Astronomy Data Center
国立天文台 天文データセンター

国立天文台 天文データセンター (ADC)の活動：現状と今後

国立天文台 天文データセンター
小杉城治、古澤久徳
2022年9月22日@光天連シンポ

はじめに

- 「2030年代の天文学と光赤外地上・スペース計画」に向けて、大学共同利用機関である国立天文台の天文データセンター(ADC)が果たすべき役割や期待について、コミュニティの意見を伺いたい。

- 有限のリソースの取り合いになるため、現時点で全ての期待に完全に答えることは難しいが、コミュニティの意見を取り入れつつ優先度を付けて進める。10年後を見据えて、必要なリソースをコミュニティと協力して獲得していきたい。

ADC活動に関連（影響）する最近の動き

1．ADCユーザーズミーティング 2021年5月19日と6月24日
- 次期計算機システムに向けて -- 共同利用計算機リプレースへの戦略 --
- ユーザースミーティングの意見を仕様書に反映
- 現在ADCの計算機リプレースの手続き中

2．光赤天連提言（将来に向けた光赤外天文観測データアーカイブの在り方への提言 2021年9月15日）

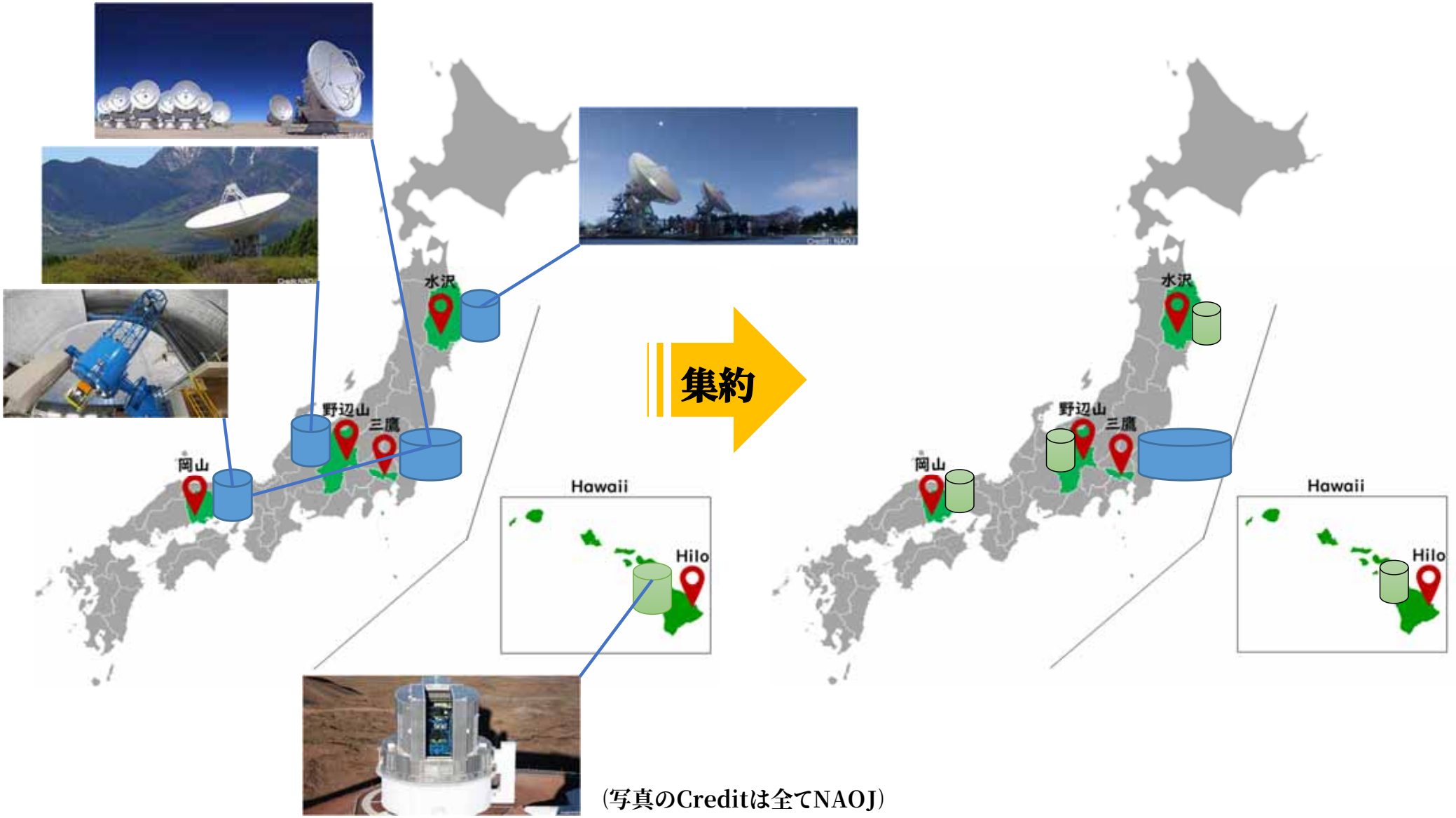
3．国立天文台科学戦略委員会のもとに国立天文台データアーカイブWG(2021年4月1日～2022年3月31日）を設置
- 科学戦略委員会の諮問に対し、答申と国立天文台データポリシーの改訂を提言(2022年3月31日)
- 科学戦略委員会に2022年7月中旬に説明

最近の動き1．計算機リプレースのステータス

- 新システム稼働開始予定
  - 2023年9月から6年間（計算機納期等の影響で2023年3月から半年延期）
  - 対象システムは、大規模アーカイブシステムと多波長解析システム
  - 大規模アーカイブシステム
    - アーカイブデータは、時間と共に種類と量が増え続ける
  - 多波長解析システム
    - そのような過去から現在までの様々なアーカイブデータを効率よく処理することが要求される
- 新システムのコンセプト
  - アーカイブデータは増え続けるが、コストは抑える
  - レンタル計算機は三鷹に集約
  - レンタル計算機と買取計算機のハイブリッド構成（確実に運用したい部分はレンタルで、運用がしやすい部分は買取計算機で）
  - データの遠隔地2次バックアップはクラウド活用を検討中

