

応用心理学 I
データ収集後探索的解析（テキストマイニング） その 2

テキストマイニング
講義編

明治大学 研究・知財戦略機構
佐藤浩輔

位置づけ

- 7/25 午後 Rブートキャンプ
- 7/26 午前 質的データの量的コーディング*
- 7/26 午後 (前半) **テキストマイニング講義編←ここ**
(後半) テキストマイニング実習編

*中分遙先生（高知工科大学）担当

講義の概要

- この授業は
 - 計量テキスト分析/テキストマイニング/自然言語処理とはなにか
 - テキストを扱う社会科学の研究をどのように計画するか
 - テキストを分析することで何がわかるか
 - テキストデータをどのように処理するか
 - テキストをどのように分析するか
 - 結果をどのように報告するか

について

300分で説明から実習までやってしまおうという
大変無謀な野心的な講義

講義の流れ

- 講義編
 - テキストマイニング(計量テキスト分析)と自然言語処理の概要を学び、何ができるかを知る
 - テキストを用いた研究のデザイン、研究計画の立て方を学ぶ
- 実習編
 - 実際にデータを扱いながら学ぶ
 - ①前処理を行い、分析ができるようデータを加工する
 - ②分析を行い、解釈する

計量テキスト分析/テキストマイニングとは何か
それを用いていったい何ができるか

講義編のアウトライン

- 前半
 - 計量テキスト分析・テキストマイニング・自然言語処理
 - それぞれの用語の整理
 - 特に、自然言語処理とは何か
 - なぜ量的手法が必要か
 - 人文・社会科学での応用例
- 後半
 - 計量テキスト分析・テキストマイニングの研究デザイン
 - 研究の立案から
 - 研究に関わる誤差
 - データの収集法
 - 分析手法について

計量テキスト分析・テキストマイニング・
自然言語処理

計量テキスト分析とテキストマイニング

- 計量テキスト分析 quantitative text analysis
 - テキストデータを量的 quantitative な手法を用いて分析すること
 - 社会科学の内容分析 content analysis の流れをくむ(樋口, 2006, 2014)
- テキストマイニング text mining
 - 大量のテキストデータから (機械を用いて) 価値のある情報を取り出すこと
 - 工学、マーケティングの流れをくむ
 - データマイニング:
 - mining: *Mining is the industry and activities connected with getting valuable or useful minerals from the ground, for example coal, diamonds, or gold.* --Collins COBUILD English dictionary
 - 大量のデータの中から価値のある情報を取り出す技術
 - cf. Webマイニング: 大量のWebデータの中から
 - 探索的な手法というニュアンス
 - 価値のある情報が埋まっているとは限らない

自然言語処理

- テキストマイニング/計量テキスト分析
→**自然言語処理技術**を用いてテキストデータから情報を抽出
- 自然言語処理(Natural Language Processing: NLP)
 - 構造化されていない自然言語を扱うための技術
 - 自然言語：普通の人が使うような言葉や文章
vs. 形式言語：人工的に作られた言葉(e.g. プログラム言語)
 - 自然言語を処理して様々な情報を抜き出したり生成したりする

自然言語処理の技術

- 基礎技術
 - 形態素解析
 - 構文解析
 - 意味解析
 - 固有表現抽出
- 応用技術
 - 文書分類
 - 自然言語理解
 - 自然言語生成
- 実社会への応用例
 - 検索エンジン
 - 自動翻訳
 - 質問応答・チャットボット

Google

Google 検索

I'm Feeling Lucky

<https://www.google.co.jp/>

文 テキスト

ドキュメント

言語を検出する

日本語

英語

韓国語



吾輩は猫である。名前はまだ無い。



どこで生れたかとんと見当がつかぬ。何でも薄暗いじめじめした
所でニヤーニヤー泣いた。声がけはなかった。アハフ。五年目だ。
始めて人間というも

Wagahaihanekodearu. N
Nani demo usugurai jim
iru. Wagahai wa koko de



英語

日本語

韓国語



I am a cat. There is no name yet.



I have no idea where I was born. I remember only that I was crying
in a place where it was dull bullying anything. For the first time here
I saw a human being.





●●●● au

9:53

97%

“面白い話”

タップすると編集できます

わかりました...

昔々、遙か彼方の仮想銀河
に、Siriという若くて知的な工
ージェントが住んでいまし
た。

ある晴れた日、Siriはパーソナ
ルアシスタントとしてAppleに

既読
12:25

探偵ごっこ

郊外の道路でオートバイ
が横転し、運転者がケガ
をした。捜査を進めた結
果、被疑者AとBが浮かび
上がったが、どちらが犯
人かはわからない。2人は
事件に対し、こう供述し
た。



