサイトをデータ研磨でクラスタにし、ユーザ行動の意味をとらえやすくしたことにより、学習アルゴリズムの精度が向上

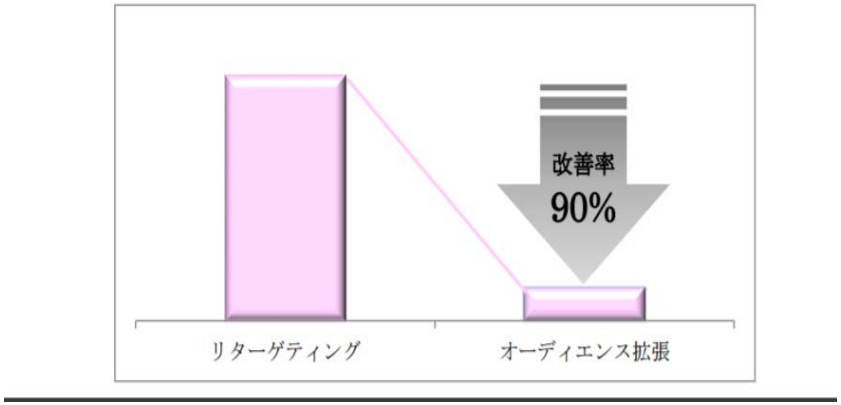
ターゲッティングの改善率が90%！

PLATFORM ONE

News Release

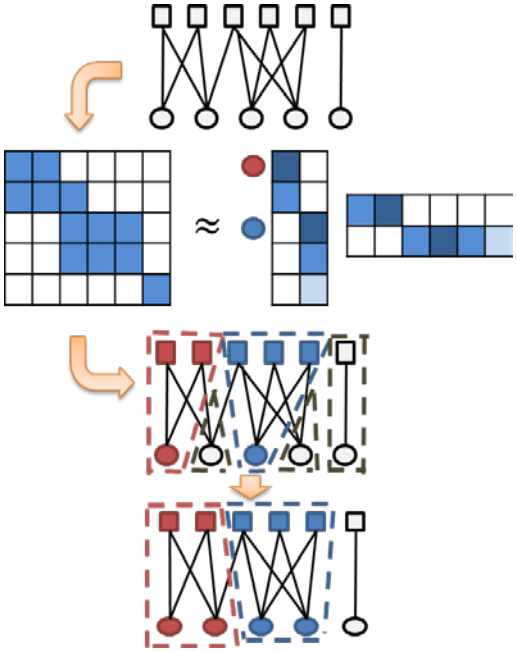
- サイトは、そのサイトを訪問したユーザの集合の類似度でクラスタリング

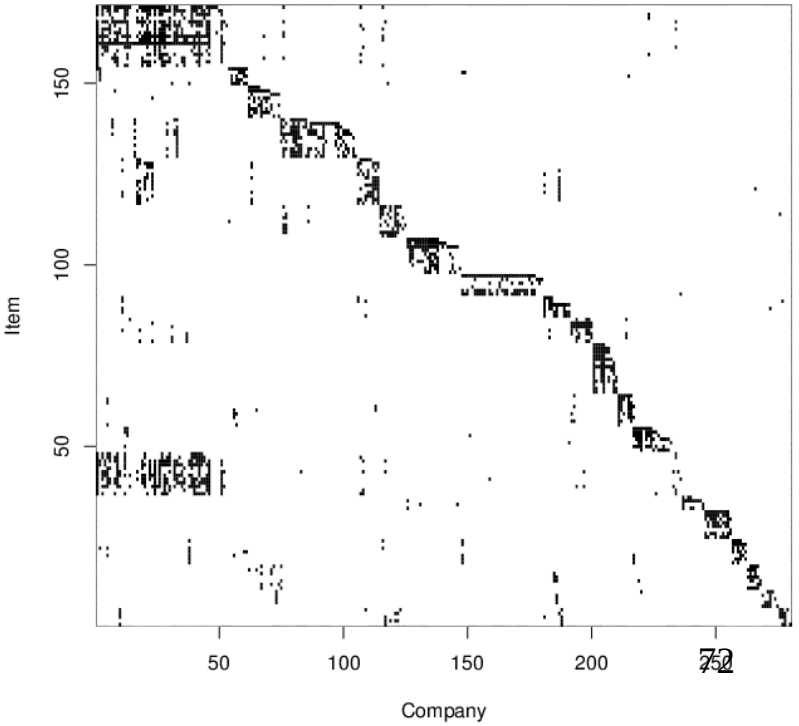
化粧品広告主A（無料サンプル配布案件;初回訪問コンバージョン率80%）におけるCPA