恒久データアーカイブストレージの集約計画

一時保管用ストレージは各サイトに設置



（写真のCreditは全てNAOJ）

計算機リプレース：主なシステムの仕様

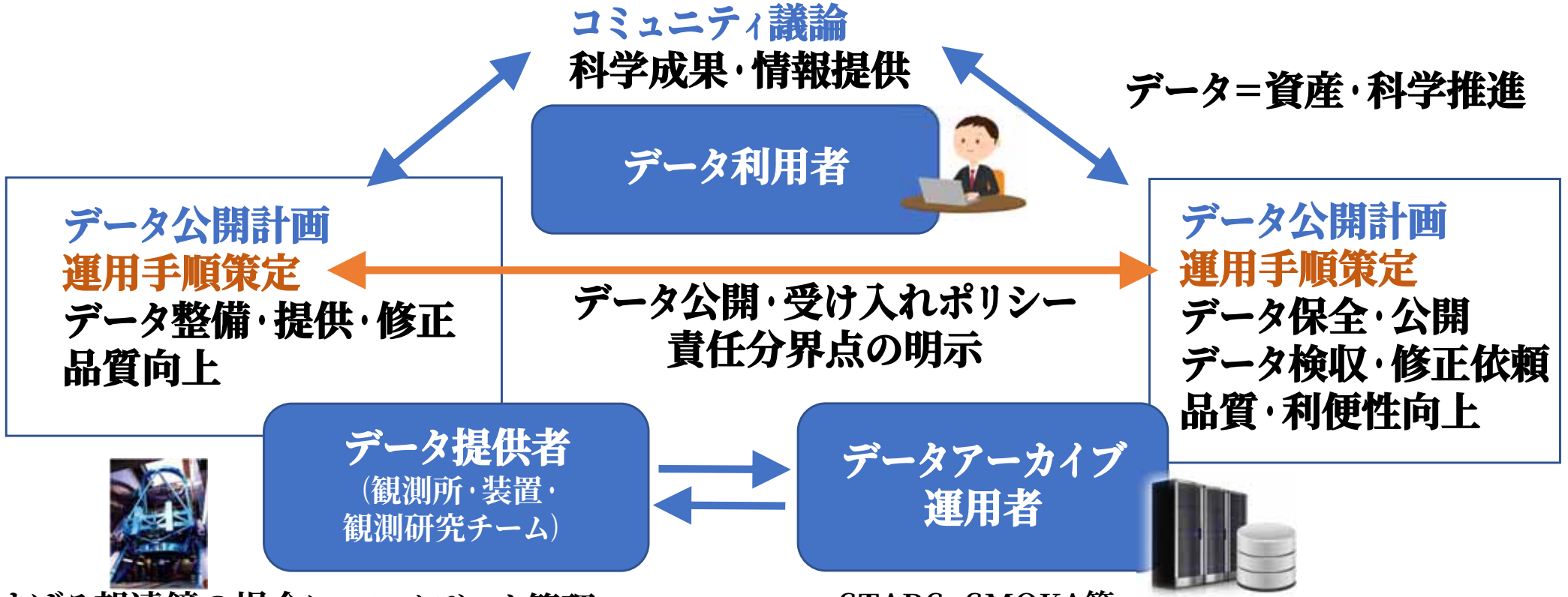
- 多波長データ解析システム
  - 物理コア数：総物理コア数は 現システム同等以上(≧544)、1ホスト当り32以上
  - メモリ量：総メモリ量は現システム同等以上(≧7.3TB)、1ホスト当り384GB以上
  - 作業用ストレージ領域 4 PiB以上 (現システム2.5PB以上)
  - 上記のうち作業用共有ストレージ領域3 PiB以上 (現システム1.7PB以上)

- 大規模アーカイブシステム
  - SMOKAアーカイブシステム
    - テープ2500本以上収納可能なライブラリ（LTO-8で埋めた場合30PB以上）（現システム6PB以上）
    - 総ディスク容量2.5PB以上（現システム2.7PB以上）
  - ALMAアーカイブシステム
    - 総ディスク容量4PB以上（現システム2PB以上）
    - データベース領域は買取計算機で運用

最近の動き2．将来に向けた光赤外天文観測データアーカイブの在り方への提言（光赤天連 2021年9月15日）

http://gopira.jp/seimei.html

- コミュニティ（データ利用者、データ提供者、アーカイブ運用者）が目標を共有・協力し、安定で有用なアーカイブを実現する
- 生データと解析情報を長期に保全・公開し、再利用価値の向上で科学成果を促進する
- 国立天文台は、他機関と協調して長期的なデータ利用拠点として機能する



国立天文台科学戦略委員会のもとに設置された国立天文台データアーカイブWG

最近の動き3．国立天文台データアーカイブWG答申

いずれ国立天文台から答申内容等が公開されると思われる

(2022年3月31日)

- 諮問1：大学共同利用機関として持つべき天文データアーカイブのあり方
- 答申1

1．大学など国立天文台外の観測データのアーカイブ公開

1．天文コミュニティと国立天文台の間でデータ受入を審議する委員会を設け、科学的な価値判断に基づいて優先度を付けて受け入れる仕組みを整える

2．データ提供者あるいは提供されるデータが守るべきルールを明確にする

3．国立天文台はデータ提供者に対して十分なサポートを提供する

2．全ての観測データを保管することが非現実的な巨大データの扱い

1．必要な情報だけを抽出し（情報の取捨選択は事前に十分に検証）、データ容量を圧縮した上でアーカイブ保存

2．巨大データの時代は生データまで戻れない時代である＊

3．ワンストップアーカイブサービスに向けて

4．国立天文台が構築・運用の一部を担う国際プロジェクトのアーカイブとの関係

＊ それ故、結果の検証ができることが必須。そのために必要なデータや処理済みデータの再利用ができるような仕組みを、装置の計画時から十分に議論して検討しておかなければならない

国立天文台データアーカイブWG答申(2022年3月31日)

