Proceedings of the Twenty-Sixth RAMP Symposium
Hosei University, Tokyo, October 16-17, 2014

ERATO 湊離散構造処理系プロジェクトの主な研究成果と近況について

Principal Results and Resent Activities of ERATO MINATO Discrete Structure Manipulation System Project

湊　真一[1*]

Shin-ichi Minato

概要　計算機が扱う問題の多くは、離散構造の処理を基盤としている．近年，論理や集合のような基本データ構造を効率よく処理する「BDD」「ZDD」と呼ばれるデータ構造とアルゴリズムが様々な分野で活用されている．このような技法をベースとして，種々の離散構造を統合的に演算処理する技法を体系化し，分野横断的かつ大規模な実問題を高速に処理する技術基盤を構築することを目標として，「ERATO 湊離散構造処理系プロジェクト」が 2009 年 10 月に採択された。今年度までの 5 年間に，多くの興味深い研究成果が得られている．本講演では，これまでのプロジェクトの主な研究成果と最近の話題，および今後の展望について述べる．

キーワード　BDD, ZDD, 二分決定グラフ, 離散構造, アルゴリズム, データ構造, 列挙索引化

1　北海道大学大学院情報科学研究科 / JST ERATO 湊離散構造処理系プロジェクト，
〒060-0814 札幌市北区北 14 条西 9 丁目
Graduate School of Information Science and Technology, Hokkaido University /JST ERATO MINATO Discrete Structure Manipulation System Project,
North 14 West 9, Sapporo 060-0814, Japan

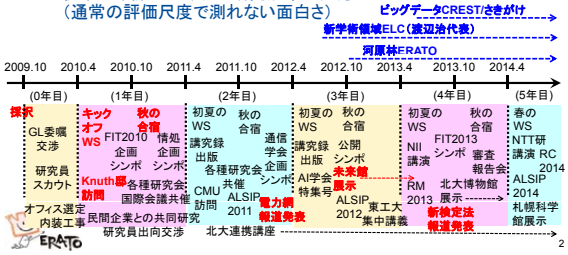
∗　E-mail address: minato@ist.hokudai.ac.jp

ERATO湊離散構造処理系プロジェクトの主な研究成果と近況について

湊　真一
北海道大学 / JST ERATO

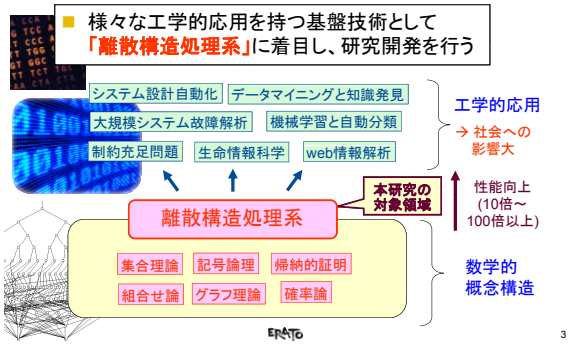
湊ERATOプロジェクトの活動経過

- 2010年度より研究活動を開始して、約5年が経過。
  - 今年度が最終年度。(来年度1年延長が決定-軟着陸期間)
  - 開放的な組織で、多様な研究者が集まる場を形成
  - 従来の常識から外れた研究成果の出方
    (通常の評価尺度で測れない面白さ)



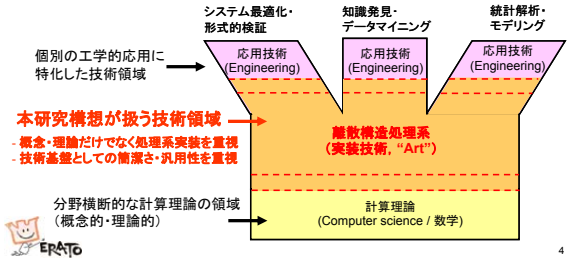
本プロジェクトの基本構想

- 様々な工学的応用を持つ基盤技術として
  「**離散構造処理系**」に着目し、研究開発を行う



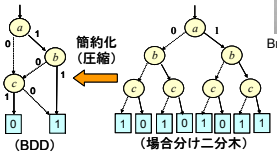
本研究構想の対象領域

BDD/ZDD技術の新しい切り口として、様々な離散構造を**統合的に演算処理**する技法を体系化し、**分野横断的かつ大規模な**実問題を高速に処理するための技術基盤を構築する。

