CENTER FOR ASTROPHYSICS
HARVARD & SMITHSONIAN

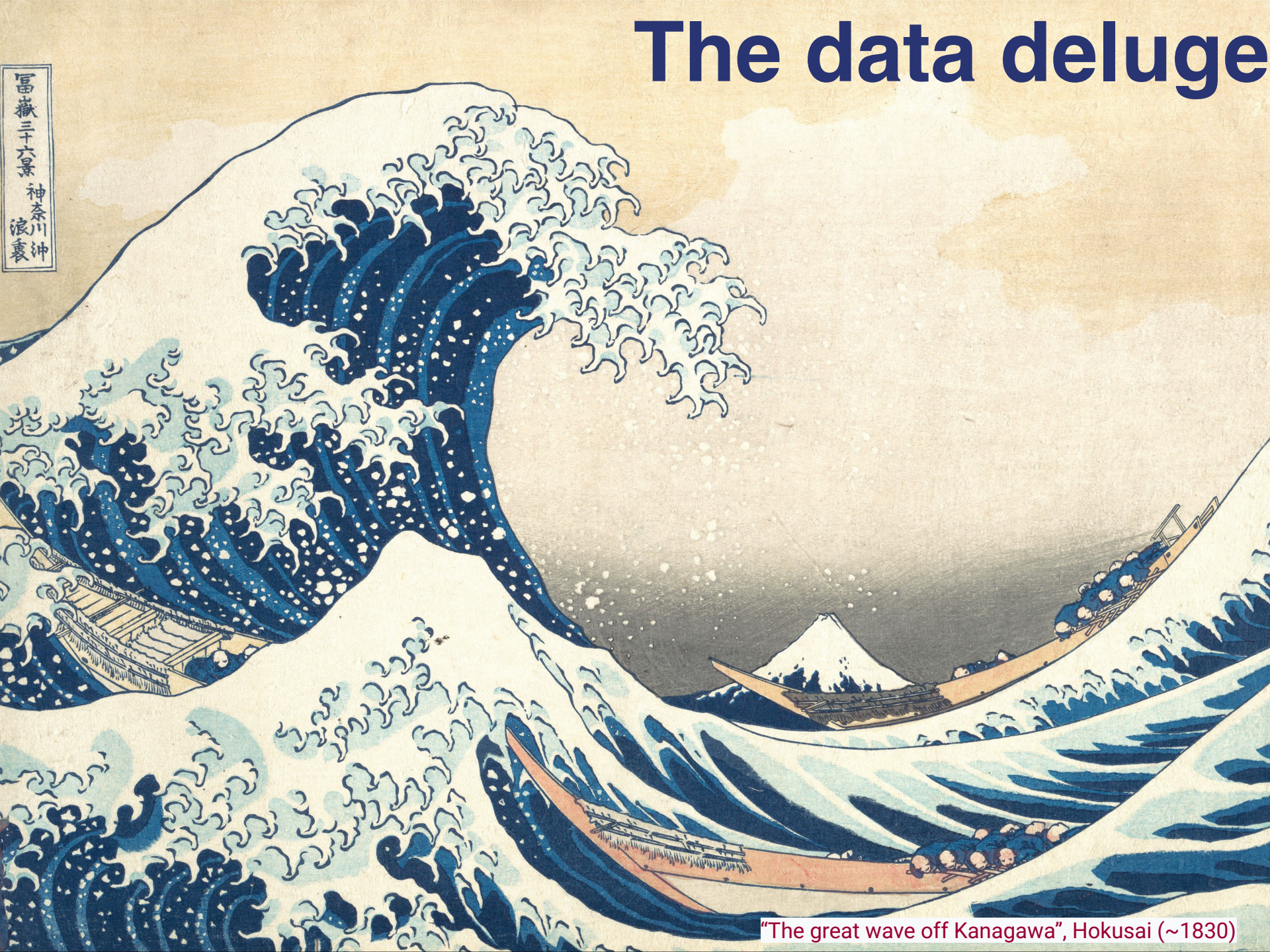
Astronomy in the next decades: steps towards a benevolent (archival) panopticon

Raffaele D'Abrusco
and the Archive Operations team

Chandra X-ray Center

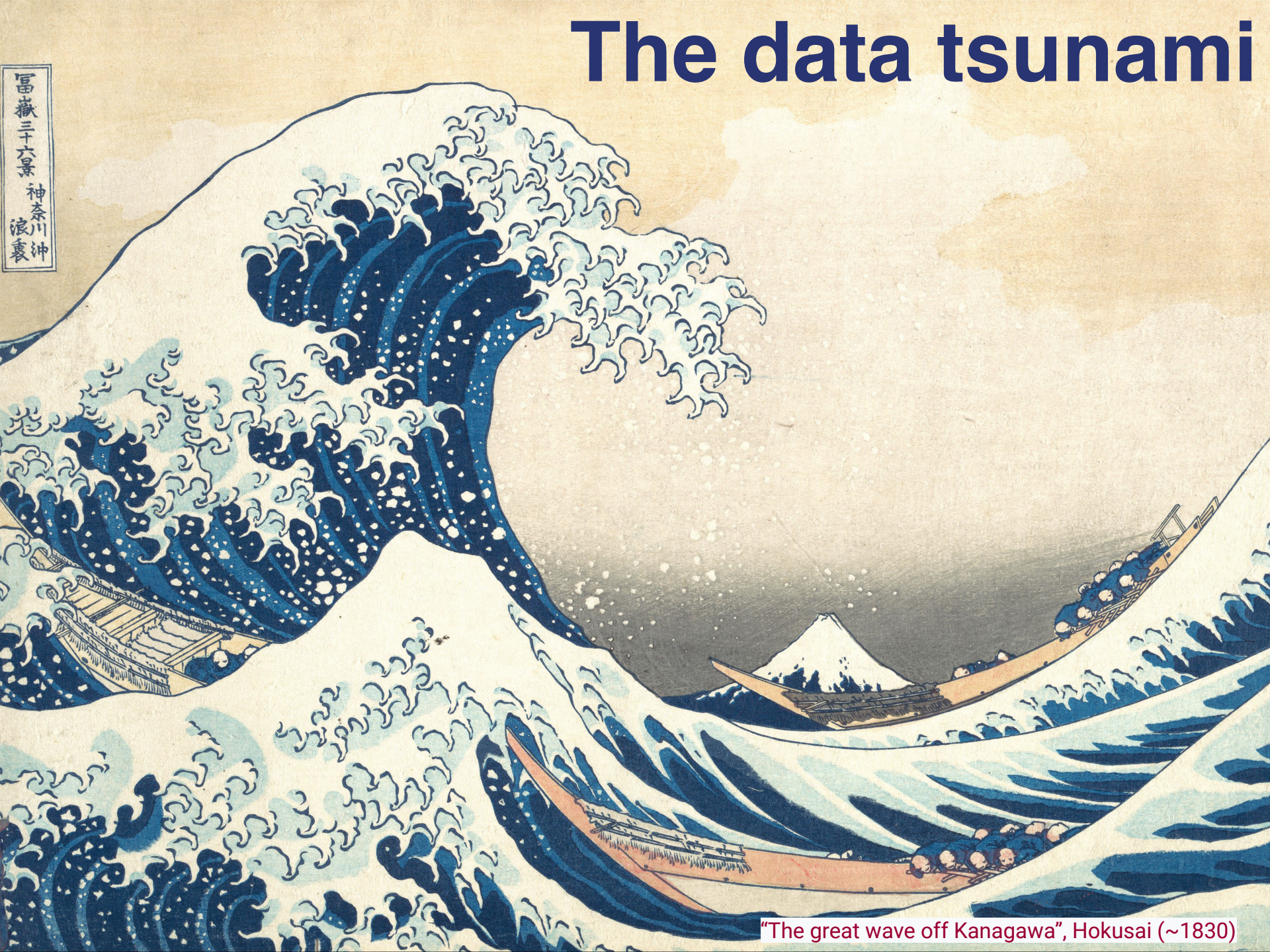
November 21, 2019

The data deluge



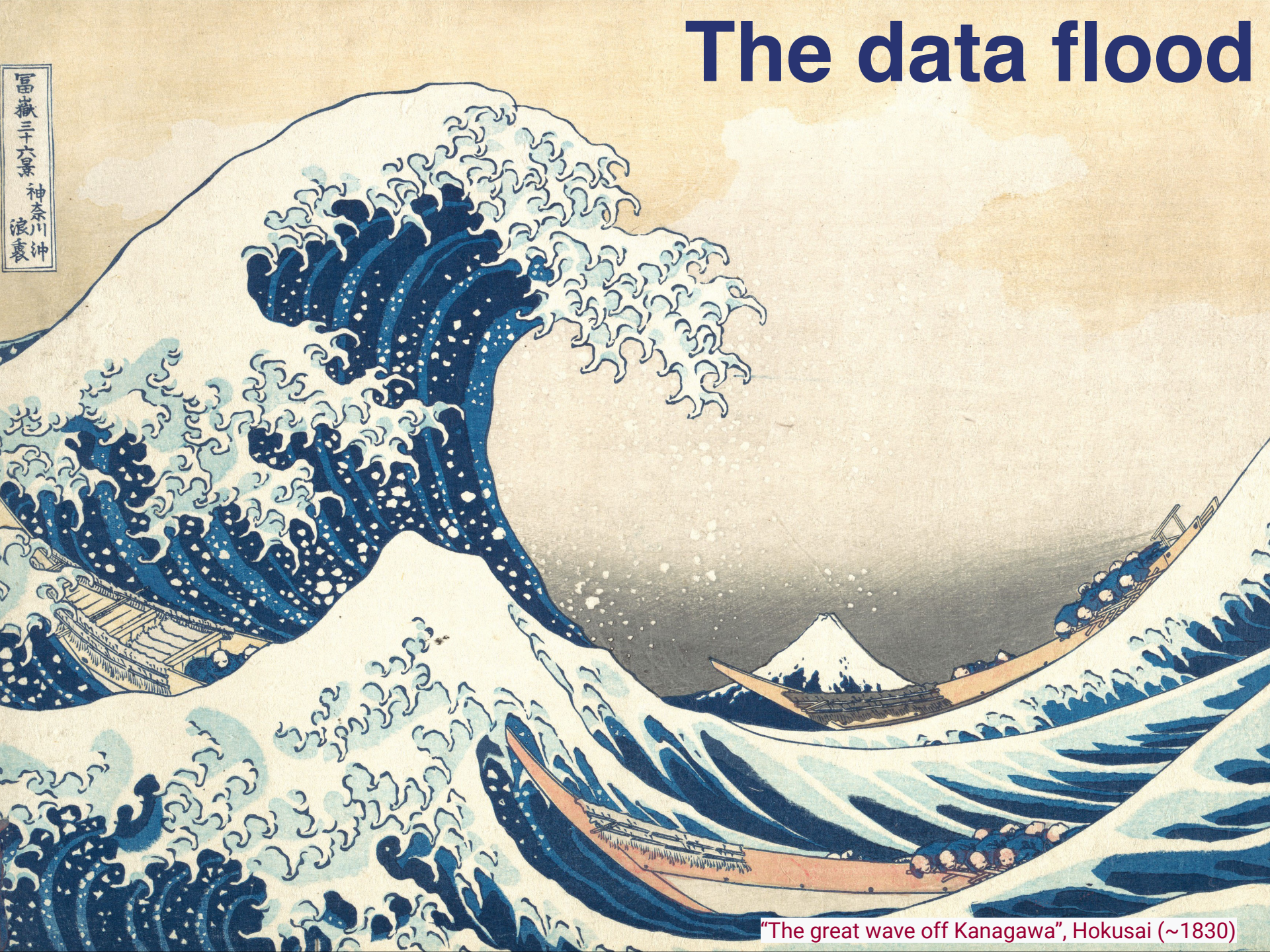
"The great wave off Kanagawa", Hokusai (~1830)

The data tsunami



"The great wave off Kanagawa", Hokusai (~1830)

The data flood



"The great wave off Kanagawa", Hokusai (~1830)

The (excessive) data cornucopia

Good times

>>> **Data accessibility never been better**

Archives/data centers/data providers focusing on accessibility

- Dealing with scale of data effectively
- Working on minimizing data transfer
- Compartmentalizing access to different levels of data hierarchy

