The Emergence of
Computational Archival Science (CAS)

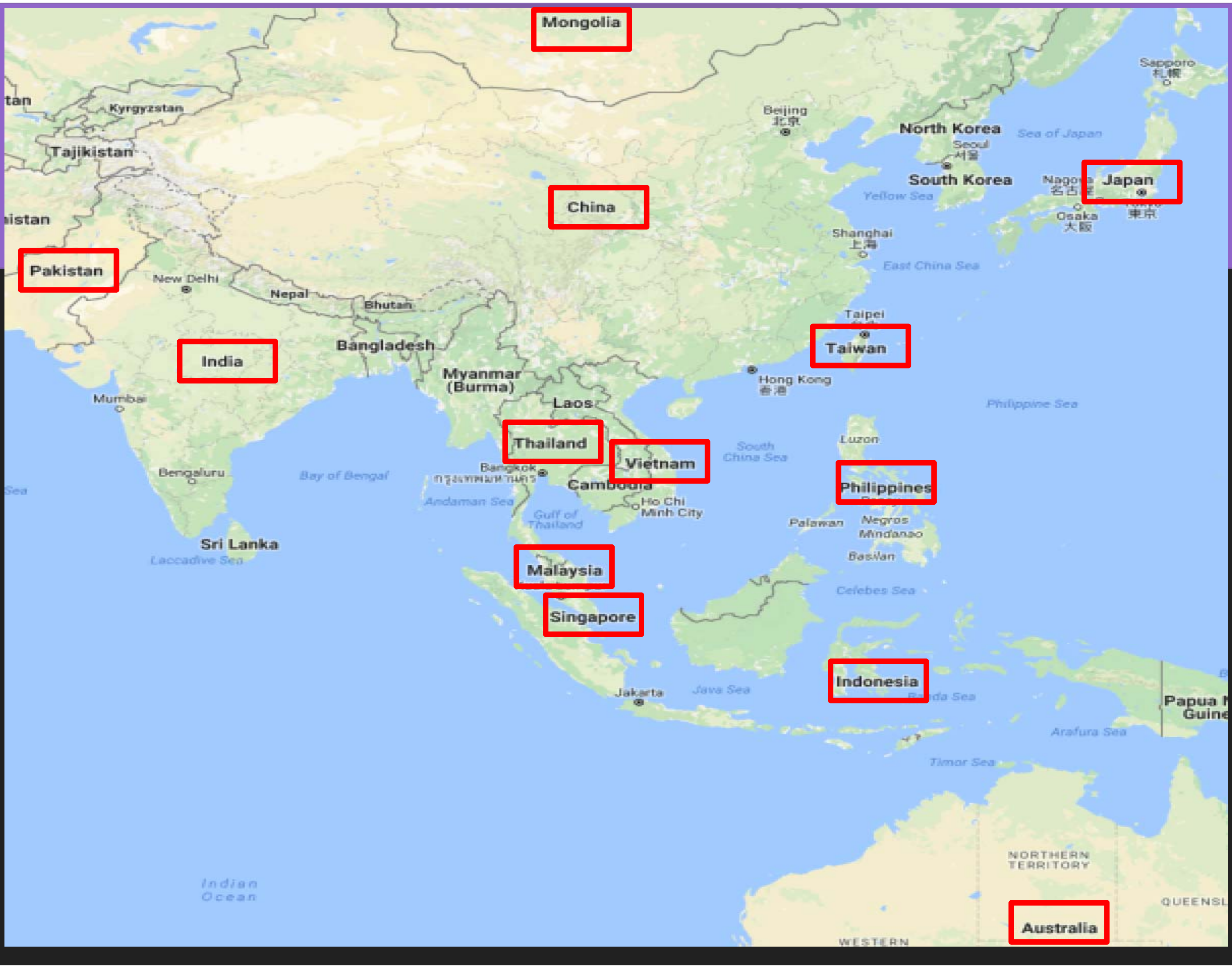
Richard **MARCIANO**

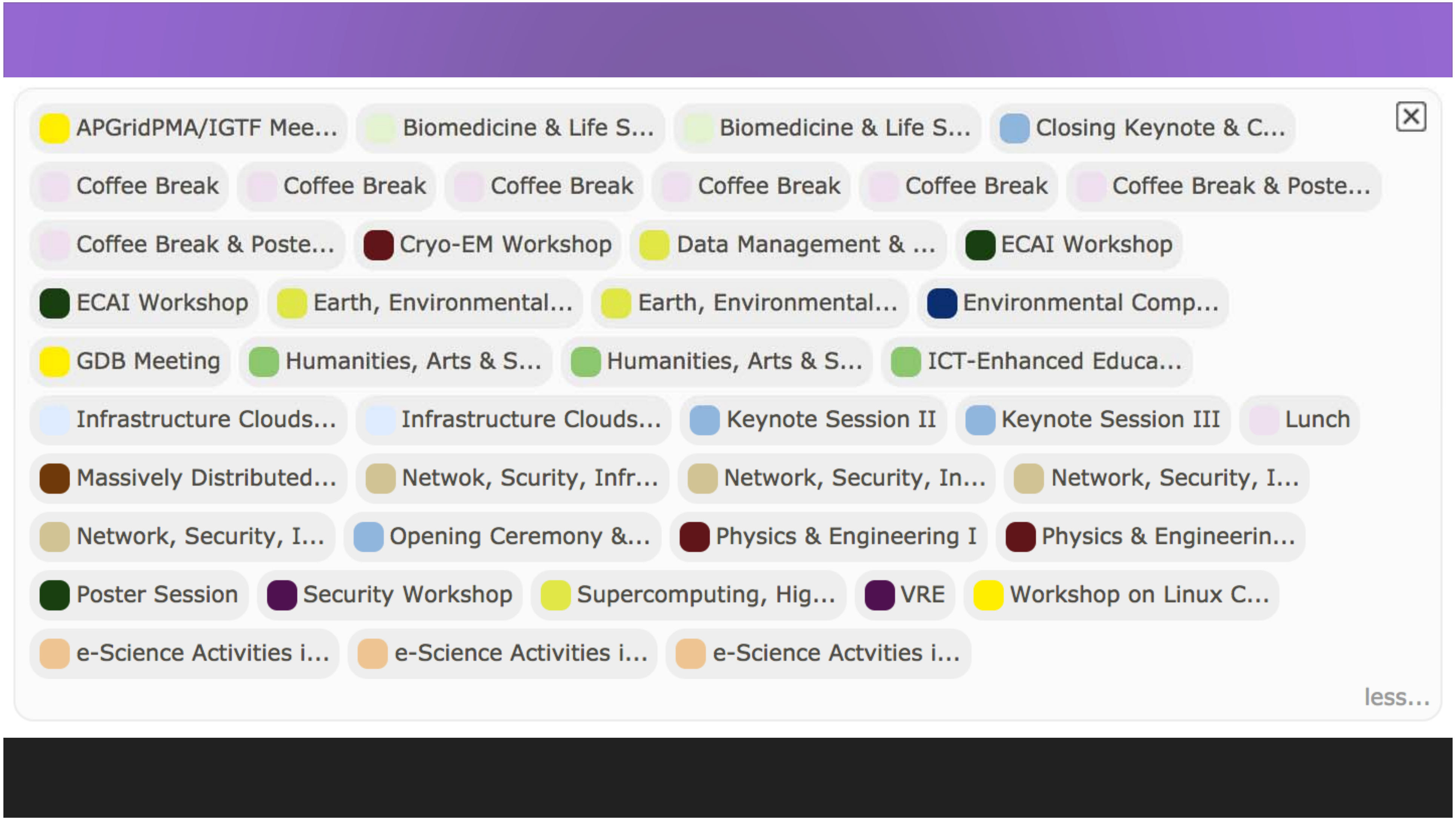
marciano@umd.edu
University of Maryland iSchool

UNIVERSITY OF MARYLAND

COLLEGE OF INFORMATION STUDIES *iSchool*

dcic digital curation innovation center

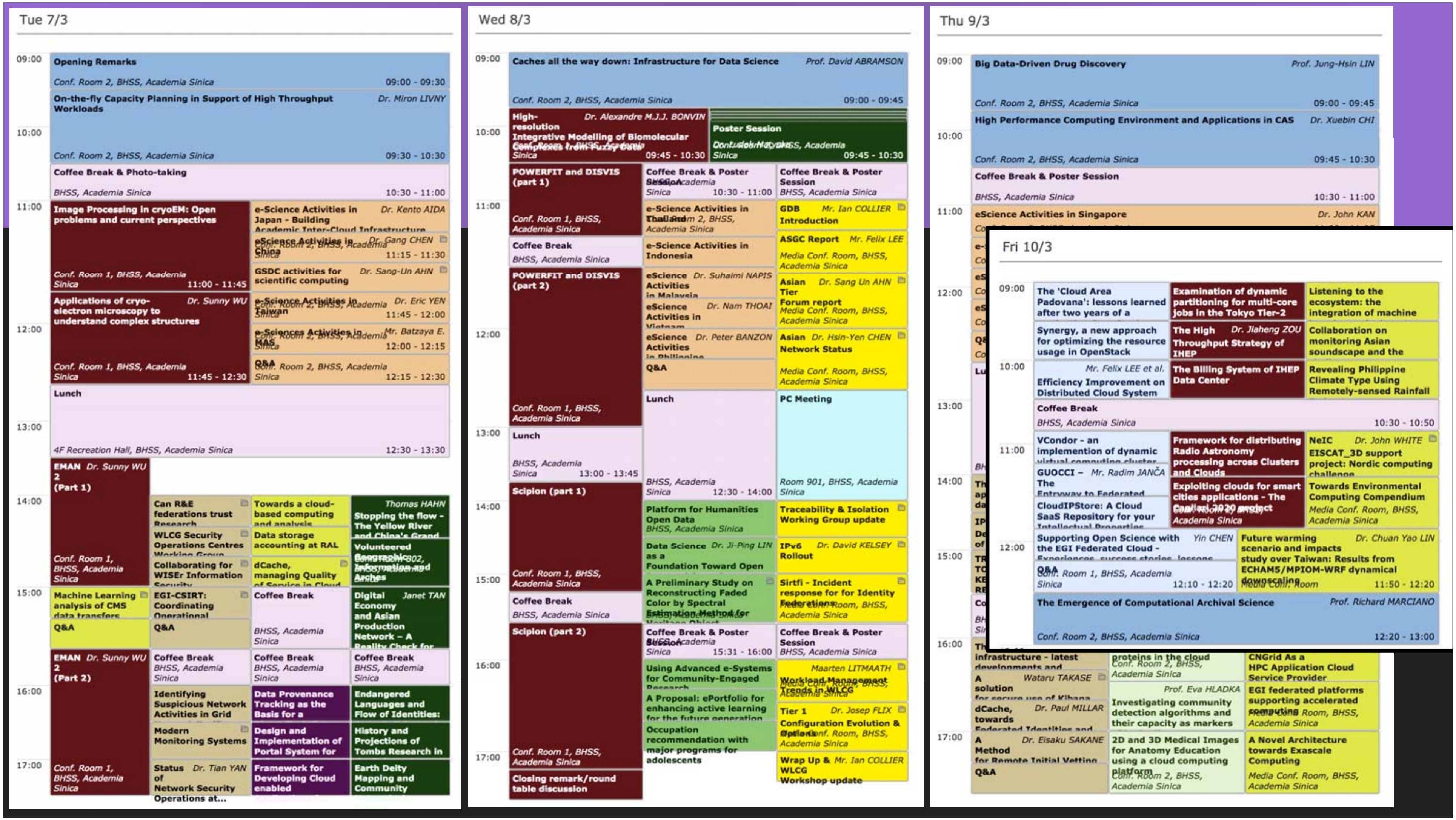

Mongolia
tan
Kyrgyzstan
Tajikistan
Beijing
北京
North Korea
Sea of Japan
Sapporo
札幌
Seoul
서울
South Korea
Nagoya
名古屋
Japan
東京
Osaka
大阪
Yellow Sea
China
istan
Shanghai
上海
East China Sea
Pakistan
New Delhi
Nepal
Bhutan
Taipei
Taiwan
India
Bangladesh
Myanmar
(Burma)
Hong Kong
香港
Laos
Philippine Sea
Mumbai
Thailand
Vietnam
South
China Sea
Luzon
Bengaluru
Bay of Bengal
Bangkok
กรุงเทพมหานคร
Cambodia
Philippines
Sea
Andaman Sea
Ho Chi
Minh City
Gulf of
Thailand
Palawan
Negros
Mindanao
Sri Lanka
Laccadive Sea
Basilan
Malaysia
Celebes Sea
Singapore
Indonesia
Jakarta
Java Sea
Banda Sea
Papua N
Guine
Arafura Sea
Timor Sea
Indian
Ocean
NORTHERN
TERRITORY
QUEENSL
Australia
WESTERN



☒

🟨 APGridPMA/IGTF Mee...
🟩 Biomedicine & Life S...
🟩 Biomedicine & Life S...
🟦 Closing Keynote & C...
⬜ Coffee Break
⬜ Coffee Break
⬜ Coffee Break
⬜ Coffee Break
⬜ Coffee Break
⬜ Coffee Break & Poste...
⬜ Coffee Break & Poste...
🟥 Cryo-EM Workshop
🟨 Data Management & ...
🟩 ECAI Workshop
🟩 ECAI Workshop
🟨 Earth, Environmental...
🟨 Earth, Environmental...
🟦 Environmental Comp...
🟨 GDB Meeting
🟩 Humanities, Arts & S...
🟩 Humanities, Arts & S...
🟩 ICT-Enhanced Educa...
⬜ Infrastructure Clouds...
⬜ Infrastructure Clouds...
🟦 Keynote Session II
🟦 Keynote Session III
⬜ Lunch
🟫 Massively Distributed...
🟫 Netwok, Scurity, Infr...
🟫 Network, Security, In...
🟫 Network, Security, I...
🟫 Network, Security, I...
🟦 Opening Ceremony &...
🟥 Physics & Engineering I
🟥 Physics & Engineerin...
🟩 Poster Session
🟪 Security Workshop
🟨 Supercomputing, Hig...
🟪 VRE
🟨 Workshop on Linux C...
🟧 e-Science Activities i...
🟧 e-Science Activities i...
🟧 e-Science Actvities i...

less...



Tue 7/3

09:00
Opening Remarks
Conf. Room 2, BHSS, Academia Sinica 09:00 - 09:30

On-the-fly Capacity Planning in Support of High Throughput Workloads — Dr. Miron LIVNY
Conf. Room 2, BHSS, Academia Sinica 09:30 - 10:30

10:00

Coffee Break & Photo-taking
BHSS, Academia Sinica 10:30 - 11:00

11:00
Image Processing in cryoEM: Open problems and current perspectives
Conf. Room 1, BHSS, Academia Sinica 11:00 - 11:45

e-Science Activities in Japan - Building Academic Inter-Cloud Infrastructure — Dr. Kento AIDA

eScience Activities in China — Dr. Gang CHEN
Conf. Room 2, BHSS, Academia Sinica 11:15 - 11:30

GSDC activities for scientific computing — Dr. Sang-Un AHN

Applications of cryo-electron microscopy to understand complex structures — Dr. Sunny WU
Conf. Room 1, BHSS, Academia Sinica 11:45 - 12:30

e-Science Activities in Taiwan — Dr. Eric YEN
Conf. Room 2, BHSS, Academia Sinica 11:45 - 12:00

12:00

e-Sciences Activities in MAS — Mr. Batzaya E.
Conf. Room 2, BHSS, Academia Sinica 12:00 - 12:15

Q&A
Conf. Room 2, BHSS, Academia Sinica 12:15 - 12:30

Lunch
4F Recreation Hall, BHSS, Academia Sinica 12:30 - 13:30

13:00

EMAN 2 (Part 1) — Dr. Sunny WU
Conf. Room 1, BHSS, Academia Sinica

14:00
Can R&E federations trust Research

Towards a cloud-based computing and analysis

Stopping the flow - The Yellow River and China's Grand — Thomas HAHN

WLCG Security Operations Centres Working Group

Data storage accounting at RAL

Volunteered Geographic Information and Arches
Conf. Room 802, BHSS, Academia Sinica

Collaborating for WISEr Information Security

dCache, managing Quality of Service in Cloud

15:00
Machine Learning analysis of CMS data transfers

EGI-CSIRT: Coordinating Operational

Coffee Break
BHSS, Academia Sinica

Digital Economy and Asian Production Network – A Reality Check for — Janet TAN

Q&A

Q&A

EMAN 2 (Part 2) — Dr. Sunny WU
Conf. Room 1, BHSS, Academia Sinica

Coffee Break
BHSS, Academia Sinica

Coffee Break
BHSS, Academia Sinica

Coffee Break
BHSS, Academia Sinica

16:00
Identifying Suspicious Network Activities in Grid

Data Provenance Tracking as the Basis for a

Endangered Languages and Flow of Identities:

Modern Monitoring Systems

Design and Implementation of Portal System for

History and Projections of Tombs Research in

17:00
Status of Network Security Operations at... — Dr. Tian YAN

Framework for Developing Cloud enabled

Earth Deity Mapping and Community

Wed 8/3

09:00
Caches all the way down: Infrastructure for Data Science — Prof. David ABRAMSON
Conf. Room 2, BHSS, Academia Sinica 09:00 - 09:45

10:00
High-resolution Integrative Modelling of Biomolecular Complexes from Fuzzy Data — Dr. Alexandre M.J.J. BONVIN
Conf. Room 1, BHSS, Academia Sinica 09:45 - 10:30

Poster Session — Dr. Ludek Matyska
Conf. Room 1, BHSS, Academia Sinica 09:45 - 10:30

POWERFIT and DISVIS (part 1)
Conf. Room 1, BHSS, Academia Sinica

Coffee Break & Poster Session
BHSS, Academia Sinica 10:30 - 11:00

Coffee Break & Poster Session
BHSS, Academia Sinica

11:00
e-Science Activities in Thailand
Conf. Room 2, BHSS, Academia Sinica

GDB Introduction — Mr. Ian COLLIER

Coffee Break
BHSS, Academia Sinica

e-Science Activities in Indonesia

ASGC Report — Mr. Felix LEE
Media Conf. Room, BHSS, Academia Sinica

POWERFIT and DISVIS (part 2)
Conf. Room 1, BHSS, Academia Sinica

eScience Activities in Malaysia — Dr. Suhaimi NAPIS

Asian Tier Forum report — Dr. Sang Un AHN
Media Conf. Room, BHSS, Academia Sinica

eScience Activities in Vietnam — Dr. Nam THOAI

12:00
eScience Activities in Philippine — Dr. Peter BANZON

Asian Network Status — Dr. Hsin-Yen CHEN
Media Conf. Room, BHSS, Academia Sinica

Q&A

Lunch
BHSS, Academia Sinica 12:30 - 14:00

PC Meeting
Room 901, BHSS, Academia Sinica

13:00
Lunch
BHSS, Academia Sinica 13:00 - 13:45

Scipion (part 1)
Conf. Room 1, BHSS, Academia Sinica

14:00
Platform for Humanities Open Data
BHSS, Academia Sinica

Traceability & Isolation Working Group update

Data Science as a Foundation Toward Open — Dr. Ji-Ping LIN

IPv6 Rollout — Dr. David KELSEY

15:00
A Preliminary Study on Reconstructing Faded Color by Spectral Estimation Method for Heritage Object
BHSS, Academia Sinica

Sirtfi - Incident response for for Identity Federations
Media Conf. Room, BHSS, Academia Sinica

Coffee Break
BHSS, Academia Sinica

Scipion (part 2)
Conf. Room 1, BHSS, Academia Sinica

Coffee Break & Poster Session
BHSS, Academia Sinica 15:31 - 16:00

Coffee Break & Poster Session
BHSS, Academia Sinica

16:00
Using Advanced e-Systems for Community-Engaged Research

Workload Management Trends in WLCG — Maarten LITMAATH
Media Conf. Room, BHSS, Academia Sinica

A Proposal: ePortfolio for enhancing active learning for the future generation

Tier 1 Configuration Evolution & Options — Dr. Josep FLIX
Media Conf. Room, BHSS, Academia Sinica

Occupation recommendation with major programs for adolescents

17:00
Wrap Up & WLCG Workshop update — Mr. Ian COLLIER

Closing remark/round table discussion

Thu 9/3

09:00
Big Data-Driven Drug Discovery — Prof. Jung-Hsin LIN
Conf. Room 2, BHSS, Academia Sinica 09:00 - 09:45

High Performance Computing Environment and Applications in CAS — Dr. Xuebin CHI
Conf. Room 2, BHSS, Academia Sinica 09:45 - 10:30

10:00

Coffee Break & Poster Session
BHSS, Academia Sinica 10:30 - 11:00

11:00
eScience Activities in Singapore — Dr. John KAN

12:00

13:00

14:00

15:00

16:00
infrastructure - latest developments and

proteins in the cloud
Conf. Room 2, BHSS, Academia Sinica

CNGrid As a HPC Application Cloud Service Provider

A solution for secure use of Kibana — Wataru TAKASE

Investigating community detection algorithms and their capacity as markers — Prof. Eva HLADKA

EGI federated platforms supporting accelerated computing
Media Conf. Room, BHSS, Academia Sinica

dCache, towards Federated Identities and — Dr. Paul MILLAR

17:00
A Method for Remote Initial Vetting — Dr. Eisaku SAKANE

2D and 3D Medical Images for Anatomy Education using a cloud computing platform
Conf. Room 2, BHSS, Academia Sinica

A Novel Architecture towards Exascale Computing
Media Conf. Room, BHSS, Academia Sinica

Q&A

Fri 10/3

09:00
The 'Cloud Area Padovana': lessons learned after two years of a

Examination of dynamic partitioning for multi-core jobs in the Tokyo Tier-2

Listening to the ecosystem: the integration of machine

Synergy, a new approach for optimizing the resource usage in OpenStack

The High Throughput Strategy of IHEP — Dr. Jiaheng ZOU

Collaboration on monitoring Asian soundscape and the

10:00
Efficiency Improvement on Distributed Cloud System — Mr. Felix LEE et al.

The Billing System of IHEP Data Center

Revealing Philippine Climate Type Using Remotely-sensed Rainfall

Coffee Break
BHSS, Academia Sinica 10:30 - 10:50

11:00
VCondor - an implemention of dynamic virtual computing cluster

Framework for distributing Radio Astronomy processing across Clusters and Clouds

NeIC EISCAT_3D support project: Nordic computing challenge — Dr. John WHITE

GUOCCI – The Entryway to Federated — Mr. Radim JANČA

Exploiting clouds for smart cities applications - The Cagliari 2020 project
Conf. Room 1, BHSS, Academia Sinica

Towards Environmental Computing Compendium
Media Conf. Room, BHSS, Academia Sinica

CloudIPStore: A Cloud SaaS Repository for your Intellectual Properties

12:00
Supporting Open Science with the EGI Federated Cloud - Experiences, success stories, lessons — Yin CHEN

Future warming scenario and impacts study over Taiwan: Results from ECHAM5/MPIOM-WRF dynamical downscaling — Dr. Chuan Yao LIN
Media Conf. Room 11:50 - 12:20

Q&A
Conf. Room 1, BHSS, Academia Sinica 12:10 - 12:20

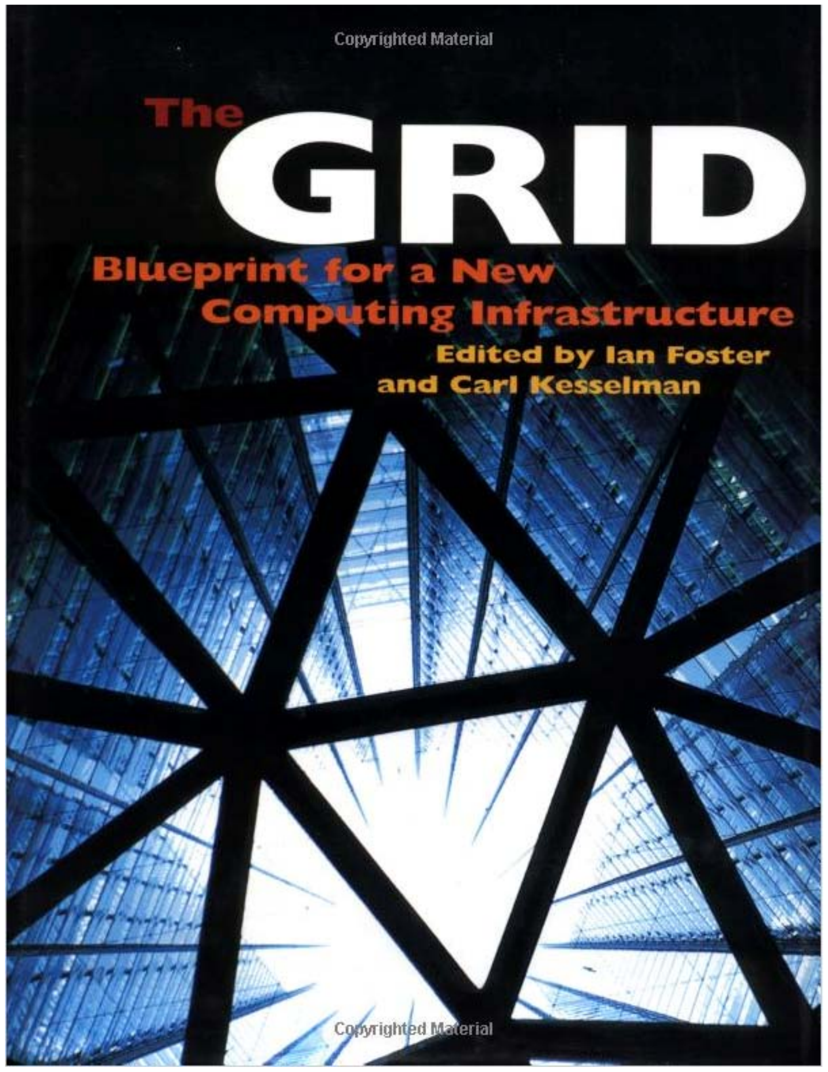
The Emergence of Computational Archival Science — Prof. Richard MARCIANO
Conf. Room 2, BHSS, Academia Sinica 12:20 - 13:00

ISGC Topics

1. Applications from the Virtual Research Communities and Industry
    1. Physics & Engineering applications
    2. Biomedicine & Life Sciences applications
    3. Earth & Environmental Sciences & Biodiversity applications
    4. **Humanities, Arts, and Social Sciences applications**

2. Technologies that Provide Access and Exploitation of Different Site Resources and Infrastructures
    5. Virtual Research Environment (including Middleware, tools, services, workflow, etc.
    6. **Big Data & Data Management**

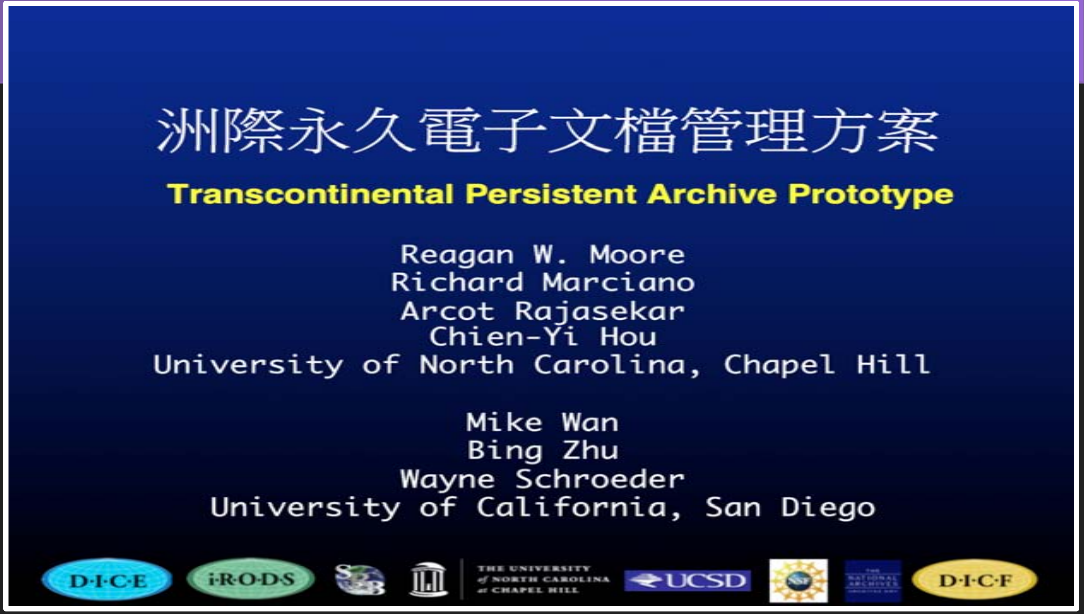
3. Infrastructure for Research
    7. Networking, Security, Infrastructure & Operations
    8. **Infrastructure Clouds and Virtualization**
    9. **Business Models, Policy, and Sustainability**
    10. Massively Distributed Computing and Citizen Sciences
    11. Supercomputing, High Throughput, Accelerator Technologies and Integration



The GRID

Blueprint for a New Computing Infrastructure

Edited by Ian Foster and Carl Kesselman



洲際永久電子文檔管理方案

Transcontinental Persistent Archive Prototype

Reagan W. Moore
Richard Marciano
Arcot Rajasekar
Chien-Yi Hou
University of North Carolina, Chapel Hill

Mike Wan
Bing Zhu
Wayne Schroeder
University of California, San Diego

D·I·C·E i·R·O·D·S THE UNIVERSITY of NORTH CAROLINA at CHAPEL HILL UCSD D·I·C·F

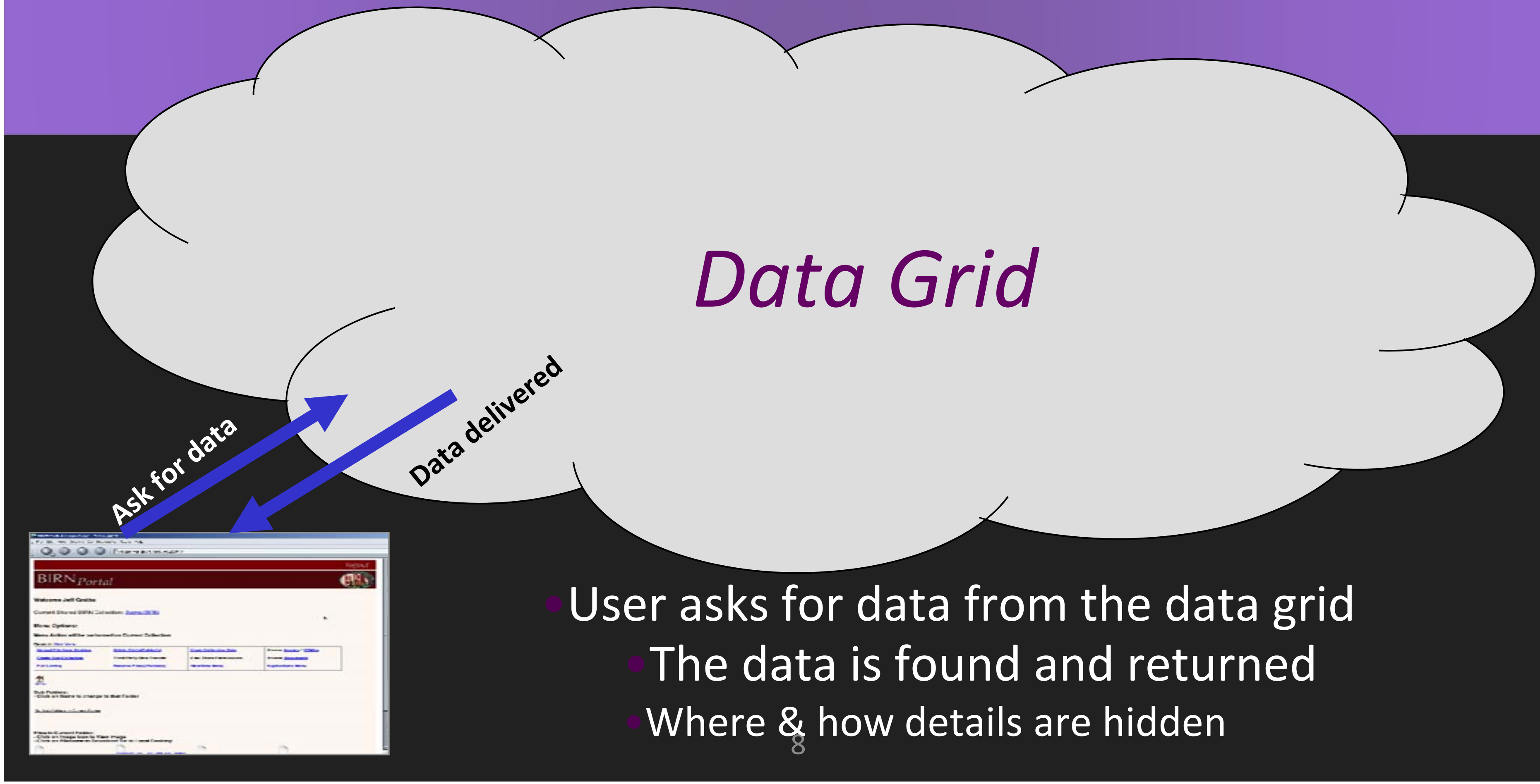
with: Chien-Yi HOU, Sheau-Yen CHEN

What are Data Grids?

Data Grids are middleware services

- Sitting between the applications and data providers
- Providing transparent and uniform access to diverse types of digital assets
  - Files, databases, streams, web, programs,…
  - Documents, images, data, sensor packets, tables,…
- From heterogeneous resources
  - File Systems, tape archives, sensor streams,…
- Distributed over a wide area network
  - Multiple administrative and security domains
- With users unaware of physical attributes of the data access
  - System addresses, paths, protocols, …

Using a Data Grid – *in Abstract*



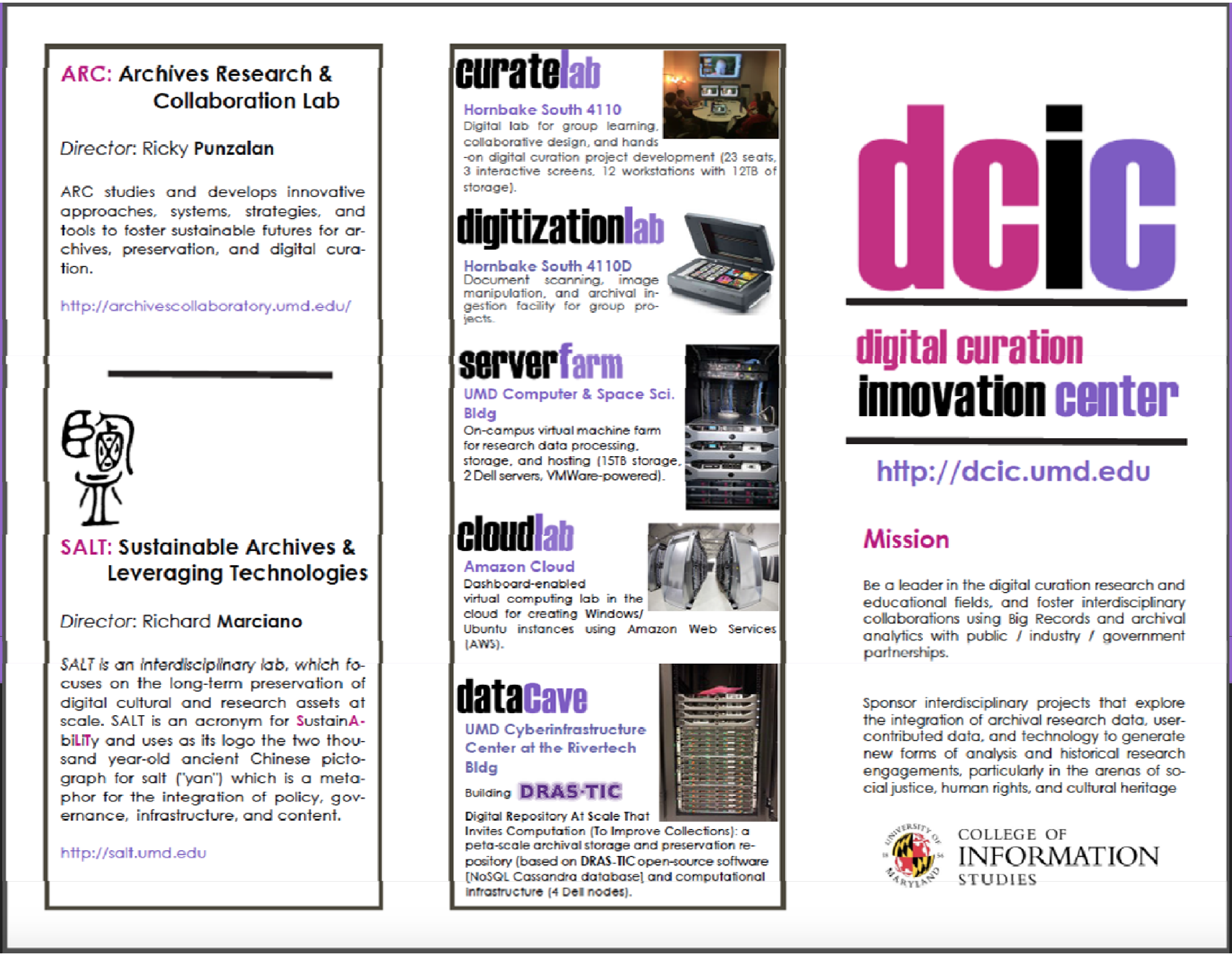
- User asks for data from the data grid
  - The data is found and returned
  - Where & how details are hidden

Digital Curation Innovation Center (DCIC) @ U. Maryland (USA)

Mission:

- Be a leader in the digital curation research and educational fields, and foster interdisciplinary partnerships using **Big Records and archival analytics** through public / industry / government collaborations.

- Sponsor interdisciplinary projects that explore the integration of archival research data, user-contributed data, and technology to **generate new forms of analysis and historical research engagements.**



ARC: Archives Research & Collaboration Lab

Director: Ricky **Punzalan**

ARC studies and develops innovative approaches, systems, strategies, and tools to foster sustainable futures for archives, preservation, and digital curation.

http://archivescollaboratory.umd.edu/

SALT: Sustainable Archives & Leveraging Technologies

Director: Richard **Marciano**

*SALT is an interdisciplinary lab, which fo*cuses on the long-term preservation of digital cultural and research assets at scale. SALT is an acronym for SustainA-biLiTy and uses as its logo the two thousand year-old ancient Chinese pictograph for salt ("yan") which is a metaphor for the integration of policy, governance, infrastructure, and content.

http://salt.umd.edu

curatelab

Hornbake South 4110

Digital lab for group learning, collaborative design, and hands-on digital curation project development (23 seats, 3 interactive screens, 12 workstations with 12TB of storage).

digitizationlab

Hornbake South 4110D

Document scanning, image manipulation, and archival ingestion facility for group projects.

serverfarm

UMD Computer & Space Sci. Bldg

On-campus virtual machine farm for research data processing, storage, and hosting (15TB storage, 2 Dell servers, VMWare-powered).

cloudlab

Amazon Cloud

Dashboard-enabled virtual computing lab in the cloud for creating Windows/ Ubuntu instances using Amazon Web Services (AWS).

dataCave

UMD Cyberinfrastructure Center at the Rivertech Bldg

Building **DRAS-TIC**

Digital Repository At Scale That Invites Computation (To Improve Collections): a peta-scale archival storage and preservation repository (based on DRAS-TIC open-source software [NoSQL Cassandra database] and computational infrastructure (4 Dell nodes).

dcic

digital curation innovation center

http://dcic.umd.edu

Mission

Be a leader in the digital curation research and educational fields, and foster interdisciplinary collaborations using Big Records and archival analytics with public / industry / government partnerships.

Sponsor interdisciplinary projects that explore the integration of archival research data, user-contributed data, and technology to generate new forms of analysis and historical research engagements, particularly in the arenas of social justice, human rights, and cultural heritage

COLLEGE OF
INFORMATION
STUDIES

Projects

Cyberinfrastructure for the curation & management of digital assets at scale:

"Brown Dog"

A CIC Big-10 $10.5M NSF/DIBBs-funded collaboration with U. of Illinois NCSA Supercomputing Center and industry partners (NetApp and Archive Analytics Solutions). This project aims to help accelerate the development of digital curation processes and services and create a data observatory to provide access to Big Records training sets and teach students practical digital curation skills.

"Curate Cloud"

A $300K IMLS-funded project that helped launch a new online professional education certificate for digital curation professionals, the Curation and Management of Digital Assets (CMDA). Curate Cloud is also developing an open-source research and educational platform, the VCL (Virtual Computing Lab), to remove barriers to access for curation tools and resources.

Digital Curation training:

Digital Curation Fellowships

The iSchool has several Fellowship opportunities for students in digital curation and archives. These include a collaboration with the **National Agricultural Library** (NAL); extensive project work with the **National Park Service** (NPS); and a scholarship established in honor of **Bruce Ambacher**, retired senior archivist and iSchool faculty member.

