

3. 高次元データを理解する

主成分分析, t-SNE, UMAP, ...

梅本 和俊

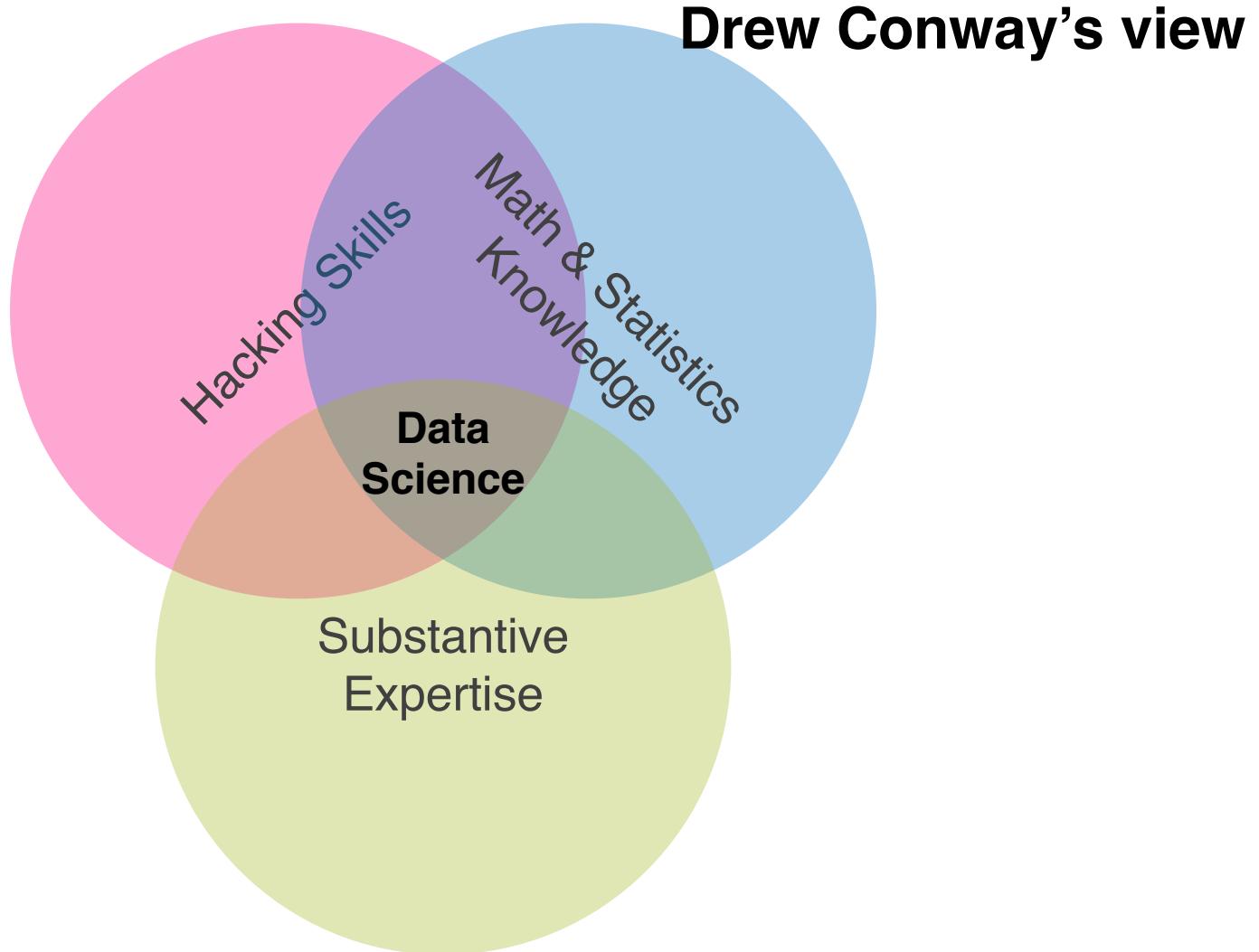
情報通信研究機構／東京大学
umemoto@tkl.iis.u-tokyo.ac.jp