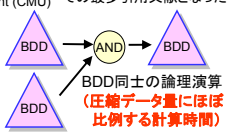


BDD(二分決定グラフ)

離散構造の最も基本的なモデルである「**論理関数**」の処理技法



1986年に画期的なBDD演算アルゴリズムを提案。以後急速にBDD技術が発達。
(長期間、情報科学の全分野での最多引用文献となった)



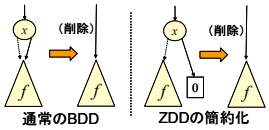
BDD同士の論理演算
(圧縮データ量にほぼ比例する計算時間)

近年のPC主記憶の大規模化により、BDDの適用範囲が拡大
(特に2000年以降)

- 場合分け二分木グラフを簡約化(データ圧縮)
- **多くの実用的な論理データをコンパクトかつ一意に表現。**(数十～数百倍以上の圧縮率が得られる例も)

ZDD(ゼロサプレス型BDD)による集合族表現

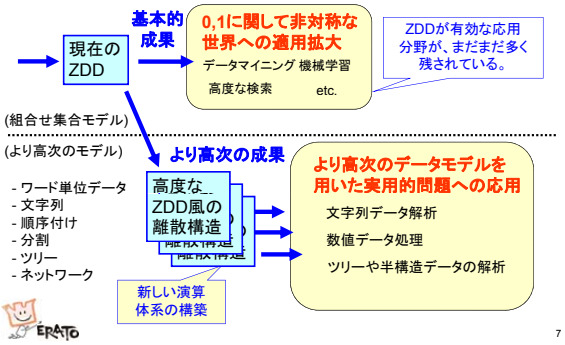
- 「**集合族**」を効率的に表現するためのBDDの改良 [Minato1993]
  - 集合族: アイテムの組合せを要素とする集合。論理関数と1対1に対応。様々な工学的応用がある。
  - **通常と異なる簡約化規則**を新規に考案。
  - **疎な集合の族を扱う場合に著しい効果が得られる。**
    (例: 商店の陳列アイテム数に比べて1顧客の購入点数は極めて少ない。)



- ZDDはBDDの改良技術として現在、世界的に広く使われている。
  - 最近では、データマイニング分野に応用されて、画期的な有効性が示されている。(数百倍のデータ圧縮率・数十倍の処理高速化)
  - 他にも応用例は増えつつある。

Dataset name: "mushroom"		CPU time (sec.)	
Min. support	#Freq. itemset	by ZDD	Prev.
1,000	123,287	0.50	0.64
500	1,442,504	1.32	3.29
300	5,259,786	2.25	9.96
200	18,094,822	3.21	31.63
100	66,076,586	5.06	114.21
70	133,336,056	7.16	277.15
50	198,169,866	8.17	357.27

本研究プロジェクトの技術面のポイント



研究実施場所とメンバ

- 札幌メインオフィス（北大・情報科学研究科）
  - ポスドク・研究員5名
  - 事務スタッフ3名、技術員3名、RA7名
  - 北大情報系の研究者と連携
- 東京サテラボ（東工大）
  - GL：東大・津田宏治氏
  - ポスドク1名
  - 東工大・東大・東京地区研究者と連携
- 大阪サテラボ（梅田駅北口）
  - GL：阪大・鷲尾隆 教授
  - ポスドク1名、技術員1名、RA2名
  - 関西地区研究者と連携

各地区は高品質TV会議システムで常時接続

フルタイムの若手研究者は極力、札幌に常駐させて、活性化・相乗効果を図る。

初夏のERATOワークショップ / 秋のERATO合宿

- 初夏のERATOワークショップ（毎年5~6月に開催）
  - 北大ERATOオフィスで開催（1泊2日）
  - 正規メンバを含む40~60名の研究者が参加。
  - 基調講演3~4件とポスター発表30~40件
- 秋のERATO合宿（毎年10~11月に開催）
  - 北海道内の旅館/ホテルを半貸切で会合（2泊3日）
    （北広島、支笏湖、夕張、登別）
  - 正規メンバを含む40~50名の研究者が参加
  - 講演（Q&Aつき）約25件とポスター発表 約15件。
  - 深夜～未明まで続く熱いディスカッション

