データ解析

日時：木曜2限・3限

場所：A33

担当：瀧川 一学 (大規模知識処理)

工学部 情報エレクトロニクス学科 情報理工学コース

大規模知識処理研究室

http://art.ist.hokudai.ac.jp/

みなと しんいち
湊　真一 教授

1965 石川県生。京都大学出身
1990 NTT研究所 研究員
1997 スタンフォード大 客員研究員 (1年間)
2004 北海道大学 助教授
2010 北海道大学 教授

たきがわ いちがく
瀧川 一学 准教授

1977 香川県生。北海道大学出身
2005 京都大学 助教
2010 ボストン大 客員研究員 (3ヶ月)
2012 北海道大学 特任助教
2014 北海道大学 准教授

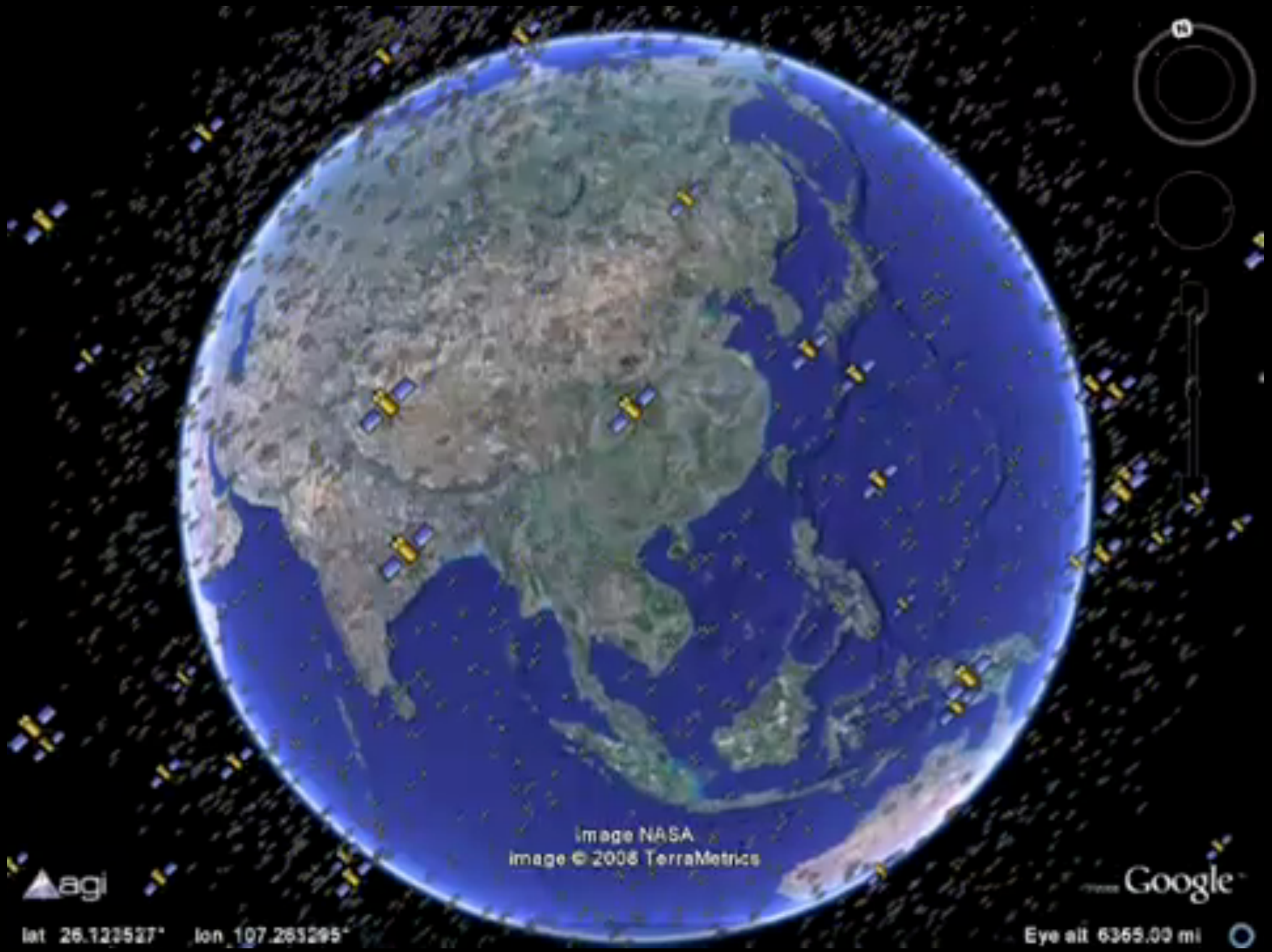
ScienceとEngineeringをつなぐ “Art” を求めて

データ解析

全てがデータになる社会へようこそ。

そして、我々はそれを生き抜かなければならない。

人工衛星はどれくらいあるのか？



スマートグラス/ウェアラブルデバイス

東芝

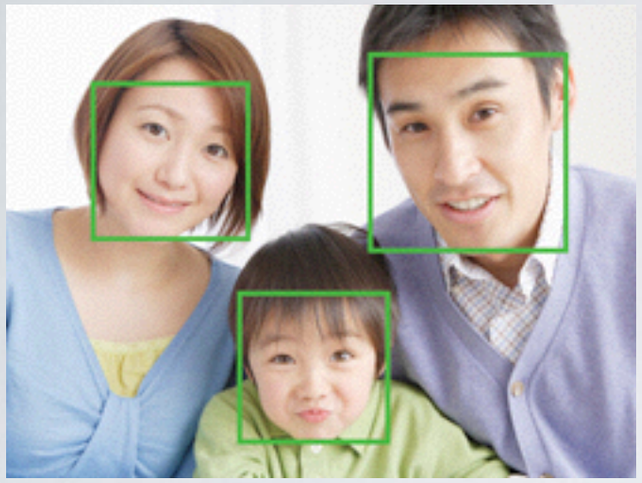
NTT docomo

EPSON

Samsung

JiNS

SONY



顔検出

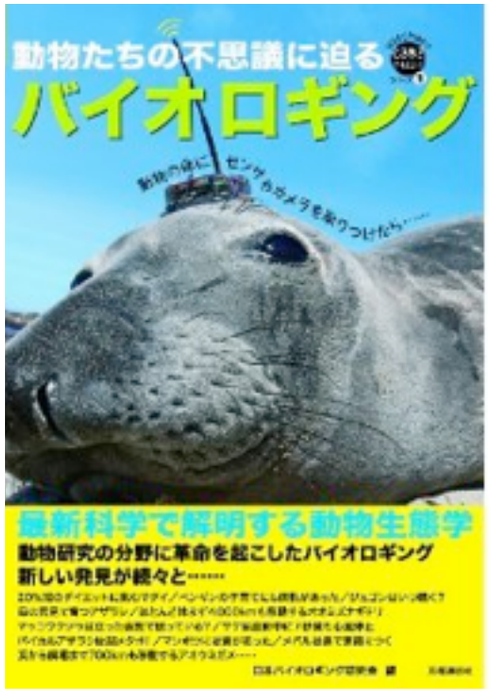
植木鉢ロボ

手書き文字認識

音声認識

「全録」の時代？

ライフログ



バイオロギング

- テレビ放送の"全録"
- 防犯カメラ映像の利活用
- 子供が言語を覚えるまでに聞いた全ての音の記録

データ大氾濫！

2011

1 NEW DEFINITION IS ADDED ON urban

1,600+ READS ON Scribd.

13,000+ HOURS MUSIC STREAMING ON PANDORA

12,000+ NEW ADS POSTED ON craigslist

370,000+ MINUTES VOICE CALLS ON skype

98,000+ TWEETS

20,000+ NEW POSTS ON tumblr.

THE LARGEST SOCIAL READING PUBLISHING COMPANY!!

320+ NEW twitter ACCOUNTS

13,000+ iPhone APPLICATIONS DOWNLOADED

100+ NEW Linked in ACCOUNTS

1 associatedcontent NEW ARTICLE IS PUBLISHED

THE WORLD'S LARGEST COMMUNITY CREATED CONTENT!!

QUESTIONS ASKED ON THE INTERNET...

100+ Answers.com

40+ YAHOO! ANSWERS

IN 60 SECONDS...