Good times

>>> **Data accessibility never been better**

Archives/data centers/data providers focusing on accessibility

- Dealing with scale of data effectively
- Working on minimizing data transfer
- Compartmentalizing access to different levels of data hierarchy

>>> **Virtual Observatory significant contribution**

Standardization of data and metadata

- Data models
- Common observational metadata summary

Good times

>>> Data accessibility never been better

Archives/data centers/data providers focusing on accessibility

- Dealing with scale of data effectively
- Working on minimizing data transfer
- Compartmentalizing access to different levels of data hierarchy

>>> Virtual Observatory significant contribution

Standardization of data and metadata

- Data models
- Common observational metadata summary

Interoperability

- of tools - SAMP (Simple Application Messaging Protocol)
- of missions - to coordinate observational response

Good times

>>> Data accessibility never been better

Archives/data centers/data providers focusing on accessibility

- Dealing with scale of data effectively
- Working on minimizing data transfer
- Compartmentalizing access to different levels of data hierarchy

>>> Virtual Observatory significant contribution

Standardization of data and metadata

- Data models
- Common observational metadata summary

Interoperability

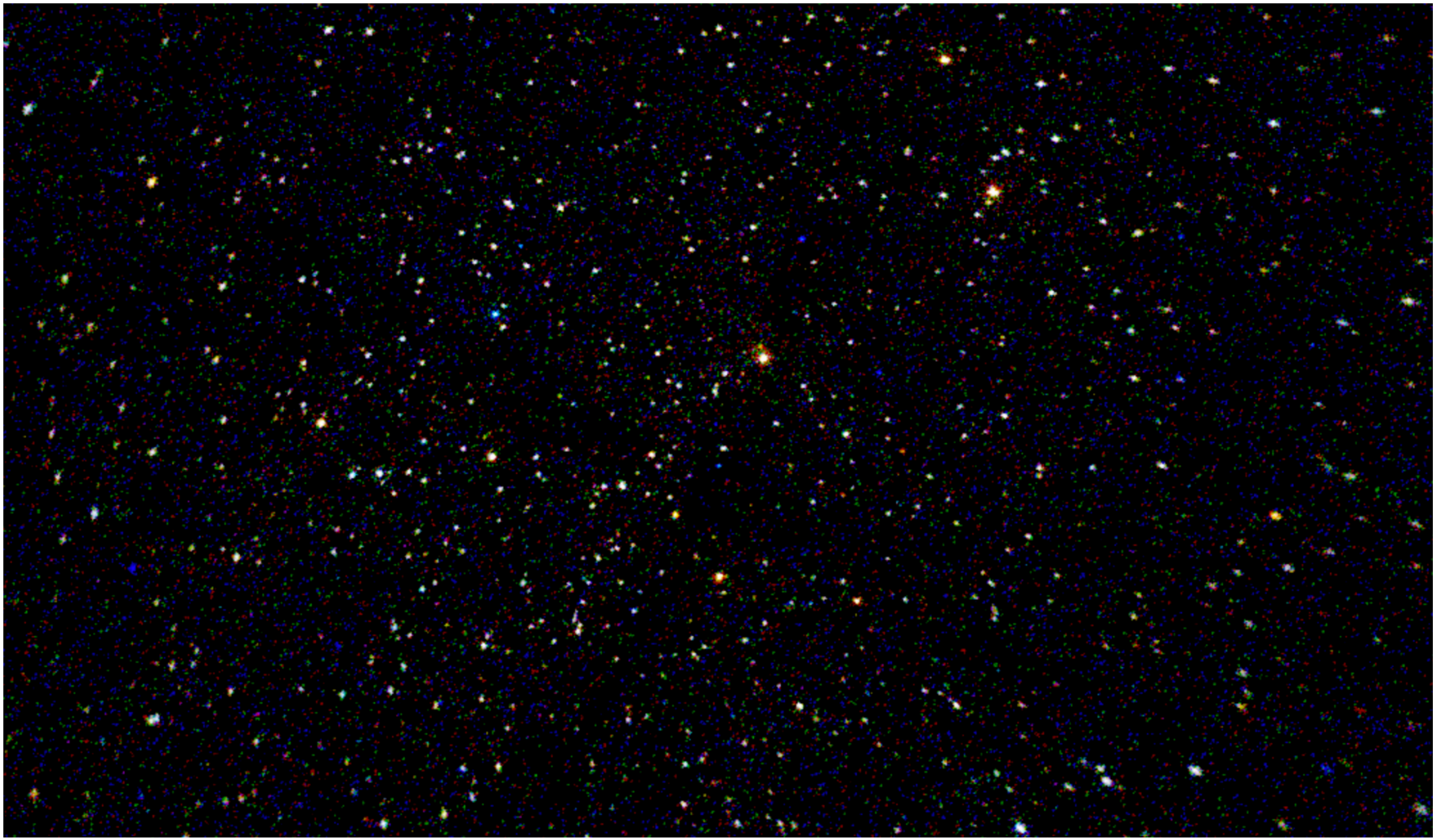
- of tools - SAMP (Simple Application Messaging Protocol)
- of missions - to coordinate observational response

Focus on visual exploration of data

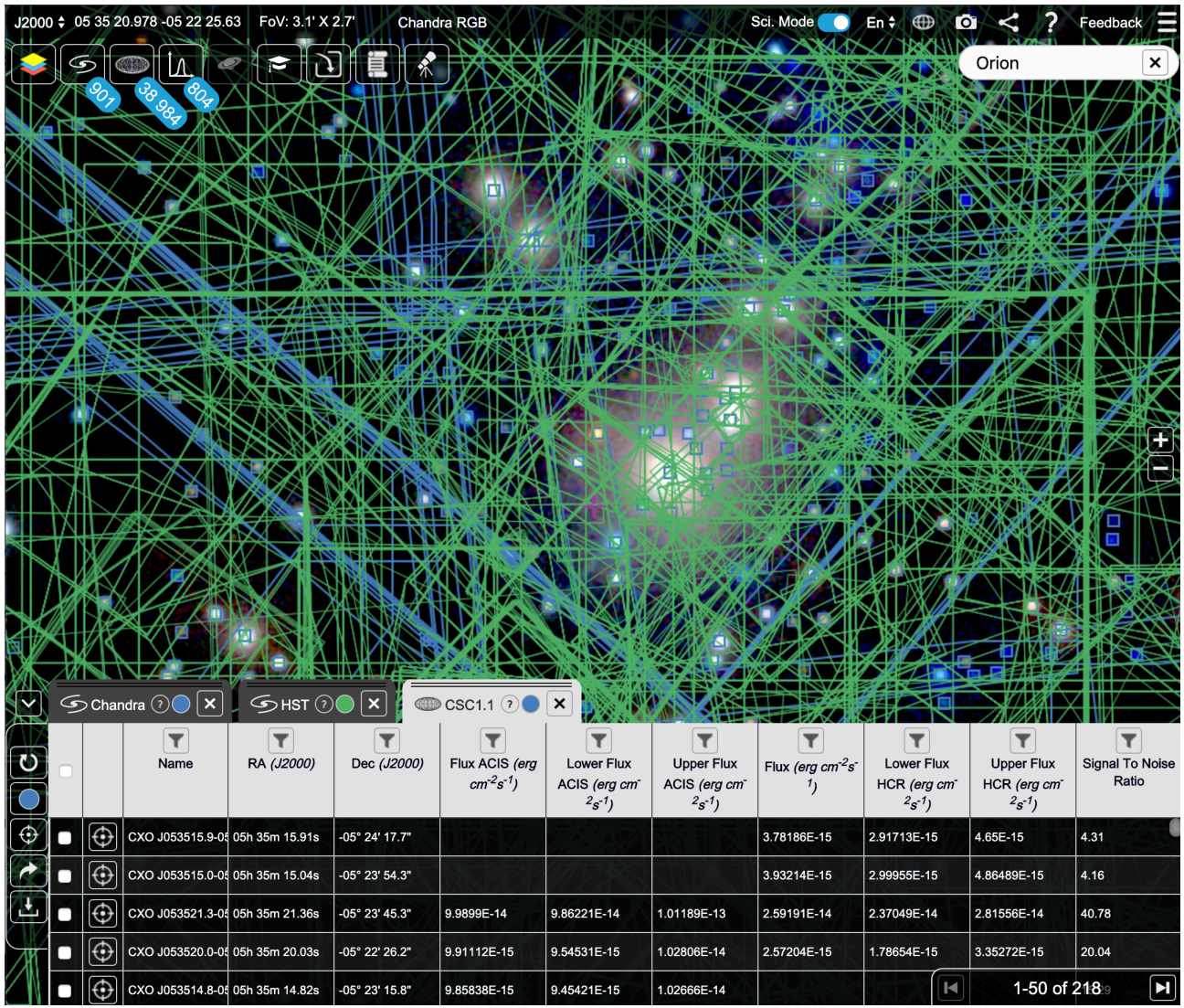
- Support for effective, scalable visualization
- Layered data visualization -> data exploration!

HiPS

Compressed, scalable, efficient data visualization standard



Exploration/visualization



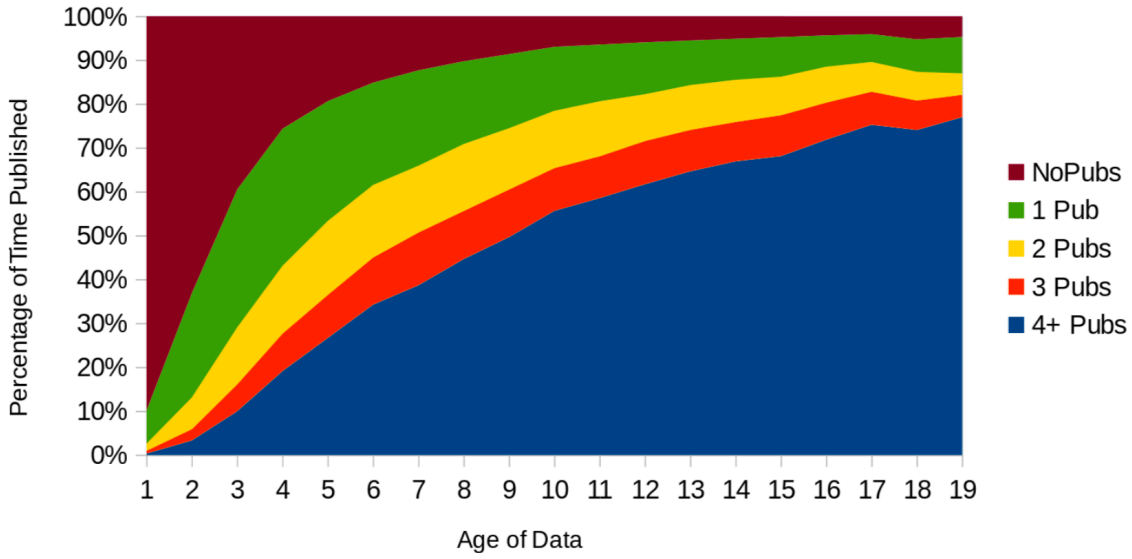
Data discovery quick, interactive and appealing

Lowered entrance threshold for data discovery

Made interactive data exploration pleasant!