- 発表者の質・量ともに素晴らしい
- 未発表の新鮮なアイデアを真剣に議論
- 新しい共同研究に発展することも

ERATOセミナと講究録

- ERATOセミナ：ほぼ毎週・年間35～40回実施。これまでに**150回を突破。**
  - 国内外の第一線の研究者による講演と意見交換
  - セミナの後に有志で懇親会を実施
- セミナ講演およびワークショップ・合宿での講演資料を再編集して「講究録」として毎年出版している。CD-ROMは関係者に無償配布。
  - 講演者の研究分野が幅広く、新しい研究の種が含まれている。

本プロジェクトの主な研究成果

特筆すべき成果

- フロンティア法と電力網への応用
  - **社会経済への波及効果大**
    電力網の網羅的解析を世界で初めて実現
- ビッグデータから新たな科学的発見をもたらす統計手法
  - **実験系のあらゆる学問への波及効果大**
    PNASに掲載。将来ものすごく引用される可能性あり
- 「フカシギの数え方」の展示と数え上げ世界記録
  - **青少年・一般市民への波及効果大**
    展示来場者23万人、YouTube155万ビュー

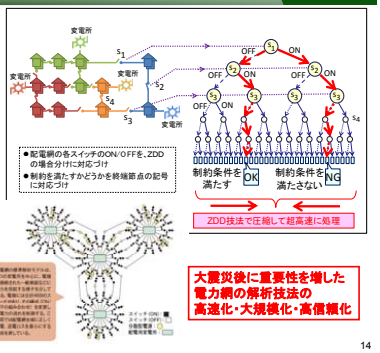
電力網への応用（2012年2月プレスリリース）

- 林 泰弘 教授（早稲田大）との共同研究
  - 電力系スマートグリッド業界のリーダ的存在
    （経産省スマートハウス標準化検討会座長、他多数の要職）
  - 電力網最適化の研究で1990年代より湊と協力関係
- 大震災後、より緊急性の高い研究課題に
  - 今後長期的に不足する電力を自然エネルギーで
    補うために必須の電力網解析・制御技術を支援

情報科学の研究者集団として
我が国の苦境を克服するため
できる限り貢献したい。

→ ERATOプロジェクト
での取り組みを加速

フロンティア法と電力網への応用



大震災後に重要性を増した
電力網の解析技法の
高速化・大規模化・高信頼化

標準的な電力網モデルへの適用結果

・標準的な電力網モデル（スイッチ468個）でグラフ的制約と電気的制約を共に満たす解を全て求め、圧縮して表現することに成功
・電力の損失を最小にする組合せを求めることにも成功

圧縮データ:約110万個（779MB）　　実行時間:数十分
解の個数:約10^{63}（2136那由他8201阿僧祇3834恒河沙8532極9116載8261正2214澗8049溝560穣9817杼8392垓4438京5235兆3981億8952万1540）通り

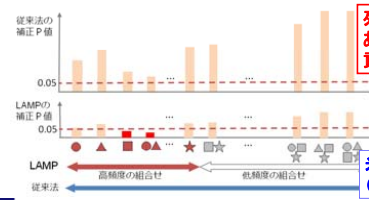
- 普通の方法で探したのでは絶対終わらない数だが、
  ZDDで圧縮してまとめて、正確に全て探し出している
- 計算結果は計算機メモリに索引化されて保存

Takeru Inoue, Keiji Takano, Takayuki Watanabe, Jun Kawahara, Ryo Yoshinaka, Akihiro Kishimoto, Koji Tsuda, Shin-ichi Minato, and Yasuhiro Hayashi:"Distribution Loss Minimization With Guaranteed Error Bound," **IEEE Trans. Smart Grid**, Vol. 5, No. 1, pp. 102-111, Jan. 2014.

