情報リテラシ 第一

2023年度1Q 5c/6c（IL1） 木曜日

担当: 地引

TA: 増井

【重要】第１回小テスト（成績評価の対象になります）

- 課題の掲示先：

  ≫ 2023 年度情報リテラシ第一 5c/6c ページ

  → “2.課題” → “1.第１回小テスト”

  → “情報リテラシ第一 課題「メール」”

  → Google Forms による回答

- 提出：5/9（火）まで（注意をよく確認して下さい）

テーマ 3 の講義内容（1）

- 情報検索（本日は後者がテーマ）
  - **人為的な規則**に従い、**整理**が試みられている情報群
    - ≫ 多くの人に知見を提供できる情報 ⇒ 文献として整理されている。
    - ≫ 但し、整理されているとは言え、その蓄積量は膨大
    - ≫ 効率的な検索技術の仕組みを理解する必要がある。
  - 全体を対象とした整理規則が存在しない情報群
    - ≫ Web ページに掲載されている情報など
      ⇒ 情報同士の**局所的な関連性に基づく繋がり**がある。
    - ≫ **繋がり方の科学**を理解する必要がある。

テーマ 3 の講義内容（2）

- Web ブラウザの利用（Google Chrome を例として説明します）
- ホームページからの情報収集
- 情報検索の考え方（繋がり方の科学）

本テーマは、分量が多いので、
時間内にできなかった項目は、
次回以降で取り上げます。

Web ブラウザの利用

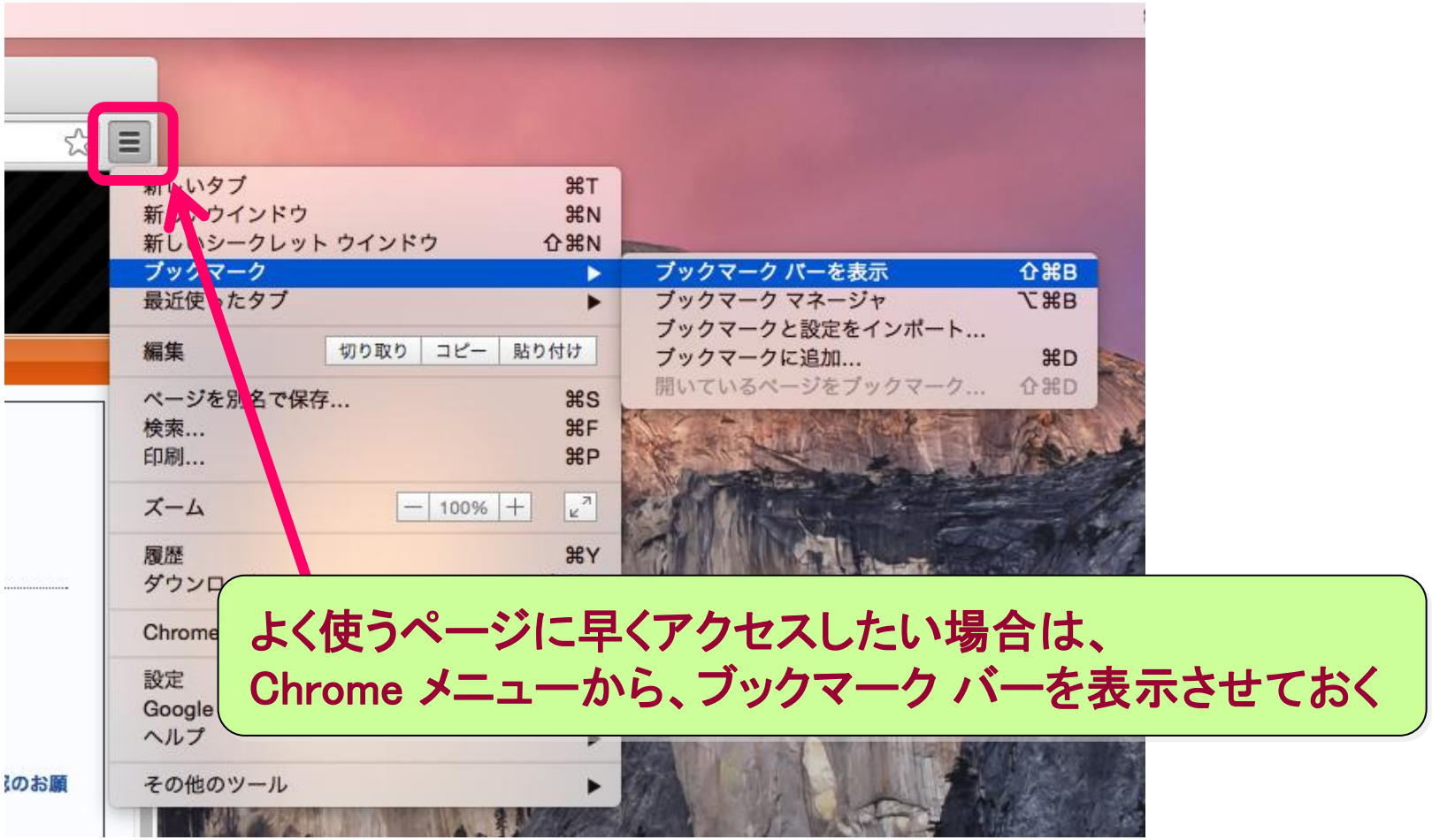
WWW について

- WWW： World Wide Web
- 全世界的な情報共有システム
- インターネットを介してアクセス
- ハイパー テキスト形式
- ブラウザで閲覧
- URL（Uniform Resource Locator）で場所を指定

Chrome の基本的な使い方

ブックマークの作成（1）

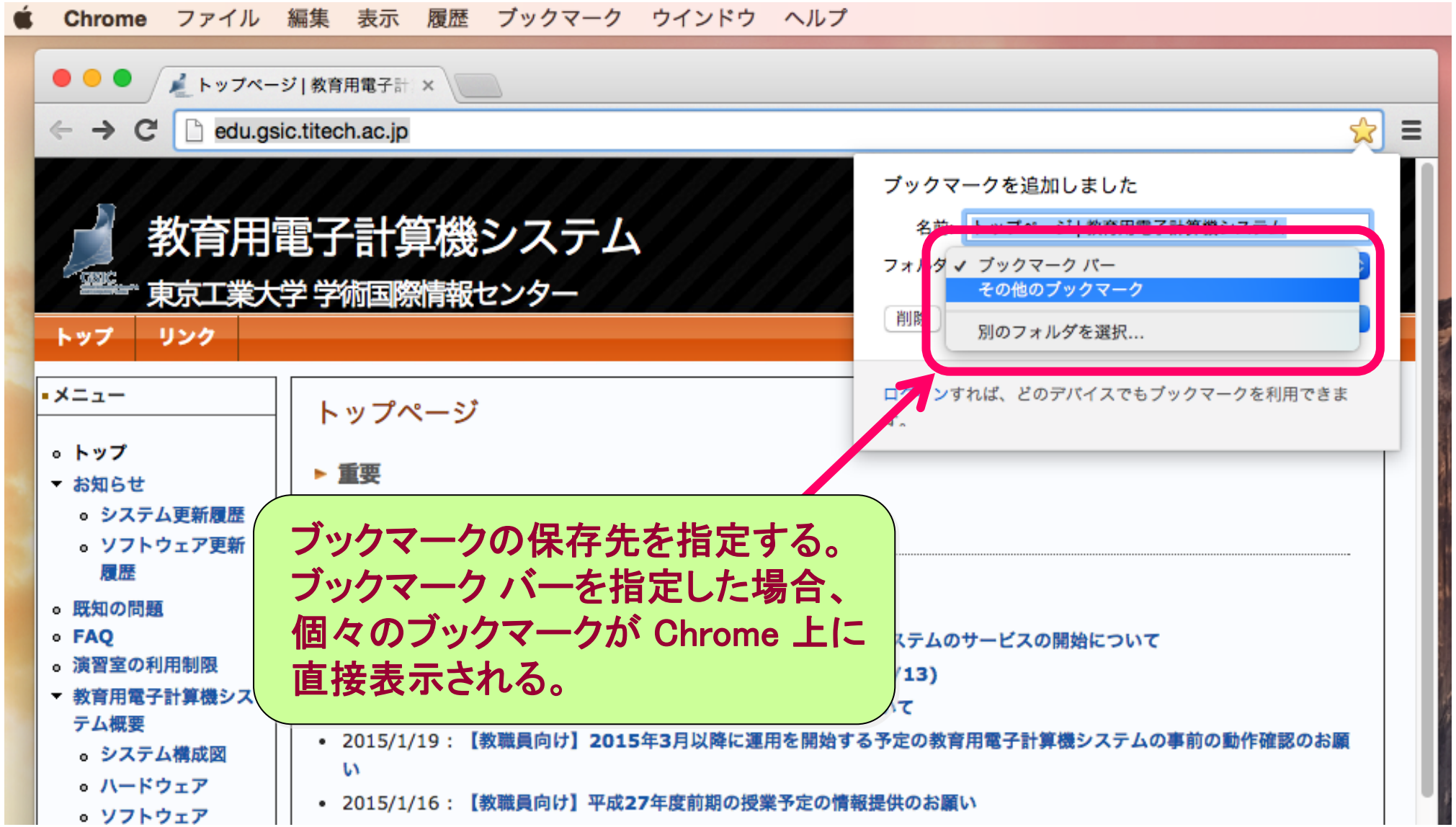
- **有益なページは覚えておくと大変便利**
  - Chrome には URL を手軽に覚えておく機能（**ブックマーク**）があります。



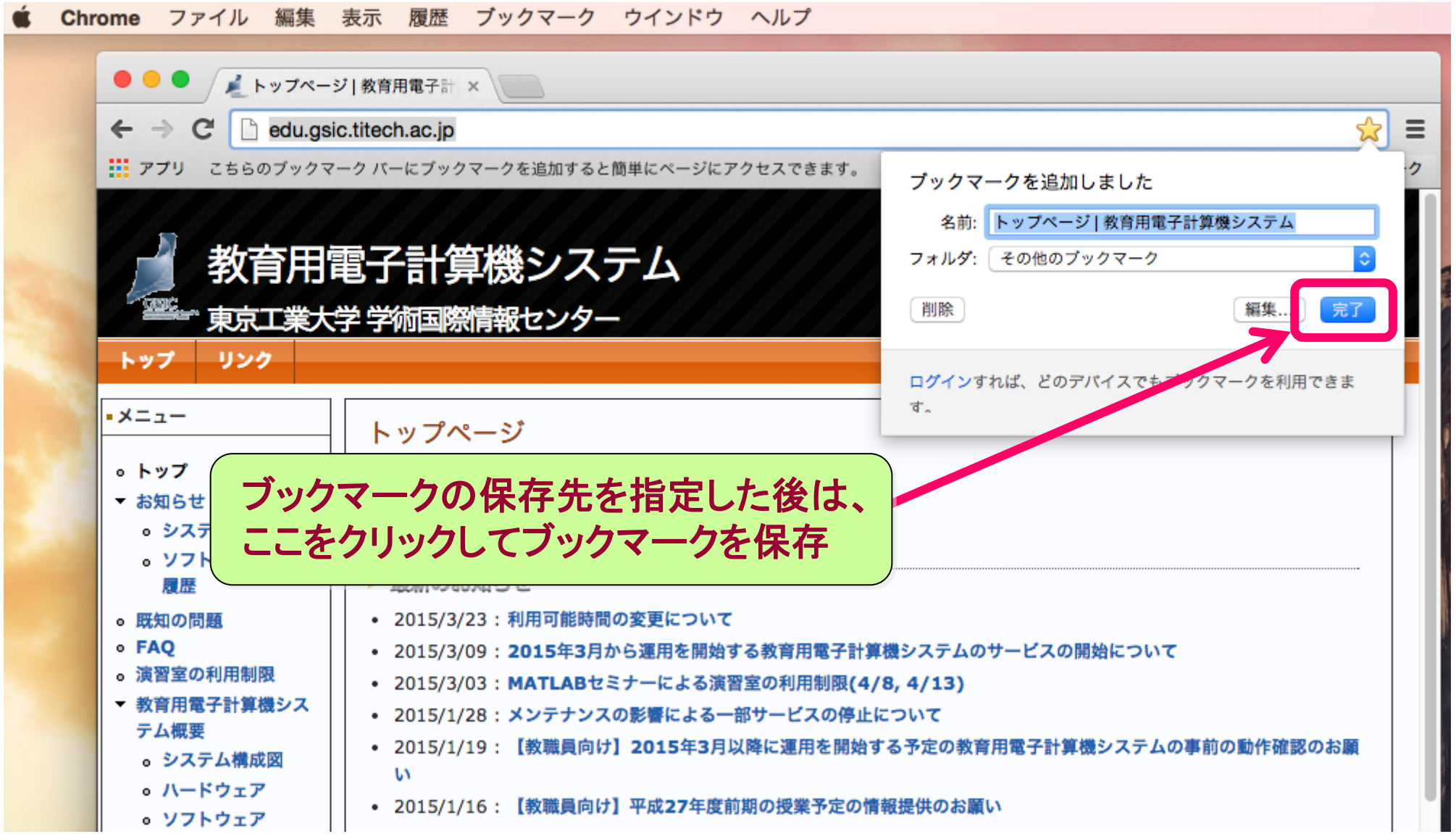
ブックマークの作成（2）


ブックマークに登録したいページを
表示させた状態でこのボタンをクリック
ブックマークの保存先を指定
するには、このボタンをクリック
ブックマークを追加しました
名前: トップページ | 教育用電子計算機システム
フォルダ: ブックマーク バー
削除
編集
完了
教育用電子計算機システム
東京工業大学 学術国際情報センター
トップ
リンク
メニュー
トップページ
重要
最新のお知らせ

