アルゴリズムとプログラミング実践講座

http://akashi.ci.i.u-tokyo.ac.jp/mary/lectures/algorithm/

火曜　13:00 -- 14:30
I-REF 棟 6階　ヒロビー

稲葉真理　with 浅井大史・手塚宏史

レポートの〆切について

第一回・第二回（行列アクセス時間の計測）は
5月18日（日）〆切ます。
〆切後、いくつかのプログラムを　講義のweb
ページで公開する予定です。

第三回（単語の頻度カウント）レポートの〆切は、
もう少し先に設定する予定です。
先の課題の事情もあるので、単語を数えるだけでも・・・

（繰り返しになりますが）

この実践講座の背景

- 今年が初年度
- なぜ始めようと思ったか？
  - 研究室でも学生のバックグラウンドがまちまち
    - 研究室には、法学部・農学部出身者
  - アルゴリズムの基礎知識はあったほうが良い
    - 道具としてのアルゴリズム。 → 動作理解は楽しい
      - しめじソートとじゃがいもソート
    - 食わず嫌いの原因の一つは「証明」 → これは省く
  - 計算機実験で知っていてほしい基礎知識がある。
    - 実験の再現性（と効率）→ 自動化（手打ちは止めましょう）

（繰り返しになりますが）

この実践講座について

- ターゲット
  - 学部で情報系の授業を履修しなかった人
  - 学部で受けた情報系の授業の復習をしたい人

  かつ、アルゴリズムを使う（ ≒ プログラムを書く）予定がある人

- 目標
  - プログラムを書くとき、適材適所のツール（アルゴリズムを含む）が使えるように意識できること。

方針について

- 変わったところ
  → （できるだけ）深堀りもできるように
  → 全員のソースコードレビューは無理
  → （そのかわり）面白いコード・レポートの紹介

- 変わらないところ
  → 研究の助けになりそうなこと
  → 深堀りしなければ、あまり負担にならない

先週のレポートから
Hash 性能比較

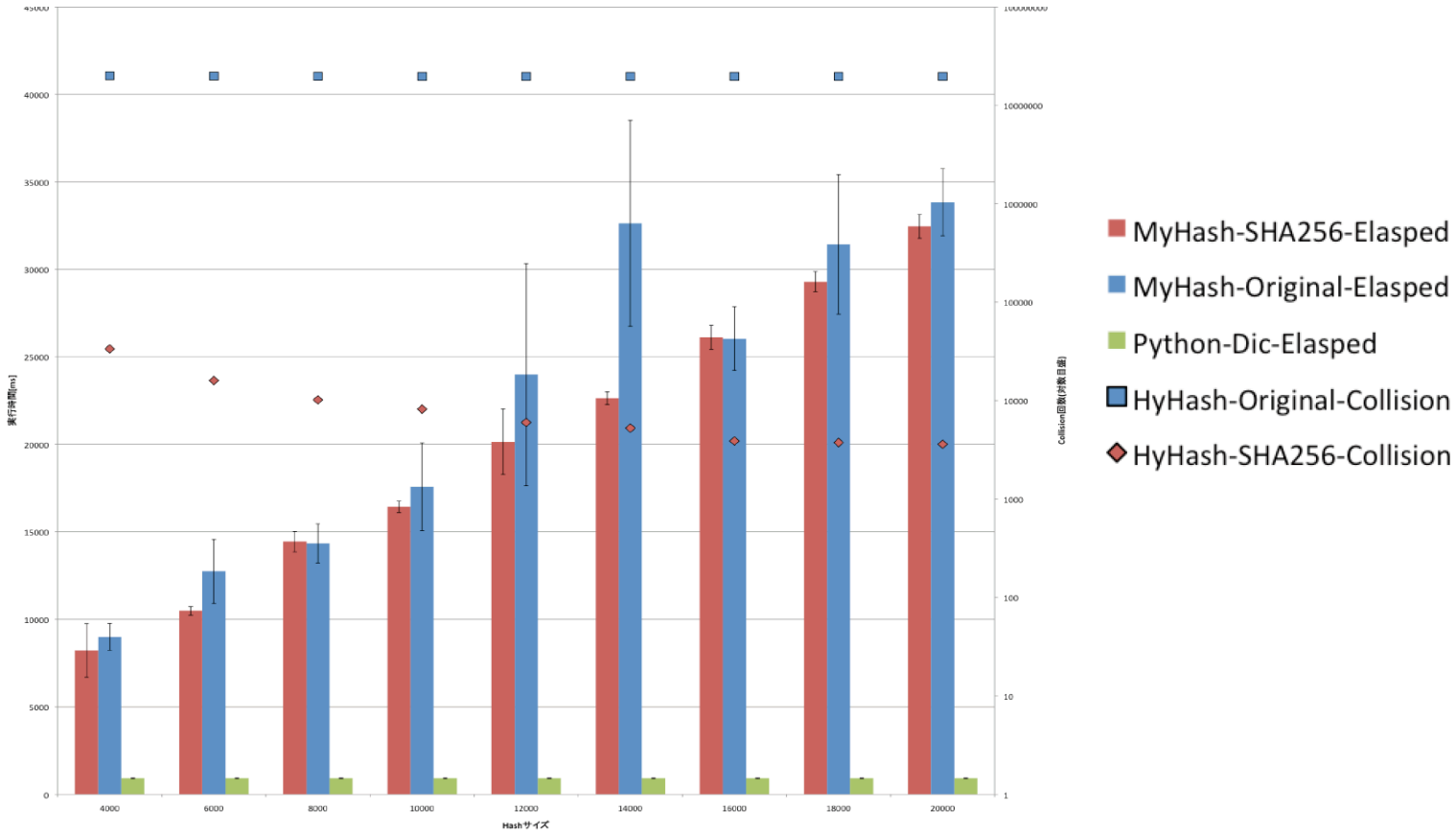


図 2　『銀河鉄道の夜』処理時間結果

先週のレポートから
HASH 詳細比較

表 1　Hash の性能比較表

	Hash サイズ	4000	6000	8000	10000	12000	14000	16000	18000	20000
MyHash-Original	平均	8992.730176	12741.45941	14335.29259	17573.66676	23978.81612	32632.82053	26036.5001	31422.89579	33838.72455
	標準偏差	763.8058074	1826.315298	1109.242106	2513.124672	6351.482293	5888.244205	1818.263057	3985.631608	1911.456245
	最小値	8083.519936	10536.31902	12927.26684	15083.37998	18226.41206	20820.20783	23223.1431	27122.13492	31300.70496
	最大値	10996.9449	17748.49486	16487.31494	25093.09006	41331.10714	40461.10106	31093.045	41652.20499	37741.34588
	Collision	19922520	19808200	19737540	19669920	19669920	19669920	19669920	19669920	19669920
MyHash-SHA256	平均	8219.993198	10483.2683	14435.00623	16420.87674	20140.69862	22625.73255	26114.69873	29295.25251	32451.37436
	標準偏差	1527.433578	239.5433393	578.6887398	333.3838216	1860.860194	361.911067	686.6431382	570.9393648	682.6236075
	最小値	7465.903997	10204.47397	14048.29693	16092.96203	19044.82913	22307.75499	25446.06996	28797.84584	31673.455
	最大値	14067.46793	11314.04781	16213.43184	17626.7221	27020.10393	23656.97503	27850.12603	30914.15501	34223.46687
	Collision	33440	15980	10160	8180	5980	5240	3880	3740	3600
Python-Dic (サイズはなし)	平均	911.9447947								
	標準偏差	21.46948364								
	最小値	887.6199722								
	最大値	967.3838615								

