2012年11月9日
ROISシンポジウム2012

ますます多様化・大規模化する生命情報に求められる情報技術とは？

東京大学
大学院新領域創成科学研究科

情報・システム研究機構
国立遺伝学研究所DDBJセンター

科学技術振興機構
バイオサイエンスデータベースセンター

高木利久

問題意識

- 生命情報は果たして「ビッグデータ」なのか？
- そもそも「ビッグデータ」とは何を意味するのか？
- 新型シークエンサー（NGS)データはビッグデータ？
- ビッグデータだとして何が有り難いのか？
- ビッグデータ用の情報技術を導入すればいい？

- 生命情報とビッグデータの共通点、相違点は？
- このように考えてみたほうが
  必要な技術、資源がより具体化明確化できるのは？
- もちろん生命情報だけでなくビッグデータの視点重要

本日の話の内容

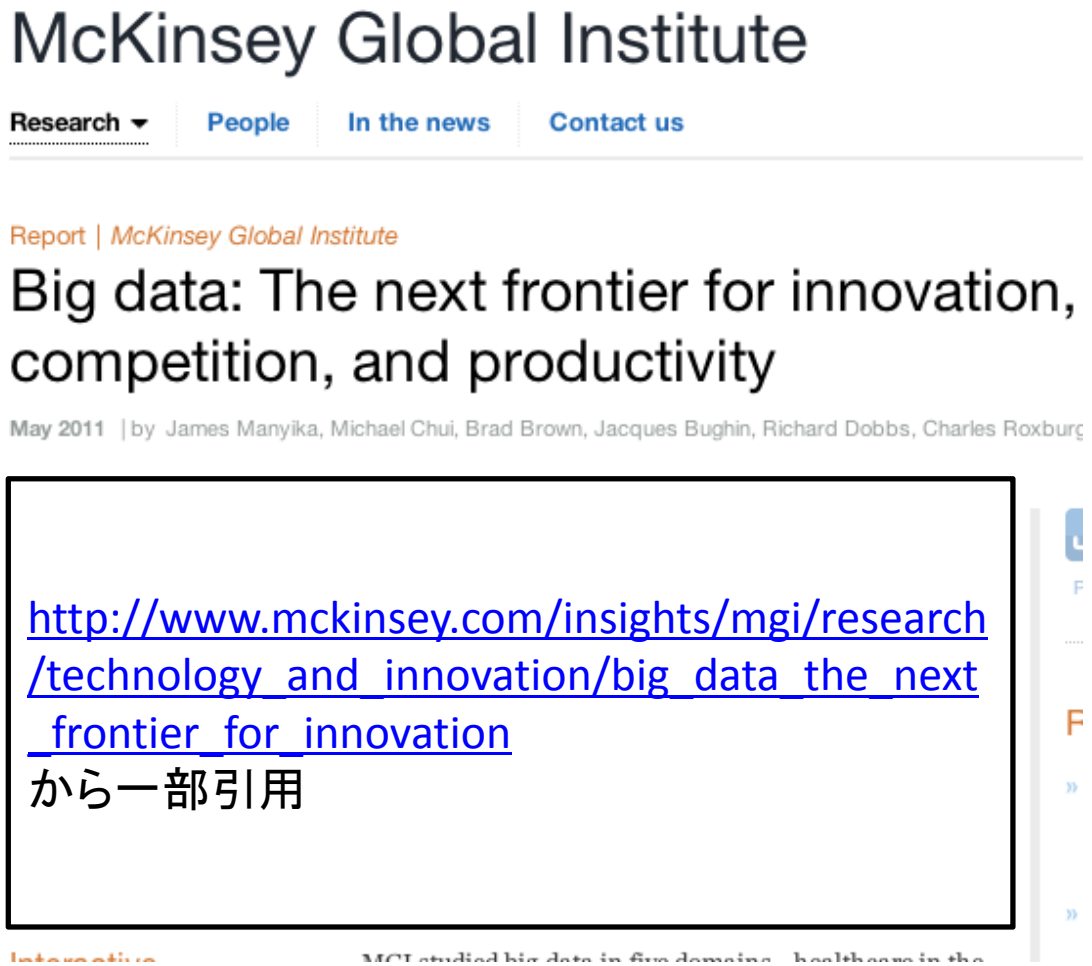
- 生命情報に求められる具体的な情報技術は？
- 生命情報のデータ、DB、解析の現状を伝える
- 「NGSデータ＝ビッグデータ」とは違う視点の提供

  - ビッグデータとは（ビジネス、サイエンス）
  - 遺伝研DDBJスパコンから眺めたビッグデータの現状
  - NGSデータとその解析
  - NGS以外のデータの活用の重要性、求められる技術
  - 統合データベースプロジェクト（RDF化）
  - 第4の科学（データ中心科学）の課題
  - ヒト生命情報統合研究に向けて

ビッグデータ

http://www.datadeluge.com/2010/03/economist-data-deluge.html
から挿絵引用

Feb 25th 2010



http://www.mckinsey.com/insights/mgi/research/technology_and_innovation/big_data_the_next_frontier_for_innovation
から一部引用

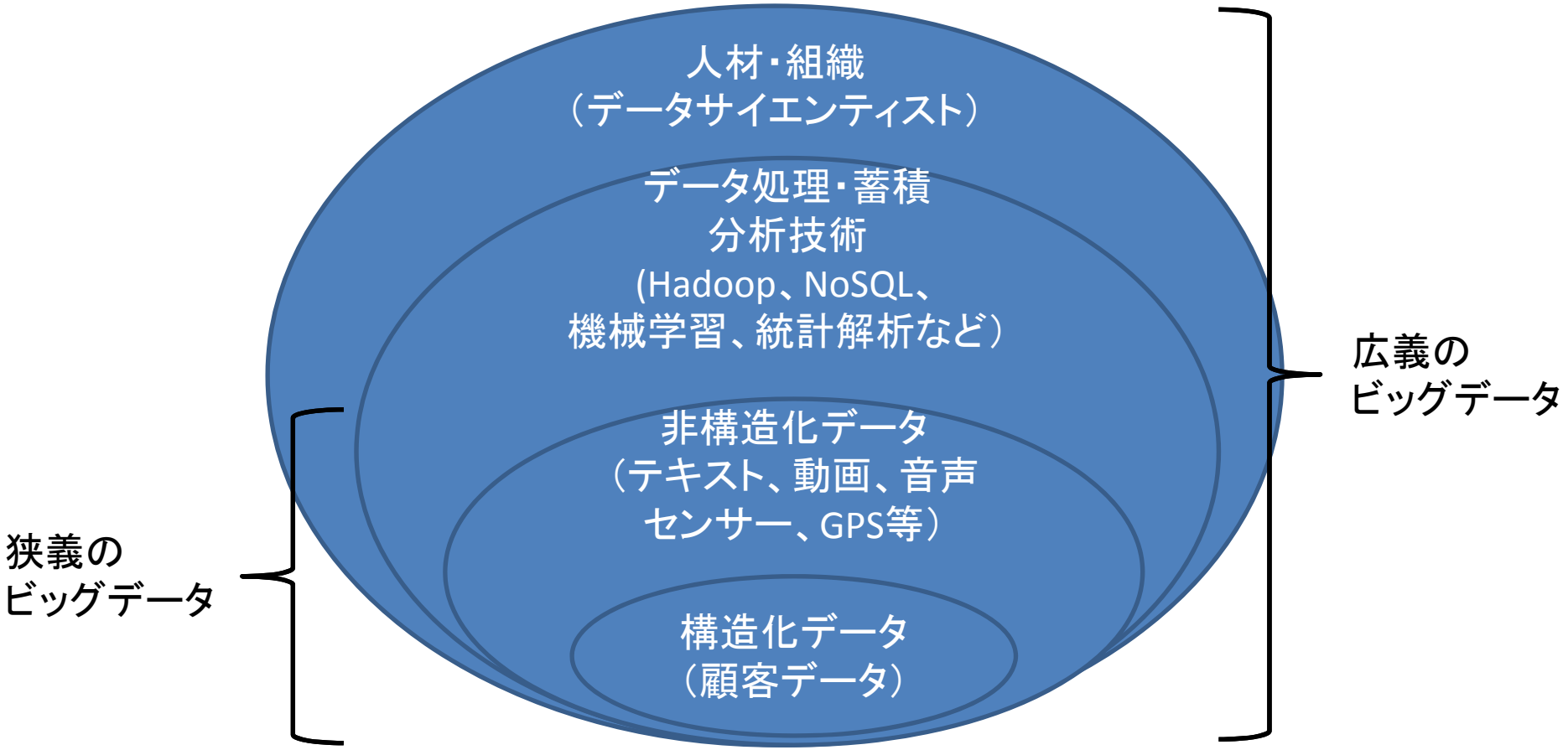
May 2011

ビッグデータの定義

- Volume（データ量）
- Velocity（データ発生頻度）
- Variety（データ多様性）（非構造化＋構造化）

- ビッグデータの定義
  - これまでの技術では扱うのが困難な大量データ

広義のビッグデータ



http://www.nri.co.jp/publicity/mediaforum/2012/pdf/forum174_1.pdfとして公開されている
野村総合研究所　城田氏の資料のp6の図から抜粋・改変