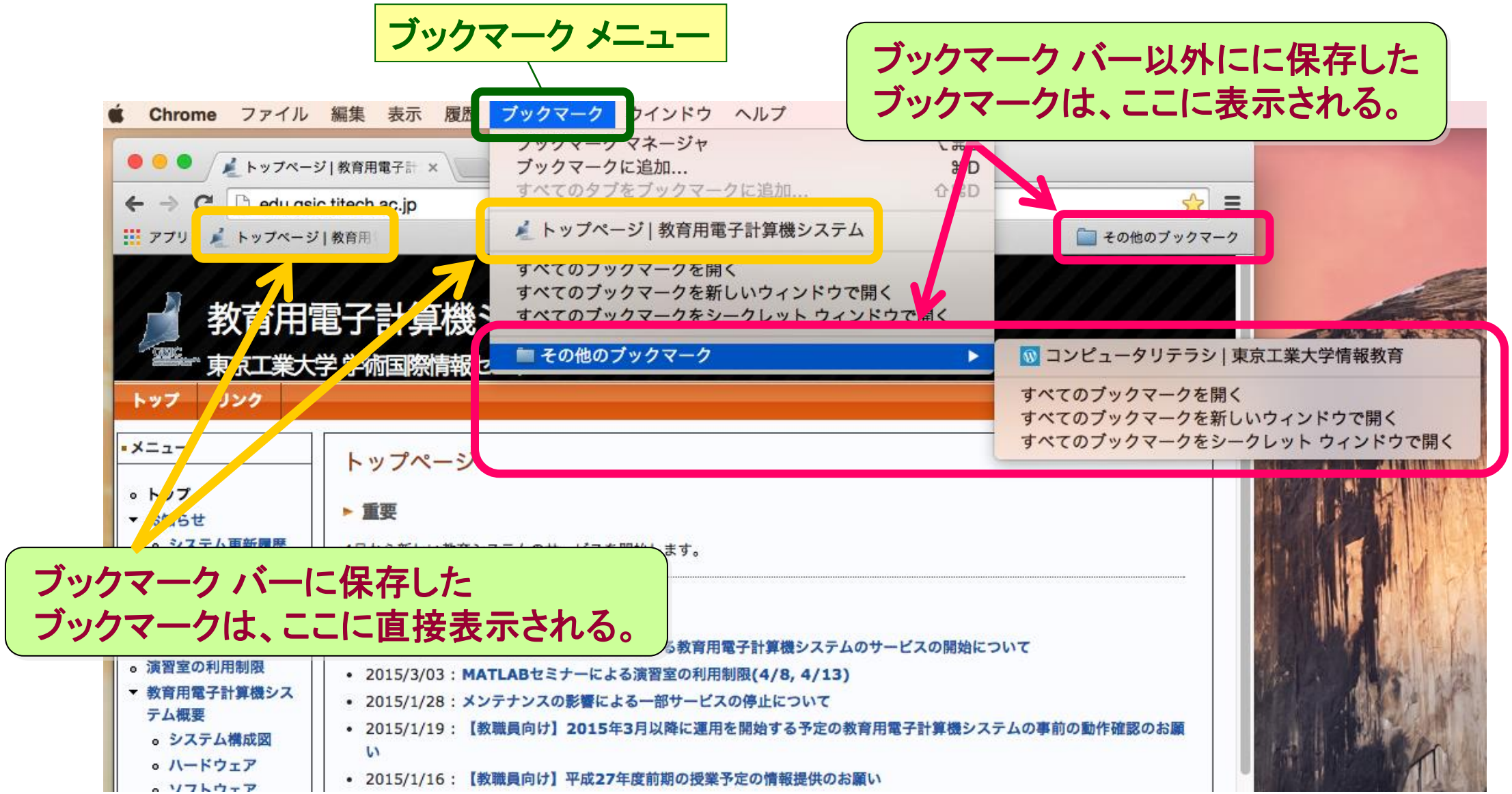
ブックマークの作成（3）


Chrome ファイル 編集 表示 履歴 ブックマーク ウインドウ ヘルプ
トップページ | 教育用電子計
edu.gsic.titech.ac.jp
教育用電子計算機システム
東京工業大学 学術国際情報センター
トップ
リンク
メニュー
トップ
お知らせ
システム更新履歴
ソフトウェア更新履歴
既知の問題
FAQ
演習室の利用制限
教育用電子計算機システム概要
システム構成図
ハードウェア
ソフトウェア
トップページ
重要
2015/1/19：【教職員向け】2015年3月以降に運用を開始する予定の教育用電子計算機システムの事前の動作確認のお願い
2015/1/16：【教職員向け】平成27年度前期の授業予定の情報提供のお願い
ブックマークを追加しました
名前
フォルダ
ブックマーク バー
その他のブックマーク
別のフォルダを選択...
削除
ログインすれば、どのデバイスでもブックマークを利用できます。
ブックマークの保存先を指定する。
ブックマーク バーを指定した場合、
個々のブックマークが Chrome 上に
直接表示される。

ブックマークの作成（4）

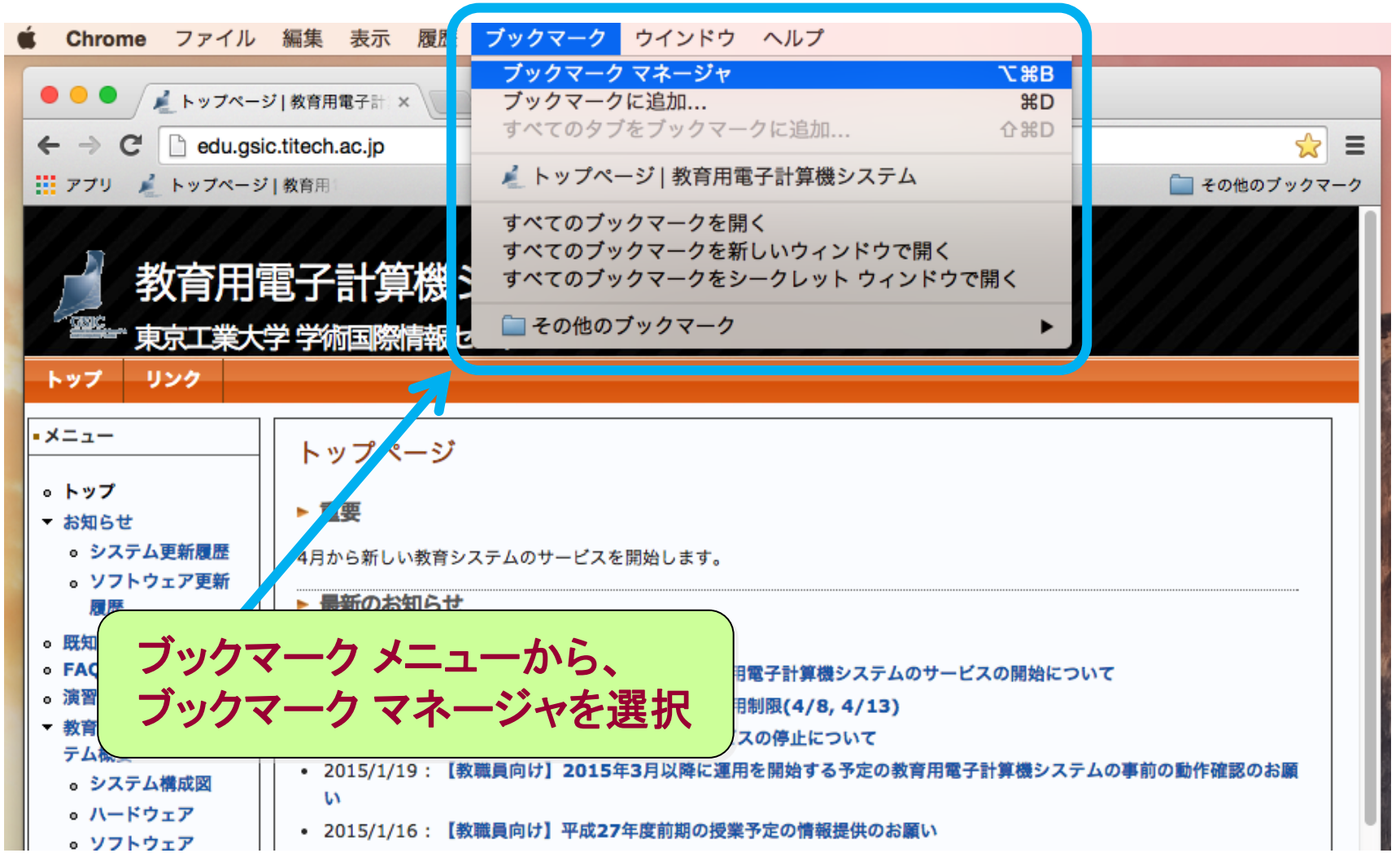

Chrome ファイル 編集 表示 履歴 ブックマーク ウインドウ ヘルプ
トップページ | 教育用電子計
edu.gsic.titech.ac.jp
アプリ こちらのブックマーク バーにブックマークを追加すると簡単にページにアクセスできます。
ブックマークを追加しました
名前: トップページ | 教育用電子計算機システム
フォルダ: その他のブックマーク
削除
編集...
完了
ログインすれば、どのデバイスでもブックマークを利用できます。
教育用電子計算機システム
東京工業大学 学術国際情報センター
トップ
リンク
メニュー
トップ
お知らせ
既知の問題
FAQ
演習室の利用制限
教育用電子計算機システム概要
システム構成図
ハードウェア
ソフトウェア
トップページ
ブックマークの保存先を指定した後は、
ここをクリックしてブックマークを保存
2015/3/23：利用可能時間の変更について
2015/3/09：2015年3月から運用を開始する教育用電子計算機システムのサービスの開始について
2015/3/03：MATLABセミナーによる演習室の利用制限(4/8, 4/13)
2015/1/28：メンテナンスの影響による一部サービスの停止について
2015/1/19：【教職員向け】2015年3月以降に運用を開始する予定の教育用電子計算機システムの事前の動作確認のお願い
2015/1/16：【教職員向け】平成27年度前期の授業予定の情報提供のお願い

ブックマークの作成（5）

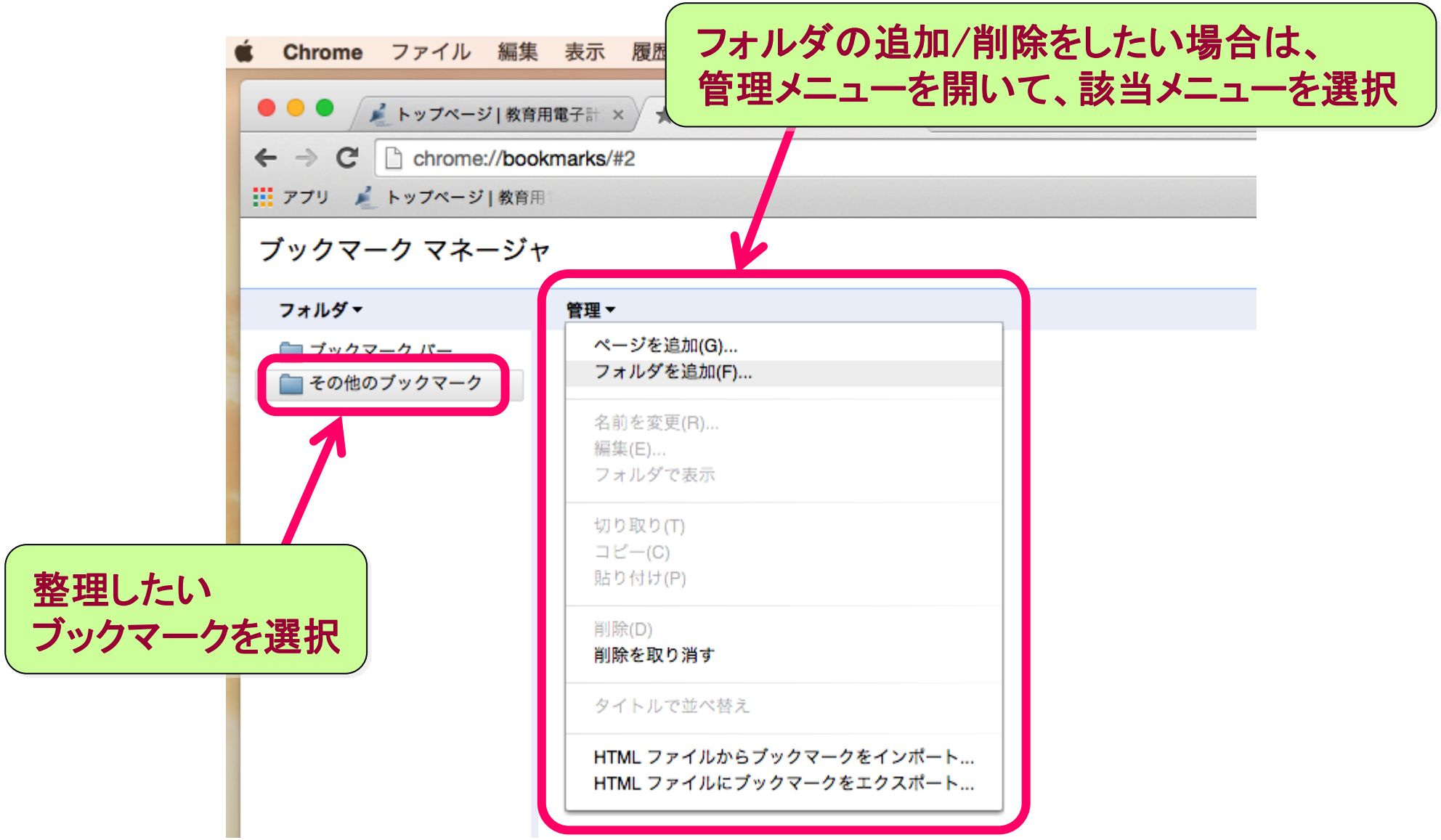

ブックマーク メニュー
ブックマーク バー以外にに保存した
ブックマークは、ここに表示される。
ブックマーク バーに保存した
ブックマークは、ここに直接表示される。
Chrome ファイル 編集 表示 履歴 ブックマーク ウインドウ ヘルプ
ブックマークに追加...
トップページ | 教育用電子計算機システム
すべてのブックマークを開く
すべてのブックマークを新しいウィンドウで開く
その他のブックマーク
コンピュータリテラシ | 東京工業大学情報教育
すべてのブックマークを開く
すべてのブックマークを新しいウィンドウで開く
すべてのブックマークをシークレット ウィンドウで開く
トップページ | 教育用電子計算機システム
その他のブックマーク
トップ
メニュー
重要
演習室の利用制限
教育用電子計算機システム概要
システム構成図
ハードウェア