先週のレポートから
銀河鉄道の夜

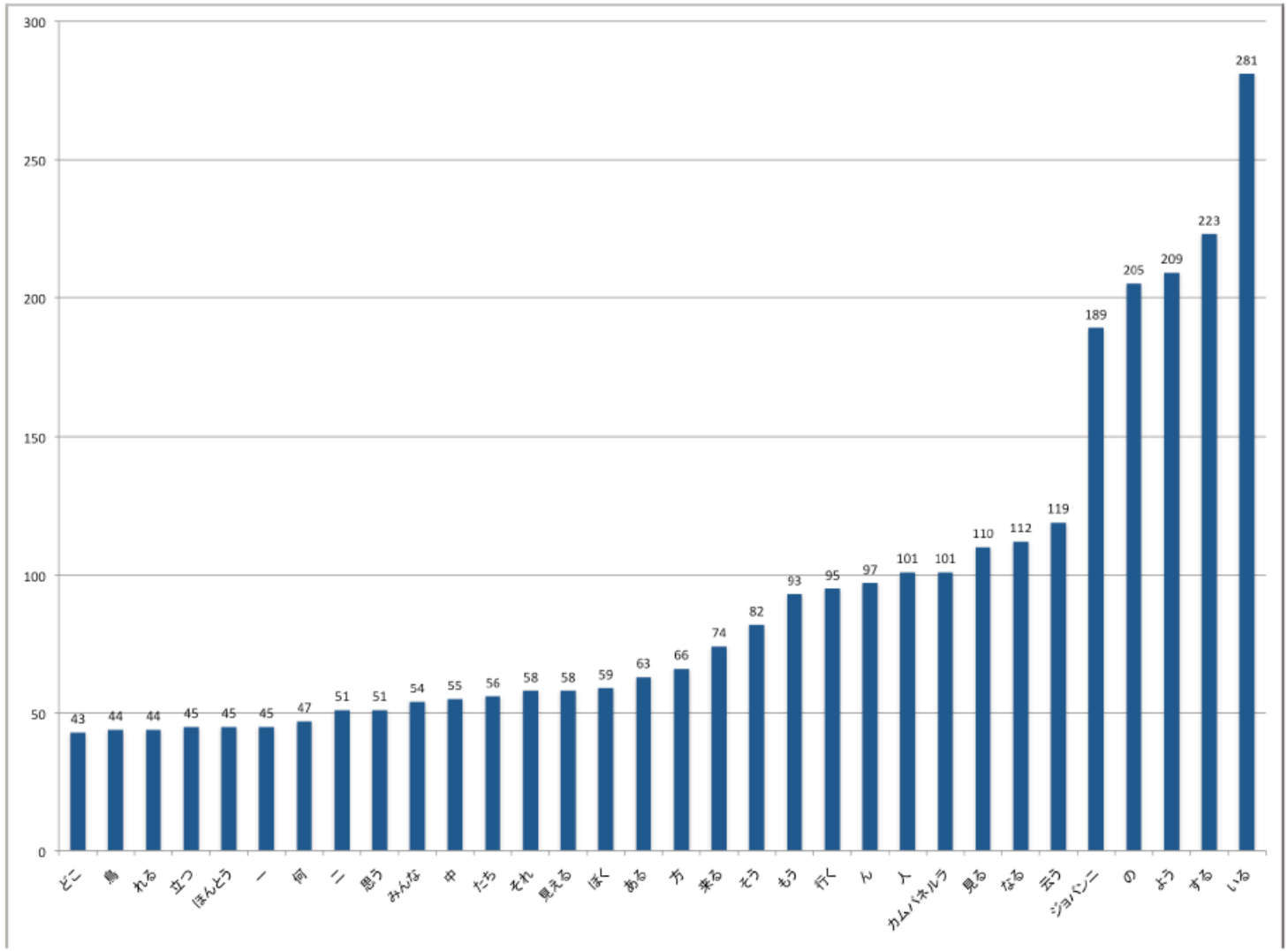


図 1　『銀河鉄道の夜』単語カウント結果のヒストグラム（$n > 40$）

先週のレポートから
Trie, Hash, map(C++) の比較

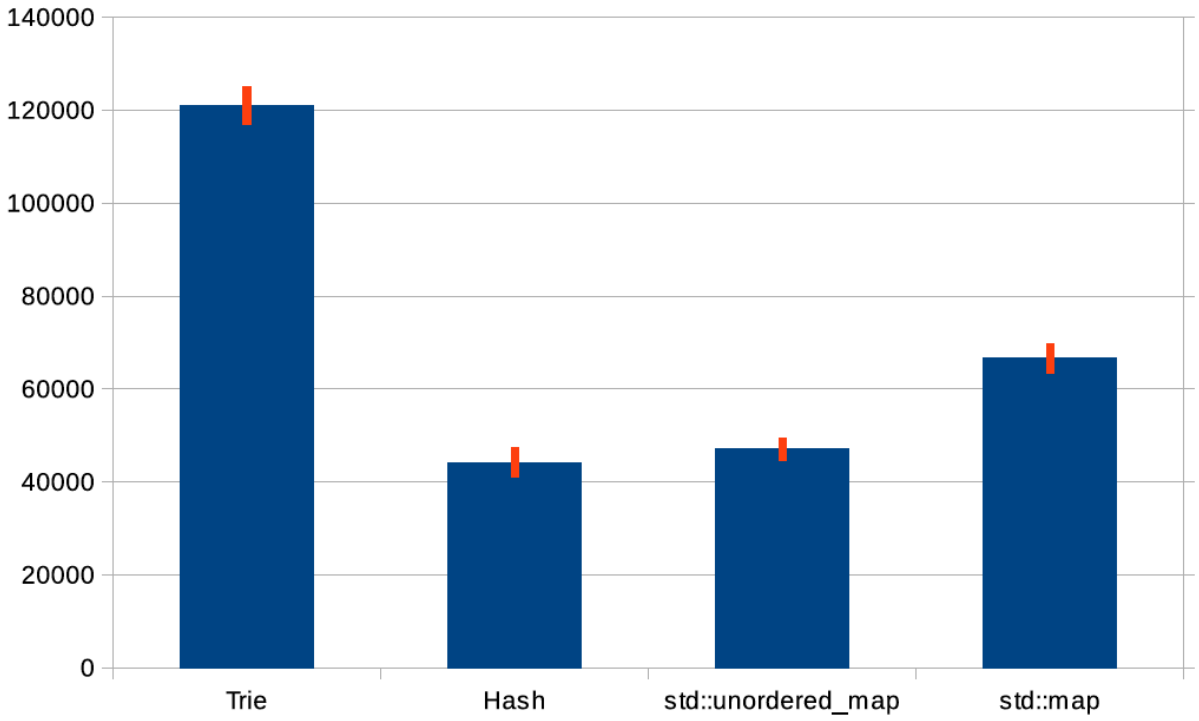


図 1　それぞれの実装において頻度カウントにかかった時間 (usec)（出力含まず）

先週のレポートから

2.2 「吾輩は猫である」に利用されている単語

先の実験で出力した単語の頻度カウントを見ると、空白を除くと 11953 個の単語が 205980 回利用されていることが分かった[*3]。これは、一つの単語が平均 17.2 回程度使われている事を意味する。これらのデータを元に、上位 100 位までの単語の出現頻度を図にしたものが、図 3 である。この図を見て分かるように、「吾輩は猫である」の 4.4% は「の」であり、3.6% は「。」であることが分かる。「吾輩は猫である」という名作[*4]を執筆するには、「の」及び「。」の使い方さえ理解できれば、8.0% 程度は書けてしまうということは間違いないだろう。また、上位 27 単語（0.2%）で出現単語回数の 50% を占めている。27 単語を理解すれば、日本語の本質の 50% を理解できるということは、この結果から言うまでもない。また、頻度についてジニ係数を計算[*5]した所、0.88 となり極めて不均衡な状態であった。

先週のレポートから
夏目漱石比較

2.3 「三四郎」に利用されている単語

先と同じ解析を夏目漱石著「三四郎」についても行った。空白を除くと、6242 個の単語が、106424 回利用されており、平均一つの単語に付き 17.0 回である。一般に物語が長くなると同じ単語を頻繁に利用し回数が伸びそうであるが、「吾輩は猫である」と「三四郎」で回数が殆ど変わらないというのは驚きである[*6]。

また、上位 10 単語について書き出すと「。」「、」「の」「た」「て」「は」「に」「を」「だ」「と」であった。殆どが助詞、助動詞であるというのは「吾輩は猫である」と同様の傾向である。また、ジニ係数についても 0.88 であり、これも同様の傾向である。

2.4 「それから」に利用されている単語

夏目漱石著「それから」については、空白を除き、6908 個の単語が 119484 回利用されており、平均 17.3 回であった。これもほぼ同等の平均利用回数を示している。

この作品の上位 10 単語は「、」「た」「。」「の」「に」「て」「は」「を」「と」「が」であり、ジニ係数は 0.88 であった。ほぼ先の 2 作品と同様の傾向である。

2.5 「手紙」に利用されている単語

比較的短い作品である夏目漱石著「手紙」について同様の解析を行った。空白を除き、1427 個の単語が、7175 回利用されており、平均 5.03 回利用されている。

この作品の上位 10 単語は「、」「の」「た」「に」「て」「。」「は」「を」「だ」「が」であり、ジニ係数は 0.73 であった。

先週のレポートから
太宰治・そして自分のレポート

2.6　「人間失格」に利用されている単語

別の作者の作品についても調べる。太宰治著「人間失格」については、空白を除き、4720 個の単語が、45854 回利用されており、平均 9.71 回利用されている。

この作品の上位 10 単語は「、」「の」「に」「て」「た」「は」「。」「を」「だ」「が」であり、ジニ係数は 0.83 であった。

2.7　まとめ

様々な作品についてまとめてみると、一般的な傾向として、やはり短い作品は平均単語使用回数が少なく、長い作品ほど多くなる傾向があるようである。しかしながら、一定の長さを超えるとほぼ安定して 17 回程度になるようである。これは、様々な異なる単語を利用しようとする意識が働くからではないかと思われる。これは、少なくとも夏目漱石と太宰治については変化しない傾向であると考えられる。

ちなみに、先の段落までについて今までと同様の解析を行うと、462 個の単語が 2391 回利用されており、平均 5.17 回である。また、ジニ係数は 0.68 であり、上位 10 単語は「、」「の」「する」「を」「。」「だ」「に」「は」「「」「」」であった。単語の平均利用回数的には「手紙」よりも長い文章であるべきであろうが、如何せん文章能力が無いので、このような短い文章でも単語の平均利用回数が多いと思われる。私自身の日本語の語彙数の増加が望まれる。