サイエンスにおけるビッグデータ

4 September 2008

Column

Big data: Data wrangling p15
Collecting and releasing environmental data have stirred
Washington, says David Goldston, and will continue to do

David Goldston

doi:10.1038/455015a

Full Text | PDF (107K)

News Features

Big data: Welcome to the petacentre p16
What does it take to store bytes by the tens of thousands
Doctorow meets the people and machines for which it's al

doi:10.1038/455016a

Full Text | PDF (2,425K)

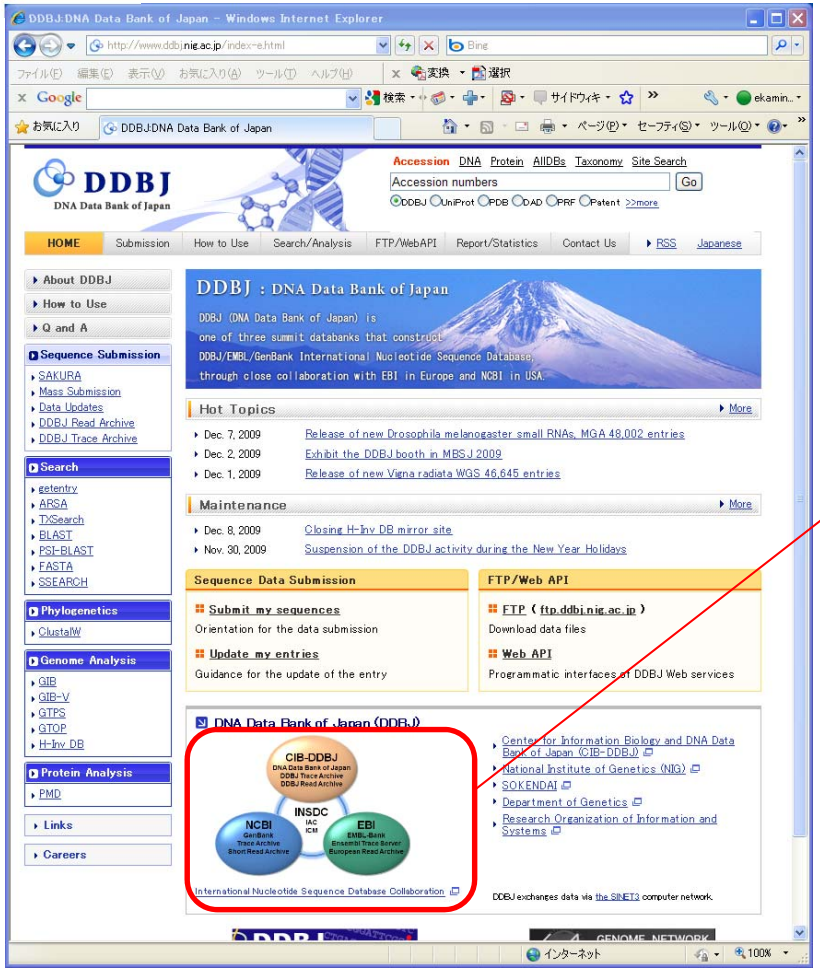
Big data: The next Google p8
Ten years ago this month, Google's firs

Big data: Wikiomics p22
Pioneering biologists are trying to use wiki-type web page

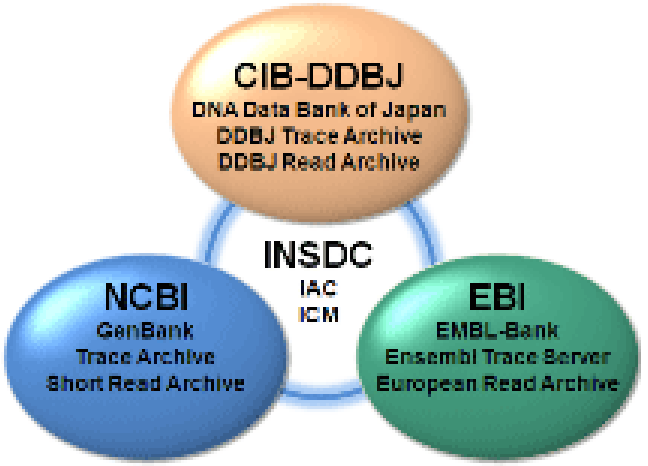
ここに掲載された記事はすべてNature誌のウェブサイトから引用（自由に閲覧可能）
http://www.nature.com/news/2008/080903/full/455008a.html
http://www.nature.com/news/2008/080903/full/455015a.html
http://www.nature.com/news/2008/080903/full/455016a.html
http://www.nature.com/news/2008/080903/full/455022a.html

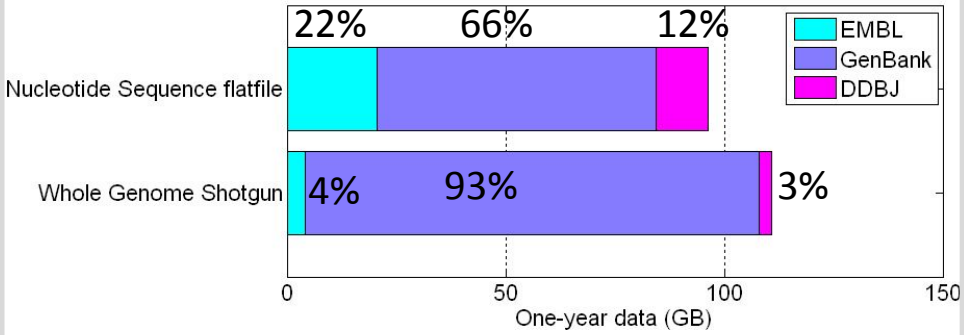
DDBJ(DNA Data Bank of Japan) は、国際塩基配列データベース(INSDC)を構築

http://www.ddbj.nig.ac.jp/index-e.html



日本・欧州・米国の3極で毎日データ交換





3極登録配列の約10%に貢献

DDBJ
DNA Data Bank of Japan

遺伝研スパコン機器構成

2012年度導入機器

Thin計算ノード

台数：352ノード [117TFLOPS以上(CPUのみ)]
CPUコア/node: 16コア
メモリ/node: 64GB
【PCクラスタ(HP Proliant SL230 Gen8)】

Fat計算ノード

台数：1台 [8.171TFLOPS]
CPUコア/node: 784コア
メモリ/node: 10TB
【SMPサーバ(SGI Altix UV1000)】

Medium計算ノード

台数：2台 [1.536TFLOPS]
CPUコア/node: 80コア
メモリ/node: 2TB
【PCサーバ(HP Proliant DL980 G7)】

高速領域

合計容量：約2PB
【Lustre FileSystem(DDN SFA10000)】

省電力領域

合計容量：約3PB
【NFS FileSystem(NEXSAN E60/E60X)】

ノード間相互接続網
【InfiniBand 4xQDR】

Rank:170th in Top500 (Nov,2011)
(Rank:11 th in Japan)

2014年度導入機器

Thin計算ノード

台数：352ノード [117TFLOPS以上(CPUのみ)]
CPUコア/node: 16コア
メモリ/node: 64GB
【PCクラスタ(未定)】

Medium計算ノード

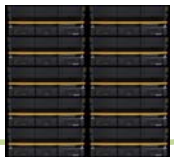
台数：2台 [1.536TFLOPS]
CPUコア/node: 80コア
メモリ/node: 2TB
【PCサーバ(未定)】

高速領域

合計容量：約5PB
【Lustre FileSystem(未定)】

省電力領域

合計容量：約2.5PB
【NFS FileSystem(未定)】



2014年増強後の総計

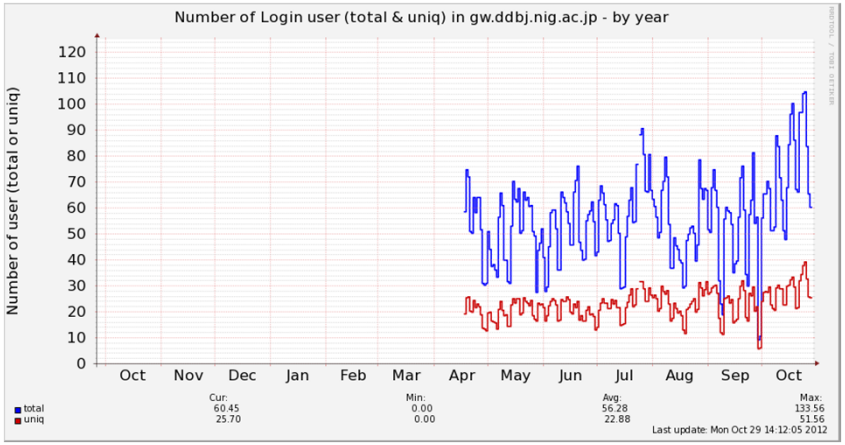
Thinノード	Fatノード	Medノード	高速領域	省電力領域	CPUピーク性能
704ノード 11,264コア	1台 784コア	4台 320コア	7PB	5.5PB	245TFLOPS以上