Finding dense subgraphs and Bid rigging
密部分抽出と地方公共団体競争入札

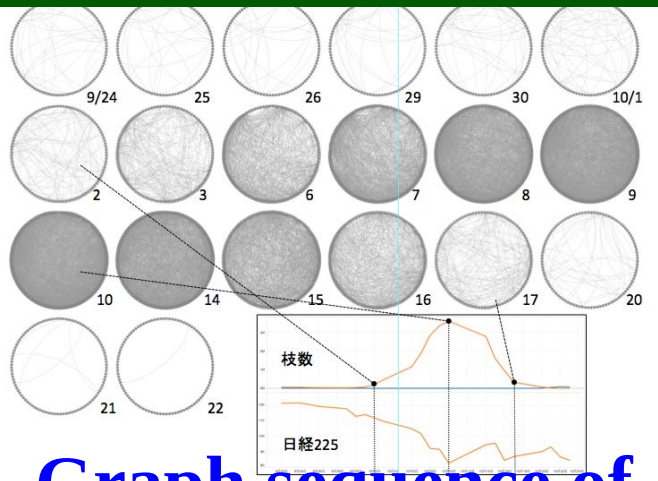
- Extracting particles from real big data
  Chiba-city fiscal 2005 (172 items, 276 companies)
  to finding sets of items bid by groups of companies
  and groups of companies bidding a set of items simultaneously
- Combination of NMF for matrix computation and merging clusters according to “density first search”



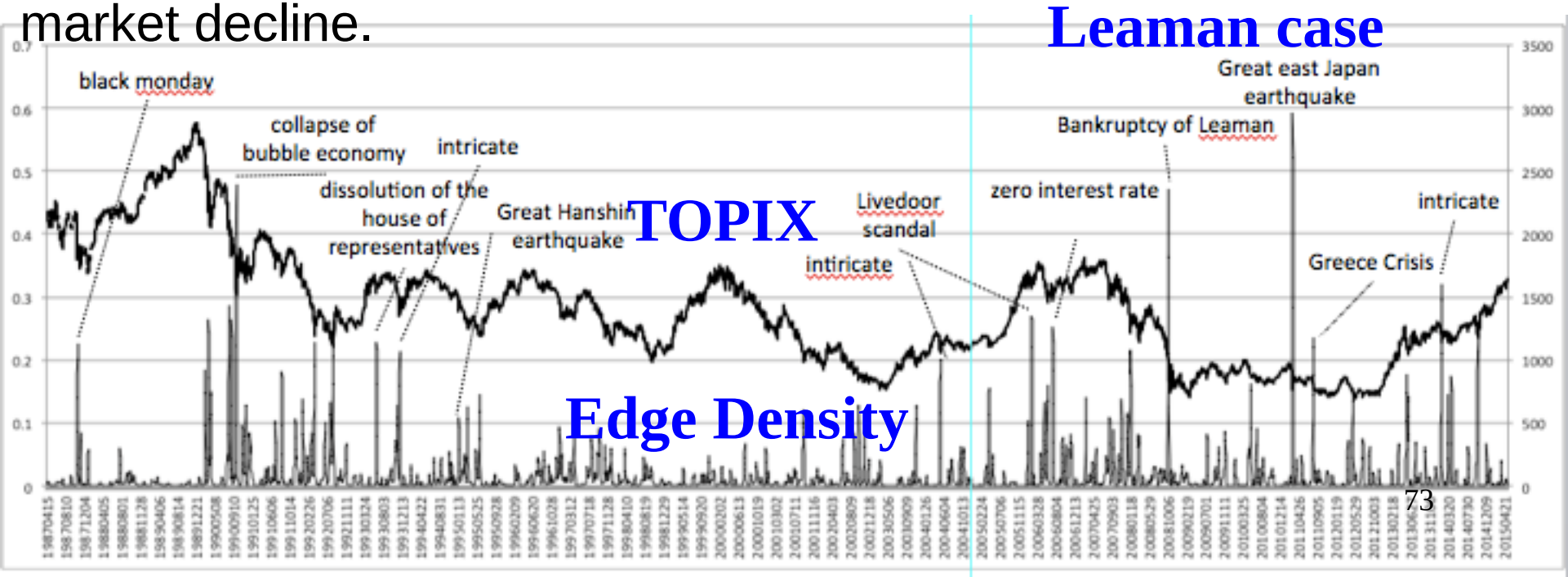


Herding in Finance

- Model building of "herding" phenomenon using individual stock price data.
- Representation by graph sequence
- Similarity graph with pairs of equities having price co-movement
- Edge density predict major bottoms of market decline.



