



# 独自開発の人工知能応用技術で ビッグデータを解析する 日本発ベンチャー UBI C

— 訴訟大国・米国に挑み磨かれた技術とは



株式会社UBIC 機関投資家向けセミナー  
2014年3月6日



# UBICの技術的特徴

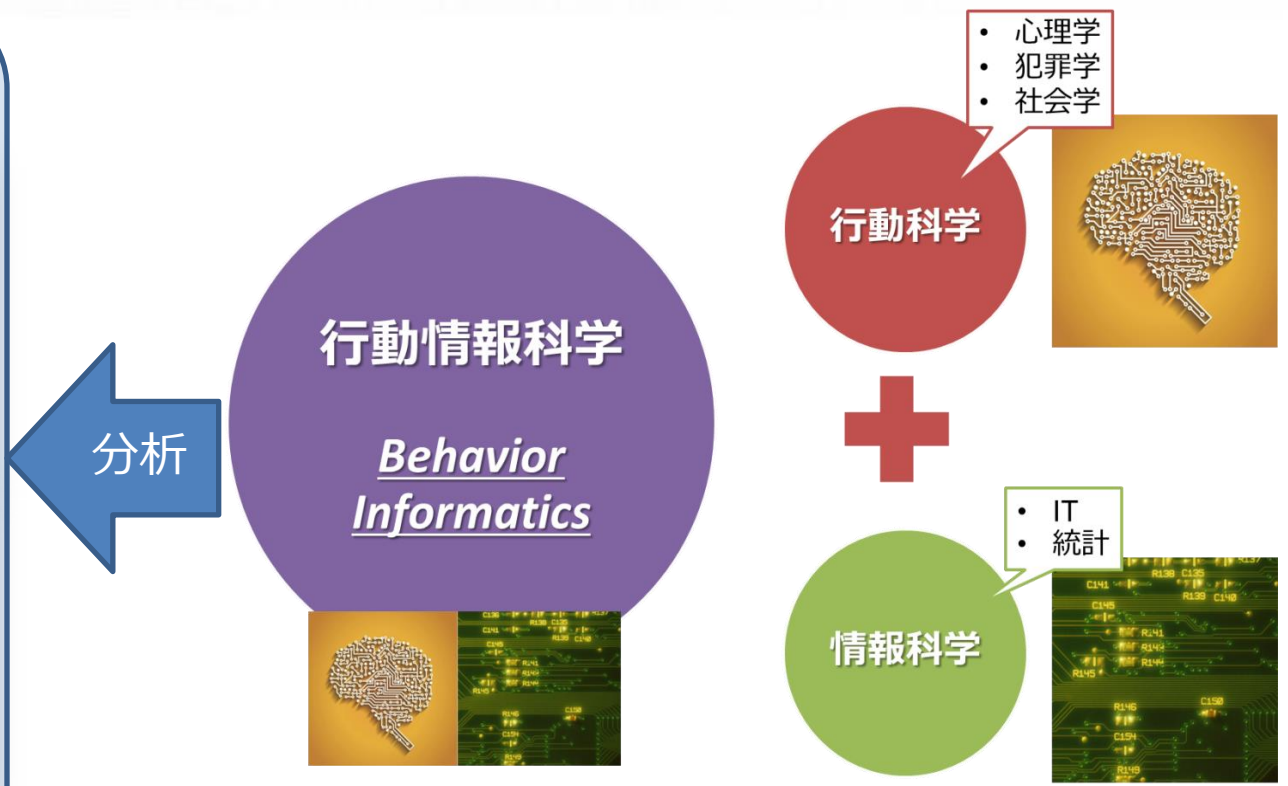
## Big Data



思考結果

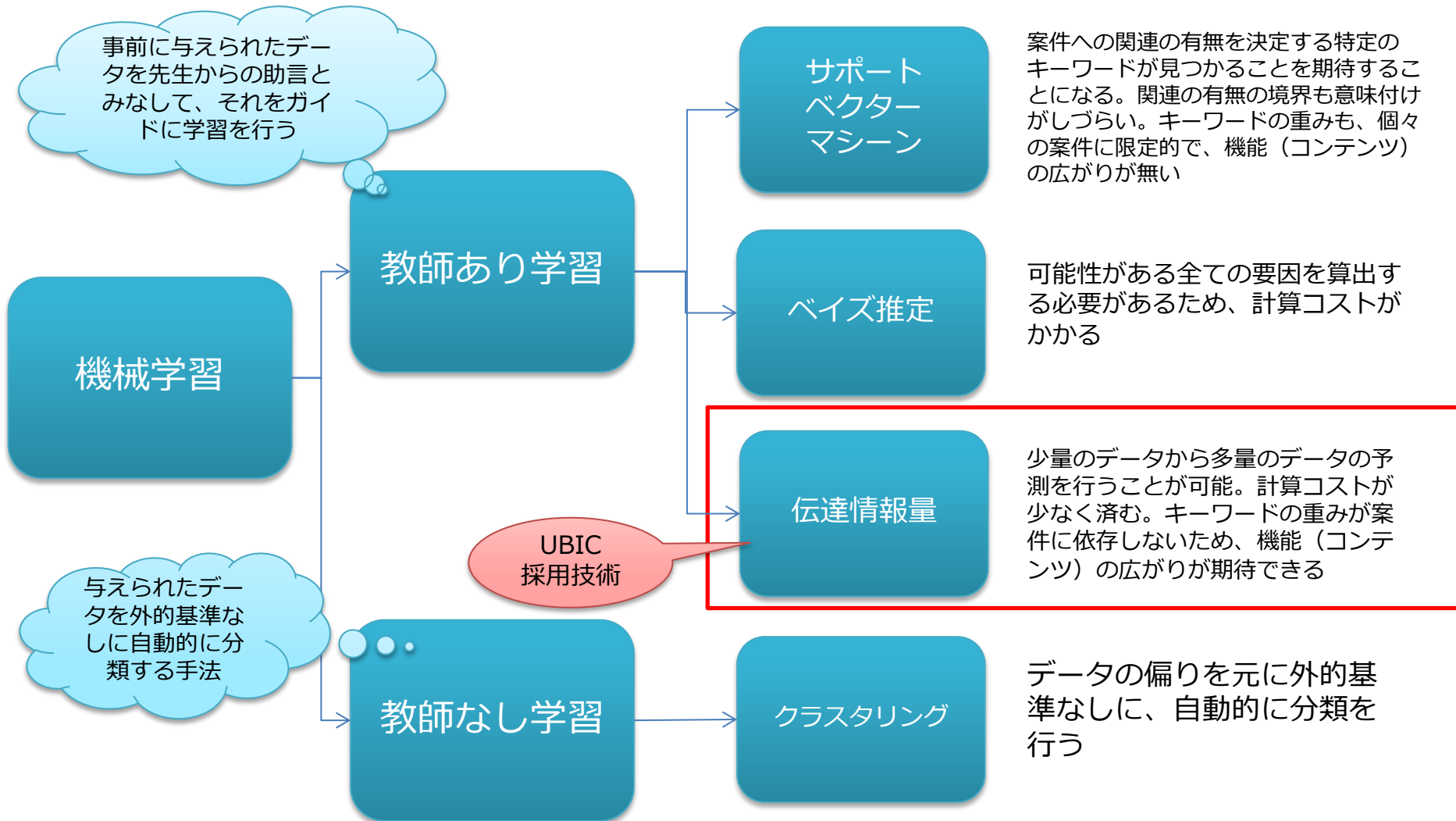


行動結果



ビッグデータを 人の思考と行動の結果 による データの集合体 として捉え、 行動科学 と 情報科学 を融合させた 行動情報科学 によって分析

## 教師あり学習 (*Supervised learning*)





# UBICの技術力の源泉：ユニークな開発チーム UBIC

## 武田 秀樹 (UBIC CTO)

1996年 早稲田大学を卒業、専攻は哲学。  
複数のシステム開発会社で経験し、  
2009年、UBICのCTOに就任。  
UBICでユニークな開発チームを立ち上げ、UBIC  
の誇るLit I Viewのプラットフォームの開発、人工  
知能「Virtual Data Scientist」の開発の指揮を  
とる。



## 蓮子 和巳 博士 (理学)

1999年 東北大学大学院理学研究科物理学専攻  
博士課程修了。在籍中、スタンフォード  
大学の関連研究所で研究。  
同年、東京大学素粒子物理国際研究センター  
特別研究員。次世代粒子加速器 (LHC) 実験  
装置の設計、開発および新粒子探索の研究に  
従事。在籍中、欧州原子核研究機構  
(CERN) で研究。  
2002年 独立行政法人理化学研究所 研究員。  
ブルックヘブン国立研究所 (BNL)などで粒子  
加速器を用いた量子力学的現象 (スピン現  
象) の研究に従事。



## ハルスコウ ヤコブ 計算言語学博士

2007年 コペンハーゲンビジネススクール計  
算言語学テキストマイニング専攻 博士課程  
修了。  
アジア言語を含む4か国語を操る語学力およ  
びテキストマイニングと言語学の造詣を生か  
し、ネットワーク上のコミュニケーションの  
分析精度を如何に向上させるかを主な研究  
テーマとする。