6,600+ NEW PICTURES ARE UPLOADED ON flickr

600+ NEW VIDEOS

50+ WordPress DOWNLOADS

25+ HOURS TOTAL DURATION

70+ DOMAINS REGISTERED

695,000+ facebook STATUS UPDATES

125+ PLUGIN DOWNLOADS

60+ NEW BLOGS

168 MILLION EMAILS ARE SENT

694,445 SEARCH QUERIES

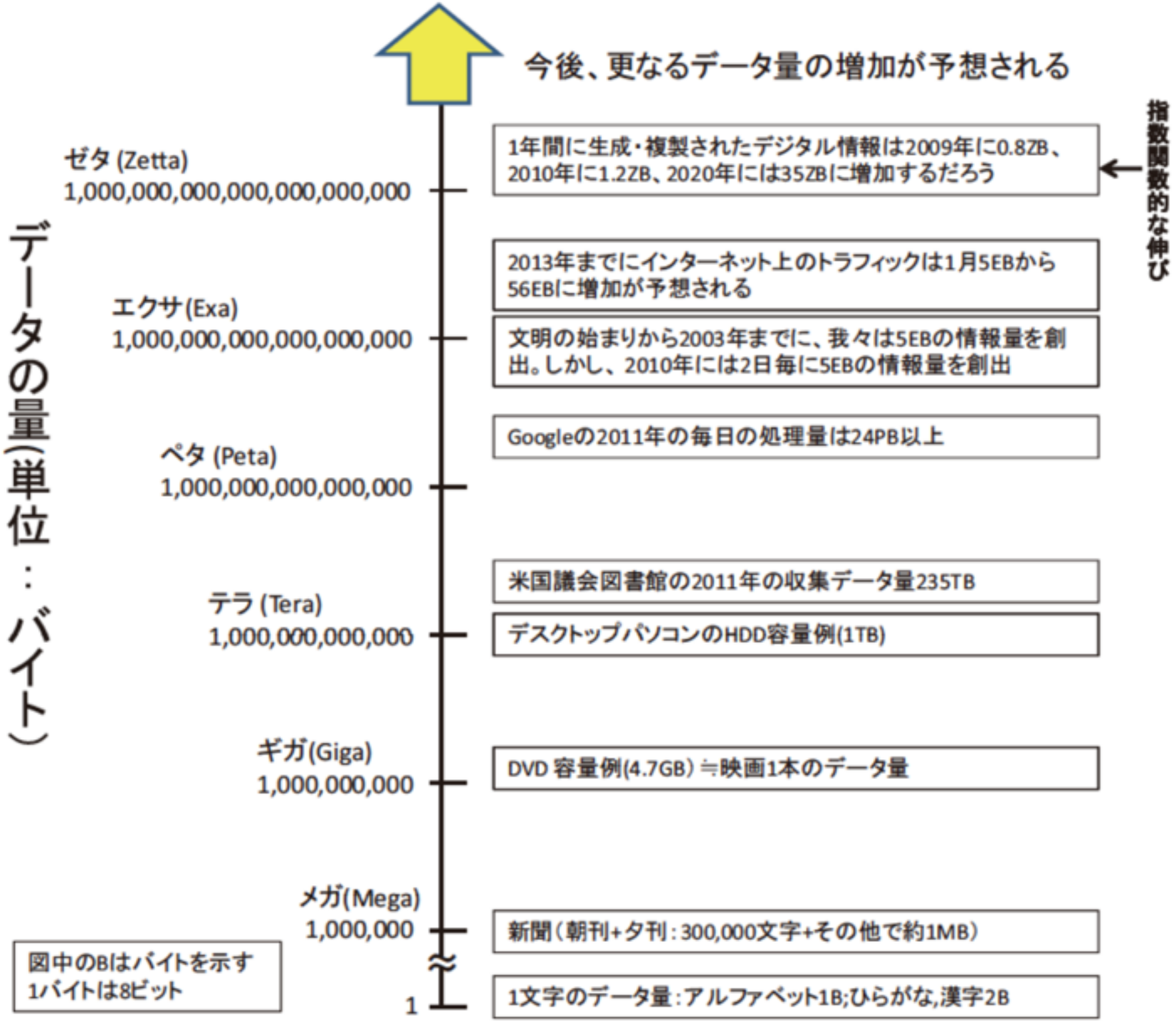
1,700+ Firefox DOWNLOADS

1,500+ BLOG POSTS

79,364 WALL POSTS

510,040 COMMENTS

GO-Globe.com



出典: 国立科学技術政策研究所 科学技術動向研究(2012)「米国政府のビッグデータへの取り組み」

溢れかえるデータをどうする！？

Science 2011 Feb

ERSPECTIVE

More Is Less: Signal Processing and the Data Deluge

Richard G. Baraniuk

The data deluge is changing the operating environment of many sensing syste
to data-rich—so data-rich that we are in jeopardy of being overwhelmed. Ma
exploiting the data deluge require a reinvention of sensor system design and
the potential pay-offs are huge, as the resulting sensor systems will
and powerful new tools for scientific discover

Nature 2010 Oct

TECHNOLOGY

Science in the digital age

The goals of science have not changed since the early days of the Lindau meeting, yet the way they are pursued has.

BY NED STAFFORD

The year was 1947—no high-powered personal computers and no internet for seemingly magical keyword searches of massive databanks. So when Oliver Smithies, a 22-year-old science student at Oxford University, was assigned an essay
exclusion principle
scient
doing
He went to the

page after page with his eyes. When Smithies, who 60 years later won the 2007 Nobel Prize in Physiology or Medicine, finally finished researching his essay, he had the pleasure of writing it by hand.
single change between
is available.
image had

goals of science remain unaltered in this new world, but the paths that scientists traverse to reach them have changed, triggering a further cascade of new developments.
"This
been imagined twenty
, a science historian
tory of Natural Sci-
of Sciences, in Bei-
terminal, scientists
most all important
ifferent languages
ntries and fields."
o bypass tradi-
uickly and easily

The Economist 2010 Feb

The Economist

Obama the warrior
Misgoverning Argentina
The economic shift from West to East
Genetically modified crops blossom
The right to eat cats and dogs

