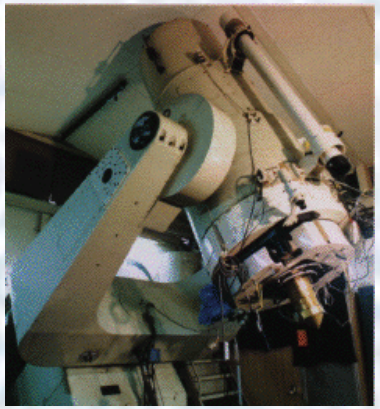
100年後に残すデータ

国立天文台天文データセンター 小杉城治

自己紹介　小杉 城治

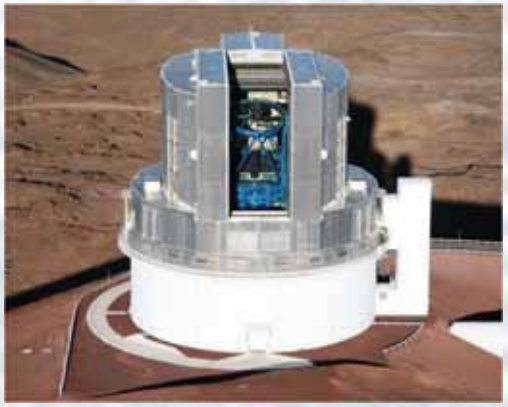
•天文学のデジタル化と共に歩んできました

- 大学院時代
  - 京都大学大宇陀観測所60cm RC望遠鏡＋手作り分光器＋冷却CCD
  - 岡山天体物理観測所188cm望遠鏡+新カセグレン分光器＋冷却ＣＣＤ
- すばるプロジェクト時代
- アルマプロジェクト時代
- 天文データセンター時代



Credit:京都大学

Credit:佐々木敏由紀



Credit:国立天文台

Credit:NAOJ/ESO/NRAO

自己紹介（大学院時代）1988～

- 望遠鏡や観測装置制御のデジタル化、検出器のデジタル化
- 観測データ ＝ デジタルデータ が始まった時代

Spectro-Nebulagraph: A Tridimensional-Spectroscopic System Based on a Local Area Network of Personal Computers

GEORGE KOSUGI[1] AND HIROSHI OHTANI

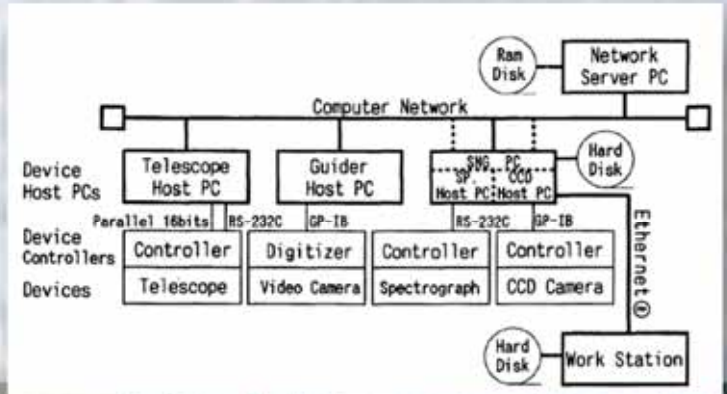
Department of Astronomy, Faculty of Science, Kyoto University, Sakyo-ku, Kyoto 606-01, Japan
Electronic mail: george@optik.mtk.nao.ac.jp, ohtani@kusastro.kyoto-u.ac.jp

TOSHIYUKI SASAKI,[1] HISASHI KOYANO, AND YASUHIRO SHIMIZU
Okayama Astrophysical Observatory,[2] Kamogata-cho, Asakuchi-gun, Okayama 719-02, Japan
Electronic mail: sasaki@opal.mtk.nao.ac.jp, koyano@kibi.oao.nao.ac.jp, shimizu@kibi.oao.nao.ac.jp

MICHITOSHI YOSHIDA[3]

Department of Astronomy, Faculty of Science, Kyoto University, Sakyo-ku, Kyoto 606-01, Japan

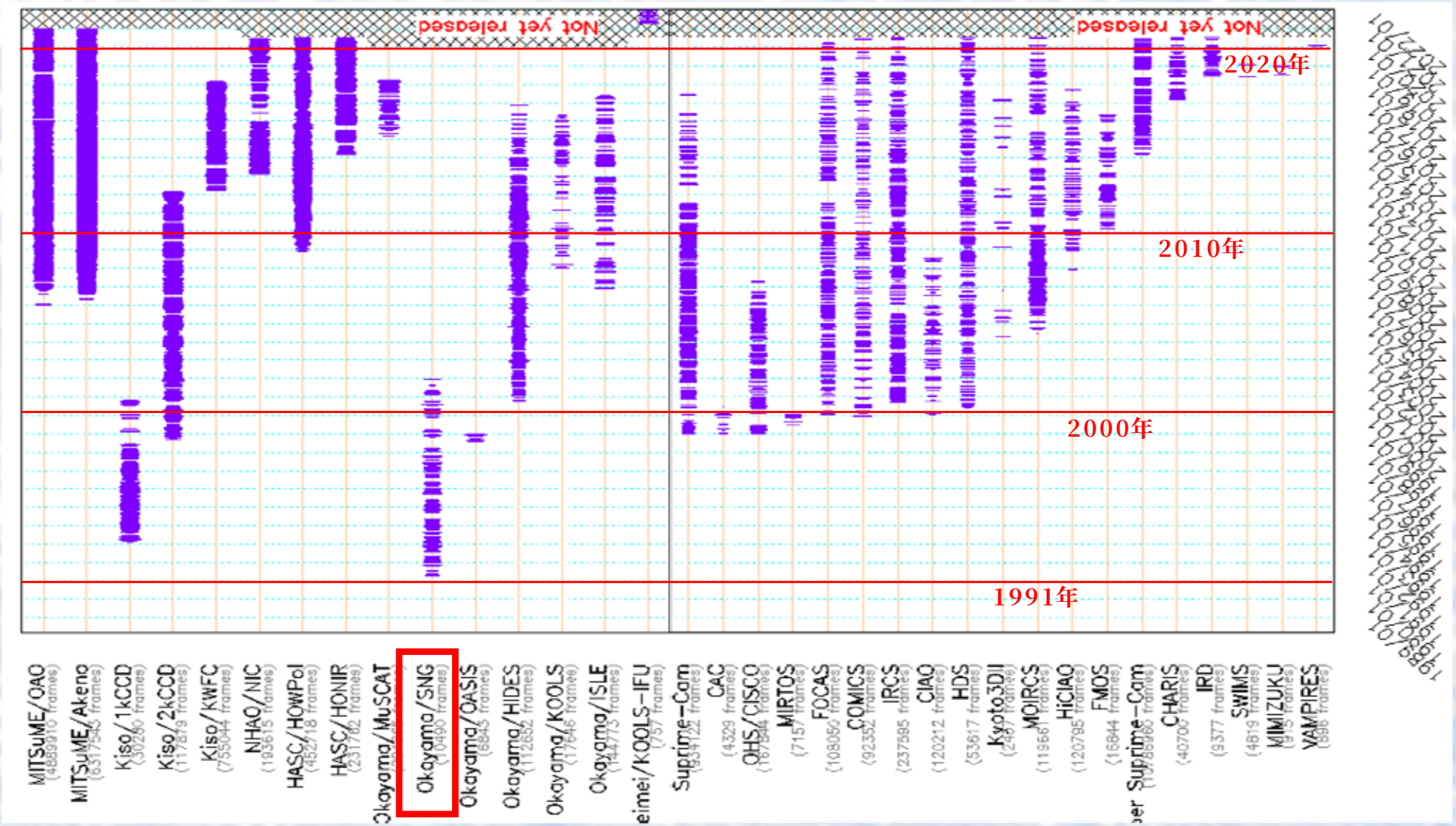
図3.12　SNGの観測制御画面：左図では望遠鏡を駆動するコマンドが実行中で、右図では分光器とCCDカメラを制御するコマンドが実行中



