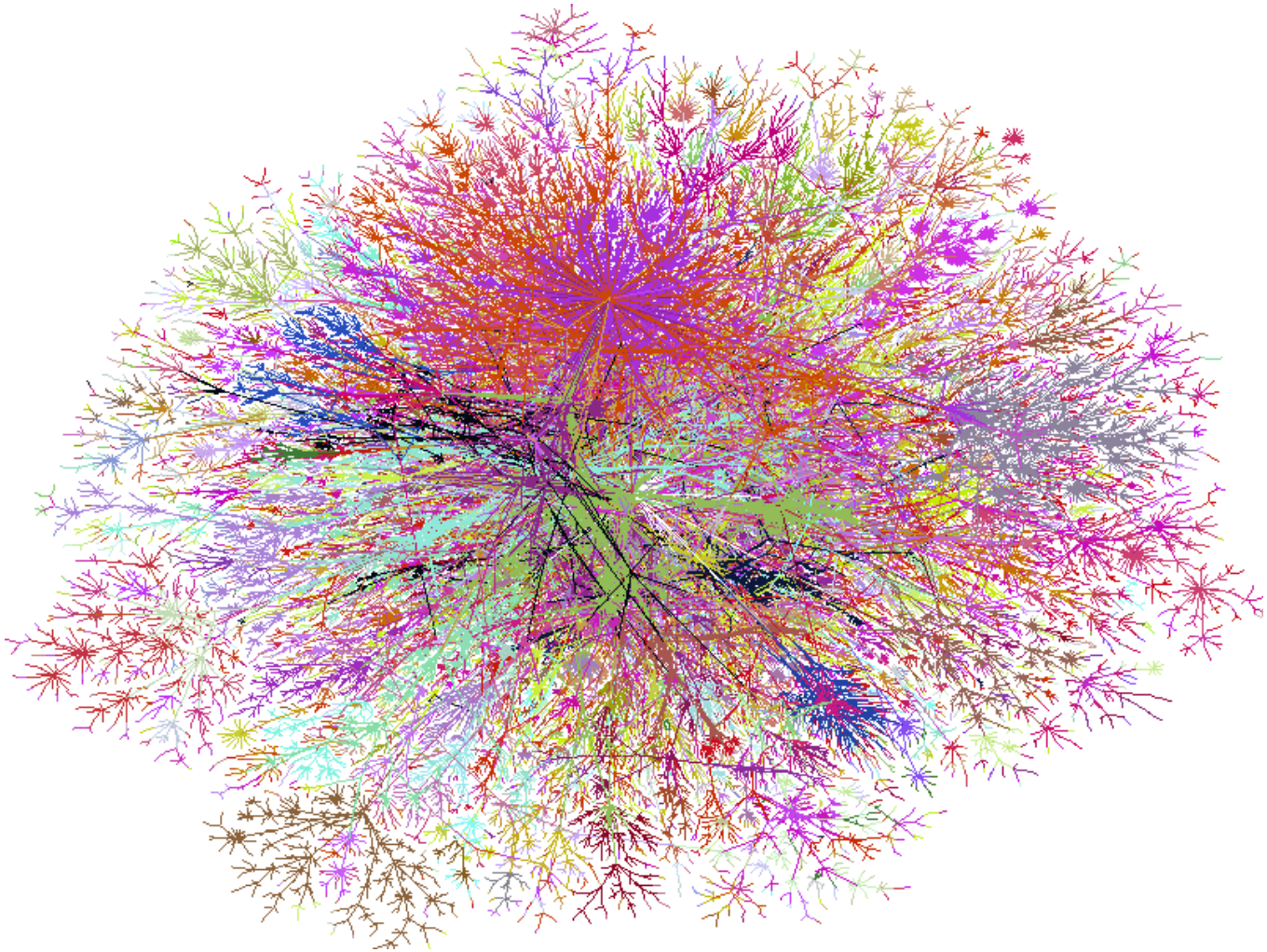
インターネット計測とデータ解析 第 1 回

長 健二朗

2014 年 4 月 7 日

はじめに

世界中にはり巡らされたインターネットの全体像とは？



lumeta internet mapping http://www.lumeta.com

http://www.cheswick.com/ches/map/

はじめに (つづき)

世界中にはり巡らされたインターネットの全体像とは？

- 誰も把握できていない
- でも、誰もが知りたい

本授業のテーマ

- いろいろな切口からインターネットとデータ解析を考える
  - 容易に計測できないものをどう計るか
  - 大量データからいかに情報を抽出する

このようなアプローチの仕方は今後の情報社会でますます重要となってくる

- ネットワーク系の計測およびアプリケーションでのデータ解析

インターネット計測とデータ解析

インターネット計測とデータ解析
(Internet measurement and data analysis)

