多項目データの因果関係を説明可能な人工知能データ解析技術

京都大学　大学院医学研究科
人間健康科学系専攻
特定准教授　玉田　嘉紀

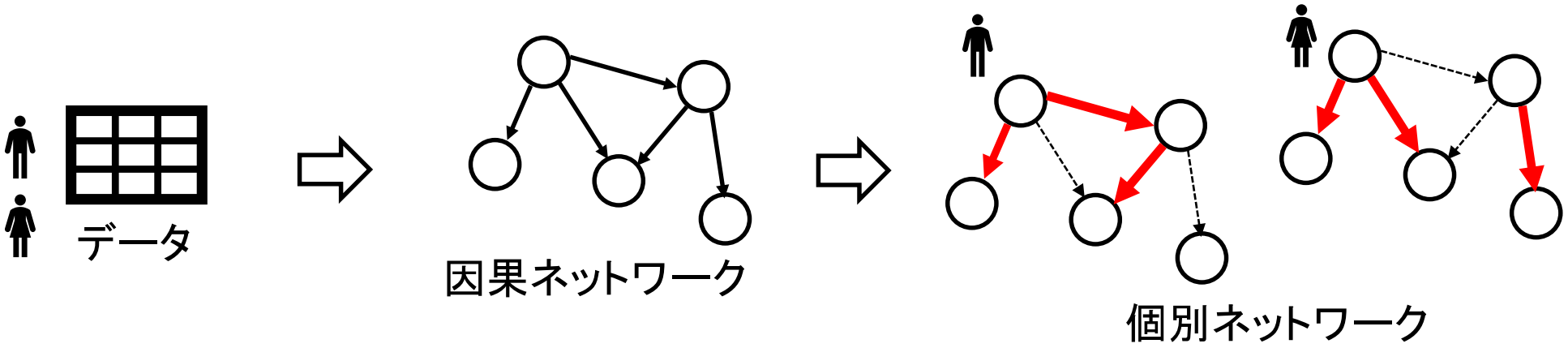
令和2年7月21日

社会的背景

- 大量のデータが日々蓄積される「ビッグデータ時代」では、データから自動的に知識やルールを獲得し、予測や分析を行うことができるAI（人工知能）技術が求められています。
- 深層学習（ディープラーニング）に代表される人工知能技術は高精度に予測が可能ですが、いわゆる「ブラックボックス」と呼ばれ、機械による判断の理由を説明することが困難です。
- また従来のAI技術はデータに共通の特徴をとらえるため、個別のデータサンプル（個人）の解析ができません。
- 本技術は多項目間の因果関係を自動で学習・推定し、人間が理解できるネットワーク図として提示できるだけでなく、個々のデータサンプルがどのように解釈されうるかを図示できる新しい技術です。

概要

- 多項目データから推定・学習した項目間因果ネットワークを用いて、個人ごと・サンプルごとの項目間の因果関係をネットワーク図として説明可能な人工知能技術です。



- 事前に学習した因果ネットワークモデルがあれば、単一のサンプル（ケース）あるいはごく少数のサンプルのデータからでもネットワークを図示することが可能です。

概要 （続き）

- 因果ネットワークのモデルとして *B*-Spline ノンパラメトリック回帰などを用いた連続値型ベイジアンネットワークを用います。項目（変数・属性）の値の型（離散・連続）を問いません（従来技術）。

- 新技術「枝貢献量(ECv)」により、個々のサンプル（ケース）がモデルの中でどのように使われるか定量化することができます。これにより、サンプルごとの因果ネットワークを用いた個別データの解析を可能とします。

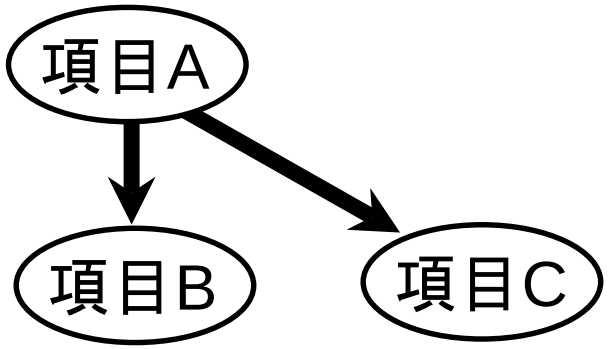
従来技術の解説1：ベイジアンネットワーク

ベイジアンネットワークは項目（変数）間の確率的な因果の関係性を、変数を表す点（ノード＝変数）とそれらを枝（エッジ）で結んだ**ネットワーク**で表現したものです。

（数学的には、変数間の条件付き独立性を仮定することで、全変数の同時確率が局所確率の積で表すことができ、それをネットワークとして図示できる、というモデルになります。）

例

「項目 A が、B と C の原因である」
という関係性を表すベイジアンネットワーク



ネットワークによる表現

$$\underbrace{\Pr(A,B,C)}_{\text{同時確率}} = \Pr(A) \times \Pr(B|A) \times \Pr(C|A)$$

数学的表現

この例では $\Pr(B,C|A) = \Pr(B|A) \times \Pr(C|A)$（条件付き独立）を仮定している

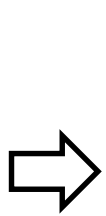
従来技術の解説2：ベイジアンネットワーク

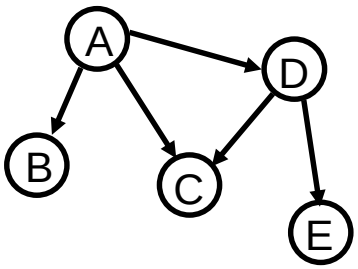
- 多項目の計測されたデータを用いて、そのデータに適合するベイジアンネットワークを推定・学習することができます。つまりデータから項目間の関係性（因果関係）を推定・学習できます。

- ベイジアンネットワークは、教師なし機械学習（AI:人工知能）技術の一つです。学習されたモデルがネットワーク（数式）で陽に表されることから、人間がモデルを見て理解できる「説明可能AI技術」の一つとされています。