岡山天体物理観測所188cm望遠鏡でSNG観測システムを開発し共同利用

自己紹介（大学院時代）1988～

SNG観測データは、1991年よりアーカイブ



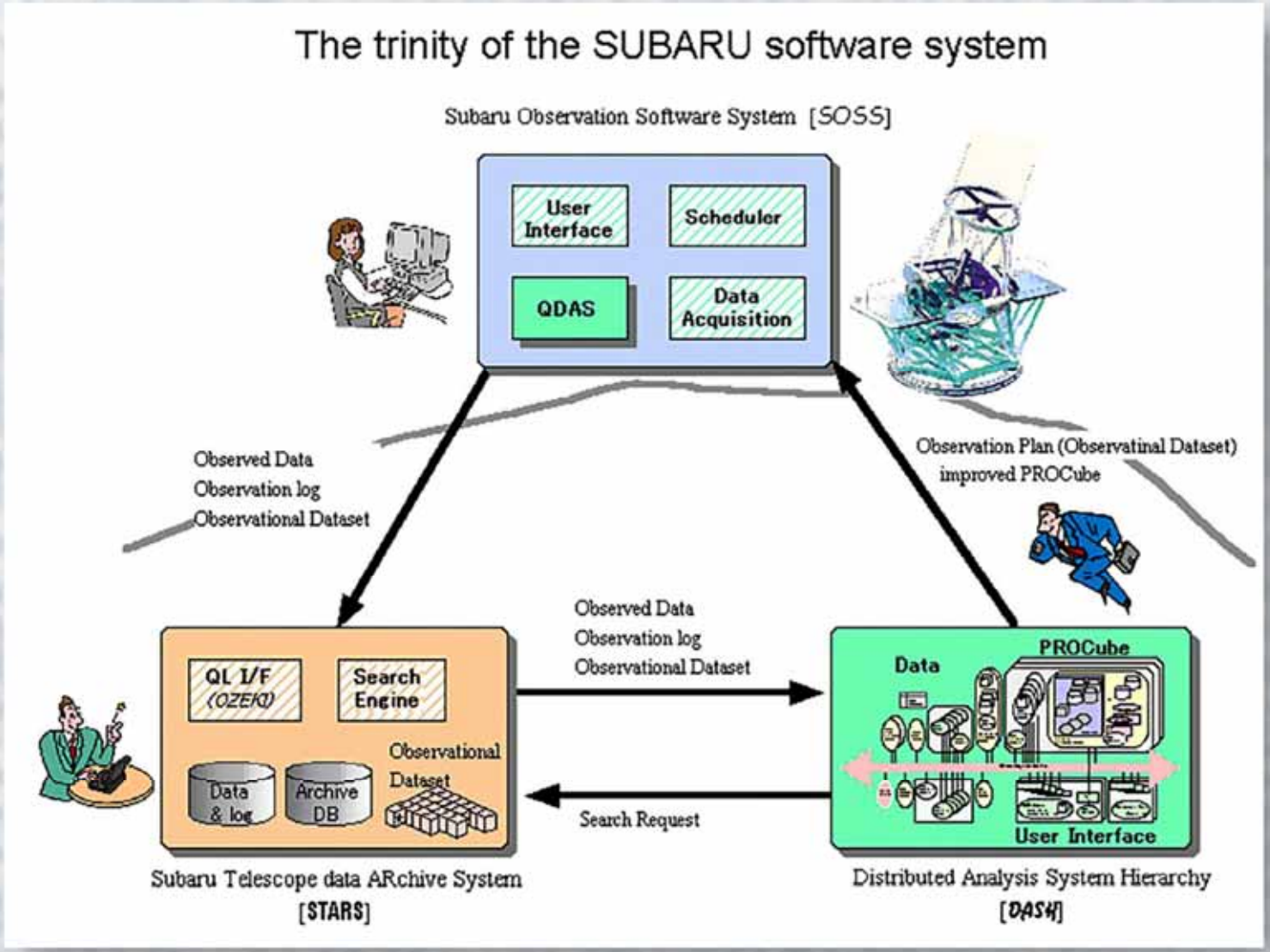
SMOKAホームページ(https://smoka.nao.ac.jp/)より

自己紹介（すばるプロジェクト時代）1996～

- すばる観測制御システムの要件定義、立ち上げ、運用
- 第1期観測装置FOCASの制御ソフトウェア、解析ソフトウェア開発
- 望遠鏡や観測装置の性能評価、科学観測

