数学特別講義 ： 大量の非定型データからマイニング

池田, 大輔
九州大学システム情報科学研究院准教授

九州大学
KYUSHU UNIVERSITY

大量の非定型データからのマイニング

国立大学法人 九州大学
大学院システム情報科学研究院
情報学部門 知能科学講座

池田 大輔

daisuke@inf.kyushu-u.ac.jp

2010.11.8〜11 集中講義@北大．理．数

九州大学
KYUSHU UNIVERSITY

研究室紹介

2010.11.8〜11 集中講義@北大．理．数

自己紹介

池田 大輔

いけだ　　だいすけ

九州大学 理学部 物理学科 卒業

九州大学 総合理工学研究科 情報システム学専攻

九州大学 大型計算機センター 助手、講師

- (後の情報基盤センター→情報基盤研究開発センター)

九州大学 附属図書館 助教授

九州大学 システム情報科学研究院 准教授

背景：情報コミュニケーションの変化

いつでもどこでもアクセス可能

双方向のコミュニケーション

マスではなく利用者ベースのメディア

社会に大きな変化を与えている

特に学術情報のコミュニケーション基盤

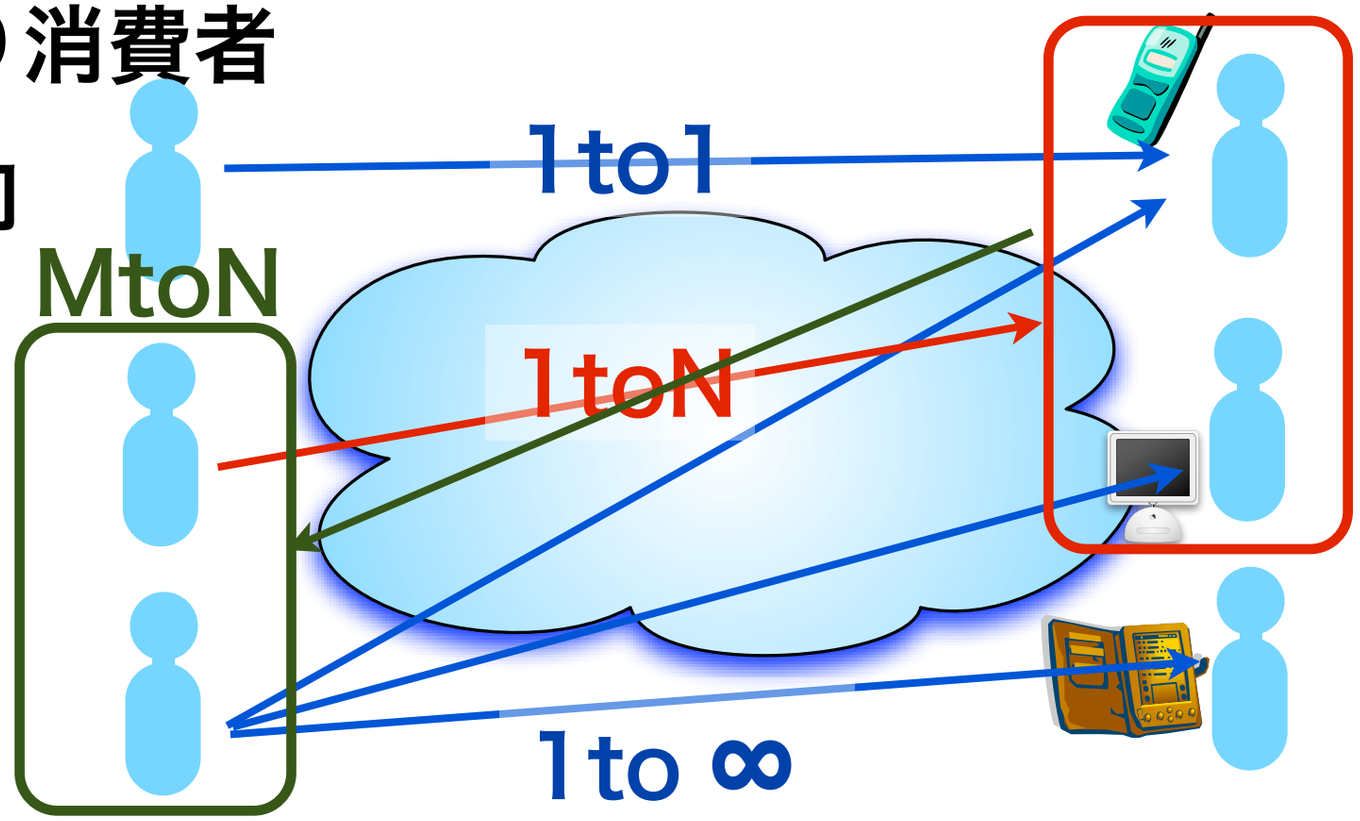
研究者は情報の生産者であり消費者

情報の方向：一方向／双方向

人数：様々な粒度が存在

情報消費のスピードが速い

→ 新たな知の創発



研究における情報コミュニケーションが
今後の情報コミュニケーション基盤研究の
題材として好適である

研究テーマ：(学術)情報基盤の構築

情報基盤における(自動的な)知識発見・マイニング(Mining)

- データから有用な情報や知識を機械的に抽出するアルゴリズムの構築。
- 例：
  音楽データ自身の類似度を計算し、似た曲を探す。
  曲の購入履歴を元に「この曲を買った人は...」と推薦する。
  マイクロブログ(twitterなど)データから、こんな曲が流行る、と予想する。
  曲のレビュー掲示板における「荒し」を検出する。

情報基盤そのものに関する研究

- コンテンツが集るシステム作り。
- 例：
  曲の購入履歴が残らず、プライバシー保護が可能な認証システム。
  グループで簡単に曲に関する感想をシェアするシステム。
  大量の曲を保存し、素早く検索できるシステム。

授業概要

この集中講義では、
Web上のデータを対象にしたマイニングを説明します。
まず、①計算機科学の基本的な概念の説明と、
講義全体を概観し、
次に、②マイニングの基礎となるデータベースや
情報検索、機械学習の手法を説明します。
その後、③スパム検出を題材にマイニングの基本を学び、
最後に④これらの技術に関する最新動向を紹介します。

- キーワード：テキストマイニング、Webマイニング、アルゴリズム、データモデル、データ構造

授業目的

マイニングの手法は、
統計的な手法など、古くからある手法も用いられますが、
大きな違いは莫大なデータを扱うことです。
単純な方法では、現実的な時間での処理ができませんので、
様々な工夫が必要です。
この講義で、計算機科学における処理の一連のプロセス、
すなわち、対象の抽象化や単純化（モデリング）、
データ構造、アルゴリズム等における工夫を理解し、
理論が実用に与える影響を体感してもらいます。