ブックマークの整理（1）

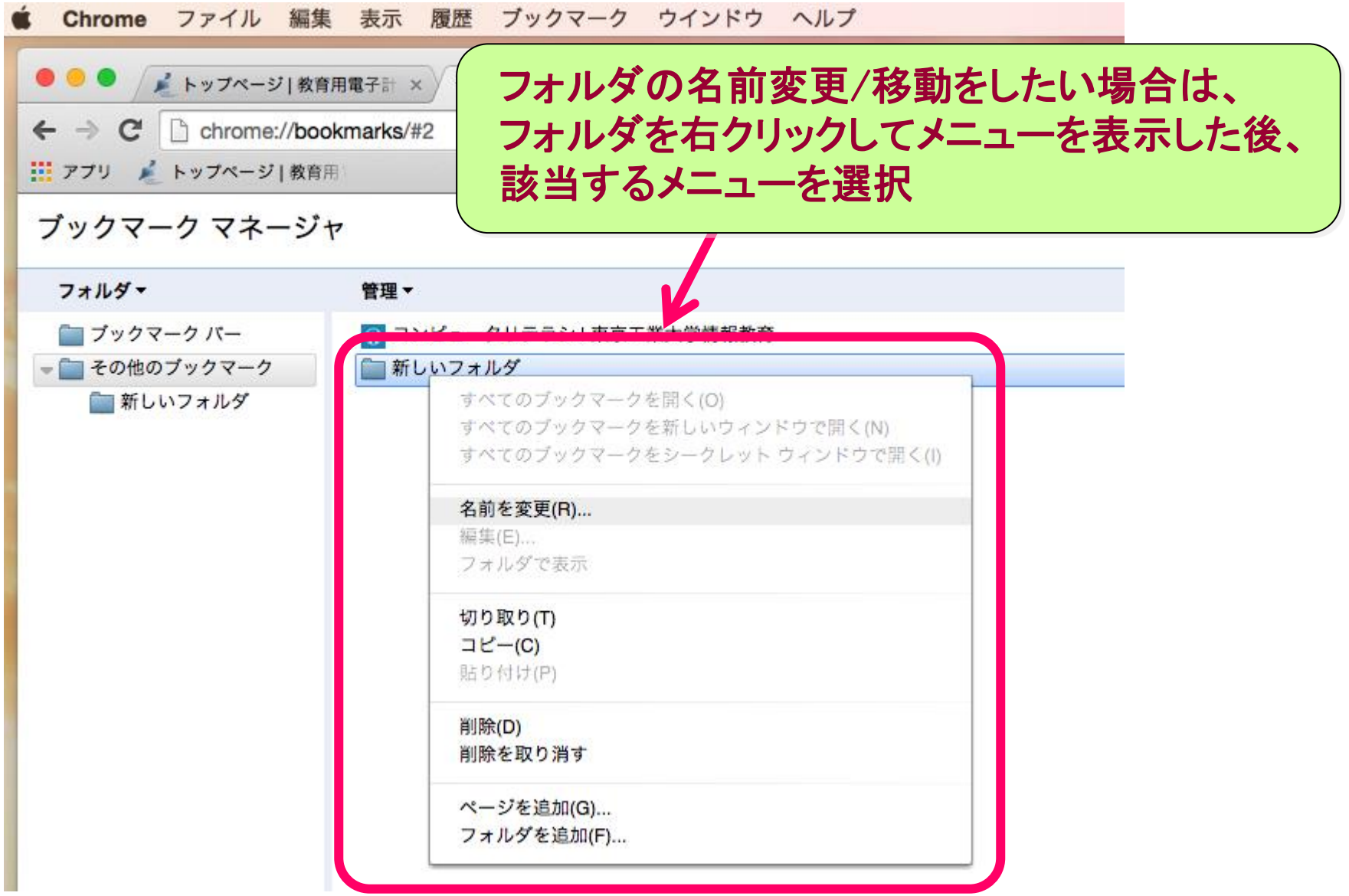
- メールと同様、ブックマークも整理が重要です。



ブックマークの整理（2）

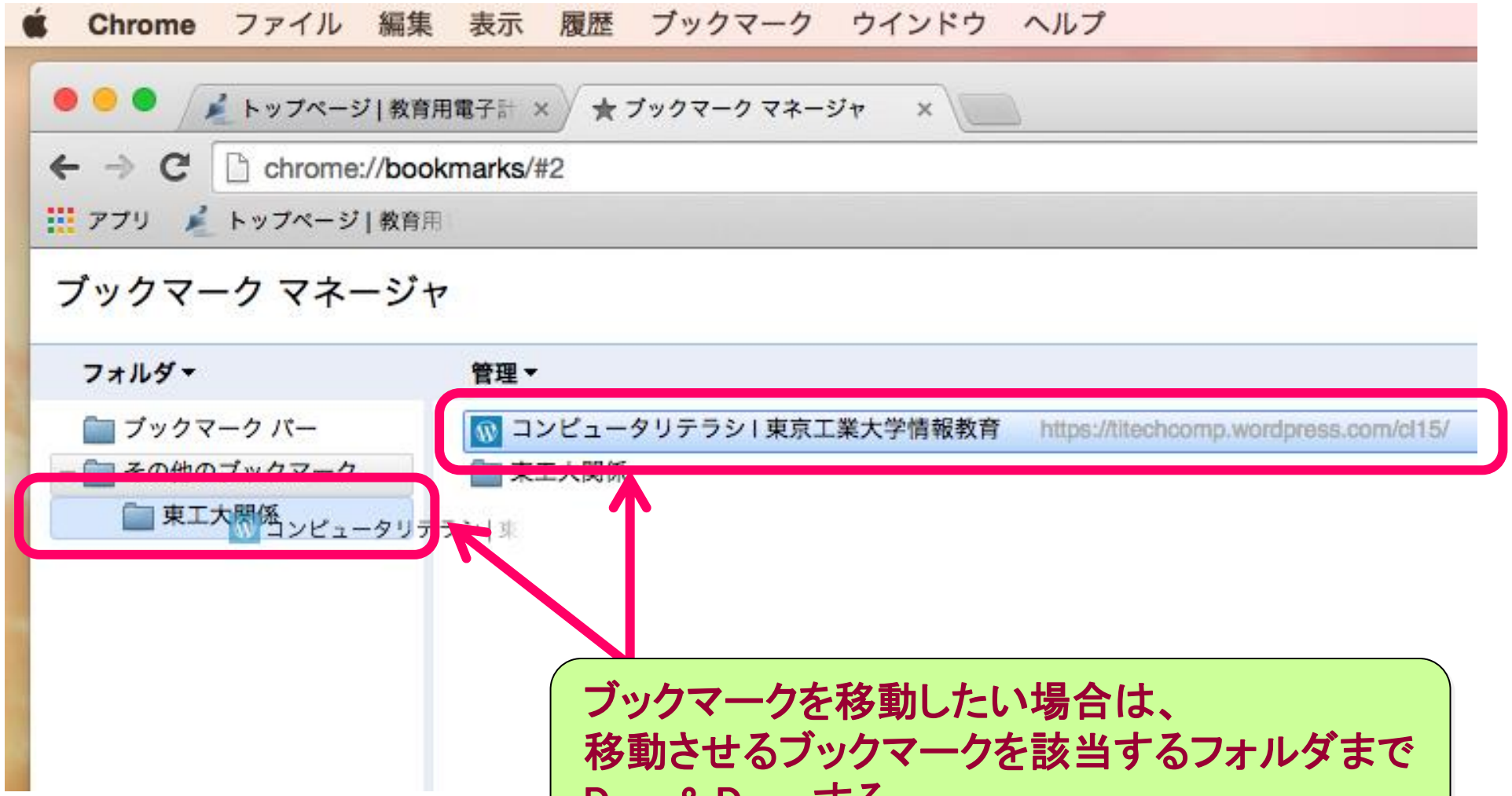

フォルダの追加/削除をしたい場合は、
管理メニューを開いて、該当メニューを選択
Chrome ファイル 編集 表示 履歴
トップページ | 教育用電子計
chrome://bookmarks/#2
アプリ
トップページ | 教育用
ブックマーク マネージャ
フォルダ
その他のブックマーク
管理
ページを追加(G)...
フォルダを追加(F)...
名前を変更(R)...
編集(E)...
フォルダで表示
切り取り(T)
コピー(C)
貼り付け(P)
削除(D)
削除を取り消す
タイトルで並べ替え
HTML ファイルからブックマークをインポート...
HTML ファイルにブックマークをエクスポート...
整理したい
ブックマークを選択

ブックマークの整理（3）



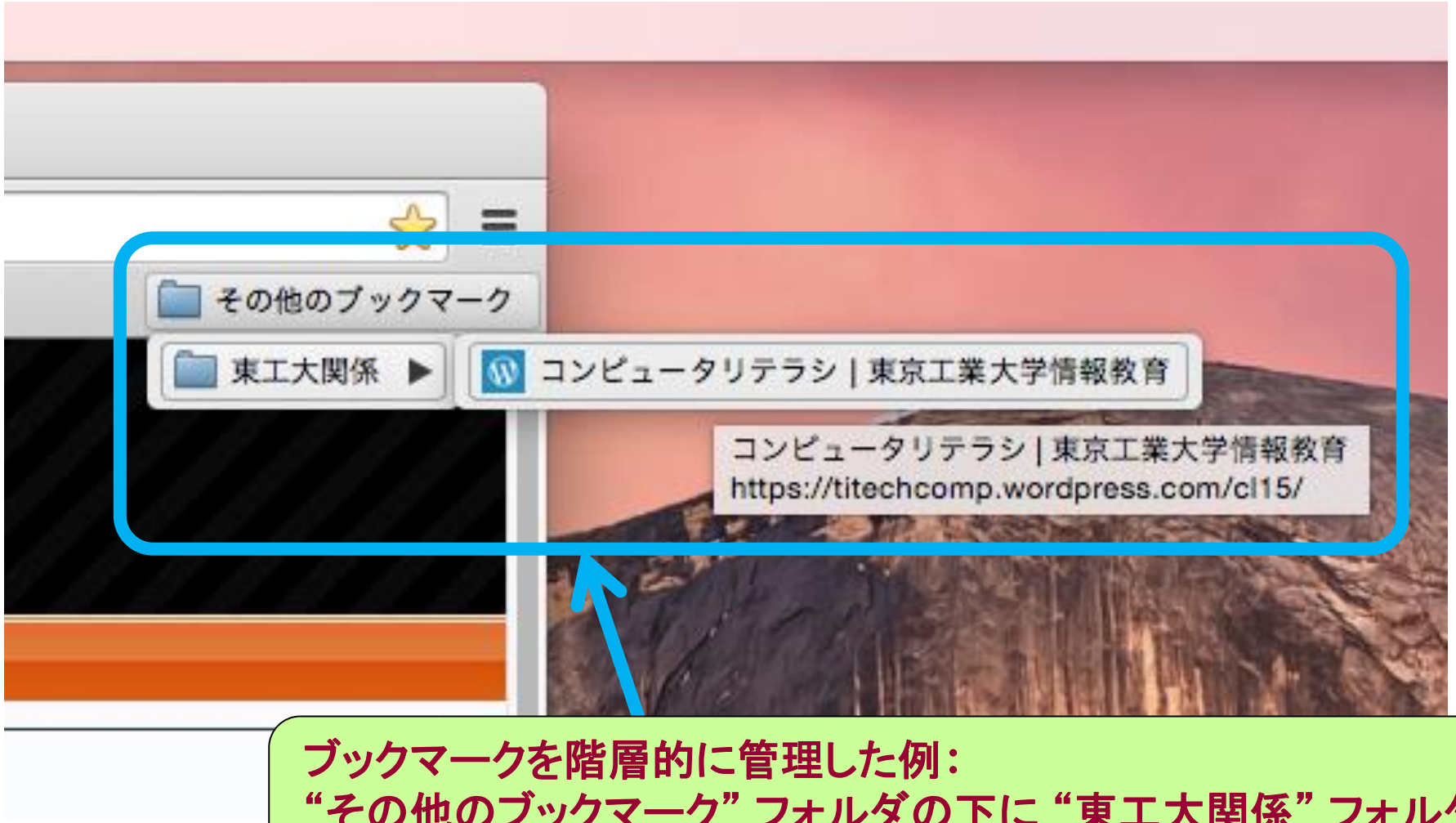
フォルダの名前変更/移動をしたい場合は、フォルダを右クリックしてメニューを表示した後、該当するメニューを選択

ブックマークの整理（4）



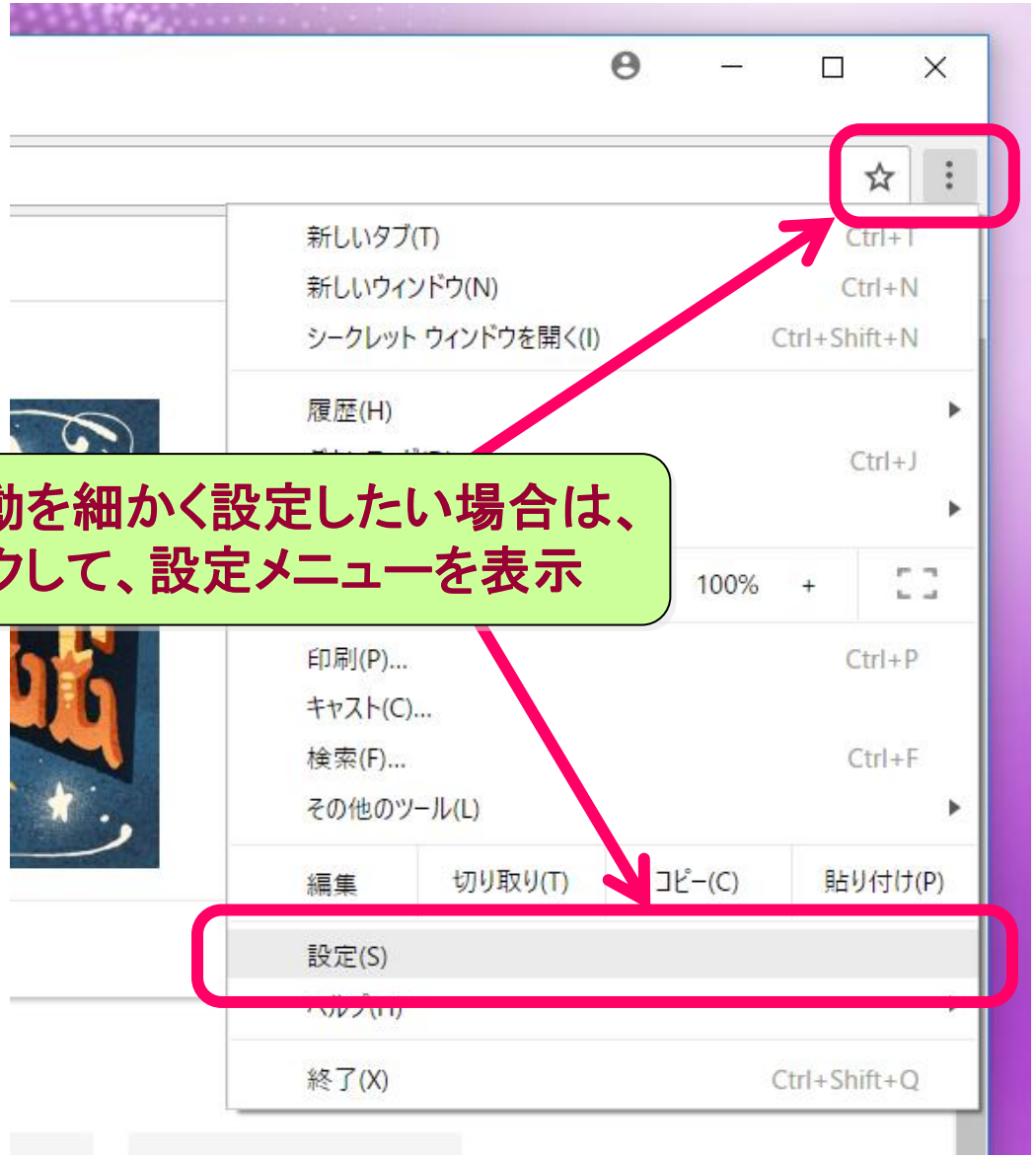
ブックマークを移動したい場合は、
移動させるブックマークを該当するフォルダまで
Drag & Drop する。

ブックマークの整理（5）



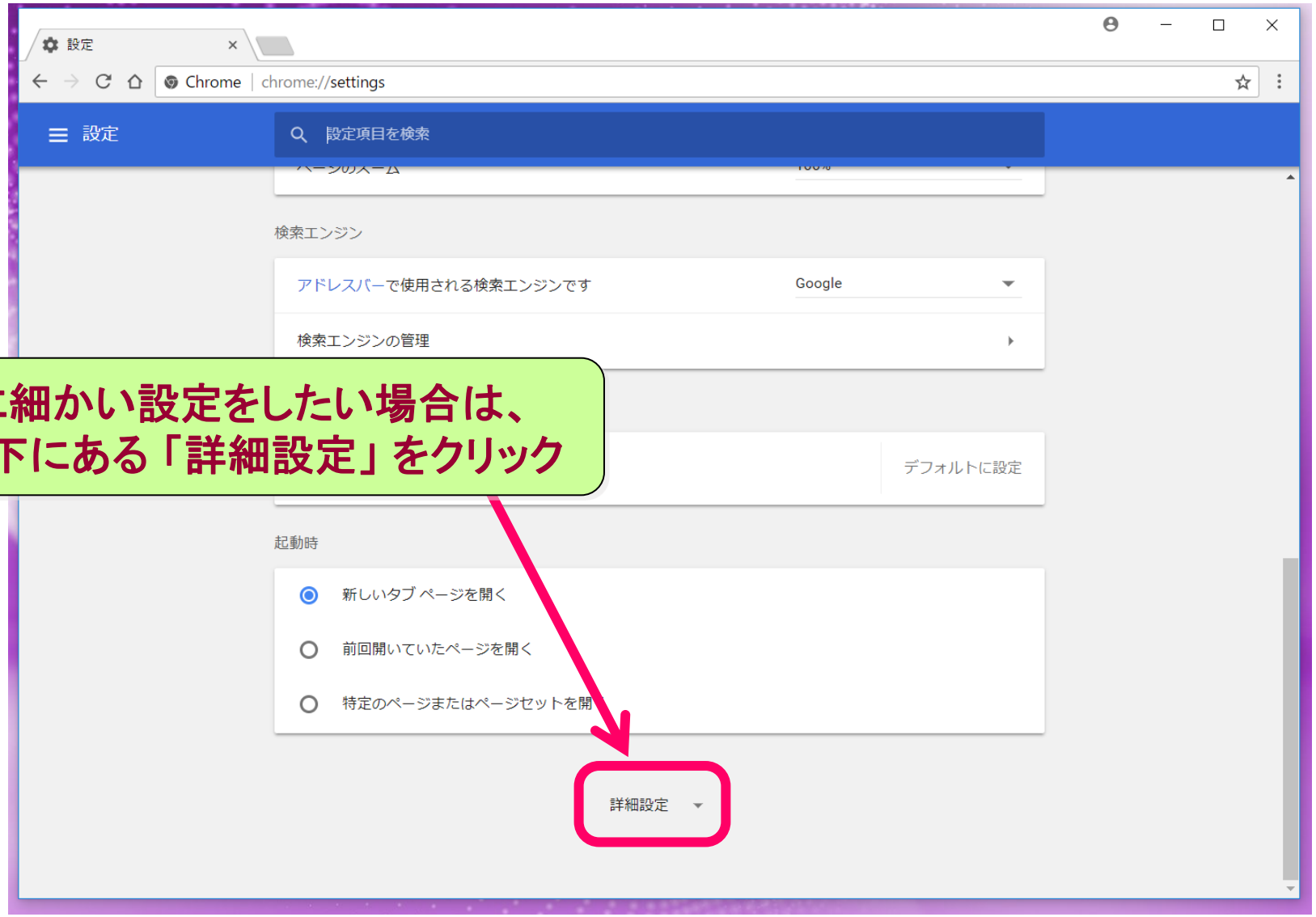
ブックマークを階層的に管理した例：
“その他のブックマーク” フォルダの下に “東工大関係” フォルダを作成し、その中へ “コンピュータリテラシ” ブックマークを格納している。