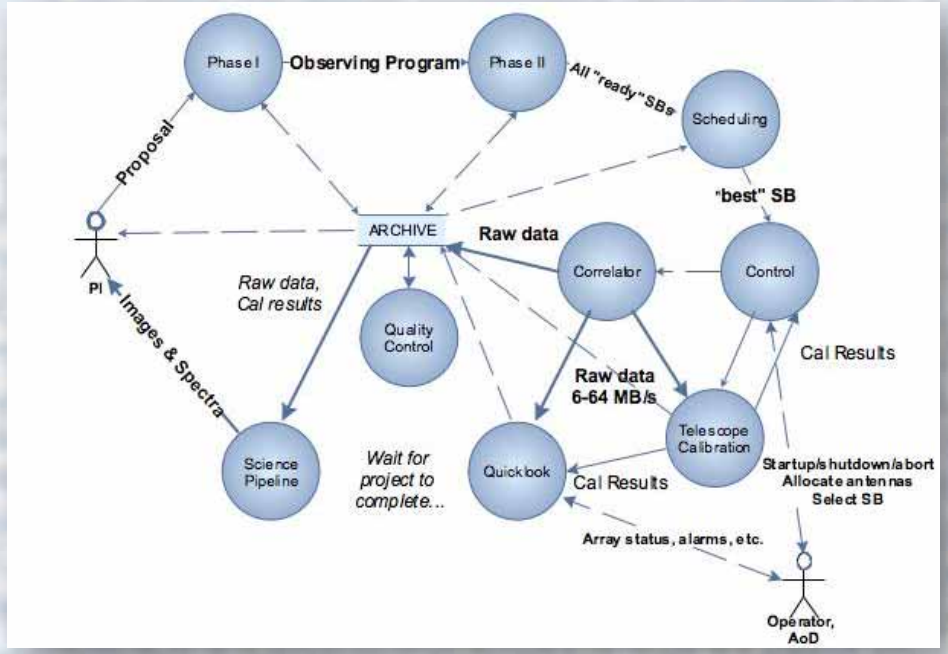
自己紹介（すばるプロジェクト時代）1996～

- すばる望遠鏡のソフトウェアシステムが目指したもの



自己紹介（アルマプロジェクト時代）2005～

- 日米欧国際チームCIPT(Computing Integrated Product Team)活動
  - データ解析パイプライン開発 (単一鏡モード)
  - ACAアンテナや相関器の立ち上げ、性能評価
- CIPT-> ICT (Integrated Computing Team)
  - マネージャー



ALMA Data Flow Model

自己紹介（天文データセンター時代）現在

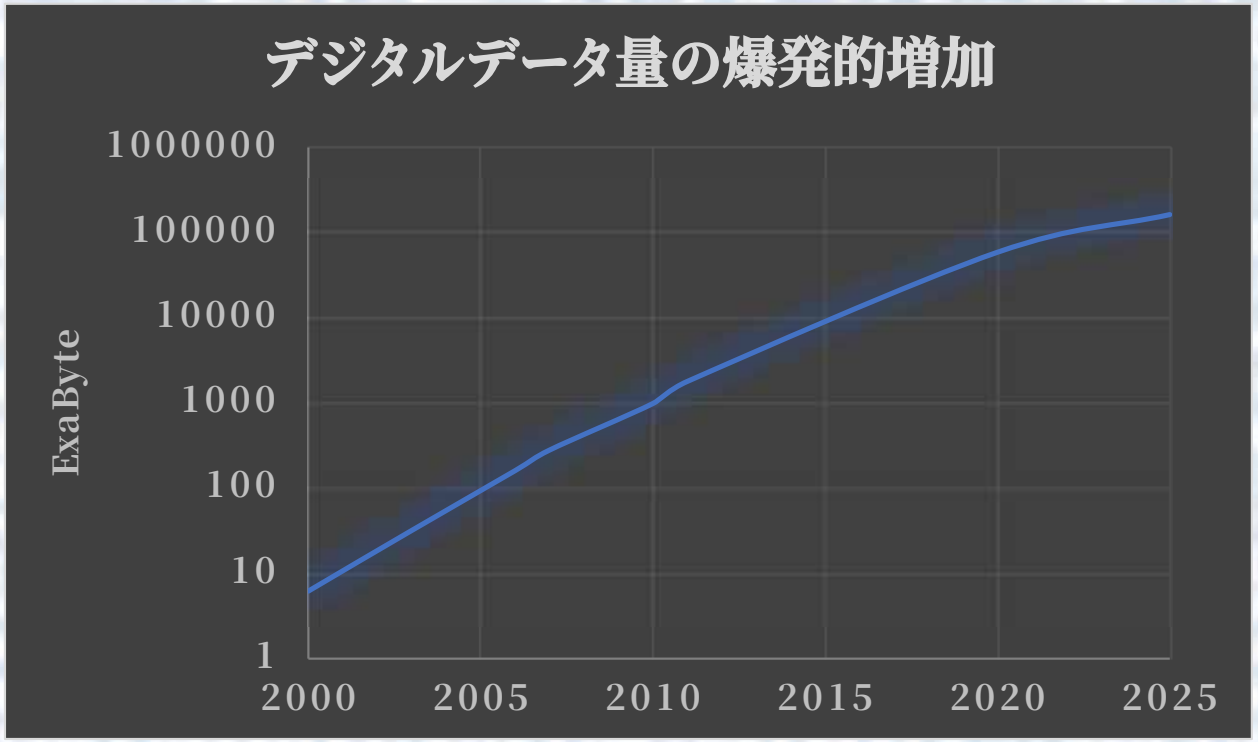
- 昨年度の特別公開日用インタビューで「現在、及び、未来の人類に観測データを届けるのが仕事」と話をした。

