アルゴリズムとデータ構造

グループワーク

成果発表会スケジュール

- 第一部（14:45～15:20）

  開会，各種の案内，ライトニングトーク

  - 全員**56号館101号室**に集合
- 第二部（15:25～17:25）
  - ミニポスター発表
  - G1～G10の発表者は，56号館101号室で発表
  - G11～G21の発表者は，56号館102号室で発表
  - 発表を聞く人は，各自で適宜教室を移動
- 第三部（17:30～18:00）
  - 講評，結果発表，まとめ
  - 全員56号館101号室に集合

成果発表会スケジュール

- 第一部（14:45～15:20）
  - 開会，各種の案内：14:45～14:55
  - ライトニングトーク：14:55～15:20

成果発表会スケジュール

- 第二部（15:25～17:25）
  - ミニポスター発表
    - 15:25～16:05: スロット 1
    - 16:05～16:45: スロット 2
    - 16:45～17:25: スロット 3
  - 適宜，休憩，発表評価入力（Moodleから）
  - 発表がよかったグループは17:30までに入力すること.

成果発表会スケジュール

- 第三部（17:30～18:00）
  - 講評，結果発表，まとめ

成果発表会当日の名札について

- 当日は教室の入り口で名札入れを配布します．**氏名，グループ名を記入した紙（縦5cm，横9cm以内の大きさ）を発表会の当日もってきてください．**

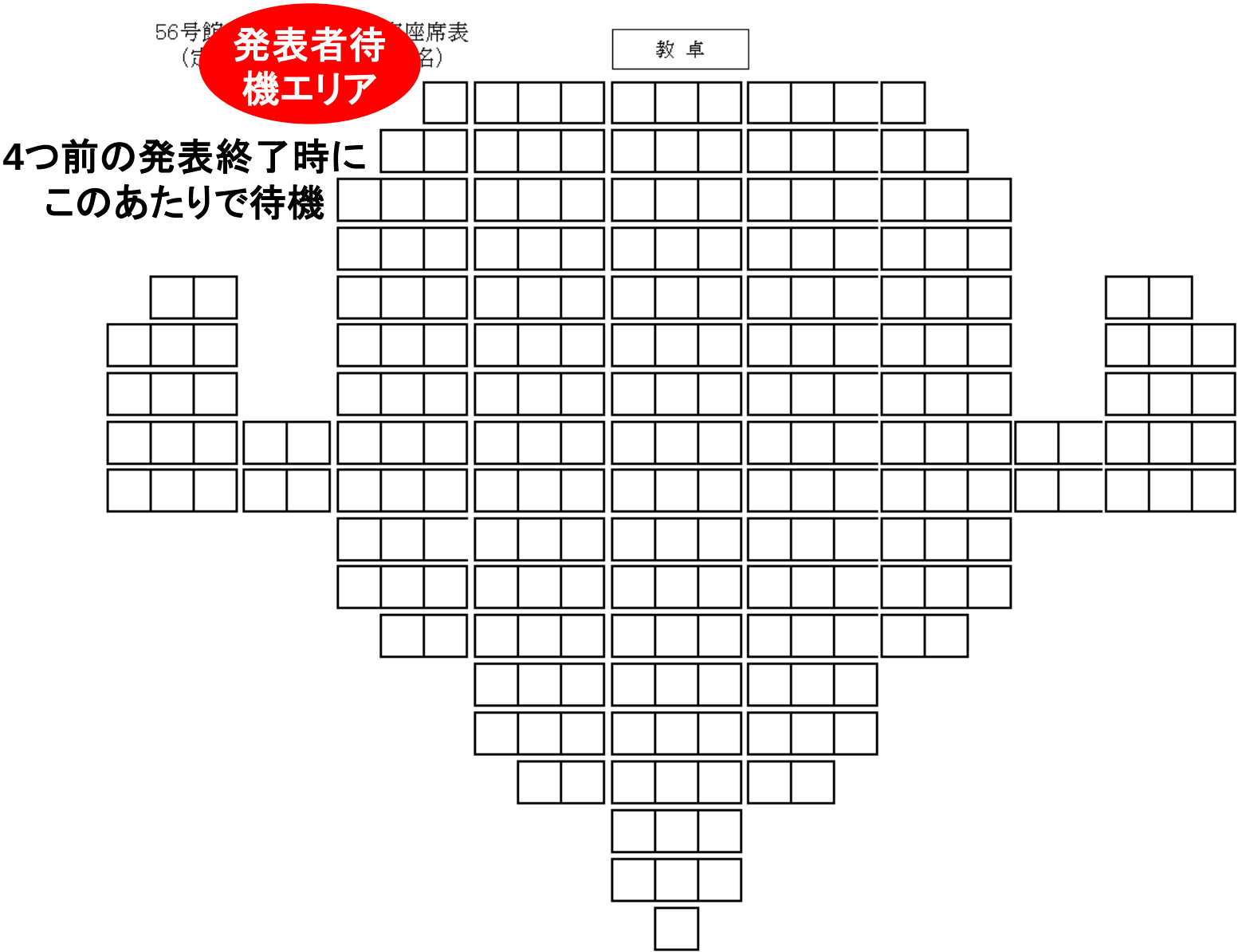
ライトニングトーク

- グループの代表者が壇上で口頭発表
- グループID順に発表
  - **教壇のPCに発表資料が保存されています.**
- 持ち時間： **1分（厳守）** **※1分で打ち切ります**
- 各グループの発表者は，自分の発表の4つ前のグループの発表が終わったら，壇上の近くで待機すること．G1～G4の発表者は授業開始時に壇上近くで待機すること.
- ppt, pdf以外（例えばkeynote）で発表したい人は，授業時間中にご相談ください.

ライトニングトーク

- 非常に短い時間での発表のため，ポイントを絞ってアピールすること.
  - 問題は全員知っているので説明不要.
  - 何に注目して，どんな工夫をしたのか？
  - 全部を説明しようとするのではなく，聴衆が後でミニポスター発表に足を運びたくなるようなアピールを心掛ける.
  - イメージを伝えやすいように，図やアニメーションなどを効果的に使う戦略もよい.