遺伝研スパコンユーザ登録数一覧（2012/10/30）

	外部ユーザ数	遺伝研所属ユーザ数	総ユーザ数
一般研究ユーザ	**104**	**29**	**133**
一般研究ユーザ - 大規模 -	**39**	**22**	**61**
Webサービスユーザ	**159**	**6**	**165**
DDBJ pipelineユーザ	**54**	**11**	**65**
業務ユーザ	0	69	69
スパコン管理者	0	17	17
全ユーザ合計数	356	154	510

ログインユーザ数の推移

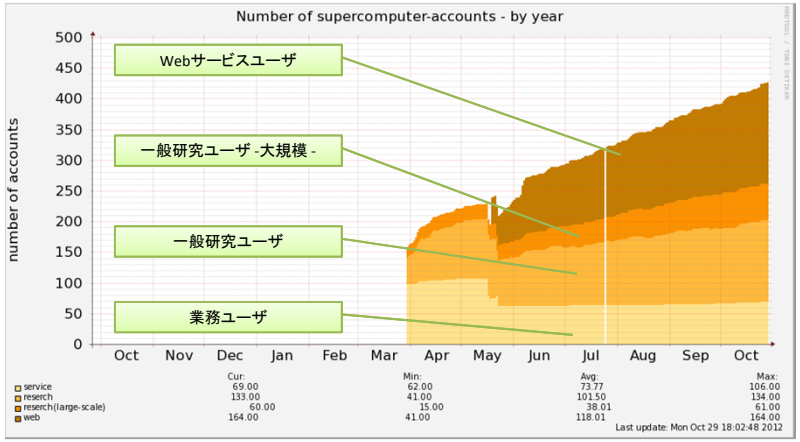
(青：総数、赤：ユニーク数)



プロット単位：1日

ユーザ数の推移

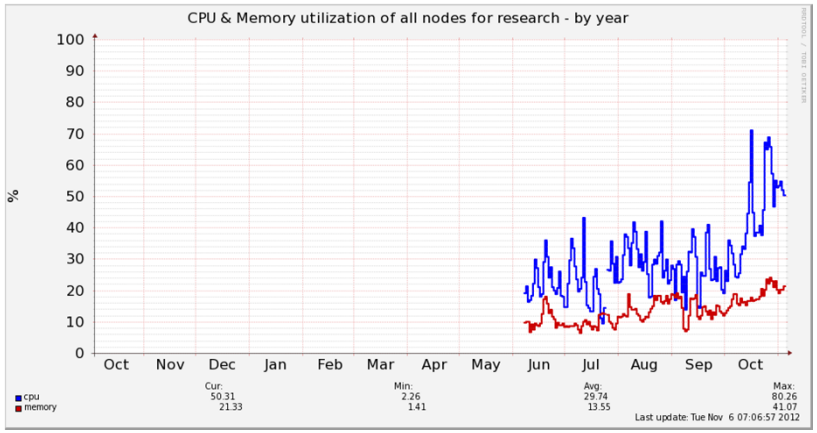
(DDB Jpipelineユーザ、スパコン管理者を除く)



プロット単位：1日

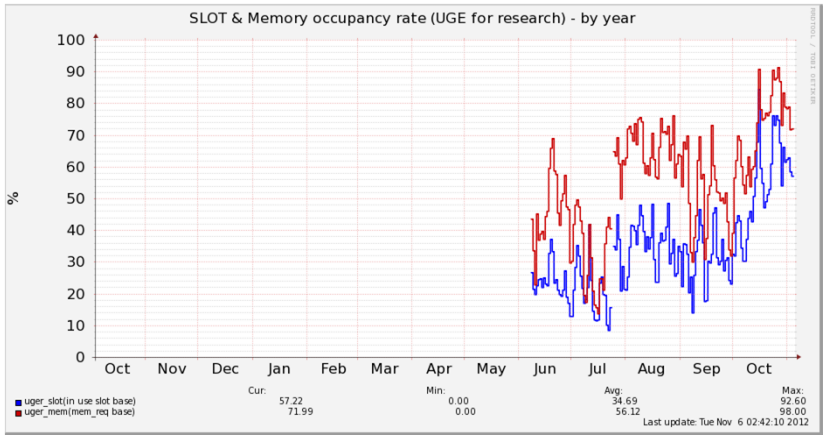
遺伝研スパコン　CPU稼働率

研究用UGE CPU&メモリ使用率　6月～



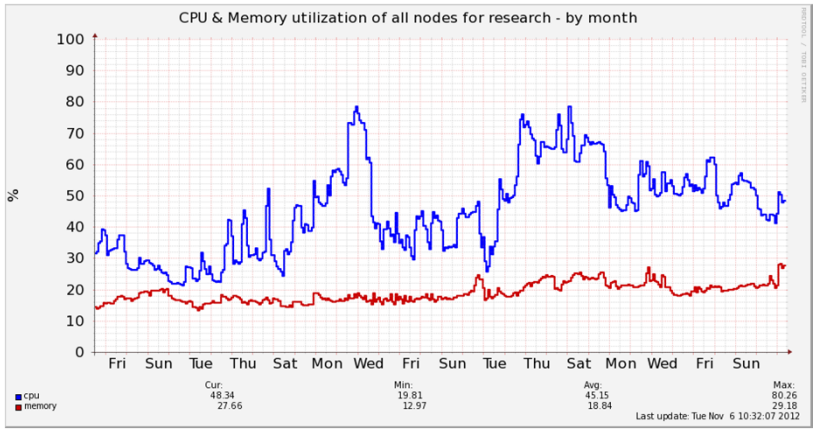
プロット単位：1日

研究用UGEスロット＆メモリ要求率　6月～



プロット単位：1日

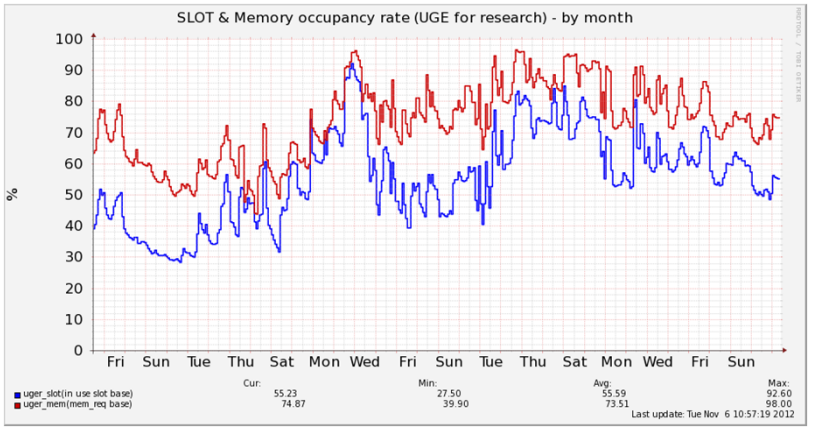
研究用UGE CPU&メモリ使用率　10月



プロット単位：2h

研究用UGEに割り当てられている全計算ノードのCPUとメモリの実使用率です。
(青：CPU利用率　赤：メモリ利用率　)