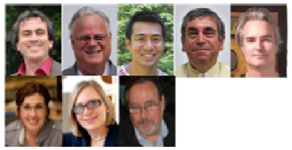
Interdisciplinary Research Teams

Gain new digital skills, conduct interdisciplinary research, explore professional development opportunities at the intersection of archives, big data, and analytics through a number of project themes: *Refugee Narratives, Community Displacement, Racial Zoning, Cyberinfrastructure for Digital Curation, Movement of People, Citizen Internment.*

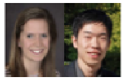
People

Research Staff:

Richard **Marciano**	*Director & SALT Lab Director*
Michael **Kurtz**	*Associate Director*
Ricardo **Punzalan**	*Research Associate & ARC Lab Director*
Ken **Heger**	*Research Associate & DigitizationLab Director*
Greg **Jansen**	*Research Software Architect*
Maria **Esteva**	*Affiliate Professor*
Victoria **Lemieux**	*Affiliate Professor*
William **Underwood**	*Affiliate Professor*



Mary **Kendig** / Myeong **Lee**	*Graduate Research Assistants*



Research Affiliates:

U. Maryland

Tammy **Clegg**, Nick **Diakopoulos**, Jesse **Johnston**, Trevor **Owens**, Jenny **Preece**, Katie **Shilton**

External

Bruce **Ambacher**, Natalie **Baur**, John **Burns**, Andrew **Lau**, Scott **Madry**

Postdoctoral Fellows:

Morgan **Daniels**, Kathryn **Gucer**, Adam **Kriesberg**

(Advisor: Punzalan)

Students (Undergraduate, Master's [MLIS, MIM, HCIM], Doctoral):

Maddie **Allen**, Saba **Al-Dughaither**, Vinita **Atre**, Myuresh **Amdekar**, Richard **Bool**, Arpit **Chandra**, Shiyun **Chen**, John **Dela Cruz**, Anne **Dempsey**, Shaina **Destine**, Kelsey **Diemand**, Pal **Doshi**, Erin **Durham**, Will **Froliklong**, Alicia **Geller**, Karishma **Ghiya**, Janet **Glazier**, Rajesh **Gnanasekaran**, Rhett **Greenfield**, Allison **Gunn**, Ashley **Haddix**, Scott **Harkless**, Tarra **Hausmann**, Eric **Hung**, Hardik **Jhaveri**, Ruchira **Kapoor**, Amar **Kurane**, Yuting **Liao**, Zhenye **Ma**, Sheryl **Mathias**, Paridhi **Mathur**, Martin **Moreno**, Jennifer **Proctor**, Brian **Redford**, Darlene **Reyes**, Benjamin **Sagey**, Sohan **Shah**, Jay **Sheth**, Niraj **Shirame**, Edel **Spencer**, Akash **Udani**, Sydney **Valle**, Jennifer **Wachtel**, Melissa **Wertheimer**, Meaghan **Wilson**, Jiahui **Wu**, David **Zhang**, Xinyun **Zhang**

Doctoral Students:

Andrew **Caserlano**, Will **Thomas**	*(Advisor: Marciano)*
Diane **Travis**	*(Advisors: Butler/Marciano)*
Edward **Summers**, Amy **Wickner**	*(Advisor: Punzalan)*

Projects

Justice, Human Rights, & Cultural Heritage:

Overseas Pension Project

A student- and professional society-driven project to collect information documenting payment of pensions to American veterans living overseas. The project creates datasets documenting migration patterns, the flow of money, health conditions, and family connections prior to World War I.

International Research Portal Project (IRP²)

This project will improve access to an important tool which identifies and locates looted art and other cultural assets found on the *International Research Portal for Records Related to Nazi-Era Cultural Property.*

Mapping the Voyage of the St. Louis

In 1939, 937 passengers (mostly Jews) fled Germany aboard the SS St. Louis ship, heading to Cuba, where they were turned away and forced to return to Europe where 254 were killed during the Holocaust. The project looks at mapping individual and collective stories through graph database techniques.

Japanese-American WWII Camps

Building on a UMD FIA Seed Grant, the project explores the integration of archival and user-contributed data using social networking graphs to link people, places, and events. Using WWII Camp data.

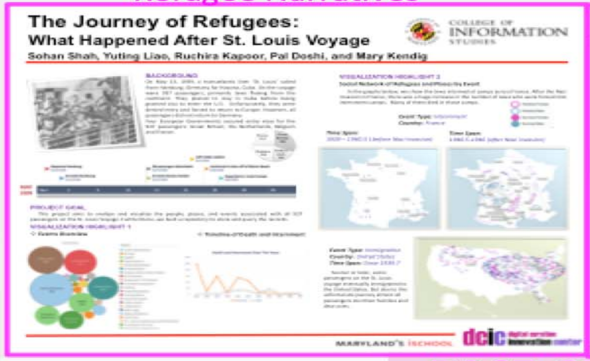
Mapping Inequality

A project with Johns Hopkins, Virginia Tech, and U. of Richmond where a national collection of New Deal *redlining* records is being crowdsourced (these unique records capture racial, ethnic, and economic conditions).

The Human Face of Big Data

A student-led project that will create access and collaborative opportunities around historically and socially- significant heterogeneous datasets rooted in urban renewal housing records for a number of cities.

Refugee Narratives



The Journey of Refugees: What Happened After St. Louis Voyage

Sohan Shah, Yuting Liao, Ruchira Kapoor, Pal Doshi, and Mary Kendig

COLLEGE OF INFORMATION STUDIES

MARYLAND'S iSCHOOL

Racial Zoning

Mapping Inequality: Redlining in New Deal America

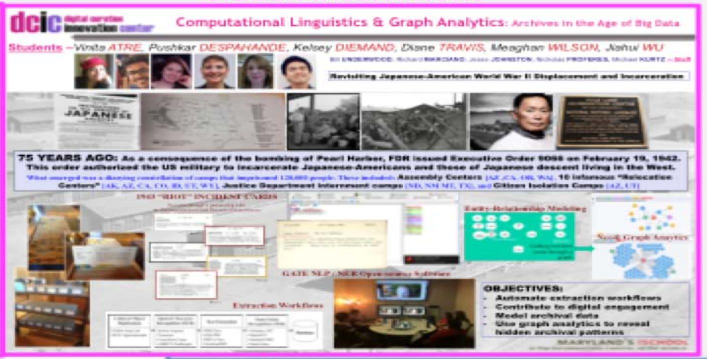
Goal 1: Visualizing Historical Data on the Map

Goal 2: Transcribing and Structuring Scanned Documents

Solution 1: OCR (ABBYY). A Fully Automated Way

Solution 2: SCRIBE. A Crowdsourcing Technique

Citizen Internment



Computational Linguistics & Graph Analytics: Archives in the Age of Big Data

Students – Vinita ATRE, Pushkar DESPAHANDE, Kelsey DIEMAND, Diane TRAVIS, Meaghan WILSON, Jiahui WU

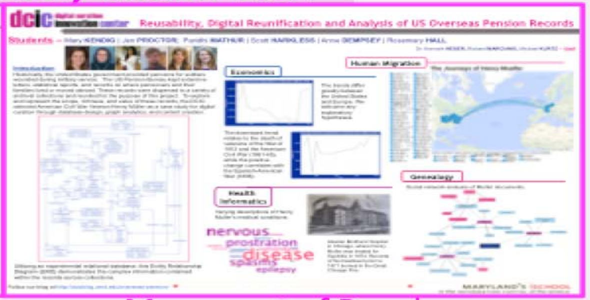
Revisiting Japanese-American World War II Displacement and Incarceration

75 YEARS AGO: As a consequence of the bombing of Pearl Harbor, FDR issued Executive Order 9066 on February 19, 1942. This order authorized the US military to incarcerate Japanese-Americans and those of Japanese descent living in the West.

OBJECTIVES:
- Automate extraction workflows
- Contribute to digital engagement
- Model archival data
- Use graph analytics to reveal hidden archival patterns

RACE, SPACE, and PLACE

Human Face of Big Data: Urban Renewal in Asheville, NC



Reusability, Digital Reunification and Analysis of US Overseas Pension Records

Community Displacement

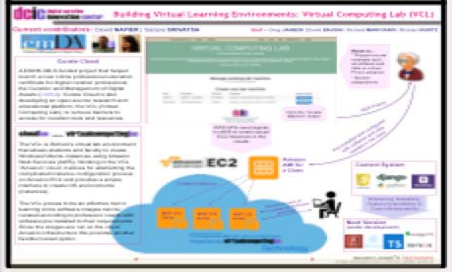
Movement of People

dcic digital curation innovation center

INNOVATIVE CYBER-INF.

Virtual Learning

Brown Dog: Cloud Services for Auto-Curation



Building Virtual Learning Environments: Virtual Computing Lab (VCL)

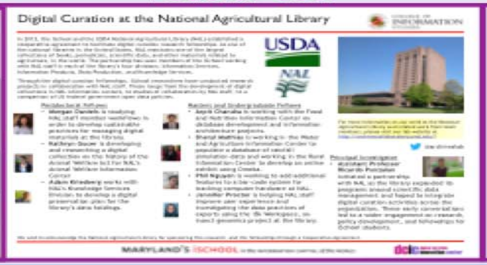
Curation At Scale

DIGITAL CURATION

Local Gov.

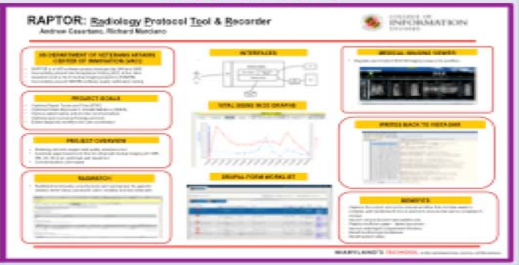
Historic Survey of the Eastport Community: a Partnership for Action Learning in Sustainability (PALS)

Federal Gov.



Digital Curation at the National Agricultural Library

Health Care



RAPTOR: Radiology Protocol Tool & Recorder

Andrew Casertano, Richard Marciano

Pursuing a CAS Training / Teaching Agenda

There is a need to :

- create innovative classes that emphasize new modes of collaboration, and interdisciplinary work.
- blend elements of archival thinking and computational thinking:
  - problem solving that uses modeling, decomposition, pattern recognition, abstraction, algorithm design, and scale.
- develop inter-disciplinary iSchools with faculty from Computer Science, Archival Science, and Data Science.
- develop extensive hands-on experience working with cyberinfrastructure to carry out archival functions.

How Each Project Is Related to Computational Archival Science (CAS) Themes:

Project	Computational Linguistics	Data Modeling & Evolutionary Prototyping	Graph Analytics	Crowdsourcing	GIS
1. Human Face of Big Data [Community Displacement]		X		X	X
2. Mapping Inequality [Racial Zoning]	X			X	X
3. St. Louis Voyage [Refugee Narratives]		X	X		X
4. World War II Japanese Camps [Citizen Internment]	X	X	X	X	X

IEEE Big Data 2016 “Computational Archival Science: digital records in the age of big data

http://dcicblog.umd.edu/cas/ieee_big_data_2016_cas-workshop/

Dec. 2016 workshop

http://dcicblog.umd.edu/cas

April 2016 workshop

Upcoming… IEEE Big Data 2017 in BOSTON

- Mark Hedges, Tobias Blanke, KCL
- Bill Underwood, GTRI (now UMD)
- Victoria Lemieux, UBC
- Maria Esteva, TACC
- Richard Marciano, Michael Kurtz, UMD

- **Application of analytics to archival material**, including text-mining, data-mining, sentiment analysis, network analysis.
- **Analytics in support of archival processing**, including appraisal, arrangement and description.
- **Scalable services for archives**, including identification, preservation, metadata generation, integrity checking, normalization, reconciliation, linked data, entity extraction, anonymization and reduction.
- **New forms of archives**, including Web, social media, audiovisual archives, and blockchain.
- **Cyber-infrastructures for archive-based research** and for development and hosting of collections
- **Big data and archival theory** and practice
- Digital curation and preservation
- Crowdsourcing and archives
- **Big data and the construction of memory** and identity
- Specific big data technologies (e.g. NoSQL databases) and their applications
- Corpora and reference collections of big archival data
- Linked data and archives
- Big data and **provenance**
- Constructing big data research objects from archives

Our working definition of Archival Computational Science (CAS):

An interdisciplinary field concerned with the application of computational methods and resources to large-scale records/archives processing, analysis, storage, long-term preservation, and access, with aim of improving efficiency, productivity and precision in support of appraisal, arrangement and description, preservation and access decisions, and engaging and undertaking research with archival material.

e.g.: NSF/SBE RIDIR, LOC National Digital Initiative, IMLS Always Already Computational, etc.

DataNet: $62.6M

SEAD:	$8.1M (Michigan): data curation software and services for the "long tail" of small- and medium-scale data producers in sustainability science.
Terra Populus:	$8.2M (Minnesota) -- Build tools for data integration across the domains of social science and environmental data
DFC:	$8.3M (North Carolina) -- Use the integrated Rule-Oriented Data System (iRODS) to provide data grid infrastructure for science and engineering.
DataONE:	$27.9M (New Mexico) -- $20M + $7.9M Oct. 2014) -- platform for collaborative environmental and ecological science
DataConservancy:	$10M – Johns Hopkins U. – 2009-2014

NSF Office of Advanced CI (part of CISE)

DataNet &DIBBs:

~ $178M of National Investments ($115M for DIBBs)

DIBBS: 3 1/2 years: $115 M

======= Fall 2013: $32.8M

Brown Dog	-- University of Illinois at Urbana-Champaign / U. Maryland	-- $10.5M
Data Exacell	-- Carnegie Mellon University	-- $8.9M
SkyServer	-- Johns Hopkins University	-- $8.9M
GABBs	-- Purdue University	-- $4.5M

======= Fall 2014: $20.8M

-**Building a Modular Cyber-Platform for Systematic Collection, Curation, and Preservation of Large Engineering and Science Data** -- Purdue University -- $1.5M
-**User Driven Architecture for Data Discovery** -- Corporation for National Research Initiatives (NRI) -- $1.5M
-**Collaborative Research: Cyberinfrastructure for Interpreting and Archiving U-series Geochronologic Data** -- College of Charleston -- $580K
-**T2-C2: Timely and Trusted Curator and Coordinator Data Building Blocks** -- University of Illinois at Urbana-Champaign -- $1.5M
-**Scalable Capabilities for Spatial Data Synthesis** -- University of Illinois at Urbana-Champaign -- $1.5M
-**Domain-Aware Management of Heterogeneous Workflows: Active Data Management for Gravitational-Wave Science Workflows** -- Syracuse University -- $750K
-**SPIDAL: Middleware and High Performance Analytics Libraries for Scalable Data Science** -- Indiana University -- $5.1M
-**Ubiquitous Access to Transient Data and Preliminary Results via the SeedMe Platform** -- University of California-San Diego -- $1.3M
-**DIBBs for Intelligence and Security Informatics Research Community** -- University of Arizona -- $1.5M
-**STORM: Spatio-Temporal Online Reasoning and Management of Large Data** -- University of Utah -- $1.2M
-**Systematic Data-Driven Analysis and Tools for Spatiotemporal Solar Astronomy Data** -- Georgia State University Research Foundation -- $1.5M
-**An Infrastructure for Computer Aided Discovery in Geoscience** -- Massachusetts Institute of Technology -- $1.4M
-**Porting Practical Natural Language Processing (NLP) and Machine Learning (ML) Semantics** -- University of Colorado at Boulder -- $1.5M

======= 2015: $27.5

-**Tripal Gateway, a Platform for Next-Generation Data Analysis and Sharing** -- Washington State University – $1.5M
-**An Integrated System for Public/Private Access to Large-Scale, Confidential Social Science Data** --Duke University -- $1.5M
-**LearnSphere: Building a Scalable Infrastructure for Data-Driven Discovery and Innovation in Education** -- Carnegie-Mellon University -- $4.8M
-**An Infrastructure Supporting Collaborative Data Analytics Workflow Design and Management** -- Carnegie-Mellon University -- $1M
-**DNI: Give Your Data the Edge: A Scalable Data Delivery Platform** -- University of Arizona -- $3.8M
-**DNI: Multi-Institutional Open Storage Research InfraStructure (MI-OSiRIS)** --University of Michigan Ann Arbor -- $4.9M
-**DNI: Data Analysis and Management Building Blocks for Multi-Campus Cyberinfrastructure through Cloud Federation** -- Cornell University -- $5M
-**DNI: The Pacific Research Platform** -- University of California-San Diego -- $5M

======= 2016: $31.4M

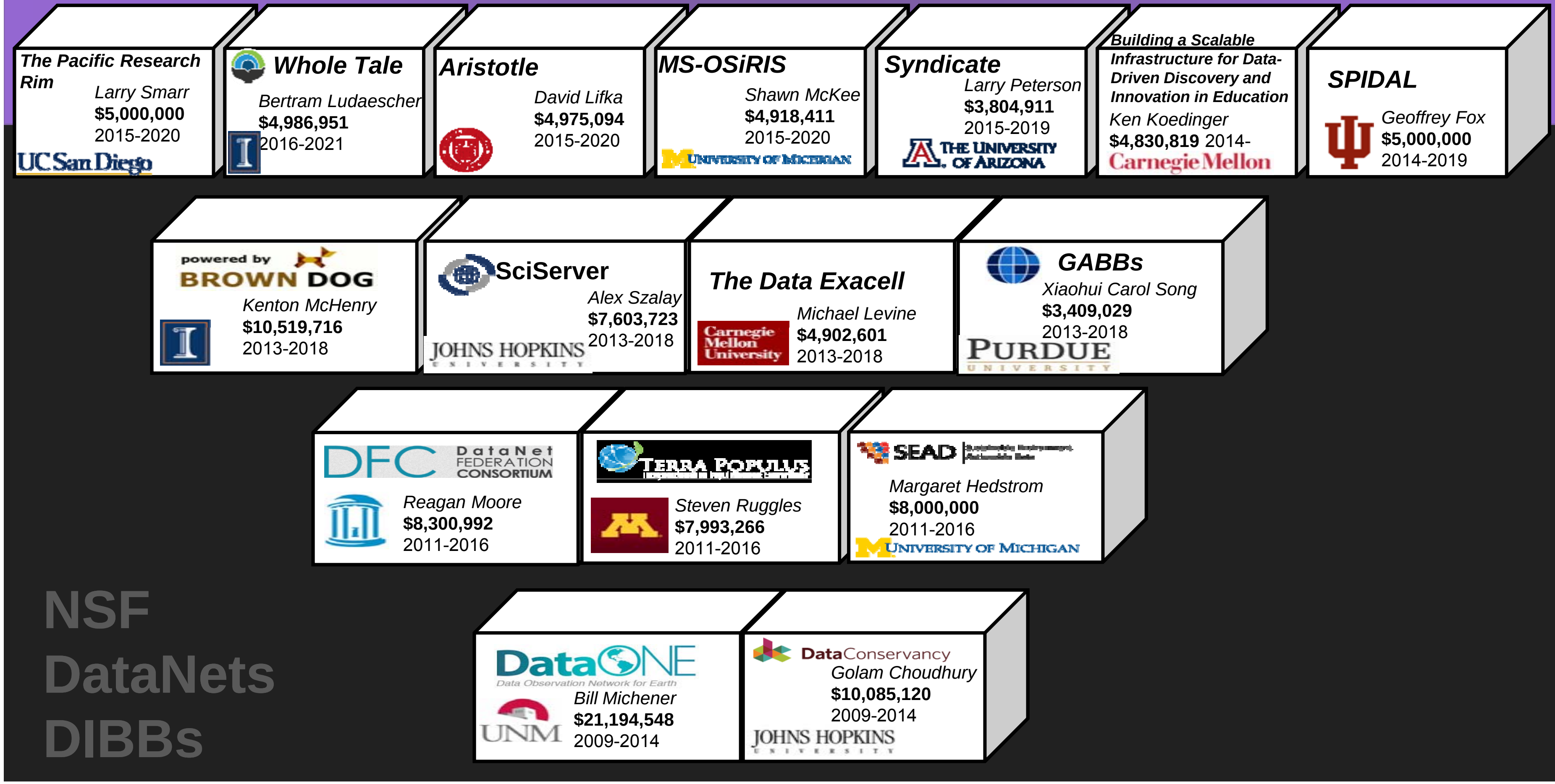
-**EI: Virtual Data Collaboratory: A Regional Cyberinfrastructure for Collaborative Data Intensive Science** – Rutgers -- $4M
-**EI: Data Laboratory for Materials Engineering –** SUNY at Buffalo-- $2.9M
-**EI: mProv: Provenance-based Data Analytics cyberinfrastructure for High-frequency Movile Sensor data –** U. Memphis -- $4M
-**EI: Merging Science and Cyberinfrastructure Pathways: The Whole Tale** -- University of Illinois at Urbana-Champaign -- $5M
-**PD: Ontology-Enabled Polymer Nanocomposite Open Community Data Resource** -- Rensselaer Polytechnic Institute -- $500K
-**EI: The Local Spectroscopy Data Infrastructure (LSDI)** – UC Berkeley -- $3.9M
-**EI: VIFI:Virtual Information-Fabric Infrastructure (VIFI) for Data-Driven Decisions from Distributed Data** -- UNC Charlotte -- $4M
-**PD: Metadata Toolkits for Building Multi-Faceted Data – Relationship Models** – MIT -- $500K
-**EI: Continuous Capture of Metadata for Statistical Data** – U. Michigan -- $2.6M
-**EI: North East Storage Exchange** – Harvard U. -- $4M

======= 2017: $3.2M so far…

-**EI: Vizier, Streamlined Data Curation** – SUNY at Buffalo -- $2.7M
-**PD: Accelerating Comparative Metagenomics through an Ocean Cloud Commons** – U. Arizona -- $500K

This solicitation includes two classes of science data pilot awards:

- **Early Implementations** are large "at scale" evaluations, building upon cyberinfrastructure capabilities of existing research communities or recognized community data collections, and extending those data-focused cyberinfrastructure capabilities to additional research communities and domains with broad community engagement.
- **Pilot Demonstrations** address advanced cyberinfrastructure challenges across emerging research communities, building upon recognized community data collections and disciplinary research interests, to address specific challenges in science and engineering research.



The Pacific Research Rim
Larry Smarr
$5,000,000
2015-2020
UC San Diego

Whole Tale
Bertram Ludaescher
$4,986,951
2016-2021

Aristotle
David Lifka
$4,975,094
2015-2020

MS-OSiRIS
Shawn McKee
$4,918,411
2015-2020
University of Michigan

Syndicate
Larry Peterson
$3,804,911
2015-2019
The University of Arizona

Building a Scalable Infrastructure for Data-Driven Discovery and Innovation in Education
Ken Koedinger
$4,830,819 2014-
Carnegie Mellon

SPIDAL
Geoffrey Fox
$5,000,000
2014-2019

powered by BROWN DOG
Kenton McHenry
$10,519,716
2013-2018

SciServer
Alex Szalay
$7,603,723
2013-2018
Johns Hopkins University

The Data Exacell
Michael Levine
$4,902,601
2013-2018
Carnegie Mellon University

GABBs
Xiaohui Carol Song
$3,409,029
2013-2018
Purdue University

DFC DataNet Federation Consortium
Reagan Moore
$8,300,992
2011-2016

Terra Populus
Steven Ruggles
$7,993,266
2011-2016

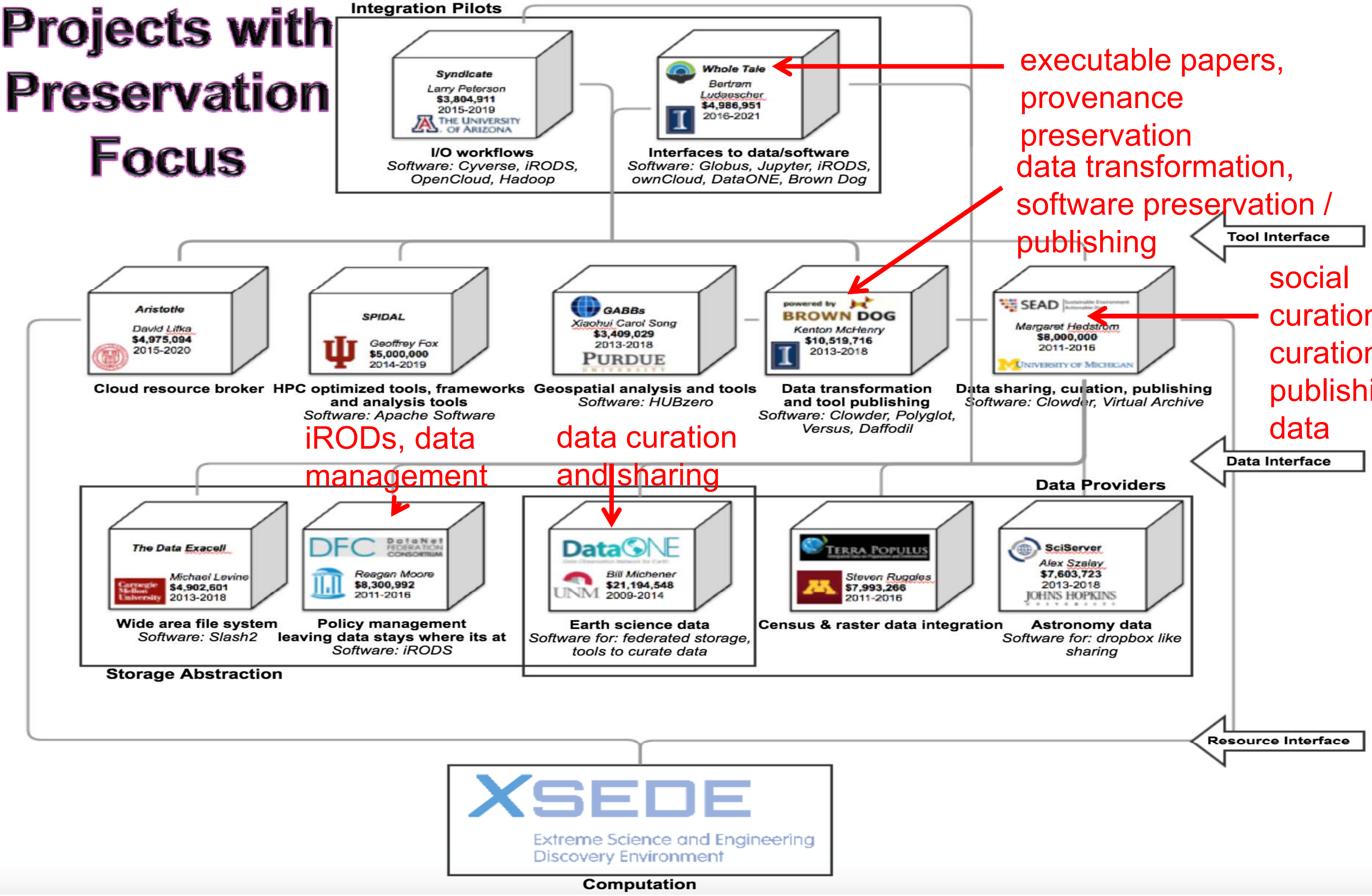
SEAD
Margaret Hedstrom
$8,000,000
2011-2016
University of Michigan

DataONE
Data Observation Network for Earth
UNM
Bill Michener
$21,194,548
2009-2014

DataConservancy
Golam Choudhury
$10,085,120
2009-2014
Johns Hopkins University

NSF DataNets DIBBs

Projects with Preservation Focus



Integration Pilots

Syndicate
Larry Peterson
$3,804,911
2015-2019
The University of Arizona

I/O workflows
Software: Cyverse, iRODS, OpenCloud, Hadoop

Whole Tale
Bertram Ludaescher
$4,986,951
2016-2021

Interfaces to data/software
Software: Globus, Jupyter, iRODS, ownCloud, DataONE, Brown Dog

Tool Interface

Aristotle
David Lifka
$4,975,094
2015-2020

Cloud resource broker

SPIDAL
Geoffrey Fox
$5,000,000
2014-2019

HPC optimized tools, frameworks and analysis tools
Software: Apache Software

GABBs
Xiaohui Carol Song
$3,409,029
2013-2018
Purdue University

Geospatial analysis and tools
Software: HUBzero

powered by Brown Dog
Kenton McHenry
$10,519,716
2013-2018

Data transformation and tool publishing
Software: Clowder, Polyglot, Versus, Daffodil

SEAD
Margaret Hedstrom
$8,000,000
2011-2016
University of Michigan

Data sharing, curation, publishing
Software: Clowder, Virtual Archive

Data Interface

Data Providers

The Data Exacell
Carnegie Mellon University
Michael Levine
$4,902,601
2013-2018

Wide area file system
Software: Slash2

DFC DataNet Federation Consortium
Reagan Moore
$8,300,992
2011-2016

Policy management
leaving data stays where its at
Software: iRODS

Storage Abstraction

DataONE
UNM
Bill Michener
$21,194,548
2009-2014

Earth science data
Software for: federated storage, tools to curate data

Terra Populus
Steven Ruggles
$7,993,266
2011-2016

Census & raster data integration

SciServer
Alex Szalay
$7,603,723
2013-2018
Johns Hopkins University

Astronomy data
Software for: dropbox like sharing

Resource Interface

XSEDE
Extreme Science and Engineering Discovery Environment

Computation

executable papers, provenance preservation

data transformation, software preservation / publishing

social curation, auto curation, publishing data

iRODs, data management

data curation and sharing

Big Data Curation
& Large-Scale Collaboration

NCSA Brown Dog

“The Super Mutt”

Public API for

- Format Migration
- Feature Extraction

Web Scale

The Problem Addressed by Brown Dog

- Large collections of **un-curated** and/or **unstructured** digital data (“long-tail” data)
  - Many file formats
  - No metadata
  - No useful filenames
  - No useful directory structure
  - No textual contents

What Is Needed

- Means of indexing data contents so that large collections of data can be searched and desired data found
  - An ability to compare data

Brown Dog Data Transformation Services

- The Data Access Proxy (DAP)
  - http://dap.ncsa.illinous.edu/conversion/:output/:file
  - File in, File out

- The Data Tilling Service (DTS)
  - http://dts.ncsa.illinois.edu/extraction/:domain/:file
  - File in, JSON out
  - JSON can contain metadata, tags, signatures, links to derived data products, etc…

Brown Dog Use Cases

- Addressed specifically here:
  1. Biology/Ecology
  2. Civil and Environmental Engineering
  3. Social Science

- Testbed data:
  4. UMD CI-BER testbed, at the U. Maryland iSchool

Brown Dog

The Data Tilling Service (DTS)

Data Tilling

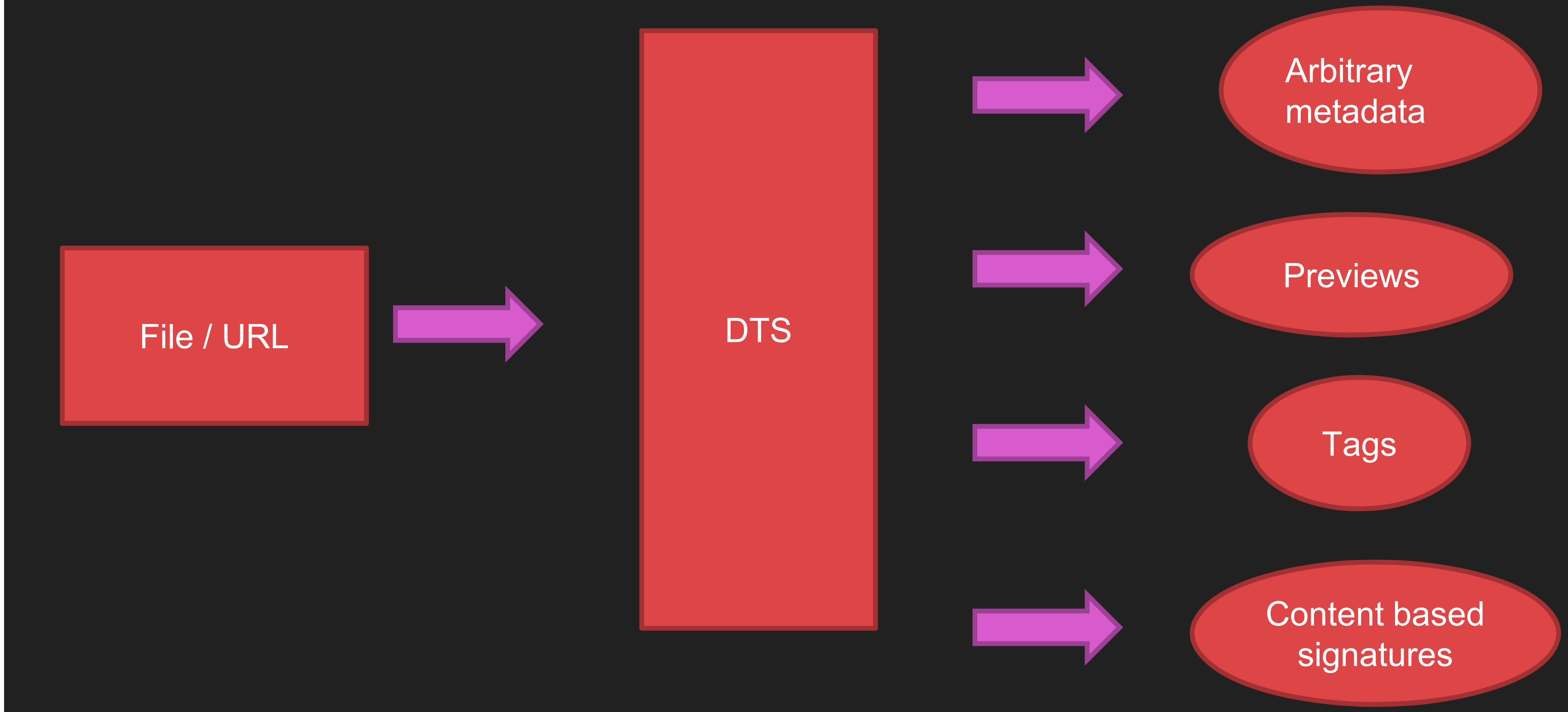
- Data Tilling (v): To prepare and cultivate (*data*) for *analysis*
- Data Tillage (n): Is the *computational* preparation of *data* by *algorithmic* agitation of various types, such as digging, stirring, and overturning

+

(Pre) Data Analysis

- Not necessary *data cleaning*
- More like *metadata extraction*
- Not full analysis / Not perfect results
- Apply as many methods as possible
- Support the user in finding the metadata they need

Extracting Information from Raw Files



File / URL

DTS

Arbitrary metadata

Previews

Tags

Content based signatures

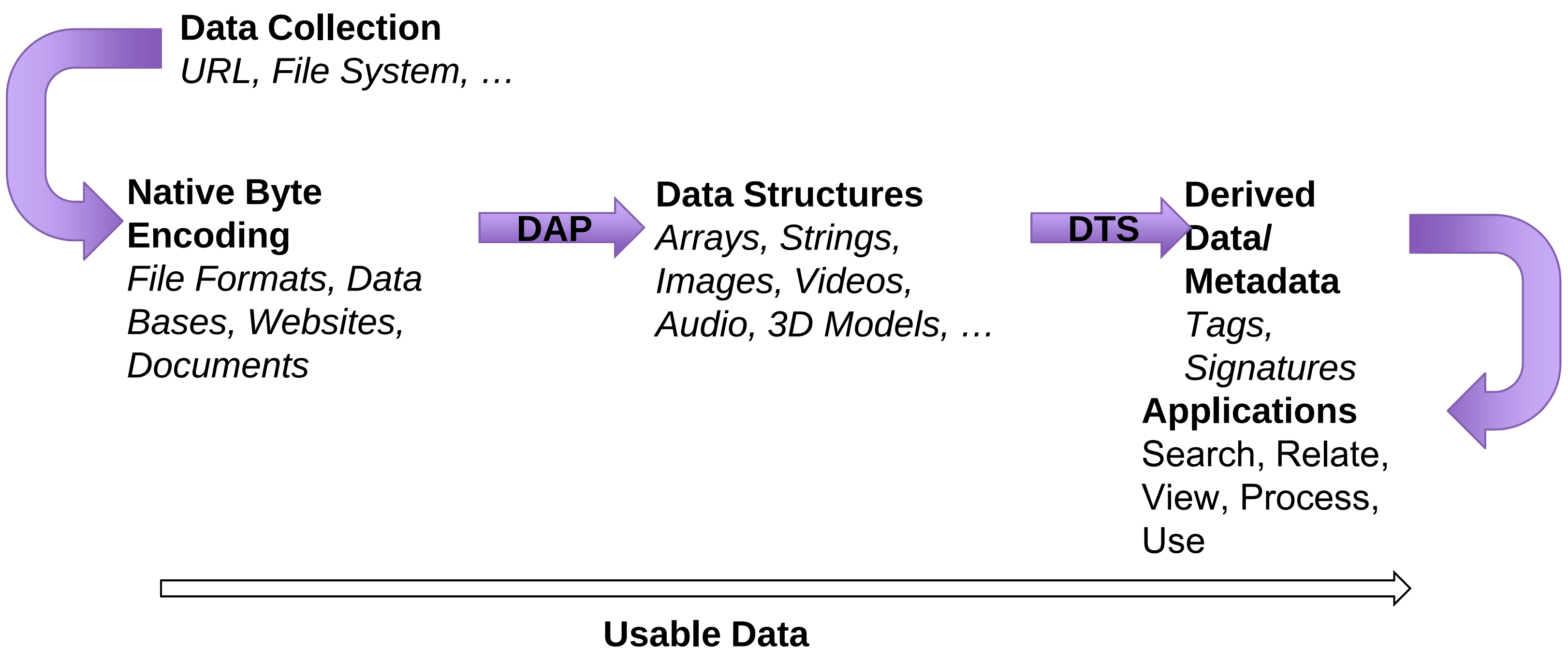
Extractors

[M3d] User Case Extractors

- Floodplain extraction
- Pond extraction from aerial photos
- Gap filled versions of the data
- Text extraction from digitized documents (in particular numerical values)
- River locations from hand drawn maps
- River locations from aerial photos
- Route/image extraction
- Geolocation
- Green Index extractor
- Human preference from images
- Sentiment analysis from text
- Data extraction from articles (e.g. tables)

[M3d] Other Extractors

- OpenCV
  - Faces, eyes
- Tika
  - Language detection
- Simple Summary
  - Summaries
- Cell Profiler
  - Human, yeast, fly, tumor, …
- Tesseract
  - Text extraction from images
- CMU Sphinx
  - Speech recognition
- VLFeat
  - Plane, motorcycle, …



Data Collection
URL, File System, …

Native Byte Encoding
File Formats, Data Bases, Websites, Documents

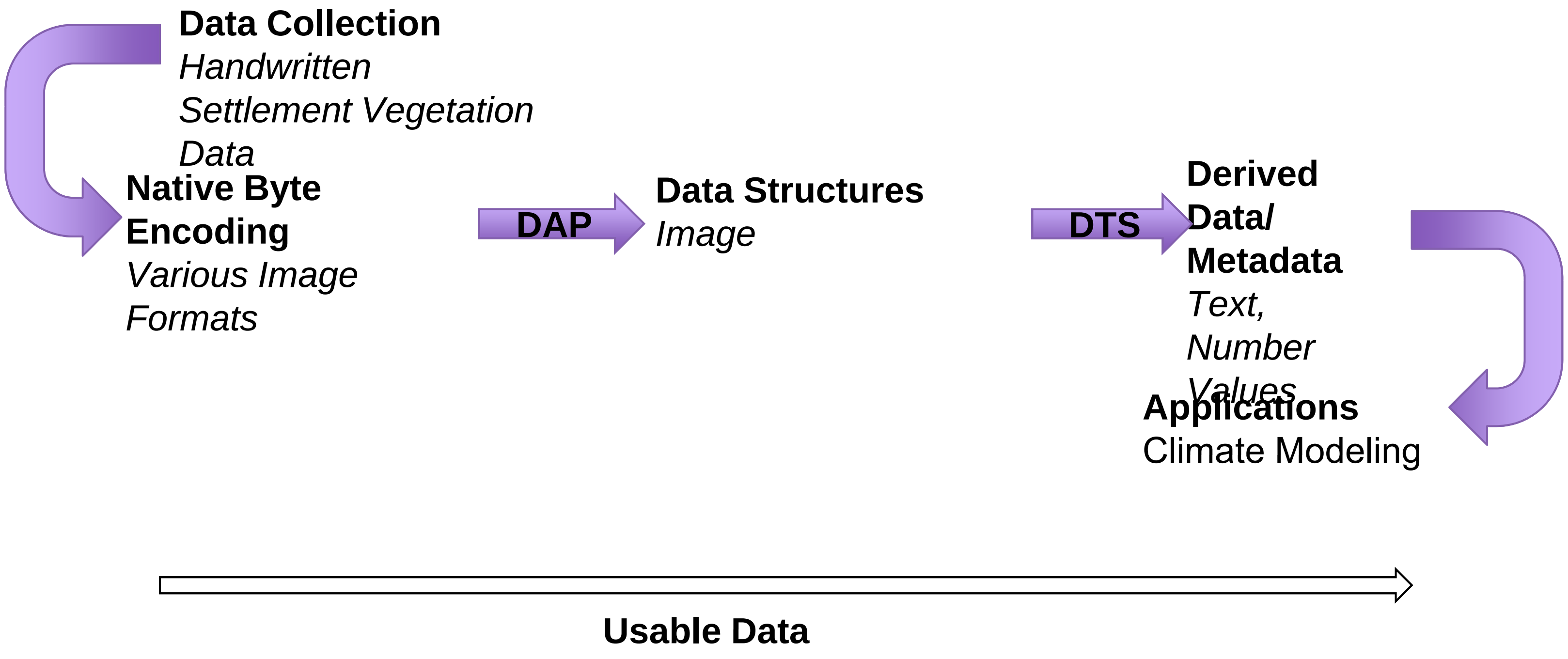
Data Structures
Arrays, Strings, Images, Videos, Audio, 3D Models, …

Derived Data/ Metadata
Tags, Signatures

Applications
Search, Relate, View, Process, Use

Usable Data

M. Dietze, Ecology, Boston University
A. Desai, Ecology, University of Wisconsin
D. LeBauer, Ecology, University of Illinois Urbana-Champaign



Data Collection
Handwritten Settlement Vegetation Data

Native Byte Encoding
Various Image Formats

DAP

Data Structures
Image

DTS

Derived Data/ Metadata
Text, Number Values

Applications
Climate Modeling

Usable Data

Chs. lks.
a B. Oak 12 in. diam. brs. N. 85 E. 15 lks. dis.
a Do 12 " " " S. 40 W. 20 " "
79.87 to Sec. Cor. April 24th

North Between Sections 10 & 11
15.00 Left Prairie
22.75 a Gum 14 in. diam.
28.50 Road to Lafiana
30.00 a Marsh
36.50 Canal Line
40.00 Set qr. Sec. post fr. wh.
a W. Oak 10 in. diam. brs. S. 18 E. 365 lks. dis.
a B. Oak 15 " " " S. 35 W. 300 " "
42.50 Pigeon River 110 lks. wide co. S. W.
47.00 Left Marsh
72.70 a W. Oak 12 in. diam.
80.00 Set post cor. to Secs. 2. 3. 10 & 11 fr. wh.
a W. Oak 18 in. diam. brs. S. 15 W. 200 lks. dis.
a Do 12 " " " S. 65 W. 114 " "
Land &c same

East On Random Betw. Secs. 2 & 11.
40.00 Set Temp. post
48.00 Wet Prairie
58.00 Left Do
79.88 Intersected N. & S. Line 16 lks. N. of post
Land &c same

West On True Line
39.94 Set qr. Sec. post fr. wh.
a W. Oak 12 in. diam. S. 35 E. 32 lks. dis.
a Do 8 " " S. 85 W. 32 " "
79.88 to Sec. Cor.