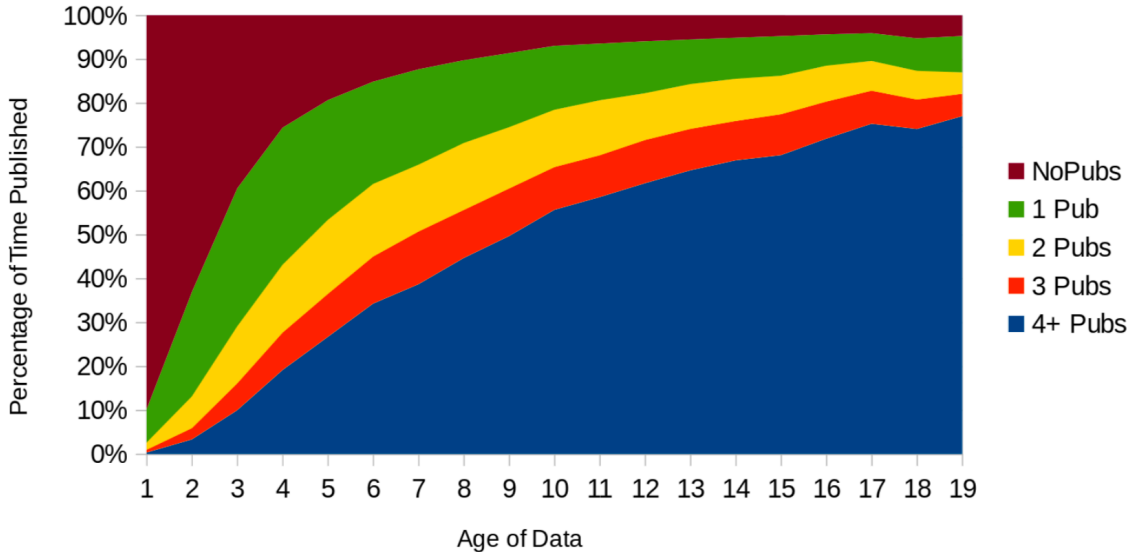
CENTER FOR ASTROPHYSICS
HARVARD & SMITHSONIAN

Chandra archival science

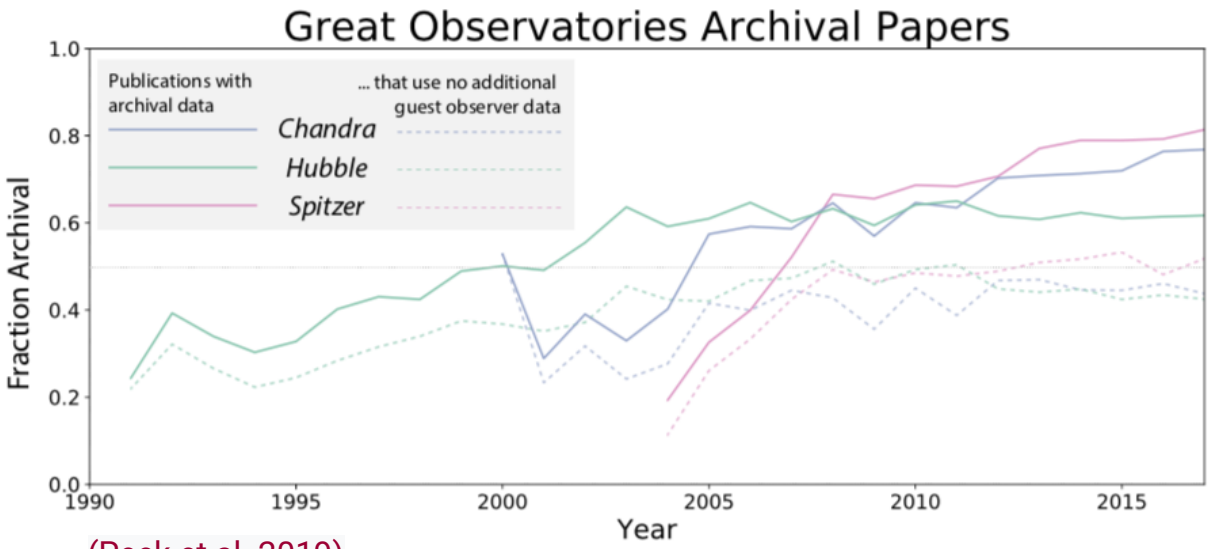


Chandra data are **heavily re-used** over their archival life

Chandra archival science



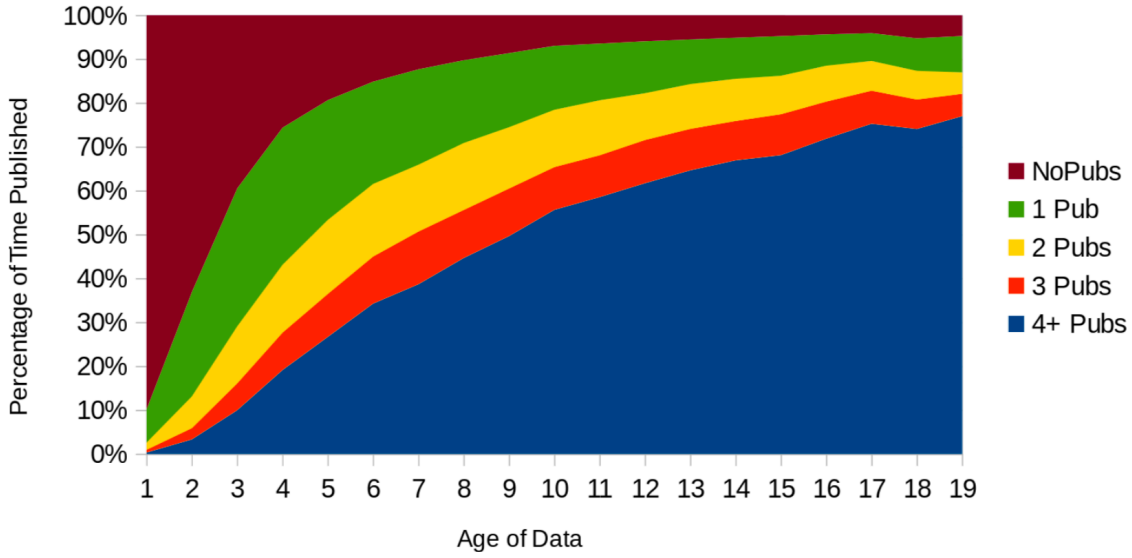
Chandra data are **heavily re-used** over their archival life



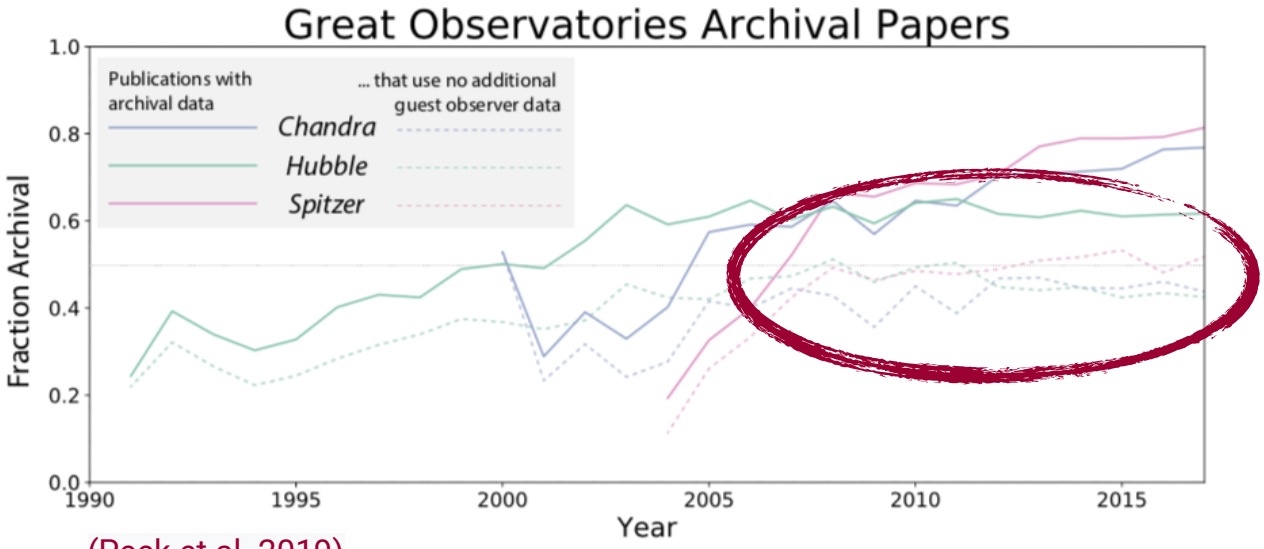
(Peek et al. 2019)

CENTER FOR ASTROPHYSICS
HARVARD & SMITHSONIAN

Chandra archival science



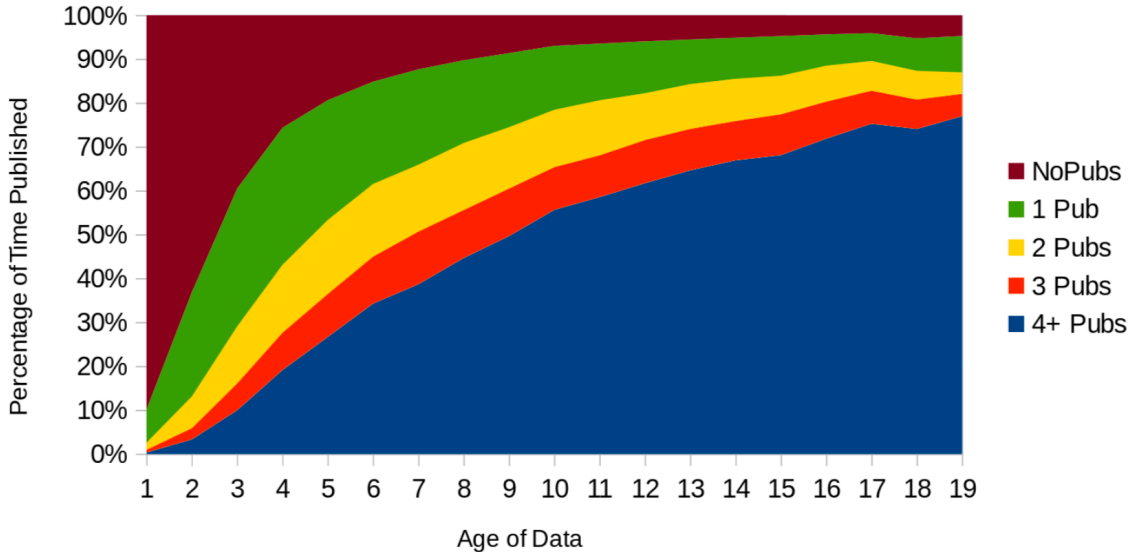
Chandra data are **heavily re-used** over their archival life



(Peek et al. 2019)