講義計画

8日4限 (1)導入：マイニングの例、
(2) 計算と計算量：効率の定義
実習：環境構築と簡単なサンプルプログラム

9日3限 大量データの蓄積と利用：データベースと情報検索

9日4限 実習：検索におけるデータ構造とアルゴリズム

10日3限 データの活用：マイニングへ

10日4限 実習：機械学習とマイニングによるスパム検出

11日3限 情報検索、DB、マイニングに関する最新動向

- コンピュータ囲碁、Googleにおける大量データ処理基盤、インデックスレスの検索システムShunsaku等

導入：マイニングの例

このコマの目的

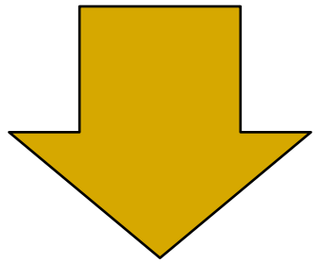
例を通してマイニングの概略を理解する。

マイニングの普及の歴史的な背景を理解する。

計算の効率を測る(漸近的)計算量理論の基礎を理解する。

科学をささえるもの

理論・実験



理論・実験・シミュレーション

第4の柱：data intensive science

NASAの極超音速実験機 X-43Aまわりの圧力分布の可視化結果。赤色部分が高圧となっている。

データ増加の裏側：データ取得機会の増加

- センサー、データセンシング技術の向上、衛星等による観測データの充実。
- 自動認識技術による物体識別子のデジタル化。
  - バーコード、二次元バーコード、RFIDタグ、ICカードなど。
- 大量高速な計算により、新たなデータが産み出されるようになった。
- 業務のICT化、Web化により、そもそもデータがボーンデジタルになった。
  - 顧客の注文がWebフォームから。
- Web、メール、ブログ、マイクロブログ等のコミュニケーションも充実。