- 担当教員: 長 健二朗 <kjc@sfc.keio.ac.jp>
- TA: 空閑 洋平 <sora@sfc.wide.ad.jp>
- SA: TBA
- URL: http://web.sfc.keio.ac.jp/~kjc/classes/sfc2014s-measurement/
- 授業サポートメール (教員、TA、SA に届く): <imda2014s@sfc.wide.ad.jp>
- 教材・参考文献: 講義資料をオンライン配布
- プログラミングによるデータ解析演習を重視
- 提出課題・成績評価の方法: 2 回の課題提出 (20%づつ) と学期末レポート提出 (60%)

授業のねらい

(学生に身につけて欲しいこと)

- データのばらつきについて理解し、データ処理とグラフ化を習得
  - 卒論や他のレポートを書くときに役立つはず
- 大量データを処理するプログラミング技術を習得
  - 既成のパッケージソフトウェア依存では限界
- 統計データを疑う力をつける
  - 作為的な統計データや情報操作の氾濫
  - (オンラインプライバシーに関するリテラシー向上)

- プログラミングとデータ処理体験
  - 手と頭を使って試行錯誤をすることで身につく肌感覚
  - 書籍を読む、講義を聴くだけでは得られない直観

最近"Big Data"が騒がれている

the WHITE HOUSE PRESIDENT BARACK OBAMA

BLOG | PHOTOS & VIDEO | BRIEFING ROOM | ISSUES | the ADMINISTRATI

Home • The Administration • Office of Science and Technology Policy

Office of Science and Technology Policy

About OSTP | OSTP Blog | Pressroom | Divisions | R&D Budgets | Resource Library | NST

Big Data is a Big Deal

Subscribe

Posted by Tom Kalil on March 29, 2012 at 09:23 AM EDT

E-Mail | Tweet

[Editor's Note: Watch
http://live.science360

Today, the Obama Admi
our ability to extract kno
promises to help accele
transform teaching and l

To launch the initiative,
commitments that, toget
glean discoveries from h
address the challenges

We also want to challen
most of the opportunitie
President calls an "all ha

Some companies are al
research. Universities a
generation of "data scie
bono data collection, an
forum to highlight new p

Tom Kalil is Deputy Di

The Economist

Log in | Register | Subscribe

World politics | Business & finance | Economics | Science & technology | Culture

Current issue | Previous issues | Special reports | Politics this week | Business this

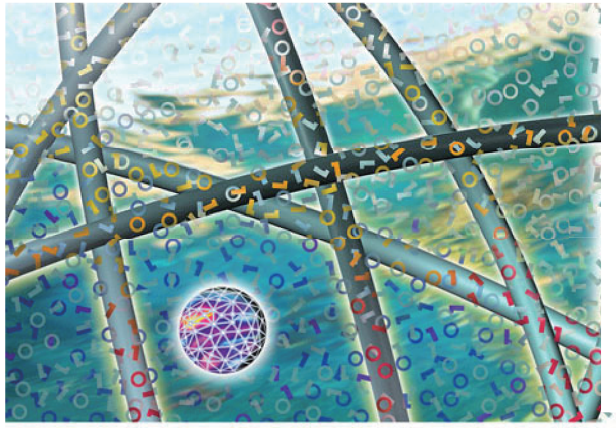
Special report: Managing information

Data, data everywhere

**Information has gone from scarce to superabundant. That b
benefits, says Kenneth Cukier (interviewed here)—but also**

Feb 25th 2010 | from the print edition

Like 30



The New York Times

SundayReview | The Opinion Pages

WORLD | U.S. | N.Y. / REGION | BUSINESS | TECHNOLOGY | SCIENCE | HEALTH

NEWS ANALYSIS

The Age of Big Data

By STEVE LOHR

Published: February 11, 2012

GOOD with numbers? Fascinated by data? The sound you hear is opportunity knocking.

Enlarge This Image

Chad Hagen

Mo Zhou was snapped up by I.B.M. last summer, as a freshly minted Yale M.B.A., to join the technology company's fast-growing ranks of data consultants. They help businesses make sense of an explosion of data — Web traffic and social network comments, as well as software and sensors that monitor shipments, suppliers and customers — to guide decisions, trim costs and lift sales. "I've always had a love of numbers,"

McKinsey Global Institute

Research | People | In the news | Contact us

Report

Big data: The next frontier for innovation, competition, and productivity

May. 2011 | by James Manyika, Michael Chui, Brad Brown, Jacques Bughin, Richard Dobbs, Charles Roxburgh

Download » Executive Summary PDF–922KB » Full Report PDF–6MB » Kindle MOBI–4MB » eBook EPUB–3MB