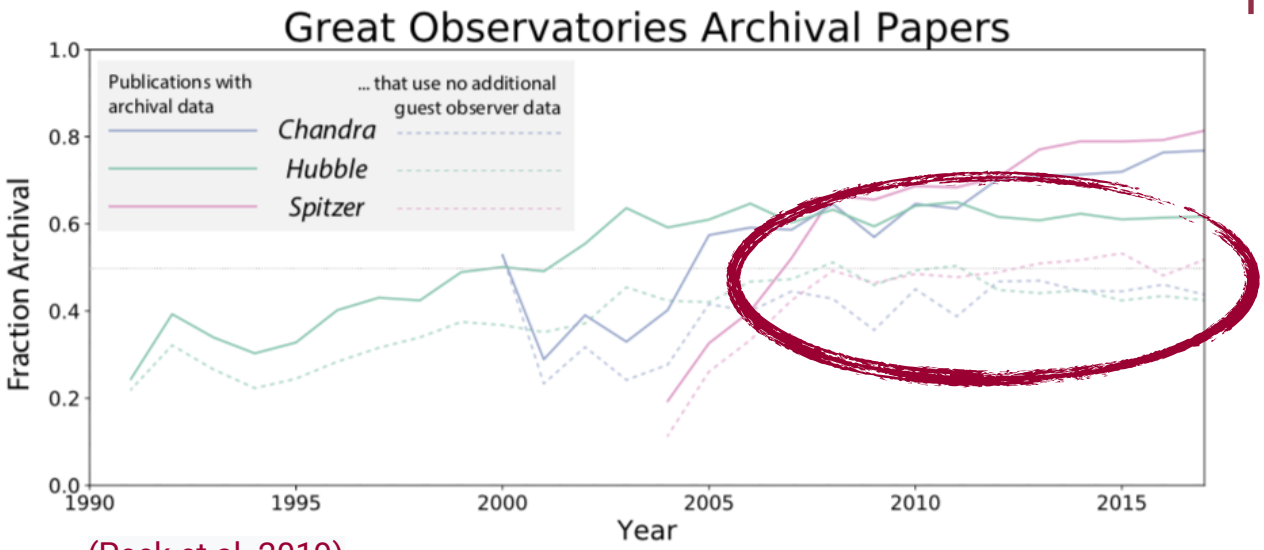
CENTER FOR ASTROPHYSICS
HARVARD & SMITHSONIAN

Chandra archival science



Chandra data are **heavily re-used** over their archival life

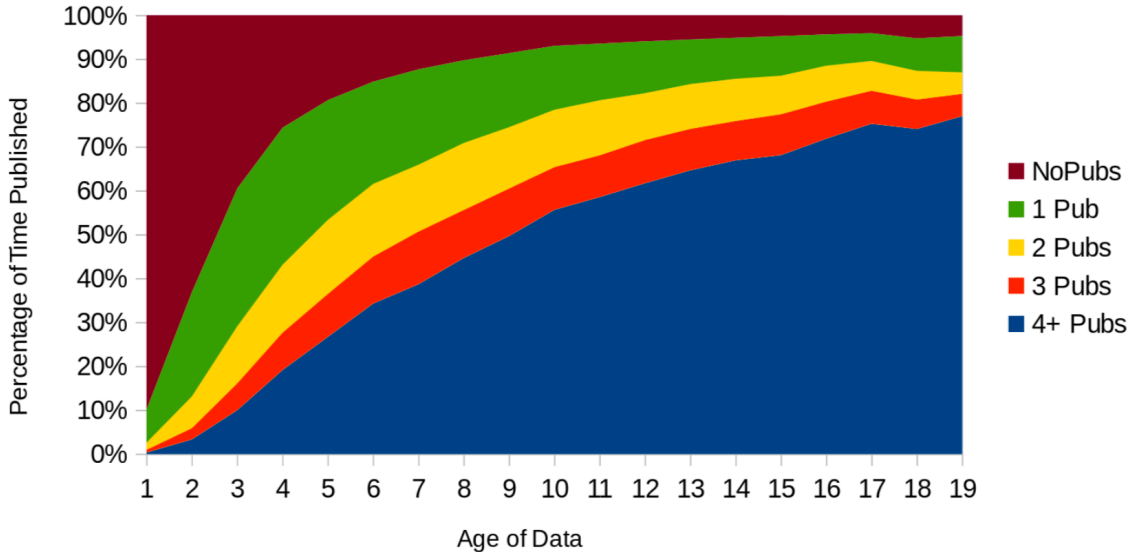
Fraction of archival *Chandra* papers plateaued in the last ~ 5 years



(Peek et al. 2019)

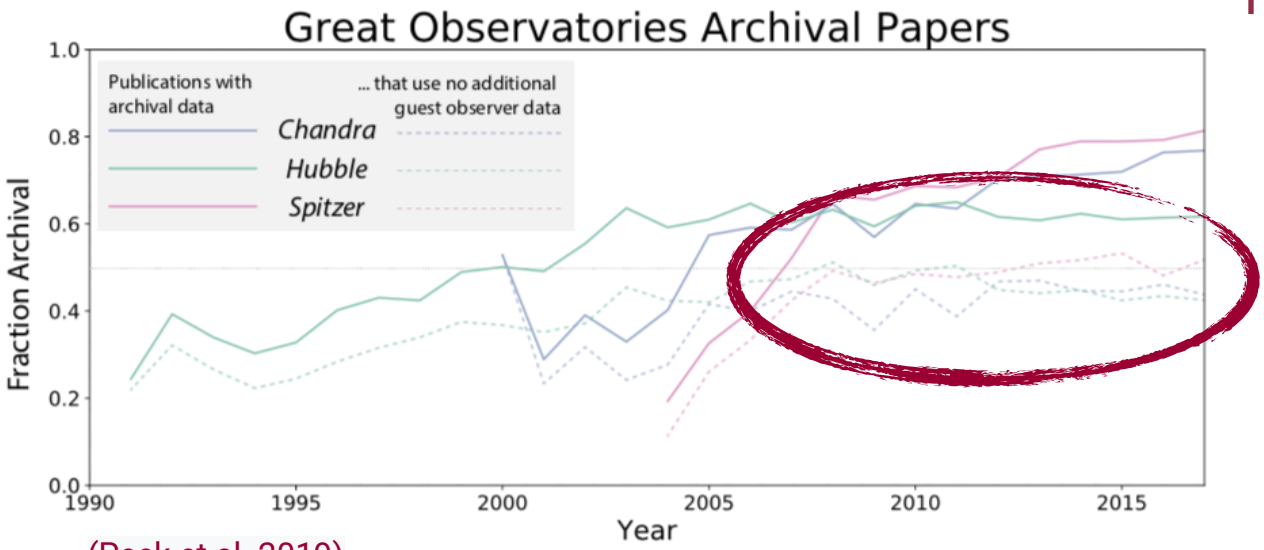
CENTER FOR ASTROPHYSICS
HARVARD & SMITHSONIAN

Chandra archival science



Chandra data are **heavily re-used** over their archival life

Fraction of archival *Chandra* papers plateaued in the last ~ 5 years

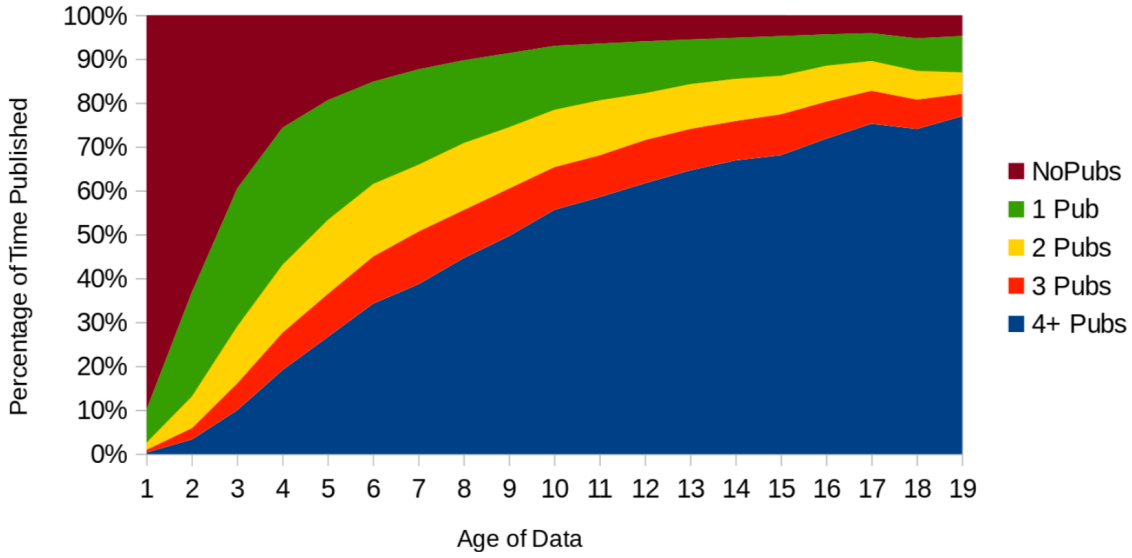


(Peek et al. 2019)

- Science content exhausted?
- Better data around?
- Untapped/able reservoirs?
- Low hanging fruits all picked?

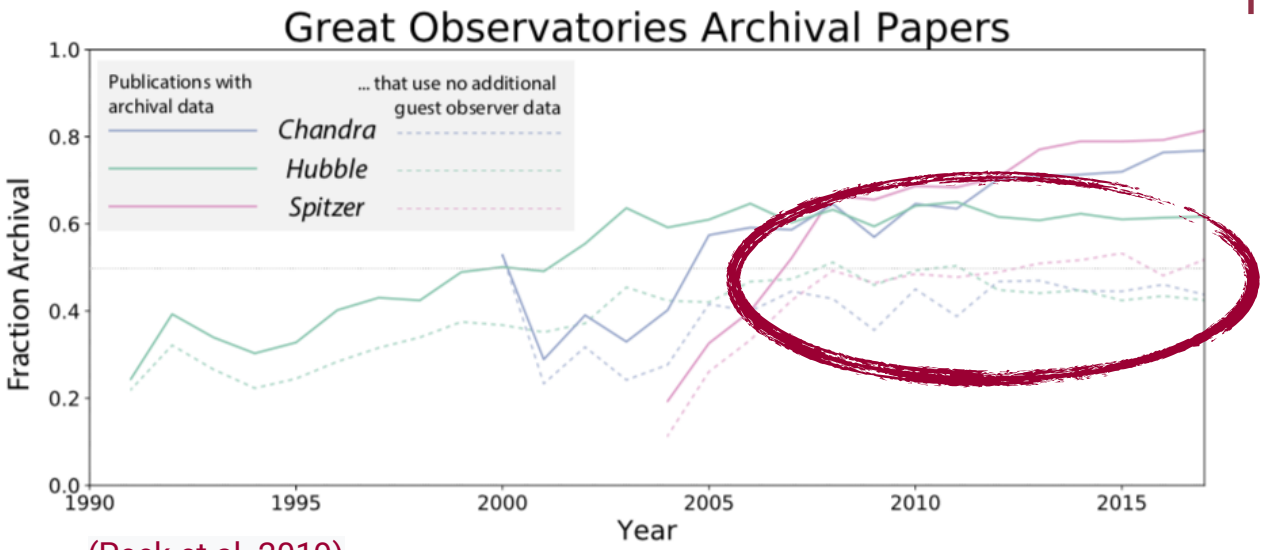
CENTER FOR ASTROPHYSICS
HARVARD & SMITHSONIAN

Chandra archival science



Chandra data are **heavily re-used** over their archival life

Fraction of archival *Chandra* papers plateaued in the last ~ 5 years



(Peek et al. 2019)

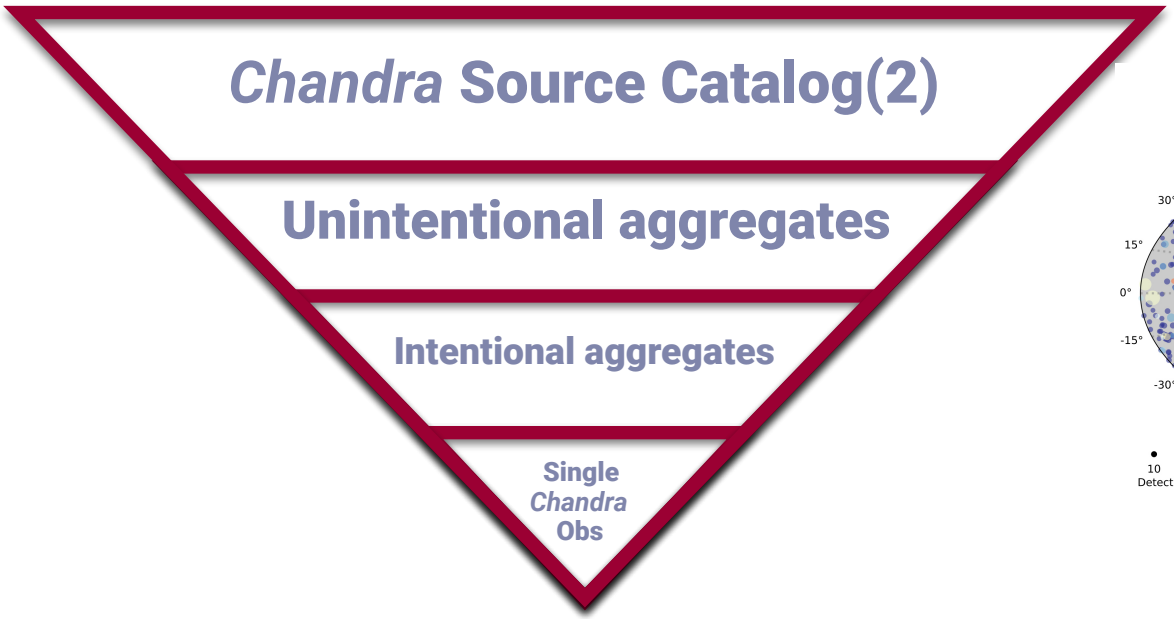
- ~~Science content exhausted?~~
- ~~Better data around?~~
- Untapped/able reservoirs?
- Low hanging fruits all picked?