さて、本当にできるのだろうか？

増え続けるデータ

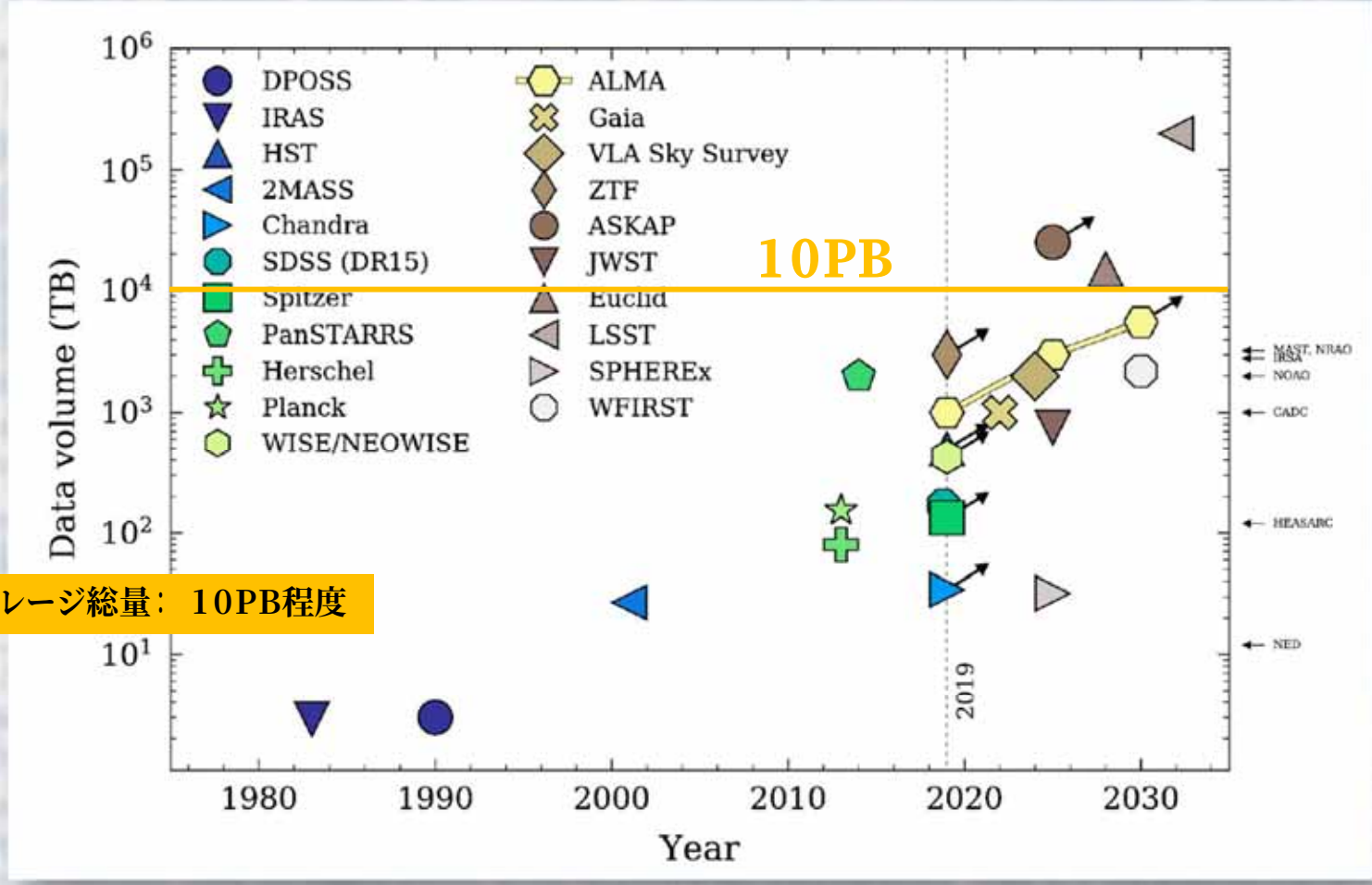
データ爆発の時代

- 2020年の世界データ総量は50 Zeta Bytes
  - Zeta-byte = 1,000 Exa-byte = 1,000,000 Peta-byte
  - 参考：ALMAの年間データ生成量は200TB = 0.2PB



天文データの動向

- LSST (平均15TB/夜, サーベイ完了時の解析済データは数百PB)
- ngVLA (平均7.6GB/s, 休みなく動けば ～240PB/yr)
- SKA1(～600PB/yr)

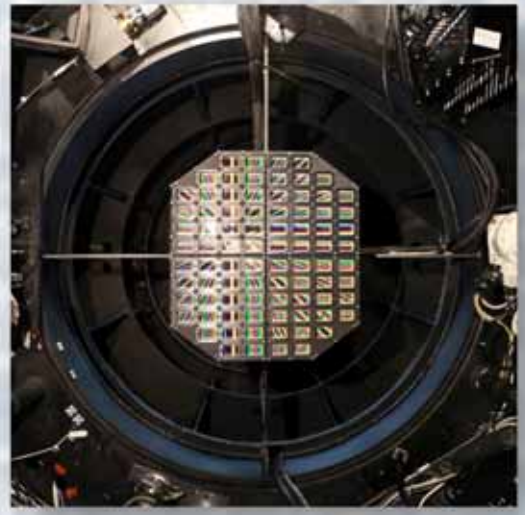


国立天文台アーカイブストレージ総量： 10PB程度

Desai V. et al., 2019

日本の中小望遠鏡データでもデータ爆発

- CMOSセンサー
  - トモエゴゼン：木曽シュミット ＋ モザイクCMOSカメラ
    - 84 CMOSセンサー(2K x 1.1K)
    - 2フレーム/秒で一晩観測すると30TB/夜 ～10PB/yr
      （但し晴天率や観測プログラムによる）



Credit:東京大学

  - TriCCS：せいめい望遠鏡 ＋ 可視3色同時CMOSカメラ
    - 3 CMOSセンサー(2.2K x 1.3K, 最大98fps)
    - 10fps 8時間観測で5.2TB/夜

Credit:京都大学

生データを全ては残せない時代に

- トモエゴゼンのデータは国立天文台SMOKAアーカイブに保管されつつある

  ただし、データ量の制約から、複数フレームを積分したデータのみ

技術的に可能でも、予算的に困難な場合もある

必要な情報をアーカイブするためには、システムズエンジニアリング的なアプローチが必要

- 例えばTriCCSは一晩5TB以上のデータを取れるが、1日かけても国立天文台三鷹のSMOKAアーカイブにネットワークで送りきれない
- データ転送、データ処理、データ保管、のボトルネックを除去
- 全プロセスを含む全体システムとしての最適化を進める

国立天文台観測データポリシー

- 観測データは宇宙の歴史的記録
- 観測データには寿命がない
- 観測データの完全な理解は、後世まで含めたあらゆる世代に委ねられるべき

1．国立天文台の観測データは国立天文台に帰属する

2．国立天文台は、観測データを利用可能なデジタル形式で**永続的に保管**する

3．国立天文台は、観測データを**利用しやすい形式で公開**する

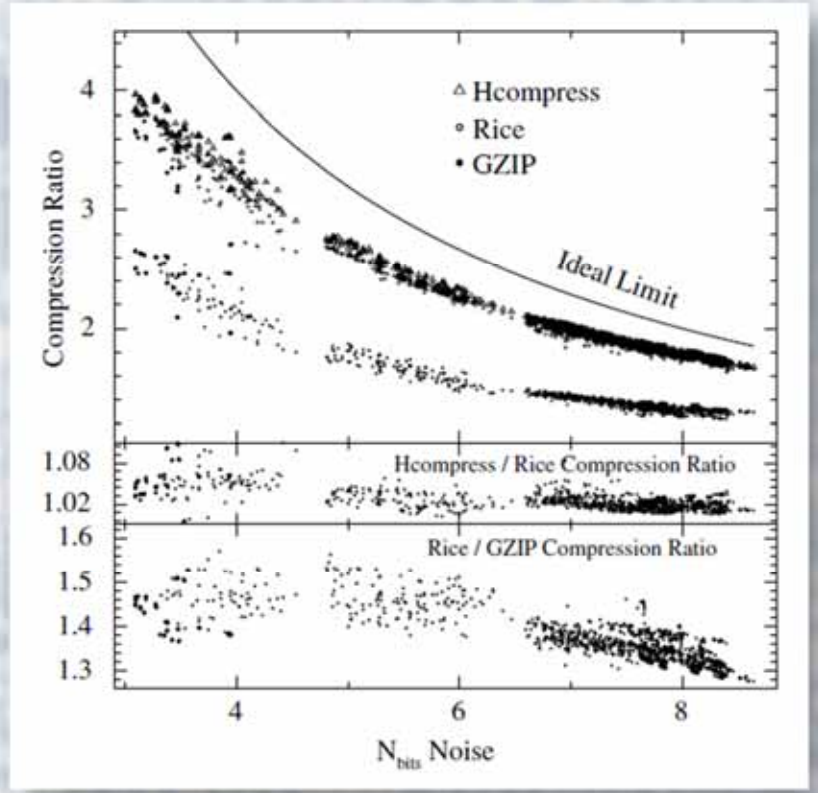
何を残して何を捨てるか：データ圧縮

- 何も捨てない
- lossless圧縮（gzip, Rice等）で半分程度にまで圧縮可
  - 最近の天文データ解析ソフトは圧縮形式のデータにも対応
  - 圧縮率はノイズレベル次第