The amount of data in our world has been exploding, and analyzing large data sets—so-called big data—will become a key basis of competition, underpinning new waves of productivity growth, innovation, and consumer surplus, according to research by MGI and McKinsey's Business Technology Office. Leaders in every sector will have to grapple with the implications of big data, not just a few data-oriented managers. The increasing volume and detail of information captured by enterprises, the rise of multimedia, social media, and the Internet of Things will fuel exponential growth in data for the foreseeable future.

big data

- big data: 大量の非定型データから隠れた価値のある情報を引き出す技術の総称
  - 新たなビジネスモデルの構築や経営改革に繋げる
- 技術は以前から使われている
  - 検索ランキング、オンラインストアのお勧めシステムなど
  - さらには、クレジットカードの不正使用検出、保険制度など
- インターネット計測: 大量かつ不完全なデータからインターネットを把握する試み
  - 統計的な手法による推測

データの時代

- あらゆる分野でデータ革命とよばれる技術革新が進行中
- これまで難しかった応用が可能に
  - 膨大なデータへのアクセス、常に更新されるデータの解析、非線形モデルへの応用など
- 科学技術のあらゆる分野でビッグデータ解析が重要な研究手法に

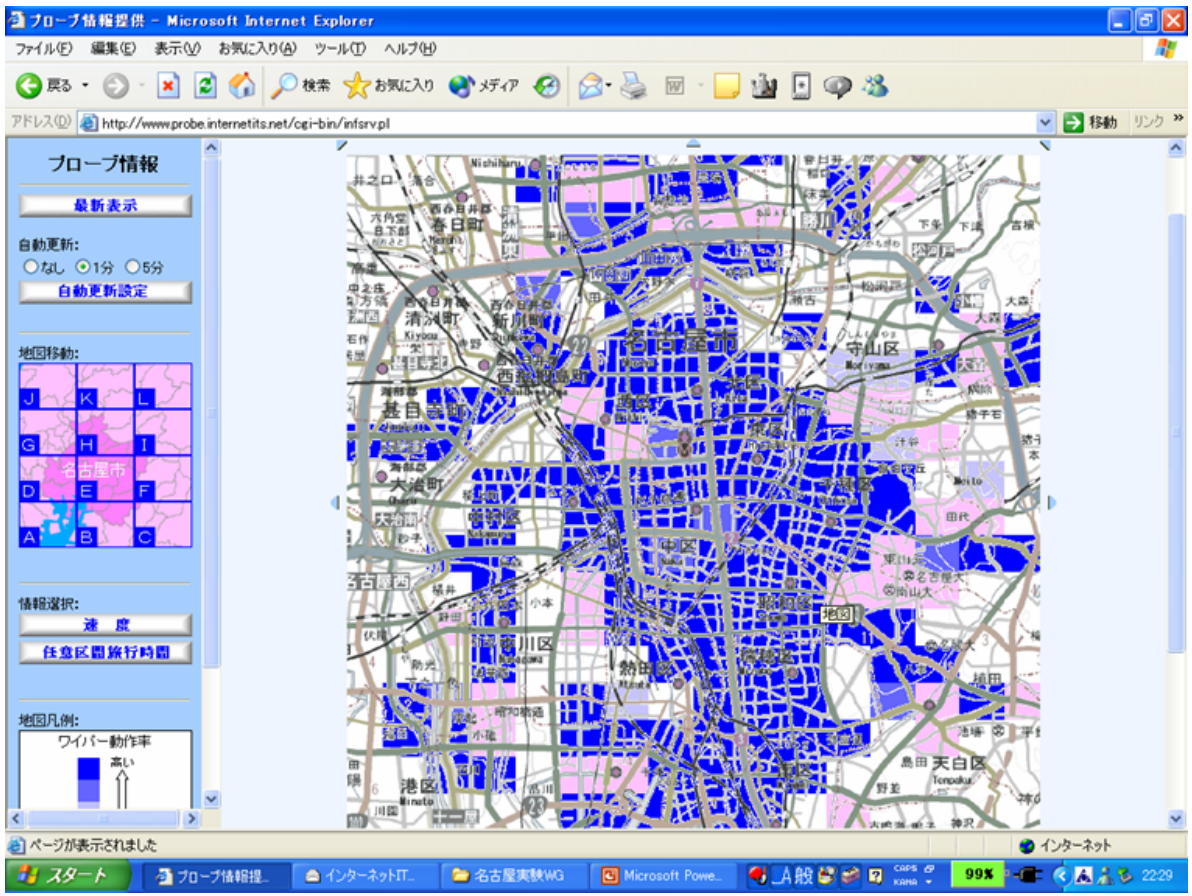
科学での応用例

e-サイエンス: 研究手法のパラダイムシフト

- 理論
- 実験
- シミュレーション (コンピュータ)
- データによる発見 (ビッグデータ)