Amazon.co.jp : Echo 第2...
amazon.co.jp



スマートスピーカー(AIスピーカー)徹...
yuki-no-yabo.com



価格.com - スマートスピーカー...
kakaku.com



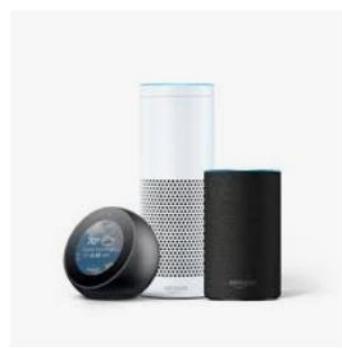
スマートスピーカー(AIスピ...
watch.impress.co.jp



価格.com - スマートスピーカ...
kakaku.com



スマートスピーカー・AIスピーカーでできること・選び方...
e-earphone.jp



スマートスピーカー(AIスピ...
watch.impress.co.jp



あらゆるスマートスピーカーを徹...
moov.ooo



最新スマートスピーカー徹底比較】「Amazon Ech...
robotstart.info

なぜテキストマイニングか

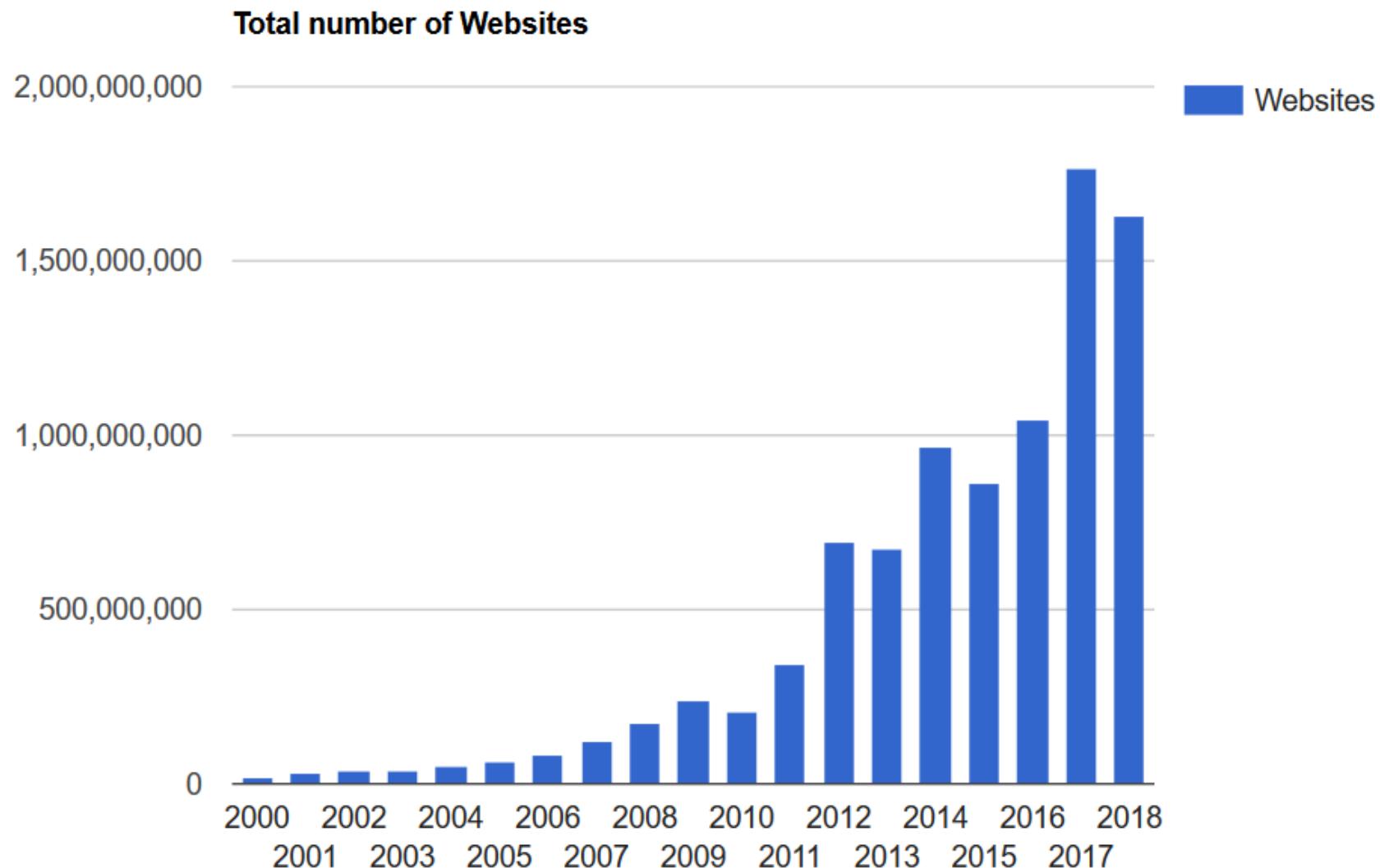
- 言語データの蓄積 + データ処理環境の整備

→言語データを手軽に大規模に利用できるようになった

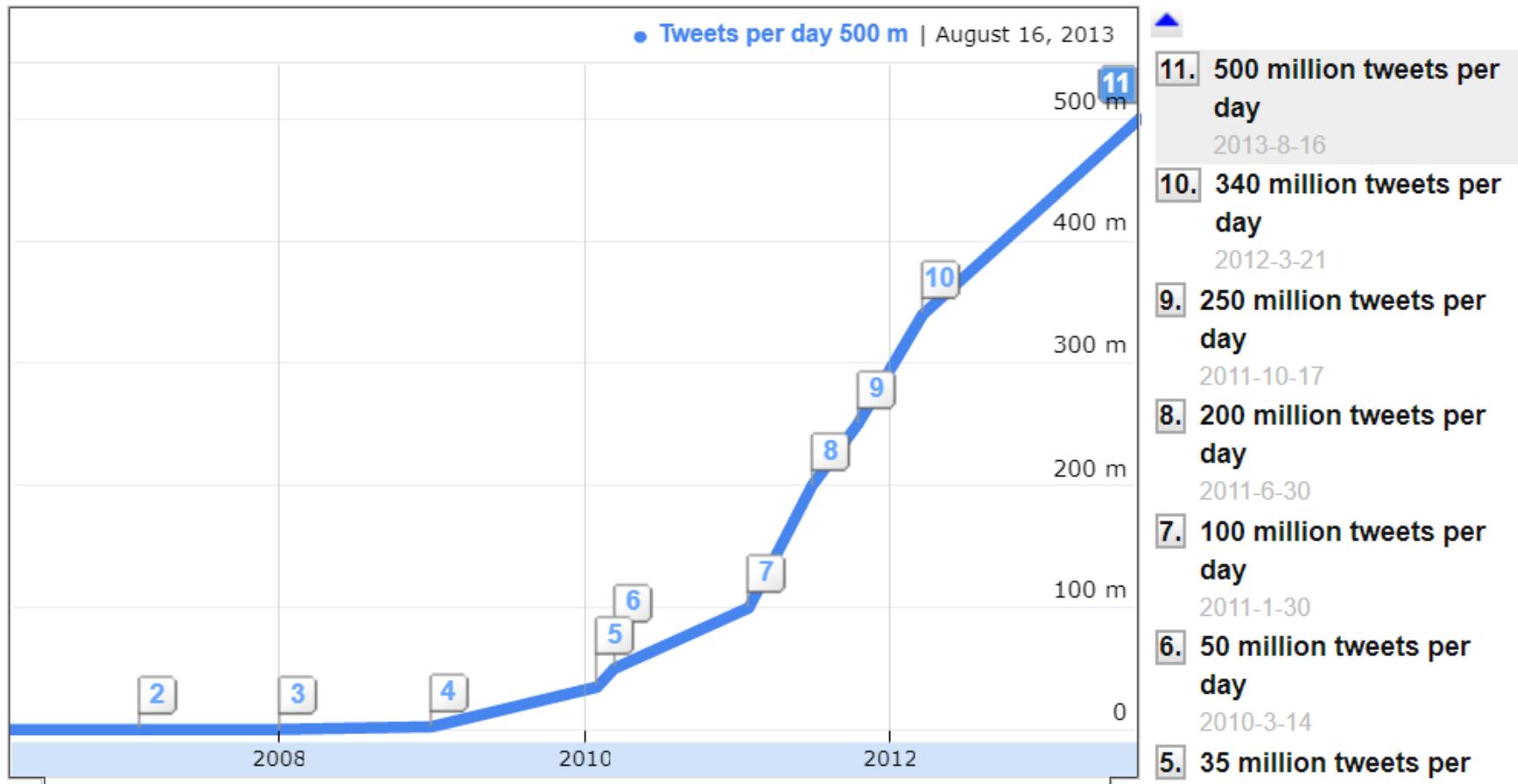
言語データの蓄積

- 人間同士のやりとり：言語情報
 - 文字情報
 - 文書・書籍
 - SNS・チャット
 - 電子メール
 - Webサイト
 - 音声情報
 - 会話

インターネット上のWebサイトの数



Twitterにおける一日あたりのtweetの数



データ処理環境の整備

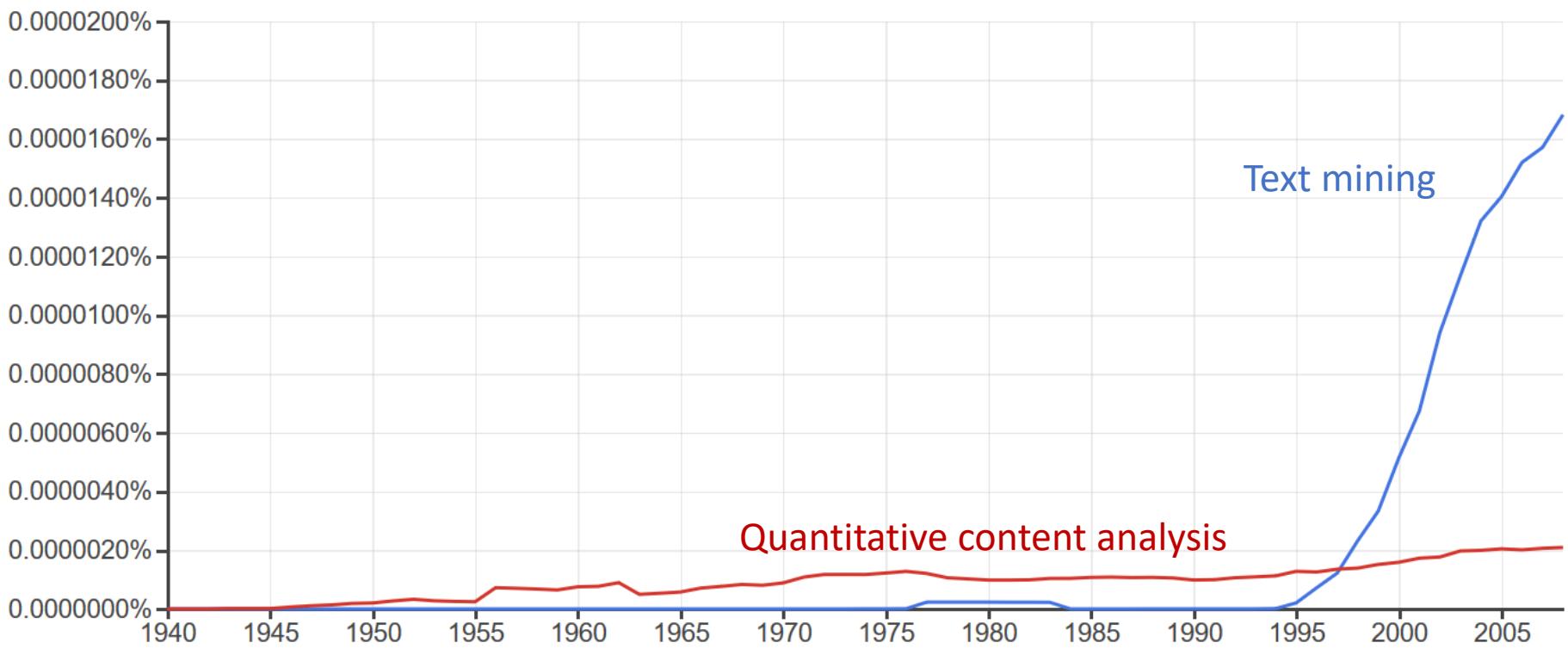
- ハードウェア能力の向上
 - 保存容量の増加：大量のデータを蓄積できる
 - 計算機の処理速度：大量のデータを処理できる
 - 通信速度の向上：大量のデータをやりとりできる
 - 通信インフラの整備：誰でもインターネットにアクセスできる
- データ処理技術の発展
 - データマイニング/統計的手法（特に機械学習）の発展
 - データセットの整備
 - 扱えるデータの増加
 - 自然言語処理技術の発展
 - 言語データを（一定の精度で）大量に、自動的に処理できる

Google Books Ngram Viewer

Graph these comma-separated phrases: case-insensitive

between and from the corpus with smoothing of .

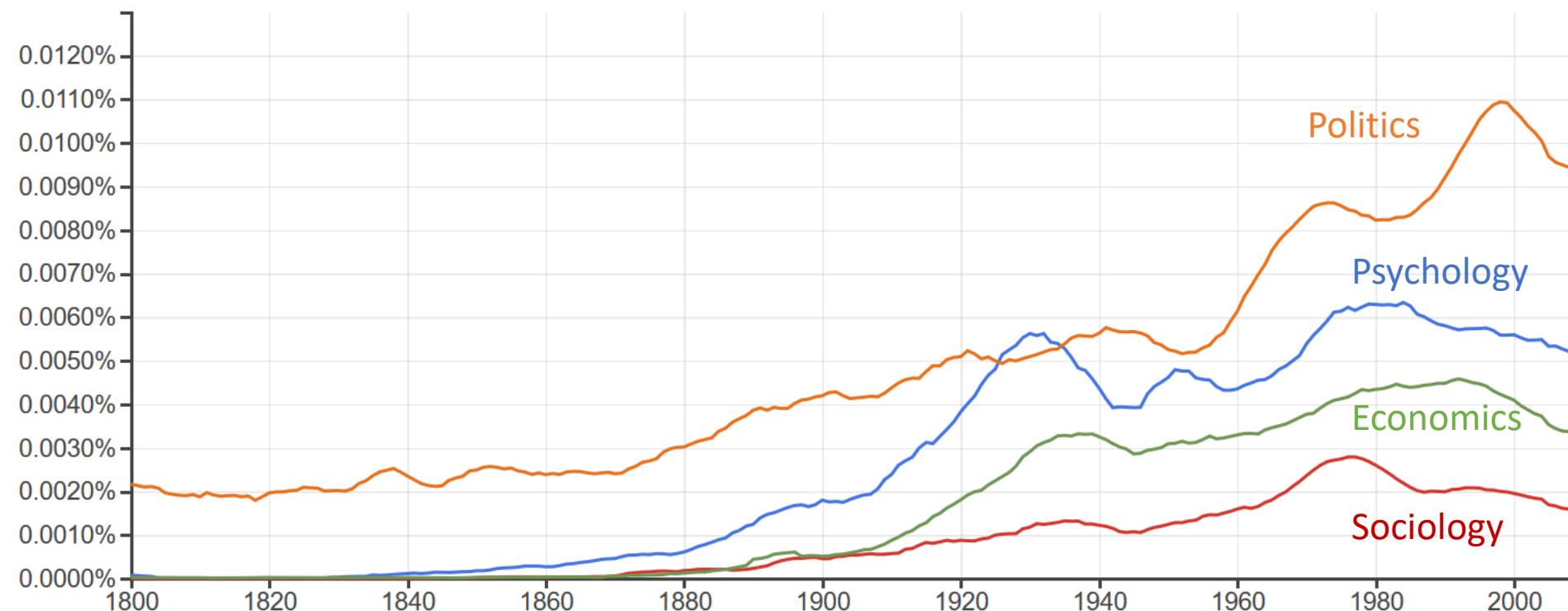
Search lots of books



Google Books Ngram Viewer

Graph these comma-separated phrases: case-insensitive

between and from the corpus with smoothing of



計量できると何が嬉しいか

- 質的なものを量的に扱える
 - 処理の効率性
 - 機械的かつ大量に処理できる
 - 手続きの明瞭性
 - 同じデータに同じ手続きを適用すれば、同じ結果が得られるはず
→検証可能である
- 量的に分析することで、質的な分析では見えてこないものを発見できる
 - 質的な分析と相互に補完しあえる

人文学分野における応用例

- デジタル人文学 digital humanities :
情報処理の技術を人文学の研究に応用
 - 文学
 - Distant reading (Moretti, 2013)
 - 精読 close reading に対して、情報処理技術で大量の文献を扱う
 - 計量文体学 stylometry / stylometrics (村上, 2002)
 - 文体を量的に扱う
 - 言語学
 - 計算言語学 computational linguistics
 - 歴史学
 - Digital history
 - 民俗学/民話学
 - 計算民話学 computational folkloristics (Abello et al. 2012; Tangherlini, 2016)
 - 民話の自動タグ付けやデータベースに活用

社会科学分野における応用例

- 計算社会科学: computational social sciences

Webのソーシャル化や実空間での様々な行動センシングが進行している現在、人々の自発的な情報行動やコミュニケーションなどの詳細はデジタルに記録・蓄積されるようになりました。このような大規模社会データを情報技術によって取得・処理し、分析・モデル化して、人間行動や社会現象を定量的・理論的に理解しようとする学問が「計算社会科学」(Computational Social Science)です。

計算社会科学はその目的の達成の方法論として、大規模社会データ分析研究、社会シミュレーションによる理論的研究、バーチャルラボによる実験的研究などを用いています。

<https://css-japan.com/about/>

Figure 3. “Censorship Magnitude,” The Percent of Posts Censored Inside a Volume Burst Minus Outside Volume Bursts.

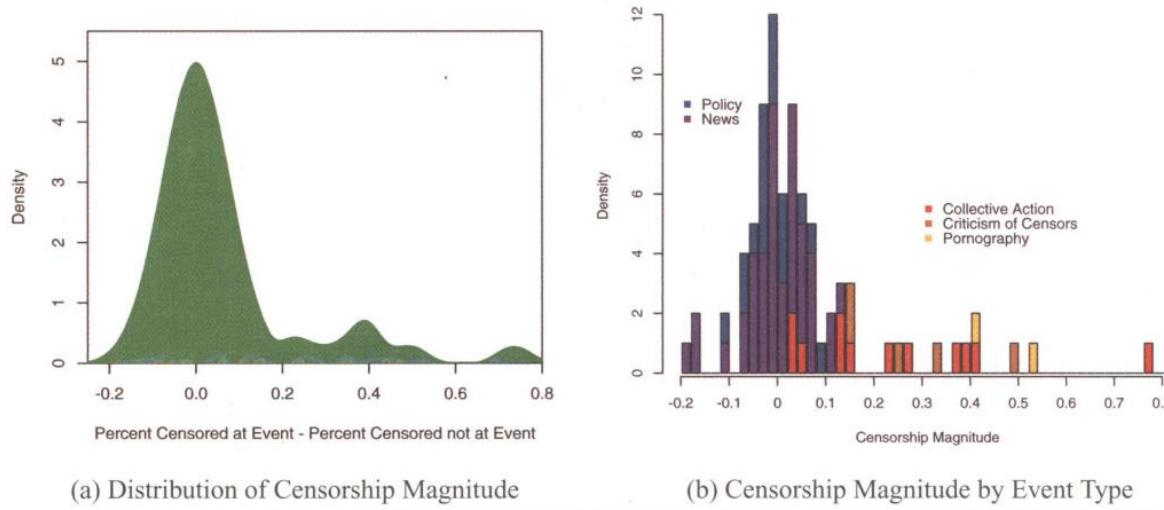
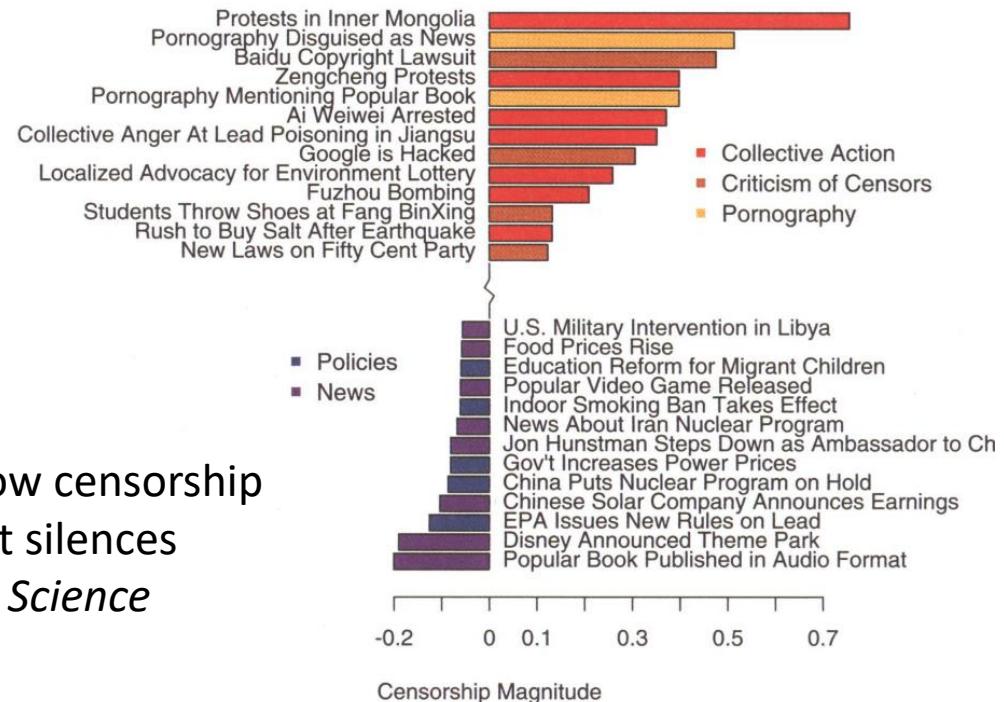


Figure 4. Events with Highest and Lowest Censorship Magnitude



King, G., Pan, J., & Roberts, M. E. (2013). How censorship in China allows government criticism but silences collective expression. *American Political Science Review*, 107(2), 326–343.