- より圧縮率を高めたければ

  非可逆圧縮！

以前は非可逆圧縮への天文学者のイメージは良くなかった



W.D.Pence et al., 2009, PASP

何を残して何を捨てるか：データ圧縮

- 何かを捨てる ＝ 非可逆圧縮

トモエゴゼン（CMOSカメラ）

- 画像の積算によるデータ圧縮 ： 時間軸情報を捨てている
- CMOSカメラは時間軸天文学を切り拓く ： 時間軸情報を残したい
  - Robust PCAによりLow-rank行列とスパース行列に分解し、transient天体情報を残したままmovieを高圧縮(～1/10)

残したい情報の素性がわかっていれば、その情報を残して高圧縮することは可能

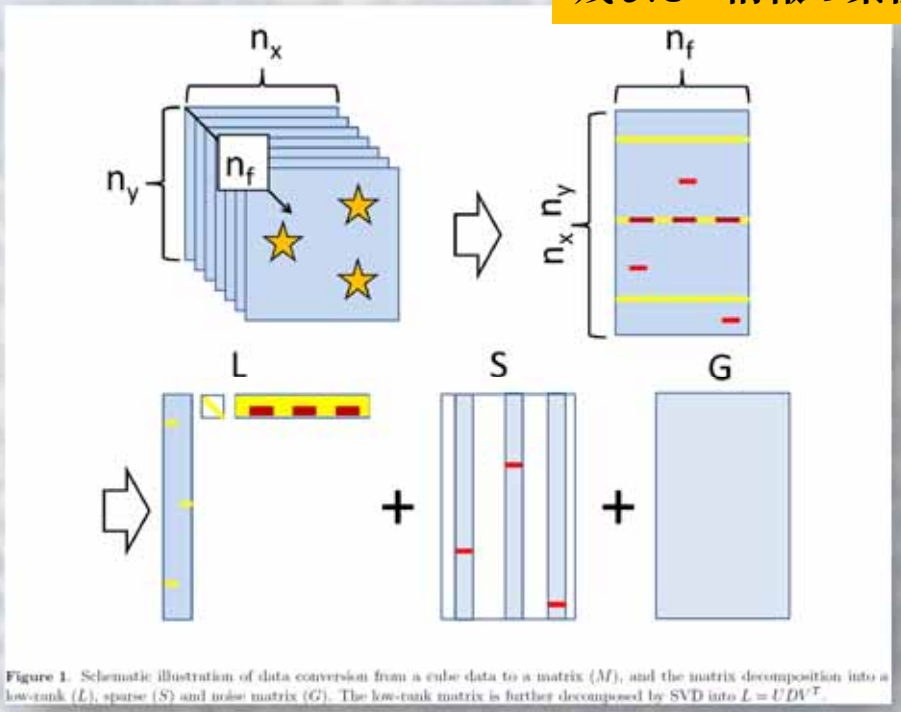


Figure 1. Schematic illustration of data conversion from a cube data to a matrix (M), and the matrix decomposition into a low-rank (L), sparse (S) and noise matrix (G). The low-rank matrix is further decomposed by SVD into $L = UDV^T$.

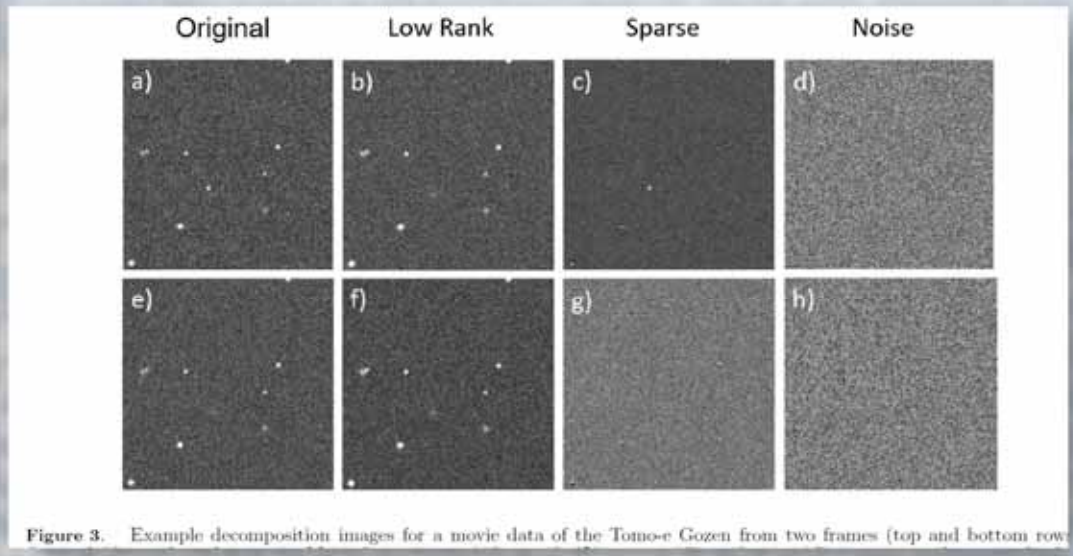


Figure 3. Example decomposition images for a movie data of the Tomo-e Gozen from two frames (top and bottom row

Morii et al., 2017, ApJ

大量データを残すための技術

- データ圧縮には情報理論や統計数理のドメイン知識・技術が必要
- 長期データ保管にはITのドメイン知識・技術が必要
- 大量データを活用するには、AIのドメイン知識・技術も必要となろう

