http://www.jaist.ac.jp/ks/labs/ho

データの中にある知識を発見する機械学習・データマイニング手法

Machine learning and data mining methods for discovering knowledge in data

ホー研究室

教授: ホーツーバオ

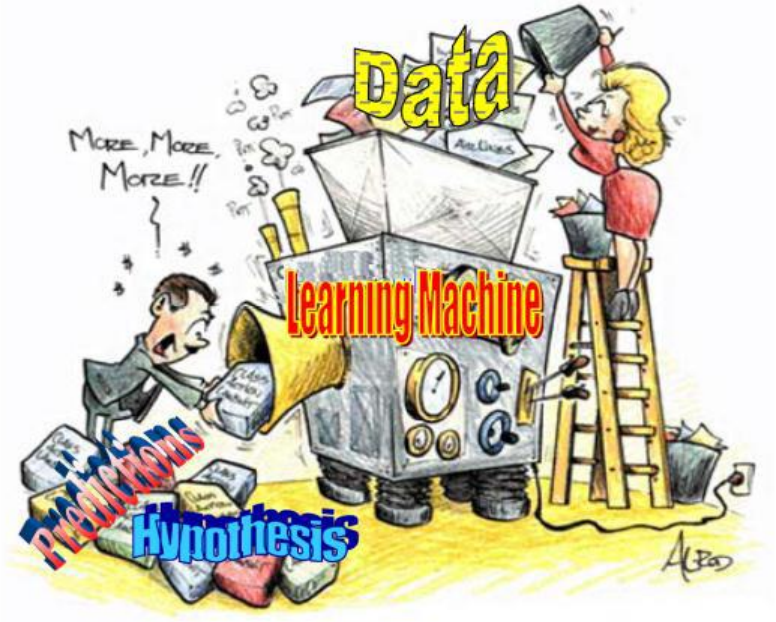
Ho Tu Bao (ホーツーバオ)

YEAR	EDUCATION
1987	Bachelor in Applied Mathematics, Hanoi Univ. of Technology
1984	Master in Computer Science (AI), University Paris 6, France
1987	PhD in Computer Science (AI), University Paris 6, France

YEAR	INSTITUTION
1991	Associate Professor, Vietnam Academy of Science & Technology
1993	Visiting Associate Professor, Information Science (JAIST)
1998	Professor, Knowledge Science (JAIST)

PROFESSIONAL ACTIVITIES
Chair of Steering Committee of Pacific-Asia Knowledge Discovery & Data Mining (PAKDD)
Chair of Steering Committee of Asia Conference on Machine Learning (ACML)
Chair of Steering Committee of IEEE RIVF Conference on ICT (IEEE RIVF)
Member of Steering Committee of Pacific Rim Inter. Conf. on Artificial Intelligence (PRICAI)
Member of Editorial Boards of several International Journals (5)

機械学習･データマイニング
Machine learning and Data Mining



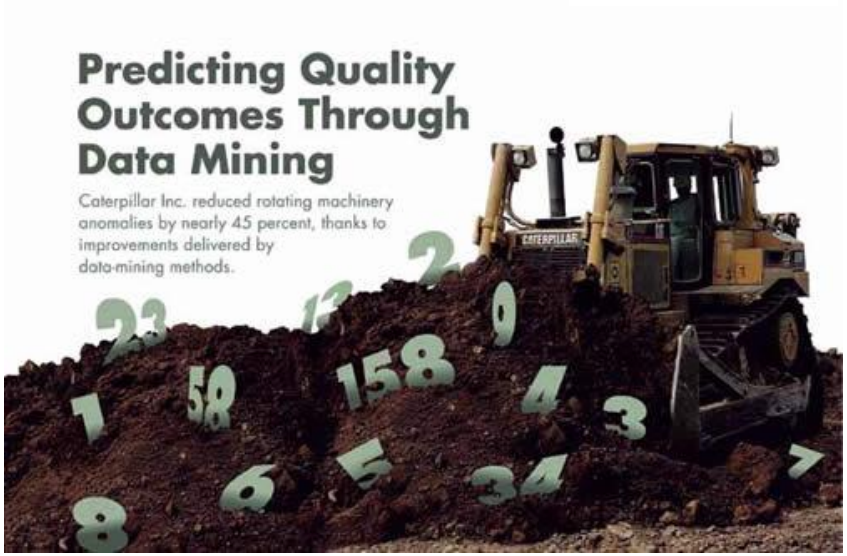
(from Eric Xing, Stanford University)

機械学習とは、人間が学習するように、機械に学習能力をもたせることを目的とする研究分野.

Make computers with learning ability as human

データマイニングの目的：
大規模なデータから未知で有用な知識を発見すること.