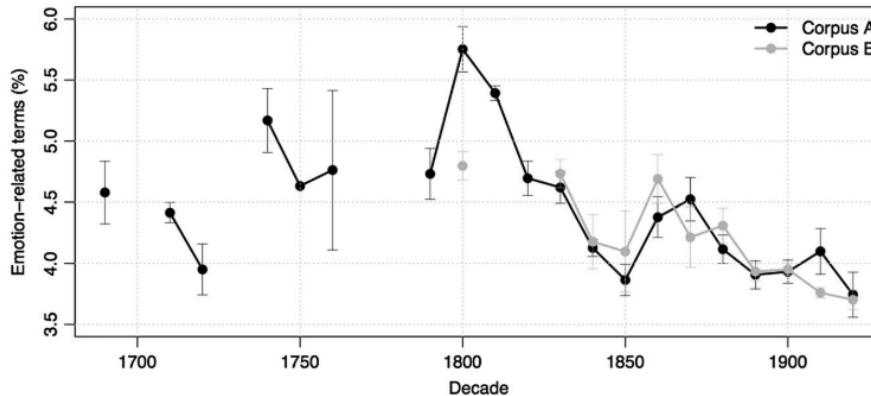


Figure 2. Emotionality changes in Anglophone literature, for the two “small data” corpora. Error bars represent 95% confidence intervals.

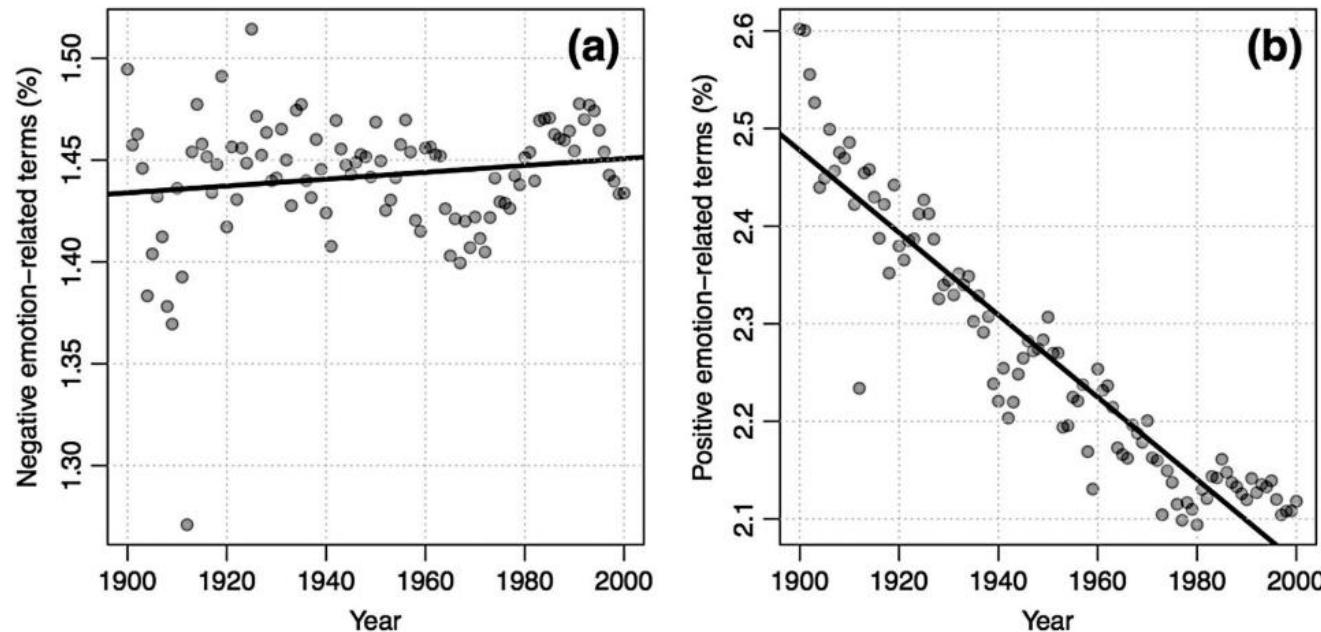


Figure 3. Emotionality changes in Anglophone literature, for the Google Books corpus. (a): negative emotions-related terms. (b): positive emotions-related terms. Solid lines represent linear regressions of the data.

Morin, O., & Acerbi, A. (2017). Birth of the cool: a two-centuries decline in emotional expression in Anglophone fiction. *Cognition and Emotion*, 31(8), 1663–1675.

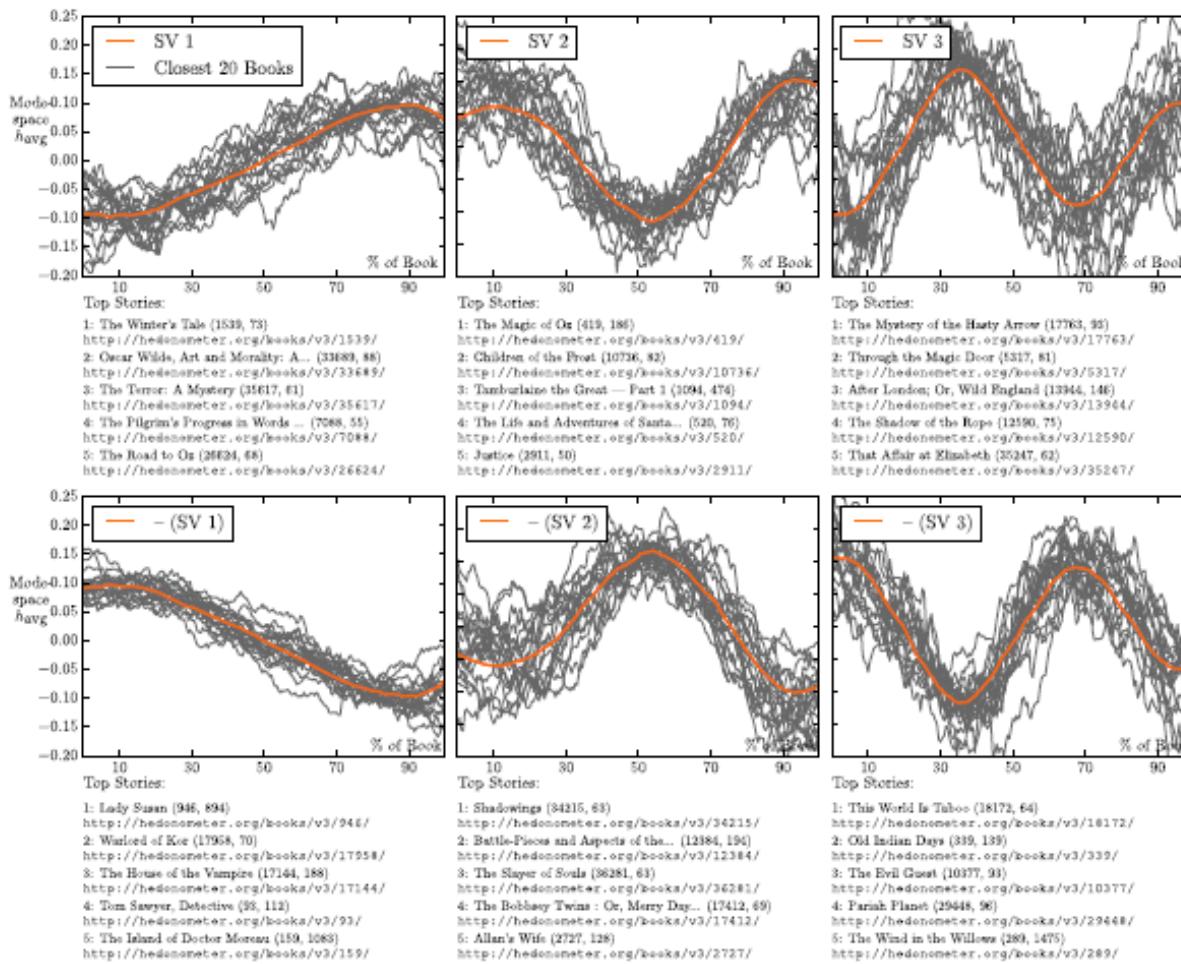


Figure 4 First 3 SVD modes and their negation with the closest stories to each. To locate the emotional arcs on the same scale as the modes, we show the modes directly from the rows of V^T and weight the emotional arcs by the inverse of their coefficient in W for the particular mode. The closest stories shown for each mode are those stories with emotional arcs which have the greatest coefficient in W . In parentheses for each story is the Project Gutenberg ID and the number of downloads from the Project Gutenberg website, respectively. Links below each story point to an interactive visualization on <http://hedonometer.org> which enables detailed exploration of the emotional arc for the story.

小まとめ

- 計量テキスト分析・テキストマイニング
 - 自然言語処理技術を用いて、テキストから価値のある情報を抽出することができる
 - 質的なものを量的に扱える
 - コンピュータを用いて、高速に大量に情報を処理することができる
 - 人文学・社会科学の様々な領域に広がっている

計量テキスト分析・テキストマイニングの 研究デザイン

研究の流れ

- 研究計画の立案
 - 研究目的の設定
 - データ収集手法の決定
- データ収集
- 分析
 - 前処理・クリーニング
 - 分析
 - 検証
- アウトプット
 - 報告・発表
 - レポート・論文・学会発表, etc.

←研究デザイン

研究計画の立案

研究デザインの重要性

**“Theories without facts may be barren, but
facts without theories are meaningless.”**

- K.E. Boulding(1941)

「事実のない理論は不毛であるが、理論のない事実は無意味である」

研究デザインの重要性

- 保存できるデータは有限
 - 潜在的に測定可能な変数の数は膨大
 - 「あらゆる」データを保存することは不可能
 - 取捨選択が必要：何が重要か
 - 理論(知識)によってフィルタする
 - データから言えることは多くない
 - 測定の問題
 - 一般化可能性
 - "Garbage in, Garbage out"
 - 「ゴミを入れればゴミが出てくる」
 - 何のデータをどうやって取るか、の計画が重要
- ⇒研究デザインが必要

研究計画の立案

- 「So What?」な研究にならないために
 - その研究のオーディエンスは誰か
 - 目的を明確にする：その研究をすると何が嬉しいのか
 - 理論的な価値：理論的な意味がある
 - 応用的な価値：何かの役に立つ
 - 資料的な価値：そのデータを取ること自体に価値がある
 - 分野によって関心が異なる
 - →色々な分野の研究を知ってセンスを磨く

研究目的の設定

- 仮説**検証**型研究
 - 特定の仮説を検証するための研究
 - 検証的な手法と相性がよい
- 仮説**生成**型研究
 - (意味のある) 仮説を生成するための研究
 - 事前の仮説はないものの、関心のある変数の分布や変数間の関連を調べる
 - 探索的な手法と相性がよい
 - 探索的データ解析(EDA: Exploratory Data Analysis; Tukey, 1977)
 - データマイニング

仮説検証と仮説生成

- ひとつの研究の中で組み合わせてもよい
 - e.g. 仮説検証パートと、主要な変数との関連を探索的に調べるための質問群
 - 重要なのは、それぞれこの項目は**どのような目的のためにとるのか**、を意識すること
 - 研究に使えるリソースは有限
 - 何が重要で何が重要でないか、優先順位を明らかにする
 - 探索的な分析で出た結果
 - 議論するに十分な測定精度がないかもしれない
 - 統計的なアーティファクトかもしれない
 - 多重検定の問題

目的を設定したら

- その目的を果たすためのプランを考える
 - 考慮すべき事項：測定の問題
 - 信頼性と妥当性
 - 誤差
 - データ収集法の選定
 - 実験, 調査, etc.
 - 収集後の分析方法

測定の問題

構成概念(construct)

各種心理構成概念：
感情, 態度, 動機
欲求, 認知, 知能, ...

現象を解釈・推論

頭の中の世界

現実世界



研究者

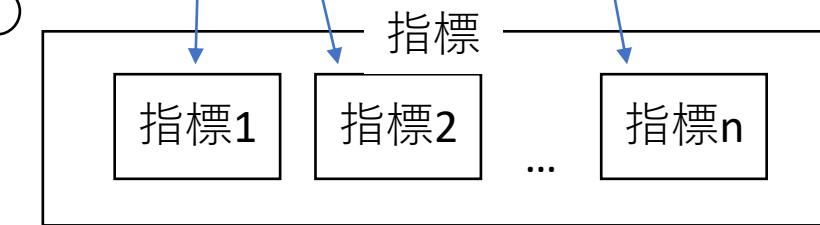
できごと

モデル・仮説を構築

構成概念X

構成概念Y

操作的に定義



測定

実測値

指標1

指標2

...