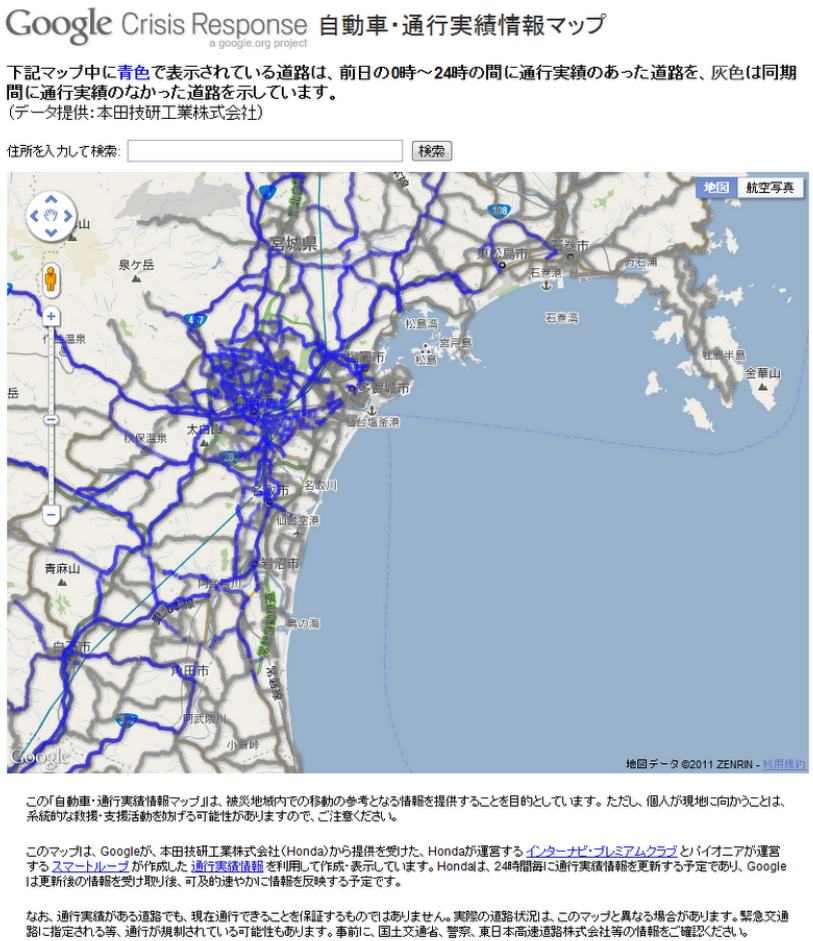
例: インターネット自動車実験

- WIDE プロジェクトが 2001 年に名古屋で実施
  - 1570 台のタクシーから位置、スピード、ワイパー動作状況を取得
  - ワイパーの動作情報から詳細な降雨状況の把握が可能に



東日本大震災での活用

- 前述のシステムは ITS の一部として利用中
- 地震の 3 日後に利用可能な道路情報が公開される
  - ホンダ (トヨタ, 日産) によるデータ提供



Google's Chief Economist Hal Varian on Statistics

The McKinsey Quarterly, January 2009

"I keep saying the sexy job in the next ten years will be statisticians. People think I'm joking, but who would've guessed that computer engineers would've been the sexy job of the 1990s? The ability to take data — to be able to understand it, to process it, to extract value from it, to visualize it, to communicate it — that's going to be a hugely important skill in the next decades, not only at the professional level but even at the educational level for elementary school kids, for high school kids, for college kids. Because now we really do have essentially free and ubiquitous data. So the complimentary scarce factor is the ability to understand that data and extract value from it."

計測とデータ分析はあくまで道具

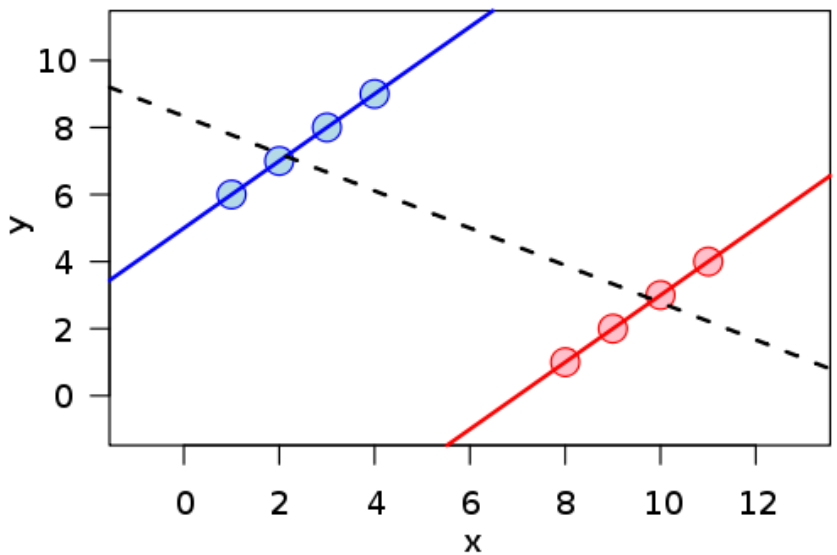
- 最近のビッグデータの話題はツールや手法が強調されがち
- データ解析はあくまでツール
  - 仮説を立てて、データで検証
  - 結果が予想と異なれば、そこから新たな疑問へ
  - このプロセスの繰返しから、役立つ情報や興味深い事実の発見
- 目的を持たずにデータを集め CPU を回し解析してもムダ
- 逆にデータから何を得たいかがはっきりすれば、やるべきことは見えてくる