図書に貼付されたバーコードとRFIDタグ

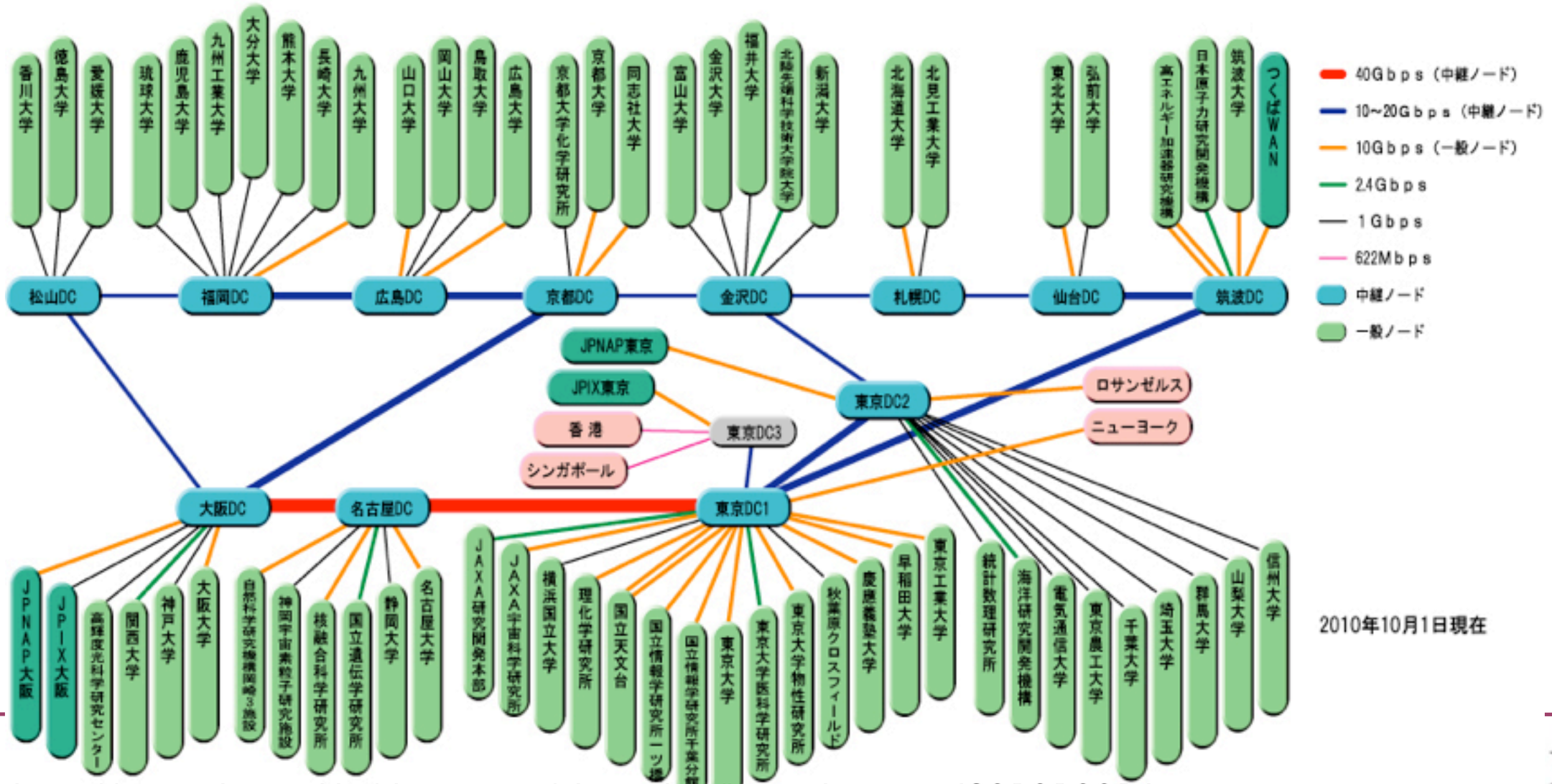

RFIDリーダで複数の図書を読み取り

RFIDリーダ組み込みの書架

データ増加の裏側：データ転送基盤の充実

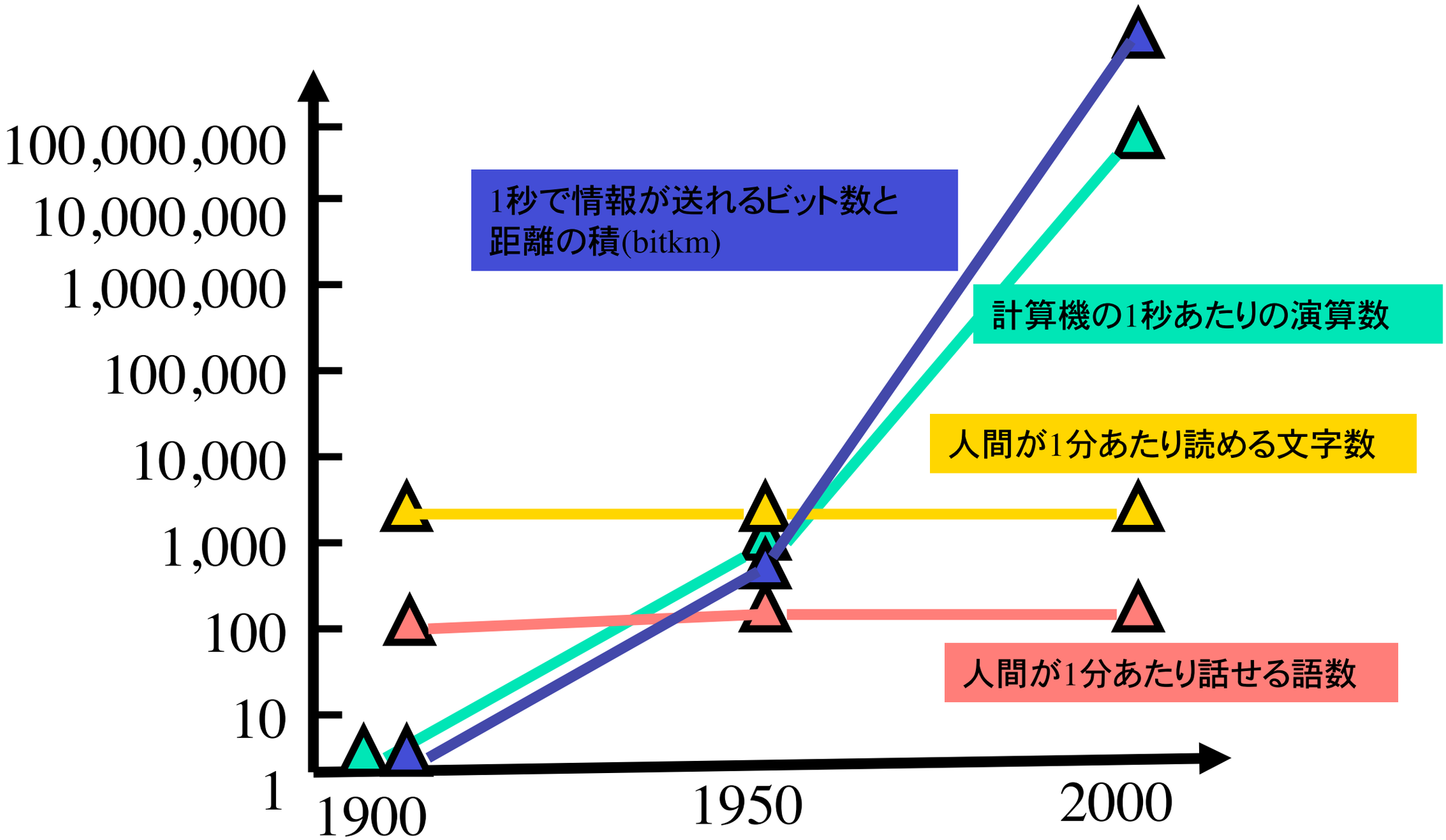
インターネットによるデータ転送が可能に。

- 観測データ、POSデータ、テキストデータ等を独自のインフラなしで自由に安価にやりとりできる。



http://www.sinet.ad.jp/document/sinet_doc/kousei_syousai20101001.jpg

データの増加量



九州大学安浦寛人理事より

(データ)マイニング普及の裏側(1/2)

計算機の高性能化。

- CPUに加え、メモリやディスクの大容量化により大量データ処理が現実的に。
  - ムーアの法則はまだ健在。
- マルチコア等による並列的な処理が可能で、より大量データ処理が可能に。

要素技術の（理論的な）研究が進んでいた。

- ベイズ統計、機械学習、パターン認識等。

プログラミング言語、ライブラリ、ツール等の普及。

- Perl等の言語、統計解析用のツール(RやOctave)、機械学習のライブラリ。

(データ)マイニング普及の裏側(2/2)

インターネットによるデータ解析結果の送信。

- 一昔前のディスクへのアクセス速度と変わらない(くらいの)速さ。
- 個々のコンピュータで計算する必要がなく、結果を受けとるだけでもよい。
- →クラウド化。

社会インフラの中抜き化(効率化)。

- コンビニやamazon.com等。

point of sales

データマイニングの例：POSデータ

典型的な関係データベースからの知識発見

- Knowledge Discovery from Database (KDD)
- (関係)データベースは明日詳しく説明する予定。

商品に関する属性、顧客の属性等をデータベースに格納。

- 商品情報はバーコードにより入力。
- 顧客の情報はポイントカード、ICカード等により取得。
- インターネットを経由して、逐次、データは集められる。
- データは在庫管理、顧客管理、販促キャンペーン等の決定に利用される。

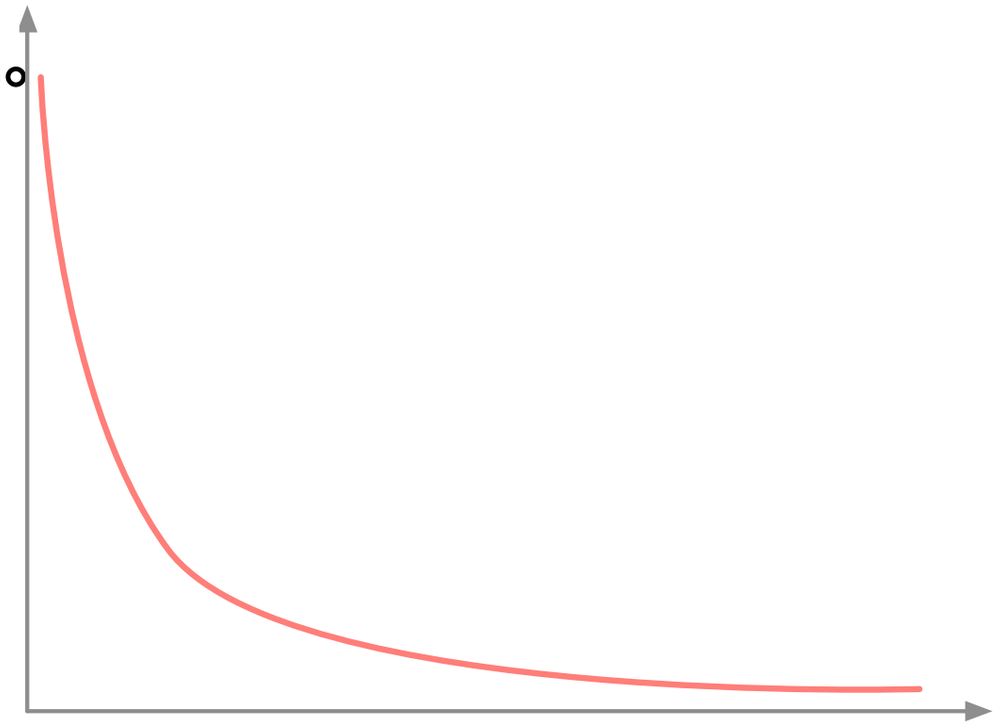
例：売れる商品の組み合わせ、顧客層と商品の組み合わせ等

- 週末の夜にビールとおむつの組み合わせがよく売れる等。

具体例：amazon.com

本を主な対象にしたe-コマース最大手。

- 顧客の購買データを最大限に活用し、アイテムの推薦を行う。
  - 「このページを見た人は」「この本を買った人は」
- 日本では、(雑誌を除いて)本が売れない時代と言われていたのに対し、ロングテールに着目し、客単価を3倍に上げた。
  - 他の店より3倍お金を落としてくれるような仕組みを作り、送料無料が可能に。
- 参考文献：
  アマゾン・ドット・コムの光と影：潜入ルポ
  横田 増生、情報センター出版局