データサイエンスが熱い



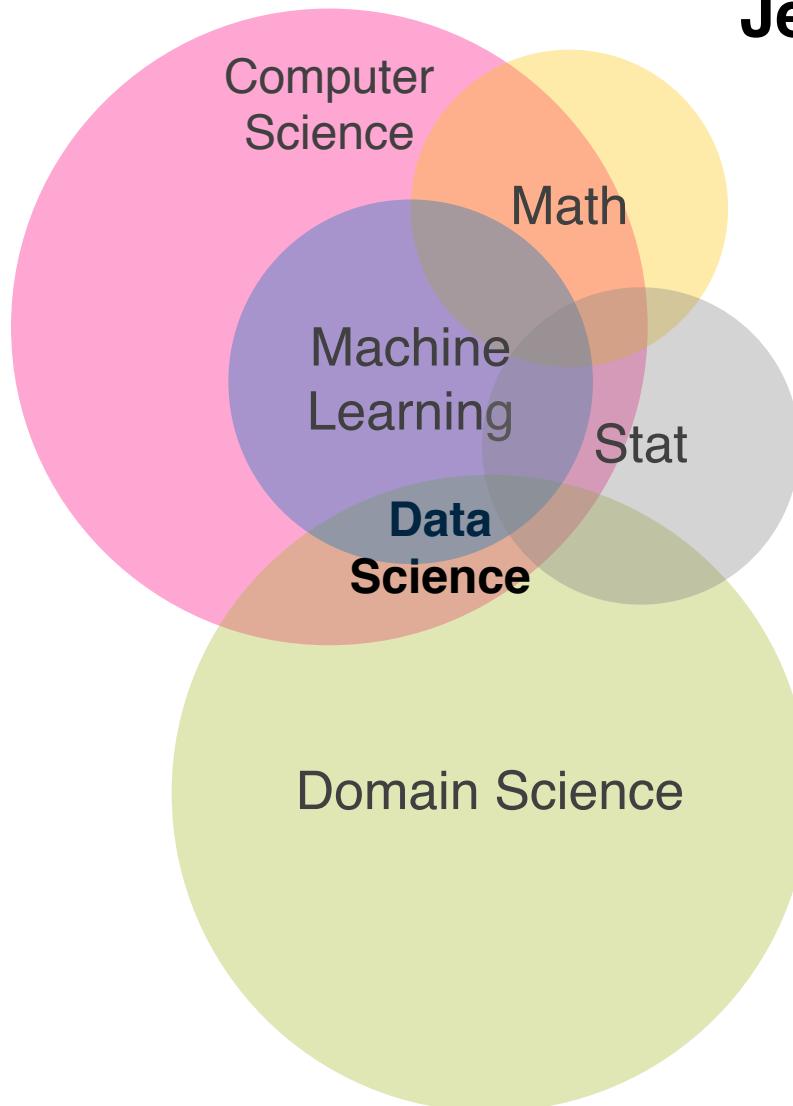
そもそもデータサイエンスとは？

そもそもデータサイエンスとは？



そもそもデータサイエンスとは？

Jeffrey Ullman's view



担当回

3. 10/21 (月)

- › 高次元データの理解: 主成分分析, t-SNE, UMAP, ...

6. 11/18 (月)

- › (仮) 隠れた値の推定: トピックモデル, クリックモデル, ...

8. 12/02 (月)

- › (仮) グラフデータの分析: PageRank, TwitterRank, 影響最大化

(今日扱う) データ

- 全体のデータ: $\mathcal{D} = \{x_n\}_{n=1}^N = \{x_1, \dots, x_N\}$
- 個々のデータ: $x_n = [x_{n1}, \dots, x_{nD}]$
- データの各次元の値: $x_{nd} \in \mathbb{R}$ (つまり実数値)

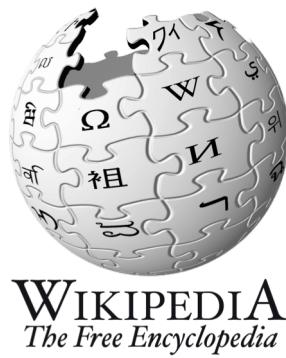


(ユーザ情報を無視するなら) $N = 4$

	ユーザID	国語	英語	数学	物理
D 4	1	80	90	60	40
	2	90	90	80	90
	3	60	40	70	80
	4	70	40	60	40

高次元データ is everywhere

文書データ（次元: 単語）



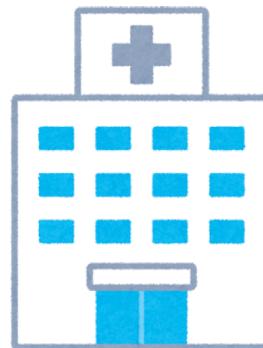
画像出典: https://en.wikipedia.org/wiki/Main_Page