- **諮問1：大学共同利用機関として持つべき天文データアーカイブのあり方**
- **答申1**

    1．大学など国立天文台外の観測データのアーカイブ公開

    2．全ての観測データを保管することが非現実的な巨大データの扱い

    3．ワンストップアーカイブサービスに向けて

        1．国立天文台が運用する全てのデータアーカイブにVOインターフェースを整備する

        2．国立天文台が運用する複数のアーカイブの集約を目指す（アーカイブ間の連携、インフラや管理機能の共通化とスケールメリットの享受、ノウハウの蓄積）

    4．国立天文台が構築・運用の一部を担う国際プロジェクトのアーカイブとの関係

        1．国際プロジェクトの最先端の観測データと密接に連携させることで、国立天文台が保管公開している観測データの多様性を生かす

国立天文台データアーカイブWG答申(2022年3月31日)

- **諮問2：現在および未来の天文コミュニティのデータ利活用を促進させるための施策**
- **答申2**

1．DOIの活用による持続可能性の担保

1．データ*にDOIを付与してトレーサビリティを確保し、データの整備を進める組織的な取組を可視化・計量化して組織の正当な評価へとつなげる

2．国立天文台は共同利用機関として天文観測データの機関レポジトリの役割を担う

3．DOIの発行、登録、確認には事務作業が発生するため、リソースの確保も必要

2．巨大データを利活用する仕組みの構築

1．巨大データと計算資源を近くに配置して、研究者がリモートからインタラクティブで探索的な処理を実施できるサイエンスプラットフォームという仕組みを整備

2．較正処理済データ、或いは、将来較正処理を行うのに必要なツールや文書の整備

3．国際プロジェクトとの関係

1．国際プロジェクトのアーカイブと国立天文台のアーカイブやサイエンスプラットフォームとのリソース共有やインターオペラビリティを確保する

2．国立天文台の科学データ資産とのシナジーを追求する

＊ 一定の質が保証がされたデータセット

最近の動き３’国立天文台の観測データポリシー改訂提案の要旨
（国立天文台アーカイブWG 2022年3月31日）

国立天文台データアーカイブWG答申と整合

第1項　国立天文台の観測データは国立天文台に帰属する

- 「帰属」とは、データの管理や利用に関する運用ルールを当該機関が規定する、という意味で用いている。複数の機関に帰属する場合は、双方の合意の上でデータ占有権などの運用ルールが規定される。

第2項　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

- 観測データを全て保管し続けることが技術的･経済的に困難な場合はダウンサイジング可

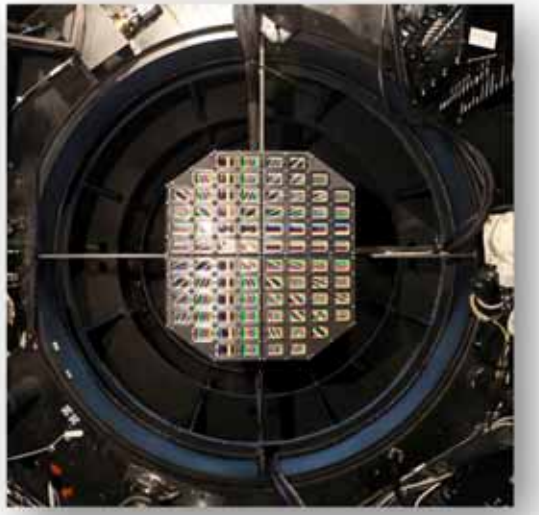
第3項　国立天文台は、観測データを利用しやすい形式で公開する

第4項　国立天文台は、観測データを国内大学や研究機関から受け入れて公開する

- 国立天文台は大学共同利用機関として、科学的研究を目的として国内大学等で作成された公開を前提とした観測データを、可能な範囲で受け入れる
- 国内大学等は、公開される観測データを前提知識なしで研究者が利用できるよう、較正処理の実施やそのためのツールの作成･公開を進めるよう努める
- 国立天文台が受け入れた観測データは、作成した国内大学等および国立天文台に帰属する
- 観測データを国内大学等で保管する場合でも、メタデータを国立天文台に集約することで、データ検索やデータ連係などの利便性の向上を期待する

観測データを取り巻く状況（データの巨大化）

- 日本の中小望遠鏡データでもデータ爆発：CMOSセンサーの登場
  - 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