North Between Sections 2 & 3
23.10 a B. Oak 14 in. diam.

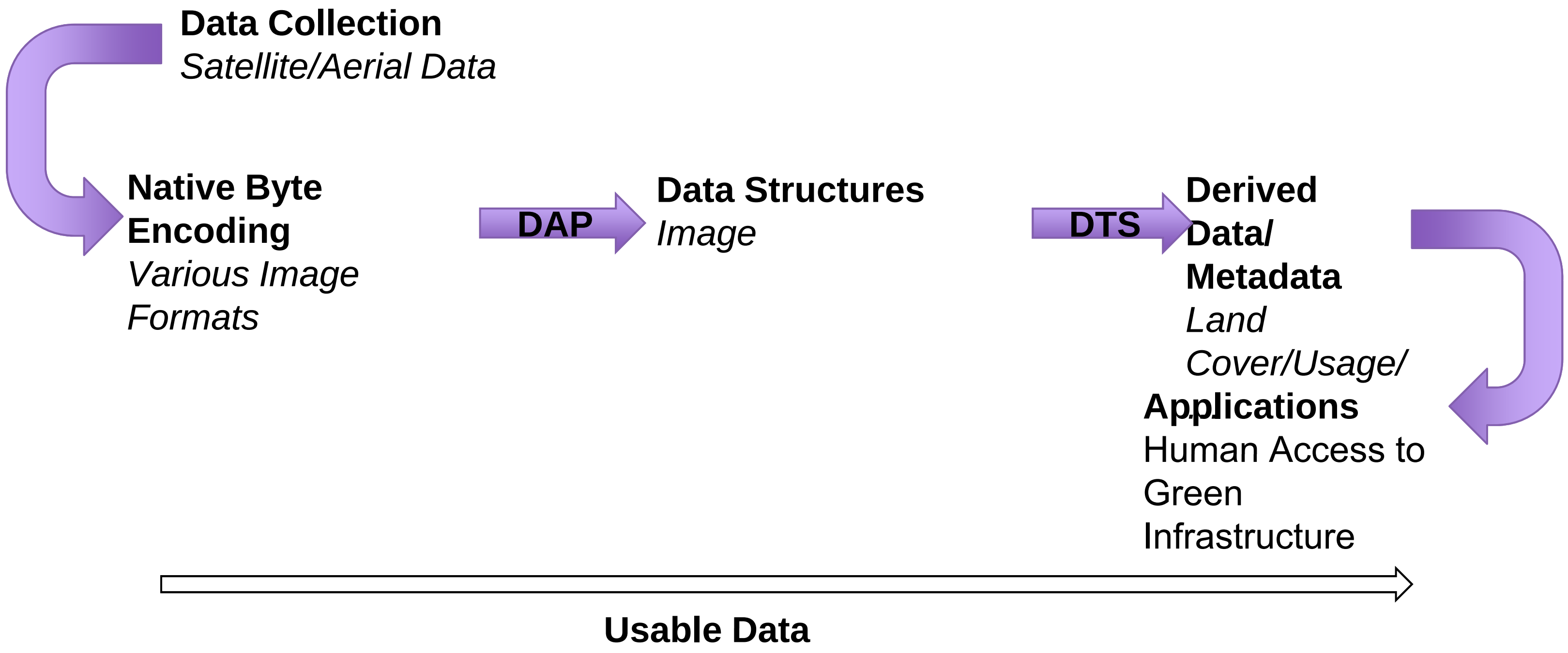
Chs. lks.
25.00 a Marsh
36.00 Left Do
40.00 Set qr. Sec. post fr. wh.
a W. Oak 18 in. diam. brs. N. 72 W. 75 lks. dis.
a R. Oak 20 " " " S. 38 E. 40 " "
81.14 Intersected N. Boundy. 1.75 W. of post
Set post at intersection fr. wh.
a Hickory 10 in. diam. brs. S. 39 E. 250 dis.
a B. Oak 12 " " " S. 35 W. 275 "
Land &c same April 25th

North Between Sections 33 & 34
33.50 a W. Oak 16 in. diam.
40.00 Set qr. Sec. post fr. wh.
a B. Oak 20 in. diam. brs. N. 58 W. 28 lks. dis. &
a Do 20 " " " S. 45 E. 18 " "
69.80 a W. Oak 18 " "
80.00 Set post cor. to Secs. 27. 28. 33 & 34 fr. wh.
a B. Oak 12 in. diam. brs. S. 47 E. 90 lks. dis.
a Do 40 " " " N. 70 W. 70 " "
Land rolling 1st Rate — Timber
Ash, Walnut, Poplar, Oak &c —

East On Random Betw. Secs. 27 & 34
36.00 a Creek 50 lks. wide co. N. W.
40.00 Set Temp. post
80.00 Intersected N. & S. Line 10 lks. S. of post
Land level 2nd Rate Timber
B. & W. Oak Sassafras & Grapevine

West On True Line
40.00 Set qr. Sec. post fr. wh.
an S. Ash 20 in. diam. brs. S. 40 E. 22 lks. dis.
a Sycamore 20 " " " N. 29 W. 50 " "
80.00 to Sec. Cor.

B. Minsker, Civil & Env. Engineering, University of Illinois Urbana-Champaign
A. Schmidt, Civil & Env. Engineering, University of Illinois Urbana-Champaign
B. Sullivan, Landscape Architecture, University of Illinois Urbana-Champaign



Data Collection
Satellite/Aerial Data

Native Byte Encoding
Various Image Formats

DAP

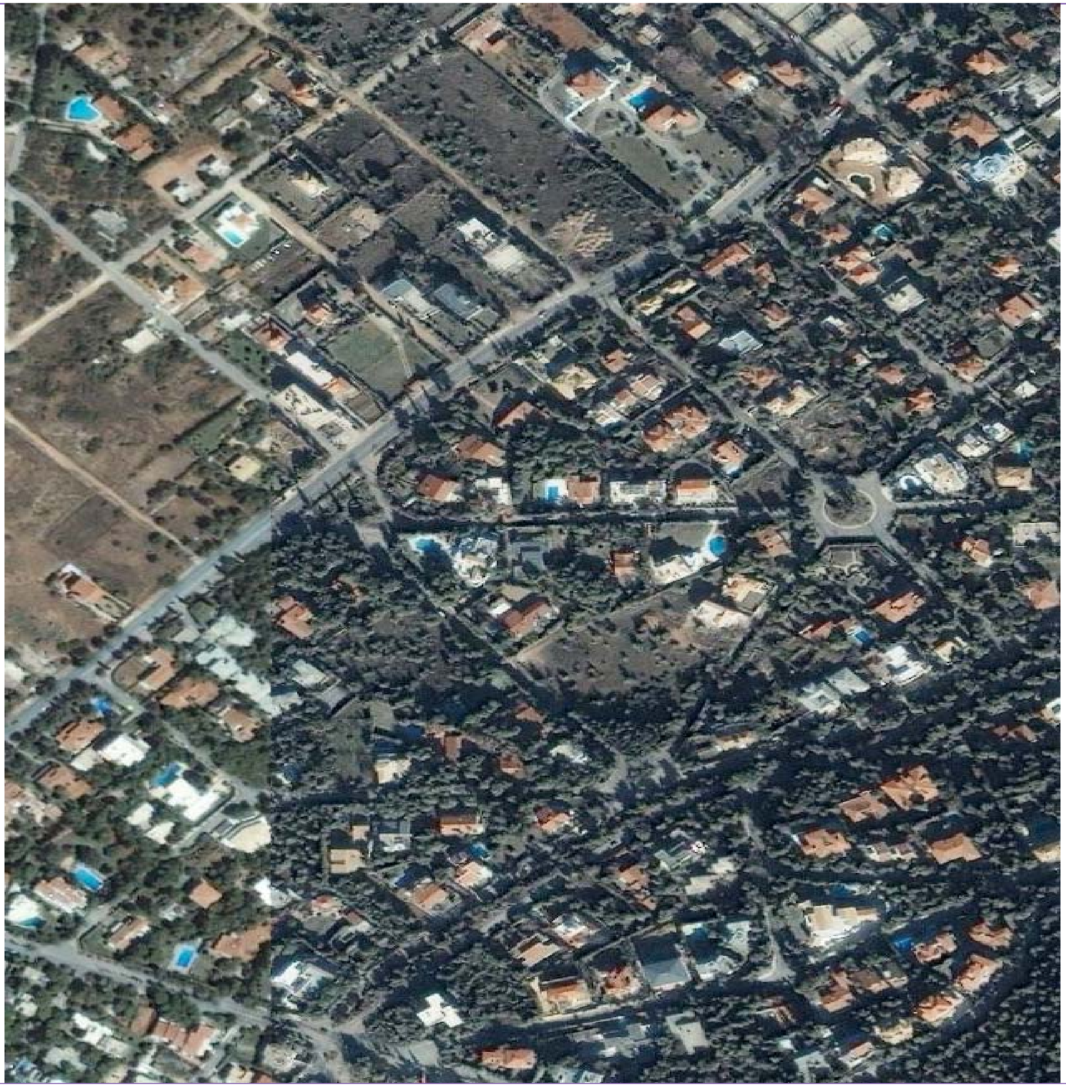
Data Structures
Image

DTS

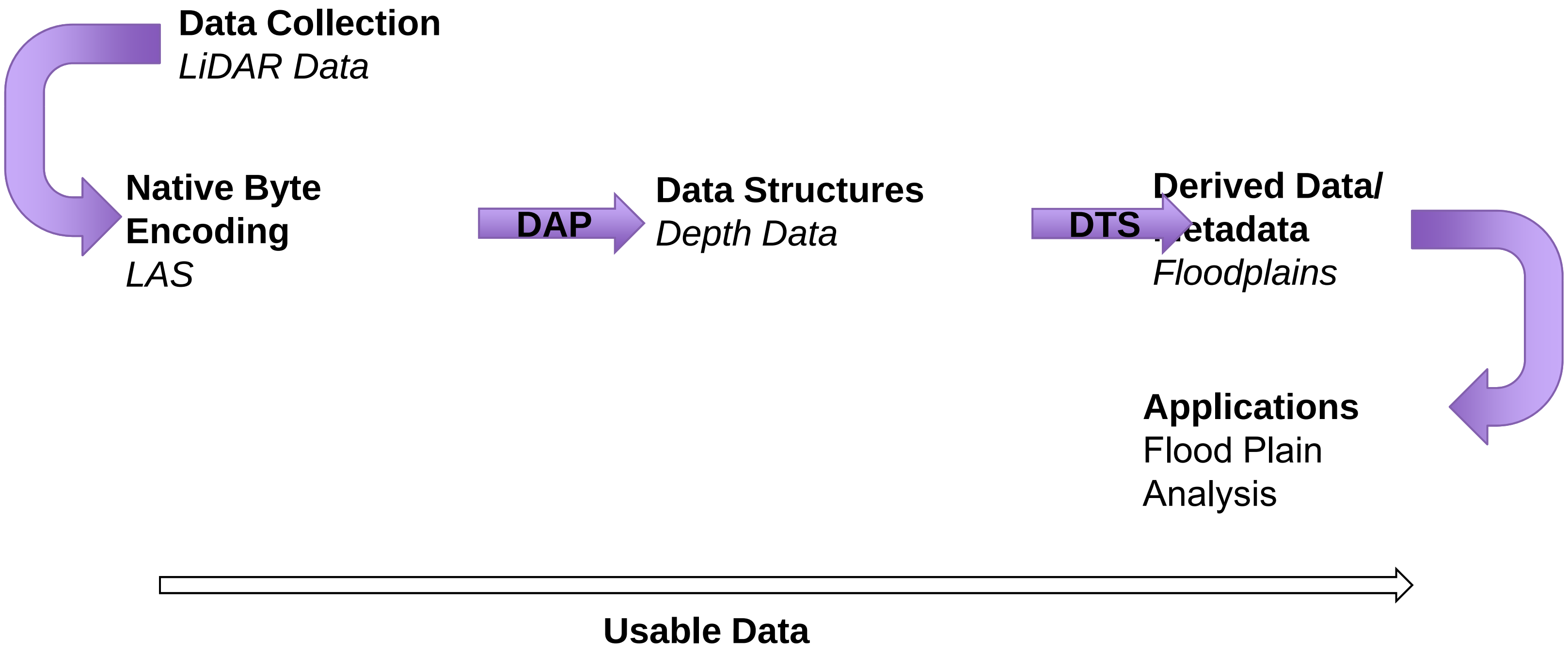
Derived Data/ Metadata
Land Cover/Usage/

Applications
Human Access to Green Infrastructure

Usable Data



P. Kumar, Civil & Env. Engineering, University of Illinois Urbana-Champaign



Data Collection
LiDAR Data

Native Byte Encoding
LAS

DAP

Data Structures
Depth Data

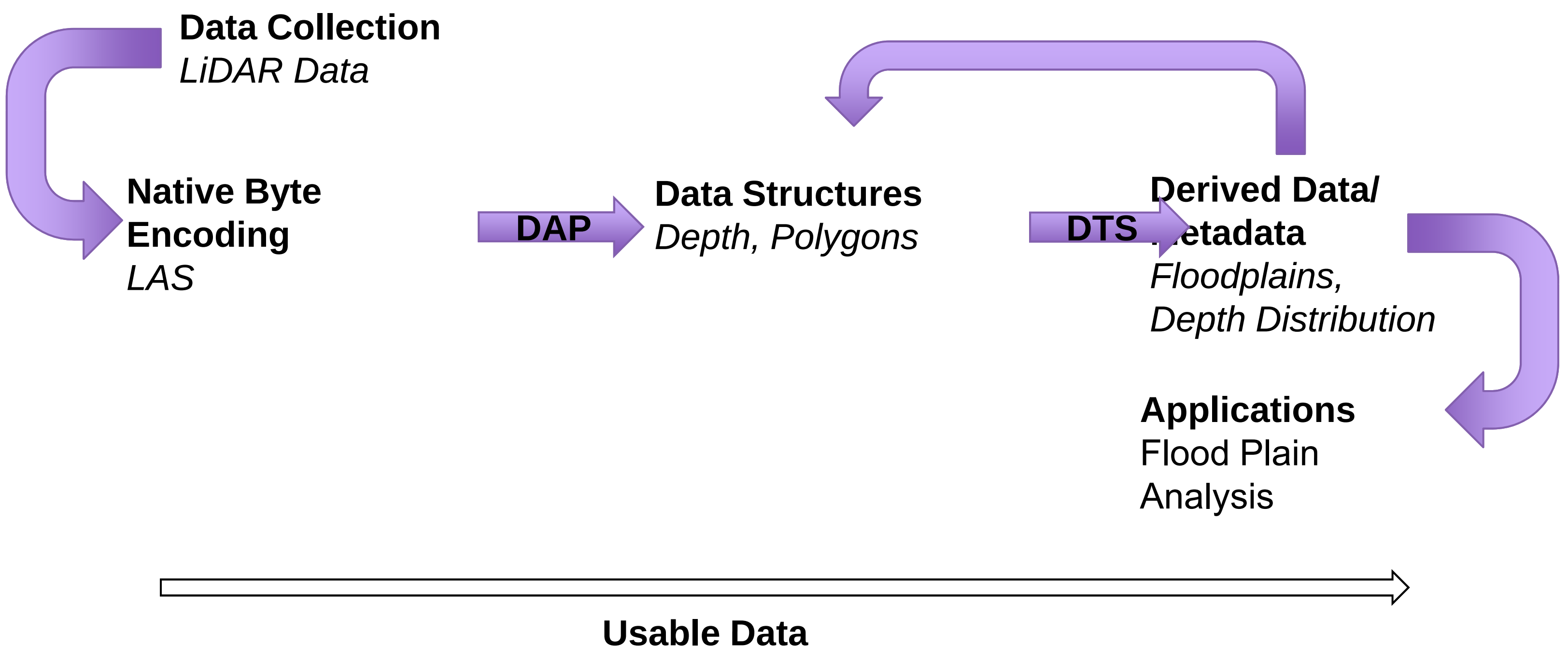
DTS

Derived Data/ Metadata
Floodplains

Applications
Flood Plain Analysis

Usable Data

P. Kumar, Civil & Env. Engineering, University of Illinois Urbana-Champaign



Data Collection
LiDAR Data

Native Byte Encoding
LAS

DAP

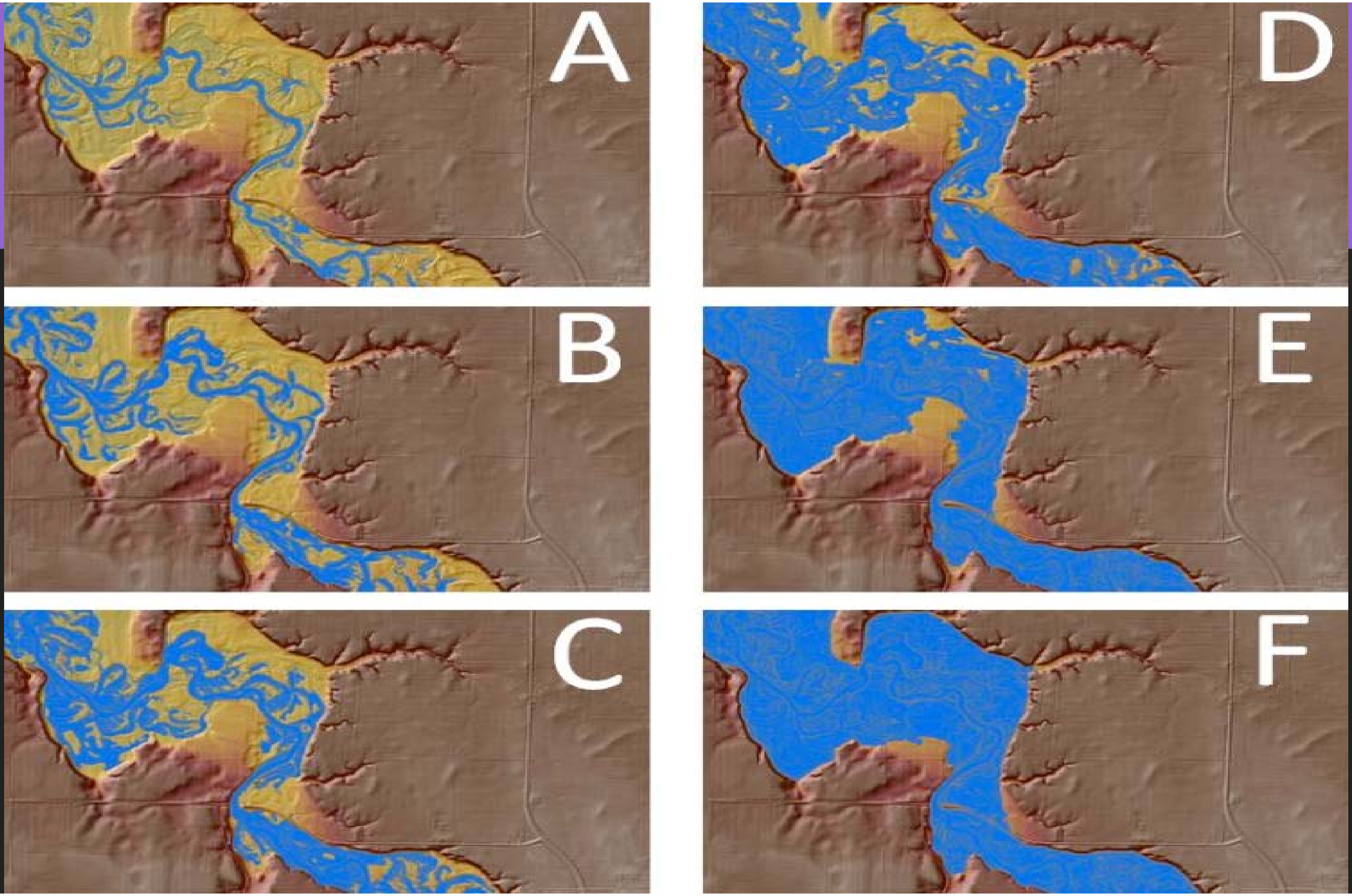
Data Structures
Depth, Polygons

DTS

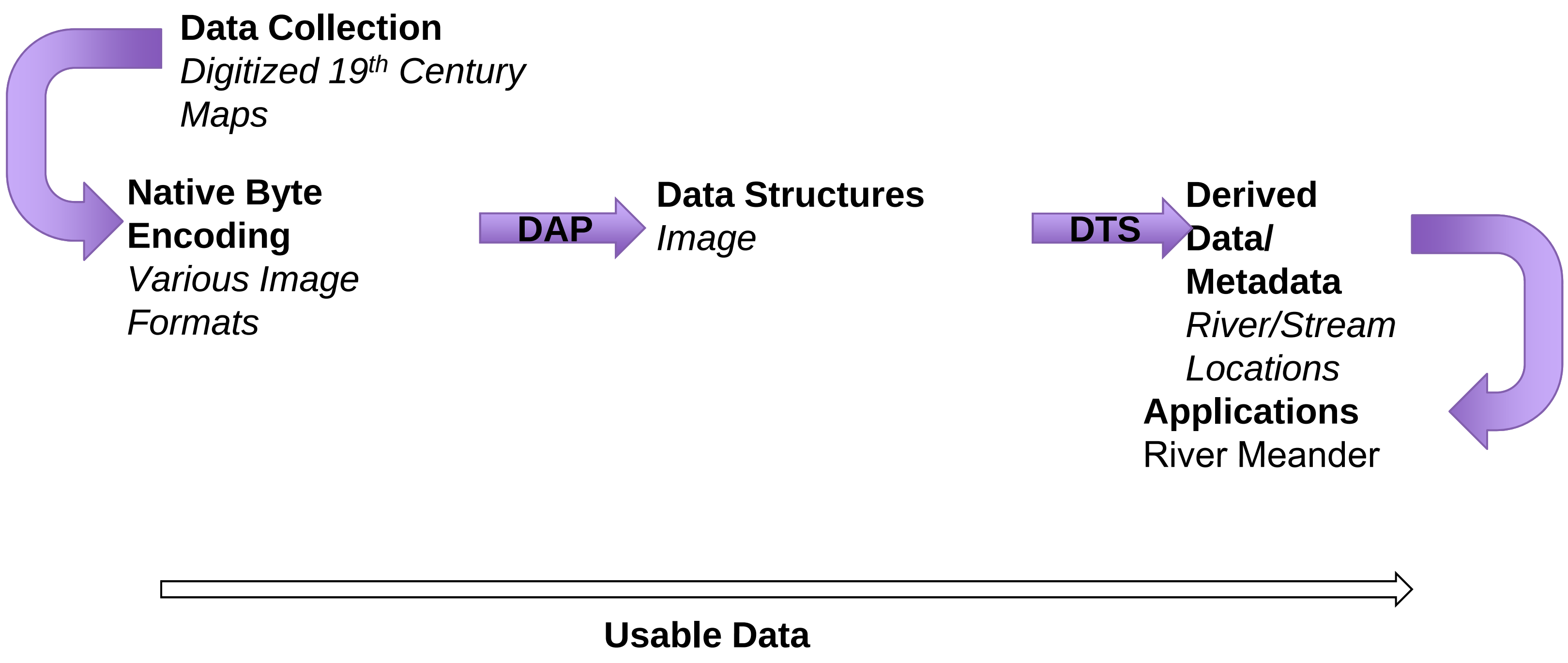
Derived Data/ Metadata
Floodplains, Depth Distribution

Applications
Flood Plain Analysis

Usable Data


A
B
C
D
E
F

P. Kumar, Civil & Env. Engineering, University of Illinois Urbana-Champaign



Data Collection
Digitized 19th Century Maps

Native Byte Encoding
Various Image Formats

DAP

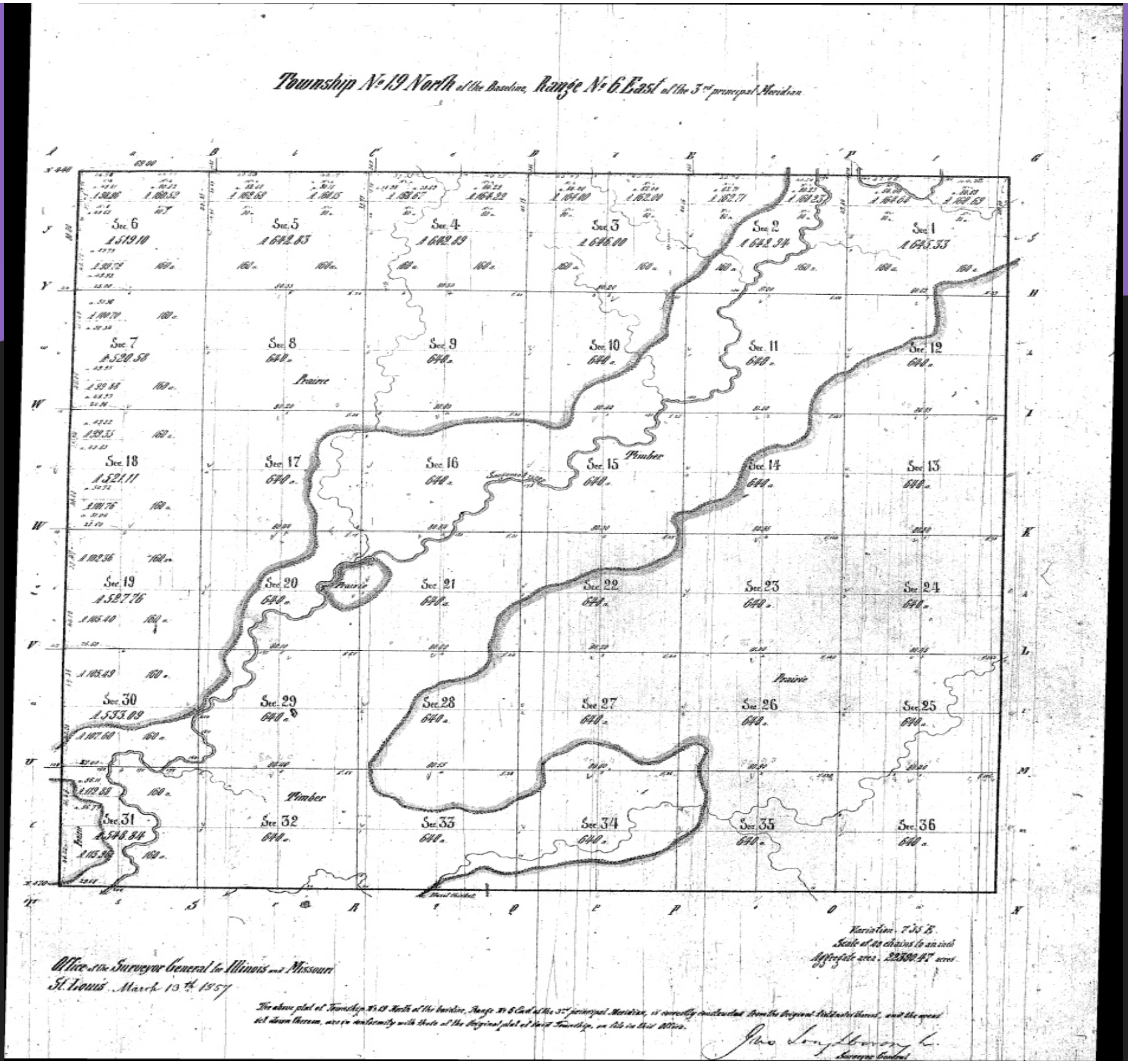
Data Structures
Image

DTS

Derived Data/ Metadata
River/Stream Locations

Applications
River Meander

Usable Data



Township No 19 North of the Baseline, Range No 6 East of the 3rd principal Meridian

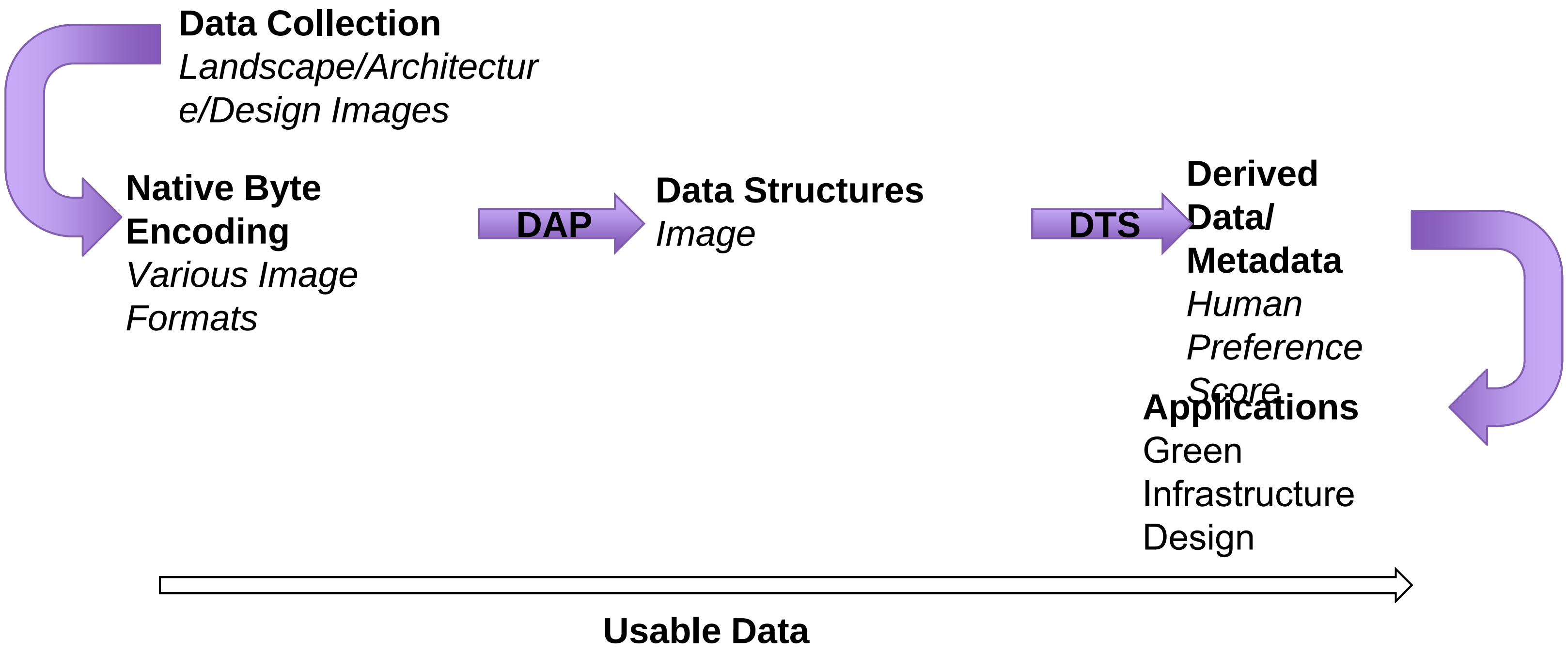
Variation 7°35' E.
Scale of 40 chains to an inch
Aggregate area: 22380.47 acres

Office of the Surveyor General for Illinois and Missouri
St. Louis March 13th 1857

The above plat of Township No 19 North of the baseline, Range No 6 East of the 3rd principal Meridian, is correctly constructed from the Original field notes thereof, and the areas set down thereon, are in conformity with those of the Original plat of said Township, on file in this Office.

Jno Loughborough
Surveyor General

B. Minsker, Civil & Env. Engineering, University of Illinois Urbana-Champaign
A. Schmidt, Civil & Env. Engineering, University of Illinois Urbana-Champaign
B. Sullivan, Landscape Architecture, University of Illinois Urbana-Champaign



Data Collection
Landscape/Architecture/Design Images

Native Byte Encoding
Various Image Formats

DAP

Data Structures
Image

DTS

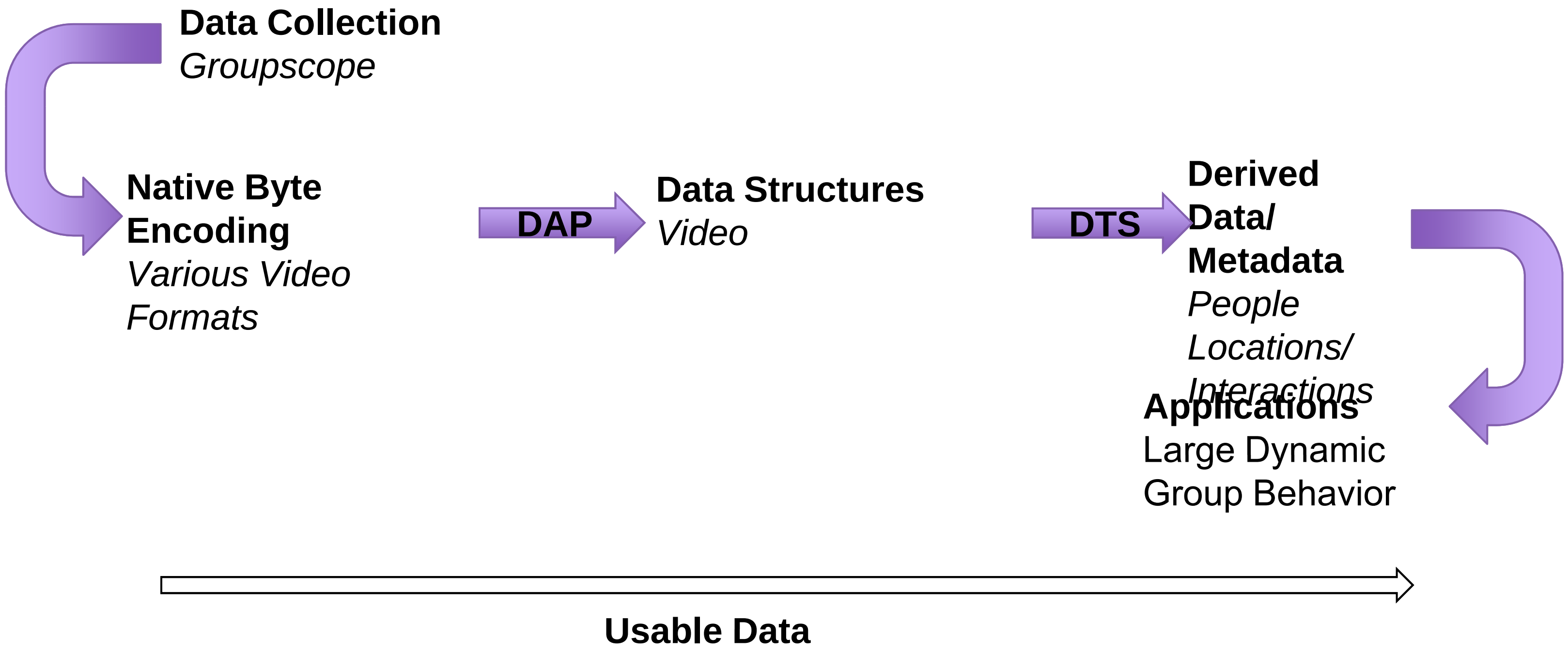
Derived Data/ Metadata
Human Preference Score

Applications
Green Infrastructure Design

Usable Data

BROADWAY
UNITED METHODIST CHURCH

M. Poole, Social Science, University of Illinois Urbana-Champaign
F. Pena-Mora, Civil & Env. Engineering, Columbia University
D. Espelage, Education, University of Illinois Urbana-Champaign



Data Collection
Groupscope

Native Byte Encoding
Various Video Formats

DAP

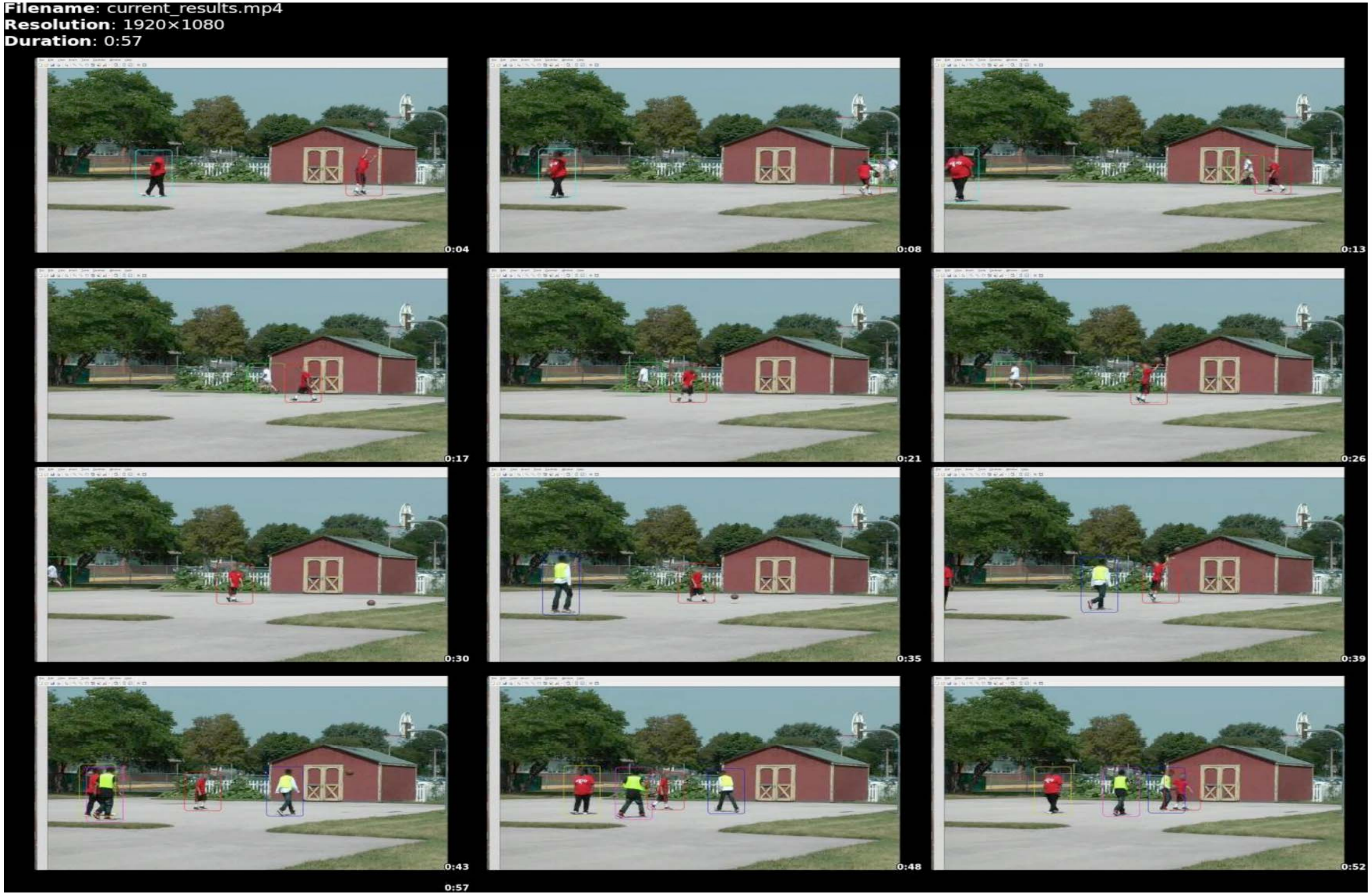
Data Structures
Video

DTS

Derived Data/ Metadata
People Locations/ Interactions

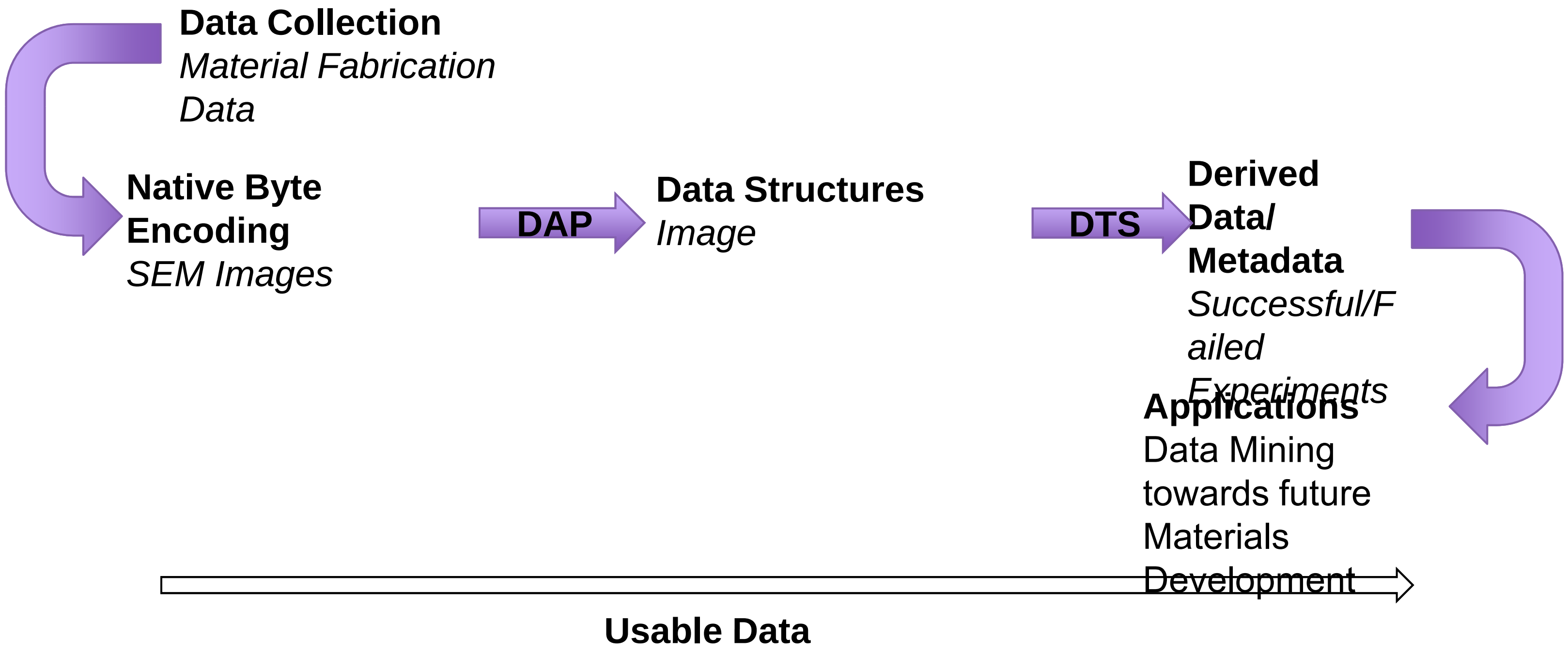
Applications
Large Dynamic Group Behavior