先週のふりかえり（グラフ）

線の形状や長さは無視、
どの点とどの点がつながれているのか
（つながれてないのか）だけに着目
→ グラフ という数学的概念

お約束の記法

グラフ $G = (V,E)$

V: 頂点集合

E: 辺集合

グラフのいくつかの基礎的な性質

- 次数： ある頂点につながっている辺の数

- 連結・連結度　どのくらい、頑丈に、つながっているか？
  - いくつ**頂点**を取ってもつながっているか？
  （注： 辺ではない）

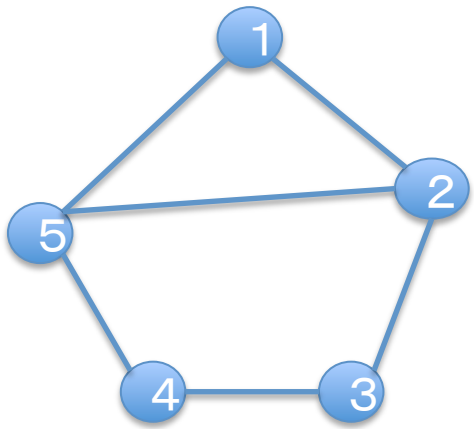
- 疎・密： 辺の数|E|は　$|V|^2$以下
  次数の平均が定数なら　O(|V|) 程度

グラフの表現 (バリエーションあり)

- エッジリスト　(1,2)(2,3)(2,5)(3,4)(4,5)(5,1)

- 隣接リスト

1	2,5
2	1,3,5
3	2,4
4	3,5
5	1,2,4



- 隣接行列

$$\begin{pmatrix} 0 & 1 & 0 & 0 & 1 \\ 1 & 0 & 1 & 0 & 1 \\ 0 & 1 & 0 & 1 & 0 \\ 0 & 0 & 1 & 0 & 1 \\ 1 & 1 & 0 & 1 & 0 \end{pmatrix}$$

グラフドローイング

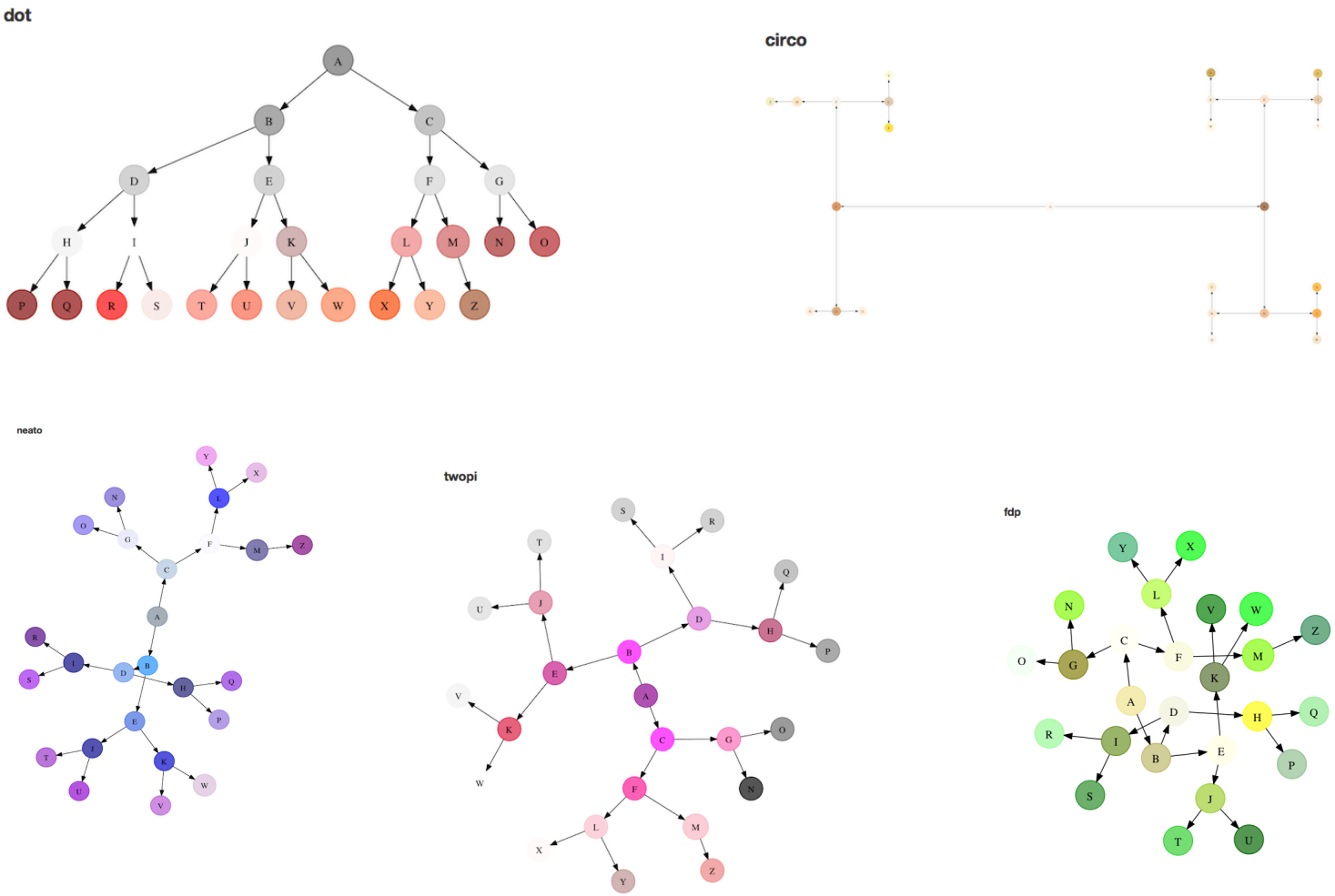
- どうやったらきれい？
  - エッジが少ない方が自由度は高い
    → たとえばTREE
  - 頂点に面積があると難しくなる （ラベルなど

  - そもそも、きれいって？
    → 色々な CRITERIA がある
    （はやるかどうかは、多数決）

グラフドローイング

- いろいろなクライテリア
  - 交差する辺の数を最小化
  - 線は直線で
  - グリッド上で辺が曲がる回数を最小化
  - できるだけ各ノードを離す（バネモデル）
  - 中心からできるだけ放射状に
  - 頂点は環状に、球状に
  - 近いノードは近く
  - 階層構造が見えるように描画
  - クラスターが見えるように描画

同じグラフの色々なレイアウト（graphviz)



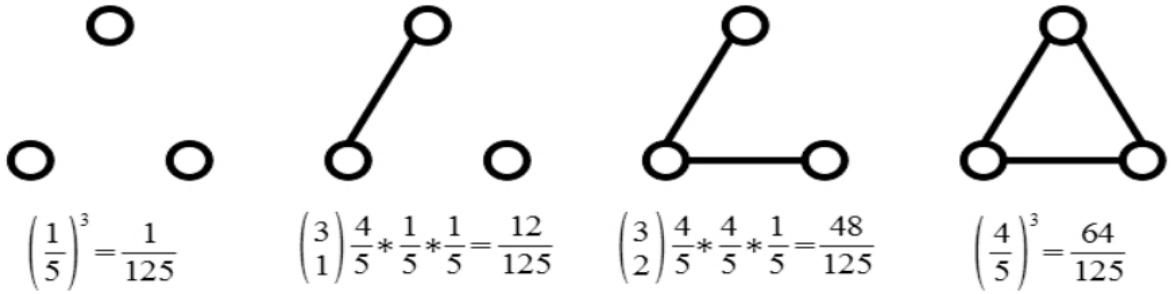
http://melborne.github.io/2013/04/02/graphviz-layouts/

ちょっと変わったグラフ

Random Graph

- 頂点数　n
- 確率 p ，0<=p<=1
- 完全グラフの辺を　確率 pで残したグラフ

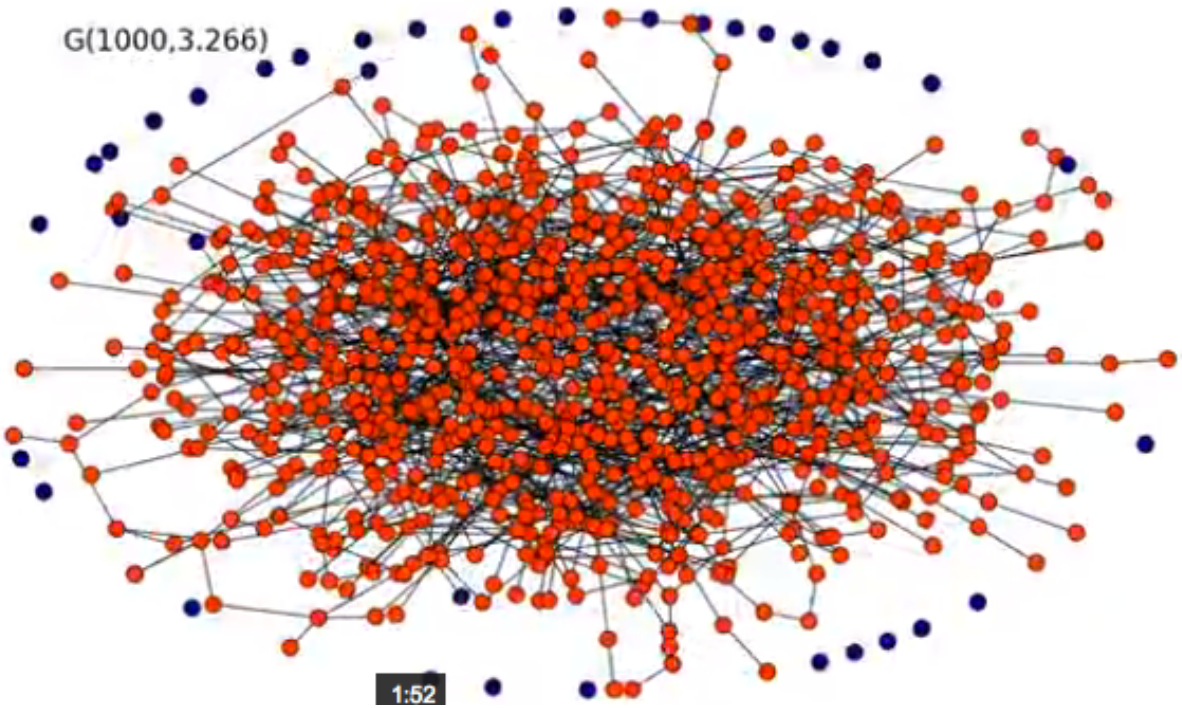
- 例　n=3, p=0.8



wikipedia より♪
http://upload.wikimedia.org/wikipedia/commons/f/fa/Random_3.png

The evolution of the G(n,p) random graph

- https://www.youtube.com/watch?v=mpe44sTSoF8



レポート・成績について

- ほぼ毎回、プログラミング課題を出題する予定
  - 効率の良い計算機実験のためのツールを使ってみる
  - アルゴリズムの実装
  - ライブラリの利用・・・　など