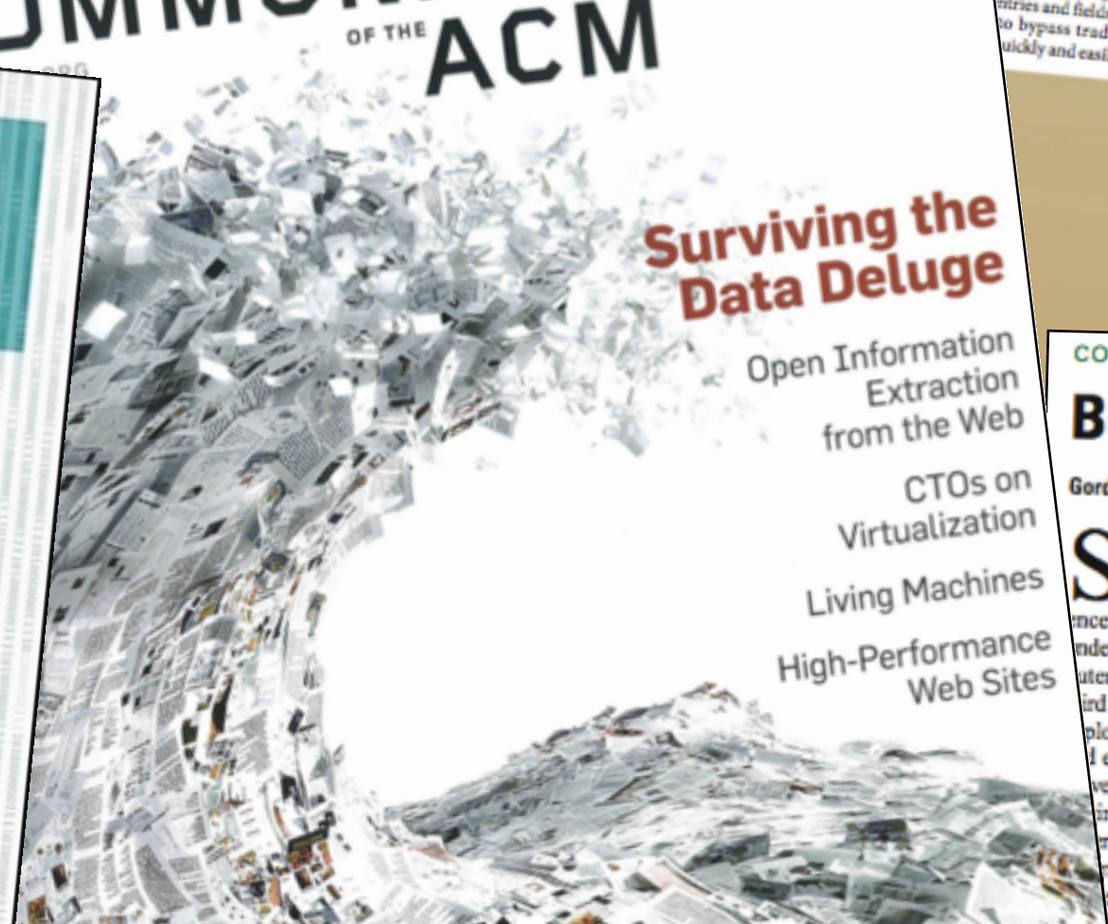
The data deluge

AND HOW TO HANDLE IT: A 14-PAGE SPECIAL REPORT

Commun ACM 2008 Dec

COMMUNICATIONS OF THE ACM

12/08 VOL.51 NO.12



Surviving the Data Deluge

Open Information Extraction from the Web

CTOs on Virtualization

Living Machines

High-Performance Web Sites

IEEE Spectrum 2011 Feb

The Coming Data Deluge

The new model is for the data
generated
and for the
computers
pipeline
ies or

Science 2009 Mar

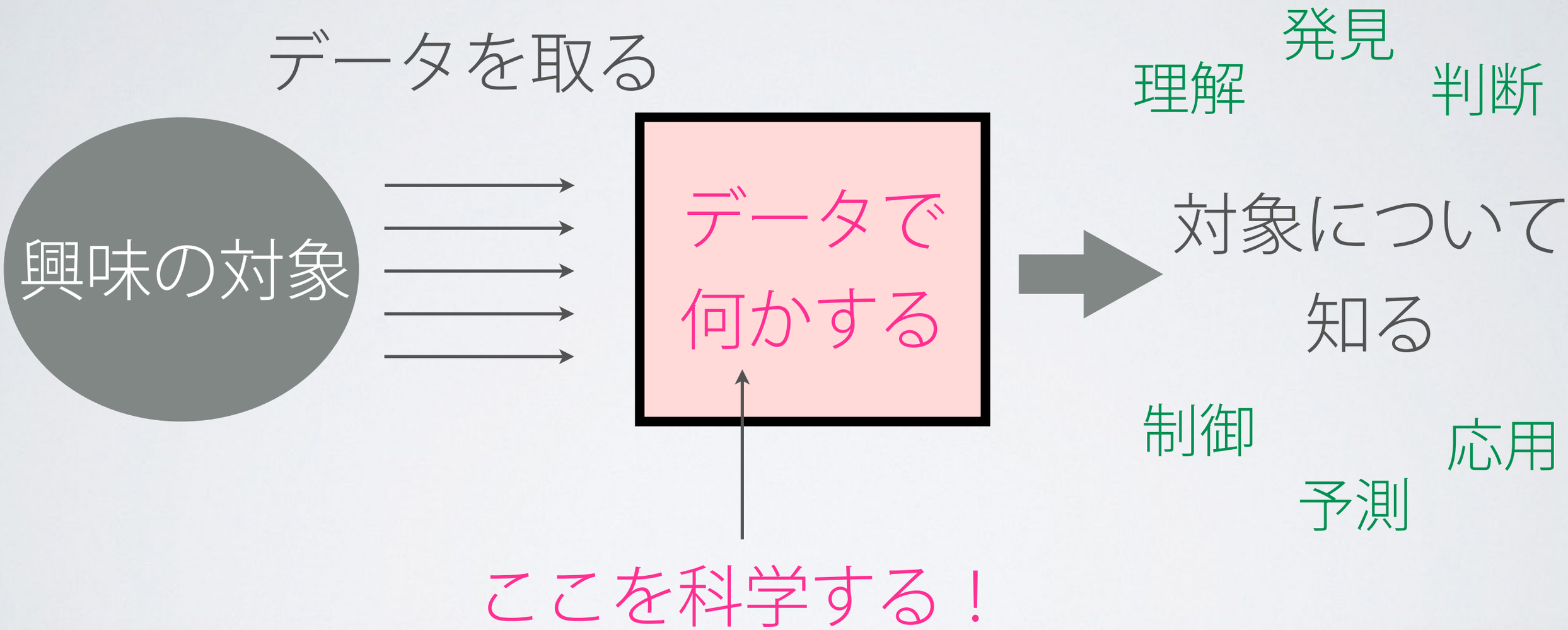
COMPUTER SCIENCE

Beyond the Data Deluge

Gordon Bell,[1] Tony Hey,[1] Alex Szalay[2]

Since at least Newton's laws of motion in the 17th century, scientists have recognized experimental and theoretical science as the basic research paradigms for understanding nature. In recent decades, computer simulations have become an essential third paradigm: a standard tool for scientists to explore domains that are inaccessible to theory and experiment, such as the evolution of the universe, car passenger crash testing, and predicting climate change. As simulations and experiments yield ever more data, a fourth paradigm is emerging, consisting of the techniques and technologies needed to perform