Finding knowledge in large datasets

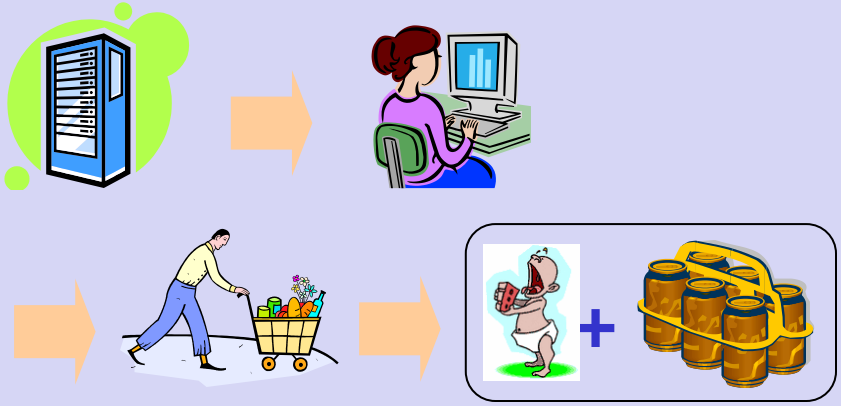


データから知識を創り出す例
Finding Knowledge from Data

マーケット・バスケット分析（IBM）

Super market data

“Young men buy diaper and beer together”
「紙おむつを買う男性は缶ビールを一緒に買うことが多い」



（解釈：顧客像）紙おむつを買うように頼まれた男性がついでに自分用の缶ビールを購入していた
→ 今後の陳列に活かすことのできる知識.

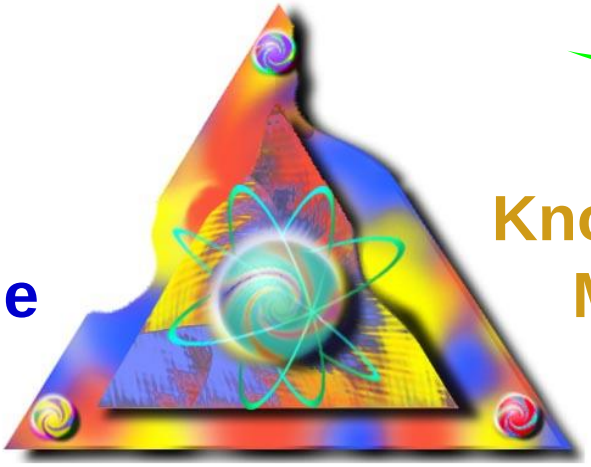
- Identify which genes cause a given disease.
  所与の疾患の原因となる遺伝子を同定する.
- Mining risks in financial market and investment.
  金融市場や投資におけるリスクをマイニングする.
- Mining opinions about society.
  社会に関するオピニオンをマイニングする.
- Retrieve documents by topics but not keywords.
  トピックによる文書検索.

機械学習・データマイニングと知識科学
Machine Learning and Data Mining at Knowledge Science

Human and machine create and use knowledge in its diversity

Systems Knowledge

Service Knowledge



Social Knowledge

Knowledge Media

機械学習・データマイニング

Human's ability to create and use knowledge

Knowledge science is science of creating and using knowledge by human and by machine

Machine support people to create and use knowledge

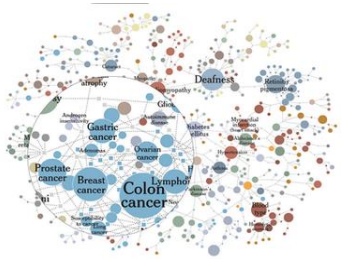
三つの飛躍的ITテクノロジー
Three emerging IT technologie

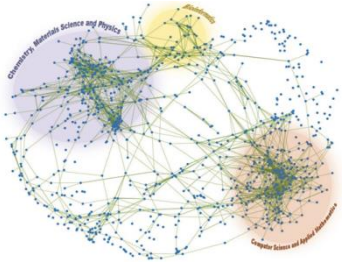
cloud computing

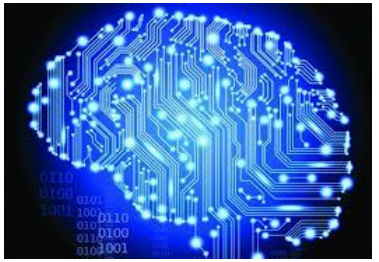
Smart-devices

Big Data are data sets that are **too large** and **complex** that cannot be well managed and analyzed with conventional IT techniques.

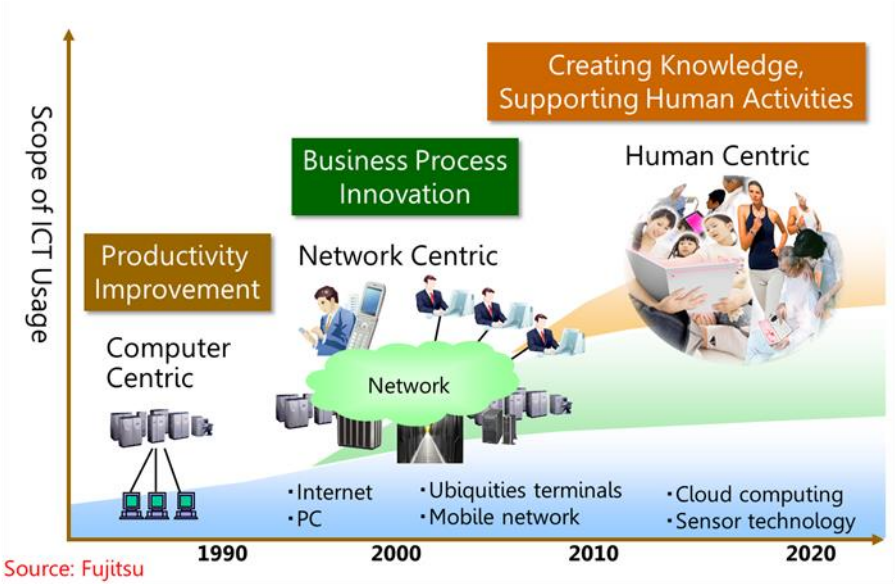
ビッグデータとは**巨大**かつ**複雑**なデータでこれまでのIT技術では管理点解析が困難データの集合