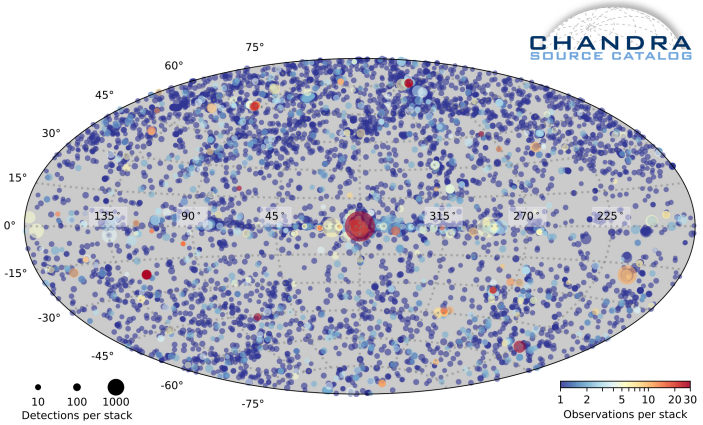
CENTER FOR ASTROPHYSICS
HARVARD & SMITHSONIAN

The power of aggregation

The ***Chandra*** Source Catalog 2 (CSC2) is the culmination of a data aggregation process that has been possible thanks to ***Chandra*** longevity

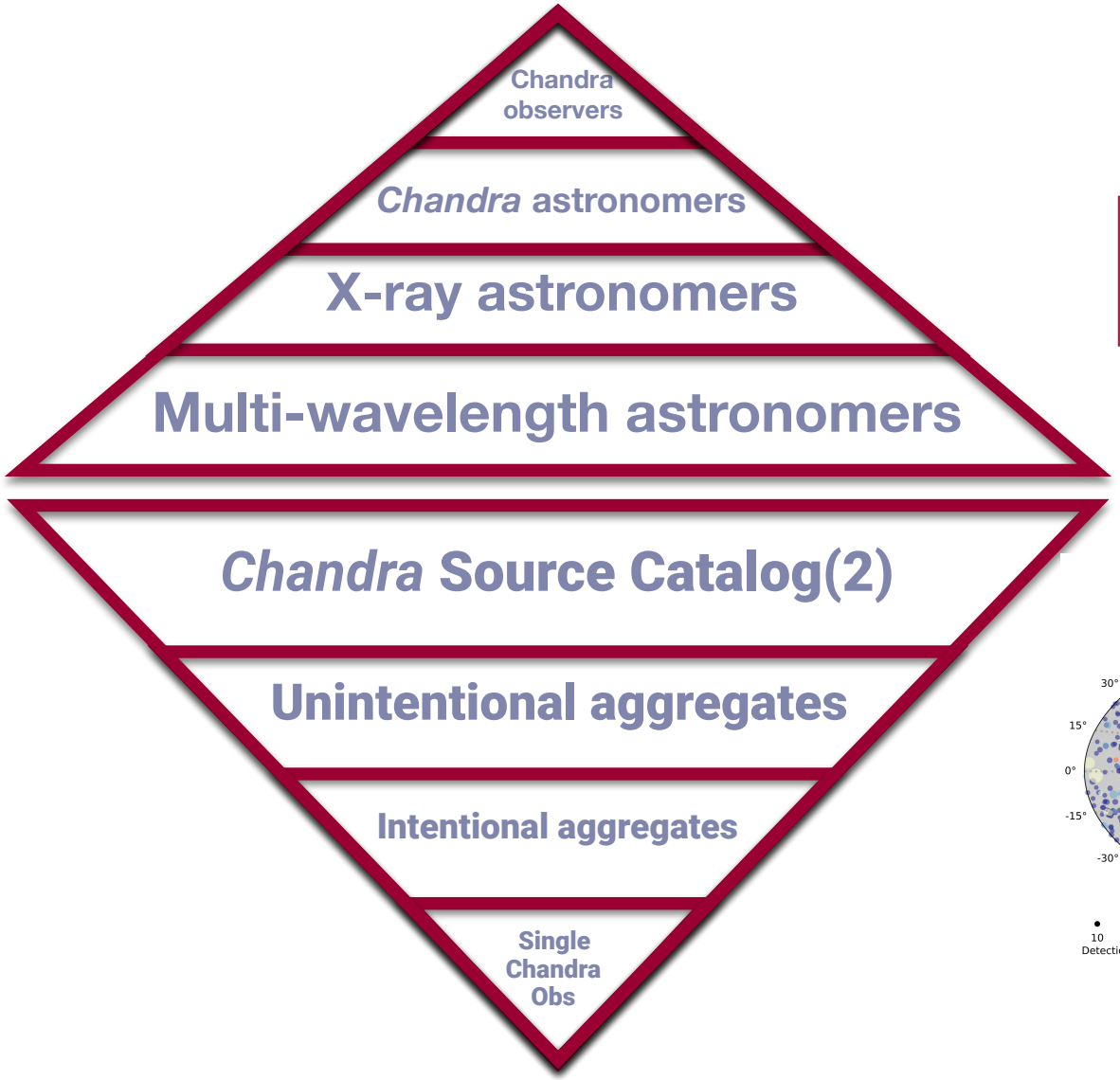
The CSC2 completes the ***Chandra*** archive and its ***vertical expansion***, enhancing the legacy of the mission.





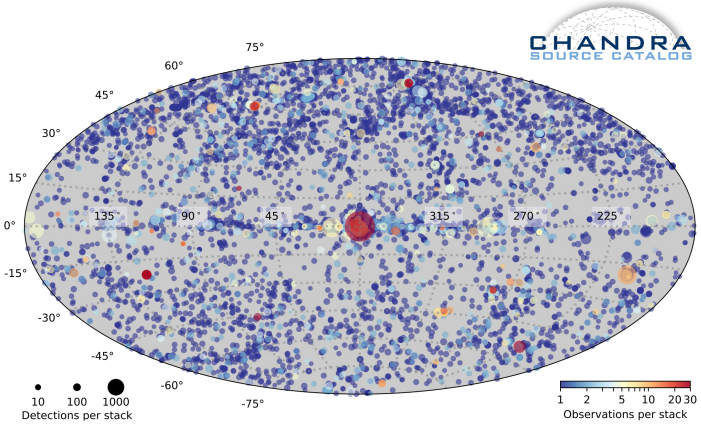
CENTER FOR ASTROPHYSICS
HARVARD & SMITHSONIAN

The power of aggregation



The *Chandra* Source Catalog 2 (CSC2) is the culmination of a data aggregation process that has been possible thanks to *Chandra* longevity

The CSC2 completes the *Chandra* archive and its ***vertical expansion***, enhancing the legacy of the mission.



CENTER FOR ASTROPHYSICS
HARVARD & SMITHSONIAN

Different goals

>>> **High priority projects**

Projects that **need to be completed**

- Core science (“what you do as scientist”)
- Projects with commitments (PhDs, postdocs, funded projects)
- High impact, time sensitive projects

Different goals

>>> High priority projects

Projects that **need to be completed**

- Core science (“what you do as scientist”)
- Projects with commitments (PhDs, postdocs, funded projects)
- High impact, time sensitive projects

>>> Lower priority projects

Projects that **might be interesting**, but are not your core science

- New, risky ideas (failure is a possibility)
- Side projects that would be nice to purse, if promising

Different goals

>>> High priority projects

Projects that **need to be completed**

- Core science (“what you do as scientist”)
- Projects with commitments (PhDs, postdocs, funded projects)
- High impact, time sensitive projects

>>> Lower priority projects

Projects that **might be interesting**, but are not your core science

- New, risky ideas (failure is a possibility)
- Side projects that would be nice to purse, if promising

Resources available are limited

- Stifled by large amount of effort required
- Encouraged when results come promptly

Different goals

Need-driven

>>> High priority projects

Projects that **need to be completed**

- Core science (“what you do as scientist”)
- Projects with commitments (PhDs, postdocs, funded projects)
- High impact, time sensitive projects

>>> Lower priority projects

Projects that **might be interesting**, but are not your core science

- New, risky ideas (failure is a possibility)
- Side projects that would be nice to purse, if promising

Resources available are limited

- Stifled by large amount of effort required
- Encouraged when results come promptly

Mostly archival

Opportunity driven

A question

> What are all the ***Chandra*** ACIS observations longer than 20 ks that contain (extragalactic) globular clusters (GCs), located 30° above or below the Galactic plane, observed in the 2008-2015 period and whose ETG hosts have IFU data?

A question

What are all the ***Chandra*** ACIS observations longer than 20 ks that contain (extragalactic) globular clusters (GCs), located 30° above or below the Galactic plane, observed in the 2008-2015 period and whose ETG hosts have IFU data?

Chandra ACIS observations longer than 20 ks, located at b ≥ 30° or b ≤ 30° degrees, observed in the 2008-2015 period

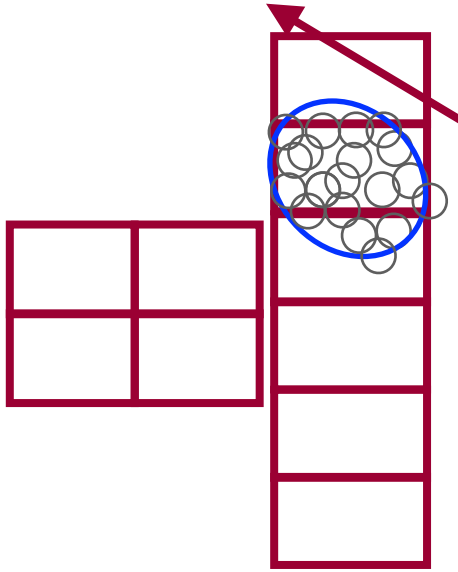
Catalog(s) of GCs observed with HST

List of ETGs observed with IFU instruments

A question

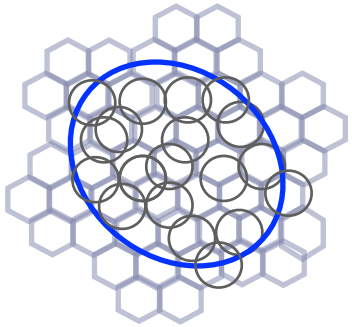
What are all the ***Chandra*** ACIS observations longer than 20 ks that contain (extragalactic) globular clusters (GCs), located 30° above or below the Galactic plane, observed in the 2008-2015 period and whose ETG hosts have IFU data?

Chandra ACIS observations longer than 20 ks, located at b ≥ 30° or b ≤ 30° degrees, observed in the 2008-2015 period



Catalog(s) of GCs observed with HST

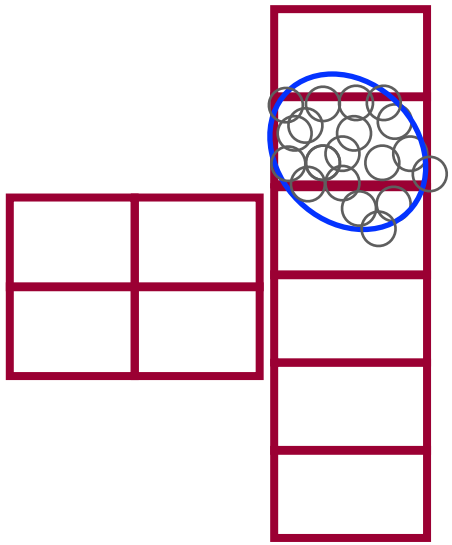
List of ETGs observed with IFU instruments



A question

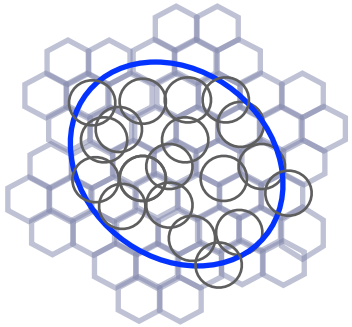
What are all the *Chandra* ACIS observations longer than 20 ks that contain (extragalactic) globular clusters (GCs), located 30° above or below the Galactic plane, observed in the 2008-2015 period and whose ETG hosts have IFU data?