画像データ（次元: 画像特徴量）



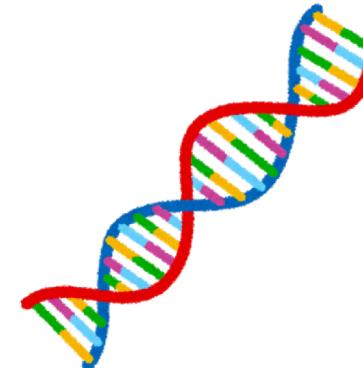
画像出典: <https://www.silhouette-illust.com/illust/15878>

医療データ（次元: 疾患, 医薬品, …）



画像出典: https://www.irasutoya.com/2012/12/blog-post_4261.html

ゲノムデータ（次元: 遺伝情報）



画像出典: <https://www.irasutoya.com/2015/04/dna.html>

高次元データを扱うことの難しさ

- データの傾向を人間が把握しづらい
 - › 人間で可視化で理解できるのは3次元まで
 - › 各次元をしらみつぶしに調べる？
 - › 似ていそうな次元やデータ点にどうやって当たりをつける？
 - › ...
- 学習
 - › 次元数が増えるほど一般的に学習に時間がかかる
 - › データ件数に比べて次元数が多すぎると学習できない
 - › 次元間に相関があると独立性を仮定するアルゴリズムは使えない

次元削減

- 入力: 膨大な数 D の次元からなるデータ $x = [x_1, \dots, x_D]$
- 出力: x を少数 D' ($D' < D$) の次元で表したデータ $y = [y_1, \dots, y_{D'}]$

● 成績理解

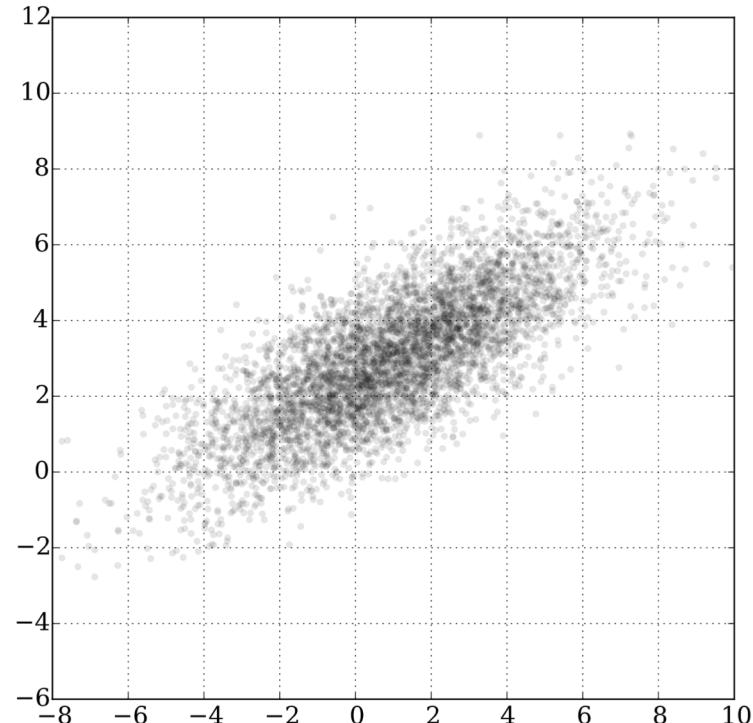
- › 入力: [国語の点数, 数学の点数, ...]
- › 出力: [総合成績, 文系／理系]

● 健康評価

- › 入力: [身長, 体重, ...]
- › 出力: [大きさ, 肥満度, ...]

