

# UDAC研究活動の紹介

データアナリティクス研究部門 教授 志賀元紀

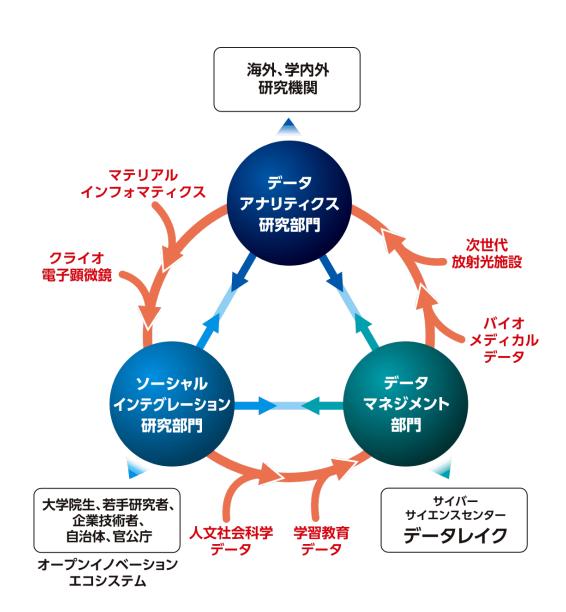


## データアナリティクス研究部門

#### 1. 最先端のデータ解析法の開発

- 大規模なサイズ
- 多様な形式
- ・応用に適した解析・理論

- 2. 学内外・応用分野との連携 物質材料科学、分子生物学など (ゲノム科学、医療創薬)
- 3. 社会実装に向けて





## 多様な形式のデータ

- 表形式などの構造化データ
- 高次元配列(テンソル)データ
- グラフ構造

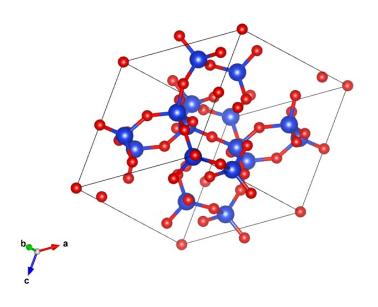
#### 文字列(ゲノム配列)



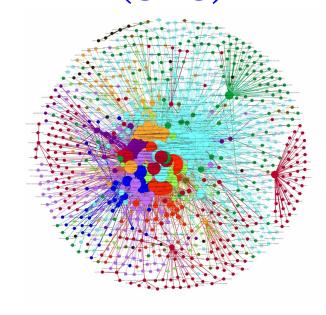
#### 多次元配列(顕微計測)

# 成分スペクトル 成分1 成分2 エネルギー損失(eV)

#### グラフ構造 (化合物)



(SNS)



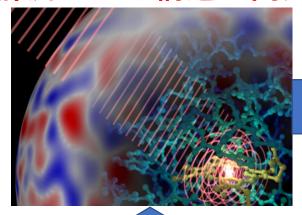
- データ形式に合わせた解析理論
- ドメイン知識を取り入れた応用解析

の両立が必要不可欠!

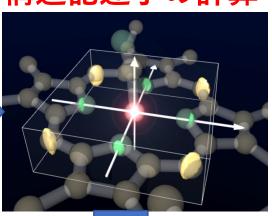


# 物質・材料科学におけるデータ解析

#### 計測による構造の同定



#### 構造記述子の計算



#### 新規材料合成・探索



#### 物性予測と自動設計

元素A C

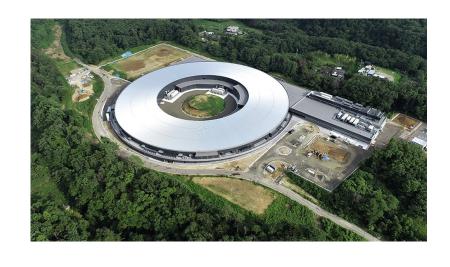
#### データ科学の役割:

- 計測データから 精密かつ効率的な構造同定
- **構造データ**から 構造秩序・物性に関する 記述子の設計・計算
- 構造記述子から 正確な物性予測、さらに、 効率的な実験計画の提示



# 微細構造に対する様々な計測技術

• 放射光 (軟X線 or 硬X線) 、 中性子線、電子線の相補的な利用 それぞれメリット・デメリット、得意な材料が異なる

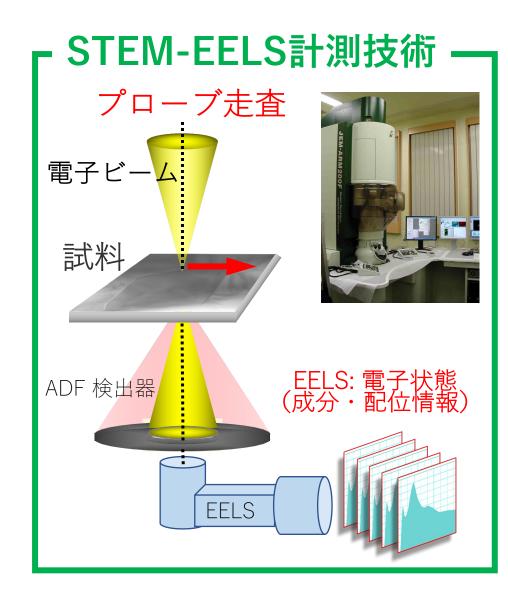




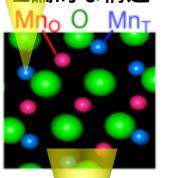
- ・様々な計測データ形状 実空間像、特性スペクトル、回折像(波形、2次元図形)
- 様々なノイズレベル S/Nの良いものがあれば、非常に悪いものもある

