平成29年8月24日

Tohoku Forum for Creativity Symposium
データとインテリジェンス

ヨッタバイトスケールの巨大情報量と
そのインパクト
―激増する情報量の壁を越えて―

村岡裕明

東北大学
電気通信研究所
ヨッタインフォマティクス研究センター

概要

- 情報科学技術の進展と社会への浸透に伴い、人類が生成する情報量は爆発的な増加を続けている。情報ストレージの高密度化や半導体技術の進歩や新メモリの発明などがこれを支えているが、情報の伸びには追いつかない
- その巨大化した情報量は、2030年までには1ヨッタバイト（1兆バイトの1兆倍）に達する見通しで、データが溢れる中でその価値を引き出す「ビヨンドビッグデータ」時代が到来する。
- シャノンが情報量を定量化して以来情報学は大きく発展してきたが、今後の新しいアプローチとして、情報の「質」や「価値」に着目することで巨大情報を扱う新しいパラダイムを拓けないか。
- このための新しいインフォマティクスの構築を目指す。情報の質や価値を扱うには人間の抽象的な知に基づく必要があり、理工学だけの取組ではなく人文科学や社会科学の連携する。

情報とは

1. 明治9年　酒井忠恕 「敵情の報知」から*
2. 情報の定義
   - あいまいさを減らして行動を決める手がかり
3. シャノンの情報理論（1949年）→質や価値は放棄して確率によって定量化した情報「量」に限定した
   - 1ビットの情報量：　等確率の2通りのうちのいずれかを指定できる（あいまいさがなくなる）
   - 生起確率Pの事象aの持つ情報量Iは、