量子力学、計算言語学などのビッグデータを解析  
する究極の手法を選択

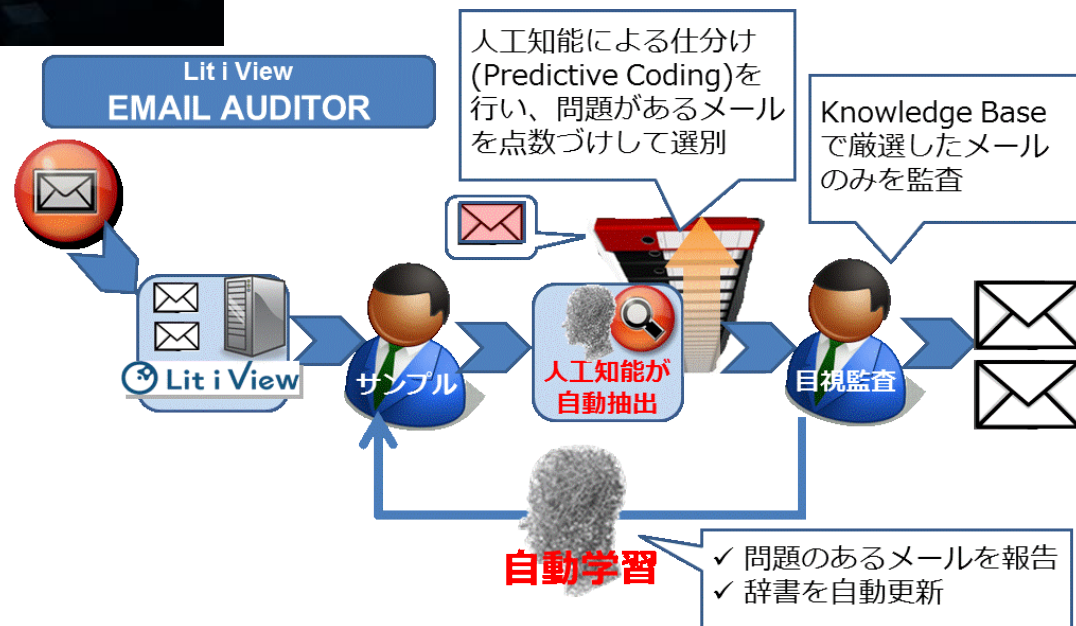


# UBICのソリューション



これまで不可能であった  
メール常時監視を  
**人工知能が**  
正確かつ低コストで実現

カルテル、  
知的財産評価、  
情報漏えい等の  
調査に活用





## 医療分野



### 自動鑑別診断システム



## 知的財産分野



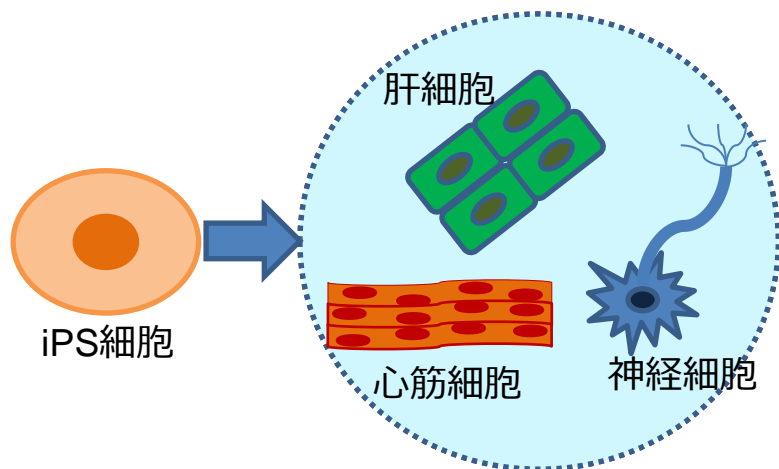
### 人工知能応用 知財分類システム



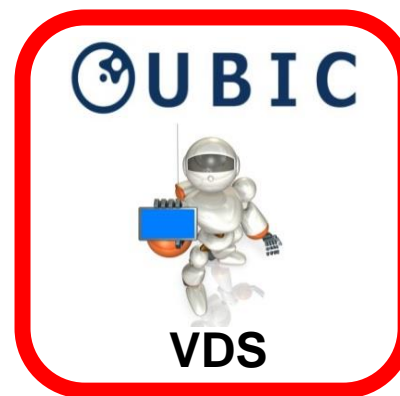


## Virtual Data Scientist

### ビッグデータ解析技術 におけるiPS細胞



人工知能VDS



医者



弁護士



警察官



# **UBIC** 機関投資家向けセミナー

UBICのビッグデータ解析技術はなぜ“UBICにしか”開発できないのか？



**NASDAQ**<sup>®</sup>  
LISTED

Date: 6 Mar. 2014

Tokyo | Osaka | Nagoya | Seoul | Taipei | Hong Kong | Silicon Valley | Washington DC | New York | London



Lit i View  
**EMAIL AUDITOR**  
Eメール監査

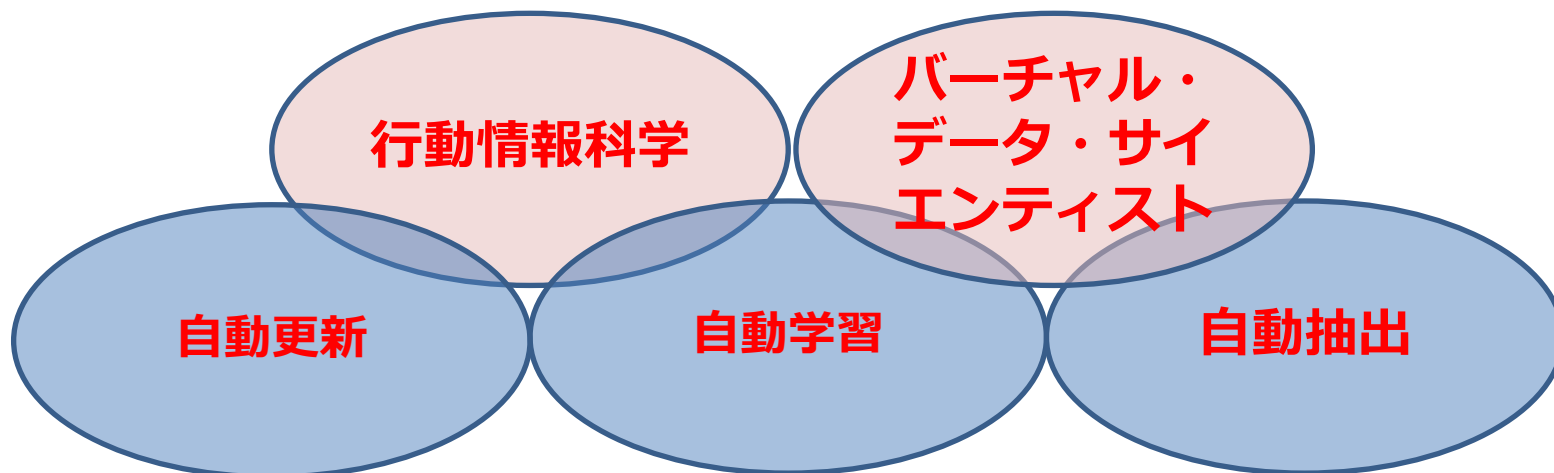


# メール監査を自動化することにより...

✓ 人的監査と比較し、4000倍の速度に  
⇒ 50時間を1分に

✓ 大幅なコスト削減

✓ Emailの常時監査を可能に



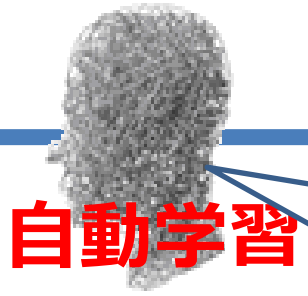
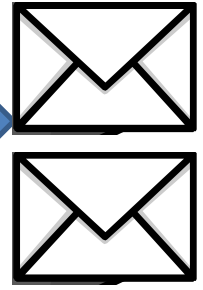
# Email監査システム



Lit i View  
EMAIL AUDITOR

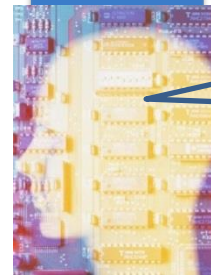
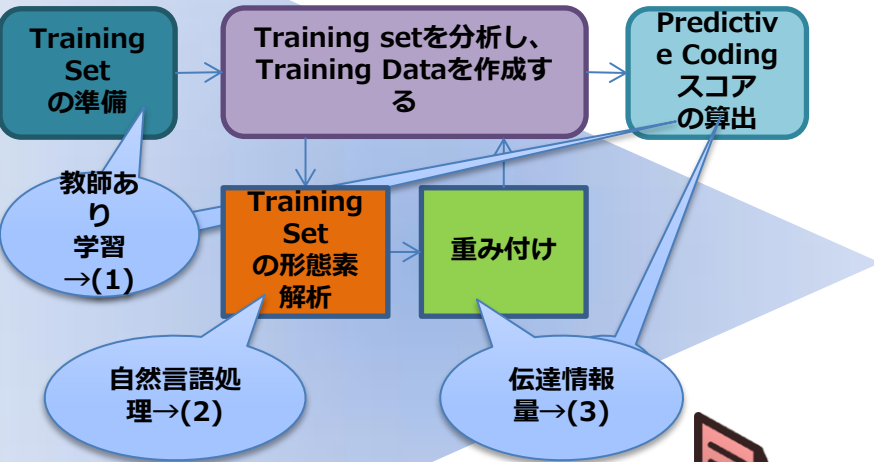
人工知能による仕分け  
(Predictive Coding)  
を行い、問題がある  
メールを点数づけして  
選別