	項目A	項目B	項目C	項目D	…
サンプル1					
サンプル2					
…					

多項目入力データ



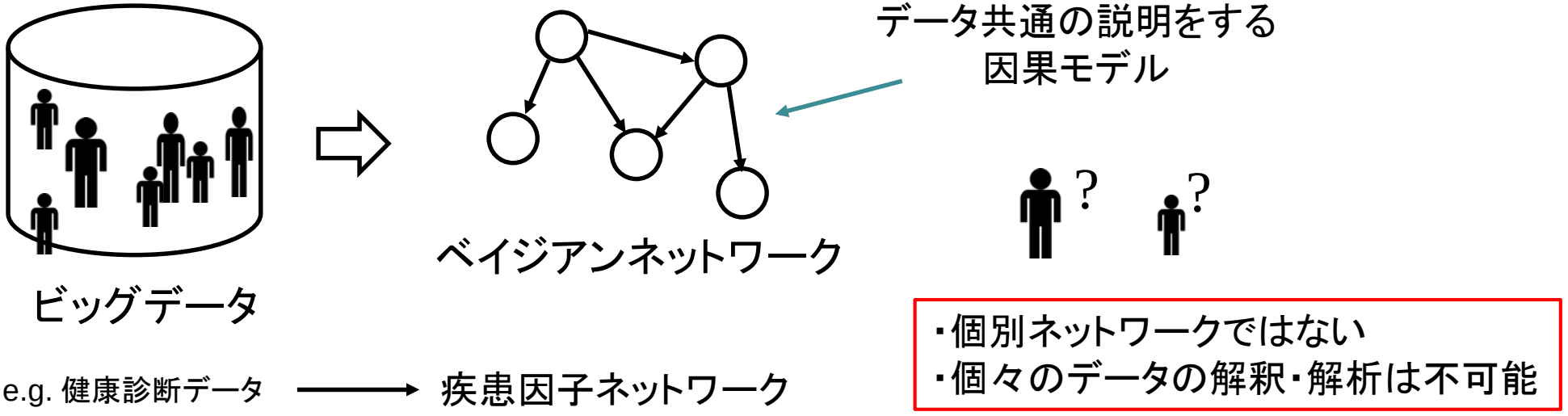


ベイジアンネットワーク

従来技術の問題点1

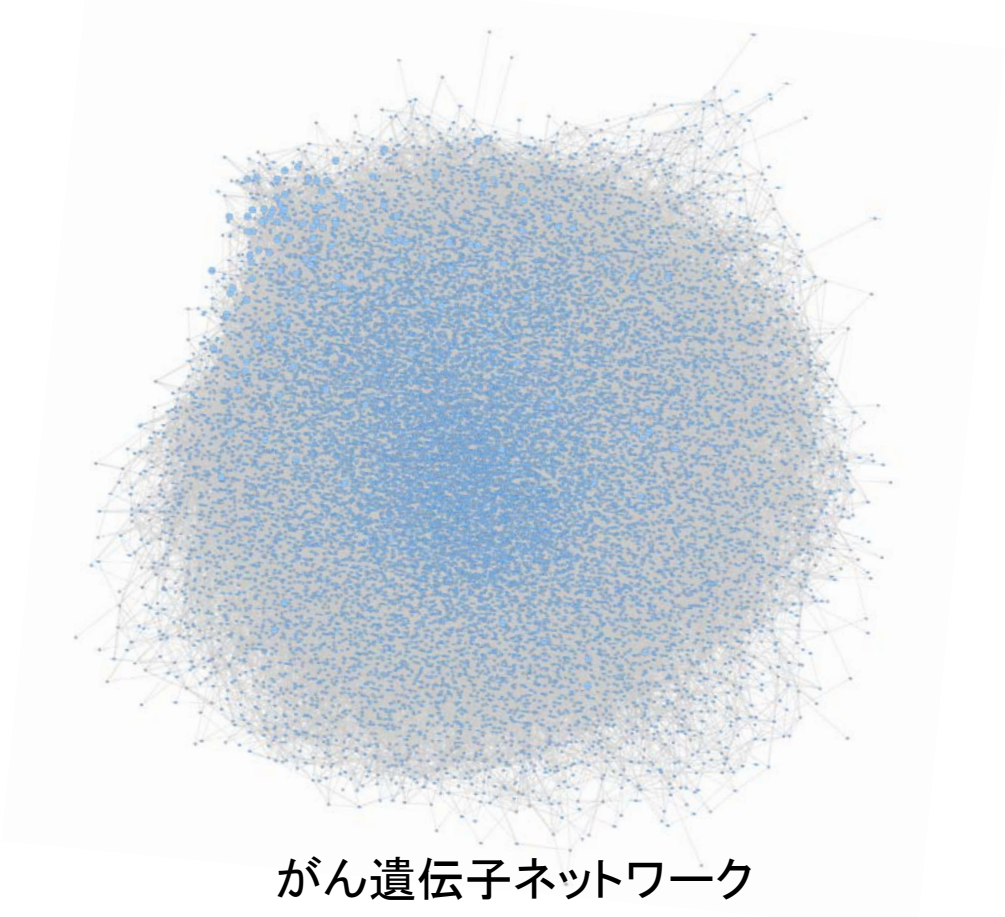
- 従来技術であるベイジアンネットワーク推定は多項目データからそのデータに含まれるサンプル(ケース)に共通する項目間の関係性を推定・学習するものです。

- したがって、個々のサンプル(ケース)についてのネットワークを得ることや評価・解析ができません。



従来技術とその問題点2

- また、超多項目のデータからでは、「毛玉」のようなネットワークが得られるだけで、そこから個々のサンプルや特定の条件間で意味のある重要なネットワークがどの部分かがわかりません。



がん遺伝子ネットワーク

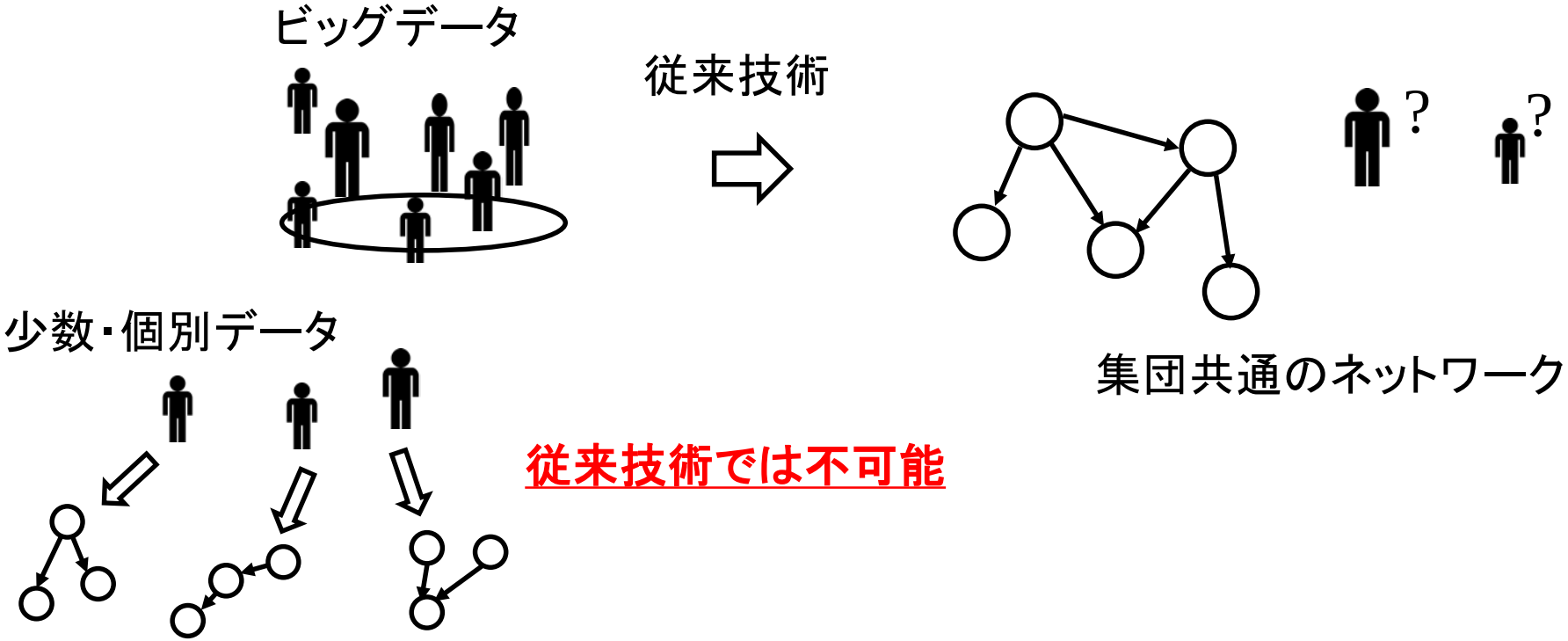
公開がん遺伝子発現データから推定した遺伝子間の発現制御の依存関係を推定した「遺伝子ネットワーク」の実例