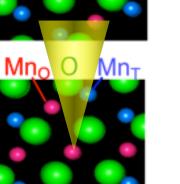


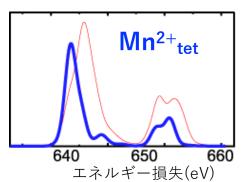
# 電子顕微鏡による計測データの例

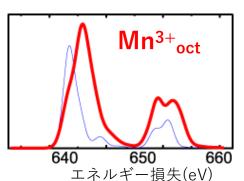












局所構造(原子・分子・配位数)

⇒ スペクトル形状が対応する。

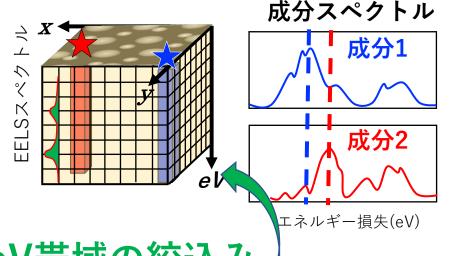
スペクトル形状⇔局所構造 を予測できる。



## 機械学習による自動解析が必要!

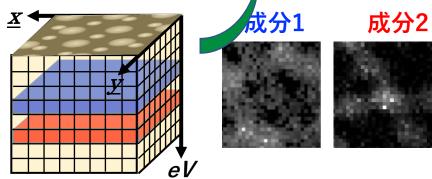
#### マニュアル情報抽出の場合:

#### ①成分位置の絞込み



異なる成分の領域をピックアップして 成分ごとのスペクトルを同定する。

②eV帯域の絞込み

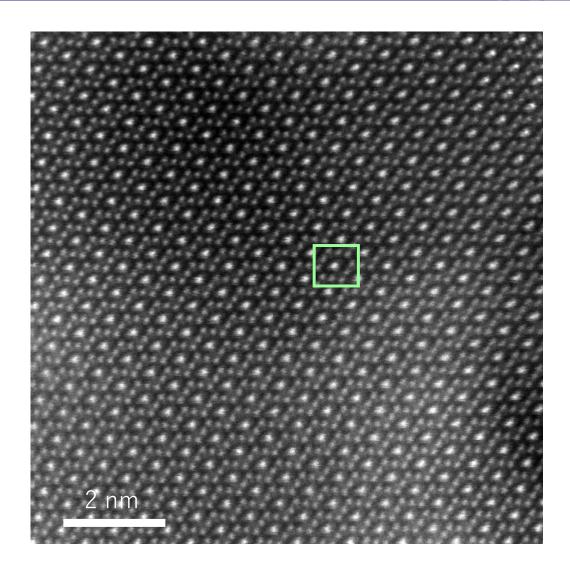


成分ごとの特徴的なピークを見つけピーク強度に基づき空間分布を描く。

S/Nの悪いデータに対して、機械学習による情報抽出が必要

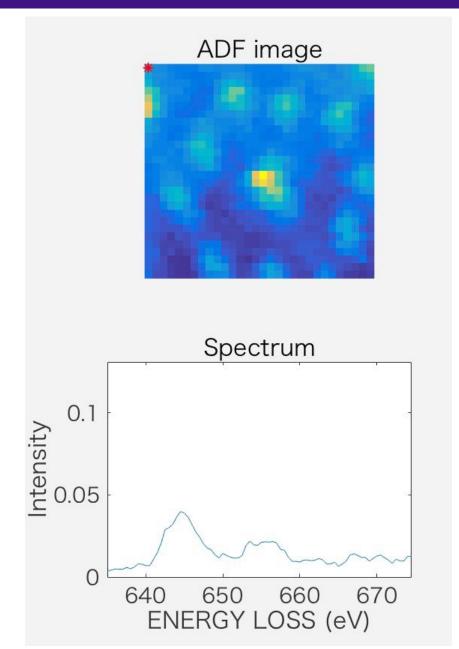


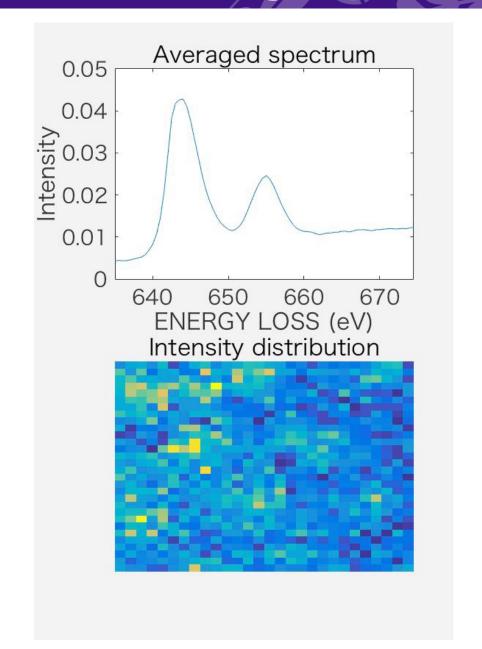
# 原子解像度STEM-EELS-SI (Mn<sub>3</sub>O<sub>4</sub>)