56-101座席表＆発表者待機エリア



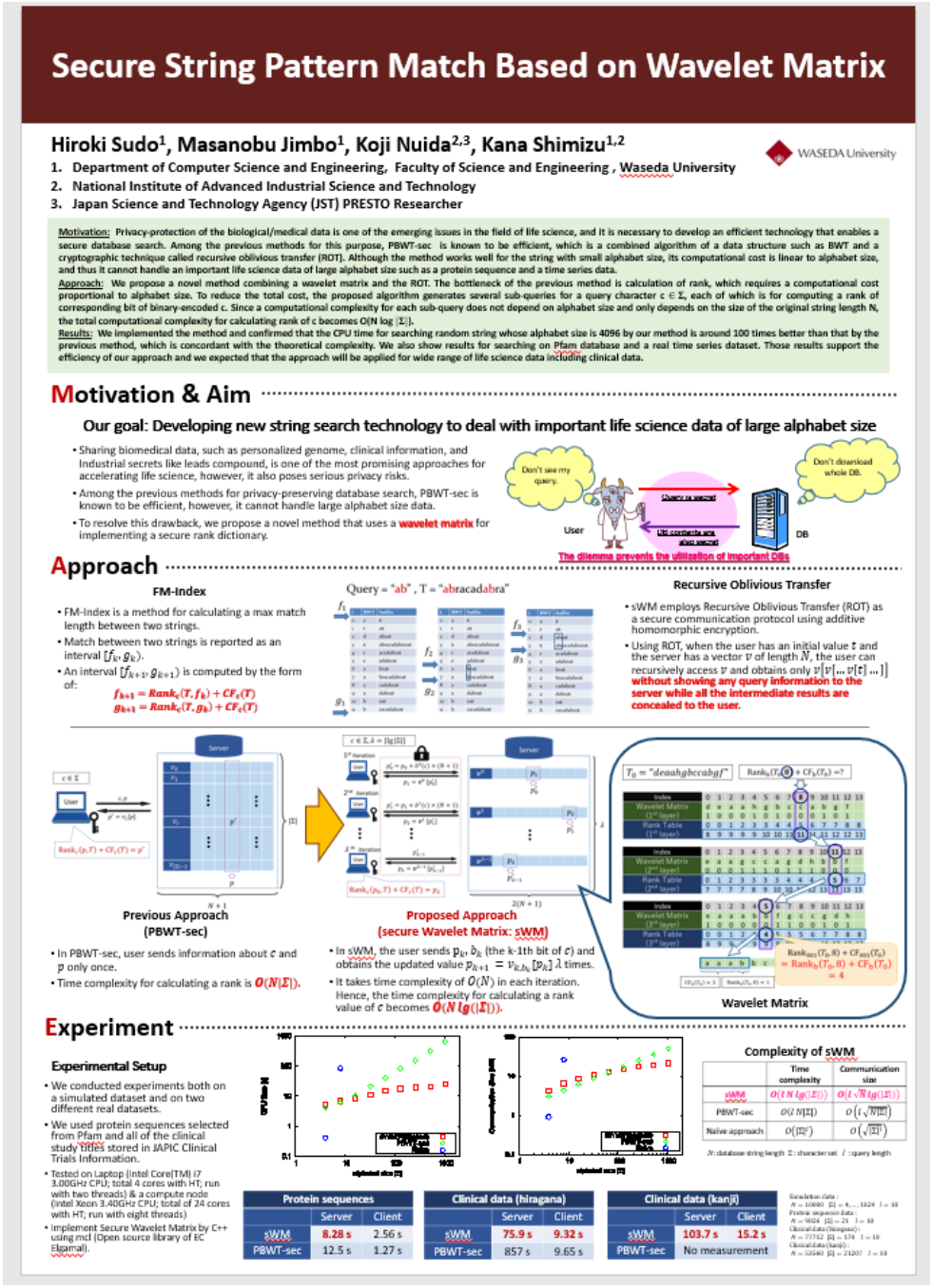
4つ前の発表終了時に
このあたりで待機

ミニポスター発表

- 発表者側の準備
  - A3用紙2枚に発表内容をまとめる．（ミニポスターと呼ぶ）
  - 自分のグループの机にミニポスターを置く
- 聴衆は，発表を聞きたいグループの場所に行く．
- 発表者は，来訪する聴衆に対して，ミニポスターを使って説明をする．補助的にノートPCを使用してもよい．

ポスターの例

- 例は通常のポスターですが，今回は都合上，A3サイズ2枚で作成してください．
- なお，日本語でOKです．
- 全グループのポスターはMoodle上で事前に共有します．



Secure String Pattern Match Based on Wavelet Matrix

Hiroki Sudo[1], Masanobu Jimbo[1], Koji Nuida[2,3], Kana Shimizu[1,2]

1. Department of Computer Science and Engineering, Faculty of Science and Engineering , Waseda University
2. National Institute of Advanced Industrial Science and Technology
3. Japan Science and Technology Agency (JST) PRESTO Researcher

WASEDA University

Motivation: Privacy-protection of the biological/medical data is one of the emerging issues in the field of life science, and it is necessary to develop an efficient technology that enables a secure database search. Among the previous methods for this purpose, PBWT-sec is known to be efficient, which is a combined algorithm of a data structure such as BWT and a cryptographic technique called recursive oblivious transfer (ROT). Although the method works well for the string with small alphabet size, its computational cost is linear to alphabet size, and thus it cannot handle an important life science data of large alphabet size such as a protein sequence and a time series data.