その他の設定（1）



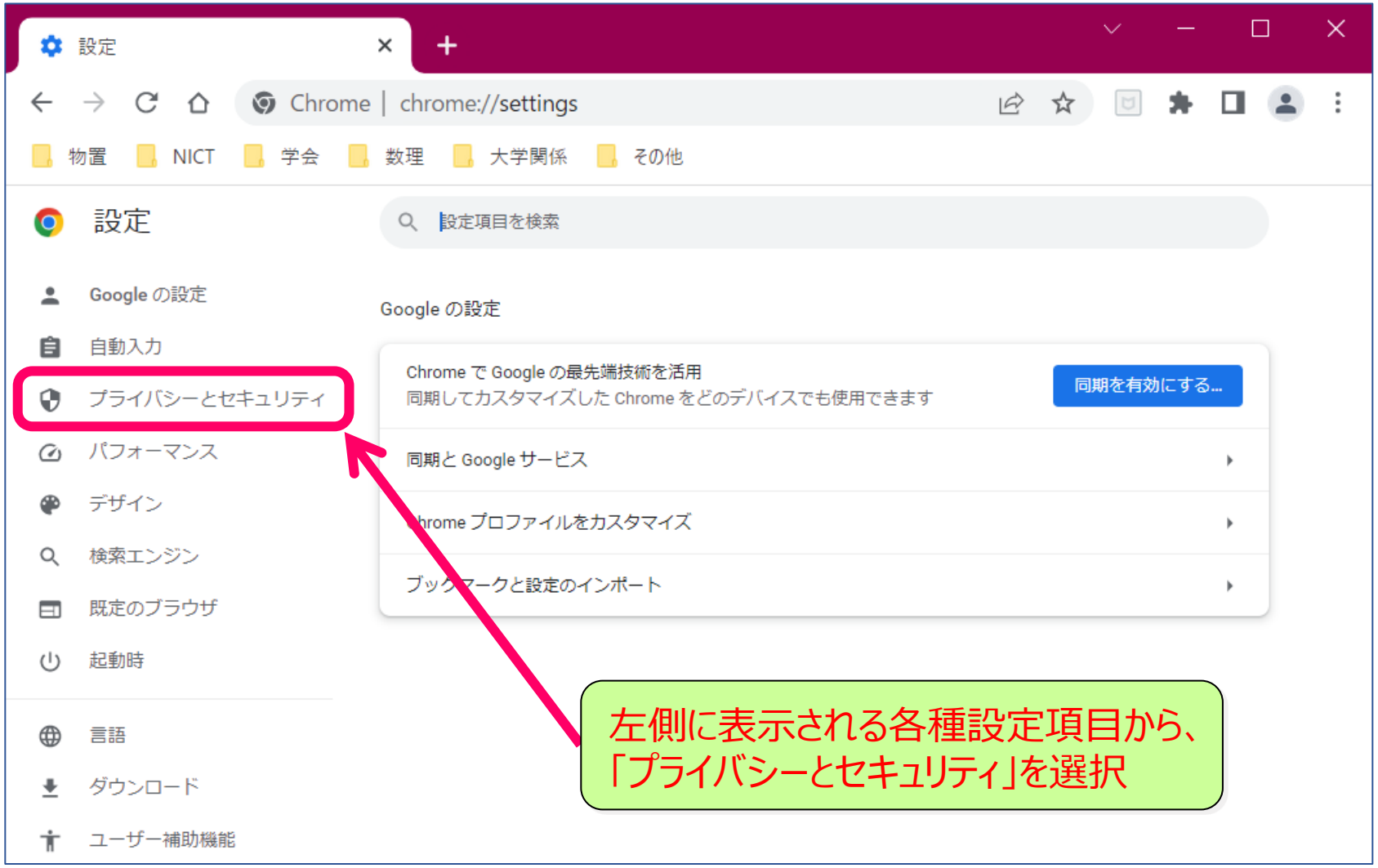
Chrome の挙動を細かく設定したい場合は、
これらをクリックして、設定メニューを表示

その他の設定（2）



さらに細かい設定をしたい場合は、
一番下にある「詳細設定」をクリック

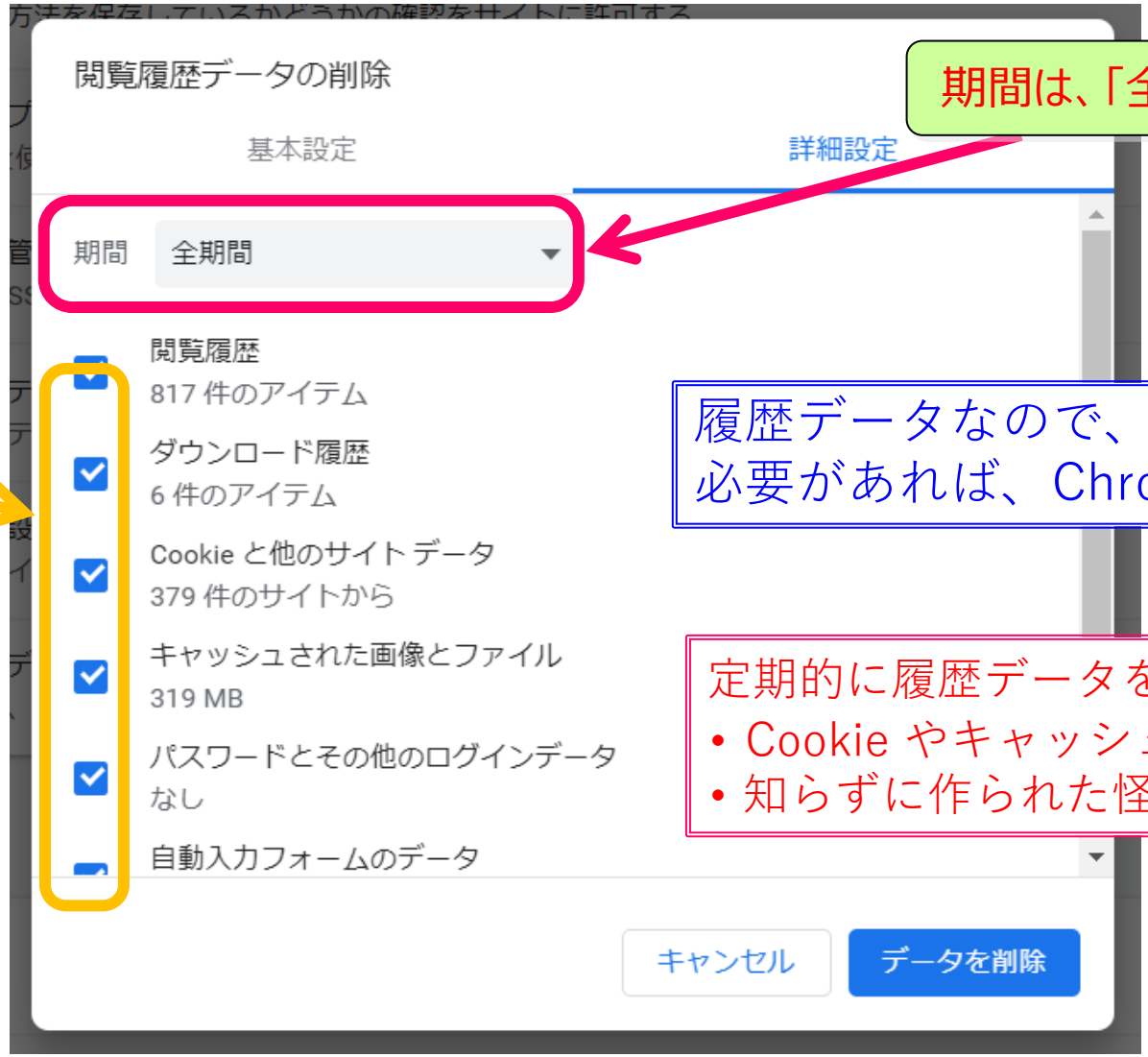
定期的な閲覧履歴データの削除（1）



定期的な閲覧履歴データの削除（2）

詳細設定より、「閲覧履歴データの削除」を探して選択

定期的な閲覧履歴データの削除（3）



期間は、「全期間」を選択

全項目を選択

履歴データなので、全て消しても大丈夫。
必要があれば、Chrome が取りに行きます。

定期的に履歴データを削除する理由：
- Cookie やキャッシュなどに潜むウィルスを掃除
- 知らずに作られた怪しいデータを掃除

Cookie について

- Web サーバが、閲覧者の PC へ作成するファイル
  - 内容: 会員番号, パスワード, 買い物カゴ, クレカ番号など
  - 2 回目以降の接続で便宜を図るためだが、危険でもある。

- Cookie の危険性
  - Web ページを介して買い物をする機会は多い一方で、毎回煩雑な手続きをするのは面倒 ⇒ **Cookie には、致命的な情報が含まれる場合が多い。**
  - 扱う情報は致命的だが、セキュリティは杜撰な場合が多い。
    - ≫ Web サーバ側, ブラウザ側の両方で対策をしていく必要がある。
    - ≫ しかし、面倒が増えるのは‥**（ここにも安全と便宜とのバランス）**
  - **定期的な Cookie の削除**
    - ≫ 脆弱な or 余分な or 古い Cookie を淘汰していくという意味

参考：セキュリティ ホールを突いて情報漏洩させる方法（1）

例えば、自分が自分のブラウザ上でブログを作成すると（サイトに送ると）、そのブログは自分のブラウザに表示されますね。文章ではなくスクリプトだったとしても、そのまま表示されてしまうわけです（入力できるHPは、全て事情は同じです）。

(1). ユーザが悪意ある Web サイトを閲覧する。

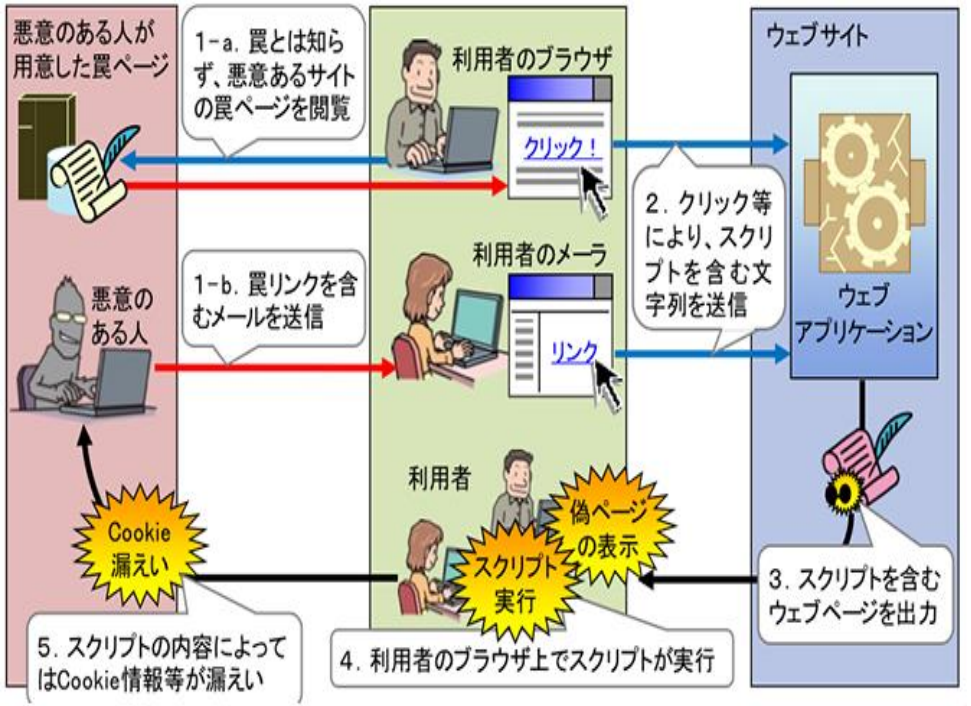
(2). Web ページには、悪意あるスクリプト（プログラム）が埋め込まれている。

(3). 例えば、そこには脆弱ブログ サイトへのリンクがあり、URL にはブログへ入力する文章の代わりにスクリプトが付いている。リンクを踏むと、これがブログへの入力データとして脆弱サイトへ送られる。

(4). この脆弱サイトには、ブラウザからのアクセス時にスクリプトが送られたとしても、それを排除しないという欠陥がある場合、そのまま受け入れてしまう。（例えば、ブラウザへの悪意ある要求など）。