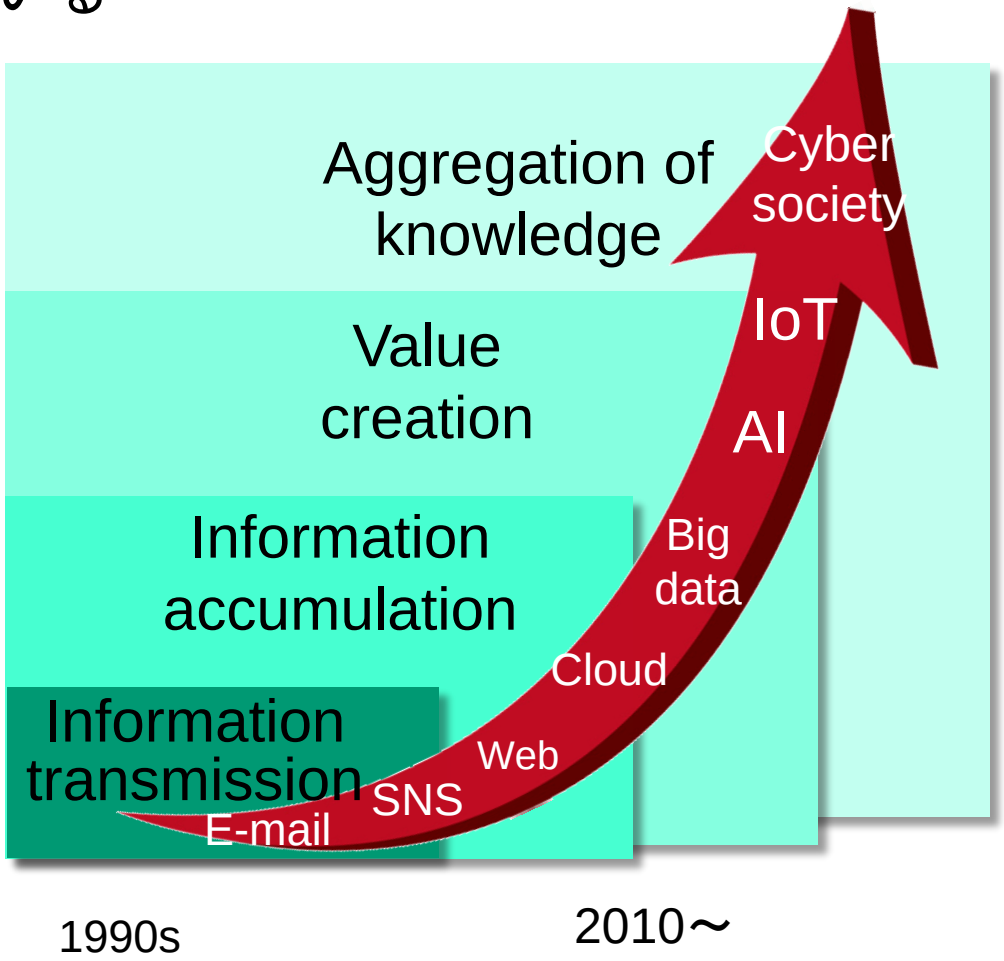
$$I(a) = \log_2 \frac{1}{P(a)}$$

$$P = \frac{1}{8} = \frac{1}{2^3}$$

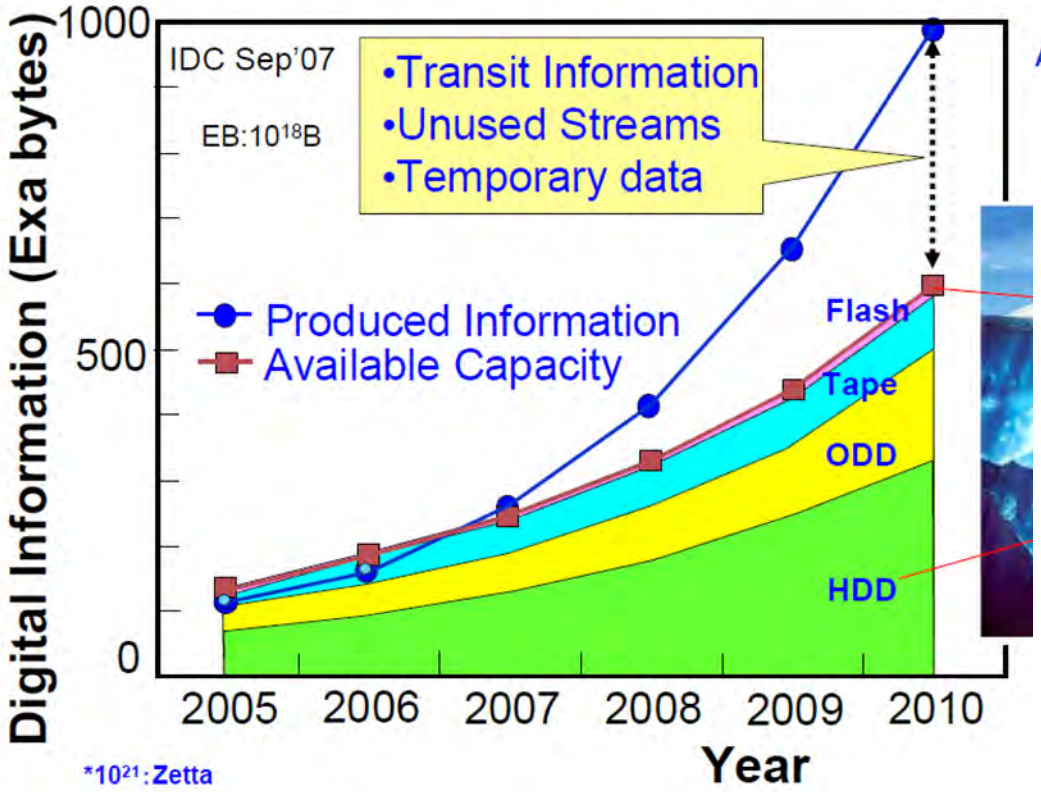
*石井健一郎 「情報」を学び直す　NTT出版

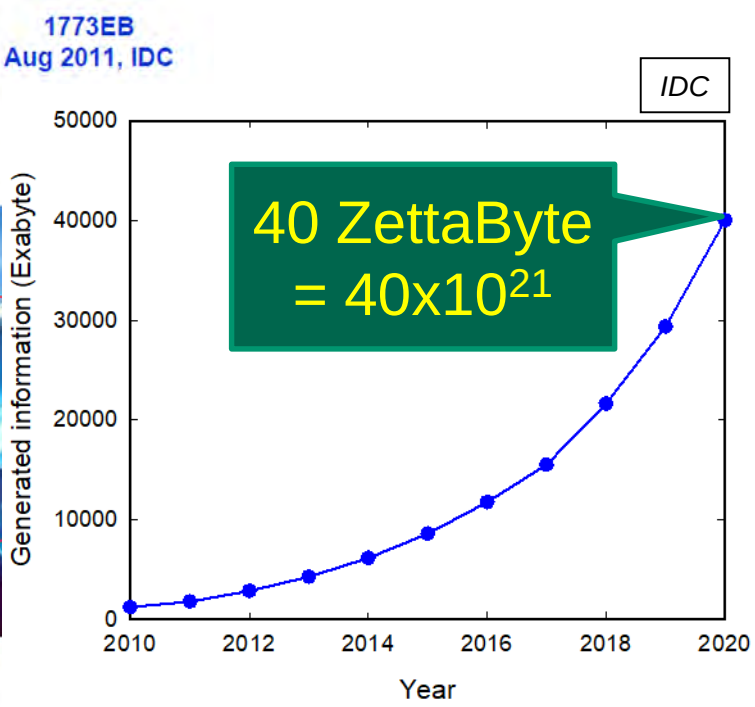
インターネットがICTを変革する

- ICTは、ビッグデータ、AI、IoTを軸にイノベーションが続く
- 次々と起こるパラダイムシフトが生成情報量の拡大を生み出している



インターネットの世界における巨大情報量





Y. Shiroishi et al., FA-01, Intermag 2009

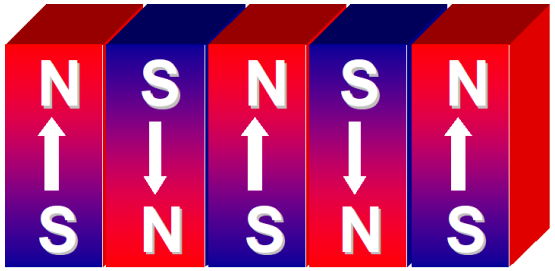
アボガドロ数＝6x10^23
宇宙の恒星の数＝7x10^22

- 巨大な情報量が生み出されている。
  - 1773エクサバイト＝1.773×10^21バイト＝1773兆の100万倍
- 世界一の図書館の情報量のさらに100万倍がインターネットを流通
- その大半を磁気を使って保存している

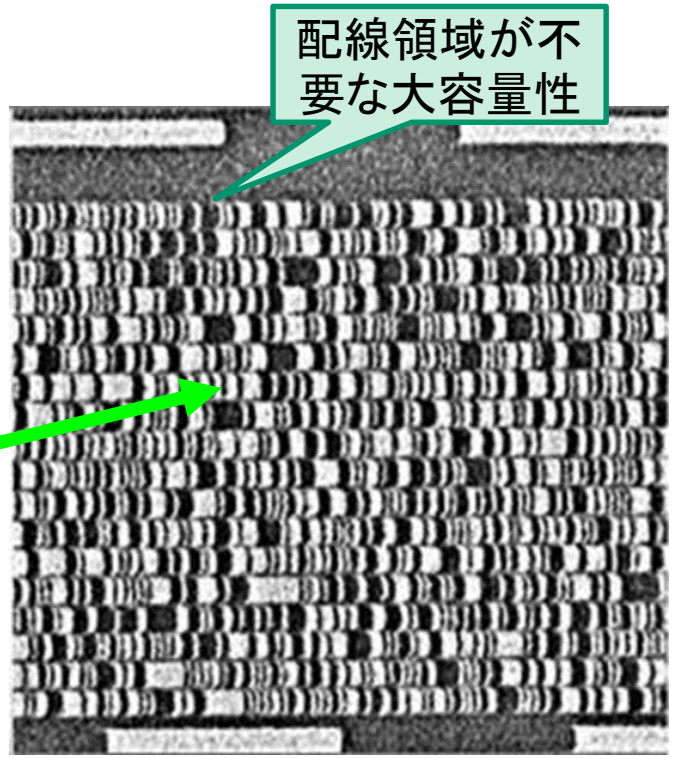
ビッグデータを支える情報ストレージ技術

垂直磁気記録による高密度ハードディスク

- 2状態を持つ物理量を1ビットのメモリに対応。
- 素磁石をディスク面に垂直に配置して安定な高密度記録を実現。(1977年　東北大学岩崎俊一)
- 配線領域が不要な大容量性
- 大容量ほど微小な磁石を形成する必要がある。

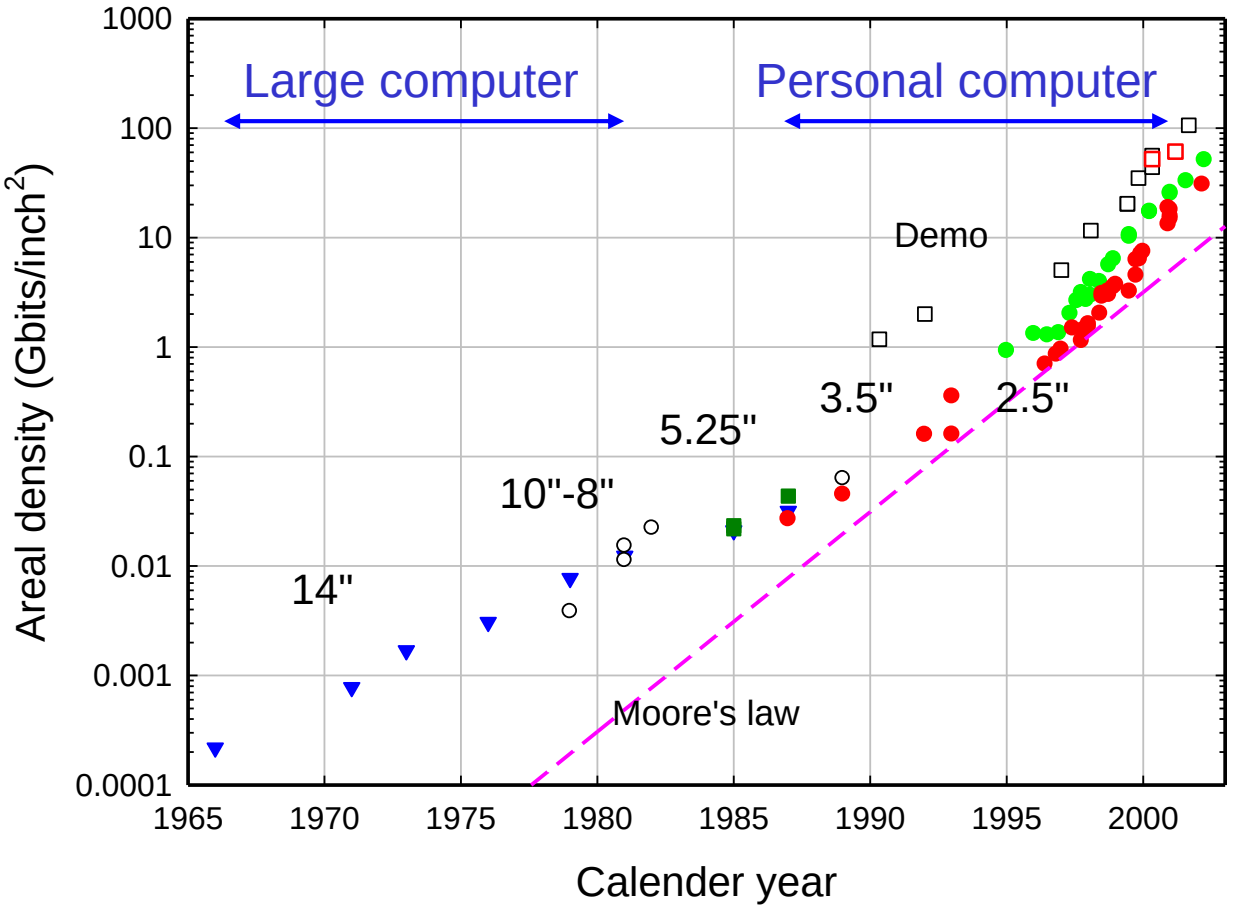


→　2.5インチ1TB= 8×10^{12} bit → 直径63mmディスク両面で8兆個の磁石を配置→　790 nm^2の超微小磁石



垂直記録磁化パターン
(日立製作所・東北大学)

