

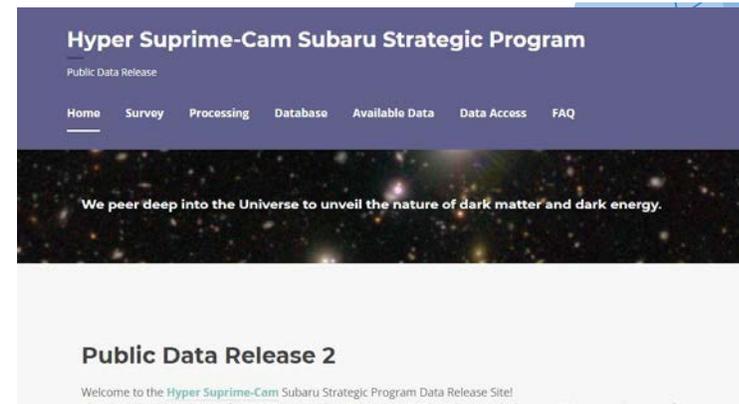
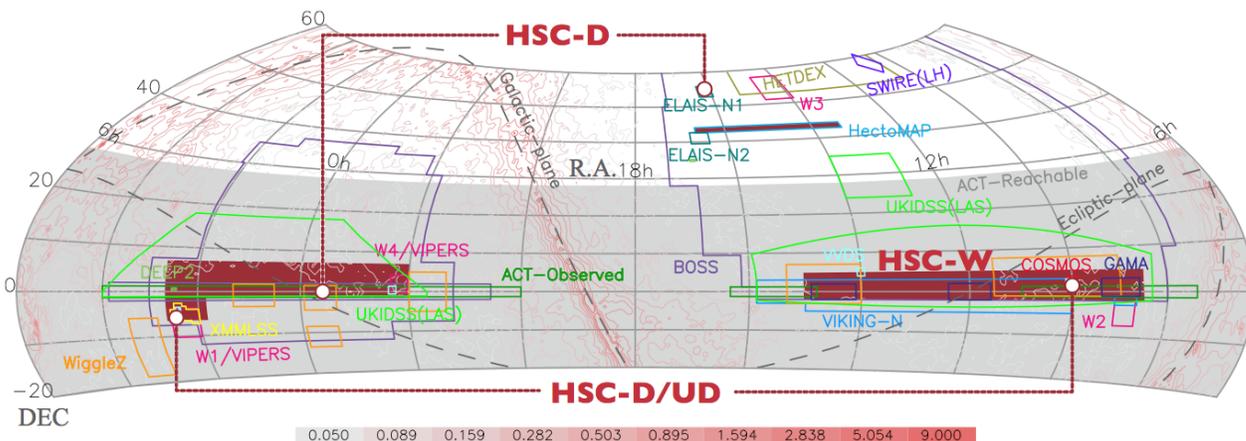
# HSC(-SSP)データアーカイブ

2020.1.29 データアーカイブワークショップ  
古澤 久徳@国立天文台 (ADC・ハワイ観測所)

