TheYeastMap

生物学のための、検索ではない
知識探索方法の確立

環境生命科学研究科
D2 佐伯望

今日の内容

1. 研究のキッカケ

２. 研究動機　私の研究領域が抱える問題点　⬅ここを重点

3. 研究内容　TheYeastMap

4. 質問・コメント

今日の目標

私が「問題として考えていること、解決したいこと」を伝える

佐伯　望　(サエキ　ノゾム)

出身　兵庫県豊岡市但東町, 日本

学士　神戸市立工業高等専門学校　応用化学専攻
「量子化学計算」

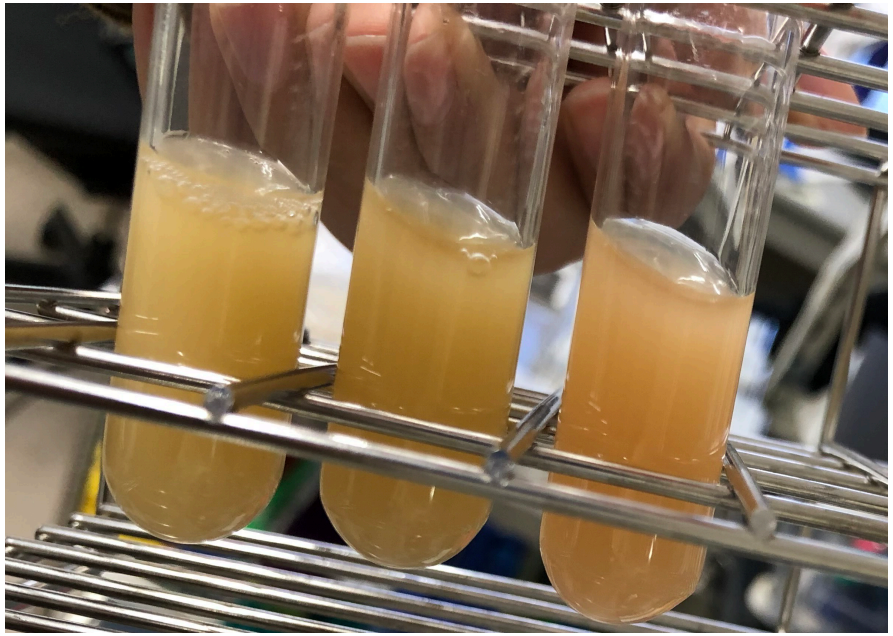
**生物学に興味があり、
システムズバイオロジーという
研究領域を知る**

修士　岡山大学　環境生命科学研究科
「システムズバイオロジー」

現在　岡山大学　環境生命科学研究科　博士課程在学中　JSPS DC
「システムズバイオロジー」と「細胞のストレス適応」

研究のキッカケ

研究は実験中心



〜修士時代
お金がなかったので、
自然言語処理やコンピューター関係の
仕事を個人事業主としてやっていた

コロナ時代　2020年
研究室がコロナで閉鎖されたため、
家でできる研究活動として、
自然言語処理を研究に応用してみたら、
指導教官の受けが良かった