高面密度化の経緯



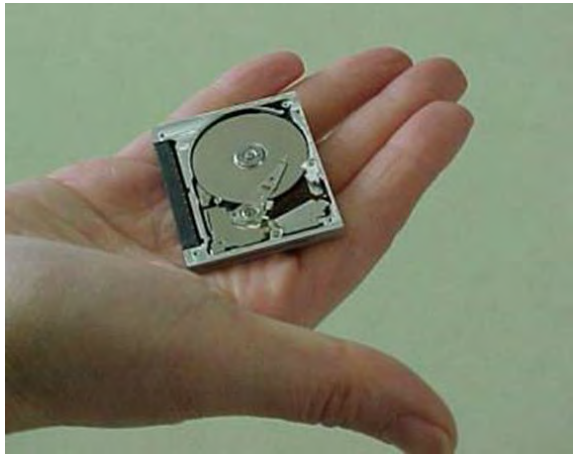
- ディスクの小型化と高密度化が同時進行
- 40年間で100万倍の面密度→ ビットの大きさは100万分の一

ハードディスク装置の進歩

- 巨大な機械から手のひらサイズの小型ディスクへ
- 記憶できるデータ量は2000倍

世界で初めての
ハードディスク装置
（IBM社 RAMAC
1956年）容量5 MB

大型計算機用
ディスク装置
（IBM社 3380
1980年）容量
2.5 GB

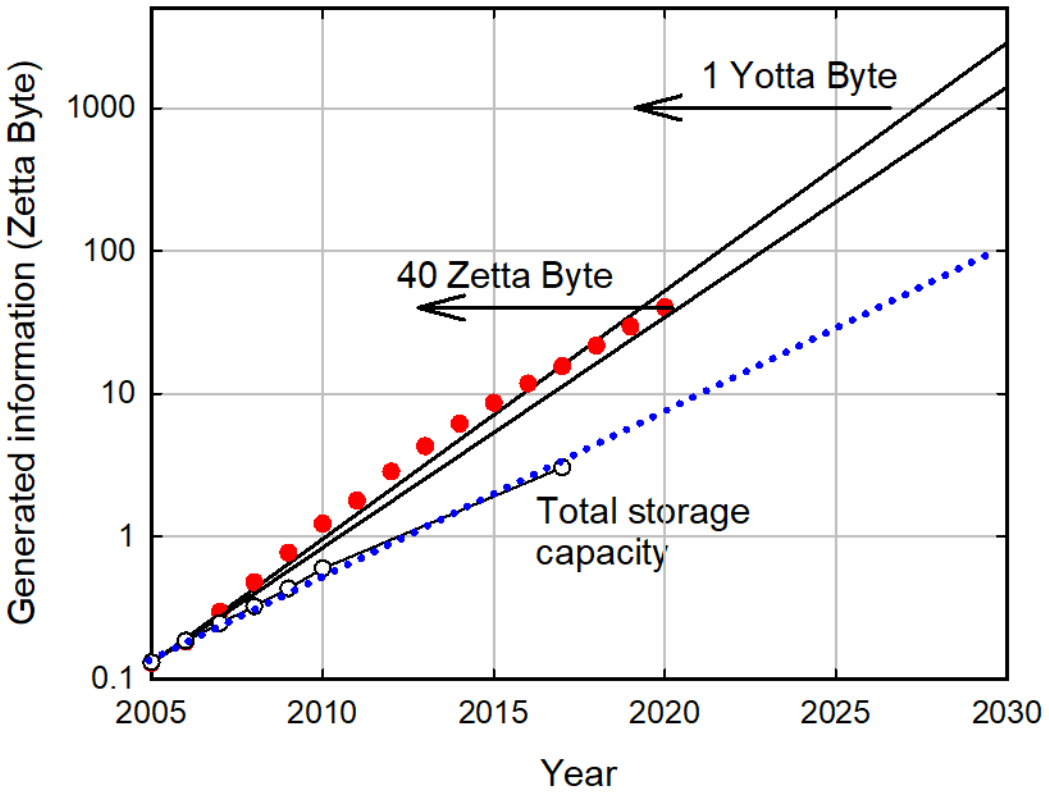


1インチ型小型垂直ハードディスク装置（2006年）10 GB

巨大容量データストレージ

- 社会インフラとしての重要性から大容量ストレージサーバの規模は拡大の一途
- HDDを並列に用いる大容量性と高速性を実現
- 機器が故障したり災害があってもデータを守ることが重要

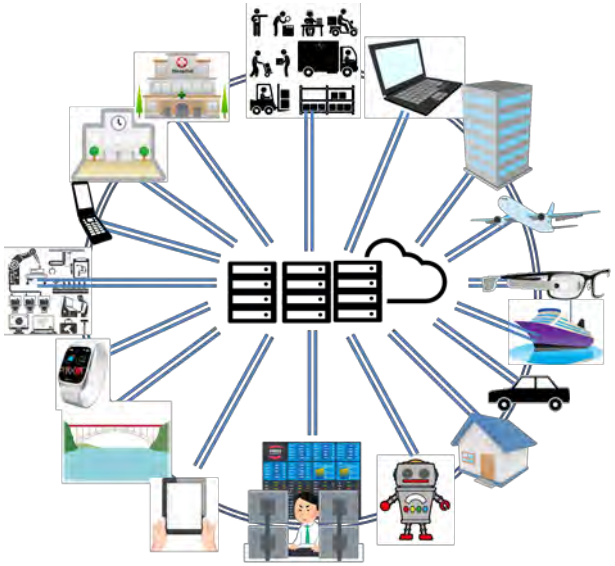
2030年までにヨッタバイトの情報量に達する

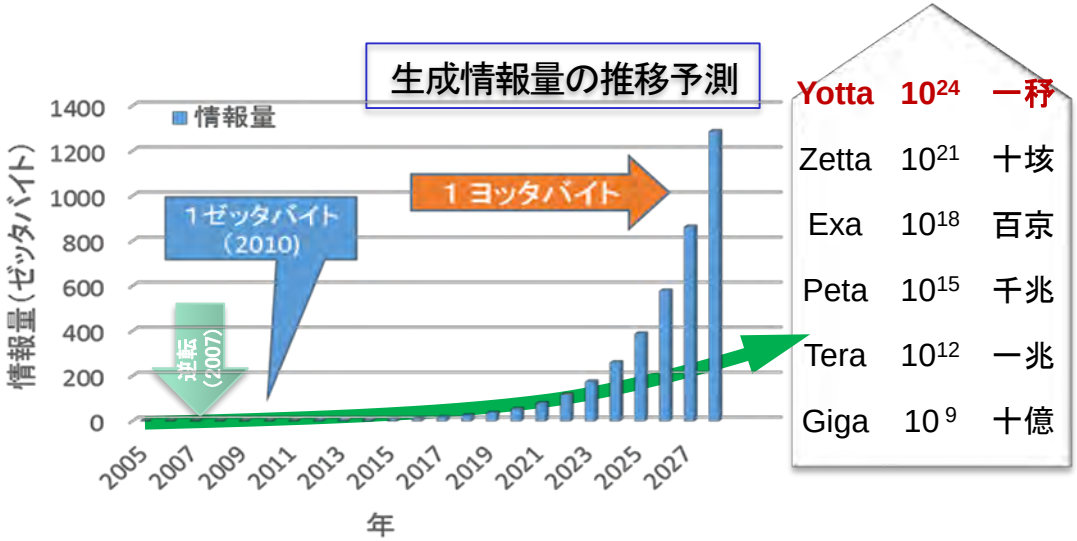


IDC Dec 2012, "Digital universe in 2020"