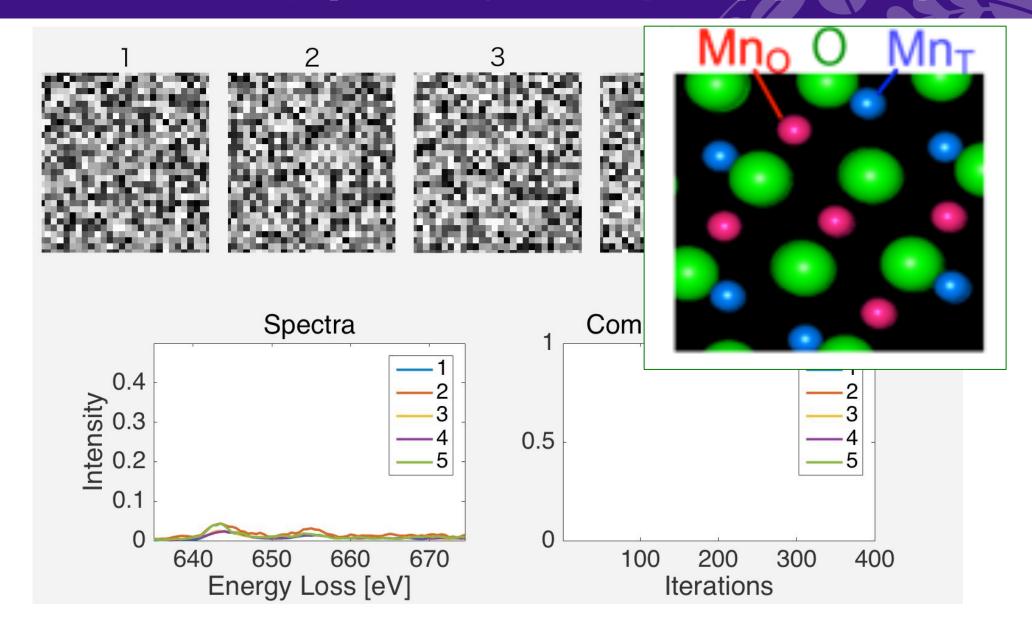
# 原子解像度STEM-EELS (Mn<sub>3</sub>O<sub>4</sub>)







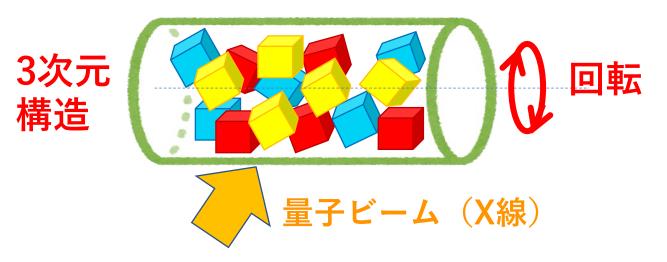
# 開発法による化学成分分布の自動解析



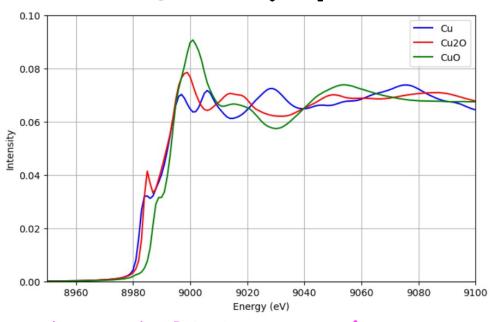


## 放射光計測データの例

#### CT- XAFS計測データ



#### XAFSスペクトル



#### <u>試料を回転させながらXAFSイメージング計測するデータ</u>

4次元のビッグデータ (=xy平面×スペクトル帯域×回転軸)
 1計測のデータサイズは約200GB

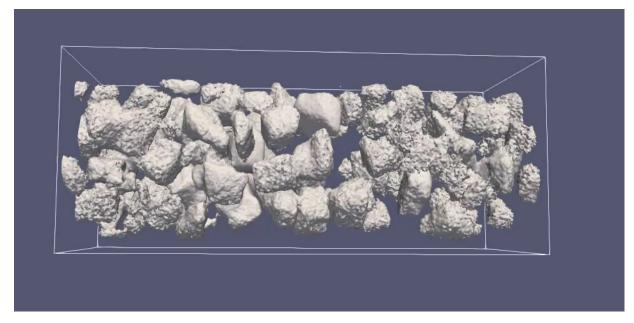
・ 計測データから3次元再構成+成分同定が必要

株式会社デンソーとの共同研究(あいちシンクロトロン光センターで計測)

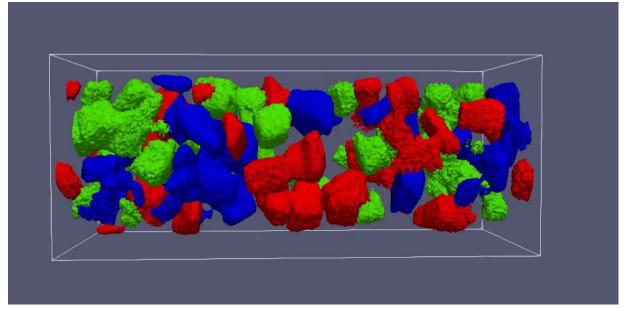


#### 3次元構造の同定結果 (開発中の機械学習法)

#### 単色CTによる再構成結果 (約1GBのデータ)



CT-XAFSによる再構成結果 (約200GBのデータ)



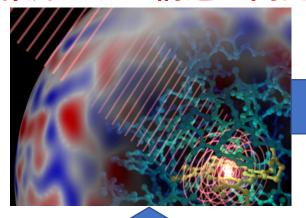
3次元構造のみを同定できる。 (化学成分を判定できない) XAFSの特性スペクトルによって、 化学成分の3次元配置も同定できる

さらに、従来手順より数十倍以上も高速な解析法を実現しつつある。

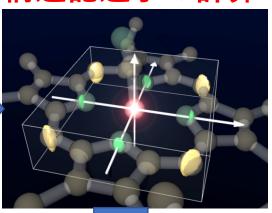


# 物質・材料科学におけるデータ解析

#### 計測による構造の同定



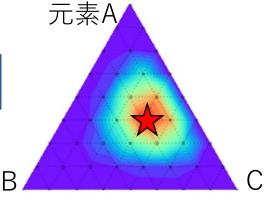
#### 構造記述子の計算



#### 新規材料合成・探索



#### 物性予測と自動設計



#### データ科学の役割:

- **計測データ**から 精密かつ効率的な構造同定
- **構造データ**から 構造秩序・物性に関する 記述子の設計・計算
- 構造記述子から 正確な物性予測、さらに、 効率的な実験計画の提示