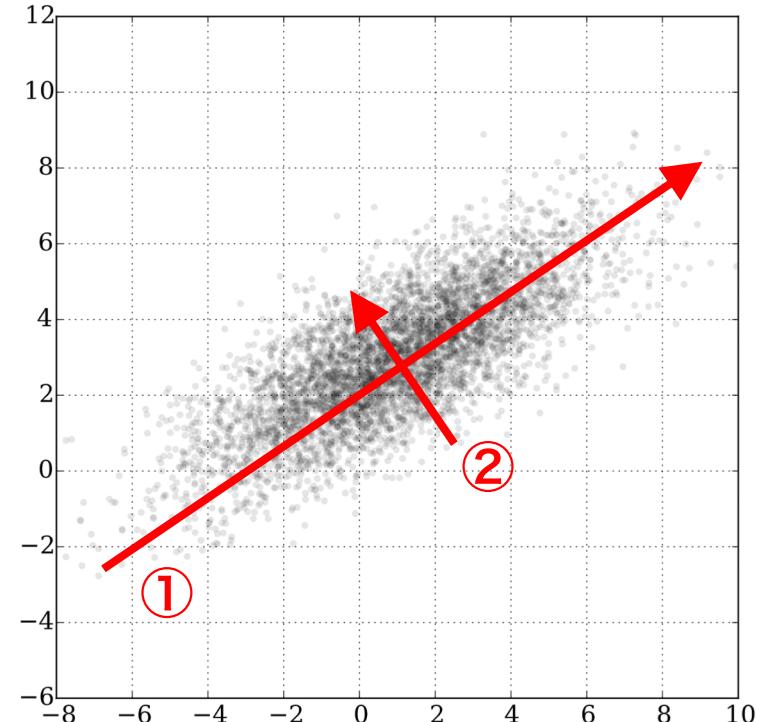
どうすれば？

1. 5つの次元（車, car, バイク, 図書館, 本）を無理やり2次元にするならどれをまとめる？
2. 右の2次元データを1次元に削減するならどこに軸を取る？
3. もう1つ軸を加えるならどこに取る？



どうすれば？

1. 5つの次元 (車, car, バイク, 図書館, 本) を無理やり2次元にするならどれをまとめる?
› (車・car・バイク, 図書館・本)
2. 右の2次元データを1次元に削減するならどこに軸を取る?
3. もう1つ軸を加えるならどこに取る?



主成分分析 | ポイント

- 新次元（主成分）を旧次元の線形結合で作る
 - › $y_{d'} = w_1x_1 + \cdots + w_Dx_D$
 - › 乗り物 = 0.6×車 + 0.4×car + 0.3×バイク + …
- 主成分間には相関がない (≒全く別の事柄を表現)
 - › OK: 「乗り物」成分, 「勉強」成分
 - › NG: 「乗り物」成分, 「飛行機」成分
- データの散らばりが大きくなるように主成分を選択
 - › 分散を最大化→情報量を最大化→情報損失を最小化
- 主成分の寄与率が分かる
 - › この新次元だけで全データの何%を説明可能か

主成分分析 | 計算方法 | 前処理

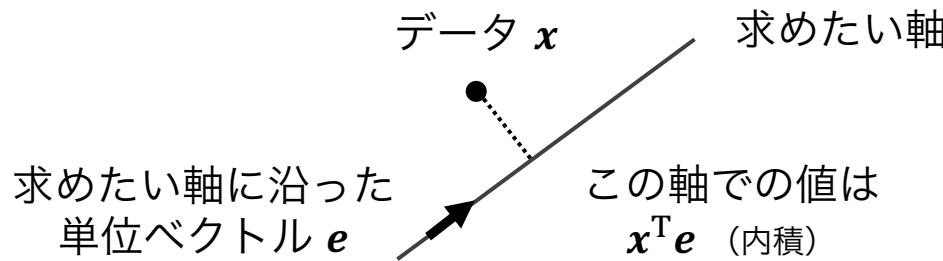
- 各次元に関して平均が0になるよう調整

$$x_{nd} \leftarrow x_{nd} - \bar{x}_d = x_{nd} - \frac{1}{N} \sum_{n=1}^N x_{nd}$$

- 調整後の全体データを行列 X と表現

$$X = \begin{pmatrix} x_{11} & x_{12} & \cdots & x_{1D} \\ x_{21} & x_{22} & \cdots & x_{2D} \\ \vdots & \vdots & \ddots & \vdots \\ x_{N1} & x_{N2} & \cdots & x_{ND} \end{pmatrix}$$

主成分分析 | 計算方法 | 軸が既知なら...