- 年率140%の急激な情報量の増加が続き、このままのペースが続けば2030年までには総情報量は1ヨッタバイトを超える（10の24乗バイト：70億人が一人当たり140 テラバイトを持つ）
- ストレージ容量の伸びは年率120%程度にとどまる

生成情報量の急増：ビヨンドビッグデータ





ICT技術の進歩を上回るペースで激増

- インターネット、クラウド、SNS、IoT、センサネット等 2015年の100億個から15年で**10兆個**を超える*
- 生成される情報量：10年で40倍
- モバイル通信トラヒック量の増大：10年で1000倍

*ローム社「NEハンドブックシリーズ．センサーネットワーク」

ストレージ容量(蓄積可能情報量)の限界

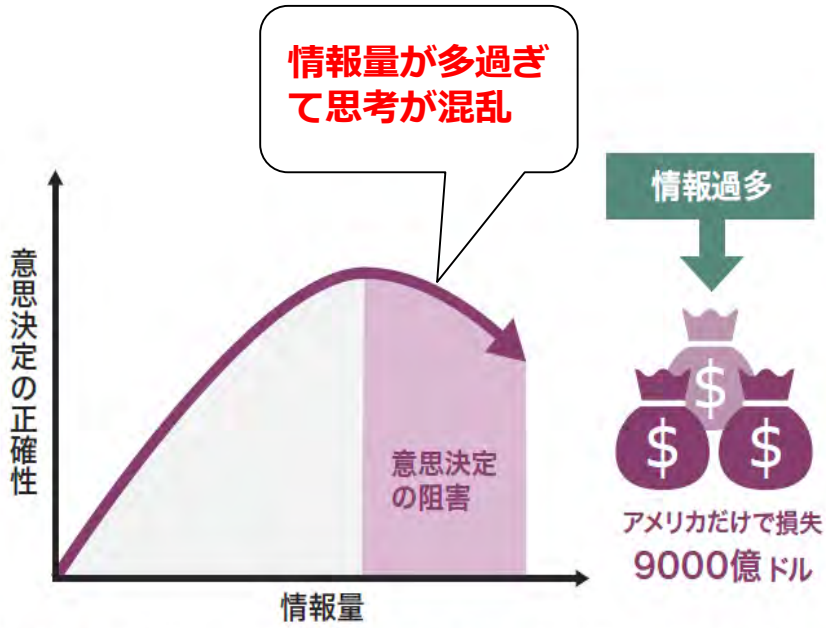
- 生成情報量は、2030年に1ヨッタバイト（1兆バイトの1兆倍）に達する
- ストレージ開発技術は10年で10倍程度
- データセンター容量、スペース、ネットワークトラヒックも限界。 消費電力、コストの増大

情報通信環境の課題

通信	演算	蓄積
送れない！	計算できない！	取っておけない！

社会と個人の負担が増加： 情報オーバーロード*

*J.B. Strother et al, “Information overload”, IEEE Press 2012.



参照 | Eppler, M. J. & Mengis, J. “The Concept of Information Overload: A Review of Literature from Organization Science, Accounting, Marketing, MIS, and Related Disciplines,” *The Information Society*, 20, 2004, pp. 325-344.
Spira, J.B. & Burke, C. “Intel’ s War On Information Overload : A Case Study” Basex., 2009.

量的オーバーロード

容量やネットワークに負荷がかかり、演算や通信が困難になる。

保存や管理に多大なコストがかかり、情報を蓄積することができない。

（参考）1ヨッタバイト保存に必要なコスト（5万円／100TBのHDD 1千億台の場合）
価格≒500兆円　消費電力≒東京都の総消費電力

知的オーバーロード

必要な情報を選択したり、有用な情報を検索するのが困難になる。

論理的に複雑なため意思決定や本質的理解が妨げられる。

過剰な情報による知的生産性の低下は米国だけで年間9千億ドルの損失。

情報の質

1ビットの情報量

1. 1ビットの情報は（等確率の）2通りのうちのどちらかが分かる情報量。人為的に制御できる2通りの状態を持つ手段で1ビットの情報を担う。

   → 電気のプラスとマイナス、磁気のN極とS極、など。

2. 災害時の家族の安否も、遠い天体の惑星の有無もいずれも1ビットで表現。現状の情報工学では等しくリソースを費やす。その価値によって扱いを変えることはない。

3. もし、価値の高い情報だけに限定して扱うことにすればオーバーロードの弊害は抑えられる。

情報の優先付け： 情報トリアージ*

*トリアージ：仏語のTrier（選別する）が語源であり、羊毛の品質をクラス分けやコーヒー豆の選別作業時に使われたと言われる。

これまでの情報優先付け

アクセス統計による
メモリ・ストレージの階層化

頻繁に参照される情報を（大事な情報に違いないので）優先的に処理

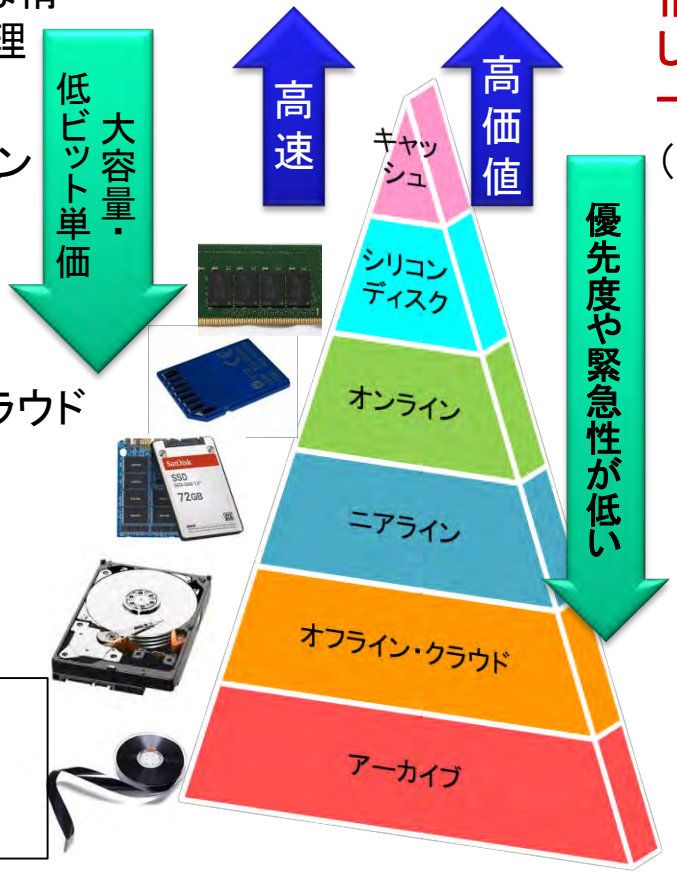
- 頻繁に読出し
  →高速キャッシュとオンライン
- 時々必要
  →ニアラインディスク
- 緊急性は低いが要保存
  →オフライン(テープ等)、クラウド
- 当面不要
  →アーカイブ