Knowledge Base  
で厳選したメール  
のみを監査

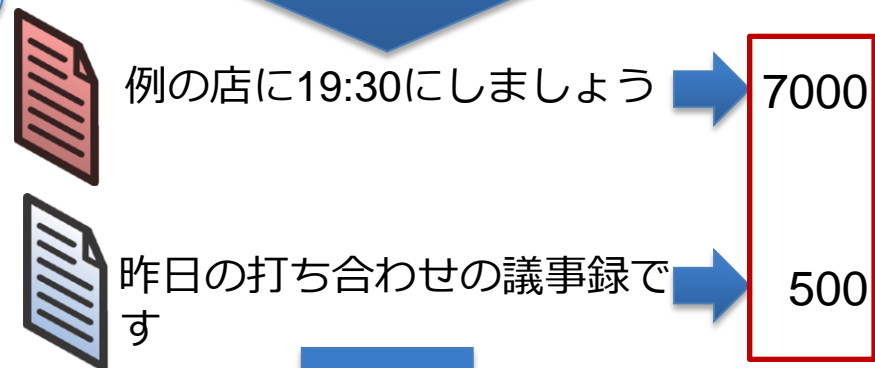


- ✓ 問題のあるメールを報告
- ✓ 辞書を自動更新

未レビューの文書に対し、  
Predictive Codingを行い、  
スコアを付ける



バーチャルデータサイエンティスト

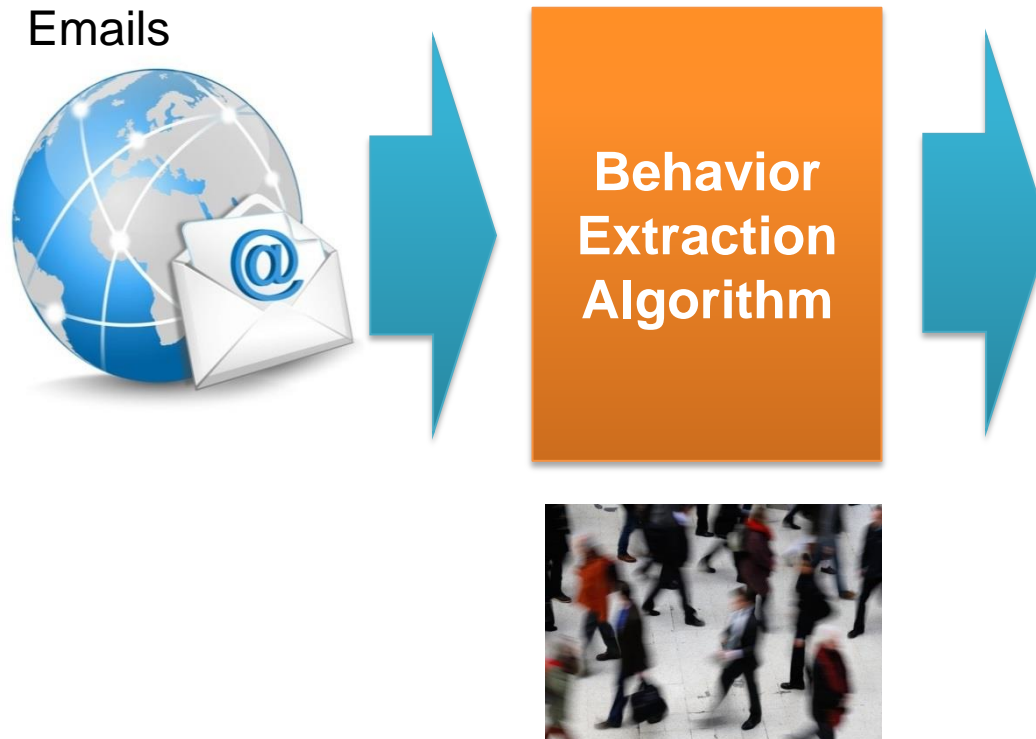


スコアの高い文書から優先的にレビュー



## 人間行動抽出機能

人のどのような行動に着目すべきかを  
Virtual Data Scientistが示唆

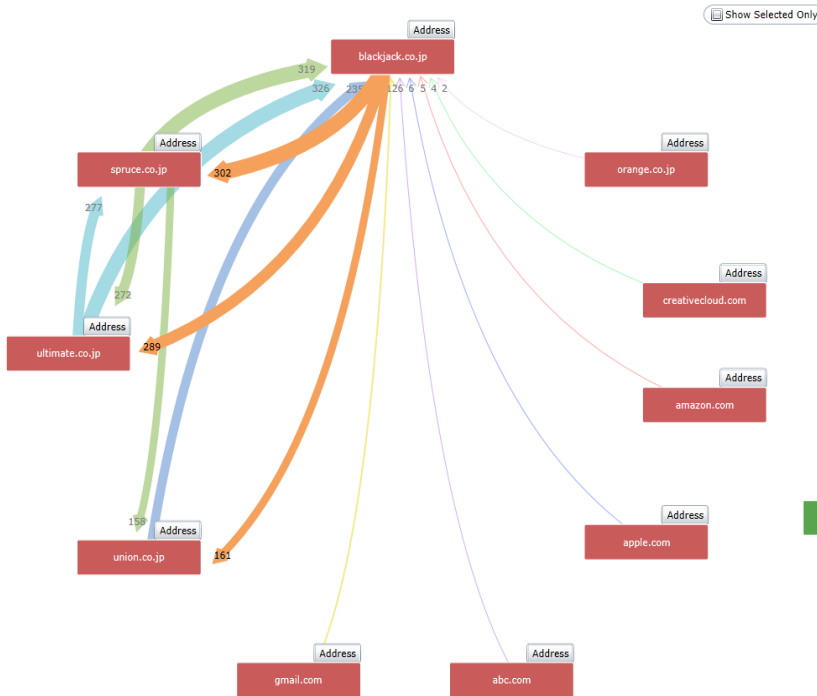


## List of Behavior

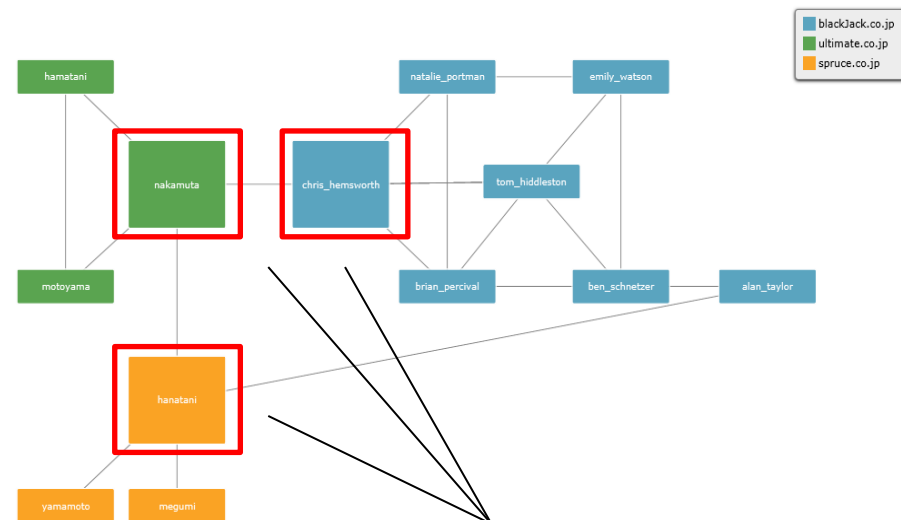
	Keyword1	Keyword2		
1	情報	Get	入手	Information
2	情報	Gather	収集	Information
3	情報	Provide	提供	Information
4	価格	Negotiate	交渉	Price
5	打ち合わせ	Have	実施	Meeting
6	サンプル	Deliver	納入	Sample
7	価格	Reply	回答	Price
8	技術Technology		交流	Exchange
9	新規	New	受注	Order
10	内部	Discuss	検討	Internally

**Extract Verb +  
Noun**

## 組織間相関図



Virtual Data Scientistは  
人物間だけでなく、  
組織間の関係も分析



組織間の主要窓口

# 誰が開発したのか？



# UBICの技術力の源泉：ユニークな開発チーム UBIC

## 武田 秀樹 (UBIC CTO)

1996年 早稲田大学を卒業、専攻は哲学。  
複数のシステム開発会社で経験し、  
2009年、UBICのCTOに就任。  
UBICでユニークな開発チームを立ち上げ、UBIC  
の誇るLit I Viewのプラットフォームの開発、人工  
知能「Virtual Data Scientist」の開発の指揮を  
とる。



## 蓮子 和巳 博士 (理学)

1999年 東北大学大学院理学研究科物理学専攻  
博士課程修了。在籍中、スタンフォード  
大学の関連研究所で研究。  
同年、東京大学素粒子物理国際研究センター  
特別研究員。次世代粒子加速器 (LHC) 実験  
装置の設計、開発および新粒子探索の研究に  
従事。在籍中、欧州原子核研究機構  
(CERN) で研究。  
2002年 独立行政法人理化学研究所 研究員。  
ブルックヘブン国立研究所 (BNL)などで粒子  
加速器を用いた量子力学的現象 (スピン現  
象) の研究に従事。



## ハルスコウ ヤコブ 計算言語学博士

2007年 コペンハーゲンビジネススクール計  
算言語学テキストマイニング専攻 博士課程  
修了。  
アジア言語を含む4か国語を操る語学力およ  
びテキストマイニングと言語学の造詣を生か  
し、ネットワーク上のコミュニケーションの  
分析精度を如何に向上させるかを主な研究  
テーマとする。