指標n

実験・調査

信頼性と妥当性

- 信頼性 reliability
 - 同じ条件で同じ測定を繰り返したとき、同じ結果が得られるかどうか
 - 信頼性のない例：誤差が $\pm 50\text{kg}$ ある体重計
 - 内的一貫性 internal consistency
 - Cronbach's α : 同じものを測っているなら下位項目間の相関は高いはず
- 妥当性 validity
 - 測定したいものが実際に測定できているかどうか
= 構成概念と測定指標との関連
 - 妥当性のない例：「体力」の指標としての国語のテストの成績

様々な妥当性

- 構成概念妥当性 construct validity: 構成概念を実際に測定できているか
 - 構成概念 construct
 - 現象を説明するために導入される仮説的/仮設的な概念
 - **構成概念そのものを直接観察することはできない**
 - 事象から推察ないし推論される
 - 構成概念を直接測ることはできないので、研究においては何らかの形で**操作的に**定義する必要がある
(e.g. ○○の指標として××を測定する)
 - 基準関連妥当性 criterion-related validity
 - 関連すべき他の基準とどれだけ関連しているか
 - 内容的妥当性 content validity
 - 測定が測りたい構成概念を満遍なくカバーしているか
- 外的妥当性 external validity
 - その研究の外(e.g. 日常場面)でも結果が妥当するか

研究における誤差

- 誤差：真の値からの「ずれ」
 - 偶然誤差 random error
 - ランダムに生じる真の値からのずれ（ばらつき）
 - サンプルサイズを増やせば減らすことができる
 - 系統誤差 systematic error
 - ランダムでない真の値からのずれ
 - サンプルサイズを増やしても減らすことができない
 - 系統誤差の3つのカテゴリ
 - 選択バイアス selection bias
 - 情報バイアス information bias
 - 交絡 confounding

研究における誤差

選択バイアス

- 研究対象者を選定する際に生じるバイアス
(=標本の性質に関わるバイアス)
 - 標本抽出バイアス **sampling bias**
 - 抽出した参加者は母集団を代表していないかもしれない
 - 自己選択バイアス **self-selection bias** /
参加バイアス **participation bias**
 - 研究に参加してくれるのは特殊な人かもしれない
 - 特定の人々は研究に参加してくれないかもしれない

研究における誤差

情報バイアス

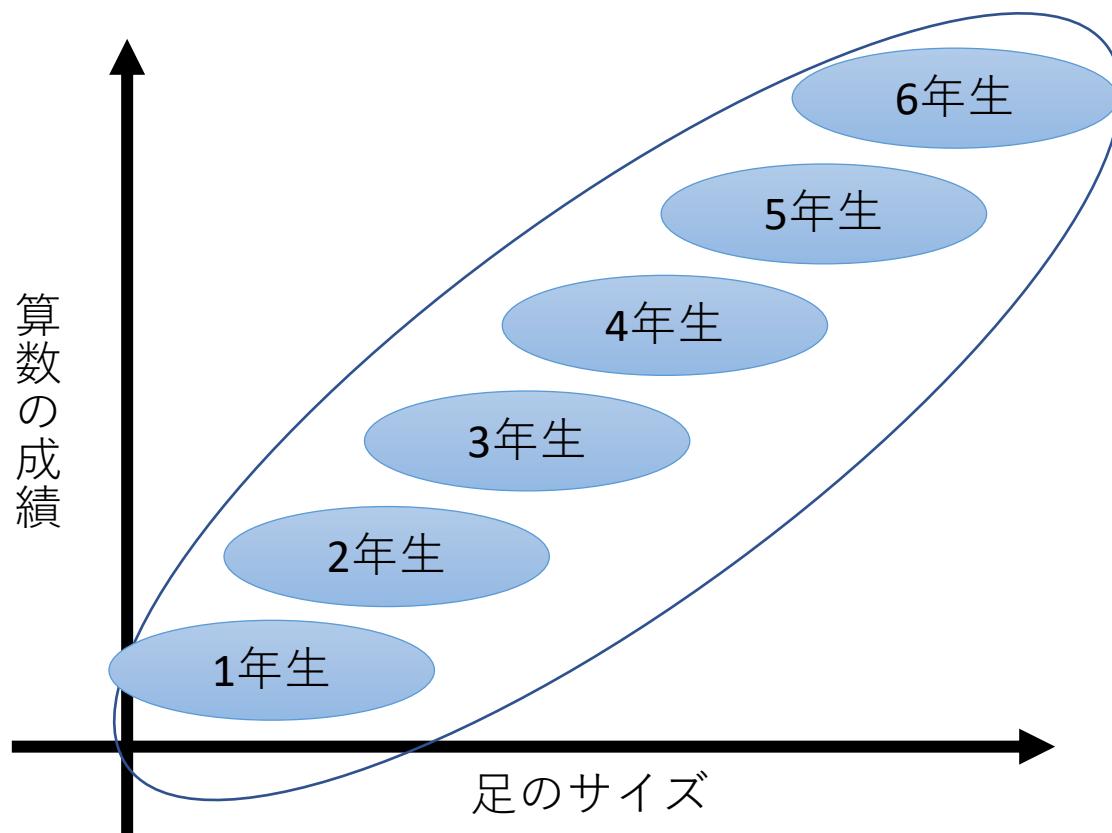
- 研究対象者からデータを得る際に生じるバイアス
（=測定に関わるバイアス）
 - 実験者効果 experimenter effect / 要求特性 demand characteristics
 - 実験者が仮説を知っていると、仮説を支持する方向に（意図せずに）参加者を誘導してしまう (Rosenthal, 1966)
 - 実験参加者は、実験者の期待する回答をしてしまう傾向がある (Orne, 1962)
 - 社会的望ましさ social desirability / 報告バイアス reporting bias
 - 望ましくない回答は抑制される
 - 想起バイアス recall bias
 - 群によって記憶の正確さが違うかもしれない
 - 誤分類 misclassification
 - 測定の精度や方向性が群間で異なると問題になることがある

研究における誤差

交絡

- ・他の変数の効果が混ざってしまうために生じるバイアス
- ・第三の変数の影響を受けること
- ・例：足の大きさと成績
 - ・「ある小学校の児童全体を対象に、算数の能力テストと足のサイズの計測を行った。その結果、強い有意な正の相関がみられた。ゆえに足の大きさと算数の能力は関係しているといえる」
⇒ ? ? ? ?

考えられる可能性



⇒ 「学年」が「成績」と「足のサイズ」両方に影響を与えている

研究における誤差

交絡への対応

- 研究デザインで対応
 - 無作為割り当てる (実験)
 - 交絡要因が実験群と対象群で異なるよう群を無作為に割り当てる
 - 交絡要因が同じと思われる集団を対象に分析する
 - e.g. 職業コホート
- 分析で対応
 - 層化 stratification して分析する
 - 交絡をもたらすと思われる変数(e.g. 年齢)によって層に分け、グループごとに分析する

手法の選定

- 心理学における主な選択肢
 - 実験
 - 調査
 - ビッグデータ

実験の特徴

- 実験 experiment
 - pros
 - 自由度が高い
 - 条件の無作為割り当てができる
 - 交絡を抑えられる
 - 因果関係を議論できる
 - 様々な外的要因を統制できる
 - cons
 - コストが高い
 - 参加者を確保する
 - 謝金・謝礼が必要
 - 多人数を対象にするのには向かない

- 生理指標・行動指標
 - 信頼のおける指標
 - コストが高い
 - 一度にたくさんの変数を測るのが難しい
- 質問紙実験
 - 容易に実施可能
 - 一度にたくさんの変数が測れる
 - 結果が細かいワーディングに左右される
 - 態度と行動は一貫しない(LaPiere, 1934)
→行動データではない
 - 実際の人間の振舞いを表しているとは限らない
→妥当性を担保する工夫が必要

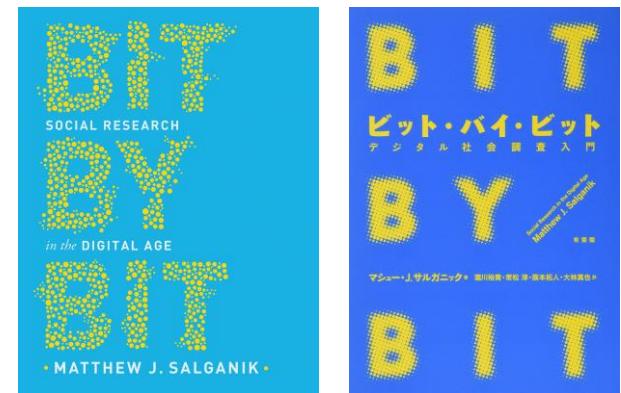
調査の特徴

- 調査 survey
 - pros
 - 標本が特定の母集団を代表するように設計できる
 - "一般の"人々の回答を得られる
 - 多人数を対象にできる
 - cons
 - 交絡がある
 - コストが高い
 - 回答率が低い場合には代表性が失われる
 - 基本的には一時点の調査では因果関係を議論できない

- 社会調査
 - 何らかの抽出台帳(e.g. 住民基本台帳)から無作為抽出した人々を対象に調査票に回答してもらう
 - 訪問留置法・郵送法など様々な手段があるが一般に高コスト
 - 回答率を確保することが難しい
- Web調査
 - 調査会社に依頼し、企業の確保・整備しているアンケートモニターを対象に収集してもらう
 - 性別・地域等が柔軟に設計できる
 - 回答率が高い
 - 調査の中では比較的低コスト
 - 選択バイアスを考慮すべき
 - 質問紙よりも回答者が適当に答えがちな傾向がある
- その他
 - パネル調査
 - 同じ回答者に期間をおいて何度も調査に参加してもらう
 - コホート研究
 - 特定の集団を長期間追跡

ビッグデータの特徴

- ビッグデータの10の特徴(Salganik, 2017)
 - 研究にとって有益な特徴
 - 巨大さ
 - 常時オン
 - 非反応性
 - 問題となる特徴
 - 不完全性
 - アクセス不能性
 - 非代表性
 - ドリフト
 - アルゴリズムによる交絡
 - 汚染
 - センシティブ



Salganik (2017) *Bit by bit*

ビッグデータの特徴①

巨大さ

- データセットが巨大であること有益な研究
 - まれなできごとの研究
 - 不均質性 heterogeneityの研究
 - 実験の処理の効果の違い
 - 地域の特性の違い(e.g. Chetty et al., 2014)
 - 微小な差異の検出
 - 1%の差異が意味を持つ分野もある
- 落とし穴
 - 系統誤差に注意しなければならない
 - 偶然誤差は減っても系統誤差は減らない

ビッグデータの特徴②

常時オン

- 絶えずデータを収集
 - 時系列データを取ることができる
 - 予期せぬ出来事の研究
 - 歴史的なできごと・事件
 - リアルタイム推定が可能になる

ビッグデータの特徴③

非反応性

- 社会科学における「反応性 reactivity」
 - 人は観察されると行動を変える(Webb, 1966)
 - 実験者効果
- オンラインのデータ
 - データをとられることを人々が通常意識していないという意味で、非反応的
 - 落とし穴
 - 非反応的であるからといって、そのままの態度や行動を表しているわけではない
 - 社会的望ましさなどといった要因の影響はなお残る

ビッグデータの特徴④

不完全性

- 欲しい情報が入っていない
- 研究上の構成概念と対応するか
 - 構成概念妥当性

ビッグデータの特徴⑤

アクセス不能性

- データが存在しても研究者がアクセスできるとは限らない
 - 政府や自治体、企業の中にあるデータ

ビッグデータの特徴⑥

非代表性

- ビッグデータの多くは非代表的
→母集団を代表してはいない
 - 研究結果を一般化できるか？

ビッグデータの特徴⑦

ドリフト

- ドリフト(浮動)
 - 時間にともなうシステムの変化
 - どのようなシステムか
 - 誰が使うのか
 - どのように使うのか

ビッグデータの特徴⑧

アルゴリズムによる交絡

- システム上の行動：人間のありのままの行動ではない
 - システム設計者の企図によって人工的な結果 (artifact)が生じる
 - Ugander(2011): Facebookにおけるネットワーク
 - 友達の人数は「20」が突出して多い
 - 友人を20人になるまで増やすようシステムがうながす仕組みがある
 - 「友達の友達」同士は友達になりやすい
 - 社会ネットワークにおいては推移性 **transitivity** として知られる現象
 - 社会理論を知っている設計者がシステムに理論を組み込んでいる(遂行性 **performativity**)

ビッグデータの特徴⑨

汚染

- スパムやボットなど、人間の行動を反映しないデータが紛れ込んでいる
 - Back, Kühner, & Egloff(2010): 9.11後のSNS上のメッセージを分析
→「9.11後に怒りの感情がSNS上で増加している」
 - Pury(2011): 「Backらの結果は誤り」
 - Backらの結果はBotの仕業
 - Botの投稿を取り除くとBackらの結果は再現されない
→人工的結果(artifact)
 - その後、Backら自身の再集計後の分析でも結果は再現されず(Back, Kühner, & Egloff, 2011)