ロングテール

データ活用の特徴

- ユーザのレビューを収集。
- アフェリエイトにより、外部サイトでの販売とアクセスの確保。
- ページの閲覧履歴もチェック。

マイニングの例：ゲノム配列

様々な種のゲノム配列がスキャンされ、テキストデータとして公開。

- DNA配列は{A, C, G, T}、アミノ酸配列は20種の文字からなる。

データが揃い、これから様々な解析が期待される。

- 個別の遺伝子的な違いを考慮したテーラーメイド医療など。

情報系のアルゴリズムやデータ構造、機械学習、マイニング等の専門家がbioinformaticsに参入。

- お金(助成金)も大きく入っている。
- 開発者等のコミュニティも大きいらしい。

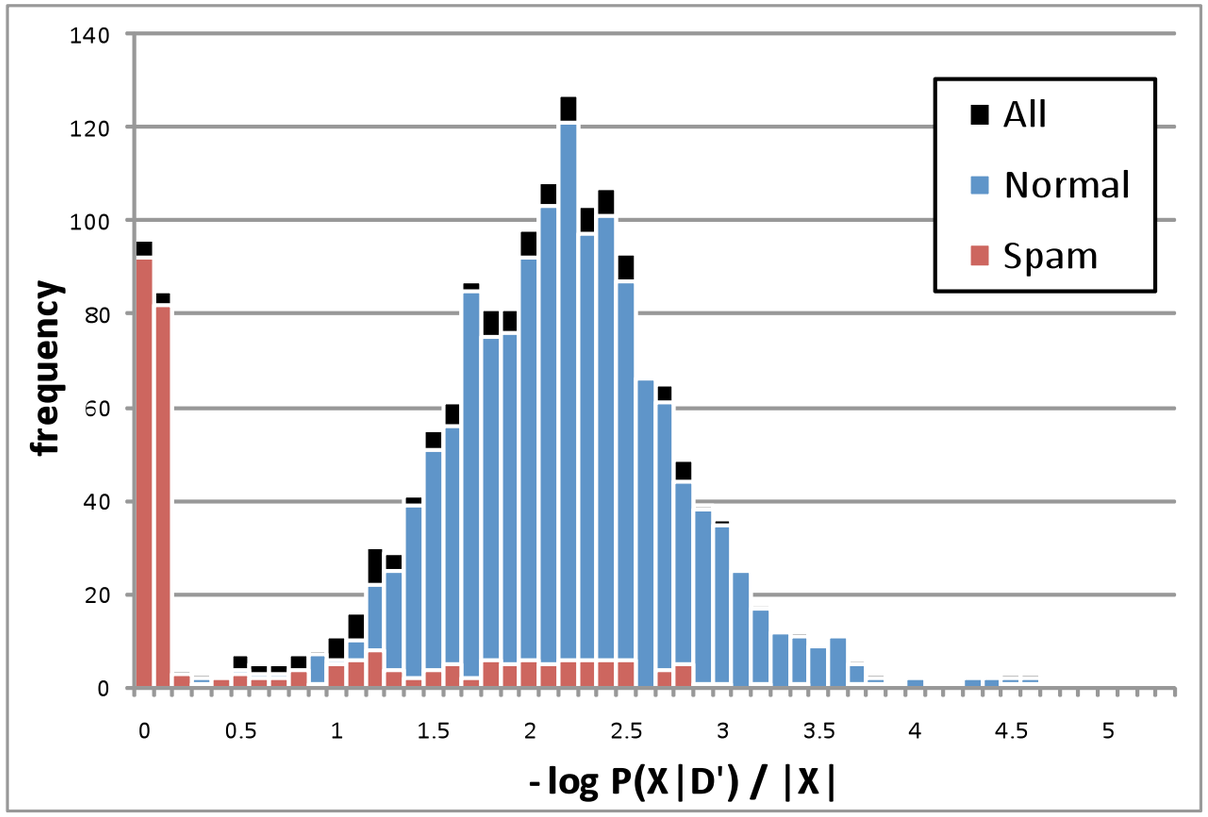
マイニングの例：テキストデータ

和歌の類似性検出

– 従来知られてなかった国文学上の発見へ。

スパム検出

– 特徴検出から不自然なテキストの検出へ。




研究紹介

コンピューターで和歌を読み解く

竹田 正幸

似た和歌を探せ

平安末期から鎌倉にかけての歌人、藤原定家に次のような和歌があります。

白妙の／衣ほすてふ／夏の来て／かきねもたわに／咲ける卯の花

「白妙の衣ほすてふ」といえば、すぐに思い浮かぶ歌がありませんか？ そう、持統天皇の作とされる百人一首のあの歌です。

春過ぎて／夏来にけらし／白妙の／衣ほすてふ／天の香具山

定家の歌は、この持統天皇の歌を下敷きにして作られました。このように、よく知られた歌をもとにして新たに歌を作ることを「本歌取り」といいます。本歌取りは、もとになる古い歌を皆が知っていることを前提にしたものなので、それと知らずに新しい歌だけを見ても、表現の深みがうまく見えてきません。

そこで、計算機を使って、表現の「似た」歌の対を取り出すことを考えました。具体的には、二つの和歌集の間で、すべ

図1：編集距離
たとえば、英単語 necessarily を neccesarry と打ち間違えたとする。図において、直線で結ばれた2文字は文字の一致を、異なった2文字が上下向き合っているものは文字の置換を、一方がハイフンで他方が文字のものは、文字の挿入あるいは削除を表す。このことは、文字 c を挿入し、文字 s を削除し、文字 l を文字 r で置き換えれば、英単語 necessarily を neccesarry へ変換できる、ということを意味している。このように、一方の文字列を他方の文字列に変換するには「挿入・削除・置換」が最低何回必要か、という観点から、二つの文字列間の非類似度を測ったものが編集距離である。この場合、編集距離は3となる。

春過ぎて 夏来にけらし 白妙の 衣ほすてふ 天の香具山 （「百人一首」2番、持統天皇）
大意：いつの間にか、春が過ぎて夏がやって来たらしい。白妙の衣を干すという、天の香具山に。

白妙の 衣ほすてふ 夏の来て かきねもたわに 咲ける卯の花 （「新選和歌」280番、藤原定家）
大意：（持統天皇の歌に詠まれた）白妙の衣を干すという夏が来て、垣根には、たわわに白い花の花が咲いている。