Approach: We propose a novel method combining a wavelet matrix and the ROT. The bottleneck of the previous method is calculation of rank, which requires a computational cost proportional to alphabet size. To reduce the total cost, the proposed algorithm generates several sub-queries for a query character $c \in \Sigma$, each of which is for computing a rank of corresponding bit of binary-encoded c. Since a computational complexity for each sub-query does not depend on alphabet size and only depends on the size of the original string length N, the total computational complexity for calculating rank of c becomes $O(N \log |\Sigma|)$.

Results: We implemented the method and confirmed that the CPU time for searching random string whose alphabet size is 4096 by our method is around 100 times better than that by the previous method, which is concordant with the theoretical complexity. We also show results for searching on Pfam database and a real time series dataset. Those results support the efficiency of our approach and we expected that the approach will be applied for wide range of life science data including clinical data.

Motivation & Aim

Our goal: Developing new string search technology to deal with important life science data of large alphabet size

- Sharing biomedical data, such as personalized genome, clinical information, and industrial secrets like leads compound, is one of the most promising approaches for accelerating life science, however, it also poses serious privacy risks.
- Among the previous methods for privacy-preserving database search, PBWT-sec is known to be efficient, however, it cannot handle large alphabet size data.
- To resolve this drawback, we propose a novel method that uses a **wavelet matrix** for implementing a secure rank dictionary.

Don't see my query. / Don't download whole DB. / Query is secret / DB contents are also secret / User / DB

The dilemma prevents the utilization of important DBs

Approach

FM-Index

- FM-Index is a method for calculating a max match length between two strings.
- Match between two strings is reported as an interval $[f_k, g_k)$.
- An interval $[f_{k+1}, g_{k+1})$ is computed by the form of:

$$f_{k+1} = Rank_c(T, f_k) + CF_c(T)$$
$$g_{k+1} = Rank_c(T, g_k) + CF_c(T)$$

Query = "ab" , T = "abracadabra"

Recursive Oblivious Transfer

- sWM employs Recursive Oblivious Transfer (ROT) as a secure communication protocol using additive homomorphic encryption.
- Using ROT, when the user has an initial value t and the server has a vector v of length N, the user can recursively access v and obtains only $v[v[\dots v[t]\dots]]$ **without showing any query information to the server while all the intermediate results are concealed to the user.**

Previous Approach (PBWT-sec)

- In PBWT-sec, user sends information about c and p only once.
- Time complexity for calculating a rank is $O(N|\Sigma|)$.

Proposed Approach (secure Wavelet Matrix: sWM)

- In sWM, the user sends p_k, b_k (the k-1th bit of c) and obtains the updated value $p_{k+1} = v_{k,b_k}[p_k]$ λ times.
- It takes time complexity of $O(N)$ in each iteration. Hence, the time complexity for calculating a rank value of c becomes $O(N \lg(|\Sigma|))$.

T_0 = "deaahgbccabgf"

Wavelet Matrix

Experiment

Experimental Setup

- We conducted experiments both on a simulated dataset and on two different real datasets.
- We used protein sequences selected from Pfam and all of the clinical study titles stored in JAPIC Clinical Trials Information.
- Tested on Laptop (Intel Core(TM) i7 3.00GHz CPU; total 4 cores with HT; run with two threads) & a compute node (Intel Xeon 3.40GHz CPU; total of 24 cores with HT; run with eight threads)
- Implement Secure Wavelet Matrix by C++ using mcl (Open source library of EC Elgamal).