現在のストレージシステムでも
情報の**優先化や淘汰を行う**
→統計量を使う

情報の重要性に応じて、アクセスの高速化やデータの堅牢性を制御すべき

これからの情報トリアージ

質と価値に基づく情報トリアージへ



情報の「質」に応じて「価値」を判断し、その規範によって情報をトリアージ

（情報トリアージ分散ストレージ）

- 高価値情報は高速アクセスや高い保全性を確保する装置
- 公共情報はデータセンターへ
- 個人情報はローカルストレージ
- IoTセンサ情報はエッジで処理
- 不要情報の廃棄とアーカイブ

情報の質を見て価値を判断する情報エンジンを用いる**トリアージ情報処理の利用**

High Value Data （IDC）

https://japan.emc.com/leadership/digital-universe/2014iview/high-value-data.htm

- データドリブンを実現するにはデジタルユニバースの情報は多過ぎる上に拡散され過ぎていることが障害である。これを解決するにはターゲットリッチな情報を選ぶことであり、このために5つの“主観的な”尺度を導入する。
  - ✓ 容易なアクセス： データを入手できるか？誰かに占有されていないか？独自仕様の組み込みシステムに閉じ込められた情報ではないか？
  - ✓ リアルタイム： リアルタイム性は？決断や行動に使うには古過ぎないか？
  - ✓ フットプリント： 多くの人や組織の主要部分や多くの顧客に影響があるか？
  - ✓ 変革的： 解析やアクションにより企業や社会を有意義に変えることができるか？
  - ✓ シナジー： 上記の属性のうち一つだけではなくいくつかを兼ね備えているか？
- 2014年に “target rich data”は6%、2020年で11%に過ぎない。

情報の質に関する課題

- 不適切なデータによる人工知能の学習

「マイクロソフトの人工知能テイが差別的な発言」
（日本経済新聞2016.3.25）

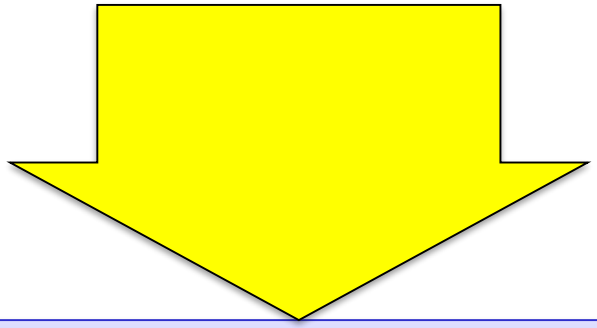
- 不適切な動画のアップロードが多発→人手によってこれを検知して削除

「投稿監視7割増7500人で　～フェイスブック　不適切動画に対応～」
（日本経済新聞2017.5.5）

- DeNAのキュレーションサイトでの不適切な引用

（読売オンライン2016.12.13）

- フィルタバブル（知的孤立）：　検索エンジンがユーザに応じて選択的に情報を提供するため多様な情報から隔離される

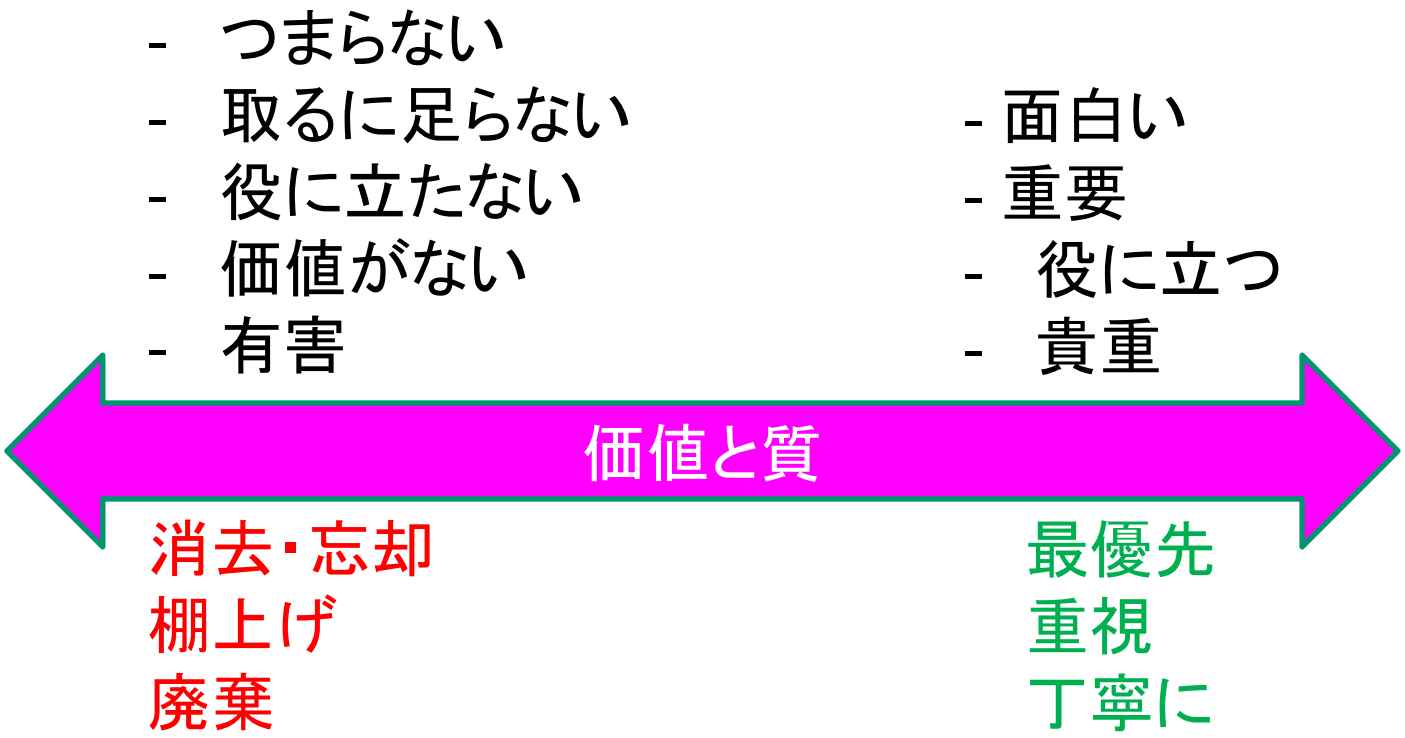


- ✓ 情報の量だけではなく「質」を考慮した情報利用がなされるべき
- ✓ 人間の判断による質の評価基準の抽出

質と価値を“測る”