現在　2021年
Cypherに応募して、採択された
副研究テーマとして研究を開始

データサイエンス、AI

自然言語処理
Natural Language Processing

自然言語(日本語や英語)をコンピューターに処理させる技術。

EX)
検索エンジン（Google、百度 etc）
機械翻訳（Google翻訳、DeepL etc）
音声認識 （Siri, Alexa etc）

本研究では、主に「統計的自然言語処理」、「情報抽出」の分野を扱う。

自然言語処理とAI

特に科学において期待されること

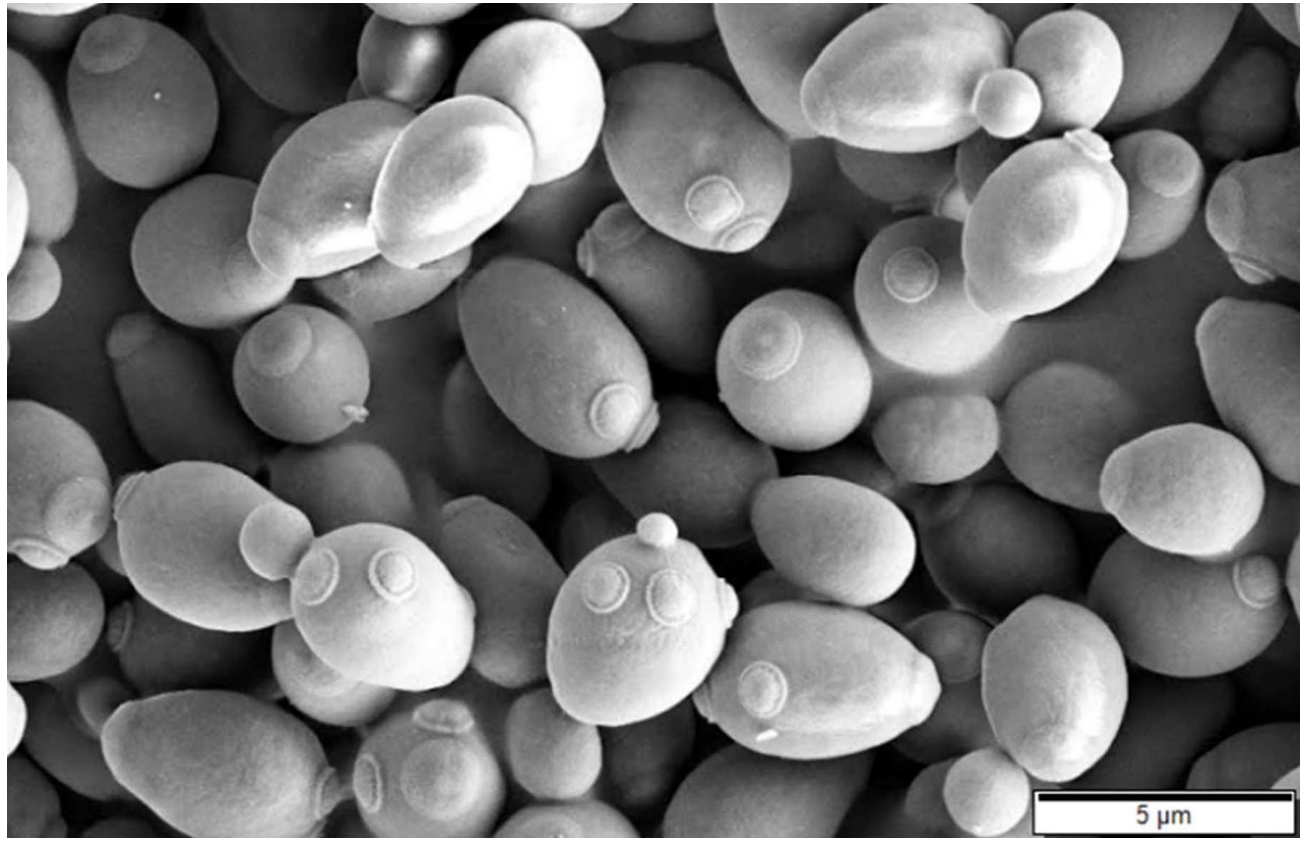
・論文の査読の自動化
・AIによる科学的知識の発見

「各論文を読み込み、重要な概念を識別する」
「キーワードを概念、区分ごとに整理する」
「さまざまなキーワード、キーフレーズの間の関係性を特定する」

現在、さまざまな研究がなされているが、2022年現在では本格的な実用化にはまだ至っていない。

画像認識の分野と比べて、まだ日の目を見ていない分野。

研究対象　「出芽酵母」



Wikipedia, “Saccharomyces cerevisiae”

食品

お酒、パン　など

基礎研究

「真核細胞のモデル」

EX)