- 射影値の平均は前処理により

$$\frac{1}{N} \sum_{n=1}^N x_n^T e = \frac{1}{N} \sum_{n=1}^N \left(\sum_{d=1}^D x_{nd} e_d \right) = \frac{1}{N} \sum_{d=1}^D e_d \left(\sum_{n=1}^N x_{nd} \right) = 0$$

- なので分散は

$$\frac{1}{N} \sum_{n=1}^N (x_n^T e - 0)^2 = \frac{1}{N} (\mathbf{X}e)^T (\mathbf{X}e) = e^T \frac{\mathbf{X}^T \mathbf{X}}{N} e$$

$$\begin{pmatrix} x_1^T e \\ \vdots \\ x_N^T e \end{pmatrix} = \mathbf{X}e$$

分散共分散行列 (以降, V)

主成分分析 | 計算方法 | 分散の最大化

- 単位ベクトルのノルム制約を考慮した次式を最大化

$$J = \mathbf{e}^\top \mathbf{V} \mathbf{e} - \lambda (\mathbf{e}^\top \mathbf{e} - 1)$$

cf. ラグランジュの
未定乗数法

- J を \mathbf{e} で偏微分した値が最大時には0となるべきなので

$$\frac{\partial J}{\partial \mathbf{e}} = 2\mathbf{V}\mathbf{e} - 2\lambda\mathbf{e} = 2(\mathbf{V} - \lambda\mathbf{I})\mathbf{e} = \mathbf{0}$$

固有値問題

- λ は $\mathbf{X}^\top \mathbf{X}$ の固有値の1つ
- \mathbf{e} はそれに対応する固有ベクトル

- 分散に代入すると

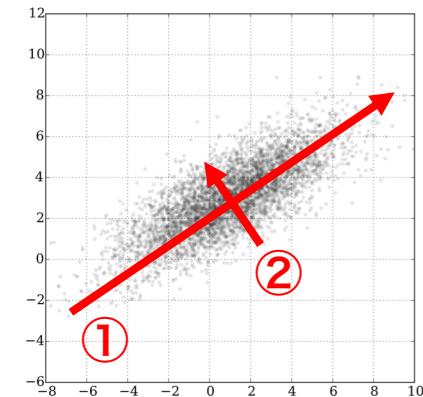
$$\mathbf{e}^\top \mathbf{V} \mathbf{e} = \mathbf{e}^\top (\lambda \mathbf{e}) = \lambda$$

第1軸の単位ベクトル \mathbf{e}

分散共分散行列の最大固有値 λ_1 に対する固有ベクトル

主成分分析 | 計算方法 | 残りの軸

- 残りの軸の候補
 - = 既存軸で説明できない軸
 - = 既存軸と直交する軸
- 対称行列の固有ベクトルは互いに直交
- 詳しい説明: <http://manabukano.brilliant-future.net/document/text-PCA.pdf>
- 固有値分解でなく特異値分解により主成分分析を行うことも



第 n 軸の単位ベクトル e

分散共分散行列の d 番目に大きい固有値 λ_d に対応する固有ベクトル

主成分分析 | 計算方法 | 寄与率

- 主成分の第 d 番目の主成分の寄与率
(元データの特徴をどれだけ捉えられているか)

$$C_d = \frac{\lambda_d}{\sum_{d'=1}^D \lambda_d} = \frac{\lambda_d}{\text{tr}(V)}$$

- 寄与率の小さな主成分の追加
 - › 些末な軸にまで無意味な解釈を与えててしまう
- 寄与率の大きな主成分の除去
 - › 重要な軸による変動を見逃してしまって

主成分分析 | 応用例

Google 検索結果
検索語: チャリ

- 自転車 | 製品情報、価格比較、通販 - 価格.com**
<https://kakaku.com> › bicycle ▾
自転車を買うなら、まずは価格.com自転車をチェック！ 全国の通販サイトの販売価格情報をはじめ、スペック検索、クチコミ情報、ランキングなど、さまざまな視点から商品を比較・検討できます！
電動自転車・シティサイクル・ママチャリ・クロスバイク・折りたたみ自転車・ミニベロ
- サイクルベースあさひ 総合サイト**
<https://www.cb-asahi.co.jp> ▾
サイクルベースあさひ お得なセールやキャンペーン、イベント情報から、自転車メンテナンス、カスタム、使い方や選び方などのお役立ち情報、自転車ブランド、製品情報まで！ 自転車情報が目白押し！！ サイクルベースあさひの総合サイトです。
- ヨドバシ.com - 自転車 通販【全品無料配達】 - ヨドバシカメラ**
<https://www.yodobashi.com> › アウトドア・スポーツ用品 ▾
自転車の通販ならヨドバシカメラの公式サイト「ヨドバシ.com」で！ 自転車パーツや自転車アクセサリなど人気の商品を多数取り揃えています。ご購入でゴールドポイント取得！ 今なら日本全国へ全品配達料金無料、即日・翌日お届け実施中。
- 東京都世田谷区の自転車の中古あげます・譲ります | ジモティー**
...
<https://jmtv.jp> › 売ります・あげます › 自転車 › 東京都の自転車 ▾
【ジモティー】東京都世田谷区で出品されている中古の自転車が掲載されています。ジモティーの東京都世田谷区の自転車のページは、東京都世田谷区付近で取引希望の中古の自転車が集まっているページです。東京都世田谷区付近で中古の自転車、格安の ...
- 自転車とは (ジテンシャとは) [単語記事] - ニコニコ大百科**
<https://dic.nicovideo.jp> › 自転車
いわゆるママチャリ。細分すると様々な種類がある。安い物だと一万円前後と非常に安価で普及しており、最低限の変速機や泥よけ、買い物かごなど日常生活用のパーツも基本的に装着済みで快適。ただし、速度効率や安定性などは下記の特化車種に及ばない ...