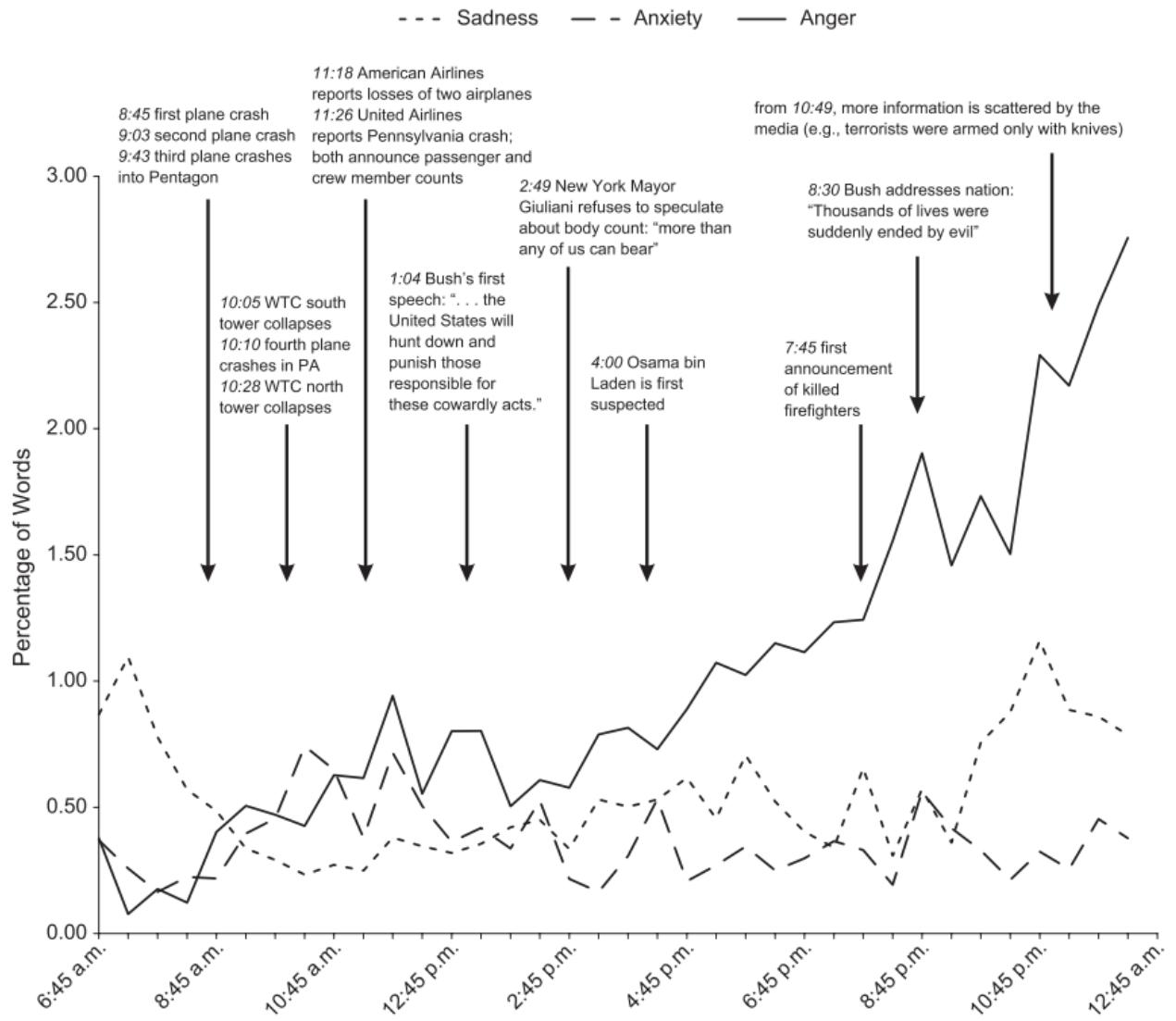
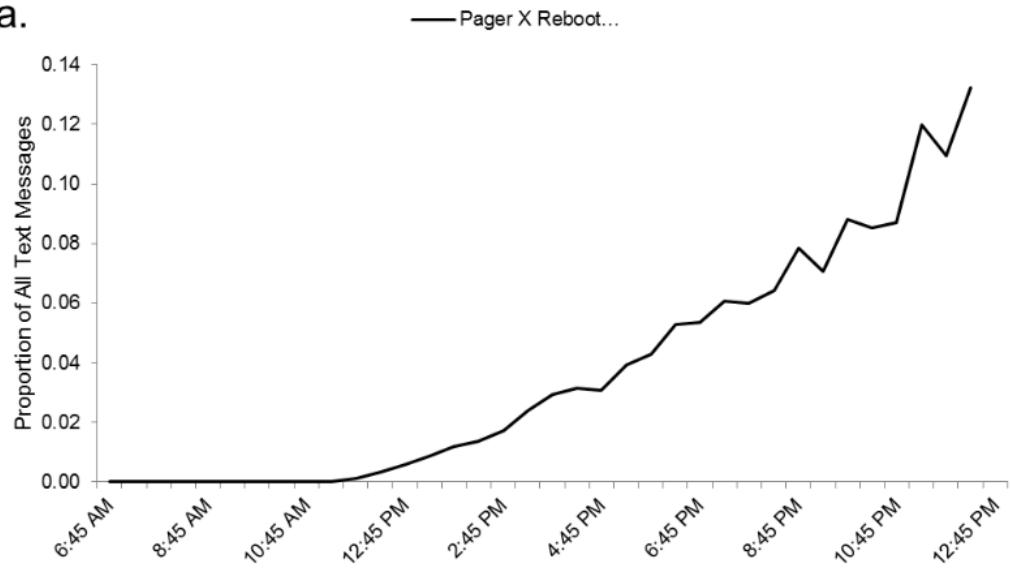


Fig. 1. The timeline of sadness, anxiety, and anger on September 11 as expressed in messages sent to text pagers. Each data point represents the mean percentage of words related to the specific negative emotion, averaged across 30 min. The time slots start at 6:45 a.m. to 7:14 a.m. on September 11, 2001, and end at 12:15 a.m. to 12:44 a.m. on September 12, 2001. Exact times and brief descriptions of the most important events of September 11 are included above the timelines. WTC = World Trade Center

a.



b.

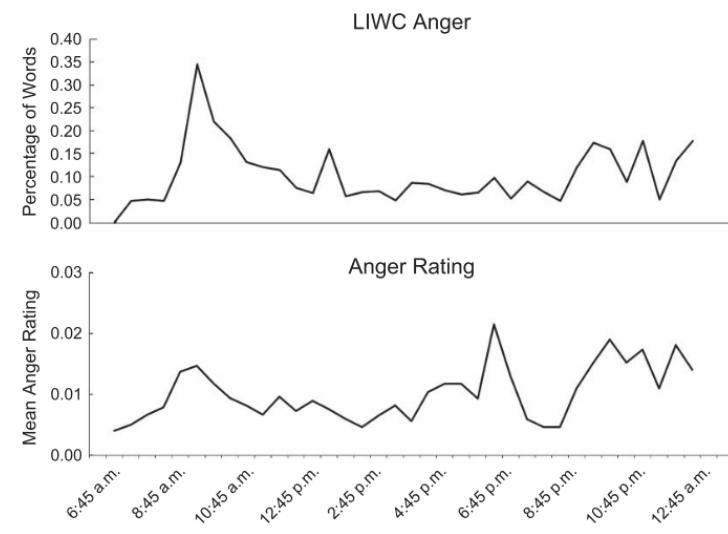
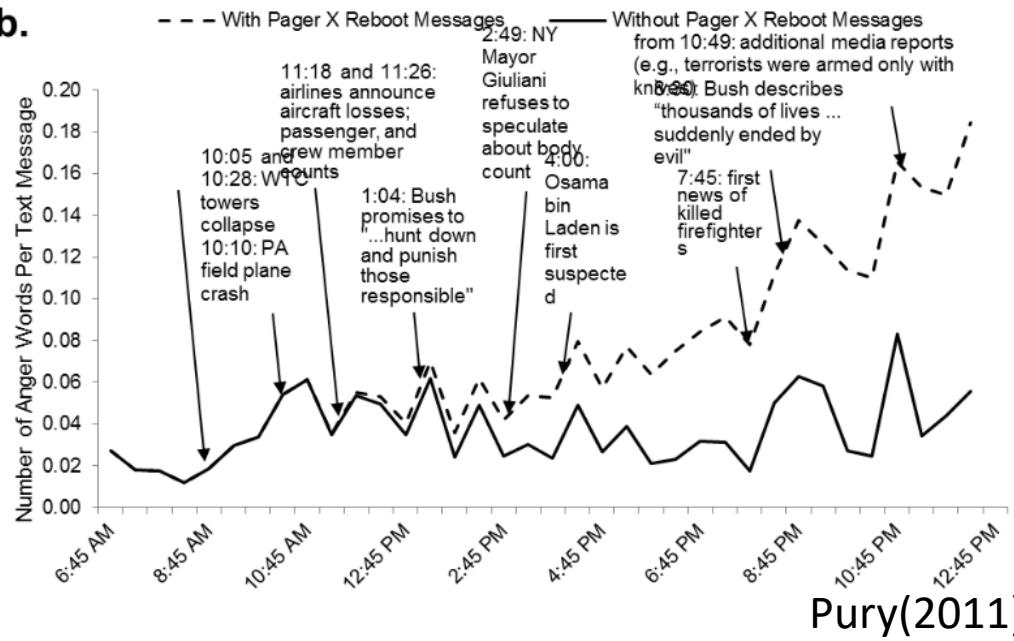


Fig. 1. A revised timeline of anger as expressed in 37,606 social messages sent to text pagers on September 11, 2001. The graphs show (a) the mean percentage of words related to anger (as classified by Linguistic Inquiry and Word Count; Pennebaker, Francis, & Booth, 2001) and (b) the mean anger rating (0 = no anger, 1 = some anger, 2 = strong anger; averaged across three raters for each message) across time slots starting at 6:45 a.m. to 7:14 a.m. on September 11, 2001, and ending at 12:15 a.m. to 12:44 a.m. on September 12, 2001.

Back, Kühner, & Egloff (2011)

ビッグデータの特徴⑩

センシティブ

- 個人のセンシティブな情報が含まれる
 - 複数のデータをつなげることで個人が特定できてしまうかもしれない

テキストデータの収集法

- 新たに入手する
 - 実験や調査の実施
- 既存のデータの活用・Web上での収集
 - 文書・書籍
 - 公開データ
 - データを持っている団体との接触
 - Web API
 - スクレイピング

テキストデータの収集法

実験・調査による収集

- 実験・調査
 - 文章データ
 - 記述式・自由回答
 - 日記法
 - 音声・映像データ
 - 面接（インタビュー）
 - 討議などの録音
- 注意点
 - 事前に電子化する方法を考えておく
 - 必ずしも取りたいデータが取れるとは限らないので、あくまで副次的な方法として考えておく

テキストデータの収集法

文書・書籍からの収集

- 手動入力
 - 手間がかかる
- OCRによるスキャン
 - 光学的自動文字認識 Optical Character Recognition
 - 日本語の文章は弱い
- 入力チェックがある分、データクリーニングに時間がかかることに注意

テキストデータの収集法

公開データの利用

- 公開されているデータセットの例
 - 各種オープンデータ
 - 国・地方公共団体・官公庁のオープンデータ
 - 研究用データセット
 - 各種言語資料（コーパス）
 - 情報学研究データリポジトリ
<https://www.nii.ac.jp/dsc/idr/datalist.html>
 - パブリックドメインの文学作品
 - Project Gutenberg
 - 青空文庫
 - その他
 - Wikipedia
 - Kaggle

テキストデータの収集法

データを持っている団体との接触

- 国や地方公共団体・企業
 - 様々なデータを持っている
 - その多くは通常アクセスできない
- アクセスできる可能性：ゼロではない
 - お願いしてみる
 - 共同研究

テキストデータの収集法

Web APIを用いた収集

- API: Application Programmable Interface
 - 他のプログラムからアクセスするために提供されているツール群
- Webサービスの中にはAPIを通じて様々な情報を取得できるものがある
 - Twitter
 - Instagram
 - Facebook

テキストデータの収集法

スクレイピング

- スクレイピング(scraping: こそげ落とす)
 - Webページを取得し、意味のある情報を抽出する
 - すべてのWebページにAPIが用意されているわけではない
→ダウンロード・加工してデータに
 - Webページ：HTMLで記述されている
 - 様々なツールがある

テキストデータの収集法

データ取得時の注意

- 公式の取得法があればそれを使う

クローラを使わない [編集]

記事を大量にダウンロードするためにクローラを使わないで下さい。強引なクローリングは、ウィキペディアが劇的に遅くなる原因となります。

ウィキペディアのデータベースから自動的にデータの収集がなされた場合、システム管理者によってあなたのサイトからウィキペディアへのアクセスを禁止する措置が取られることもあります。またWikimedia Foundationが法的措置を検討することもあります。

<https://ja.wikipedia.org/wiki/Wikipedia:データベースのダウンロード>

- 取得先に過度の負荷をかけないようにする
 - 短期間・高頻度にアクセスすると攻撃とみなされるかもしれない

データの収集については
8月or9月開催予定のセミナーで詳しく話します

どういう分析をするか

- 質的なデータ（自然言語）を量的なデータに変換する
 - 頻度
 - 分布
 - 各種指標・統計量

どういう分析をするか

頻度

- 文書や文を単位に頻度を算出する
 - 文字
 - 単語
 - トークン token : ひとつひとつの単語の出現「延べ語数」
 - タイプ type : 単語の種類「異なり語数」
 - 共起
 - 単語同士が文や文書に同時に登場する回数
 - n-gram
 - 連続するn個の単語
 - 機械学習によるタグ付け
 - 感情分析などによる「感情」の判定
 - 各種分類器による判定

どういう分析をするか

分布

- 各種要素の頻度の分布
 - 長さ
 - 単語の長さ
 - 文の長さ
 - 単語の種類
 - 品詞
 - 識別語
 - 機能語
 - その他
 - 語彙・漢字・仮名・読点・文節・音韻・文頭文字

どういう分析をするか

指標・統計量の例

- TF-IDF (Term-Frequency / Inverse DFrequency)
 - 文書における単語の重要度
- 類似度
 - 特徴ベクトル間の類似度
 - Pearsonの積率相関・Spearmanの順位相関・コサイン類似度
 - 集合同士の類似度
 - Jaccard係数
 - 文字列同士の類似度
 - Levenshtein距離（編集距離）
- 相互情報量 **mutual information**
 - 共起の重要度
- TTR (**t**oken **t**ype **r**atio)
 - 延べ語数・異なり度数。語彙の多様性
- Simpson's D
 - 繰り返し表現の多さ
- 各種スコア (e.g. 感情分析)
 - LIWC(Linguistic Inquiry and Word Count)