1996年　真核生物初の全ゲノム決定

Goffeau et al Science 2016

2001年　細胞周期　ノーベル賞

Nurse博士、Hartwell博士が酵母から

Hunt博士はウニから

2016年　オートファジー　ノーベル賞

大隅博士

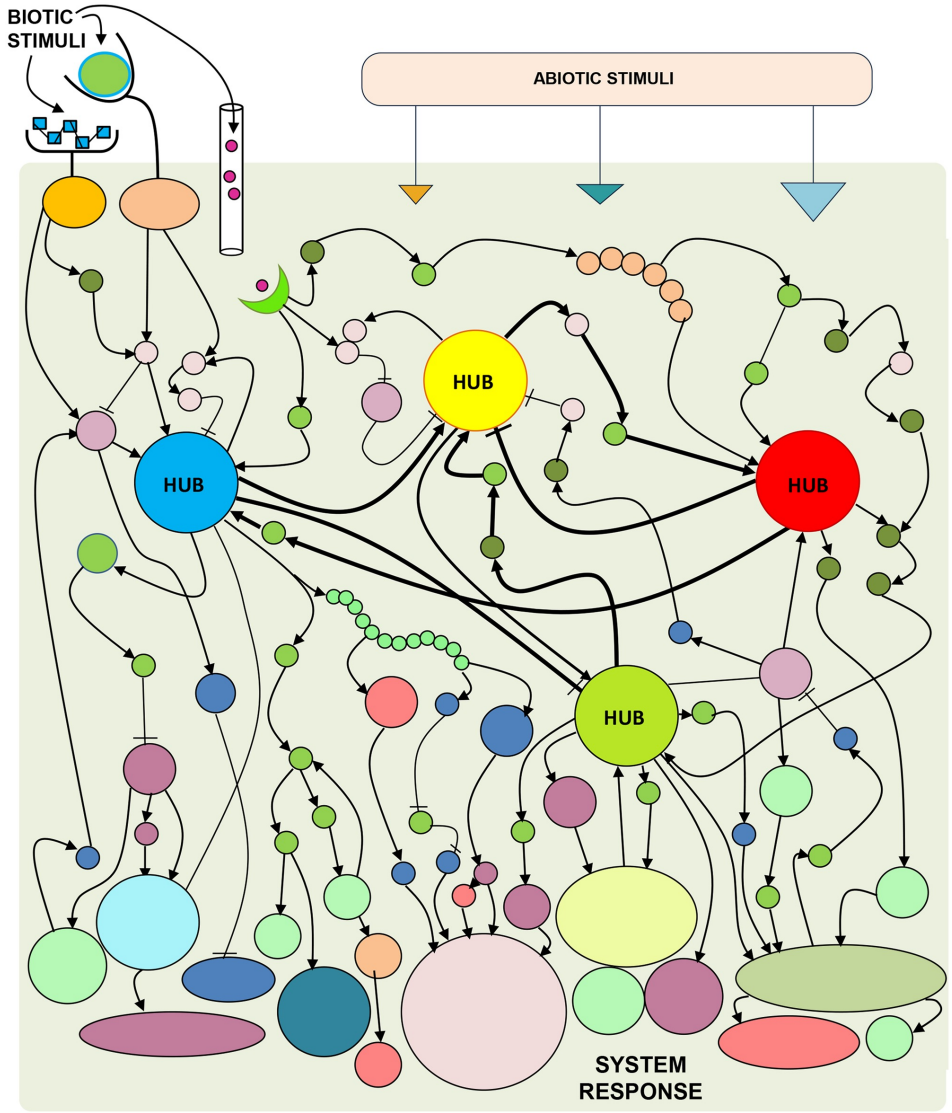
研究対象 「出芽酵母」

Life with 6000 Genes

A. Goffeau,* B. G. Barrell, H. Bussey, R. W. Davis, B. Dujon,
H. Feldmann, F. Galibert, J. D. Hoheisel, C. Jacq, M. Johnston,
E. J. Louis, H. W. Mewes, Y. Murakami, P. Philippsen,
H. Tettelin, S. G. Oliver

出芽酵母は分子生物学の発展の中で、次々に遺伝子の機能が調べられ、報告されてきた。2022年現在、出芽酵母の約6000遺伝子のうち、90%以上の遺伝子は機能や表現型などの情報が付与されている。出芽酵母では分子生物学は終わったと述べる先生もおられる。

生命（細胞）の統合的理解を目指して
ーシステムズバイオロジー



Hillmer, PLOS Pathog, 2015

生命（細胞）の統合的理解を目指して
ーシステムズバイオロジー

シーケンス技術、オミクス技術、
バイオインフォマティクス、ロボティクス、ラボオートメーションの発展

日々、膨大なデータが生み出されている

WORLD VIEW | 13 September 2021

Biology must generate ideas as well as data

Data should be a means to knowledge, not an end in themselves.

Paul Nurse

Accepting a Nobel prize nearly two decades ago, my old friend Sydney Brenner had a warning for biology. "We are drowning in a sea of data and starving for knowledge," he said. That admonishment, from one of the founders of molecular biology, who established the nematode worm *Caenorhabditis elegans* as a model organism, is even more relevant to biology today.