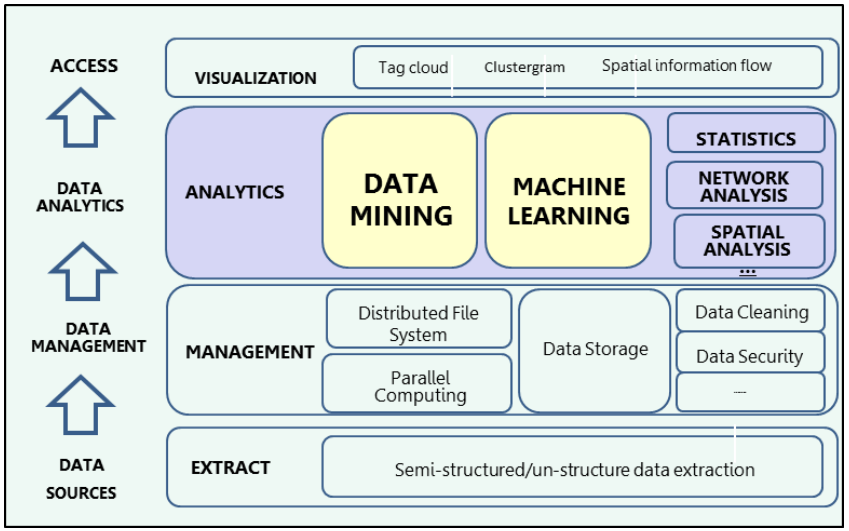


ビッグデータはどの様に私たちに大きな価値をもたらすか？
How can Big Data bring us Big Value?



Machine learning and data mining are the key technologies for analyzing Big Data

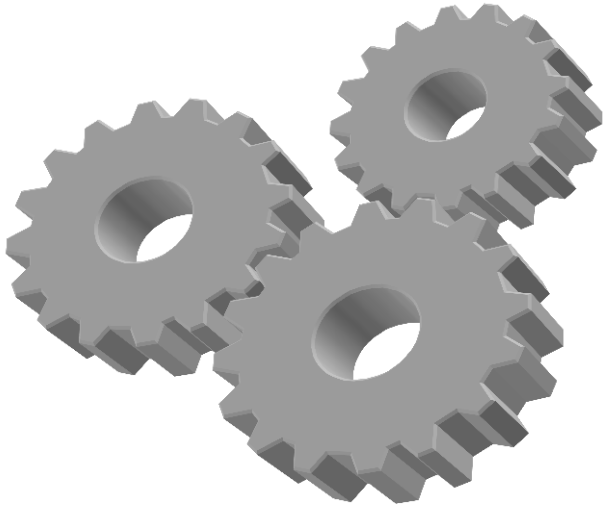
機械学習とデータマイニングはビッグデータ解析にとってのキーテクノロジー

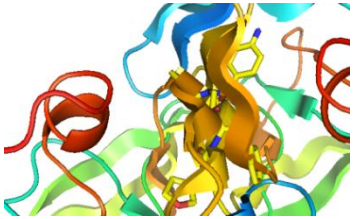


ホー研究室の研究の方向性
Research directions

Text & Web mining

大量のテキストデータやウェブページから知識をみつける





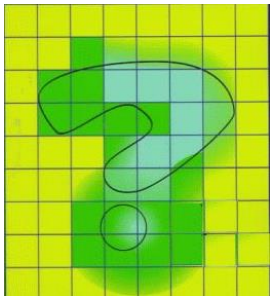
Scientific data mining

医学・生物学・物理学・化学・経済経営等のデータから知識を見つける

(e.g., Amazon, Seven-Eleven stores)

Basic research

データマイニングの難しい課題に対する先進的な技法(カーネル手法,類似性の評価，不均衡データ,ソフトコンピューティングなど）



当研究室が取組む3つの研究対象
Research directions in more details

Automate the understanding of the text meaning in huge corpora (topic modeling)

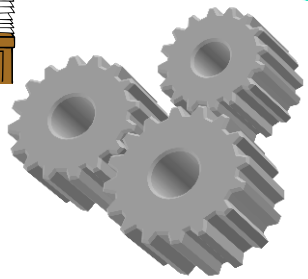
大規模コーパス内のテキスト内容理解の自動化.