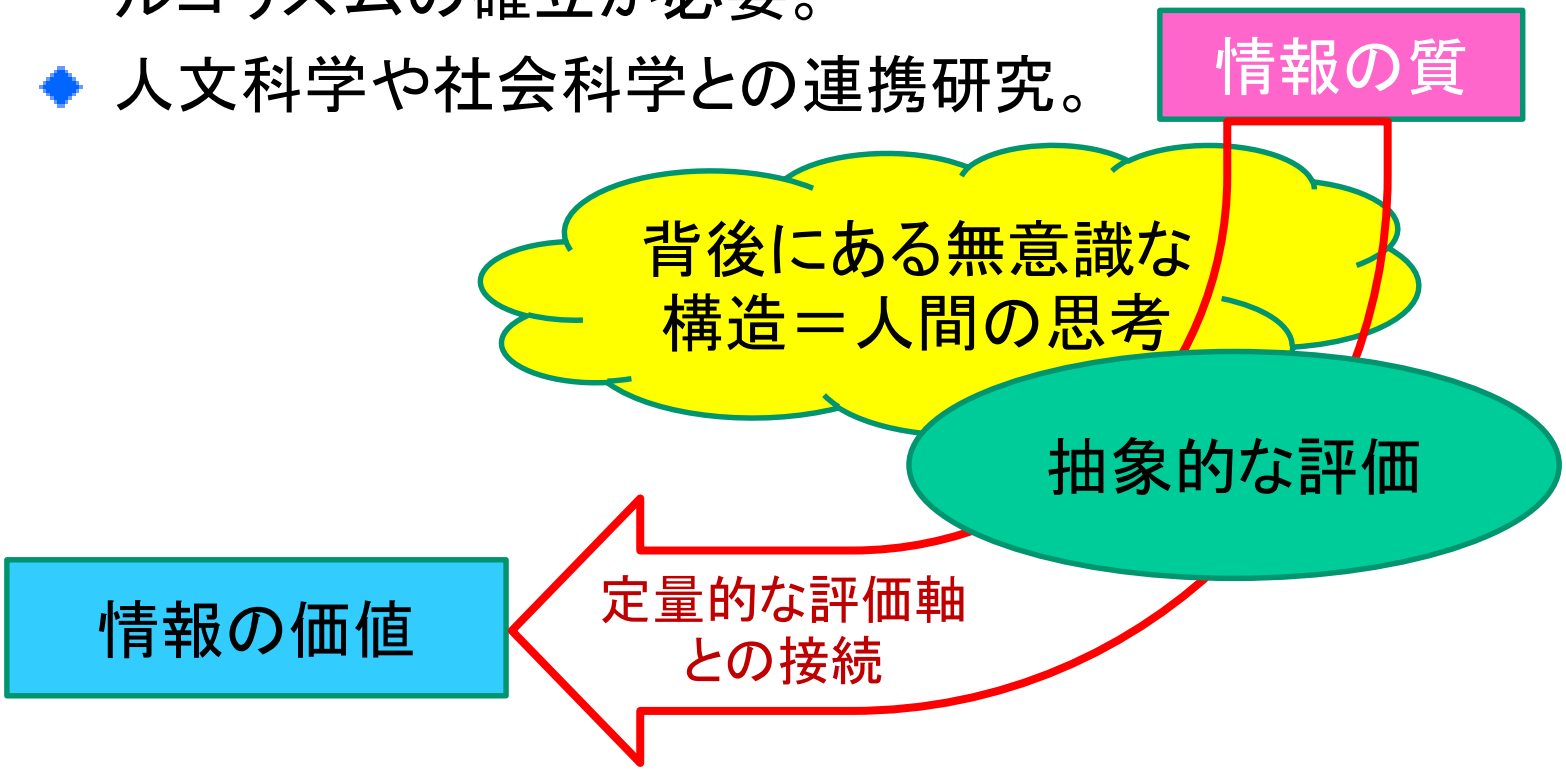
情報の質と価値

- 役に立たない情報を欲しがる人はいない。もしシステムが情報の質や価値を理解できれば自動的に情報を選ぶことができる。
- 情報の優先化が可能になる→情報のトリアージ
- 情報処理（伝送・蓄積・演算）の優先度を情報の価値に応じて決めることで無暗に大きな情報を相手にしなくて済む。
- その基準はしばしば抽象的な評価に基づく。



情報の抽象性

- 人間にとっての情報の質や価値は抽象的。
- 抽象性には無意識（暗黙）の理由が背後にある。
- 無意識性を網羅的な多数の定量的な評価の組み合わせを通じて評価して、抽象的な情報の価値をシステムに”判断”させるアルゴリズムの確立が必要。
- 人文科学や社会科学との連携研究。