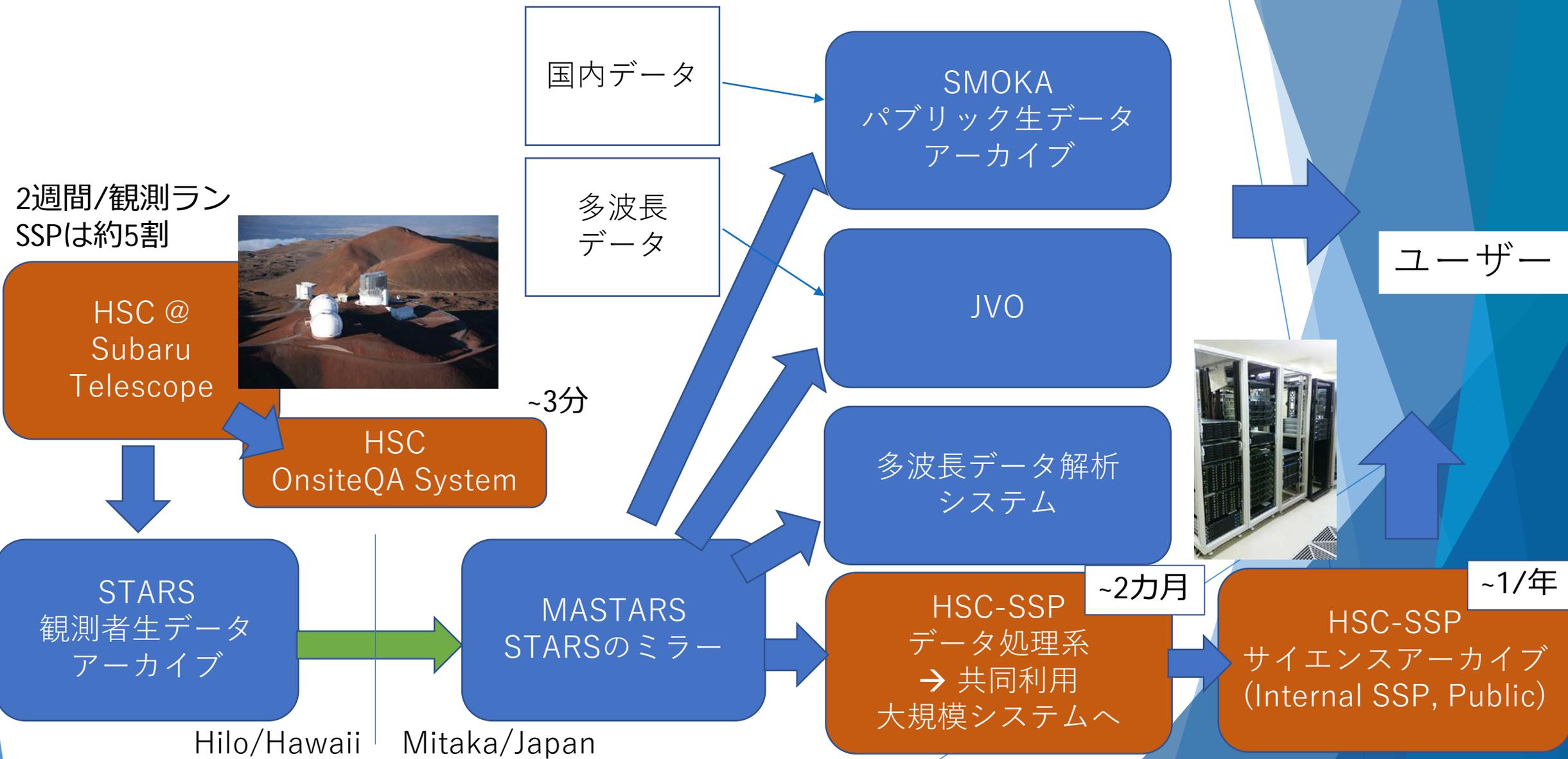
# HSCすばる戦略枠プログラム(SSP)

- ▶ マルチバンド撮像サーベイ (grizy + 4 narrowbands)
- ▶ 期間: 300+30 夜 (2014.3-present; 5年+α ~S20B)
- ▶ 領域: 1400 sq. degree fields around equator
- ▶ Data volume: File:1PB file, DB:20TB (coadd)/200TB(ccd)
- ▶ プロジェクトの将来連携: PFS, Euclid, WFIRST?, ...
- ▶ 国際共同研究: 日本 (NAOJ, K-IPMU etc), Princeton, 台湾
- ▶ 国立天文台: データ解析 and データアーカイブ (生データ・処理済み)