- ３回以上、レポートのファイルとプログラムのソースを添付し、メールで提出のこと
  - E-mail: algorithm2014@edu.jar.jp
  - サブジェクト 「アルゴリズムとプログラム実践講座・レポート」
  - 学生証番号と名前は、 メールの本文にも書いてください。
  - 〆切：次の授業の直前の日曜日深夜 （講評の都合上。〆切後も受付ます）
  - プログラムやレポートは（見本として）公開することがあります。適宜、作者名やコピーライトをいれておいてください。公開不可の場合は、プログラムの冒頭にその旨、コメントをいれておいてください。
  - 質問・作問提案も歓迎 （作問については採用の場合は別途加点）
  - サンプルプログラムは「初心者向け」です。 **上級者は無視してください。**

推奨環境など

- Linux, Mac,（Windows+Cygwin）
- 仮想マシン環境（VMware, VirtualBox, Parallels)
  - 余裕があれば、いろいろな組み合わせを試して比較してみると面白いと思います

- 言語
  - 自由。ただし、一般的でない言語については、上記いずれかのOS 上にインストール可能なもの

第四回　課題

Sample programs

graph.awk

グラフの生成・ツールを使った表示

1. ランダムグラフや WSモデル（後述）を生成するプログラムを作成する。ただしノード数 n　および確率p は指定できるようにすること。

2. ノード数いくつくらいまで生成できるか確認せよ

3. 生成したグラフそれぞれの最大次数、平均次数を求めよ

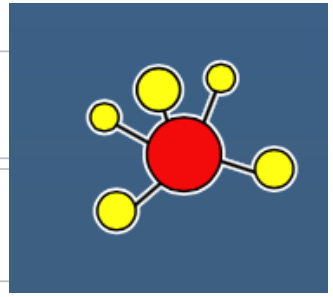
4. 生成したグラフを、グラフ可視化ツール（あるいは自作ツール）で表示する。

この課題の狙うところ

- パラメータで性格が変わるグラフの生成

- 平均次数が小さい巨大グラフの生成
（どこまで大きいグラフが生成できるか）

- グラフを表示するツールを使ってみる



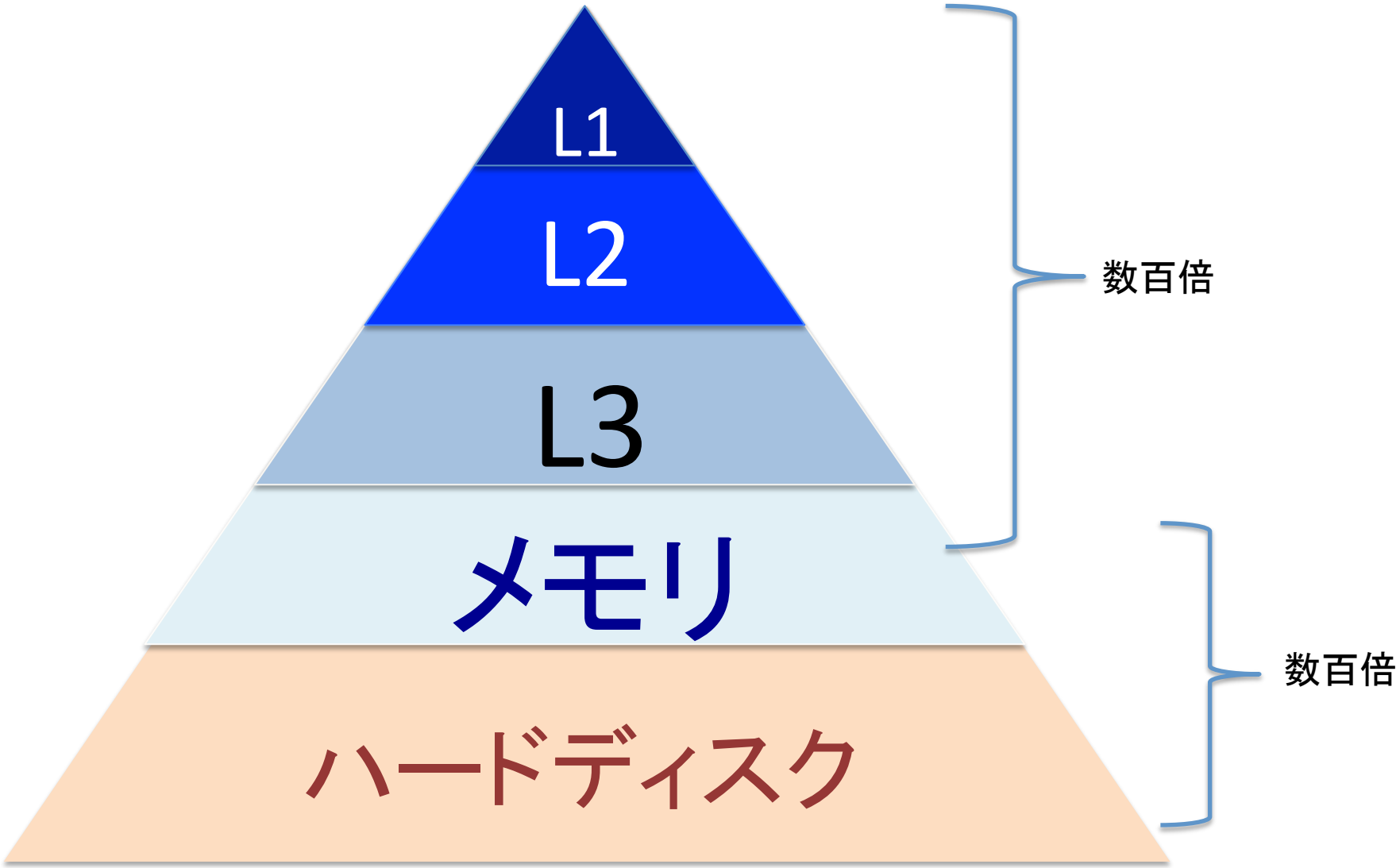
ところで

世界の人口が数十億
Web ページの数も、数億から数百億？

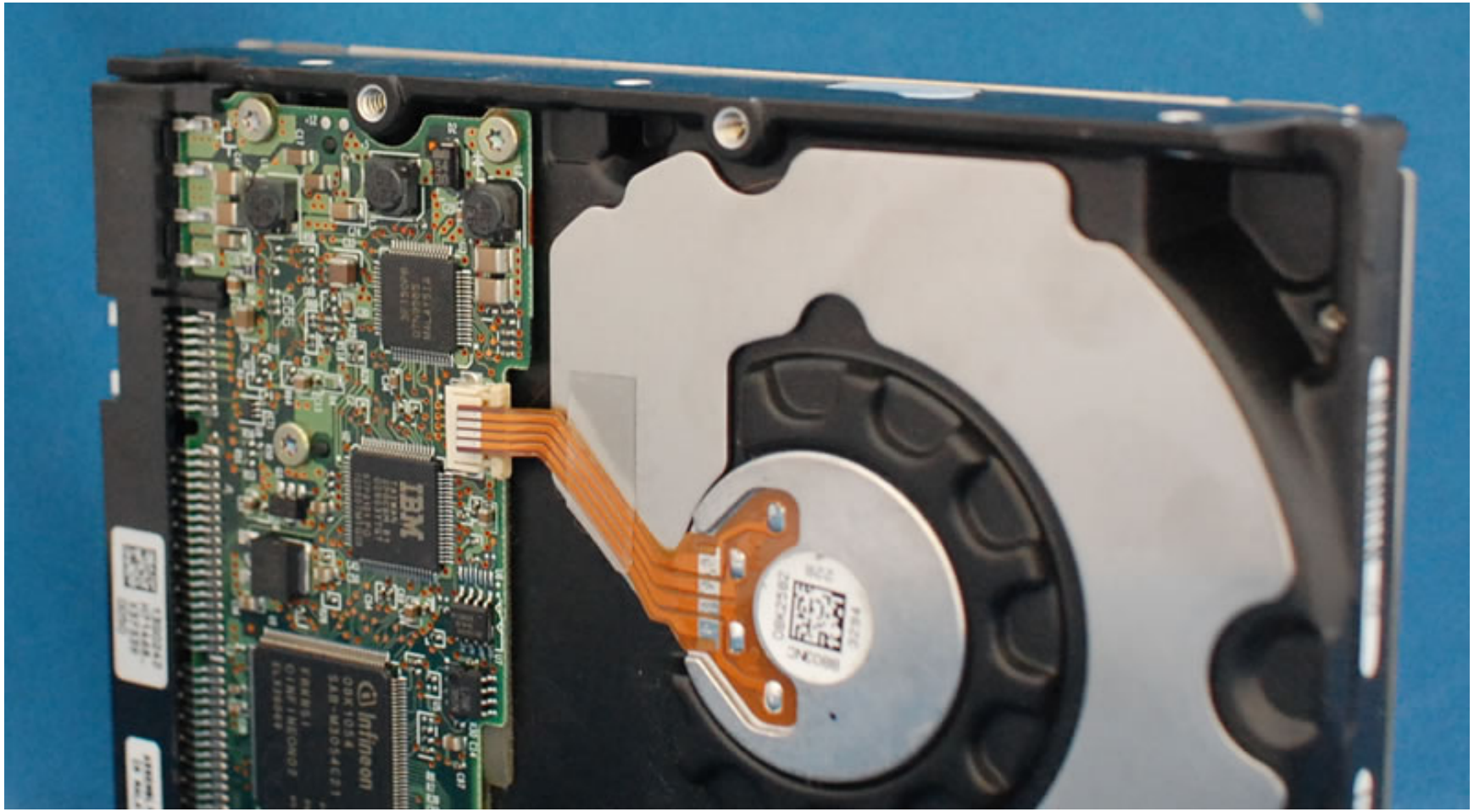
ノードが、そもそもメモリーにのらない。

→ **どうしましょ？**

（おまけ）メモリー階層

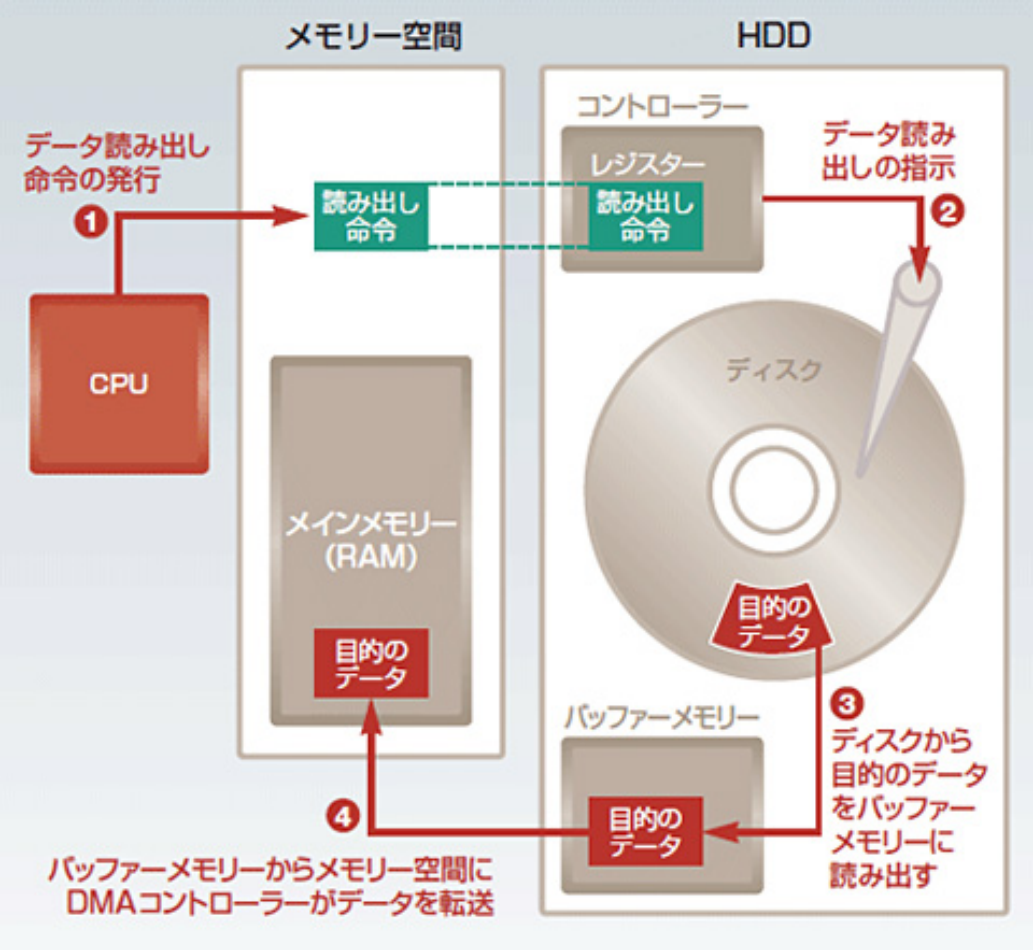


（おまけ）ハードディスクの基礎知識



http://hddnavi.jp/focus/controller_change2.html より

（おまけ）ハードディスクの基礎知識



IT Pro 2007/8/31 倉田雅弘氏の記事より
http://itpro.nikkeibp.co.jp/article/lecture/20070824/280199/