このネットワークは 19,849 遺伝子（ノード・項目）からなり、推定された枝は **154,369本** である。

これを解釈・解析することは非常に困難。

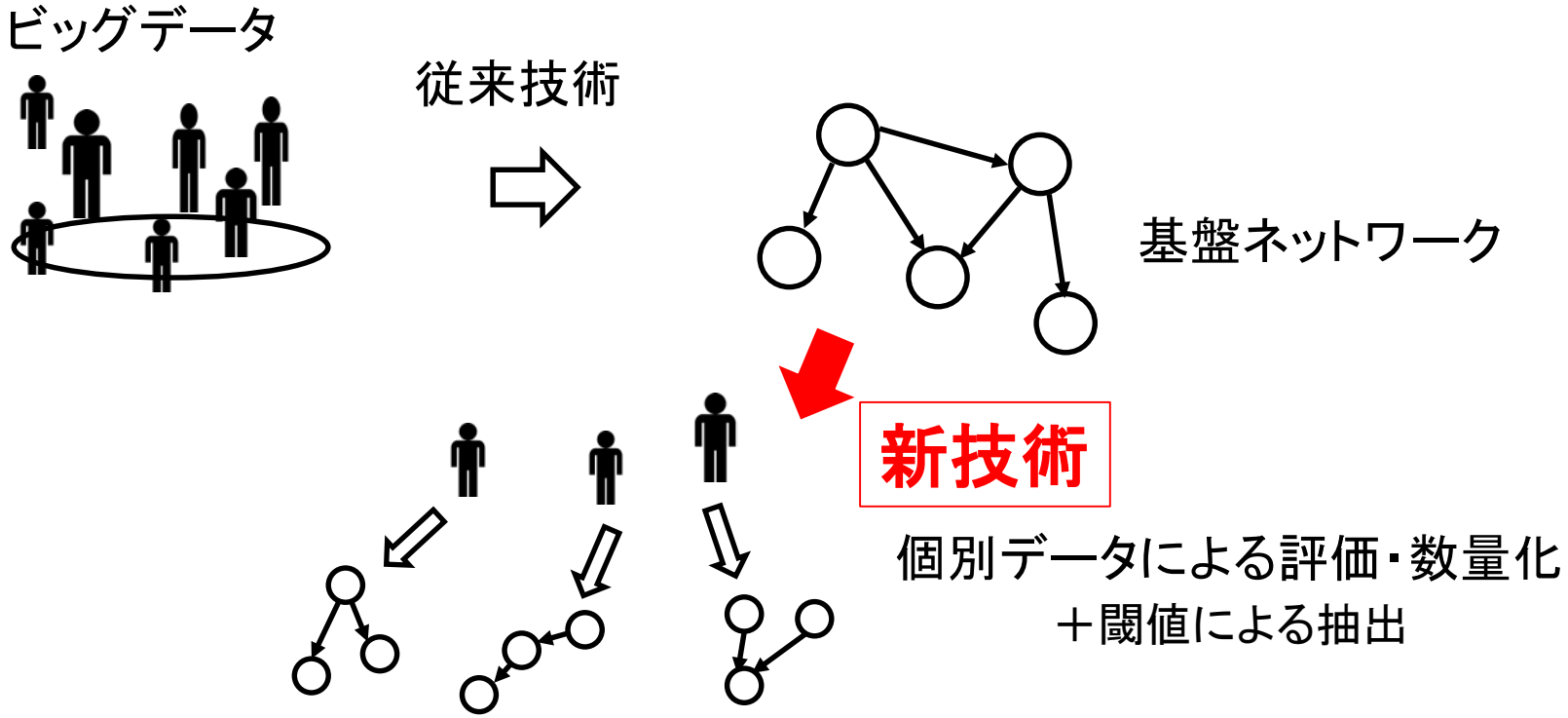
従来技術とその問題点3

- また、ネットワーク構造の推定・学習に大規模のデータが必要であり、少数のサンプル（ケース）しかない場合では、そもそもネットワーク解析・因果分析は不可能でした。