我々はデータの海に溺れ
知識に飢えている

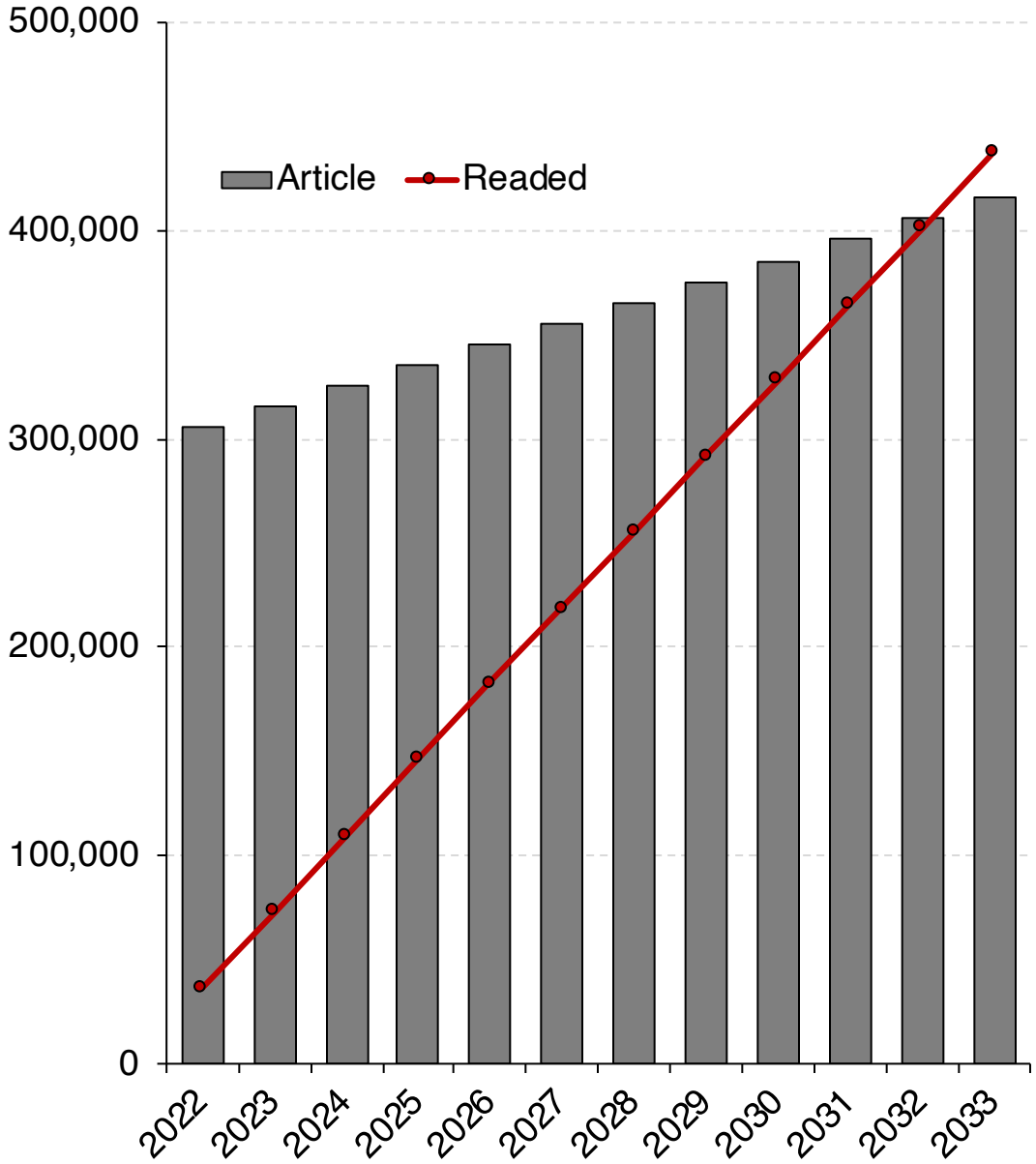
論文の数も増え続けている

EX)
PubMedで “Yeast” と検索すると
305,634 results （20220124)

さらに、
毎年1万報以上のペースで増加している

365日毎日100報論文を読むと、
(不眠不休で1報あたり15分以内)
約10年で追いつける

酵母の論文だけでは、研究はできない
本当はもっと膨大な文献がある



論文の数も増え続けている

EX)
PubMedで “Cell” と検索すると
7,497,918 results （20220124)

さらに、
毎年30万報以上のペースで増加し、年々加速しているている

365日毎日3000報論文を読むと、
(不眠不休で1報あたり30秒以内)
約10年で追いつける

膨大な情報を生かしきれているか?
膨大な情報量に圧倒されていないか?