Text & Web mining

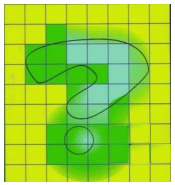
Scientific data mining

Establish models and computing methods in life science (biomedicine) and materials science

生命科学(生物医学)・材料科学分野におけるモデルや計算手法の確立.



Basic research



Methods of sparse modeling, dimensionality reduction, graphical models … for Big Data

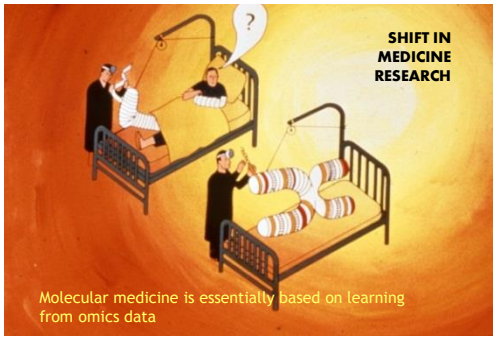
スパースモデリング, 次元縮退, グラフィカルモデル等のビッグデータ手法.

OUR TARGET

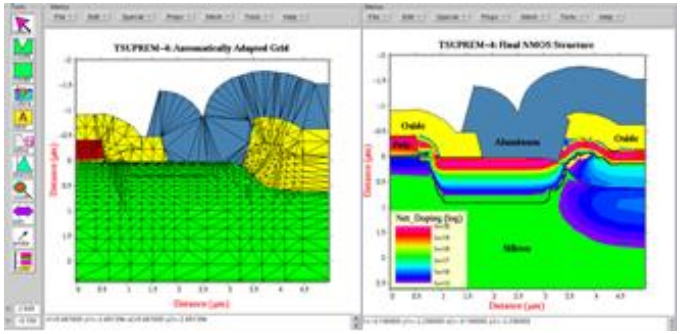
- Make breakthrough in machine learning and data mining research.
  機械学習とデータマイニング分野におけるブレイクスルーを起こすこと.
- To promote data science in knowledge science.
  データマイニングを知識科学のツールとすること.

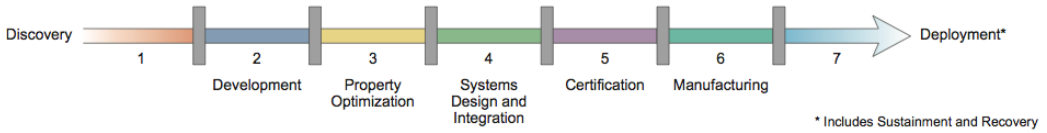
科学的なブレイクスルー
Scientific breakthroughs

- Life science, biomedicine
  (ライフサイエンス,生物医学)
  - ➔ Genomics medicine
    (ゲノミクス医療) :
    Combine clinical medicine with molecular biology.
  - ➔ Big data for promotion of life innovation: Exploiting electronic medical records (EMRs)
    電子カルテ(EMRs)の有効利用.