特筆すべき成果

- フロンティア法と電力網への応用
  - **社会経済への波及効果大**
    電力網の網羅的解析を世界で初めて実現
- ビッグデータから新たな科学的発見をもたらす統計手法
  - **実験系のあらゆる学問への波及効果大**
    PNASに掲載。将来ものすごく引用される可能性あり
- 「フカシギの数え方」の展示と数え上げ世界記録
  - **青少年・一般市民への波及効果大**
    展示来場者23万人、YouTube155万ビュー

ビッグデータから新たな科学的発見をもたらす統計手法



列挙アルゴリズム技術が
あらゆる実験系科学に
貢献できる道が開かれた

米国科学アカデミー紀要
（2013年8月）

PNAS

Statistical significance of combinatorial regulations

Aika Terada[a,b,c], Mariko Okada-Hatakeyama[d], Koji Tsuda[c,e,1], and Jun Sese[a,b,1]

[a]Department of Computer Science and [b]Education Academy of Computational Life Sciences, Tokyo Institute of Technology, Meguro-ku, Tokyo 152-8550, Japan; [c]Minato Discrete Structure Manipulation System Project, Exploratory Research for Advanced Technology, Japan Science and Technology Agency, Sapporo, Hokkaido 060-0814, Japan; [d]Laboratory for Integrated Cellular Systems, RIKEN Center for Integrated Medical Sciences (IMS-RCAI), Yokohama, Kanagawa 230-0045, Japan; and [e]Computational Biology Research Center, National Institute of Advanced Industrial Science and Technology, Koto-ku, Tokyo 135-0064, Japan

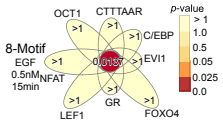
Edited by Wing Hung Wong, Stanford University, Stanford, CA, and approved July 3, 2013 (received for review February 4, 2013)

More than three transcription factors often work together to enable cells to respond to various signals. The detection of combinatorial regulation by multiple transcription factors, however, is not only deliberately excluding such tests. Here, we propose an efficient branch-and-bound algorithm, called the "limitless arity multiple-testing procedure" (LAMP). LAMP counts the exact number of

LAMP (Limitless Arity Multiple testing Procedure)

- ERATO東京グループの成果（寺田・瀬々・津田）
- 科学雑誌では、検定値（P値）が十分小さくないと、発見として認められない
- 多重検定補正（ボンフェローニ法）：n個の要因候補から発見する場合、0.05/n をP値の閾値として用いる
- 組合せ要因の数は膨大　⇒　発見は非常に困難
- 列挙アルゴリズムで出現頻度の低い組合せ要因を除く：発見力の大幅向上

これまで見過ごされてきた、
乳がん細胞の分化に関与する
転写因子の組合せを発見！



LAMP法のキーアイデア

- **「出現頻度が非常に低い組合せパタンは、決して有意にならないので、仮説集合から除外しても判定に影響しない。」**
  - あるパタンの出現頻度をλとすると、それが全部正例に入るという最も極端な偏りを起こす確率が ${}_pC_\lambda$ / ${}_mC_\lambda$ なので、これが偶然のデータ偏りの確率値（P値）の下限となる。（以降、これをf(λ)と書く）
- **P値の閾値がf(λ)より小さければ、出現頻度λ以下のパタンは決して有意にならないので、仮説集合から除外できる。**
  - 出現頻度λ以上のパタンの総数k_λを数え上げてこれを仮説集合とすると、補正した閾値は、α / k_λとなる。
  - α / k_λ < f(λ) となるような、最大の頻度λ_{max}を見つけ出せばよい。

19

LCMアルゴリズムに基づくLAMPの計算

- LCM: NII宇野先生が開発した頻出パタンマイニング手法
  - 世界最高速の実装
- LAMP法のホームページでプログラムが公開されている。
  - 内部で宇野先生のLCM実装を呼び出している。
- 結構、計算時間がかかっていて、少し大きなデータになると計算に何日もかかる、という問題を抱えていた。
  → 最近、新しいアルゴリズムの開発に成功（ECML/PKDD2014に採択）
  - 従来、頻度1きざみで何回もLCMを呼び出して探索していた。
    → 1回の深さ優先探索でLAMPの閾値を求めることが可能に