Usable Data



Filename: current_results.mp4
Resolution: 1920×1080
Duration: 0:57

K. Nahrstedt, Computer Science, University of Illinois Urbana-Champaign
J. Rogers, Material Science, University of Illinois Urbana-Champaign



Data Collection
Material Fabrication Data

Native Byte Encoding
SEM Images

DAP

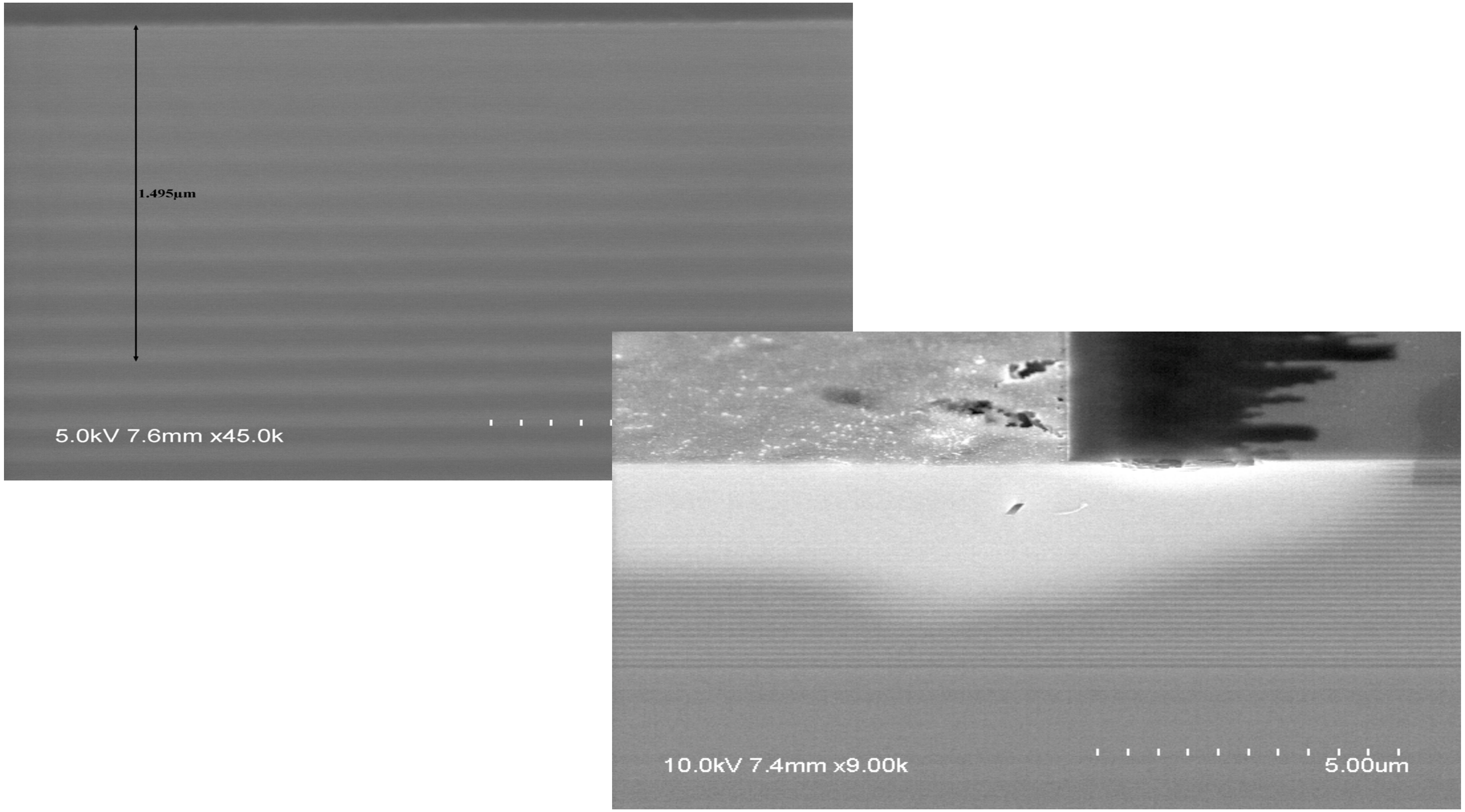
Data Structures
Image

DTS

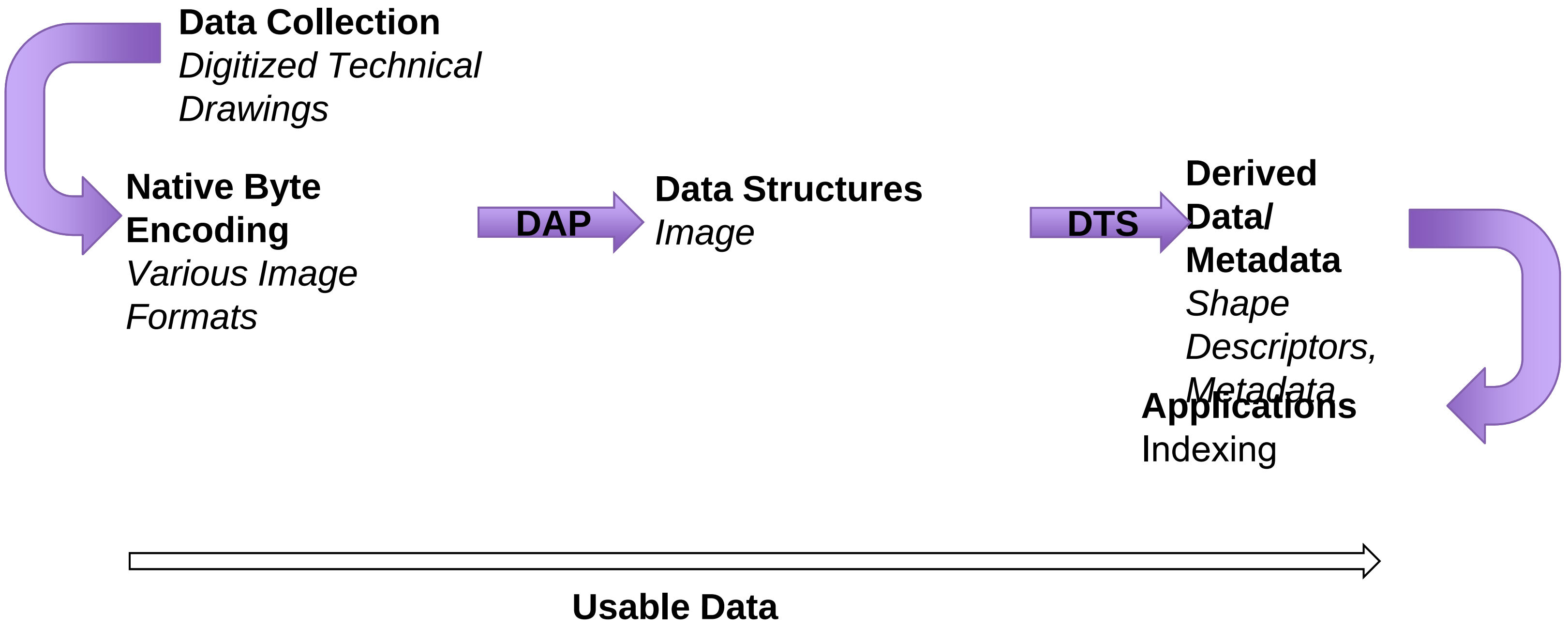
Derived Data/ Metadata
Successful/Failed Experiments

Applications
Data Mining towards future Materials Development

Usable Data


1.495µm
5.0kV 7.6mm x45.0k
10.0kV 7.4mm x9.00k
5.00um

Industry



Data Collection
Digitized Technical Drawings

Native Byte Encoding
Various Image Formats

DAP

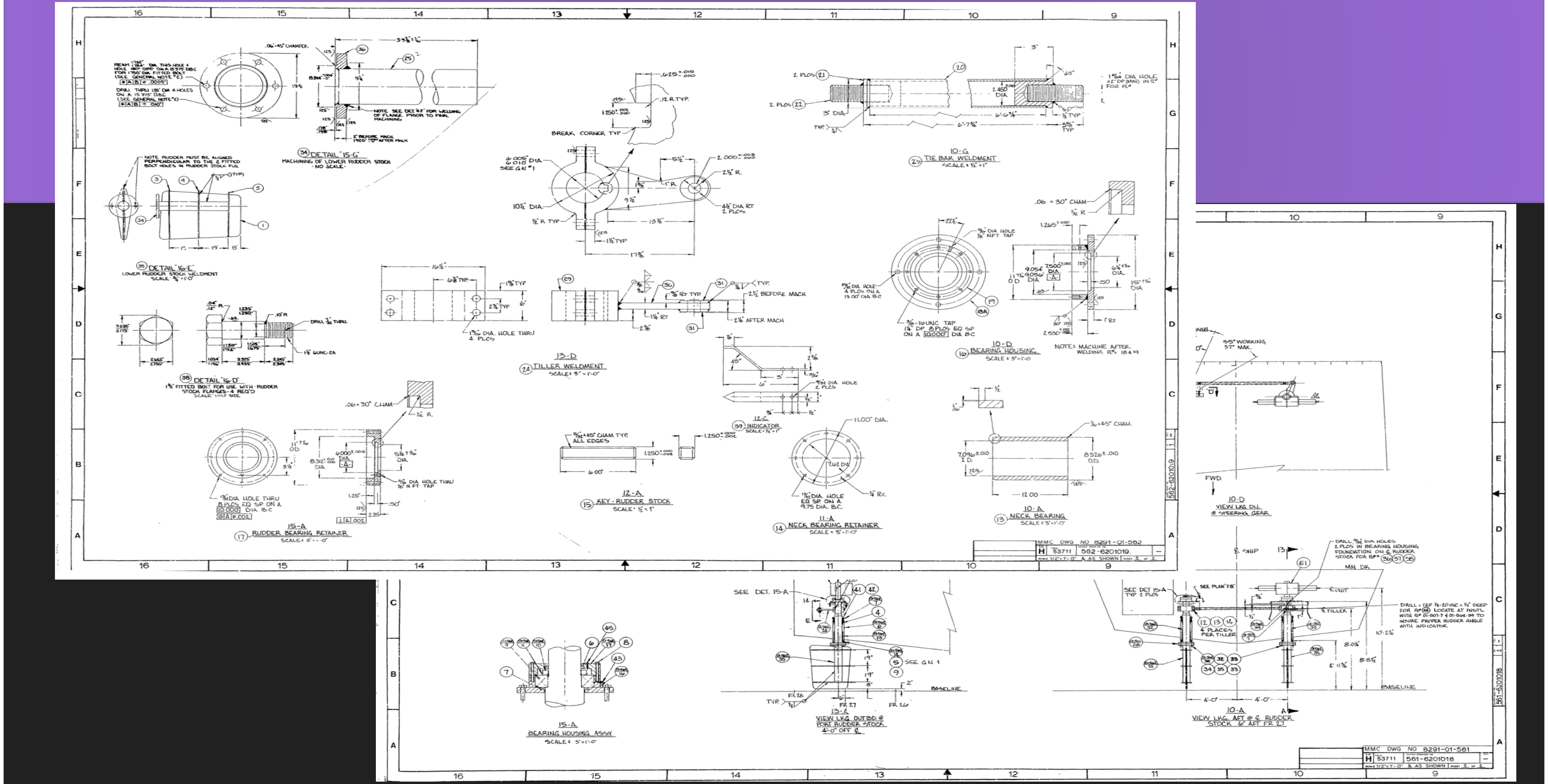
Data Structures
Image

DTS

Derived Data/ Metadata
Shape Descriptors, Metadata

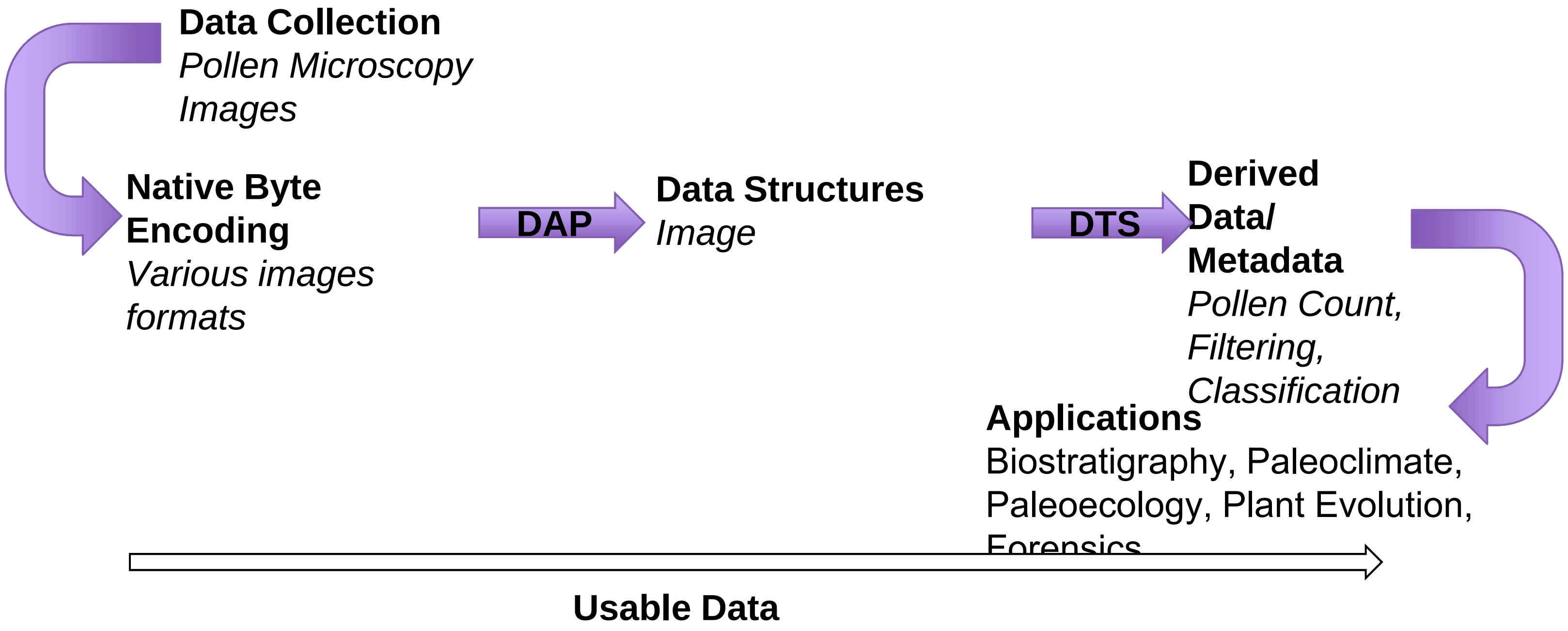
Applications
Indexing

Usable Data



DETAIL "15-G"
MACHINING OF LOWER RUDDER STOCK
NO SCALE
DETAIL "16-E"
LOWER RUDDER STOCK WELDMENT
DETAIL "16-D"
1⅜" FITTED BOLT FOR USE WITH RUDDER STOCK FLANGES - 4 REQ'D
BREAK CORNER TYP.
13-D
TILLER WELDMENT
SCALE: 3" = 1'-0"
10-G
TIE BAR WELDMENT
10-D
BEARING HOUSING
SCALE: 3" = 1'-0"
NOTE: MACHINE AFTER WELDING P/S 18 & 19
12-C
INDICATOR
15-A
RUDDER BEARING RETAINER
SCALE: 6" = 1'-0"
12-A
KEY - RUDDER STOCK
SCALE: ½" = 1"
11-A
NECK BEARING RETAINER
SCALE: 3" = 1'-0"
10-A
NECK BEARING
SCALE: 3" = 1'-0"
MMC DWG NO 8291-01-562
53711 562-6201019
15-A
BEARING HOUSING ASSY
SCALE: 3" = 1'-0"
13-A
VIEW LKG OUTBD @ PORT RUDDER STOCK 4'-0" OFF ℄
10-A
VIEW LKG AFT @ ℄ RUDDER STOCK 6" AFT FR 27
10-D
VIEW LKG DN @ STEERING GEAR
35° WORKING 37° MAX.
BASELINE
MN DK.
MMC DWG NO 8291-01-561
53711 561-6201018

S. Punyasena, Biology, University of Illinois Urbana-Champaign



Data Collection
Pollen Microscopy Images

Native Byte Encoding
Various images formats

DAP

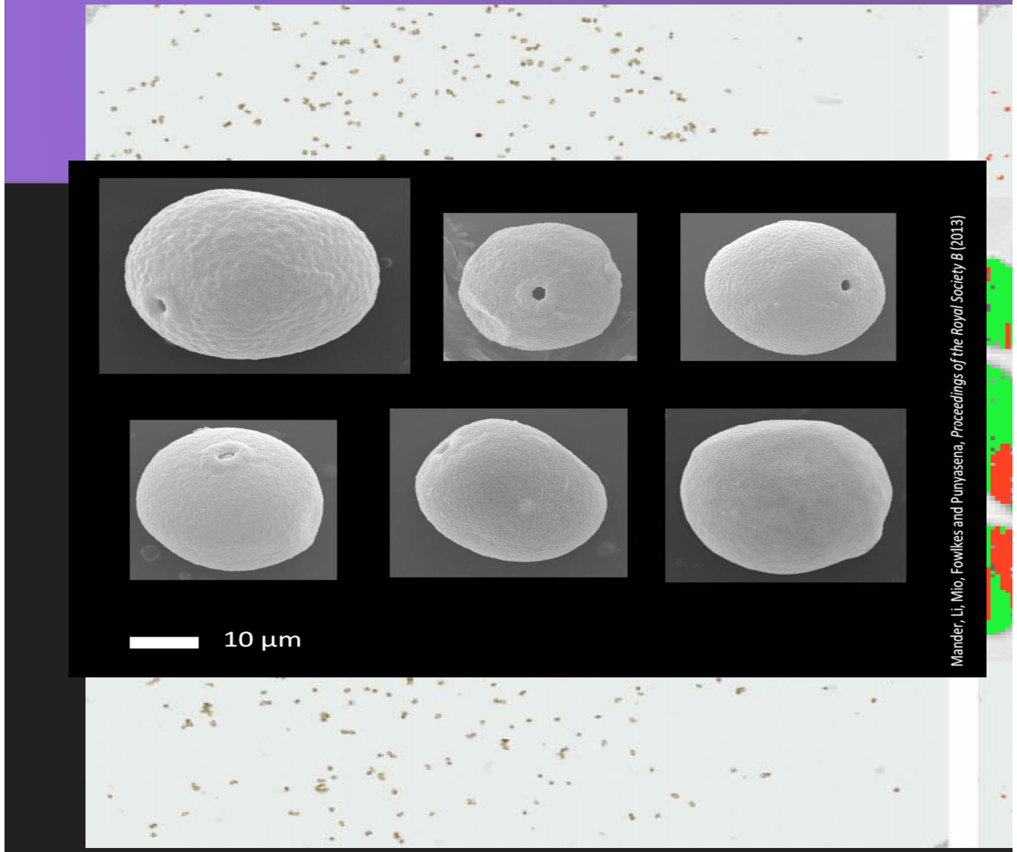
Data Structures
Image

DTS

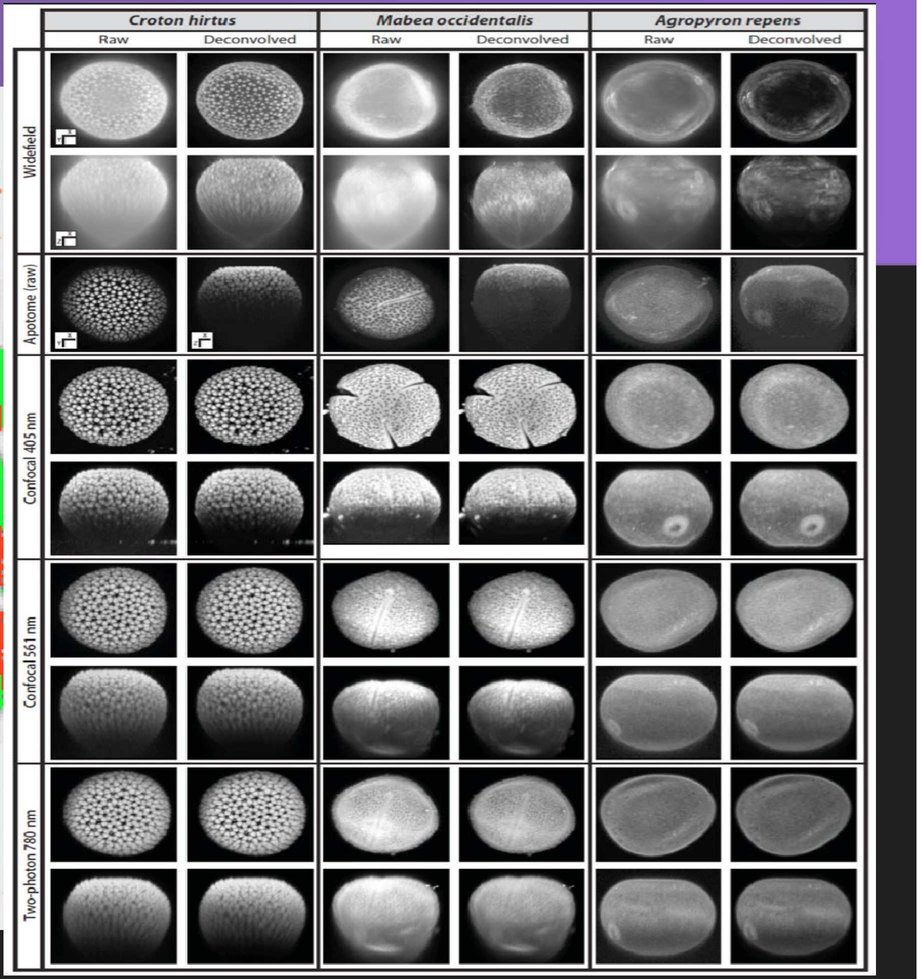
Derived Data/ Metadata
Pollen Count, Filtering, Classification

Applications
Biostratigraphy, Paleoclimate, Paleoecology, Plant Evolution, Forensics

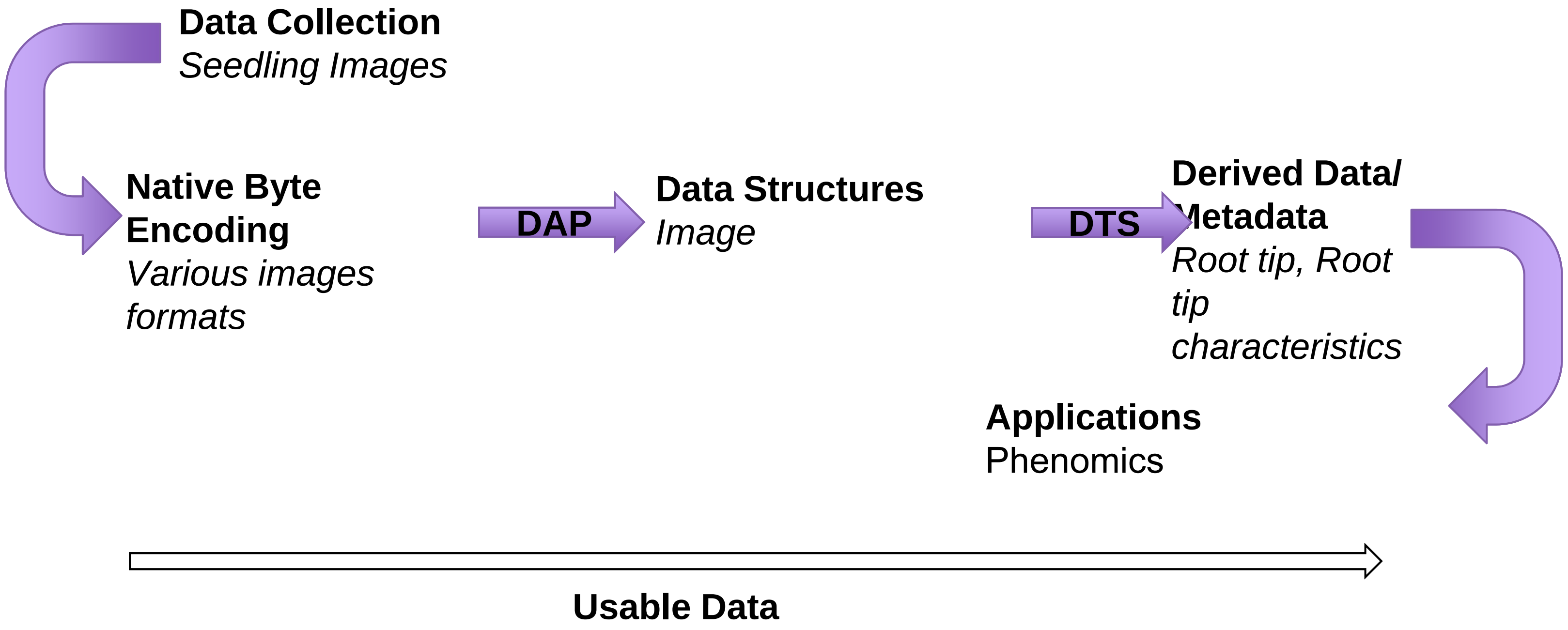
Usable Data



Mander, Li, Mio, Fowlkes and Punyasena, *Proceedings of the Royal Society B* (2013)



E. Spalding, Botany, University of Wisconsin



Data Collection
Seedling Images

Native Byte Encoding
Various images formats

DAP

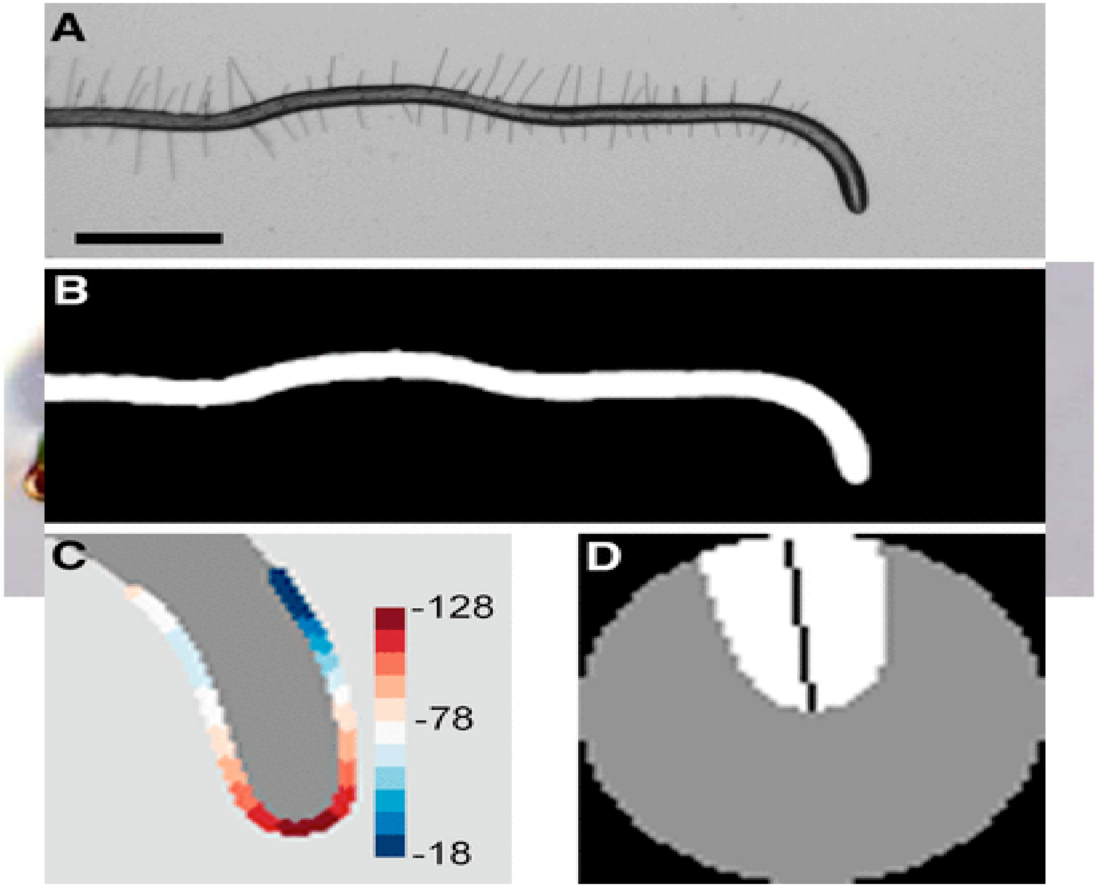
Data Structures
Image

DTS

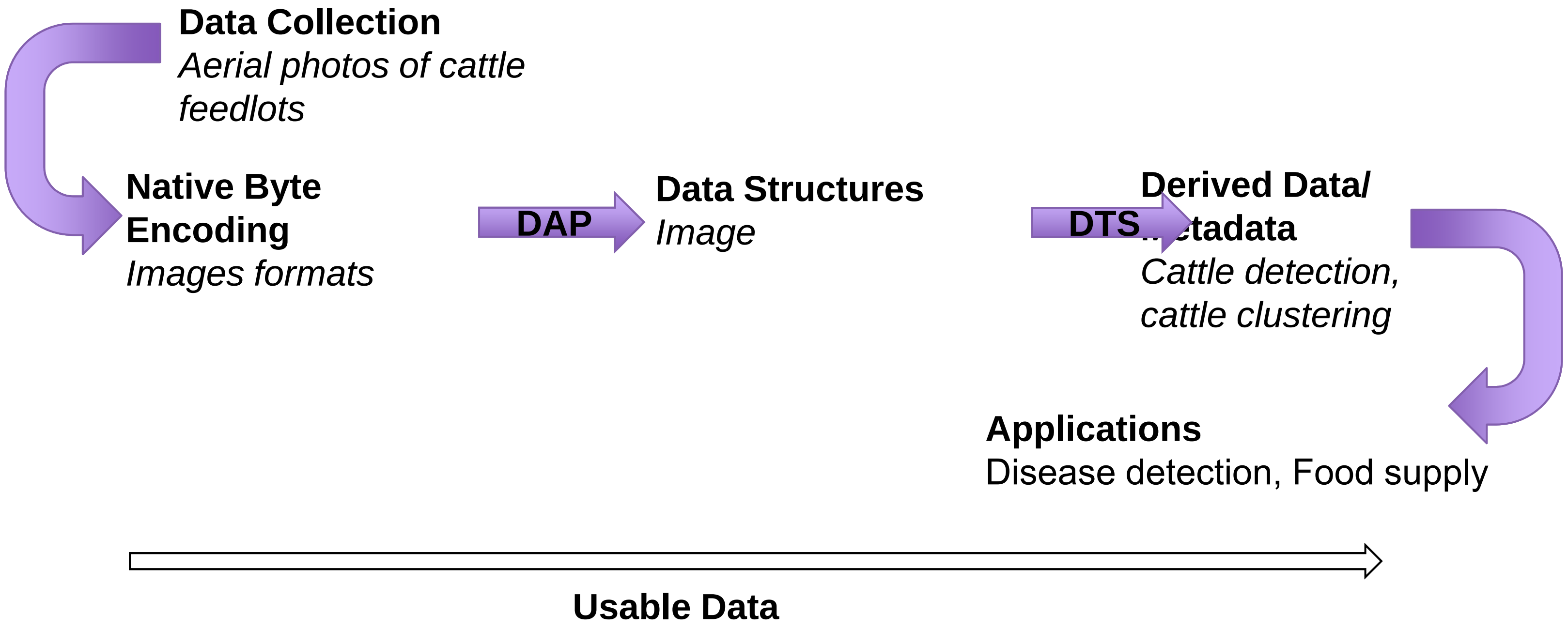
Derived Data/ Metadata
Root tip, Root tip characteristics

Applications
Phenomics

Usable Data


A
B
C
-128
-78
-18
D

Stephan Joslyn, Veterinary Medicine, University of Illinois Urbana-Champaign



Data Collection
Aerial photos of cattle feedlots

Native Byte Encoding
Images formats

DAP

Data Structures
Image

DTS

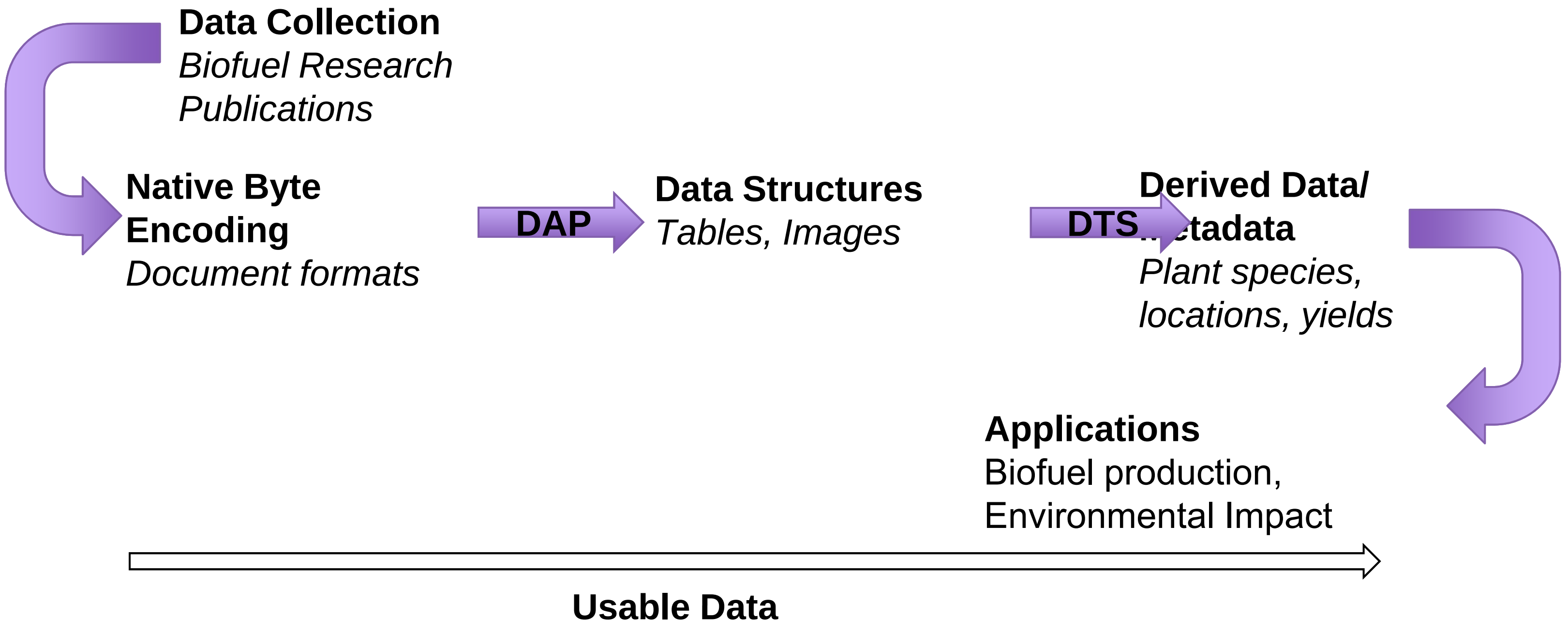
Derived Data/ Metadata
Cattle detection, cattle clustering