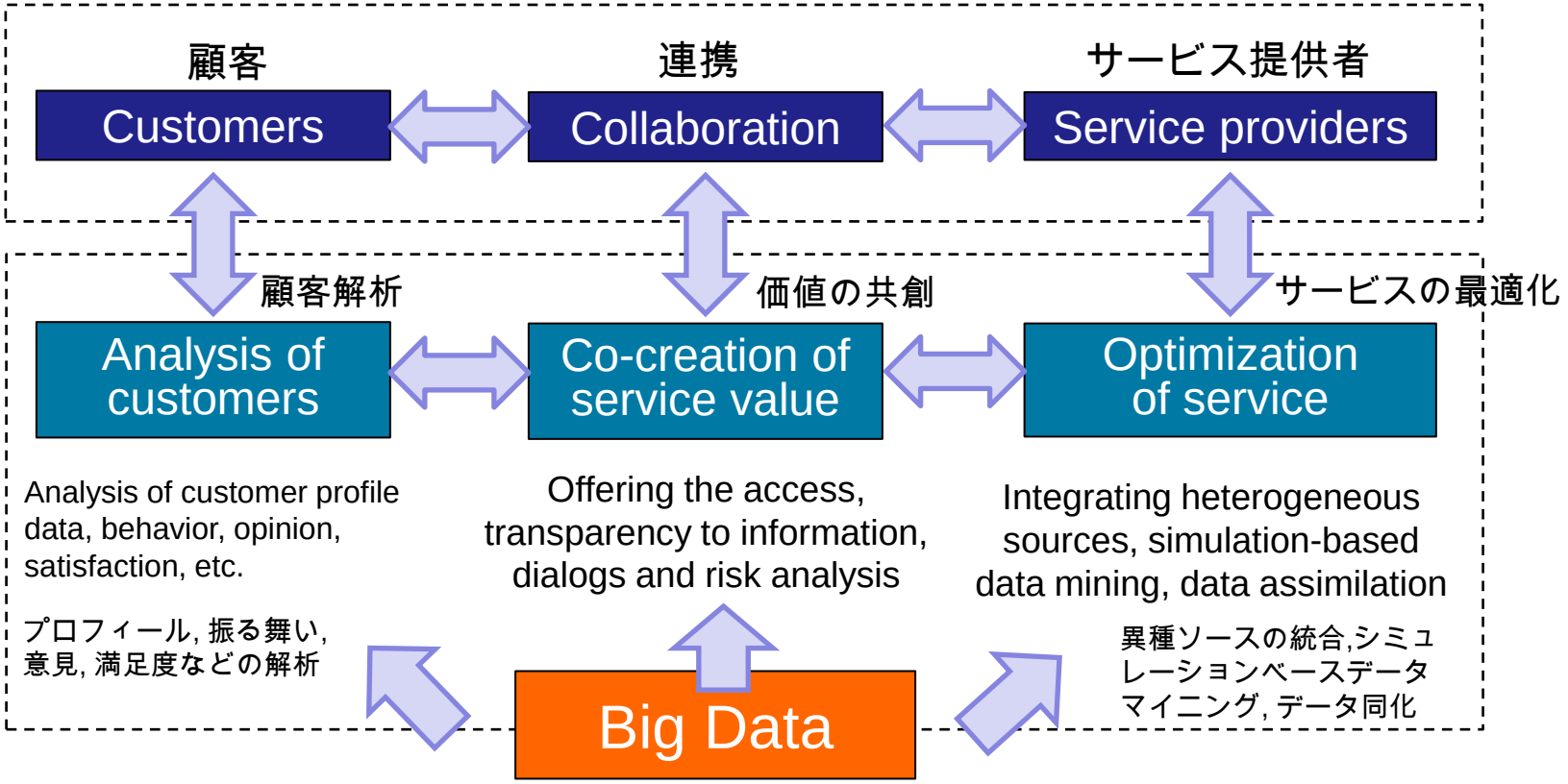


- Materials genome initiative
  - ➔ Contribution to shorten the materials development cycle from its current 10-20 years to 2 or 3 years
    材料開発のサイクルを10倍にする





サービスサイエンスにおけるビッグデータ
Big Data in a Service Science: key idea



Key word: Customer Relationship Management (CRM)

キーワード: 顧客関係管理(CRM)

プロジェクト Projects

- 科学研究費特定領域(C)「ゲノム情報科学の新展開」 (2004-2007)
- COE知識科学に基づく科学技術の創造と実践 (2003-2008)
- 「計算物理学とデータマイニングの融合による結晶学への現実的・効率的アプローチ (2005-2007)
- 科学研究費基盤研究(C)解釈指向マイニングによる診療情報からの医学的知見の発見(2009-2012)
- 科学研究費基盤研究 (B)「多種情報源からのデータマイニング手法による肝炎知識の発見」 (2004-2007)
- 科学研究費基盤研究(B)「科学 データのための先進的計算手法」 (2007-2010)
- 科学研究費基盤研究(B) 「算アプローチによる肝炎の病態・治療に関する 分子機構の解明」 (2011-2014).

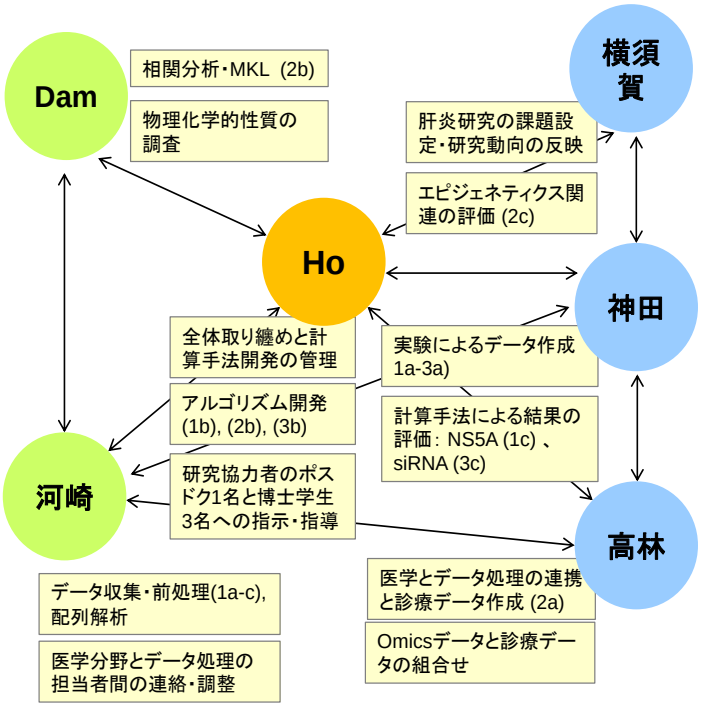


図5　研究課題と分担

生物学データのマイニング
Working on genomic medicine

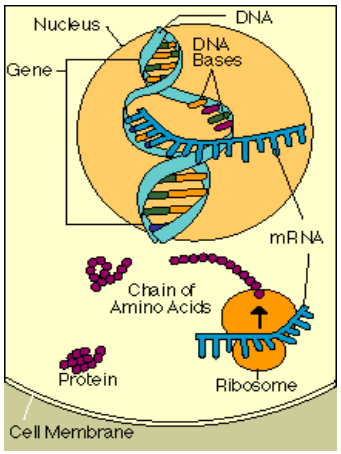
Molecular biology knowledge and omics data
are playing essential role in medical research (biomedicine).