20

特筆すべき成果

- フロンティア法と電力網への応用
  - **社会経済への波及効果大**
    電力網の網羅的解析を世界で初めて実現
- ビッグデータから新たな科学的発見をもたらす統計手法
  - **実験系のあらゆる学問への波及効果大**
    PNASに掲載。将来ものすごく引用される可能性あり
- 「フカシギの数え方」の展示と数え上げ世界記録
  - **青少年・一般市民への波及効果大**
    展示来場者23万人、YouTube155万ビュー

ERATO

21

「フカシギの数え方」の展示と数え上げ世界記録

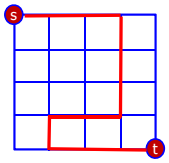
YouTube
未来館チャンネルで
1,557,869 ビュー
（2014年8月17日現在 ）

情報科学の重要性を小中高生や一般市民に強く印象付けることに成功

22

“seif-avoiding walks”の数え上げ

- 最短経路の数え上げは簡単
  (→ ${}_{2n}C_n$; 高校で習う問題)
- 最短でない経路を許すと突然難しくなる。
  （計算式や漸化式は見つかっていない）
  - おねえさんが25万年かかった結果がアニメの中に表示されているため、「計算式教えて」というコメント多数。
  - 残念ながら計算式は知られていない。効率良く数え上げるしかない。



23

「おねえさんの問題」の世界記録

- ERATO研究員の岩下氏が世界記録を更新（2013年12月）
  - **n=26までの数え上げに成功**
  - ノルウェーの大学が記録したn=24（2013年2月）を2段階上回る。
  - 様々な工夫を凝らした末に、
    主記憶2TBの計算機2台を約1週間占有して計算。

ERATO

24

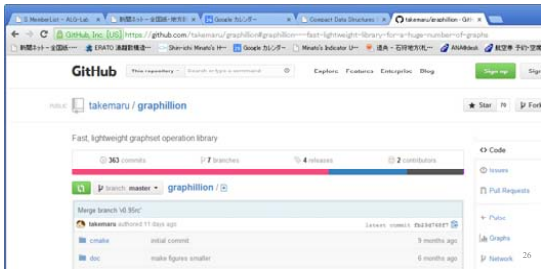
情報処理学会誌　解説記事（2013年11月号）

読者アンケート評価
2013年の全記事中で第1位

25

Open software: “Graphillion.org”

- Toolbox for ZDD-based graph enumeration.
  - – Easy interface using Python graph library.



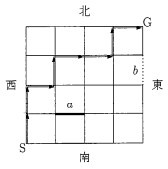
26

Graphillion チュートリアル動画

27

今年の大学入試問題　（2014年度 北大理系 数学）

4　図のような格子状の道路がある。S 地点を出発して，東または北に進んで G 地点に到達する経路を考える。ただし太い実線で描かれた区間 a を通り抜けるのに 1 分，点線で描かれた区間 b を通り抜けるのに 8 分，それ以外の各区間を通り抜けるのに 2 分かかるものとする。たとえば，図の矢印に沿った経路では S を出発し G に到達するまでに 16 分かかる。



(1)　a を通り抜ける経路は何通りあるか。

(2)　a を通り抜けずに b を通り抜ける経路は何通りあるか。

(3)　すべての経路から任意に 1 つ選んだとき，S 地点から G 地点に到達するのにかかる時間の期待値を求めよ。

ERATO

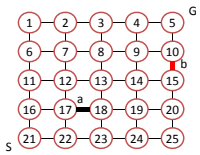
28

```
>>> from graphillion import GraphSet
>>> import graphillion.tutorial as tl
>>> GraphSet.set_universe(tl.grid(4, 4))
>>> S, G, a, b, L = 21, 5, (17,18), (10,15), 8
>>> P = GraphSet.paths(S, G).len(L)

>>> '(1)', len(P.including(a))
(1) 20

>>> '(2)', len(P.excluding(a).including(b))
(2) 9

>>> def E(P):  # 集合 P のパスを通る時間の期待値
...     L = 0.
...     for p in P:
...         for e in p:
...             if   e == a: L += 1
...             elif e == b: L += 8
...             else       : L += 2
...     return L / len(P)
...
>>> '(3)', E(P)
(3) 17.0
```