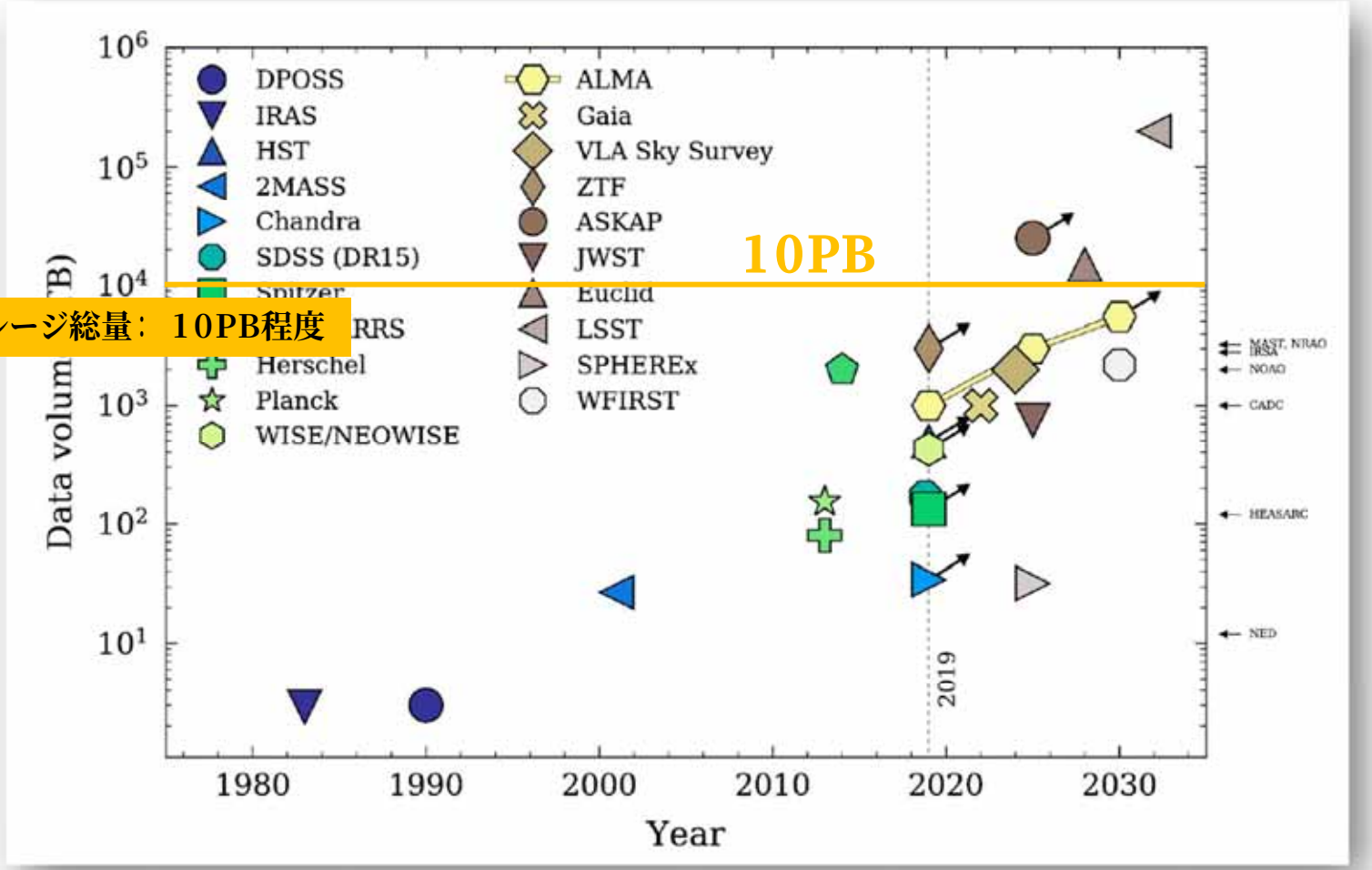
  - TriCCS：せいめい望遠鏡 + 可視3色同時CMOSカメラ
    - 3 CMOSセンサー(2.2K x 1.3K, 最大98fps)
    - 10fps 8時間観測で5.2TB/夜

Credit:京都大学

世界の天文データの動向

- 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

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が
  - データ量の制約から、複数フレームを積分したデータを保管・公開
  - 検証のために元のフレームレートの生データも一部保管
  - 保全、公開に重要なデータを十分議論の上で決める