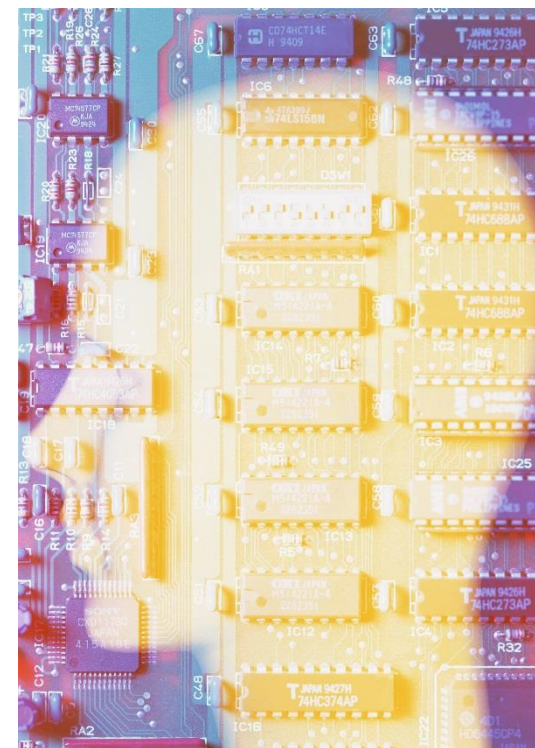
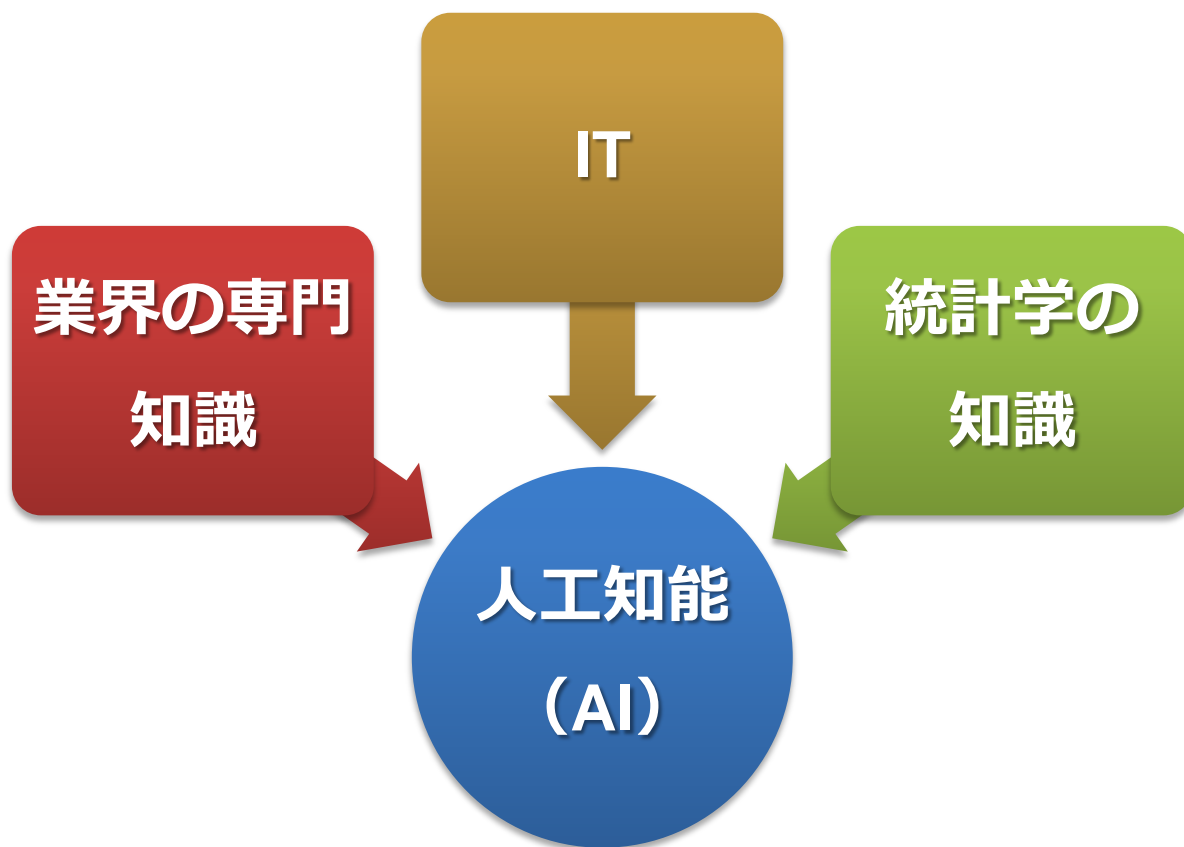
量子力学、計算言語学などのビッグデータを解析  
する究極の手法を選択

## 武田 秀樹 (UBIC CTO)

1996年 早稲田大学を卒業、専攻は哲学。  
複数のシステム開発会社で経験し、  
2009年、UBICのCTOに就任。  
UBICでユニークな開発チームを立ち上げ、  
UBICの誇るLit | Viewのプラットフォームの開発、  
人工知能「Virtual Data Scientist」の開発の指揮をとる。

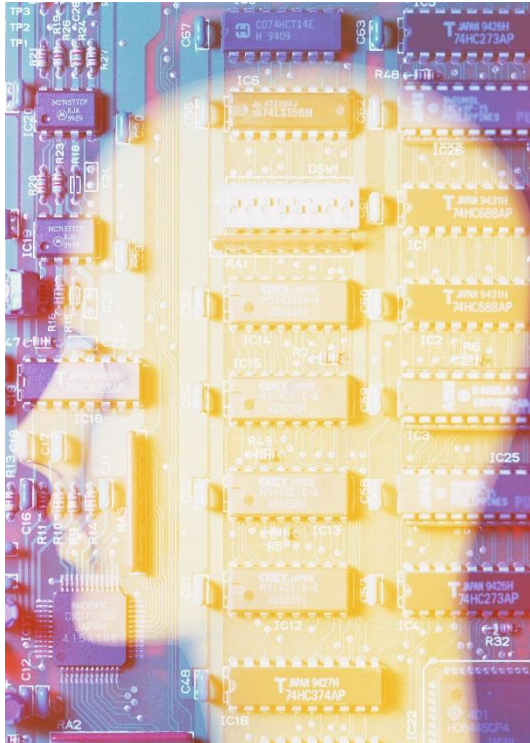


1. ユニークなコンセプト
2. サービスによるノウハウの蓄積
3. 研究→開発の速さ



**人工知能によって人の思考・行動を  
分析し、未来を予測**





人工知能



人間の知的判断を代替するシステム



何を目的にしたシステムか？



情報の仕分け・証拠の発見

## Big Data



分析

行動情報科学

*Behavior Informatics*



行動科学

- 心理学
- 犯罪学
- 社会学



情報科学

- IT
- 統計



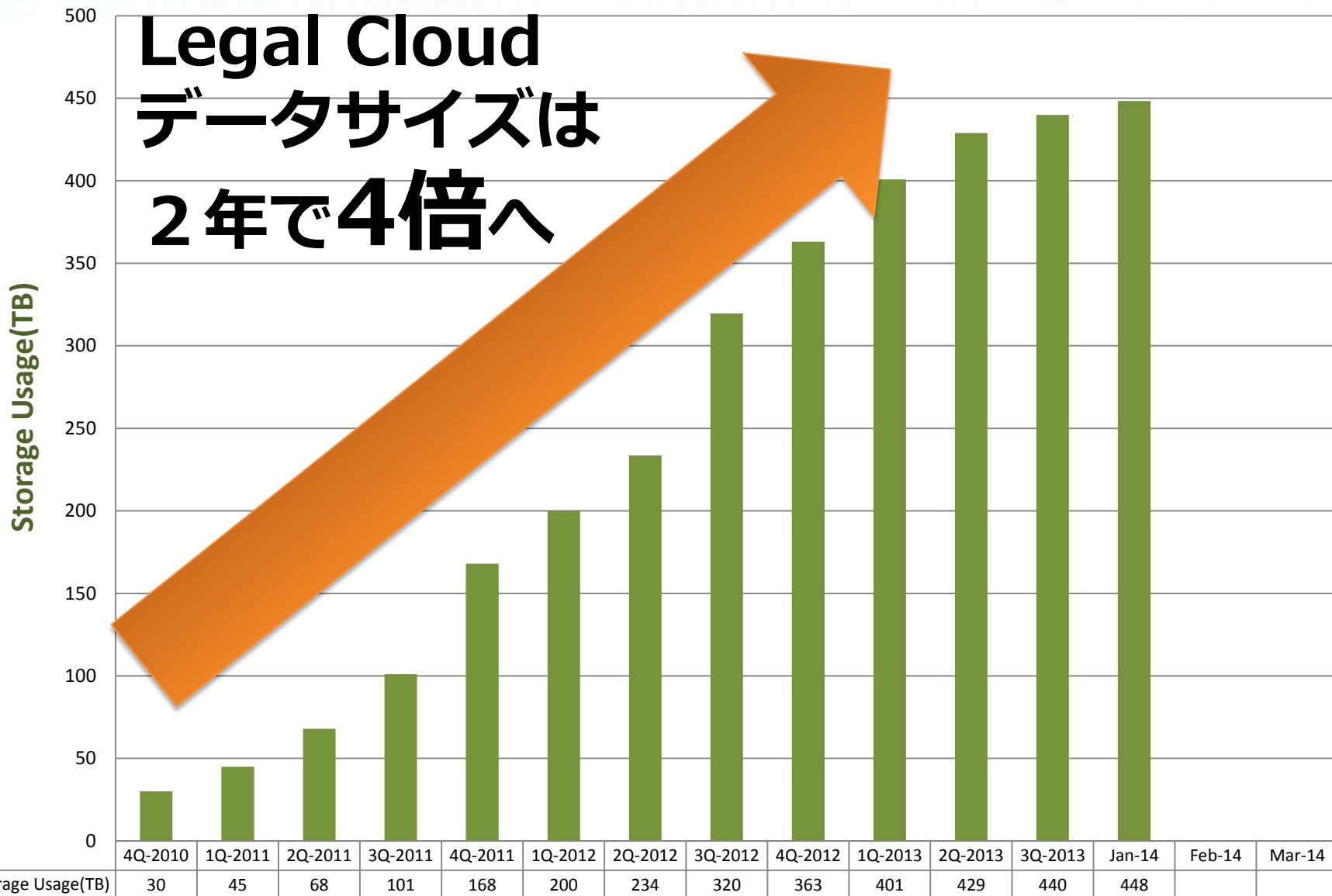
ビッグデータを人の思考と行動の結果によるデータの集合体として捉え、行動科学と情報科学を融合させた行動情報科学によって分析

# Legal Cloud



Copyright(C) 2007 TSUKUI International Inc. All Rights Reserved.