## マテリアルズインフォマティクス人材育成が急務!



材料科学者







データ科学者

マテリアルズインフォマティシャン

=材料科学とデータ科学のどちらにも精通した研究者

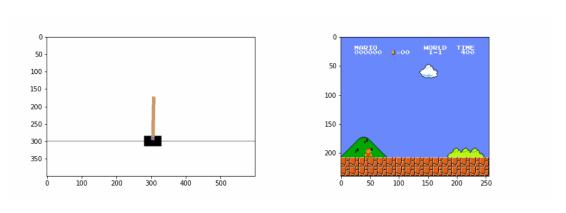


#### 山田和載教授の研究紹介

ソーシャルインテグレーション研究部門所属.

2022年6月まで東北大学情報科学研究科, 2022年7月より未踏スケールデータアナリティクスセンターの専任教員.





https://yamada-lab.gr.jp/

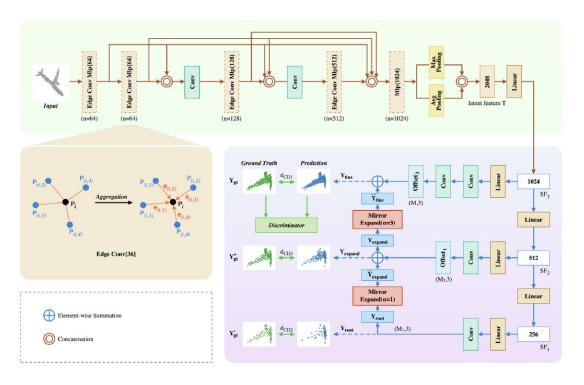


# 三次元点群補完の方法の開発

人工知能ブーム以来,画像,自然言語等様々なタイプのデータが機械学習の解析の対象となったが, 3次元データも重要な解析対象.高性能な3次元点群の補完の手法を開発した研究.



より正確な3次元点群の取得を補助する方法.

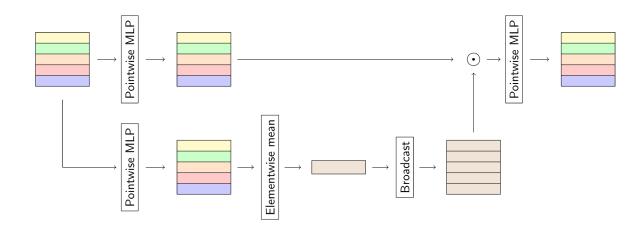


Cosmos Propagation Network: Deep Learning Model for Point Cloud Completion, *Neurocomputing*, 2022.



# 高速に文脈を把握する方法の開発

アテンション機構という文字列の文脈を把握する方法が 実は all you need ではないのでは?という動きがあっ て (MLP-Mixer, gMLP), 同じくアテンションを用い ずに高速に文脈把握をする方法を開発した研究.

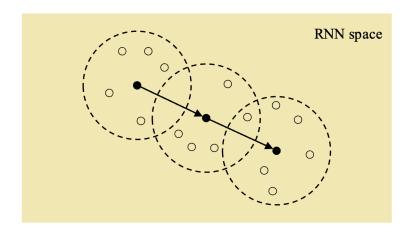


Method	IMDb		Reuters		WikiText-103	
	Total time	Time / epoch	Total time	Time / epoch	Total time	Time / epoch
Parameter	size: 64					
Baseline	n/a	n/a	n/a	n/a	n/a	n/a
	n/a	n/a	n/a	n/a	n/a	n/a
Attention	316	15.3	952	4.35	44100	1100
	13.8	0.0108	111	0.00920	1600	1.73
Linear	64.1	2.14	233	0.720	1265	65.9
	5.92	0.00656	50.8	0.00213	85.19	3.80
Relation	32.9	1.73	145	0.579	1409	50.3
	0.103	0.00543	10.9	0.00727	34.8	1.24
Parameter	size: 128					
Baseline	n/a	n/a	n/a	n/a	n/a	n/a
	n/a	n/a	n/a	n/a	n/a	n/a
Attention	347	17.2	563	4.93	44200	1210
	7.69	0.00800	29.0	0.00650	1370	1.68
Linear	107	3.73	172	1.27	2050	108
	3.04	0.0554	13.4	0.00249	67.6	3.56
Relation	54.1	2.85	115	0.971	2080	81.1
	0.352	0.0185	7.07	0.0152	59.0	2.97
Parameter	size: 256	100	200	40	26.7	
Baseline	n/a	n/a	n/a	n/a	n/a	n/a
	n/a	n/a	n/a	n/a	n/a	n/a
Attention	431	22.2	445	5.65	51000	1540
	11.7	0.0539	14.2	0.0360	1110	1.58
Linear	166	6.12	190	2.09	3520	185
	13.1	0.00934	9.36	0.0112	69.6	3.66
Relation	86.1	4.53	134	1.59	3400	137
	0.0957	0.00505	3.89	0.0300	107	2.88

Relation is an option for processing context information, *Frontiers in Artificial Intelligence*, 2022.

## 再帰型ニューラルネットワークの開発

文字列を処理するためのニューラルネットワークである再帰型ニューラルネットワークのパラメータ サイズを小さくした新型を作った研究



ランダムな変数の組み合わせから良い再帰型ニューラルネットワークを探した.