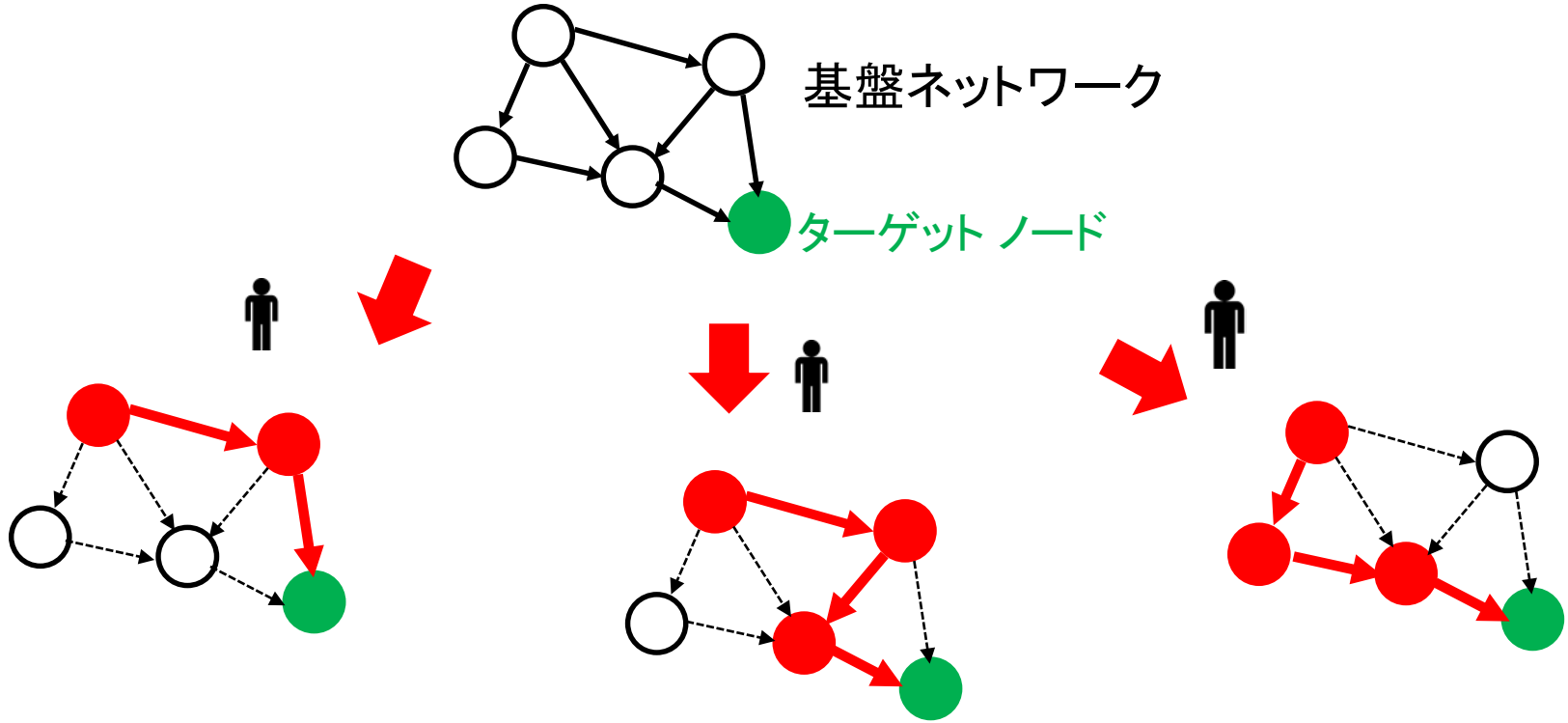
新技術の特徴1

- 個々のサンプル（ケース）について、項目間の各関係性（枝）ごとにその重要度を定量化できます（＝枝貢献量：ECv）。
  ※ネットワーク自体は同一または他のデータから推定する必要があります。



新技術の特徴2

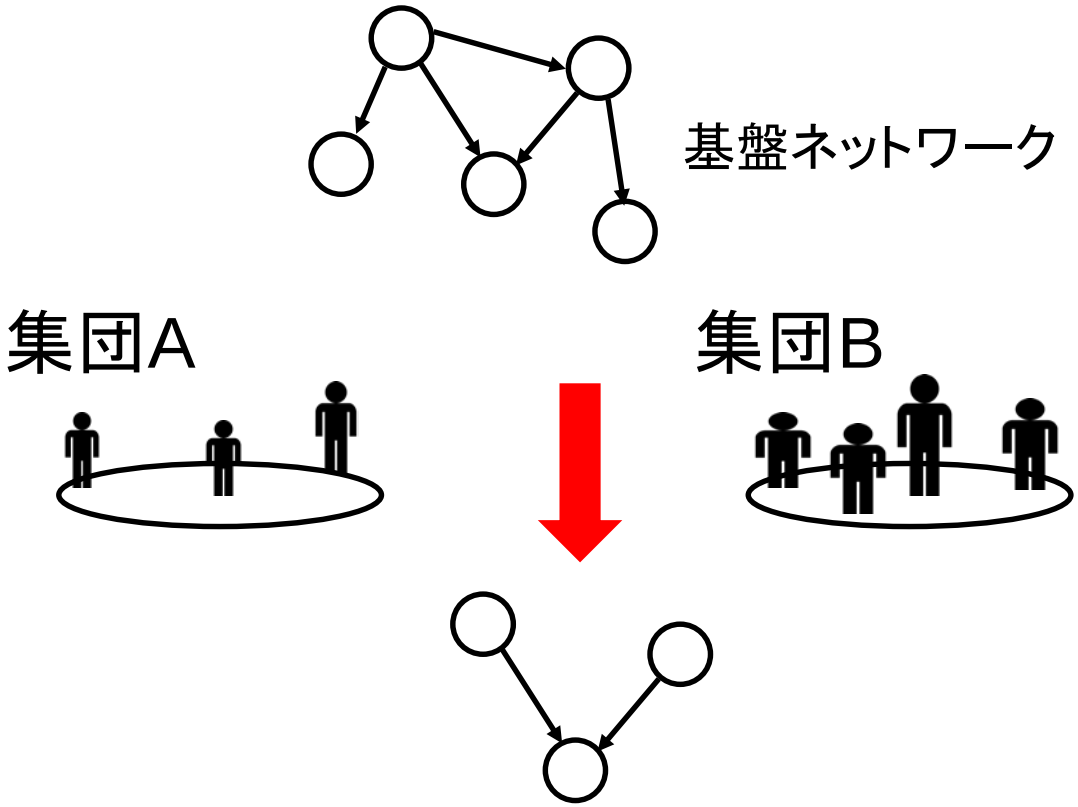
- サンプルごとに、特定ノードへの因果パス（経路）を推定できます（＝パス貢献量）。