膨大なデータと文献の “Tsunami”に呑み込まれず
統合し、知識とするにはどうすれば良いのか?

データサイエンス・AIの利用は
この問題を解決にならないか??

＊“Tsunami”の表現は、前述の記事でNurse博士が用いた比喩

データサイエンス・AIの利用は
情報のTsunami問題を解決にならないか??

我々が持つ2つの情報学ツール

・検索
・データベース

検索とは

けん-さく【検索】
[名](スル)調べて探しだすこと。特に、文献・カード・ファイル・データベース・インターネットなどの中から必要な情報を探すこと。「検索の便を図る」「索引で関係事項を検索する」 デジタル大辞林

search /səːtʃ/ ♫
▶ verb [no object] try to findsomething by looking or otherwiseseeking carefully and thoroughly:
Oxford Dictionary of English

膨大な情報から、必要な情報を見つけ出す上では有用。
情報の統合はできない。

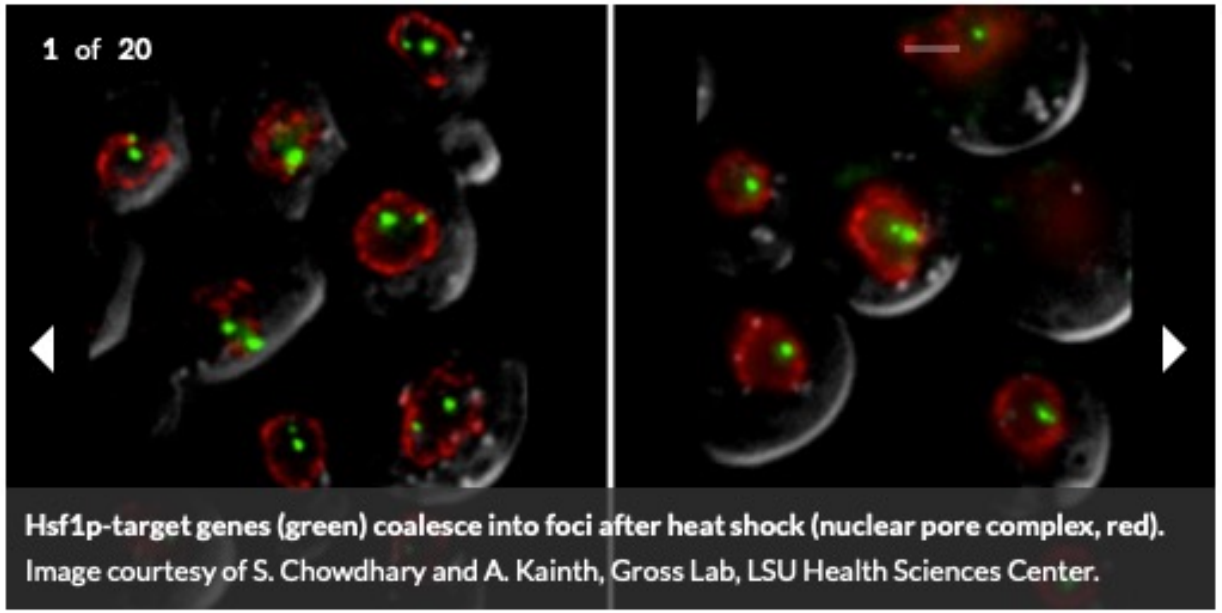
データベース　Saccharomyces Genome Database

膨大な情報が、各遺伝子ごとに整理されている。

SGD Saccharomyces GENOME DATABASE

Analyze　Sequence　Function　Literature　Community

search: actin, kinase, glu



Hsf1p-target genes (green) coalesce into foci after heat shock (nuclear pore complex, red).
Image courtesy of S. Chowdhary and A. Kainth, Gross Lab, LSU Health Sciences Center.