#### LSTM

$$\mathbf{v}_1 = \sigma(\mathbf{W}_{1a}\mathbf{u}_t + \mathbf{W}_{1b}\mathbf{h}_{t-1} + \mathbf{b}_1)$$

$$\mathbf{v}_2 = \sigma(\mathbf{W}_{2a}\mathbf{u}_t + \mathbf{W}_{2b}\mathbf{h}_{t-1} + \mathbf{b}_2)$$

$$\mathbf{v}_3 = \sigma(\mathbf{W}_{3a}\mathbf{u}_t + \mathbf{W}_{3b}\mathbf{h}_{t-1} + \mathbf{b}_3)$$

$$\mathbf{v}_4 = \tau(\mathbf{W}_{4a}\mathbf{u}_t + \mathbf{W}_{4b}\mathbf{h}_{t-1} + \mathbf{b}_4)$$

$$\mathbf{s}_t = \mathbf{v}_1 \odot \mathbf{v}_4 + \mathbf{v}_2 \odot \mathbf{s}_{t-1}$$

$$\mathbf{h}_t = \mathbf{v}_3 \odot \tau(\mathbf{s}_t)$$

#### S-LSTM

$$\mathbf{v}_1 = \sigma(\mathbf{W}_{1a}\mathbf{u}_t + \mathbf{W}_{1b}\mathbf{h}_{t-1} + \mathbf{b}_1)$$

$$\mathbf{v}_2 = \tau(\mathbf{W}_{2a}\mathbf{u}_t + \mathbf{W}_{2b}\mathbf{h}_{t-1} + \mathbf{b}_2)$$

$$\mathbf{s}_t = (\mathbf{1} - \mathbf{v}_1) \odot \mathbf{v}_2 + \mathbf{v}_1 \odot \mathbf{s}_{t-1}$$

$$\mathbf{h}_t = \tau(\mathbf{s}_t)$$

#### GRU

$$\mathbf{v}_1 = \sigma(\mathbf{W}_{1a}\mathbf{u}_t + \mathbf{W}_{1b}\mathbf{h}_{t-1} + \mathbf{b}_1)$$

$$\mathbf{v}_2 = \sigma(\mathbf{W}_{2a}\mathbf{u}_t + \mathbf{W}_{2b}\mathbf{h}_{t-1} + \mathbf{b}_2)$$

$$\mathbf{v}_3 = \tau(\mathbf{W}_{3a}\mathbf{u}_t + \mathbf{W}_{3b}(\mathbf{v}_2 \odot \mathbf{h}_{t-1}) + \mathbf{b}_3)$$

$$\mathbf{h}_t = (\mathbf{1} - \mathbf{v}_1) \odot \mathbf{h}_{t-1} + \mathbf{v}_1 \odot \mathbf{v}_3$$

#### **MGU**

$$\mathbf{v}_1 = \sigma(\mathbf{W}_{1a}\mathbf{u}_t + \mathbf{W}_{1b}\mathbf{h}_{t-1} + \mathbf{b}_1)$$

$$\mathbf{v}_2 = \tau(\mathbf{W}_{2a}\mathbf{u}_t + \mathbf{W}_{2b}(\mathbf{v}_1 \odot \mathbf{h}_{t-1}) + \mathbf{b}_2)$$

$$\mathbf{h}_t = (\mathbf{1} - \mathbf{v}_1) \odot \mathbf{h}_{t-1} + \mathbf{v}_1 \odot \mathbf{v}_2$$

$$\mathbf{YamRNN}$$

$$\mathbf{v}_{1} = \sigma(\mathbf{W}_{1a}\mathbf{u}_{t} + \mathbf{W}_{1b}\mathbf{h}_{t-1} + \mathbf{b}_{1})$$

$$\mathbf{v}_{2} = \tau(\mathbf{W}_{2}\mathbf{u}_{t} + \mathbf{b}_{2})$$

$$\mathbf{h}_{t} = (1 - \mathbf{v}_{2})^{2} - (\mathbf{h}_{t-1} + \mathbf{v}_{2}^{2}) \odot \mathbf{v}_{1}$$

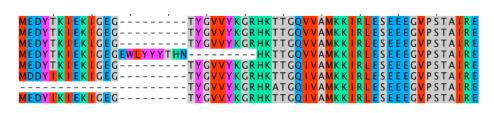
Developing a novel recurrent neural network architecture with fewer parameters, *Interdisciplinary Information Sciences*, 2021.



## 生物学的文字列配列法MAFFTの開発

生命科学分野においてある生物の文字列と別の文字列を揃えることはとても意義深い。MAFFTは元々は2002年に加藤和貴によって作られた3本以上の生物学的文字列をFFTを用いて揃える方法。

#### MAFFT version 7 Multiple alignment program for amino acid or nucleotide sequences Download version This service was unavailable due to scheduled maintenance, 7:00 - 21:00, Feb. 20, 2022, JST. Mac OS X Hardware was upgraded, Jan 16, 2022. There should be no change in user interface. If you notice a Windows Linux To avoid overload, try a light-weight option, for MSA of Source Online version For a large number of short sequences, try an experimental service. mafft --add Experimental service for aligning raw reads (2019/Aug) Merge Phylogeny Rough tree Multiple sequence alignment and NI / UPGMA phylogeny Merits / limitations Algorithms Tips Paste protein or DNA sequences in fasta format. Example Benchmarks **Feedback** MAFFT original website (https://mafft.cbrc.jp/)



8本の生物学的文字列の整列

全科学分野において歴史上,最も引用された論文トップ10位以内のClustalシリーズより単年度引用数が多い。以下の3本の論文の引用数は4800回。

MAFFT online service: multiple sequence alignment, interactive sequence choice and visualization, Briefings in *Bioinformatics*, 2019

Parallelization of MAFFT for large-scale multiple sequence alignments, *Bioinformatics*, 2018

Application of the MAFFT sequence alignment program to large data - reexamination of the usefulness of chained guide trees, *Bioinformatics*, 2016



# ソーシャルインテグレーション活動の紹介

教授 山田和載



#### 本日話すこと

- ・ソーシャルインテグレーション研究部門とは.
- ソーシャルインテグレーション研究部門で取り組む人材育成.