Mar / 2014





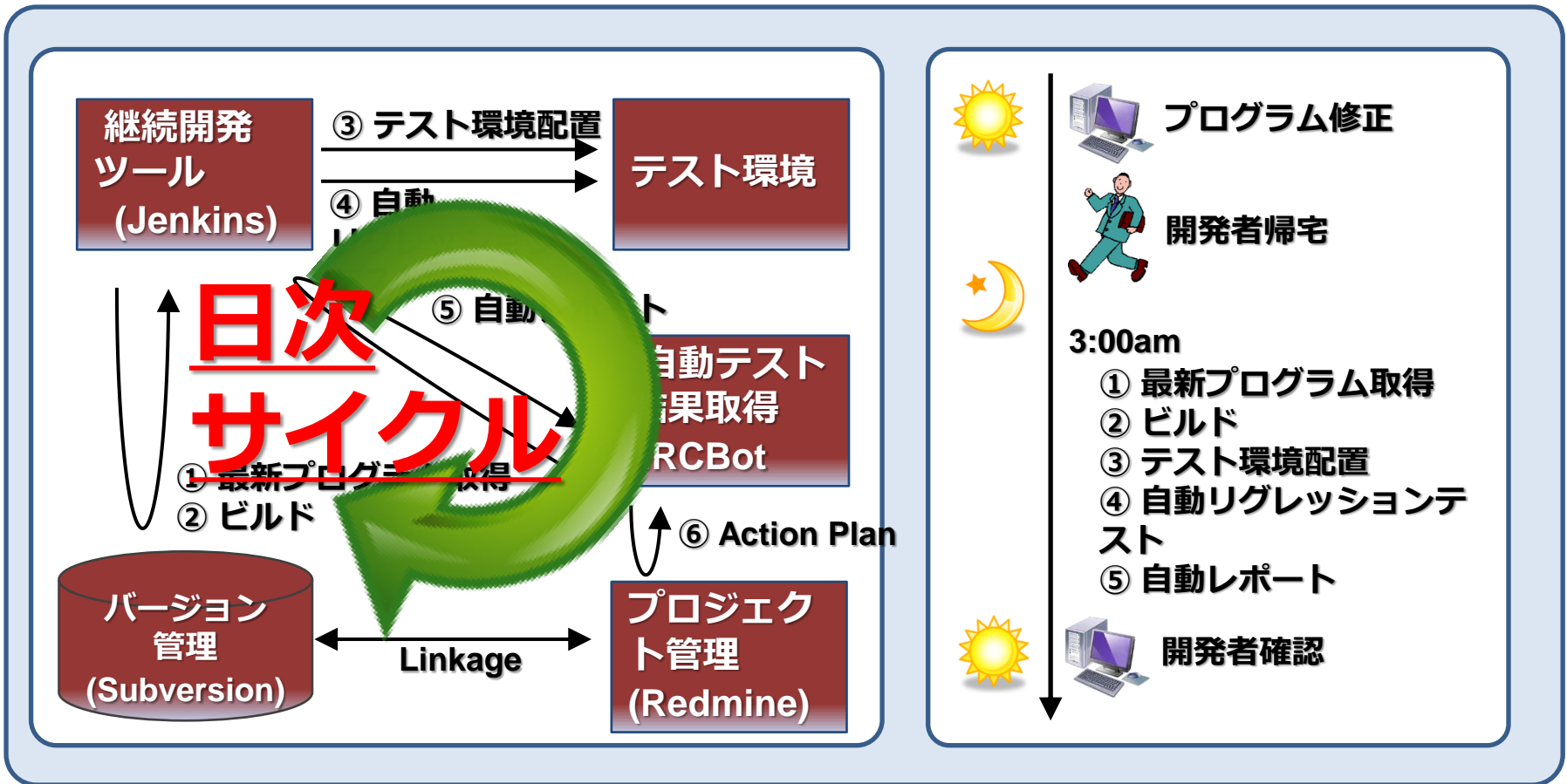


## UBIC カルテル辞書

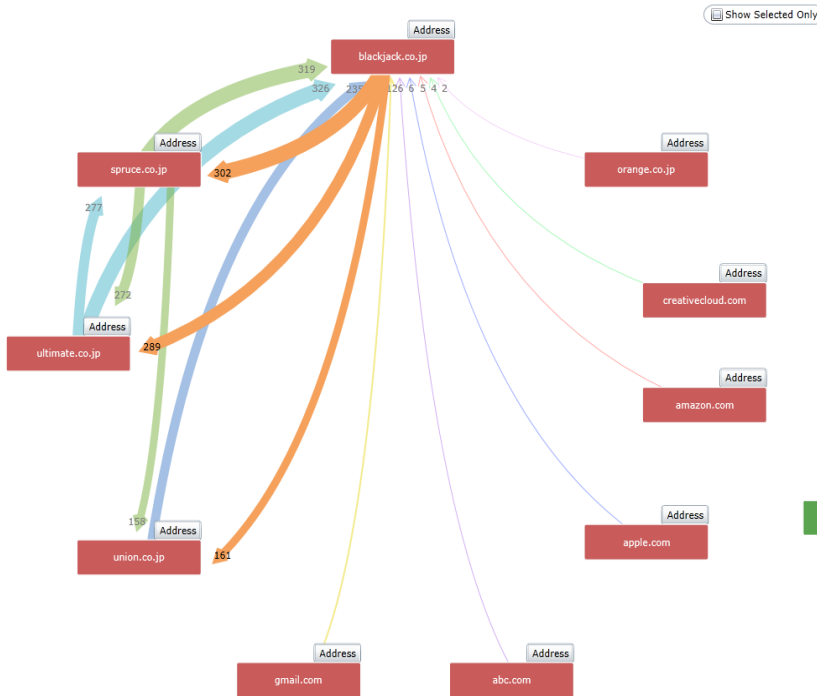
専門家の知識を蓄積

価格	price	競合	compete
見積り	estimation	提示	propose
製品	product	対応	accommodate
メーカー	maker	出荷	ship
価格	price	提出	propose
見積もり	estimation	依頼	ask