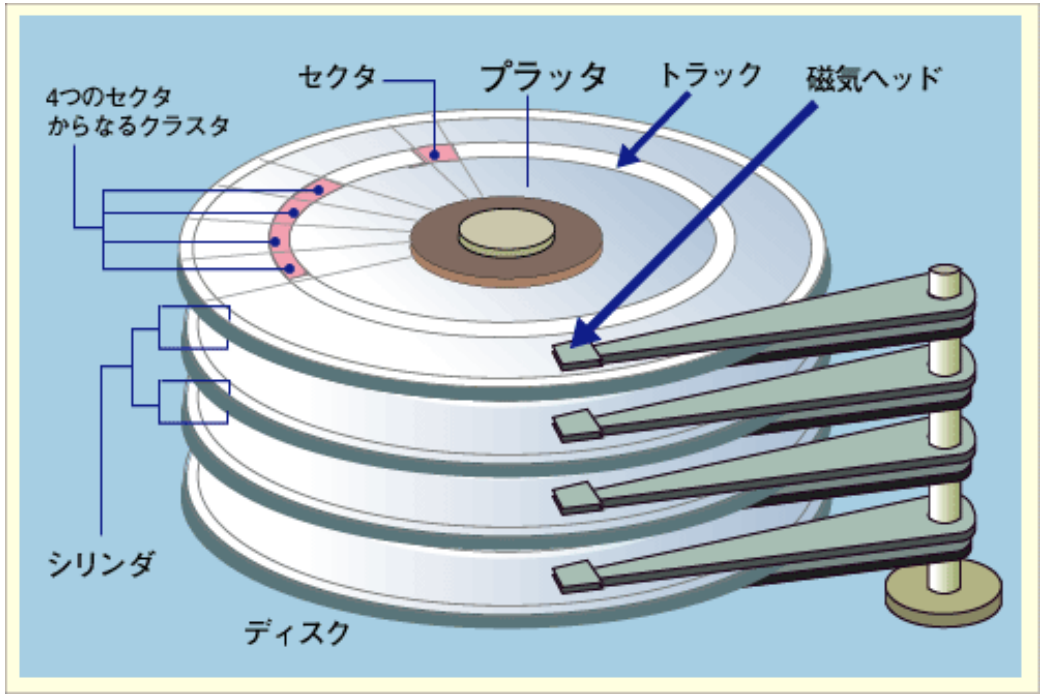
ハードディスクの部分的な不調

太古の昔は、メインコンピュータが管理

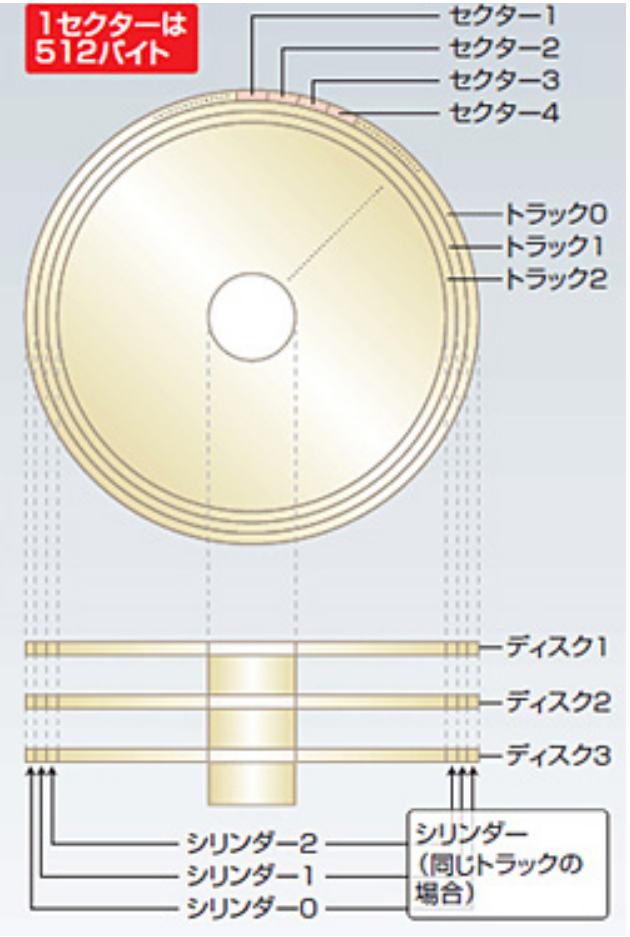
→ defect list で、アクセスしてはいけない
セクターを管理していた

今は、（たぶん）「物理アドレス」と「論理アドレス」
が分かれていて、計算機は「論理アドレス」で
アクセス、HDD のコントローラが交替セクタを扱う

（おまけ）ハードディスクの基礎知識



http://www.atmarkit.co.jp/fwin2k/experiments/defragment/defragment_1.html



（注）最近のディスクは1セクタ4k

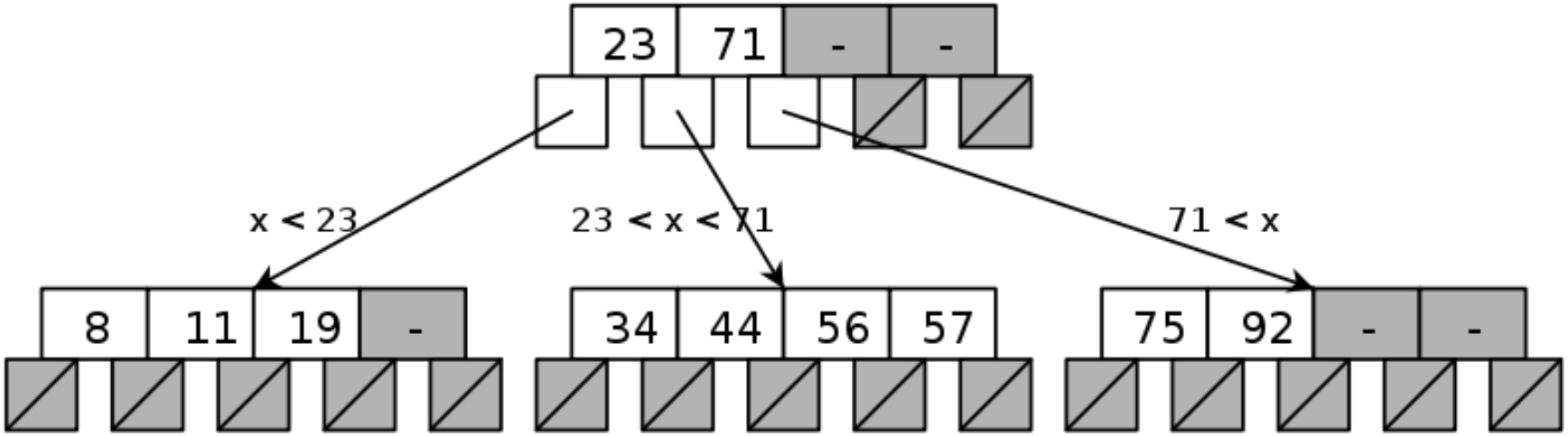
IT Pro 2007/9/3 倉田雅弘氏の記事より
http://itpro.nikkeibp.co.jp/article/lecture/20070824/280261/