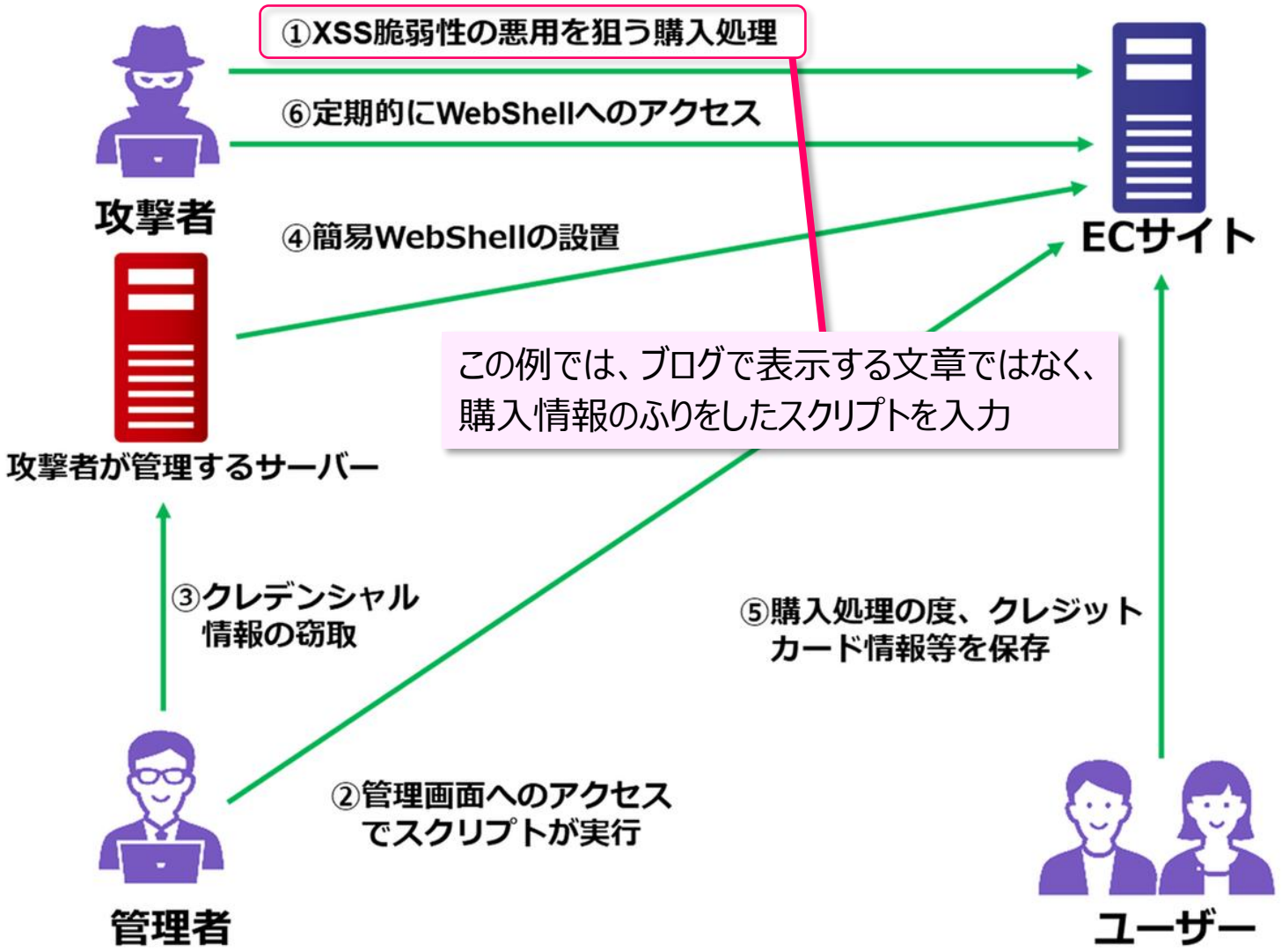
(5). 一般にブラウザは、Web サイトから HTML 言語でページ構成を受け取り、これを実行して画面を作成（表示）している。

(6). 脆弱サイトから送られた構成情報に悪意のスクリプトが含まれていた場合、画面作成時に実行されてしまう。



図出展：IPA 安全なウェブサイトの作り方

参考：セキュリティ ホールを突いて情報漏洩させる方法（2）



ECサイト:

電子商取引をするサイト全般。

XSS:

正式名称は、クロスサイト スクリプティング (Cross-Site Scripting)。

全スライドのように、入力欄のある HP を利用して攻撃する手法。

フィッシングではない (騙しているわけではなく、システムの穴を突いている) ので、対策が難しい。左の例では、EC サイトの作成業者 or サイト構築ツールに問題があり、直接の当事者 (管理者/ユーザー) は気付きにくい (責任者が出て来ない)。

古くからある手法だが、ECサイトの増加に伴い、最近再び流行りだして来た。

(図出展: JPCERT/CC Eyes)

ホームページからの情報収集

インターネットを用いた情報収集

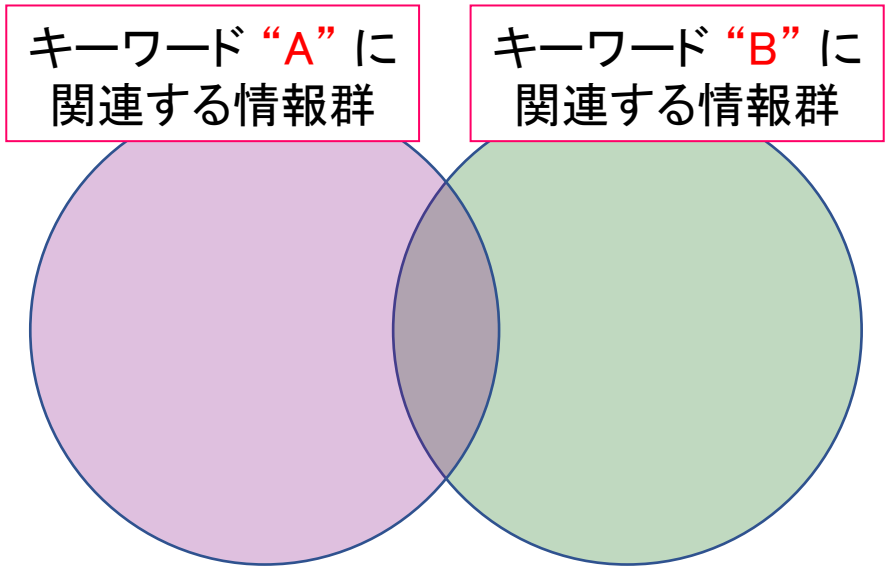
- 様々な組織が様々な情報を発信
  - 組織や情報の有益性は千差万別
  - どうすれば有益な情報を収集できるか？
  - 有益な URL を覚えておくというのも一つの手だけど‥
    ≫**一番最初**は、どうやって見つけたらよいの？

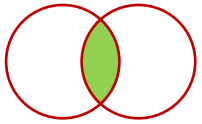
- 検索エンジンの利用
  - 自分が欲しい情報に関連のありそうな URL を教えてくれる。
  - 検索サイトの URL さえ覚えておけば、何とかなる。
    ≫とても便利／効率的

検索サイトの歴史

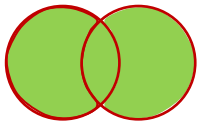
- https://www.yahoo.co.jp
  - 基本的に人手による情報収集
  - それなりに厳選された情報
- https://www.goo.ne.jp
  - 情報収集プログラムによる自動収集
  - 網羅的（検索を絞る工夫を）
  - 現在では Google に押されポータル化（Yahoo 等と同じ）
- https://www.google.co.jp
  - 情報検索に特化（お勧め情報などは表示しない）
  - goo よりは情報の質が高い
  - 検索エンジンのデファクト スタンダードへ

キーワードによる情報検索のイメージ

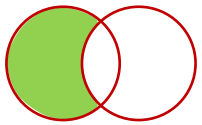




AとBの両方に関連する情報（A∩B）



AとBのどちらかに関連する情報（A∪B）



A だけに関連する情報（A∩$\overline{B}$）

キーワードとの関わり方に
基いた情報の３分類

どの集合に属する情報を
取り出すか ≒ 情報検索

効率的な検索の基本（A∩B）

あなたは、今度友人の結婚式に呼ばれることになりました。
しかし、御祝儀を幾ら持っていくべきなのか分かりません。
検索サイトで、どのような情報検索をすればよいでしょうか？

- 「結婚」や「御祝儀」といった単一の単語では欲しい情報を絞れない。
- 欲しい情報は、「結婚」の「御祝儀」の「相場」なので、「結婚」，「御祝儀」，「相場」の三つを同時に含むページを検索
  - これを**“And 検索”**と呼びます。
- Google では、スペースを空けて複数の単語を並べる。
  - 「結婚　御祝儀　相場」と指定

効率的な検索の基本（ＡＵＢ）

あなたは、あるニュース〇〇が本当の出来事かフェイク情報かを確かめたいと考えています。
検索サイトで、どのような情報検索をすればよいでしょうか？

- あるニュース〇〇を、信用のあるニュース サイトだけから検索したい（フェイク サイトは除きたい）。
- 日経新聞，CNN，時事通信の**何れか**に掲示されているニュース〇〇を検索し、もし掲示されていれば、この順で表示したい。
  - これを**“Or 検索”**と呼びます。
- Or 検索のキーワードを「｜」で連結する。
  -「〇〇　日経新聞｜CNN｜時事通信」と指定

効率的な検索の基本（A∩$\overline{B}$）

あなたは今度、自転車を購入することにしました。
せっかくなので、錆びにくい自転車を選びたいと考えています。
検索サイトで、どのような情報検索をすればよいでしょうか？

- 「自転車」+「錆びにくい」では、錆び取りや錆びさせない方法に関するページが大量に表示されてしまい、自分に必要な情報を絞れない。
- 販売情報だけに絞り、「〇〇の方法」などは**排除したい。**
  - これを**“フィルタリング”**と呼びます。
- Google では、排除したい単語の前に「**－**」を付ける。
  - 「自転車　錆びにくい　**－**方法」と指定

Google を利用する際の注意（1）

渋谷公会堂で開かれるクラシックコンサートの情報を検索しようと思い、
「渋谷公会堂　クラシックコンサート」と指定したら、
関係なさそうなサイトも数多く表示されてしまった･･。

- Google では、And 検索の場合に単語を勝手に区切ってしまう。
  - 「渋谷公会堂　クラシックコンサート」と検索すると、
    「渋谷　公会堂　クラシック　コンサート」で検索されてしまう。
- 区切って欲しくない単語は、“ ” で囲む。
  - 「“渋谷公会堂” “クラシックコンサート”」と指定

Google を利用する際の注意（2）

あなたは、人文科目のレポートでアメリカ大統領について調べることになりました。
検索サイトで、どのような情報検索をすればよいでしょうか？

- アメリカ大統領を英語で表記すると「The President」になるが、Google では、そのまま入力しても有益な情報が表示されない。
- Google では、例えば「The」のような**非常に頻度の高い単語は検索の対象にしない。**
- 頻出単語の前に（**半角**の）「**＋**」を付ける。
  - 「**＋**The　President」と指定

Google を利用する際の注意（3）

日本映画について調べるため、「日本映画」と入力したら、
日本映画と関係のないホームページまで表示されてしまった・・。

- Google では、本文だけではなく、URL やタイトルに指定された単語を含むページや、該当するページがリンクしている先のページなども表示される。
- 本文だけを検索したい場合は、単語の前に「intext：」を付ける。
  - 「intext：日本映画」と指定
- その他の特別構文機能
  - site：　　指定されたサイト内だけを検索
  - inurl：　　キーワードを URL に含むページを検索
  - intitle：　同、タイトルに含むページを検索
  - filetype：指定された種類のファイルを検索

現在では、これらは検索オプションのメニューより設定できます。

学術成果の検索

- **大前提：研究活動には経費が掛かる。**
  - 研究成果をまとめた論文はタダではない。
  - 東工大は、学術団体に論文の代価を支払っているが、その取り扱いを疑われたら、学術団体から容赦なく切られます。

- **大前提：人類の資産と言える研究成果は公開すべき。**

- Google Scholar＞ http://scholar.google.co.jp/
  - 有料論文でも、事情により（例えば、著者が意見を求めるために公開した草稿など）無料で公開している版があれば、検索してくれる。

- arXiv＞ http://arXiv.org/　（“アーカイブ”と呼ばれています）
  - 草稿を議論するために論文を公開できる著名サイト
  - 物理，数学，計算機科学が中心

繋がり方の科学

- 情報はどのように繋がっているの？
- どうやれば効率よく辿れるの？

検索サイトの仕組み（１）

- 検索は通常 １ 秒以内で完了。
- 世界中のホームページを見て回るのに １ 秒？
- 検索サイトは、どのようにして情報を収集しているか？

- 検索のたびにホームページを駆け回るわけではない。
- 常に世界中のホームページを見て回り、保存している。
- 検索を要求されると、保存している情報を検索する。

検索サイトの仕組み（2）

- 保存してある数十兆ページの情報の検索に 1 秒？
  - Google では、数十兆ページの保存に PC 数百万台を利用
- **10 万件マッチしても、欲しい情報が上位にあるのはなぜ？**

- 検索キーワードの出現数が多いページが良いページ？
  - 最初はみんなそう考えた ⇒ 結果は全然ダメ
- 良いページを定量的に表現できる指標とは何か？
  - **ヒント: 数理計算科学 ⇒ 論理モデルと人間の振舞いを橋渡し**
  - 美味しい食事を安く食べさせてくれるお店は、みんな行きたいよね。
    じゃあ、どうやって探す？ 美味しいかどうかは、どう判断する？

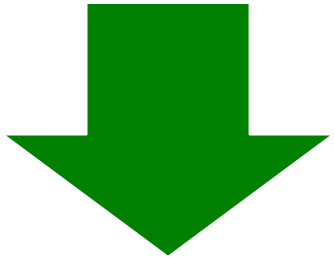
電脳空間における真実とは何か？

- 世界で一番胡散臭いのは亀井静香？
  - 2005 年に行なわれた衆議院選挙の際、検索サイトで「胡散臭い」を検索すると、亀井静香代議士のページがトップであった。
  - しかし、**亀井氏のページ内には、「胡散臭い」という単語は一言も使われていない。**
  - インターネットにおける情報検索の隙を突かれたらしい。
  - 何も知らなければデマや情報誘導されてしまう？

- インターネットにおける情報の繋がり方について知る必要がありそう。
  - **情報の 3 分類だけで、有益な情報を検索できるのか…**

情報検索が簡単な場合

- サッカーワールドカップで最も多い点を取った選手は誰でしょうか。
  - 複数の大会に参加した選手は、それらを全て合計してね。



- **キーワード or テーマが明確**（サッカー） & **結果も一意**に定まる（最も多い得点者）。
- 検索のキーワードさえ工夫すれば、検索に成功する確率が高い。

数理的なモデルを作る場合、このようなシンプル化は重要（〇〇は一定と仮定するとか）

情報検索が難しい場合

- 複数のテーマに関連した情報 or 複数の結果を比較する場合
  - 今度、東工大が町田へ移転することになり、現在の跡地に総合病院を建設する計画が検討されています。東工大の跡地に建設する規模の総合病院について、総建設費とその内訳を見積もって下さい（適切な病院の規模/設置診療科等の調査とは？ → **キーワード検索**だけで**体系化された情報を構築**できるのか？）。

- 情報と情報の繋がり方（リンク）を調べる必要あり。
  - ネットワークの科学を勉強する必要がある。

どこからモデル化すれば良いだろうか？

より良い情報に辿り着くには（１）

- これまでの失敗
  - キー ワードを元に、ページ内の文脈を様々な手法で解釈してみたが、そのキー ワードにとって、ページを有益な順に並べることはできなかった。
- 検索エンジンも突き詰めれば、人間と同じことをするしかない。
  - リンクを辿って情報を集める。
- ホーム ページの世界で、あるキー ワードに対する有益な情報は、どのような扱いをされているだろうか？
  - 一つの予想：