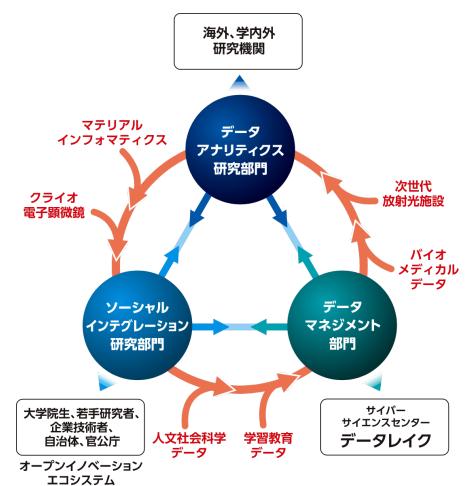
・ 学内外アカデミア研究者, ベンチャー企業, 伝統的な企業の協 働による新たなる価値の創生.



# ソーシャルインテグレーション研究部門

ソーシャルインテグレーション研究部門は、現在までに3つの部門からなる未踏センターの 1 部門. ミッションは以下の 3 (1) 点.

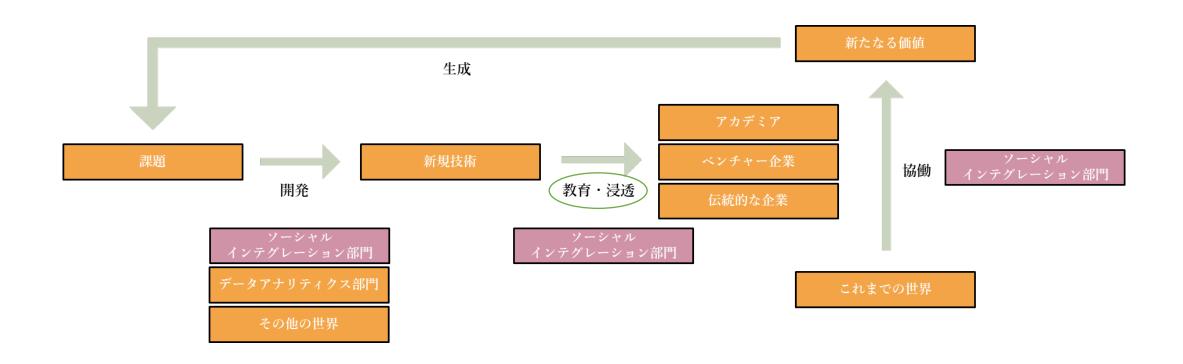
- 未踏スケールデータを解析する新たな技術を開発する (未踏各研究室のミッション).
- 新たなる解析技術を社会に浸透させる.
- <u>技術を習得させた企業や研究者による協働によっても</u> たらされる新たなる価値の創造を支援する.



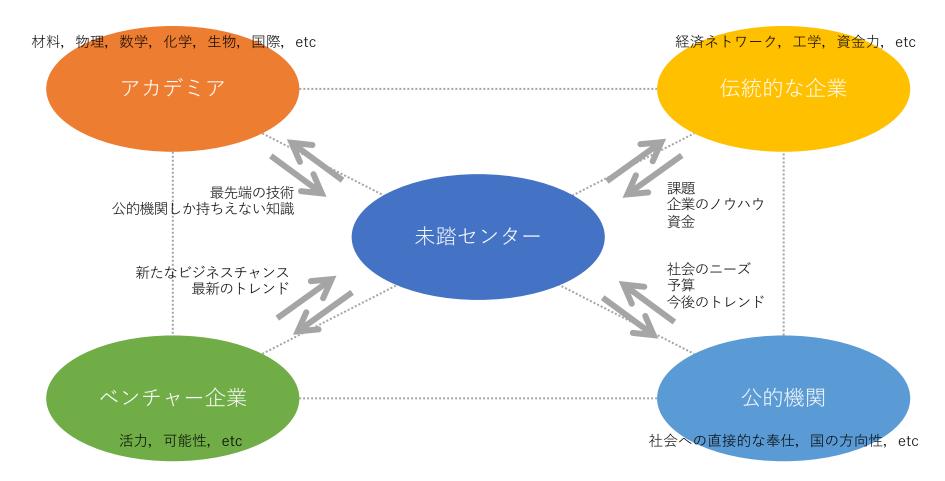


## 新規技術の教育と浸透

ソーシャルインテグレーション研究部門のミッションのひとつは開発した解析手法や, 日々開発され続ける新たな解析手法を社会に浸透させること.