Applications
Disease detection, Food supply

Usable Data

1944/07/
1944/07/13 0

David LeBauer, Ecology, University of Illinois Urbana-Champaign



Data Collection
Biofuel Research Publications

Native Byte Encoding
Document formats

DAP

Data Structures
Tables, Images

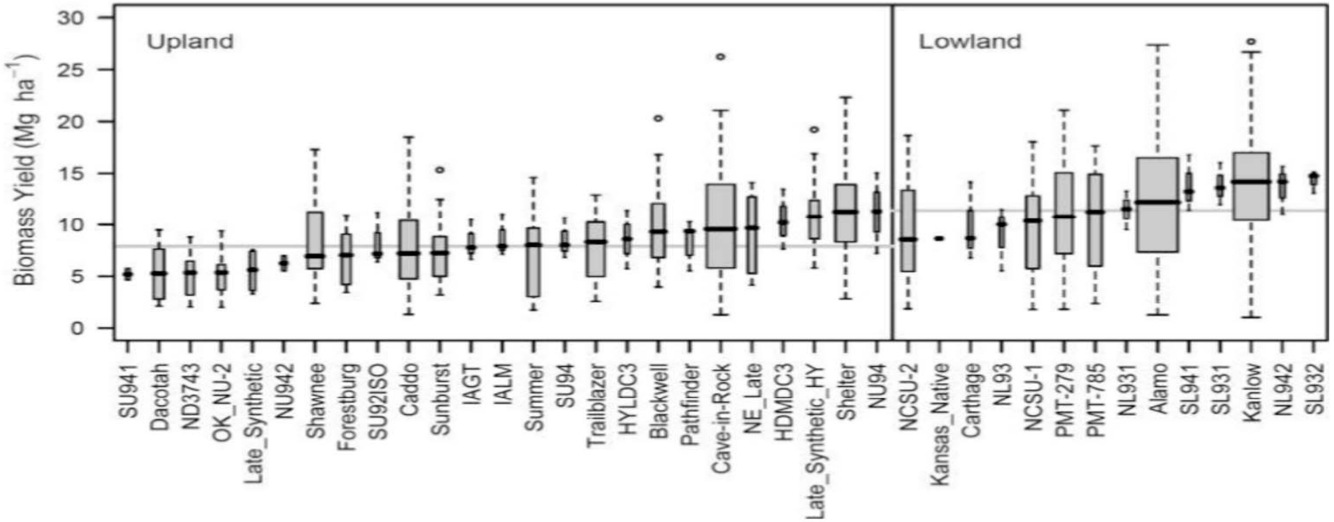
DTS

Derived Data/ Metadata
Plant species, locations, yields

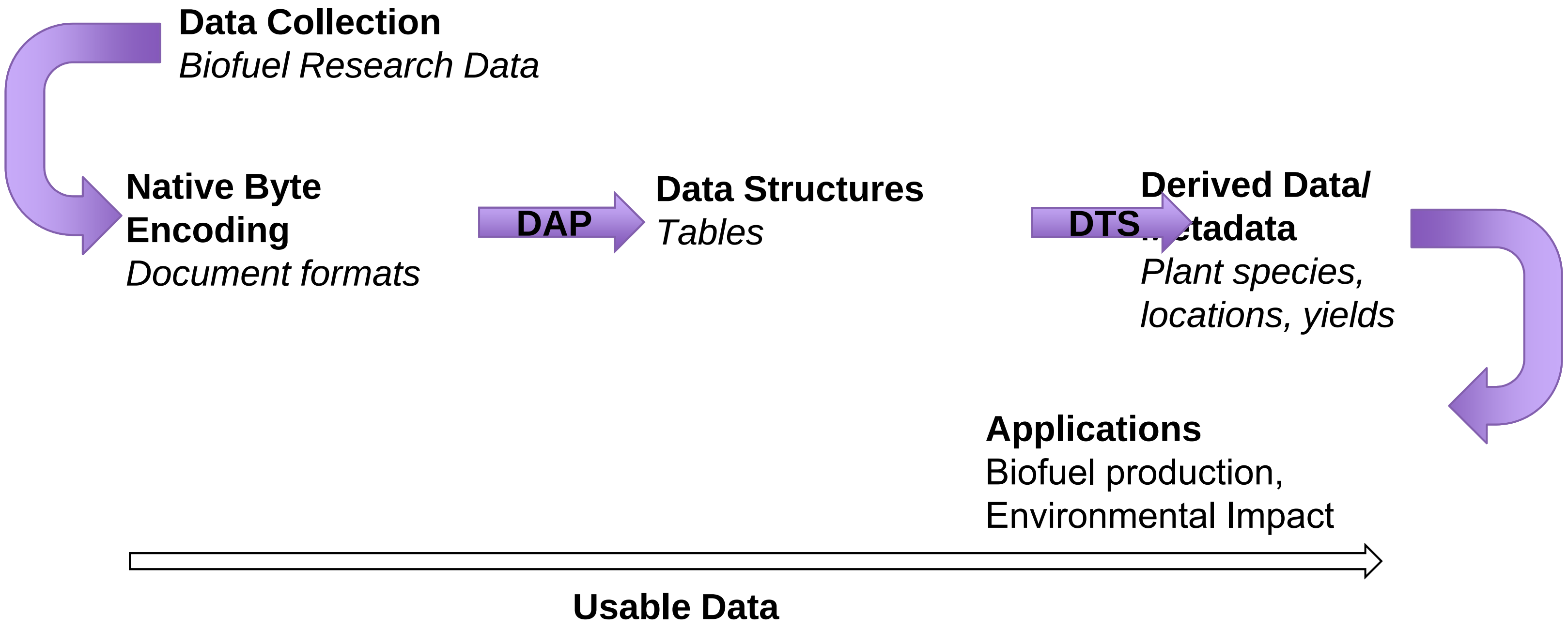
Applications
Biofuel production, Environmental Impact

Usable Data

Indicators	N_0	N_{60}	N_{120}	LSD_{05}	N_0	N_{60}	N_{120}	LSD_{05}
				Second year of growth				
				Annual biomass				
t ha^{-1}	15.8	20.0	24.7*	5.63	4.62	6.55*	6.82*	1.85
%	100	127	157	40.5	100	142	148	42.9
				Biomass weight per plant				
kg	1.66	2.26*	2.53*	0.627	0.49	0.74*	0.70*	0.208
				Third year of growth				
				Annual biomass				
t ha^{-1}	27.0	28.5	29.7	5.31	10.5	10.7	11.5	2.47
%	100	105.6	110.1	18.68	100	102.3	110.2	22.61
				Biomass weight per plant				
kg	2.05	2.18	2.25	0.396	0.79	0.81	0.87	0.187



David LeBauer, Ecology, University of Illinois Urbana-Champaign



Data Collection
Biofuel Research Data

Native Byte Encoding
Document formats

DAP

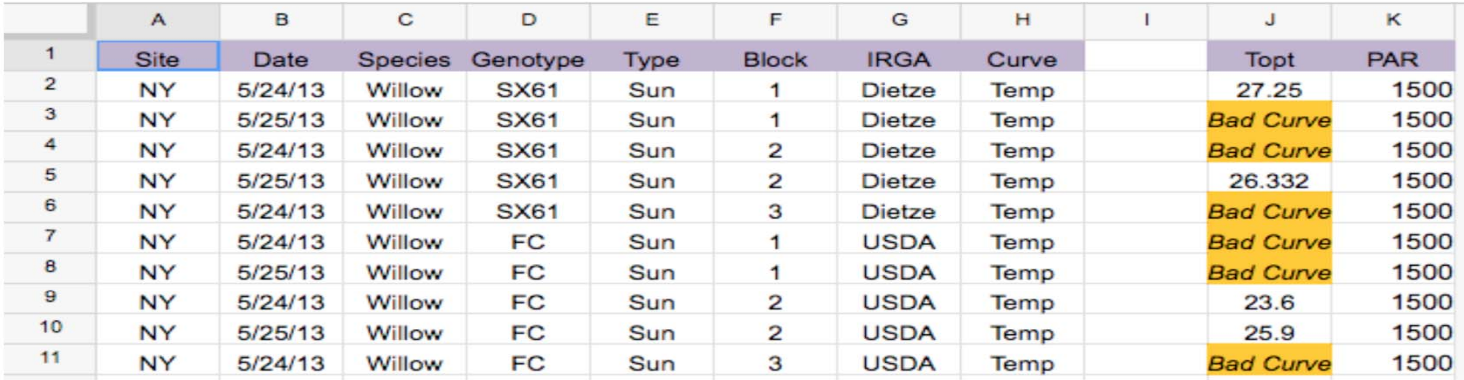
Data Structures
Tables

DTS

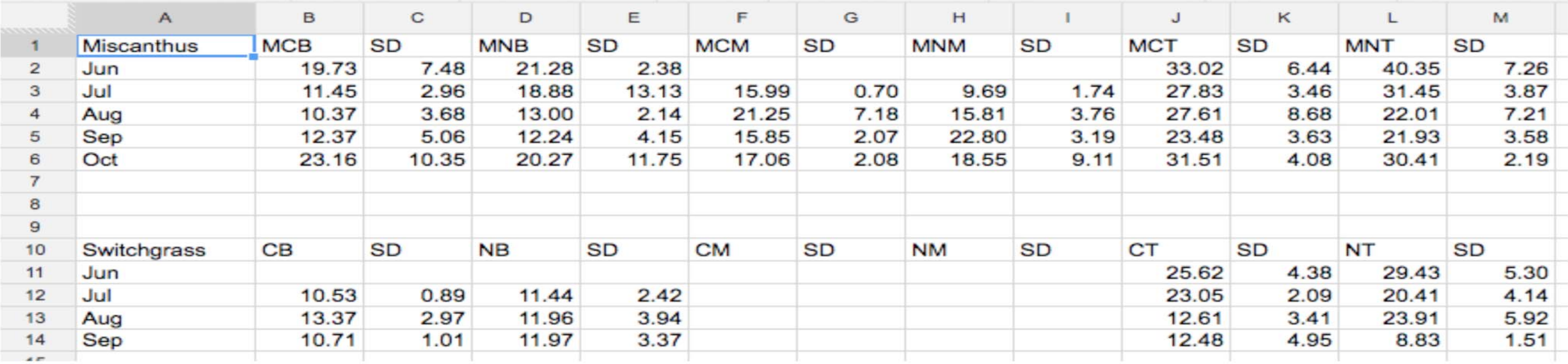
Derived Data/ Metadata
Plant species, locations, yields

Applications
Biofuel production, Environmental Impact

Usable Data

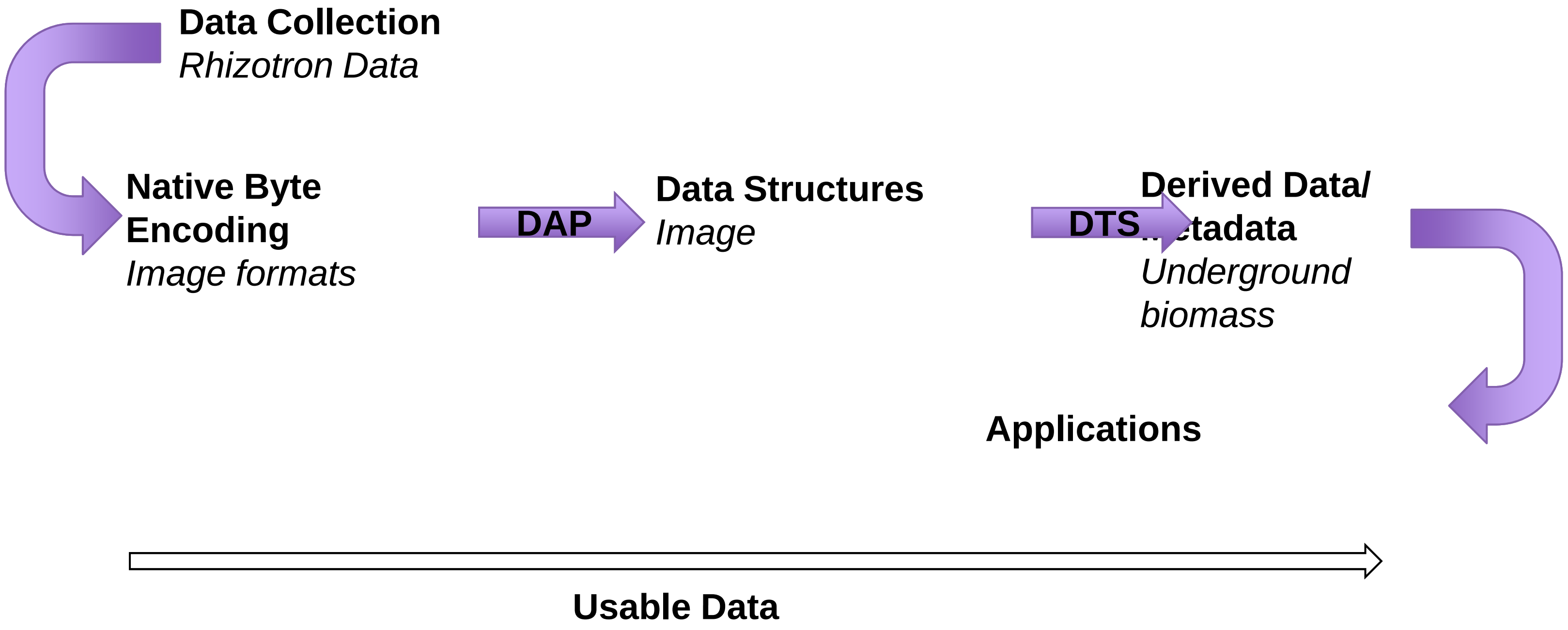


	A	B	C	D	E	F	G	H	I	J	K
1	Site	Date	Species	Genotype	Type	Block	IRGA	Curve		Topt	PAR
2	NY	5/24/13	Willow	SX61	Sun	1	Dietze	Temp		27.25	1500
3	NY	5/25/13	Willow	SX61	Sun	1	Dietze	Temp		*Bad Curve*	1500
4	NY	5/24/13	Willow	SX61	Sun	2	Dietze	Temp		*Bad Curve*	1500
5	NY	5/25/13	Willow	SX61	Sun	2	Dietze	Temp		26.332	1500
6	NY	5/24/13	Willow	SX61	Sun	3	Dietze	Temp		*Bad Curve*	1500
7	NY	5/24/13	Willow	FC	Sun	1	USDA	Temp		*Bad Curve*	1500
8	NY	5/25/13	Willow	FC	Sun	1	USDA	Temp		*Bad Curve*	1500
9	NY	5/24/13	Willow	FC	Sun	2	USDA	Temp		23.6	1500
10	NY	5/25/13	Willow	FC	Sun	2	USDA	Temp		25.9	1500
11	NY	5/24/13	Willow	FC	Sun	3	USDA	Temp		*Bad Curve*	1500



	A	B	C	D	E	F	G	H	I	J	K	L	M
1	Miscanthus	MCB	SD	MNB	SD	MCM	SD	MNM	SD	MCT	SD	MNT	SD
2	Jun	19.73	7.48	21.28	2.38					33.02	6.44	40.35	7.26
3	Jul	11.45	2.96	18.88	13.13	15.99	0.70	9.69	1.74	27.83	3.46	31.45	3.87
4	Aug	10.37	3.68	13.00	2.14	21.25	7.18	15.81	3.76	27.61	8.68	22.01	7.21
5	Sep	12.37	5.06	12.24	4.15	15.85	2.07	22.80	3.19	23.48	3.63	21.93	3.58
6	Oct	23.16	10.35	20.27	11.75	17.06	2.08	18.55	9.11	31.51	4.08	30.41	2.19
7													
8													
9													
10	Switchgrass	CB	SD	NB	SD	CM	SD	NM	SD	CT	SD	NT	SD
11	Jun									25.62	4.38	29.43	5.30
12	Jul	10.53	0.89	11.44	2.42					23.05	2.09	20.41	4.14
13	Aug	13.37	2.97	11.96	3.94					12.61	3.41	23.91	5.92
14	Sep	10.71	1.01	11.97	3.37					12.48	4.95	8.83	1.51

Brian Wee, Ecology, National Ecological Observatory Network (NEON)



Data Collection
Rhizotron Data

Native Byte Encoding
Image formats

DAP

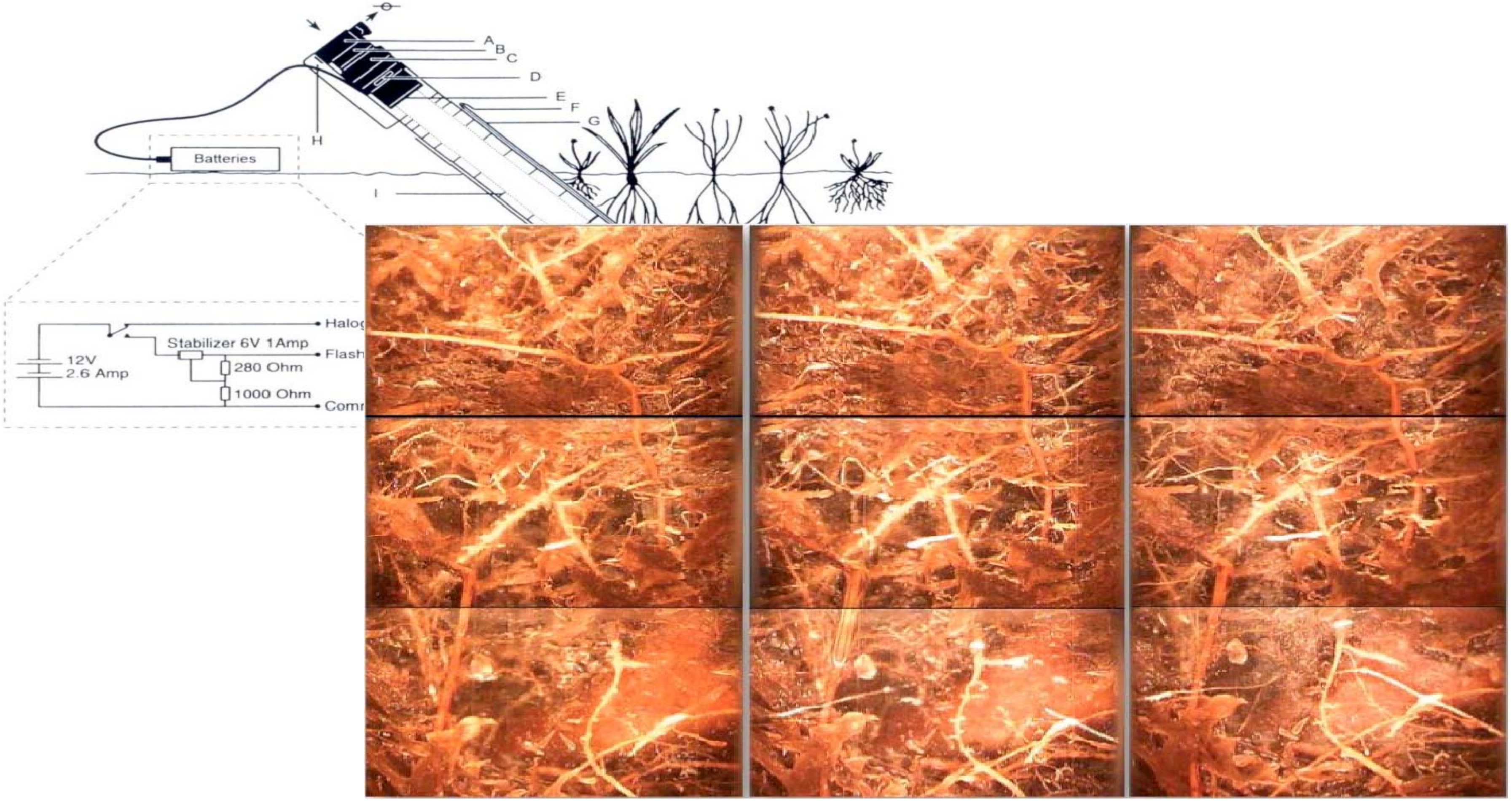
Data Structures
Image

DTS

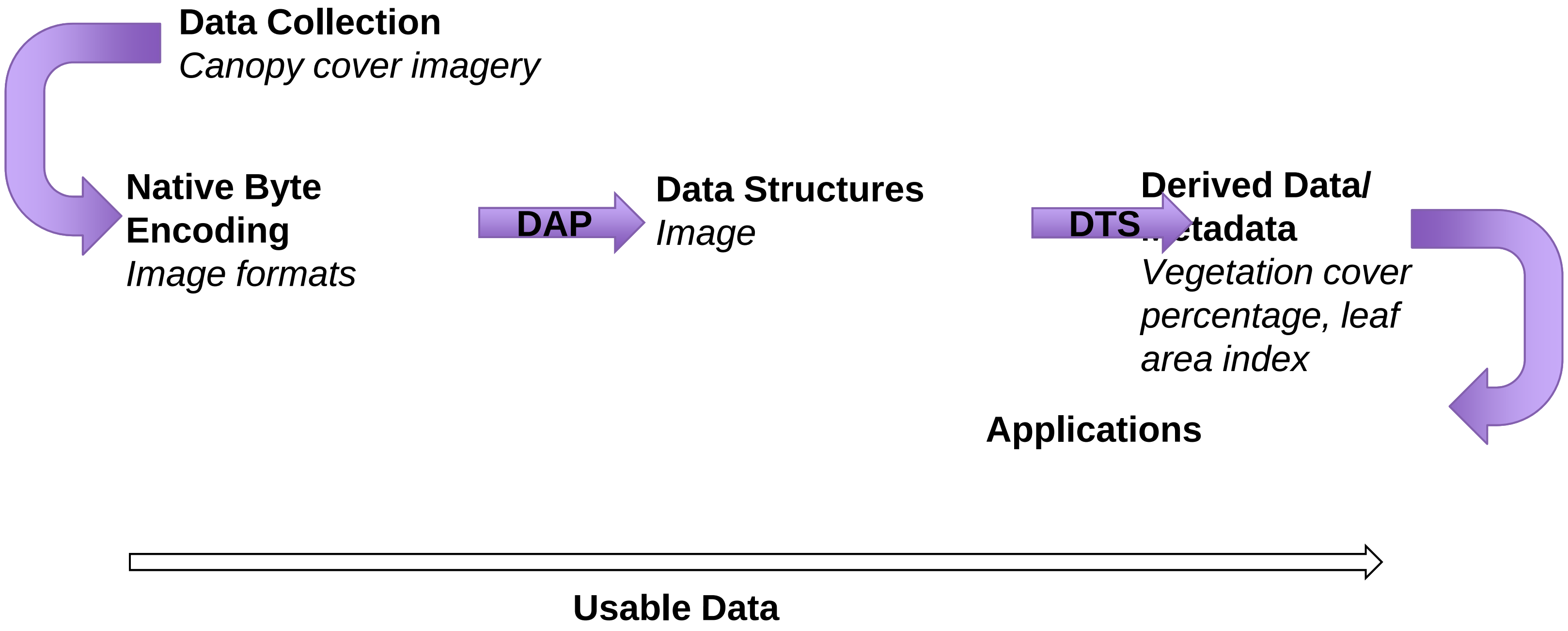
Derived Data/ Metadata
Underground biomass

Applications

Usable Data


A
B
C
D
E
F
G
H
I
Batteries
12V
2.6 Amp
Stabilizer 6V 1Amp
280 Ohm
1000 Ohm
Halog
Flash
Comr

Brian Wee, Ecology, National Ecological Observatory Network (NEON)



Data Collection
Canopy cover imagery

Native Byte Encoding
Image formats

DAP

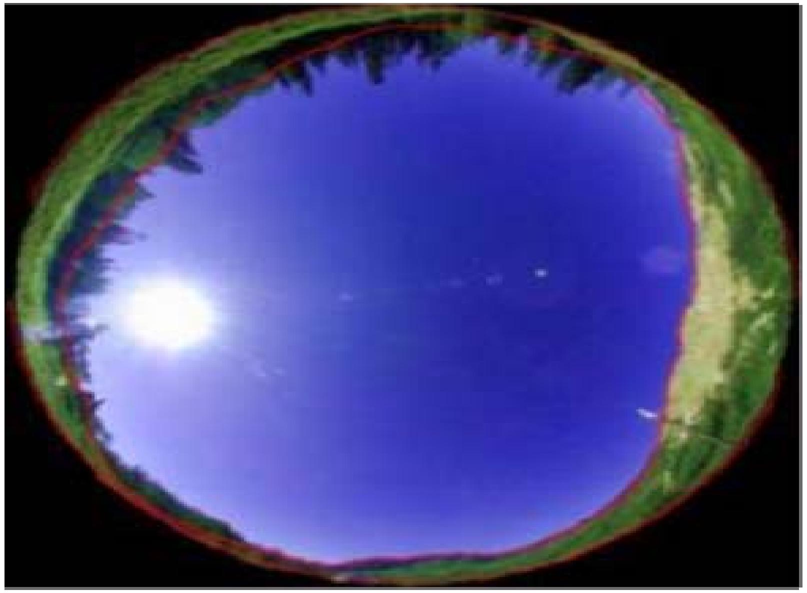
Data Structures
Image

DTS

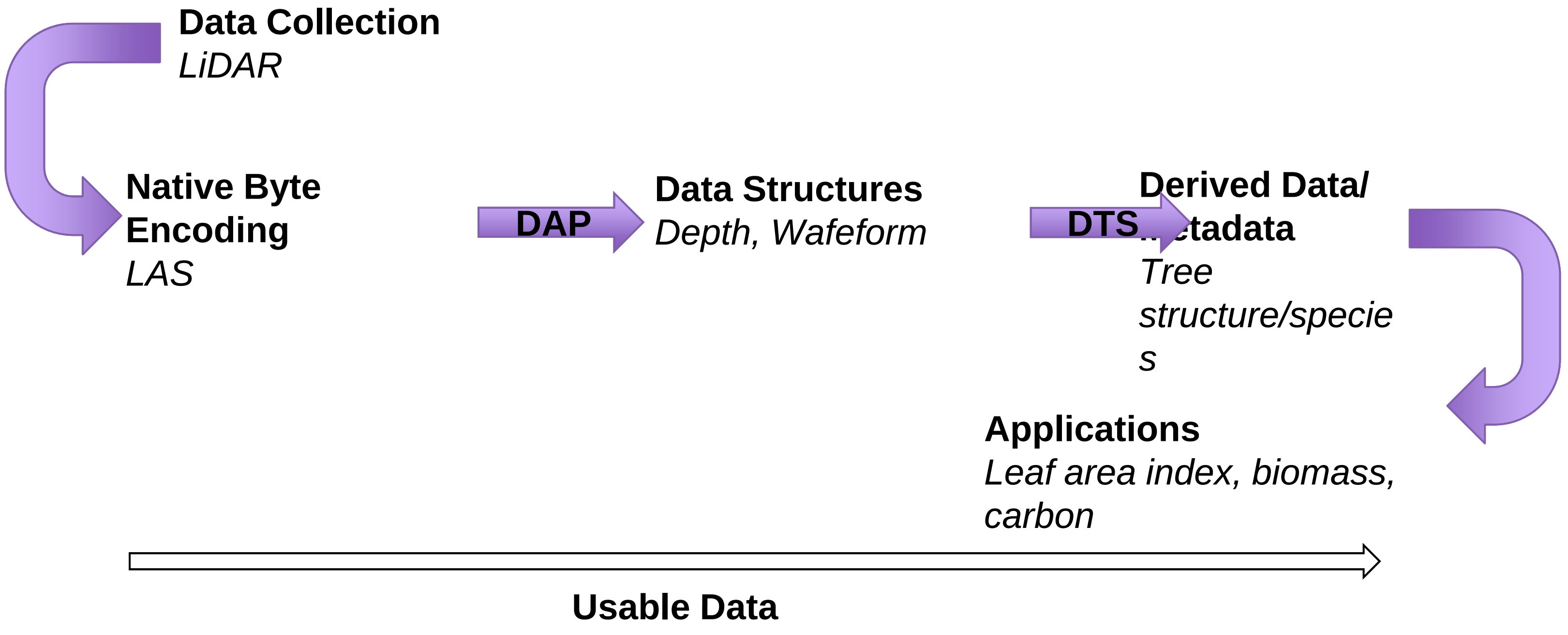
Derived Data/ Metadata
Vegetation cover percentage, leaf area index

Applications

Usable Data



Brian Wee, Ecology, National Ecological Observatory Network (NEON)



Data Collection
LiDAR

Native Byte Encoding
LAS

DAP

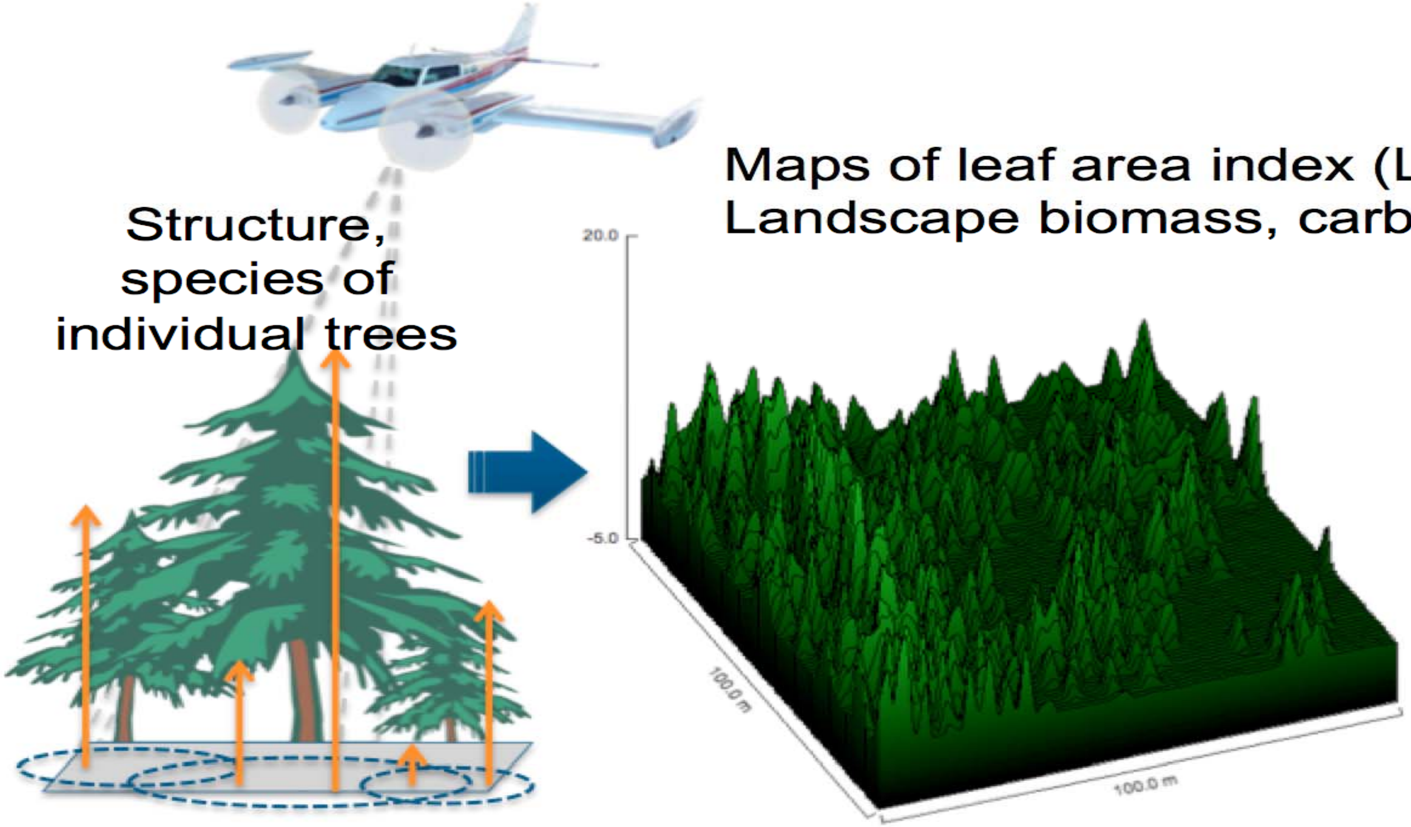
Data Structures
Depth, Wafeform

DTS

Derived Data/ Metadata
Tree structure/species

Applications
Leaf area index, biomass, carbon

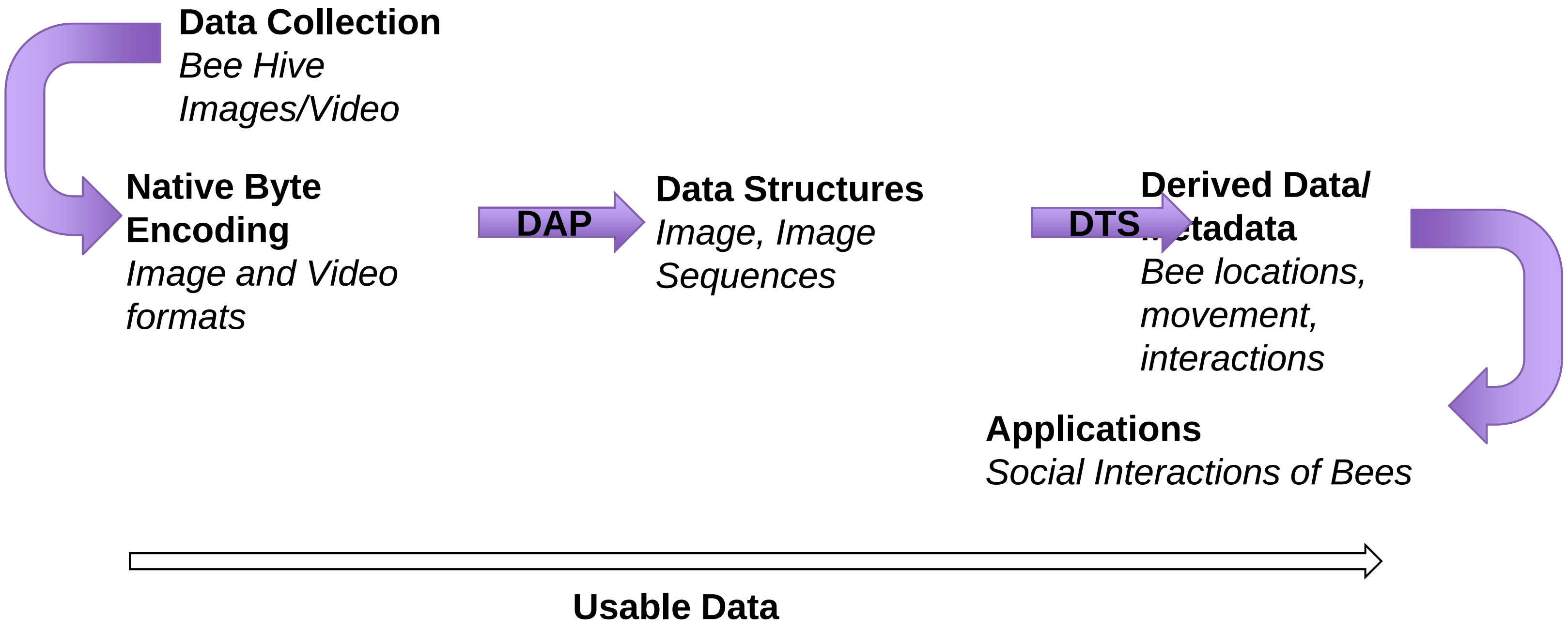
Usable Data



Structure, species of individual trees

Maps of leaf area index (LAI); Landscape biomass, carbon

Tim Gernat, Entomology, University of Illinois Urbana-Champaign
Gene Robinson, Entomology, University of Illinois Urbana-Champaign



Data Collection
Bee Hive Images/Video

Native Byte Encoding
Image and Video formats

DAP

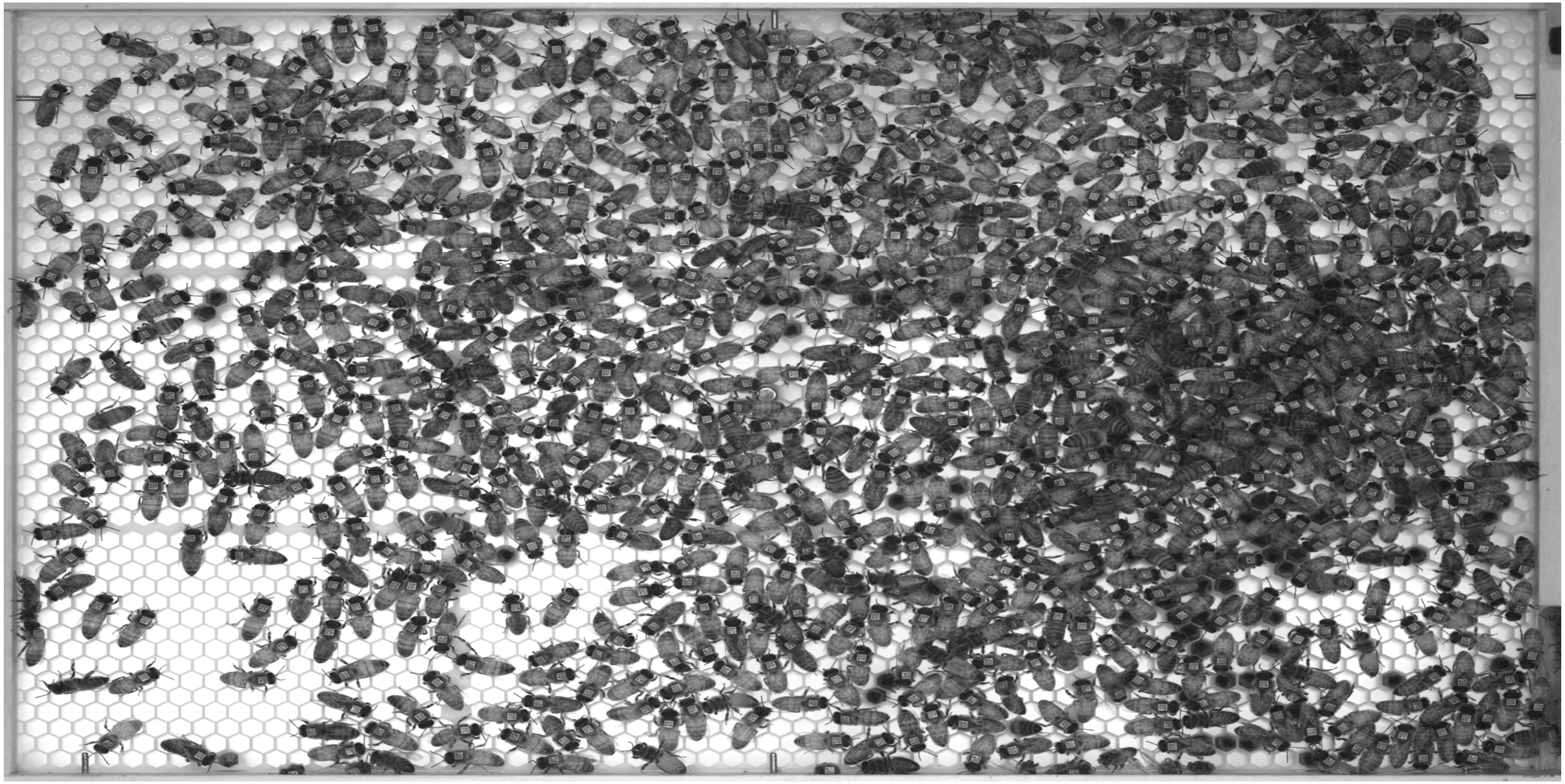
Data Structures
Image, Image Sequences

DTS

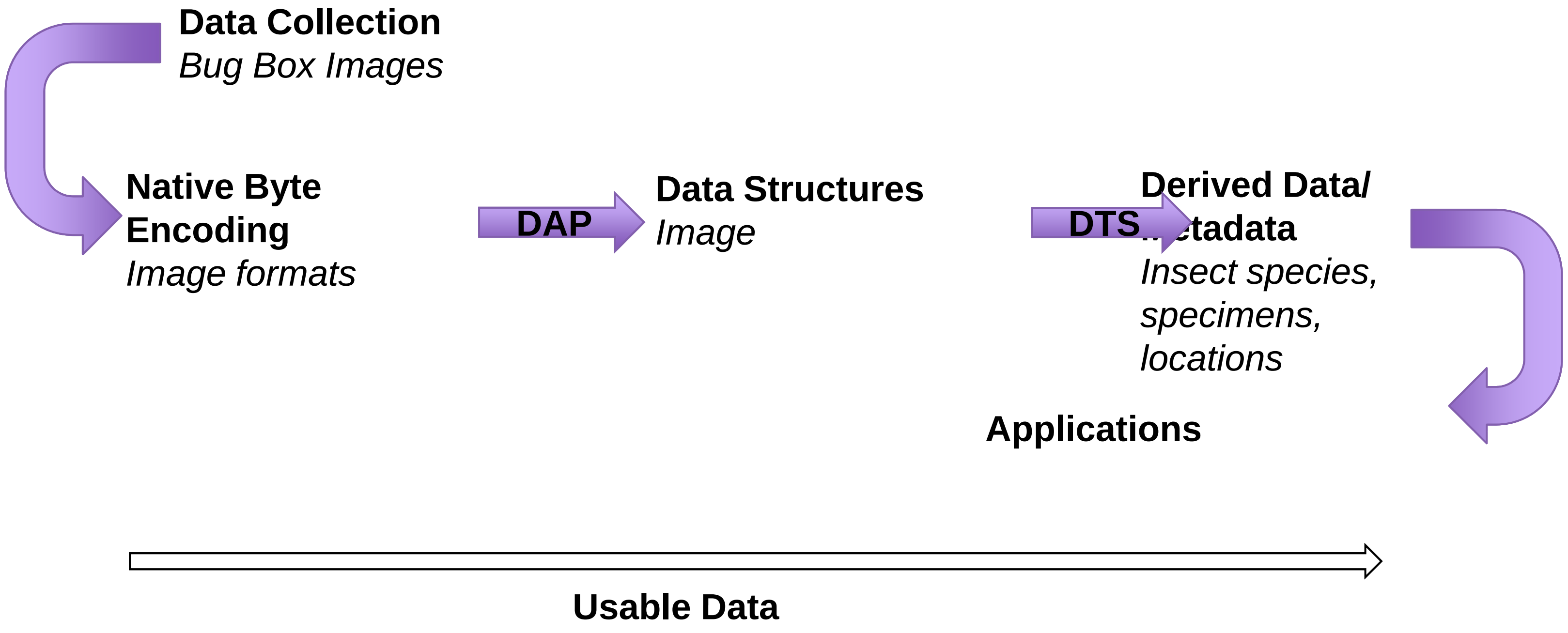
Derived Data/ Metadata
Bee locations, movement, interactions