Chandra ACIS observations longer than 20 ks, located at b ≥ 30° or b ≤ 30° degrees, observed in the 2008-2015 period



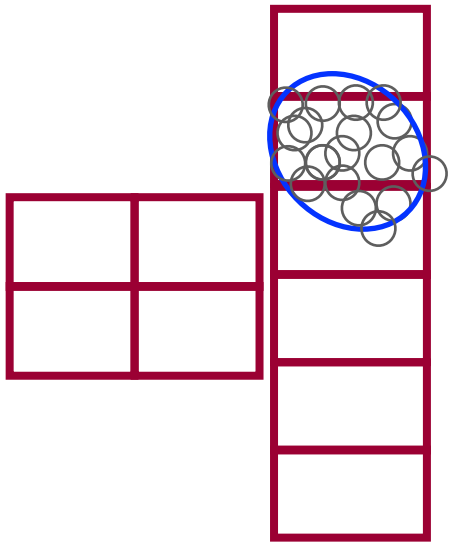
Catalog(s) of GCs observed with HST

List of ETGs observed with IFU instruments

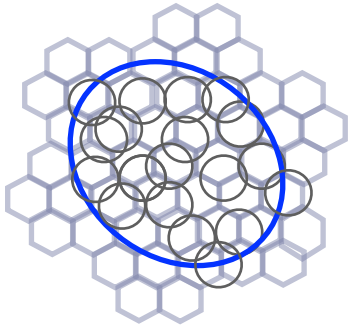


A question

Chandra ACIS observations longer than 20 ks, located at b ≥ 30° or b ≤ 30° degrees, observed in the 2008-2015 period



List of ETGs observed with IFU instruments



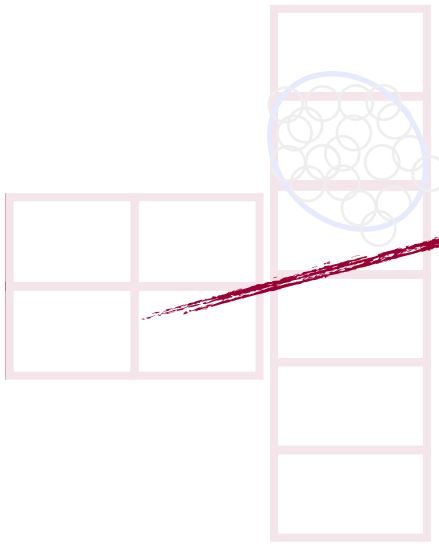
Catalog(s) of GCs observed with HST

A question

Chandra ACIS observations longer than 20 ks, located at b ≥ 30° or b ≤ 30° degrees, observed in the 2008-2015 period

List of ETGs observed with IFU instruments

Catalog(s) of GCs observed with HST



Step 1

What?

- List of HST-detected GCs catalogs
- List of galaxies with IFS observations

Where?

- Single missions/instruments archives
- VO registry, Vizier

How?

- Crossmatching

A question

Chandra ACIS observations longer than 20 ks, located at b ≥ 30° or b ≤ 30° degrees, observed in the 2008-2015 period

List of ETGs observed with IFU instruments

Catalog(s) of GCs observed with HST

Step 1

What?

- List of HST-detected GCs catalogs
- List of galaxies with IFS observations

HST instrument/filter/exposure

IFS: spectral/spatial resolution

→

Where?

- Single missions/instruments archives
- VO registry, Vizier

How?

- Crossmatching

A question

Chandra ACIS observations longer than 20 ks, located at b ≥ 30° or b ≤ 30° degrees, observed in the 2008-2015 period

List of ETGs observed with IFU instruments

Catalog(s) of GCs observed with HST

Step 1

HST instrument/filter/exposure

IFS: spectral/spatial resolution

→

What?

- List of HST-detected GCs catalogs
- List of galaxies with IFS observations

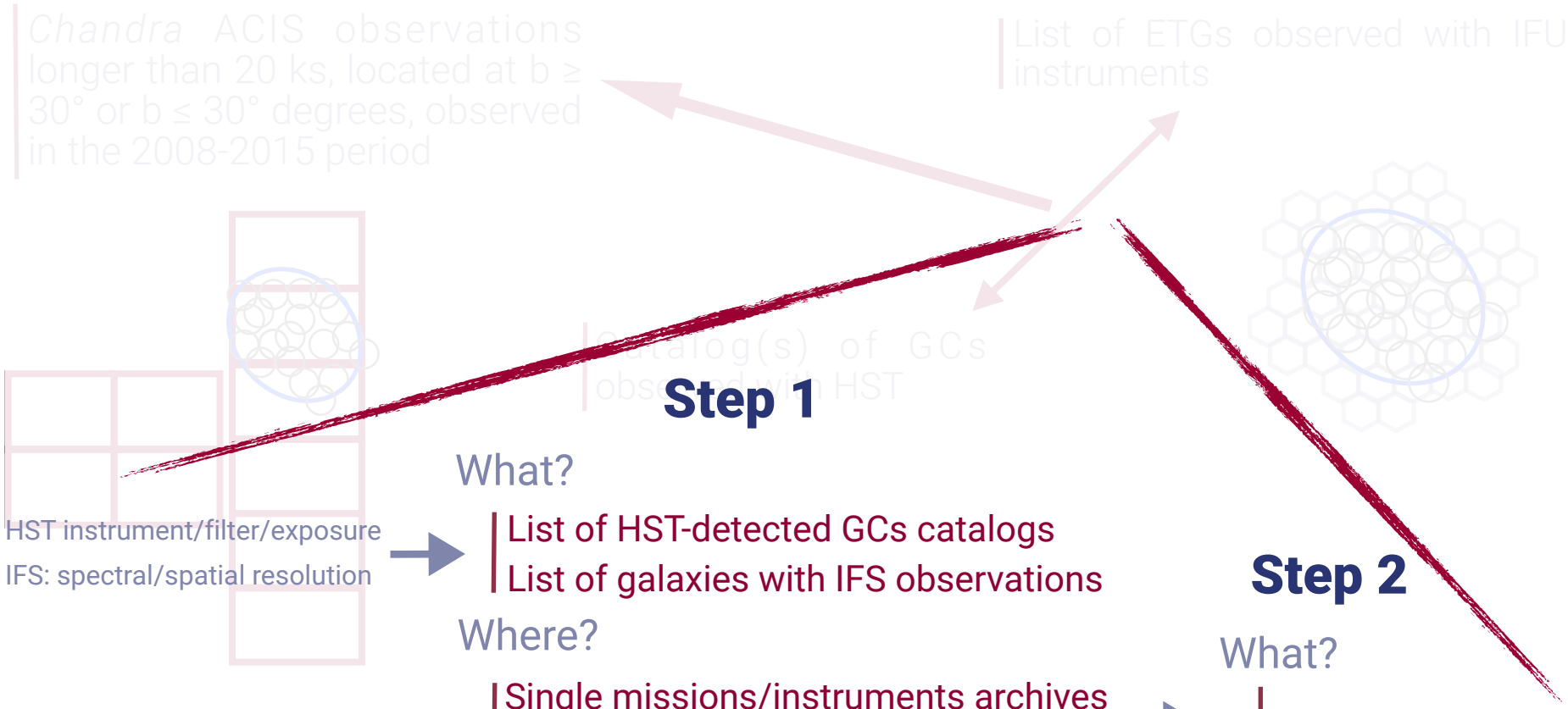
Where?

- Single missions/instruments archives
- VO registry, Vizier

How?

- Crossmatching
  - Crossmatch can be complicated
  - Crossmatch can be very complicated

A question


Chandra ACIS observations longer than 20 ks, located at b ≥ 30° or b ≤ 30° degrees, observed in the 2008-2015 period
List of ETGs observed with IFU instruments
Catalog(s) of GCs observed with HST
Step 1
What?
List of HST-detected GCs catalogs
List of galaxies with IFS observations
Where?
Single missions/instruments archives
VO registry, Vizier
How?
Crossmatching
Crossmatch can be complicated
Crossmatch can be very complicated
HST instrument/filter/exposure
IFS: spectral/spatial resolution
Step 2
What?
.....
Where?
.....
How?
.....

The real question

>>> **Where does information reside?**

Everything in its own place

- GCs, galaxies from literature, contributed datasets
- Instrumental parameters from mission/instrument archives or VO

Building pipes to transfer data from place to place

- Possible, but time consuming and laborious
- Hindering data exploration/collection

A solution

>>> **Thinking of sources as attributes of observations**

All that lives in a region of sky covered by ***Chandra*** observations, pertains to those observations

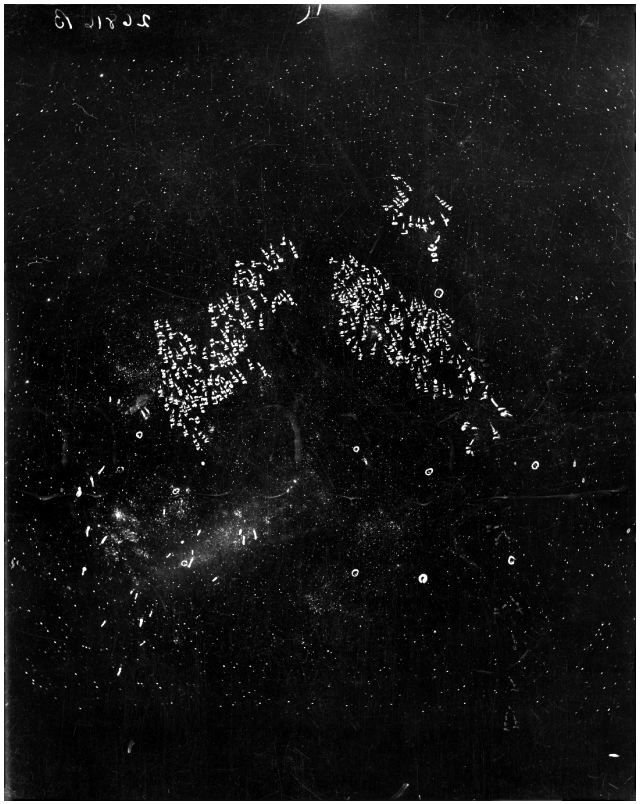
- Annotating ***Chandra*** with external knowledge
- Focusing on sources and their multi-wavelength properties
- X-ray based annotation (***CSC*** sources/properties) as further step

In general

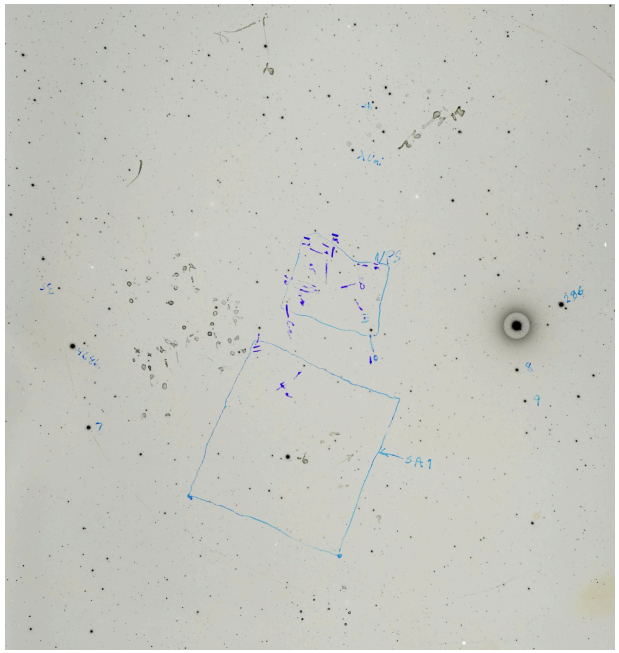
Data fusion, aka data integration in some fields, is the process of integrating multiple data sources to produce more consistent, accurate, and/or useful information than that provided by any individual data source.

In general

Data fusion, aka data integration in some fields, is the process of integrating multiple data sources to produce more consistent, accurate, and/or useful information than that provided by any individual data source.