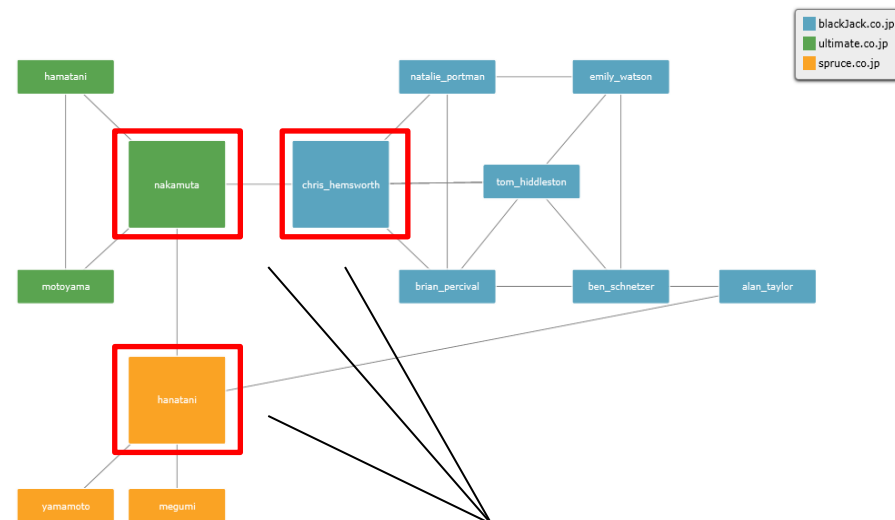
# 独自の自動テストシステムの構築により、開発時間短縮と高品質維持を両立



## 組織間相関図



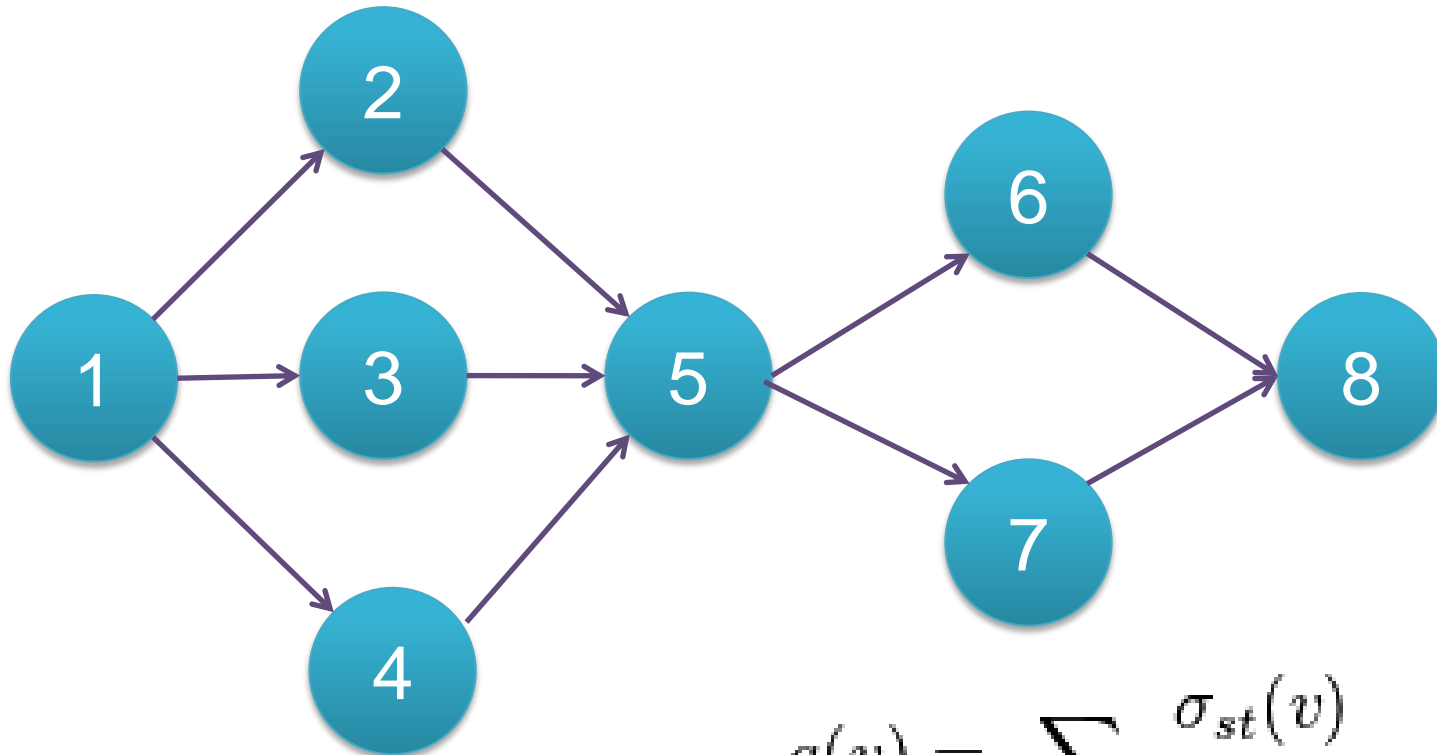
Virtual Data Scientistは  
人物間だけでなく、  
組織間の関係も分析



組織間の主要窓口

## 例1. 社会学 (社会ネットワーク分析)

人間関係をどの程度媒介しているのかを分析。特定のコミュニケーション・行動だけに絞り、その人間関係を把握、また、組織間のコミュニケーションの窓口になっている人間のコミュニケーションを把握



$$g(v) = \sum_{s \neq v \neq t} \frac{\sigma_{st}(v)}{\sigma_{st}}$$



分析対象そのものの構造を定性的に分析。対象の構造と人間の行動を統計的に評価し、分析を自動化する

## 例2. 犯罪学

カルテルの構造を分析、人間の行動を分析、その構造特有の行動やコミュニケーションを分析することで、状況を自動的に把握

# ハルスコウ ヤコブ 計算言語学博士

2007年 コペンハーゲンビジネススクール計算言語学テキストマイニング専攻 博士課程修了。  
アジア言語を含む4か国語を操る語学力およびテキストマイニングと言語学の造詣を生かし、ネットワーク上のコミュニケーションの分析精度を如何に向上させるかを主な研究テーマとする。



1. 自然言語処理を駆使し、非構造化データを構造化
2. マルチリンガルを活かし、言語の特徴に応じた自然言語処理を行う

## 人間行動抽出機能

人のどのような行動に着目すべきかを  
Virtual Data Scientistが示唆

Emails



## List of Behavior

	Keyword1	Keyword2		
1	情報	Get	入手	Information
2	情報	Gather	収集	Information
3	情報	Provide	提供	Information
4	価格	Negotiate	交渉	Price
5	打ち合わせ	Have	実施	Meeting
6	サンプル	Deliver	納入	Sample
7	価格	Reply	回答	Price
8	技術Technology		交流	Exchange
9	新規	New	受注	Order
10	内部	Discuss	検討	Internally



**Extract Verb +  
Noun**

# バベルの塔

問題： **国際化** => **多言語・多文化的な通信の分析**  
が必要となります。

解決：コンピューターによる言語解析。  
**自然言語処理 (NLP)** という技術です。



# 言語解析って簡単な三角に過ぎませんか？

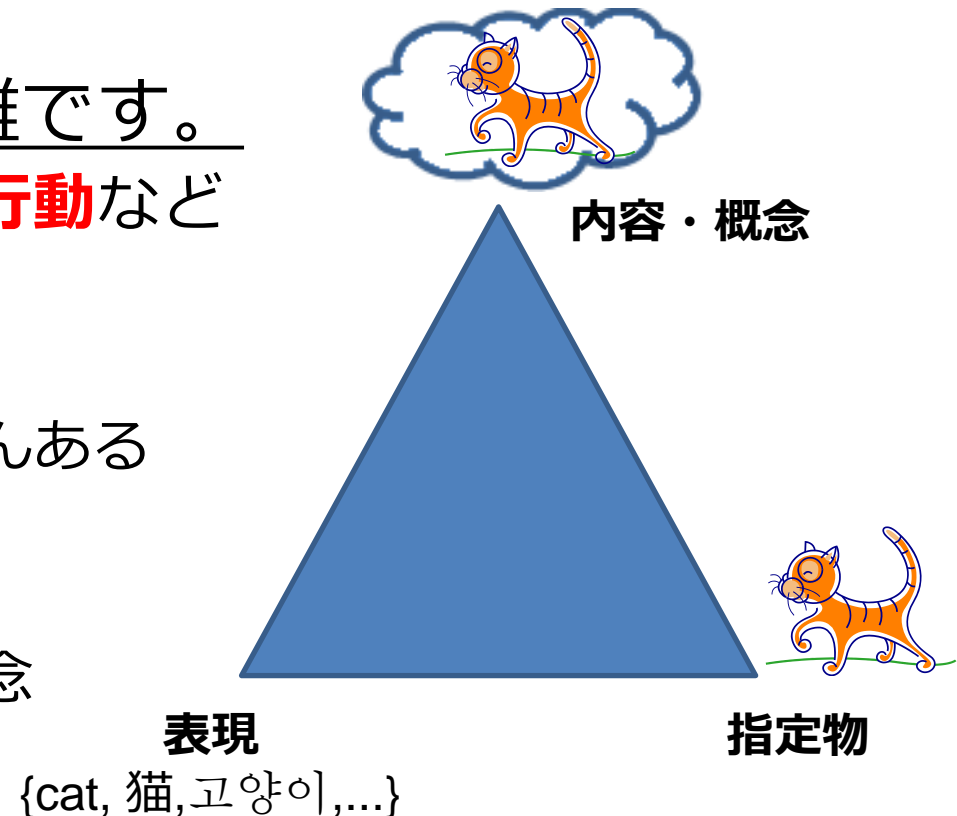
言語の記号 [sign] = 表現 [form] + 内容 [meaning]

- 表現 = 音声|文字|身振り|...

違います。三角より複雑です。

言語は人間の感情、**思考**、**行動**なども表す。

- 指定物のない概念はたくさんある
- 概念は**文化**による異なる
- 概念は**文脈**による異なる
- 複数の表現 <—> 複数の概念  
=> **不確定さ・曖昧さ**

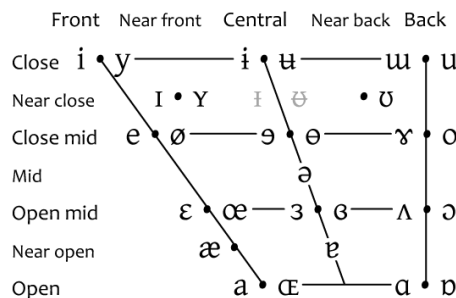


# 計算言語学 (NLP) の役割

非構造化  
データ



構造化  
データ



言語学

+



コンピューター  
サイエンス

UBIC技術の独自性の一つは多言語のNLPノウハウです。

# UBICの製品と言語学のレベル

レベル	研究対象	NLPツール	UBIC製品名	他社(eDiscovery)
音声学 [phonetics]	音素	音声認識エンジン	Voice Auditor	CJK非対応
形態論 [morphology]	品詞 (名詞、動詞など)	形態素分析エンジン	Predictive Coding (TM)	CJK非対応
統語論 [syntax]	語順、主語、 目的語など	チャンキング、 パーシング		
語彙論 [lexicology]	単語の関連、 類義語、対義語	辞書作り	Case Knowledge Dictionary	
意味論 [semantics]	概念、 オントロジー	概念検索、共起、 検索拡張、 固有表現抽出	Behavior Extractor	
語用論 [pragmatics]	文脈、常識的知識は 言語にどんな影響を 与えるか	対話システム、 代名詞の照応解決、 発話意図推定	Case Counselor	
文体論 [stylistics]	意味合い	感情抽出		

# 自然言語の類似点と相違点

特徴	English	中文	日本語	한국어	デンマーク語
音調言語	no	yes	<b>no</b>	no	<i>no</i>
語順	SVO	SVO (free)	<b>(S)OV</b>	(S)OV	SVO + VO
文字	ローマ字	漢字だけ	<b>漢字、ひらがな、カタカナ、ローマ字</b>	ハングル	ローマ字+3
大文字・小文字	有り	無い	<b>無い</b>	無い	有り
単数・複数	有り	無い	<b>無い</b>	無い	有り
定性 [definiteness]	有り	無い	<b>無い</b>	無い	有り
テンス（時制）	有り	無い	<b>有り</b>	有り	有り
代名詞	多い	少ない	<b>少ない</b>	少ない	多い
敬語、丁寧語	少ない	多少	<b>多い</b>	多い (7 levels)	少ない



# NLPは再現率とプレジションを最大化

	語幹	品詞
Late	late	RB
Wednesday	Wednesday	NNP
,	,	,
lawmakers	lawmaker	NNS
sent	<b>send</b>	VBD
a	a	DT
bill	bill	<b>NN</b>
to	to	TO
<b>&lt;person&gt;</b>		
<b>Gov.</b>	Gov.	NNP
<b>Gray</b>	Gray	NNP
<b>Davis</b>	Davis	NNP
<b>&lt;/person&gt;</b>		
that	that	IN
...	...	...