サンプル（人）ごとにターゲットへ至る因果の経路は異なる

新技術の特徴3

- 特定のサンプル（ケース）間の比較により、差のある（サブ）ネットワークを抽出・定量化できます。

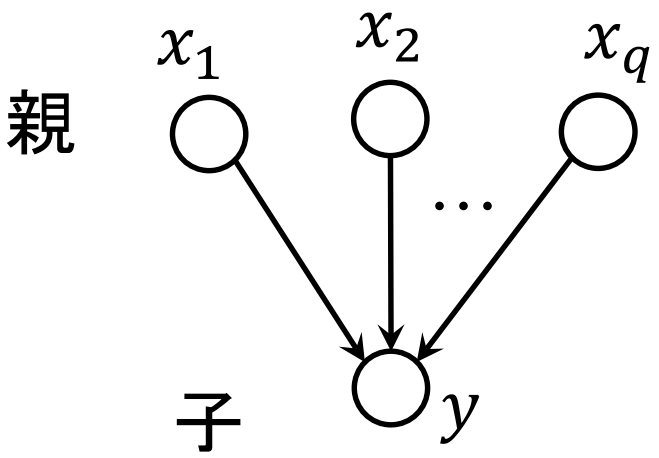


集団の分け方:
疾患の有無
薬剤の適応
予後の良悪
etc

AとBの違いを特徴付けるサブネットワークを抽出

新技術の技術的解説1：枝貢献量 ECv

回帰モデルによる連続値型ベイジアンネットワーク



ある子変数 y は次のように表わされる

$$y = m_1(x_1) + m_2(x_2) + \cdots + m_q(x_q) + \varepsilon$$

$m_k(\cdot)$: *B*-Spline 曲線など

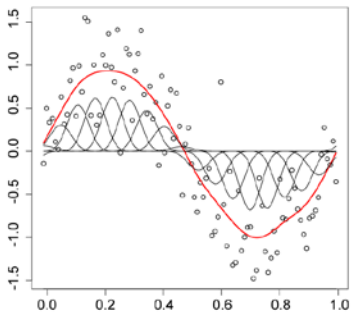
q: 親の数

ε: ノイズ項

つまり、子 y を、その親の値を関数 $m(\cdot)$ で変換した値の和で表す。

B-Spline による $m(\cdot)$ の例

$$m(x) = \sum_{l=1}^{M} \gamma_l b_l(x)$$



係数

B-spline 曲線

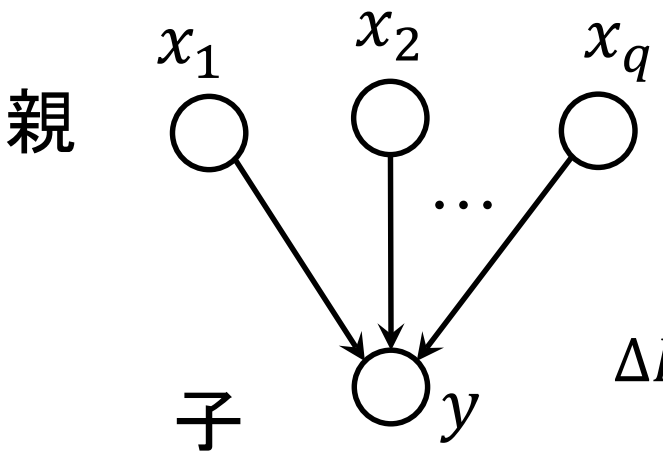
定義

x_j の y への枝貢献量 ECv (Edge Contribution value) を次のように定義する:

$$ECv(x_j \to y) = m_j(x_j)$$

→枝のサンプルごとの定量化

新技術の技術的解説2：ΔECv

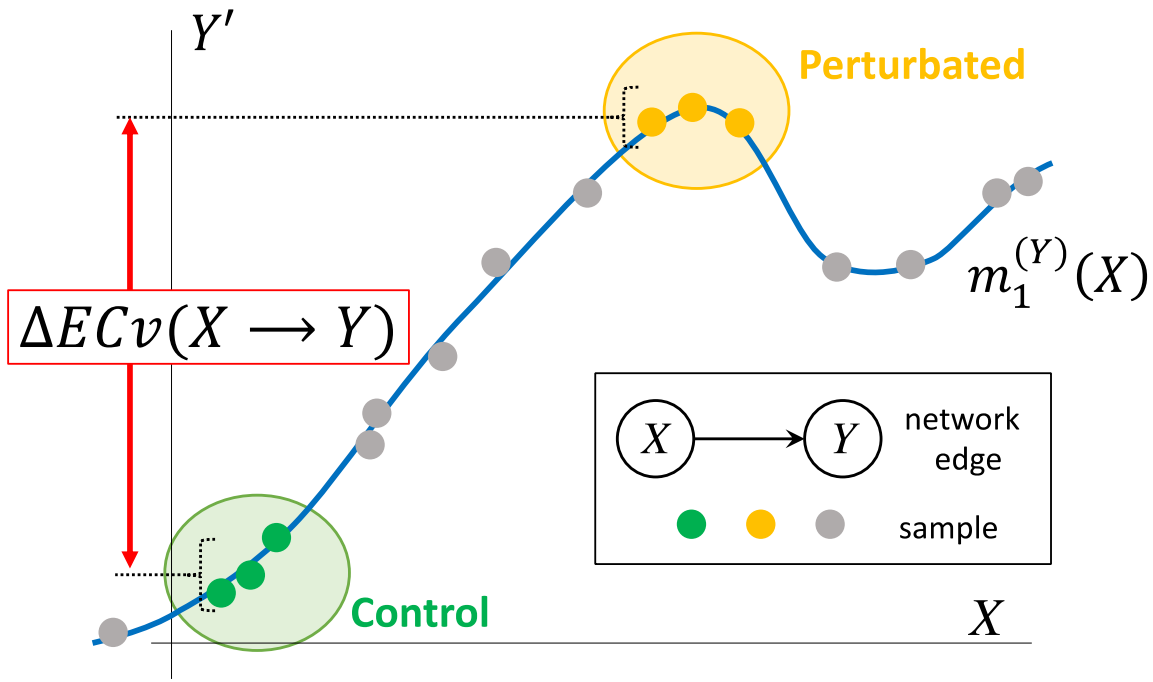


2サンプル A と B（e.g. コントロールと対象サンプル等）の ΔECv（デルタECv）を次のように定義する。

$$\Delta ECv\left(x_j \rightarrow y, A, B\right) = \left|ECv\left(x_j^A \rightarrow y^A\right) - ECv(x_j^B \rightarrow y^B)\right|$$

ただし、A, B はサンプル番号、x^AはサンプルAでの変数xの計測値

新技術の技術的解説2：ΔECvの図的説明



$$Y = m_1^{(Y)}(X) + \cdots$$

枝 $X \longrightarrow Y$ に対するΔECvの図示。
青線 ～ ：Y に対する親 X の回帰モデル $m_1^{(Y)}(X)$

この図では Control サンプル群（緑点）と Treated サンプル群（黄点）の二つの
サンプル集合間のΔECvを赤線の長さで表している。
個人（1つのサンプル）も同様に全サンプル平均や他のサンプルに対するΔECvを計算可能

新技術の技術的解説3：RCとRCr

ECvを用いた相対的な枝の値を以下のように定義する

x_jのyへの相対貢献度 RC (Relative Contribution) を次のように定義する：

$$RC(x_j \to y) = \frac{|m_j(x_j)|}{\max_{0<k\leq q} |m_k(x_k)|}$$

※$RC(x_j \to y)$は0から1の値になる

yを決めるモデル上の重要因子を定量化可能

x_jのyへの相対貢献率 RCr (Relative Contribution) を次のように定義する：

$$RCr(x_j \to y) = \frac{|m_j(x_j)|}{\sum_{k=1}^{q} |m_k(x_k)|}$$

※$RCr(x_j \to y)$は0から1の値になる

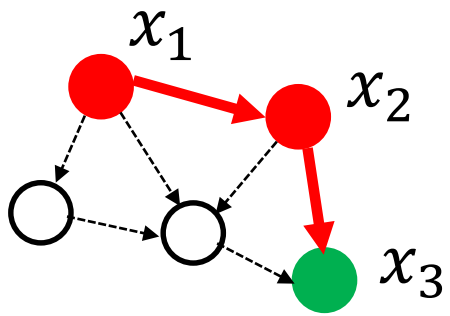
新技術の技術的解説4:パス貢献量

サンプルごとの因果パス(経路)の「パス貢献量」を RCr を用いて
以下のように定義する

$$PC(x_1, x_2, \ldots, x_n) = \sqrt[(n-1)]{\prod_{i=2}^{n} RCr(x_{i-1} \rightarrow x_i)}$$

(パス上の枝のRCr値の相乗平均)(RCも可)

例



パス $x_1 \rightarrow x_2 \rightarrow x_3$ のパス貢献量:

$$PC(x_1, x_2, x_3) = \sqrt[2]{RCr(x_1 \rightarrow x_2) \cdot RCr(x_2 \rightarrow x_3)}$$