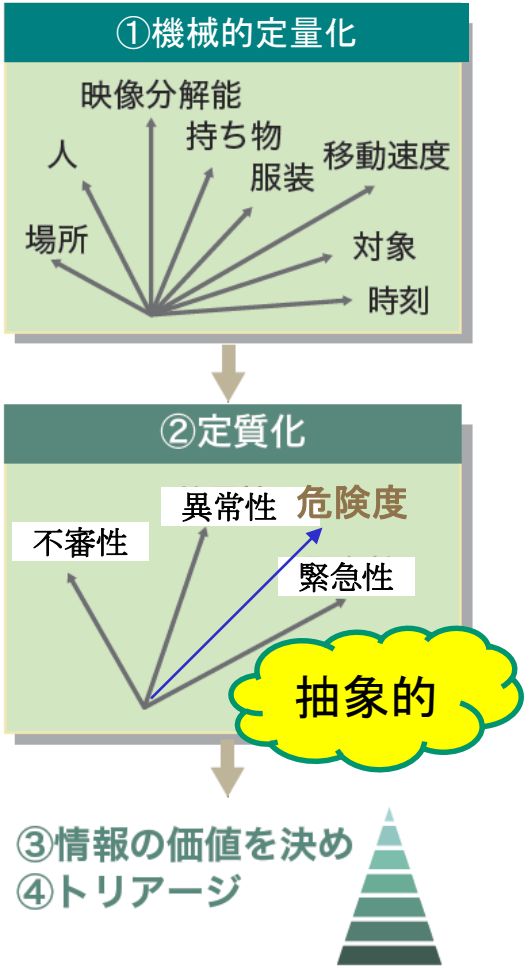
多次元ベクトルの評価軸による情報の価値判断

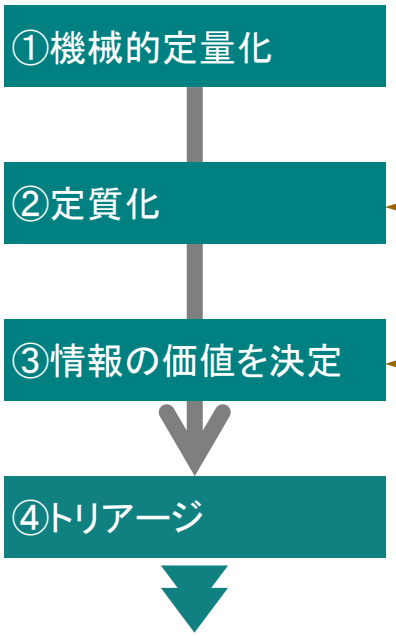
環境カメラの
セキュリティ管理のケース

カメラ画像からセキュリティ管理のため異常やリスクを検知することを目的としたトリアージの例

※動画・画像データは日本のビッグデータ流通量の58.4%を占めるため、トリアージのターゲットとして重要

参照| 総務省2015『平成27年版情報通信白書』





知識構造の利用

ヒトの評価に対応する画像特徴の同定、モデル化

- ○犯罪者心理、社会環境などに関する人文社会科学の知見
- ○警備員、警察官等専門家による評価
- ○利用者の評価

・価値の高い情報消失防止
・必要な情報の優先的処理

代表研究例①　監視カメラ画像からの危機予測

◎人の意図予測技術によるIoT画像の人間的判断AIの実現

①

定質化

②　不審性　異常性

価値評価

③
④危険度

応用例

犯罪防止
危険予測

展開例

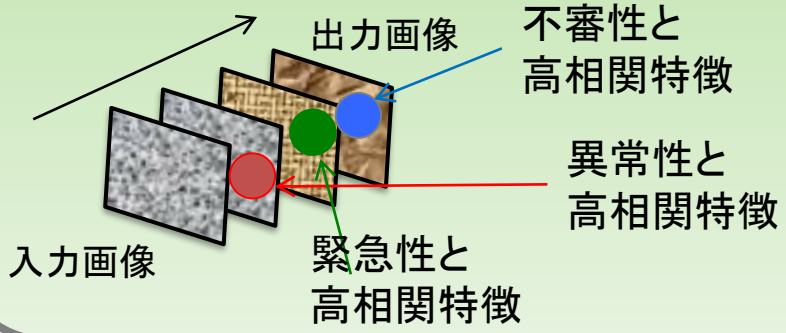
安全・安心な生活環境の構築

犯罪心理学に基づく防犯・テロ防止、弱者（迷子、認知症・急病人等）保護、映像・アート等の真贋・来歴判定

定質化計算：

画像特徴から「質」情報を抽出

深層学習（先端的AI 技術）の結果から質判断に関するヒトの評価の基礎にある特徴を推定

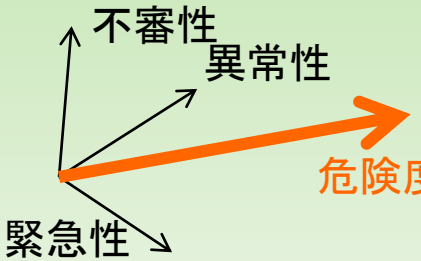


価値計算：

「質」情報から価値を決定

統計モデルから価値に関するヒトの評価の基礎にある質の組み合わせを推定

->人間的判断AIの実現



人文系の知恵に基づく評価

・犯罪心理
・行動科学
・認知科学
：

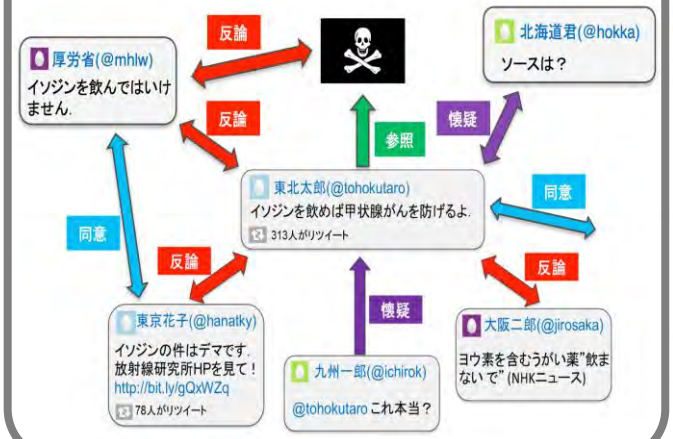
代表研究例② Web情報の信頼度

◎ 自然言語処理技術により情報の信頼性を確保

①
- ○イソジンを飲めば甲状腺がんを防げる。
- ○ヨウ素を含むうがい薬“飲まないで”（NHKニュース）
- ○イソジンの件はデマです。放射線研究所のHPを見て!
- ○イソジンを飲んではいけません
- ○震災の混乱に乗じた悪質な流言に注意を新聞で読んでびっくり。
- ○ヨウ素を含む消毒剤などを飲んではいけません。-インターネット等に流れている根拠のない情報に注意-

定質化

②



価値評価

③

④ 信頼度

- ○ヨウ素を含む、
- ○震災の…
- ○ヨウ素を…
- ○イソジンを…
- ▪
- ▪
- ▪

応用例

高信頼情報の確保・フェイクニュースの排除

展開例

信頼できるインターネット環境の構築

公正な株取引実現、マーケティング　への正確な意思決定等

定質化計算：　関連図から「質」情報を抽出

自然言語解析の結果から質判断に関するヒトの評価の基礎にある関係を推定

言語構造解析による一般・専門知識の自動獲得

深層学習（AI）による構成的な意味分散表現と論理の統合

人文系の知恵を利用した評価
- ・言語学
- ・社会学
- ・倫理学
- ▪

ネット上の情報の収集・集約・可視化

価値計算：　「質」情報から価値を決定

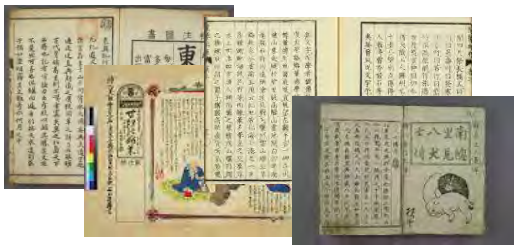
統計モデルから価値に関する専門家の評価の基礎にある質の組み合わせを推定

最先端AIの実現

代表研究例③　古典籍重要情報の発見

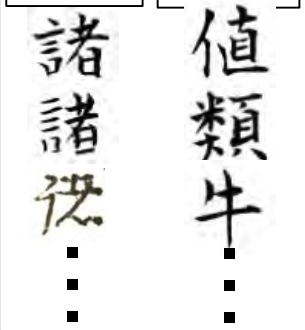
◎ 手書き文字の高速高精度認識技術の古典籍ビッグデータへのAI展開

① 国文研古典籍データ
画像データ　1ペタ（10^{16}）バイト
□源氏物語、二十一代集、
□仁義禮智信／青物談義、
□各国盛衰強弱一覧表／附図、…



定質化

② 同一文字　同一書体



価値評価

③ ④ 重要性

□方丈記
（3巻45頁）
□日本三代実録
（2巻9頁）
□吾妻鏡
（74年の記録期間中に210余件の地震記述）

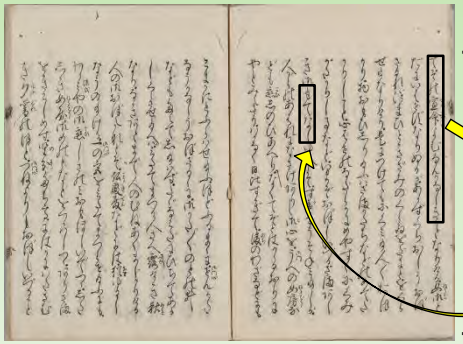
応用例

古典籍を活用した史実の分析・解読

展開例

温故知新・文化力向上：過去の知見と現代の技術のｺﾗﾎﾞﾚｰｼｮﾝによる新価値創出
日本の文化基盤の解明と独自の表現様式の国際展開、歴史的未解決事件の解明、防災減災計画への指針等

定質化計算：文字特徴の「質」に基づく文字画像生成
文字特徴分析の結果から質（字体）判断



画像抽出（質判断）

認識が困難な続け字にも対応可能

テキスト「もてなし」

検索

価値計算：「質」情報から価値を決定
研究者の評価、研究成果から価値を決め、そこに貢献する質を推定
（古典籍ビッグデータへのAI展開）
○和漢薬の効能（富山大）
○オーロラ記述（国立極地研究所）
○宝永地震海岸線沈降
（南海トラフ広域地震防災研究プロジェクト）
などを想定した調査

人文系の知恵を発展させた評価
・文学
・史学
・考古学
…

3つの研究テーマ

A：基盤テーマ
情報の定質化とトリアージ手法の開拓

- 多次元ベクトル表現による情報質定質化の基本概念構築
- 情報の価値を認識する人文社会科学規範の導入
- 情報構造化アルゴリズムとデータ処理手法
- 情報トリアージ規準の設定