どういう分析をするか

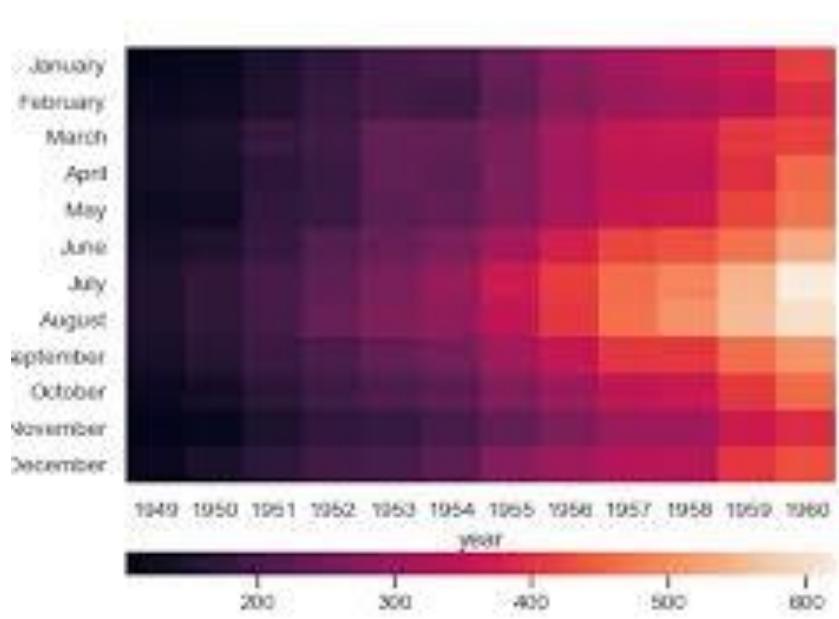
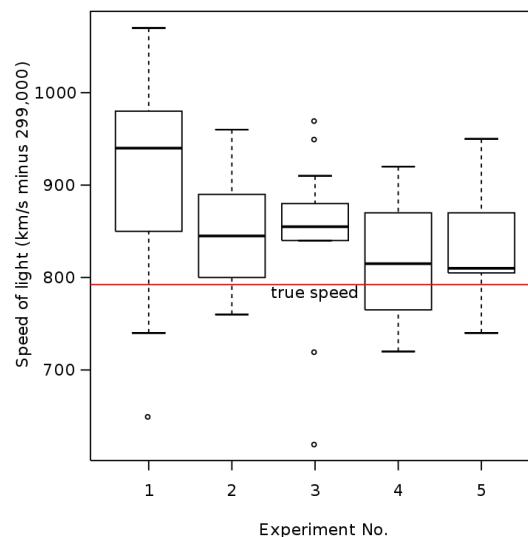
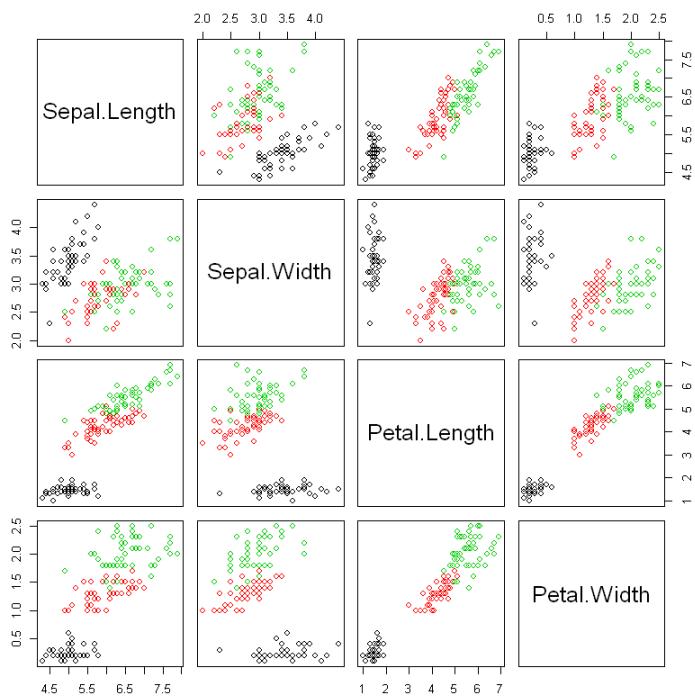
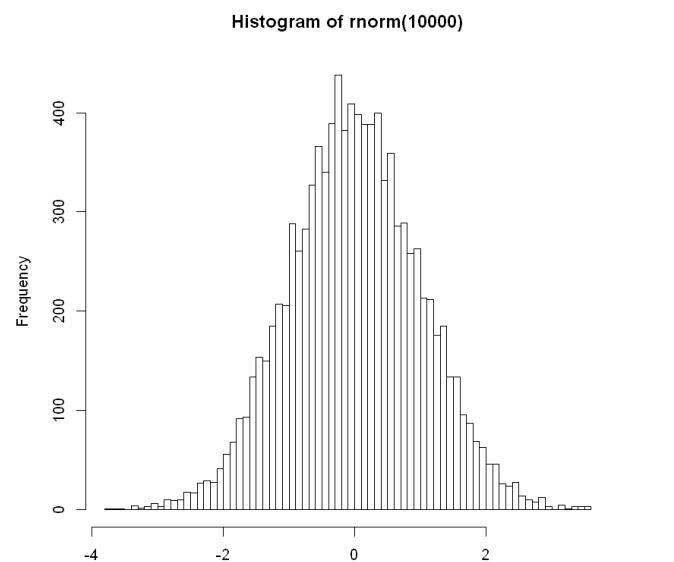
分析手法

- ①図示する
 - 各種グラフ
 - ネットワークグラフ
- ②比べる
 - カイ二乗検定：分布間の比較
 - 尤度比検定：頻度の比較
- ③まとめる
 - クラスター分析
 - 次元削減：主成分分析/因子分析
- ④分類する
 - 潜在意味解析/トピックモデル
 - 感情分析
 - ニューラルネット

どういう分析をするか

①図示する

- 可視化することで全体のパターンを把握する
 - 各種グラフ
 - ヒストグラム
 - 箱ひげ図
 - 散布図行列
 - ヒートマップ
 - ネットワークグラフ



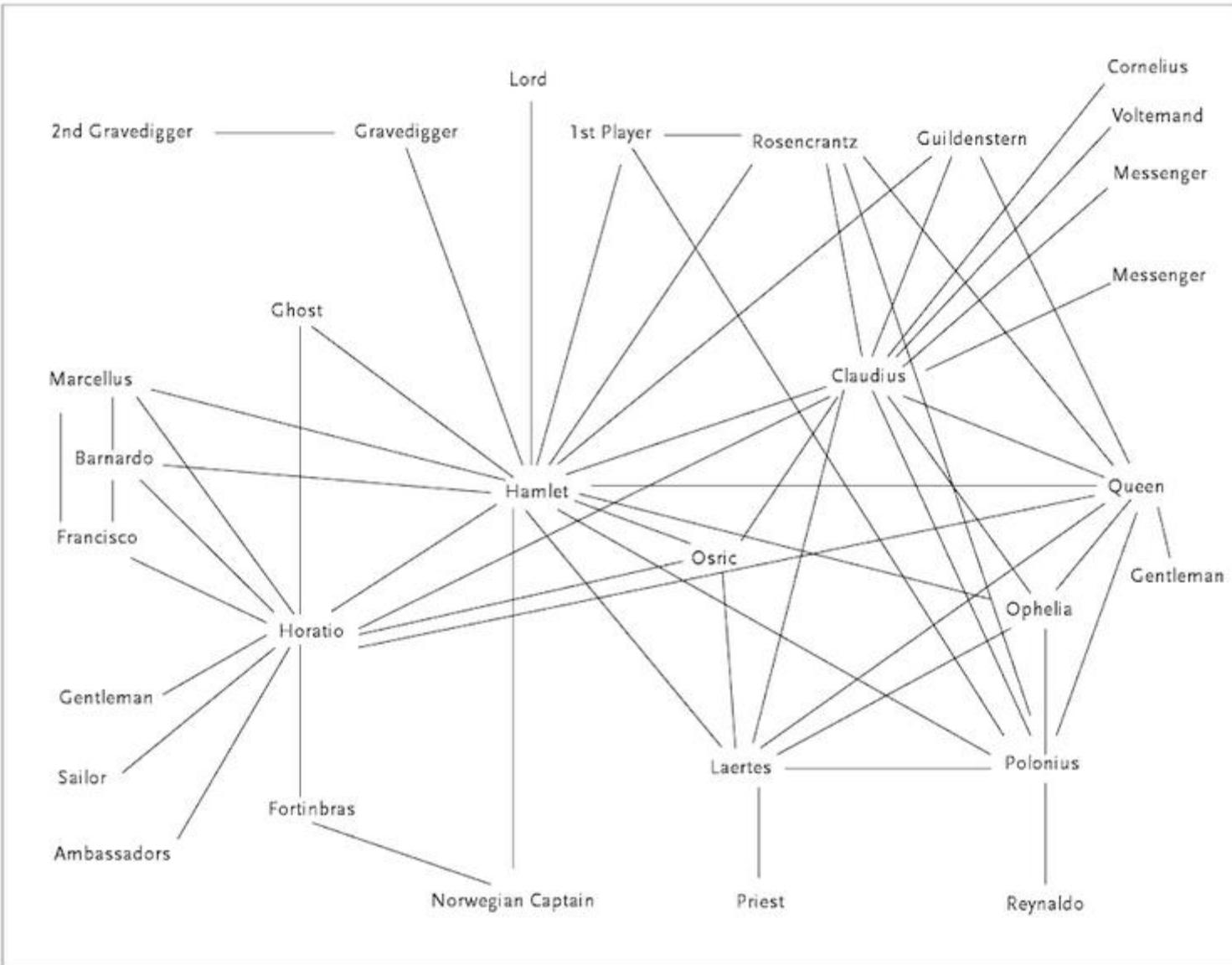


Figure 1: The Hamlet network

Moretti, F. (2013). *Distant reading*.

どういう分析をするか

② 比べる

- 頻度データ
 - 分布の比較： χ^2 二乗検定
 - 頻度の比較：尤度比検定
- 数値データ：
 - 各種パラメトリック・ノンパラメトリック検定
- 注意すべきこと
 - 検定を行う必要/必然性があるか
 - 何と何を比較しているか
 - 「母集団」に相当するものはなにか
 - 記述統計で十分な場合もある
 - 検定の多重性

どういう分析をするか

★多重検定 multiple-test

- 検定を繰り返すことによって、研究全体の**第Ⅰ種の過誤**の確率が増大してしまうこと
 - 第Ⅰ種の過誤 Type I error / α error
 - 本当は有意な差がないのに有意な差があると判断してしまう
 - 第Ⅱ種の過誤 Type II error / β error
 - 本当は有意な差があるのに有意な差がないと判断してしまう
- 真の差がなくても、たくさん検定をすれば「どこか」では統計的に有意な結果ができる
 - 擬陽性(false positive)な結果
 - 再現性のない結果を量産する一因
 - p-hacking: 納得のいく結果が出るまで「試行錯誤」を続ける
 - cherry-picking: 都合のいい結果だけを報告する
 - いずれも科学的には**無意味**な結果

- 有意確率
 - e.g. 「有意水準0.05で有意な差がみられた」
 - 帰無仮説のもとでそのデータが得られる確率
 - ×帰無仮説が正しい確率
- 真の差がなくても有意だと判定される結果を含む確率

$$\alpha_{total} = 1 - (1 - \alpha)^n$$

<i>p</i>	<i>n</i>	total <i>alpha</i>	<i>p</i>	<i>n</i>	total <i>alpha</i>
0.05	1	0.050	0.01	1	0.010
	2	0.098		2	0.020
	5	0.226		5	0.049
	10	0.401		10	0.096
	100	0.994		100	0.634

★HARKing

- **Hypothesizing After the Results are Known** (Kerr, 1997)
 - 結果が分かってから結果に合うような仮説をひねりだす行為
 - p-hackingと相性がいい
 - 当然結果の再現性は低い
 - 何がまずいか
 - 単なる第一種の過誤に過ぎないものが「理論」化される
 - 無価値な情報しか伝えていない
 - 統計を悪用してお墨付きを与えてている
 - 科学の実践として悪い例である ...etc.