考えてもらいたい点



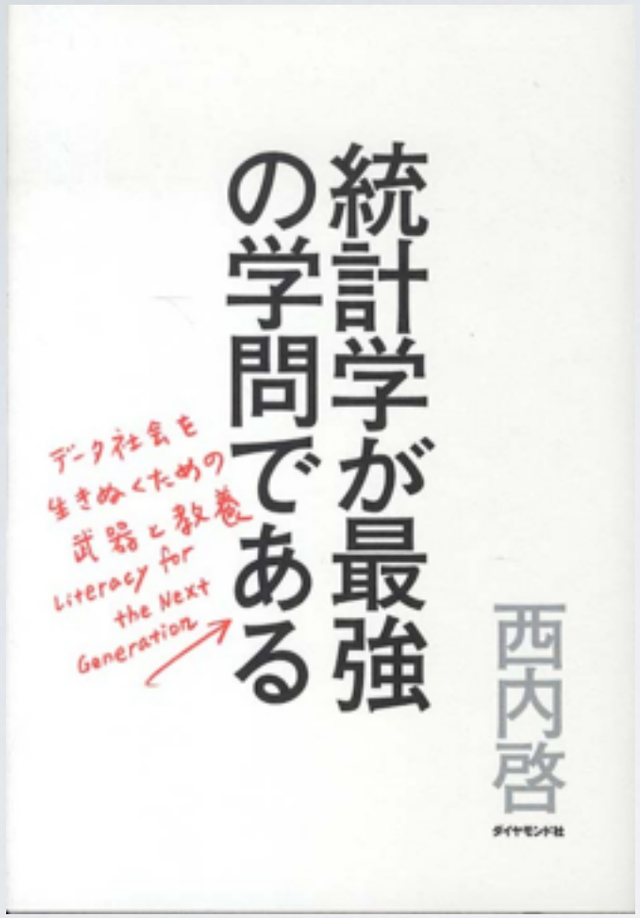
20～30年後、多様なデータを使って
何ができるようになっているだろうか？

機械による
同時通訳

自動車の
真の自動運転

個別化医療・創薬

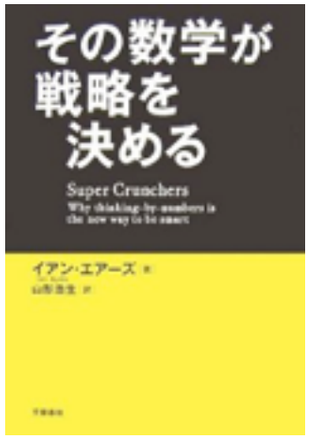
巷では空前の統計学ブームらしい（ビジネス業界）

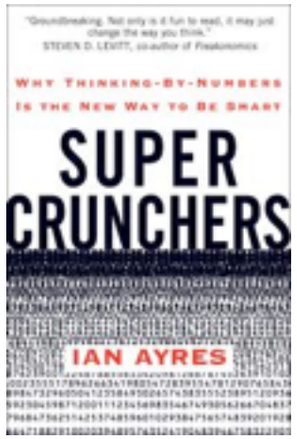


2013/1/25発売

データ分析

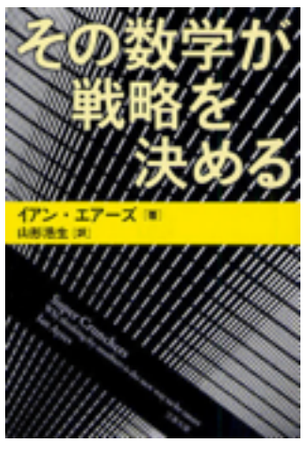
データの使い方(読み解き方)が大事

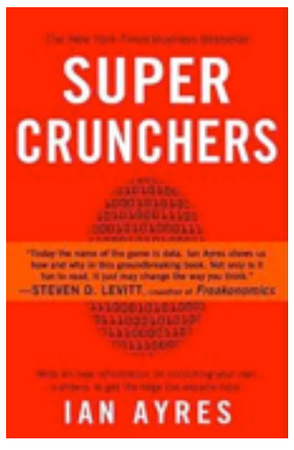




兆単位のデータ計算が専門家にとってかわる——

政治家も、評論家も、医者も、
裁判官も、映画プロデューサーも
みんな真っ青！

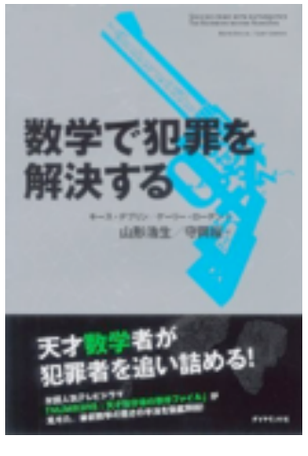




—兆単位のデータが可能にした！

一般書の帯にも
あやしげな
キャッチコピーが…

あなたに最適な結婚相手も
映画のヒットも計算できます。



クローズアップ現代

毎週月ー木曜放送 総合 午後7時30分ー午後7時56分
[再放送] 毎週火ー金曜 総合 午前0時10分ー午前0時36分(月ー木曜深夜)