(LMC, 10/18/1900, used by H. Leavitt)
(Digital Access to a Sky Century @ Harvard)



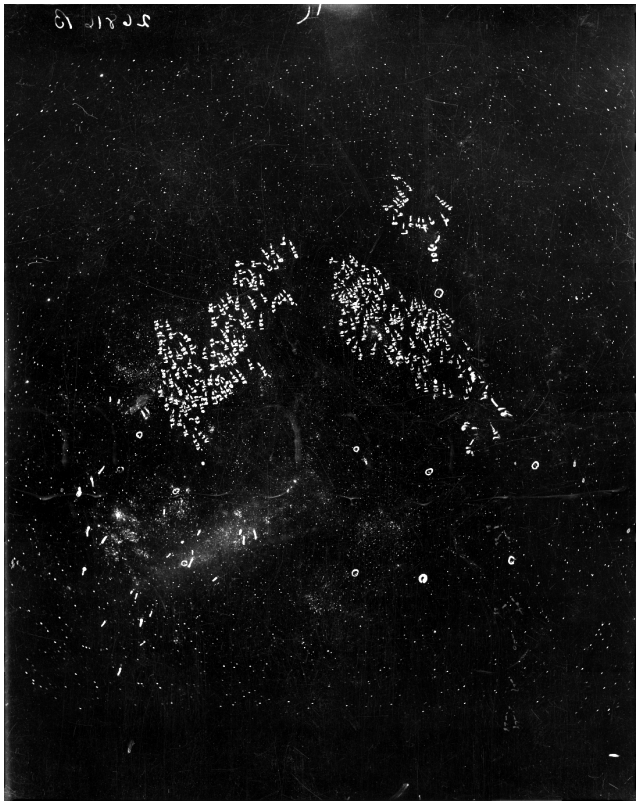
(Project PHaEDRA, Wolbach Library - Harvard)

Old tradition

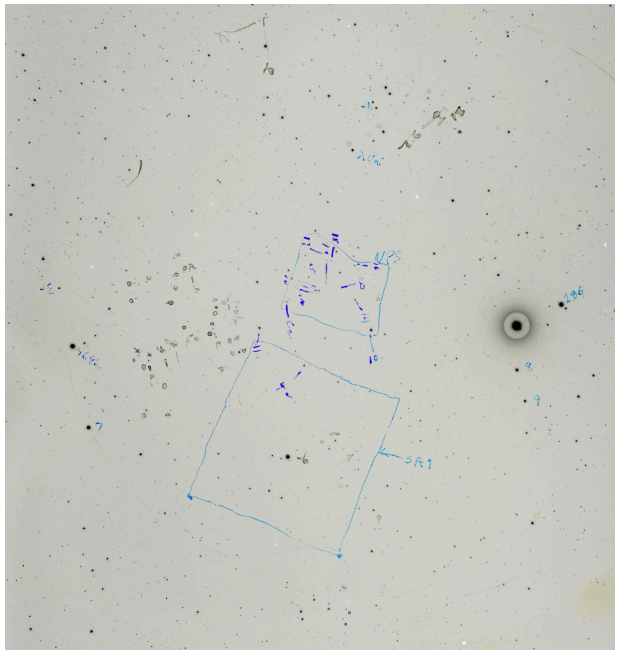
- Manual annotation of plates
- Instrumental for discovery
- Exhausting activity
- Endogenous annotation

In general

Data fusion, aka data integration in some fields, is the process of integrating multiple data sources to produce more consistent, accurate, and/or useful information than that provided by any individual data source.



(LMC, 10/18/1900, used by H. Leavitt)
(Digital Access to a Sky Century @ Harvard)



(Project PHaEDRA, Wolbach Library - Harvard)

Old tradition

- Manual annotation of plates
- Instrumental for discovery
- Exhausting activity
- Endogenous annotation

New ways

- Digital nature of all data
- Technology revolution
- Collaborative environment
- Multiwavelength revolution

Annotating the *Chandra* archive

>>> Spatial criterion

Collection of source properties within the field of view (fov) of all archival ***Chandra*** observations

- Annotation of single fov's independent of overlap
- Annotations stored in a table of ***Chandra*** observational DB
- Many-to-many indexed relationships between sources and observations
- Regular updates to the annotations

Annotating the *Chandra* archive

>>> Spatial criterion

Collection of source properties within the field of view (fov) of all archival ***Chandra*** observations

- Annotation of single fov's independent of overlap
- Annotations stored in a table of ***Chandra*** observational DB
- Many-to-many indexed relationships between sources and observations
- Regular updates to the annotations

>>> Annotations' provider

CDS Astronomical Database (**SIMBAD**)

- Queried by spatial search (polygon search)
- Leveraging CDS object classification for "class searches"
- Non-invasive queries (so far)

Annotating the *Chandra* archive

>>> Spatial criterion

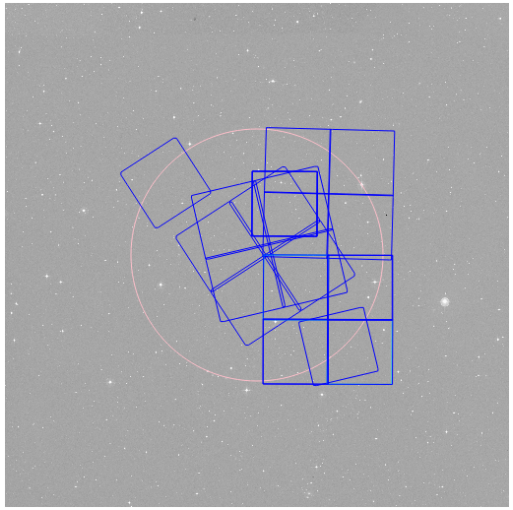
Collection of source properties within the field of view (fov) of all archival ***Chandra*** observations

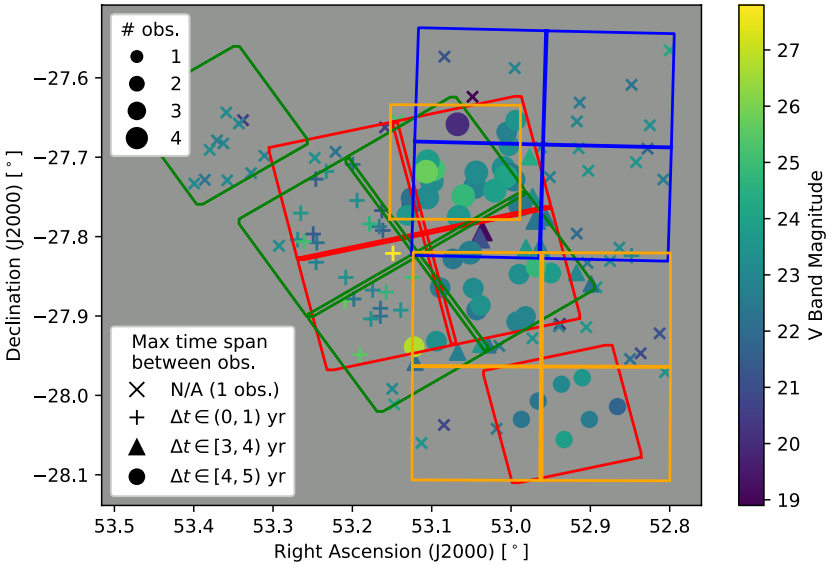
- Annotation of single fov's independent of overlap
- Annotations stored in a table of ***Chandra*** observational DB
- Many-to-many indexed relationships between sources and observations
- Regular updates to the annotations

Testing, development and plans

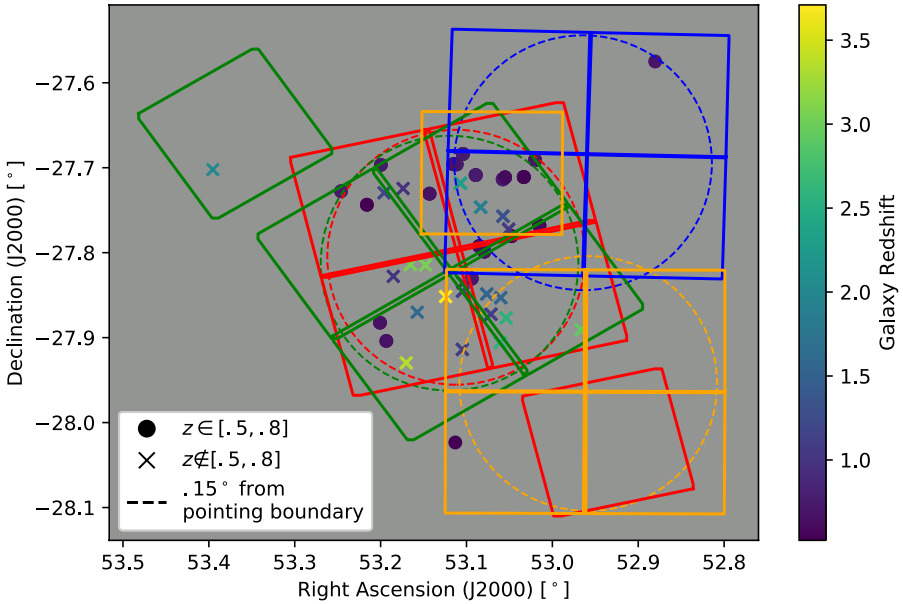
- Validating use cases collected for requirements
- Measuring performances on test cases
- Incorporating endogenous annotations (CSC)
- Annotating aggregated ***Chandra*** data (CSC stacks/contributed datasets)

Annotating the *Chandra* archive

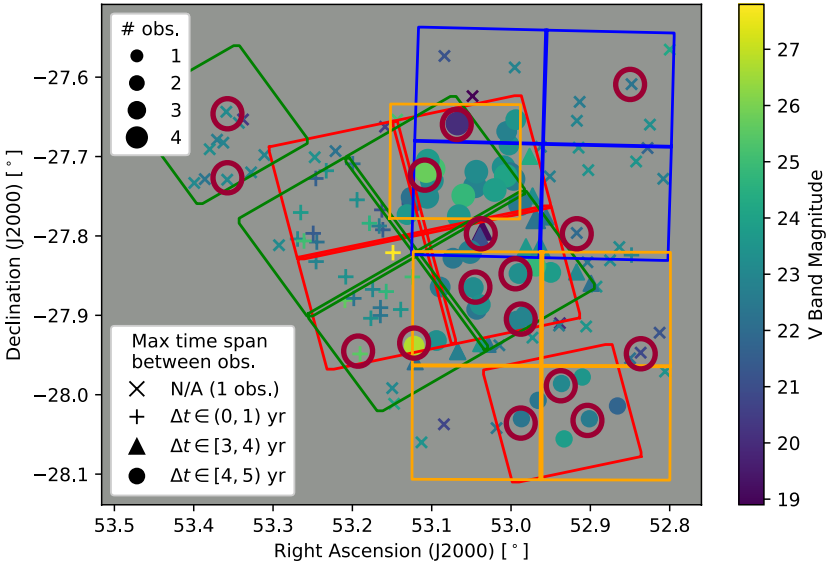




Seyfert2 galaxies with redshift in [.5, .8] within 9’ of the aimpoint



Optical quasars as a function of max Δt between ***Chandra*** observations (↑), and with additional CSC2 annotations (↓)



Annotating the *Chandra* archive

>>> Spatial criterion

Collection of source properties within the field of view (fov) of all archival ***Chandra*** observations

- Annotation of single fov's independent of overlap
- Annotations stored in a table of ***Chandra*** observational DB
- Many-to-many indexed relationships between sources and observations
- Regular updates to the annotations

Testing, development and plans

- Validating use cases collected for requirements
- Measuring performances on test cases
- Incorporating endogenous annotations (CSC)
- Annotating aggregated ***Chandra*** data (CSC stacks/contributed datasets)

Programmatic interface only

- Graphical/interactive solution should leverage existing interfaces
- TAP?

Caveats

Annotation of ***Chandra*** archival observations provides a minimal friction tool to ask fairly complex questions and tackles the problem of the "initial effort barrier", but it does not address many aspects of data discovery and analysis.

Caveats

Annotation of ***Chandra*** archival observations provides a minimal friction tool to ask fairly complex questions and tackles the problem of the "initial effort barrier", but it does not address many aspects of data discovery and analysis.

>> Quick, coarse results

Crossmatches still needed at some point of the analysis
Extended sources located close to chip margins or chip gaps
We are annotating "polygons" and basic set of metadata

Our data structure doesn't know about quantities that are more complex and/or functions of positions - exposure, PSF size & shape, etc

Caveats

Annotation of ***Chandra*** archival observations provides a minimal friction tool to ask fairly complex questions and tackles the problem of the "initial effort barrier", but it does not address many aspects of data discovery and analysis.

>> Quick, coarse results

- Crossmatches still needed at some point of the analysis
- Extended sources located close to chip margins or chip gaps
- We are annotating "polygons" and basic set of metadata
  - Our data structure doesn't know about quantities that are more complex and/or functions of positions - exposure, PSF size & shape, etc

>> What's in a name?