図2：類似度の計算
たとえば、二つの句「なつきにけらし」「なつのきて」は、いずれもパターン「なつ＊き＊」に合致する。ここで「＊」はワイルドカードといい、任意の長さの任意の文字列で置き換えてよいものとする。そして、二つの文字列の両方に合致するパターンのうち、最もスコアの高いものを考え、その値を類似度とする。そのためには、各パターンにスコアをつける仕組みが必要だ。ここでは、パターン中の文字数を数え、連続した文字の後の個数を0.9倍したものを引いた値をスコアとした。たとえば、「＊なつ＊」「なつ＊き＊」のスコアは、それぞれ、1.1、1.2となる。

著者
竹田 正幸（たけだ まさゆき）
九州大学大学院システム情報科学研究院情報理学部門助教授、博士（工学）。1989年九州大学大学院総合理工学研究科修士課程修了。同大学助手を経て、1995年より現職。科学技術振興事業団さきがけ研究21「情報と知」研究員（兼任）。専門は計算機科学で、文字列パターン照合、データ圧縮、発見科学、データマイニングなどに興味を持つ。

著者との一問一答
Q1：日頃心がけていること
A：石橋を叩いていては何も進まない。「甘い見積り」と「あげない心」が研究室のモットー。
Q2：どんな学生時代を過ごしたか
A：バブル崩壊期をひたすら遊び狂した。
Q3：ひらめきの瞬間、研究が劇的展した瞬間
A：不調が来て出張していると不意に「それ、はやってくる。その前のことは味になるとほとんど思い出せない。いたこの口癖せに似ているかもしれない。

九大広報27号
コンピューターで和歌を読み解く

ワードサラダスパムに成功!

閑話休題：図書館とのつながり

学生証・教職員証の電子マネーで利用できる自販機

自動認識技術の導入

- RFIDや全学共通ICカードの導入。
- 貸出履歴の積極的な活用：P&Pプロジェクト
  - プライバシー保護の必要性を認識。
- プライバシー保護を可能にしつつ、安全性も確保できる認証システムの開発。
  - 誰かは分からないけど、アクセスする権限がある人かどうかは確認できる。

機関リポジトリ

- インターネットとフリーのミドルウェア群(Apache, Tomcat, RDB等)を用いた学術情報に関する情報発信システム。
  - 主に大学や研究所等の機関単位で導入される。
- 学術情報は超ロングテールでは?
  - マイニング等の技術や履歴情報等の取得が重要になると予想。

マイニングの要素技術

(狭義の)データマイニングとは

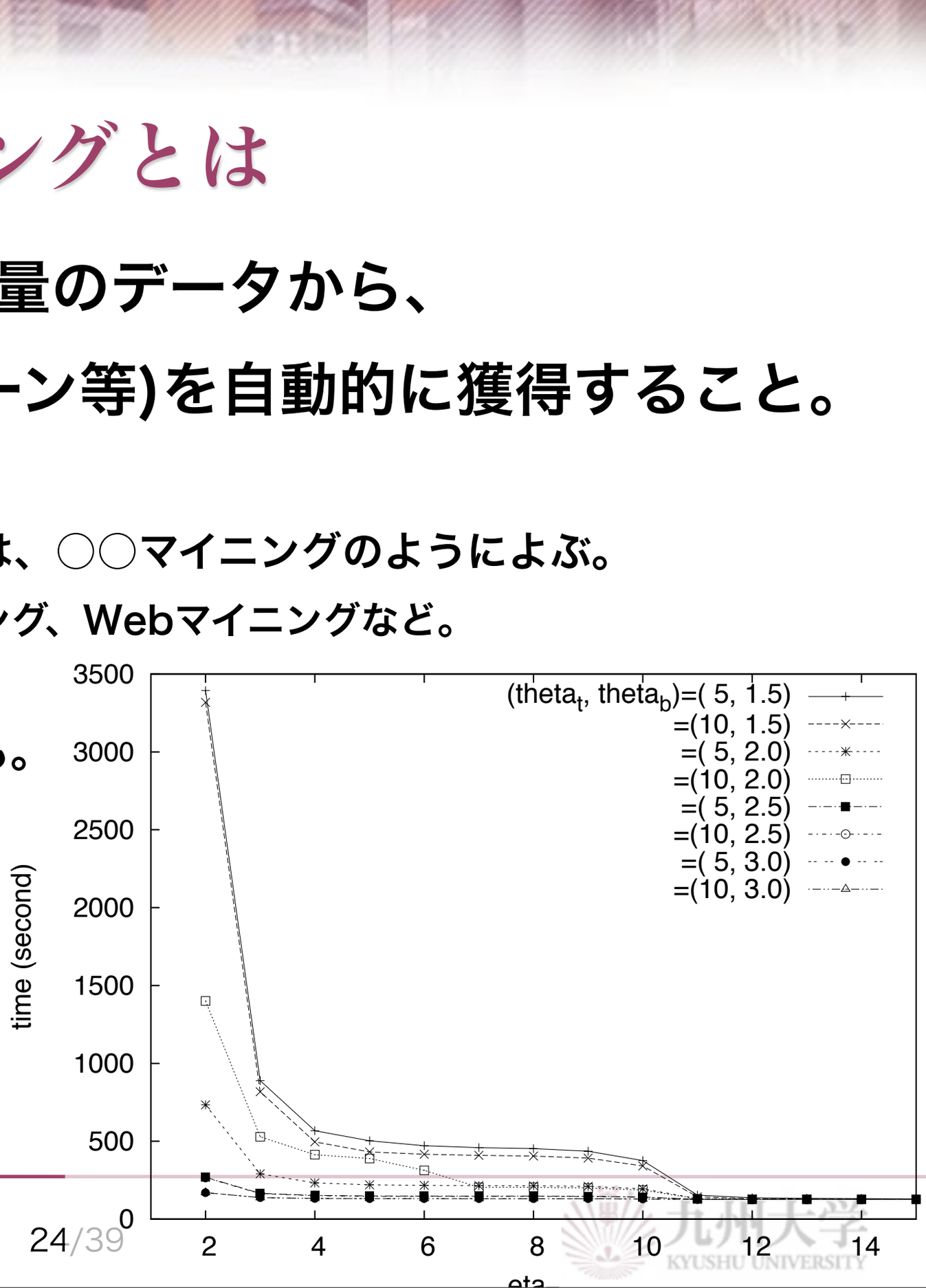
データベースに蓄積された大量のデータから、

非自明な知識(ルール、パターン等)を自動的に獲得すること。

- 基本的には関係データベースが対象。
- それ以外のデータを対象にする場合は、○○マイニングのようによぶ。
  - グラフマイニング、テキストマイニング、Webマイニングなど。
- ただし、様々なデータへの適用も
  含めてデータマイニングと呼ぶ場合も。