Applications
Social Interactions of Bees

Usable Data



Entomology



Data Collection
Bug Box Images

Native Byte Encoding
Image formats

DAP

Data Structures
Image

DTS

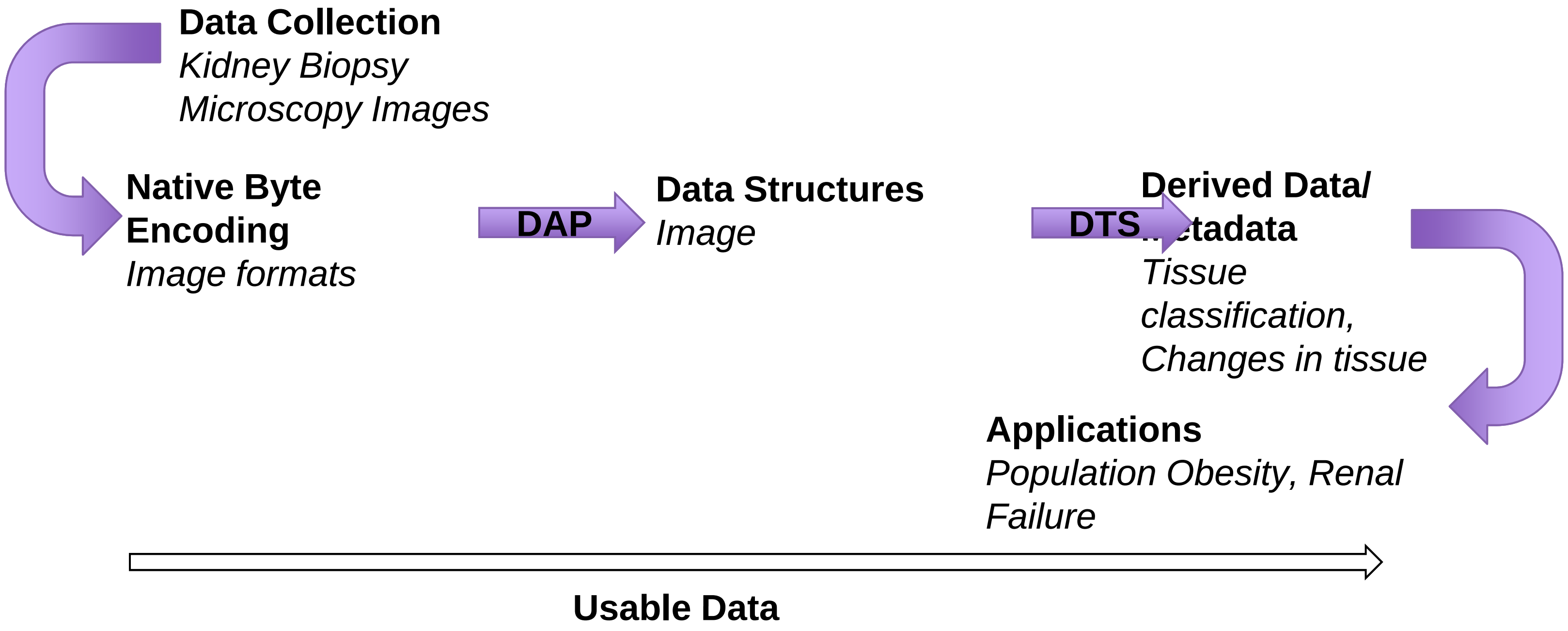
Derived Data/ Metadata
Insect species, specimens, locations

Applications

Usable Data

Div. IV
Where Insects Live
Odonata
Thysanura
Orthoptera
Dermaptera
Phasmatodea
Blattodea
Hemiptera
Lepidoptera
Mecoptera
Neuroptera
Coleoptera
Hymenoptera
Ctenostoma maculicorne det. R.L. Huber
Odontocheila iodopleura Ba det. R.L. Huber
Odontocheila nicaraguensis B det. R.L. Huber
Cicindela (Br fuscostriga det. R.L. Hube
Tetracha ignea Bate det. R.L. Hub
Cicindela (Bras nebuiosa Bate det. R.L. Huber
rilus uroguttatus S t. Hespenheide
© Ted C. MacRae 2010

Amelia Bartholomew, Medicine, University of Illinois Chicago



Data Collection
Kidney Biopsy
Microscopy Images

Native Byte Encoding
Image formats

DAP

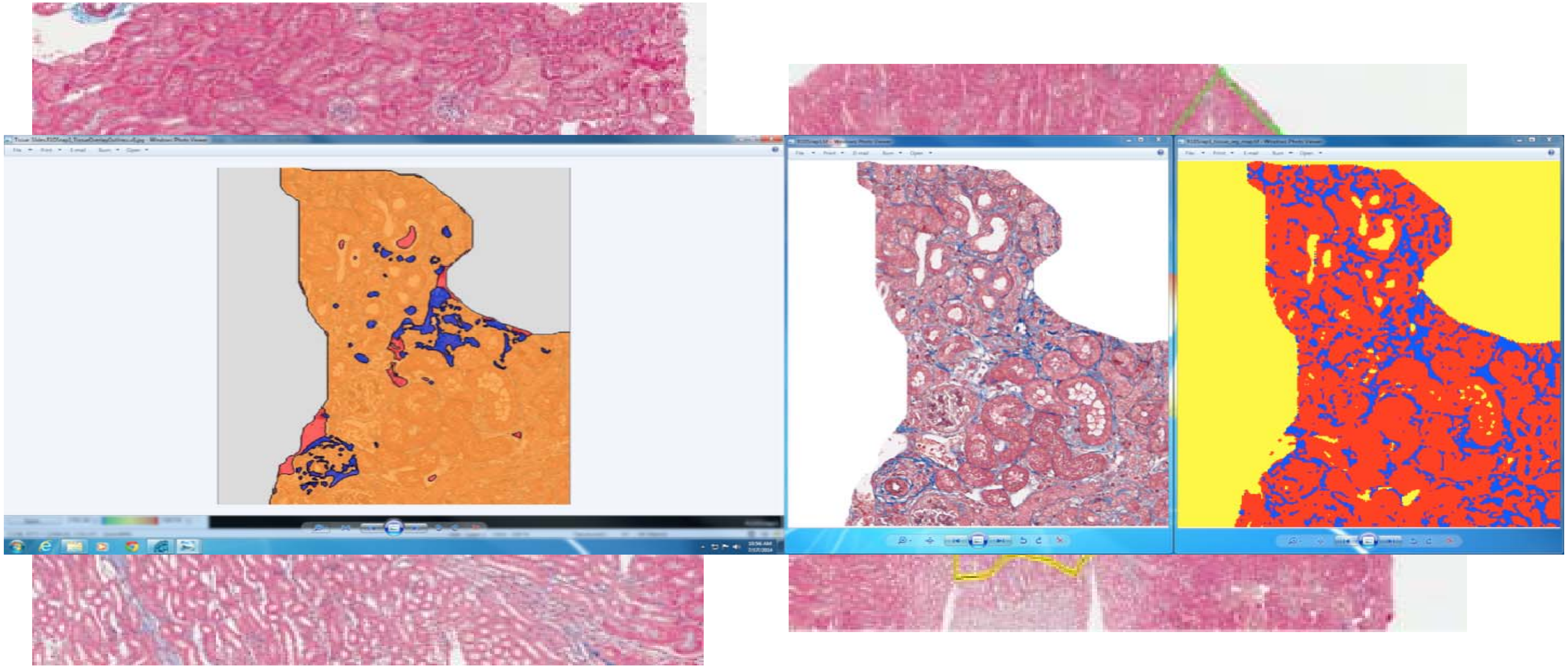
Data Structures
Image

DTS

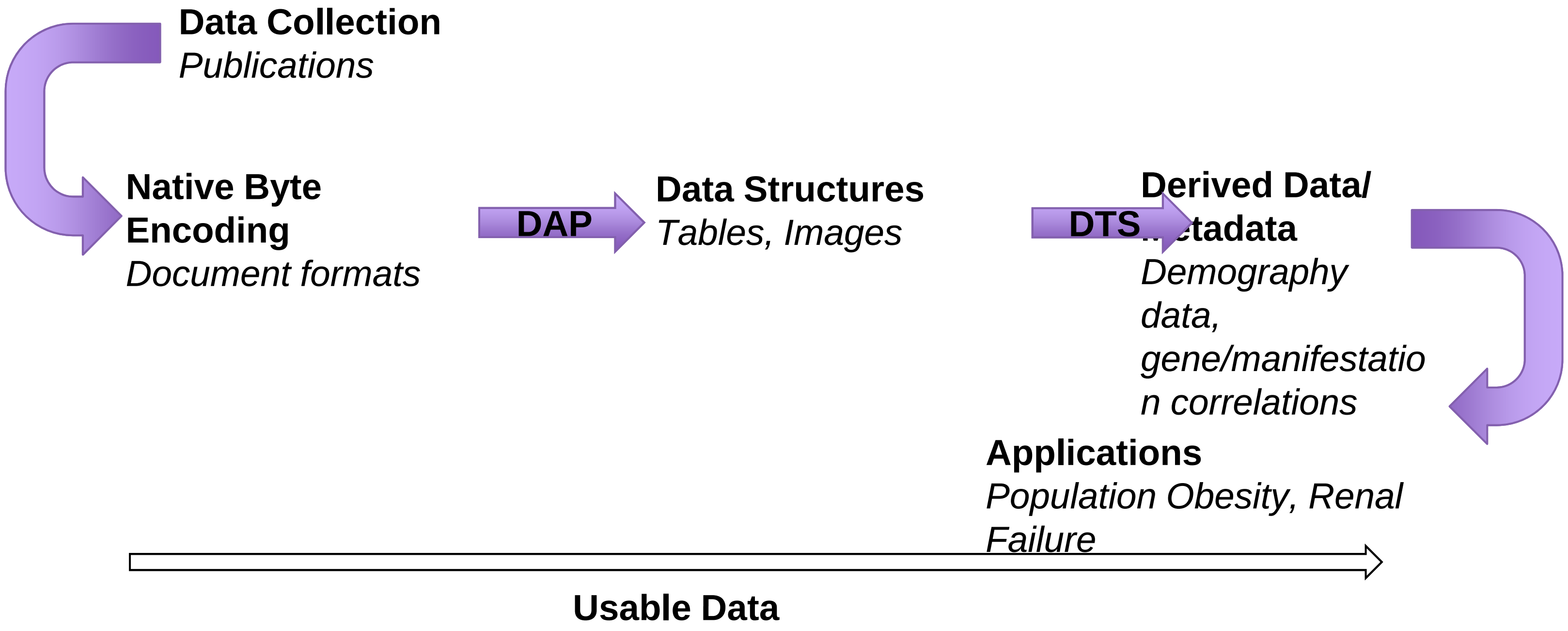
Derived Data/ Metadata
Tissue classification,
Changes in tissue

Applications
Population Obesity, Renal Failure

Usable Data



Amelia Bartholomew, Medicine, University of Illinois Chicago



Data Collection
Publications

Native Byte Encoding
Document formats

DAP

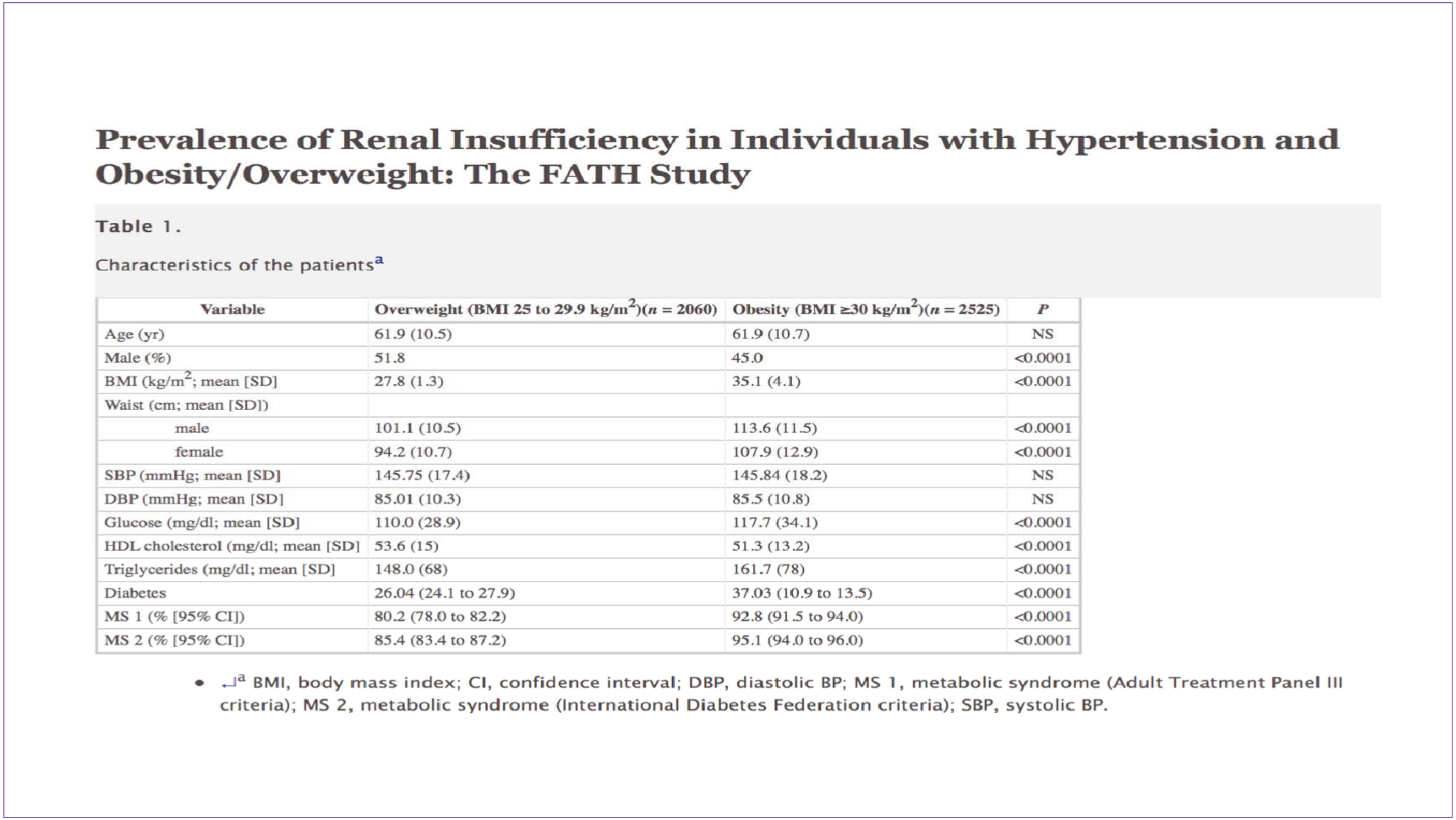
Data Structures
Tables, Images

DTS

Derived Data/ Metadata
Demography data, gene/manifestation correlations

Applications
Population Obesity, Renal Failure

Usable Data



Prevalence of Renal Insufficiency in Individuals with Hypertension and Obesity/Overweight: The FATH Study

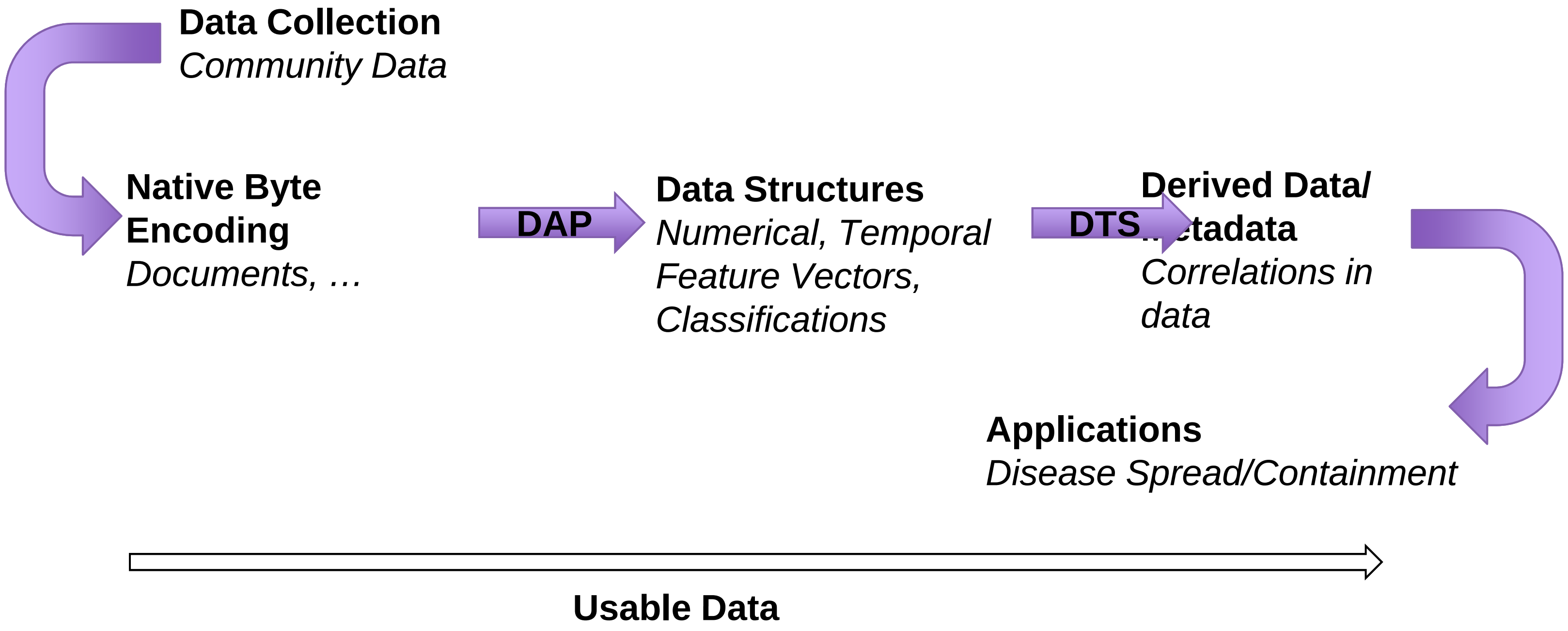
Table 1.

Characteristics of the patients[a]

Variable	Overweight (BMI 25 to 29.9 kg/m^2)(n = 2060)	Obesity (BMI ≥30 kg/m^2)(n = 2525)	*P*
Age (yr)	61.9 (10.5)	61.9 (10.7)	NS
Male (%)	51.8	45.0	<0.0001
BMI (kg/m^2; mean [SD]	27.8 (1.3)	35.1 (4.1)	<0.0001
Waist (cm; mean [SD])			
male	101.1 (10.5)	113.6 (11.5)	<0.0001
female	94.2 (10.7)	107.9 (12.9)	<0.0001
SBP (mmHg; mean [SD]	145.75 (17.4)	145.84 (18.2)	NS
DBP (mmHg; mean [SD]	85.01 (10.3)	85.5 (10.8)	NS
Glucose (mg/dl; mean [SD]	110.0 (28.9)	117.7 (34.1)	<0.0001
HDL cholesterol (mg/dl; mean [SD]	53.6 (15)	51.3 (13.2)	<0.0001
Triglycerides (mg/dl; mean [SD]	148.0 (68)	161.7 (78)	<0.0001
Diabetes	26.04 (24.1 to 27.9)	37.03 (10.9 to 13.5)	<0.0001
MS 1 (% [95% CI])	80.2 (78.0 to 82.2)	92.8 (91.5 to 94.0)	<0.0001
MS 2 (% [95% CI])	85.4 (83.4 to 87.2)	95.1 (94.0 to 96.0)	<0.0001

- ↵[a] BMI, body mass index; CI, confidence interval; DBP, diastolic BP; MS 1, metabolic syndrome (Adult Treatment Panel III criteria); MS 2, metabolic syndrome (International Diabetes Federation criteria); SBP, systolic BP.

Ian Brooks, Biochemistry, University of Illinois Urbana-Champaign



Data Collection
Community Data

Native Byte Encoding
Documents, …

DAP

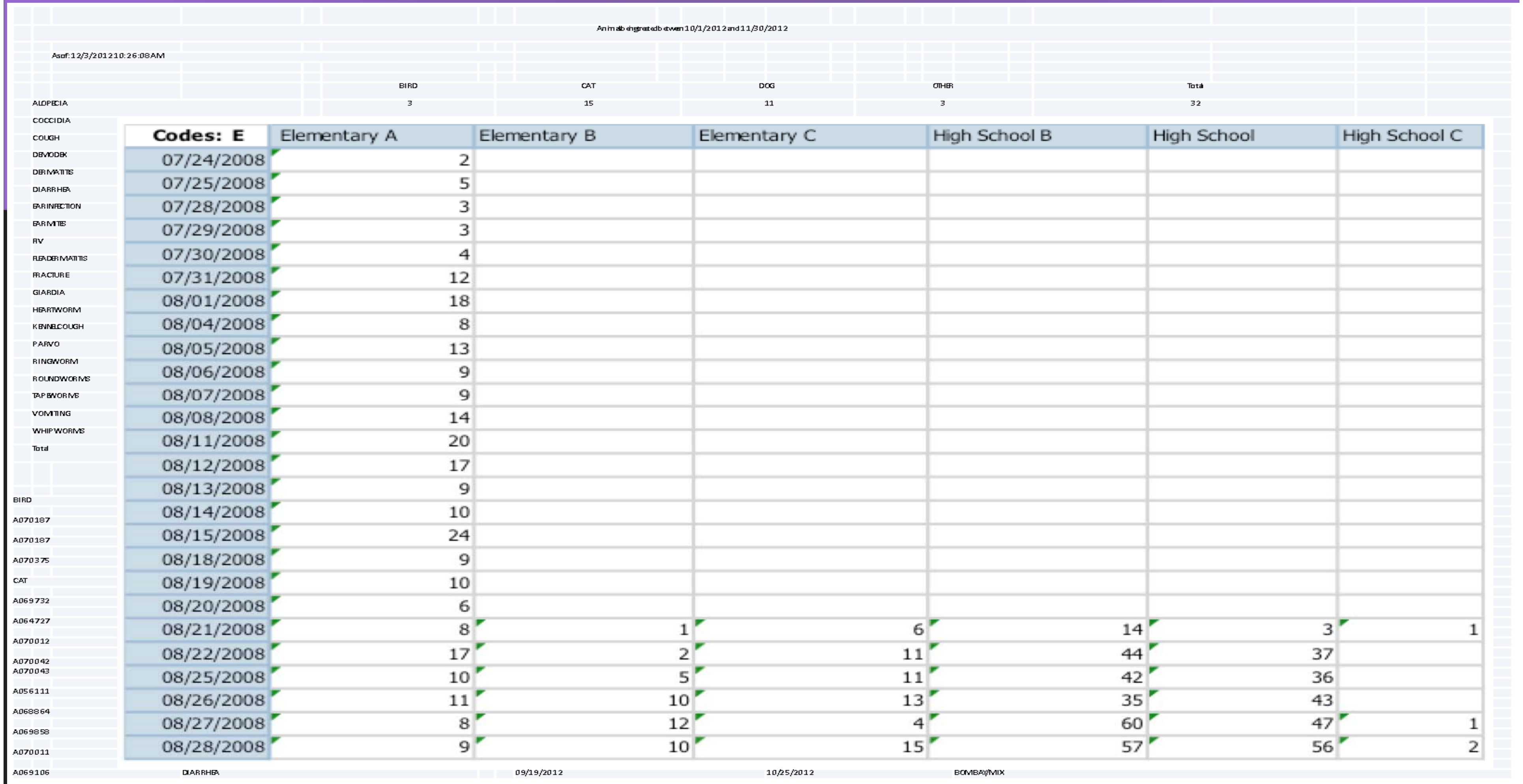
Data Structures
Numerical, Temporal Feature Vectors, Classifications

DTS

Derived Data/ Metadata
Correlations in data

Applications
Disease Spread/Containment

Usable Data



Animals being treated between 10/1/2012 and 11/30/2012

As of: 12/3/2012 10:26:08 AM

	BIRD	CAT	DOG	OTHER	Total
ALOPECIA	3	15	11	3	32
COCCIDIA					
COUGH					
DEMODEX					
DERMATITIS					
DIARRHEA					
EAR INFECTION					
EAR MITES					
FIV					
FLEA DERMATITIS					
FRACTURE					
GIARDIA					
HEARTWORM					
KENNEL COUGH					
PARVO					
RINGWORM					
ROUNDWORMS					
TAPEWORMS					
VOMITING					
WHIPWORMS					
Total					

Codes: E	Elementary A	Elementary B	Elementary C	High School B	High School	High School C
07/24/2008	2					
07/25/2008	5					
07/28/2008	3					
07/29/2008	3					
07/30/2008	4					
07/31/2008	12					
08/01/2008	18					
08/04/2008	8					
08/05/2008	13					
08/06/2008	9					
08/07/2008	9					
08/08/2008	14					
08/11/2008	20					
08/12/2008	17					
08/13/2008	9					
08/14/2008	10					
08/15/2008	24					
08/18/2008	9					
08/19/2008	10					
08/20/2008	6					
08/21/2008	8	1	6	14	3	1
08/22/2008	17	2	11	44	37	
08/25/2008	10	5	11	42	36	
08/26/2008	11	10	13	35	43	
08/27/2008	8	12	4	60	47	1
08/28/2008	9	10	15	57	56	2

BIRD

A070187

A070187

A070375

CAT

A069732

A064727

A070012

A070042

A070043

A056111

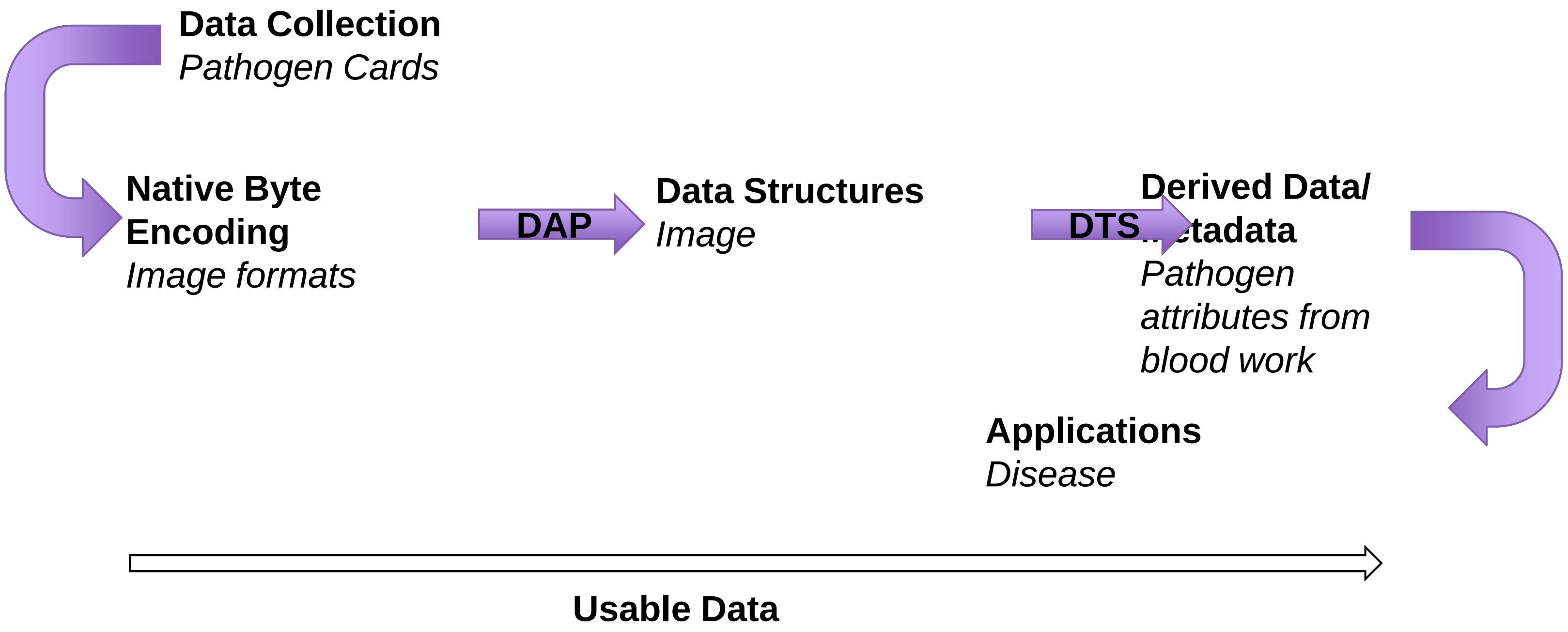
A068864

A069858

A070011

A069106	DIARRHEA	09/19/2012	10/25/2012	BOMBAY/MIX

Christopher Lynberg, Computer Science, Center for Disease Control (CDC)



Data Collection
Pathogen Cards

Native Byte Encoding
Image formats

DAP

Data Structures
Image

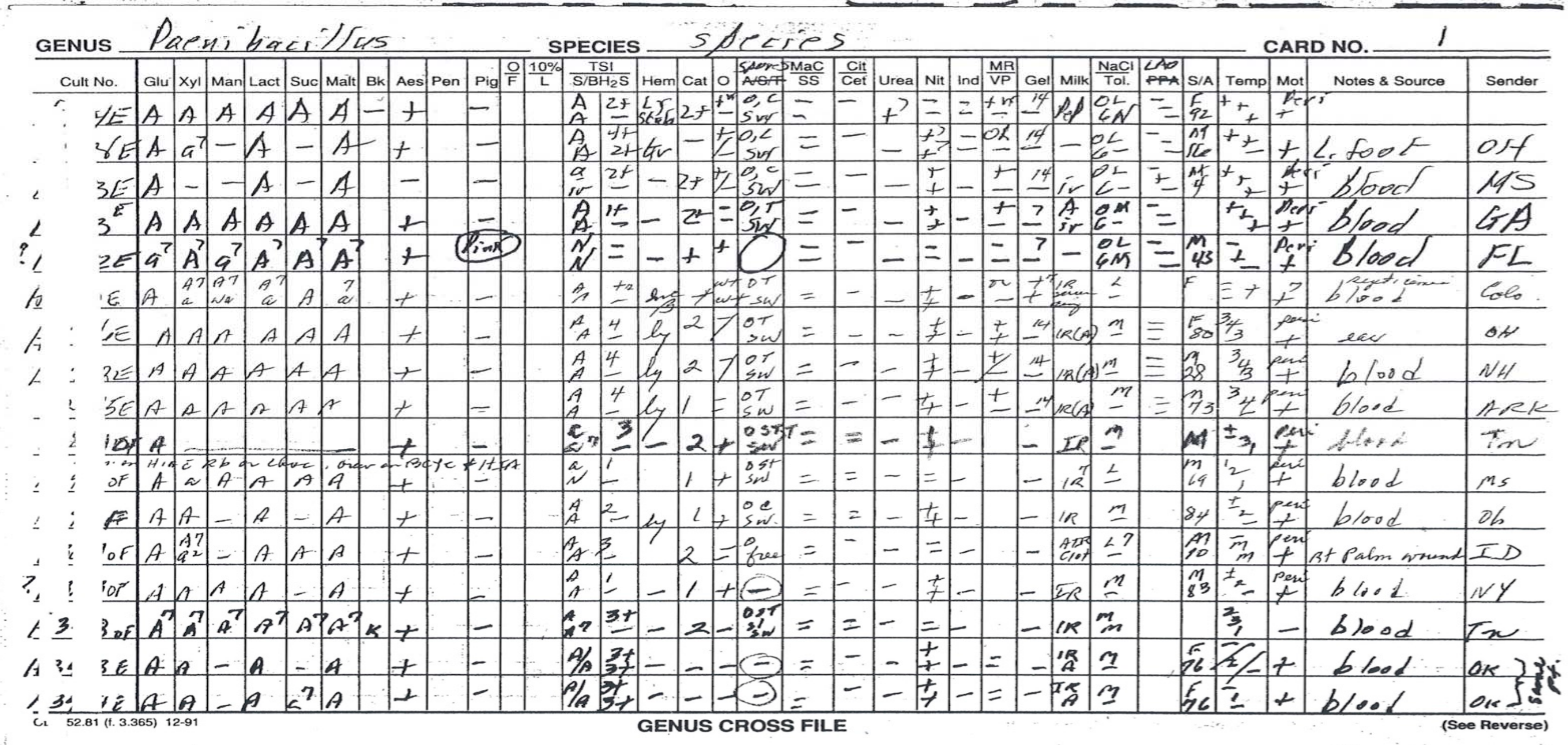
DTS

Derived Data/ Metadata
Pathogen attributes from blood work

Applications
Disease

Usable Data

Paenibacillus +
5079



GENUS Paenibacillus SPECIES species CARD NO. 1

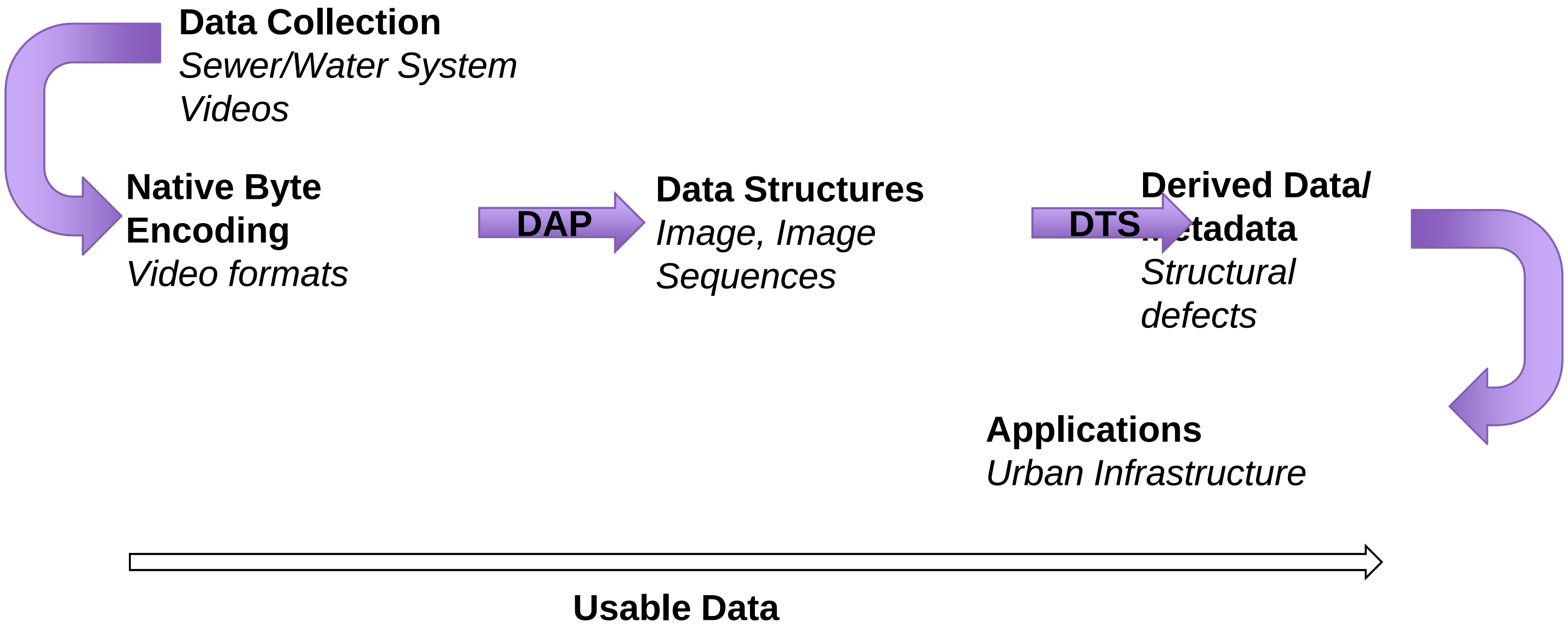
Cult No.	Glu	Xyl	Man	Lact	Suc	Malt	Bk	Aes	Pen	Pig	O F	10% L	TSI S/B	H_2S	Hem	Cat	O	~~A/G/F~~ Spore	MaC SS	Cit Cet	Urea	Nit	Ind	MR VP	Gel	Milk	NaCl Tol.	LAO ~~PPA~~	S/A	Temp	Mot	Notes & Source	Sender
4E	A	A	A	A	A	A	—	+		—			A A	2+ —	Lr Steb	2+	+w —	O,C Sw	— –	—	+?	— —	= —	+w —	14 —	Pep	OL GN	— —	F 92	++ +	Peri +		
8E	A	a?	—	A	—	A		+		—			A A	4+ 2+	Gr	—	+ —	O,L Sw	=	—	—	+? +?	— —	OL —	14 —	—	OL G—	— —	M Sl	+± —	+	L. foot	OH
3E	A	—	—	A	—	A		—		—			a iv	2+ —	—	2+	+/	O,C SW	=	—	—	+ +	—	+ —	14 —	iv	OL L—	— +—	M 4	+ + +	Peri +	blood	MS
3E	A	A	A	A	A	A		+		—			A A	1+ —	—	2+	— —	O,T SW	=	—	—	+ +	—	+ —	7 —	A ir	OM G—	— =		+ + +	Peri +	blood	GA
2E	a?	A?	a?	A?	A?	A?		+		Pink			N N	— =	—	+	+	○	=	—	—	— —	—	— —	7 —	—	OL GM	— =	M 43	— +—	Peri +	blood	FL
1E	A	A? a	A? Wa	A? a	A	? a		+		—			A A	+2 —	[illegible] β	+	wt wt	OT SW	=	—	—	+ +	—	w —	+? +	IR [illegible]	L —		F	=+	+?	[illegible] blood	Colo.
1E	A	A	A	A	A	A		+		—			A A	4 —	ly	2	/	OT SW	=	—	—	+ +	—	+ +	14 —	IR(A)	M —	≡	F 80	3/4 3	peri +	ear	OH
2E	A	A	A	A	A	A		+		—			A A	4 —	ly	2	/	OT SW	=	—	—	+ +	—	+/	14 —	IR(A)	M —	≡	M 28	3/4 3	peri +	blood	NH
5E	A	A	A	A	A	A		+		=			A A	4 —	ly	1	=	OT SW	=	—	—	+ +	—	+ —	14 —	IR(A)	M —	=	M 73	3/4 4	peri +	blood	ARK
1OF	A	—	—	—	—	—		+		—			C? S?	3 —	—	2	+	OST?T SW	=	=	—	+	—		—	IR	M —		M	±3 1	Peri +	blood	Tn
	HIAE Rb or Choc, Over on BCYE + HIA																																
0F	A	a	A	A	A	A		+		—			a N	1 —		1	+	OST SW	=	=	—	=	—		—	T IR	L —		M 69	1/2 1	peri +	blood	MS
F	A	A	—	A	—	A		+		—			A A	2 —	ly	1	+	OC SW.	=	=	—	+ +	—		—	IR	M —		84	± 2—	peri +	blood	Oh
1OF	A	A? a2	—	A	A	A		+		—			A A	3 —		2	—	O free	=	—	—	=	—		—	ATR Clot	L? —		M 10	— m m	peri +	Rt Palm wound	ID
1OF	A	A	A	A	—	A		+		—			A A	1 —	—	1	+	(—)	=	—	—	+ +	—		—	IR	M —		M 83	± 2—	Peri +	blood	NY
3 3OF	A?	A?	A?	A?	A?	A?	K	+		—			A A?	3+ —	—	2	—	OST 3/ SW	=	=	—	=	—		—	IR	M M			2 3 1	—	blood	Tn
3 3E	A	A	—	A	—	A		+		—			A/A	3+ 3+	—	—	—	(—)	=	—	—	+ +	—	=	—	IR A	M —		F 76	— ±/—	+	blood	OK }
3 1E	A	A	—	A	c?	A		+		—			A/A	3+ 3+	—	—	—	(—)	=	—	—	+ +	—	=	—	IR A	M —		F 76	— ±	+	blood	OK } [illegible]

C.L 52.81 (f. 3.365) 12-91

GENUS CROSS FILE

(See Reverse)

Sonia Giovinazzi, Civil & Natural Resource Engineering, University of Canterbury



Data Collection
Sewer/Water System Videos

Native Byte Encoding
Video formats

DAP

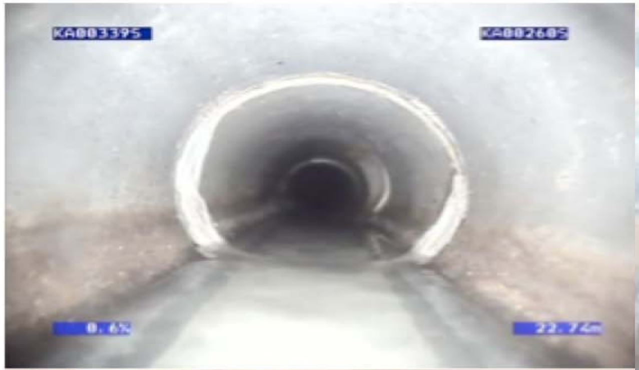
Data Structures
Image, Image Sequences

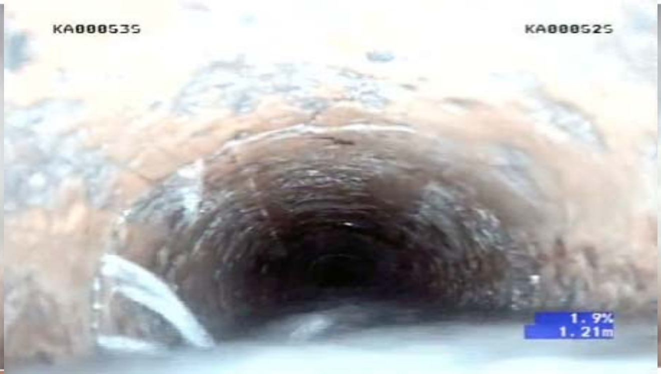
DTS

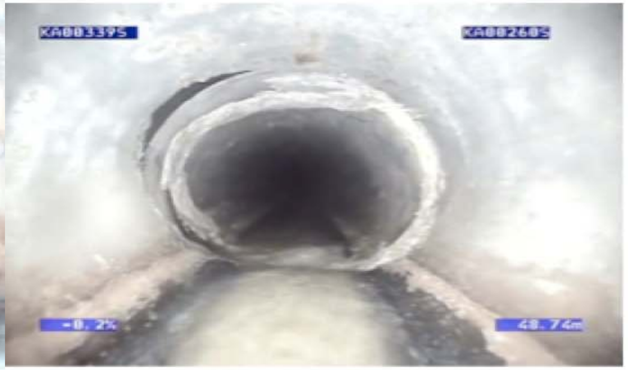
Derived Data/ Metadata
Structural defects