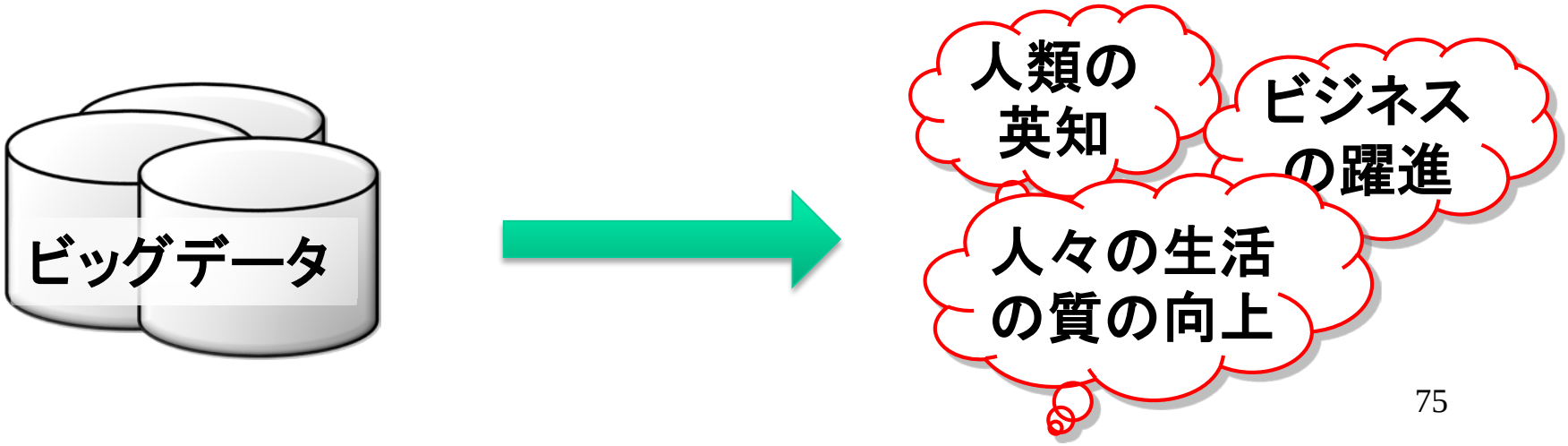
Graph sequence of Leaman case



これからのデータ解析：ある展望

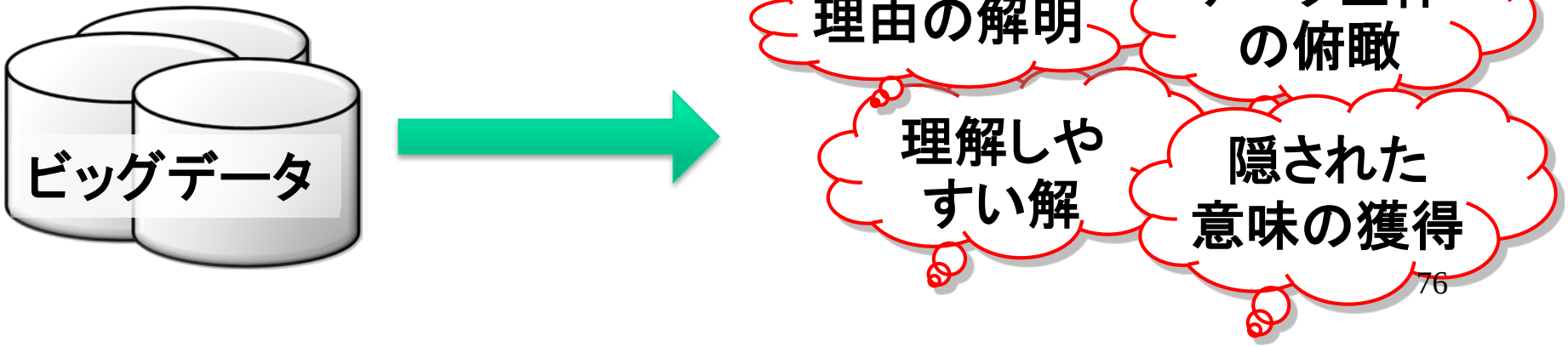
ビッグデータ時代の抱える問題点

- 猫も杓子もビッグデータ。これで世の中は大きく変わる！
- しかし、実際には箱物が多い
  ← システム導入、データ獲得、概念論、、、
- 新たな知識、感覚としての質の向上を報告したものは少ない
- ビッグデータは正しく使われているのか？



データ解析2.0

- データから物事と現象を深く理解できるようになる
- 箱物とブラックボックス技術に皮をかぶせる仕事から、
  データの意味を知り、価値を創造する仕事へ
- 小さな活動をする個人事業主やNPOもデータ解析
- 個人が自身の主観に基づく理解と解析結果を発信し、その集合知として社会が持つ「データに対する認識」が決まる時代