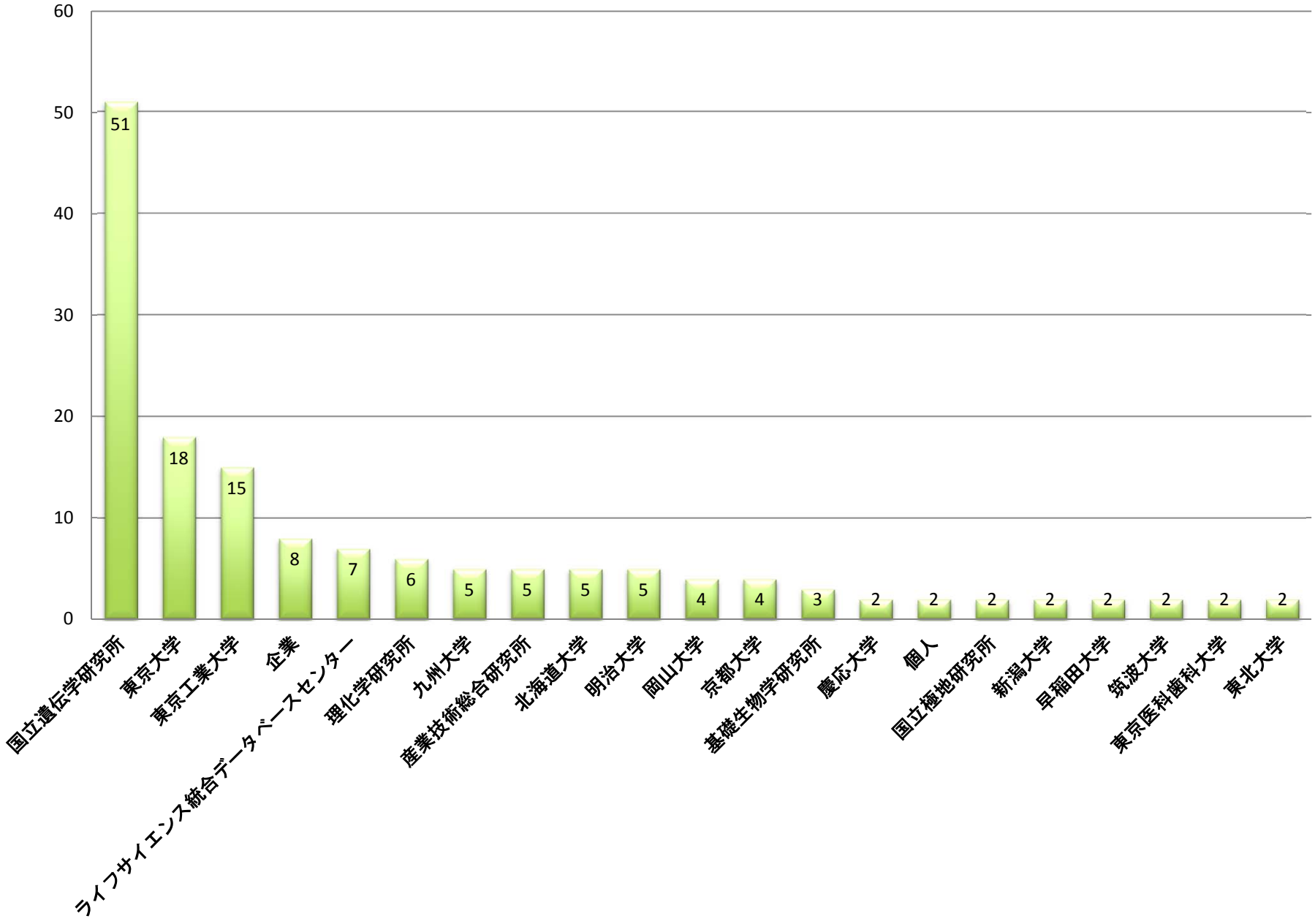
研究用UGEスロット＆メモリ要求率　10月



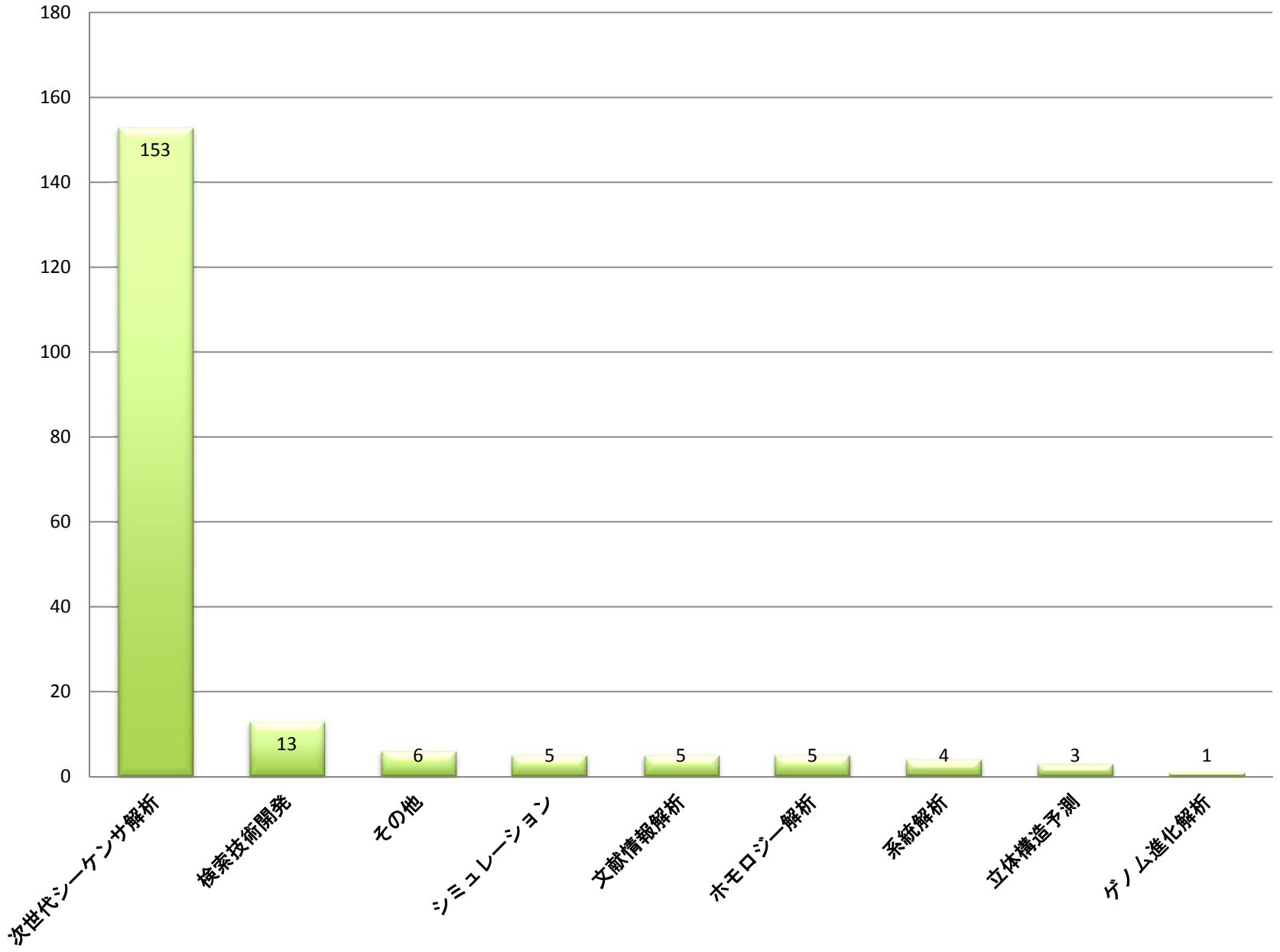
プロット単位：2h

研究用UGEに割り当てられている全ジョブスロットとメモリのうちユーザが確保している割合です。
(青：スロット要求率　赤：メモリ要求率)