アルゴリズムやデータ構造に
重点が置かれる。

- データが大量である分、
  スピードに重きを置いている、とも。



マイニングの要素技術

すでに成熟していた様々な要素技術が用いられる。

- 統計学：サンプルと母集団
- 機械学習：データの識別と分類
- アルゴリズム（データ構造）：高速化と効率化
- 情報検索：大量データの蓄積と検索→明日詳しく説明
- データベース：大量データの蓄積と検索→明日詳しく説明

統計学

データ全体を端的に表す。

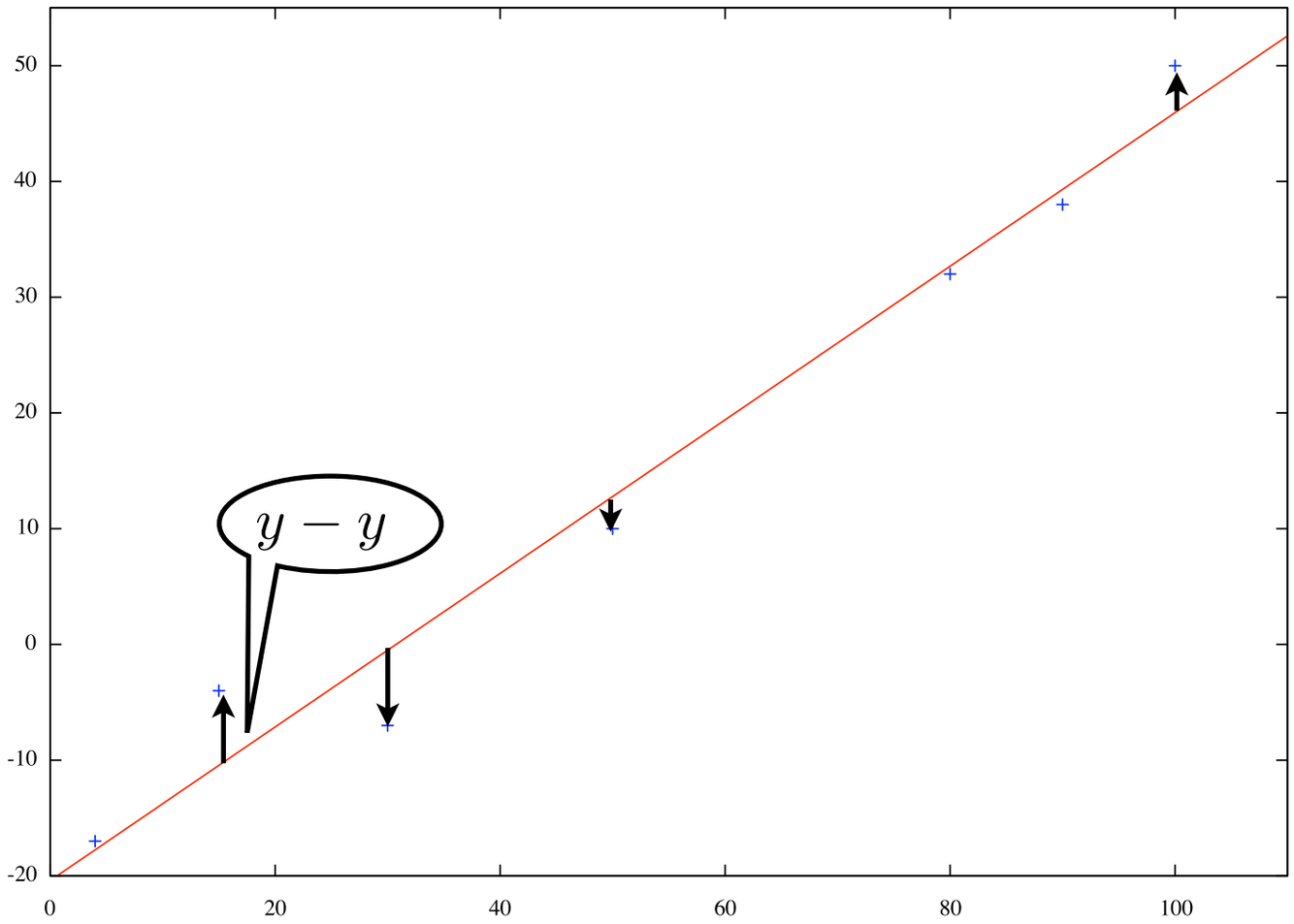
- 正規分布に従っているなら、平均値と分散で全体が分かる。

データの一部から全体を推測する。

- 回帰(regression)。

検定、ベイズ統計、
多変量解析、クラスター
分析等。

- これらの技術もマイニングの大きな要素。



機械学習：分類と識別

例の有無

- 教師あり学習、教師なし学習
  - 前者は「あるなしクイズ」、後者は「グループ分け」(明後日実習アリ)。

パターン認識（分類）

- 教師あり学習の一つ。
- 応用例にスパム検出、新たな天体の発見など。

クラスタリング

- 教師なし学習の一つ。

その他の応用例

- コンピュータ将棋、囲碁など

計算

簡単な計算モデルへ

問題：

0と1からなる(2進数の)数字の列が与えられた時 この中にある0と1の個数が同じかどうか判断する手続きを構成せよ

- **注： この数字の列は記憶できないほど非常に巨大であるとせよ**

a=1001111110101001011100010000111111110100001011110100110011101010110010011110000011111……
b=1011110100110011101010110010011110000011111001111110101001011100010000111111110100000……

問題：

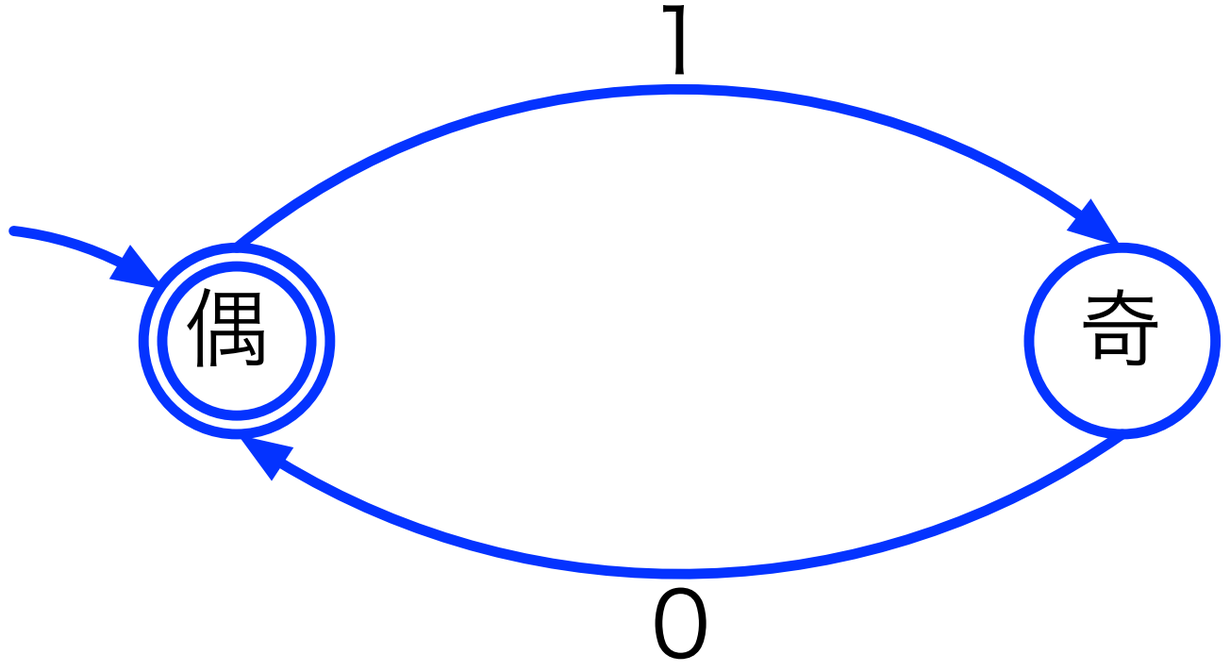
0と1からなる(2進数の)数字の列が与えられた時　この中にある1の個数が偶数かどうか判定する手続きを構成せよ