分子生物学の知見やomicsデータは医学研究においてもはや不可欠である

How machine learning and data mining creates new knowledge about biological mechanisms of life?

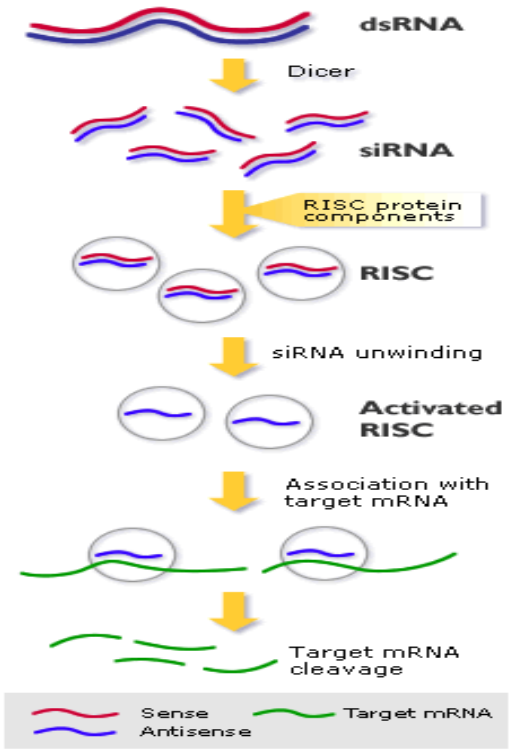
どうすれば機械学習やデータマイニングは生命の生物学的な諸機構について新たな知識を創ることができるだろうか？

RNA干渉 (RNAi)
RNA interference (RNAi)



DNA > mRNA > Protein

Fire, A., Mello, C.,
Nobel Prize 2006



- RNAi (siRNA and miRNA) is post-transcriptional gene silencing (PTGS) mechanism.

  RNAiは転写後遺伝子抑制機構(PTGS)である.

- Chemically synthesized siRNAs can mimic the native siRNAs produced by RNAi but having different ability.

  科学的に合成したsiRNAはRNAiが生成する本来のsiRANに類似するが、ことなる能力をもつ.

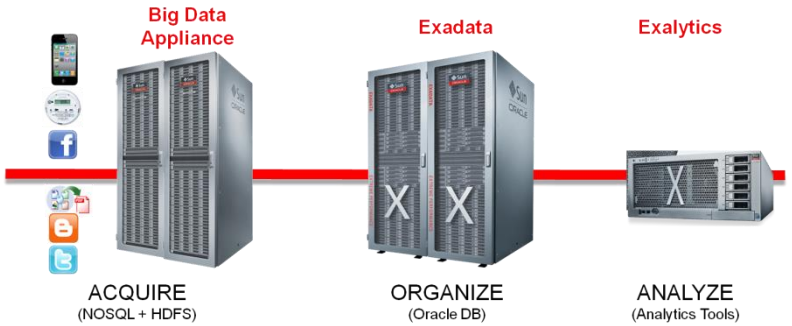
- **Problem**: Selection of potent siRNAs for silencing hepatitis viruses?

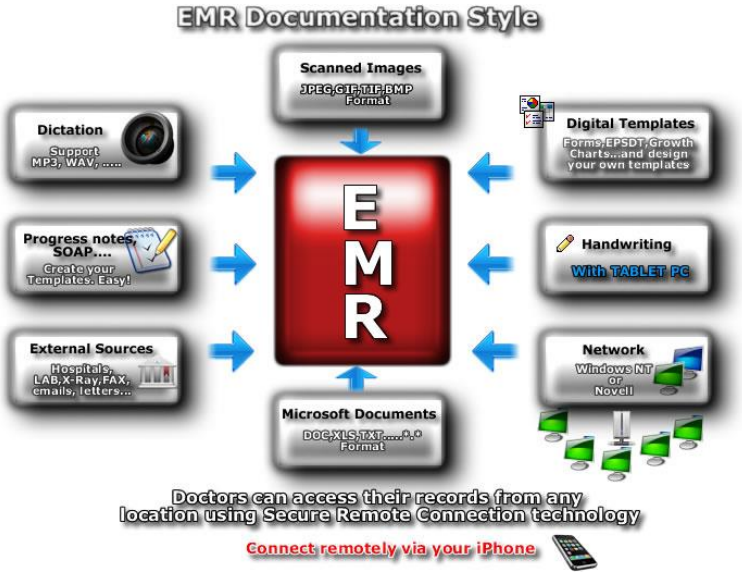
  課題：肝炎ウイルスを抑制する有力なsiRNAの選定.