（おまけのおまけ）太古の昔

動いてる軽めのHDD を
　物理的にさわったら、怒られた。
「クラッシュしたらどうするんだ！」

→　今、そもそも、ラップトップ　PC
　歩きながら、使う人いるし・・・
　ハードディスク回転中！

（おまけ）B-TREE



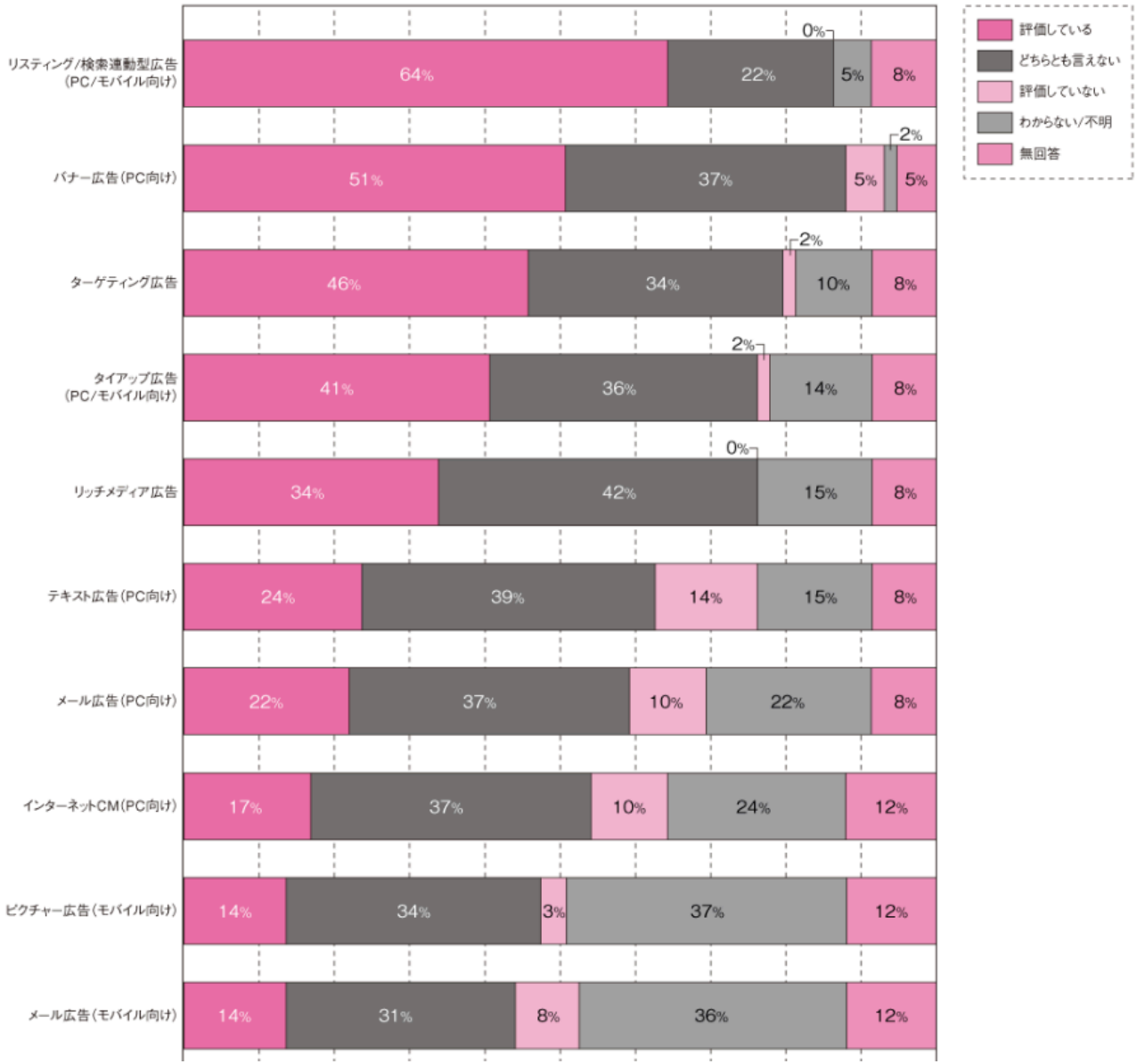
Wikipedia より

アルゴリズムのオーダーはまったく一緒
でも、現実に、性能は結構違う

講義準備していたときに
学生さんに言われた言葉

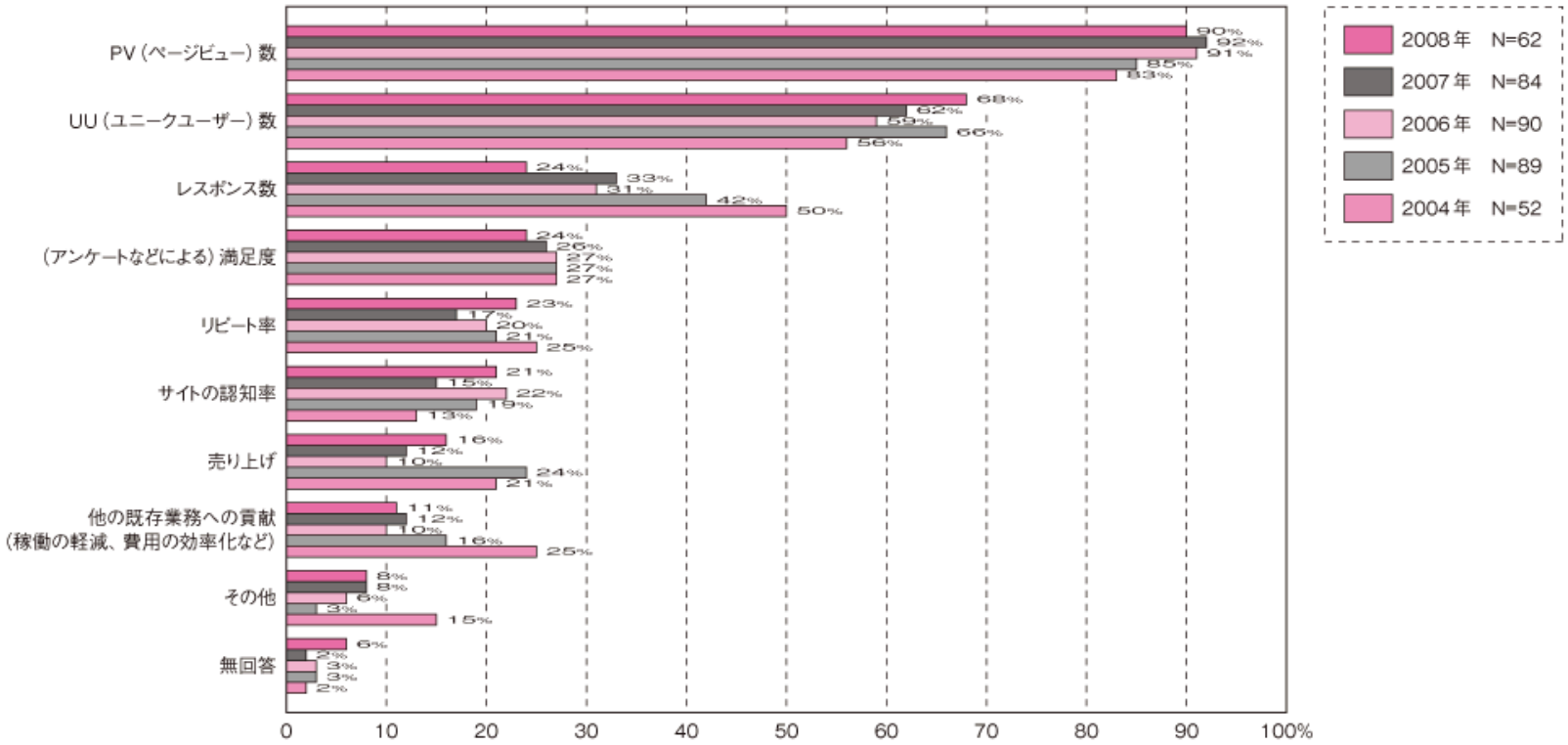
「お金が稼げるアルゴリズムの話を聞きたい」

企業側広告評価（ネット広告白書2010）



効果測定（ネット広告白書2010）

● 資料 6-1-32 大手広告主の自社サイトの効果測定指標（複数回答）



© 社団法人日本アドバタイザーズ協会
Web 広告研究会
2008 年度調査

ユーザアンケート（ネット広告白書2010）
広告後の行動

● 資料 6-2-7 広告を見た後に起こした行動（複数回答）

（%）

	テレビ広告 N＝1,419	新聞広告 N＝1,066	ラジオ広告 N＝548	雑誌広告 N＝815	パソコンでのインターネット広告 N＝1,327	ケータイでのインターネット広告 N＝832	ゲーム機でのインターネット広告 N＝144	新聞に折り込まれているチラシ N＝907	郵送されてくるダイレクトメール N＝904	交通広告 N＝829	電話帳に掲載されている広告 N＝301	屋外広告（OOH） N＝726	フリーペーパー N＝686
資料請求をした	3.4	4.8	1.8	5.2	5.8	8.4	1.4	4.3	4.3	0.4	0.7	0.8	1.6
商品やサービスを購入した	17.1	13.2	5.1	13.0	10.8	10.2	2.8	27.3	12.7	4.5	6.0	2.8	11.7
購入する商品やサービスの候補にした	18.9	15.5	9.7	18.3	11.1	10.7	3.5	20.5	14.6	11.2	6.6	8.0	16.0
お店を訪問した	8.9	7.4	1.5	8.5	3.8	5.0	2.8	21.9	7.5	2.5	5.3	3.2	18.4
電話した	2.7	3.1	2.7	2.8	2.3	1.9	2.8	4.1	2.5	0.1	18.9	0.3	5.4
クーポン券を保存したり、利用したりした	0.0	3.8	0.0	6.5	8.7	20.1	3.5	12.5	7.6	1.0	1.3	0.0	26.1
クリックして、パソコンでウェブサイトを閲覧した	0.0	0.0	0.0	0.0	34.6	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
クリックして、ケータイでウェブサイトを閲覧した	0.0	0.0	0.0	0.0	0.0	36.5	0.0	0.0	0.0	0.0	0.0	0.0	0.0
ホームページのアドレスを入力して、パソコンでウェブサイトを閲覧した	8.9	8.4	2.9	10.8	10.6	2.5	9.0	5.1	5.9	3.0	2.3	2.1	5.1
ホームページのアドレスを入力して、ケータイでウェブサイトを閲覧した	1.6	1.5	0.2	2.9	2.1	6.5	2.1	0.9	1.0	1.6	0.7	1.0	2.0
パソコンでキーワードを検索して、ウェブサイトを閲覧した	17.2	10.7	6.8	13.9	11.0	2.6	5.6	6.7	4.6	8.6	2.3	5.5	7.4