Complexity of sWM

	Time complexity	Communication size
sWM	$O(lN\lg(\lvert\Sigma\rvert))$	$O(l\sqrt{N}\lg(\lvert\Sigma\rvert))$
PBWT-sec	$O(lN\lvert\Sigma\rvert)$	$O\left(l\sqrt{N\lvert\Sigma\rvert}\right)$
Naïve approach	$O(\lvert\Omega\rvert^l)$	$O\left(\sqrt{\lvert\Omega\rvert^l}\right)$

N: database string length Σ: character set l : query length

Protein sequences	Server	Client
sWM	8.28 s	2.56 s
PBWT-sec	12.5 s	1.27 s

Clinical data (hiragana)	Server	Client
sWM	75.9 s	9.32 s
PBWT-sec	857 s	9.65 s

Clinical data (kanji)	Server	Client
sWM	103.7 s	15.2 s
PBWT-sec	No measurement	

ミニポスター作成のポイント

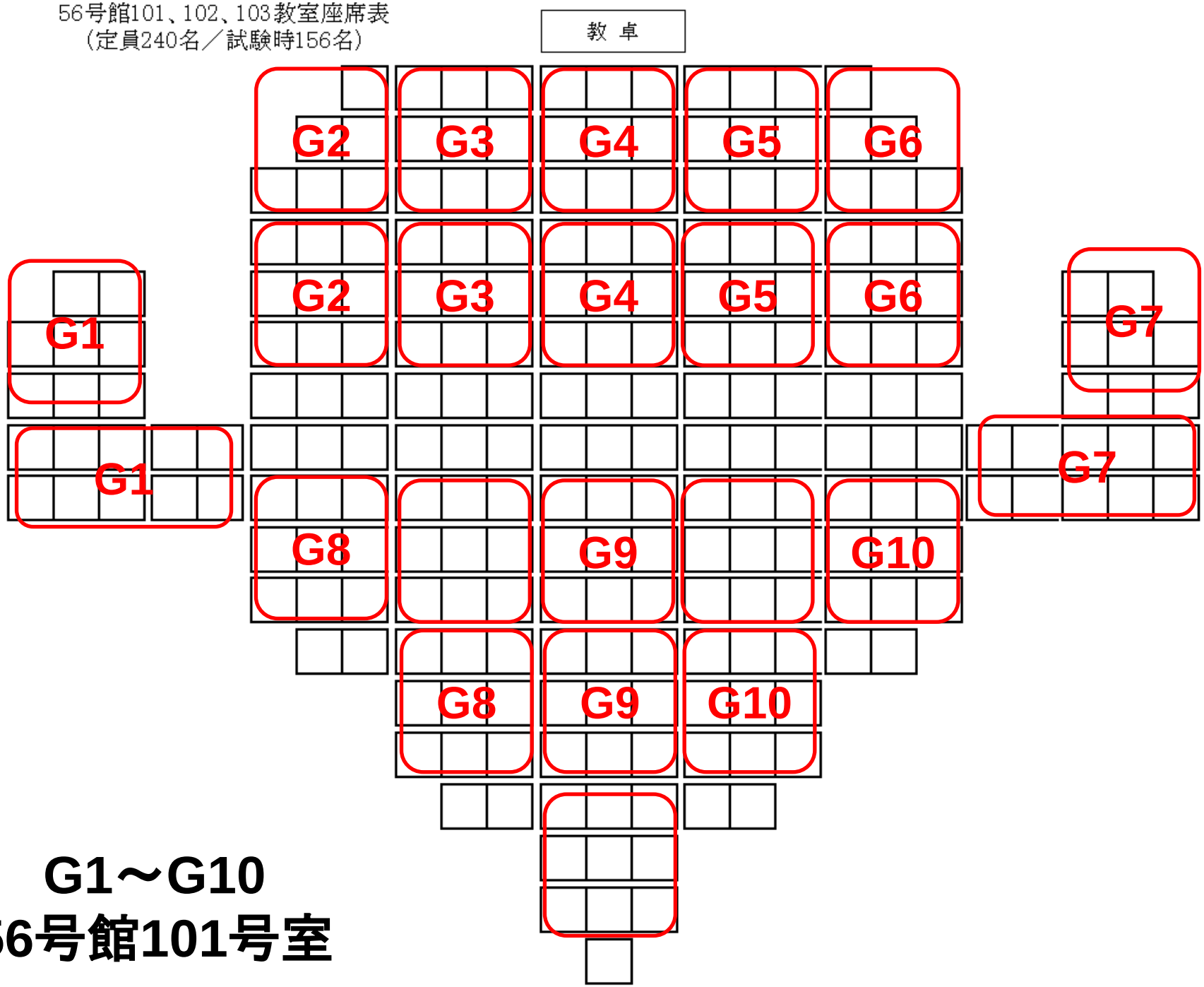
- 口頭で説明するための材料なので，文章は短めに.
- 説明に必要な図やグラフ等を中心に書く.
- 全員が知っている問題についての発表なので，今回は背景や問題の説明等は不要.
- 一度に2～3人と議論するための資料であることを想定する.