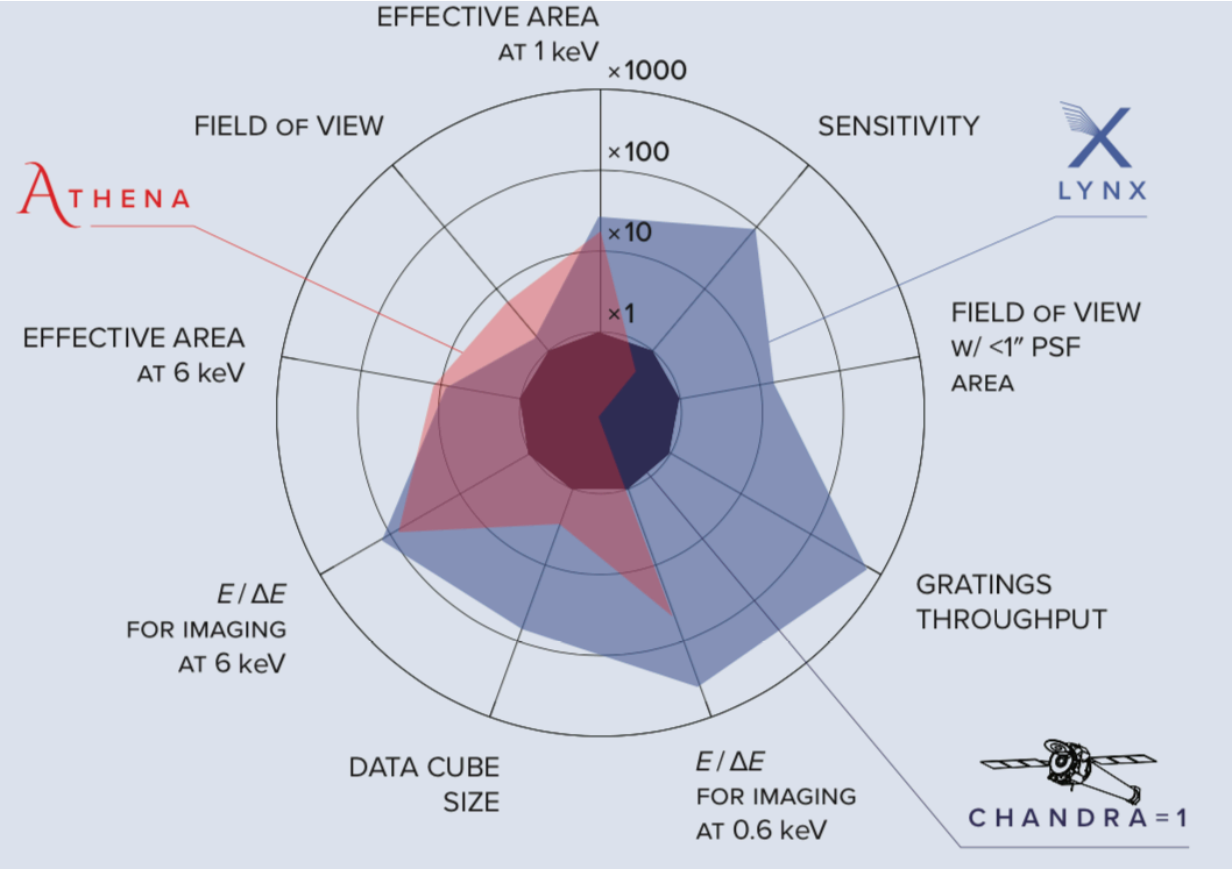
- An AGNs by any other name, wouldn't shine as bright
  - Defining classes of astronomical sources is difficult

How this helps

Archives of powerful mission like *Chandra* (and most space facilities) can provide **context** and "**baseline knowledge**" for old events or new multi-messenger signals that will trigger EM follow-ups.

How this will help

Archives of powerful mission like *Chandra* (and most space facilities) can provide **context** and "**baseline knowledge**" for old events or new multi-messenger signals that will trigger EM follow-ups.



(Lynx mission concept study report)

New missions!

- Larger fov
- Way larger EAs
- Increased sensitivity
- Better PSF behavior

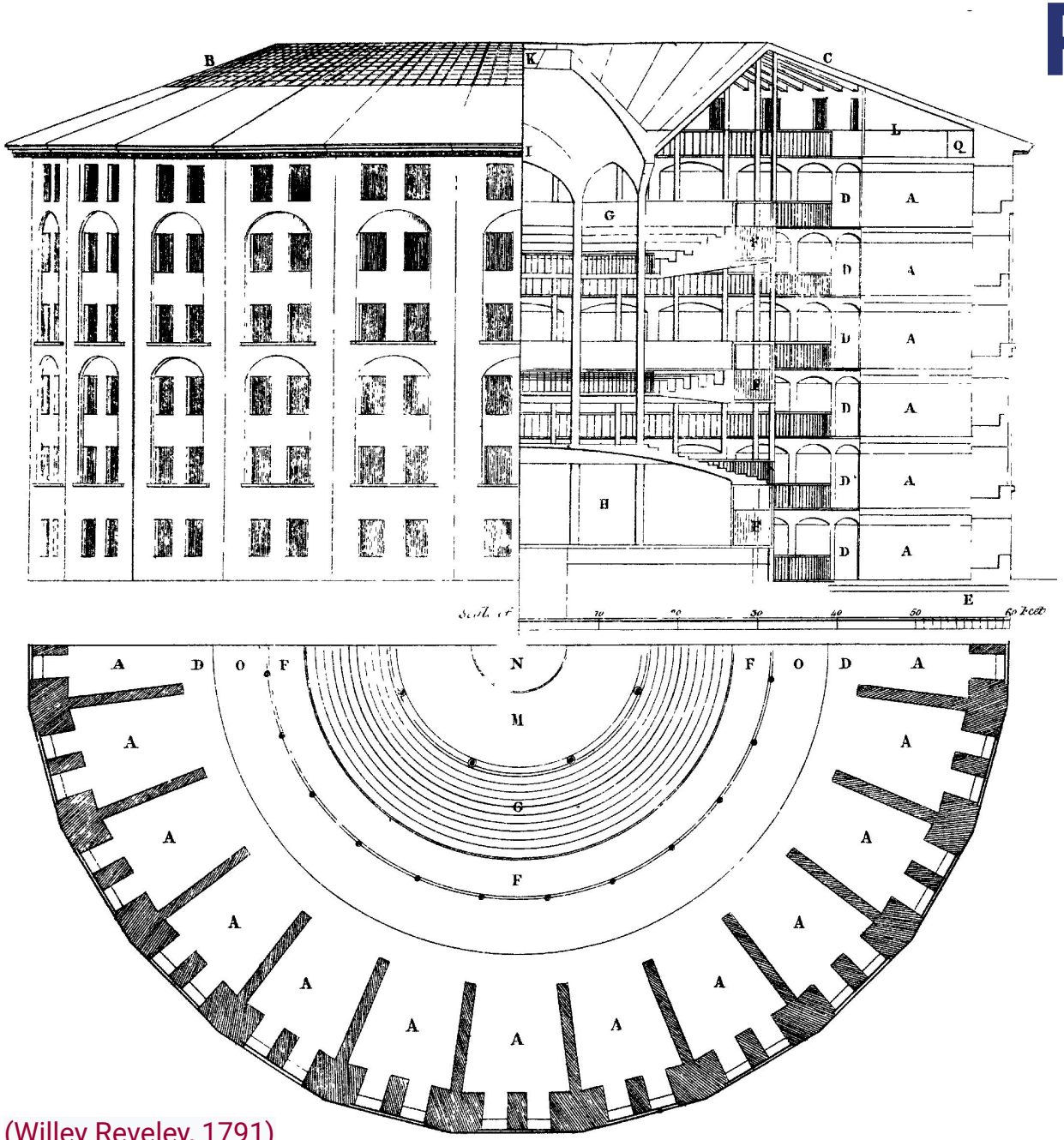
Annotations will become increasingly relevant with the next generation of X-ray missions

Panopticon

> “The Building circular – an iron cage, glazed – a glass lantern about the size of Ranelagh – The Prisoners in their Cells, occupying the Circumference – The Officers, the Centre. By Blinds, and other contrivances, the Inspectors concealed from the observation of the Prisoners: hence the sentiment of a sort of invisible omnipresence. – The whole circuit reviewable with little, or, if necessary, without any, change of place.”

Jeremy Bentham (1791). Panopticon, or The Inspection House

Panopticon



(Willey Reveley, 1791)

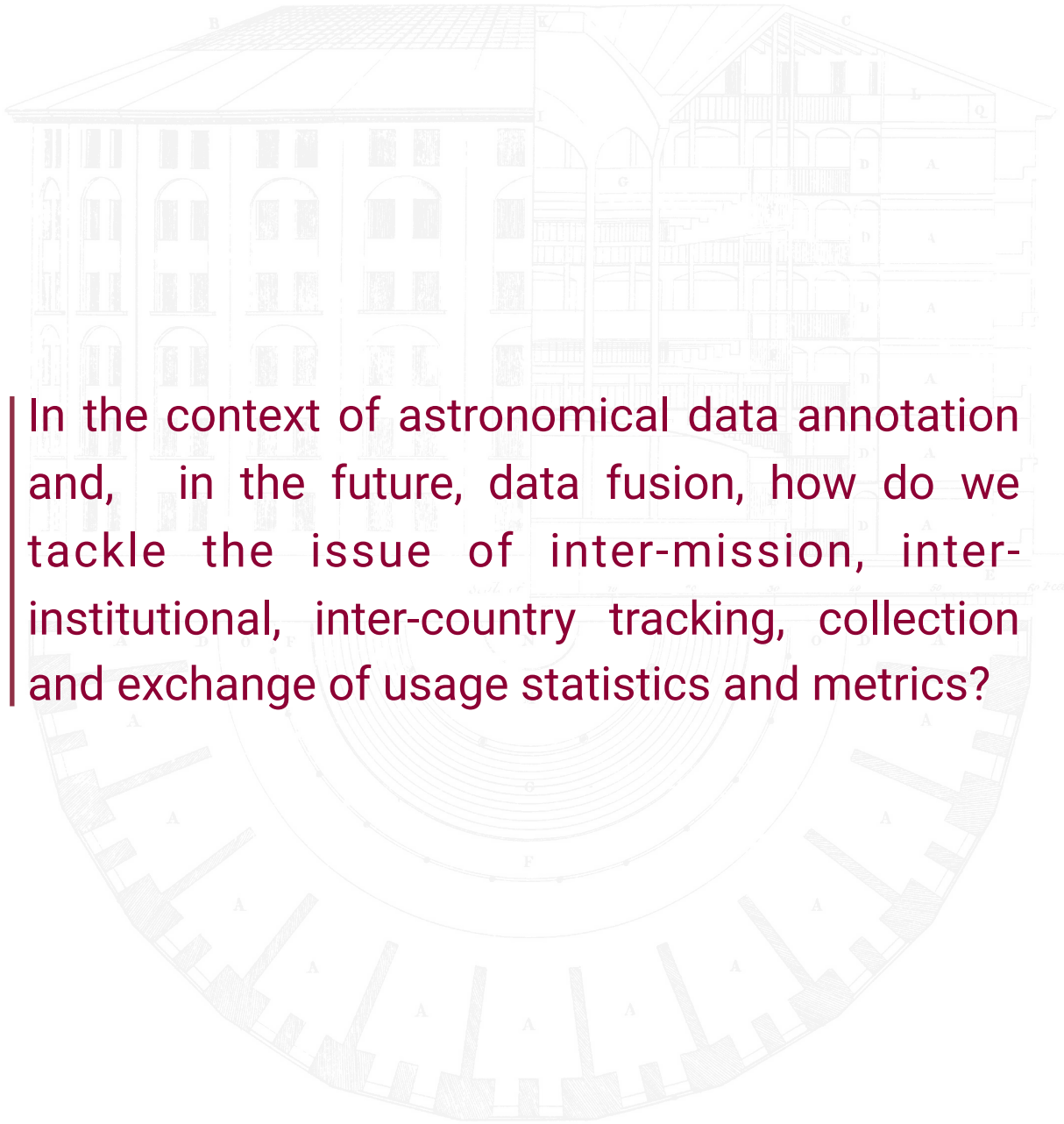
Guards are astronomers in their control tower, single “data cells” represent data providers/datasets/ archives.

All line-of-sights are free and available to the “guards” at all times (data fusion/annotation/links).

Unlike the original design of the Panopticon where inmates were not aware of guards, "data cells” should know when they are being watched.

CENTER FOR ASTROPHYSICS
HARVARD & SMITHSONIAN

Panopticon



In the context of astronomical data annotation and, in the future, data fusion, how do we tackle the issue of inter-mission, inter-institutional, inter-country tracking, collection and exchange of usage statistics and metrics?

Guards are astronomers in their control tower, single “data cells” represent data providers/datasets/ archives.

All line-of-sights are free and available to the “guards” at all times (data fusion/annotation/links).

Unlike the original design of the Panopticon where inmates were not aware of guards, "data cells” should know when they are being watched.