<https://hsc-release.mtk.nao.ac.jp/>  
日本の研究者はCoIとして参加できる



# HSC戦略枠観測プロジェクト(SSP)とその周辺のデータフロー



# HSC-SSPデータアーカイブのミッション

8-9人：開発3-4、運用2-4、解析2

## 1. データを作る

- ▶ 科学的に信頼があり、再現・検証可能な、Science-readyのデータプロダクトを作る

## 2. データを公開する

- ▶ そのようなデータプロダクトを使いやすい形でユーザに提供する

## 3. ユーザを支援する

- ▶ データを用いた科学的・教育的な活動を支援すること

- ▶ 信頼のあるデータを作るには、「装置/観測所」・「パイプライン開発」との連携が不可欠

- ▶ 観測所・データプロバイダ) 装置維持改善、データのメタ情報・品質情報の整理・修正、生データ管理

- ▶ パイプライン) データ特性・コミュニティの要求を反映

→ 装置開発の段階からこの連携を構想し、開発に参加

フィードバック

アーカイブのたいへんさ  
求めるレベルにばらつき

# データの健全性・整合性評価 オンサイトQAとデータのタグ付け

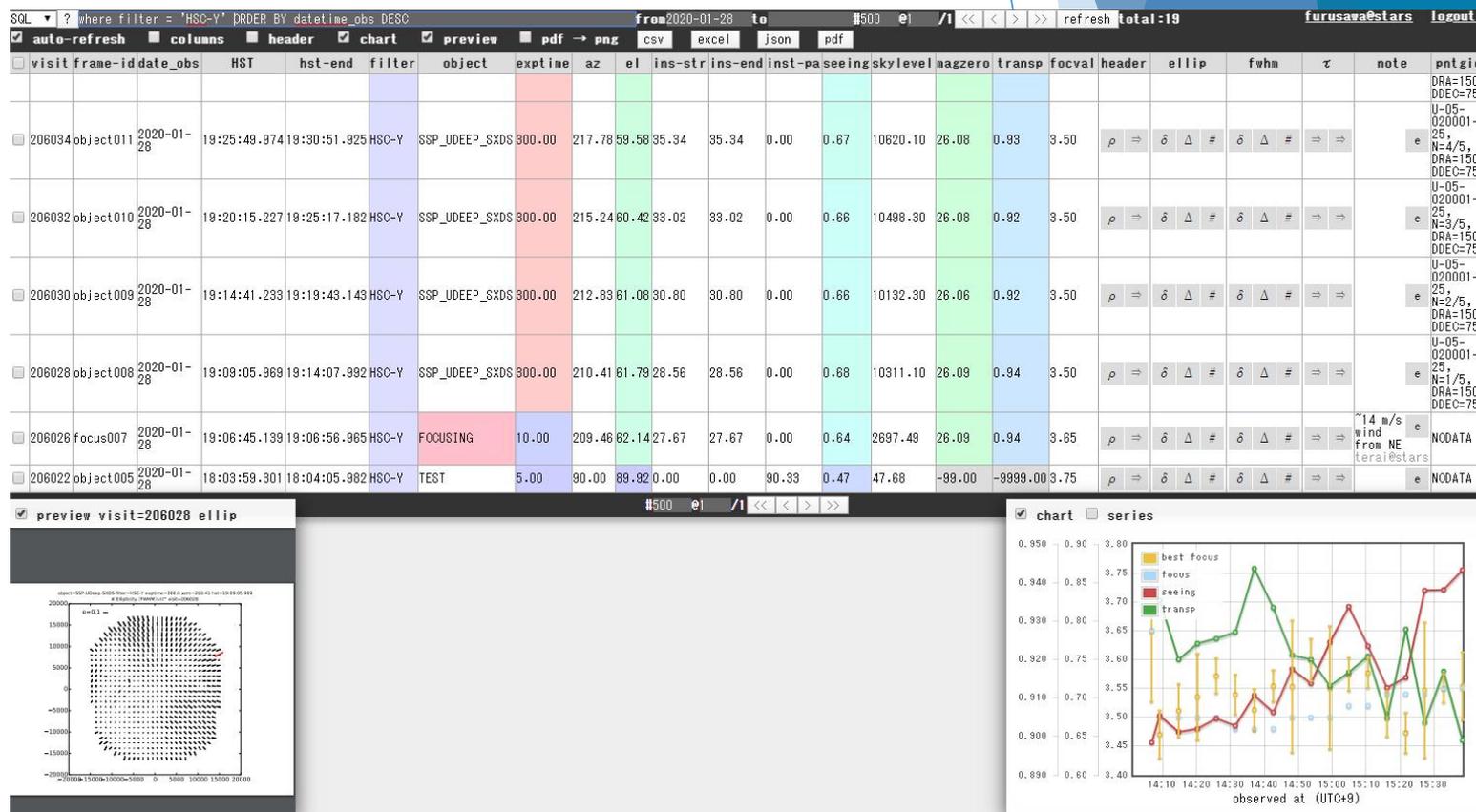
## ▶ オンサイトQAシステム

▶ シーイング・スカイレベル・  
透過率を測定

▶ DB記録・タグ付け  
→ データ解析時に利用

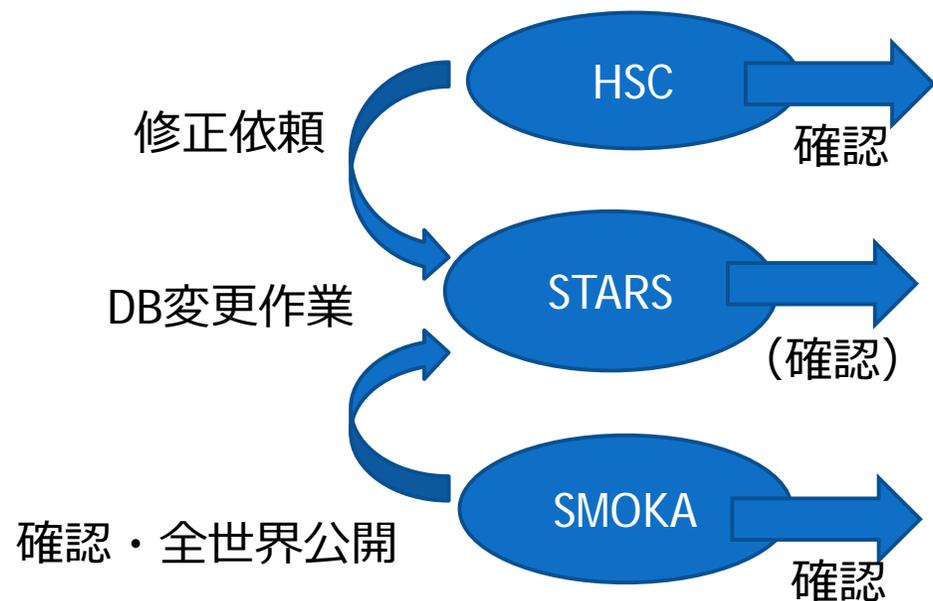
▶ キュー観測でも利用

▶ パブリックアーカイブへの  
情報付加(curation)も視野



# データの健全性・整合性評価 オンサイトQAとデータのタグ付け

- ▶ 非科学データを判別するヘッダキーワードの導入
  - ▶ テストバイアス・フラットなど
- ▶ STARS内のメタ情報修正
  - ▶ 登録後の修正の必要性
  - ▶ チェック機構の準備