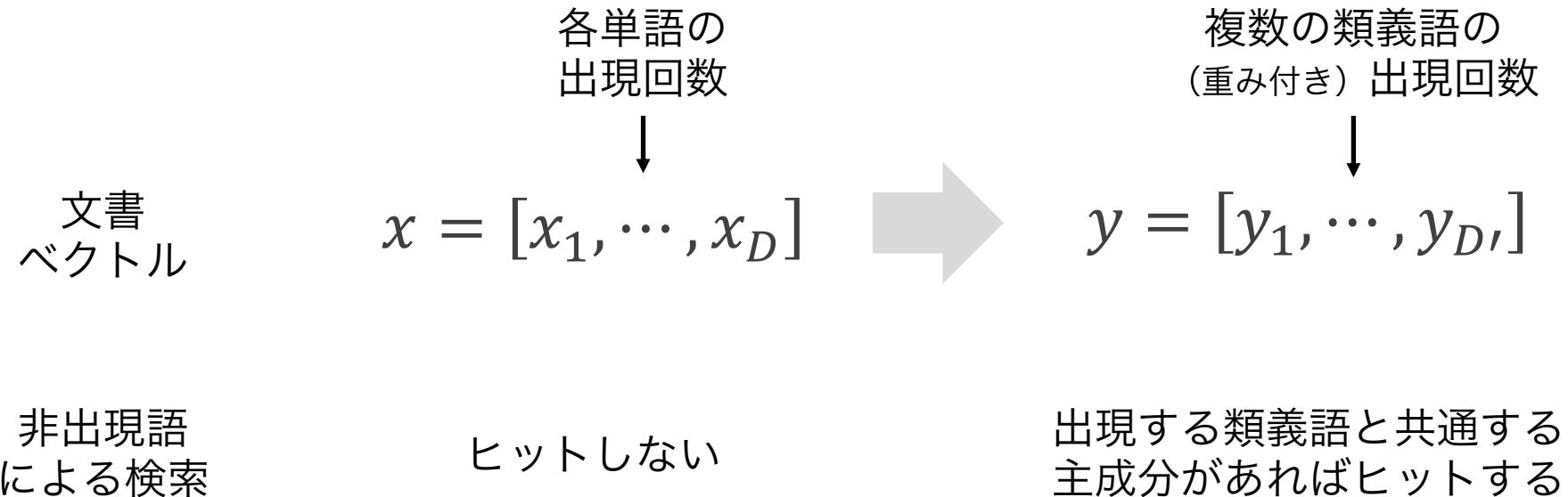
画像出典: <https://www.google.co.jp>

主成分分析 | 応用例

潜在意味解析 (Latent Semantic Analysis; LSA)

もしくは (特に情報検索では)