About SGD

The *Saccharomyces* Genome Database (SGD) provides comprehensive integrated biological information for the budding yeast *Saccharomyces cerevisiae* along with search and analysis tools to explore these data, enabling the discovery of functional relationships between sequence and gene products in fungi and higher organisms.

Explore SGD

PERSPECTIVE | 27 SEPTEMBER 2021

Model organism databases are in jeopardy FREE

Hugo J. Bellen, E. J. A. Hubbard, Ruth Lehmann, Hiten D. Madhani, Lila Solnica-Krezel, E. Michelle Southard-Smith

+ Author and article information

Development (2021) 148 (19): dev200193.

https://doi.org/10.1242/dev.200193

Model organisms (MOs), including yeast, worm (*C. elegans*), fruit fly (*Drosophila*), zebrafish, frog (*Xenopus*), mouse and rat, contribute greatly to our understanding of human development and disease. To be successful, MO research critically depends on many shared resources. Particularly important are MO stock centers and [illegible] focuses on [illegible] especially the National Human Genome Research Institute (NHGRI).

助成金予算の大幅な削減により、これらの重要なデータベースへの支援が危ぶまれていることを深く憂慮しています

We are deeply concerned that the support for these vital databases is in jeopardy due to large cuts in their grant budgets. We fear these budget cuts will slow biomedical research worldwide and create increased waste of resources due to duplication of efforts. Indeed, the cuts threaten to erode access to reliable, expertly fact-checked data and cause an increase in mis-information due to the degraded organization of knowledge and information.

データベースが持つ二つの危機

1. 人の手で管理されており、対応が追いついていない。
2. データベース管理のための予算が近い将来カットされる。

データサイエンス・AIの利用は
情報のTsunami問題を解決にならないか??

我々が持つ2つの情報学ツール
・検索エンジン
・データベース

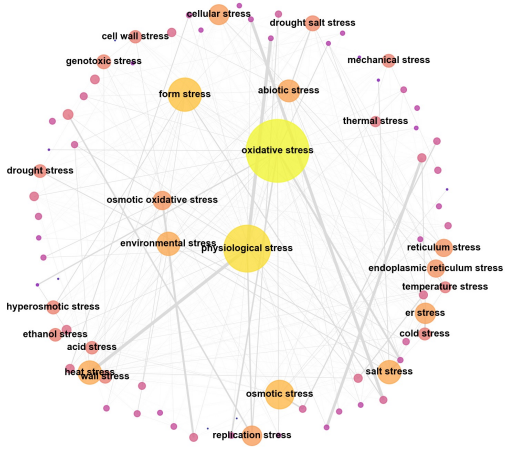
私が提案する新しいツール
・文献のメタ解析ツール “TheYeastMap”
　今の自然言語処理技術でできることをやる

酵母論文の
メタ分析ツール
TheYeastMap

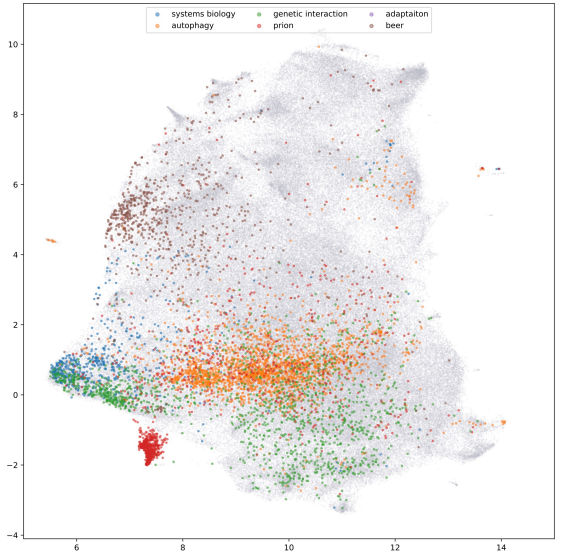
Yeastに関する約300,000報
の論文情報から
自然言語処理を用いることで
語句、概念、流行変遷等を
解析するツール