全て残すことが技術的には可能でも、予算的に困難な場合もある

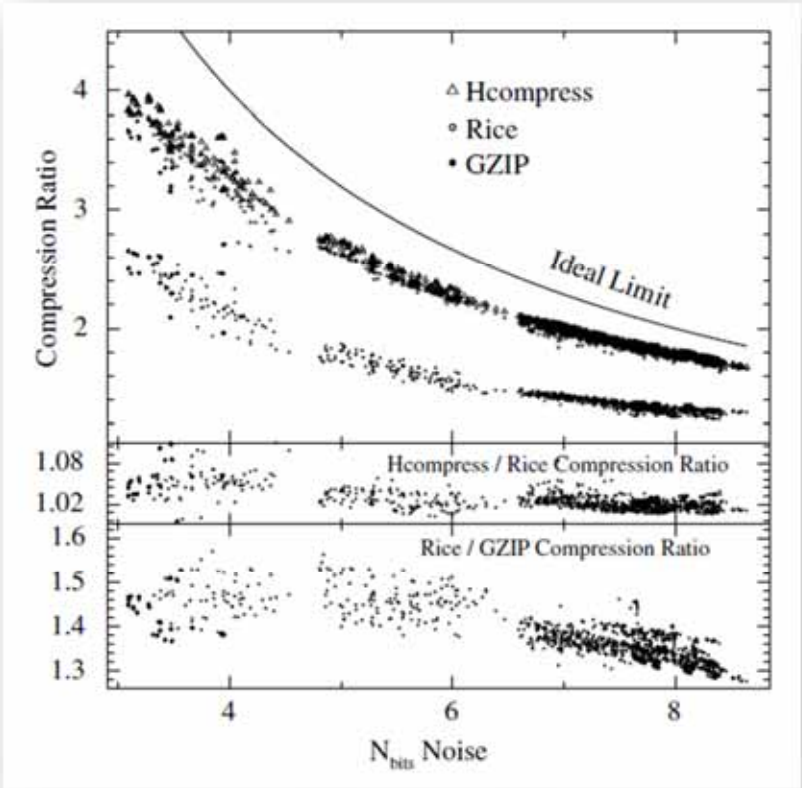
何を残して何を捨てるか：データ圧縮

- 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！

以前から非可逆圧縮に対する天文学者のイメージはあまり良くないかも



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)**

あらかじめ残したい情報の素性がわかっていれば、その情報を残すように高圧縮することは可能

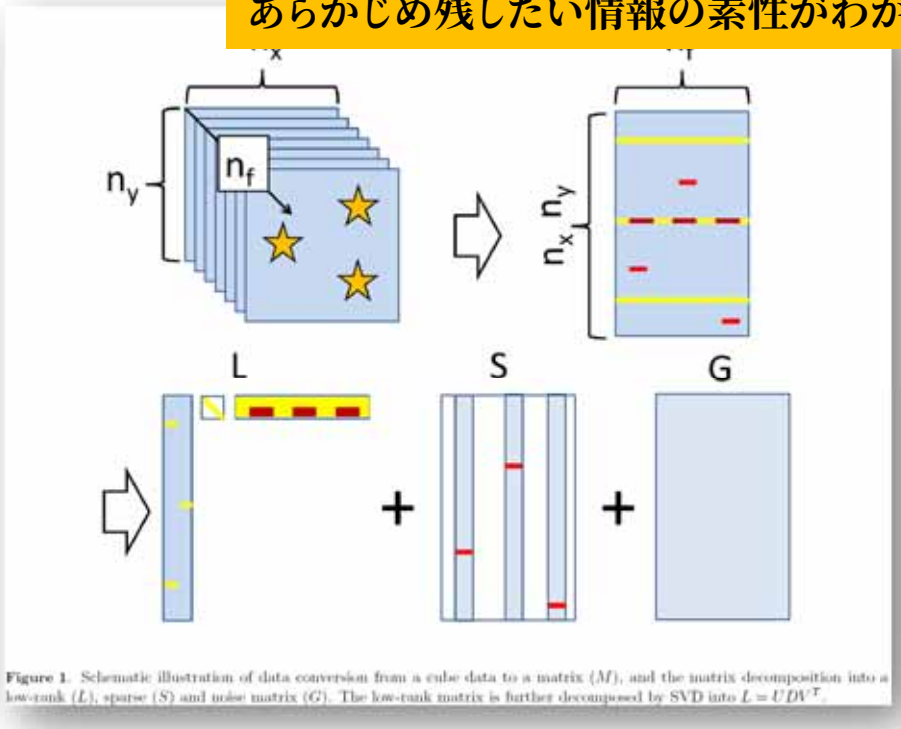


Figure 1. Schematic illustration of data conversion from a cube data to a matrix (M), and the matrix decomposition into a low-rank (L), sparse (S) and noise matrix (G). The low-rank matrix is further decomposed by SVD into $L = UDV^T$.

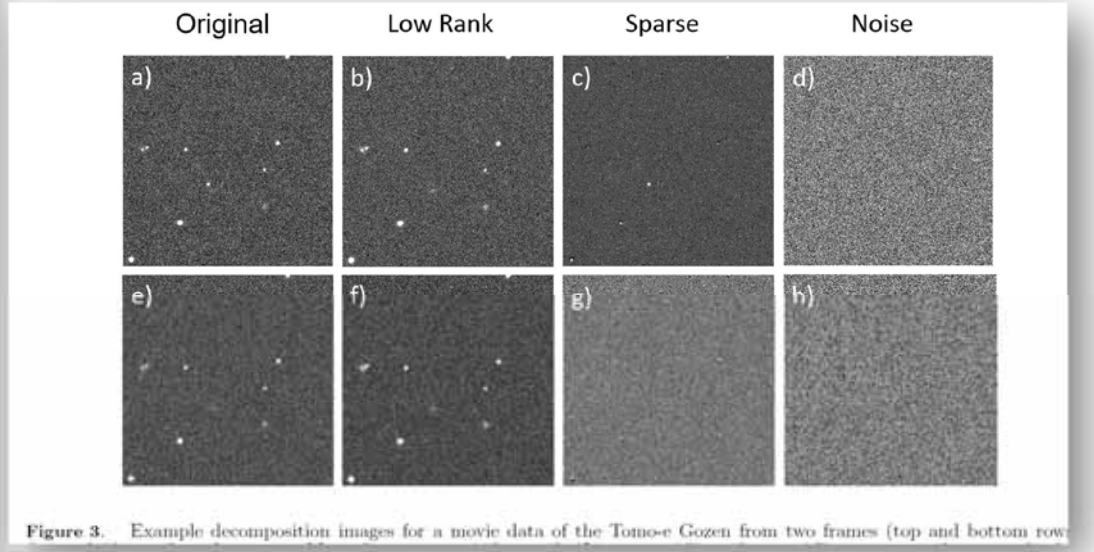


Figure 3. Example decomposition images for a movie data of the Tomo-e Gozen from two frames (top and bottom row

Morii et al., 2017, ApJ

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
  - データ量の制約から、複数フレームを積分したデータを保管・公開
  - 検証のために元のフレームレートの生データも一部保管
  - 保全、公開に重要なデータを十分議論の上で決める

全て残すことが技術的には可能でも、予算的に困難な場合もある

だから

必要な情報を残してアーカイブ保管するために、システムズエンジニアリング的なアプローチが必要

- データレートが高すぎる場合、データ取得時に非可逆圧縮を組み込むなど、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必要がある

大量データを残して活用するために

- 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

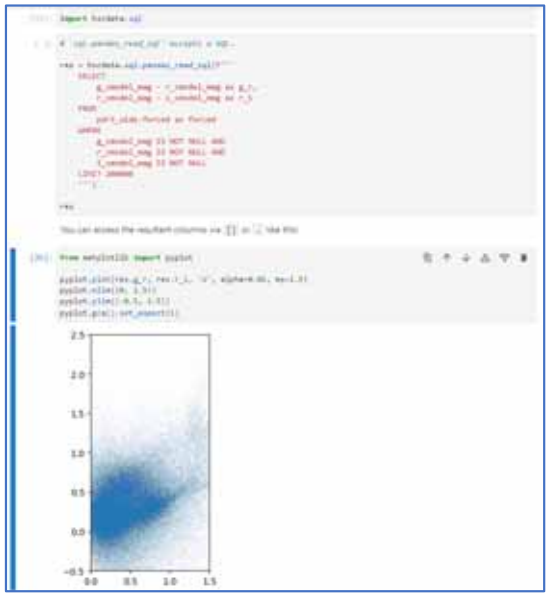
現実的にドメイン知識・技術の全てをADC内に確保できなくても、アーカイブデータを資産として活用することで、ADCは大学等研究機関や企業と研究開発協力するためのハブ、更にはデータ科学の推進拠点となることを目指したい