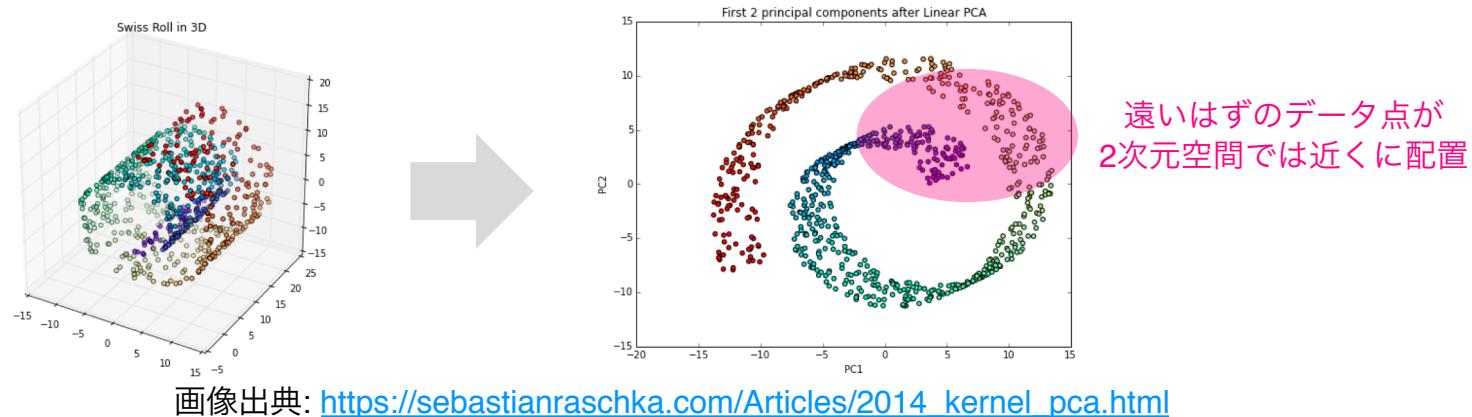
潜在意味インデキシング (Latent Semantic Indexing; LSI)



情報検索の語彙ミスマッチ問題の解消に貢献 [Deerwester+, 1990]

発展 | t-SNE [Maaten+, 2008]

- 主成分分析の問題
 - › 高次元データの構造が線形であることを想定
 - 次元間に複雑な関係性のあるデータだとうまく動かない
 - › 非類似なデータを離すことを類似なデータを近づけるよりも優先
 - 全体の構造は捉えられるが、近傍データ間の関係性は怪しいことも



- **t-distributed Stochastic Neighbor Embedding (t-SNE)**
 - › 上記の問題を解消する多次元データ可視化手法
 - › 次元削減後の空間での類似度をt分布でモデル化
 - 削減前に距離の近い／遠いデータを削減後も近く／遠くに配置

発展 | t-SNE | 利用例 [Borisov+, 2016]

ニューラルネットワークの
中間層（数百次元）の
可視化によく利用

図は元論文より引用

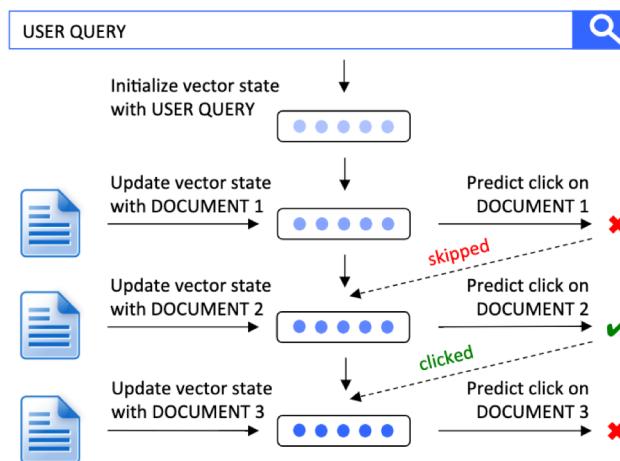


Figure 1: Modeling user browsing behavior on a SERP in the neural click model framework.