回答

○で状態を表す

- 外から矢印が入っている状態は初期状態(ここから始める)
- ◎は受理状態(この場合は偶数)

矢印で状態間の遷移を表す



動作概要：

- 初期状態から始め、入力された文字を読むごとに、矢印で示される状態へ遷移する
- 入力を読み終えた時点で、受理状態にあれば入力を受理する

コンピュータのモデルへ：有限オートマトン

状態が有限個

入力テープ(Read-Only)

0	1	0	1	1	0	1	0	0	1	0	0	0	1	1	0	0

ヘッド

有限制御部

定義：有限オートマトン(Finite Automaton)

有限オートマトンとは5つ組 $M = (\Sigma, Q, q_0, F, \delta)$ である。ただし

- Σ : アルファベット(文字の集合)
- Q : 状態の集合
- $q_0 \in Q$: 初期状態
- $F \subseteq Q$: 受理状態の集合
- $\delta : Q \times \Sigma \to Q$ 遷移関数

M によって受理される入力の集合を $L(M)$ で表す

- 単に入力の集合をL と書き、M によって受理される集合と等しいときと書く
- L は問題、M は手続き(アルゴリズム)と考えられる

オートマトンで解ける典型的な問題の例

入力中の特定の文字数の偶奇を分ける。

入力中の特定の文字数が特定の数かどうかを分ける。

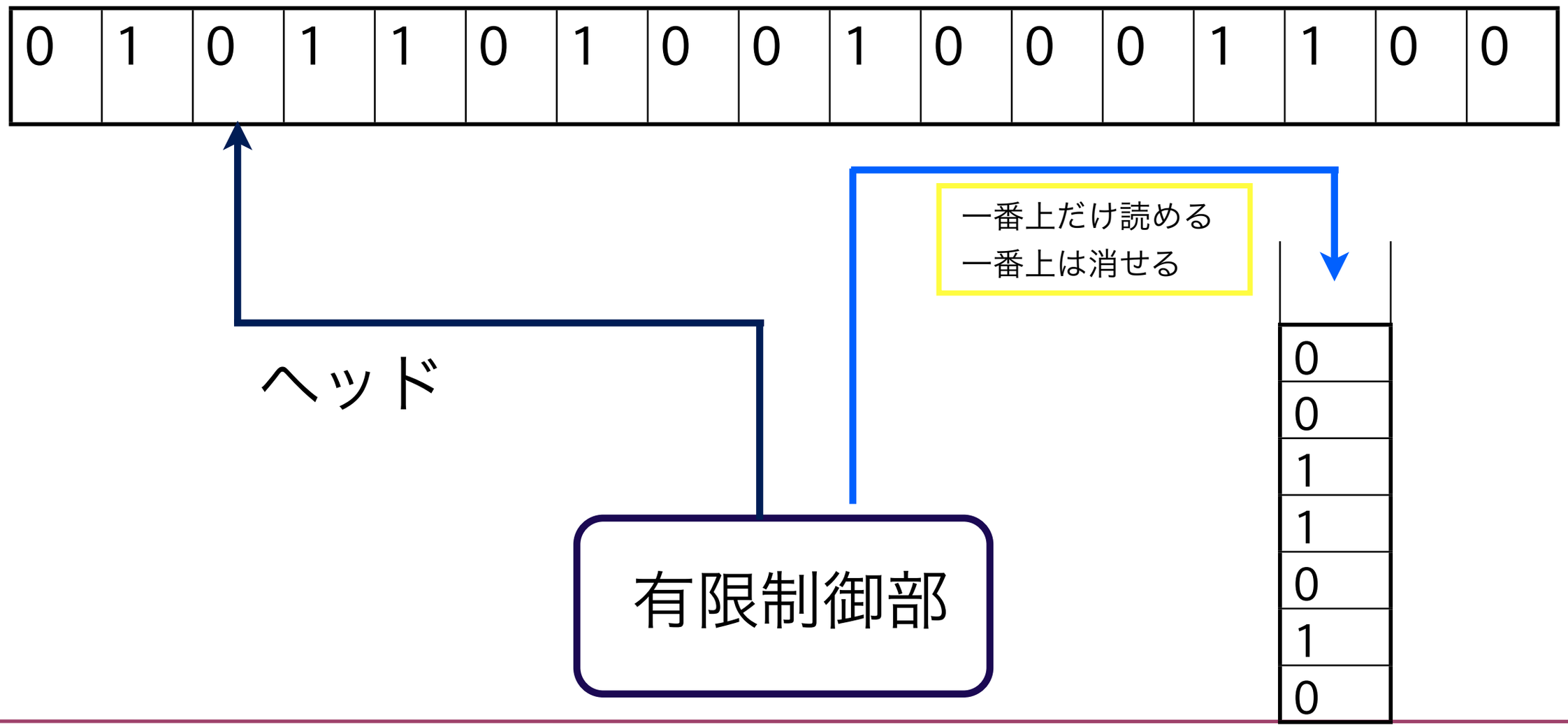
- 特定の数は定数でなければならない。

入力中のある文字と別の文字の数が等しいかどうか判定。

プッシュダウンオートマトンへ

スタック付きのオートマトン：作業のための領域が使える。

- スタック⇔後に積んだものを先に取りだせる書き込み可能なテープ



プッシュダウンオートマトンの能力

プッシュダウンオートマトンは有限オートマトンより真に能力が高い。

- プッシュダウンオートマトンが受理できる入力全体は、有限オートマトンが受理できる入力全体より真に大きい。
  - 例： $L = \{a^n b^n \mid n > 1\}$
- スタックの存在が能力の違いを生んだ。