思考プロセスの変化

- もちろん以前からデータを基に考えることは重要だった
- 情報技術によって、データに基づいて考え、考えをデータで検証する思考プロセスに変化
  - 扱えるデータの量と質、その表現方法が桁違いに
  - 文字通りデータと対話しながら考えることが可能に

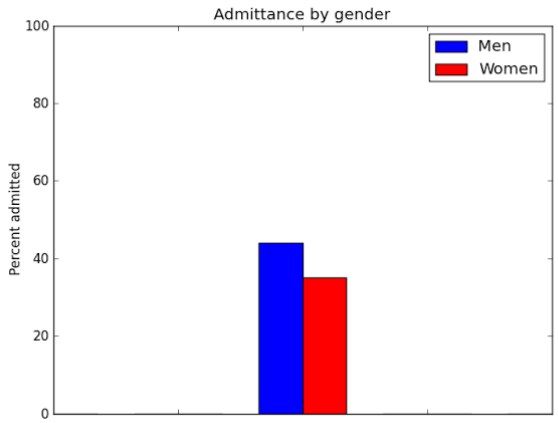
シンプソンのパラドックス

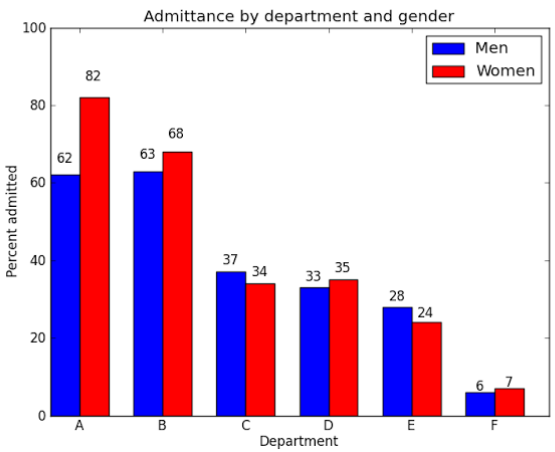
- 統計学におけるパラドックスのひとつ
  - 全体の傾向と、全体を分割した集団の傾向が事なる場合

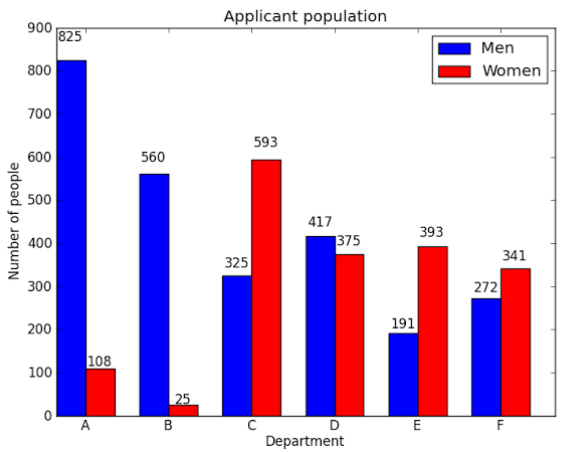


例: 1973 年の UC バークレイの女性差別訴訟

- 女性の大学院合格率が低いのは性差別であると訴訟
- 調査の結果、ほとんどの学部で女性の合格率が高いことが判明
- 女性の方が合格率の低い学部の受験者が多いことが原因







source:http://mathemathinking.blogspot.jp/2012/06/simpsons-paradox.html

自己紹介

長 健二朗 (Kenjiro Cho)

- 肩書
  - 株式会社インターネットイニシアティブ 技術研究所 所長
  - 慶應義塾大学環境情報学部 特別招聘教授 (2010-)
  - 北陸先端科学技術大学院大学 客員教授 (2002-)
  - WIDE プロジェクト ボードメンバー (2001-)
- 経歴
  - 1984 年神戸大学電子工学科卒業。同年キヤノン (株) 入社
    - ハードウェア設計から始め、OS 屋に
  - 1993 年コーネル大学コンピュータサイエンス学科修士修了
    - コンピュータサイエンス、分散システムを勉強
  - 1996 年 (株) ソニーコンピュータサイエンス研究所入社
    - 本格的にインターネット研究 (QoS 通信、計測) を開始
  - 2001 年慶應義塾大学より博士号 (政策・メディア) 取得
  - 2004 年より (株) インターネットイニシアティブ勤務
- 専門分野
  - インターネットのトラフィック計測と解析
  - データ通信サービスの品質と信頼性
  - オペレーティングシステムのネットワーク機能
  - クラウドシステムとコンテンツ配信システム