    様々なページからそのキーワードを辿った時、どこからでも近い所（リンクを何回も経由しない所）にあるのでは？

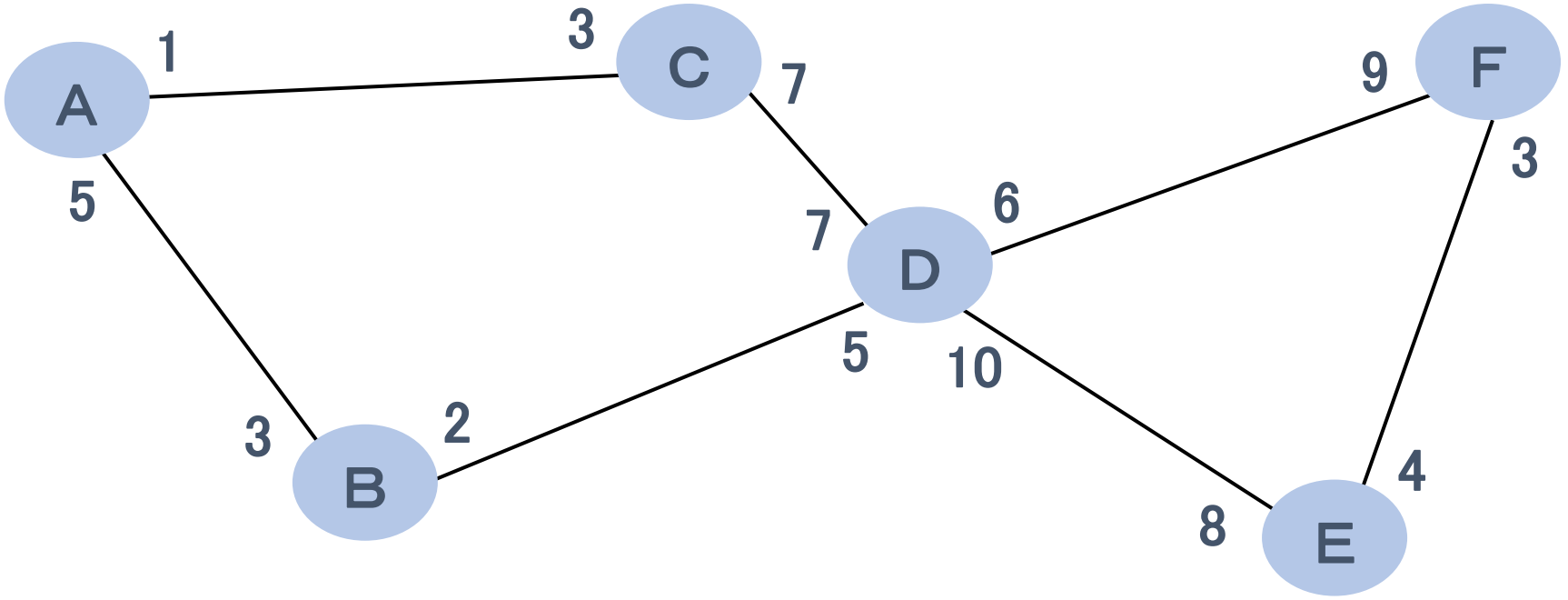
より良い情報に辿り着くには（2）

- 余計な遠回りをしないで辿り着ける方法はないか？
  - 余計な所を回る ⇒ 怪しいモノ，雑音が挟まる可能性大
    ≫調査/考察のスタートとしては自然
  - ゴールは分かっている（例えば、〇〇についての情報）。
  - 手掛かりも見えている（同、〇〇への案内（≒リンク）など）。
    ≫「どちらの案内を辿った方が効率的か？」という問題に還元できる。

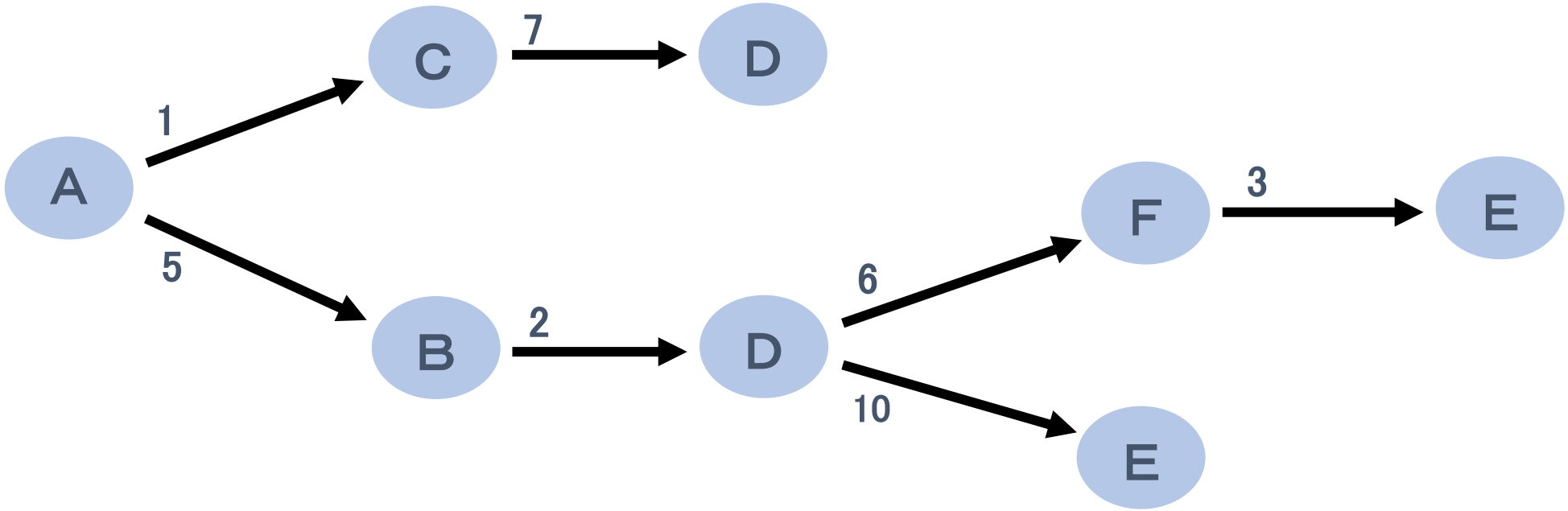
- まずは、最短経路を求める方法を探ってみよう。

解析的な最短経路の求め方

- A からスタートして、B～F に至る最短経路を求めてみよう。
- 数字は、そのリンクを通過するのに必要なコストを表わす。
  - 例えば、“A ～ F は都市” と考えると、“数字は運賃” というイメージ
  - “A ～ F は Web ページ” と考えると、“数字はリンクの適合性” など…



Dijkstra アルゴリズム（1）



- 全ての場所に至る最短経路が分かるなんて、何か凄そう。。
- この方法で十分な感じがするけれど、何か見落としている問題はないだろうか。
  - そう言えば、ホームページは数十兆ページだったなぁ。

Dijkstra アルゴリズム（2）

議論を簡単にするために、全くゼロから（誰と誰が繋がっているかも事前にわからない状態から）調べるとします。

- 最初に A を選んだ後、B ～ F から A に繋がっている X を探し、A ⇒ X のコストを調べる。
  - **B ～ F から探すので、5 回の作業が必要。**
- 次に C を選んだ後、B, D ～ F から C に繋がっている Y を探し、C ⇒ Y, A ⇒ Y のコストを調べる。
  - **B, D ～ F から探すので、4 回の作業が必要。**
- **この調子で全作業量を見積もると、5 + 4 + 3 + 2 + 1 回必要。**
- 場所が n 個所あれば、**n^2 程度の作業量**が必要
  - これを $O(n^2)$ の計算量と表わします（詳細は、後期のサイエンスで）。
- ホームページの探索には、ちょっと無理
  - 宛先が数百個所ぐらいのネットワークで利用されています（大学内ネットワークなど）。

望ましい指標とは

- 有益なページを評価する指標は、
  以下の二つの条件を兼ね備えている必要がある。

  - 有益の度合いを**定量的**に表現できる。
    ≫有益の度合いに合わせて順位を付けたい。
    ≫亀井代議士事件の罠は回避できるか？
    ≫ヒント：数理計算科学 ⇒ 論理モデルと人間の振舞いを橋渡し
    ≫（公的な案内なしで）モール内の人気店を探す方法は？

  - 膨大な Web ページを対象に、**効率良く計算**できる。
    ≫Dijkstra アルゴリズムは有益そうだけど、規模の問題がある。
    ≫こちらの方の解決は、ちょっと難しかった。
    ≫新しい繋がり方の科学が見い出されるまで、待つ必要があった。

ここまでの整理（Web 検索エンジン）

- 情報同士の**局所的な関連性に基づく繋がり**により構造化された情報群を対象とした検索
  - 全体を対象とした整理規則が存在しない

- 人間と同様に、**各 Web 上のリンクを辿りながら、情報収集**
  - データ ベース化

- 情報の繋がり方に細工をすることで、**検索エンジン自体を騙すことが可能**
  - 人間が騙され易いのと似ている。

- 辿った先の情報が有益であるかどうか、どうやって判断すれば良いのか？

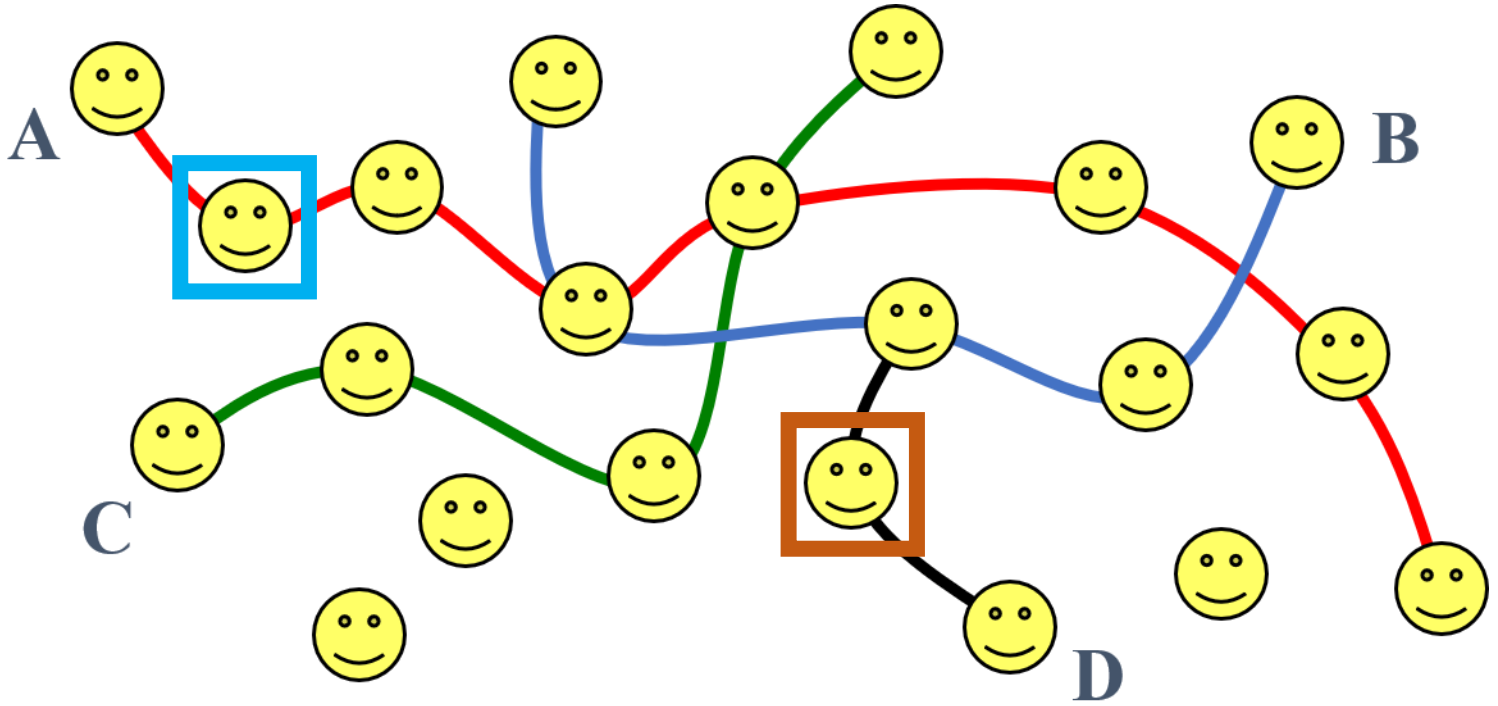
スケール フリー/スモール ワールドとページ ランク

世界の中心は自分？

- スタンレー・ミルグラムの実験
  - アイヒマン テストで有名な社会学者
  - 世の中の誰であれ、６人の**友人**を経由すれば辿り着けることを実験した。
  - 当時（1960 年代）は、郵便を使って実験した。
    ≫被験者数百人

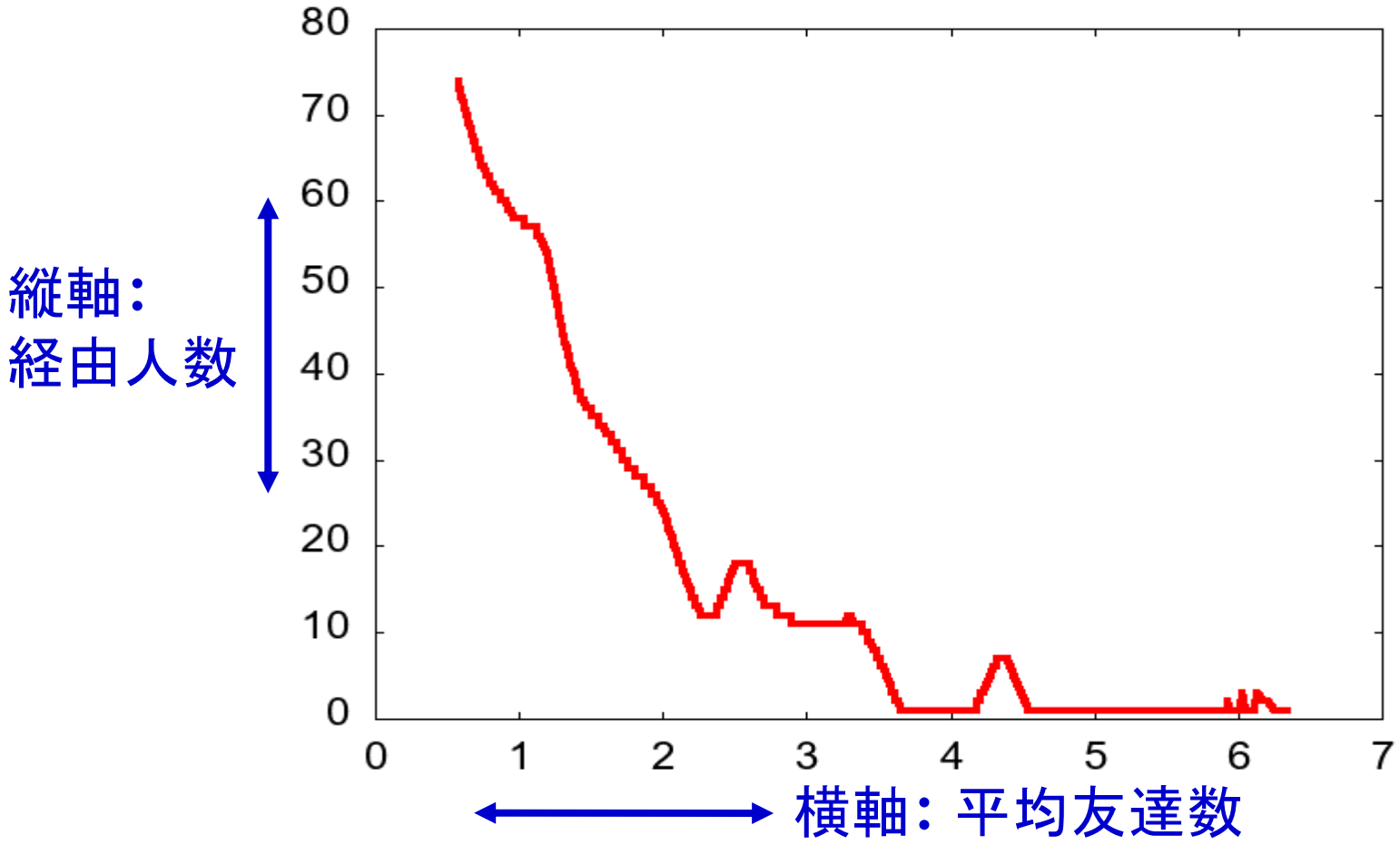
- コンピュータでシミュレーションをしてみよう。
  - **一見、手に負えない問題も、コンピュータ上に仮想環境を作ってみれば、簡単に調べられる。**
    ≫質の良い乱数生成ソフト（⇒ カオス理論, Mersenne Twister）