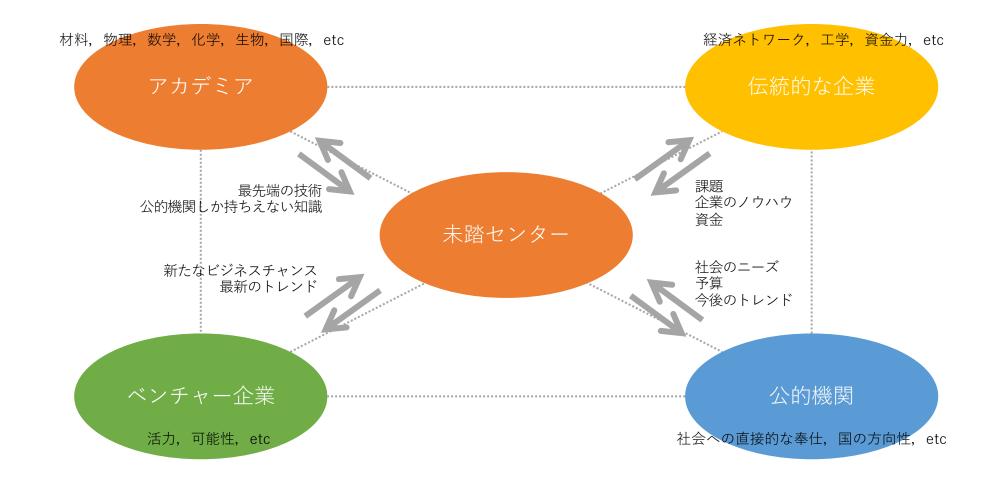


# 目指す未来



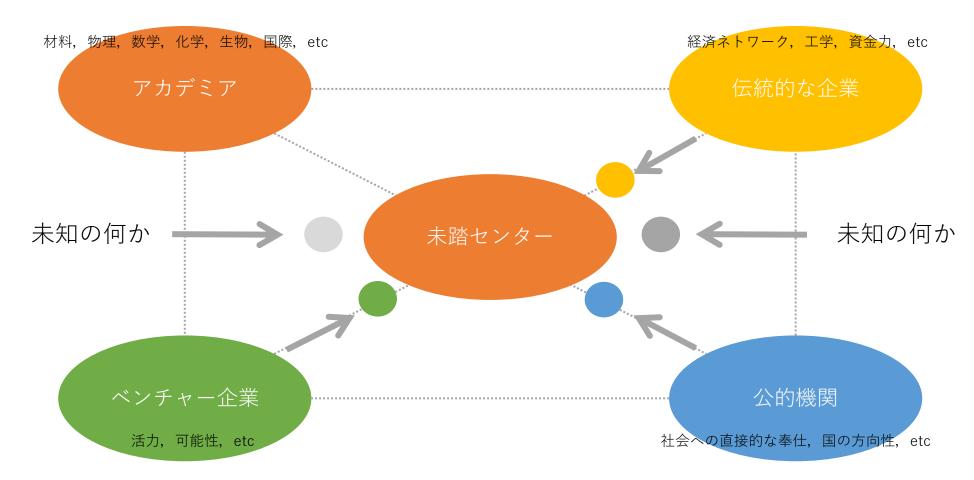
未踏センターを中心に(中心でなくてもいい)産官学を連携させ新たな価値(金と知)を創造する.

# 目指す未来の課題





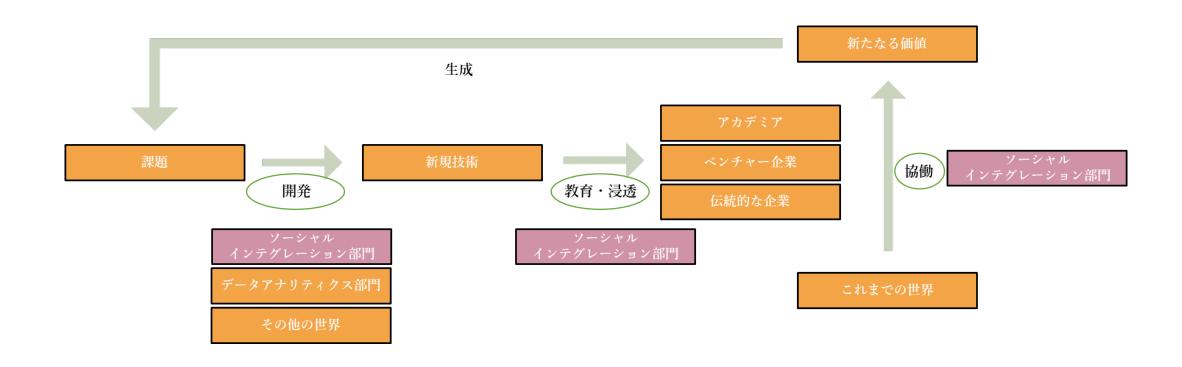
# 目指す未来



様々なフィールドから人・モノ・金を集めて役割を果たしたい.

## 今すべきこと

そのために今(アカデミアの)我々がすべきこと(できること)はビジネス界隈,公的機関の動向,社会のニーズに他の学内組織よりも大きなアンテナを張りながら以下の部分.





#### 東北大のAIMD教育の階層

#### **Senior Domain Specialist with AIMD**

#### データ科学国際共同大学院プログラム

- •情報科学研究科が6研究科と連携して提供する、留学生と日本人学生のための双方向・分野横断的な学習環境
- 実世界のビッグデータを用いて取組むプロジェクト 遂行型学習 (PBL)
- •6ヵ月あるいはそれ以上海外の研究機関に滞在して行う国際共同研究

情報科学研究科および連携する大学院群

#### **AIMD for All**

- 全ての新入生に向けてAIMDリテラシ教育を実施 (2,500名).
- ・AIMDの先端的なテーマを学び、現在進行中の研究とその最前線を知る機会を提供することで、学生のモチベーションを高める独自のコース

#### Leading Researcher with Advanced AIMD

- サイズ、分解能、モダリティの点で「未踏スケールのデータ」に基づくデータ科学を学び、研究を行う機会を提供
- アナリスト、ビジネスエキスパートとユニットを形成し、協力し合って課題解決を行なうことで、データ科学の現状とその利活用の仕方を理解する

未踏スケールデータアナリティクスセンター

スーパー研究者の養成

#### **Domain Specialists with AIMD**

- ・理工学系、人文系など、学生の専門性に応じたAIMD関連の講義を実施
- 学部生には、専門知識とAIMDスキルのダブルメジャー取得を強く推奨

各学部、学科

#### 全学教育



# 未踏センターにおける教育のターゲット

- ・未踏センターと関りが深い情報科学研究科のデータ科学国際共同大学院にて研究者の卵へのデータ科学研究は担っており、また、兄弟センターであるデータ駆動・AI教育センターでは全学教育を担っており、それらと協調効果を高めるため、未踏センターで特に教育を施したいのは各研究ドメインで活躍している現役の研究者.
- ・<u>各ドメインで活躍する研究者にデータ科学を浸透させ、スーパー研究者を育成したい</u>.