数字のカラクリ・データの真実　～統計学ブームのヒミツ～

出演者　**竹内 薫** さん
（サイエンス作家）
岩崎 学 さん
（成蹊大学教授・統計学者）

ジャンル　経済　話題・ブーム

関連タグ　IT　舞台裏　企業　ブーム

今、統計学がブームを巻き起こしている。出版界では入門書が5か月で26万部の大ヒット。都心の書店では統計学コーナーまで設置され、公開講座にはビジネスマンを中心とした受講生が殺到。そして、統計学を使いこなす「データサイエンティスト」と呼ばれる専門職は「最もセクシーな（魅力的な）職業」だとして、多くの企業から引く手あまたの状況だ。一体、人々は何を統計学に求めるのか。それは、あふれる情報の海から確かな指針を探し出す力。ビッグデータ時代と言われるが、情報には偏りやノイズがつきもので、そのままではただの「ゴミ」でしかない。データ分析から知られざる事実を解明し、未来を予測するには統計学のスキルが不可欠なのだという。さらにビジネスだけでなく、多くの人にとっても統計学的な考え方「統計リテラシー」が必要だという。人々が身につけたいと願う統計学の威力と、そこで必要とされる考え方のエッセンスを探る。

過去の関連する放送回

2012年6月25日（月）放送
危険性増す脱法ハーブ

2012年7月2日（月）放送
アイデアが世界を変える

2012年11月15日（木）放送
"おしゃべり"で老化を防げ！

2013年1月17日（木）放送
弱く、美しき者たちへ

2013年6月4日（火）放送
あなたの家が生まれ変わる

データを読み解くとはなんなのか？

(データサイエンスと統計学と多変量解析とパターン認識と機械学習と人工知能とデータマイニング)

手元にあるデータから「手元にまだないデータについて」何かを語ること！

その道具と仕組みを学びます

仮説検証：科学的事実とは何か？

医薬品の有効性

放射線の人体への影響

飛行機や建築物の安全性

因果関係を100%保証できるのか？

特定の物質の発ガン性

災害対策の十分性

食生活の健康への影響

データ分析で成功している企業

楽天

YAHOO!

Google

mixi

DeNA

twitter

GREE

Microsoft

KOMATSU

amazon

NRI 未来創発 野村総合研究所

accenture

facebook

Goldman Sachs

McKinsey&Company

教科書

この授業で学ぶこと

多変量データの統計解析の基礎

1. 多変量解析法とは
2. 統計的方法の基礎知識
3. 線形代数のまとめ
4. 単回帰分析
5. 重回帰分析
6. 数量化I類
7. 判別分析
8. 数量化2類
9. 主成分分析
10. 数量化3類
11. 多次元尺度構成法
12. クラスター分析
13. その他の方法

この授業で学ぶこと(もっと大雑把に)

線形代数を使って以下を行う方法を学ぶ

1. **回帰分析**
2. **判別分析**
3. **主成分分析**
4. **数量化**

1.2 重回帰分析とは

表 1.3 は東京のある駅の徒歩圏内の中古マンションに関するデータである.

表 1.3　中古マンションのデータ

サンプル No.	広さ x_1 (m^2)	築年数 x_2 (年数)	価格 y (千万円)
1	51	16	3.0
2	38	4	3.2
3	57	16	3.3
4	51	11	3.9
5	53	4	4.4
6	77	22	4.5
7	63	5	4.5
8	69	5	5.4
9	72	2	5.4
10	73	1	6.0

このデータに基づいて知りたいことは次の通りである.

(1) 価格は広さと築年数とによって予測できるだろうか.

(2) 予測できるとすればその精度はどのくらいか.

(3) 同じ地区で $x_1 = 70$, $x_2 = 10$, $y = 5.8$ を提示された. 価格は妥当か.

線形代数?

表のマス目にかかれた数値を足したり掛けたりするアレ

多変量解析の主題

回帰分析 ⇒ 4,5章 ------ **数量化I類** ⇒ 6章

判別分析 ⇒ 7章 ------ **数量化II類** ⇒ 8章

主成分分析 ⇒ 9章 ------ **数量化III類** ⇒ 10章

多次元尺度構成法(12章)、クラスター分析(11章)、
因子分析・パス分析・グラフィカルモデル(13章)、
正準相関分析(13章)、多段層別分析(13章)

線形代数の技術

射影行列
線形写像の像と核
直交直和分解
(線形代数学の基本定理)