解析例

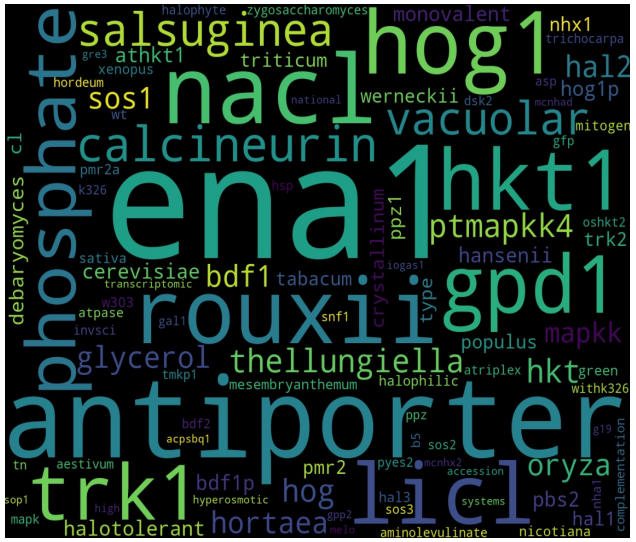
各生理学的ストレスの関連性



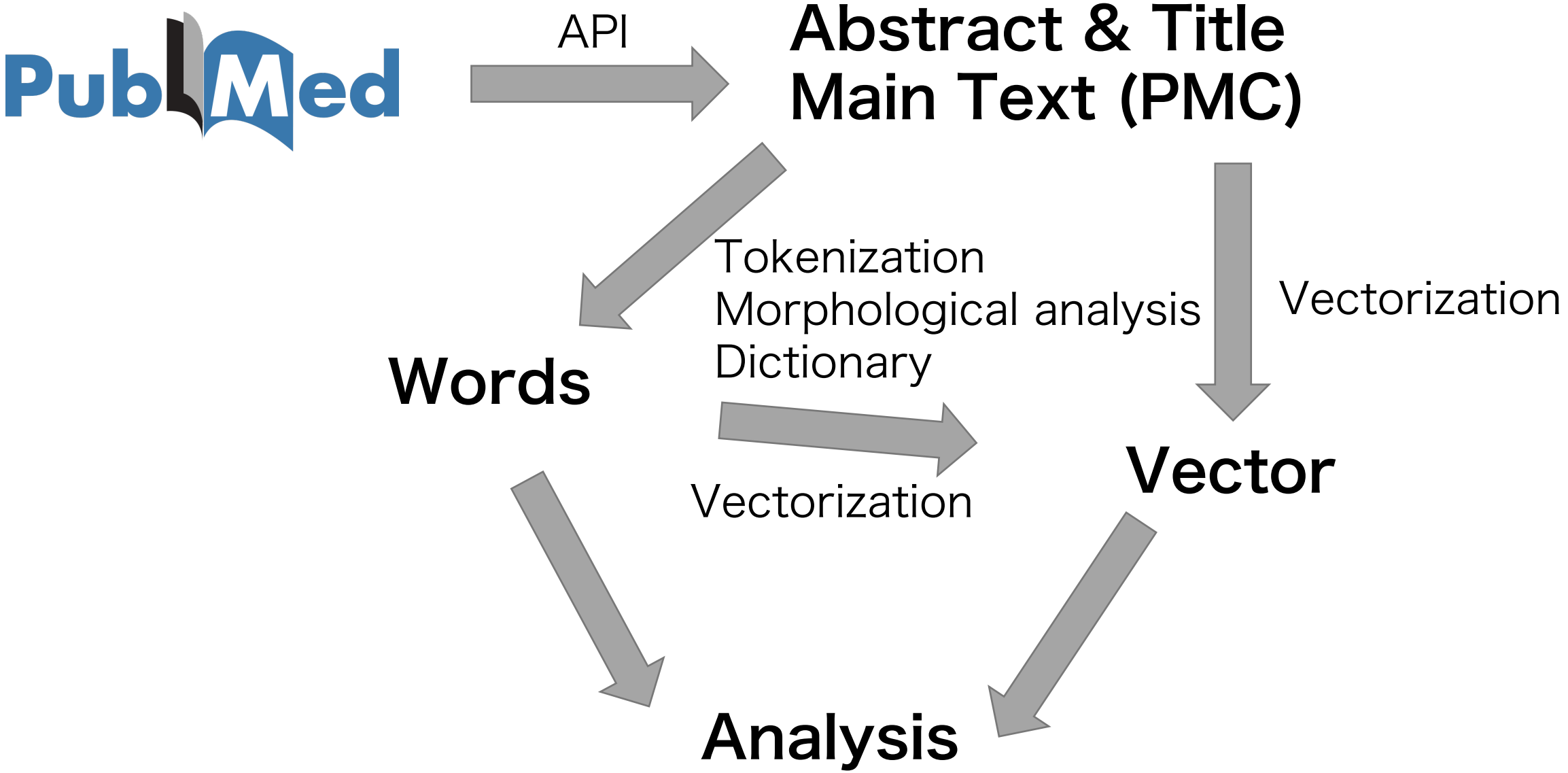
各種キーワードを持った文献の位置関係



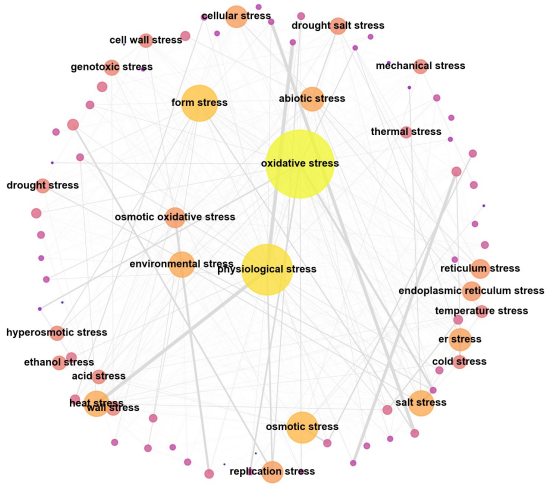
“Salt stress”と関係性の深い語句群

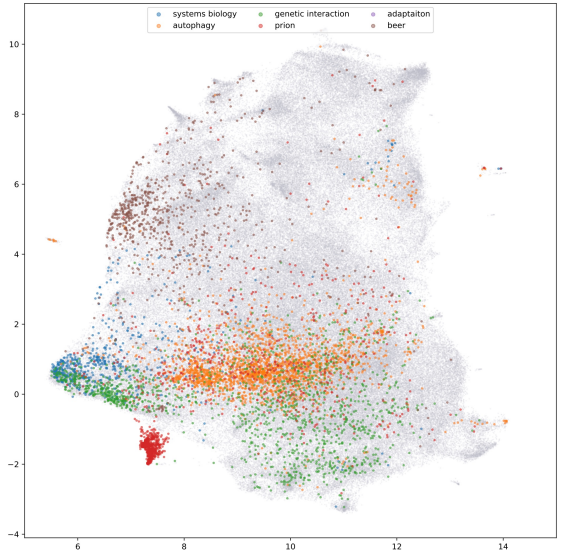


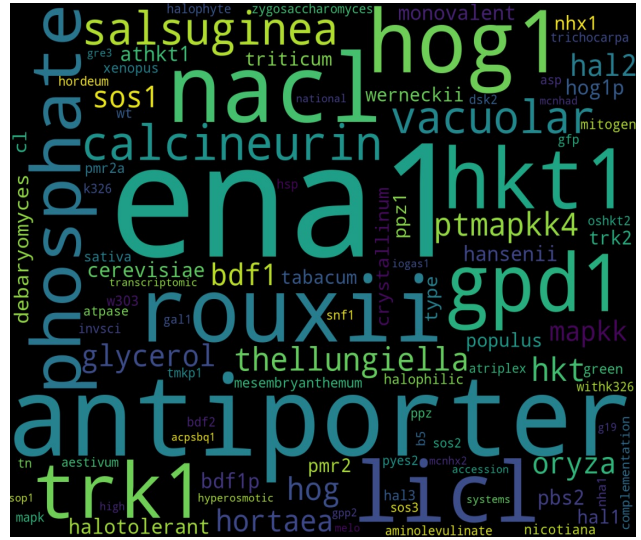
解析手順: 概要


PubMed
API
Abstract & Title
Main Text (PMC)
Tokenization
Morphological analysis
Dictionary
Vectorization
Words
Vectorization
Vector
Analysis

現在の課題







“TheYeastMap”を使って新たな知識を発見できるか?

野心: データベース構築の枠組みが作れないか

まとめ

膨大なデータと文献の “Tsunami”に呑み込まれず、
統合し、知識とするにはどうすれば良いのか?

文献のメタ解析ツール
“TheYeastMap”を開発した

“TheYeastMap”を使って新たな知識を発見できるか?

謝辞

岡山大学　竹内孔一先生
静岡大学　山本泰生先生

質問、コメントをよろしくお願いします