任意に読み書きできるテープ

入力長テープ付きオートマトン(Linear-bounded Automata)

- 入力と同じ長さで、任意の場所を読み書きできるテープが付いた有限オートマトン。
- プッシュダウンオートマトンより真に能力が高い。
  - 例：$L = \{a^n b^n c^n \mid n > 1\}$

チューリング機械 (Turing Machine)

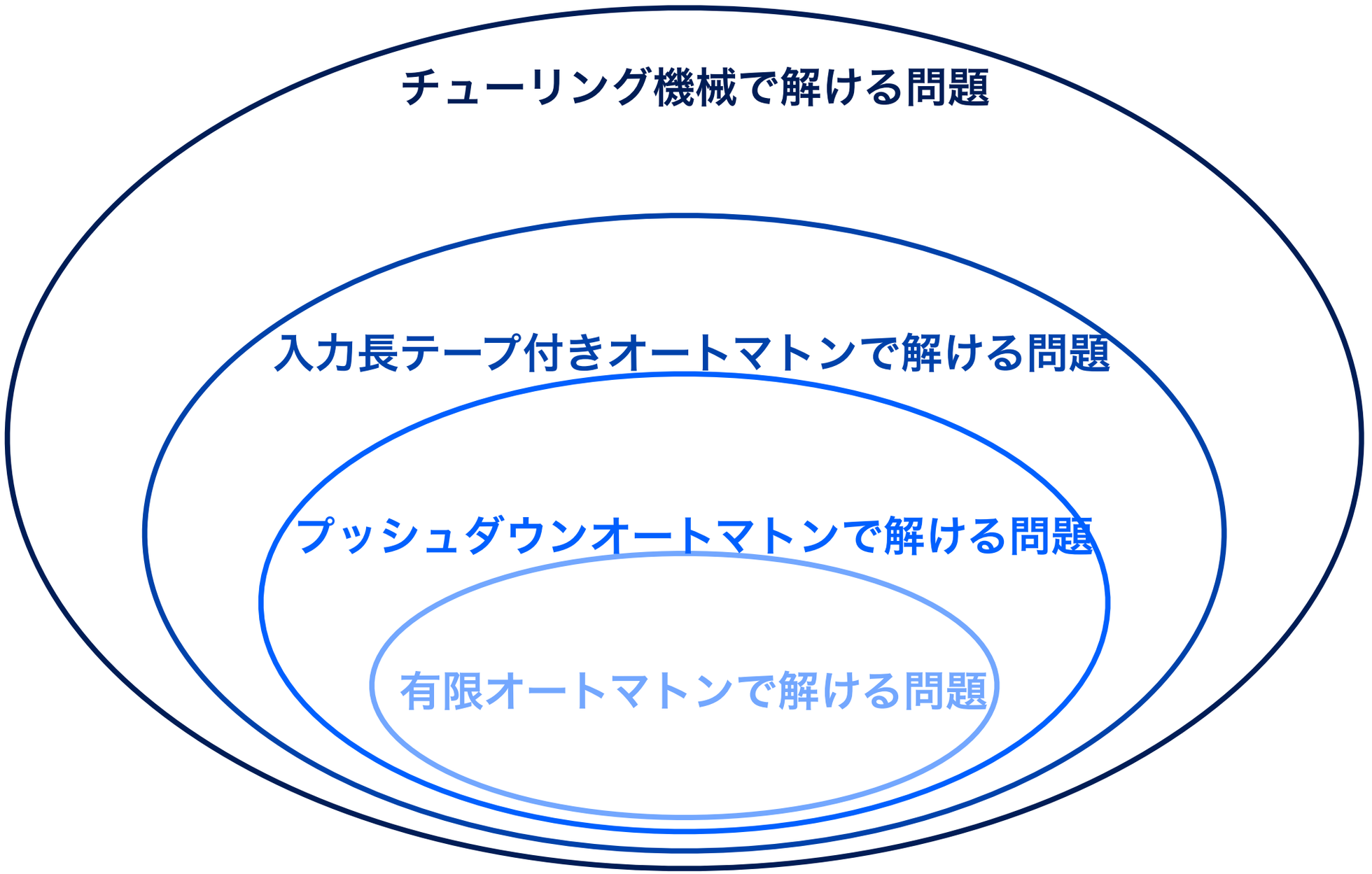
- 長さの制限がないテープが付いたオートマトン
  - 理論的には無限の長さを持っている。
- 入力長テープ付きオートマトンより真に能力が高い。
- コンピュータプログラム（アルゴリズム）に対応。

計算とは?

ある問題(集合)がチューリング機械で受理される時に、この問題は計算可能である、という。

- 他に帰納的関数、ラムダ計算など等価な計算モデルが提案されている。

チョムスキー階層


チューリング機械で解ける問題
入力長テープ付きオートマトンで解ける問題
プッシュダウンオートマトンで解ける問題
有限オートマトンで解ける問題

オーダー

$f(n), g(n)$：自然数上の関数

$f(n)$ は $g(n)$ のオーダー(order)である
⇔任意の自然数に対し、次式を満す定数Cが存在する

$$\frac{f(n)}{g(n)} < C$$

例：

$$\frac{f(n)}{g(n)} = \frac{245n^2 + 49n + 191}{n^2} = 245 + \frac{49}{n} + \frac{191}{n^2} < 485 (= 245 + 49 + 191)$$

計算量時間の評価

問題の本質的な難しさ(易しさ)や、アルゴリズムの本質的な速さが知りたい。

- 具体的なコンピュータに依存して、同じプログラムでも実行時間は異なる。

Turing機械のステップ数で評価する。

- 計算の定義における数で比較。
- ただし、入力の大きさnに対する評価をオーダーで考える。
  - nが充分大きな時の計算時間の増え方、つまり、漸近的な計算量で評価する。

例：n個の自然数の中で、最大のものを見つける。

- 変数rに0を代入する。
- rの値と一つの要素を比較し、大きい方を新たなrの値とする。
- 全ての要素に対し、比較を行う。 $O(n)$

ランダムアクセスマシン

チューリングマシンでは、ヘッドが1マスずつ動くため、
M番目のマスを読むためだけMの計算量が必要。

ランダムアクセスマシン

⇔任意のアドレスの内容をO(1)時間で読み書き可能なTM

- これにより、任意の場所の内容へのアクセスが効率的にできる。
- コンピュータのモデルは有限長テープのランダムアクセスと考えられる。

宿題：図書の目録カードを作成せよ

本を探す時に必要なメタデータ(属性)は何かを考え、

図書の目録カードを作成せよ。

- 目録カードがどういうものかは明日説明しますので、調べないで下さい。
- 要点は、後で自分が「あの本、どこにあったっけ」という時に必要なデータを考えてもらうことです。

宿題の意図：

- データベースと情報検索に必要な(情報の)属性について考えてもらう。