新技術の技術的解説5：適用可能データサイズ

ECvやΔECvの計算自体は PC で十分に対応可能です。

ベイジアンネットワーク推定は計算負荷が高く、データサイズ・使用アルゴリズムに応じて必要な計算リソースが異なります。

時系列データも適用可能です。

HC＋BSアルゴリズム：
- 2000変数程度までのベイジアンネットワークを高精度に推定するアルゴリズム
- 現行のワークステーションで計～3000時間程度。
- （サンプル数が～3000程度を想定）
- 比較的大規模のPCクラスタなどが必要

NNSR アルゴリズム
- 20000変数程度までのベイジアンネットワークを高速に推定するアルゴリズム
- 64 コア程度のシステムで～24時間
- （サンプル数が～500程度を想定）

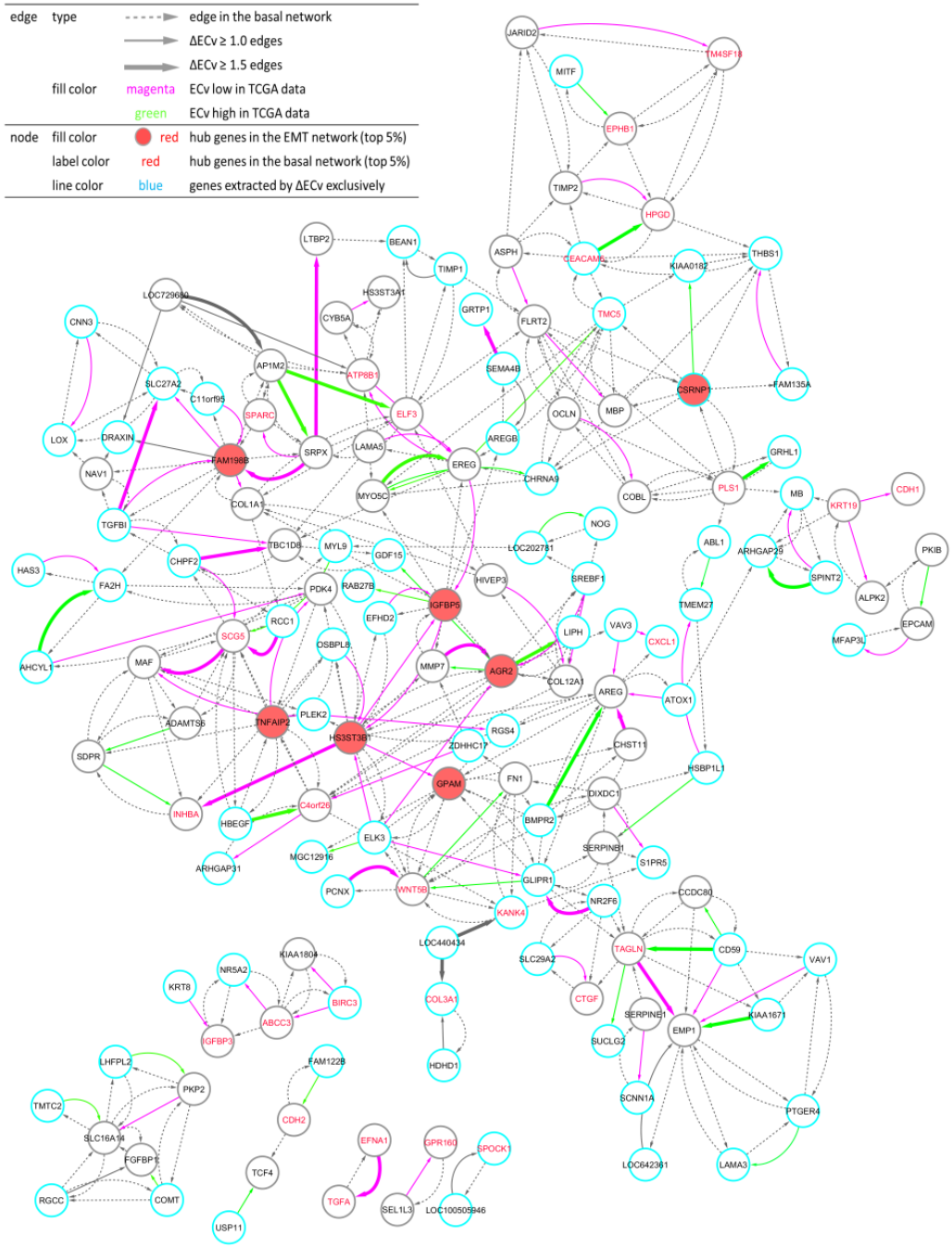
想定される用途

- 基本的に多項目のデータであれば、離散・カテゴリカル・連続値を問わず混在した項目間のベイジアンネットワークを推定し、サンプルごとのネットワーク解析が可能です。

- これまで適用した解析事例
  - 遺伝子発現データを用いた遺伝子ネットワーク推定・解析
  - 電子カルテ（主に血液検査）データを用いた急性腎疾患
  - 健康調査データ（アンケートやゲノムデータを含む）を用いた多疾患多項目ネットワーク・パス解析
  - etc

適用例：EMT 遺伝子ネットワーク解析

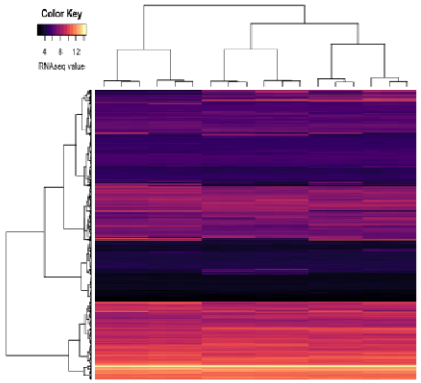
- ノード（変数）約2万
- 枝数30万程度
- 肺がんの培養細胞株の遺伝子発現データ
- がん転移に関連していると言われる EMT（上皮間葉移行）を誘導 9 したサンプルとControl9サンプルの 18 サンプル
- サンプル数が少ないため、それぞれのサンプル群でのネットワーク推定は不可能
- 18サンプルでネットワーク推定し、EMT誘導9サンプルとControl9サンプル間のΔECvを用いてサブネットワークを抽出
- 実際に EMT サンプル群を説明するサブネットワークを抽出することに成功