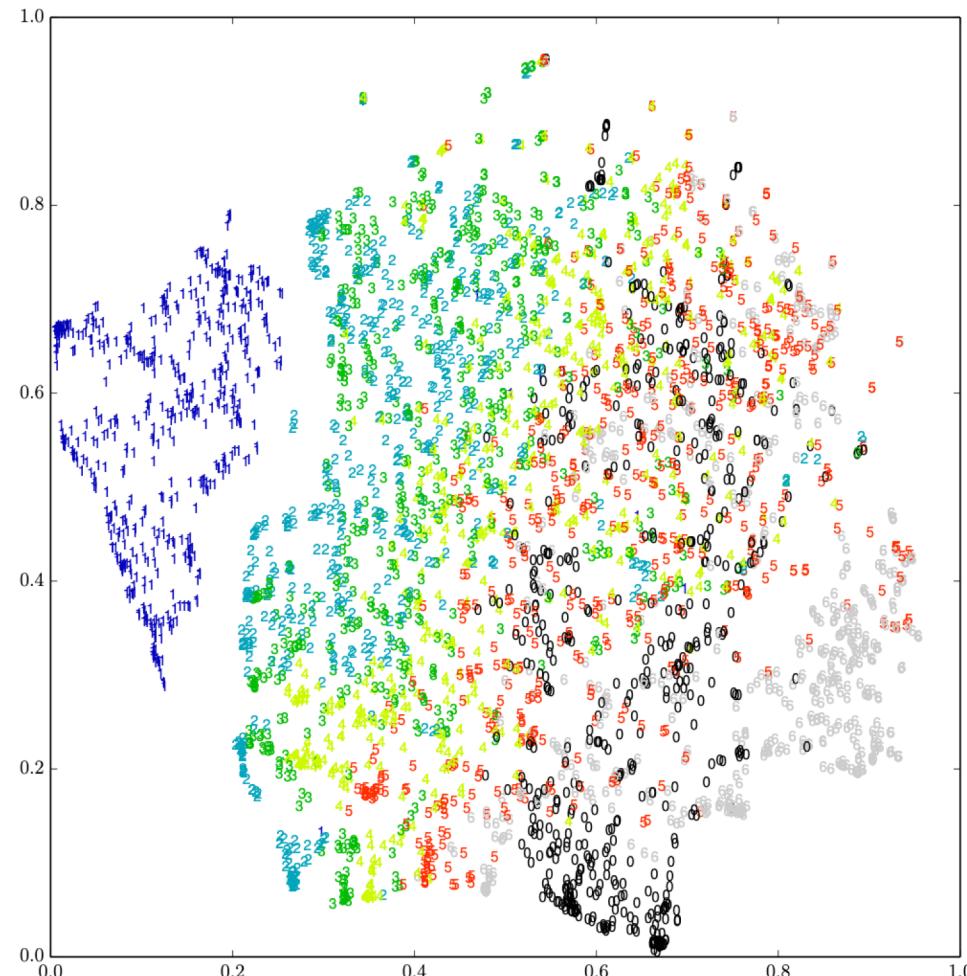


Figure 6: Two-dimensional t-SNE projections of the vector state s_7 for different distances d to the previous click. Colors correspond to distances: black 0; blue 1; blue-green 2; green 3; green-yellow 4; red 5; grey 6. (Best viewed in color.)

発展 | UMAP [McInnes+, 2018]

- **t-SNEの問題**
 - › 遅い
 - 大規模データの場合、効率性・スケーラビリティは特に重要
 - › 主に可視化用
 - 削減後の次元数が大きいと、うまくいく保証がなく、計算に時間がかかる
- **Uniform Manifold Approximation and Projection (UMAP)**
 - › 速い
 - › t-SNEと同様の可視化結果が得られる
 - › 削減後の次元数が大きくても良い
 - 可視化だけでなく機械学習用のデータ前処理としても使える

大雑把な使い分け

- **PCA**

- › 削減後の次元に意味付けをしたい

- **t-SNE**

- › 線形性がなさそうな複雑なデータ
 - › 近いデータ点を2次元でまとめてほしい

- **UMAP**

- › 速さが欲しい
 - › 可視化用途だけでなく機械学習タスクの前処理としても

座学部分のまとめ

高次元データを理解し上手く付き合うためには？

- **高次元データ**

- › 多くのデータがあてはまる：文書、画像、…
- › そのままでは扱いが難しい
 - 人間：3次元以上は理解できない
 - 機械：次元数が増えるほど計算量等の問題が

- **次元削減**

- › PCA：まずはこれ。軸（主成分）が直交
- › t-SNE：最近流行りの可視化手法。非線形でも良好な結果
- › UMAP：最新の手法。速い。機械学習の前処理にも

参考文献

- Deerwester, S., Dumais, S. T., Furnas, G. W., Landauer, T. K., & Harshman, R. (1990). Indexing by latent semantic analysis. *Journal of the American society for information science*, 41(6), 391-407.
- Maaten, L. V. D., & Hinton, G. (2008). Visualizing data using t-SNE. *Journal of machine learning research*, 9(Nov), 2579-2605.
- Borisov, A., Markov, I., de Rijke, M., & Serdyukov, P. (2016, April). A neural click model for web search. In *Proceedings of the 25th International Conference on World Wide Web* (pp. 531-541). International World Wide Web Conferences Steering Committee.
- McInnes, L., Healy, J., & Melville, J. (2018). Umap: Uniform manifold approximation and projection for dimension reduction. *arXiv preprint arXiv:1802.03426*.