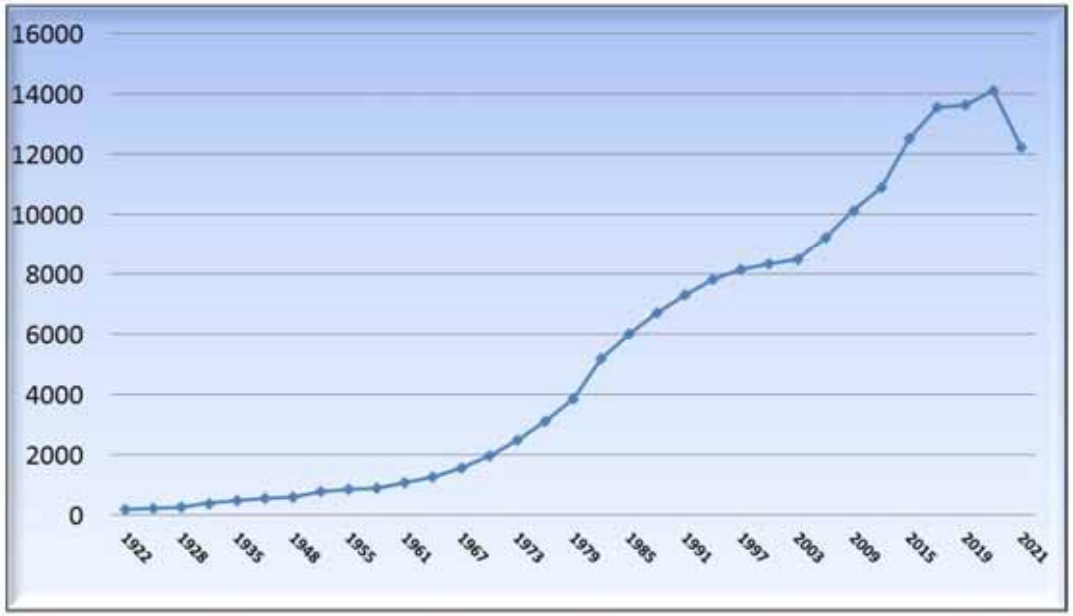
ドメイン知識・技術をどこまで（天文台側で）習得するかは、ドメインのサイズ感も判断材料となり得る（あくまで私見）

例えば、ドメインがとてつもなく大きい場合、或いは、専門性が極めて高い場合には、ドメイン知識を網羅するのは至難の業。

一方、ドメインが大きければ、協力者を見つけやすい。

天文ドメインの規模感

- 天文学者数
  - 世界でせいぜい 2～3万人？
- 天文業界の技術者数
  - 国立天文台では技術系職員の数は研究者の半分程度か



IAUホームページ（https://www.iau.org/）より

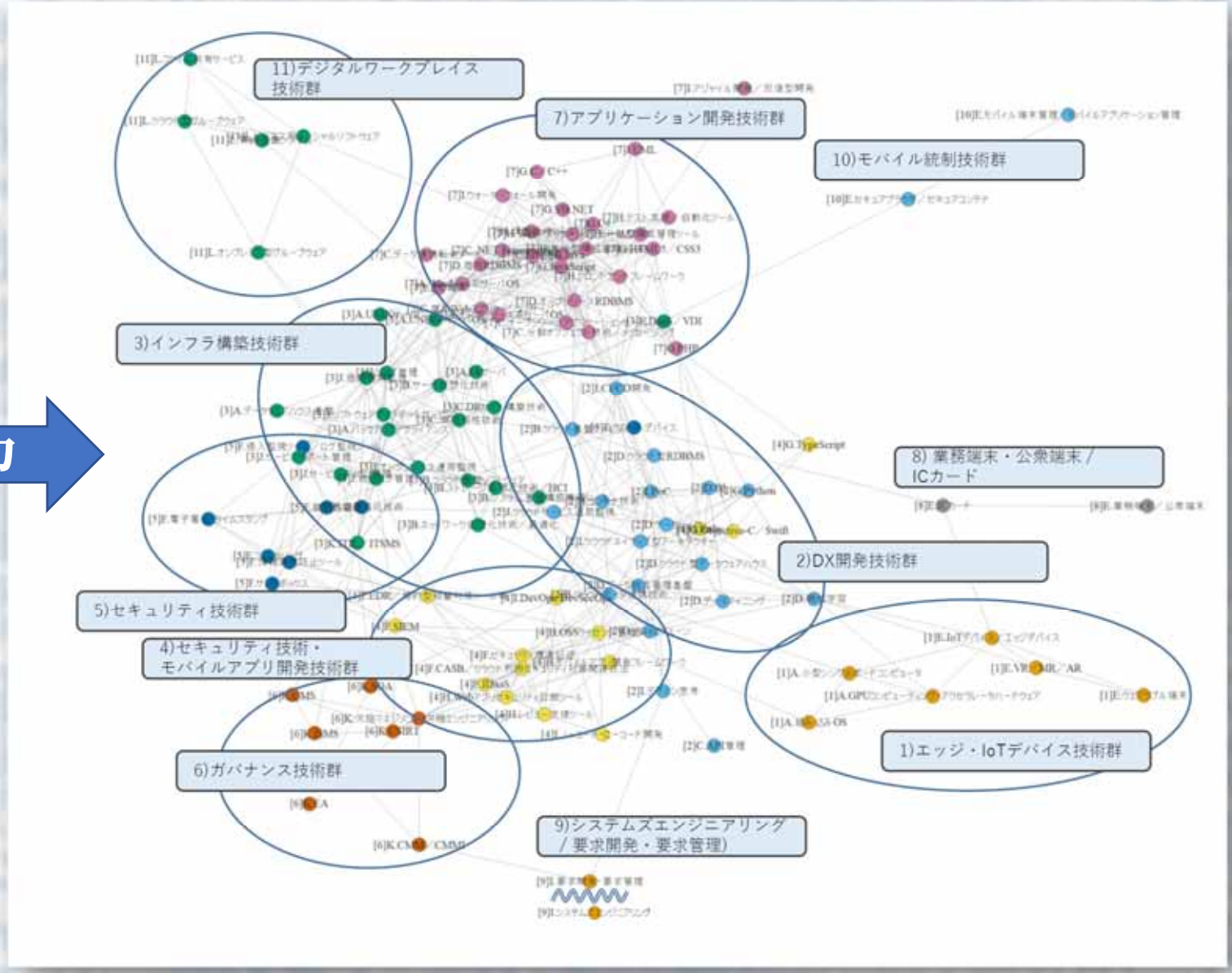
- ICT技術者数2020年
  - 世界 2100万人
  - 日本 110万人
- ICT研究者数2018年
  - 日本 17.6万人 (R2情報通信白書より)