どういう分析をするか

★多重検定対策

- 個別の分析では
 - 多重性を考慮した分析をする
 - 多重比較 multiple comparison (e.g., Tukey's HSD)
 - Bonferroniの補正
 - 比較の数を減らす
 - 次元（変数）を減らす
- 研究全体では
 - 事前にどういう分析をするか決めておく
 - 研究をpreregisterする
 - 行った分析をすべて記述する
 - 探索的な部分は正直にいう
 - 様々な角度から結果の妥当性を検証する
 - その差は実質科学的に意味のある差なのかをチェックする
 - e.g. 効果量のチェック
 - 妥当性をチェックする
 - 交差検証をする：分析用と検証用にデータに分割する
 - 追試をする：再現できることを確認する

どういう分析をするか

③まとめる

- 似たような性質を持つデータ (=行) をまとめたい
 - クラスター分析
- 似たような性質を持つ変数 (=列) をまとめたい
 - 次元削減
 - 主成分分析
 - 因子分析

どういう分析をするか

クラスター分析

- クラスター分析 cluster analysis
 - データ間の「距離」または「類似度」をもとに、データの集まり（クラスタ）を抽出する分析手法
 - 階層的手法
 - 距離の近いデータ同士からボトムアップにクラスタを統合していく
 - 欠点：データが多いと計算時間が膨大になる
 - 非階層的手法
 - K-means法
 - 計算時間が比較的少なくて済む
 - 欠点：一意に定まらない
 - 変数選択の問題：どの変数を使うか
 - みにくいアヒルの子の定理(Watanabe, 1969)：変数を増やすとどれも同じ程度似てしまう

どういう分析をするか

次元削減

- 複数の変数（次元）をデータの性質を保ったまま少ない変数で表現する
 - 主成分分析 **principal component analysis; PCA**
 - 複数の変数を数個の「主成分」に合成する
 - 主成分：データをよく説明する合成スコア
 - データの分散をもっともよく説明する軸（第1主成分）から順に直行するように軸を抜き出していく
 - 因子分析 **factor analysis**
 - 複数の変数をいくつかの「因子」に分解する
 - 因子：観測変数の背後にある潜在的な変数(e.g. 「知能」)
 - 全体に共通する因子+誤差、というモデル
 - 主成分分析とは想定するモデルが異なる
 - 因子回転の手法・解釈に任意性がある

どういう分析をするか

④分類する

- **機械学習 machine learning**
 - 機械が分類や予測などのタスクの成績をデータをもとに（自動的に）改善していく技術
- 教師あり学習：データとともに分類ラベルを与えて学習
 - 感情分析
 - 決定木分析
 - ナイーブベイズ
 - サポートベクターマシン(SVM)
- 教師なし学習：データのみから学習
 - 主成分分析
 - クラスター分析
 - 潜在意味解析

どういう分析をするか

感情分析

- 感情分析 **sentiment analysis**

- 極性語と呼ばれる単語をもとにスコアを算出
- 単純に計算する場合と、少数のデータに学習させてタグ付けする場合がある。
- 心理学的な妥当性は疑問(Basely and Mason, 2005; Panger, 2016)

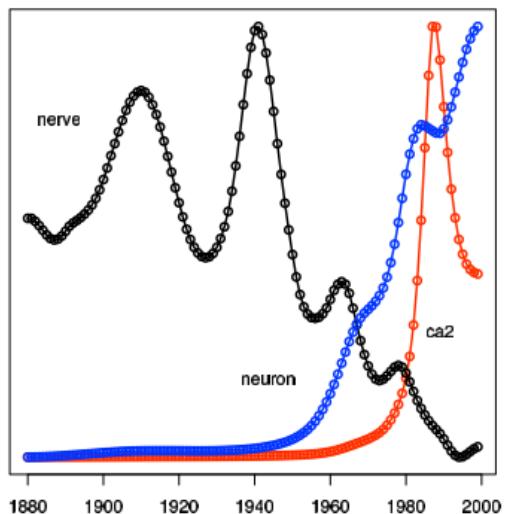
どういう分析をするか

★潜在意味解析

- Latent Semantic Analysis(LSA) / Latent Semantic Indexing(LSI)
 - 単語の生起頻度をもとに似た意味を持つ単語のグループを抽出
 - ざっくりいえば単語を主成分分析にかけるようなもの
- トピックモデル topic model
 - より統計的に洗練された手法
 - 文書群の背景にある「トピック」と、各文書がどれくらいそのトピックに該当するかを同時に推定

1881 brain movement action right eye hand left muscle nerve sound	1890 movement eye right hand brain left action muscle sound muscle sound experiment	1900 brain eye movement sound nerve active muscle left hand nerve vision sound	1910 movement brain sound nerve active nerve stimulate muscle left eye right nervous	1920 movement sound muscle active nerve stimulate fiber reaction brain response	1930 stimulate muscle sound movement response nerve frequency fiber active brain	1940 record nerve stimulate response muscle electrode active brain fiber potential	1950 respons record stimulate nerve muscle active frequency electrode potential study	1960 response stimulate record condition nerve muscle active potential stimulus nerve subject eye	1970 respons cell potential stimul neuron active nerve eye record abstract	1980 cell channel neuron response active brain stimul muscle system nerve receptor	1990 cell channel neuron ca2 active brain receptor muscle respons current	2000 neuron active brain cell fig response channel receptor synapse signal
---	---	--	--	---	--	--	---	---	--	---	---	--

"Neuroscience"



- 1887 Mental Science
 1900 Hemianopsia in Migraine
 1912 A Defence of the ``New Phrenology''
 1921 The Synchronal Flashing of Fireflies
 1932 Myoesthesia and Imageless Thought
 1943 Acetylcholine and the Physiology of the Nervous System
 1952 Brain Waves and Unit Discharge in Cerebral Cortex
 1963 Errorless Discrimination Learning in the Pigeon
 1974 Temporal Summation of Light by a Vertebrate Visual Receptor
 1983 Hysteresis in the Force-Calcium Relation in Muscle
 1993 GABA-Activated Chloride Channels in Secretory Nerve Endings

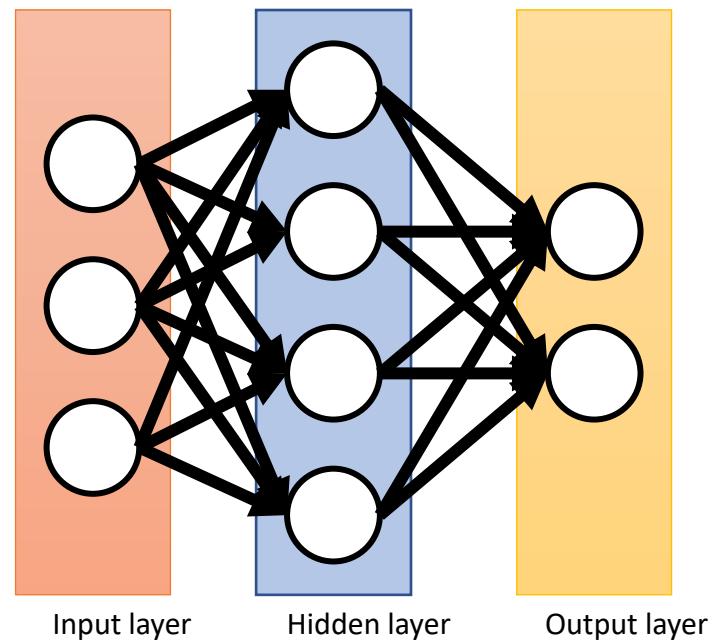
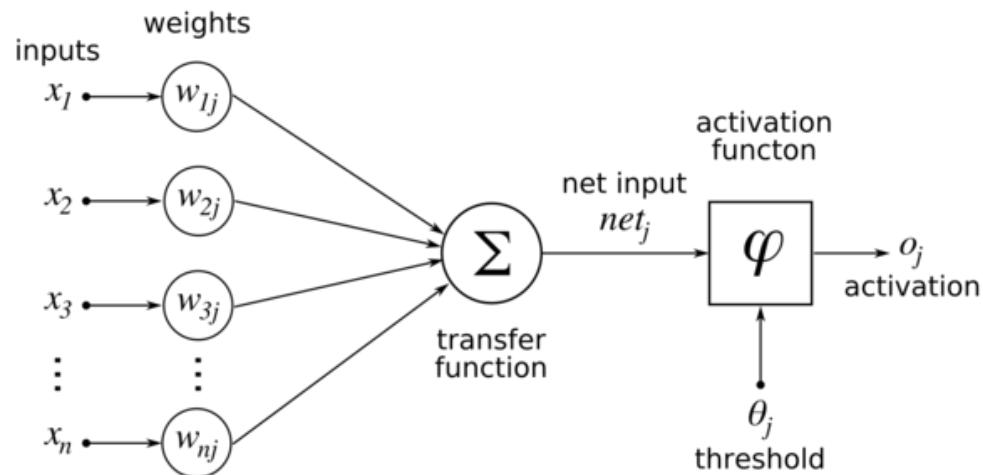
Figure 4. Examples from the posterior analysis of a 20-topic dynamic model estimated from the *Science* corpus. For two topics, we illustrate: (a) the top ten words from the inferred posterior distribution at ten year lags (b) the posterior estimate of the frequency as a function of year of several words from the same two topics (c) example articles throughout the collection which exhibit these topics. Note that the plots are scaled to give an idea of the shape of the trajectory of the words' posterior probability (i.e., comparisons across words are not meaningful).

Blei, D. M., & Lafferty, J. D. (2006). Dynamic Topic Models. In *Proceedings of the 23rd international Conference on Machine Learning*, 113–120)

どういう分析をするか

ニューラルネット

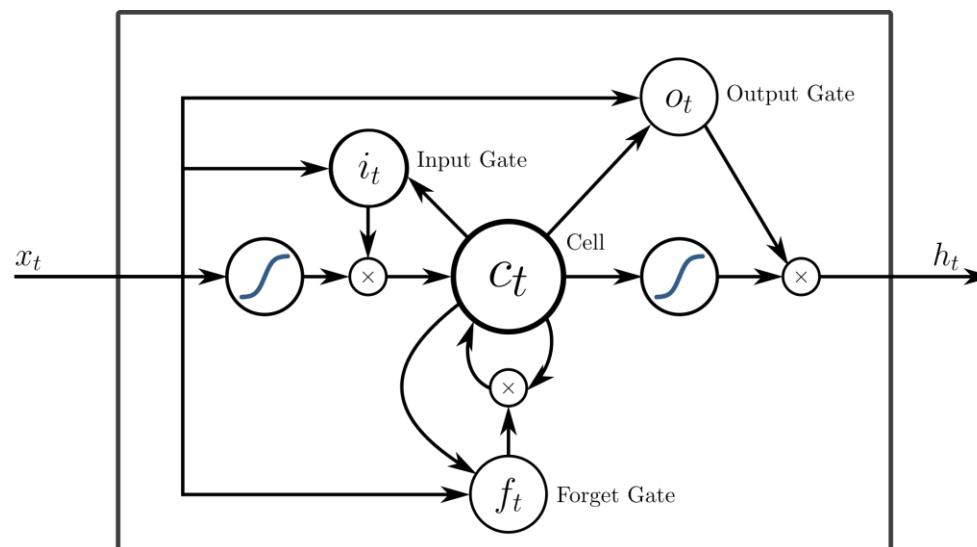
- ニューロンを模した学習器を多数組み合わせて学習させる
 - 分類や生成、様々なタスクに応用できる

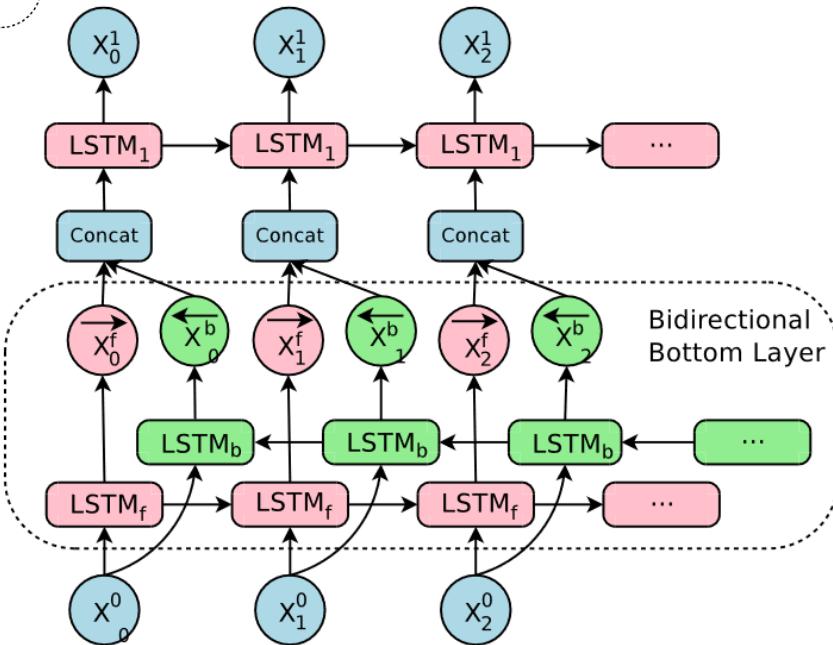
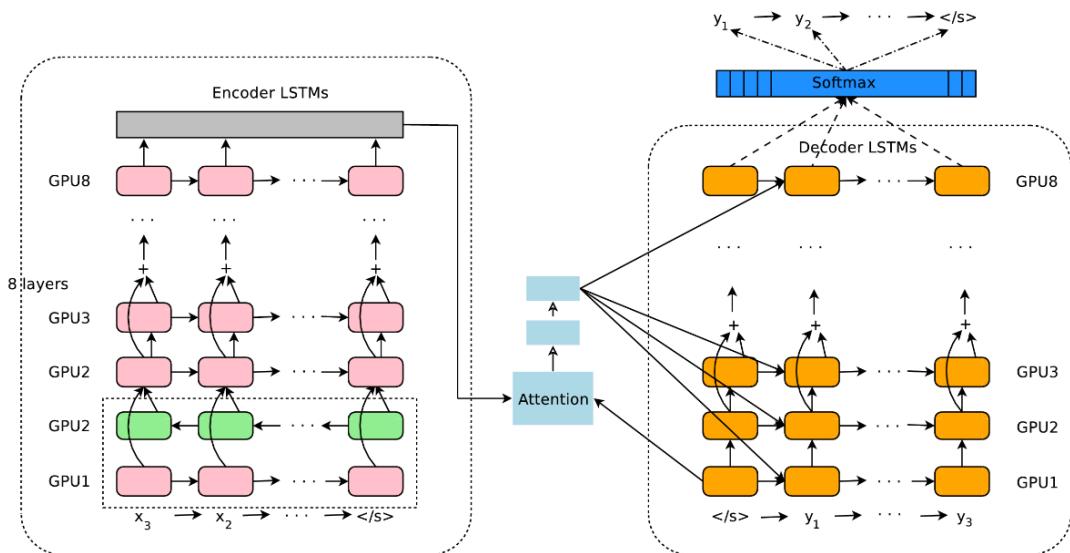