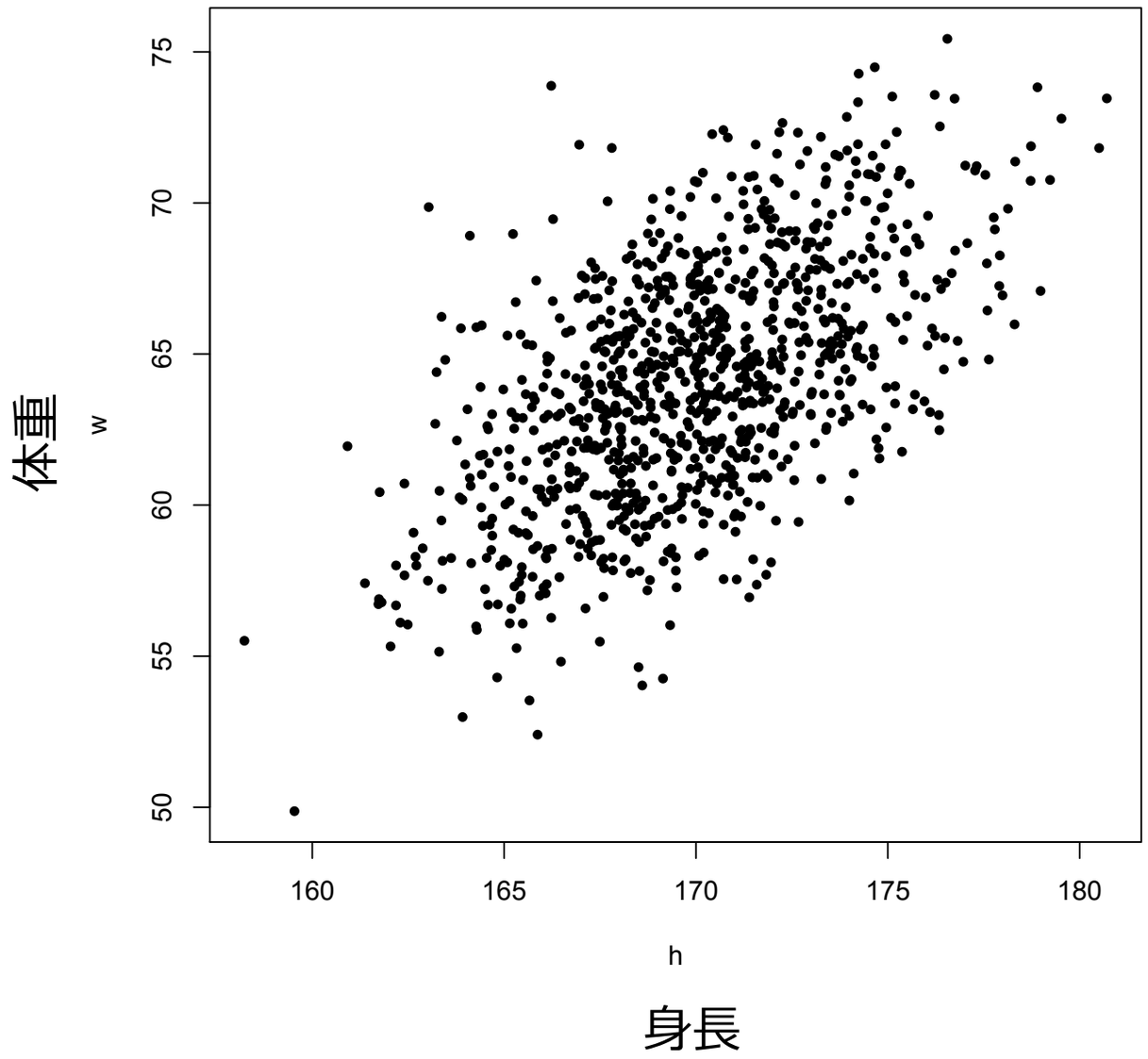
２次形式
基底変換
固有値・固有ベクトル
直交行列による対角化

多変量正規分布

標準化、マハラノビス距離

多変量のデータのイメージ (2変量の例)