発表で心掛けてほしいこと

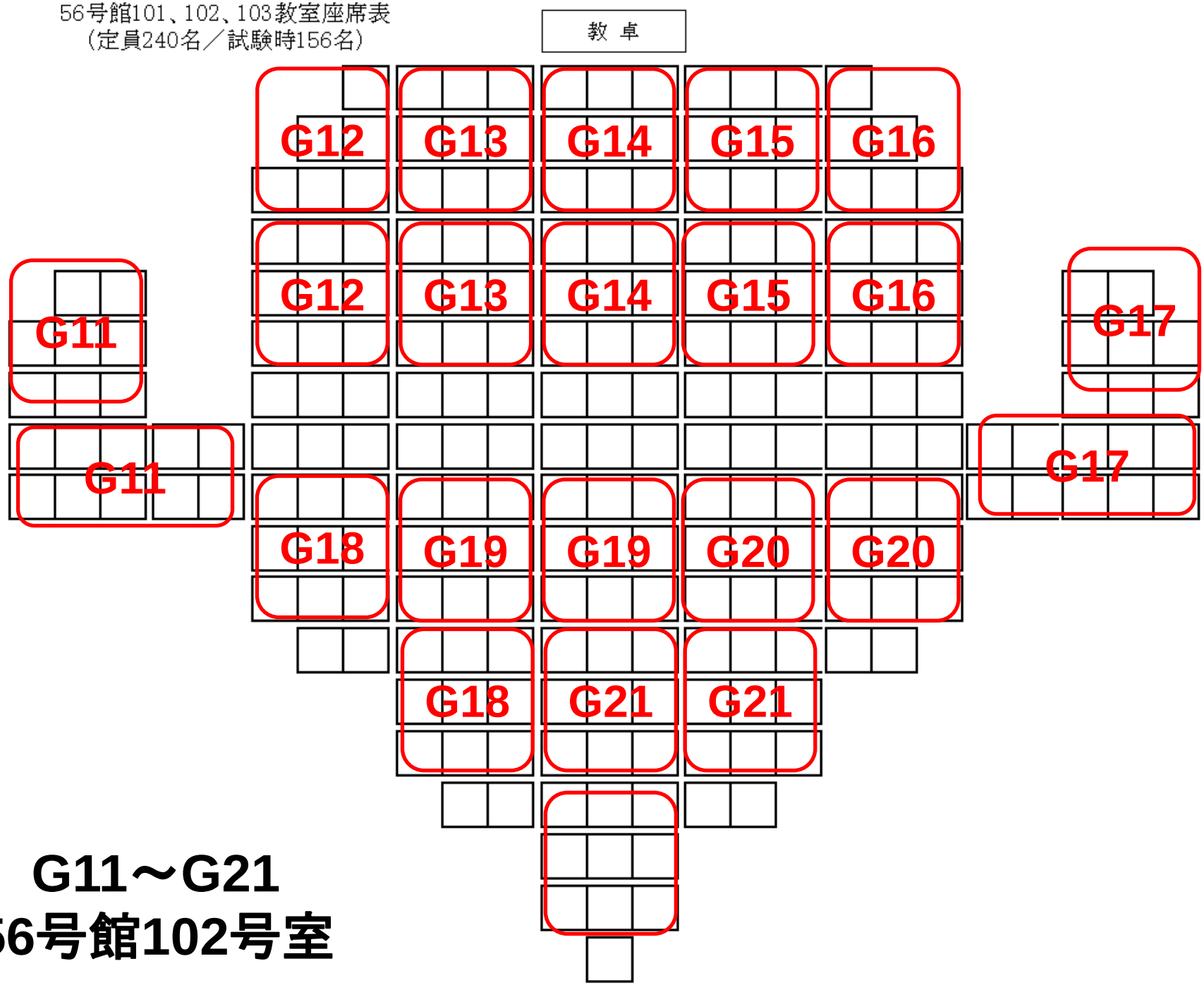
- 発表に含めるべき内容
  - どのような方針で取り組んだのか？
  - その方針を実現するためにどのような方法論を用いたのか？その方法論を用いた根拠は？
  - 実際にそれはうまくいったのか？
  - うまくいった（若しくはうまくいかなかった）要因の分析など
- ストーリーがあって、説得力のある発表を心がける事。
  - 例えば…「〇〇による予備調査の結果、〇〇の方針でやるのが良いと考え、それを実現するためには〇〇であることから、〇〇の方法と〇〇の方法を考案した。それらを実装して性能をテストしたところ、〇〇のような結果が得られた。結果を分析したところ〇〇の部分が〇〇であったが、それはデータの性質が〇〇で〇〇の方法がよく適合したためだと考えられる。」