ケータイでキーワードを検索して、ウェブサイトを閲覧した	1.8	1.3	0.7	2.7	1.9	8.3	0.0	0.8	0.4	1.9	0.3	1.1	2.3
QRコードをケータイで読み取って、ウェブサイトを閲覧した	1.7	5.3	0.0	12.8	5.5	6.3	1.4	4.6	2.9	1.8	0.7	0.3	7.4
ケータイで空メールを送信した	0.6	1.0	0.0	2.1	2.0	12.9	0.7	0.9	0.3	0.4	0.3	0.1	1.7
その他	0.3	0.7	0.4	0.5	0.2	0.6	2.1	0.7	0.6	0.4	1.3	0.0	0.1
何もしたことがない	44.7	44.5	64.1	36.1	33.8	31.3	61.8	31.8	47.9	63.1	54.8	70.9	37.0
わからない	6.3	5.6	8.8	7.9	5.0	10.3	9.0	5.0	6.7	8.8	8.6	9.0	6.4

回答者：各広告視認者

©impress R&D, 2009-2010

ユーザアンケート（ネット広告白書2010）
広告後の行動（拡大）

	テレビ広告 N＝1,419	新聞広告 N＝1,066	ラジオ広告 N＝548	雑誌広告 N＝815	パソコンでのインターネット広告 N＝1,327	ケータイでのインターネット広告 N＝832	ゲーム機でのインターネット広告 N＝144
資料請求をした	3.4	4.8	1.8	5.2	5.8	8.4	1.4
商品やサービスを購入した	17.1	13.2	5.1	13.0	10.8	10.2	2.8
購入する商品やサービスの候補にした	18.9	15.5	9.7	18.3	11.1	10.7	3.5
お店を訪問した	8.9	7.4	1.5	8.5	3.8	5.0	2.8
電話した	2.7	3.1	2.7	2.8	2.3	1.9	2.8
クーポン券を保存したり、利用したりした	0.0	3.8	0.0	6.5	8.7	20.1	3.5
クリックして、パソコンでウェブサイトを閲覧した	0.0	0.0	0.0	0.0	34.6	0.0	0.0
クリックして、ケータイでウェブサイトを閲覧した	0.0	0.0	0.0	0.0	0.0	36.5	0.0

講義準備していたときに
学生さんに言われた言葉

「お金が稼げるアルゴリズムの話を聞きたい」
→　いや、もし稼げれば・・・

（稼げるかどうかはおいておいて）
「インターネットが情報を変えた」から、
その結果　「重要になった」アルゴリズムの話

情報があふれてる世界で
（自分が）変わってきてるなと感じる事

数の暴力。
（たぶん）速度重要（リニアタイムが望ましい）

（割と）多数決の世界。
（割と）もれがあっても気にしない。
（割と）アドバーサリーは気にしないみたい。

→　確率・統計的な「説得力」の大事さは
（たぶん）増している。

情報があふれてる世界で（自分が）面白いと思うところ

（1） 探す

- 欲しい情報を探す （Information Retrieval)
  - data mining, クラスタリング，機械学習, 統計

（Google という名の広告屋さんは、とても儲かっている♪）

（2） 予測する (complex network)

- 情報がどう流れるか？
  - どう広告をうつのが有効か？ インフルエンサー

情報があふれてる世界で（自分が）面白いと思うところ

（1） 探す

- 欲しい情報を探す （Information Retrieval)
  - data mining, クラスタリング，機械学習, 統計

（Google という名の広告屋さんは、とても儲かっている♪）

（2） 予測する (complex network)