Graphillionで解いた例
by 井上 武さん
（元ERATO研究員-現在NTT）

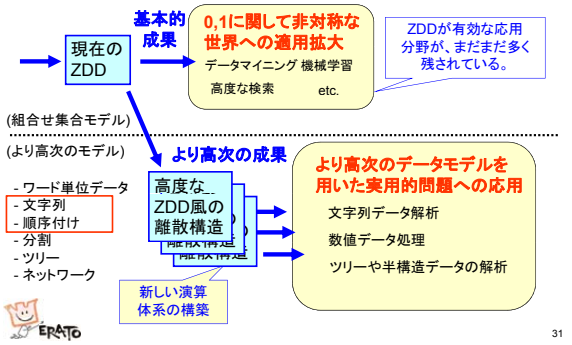
29

遠回りを許して同じところを2度通らない場合でも解ける。
（こちらの方がGraphillionの記述は簡単）

```
>>> P = GraphSet.paths(S, G)
>>> '(1)', len(P.including(a))
(1) 3322
>>> '(2)', len(P.excluding(a).including(b))
(2) 2698
>>> '(3)', E(P)
(3) 37.4081296992
```

30

本研究プロジェクトの技術面のポイント



31

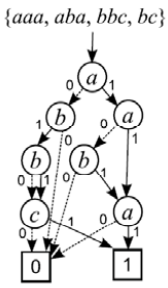
系列集合(Sets of Sequences)

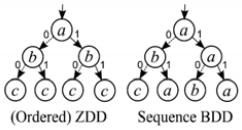
- Sets of combinations:
  - Don't consider order and duplication of items
  - "abcc" and "bca" are the same.
- Sets of sequences:
  - Distinguishes all finite sequences.
  - φ, {λ}, { ab, aba, bbc }, { a, aa, aaa, aaaa }, etc.
  - Here we exclude infinite sets such as a*.
- So many real-life applications.
  - Text search and indexing
  - Web (html/xml) data mining
  - Bio informatics

32

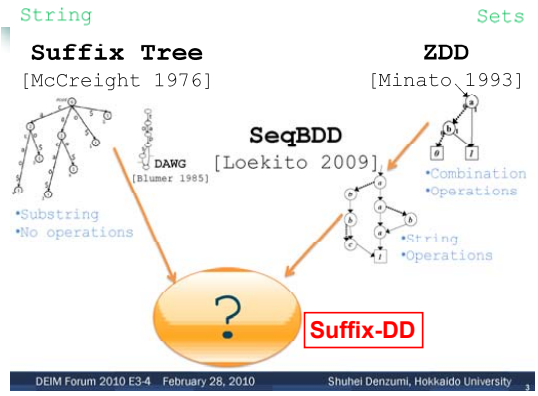
Sequence BDD (SeqBDD)

- Loekito, Bailey, and Pei (2009)
  - Same as ZDD reduction rule.
  - **Only 0-edges** keep variable ordering.
  - 1-edges has no restriction.
  - Still unique representation for a given set of sequences.
  - Each path from root to 1-terminal corresponds to a sequence.