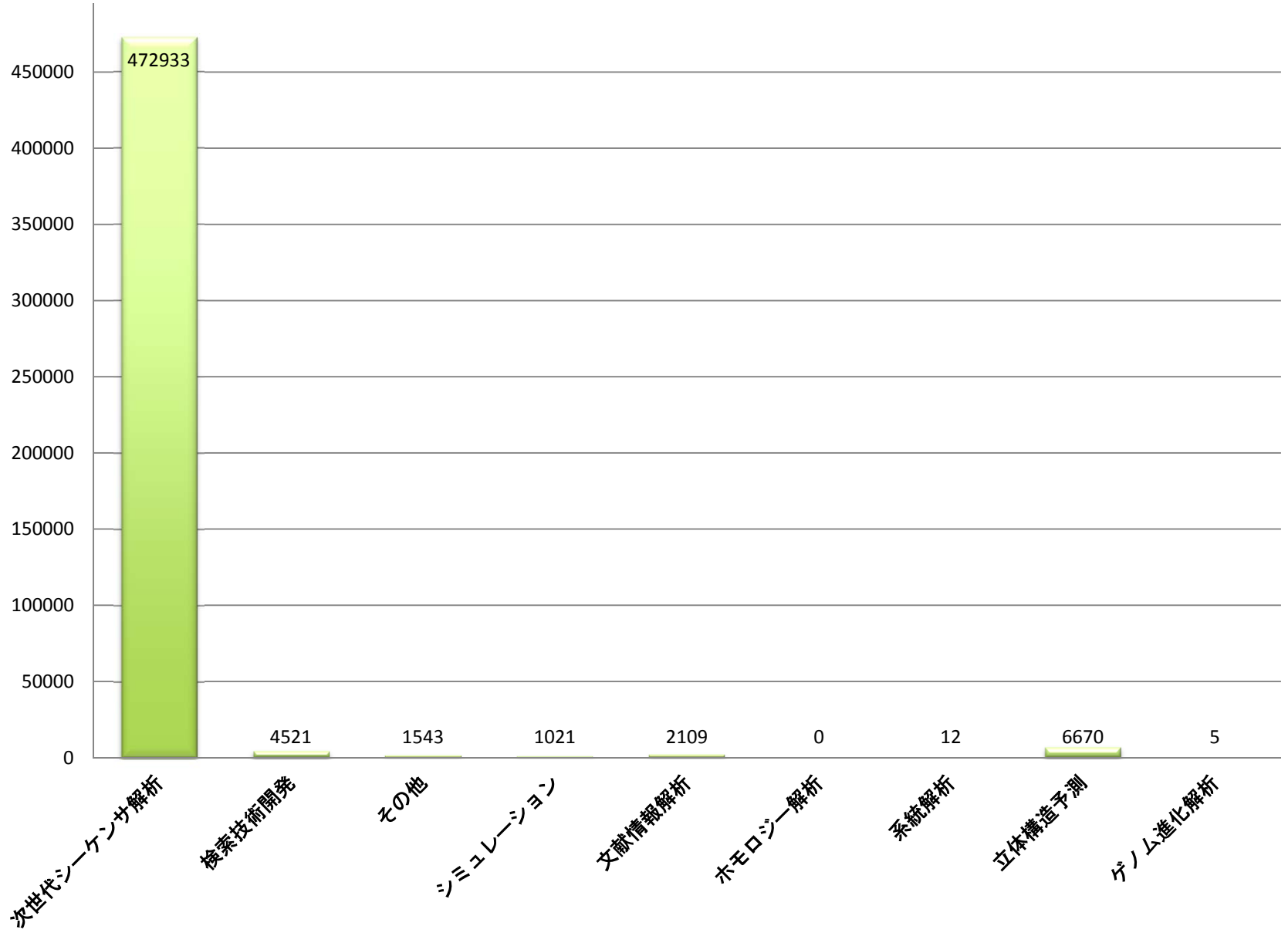
所属別　研究用ユーザ数（2名以上カウント）



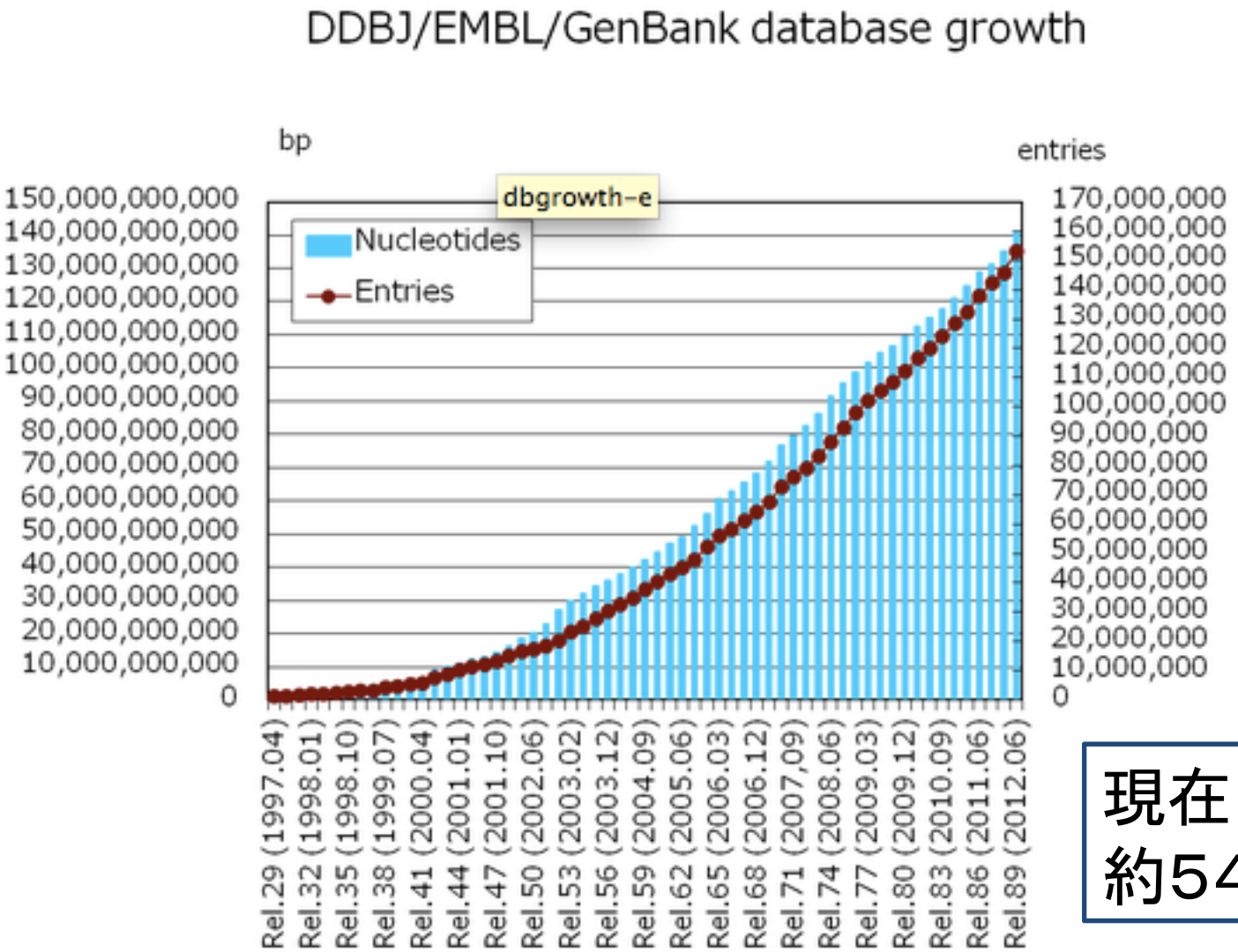
利用目的別　研究用ユーザ数(単位:人)



利用目的別　研究用ユーザ・ディスク使用量（単位：ギガバイト）

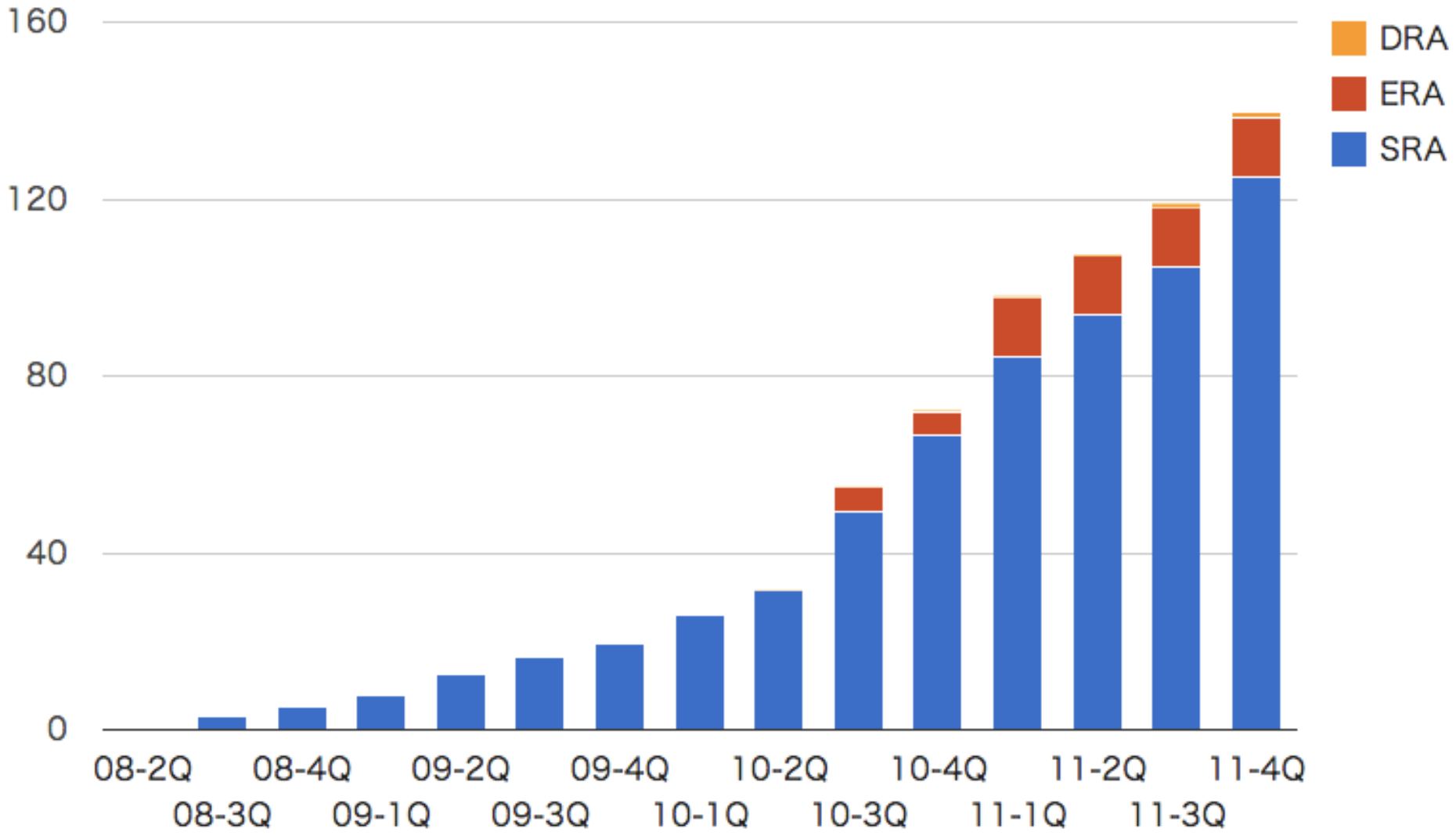


国際DNAバンクに登録された塩基数



現在のサイズ
約540GB

Sequence Read Archive推移



NGSデータの変異解析の例

シークエンスと
同時（数日）
（一検体，×30)