科目概要

インターネットによって、多様で膨大なデータが容易に取得できるようになった。そこから知見を引出し、新たなサービスを作り出すことが可能になり、ビッグデータや集合知として注目されている。しかし、これらを正しく理解し、道具として使いこなすためには、その背景にある統計、機械学習、システムに関する総合的な理解が欠かせない。

本授業は、インターネット上でのデータ取得と大規模データ解析の概要について学び、情報社会で必須となる大量情報から新たな知識獲得をするための基礎能力を身につける。

主題と目的／授業の手法など

インターネット上でのデータ収集とその解析手法について学習し、ネットワーク技術と大規模データ処理の総合的な知識と理解を得る。授業では、具体的な応用例について、その基礎技術と背景にある理論を関連づけて理解する。講義に加えて、毎回データ処理の演習を行い、習った理論をプログラムに実装してデータ処理をすることで、データ解析手法を身につける。

授業計画 (1/5)

- 第 1 回 イントロダクション (4/7)
  - ビッグデータと集合知
  - インターネット計測
  - 大規模データ解析
  - 演習: ruby 入門
- 第 2 回 データとばらつき (4/14)
  - 要約統計量 (平均、標準偏差、分布)
  - サンプリング
  - グラフによる可視化
  - 演習: gnuplot によるグラフ描画
- 第 3 回 データの収集と記録 (4/21)
  - ネットワーク管理ツール
  - データフォーマット
  - ログ解析手法
  - 演習: ログデータと正規表現

授業計画 (2/5)

- 第 4 回 分布と信頼区間 (4/28)
  - 正規分布
  - 信頼区間と検定
  - 分布の生成
  - 演習: 信頼区間
  - 課題 **1**
- 第 5 回 多様性と複雑さ (5/12)
  - ロングテール
  - Web アクセスとコンテンツ分布
  - べき乗則と複雑系
  - 演習: べき乗則解析
- 第 6 回 相関 (5/19)
  - オンラインお勧めシステム
  - 距離とエントロピー
  - 相関係数
  - 演習: 相関

授業計画 (3/5)

- 第 7 回 多変量解析 (5/26)
  - データセンシング
  - 地理的位置情報 (geo-location)
  - 線形回帰
  - 主成分分析
  - 演習: 線形回帰
- 第 8 回 時系列データ (6/2)
  - インターネットと時刻
  - ネットワークタイムプロトコル
  - 時系列解析
  - 演習: 時系列解析
  - 課題 **2**
- 第 9 回 トポロジーとグラフ (6/9)
  - 経路制御
  - グラフ理論
  - 最短経路探索
  - 演習: 最短経路探索

授業計画 (4/5)

- 第 10 回 異常検出と機械学習 (6/16)
  - 異常検出
  - 機械学習
  - スパム判定とベイズ理論
  - 演習: 単純ベイズ分類器
- 第 11 回 データマイニング (6/23)
  - パターン抽出
  - クラス分類
  - クラスタリング
  - 演習: クラスタリング
- 第 12 回 検索とランキング (6/30)
  - 検索システム
  - ページランク
  - 演習: PageRank

授業計画 (5/5)

- 第 13 回 スケールする計測と解析 (7/7)
  - 大規模計測
  - クラウド技術
  - MapReduce
  - 演習: MapReduce
- 第 14 回 まとめ (7/14)
  - これまでのまとめ
  - インターネット計測とプライバシー

ネットワーク計測とインターネット計測

- ネットワーク計測
  - 比較的限定されたネットワークにおける計測
  - ある時点のスナップショット
- インターネット計測
  - 大規模分散開放系であるインターネットにおける計測
    - 大規模分散系
    - オープンシステム (常に変化し続ける)

インターネットの計測 – 掴みどころのないものを測る

- インターネットにおける一般的な測定データの必要性
  - 例えば、一般的な利用者のトラフィック使用量分布など
- インターネットは開いた系で、つねに変化、発展、拡大
  - 中心も代表点もなく、測る場所や時間によって違う姿が観測される
  - インターネットの一般性を求める：掴みどころのないものを測る
- 現実にインターネットを運用、機器を開発、サービスを提供
  - その時点で最善の一般性を模索、将来予想し、常に見直す努力
- 技術面だけでなく、社会的、政策的、経済的な影響も考慮が必要