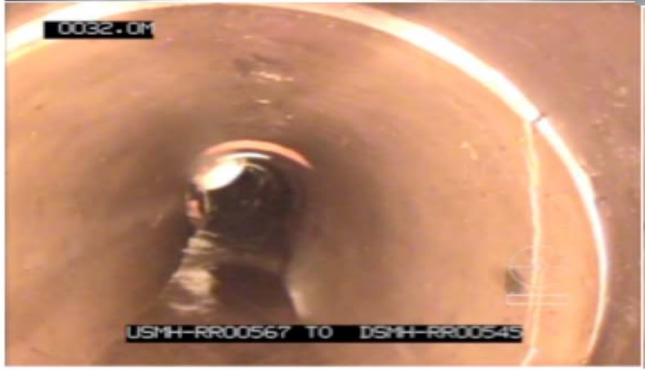
Applications
Urban Infrastructure

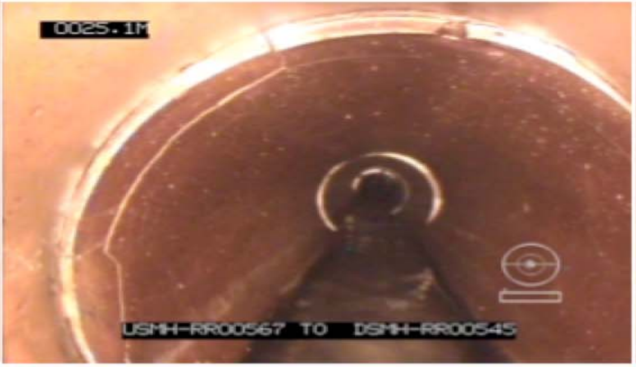
Usable Data

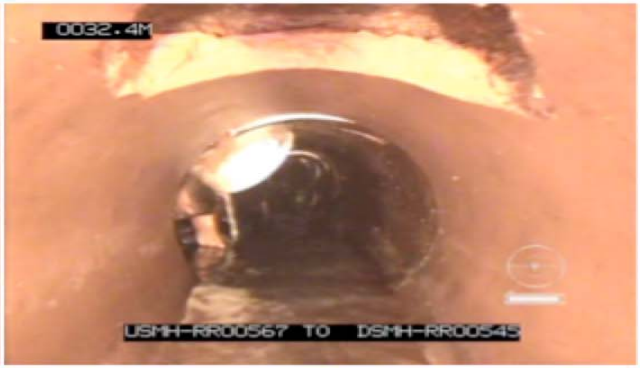

KA003395
KA002605
0.6%
22.74m


KA000535
KA000525
1.9%
1.21m

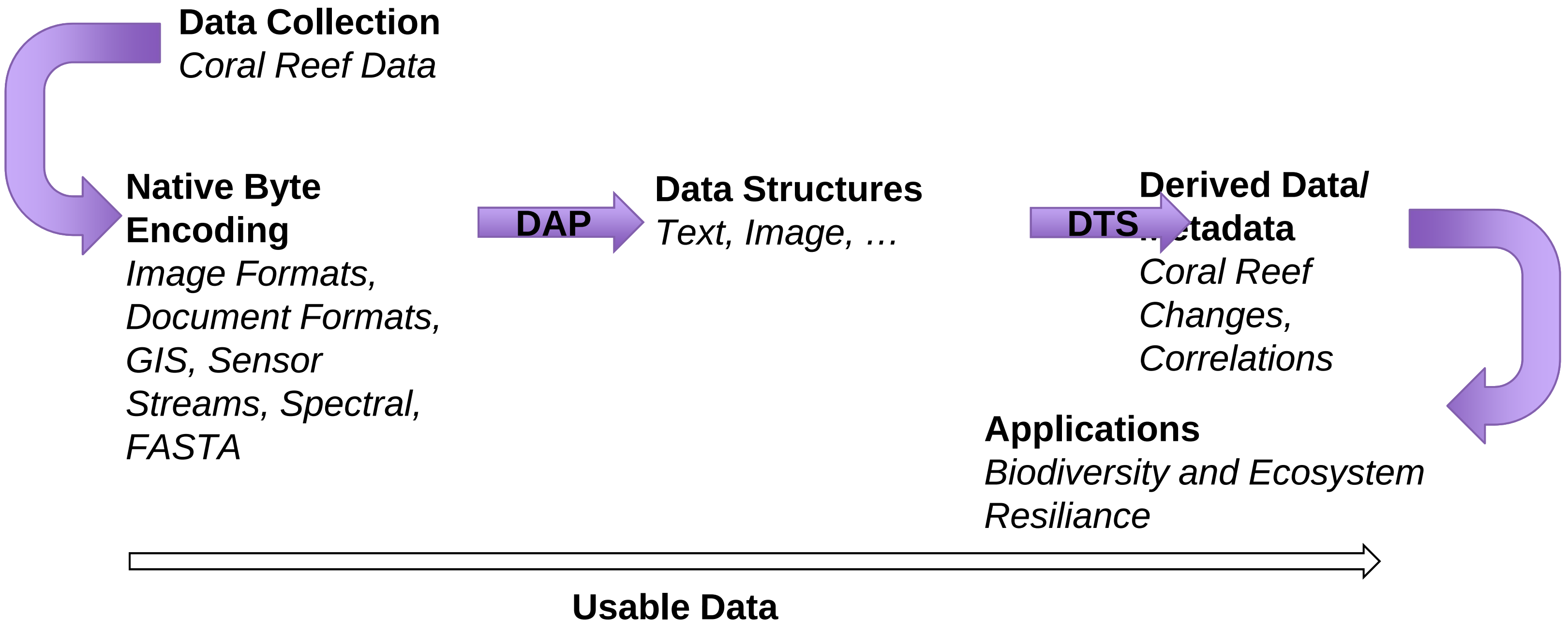

KA003395
KA002605
-0.2%
48.74m


0032.0M
USMH-RR00567 TO DSMH-RR00545


0025.1M
USMH-RR00567 TO DSMH-RR00545


0032.4M
USMH-RR00567 TO DSMH-RR00545

Ruth Gates, Marine Biology, University of Hawaii



Data Collection
Coral Reef Data

Native Byte Encoding
Image Formats, Document Formats, GIS, Sensor Streams, Spectral, FASTA

DAP

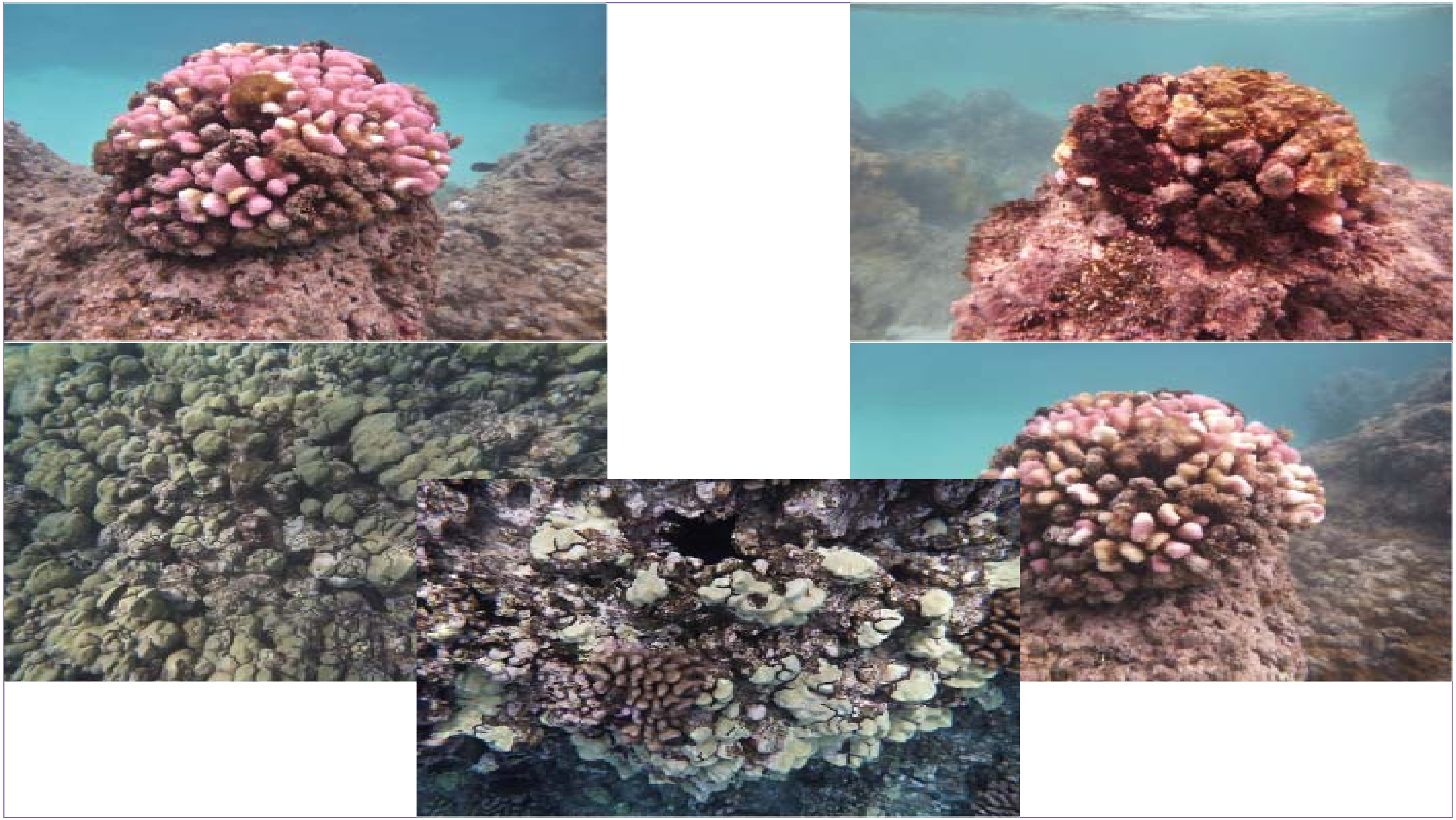
Data Structures
Text, Image, …

DTS

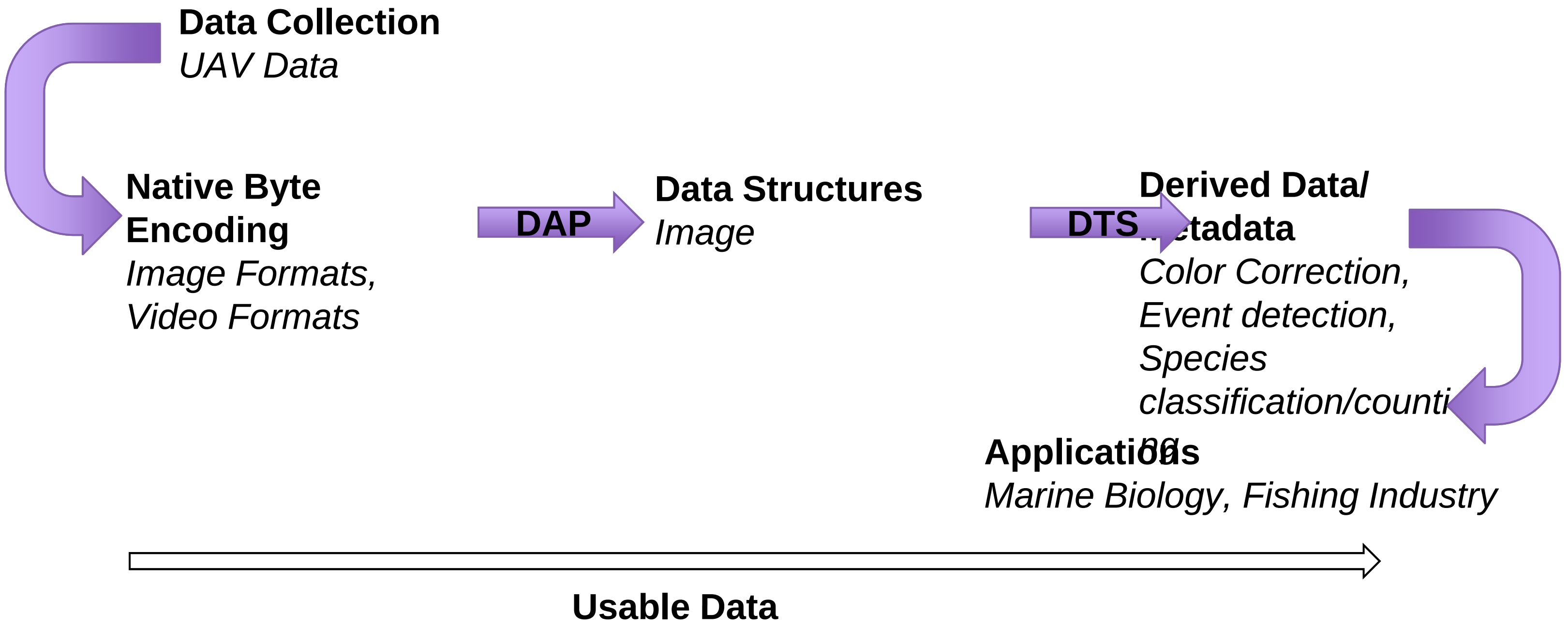
Derived Data/ Metadata
Coral Reef Changes, Correlations

Applications
Biodiversity and Ecosystem Resiliance

Usable Data



Chris German, Marine Geochemistry, Woods Hole Oceanographic Institute (WHOI)
Scott Gallager, Biology, Woods Hole Oceanographic Institute (WHOI)
James Kinsey, Mechanical Engineering, Woods Hole Oceanographic Institute (WHOI)
Joe Futrelle, Computer Science, Woods Hole Oceanographic Institute (WHOI)



Data Collection
UAV Data

Native Byte Encoding
Image Formats, Video Formats

DAP

Data Structures
Image

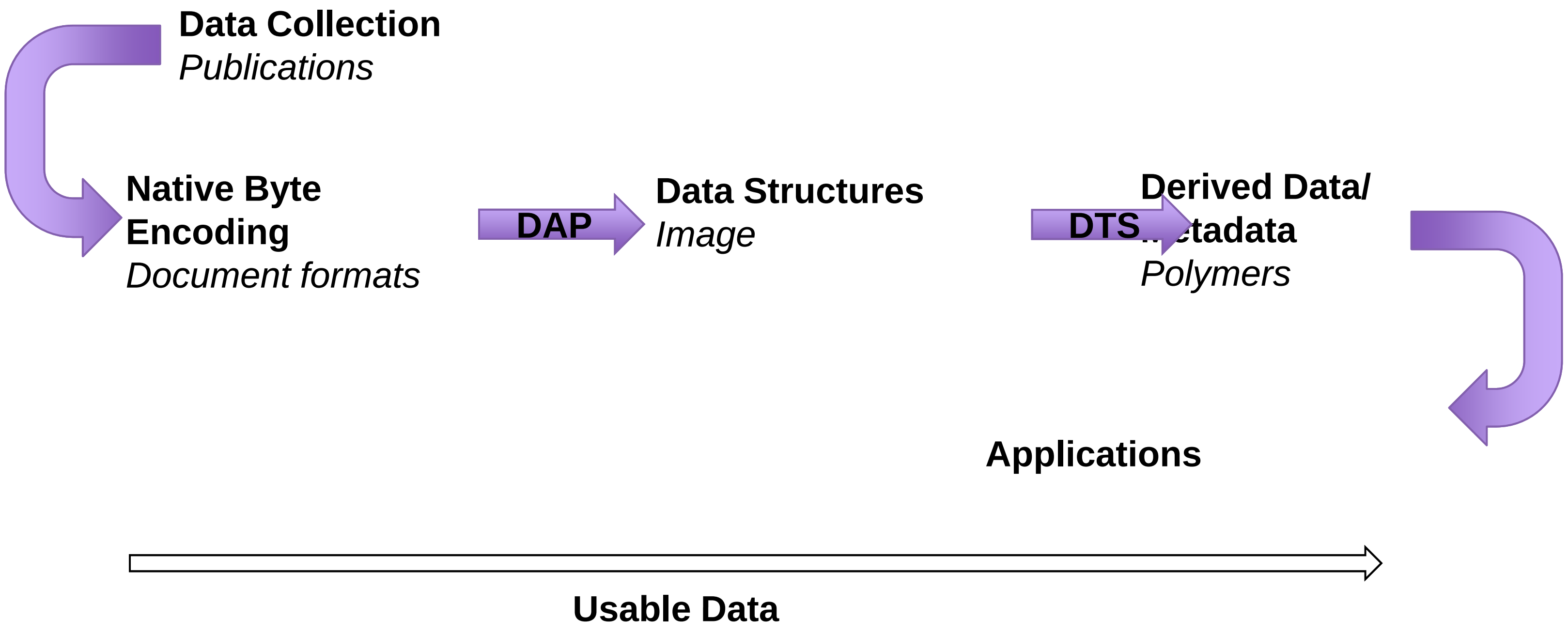
DTS

Derived Data/ Metadata
Color Correction, Event detection, Species classification/counting

Applications
Marine Biology, Fishing Industry

Usable Data

David Zeppa, Computer Science, University of California Santa Cruz



Data Collection
Publications

Native Byte Encoding
Document formats

DAP

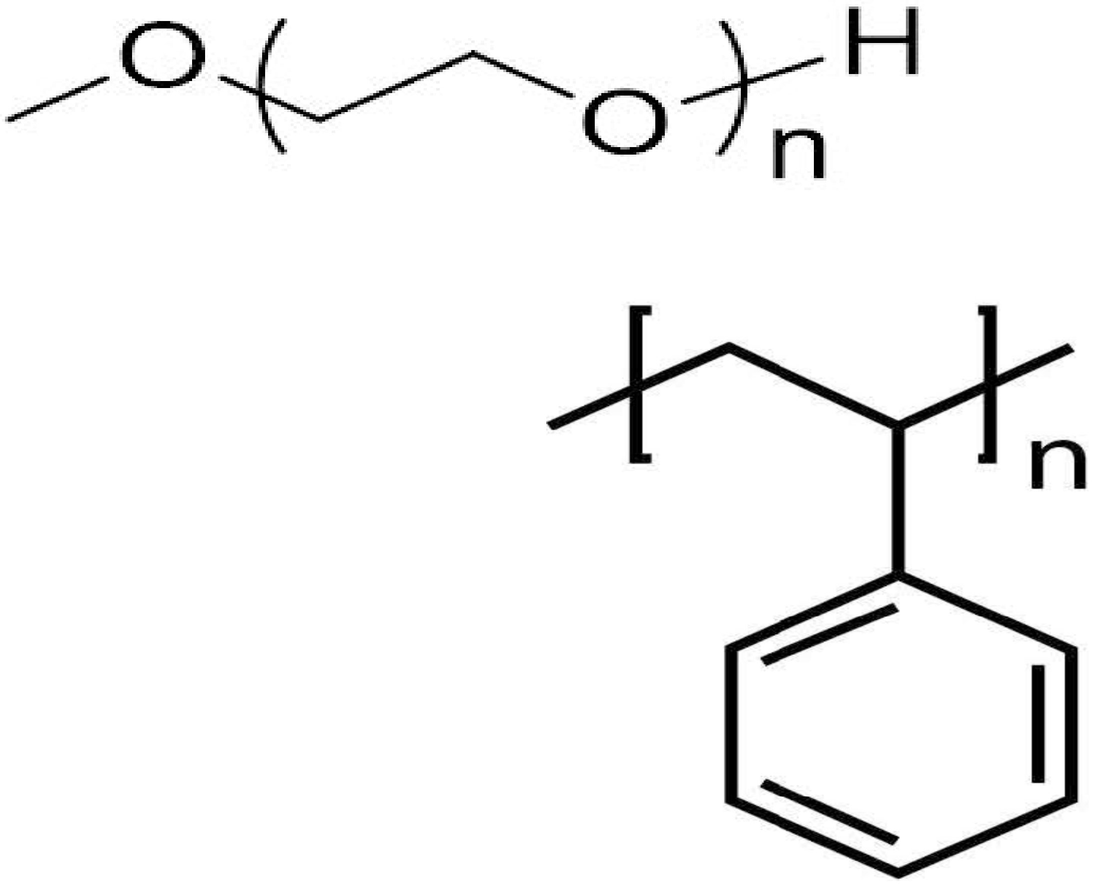
Data Structures
Image

DTS

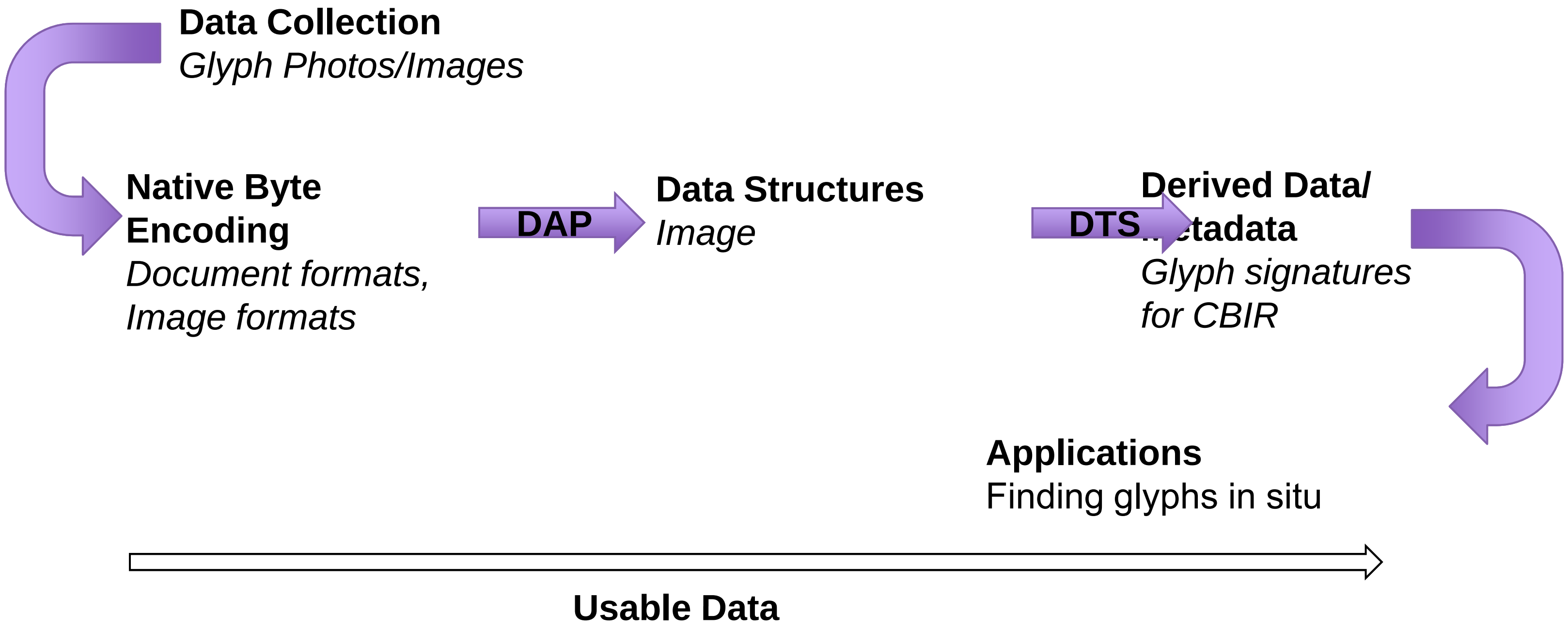
Derived Data/ Metadata
Polymers

Applications

Usable Data


O
O
H
n
n

David Zeppa, Anthropology, University of Southern Florida



Data Collection
Glyph Photos/Images

Native Byte Encoding
Document formats, Image formats

DAP

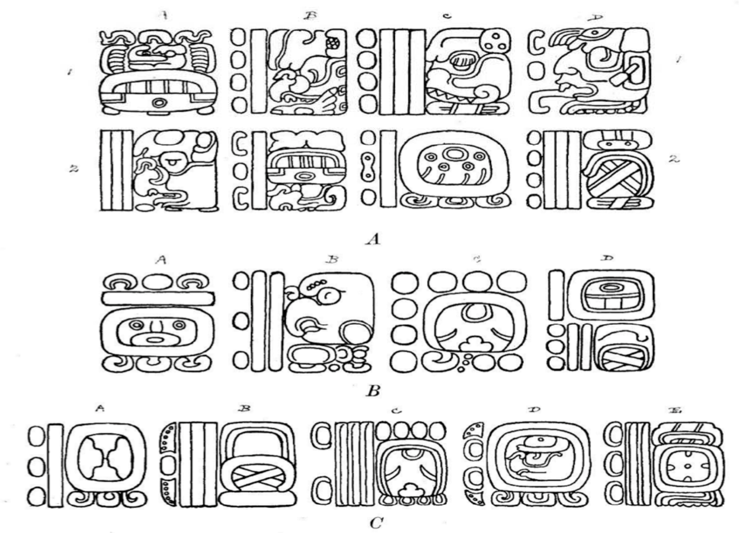
Data Structures
Image

DTS

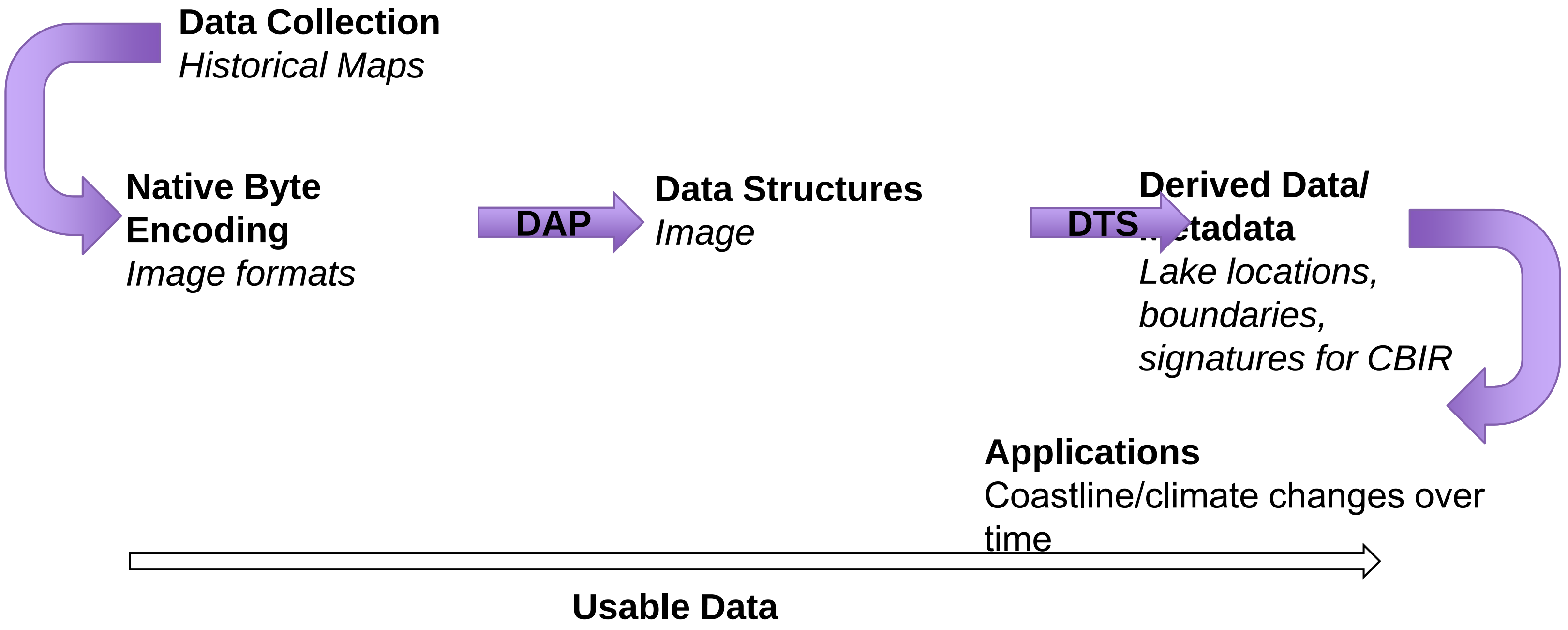
Derived Data/ Metadata
Glyph signatures for CBIR

Applications
Finding glyphs in situ

Usable Data


A
B
C
D
1
2
A
A
B
C
D
B
A
B
C
D
E
C

Robert Markley, English, University of Illinois Urbana-Champaign



Data Collection
Historical Maps

Native Byte Encoding
Image formats

DAP

Data Structures
Image

DTS

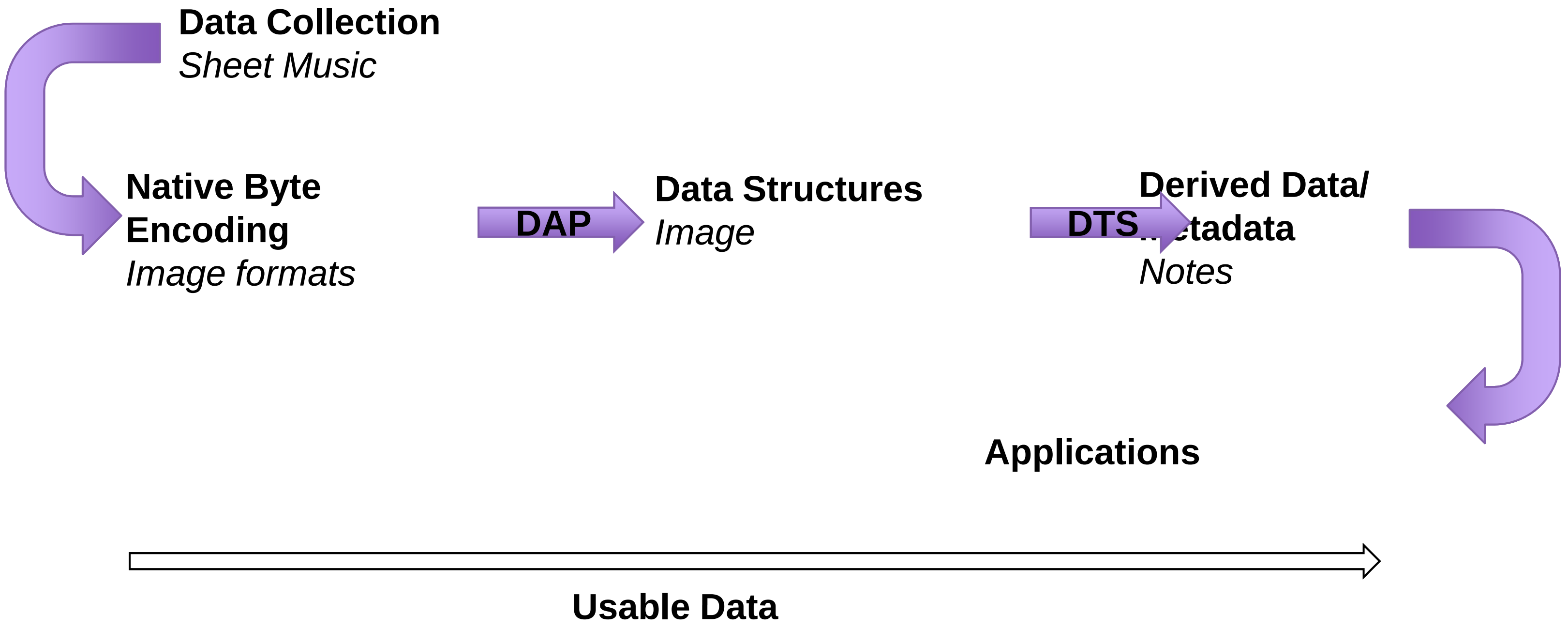
Derived Data/ Metadata
Lake locations, boundaries, signatures for CBIR

Applications
Coastline/climate changes over time

Usable Data

Great Fall
the H
L.Claire
LAKE ERIE
etroit
L. Erie
Presq
Le Beuf
Meadsville
LAKE ERIE

Ann MacNeil, Music, University of North Carolina at Chapel Hill



Data Collection
Sheet Music

Native Byte Encoding
Image formats

DAP

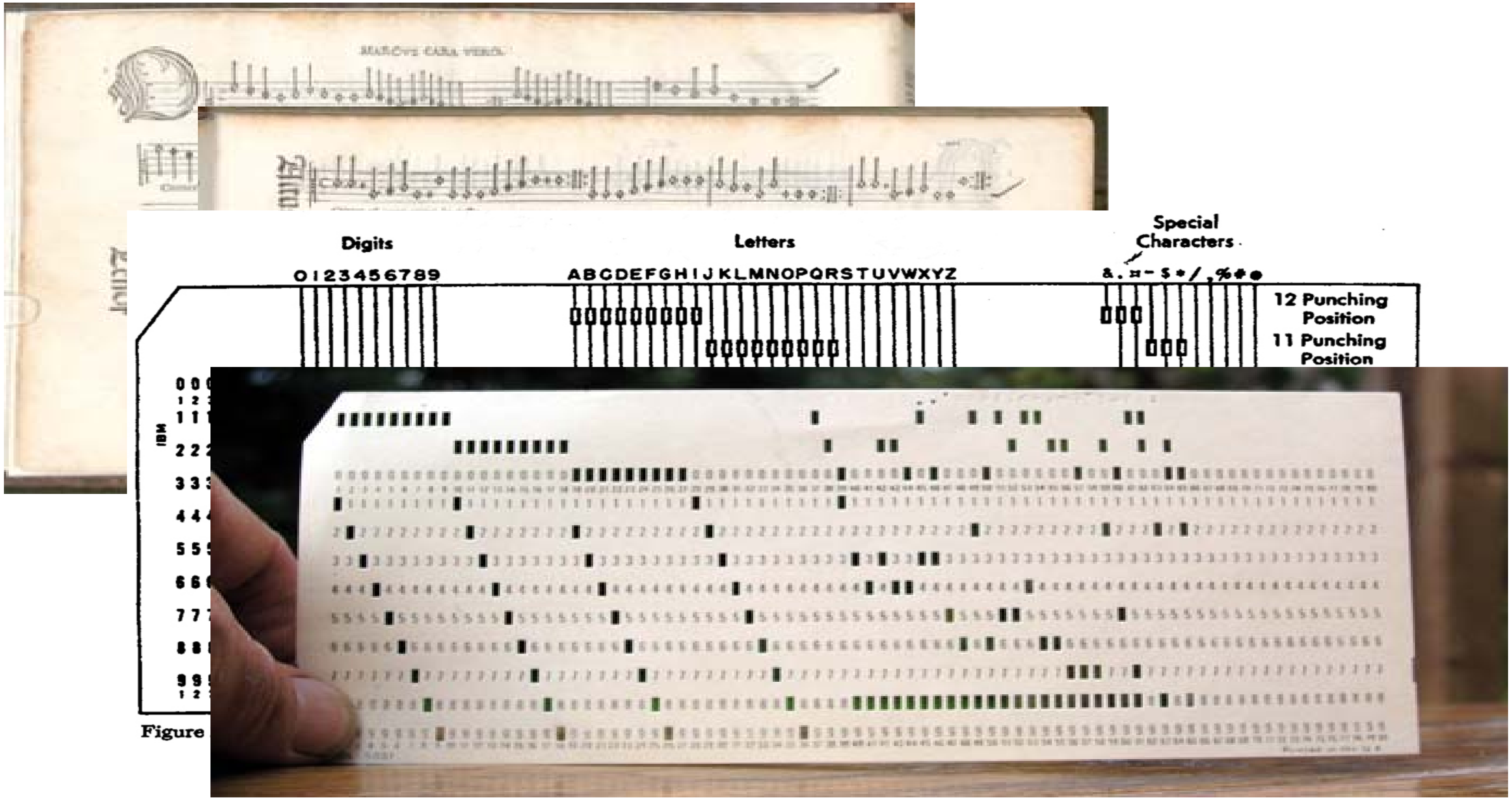
Data Structures
Image

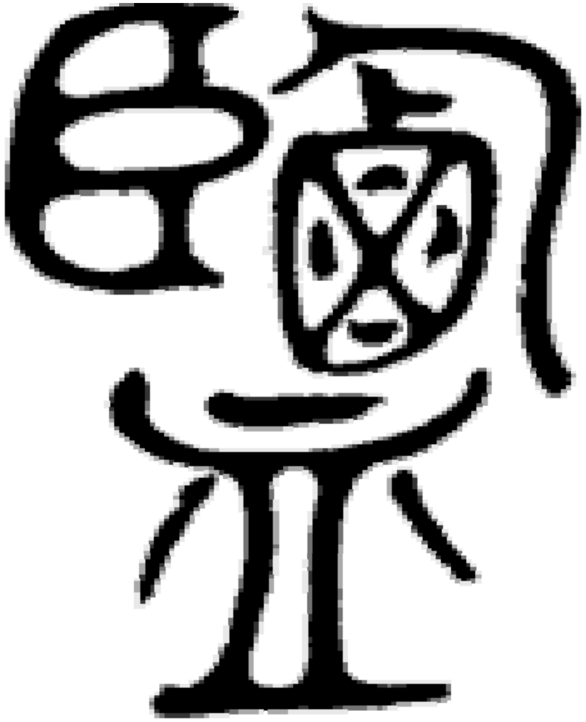
DTS

Derived Data/ Metadata
Notes

Applications

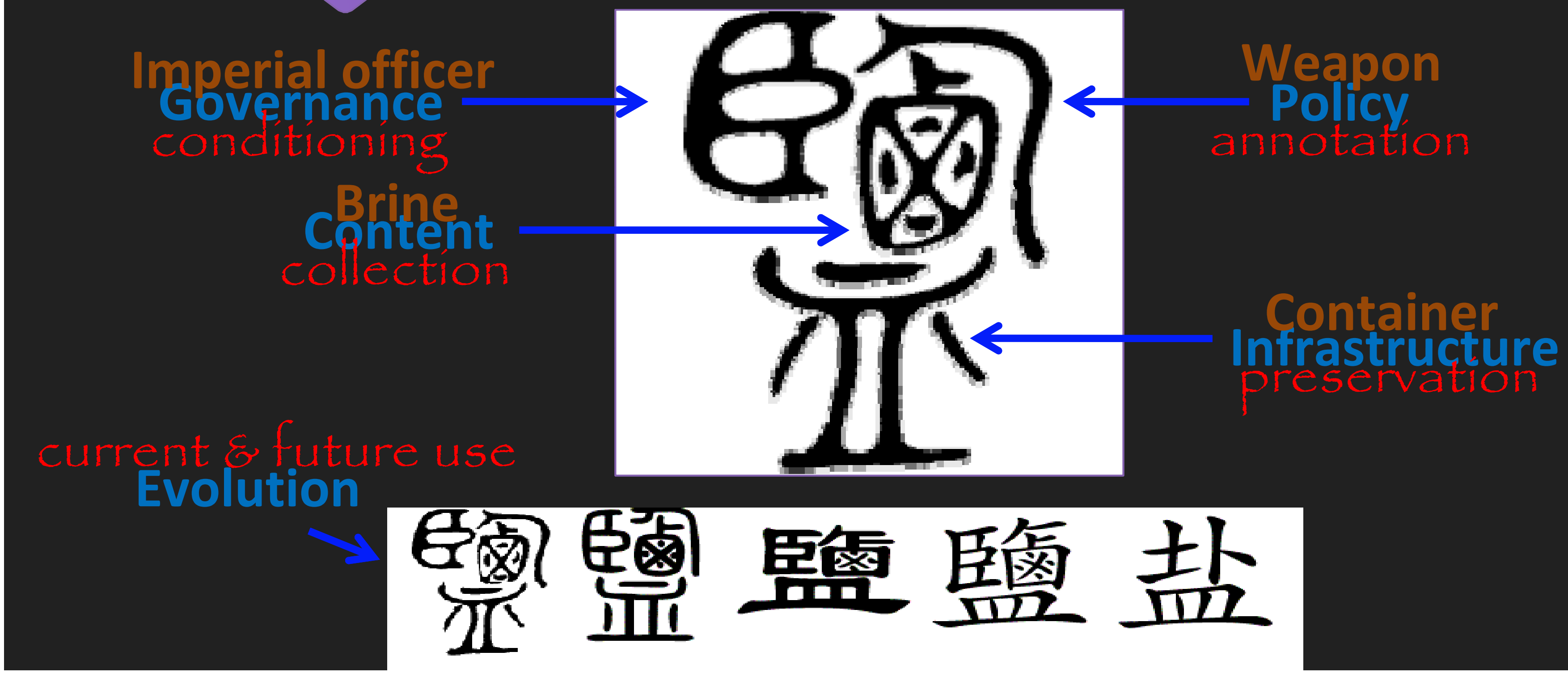
Usable Data


Digits
0123456789
Letters
ABCDEFGHIJKLMNOPQRSTUVWXYZ
Special Characters
&.¤-$*/,%#@
12 Punching Position
11 Punching Position
IBM
Figure