計算時間：数時間
(2CPU/node)

半日程度

半日程度

数時間

数時間

Base Call

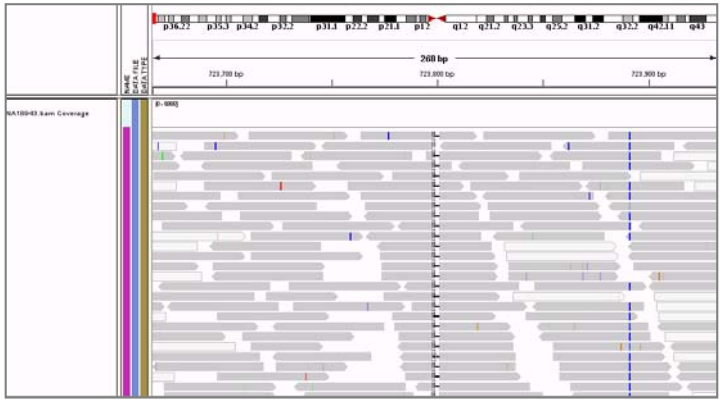
マッピング

リアライメント

リキャリブレーション

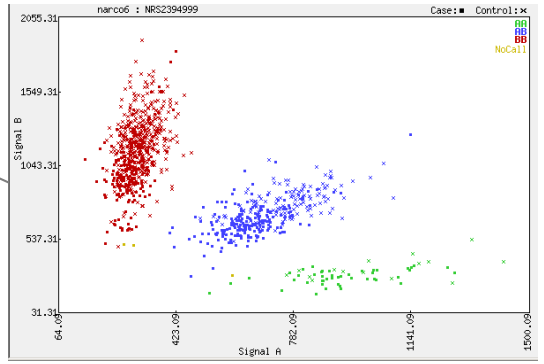
変異の検出

変異のアノテーション



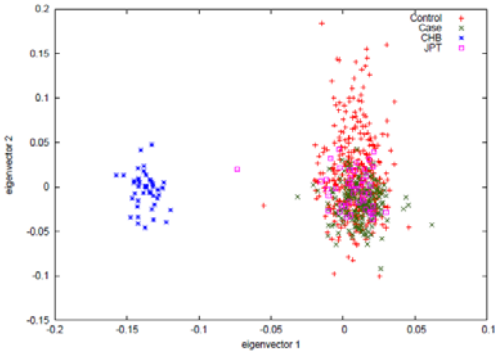
GWAS解析の例

遺伝子型の決定
（階層型ベイズなど）



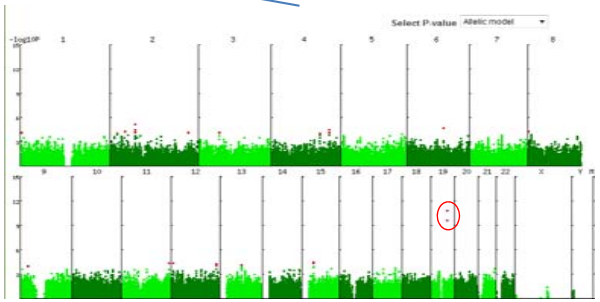
計算時間：数時間
から1日程度

品質管理
・集団としての品質管理（PCAなど）
・検体としての品質管理（検体call rate)
・SNPとしての品質管理（SNPごとのcall rate)



数時間から1日程度

・相関解析　（trend /dominant/recessive test
permutation test, 多重検定補正などを含む)
・epistasis 解析
・施設間差異の検定など



数時間から数日程度

Second stageへ

ビッグデータ利用技術のNGS適用動向

・　NGS解析の各処理は、大規模de novoアセンブルなどの一部を除き、単純並列での並列化が可能で、Hadoop(MapReduce)の適用が可能
・2009年～2012年の現状を見ると、複数のソフトウェア開発とそのオープンソース化進行中

ソフトウェア	主な利用技術	備考
Crossbow	MapReduce	マッピング
contrail	MapReduce	De novo アセンブラ
Hecate	MapReduce	De novo アセンブラ（未公表）
Hadoop-BAM	MapReduce	BAMファイルのハンドリング
Eoulsan	MapReduce	パイプライン形成
GATK	(Non Hadoop)	ツールライブラリ

・一部プロジェクトでは、Hadoopフレームワーク上に、NGSのパイプラインを構築し、これをAWS(アマゾンウェブサービス）のクラウド上に設置するということも試みている
・但し、クラウドまでの大量データ転送の問題や、課金（コスト面）の検討課題もあり、現状では発展途上の段階

NGSデータはビッグデータ？

- Volume（データ量）？
- Velocity（データ発生頻度）？
- Variety（データ多様性）（非構造化＋構造化）？
- 機械学習によるデータ分析？

- NGSデータの構造解析後の意味付け必要
  - 他のデータとの統合
  - 文献、オントロジー、生体内パスウェイなどとの統合
- データ収集からデータ分析の間に種々の処理

多様な生命情報

- ゲノム
- トランスクリプトーム
- プロテオーム
- インタラクトーム
- メタボローム
- フェノーム

- 文献
- オントロジー
- 生体内パスウェイ

ライフ分野はデータだけでなく知識も多様で膨大

データや知識の記述法

分子レベルでの実体の表現

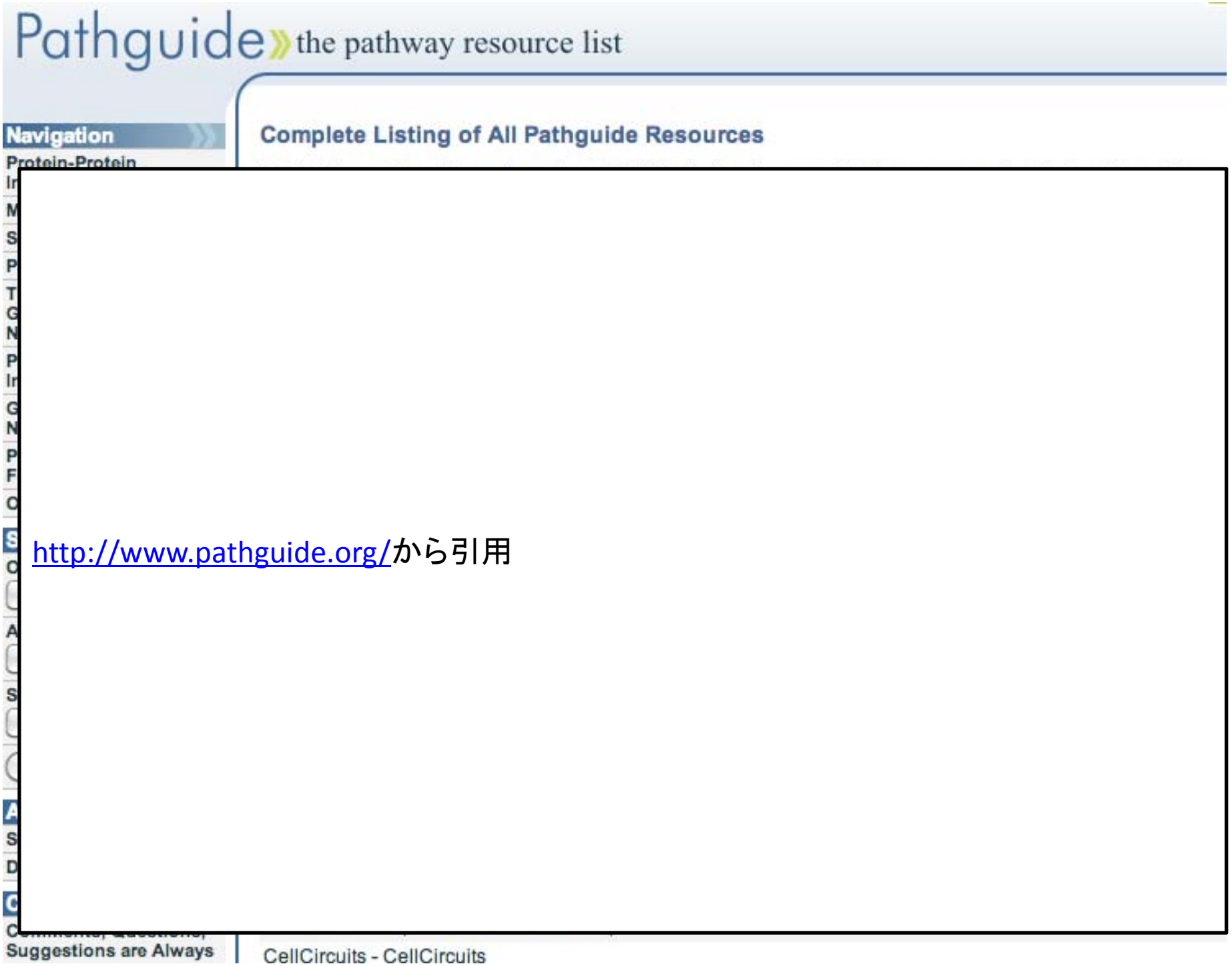
ゲノム配列、遺伝子	→	ATGCの並び
タンパク質	→	アミノ酸配列
		原子の座標配置