計測の重要性

計測はすべての技術の基礎

- ネットワークにおいては、見えないネットワークを見ようとする試み
- 運用、設計、実装、研究のすべてで必要

- しかし、インターネットの商用化、利用の拡大で難しくなってきた現状
  - トラフィック情報などは事業者の企業機密で開示されない
  - プライバシー情報の漏洩リスク

ネットワークのデータや挙動の特徴

- トラフィックやサービスの集約
  - 無数の要素の相互作用の結果、全体としてみれば個別要素の総和以上の独立な振舞い
- バラツキが大きく、偏った分布を持つ
  - 利用の偏り: 少数の利用者が大半のトラフィックを占めるなど
- さまざなな異常が日常的に発生
  - ソフトウェアのバグ、設定ミス、事故、メインテナンスなど

インターネット計測が難しい理由

従来の計測は工学的に定義された測定基準 (metric) の測定精度向上が中心。インターネットの計測は、膨大なあいまいデータから統計的手法を使って知見を引き出す。

- 大量、多様、変化するデータを扱う
- オープンな分散システムの複雑な挙動
  - 中心もなければ典型もない
  - さまざまな要因が複雑に絡む
- 動的変化
  - 適応的で障害に強いメカニズム
- さまざなな異常が日常的に発生

- いまだに体系的な理解に至っていない
  - いい教科書もない

大量データ

- インターネットの他に例をみない規模性と成長
- 解析能力を遥かに越えたデータ量
  - データサイズを小さくする必要
    - フィルタリング
    - 集約
    - サンプリング
  - 多変量の変数削減
- しかし時として詳細情報も重要
  - 大きな変化は往々にしてごく一部が引き起こす
  - 大局を見ながら、詳細にも気をくばる

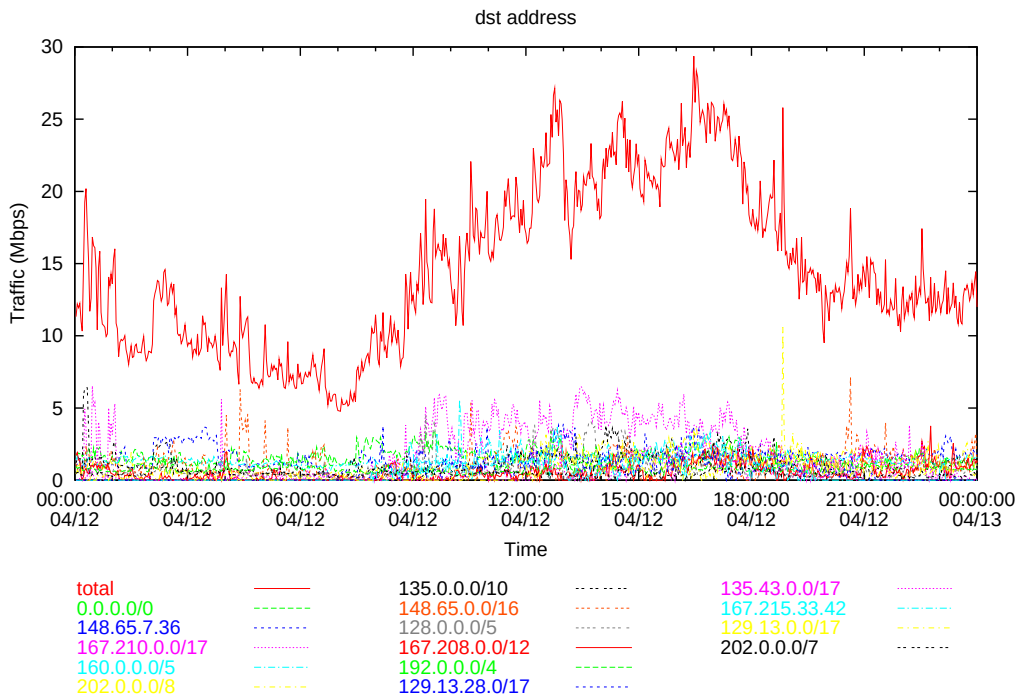
データの多様性

- 観測する場所によって異なる挙動が見える
  - 国、地域、時間
  - 企業と大学と家庭、バックボーンとアクセスネットワーク
- サービスごとに仕組みも利用者層も異なる
- 記録方法とデータフォーマット

典型的なネットワークも典型的なサービスも存在しない

時間とともに変化するデータ

- 時間帯や曜日による変化
- 長期的トレンド
  - 90 年代の web や 2000 年代の P2P ファイル共有、SNS で利用形態が大きく変化
- 将来予測は難しい