EMRs: 医療暗黙知のソース
EMRs: Medical tacit knowledge sources

Electronic medical record (EMR) is a computerized medical record created in an organization that delivers care → Promotion of life innovation.

電子カルテ(**EMR**) とは組織によって電子化された医療 → ライフイノベーションの促進

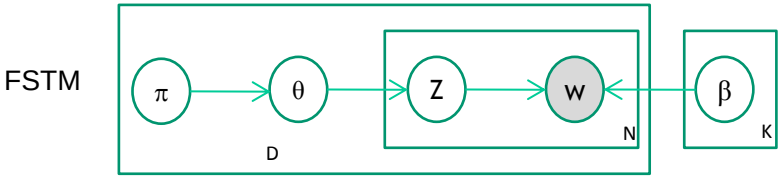




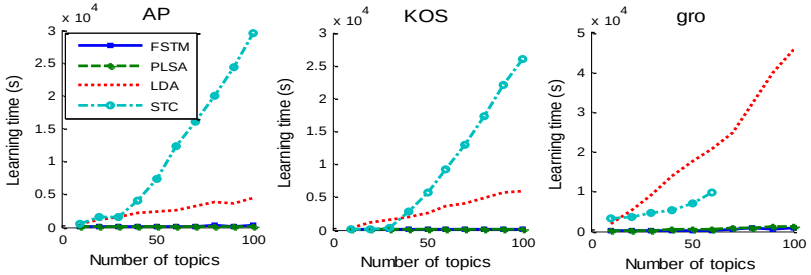
How to convert such tacit knowledge into explicit knowledge

明示的な知識にこのような暗黙知に変換する方法は？

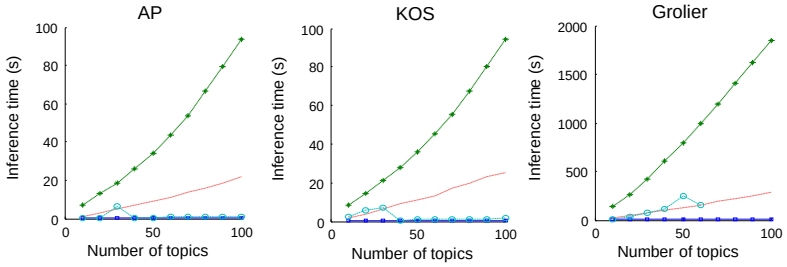
Fully sparse topic model



How fast can the models learn?



How fast can the models infer?



Sparse model vs. Dense model

- Topic modeling is the key approach to automate the text meaning (idea: a topic is a set of words with a probability distribution, and a document is mixtures of latent topics).
- Our **sparse topic model** allows dealing with big text data (millions documents and thousands topics) that current dense topic models cannot do (reducing the storage **from 23.3 Gb to 33.3 Mb** for 350,000 documents).

Sparse vs. dense	#topics: thousand & hundreds
Inference time	Linear vs. non linear
Sparse topic representation	**100 times smaller**
Sparse document representation	**350 times smaller**
Storage	**700 times smaller**

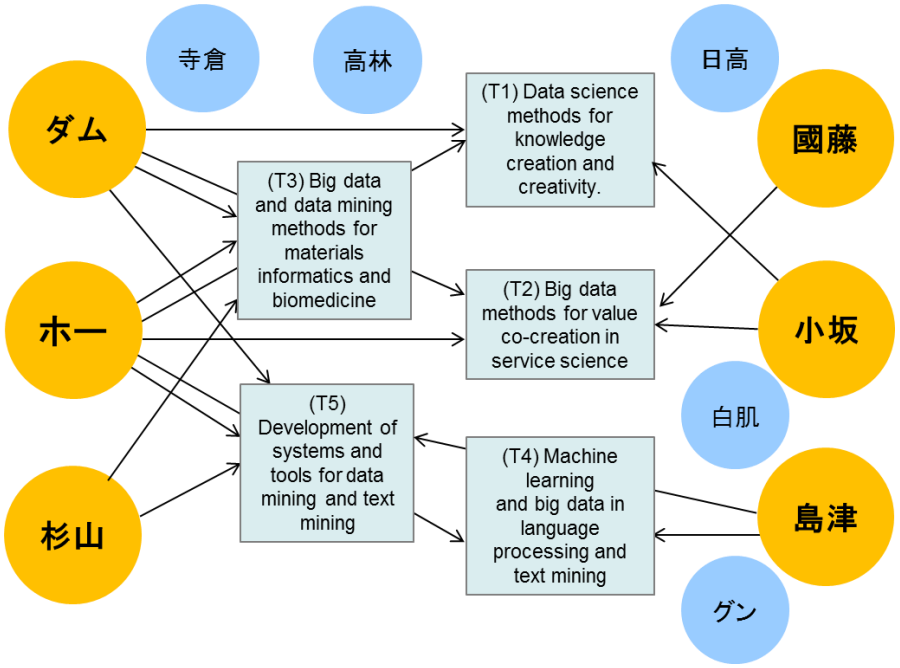
Toward data science

K112: 統計学入門, Introductory Statistics (1-1)