- **情報がどう流れるか？**
  - **どう広告をうつのが有効か？ インフルエンサー**

予測のために

現実にあった数理モデルをつくる

→ その上でシミュレーションする

→ 適当なグラフを作って
ランダムウォーク

予測のために

現実にあった数理モデルをつくる

→ その上でシミュレーションする

→ 適当なグラフを作って
ランダムウォーク

→ 割とあたる

予測のために

現実にあった数理モデルをつくる

→ その上でシミュレーションする

→ 適切なグラフを作って
ランダムウォーク

感染モデル

最初聞いたときは、眉唾と思っていたが（反省）、
人為的な要素がないだけに、結構「あたる」

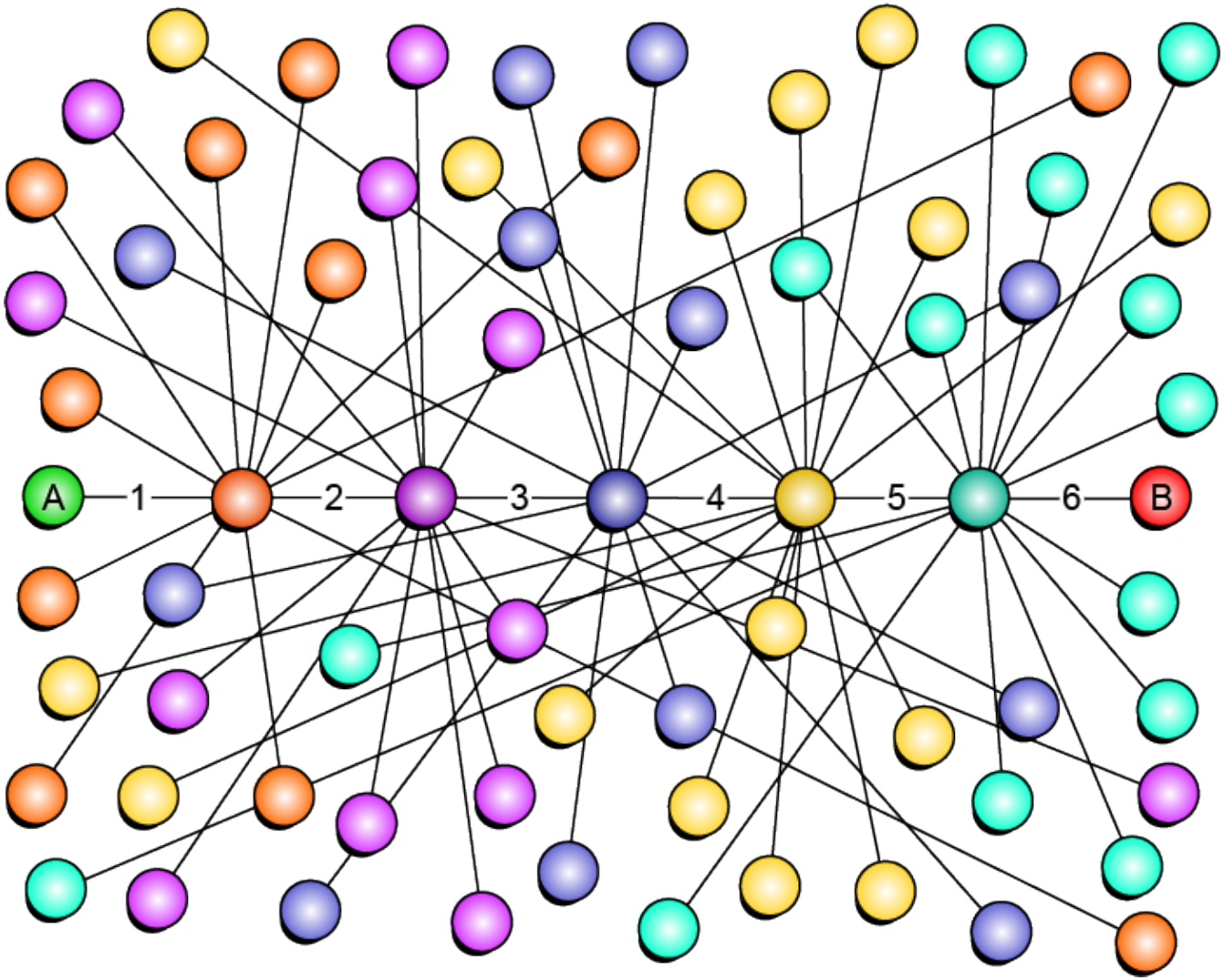
WSモデル [1998]
（「スモールワールド」序文より）

Watts が父から聞いた話

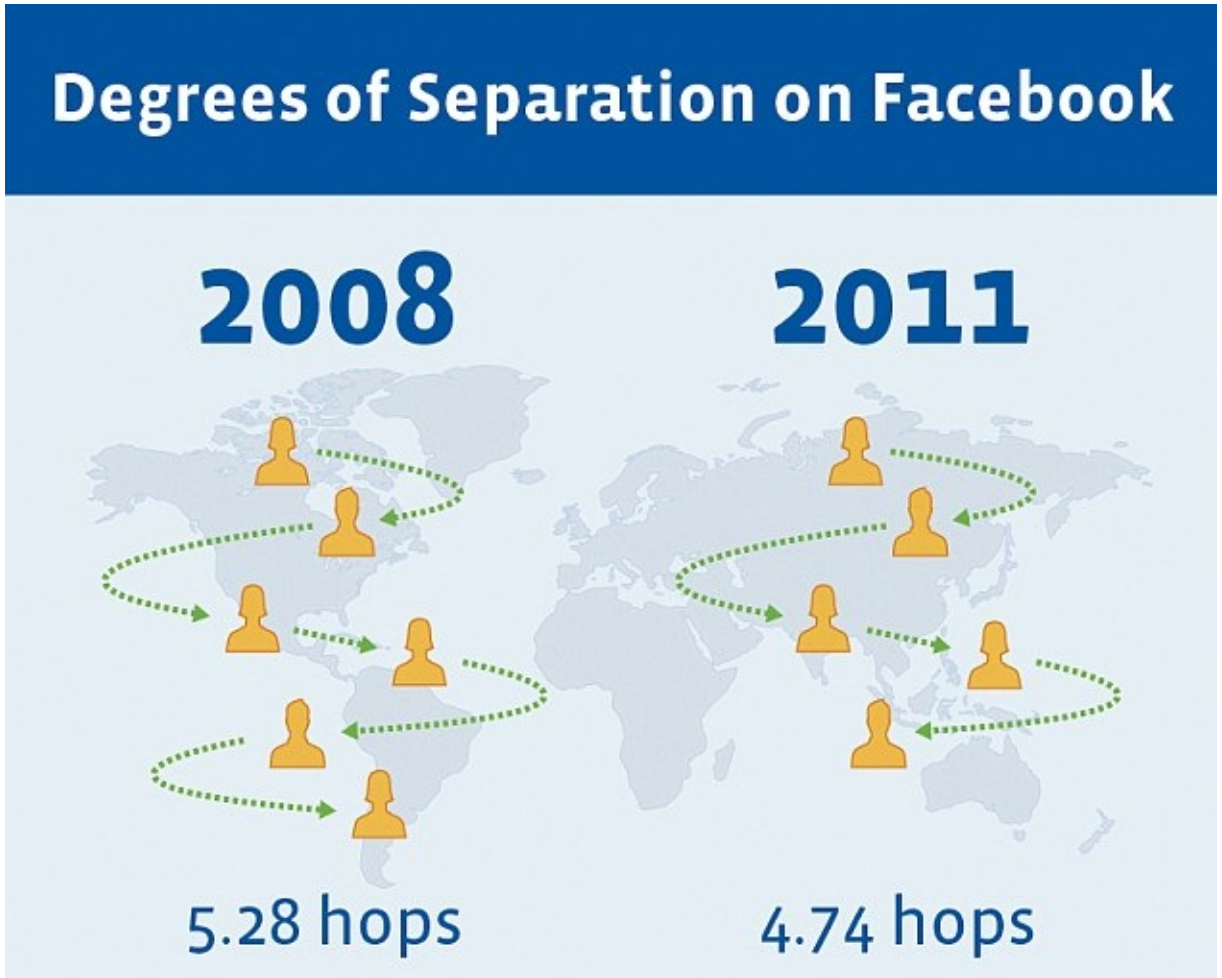
「おまえはたかだか６人の人を介すればアメリカ合衆国大統領と握手できるんだよ」

「いかなる二人であっても数人を介して握手できる世界とはどのようなものなのか、どのようなダイナミクスが隠されているのかを明らかにするという研究テーマは？」博士論文執筆中の **Watts** は　**Strogatz** 先生に意見を求めた

世界は狭い
SIX DEGREES OF SEPARATION



世界はますます狭くなった

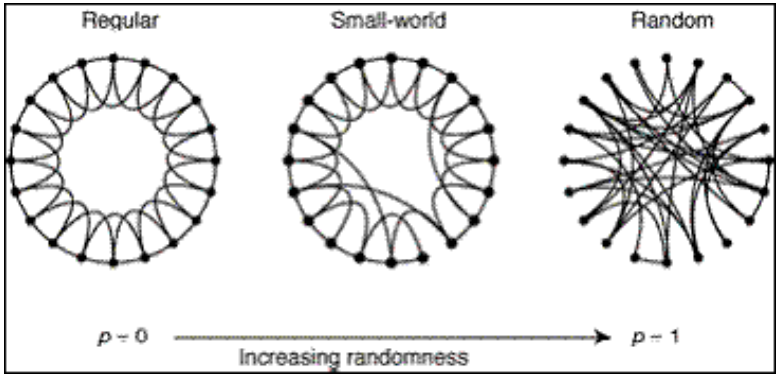


http://www.thedrum.com/news/2011/11/23/facebook-shrinks-world-four-degrees-separation

WSモデルの作り方

「複雑ネットワーク」増田直紀・今野紀雄著より

1. 平均次数　k　を決める
2. 頂点n個を輪状に置き、右左、それぞれk/2個隣までつなぐ（拡張サイクル）
3. 確率　p　で　枝を選ぶ（選ぶ枝は　pkn/2本)
4. 選んだ枝は、確率1/2 で片側の頂点を切り離し、別の頂点につなぐ。ただし多重辺にはしない。



p=0 なら拡張サイクル
p=1なら　ランダムグラフ
Albert and Barabasi, 2002

WS　モデル

こうやって、グラフを作ってみると

→　平均次数は少ないけど、
　　平均距離は、小さめなグラフが作れた

→　現実世界を良く反映してる？？

現実の世界を反映するような
数理モデルが欲しい

現実の世界のネットワークは？

感染以外に、どういう物（関係）が知りたいか？

- WWW　のリンク構造？
- インターネット（ルータ）？
- 人間関係？
- 単語同時出現？　（自然言語理解）

…

現実世界の（役に立つ）ネットワークの特徴

ネットワーク	ノード数	平均次数	
www	203549046	10.46	
インターネットルータ	150000	2.7	
論文引用	783339	8.57	
航空路線網	3880	9.7	
単語同時出現	478773	74.3	
俳優共演関係	449913	113.443	

「複雑ネットワークとその構造(矢久保考介著)」より

現実世界の（役に立つ）
ネットワークの特徴

- スケールフリー性
  → ノードの次数がベキ乗則

- スモールワールド性
  → 任意の2ノードの平均距離は短い

- クラスター性
  → あるノードと、隣接ノードの隣接ノードとの間は
  エッジがある確率が高い