既発表論文：
Tanaka et al (2020) *Biomolecules*, **10**(2), 306

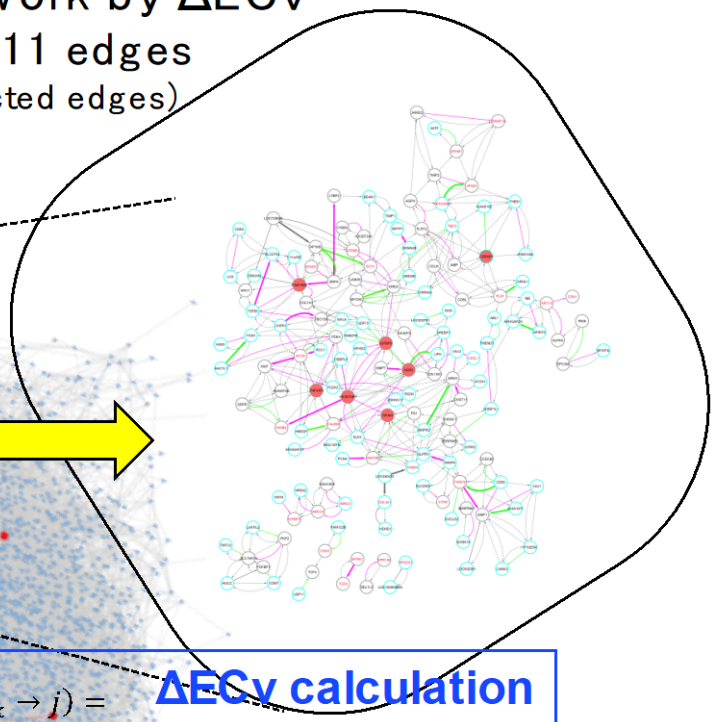
(2) ΔECvによるEMTネットワーク抽出

Control & EMT induced
gene expression data



19,849 genes 18 samples

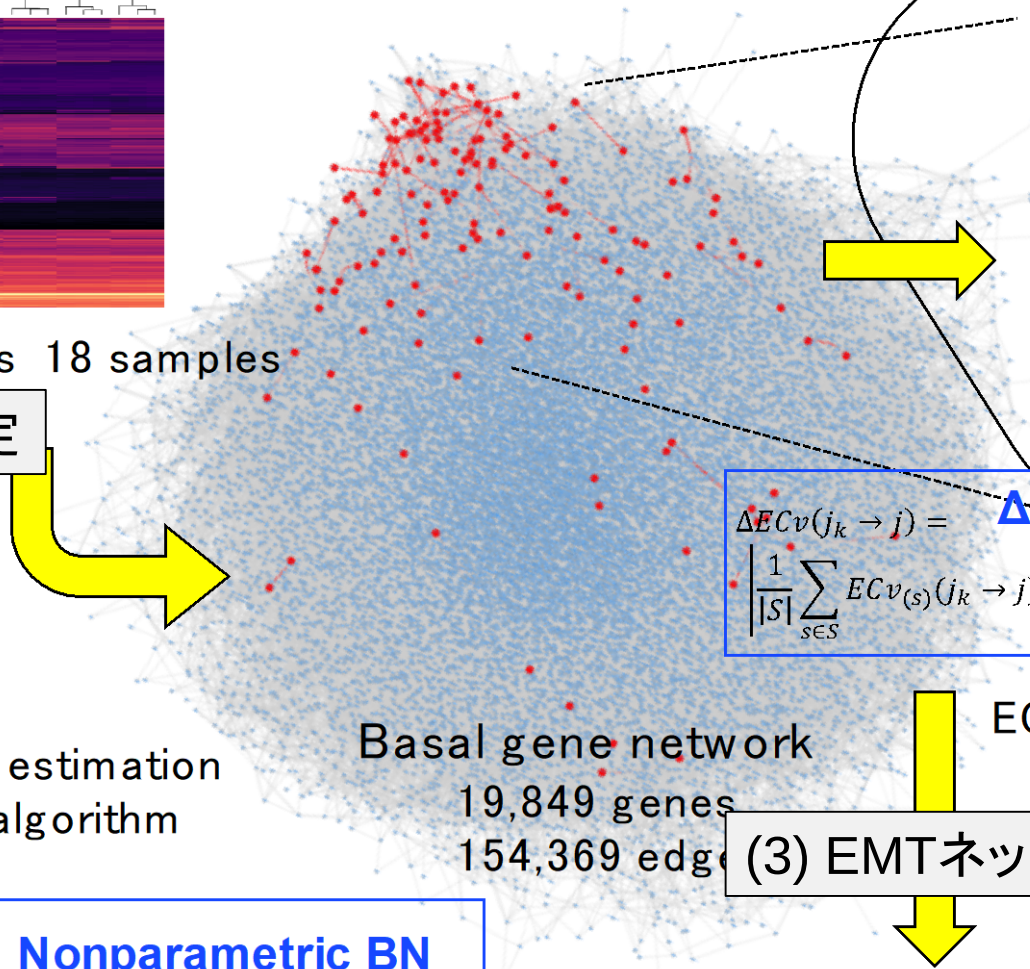
The EMT network by ΔECv
150 genes 411 edges
(incl. 120 selected edges)



(1) ネットワーク推定

Shirokane3
Supercomputer

Gene network estimation
by the NNSR algorithm

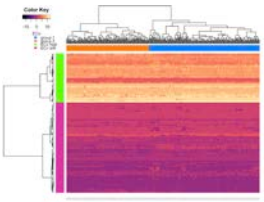


Basal gene network
19,849 genes
154,369 edge

ΔECv calculation

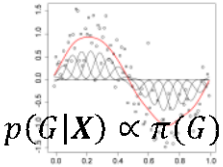
$$\Delta ECv(j_k \to j) = \left| \frac{1}{|S|} \sum_{s \in S} ECv_{(s)}(j_k \to j) - \frac{1}{|T|} \sum_{t \in T} ECv_{(t)}(j_k \to j) \right|$$

ECv clinical approach
to TCGA data



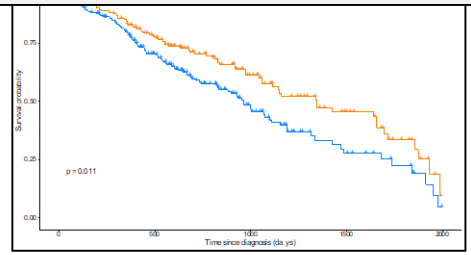
(3) EMTネットワークによる患者データ解析

Nonparametric BN



$$p(G|\boldsymbol{X}) \propto \pi(G) \int \prod_{i=1}^{n} \prod_{j=1}^{p} f(x_{ij}|pa_{ij}, \boldsymbol{\theta}_G) p(\boldsymbol{\theta}_G|\lambda) d\boldsymbol{\theta}_G$$