ミニポスター発表スケジュール

- タイムスロット：15:25～16:05: スロット１，16:05～16:45: スロット２，16:45～17:25: スロット３
- 各グループに２つの場所を用意します．
  - ミニポスターは各グループ２セット用意する．
- 各学生は，**必ず１スロットを単独で発表**する．
  - 6人グループの場合，各スロットで2名ずつが別の場所で発表することになる．
- 発表の順番は，各グループの裁量で決める．
  - 欠席が発生し発表者が4人以下の場合は，全てのスロットができるだけ埋まるように配置してください．
- 発表以外の時間は，別グループの発表を聞く．



G1～G10
56号館101号室



G11～G21

56号館102号室

提出〆切＆作業スケジュール

- 2022/12/16, 23:59
  - 提出物：　キックオフメモ
  - 提出先：　Moodle 「キックオフメモ」
- 2022/12/23, 23:59
  - 提出物：　12月23日報告
  - 提出先：　Moodle 「 12月23日報告」
- 2023/1/13, 23:59
  - 提出物：　1月13日報告
  - 提出先：　Moodle 「 1月13日報告」
- 2022/12/26, 23:59 （オプション）
  - 提出物：　中間計測用のプログラム群（cファイル1つ）
  - 提出先：　Moodle 「中間計測用提出」
- 2023/1/15, 23:59
  - 提出物：　最終評価用のプログラム群（ cファイル1つ）
  - 提出先：　Moodle 「成果物提出」

提出〆切＆作業スケジュール（続き）

- 2023/1/**18**, 23:59 ← 一日延長
  - 提出物：　発表資料
  - 提出先：　Moodle 「成果発表会用資料提出」
- 2023/1/20, 23:59
  - 作業：　発表評価（Moodle 「発表評価」）
    ※ ただし，グループ発表評価は授業時間中に行うこと．
- 2023/1/27, 23:59
  - 提出物：　報告書
  - 提出先：　Moodle
  - 作業：

- ライトニングトークのスライドは，ppt/pptxもしくは pdfで提出すること．
- ミニポスターはpdf形式で提出すること．

ミニポスターの印刷について

- 〆切までに提出されたpdfについては，教員側で印刷します．（カラーを予定）
  - 教員側での印刷が不要の場合はお知らせください．
- ポスターは，1月20日12:10～13:00までの間，清水研（63号館05-01室）の前の段ボールに置きますので各グループの代表者が受け取ってください．（持っていく際に特に断る必要はありません．）

良い発表を期待しています！