33

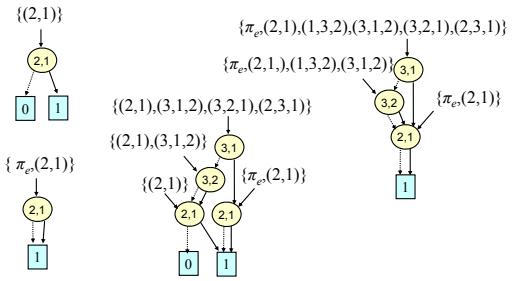


順列集合の処理

- ルービックキューブ:
  $P = \{ \pi \mid$ 基本操作による順列$\}$
  - P は18通り (= 3通り×6面) の順列の集合
  - 直積演算$P \times P$ は、2回の基本操作で得られるパタンの集合を表す。
  - P^{20} で原理的には全ての可能なパタンの集合が得られる。
    (実際は場合の数が多すぎてメモリあふれが起きる)
- 15パズル・トランプ等も順列集合で表現される
  - 最適な操作系列の解析
- あみだくじ・ソーティングネットワーク
  - 1対1のマッチング問題に適用可能
  - 順列は単射写像に対応。群論と深い関係がある。
- 情報無損失符号の設計・暗号理論への応用
  - Reversible logic設計問題への応用(量子計算とも関連)

35

Permutation Decision Diagram (πDD)

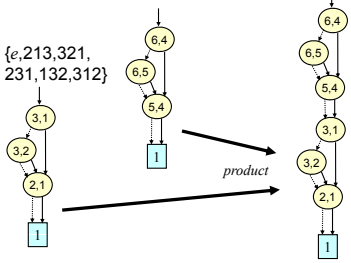


36

Product operation for disjoint permutations

{*e*,123546,123654,
123564,123465,123645}

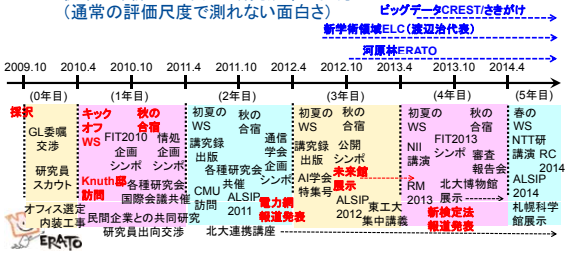
{*e*,213,321,
231,132,312}



{*e*,123546,123654,
123564,123465,123645,
213456,213546,213654,
213564,213465,213645,
321456,321546,321654,
321564,321465,321645,
231456,231546,231654,
231564,231465,231645,
132456,132546,132654,
132564,132465,132645,
312456,312546312654,
312564,312465,312645}

湊ERATOプロジェクトの活動経過

- 2010年度より研究活動を開始して、約5年が経過。
  - 今年度が最終年度。(来年度1年延長が決定-軟着陸期間)
  - 開放的な組織で、多様な研究者が集まる場を形成
  - 従来の常識から外れた研究成果の出力
    (通常の評価尺度で測れない面白さ)



今年度～来年度の活動予定

- **ERATOワークショップ(北大 4/17～18)**(招待講演:L. Simon 先生)
- **ALSIP2014主催(台湾 5/13)**(PAKDD併設ワークショップ)
- RC2014共催(京都 7/10～11)(可逆計算・量子計算)
- 通信学会IN研究会共催(北大 7/17～18)
- **ERATO合宿(礼文島 9/7～10)**
- プロジェクト最終報告会(東京 1月下旬)
- WALCOM2015基調講演(バングラデシュ 2/26～28)
- プロジェクト研究成果のpublication
  - 電子情報通信学会誌(2014年12月号)小特集(約30ページ)
    「BDD/ZDDを基盤とする離散構造処理系とその応用」
  - 専門書執筆
- **特別重点期間(2015年4月～2016年3月)**
  - 年間予算規模1/4程度、雇用研究員のソフトランディング

おわりに

- 情報科学分野のERATO組織をいかに構成するか
  - 研究する「人」と集まる「場」がすべて
  - ピラミッド型ではなく、遊牧民的なオープンな組織を構成
  - 研究テーマは一貫性を持たせた。(離散構造・BDD/ZDDを軸に)
  - 非実験系の大型プロジェクトの進め方の1つの典型例を示せたのではないか
- 「Art層」のプロジェクトは、成果の定量的評価は難しい
  - 最初のきっかけはZDDがKnuth本に取り上げられたこと
    → そもそもインパクト計測不能
  - 「Art層」の技術分野は、論文件数が少なくても大事な研究がある
  - **質的に突き抜けたERATOらしい成果をアピール**
- 計画してできるようなものでない成果:「芽」を逃さず具現化
  - 研究の方向性，オープンな協力体制，機動的リーダシップの3点