Survival analysis
by the EMT network



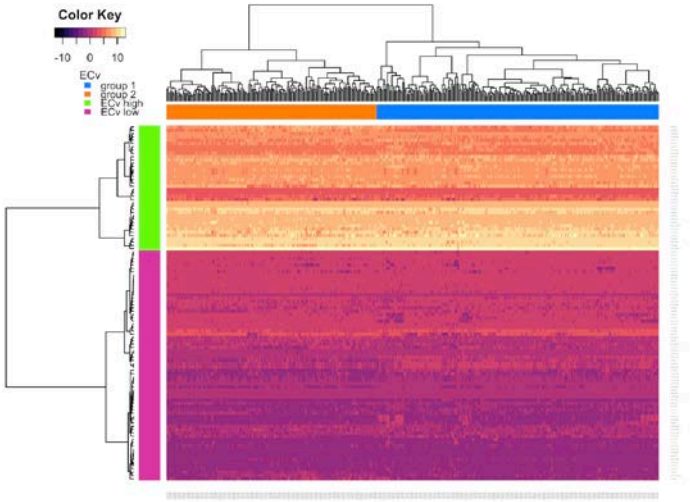
ECvパターンによる個人の特徴付け

ECv Matrix（ECv 行列）：　各行が（抽出した）ネットワークの枝
各列がサンプル（個人）の ECv

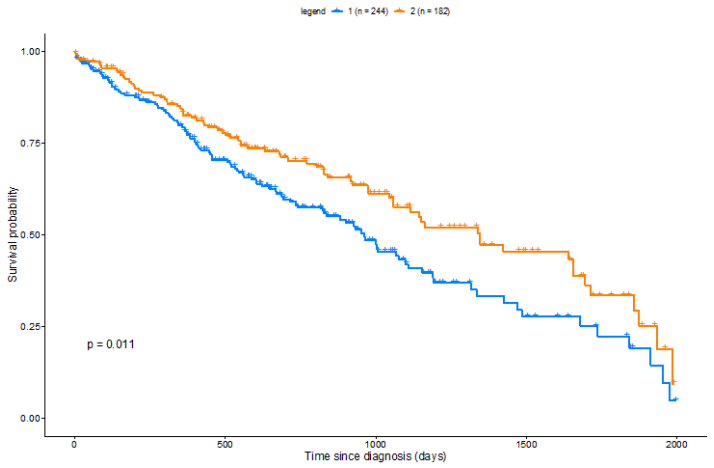
ネットワーク推定・枝抽出に使用したものとは**別のデータに適用可能**

- TCGA（ガン患者の公開データベース）の肺がんデータに適用後、クラスタリング。
- 2つのクラスタで予後（生存時間）に大きな差が出ることを実証。

個人ごとのデータの特徴づけ・分類が可能

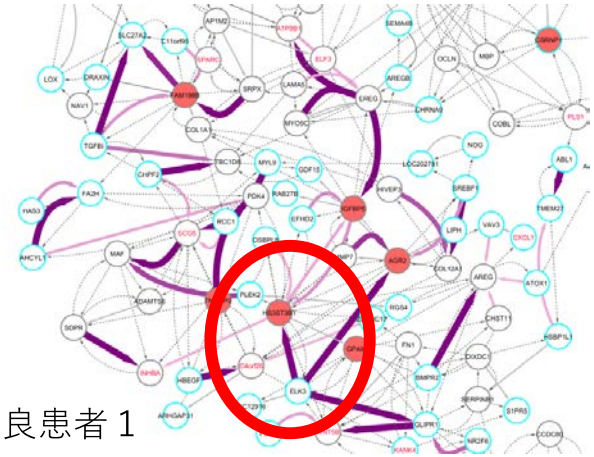


ECv行列のクラスタリングにより2群を同定

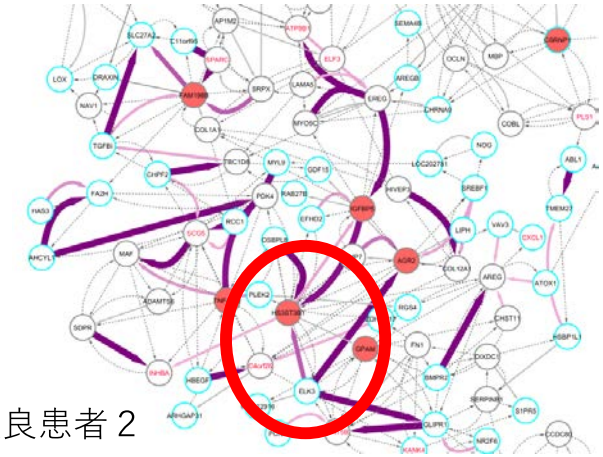


同定した2群間で有意な生存時間差を確認

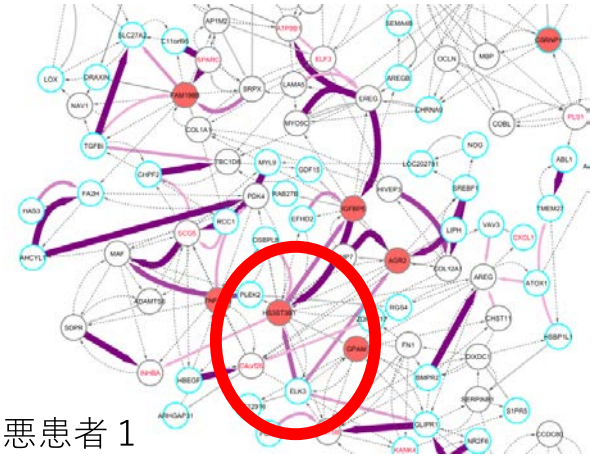
同定した予後良悪患者2名ずつの個別ネットワーク可視化・比較例



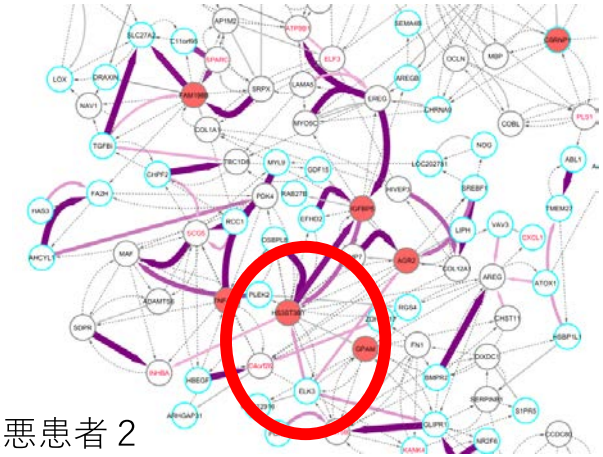
予後良患者１



予後良患者２



予後悪患者１



予後悪患者２

- 抽出したEMTネットワークを個人ごとに前述のRCを用いて可視化したものである。
- 強調部分は本ネットワークのハブ遺伝子 HS3ST3B1 への入力枝が予後良悪患者間で顕著に異なっていることを示している。この遺伝子の発現がEMTに影響を与えていることが文献上でも確認できることから重要遺伝子が正しく捉えられ、可視化も成功していることがわかる。

実用化に向けた課題

- 発現データ解析については解析事例が豊富ですでに実用的と言える。

- カルテデータや健康調査データを用いた個人疾患原因パス解析については現在データが揃いつつあるが、解析中・論文執筆中につき掲載を控えた。

- パス解析については、実用化に向けて、アノテーション付きデータによる検証が不足しているため現在研究を進めている。

本技術に関する知的財産権

- 発明の名称：特徴ネットワーク抽出装置、コンピュータプログラム、特徴ネットワーク抽出方法及びベイジアンネットワーク分析方法
- 出願番号　　：特願2020-002923
- 出願人　　：京都大学
- 発明者　　：奥野恭史、玉田嘉紀

お問い合わせ先

国立大学法人京都大学内
株式会社TLO京都
京大事業部門　技術移転チーム

TEL　075 － 753 － 9150
FAX　075 － 753 － 9169
e-mail　event@tlo-kyoto.co.jp