## 1. 活用語処理 [lemmatization]

例えば英語の送るという動詞の全ての活用形を使って検索すると**再現率**が上がります。

- [send] => {sent, sending, sends}

## 2. 形態素分析 [PoS tagging]

例えば英語では請求書と請求書を出すのは両方**bill**で書きます。その名詞と動詞を識別できれば**プレジション**が上がります。

## 3. 固有表現抽出 [NER]

固有表現と動詞の**共起**を分析することで抽出された人物、組織などの**行動**と**重要な関連**も把握できます。

# 中国語の解析事例：分析の流れ

**入力**：国务院总理李克强调研上海外高桥时提出，支持上海积极探索新机制。

As Premier **Li Keqiang** mentioned when he investigated [the idea of] the **Waigaoqiao** free-trade zone in **Shanghai**, he supports that **Shanghai** actively searches out new mechanisms.



<s id="0">		
国务院	ni	S-Ni
总理	n	0
李克强	nh	S-Nh
调研	v	0
上海	ns	B-Ns
外高桥	ns	E-Ns
时	n	0
提出	v	0
,	wp	0
支持	v	0
上海	ns	S-Ns
积极	a	0
探索	v	0
新	a	0
机制	n	0
。	wp	0
</s>		

(1) 文章の区切り => (2) セグメンテーション  
=> (3) 形態素分析 => (4) 固有表現抽出：

品詞： v = 動詞、a = 形容詞、n = 一般名詞

固有表現： ni = 組織名、nh = 氏名、ns = 地名

=> (5) 依存文法的な分析**出力**：

## まとめ：UBICの自然言語処理技術の独自性

- **非構造化データ**[unstructured data]に対応しております。
- 英語などの西洋語だけではなくて**アジア言語**にも対応しております。
- 対象言語に合わせた自然言語処理システムの**選定ノウハウ**
- 自社の言語学的なノウハウによってその**他の重要な言語**を迅速に対応可能

## 蓮子 和巳 博士 (理学)

1999年 東北大学大学院理学研究科物理学専攻  
博士課程修了。在籍中、スタンフォード大学の関連  
研究所で研究。

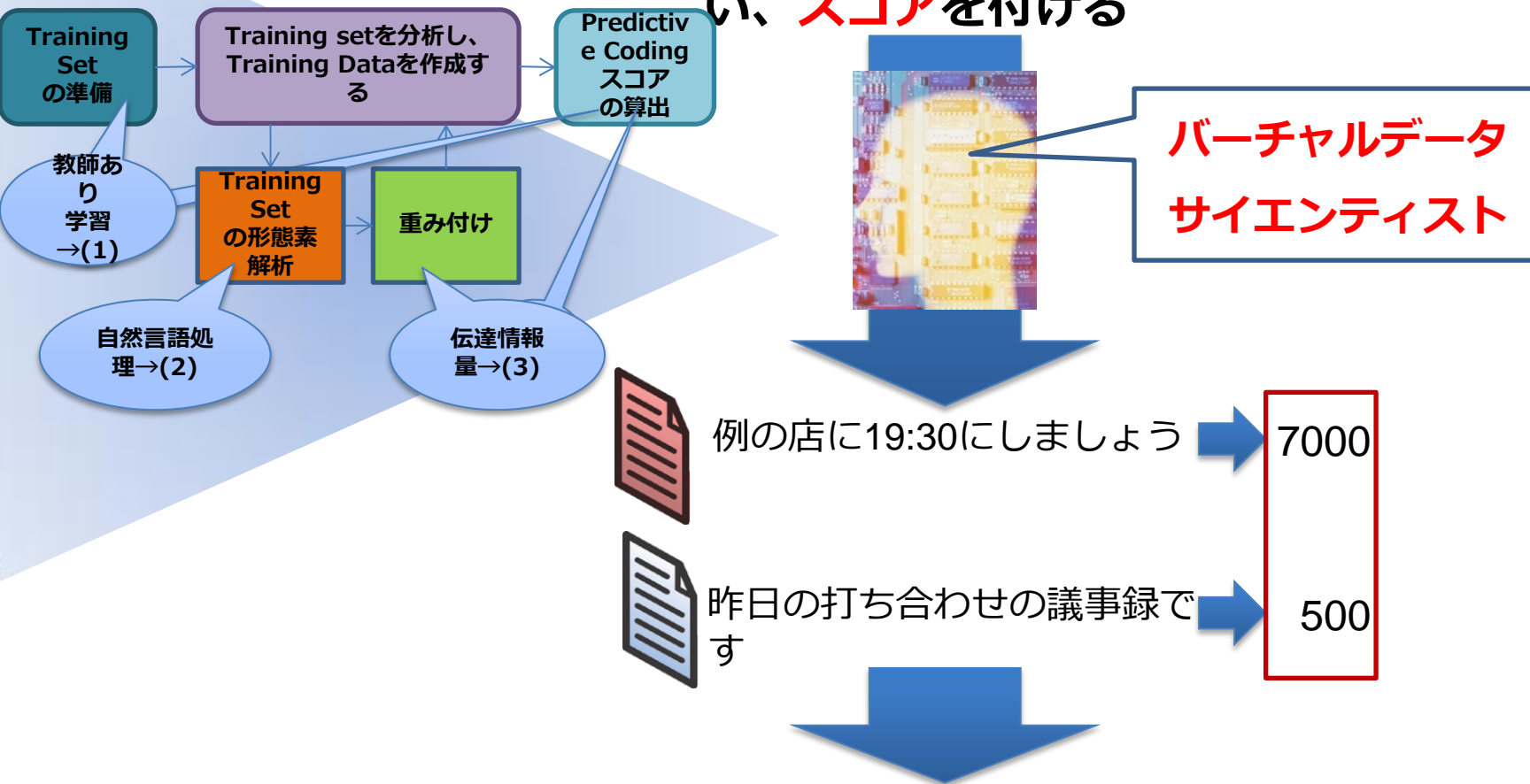
同年、東京大学素粒子物理国際研究センター 特別  
研究員。次世代粒子加速器 (LHC) 実験装置の設  
計、開発および新粒子探索の研究に従事。在籍中、  
欧州原子核研究機構 (CERN) で研究。

2002年 独立行政法人理化学研究所 研究員。  
ブルックヘブン国立研究所 (BNL)などで粒子加速  
器を用いた量子力学的現象 (スピン現象) の研究に  
従事。



1. 文書データに対する高度な統計学的  
分析処理
2. 多彩な分析経験を活かし、ニーズに  
応じた最適な手法を選択・採用

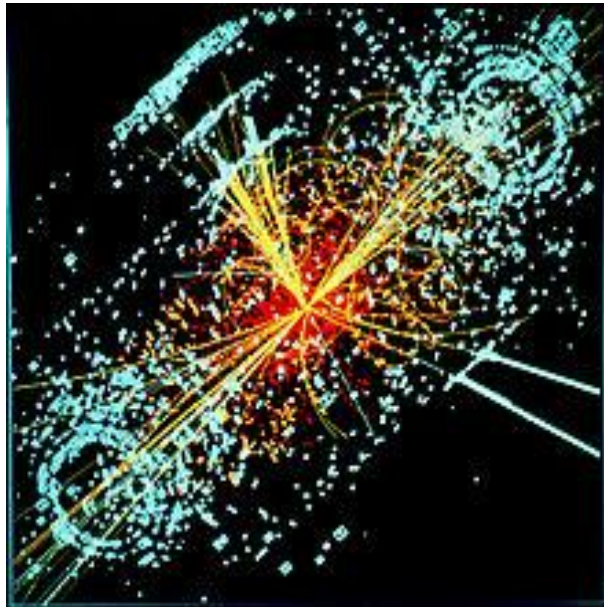
未レビューの文書に対し、  
**Predictive Coding**を行  
い、**スコア**を付ける



**スコアの高い文書から優先的にレビュー**




- 大規模な学術研究の分野では、早くから「**ビッグデータ解析**」が行われてきた。
- 素粒子物理学
  - 「ゴミ」だらけの膨大な「生データ」から「**ダイヤモンド**」の粒である**価値の高い事象**を探索する ⇒ **【元祖ビッグデータ解析】**
  - こうした極めて「稀な」現象をデータで確認するためには通常**の統計学的手法を超える高度な分析手法**が必要。



© 1997 CERN (License: [CC-BY-SA-4.0](https://creativecommons.org/licenses/by-sa/4.0/))

CERN（欧州原子核研究機構）のLHC（大型ハドロン衝突型加速器）では、100日で約4PB=4000TBの生データが記録される。LHC実験では、10兆回に1回しか生じない未知の粒子「**ヒッグス粒子**」を分析結果から確認 ⇒ 2013年ノーベル物理学賞。

技術		他社
重み付け	伝達情報量 TF/IDF	ベイズ法など
スコアリング	文書スコア 形態素自動選定	SVMなど
コーディング (自動仕分け)	Predictive Coding®	SVMなど
アプリケーション	時系列分析 関係分析など	限定的
特徴	高度・基礎的技術 汎用性が高い 様々な事例に対応	限定的 多くの事例に対応されていない