人間関係をモデル化してみよう



- m 人の集団の中で n_i 人のグループ（友達関係）を作る。
- 全く知らない人同士である A グループの□さんから D グループの□さんへ辿り着くには、何人の仲介が必要か？
  - 複数のグループに属している人は、各グループへの連絡役も行なう。
  - グループに属さない人は、取り敢えず距離が一番近そうな人に聞く。

友達関係のシミュレーション



横軸の平均友達数が1人未満とは、友達がいない人もいるという意味。この場合は、近くにいる人を適当に選んで聞くことになる。

- 100 人のネットワークで友達関係を作ってみる。
- 友達が増えると、経由人数は**劇的（指数的？）に減る。**

友達ネットワークの特徴

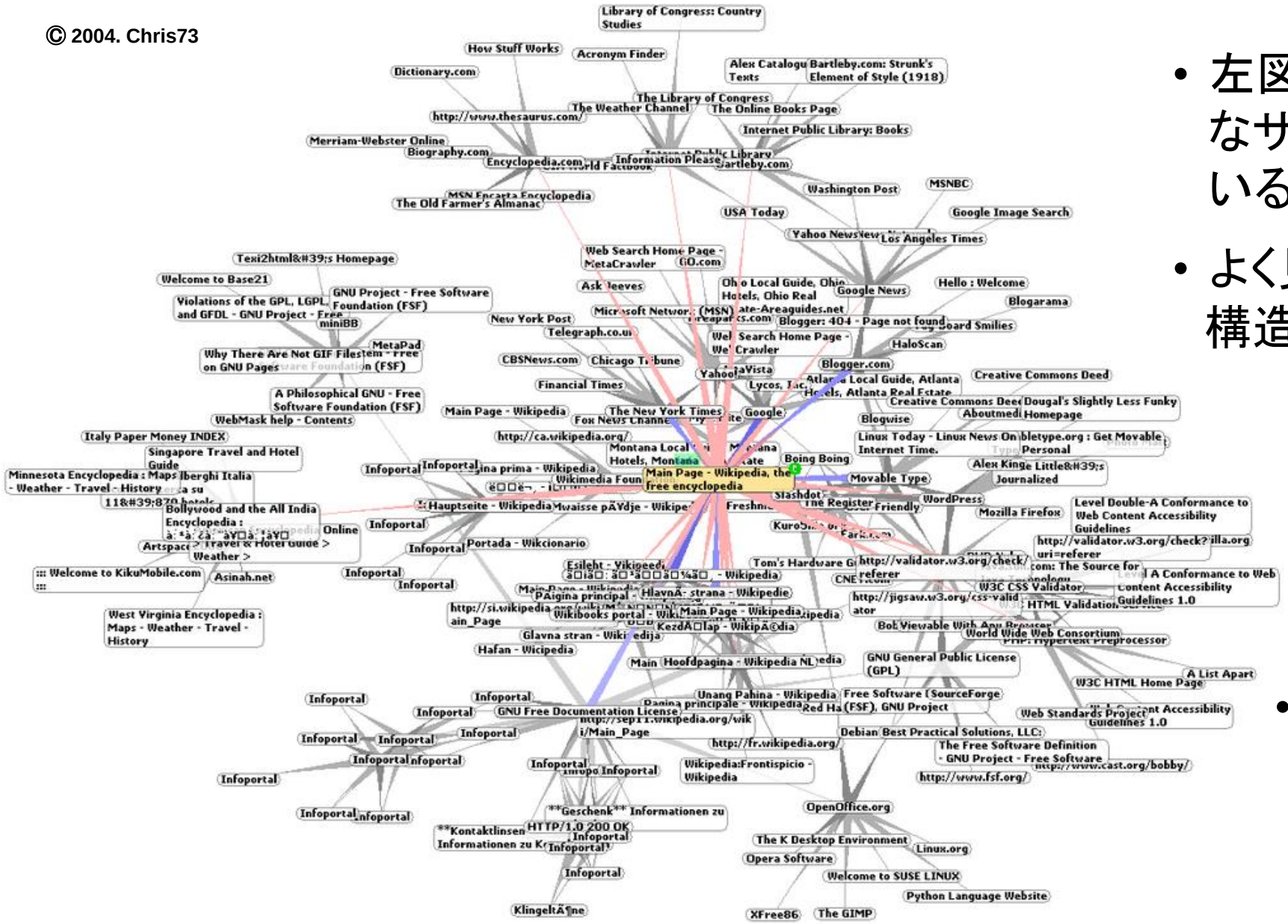
- Ｐ の友人 Ｑ と Ｒ が、また友人となる割合
  - これを**「クラスタ率」**と定義しましょう。
    - ≫ 例：友達の友達も、また友達である ⇒ クラスタ率は高い
- 先ほどのシミュレーションでは無作為に友達グループを選んだので、複数のグループに属する人もそれなりに存在したが、実際の友人関係とは似た者同士の集まりである。
  - つまり、実際の友人関係は、クラスタ率が高いはず。
  - **クラスタ率が高い集団は、閉鎖的な集団なのでは。。**
  - 友達の輪に入っていない者同士は、なかなか知り合えない。
  - でも、６人経由すれば誰にでも辿り着くけど‥（経由数は劇的に少ない）
- **クラスタ率と経由数との関係を表わす数理モデル**とは？
  - **複数のグループ間に属する人が、そう都合よく存在するのか。**

友達ネットワークの不思議

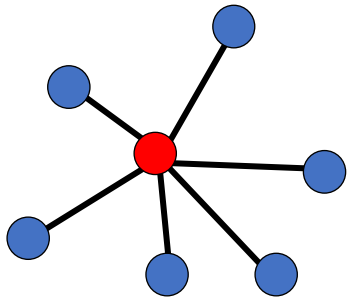
- 友達ネットワークを整理すると
  - 少ない経由数で、どこにでも辿り着ける。
  - クラスタ率（友達の友達が友達である割合）も大きい。

- どんなネットワーク構造であるかは、長い間謎であった。
  - コンピュータ技術の進歩により解明された。
  - 大規模なシミュレーション，大量のネットワークを辿ってみる 他。
    ≫ 同様な例）恒星の進化/構造などの研究
  - 出来合いのアプリをクリックしていたのでは分からない！

Wiki を中心とした Web ページの接続関係



- 左図については、Wiki が様々なサイトからリンクを張られていると読み取って下さい。
- よく見てみると、少々特徴的な構造を持っていますね。



- 上のような部品が、複数組み合わさって構成されているように見えます。

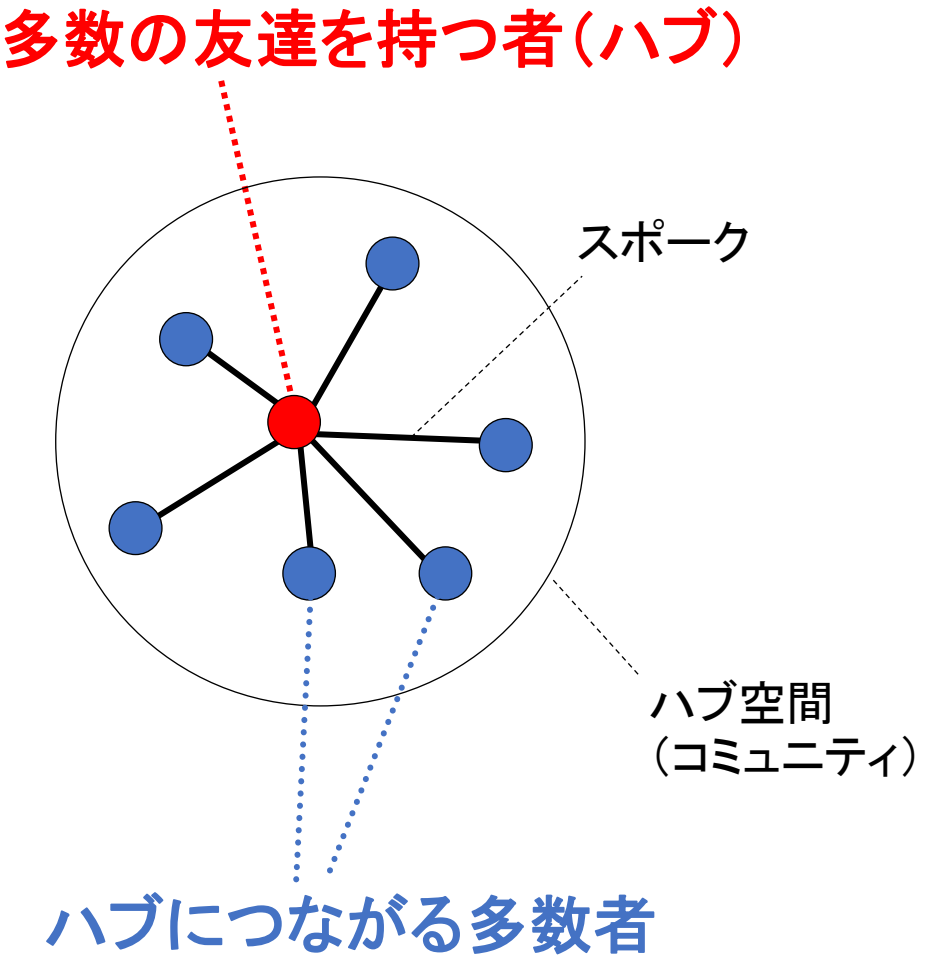
スケール フリー性

- Web のリンク関係以外も調べてみると、航空路/電力配送網/神経網/タンパク質の構造/単語の派生形など、性質の異なる様々なモノが全て同じ構造を持つことが分かった（NW に限らない，**不思議だよね？**）。

- 友達 NW で見てみると、**新しく友達関係を作る時は、その時点で一番有力そうな(Friendly な)人物と友達関係を結ぶ。**
  - 航空路を設計する時は、拠点空港とまず結ぶ。
- **強い（or 影響度の高い）モノが益々（指数的に）強くなる。**

このような性質を **スケールフリー 性** と呼ぶ（NW についてはスケールフリー NW）

スケール フリー NW の数学的モデル

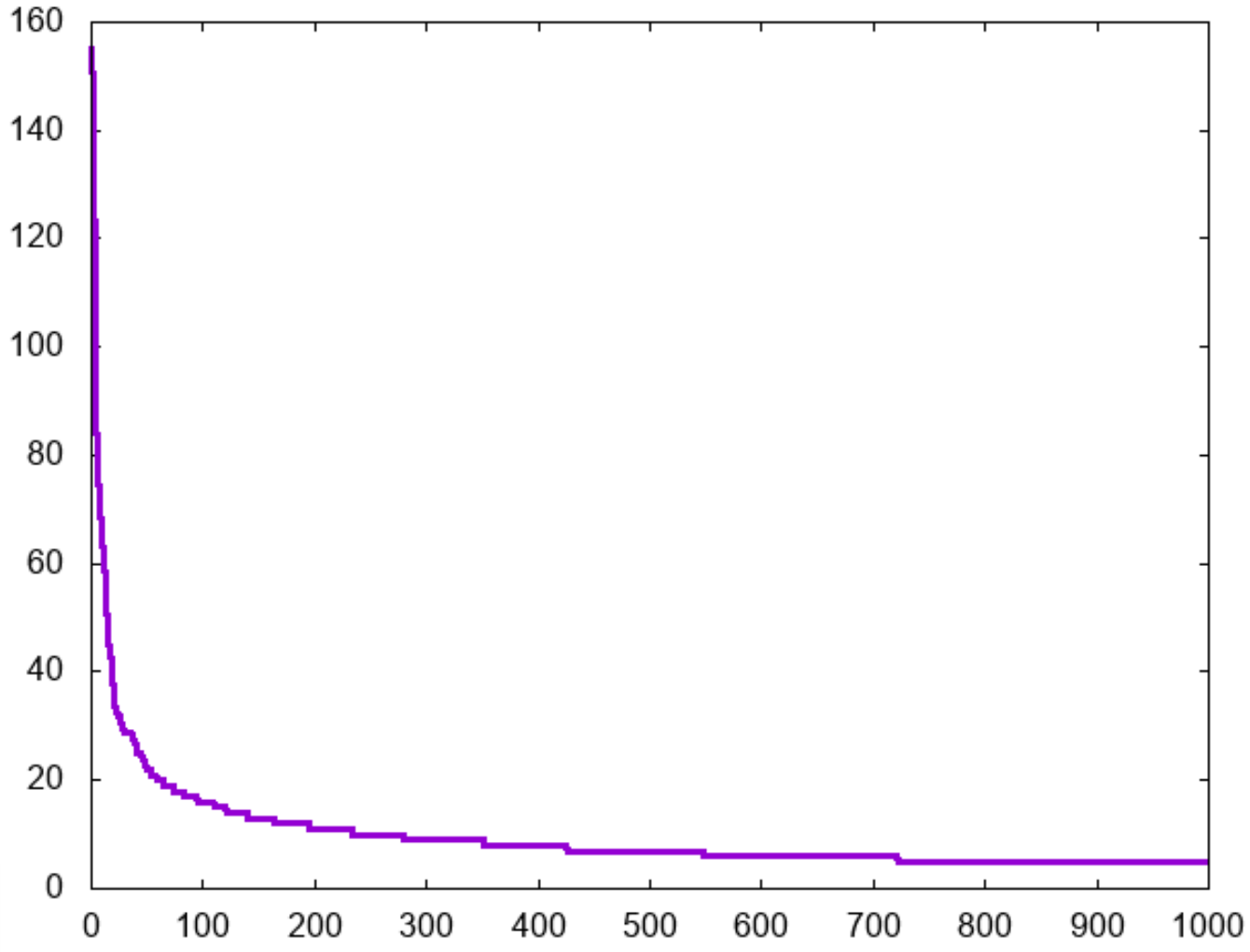


スケール フリー NW の生成：
新たにノード a_{n+1} が加わる
（リンクが増える）場合、
以下の確率 P_i で a_i に接続。
（$k_i = a_i$ が持つリンクの本数）