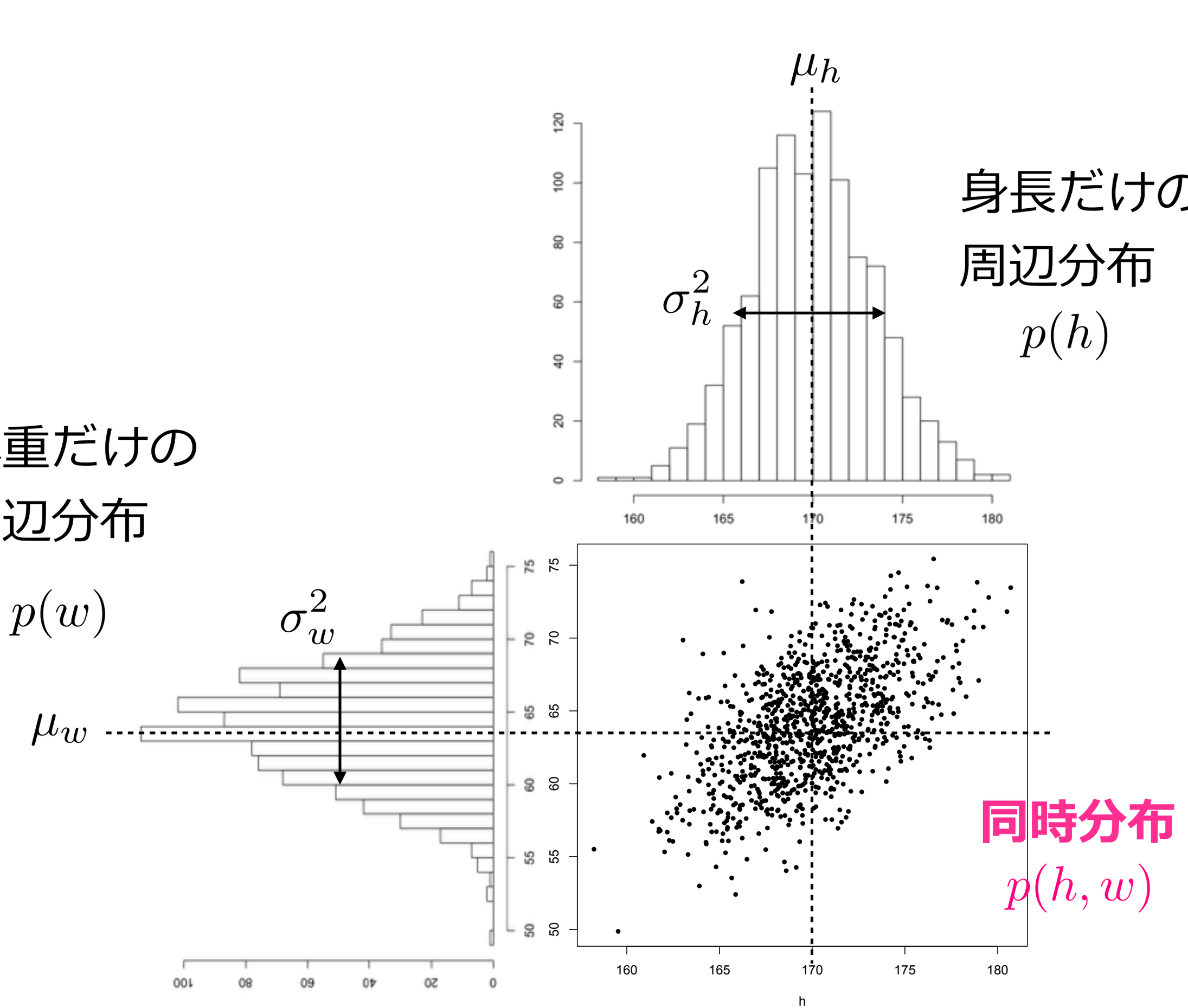
散布図



表形式データ

	身長 h	体重
1	174.0	64.1
2	169.6	65.4
3	168.4	67.4
4	171.7	63.4
5	172.1	72.3
6	167.0	63.4
7	167.0	62.5
:	:	:
999	172.7	64.9
1000	167.3	61.97

体重だけの
周辺分布



線形代数?

表のマス目にかかれた数値を足したり掛けたりするアレ

1.2 重回帰分析とは

表 1.3 は東京のある駅の徒歩圏内の中古マンションに関するデータである.

表 1.3　中古マンションのデータ

サンプル No.	広さ x_1 (m^2)	築年数 x_2 (年数)	価格 y (千万円)
1	51	16	3.0
2	38	4	3.2
3	57	16	3.3
4	51	11	3.9
5	53	4	4.4
6	77	22	4.5
7	63	5	4.5
8	69	5	5.4
9	72	2	5.4
10	73	1	6.0

このデータに基づいて知りたいことは次の通りである.

(1) 価格は広さと築年数とによって予測できるだろうか.

(2) 予測できるとすればその精度はどのくらいか.

(3) 同じ地区で $x_1 = 70$, $x_2 = 10$, $y = 5.8$ を提示された. 価格は妥当か.

表 1.3 のデータを重回帰分析で解析することにより次のことがわかる.

(1) 回帰式は次のように推定される.

$$\hat{y} = 1.02 + 0.0668x_1 - 0.0808x_2$$

築年数が同じなら広さが $1\mathrm{m}^2$ 増加するとき価格は 66.8 万円高くなり，広さが同じなら築年数が 1 年経つとき価格は 80.8 万円減少する.

(2) 自由度調整済寄与率は 0.933 であり，回帰式の精度は十分高い.

(3) $x_1 = 70$，$x_2 = 10$ を代入すると $\hat{y} = 4.89$ となる．また，信頼率 95%の予測区間を求めると $(4.21, 5.57)$ を得る．したがって，この物件が 5.8（千万円）なら，相場より高い.

表 1.3　中古マンションのデータ

サンプル No.	広さ x_1 (m^2)	築年数 x_2 (年数)	価格 y (千万円)
1	51	16	3.0
2	38	4	3.2
3	57	16	3.3
4	51	11	3.9
5	53	4	4.4
6	77	22	4.5
7	63	5	4.5
8	69	5	5.4
9	72	2	5.4
10	73	1	6.0

$$\mathbf{X} = \begin{pmatrix} 1 & 51 & 16 \\ 1 & 38 & 4 \\ 1 & 57 & 16 \\ 1 & 51 & 11 \\ 1 & 53 & 4 \\ 1 & 77 & 22 \\ 1 & 63 & 5 \\ 1 & 69 & 5 \\ 1 & 72 & 2 \\ 1 & 74 & 1 \end{pmatrix} \quad \mathbf{y} = \begin{pmatrix} 3.0 \\ 3.2 \\ 3.3 \\ 3.9 \\ 4.4 \\ 4.5 \\ 4.5 \\ 5.4 \\ 5.4 \\ 6.0 \end{pmatrix}$$

$$\mathbf{X}' = \begin{pmatrix} 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \\ 51 & 38 & 57 & 51 & 53 & 77 & 63 & 69 & 72 & 74 \\ 16 & 4 & 16 & 11 & 4 & 22 & 5 & 5 & 2 & 1 \end{pmatrix}$$

$$\hat{y} = 1.02 + 0.0668x_1 - 0.0808x_2$$

$$\begin{pmatrix} 1.02 \\ 0.0668 \\ -0.0808 \end{pmatrix} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}$$

!!

表のマス目にかかれた数値を足したり掛けたりするアレ

多変量解析の主題

回帰分析 ⇒ 4,5章 ------ **数量化I類** ⇒ 6章

判別分析 ⇒ 7章 ------ **数量化II類** ⇒ 8章

主成分分析 ⇒ 9章 ------ **数量化III類** ⇒ 10章

多次元尺度構成法(12章)、クラスター分析(11章)、
因子分析・パス分析・グラフィカルモデル(13章)、
正準相関分析(13章)、多段層別分析(13章)

線形代数の技術

射影行列
線形写像の像と核
直交直和分解
(線形代数学の基本定理)

２次形式
基底変換
固有値・固有ベクトル
直交行列による対角化

多変量正規分布

標準化、マハラノビス距離