K417: 知識創発論, Knowledge Discovery Methodology (2-1)

K619: 次世代データ分析特論, Modern Multivariate Data Analysis

Data Scientist: The Sexiest Job of the 21st Century
(Harvard Business Review, October 2012)

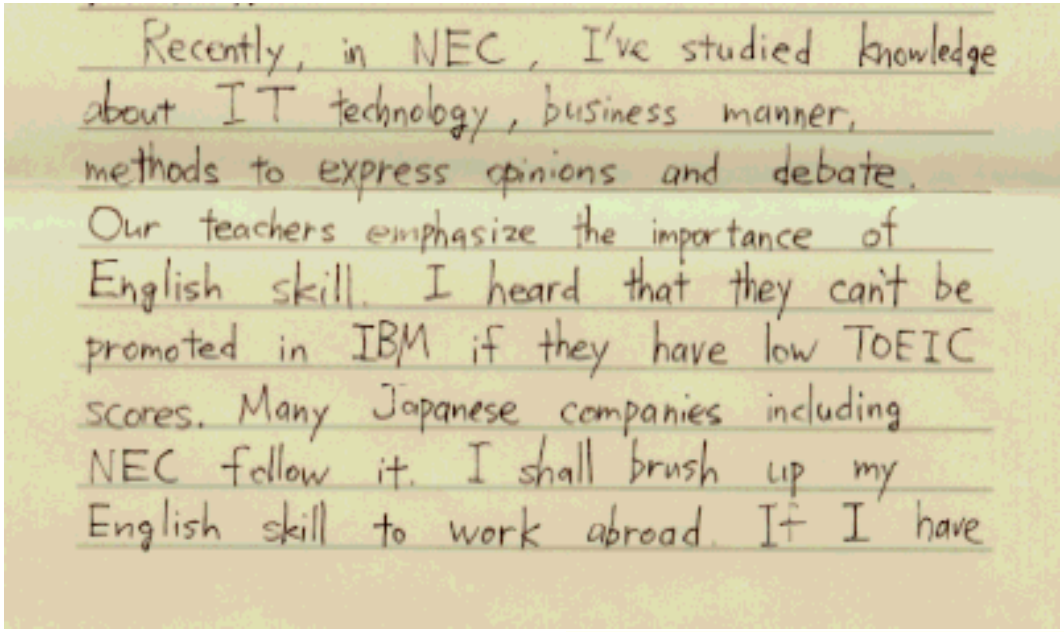


修了生の研究テーマ
Topics of master students

Name	Thesis
伊藤 博之	地球温暖化排出ガス削減における政策機構及びその重要因子の研究 (2000)
小山内 尚	東アジアの酸性・酸化性物質の動態解明へのクラスタリング手法の適用 (2000)
西田 健一郎	バッギング手法による分類システムの性能の向上 (2000)
山口 和泰	データ視覚化と決定木帰納法の統合に関する研究 (2000)
深沢 弘保	キーワード抽出アルゴリズムKFOの調査・研究とそのWebマイニングへの応用可能性 (2000)
河崎 さおり	トレランス・ラフ集合モデルに基づく階層型文書クラスタリングアルゴリズムの提案 (2000)
斎藤 昭典	胃癌患者データベースからの知識発見に関する応用研究 (2001)
冨士川 義和	知識発見プロセスにおける欠損値処理を目的とする機械学習アルゴリズムの研究 (2001)
鈴木 俊之	相関ルールマイニングにおける冗長性削減アルゴリズムに関する研究 (2002)
中田 豊久	個人の興味を映し出すWeb Communityの抽出方法の提案 (2002)
沼田 公博	時系列データにおける類似性検索へのウェブレット変換の利用 (2002)
綾川 聡司	ブースティング手法を用いた分類システムの精度向上(2003)
中本 修	Batch-Learning Self-Organizing Mapアルゴリズムを用いた検索エンジン結果判断補助のためのマッピング (2003)
村瀬 健留	量子コンピューティングとShorのアルゴリズムの研究 (2003)
Nguyen Duc Dung	Using Prior Knowledge in Rule Induction (2003)
伊藤 宏徳	決定木とニューラルネットワークの組合せによる連結学習アルゴリズムの開発 (2004)
小坂 周二	Fuzzy Support Vector Machine手法に関する研究 (2004)
永井 健太郎	背景知識を用いたテキストデータからの意味的に豊かな相関ルールの発見 (2005)
宮下 和也	帰納論理プログラミングを用いたタンパク質間相互作用に関する研究 (2005)
虫明 麿毅	チャンス発見技術を用いた新科学および技術のリスク発見 (2005)

Currently 6 PhD students and 3 master students (12 PhD and 35 masters graduated).

バイリンガル環境


Recently, in NEC, I've studied knowledge about IT technology, business manner, methods to express opinions and debate. Our teachers emphasize the importance of English skill. I heard that they can't be promoted in IBM if they have low TOEIC scores. Many Japanese companies including NEC follow it. I shall brush up my English skill to work abroad. If I have

(From letter of Fujikawa-san, graduated in March 2001, working at NEC)