# UBICの独自技術① 伝達情報量

- 伝達情報量 (transinformation) : 「重要文書」と「文書中の形態素 (キーワード)」の関連を適切に表現する基本情報。

- ベイズ (Bayes) 法 vs. 伝達情報量
  - ベイズ法 = 「形態素」を確認したときの重要文書の確率に基づく  
⇒ 一部だけしか考慮していない
  - 伝達情報量 = 「形態素の有無」と「重要文書の是非」を全て考慮  
⇒ より網羅的、完全な形で

	特定の形態素を確認	特定の形態素を確認しない
重要文書である	正	誤
重要文書でない	誤	正

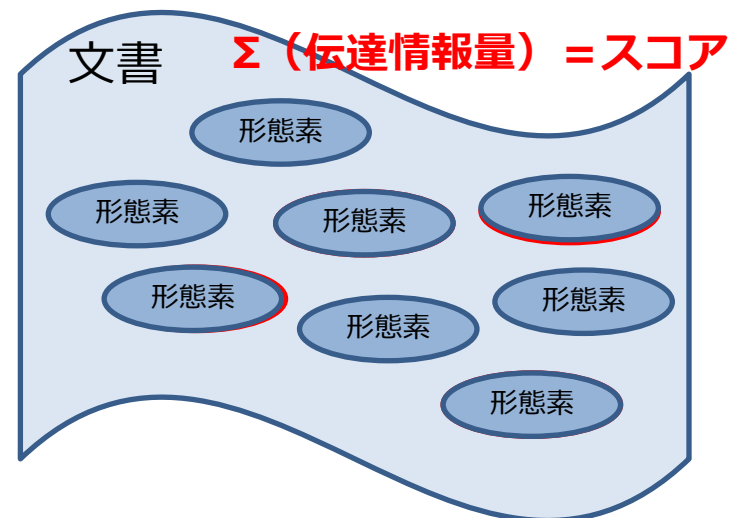
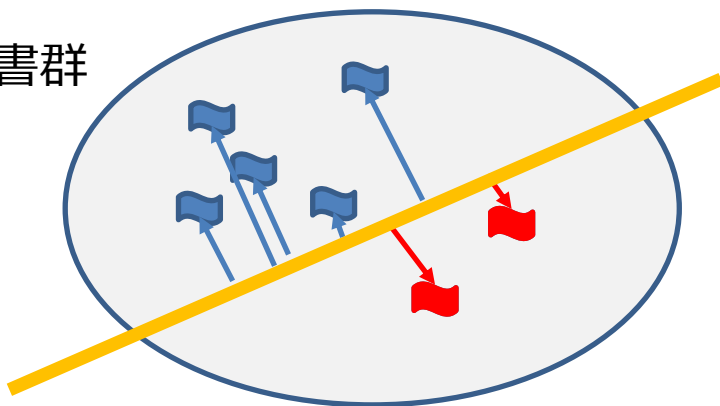
$$I(T; M) = H(T) + H(M) - H(T, M) = \sum_{m \in M} \sum_{t \in T} p(t, m) \log \frac{p(t, m)}{p(t)p(m)}$$

- H(M) 「形態素<M>を確認」する測定量の平均情報量 (エントロピー) ;
- H(T) 「ドキュメントが「タグ付けされた」<T>ものである」測定量の平均情報量 ;
- H(T, M) <T>と<M>の「結合エントロピー」  
(<T>と<M>の結合に関する情報量 (エントロピー) ) ;
- I(T; M) <M>が<T>に関して伝達する『伝達情報量』 ;

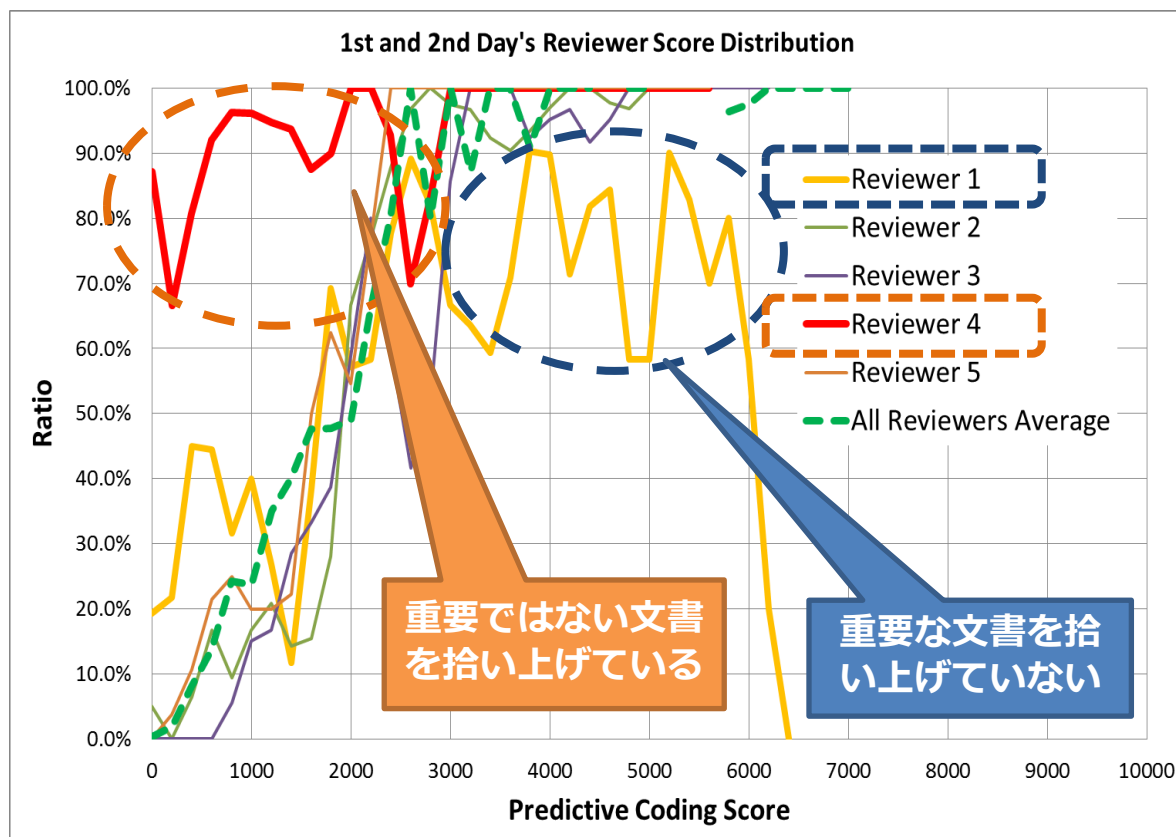
# UBICの独自技術② スコアリング

- **UBIC文書スコア**：形態素の「**重み**」 = 「**伝達情報量**」を集約。
- SVM（サポートベクターマシン） vs. UBIC文書スコア
  - SVMは、文書群を2群に分別することに特化した手法。
    - ⇒ 2群を分別する面からの距離による意味付け・解釈が困難。
    - ⇒ 計算効率も悪い。
  - UBIC文書スコアは、**必要十分な形態素を自動選択（動的選択）**。
    - ⇒ 「形態素」の関係から**意味付けも容易**。
    - ⇒ 計算効率が良い。**Knowledgeベース**で更に高度化。

文書群



- Predictive Coding®  
 自社製eディスカバリソフトの研究開発の成果。日本で唯一
- Predictive Coding® は人（レビューアー）よりも、  
より安定して確実なパフォーマンスを示す
- 基礎的 & 汎用的  
 ⇒ 多彩な応用製品





# まとめ

- 高度な分析知識と経験に基づく研究開発  
⇒ **ユニークで基礎的な技術**：  
**伝達情報量, Predictive Coding®**
  - 多彩な分析経験に基づく、**解法の「引出」**が多い  
⇒ **ビジネス課題の解決に最も適した手法を  
適宜選択・採用**
- ⇒ **UBIC研究開発の大きな強み**

## SEEDS : 課題解決のための手法提供



非構造化データを構造化データにする手法



構造化された、ノイズだらけのデータから法則性を取り出す手法



分析対象（不正行為、人間関係）の定性分析とモデル化手法

# 人間行動抽出機能

人のどのような行動に着目すべきかを  
Virtual Data Scientistが示唆

Emails



## List of Behavior

	Keyword1	Keyword2		
1	情報 Get	入手	Information	
2	情報 Gather	収集	Information	
3	情報 Provide	提供	Information	
4	価格 Negotiate	交渉	Price	
5	打ち合わせ Have	実施	Meeting	
6	サンプル Deliver	納入	Sample	
7	価格 Reply	回答	Price	
8	技術Technology	交流	Exchange	
9	新規 New	受注	Order	
10	内部 Discuss	検討	Internally	

**Extract Verb +  
Noun**



行動情報科学というコンセプトを考えた。行動を抽出できる技術を発明しよう（武田）



共起のアルゴリズムが使えるそう（蓮子）

（※共起・・・ある単語が文中に出現した時、ある特定の単語が頻繁に出現すること）



品詞を判定すれば、行動だけを抽出できる（Jakob）



テストしてみよう（武田）



Jakobさんの品詞特定プロトタイプを使い、共起のアルゴリズムを組んでみた（蓮子）



これはまさに行動抽出。人物相関図と組み合わせて使うと、人間関係の概要が分かる。ところで英語についてはどうしよう（武田）



英語も大丈夫（Jakob）



犯罪学を応用したカルテル構造分析との組み合わせを今後の研究課題にしよう（武田）



 **U B I C**