	A	B	C	D	E	F	G	H
1	START_FRAME	END_FRAME	KEYWORD	CURRENT	UPDATED	DATATYPE	COMMENT	181220100
2	HSCA15297400	HSCA15297557	T_SEEING	0	-9999	float	update for QA keyword	OK
3	HSCA15297400	HSCA15297557	T_TRANSP	0	-9999	float	update for QA keyword	OK
4	HSCA15297400	HSCA15297557	T_MAGZER	0	-99	float	update for QA keyword	OK
5	HSCA15297400	HSCA15297557	T_PSFELL	0	-9999	float	update for QA keyword	OK
6	HSCA15297400	HSCA15297557	T_PSFPA	0	-9999	float	update for QA keyword	OK
7	HSCA15297400	HSCA15297557	T_WCSRMS	0	-9999	float	update for QA keyword	OK
8	HSCA15297400	HSCA15297557	T_ZPRMS	0	-99	float	update for QA keyword	OK
9	HSCA15297400	HSCA15297557	T_N_WCS	0	-9999	float	update for QA keyword	OK
10	HSCA15297400	HSCA15297557	T_N_ZP	0	-9999	float	update for QA keyword	OK
11	HSCA15297400	HSCA15297557	T_OSLVL	0	1266.9	float	update for QA keyword	OK
12	HSCA15297400	HSCA15297557	T_FLTNSS	0	7.195	float	update for QA keyword	OK
13	HSCA15297600	HSCA15297757	T_SEEING	0	-9999	float	update for QA keyword	OK
14	HSCA15297600	HSCA15297757	T_TRANSP	0	-9999	float	update for QA keyword	OK
15	HSCA15297600	HSCA15297757	T_MAGZER	0	-99	float	update for QA keyword	OK

# サイエンスアーカイブサービス

## ▶ カタログ検索

▶ SQLによる検索

▶ 2000カラム x 7億行  
→ 時系列情報200億行へ

## ▶ 画像検索・ファイル提供

▶ 現在~500TB → 最終1PB/DR

▶ アーカイブだけで3PB →  
将来6PBくらい必要

▶ 生データは10分の1以下

## ▶ インタラクティブビューワ

The screenshot displays the Science Archive Service interface. At the top, it shows the name 'catalog-job 2016-05-31'. The main area is divided into three panels:

- Query Interface:** Contains a SQL query for selecting astronomical data from the 's15b\_udeep' database. The query filters for objects with RA between 34.0 and 36.0 degrees and declination between -5.0 and -4.5 degrees. It includes a warning about the 'LIMIT 10' and instructions to edit the schema name.
- List of tables:** A tree view showing the database structure, including tables like 'mosaic\_measlist\_deepcoadd' and 'mosaic\_forcephoto\_deepcoadd'.
- Schema browser:** A table listing the columns and their properties for the 'mosaic\_measlist\_deepcoadd' table. The columns include 'id', 'ra2000', 'decl2000', 'img\_kron', 'img\_kron\_err', 'ymag\_kron', and 'ymag\_kron\_err'.

Below the query interface, there are options for output format (CSV, CSV.GZ, SQLite3, FITS) and a checkbox for 'syntax check before enqueueing'. An 'estimate query time' button is also present.

At the bottom, a table titled 'Objectid Coordinates Magnitude stuff' displays the results of the query. The table has columns for 'id', 'ra2000', 'decl2000', 'img\_kron', 'img\_kron\_err', 'ymag\_kron', 'ymag\_kron\_err', and 'i\_y'. The first few rows of data are:

id	ra2000	decl2000	img_kron	img_kron_err	ymag_kron	ymag_kron_err	i_y
37484571888986404	35.057607259963	-4.93732873437365	25.116035214154	0.114107063765164	499.99		-474.873964785846
37484571888981864	35.0078396962661	-4.99930759079772	24.7384120063082	0.157894468090058	499.99		-475.251587993694
37484571888981830	35.0170633972484	-4.99999974304848	25.2428014388663	0.049877325305955	24.7206126105315	0.119918698842417	0.522188828164801

# サイエンスアーカイブサービス

- ▶ カタログ検索
- ▶ 画像検索・ファイル提供
- ▶ インタラクティブビューワ
  - ▶ CAS、DASと連動
  - ▶ HiPS画像表示

The screenshot displays the Science Archive Service interface, which includes a main star field view and several interactive tool windows:

- catalogs**: A table listing files with columns for name, rows, color, sparse, DL, and x. The table contains two entries: '29542.csv.gz' (10 rows, green) and 'default'.
- quarry**: A search tool with input fields for 'ra1', 'dec1', 'ra2', and 'dec2', and a 'filter' dropdown set to 'HSC-G'. It also has checkboxes for 'image', 'mask', and 'variance', and a 'rerun / tract' section.
- 29542.csv.gz**: A plot showing a distribution of points in a 2D space with axes labeled 'ra2000' and 'dec12000'.
- Fits Images**: A window showing a list of files, including '-I-9940-s18a\_wide-190301-020802.fits' (1506 1261 files selected).
- color**: A 'Color Matrix' window with a grid for selecting colors and a 'scale' slider.
- Analysis**: A window showing a spectrum plot with a peak at 130.2481 and a filter set to 'HSC-I' at 3.7947'.

# PFS・LSSTに向けた潮流

- ▶ データセットの大規模化にともない、データのより近くで処理を、というのがキーワード（トレンド）になっている
- ▶ データ解析作業のための仮想化環境を提供する準備
  - ▶ Jupyter、Container(Docker)、Kubernetes

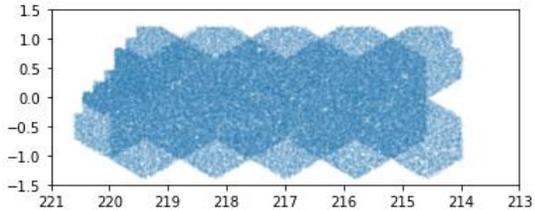
```
scidb2_20181122.gal_specLine
--WHERE
--- you probably want to specify field, (ra,dec), redshift, redshift confidence, etc.
...

# query the database. you will get the result in pandas dataframe
data = pfsdb.query_pandas(sql)
```

[4]: # Let's look at the spatial distribution

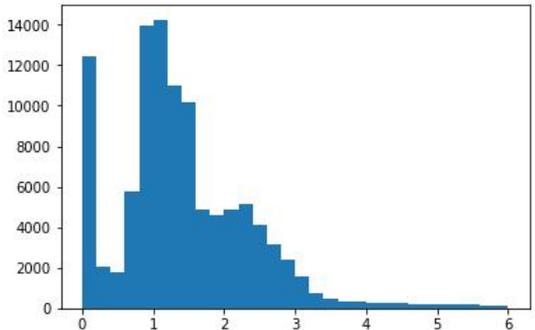
```
from matplotlib import pyplot

pyplot.plot(data.ra, data.dec, 'o', alpha=0.1, ms=1.5)
pyplot.xlim((221, 213))
pyplot.ylim((-1.5, 1.5))
pyplot.gca().set_aspect(1)
```



[5]: # now Let's Look at the redshift distribution

```
res = pyplot.hist(data.z,bins=30)
```



# データ運用面で 良かったこと・満足していないこと

- ▶ 良かった) 装置開発の初期から装置とデータ解析・アーカイブの連動を構想に入れたこと。
  - ▶ 結果、長期のサーベイの進行度をそこそこ客観的に把握でき、解析も安心して行えた。
- ▶ 良かった) HSCサイエンスアーカイブを世に出せていること
- ▶ 要改善点) 本来求めたいレベルにまでヘッダや運用モデル、システムを作りこんでから観測開始できなかったこと。
  - ▶ 実際の運用が始まると、理想に向けて大きく改変することは安全性・マンパワー双方で極めて難しい。
  - ▶ ヘッダ等の不備や本来データ運用に大いに役立つはずだった未実装のキーワードをいつまでも残してしまったり、システムも保守性が悪い
- ▶ 要改善点) データ検査や修正作業のためのルールやツールを作りこめていないこと。
  - ▶ 問題の検知漏れが起きやすく、作業効率も悪い
- ▶ 要改善点) データの品質評価に十分な時間が取れておらず、科学利用の促進のための整備・改善に注力できていない