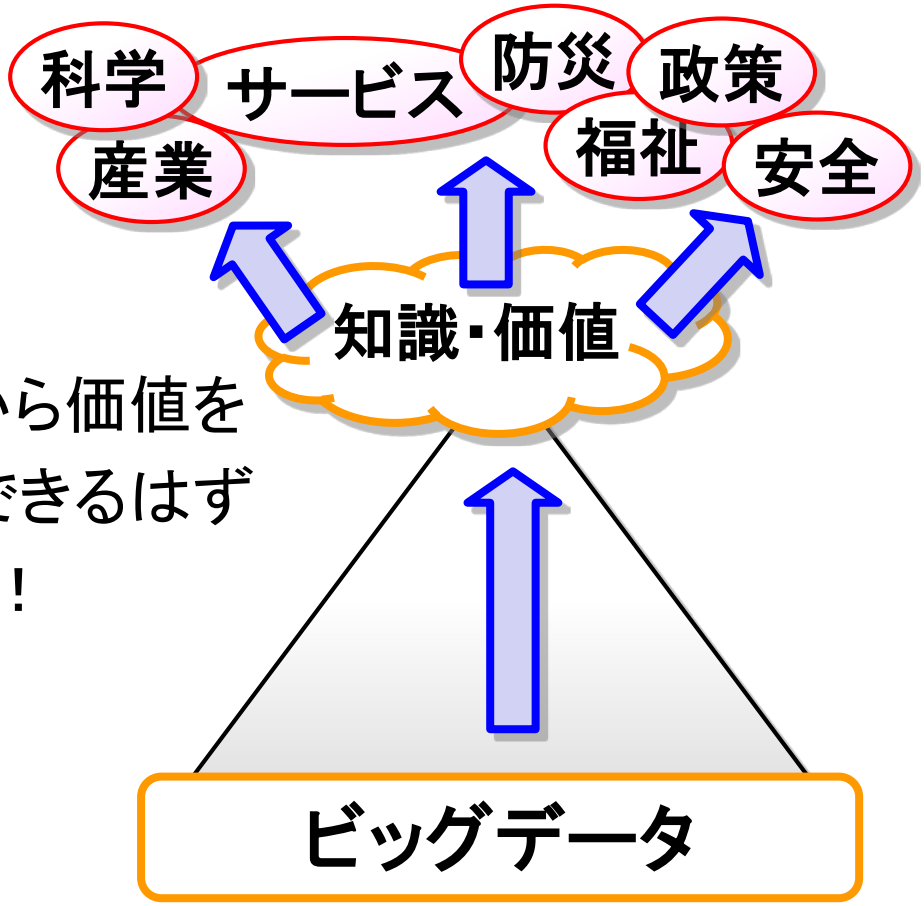
今、足りないもの

- データ解析をする、そして最終的には、価値を生み出したい
  ← 価値は**主観**によって定義される！

- 自分事として主観的に価値を創出する人が不足している

- 当事者なら、素人でも、データから価値を生み出す**「プロセス」**の設計はできるはず
  → 1つのデータから100の価値！

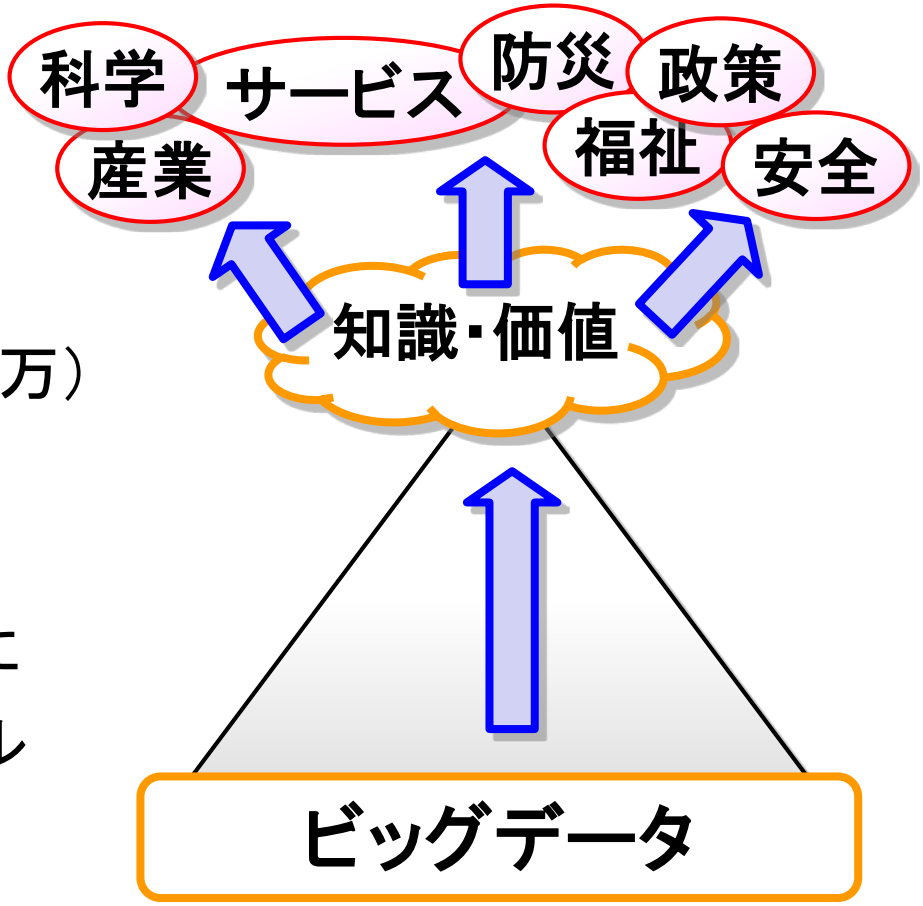
- データサイエンティスト不足が原因とは、とても思えない



世界の中の日本

- 世界を見ても、1つのデータから100の価値が生まれているとは思えない　← やはり主観不足

- ただ、欧米は市場のサイズに大きな有利さがある
（資金：　日本500万、アメリカ5000万）



- ただし、主観を生む現場の強さにおいては、日本は世界トップレベル

- 本来は、世界を牛耳っていてもいいはず。それを助ける技術開発がとても重要

参考文献

AI、データ社会をどう考えればいいか、知りたい人

情報研シリーズ　しっかり知りたいビッグデータとAI

今どきの深層学習などの技術解説ではなく、学者がやっていることでもなく、一般の人が今出現しつつあるAIを、どう見ればいいのか、という解説

ご紹介した、高性能クラスタリング技術

「データ研磨」、「データ粒子化」で検索すれば、出てきます

プロジェクトHP　http://research.nii.ac.jp/~uno/CREST/

データ研磨によるクリーク列挙クラスタリング

データ粒子化とデータ研磨を用いた未来のデータマイニング

参考文献

ご紹介した、超高速パターンマイニングアルゴリズム

頻出パターン発見アルゴリズム入門 アイテム集合からグラフまで

http://research.nii.ac.jp/~uno/papers/0806AIlecture.pdf

宇野毅明と有村博紀による公開プログラム（コード）

http://research.nii.ac.jp/~uno/codes-j.htm

列挙学校：第3コマ （スライド）

www.lab2.kuis.kyoto-u.ac.jp/keisan-genkai/reports/2007/enumeration/arimura.pdf