$$P_i = \frac{k_i}{\sum_{i=1}^{n} k_i}$$

スケール フリー生成モデルによる実例

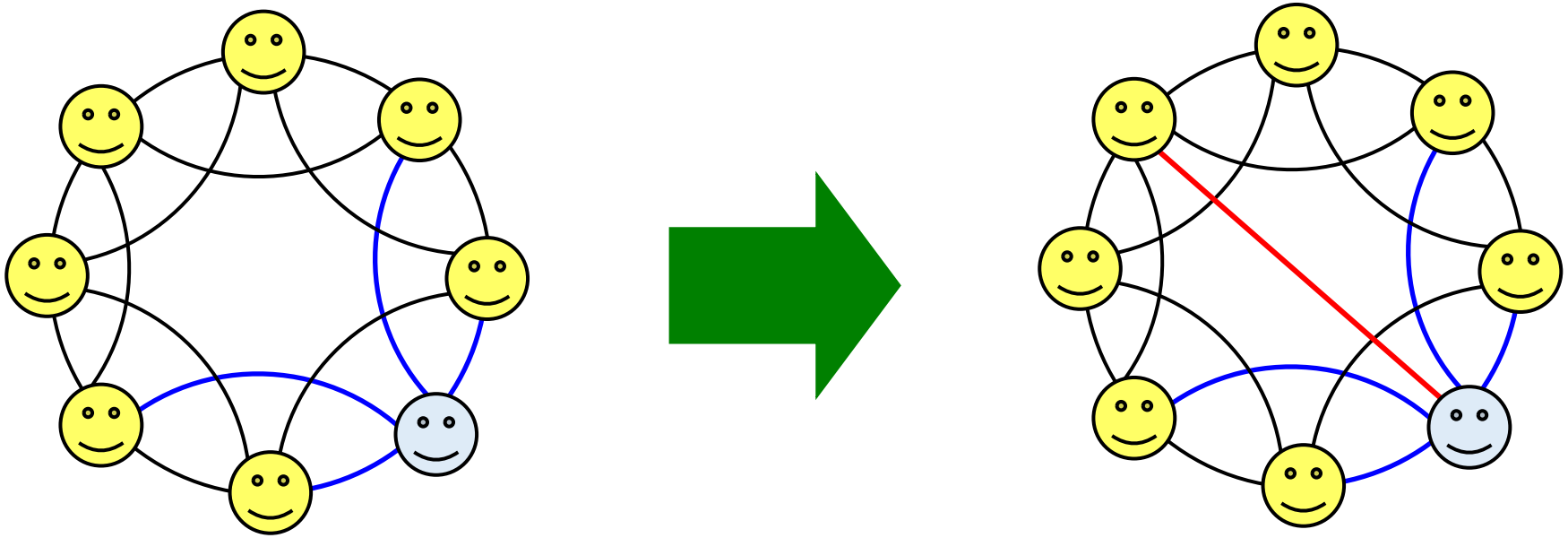
縦軸：
リンク数
（繋がりの強さ）



横軸：ノードの整理番号（a_i の i ）

全ノード数 = 1000
一つのノードが生成される時、3 本ずつの
リンクが作成されるとしたシミュレーション

友達ネットワークのタネ明かし

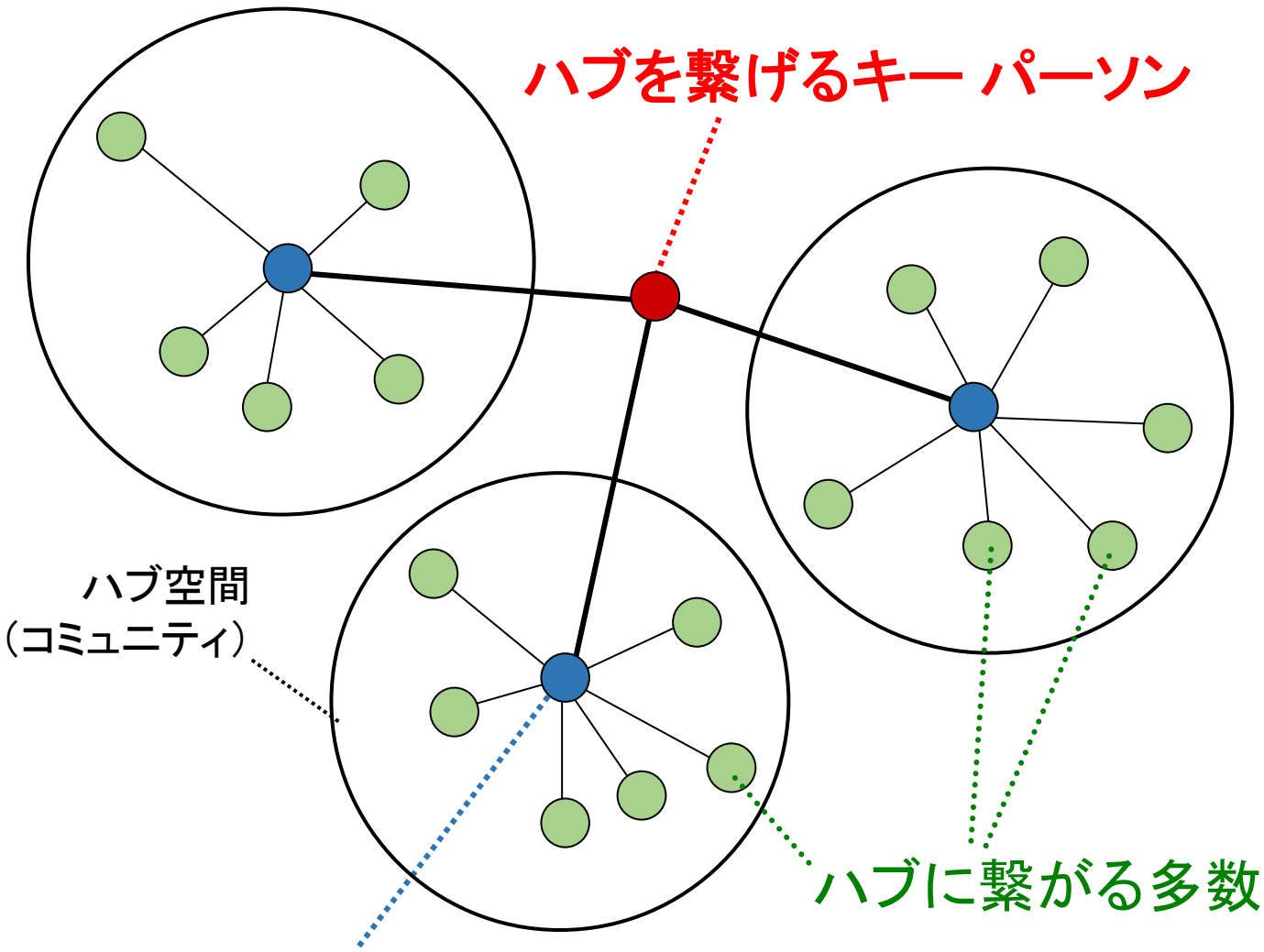


一見、孤立した（スケール フリーな）
友達集団のように見えても‥

極少数の人物は、遠くの人物と
友達関係を持っている。

- これを**弱い紐帯**と呼ぶ。
- 遠方との繋がりは、**極少数でも十分**効果がある。
  - 極少数なので、見落としがち（だから、これまでよく分からなかった）。
- このような関係を**スモール ワールド**と呼ぶ。

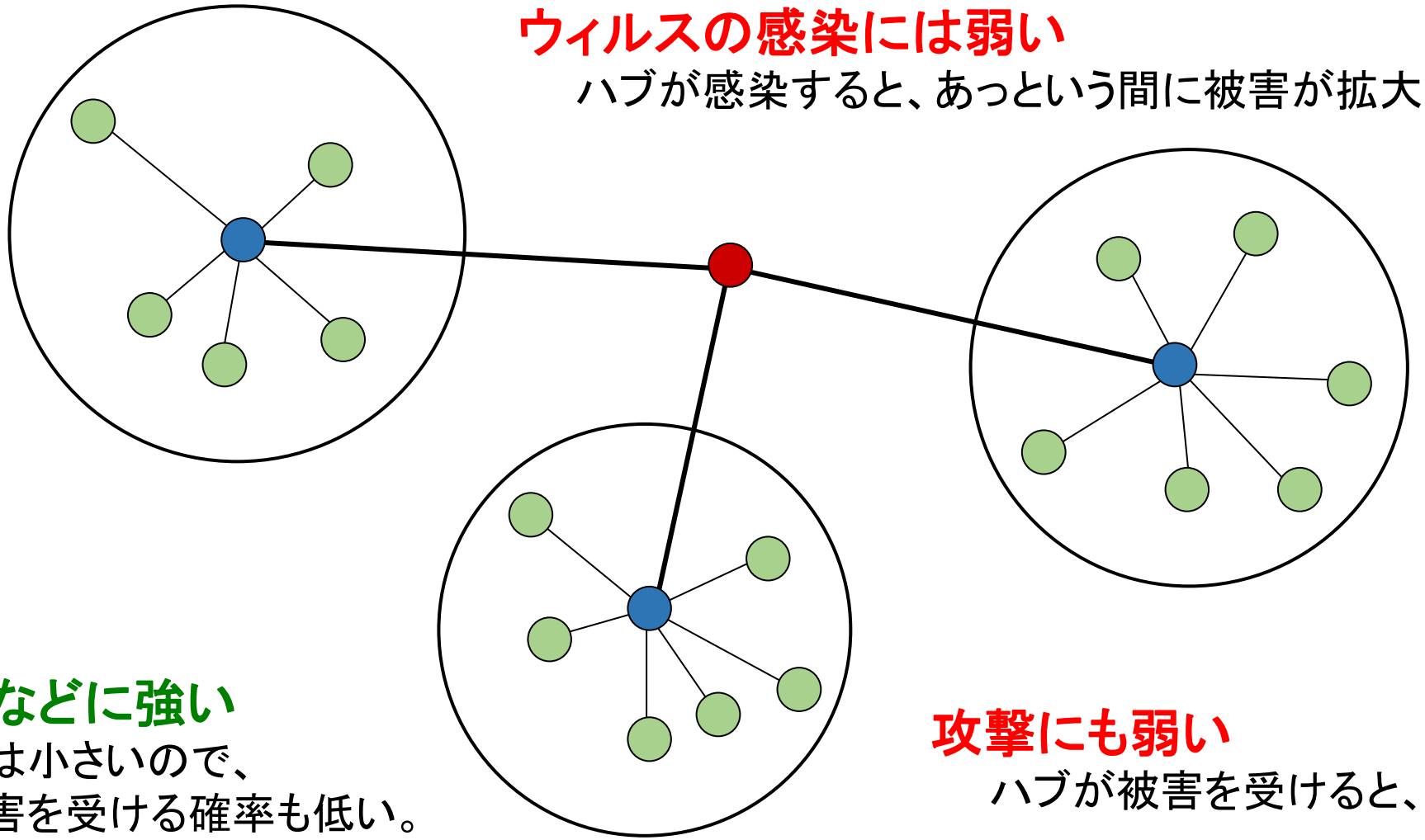
実世界における人/モノ/情報の繋がり方



- 実世界では、人/モノ/情報はハブ-スポークとキー パーソンにより、左図のような繋がり方になっています。

つまり、スケールフリー性とスモールワールド性を併せ持つ構造になっているわけです。

実世界 NW の性質



ウィルスの感染には弱い

ハブが感染すると、あっという間に被害が拡大する。

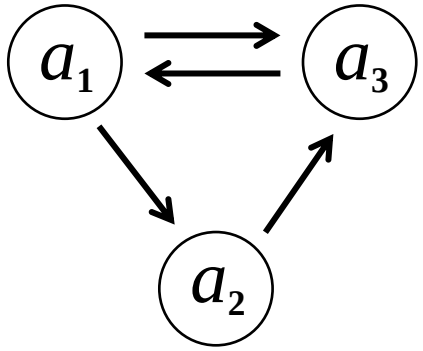
事故/災害などに強い

ハブの数は小さいので、
ハブが被害を受ける確率も低い。

攻撃にも弱い

ハブが被害を受けると、大勢が困る。

ページ ランクの原理（1）



① a_i を Web ページとして、左のようなリンクが張られていたとする（矢印の向きに注意）。
a_1 が a_2 へリンクを張って場合、
a_1 から a_2 への矢印が出る（左図では $a_2 \leftarrow a_1$）

② 上のリンク関係より、リンク $a_i \leftarrow a_j$ が存在する場合に $a_{ij} = 1$ となる下記の行列を作成し、**列成分の和が 1 となるように調整**する。

$$\begin{array}{c} \\ a_1 \\ a_2 \\ a_3 \end{array} \begin{array}{c} \begin{array}{ccc} a_1 & a_2 & a_3 \end{array} \\ \begin{pmatrix} 0 & 0 & 1 \\ 1 & 0 & 0 \\ 1 & 1 & 0 \end{pmatrix} \end{array} \Rightarrow \begin{array}{c} \\ a_1 \\ a_2 \\ a_3 \end{array} \begin{array}{c} \begin{array}{ccc} a_1 & a_2 & a_3 \end{array} \\ \begin{pmatrix} 0 & 0 & 1 \\ 1/2 & 0 & 0 \\ 1/2 & 1 & 0 \end{pmatrix} \end{array}$$

ページランクの原理（2）

③ 行列 A の列成分は**和が 1 なので、確率とみなす**ことができる。
そこで、**適当なページ**（例えば a_1）**から辿った時に**（最初の 1 回目なので初期値と呼ぶ）、
a_1 ～ a_3 に至る確率 P を計算する**（初期値も確率になるよう調整）**。

$$A=\begin{pmatrix}0 & 0 & 1\\ 1/2 & 0 & 0\\ 1/2 & 1 & 0\end{pmatrix} \Rightarrow \begin{pmatrix}0\\ 1/2\\ 1/2\end{pmatrix}=\begin{pmatrix}0 & 0 & 1\\ 1/2 & 0 & 0\\ 1/2 & 1 & 0\end{pmatrix}\begin{pmatrix}1\\ 0\\ 0\end{pmatrix}$$

④ 上の計算で得られた a_1 ～ a_3 に至る確率の分布 P_1 =（0, ½, ½）を（右式の左辺）、
リンクを一回辿った場合に至る確率分布 $P_1 = AP_0$ と考える（P_0 は初期位置に該当する分布）。
同様に、n 回辿った場合（グルグル回った場合）の確率分布 P_n を考える。

$$P_n = AP_{n-1} = A\cdot AP_{n-2} = \cdots = A^{n-1}\cdot P_1 = A^{n-1}\cdot AP_0 = A^nP_0$$

P_n ≒ P_{n-1} になるまで上の計算を繰り返した後、得られた P_n は、
普通の状態（定常状態）で、各ページを訪れる確率になっているはず。

ページランクの原理（3）

⑤ しかし、行列の n 乗計算は、一般に大変なはず。。
ここで、NW のスケール フリー 性が登場 !!
スケール フリー NW では、行列 A の成分がほとんど 0 になっている。
∵ 各ページは、ハブにしか繋がっていないから（他とは繋がらない）。
疎行列（0 成分が多い行列）から、n 乗計算をしやすい下記の行列に変形する方法が確立されている（線形代数で学ぶはず）。

$$\text{対角行列} = \begin{pmatrix} \alpha & 0 \\ 0 & \beta \end{pmatrix} \qquad \text{三角行列} = \begin{pmatrix} \alpha & \beta \\ 0 & \gamma \end{pmatrix}$$

まとめ：
多くのページからリンクを張られているサイトは、定常状態における存在確率も大きくなる。
つまり、統計的に多くの人が訪れるサイトと見なすことができる。
よって、このようなサイトほど重要であり、検索順位（ランキング）も上位にすべき。
⇒「いいね」が多いサイトほど重要というわけではないよ。

検索サイトを騙すカラクリ

- 「多くのページから参照されるサイトほど重要」と言うならば、「胡散臭い」とだけ書かれた安易なページを数多く用意して（数千ページ）、それら全てで亀井代議士のサイトにリンクを張ってやろう。

1. 検索サイトは、常に世界中のページを渡り歩いてリンク関係を調べているが、「胡散臭い」と（だけ）書かれている数多くのページが亀井代議士のサイトにリンクを張っていることが分かった。

2. おおっ、世界中の胡散臭いページが全てリンクを張っているとは、恐らくこのサイトこそ、「胡散臭い」の中心に違いない！！

実世界 NW の情報検索（作為的偏りのない体系的な情報の構築に向けて）

- 情報検索の実際は、基本的にハブ空間毎の探索になる。
  - 検索エンジンは、多くのページからリンクされているサイト（ハブ）が有益と考え、上位に表示する。
  - ハブを見つけるのは簡単
  - ハブさえ見つかれば、大半の情報を入手できるが‥
- そのハブ空間は正しいか？
  - **ハブ空間の正当性を検証するのは難しい。**
  - 例えば、自分以外が全員サクラだったら、見抜けるか？
- **異なるコミュニティ（話題）を橋渡しするキー パーソン**を探す。
  - 試してみる価値はある。
  - **or そのハブ空間が他のハブ空間からどう見られているかを検証する。**
- **有益な多くのサイトからリンクされているページこそ有益**

演習

- 講義では、ページ ランクを決定するシンプルな数理モデルを取り上げましたが、実際はもう少し様々な繋がり方の指標を使っています。

- Google の検索エンジンを対象に、他にもどのような指標が考案されているか、調べてみましょう。

- 調べた結果は、正しいかな？

今後の予定

- インターネットの仕組み
  - サーバのアドレスと URL との関係
- コンピュータ セキュリティ