B：応用テーマ
情報質評価の確立と分野別汎化

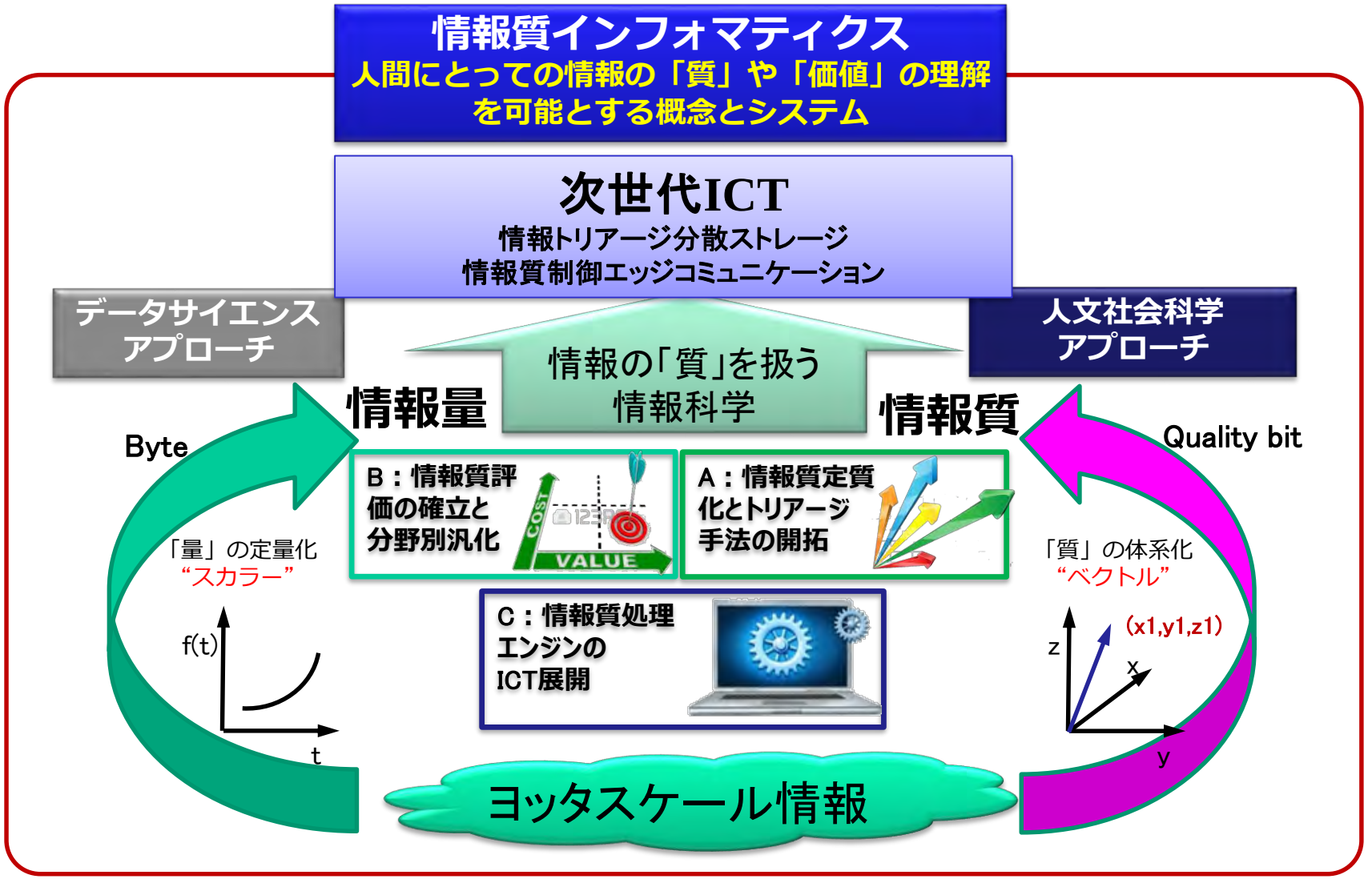
- 経済、アート、古典籍を中核とする展開と分野ごとの汎化
- 映像情報を中心とする情報の価値の定質化と情報構造化
- 情報トリアージのための文化的価値と経済的価値の統合的利用

C：展開テーマ
情報質処理エンジンのICT展開

- 情報質処理エンジンの基盤アルゴリズム開発
- 情報トリアージ手法の次世代ストレージ技術への統合
- 高速ネットワーク通信における情報クオリティアルゴリズムの具体化

情報質インフォマティクスの構築

情報「量」だけでなく情報「質」を理解し価値に応じて処理できるインフォマティクスへ

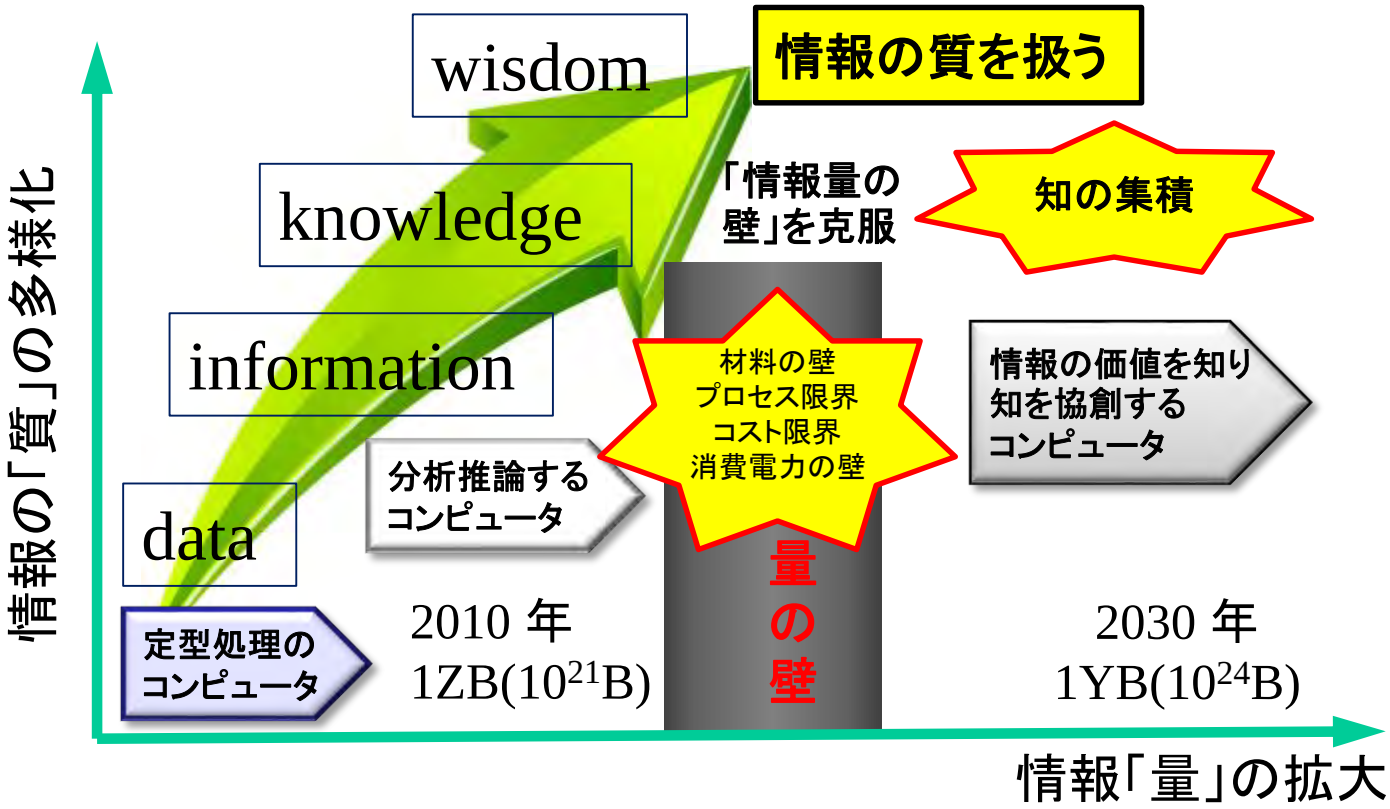


情報「量」の壁を情報「質」で超える

- 情報を「質」による価値付けで優先選別することで情報「量」の壁を超える
- 情報の質と価値の研究には人文社会科学の知を導入→文理連携の新情報学

→**人間の価値判断**に倣う情報処理を行い、**質が高く価値のある情報を集約**

→データと情報だけを扱う情報工学から知識と知のインフォマティクスへの進化（DIKW進化）



情報量（横軸）の発展だけでは情報量の壁を越えられない。情報質（縦軸）の導入により情報の価値を集約して組織と知のサイエンスを作る

まとめ

- 情報科学技術の進展と社会への浸透に伴い、人類が生成する情報量は爆発的な増加を続け、2030年までに1ヨッタバイト（1兆バイトの1兆倍）に達する可能性がある。これはストレージの蓄積可能容量が追い付かないほど大きな情報量である。
- ビッグデータを超えるこの巨大な情報量をどう扱うか。
- 一方、情報には量だけではなく質の側面がある。同じ1ビットの情報量でもその価値は異なる。重要な情報と不要な情報を相応に扱うことで情報量の拡大に対応できる。
- 価値のある情報を優先的処理するために、システムが情報の「質」を理解できるために新しいインフォマティクスが必要ではないか。
- 情報の質を扱うには人間の抽象的な知による必要があり、理工系だけの取組ではなく、人文科学や社会科学の連携のもとでの研究が必要。