★LSTM

- LSTM: Long-Short Term Memory

- 可変長の引数を扱える→文章などの データを扱える
- 複数のLSTMを組み合わせて自動翻訳を強化





Wu, Y., Schuster, M., Chen, Z., Le, Q. V., Norouzi, M., Macherey, W., ... Dean, J. (2016). Google's Neural Machine Translation System: Bridging the Gap between Human and Machine Translation. <http://arxiv.org/abs/1609.08144>

★word2vec

- 単語の分散表現（単語ベクトル）を学習
 - 分散仮説(Firth, 1957)
 - 「言葉の意味は周辺の語彙によって決まる」
 - 学習モデル
 - CBoW
 - Skip-gram
- 単語の加減算ができる

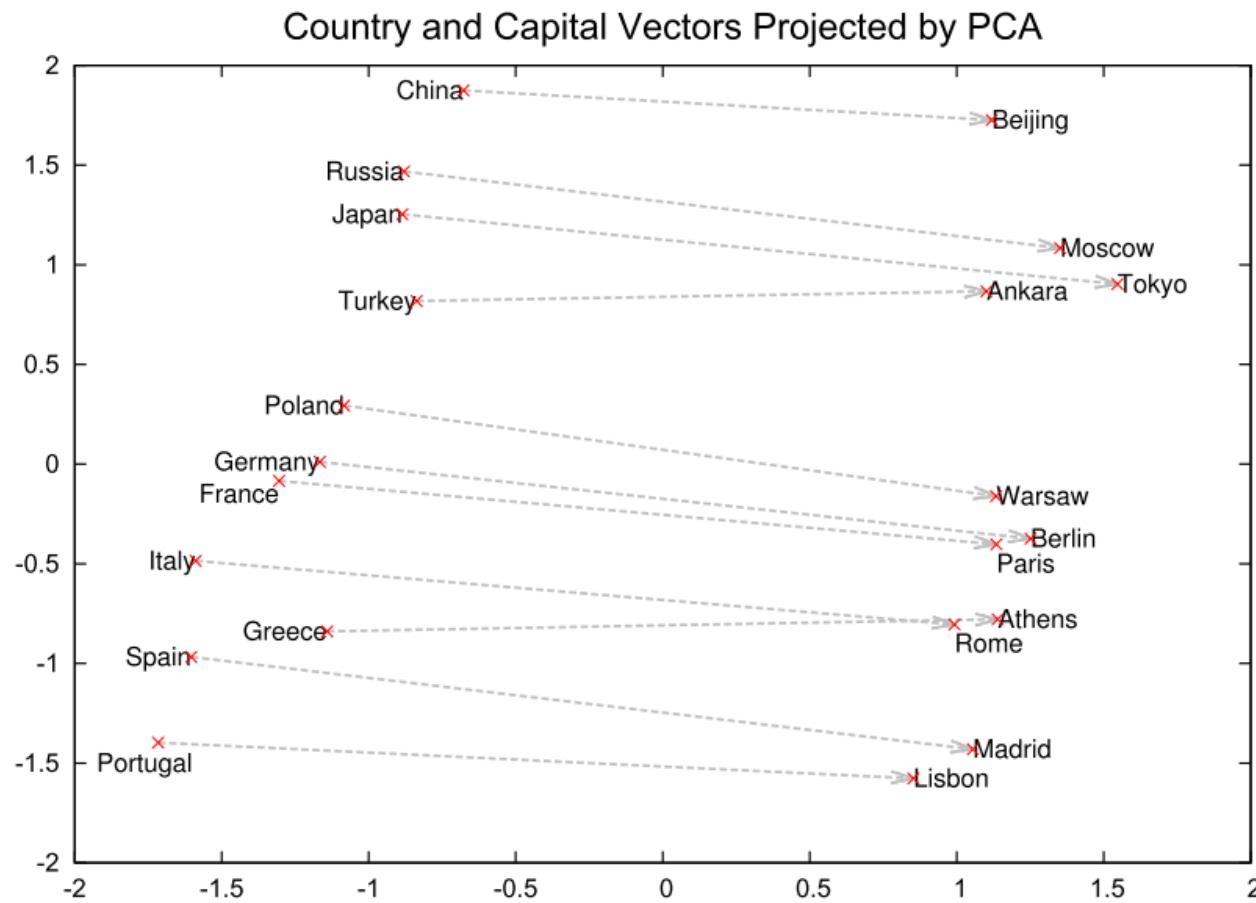


Figure 2: Two-dimensional PCA projection of the 1000-dimensional Skip-gram vectors of countries and their capital cities. The figure illustrates ability of the model to automatically organize concepts and learn implicitly the relationships between them, as during the training we did not provide any supervised information about what a capital city means.

Mikolov, T., Sutskever, I., Chen, K., Corrado, G., & Dean, J. (2013). Distributed Representations of Words and Phrases and their Compositionality. In *Advances in Neural Information Processing Systems* (pp. 3111–3119).

どういう分析をするか

分析のまとめ

- テキストを分析する選択肢は多い
 - すべてを理解する必要はない
 - 「こういう技術がある」として知っておくとよい
- 大量の変数, 大量の分析
 - 自由度が高すぎるために何をしていいかわからない
 - たくさん検定をやればどこかには統計的に「有意な」結果が存在してしまう → 多重検定
 - 統計的な検定だけにこだわらず、データ可視化手法などといった記述的な手法も視野に入れる
 - 何が知りたいことなのか、**研究デザインの段階**でよく考えておく

研究の各種制約

- 倫理
 - 通常の心理学研究の倫理綱領にもとづく(cf. 日本心理学会, 2011)
 - ビッグデータの場合はさらに注意が必要
 - 個人情報の管理は適正に
- コスト
 - 予算
 - インセンティブを使えるか（実験・調査）
 - 時間
 - 収集にかかる時間
 - 収集後の処理にかかる時間
 - データ入力
 - クリーニング・前処理
 - PCの計算時間
- その他ロジスティクス上の問題
 - 使用できるPCの性能
 - CPU
 - メモリ
 - 利用できるデータ容量 etc.

小まとめ

- 実験・調査に限らず様々な方法でテキストデータを収集することができる
 - オンラインで得られるデータは通常の実験・調査とは異なる種類の性質がある
 - 落とし穴にはまらないよう、研究デザインをしっかり立てる
- テキストデータの分析手法は多様である
 - 自由度が高い分、定型的な手法というものがない
 - 行動データではない→何を測っているのか常に意識する
 - どうすれば測りたいことを測れるのかを考える

References

- Abello, J., Broadwell, P., & Tangherlini, T. R. (2012). Computational folkloristics. *Communications of the ACM*, 55(7), 60–70. <https://doi.org/10.1145/2209249.2209267>
- Back, M. D., Küfner, A. C. P., & Egloff, B. (2011). “Automatic or the people?” Anger on september 11, 2001, and lessons learned for the analysis of large digital data sets. *Psychological Science*, 22(6), 837–838. <https://doi.org/10.1177/0956797611409592>
- Back, M. D., Küfner, A. C. P., & Egloff, B. (2010). The Emotional Timeline of September 11, 2001. *Psychological Science*, 21(10), 1417–1419. <https://doi.org/10.1177/0956797610382124>
- Beasley, A., & Mason, W. (2015). Emotional States vs. Emotional Words in Social Media. In *Proceedings of the ACM Web Science Conference on ZZZ - WebSci ’15* (pp. 1–10). New York, New York, USA: ACM Press. <https://doi.org/10.1145/2786451.2786473>
- Blei, D. M., & Lafferty, J. D. (2006). Dynamic Topic Models. In *Proceedings of the 23rd international Conference on Machine Learning* (pp. 113–120). <https://doi.org/10.1145/1143844.1143859>
- Chetty, R., Hendren, N., Kline, P., & Saez, E. (2014). Where is the land of opportunity? the geography of intergenerational mobility in the United States, 129(November), 1553–1623.
- Firth, J. R. (1957). A Synopsis of Linguistic Theory 1930-1955" in Studies in Linguistic Analysis. *The Philological Society*.
- Grimes, D. A., & Schulz, K. F. (2002). Bias and causal associations in observational research. *The Lancet*, 359, 248–252.

- Healy, K. (2015). The Performativity of Networks. *Archives Européennes de Sociologie*, 56(2), 175–205. <https://doi.org/10.1017/S0003975615000107>
- Hebb, D. (1949). The Organization of Behavior. New York. Wiley.
- King, G., Pan, J., & Roberts, M. E. (2013). How censorship in China allows government criticism but silences collective expression. *American Political Science Review*, 107(2), 326–343. <https://doi.org/10.1215/0003089X-107-2-326>
- LaPiere, R. (1934). Attitudes vs. Actions. *Social Forces*, 13(2), 230–237.
- McCulloch, W. S., & Pitts, W. (1943). A logical calculus of the ideas immanent in nervous activity. *The Bulletin of Mathematical Biophysics*, 5(4), 115–133.
- Michel, J.-B., Shen, Y. K., Aiden, A. P., Veres, A., Gray, M. K., Pickett, J. P., ... Aiden, E. L. (2011). Quantitative Analysis of Culture Using Millions of Digitized Books. *Science*, 331(6014), 176–182. <https://doi.org/10.1126/science.1199644>
- Mikolov, T., Sutskever, I., Chen, K., Corrado, G., & Dean, J. (2013). Distributed Representations of Words and Phrases and their Compositionality. In *Advances in Neural Information Processing Systems*
- Minsky, M., & Papert, S. (1969). Perceptron: an introduction to computational geometry. *The MIT Press, Cambridge, Expanded Edition*, 19(88), 2.
- Orne, M. T. (1962). On the social psychology of the psychological experiment: With particular reference to demand characteristics and their implications. *American Psychologist*, 17(11), 776.
- Panger, G. (2016). Reassessing the Facebook experiment: critical thinking about the validity of Big Data research. *Information Communication and Society*, 19(8), 1108–1126. <https://doi.org/10.1080/1369118X.2015.1093525>
- Pury, C. L. S. (2011). Automation can lead to confounds in text analysis: Back, Küfner, and Egloff (2010) and the not-so-angry Americans. *Psychological Science*, 22(6), 835–836. <https://doi.org/10.1177/0956797611408735>
- Reagan, A. J., Mitchell, L., Kiley, D., Danforth, C. M., & Dodds, P. S. (2016). The emotional arcs of stories are dominated by six basic shapes. *EPJ Data Science*, 5(1), 31. <https://doi.org/10.1140/epjds/s13688-016-0093-1>

- Rosenblatt, F. (1958). The perceptron: a probabilistic model for information storage and organization in the brain. *Psychological Review*, 65(6), 386–408.
<https://doi.org/10.1037/h0042519>
- Rosenthal, R. (1966). Experimenter effects in behavioral research.
- Rothman, K. J. (2012). *Epidemiology: an introduction*. Oxford university press.
- Sackett, D. L. (1979). Bias in analytic research. *Journal of Chronic Diseases*, 32(1–2), 51–63. Retrieved from <http://www.ncbi.nlm.nih.gov/pubmed/447779>
- Salganik, M. J. (2017). *Bit by bit: social research in the digital age*. Princeton University Press.
- Tangherlini, T. R. (2016). Big Folklore: A Special Issue on Computational Folkloristics. *Journal of American Folklore*, 129(511), 5–14. Retrieved from <http://search.ebscohost.com/login.aspx?direct=true&db=a9h&AN=113224879&site=ehost-live>
- Tukey, J. W. (1977). *Exploratory Data Analysis*. Addison-Wesley.
- Ugander, J., Karrer, B., Backstrom, L., & Marlow, C. (2011). The Anatomy of the Facebook Social Graph, 1–17. Retrieved from <http://arxiv.org/abs/1111.4503>
- Watanabe, S. (1969). Knowing and Guessing a Quantitative Study of Inference and Information.
- Webb, E. J., Campbell, D. T., Schwartz, R. D., & Sechrest, L. (1966). Unobtrusive measures: Nonreactive research in the social sciences.
- Wu, Y., Schuster, M., Chen, Z., Le, Q. V., Norouzi, M., Macherey, W., ... Dean, J. (2016). Google's Neural Machine Translation System: Bridging the Gap between Human and Machine Translation. Retrieved from <http://arxiv.org/abs/1609.08144>
- 村上征勝. (2002). 文化を計る 文化計量学序説. 朝倉書店.
- 橋口耕一. (2006). 内容分析から計量テキスト分析へ--継承と発展をめざして. 大阪大学大学院人間科学研究科紀要, 32, 1–27. <https://doi.org/info:doi/10.18910/11920>
- 橋口耕一. (2014). 社会調査のための計量テキスト分析 内容分析の継承と発展を目指して. ナカニシヤ出版.
- 橋口耕一. (2018). 計量テキスト分析およびKH Coderの利用状況と展望. 社会学評論, 68(3), 334–350.