ADCが（或いは国立天文台全体として）期待されていること（既に実施中を含む）

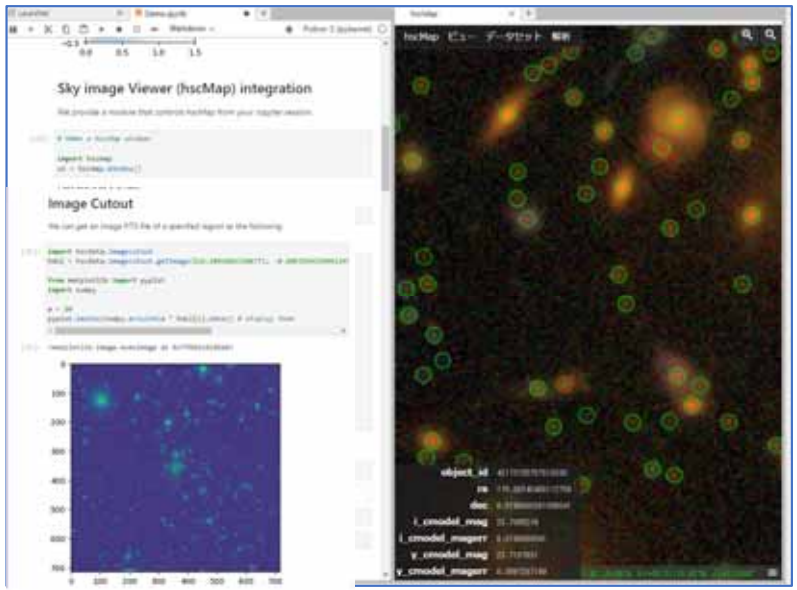
- 国立天文台が運用に絡む望遠鏡や観測装置の観測データの恒久的な保管
- 国内の大学等が運用する望遠鏡や観測装置の観測データの受入、保管と公開
- データ較正やカタログ化のソフトウェアやパイプラインの開発、運用、公開
- 観測データをすぐに研究に使えるよう、できる限りそのまま物理量として扱えるようにしてから公開（生データに加え、較正済みデータやカタログなどを公開）
- データフローシステムの設計、開発（生データが残せない高データレートの場合は、必要情報のみ抽出保管するためのシステム全体の最適化が必要）
- データへのDOIの付与と管理（機関レポジトリ）
- 観測データの利活用の促進（講習会など）
- 観測データの利活用環境の提供（サイエンスプラットフォーム等）
- データの巨大化への対応
  - 高速解析パイプラインの開発と運用、公開
  - 大規模高速解析計算機システムの構築と運用、公開
  - 大規模高速データベースの開発と運用、公開
  - データ解析、分析、可視化、公開などへの天文以外のドメイン知識の収集と活用
- 観測データを用いたデータ科学の推進拠点としての役割
- 国際プロジェクトへの貢献としての計算機環境整備と運用（国立天文台として参加が正式に決まればサポート）

サイエンスプラットフォーム構想

柔軟なプログラム環境
Jupyter+仮想化PCクラスタ

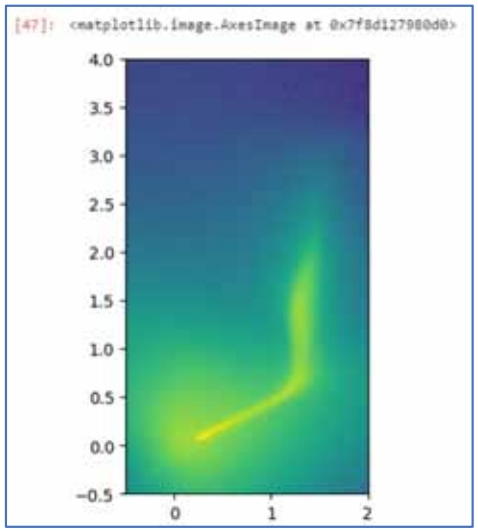


天体カタログの検索･可視化



画像ファイル(e.g., hscMap)へのアクセス

```
[47]: import numpy

def map(patch):
    where = patch('forced.i.extendedness_value') < 0.5
    patch = patch[where]
    r = patch('forced.r.psfflux_instmag')
    g = patch('forced.g.psfflux_instmag')
    i = patch('forced.i.psfflux_instmag')
    return numpy.histogram2d(
        g - r,
        r - i,
        range=((-0.5, 2), (-0.5, 4)),
        bins=(200, 400)
    )

def reduce(a, b):
    a_hist, a_xbins, a_ybins = a
    b_hist, b_xbins, b_ybinx = b
    return a_hist + b_hist, a_xbins, a_ybins

%time res = api.mapreduce(map, reduce)

hist, xedges, yedges = res
pyplot.imshow(numpy.log(1 + hist.T), origin='lower', extent=(xedges[0], xe

[00:00<00:00]
CPU times: user 200 ms, sys: 43.9 ms, total: 244 ms
```