実体間の関係や挙動の表現

分子間相互作用	→	2項関係
遺伝子発現情報	→	発現強度(数値リスト)
上記データ取得の文脈	→	オントロジー、テキスト

機能、表現型の表現

パスウェイ、ネットワーク	→	パスウェイデータベース
表現型	→	画像、動画、テキスト
概念、機能とその階層	→	オントロジー
動的な挙動	→	数理モデル、シミュレータ


Pathguide
the pathway resource list
Navigation
Complete Listing of All Pathguide Resources
http://www.pathguide.org/から引用
Suggestions are Always
CellCircuits - CellCircuits

Symbol	**MAPK1**
Full name	**mitogen-activated protein kinase 1**
Synonym(s)	ERK ERK2 ERT1 EXTRACELLULAR SIGNAL-REGULATED KINASE 2 EXTRACELLULAR SIGNAL-REGULATED KINASE II *[manual]* MAPK2 p38 p40 p41 p41mapk p42MAPK p42MAPK1 *[PUBMED:2813464]* PRKM1 PRKM2 protein kinase, mitogen-activated 1 (MAP kinase 1; p40, p41) PROTEIN KINASE, MITOGEN-ACTIVATED, 1 PROTEIN KINASE, MITOGEN-ACTIVATED, 2 PROTEIN KINASE, MITOGEN-ACTIVATED, I *[manual]* PROTEIN KINASE, MITOGEN-ACTIVATED, II *[manual]* PROTEIN TYROSINE KINASE ERK2
Gene type	**Gene**
Gene Product	ERT1 extracellular signal-regulated kinase 2 protein tyrosine kinase ERK2 ERK-2 MAP kinase 1 MAPK 2 mitogen-activated protein kinase 1 extracellular signal-regulated kinase 2 (ERK2) ERK p40

ライフ分野のデータ統合活用を阻むもの

- データの囲い込み（きちんと公開されない）
- データの権利関係がはっきりしない
- フォーマットや用語が不統一
- 信頼性のある注釈やメタデータがついてない
- どこにどういうデータがあるか分からない
- 似たようなデータベースやツールが多すぎる
- データベースやツールの使い方が分からない

- 上記の問題を解決して
  データの共有を促進し、価値を最大化するための
  統合データベースプロジェクトが発足

我が国における統合DBプロジェクト

- 内閣府主導の統合データベースプロジェクト（H18～）
  - 文科省、経産省、農水省、厚労省で実施
  - 昨年12月に四省連携のポータルサイト

- 文科省の統合データベースプロジェクト(H18～）
  - データを公共財化・コモンズ化し、その価値を最大化
  - 中核センターの設立
    - H19～ライフサイエンス統合データベースセンターDBCLS
    - H23～バイオサイエンスデータベースセンターNBDC
  - CCライセンスの採用、フォーマット、辞書、統合技術などの開発
  - カタログ、横断検索、アーカイブの構築など
  - 分野ごとのデータベース統合化進行中（現在11課題）
  - ヒト由来データの共有ガイドラインの作成
  - 統合プロジェクトの成果、サービスはポータルサイト参照されたし


LSDB
文部科学省委託研究開発事業
統合データベースプロジェクト
ホーム
DBカタログ
横断検索
ツール
ダウンロード
About us
統合ホームページへようこそ
送信
はじめての方へ：サイトの内容をムービーやリーフレットでご紹介しています。
お知らせ　一部のサービスがJSTバイオサイエンスデータベースセンターに移動しました


ポータル
Integbioデータベースカタログ new
（旧）生命科学系 データベース カタロ
生命科学系 学協会カタログ
生命科学系主要プロジェクト一覧
生物アイコン
ライフサイエンス 新着論文レビュー
ライフサイエンス 領域融合レビュー
WingPro (JSTのDBポータル)
Webリソースポータルサイト (JST解
検索
生命科学データベース横断検索
蛋白質核酸酵素 全文検索
文科省「ゲノム」研究報告書 全文検索
NBDC
National Bioscience Database Center
バイオサイエンスデータベースセンター
科学技術振興機構
文字サイズ変更 大 中 小
English | サイトマップ | サイト内検索
検 索
ホーム
NBDCについて
研究開発プログラム
公募情報
採用情報
広報
お問い合わせ先
リンク
新着情報
twitter
RSS
2012.11.08 分野別データベースにバイオイメージ関連データベースのリンクを追加しました。
2012.11.05 [記事掲載]「データベースが生命科学の未来を変える」（JSTNews 2012年11月号 特集2）
2012.11.01 [記事紹介] BioHackathon 2012報告（情報管理Vol.55,No.8）
2012.11.01 「生命科学系主要プロジェクト一覧」のデータを更新しました
2012.10.31 【メンテナンス】サーバメンテナンスのため、2012年11月22日（木）13:00～26日（月）12:00の間、横断検索システムの一部のデー
データベース横断検索
検 索
データベースのカタログ
Integbioデータベースカタログ
コンテンツ
生命科学系 学協会 カタログ
生命科学系主要プロジェクト一覧
生物アイコン
Webリソースポータルサイト
ゲノム解析ツール リンク集
ライフサイエンス 新着論文レビュー
広報
NBDC広報サイト
パンフレット(PDF:2.78MB)
アーカイブ
生命科学系データベースアーカイブ
開発ツール
TogoDB
TogoWS

H23年度採択

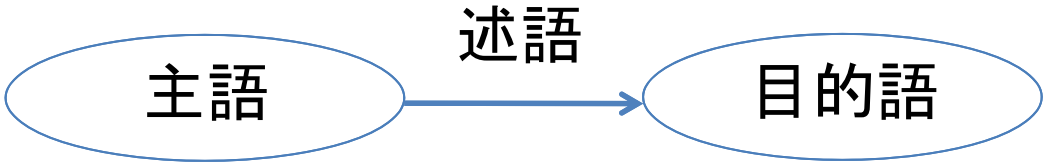
研究開発課題名	研究代表者	所属・役職
ヒト脳疾患画像データベース統合化研究	岩坪　威	東京大学 大学院医学系研究科 教授
メタボローム・データベースの開発	金谷　重彦	奈良先端科学技術大学院大学 情報科学研究科 教授
ゲノム情報に基づく疾患・医薬品・環境物質データの統合	金久　實	京都大学 化学研究所バイオインフォマティクスセンター センター長・教授
ゲノム・メタゲノム情報を基盤とした微生物DBの統合	黒川　顯	東京工業大学 大学院生命理工学研究科 教授
ゲノム情報に基づく植物データベースの統合	田畑　哲之	かずさディー・エヌ・エー研究所 副所長
ヒトゲノムバリエーションデータベースの開発	徳永　勝士	東京大学 大学院医学系研究科 教授
生命と環境のフェノーム統合データベース	豊田　哲郎	理化学研究所 生命情報基盤研究部門 部門長
蛋白質構造データバンクの国際的な構築と統合化	中村　春木	大阪大学 蛋白質研究所 教授
糖鎖統合データベースと研究支援ツールの開発	成松　久	産業技術総合研究所 糖鎖医工学研究センターセンター長
大規模ゲノム疫学研究の統合情報基盤の構築	松田　文彦	京都大学 大学院医学研究科 附属ゲノム医学センター センター長・教授