巨大な境界領域を埋める工夫が必要

- IT技術分野

天文業界

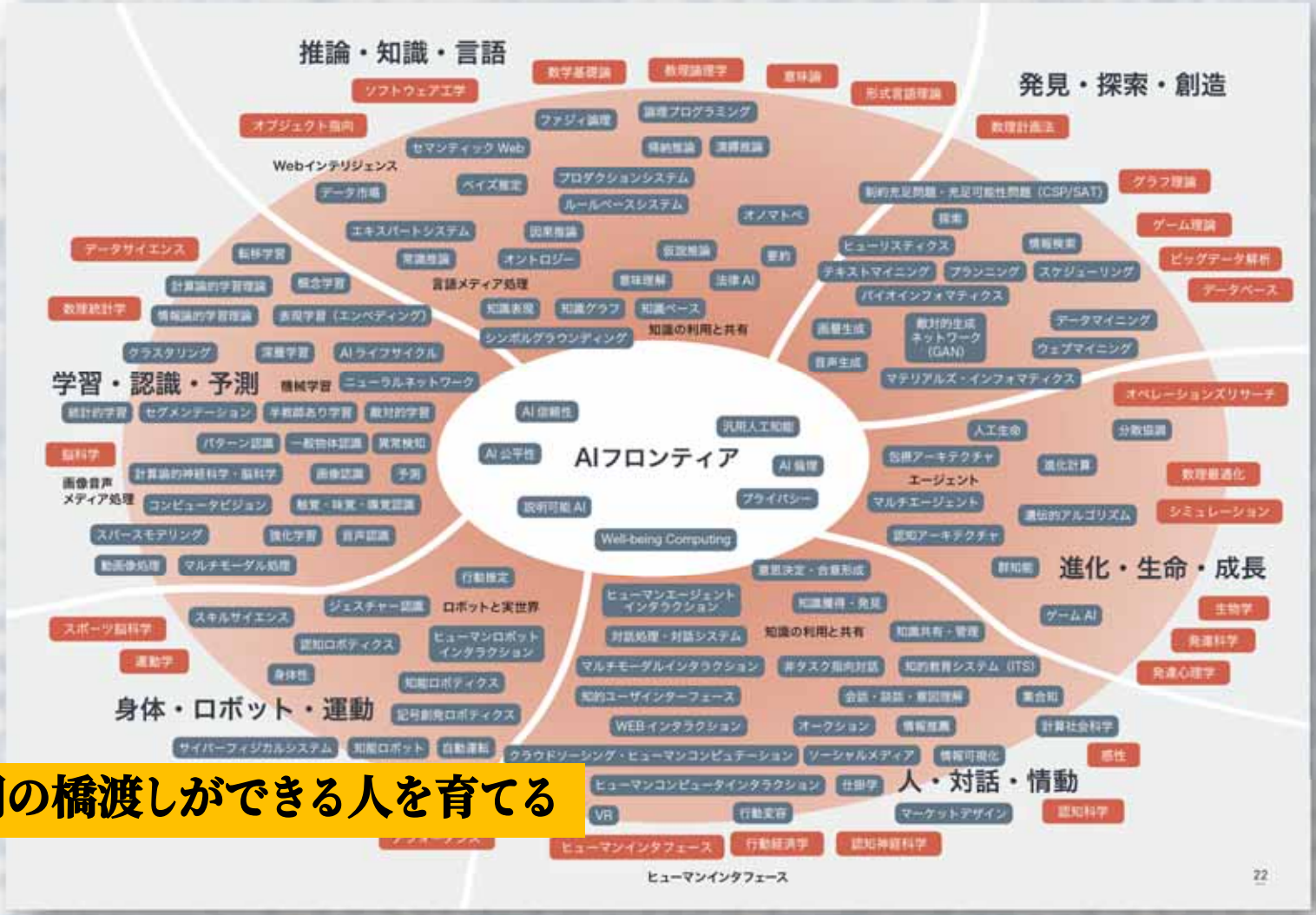
協力



一般社団法人 情報サービス産業協会ホームページより

AI技術分野も拡大中

国内AI人材数：2020年約4万人、2025年約8万人（経産省 AI人材育成の取り組み(2019年) より）



ドメイン間の橋渡しができる人を育てる

AIマップβ 2.0 人工知能学会：https://www.ai-gakkai.or.jp/

データ運用に関する技術分野

- 専門化や新技術の開発が急速に進んでいる
- 天文研究者・技術者から手が届きにくくなっている
- 境界領域を埋める工夫やコラボが必要

ビッグデータ

データの収集、取捨選択、管理及び処理に関して、一般的なソフトウェアの能力を超えたサイズのデータ集合（ウィキペディア）

総務省Society5.0： Valueにつながる4V

総務省「Society5.0」資料

Volume（容量の大きさ）

ビッグデータの第一の特徴は、その名前の通り容量が大きいことです。企業に限らず、情報技術の進化により、黙っていてもどんどんデータが集まるようになり、データ量はテラバイトからペタバイトオーダーにもなっています。データ量が大きいことだけがビッグデータの特徴だと思われがちですが、他にも以下のようなポイントがあります。

Variety（多様性、種類）

ビッグデータは、通常表計算などで扱っているように、数値化され関連づけをされたデータ（構造化データ）であるとは限りません。テキスト、音声、画像、動画などのさまざまな構造化されていないデータ（非構造化データ）もあり、これらのデータをテキストマイニングや音声、画像解析などを行ない構造化し、ビジネスに活用する動きが広まっています。

Velocity（スピード、頻度）

サーバーのアクセスログや、東京ゲートブリッジ橋梁モニタリングシステムなど、ものすごい頻度、スピードでインターネット上やセンサーからデータが生成され、取得、蓄積されています。変化の著しい現代社会では、これらのデータをリアルタイムに処理し、対応することが求められています。

Veracity（正確さ）

従来は、サンプリングによって一部のデータで全体を推測する方法が主流でした。それに対し、ビッグデータは全てのデータを取得することも不可能ではないので、正確であり推測による曖昧さや不正確さなどを排除して、本当に信頼できるデータによる意思決定が可能になります。