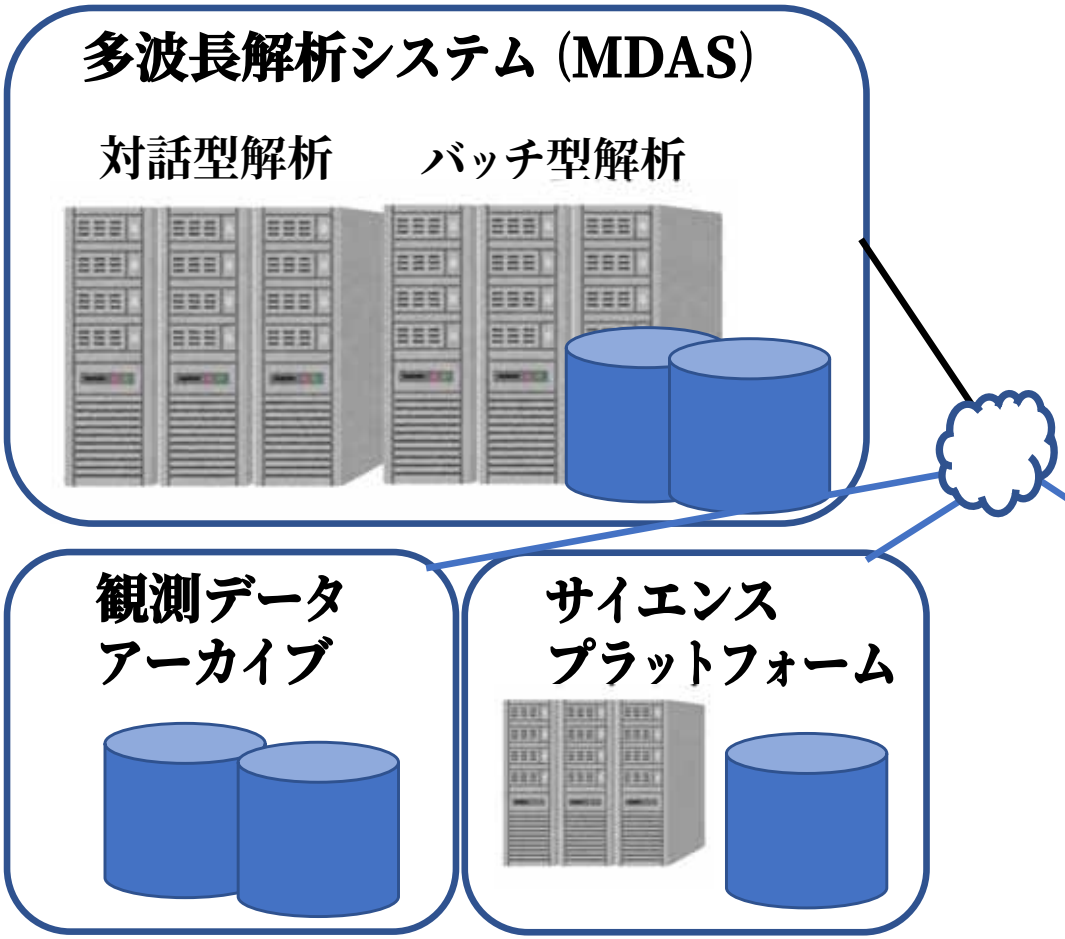


高速DB+ユーザ定義の分散処理 (将来は非同期処理も可能に)
SQLで難しい高レベルな解析の実現 (例は色が似ている近傍天体検索)

(Koike, M.)

巨大データへの技術的な対応：
大規模データ解析システムの構築と提供

1. 小規模解析のMDAS、観測生データアーカイブ、サイエンスプラットフォームと相補的に研究をサポート
2. HSCからPFS, ULTIMATEへと続く重点的なデータ解析需要に対応
3. 初代システムが5年目を迎える → 更新を進めていく



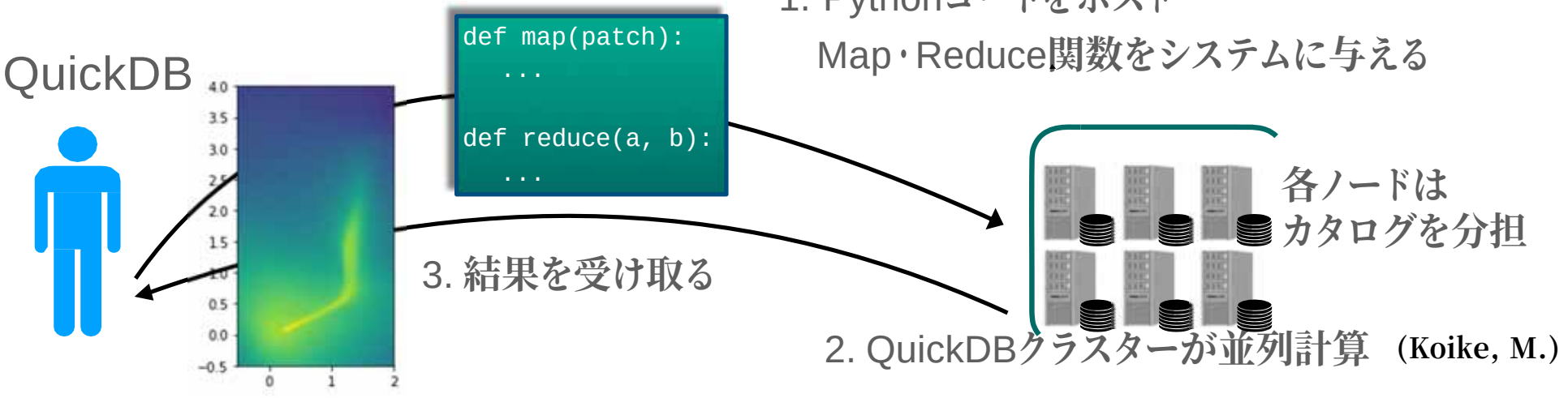
大規模観測データ解析システム（LSC）

CfCA・ハワイ観測所等とも協力し，
研究計画推進に最適な環境を構築

高速なファイルシステム(5PB)
＋計算ノード等(~2kcore)

巨大データへの技術的な対応：
高速データベース技術開発とその科学応用の推進

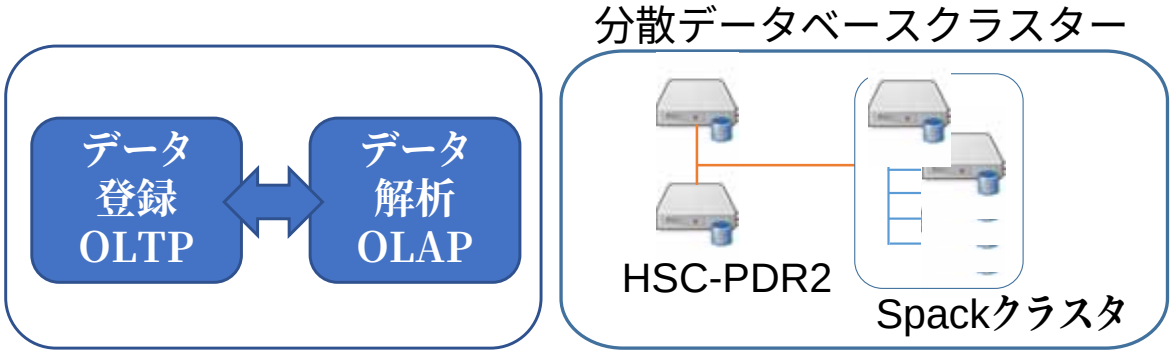
1．列指向分散データベースの構築とHSCデータへの応用



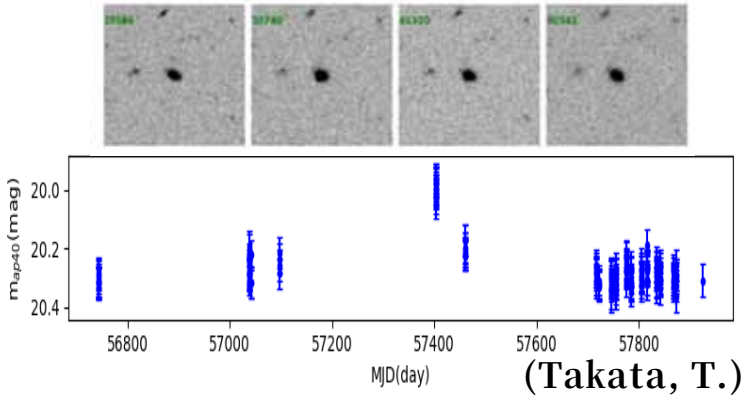
(Koike, M.)