H24年度採択

研究開発課題名	研究代表者	所属・役職
生命動態システム科学のデータベースの統合化	大浪　修一	独立行政法人　理化学研究所・生命システム研究センター　発生動態研究チーム　チームリーダー

生命情報のRDF化

- Resource Description Frameworkの略
- （主語、述語、目的語）の3つ組（トリプル）で



- 主語、述語はURI（Uniform Resource Identifier)で

- Semantic Web
- LOD　(Linked Open Data)
- Web of Data

RDF化の意義

- 生命情報DBの特徴
  - 多種多様かつ新規DBが次々と作られる
  - アップデートが頻繁
  - DB間リンクが多い、リンクの種類も多様
- RDF化することにより
  - 機械可読で統一構造で分散統合利用が可能に
  - データの追加、削除が容易
  - DB間のリンクの管理が容易、表現が豊富
- 問題点
  - エントリーが多い
  - RDF化すると容量が2，3倍

RDFストアのスケーラビリティ

データロード時間

5種類のストア (Virtuoso, OWLIM, Bigdata, 4store, Mulgara)を
CPU 2.53GHz 12 cores, メモリ64GB の PC で実験

Allie: 圧縮後ファイルサイズ 509MB,
トリプル数 0.9億

➡ 5種類のストア全て成功, ロード時間12分～90分

UniProt(一部): 圧縮後ファイルサイズ 38GB,
トリプル数 40億

➡ 3種類のストアのみ2日～4日でロード成功
残り2種は不成功

DDBJ(一部): 圧縮後ファイルサイズ 52GB,
トリプル数 80億

➡ 2種類のストアのみ2日～4日でロード成功
残り3種は不成功

現状のストアでは頻繁な大規模更新には耐えられない

RDF化の技術的課題

- RDFストアのスケーラビリティ
  - 分散化必須
  - HadoopなどによるRDFストア分散化は研究段階
- オントロジーの整備
  - 誰がどう作る？
- RDF化の普及
  - ガイドラインの策定
  - それにそった変換ツールなどの整備

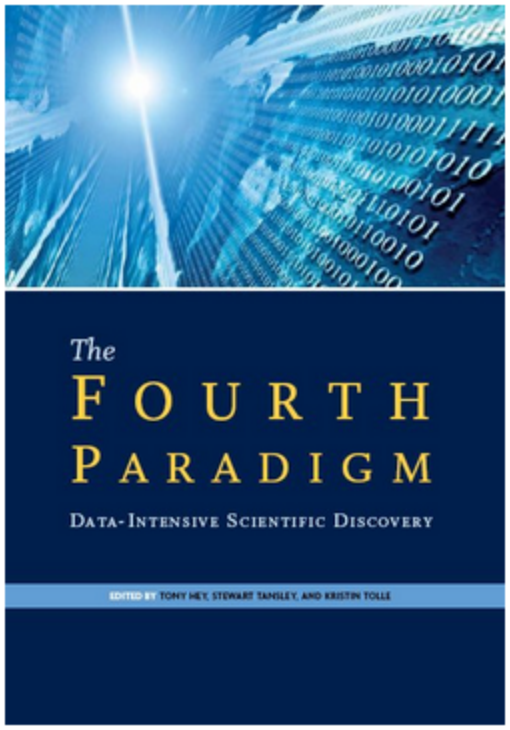
生命情報の特殊事情

- データだけでなく知識も多様
- 曖昧性、冗長性、文脈依存性など含む
- 解釈も多様
- 似ているということを扱わないといけない
- 確率的にしか捉えられないことが多々ある
- 記載の方法に多数のバリエーションある
- 要素が多数あって、その間の関係が複雑

- これらを考慮したデータ処理技術の開発必要

The Fourth Paradigm: Data-Intensive Scientific Discovery

Presenting the first broad look at the rapidly emerging field of data-intensive science



Increasingly, scientific breakthroughs will be powered by advanced computing capabilities that help researchers manipulate and explore massive datasets.

The speed at which any given scientific discipline advances will depend on how well its researchers collaborate with one another, and with technologists, in areas of eScience such as databases, workflow management, visualization, and cloud computing technologies.

In *The Fourth Paradigm: Data-Intensive Scientific Discovery*, the collection of essays expands on the vision of pioneering computer scientist Jim Gray for a new, fourth paradigm of discovery based on data-

Download

- Full text, low resoluti
- Full text, high resolut
- By chapter and essay

Purchase from Amaz

- Paperback
- Kindle version

In the News

- Sailing on an Ocean
  1s *(Science* Magazine
- A Deluge of Data Sha

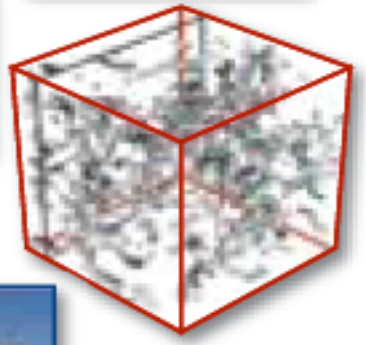
チューリング賞受賞のJim Grayが提唱した概念。2007年に不慮の死を遂げた後、Microsoft Researchが、彼の追悼の為にData-Intensive Computingをテーマとしてまとめた論考集が”The fourth Paradigm”

The Fourth Paradigm

Science Paradigms

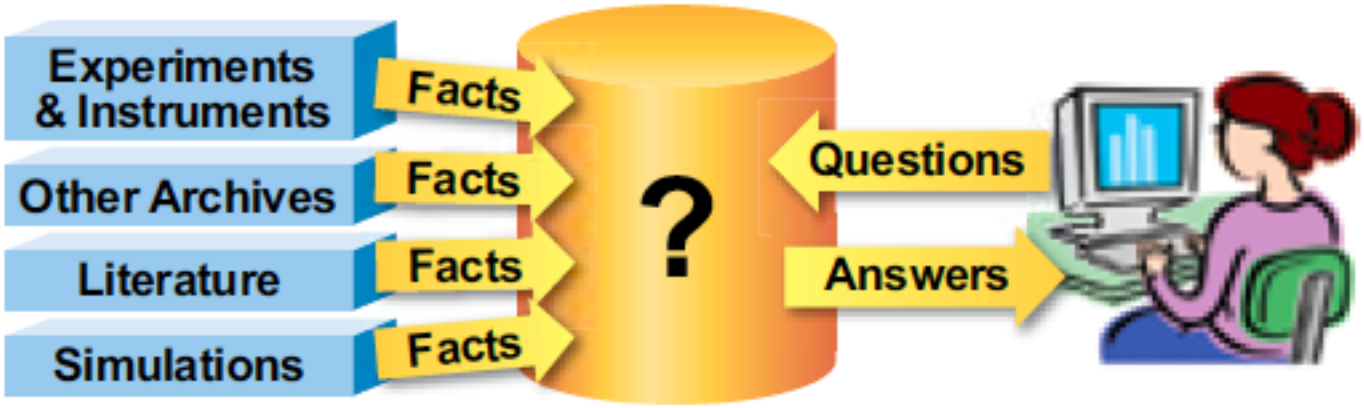
- Thousand years ago:
  science was **empirical**
  describing natural phenomena
- Last few hundred years:
  theoretical branch
  using models, generalizations
- Last few decades:
  a **computational** branch
  simulating complex phenomena
- Today: **data exploration** (eScience)
  unify theory, experiment, and simulation
  - Data captured by instruments or generated by simulator
  - Processed by software
  - Information/knowledge stored in computer
  - Scientist analyzes database/files using data management and statistics

$$\left(\frac{\dot{a}}{a}\right)^2 = \frac{4\pi G p}{3} - \mathrm{K}\frac{c^2}{a^2}$$



X-Info

- The evolution of X-Info and Comp-X for each discipline X
- How to codify and represent our knowledge



The Generic Problems

- Data ingest
- Managing a petabyte
- Common schema
- How to organize it
- How to reorganize it
- How to share it with others
- Query and Vis tools
- Building and executing models
- Integrating data and literature
- Documenting experiments
- Curation and long-term preservation

まとめ

- ビッグデータ、第4の科学：大量データが価値を生む
- 生命研究もまさにそのような状況になりつつある
- データ生産者以外がイノベータになりうる
- ライフ分野の制度の整備、研究者の意識改革必要
- NGSはビッグデータの側面あるが、専用の技術必要
- 多様な生命情報の特殊性考慮した統合化技術必要
- データ生産、DB化する前にデータ分析の設計を
- ヒト由来データの処理のガイドラインや技術必要
- 人材（統計処理だけでなく知識処理も）や組織

ヒト生命情報統合研究の拠点構築

―国民の健康の礎となる大規模コホート研究―

平成２４年（２０１２年）８月８日

日　本　学　術　会　議