インターネット計測の制約

- 多くの問題がネットワークやサービスの境界で発生
  - 組織間協調が必要だが簡単ではない
- 測定そのものが測定対象に影響を与える
- 運用者の理解と協力が不可欠
  - 運用の現状を理解して実情にあった測定方法を工夫する必要
- 測定にはあまりコストをかけられない実情
  - 最新ルータを汎用 PC で測定する測定精度の限界
- データの解析とプライバシー、企業機密
  - 外部の研究者がデータ利用する障壁
  - 第三者が解析に使える汎用のデータを蓄積し公開する努力

授業で取り上げるトピックス候補

- オンラインお勧めシステム (協調フィルタリング)
- 検索ランキング (PageRank)
- SPAM 判定
- Web サーバログ解析
- 大規模データ解析 (MapReduce)
- インターネットトポロジ、経路探索
- SNS 利用者の繋がり
- 人気キーワード抽出
- 位置情報サービス
- パケット解析
- インターネットトラフィック

まとめ

インターネットの計測とデータ解析

- 計測はすべての技術の基礎
- 掴みどころのないものを捉えようとする試み
- 技術面だけでなく、社会的、政策的、経済的な側面にも配慮

本授業のテーマ

- インターネットの計測とデータ解析を題材に
- 容易に計測できないものをどう計るか
- 大量データからいかに情報を抽出するか

Ruby 入門

Ruby とは

- オブジェクト指向プログラミングのためのインタプリタ言語
- テキスト処理やシステム管理のための豊富な機能
- 1993 年に誕生したフリーソフトウェア
- 作者: まつもと ゆきひろ
- Ruby on Rails (Web アプリケーションフレームワーク) により広く普及

Ruby 関連情報

Ruby official site: `http://www.ruby-lang.org/`
Ruby レファレンスマニュアル: `http://www.ruby-lang.org/ja/documentation/`
Ruby の歩き方: `http://jp.rubyist.net/magazine/?FirstStepRuby`

Ruby の特長

- インタプリタ言語: 実行にはコンパイル不要
- 移植性が高い: ほどんどのプラットフォームで動作
- シンプルな文法
  - 変数に型が無く、動的型付けで任意の型のデータが格納可能
  - 変数宣言が不要で、変数の種類 (ローカル変数、グローバル変数、インスタンス変数など) は変数名から分かる
- ガーベッジコレクタ: ユーザによるメモリ管理が不要
- オブジェクト指向機能
  - 全てがオブジェクト
  - クラス、継承、メソッド
  - イテレータとクロージャ
    - 制御構造や手続きをオブジェクト指向で書ける
- 強力な文字列操作/正規表現
- 組み込みで多倍長整数機能をサポート

- Ruby の欠点: オブジェクト指向インタープリタなので遅い

Ruby commands

- irb: Ruby の対話インターフェイス

```
$ irb --simple-prompt
>> puts "Hello"
Hello
```

- ruby: Ruby 本体

```
$ ruby test.rb
```

または、

```
$ ruby -e 'puts "Hello".reverse'
olleH
```

演習: ライン数をカウントするプログラム

引数ファイルのライン数をカウントする

```
filename = ARGV[0]
count = 0
file = open(filename)
while text = file.gets
  count += 1
end
file.close
puts count
```

count.rb というファイルにプログラムを書いて実行

```
$ ruby count.rb foo.txt
```

もう少し Ruby らしく書くと

```
#!/usr/bin/env ruby
count = 0
ARGF.each_line do |line|
  count += 1
end
puts count
```

参考文献

[1] Ruby official site. `http://www.ruby-lang.org/`

[2] gnuplot official site. `http://gnuplot.info/`

[3] Mark Crovella and Balachander Krishnamurthy. *Internet measurement: infrastructure, traffic, and applications*. Wiley, 2006.

[4] Pang-Ning Tan, Michael Steinbach and Vipin Kumar. *Introduction to Data Mining*. Addison Wesley, 2006.

[5] Raj Jain. *The art of computer systems performance analysis*. Wiley, 1991.

[6] Toby Segaran. (當山仁健 鴨澤眞夫 訳). 集合知プログラミング. オライリージャパン. 2008.

[7] Chris Sanders. (高橋基信 宮本久仁男 監訳 岡真由美 訳). 実践パケット解析 第 *2* 版 — *Wireshark* を使ったトラブルシューティング. オライリージャパン. 2012.

[8] あきみち、空閑洋平. インターネットのカタチ. オーム社. 2011.

[9] 井上洋, 野澤昌弘. 例題で学ぶ統計的方法. 創成社, 2010.

[10] 平岡和幸, 掘玄. プログラミングのための確率統計. オーム社, 2009.

次回予定

第 2 回 データとばらつき (4/14)

- 要約統計量 (平均、標準偏差、分布)
- サンプリング
- グラフによる可視化
- 演習: gnuplot によるグラフ描画