2．新リレーショナルDB開発と巨大時系列データ解析・変動検知の実現

トランザクション・解析両立型DB処理系の実証に参加
→ 100億行規模の自在な検索を可能に



HSC-SSPの時系列測定DBの構築
→ 変動天体研究への応用



(Takata, T.)

観測データ科学の推進・研究基盤の構築

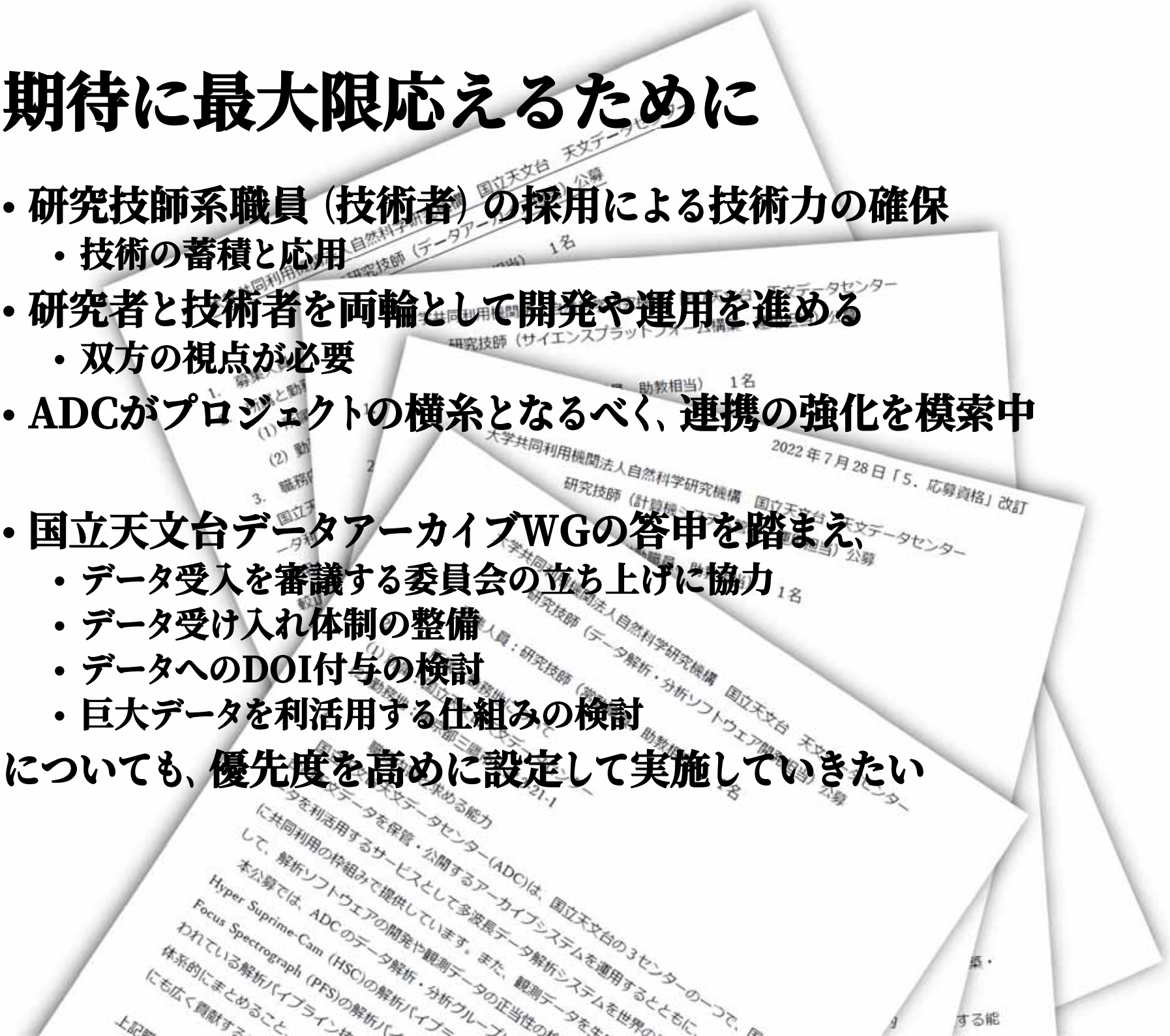
- 大規模化するデータへの技術対応による研究成果の促進
  - アーカイブと連動して多波長データを効率的に解析するユーザ環境の構築
  - 高速なデータベース技術開発
  - データ科学の応用
- 多波長データ・環境・ツール ＋ 研究者 の集約による研究基盤構築
  - 2020年代後半−2030年代の観測ミッションとの連動を目指す
    (HSC・PFS, Rubin, Euclid, Roman, ULTIMATE-Subaru, ALMA, TMT)

期待に最大限応えるために

- 研究技師系職員（技術者）の採用による技術力の確保
  - 技術の蓄積と応用
- 研究者と技術者を両輪として開発や運用を進める
  - 双方の視点が必要
- ADCがプロジェクトの横糸となるべく、連携の強化を模索中

- 国立天文台データアーカイブWGの答申を踏まえ、
  - データ受入を審議する委員会の立ち上げに協力
  - データ受け入れ体制の整備
  - データへのDOI付与の検討
  - 巨大データを利活用する仕組みの検討

についても、優先度を高めに設定して実施していきたい



Astronomy Data Center
国立天文台 天文データセンター

国立天文台 NAOJ National Astronomical Observatory of Japan

天文データセンターに期待する役割について、忌憚のないご意見をお寄せ下さい