Value（価値）

得られたデータを分析し有用な知識や知恵を導出し、モデル構築、検証し、課題解決をすることが本質的なビッグデータの価値です。

4Vと天文データアーカイブ

Volume（容量の大きさ）

あらゆる望遠鏡や観測装置のアーカイブデータを公開することで、利用できるデータ量を増やす。

★Variety（多様性、種類）

あらゆる観測モード（撮像、分光）や波長（光赤外、電波）のデータを公開することで、多様性を増やす。望遠鏡や観測装置の多様性も含まれる。

Velocity（スピード、頻度）

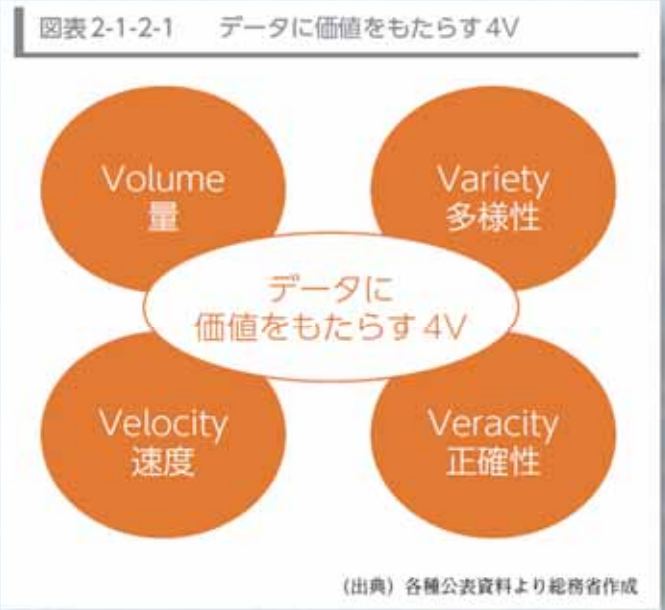
高フレームレートのCMOSデータはVelocityを充実させる。リアルタイム処理をおこなって、変動天体を検出し、フォローアップ観測に繋げることができる。

★Veracity（正確さ）

信頼できるデータをアーカイブ保管して、利用できるようにすることが必要。

Value（価値）

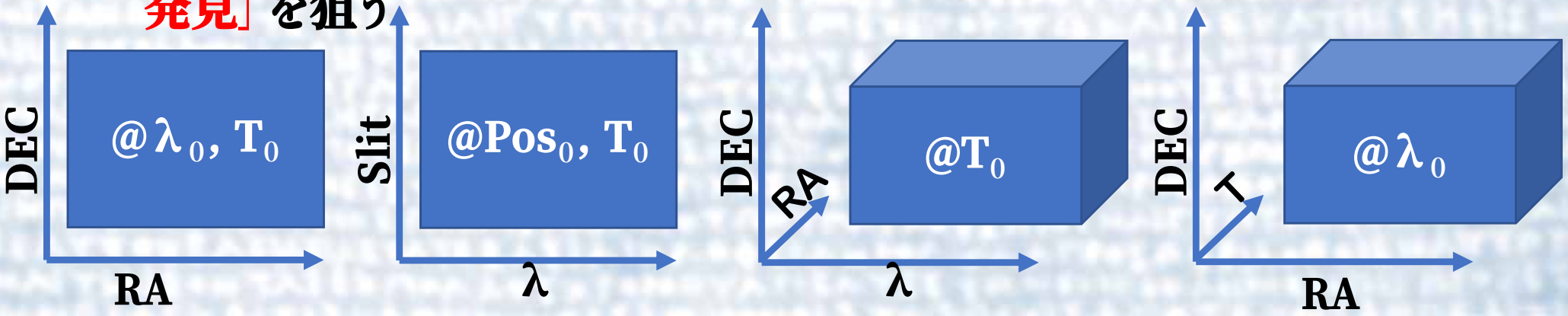
「未知の発見」へと繋げて行く。



総務省「Society5.0」資料

天文データ蓄積によるパラメータ空間の拡張

- **Variety**の充実（データを保管し続ければVarietyは広がり続ける、少なくとも時間軸方向には）
- 赤経、赤緯、波長、時間の4次元空間をアーカイブデータで埋める。それぞれの軸のスパンが大きいほど多様性が大きく、Valueを生み出す。
  - 撮像データ：赤経(RA1～RA2、ΔRA)、赤緯(Dec1～Dec2,ΔDec)の2次元データを時刻T_0、波長λ_0の場所へ
  - スリット分光データ：赤経(RA1～RA2、ΔRA)、波長(λ1～λ2、Δλ)の2次元データを赤緯Dec_0、時刻T_0の場所へ
  - 3次元分光データ：赤経(RA1～RA2、ΔRA)、赤緯(Dec1～Dec2,ΔDec)、波長(λ1～λ2、Δλ)の3次元データを時刻T_0の場所へ
  - ムービーデータ：赤経(RA1～RA2、ΔRA)、赤緯(Dec1～Dec2,ΔDec)、時刻(T1～T2、ΔT)の3次元データを波長λ_0の場所へ
- 4次元空間にマッピングされた観測データ全体から「**未知の関係性の発見**」を狙う



Veracity（正確さ）を担保するために

- 国立天文台の観測データポリシー

第3項　国立天文台は、観測データを利用しやすい形式で公開する

国立天文台は、明記された期間を経た後に、原則として全ての観測データを「研究者が利用しやすい形式」で公開する。データは特定のソフトウェアを用いなくても解析できる水準まで較正処理を進め、できる限りそのまま物理量として扱えるようにした後に公開する。

観測データポリシーに沿ったデータ運用を観測所に求めることが重要。

その望遠鏡や観測装置の開発、運用に関わった人が現場を去って装置固有キャリブレーションなどノウハウが失われてしまう前に。

観測設備は進化を続けてきたため、古いデータは情報量が少なく精度も低めであるが、時間軸方向のVarietyを大幅に広げることができ、Valueにつながる。

各データには、誤差や精度に関する情報も必要。

ビッグデータの時代には、データをそのまま扱えることが重要。

キャリブレーションや誤差解析は、100年後にデータを残す上で必須の技術。

デジタルデータは劣化しない、が

- デジタルデータは理論的には劣化しないが、デジタルデータの記録媒体は劣化したり、規格が変わって読めなくなったりして、データ自身が失われることがある。
- データが失われないように、定期的にマイグレーションする
- 或いは、クラウド上で自動的にマイグレーションしてもらう

- そもそも、100年後に天文データの標準フォーマットであるFITSが残っているのか？データ自体も、何度かフォーマット変換が必要かもしれない

まとめ

100年後に意味のあるデータを残すために

- 既存技術や予算などの制約下で、データ爆発時代にデータ（情報）を最大限残すためには、システムズエンジニアリング的なアプローチ（システムの全体最適化）が重要
- 必要な天文分野のドメイン知識・技術は、観測データのキャリブレーションや精度・誤差解析
- 天文分野外のデータ運用に関わるドメインは、規模が桁で大きい、或いは、専門性が高いため、全てを把握することは容易ではない。ドメインのガイドやドメイン人材と橋渡しができる人の育成が必要
- デジタルデータは劣化しないが、データ自身が失われることがある。定期的なマイグレーションやフォーマット変換が必要
- 上記を継続するための人材や予算が担保されることが必要

技術によって、１００年後でも使えるデータを残し、
それを使った新しいサイエンスが切り拓かれることを期待して。。。

ご静聴、ありがとうございました。