#### 各ドメインの人材がデータ科学を学ぶ必要性

なぜ研究者にデータ科学教育を?

A: 今後データ科学はリテラシー化するから(データ科学はデータサイエンティストなるものだけの学問ではない).

生命科学研究者がデータ科学を学ぶ必要ある?データサイエンティストと協力すれば良いのでは? A. 明確にNO

「理論○○学者」以外に「理論学者」とか存在する?(理論物理,実験物理,計算物理学者はいる) データ科学は科学のひとつ.各ドメインに各科学的アプローチをする研究者がいるべき(いないと深い 研究は無理).

社会人学び直しは学生時代に勉強してこなかった人がやるんでしょう?

A. 明確にNO

データ科学は(かつてから様々な解析手法は別個に存在していたもののあるものの)近年体系的にまとめられつつある学問(多分まだまとまっていない).

大学において今データ科学教育を行うことは有意義.

## ドメインデータサイエンティスト養成

様々な研究分野で活躍する研究者(大学院学生、アカデミア研究者、企業研究者)でデータ科学をこれまでに利用していない人を対象にしたデータ科学の講習会. 12月から2か月をかけて22人に. (数理・データサイエンス・AI 教育強化拠点コンソーシアムとメガバンク木下賢吾先生と共催)

- ①データ科学と機械学習
- ②Pythonの基本的な使用方法
- ③Pythonの発展的な使用方法
- ④教師あり学習法
- ⑤教師なし学習法
- ⑥TensorFlowの基本的な利用方法
- (7)多層パーセプトロン
- 8 再帰型ニューラルネットワーク
- 9アテンションネットワーク
- ⑩Hugging Faceの利用方法
- 印敵対的生成ネットワーク
- 迎強化学習法



https://udac.tohoku.ac.jp/news/event/136/

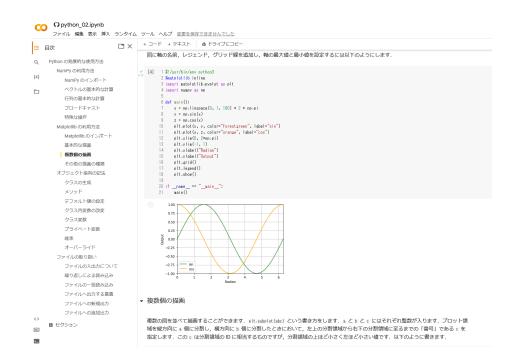


# ドメインデータサイエンティスト養成

メタバースを利用して話し合いながら実習形式で学ぶ.座学と異なり参加者等とのインタラクションが重要なため.合計24時間の講習にて全学オープンバッジ(紫)を発行.今後さらに発展予定.



学研MetaLife



Google Colaboratory



## 地元企業との連携に向けた取り組み

仙台市との共同にてSendai X-Techイノベーションプロジェクトにてリカレント教育講座を開催予定.

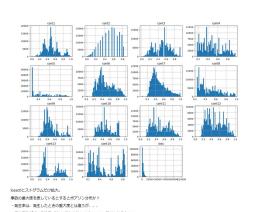
トレーニング課題

前半と後半でチームを変えながら,前半はベーシックな機械学習,後半は深層学習をやる.Kaggle の問題を利用.

用いたデータ:家屋の値段を予測する構造化されていて固定長の行列形式になっているデータ,ツイッタデータから,そのツイートが災害に関するものかどうかを予測するもの. (前回参加者30名)









#### 学部学生への先取り教育

これまでの経験から、データ科学教育には OJT が効果的であると確信を得ており、長期間にわたる PBL を行いたい. よって、授業ではなくて課外活動として「放課後Kaggleクラブ」を企画中.

kaggle



## UDACセミナー

学内外からデータ科学を利用した研究を行っている研究者や方法論の開発研究を行っている研究者を 招待しセミナーを開催中. 今後も月1回のペースで継続予定.

第1回 五十嵐康彦先生(筑波大)

第2回 土方敦司先生(東京薬科大)

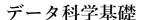
第3回 鈴木潤先生(東北大), 千葉直也先生 (オムロン・早大)





# データ科学国際共同大学院との連携

情報科学研究科,生命科学研究科,経済学研究科,工学研究科,医学系研究科,理学研究科の6研究科にて行うデータ科学の教育プログラムであるデータ科学国際共同大学院では様々なデータ科学教育の教材を開発しているがこれと協力.データ科学では他分野への横展開を担当.









gp-ds.tohoku.ac.jp

International Liaison Office.

Graduate School of Information Sc
Tehoku University.

BIG DATA SKILL-UP TRAINING

DATA SCIENCE TRAINING CAMP









# まとめ

ソーシャルインテグレーション研究部門を通じて未踏センターが実現したいことのひとつは、未踏スケールデータの解析を行う研究者らを協働させること.

色々な研究者に参加してほしい アカデミアの人材と共同研究 色々なドメインの研究者を教育したい してほしい アカデミア 伝統的な企業 未踏スケールデータ ベンチャー企業 公的機関 未踏センターを介して 課題を出してほしい 大きくなってほしい 予算を分配してほしい



未踏センターをよろしくお願いします.