SustainAbiLiTy (SALT) lab

SALT/"yan": a metaphor for digital curation



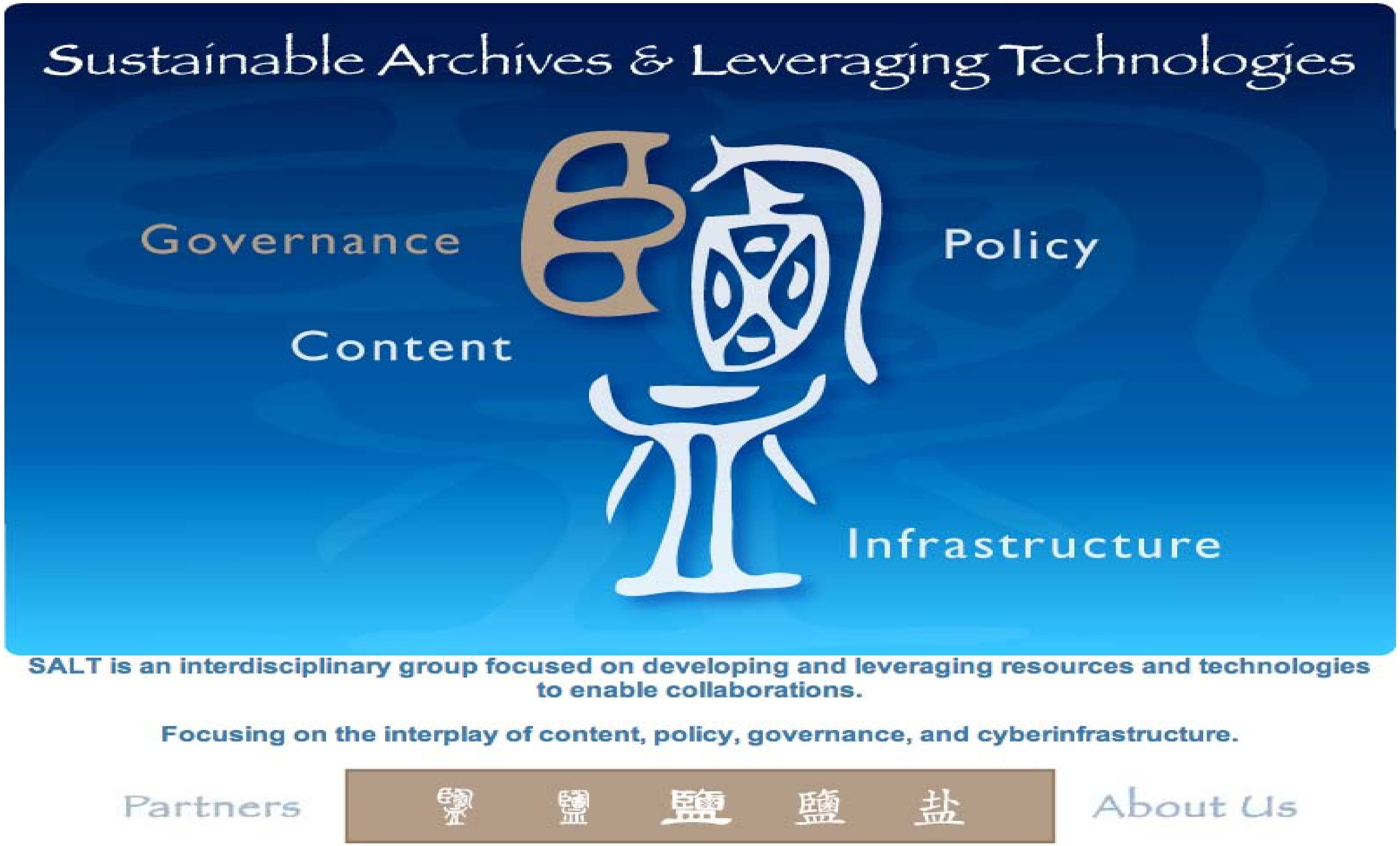
Imperial officer
Governance
conditioning

Brine
Content
collection

Weapon
Policy
annotation

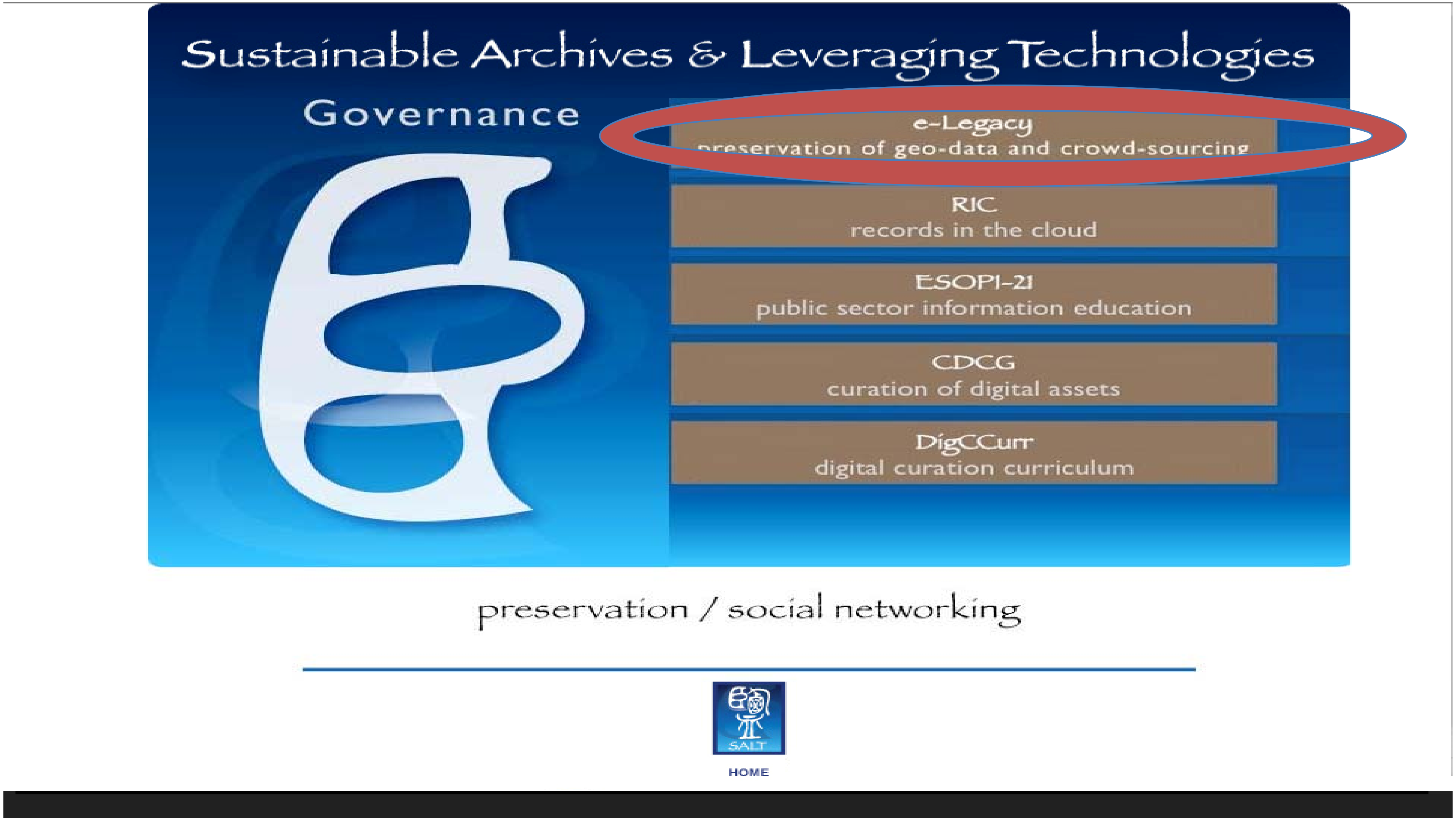
Container
Infrastructure
preservation

current & future use
Evolution



SALT is an interdisciplinary group focused on developing and leveraging resources and technologies to enable collaborations.

Focusing on the interplay of content, policy, governance, and cyberinfrastructure.



Sustainable Archives & Leveraging Technologies

Governance

e-Legacy
preservation of geo-data and crowd-sourcing

RIC
records in the cloud

ESOPI-21
public sector information education

CDCG
curation of digital assets

DigCCurr
digital curation curriculum

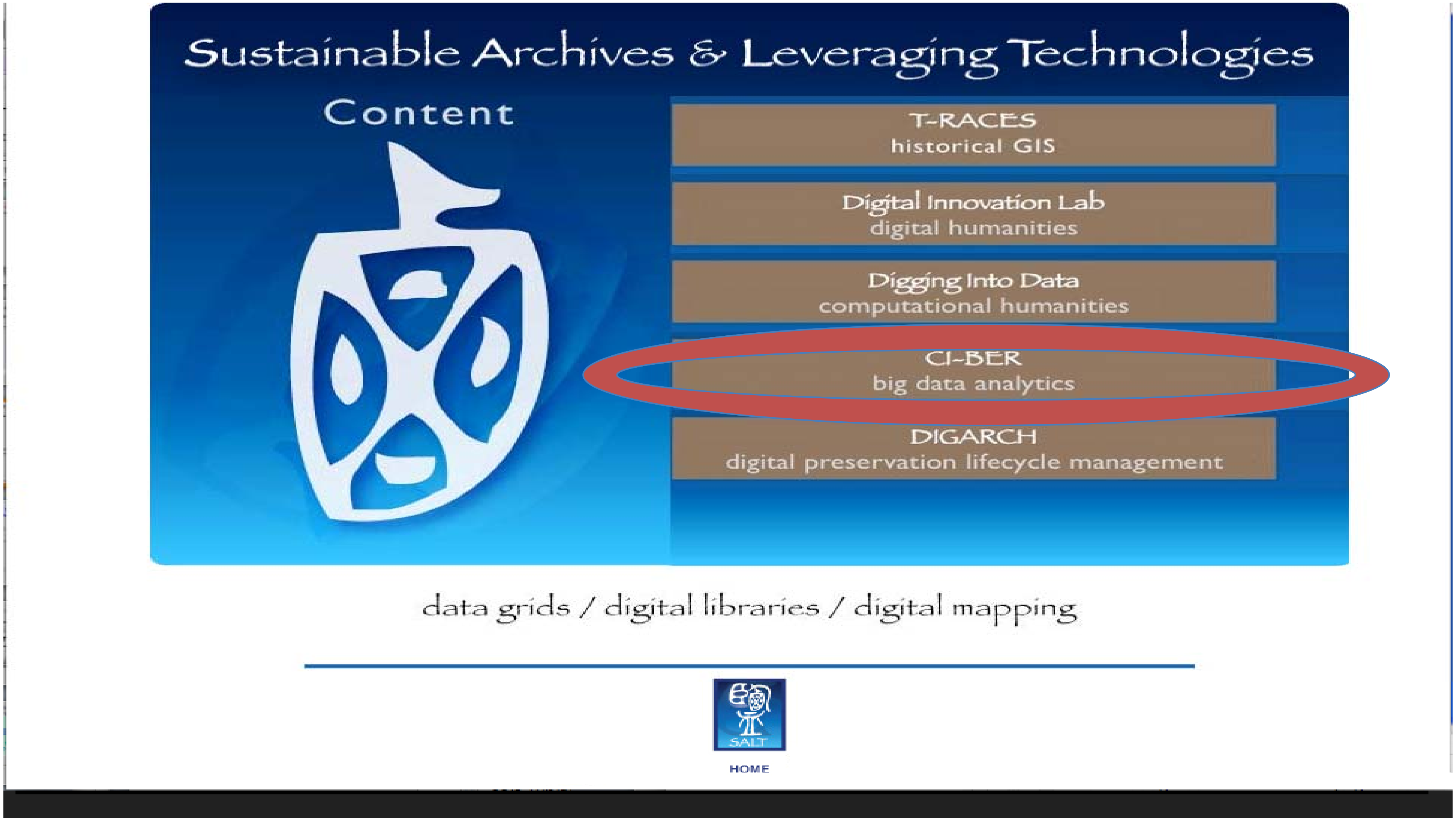
preservation / social networking

HOME

Sustainable Archives & Leveraging Technologies

SALT is an interdisciplinary group focused on developing and leveraging resources and technologies to enable collaborations.

Focusing on the interplay of content, policy, governance, and cyberinfrastructure.



Sustainable Archives & Leveraging Technologies

Content

T-RACES
historical GIS

Digital Innovation Lab
digital humanities

Digging Into Data
computational humanities

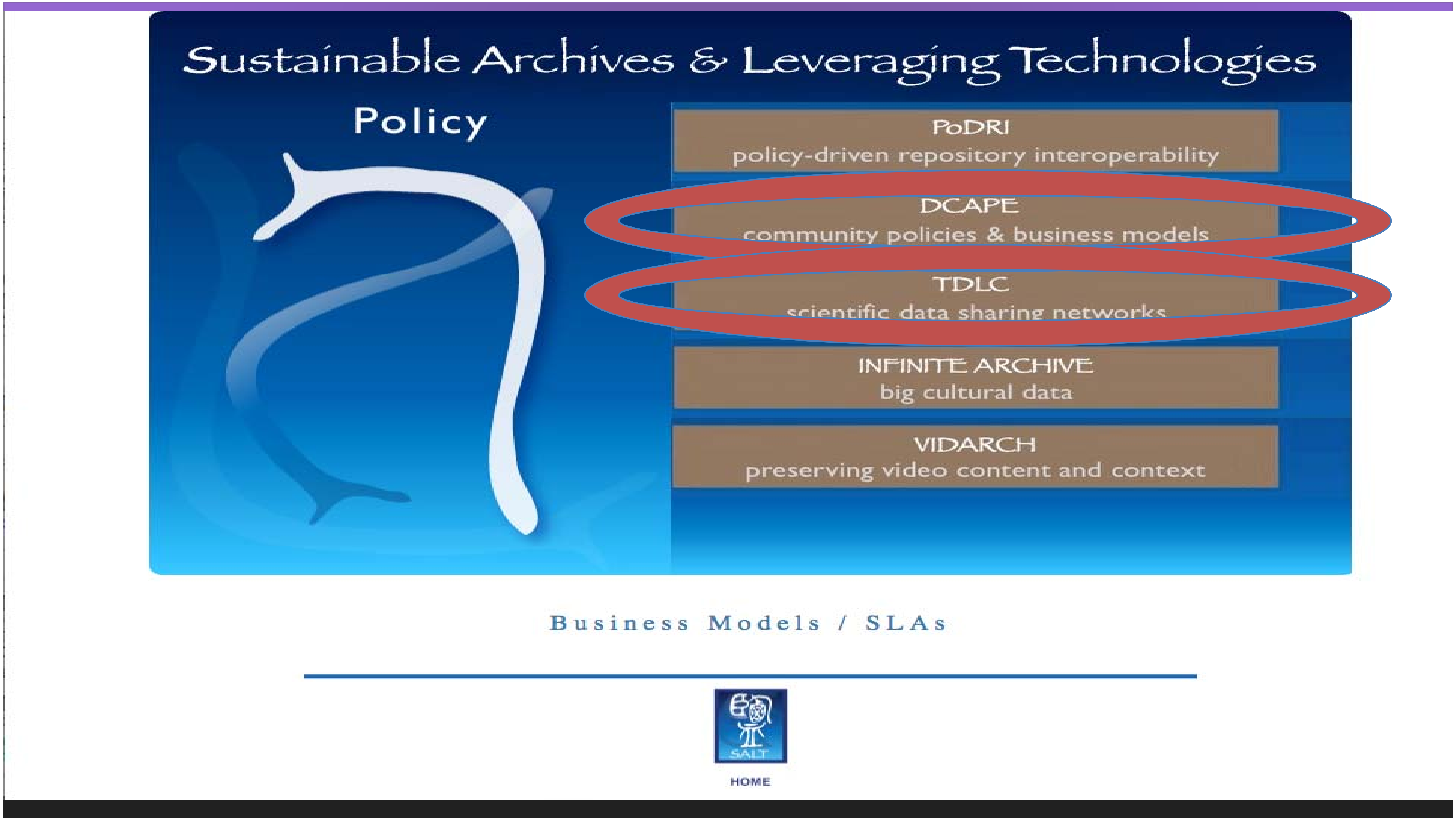
CI-BER
big data analytics

DIGARCH
digital preservation lifecycle management

data grids / digital libraries / digital mapping

SALT is an interdisciplinary group focused on developing and leveraging resources and technologies to enable collaborations.

Focusing on the interplay of content, policy, governance, and cyberinfrastructure.



Sustainable Archives & Leveraging Technologies

Policy

PoDRI
policy-driven repository interoperability

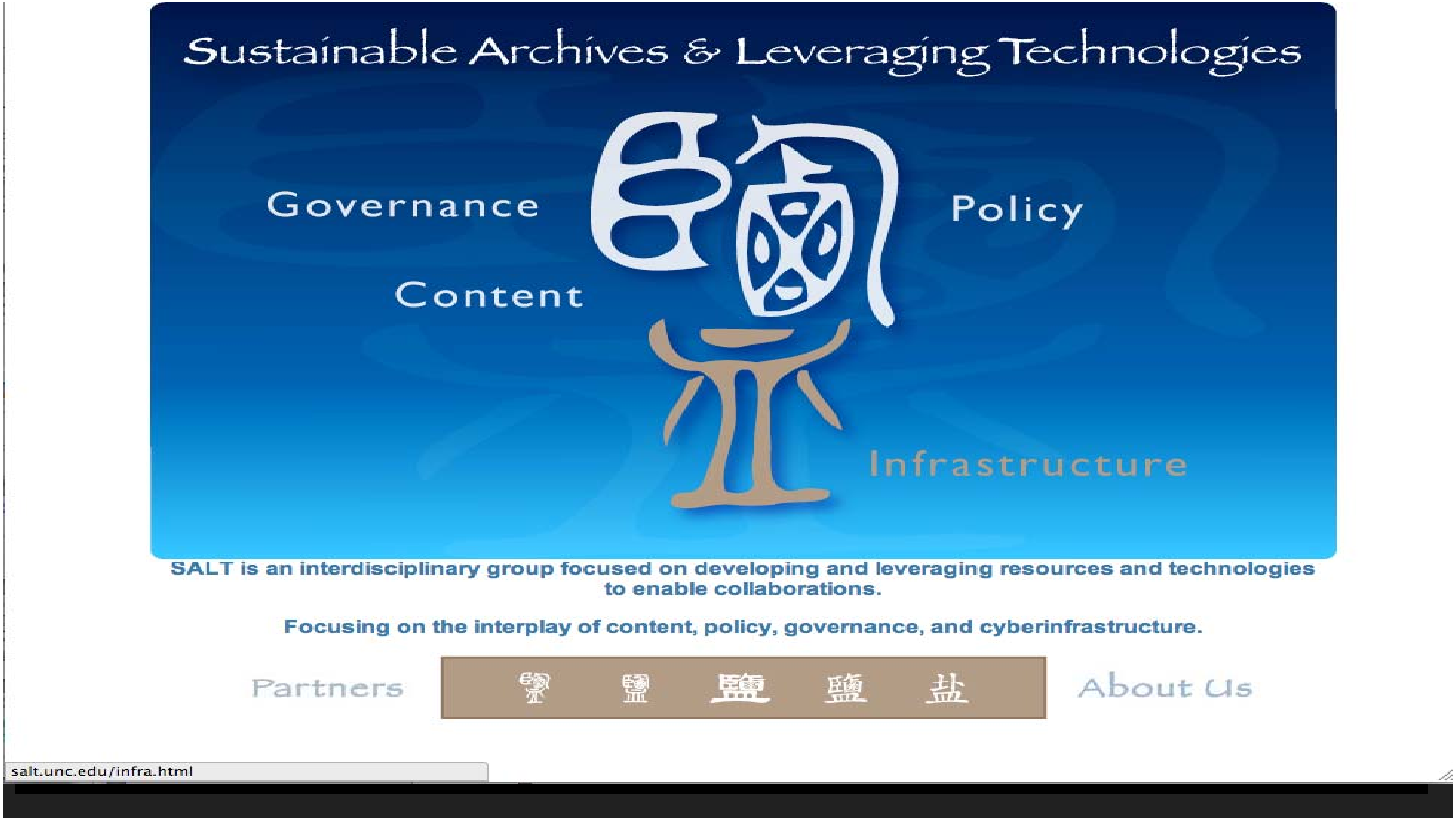
DCAPE
community policies & business models

TDLC
scientific data sharing networks

INFINITE ARCHIVE
big cultural data

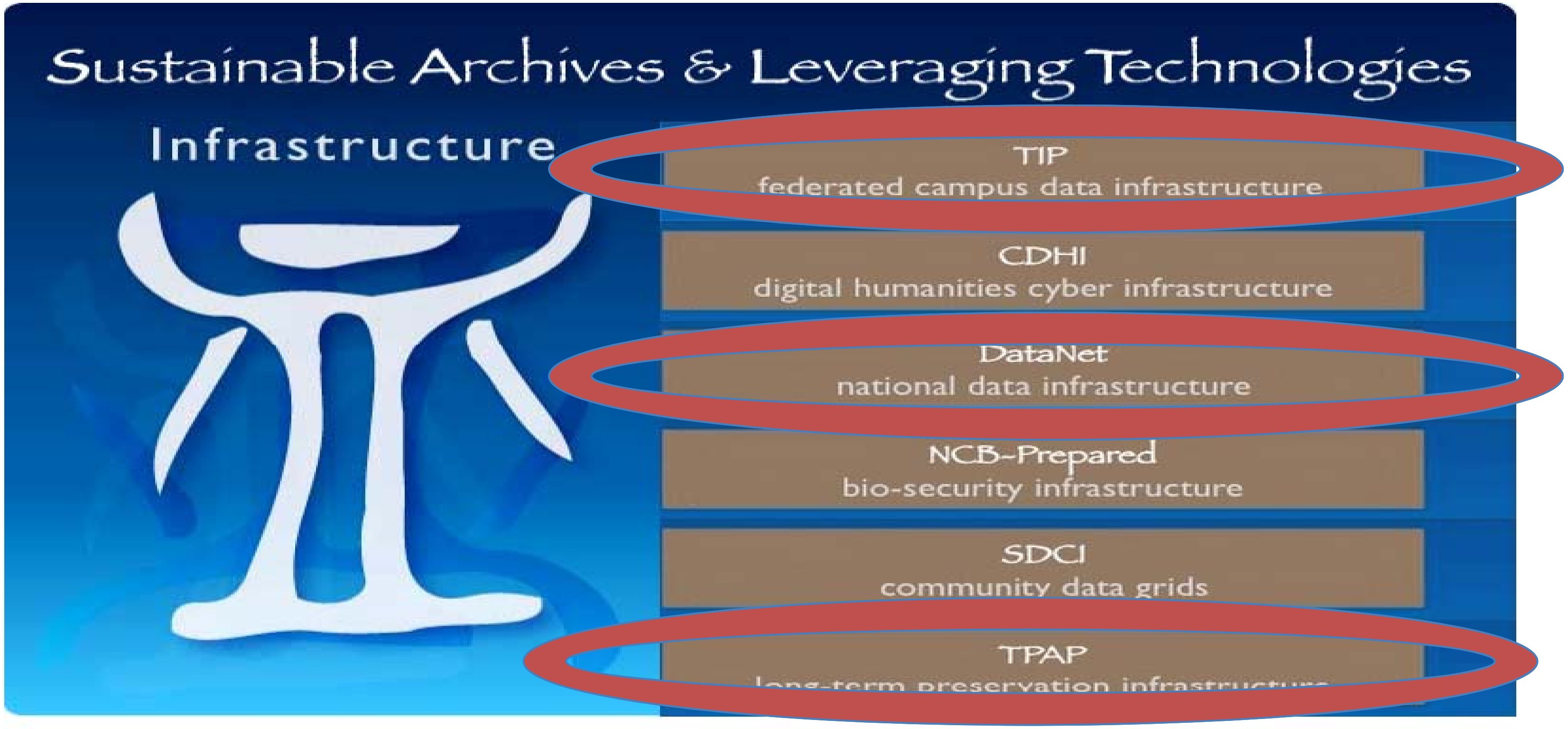
VIDARCH
preserving video content and context

Business Models / SLAs



SALT is an interdisciplinary group focused on developing and leveraging resources and technologies to enable collaborations.

Focusing on the interplay of content, policy, governance, and cyberinfrastructure.



Sustainable Archives & Leveraging Technologies

Infrastructure

TIP
federated campus data infrastructure

CDHI
digital humanities cyber infrastructure

DataNet
national data infrastructure

NCB-Prepared
bio-security infrastructure

SDCI
community data grids

TPAP
long-term preservation infrastructure

federation

SALT

HOME

Beyond iRODS

NEW!

DRAS-TIC

Digital Repository At Scale - That Invites Computation [To Improve Collections]

GOAL: Build out the open source DRAS-TIC platform into a horizontally scalable archives framework **serving the national library, archives, and scientific data management communities**

- Product of a 2-year startup by partners, Archive Analytics Solutions Ltd.
- Scaling to billions of files and beyond
- Interfaces:
  - Web client
  - Command-line client
  - REST storage API (CDMI) industry standard
- Key-value metadata
- Listener mechanism
- Python source on GitHub (Open AGPL license)
- Apache Cassandra database (CERN, eBay, GitHUB, Hulu, Instagram, Netflix, Twitter…)

- Computational Finding Aids

DRASTIC Measures

Designing Scalable Cyberinfrastructure for Metadata Extraction in Billion-Object Archives

Gregory Jansen
Richard Marciano

UNIVERSITY OF MARYLAND

dcic digital curation innovation center

NCSA

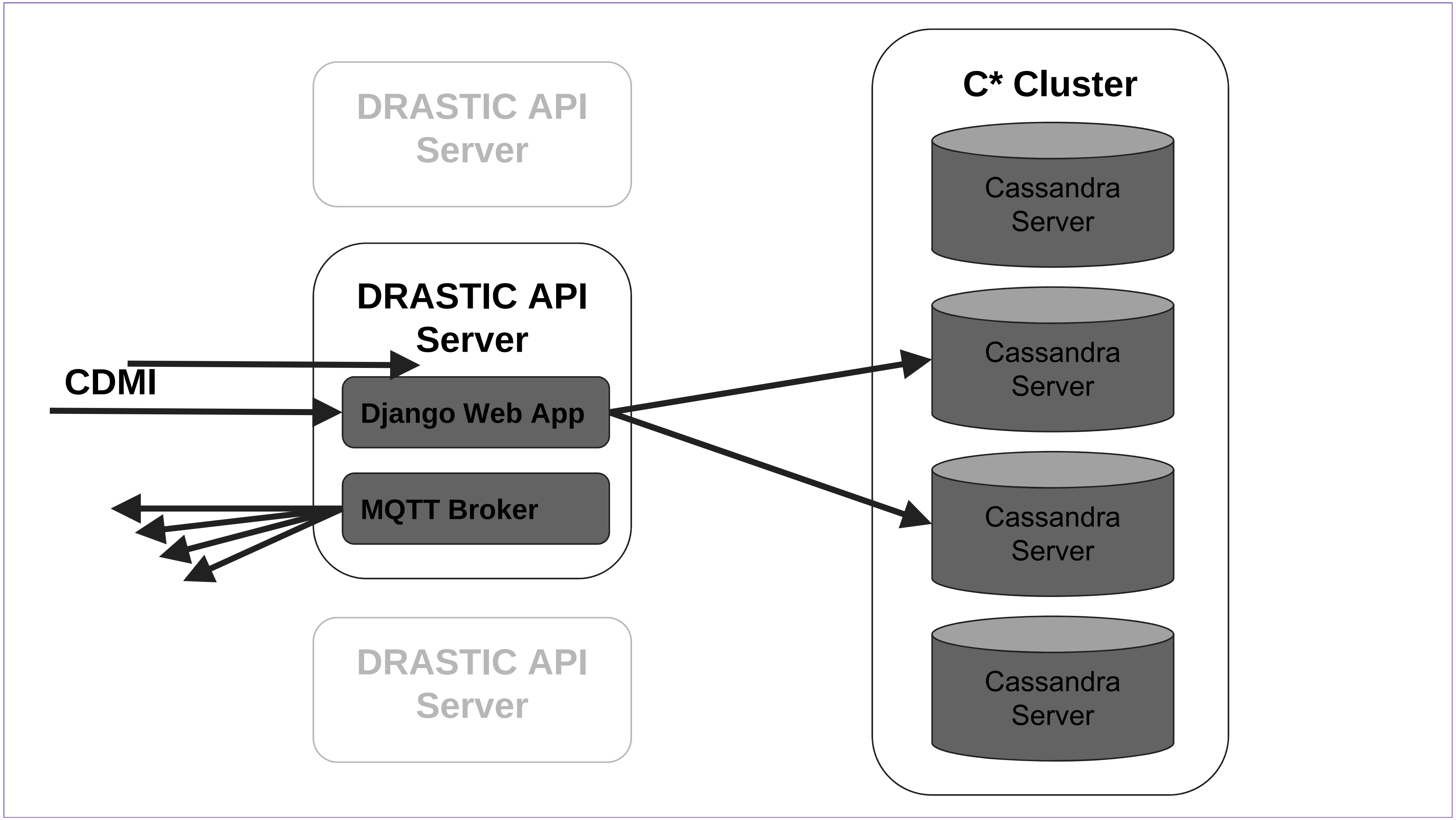
The Rest of the Talk..

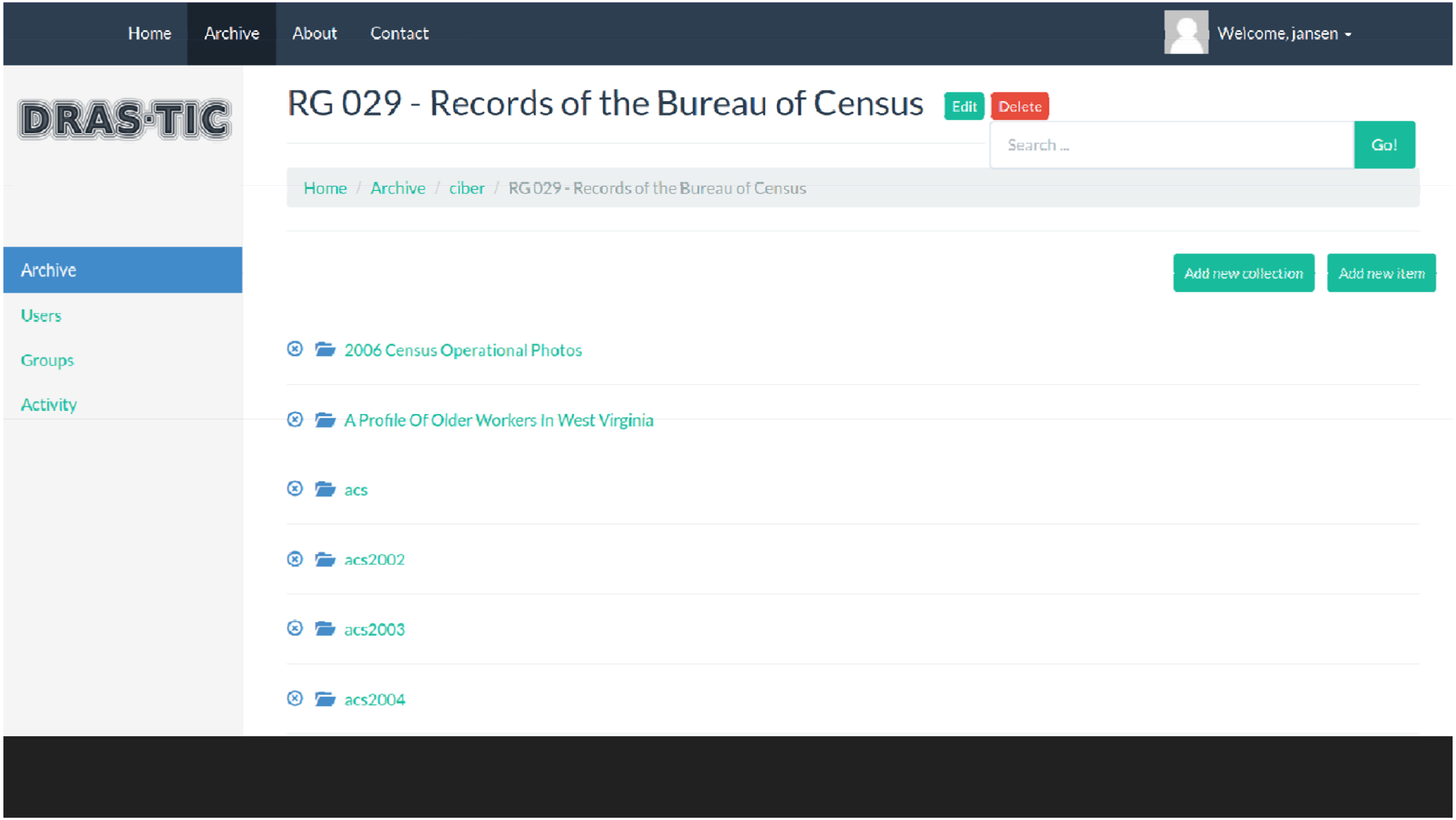
- Approaching 1 Billion files
- New DRAS-TIC Repository
- NCSA’s Brown Dog Service
- Automatic Feature Extraction & Curation
- Digging into Collections with Elasticsearch
- Projects & Opportunities

Approaching Billions at 1/10 Scale

100 Million files
72 Terabytes of data
Hundreds of file formats
Unique file formats

4 x 32 core servers
15 trays of hard drives
180 4 Terabyte drives
720 Terabytes raw storage


DRASTIC API Server
C* Cluster
Cassandra Server
DRASTIC API Server
CDMI
Django Web App
Cassandra Server
MQTT Broker
Cassandra Server
DRASTIC API Server
Cassandra Server

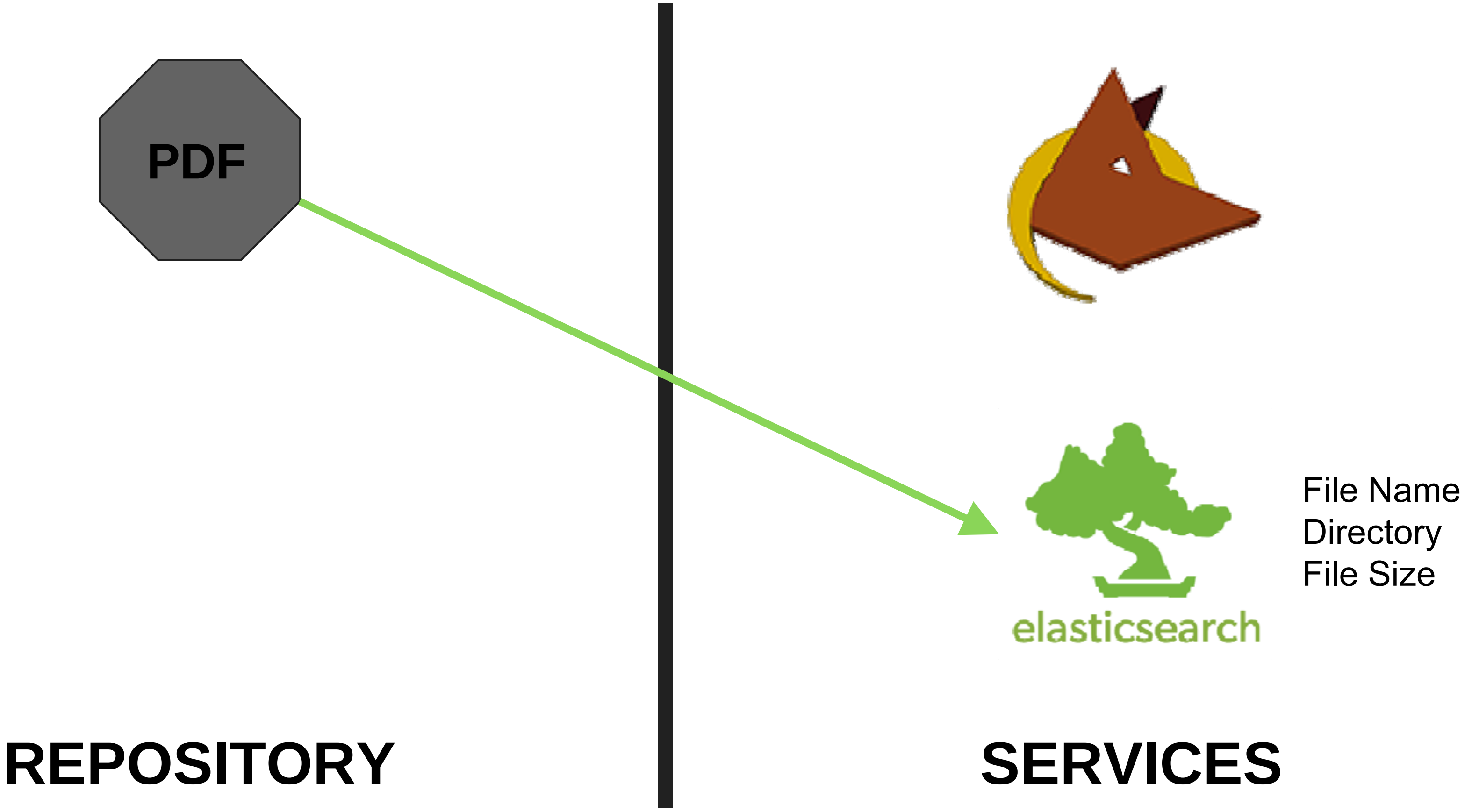


RG 029 - Records of the Bureau of Census

Home / Archive / ciber / RG 029 - Records of the Bureau of Census

- 2006 Census Operational Photos
- A Profile Of Older Workers In West Virginia
- acs
- acs2002
- acs2003
- acs2004

Workflow for a Digital Object



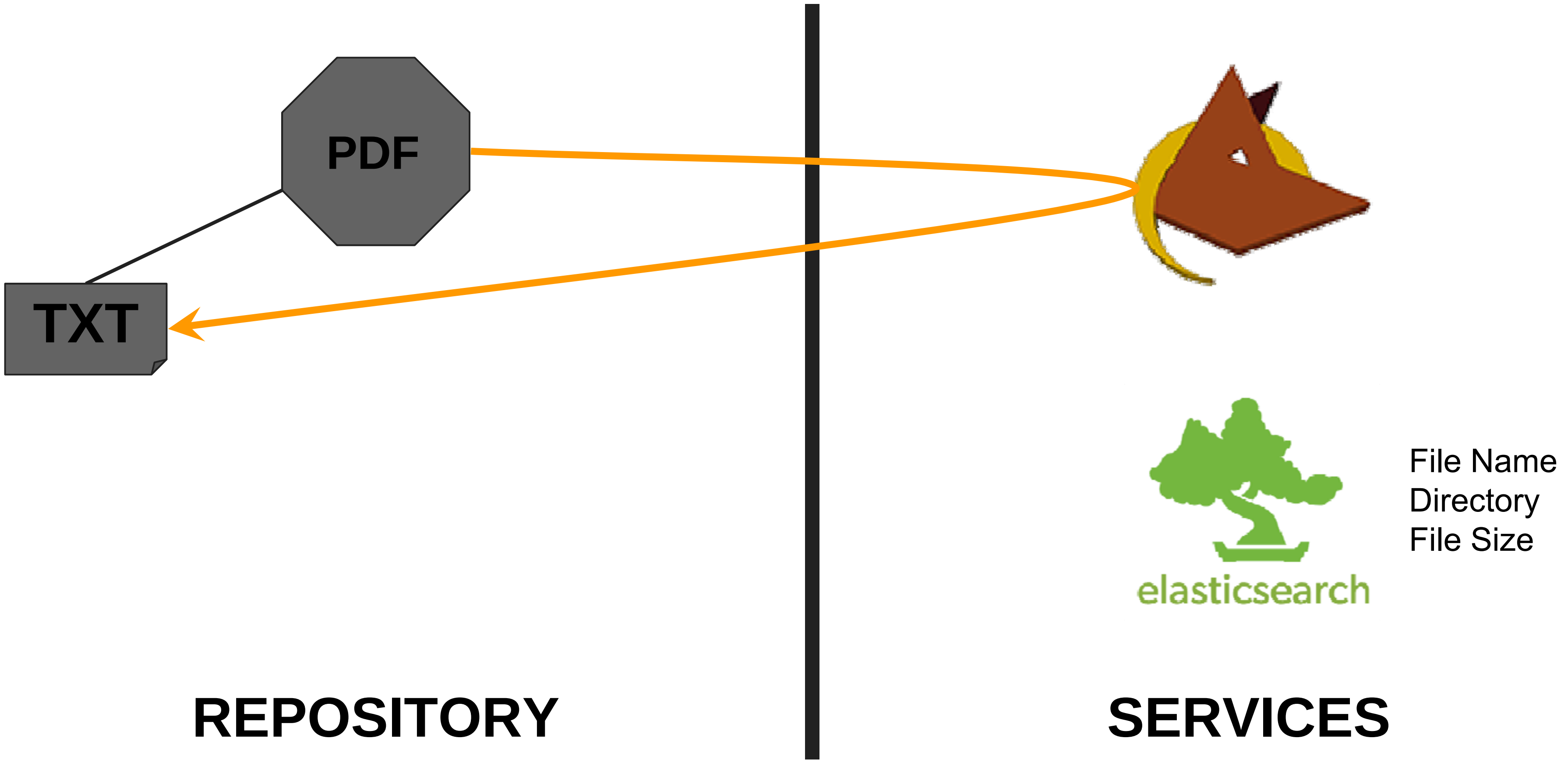
PDF

File Name
Directory
File Size

REPOSITORY

SERVICES

Text Format Conversion (PDF to TXT)



PDF

TXT

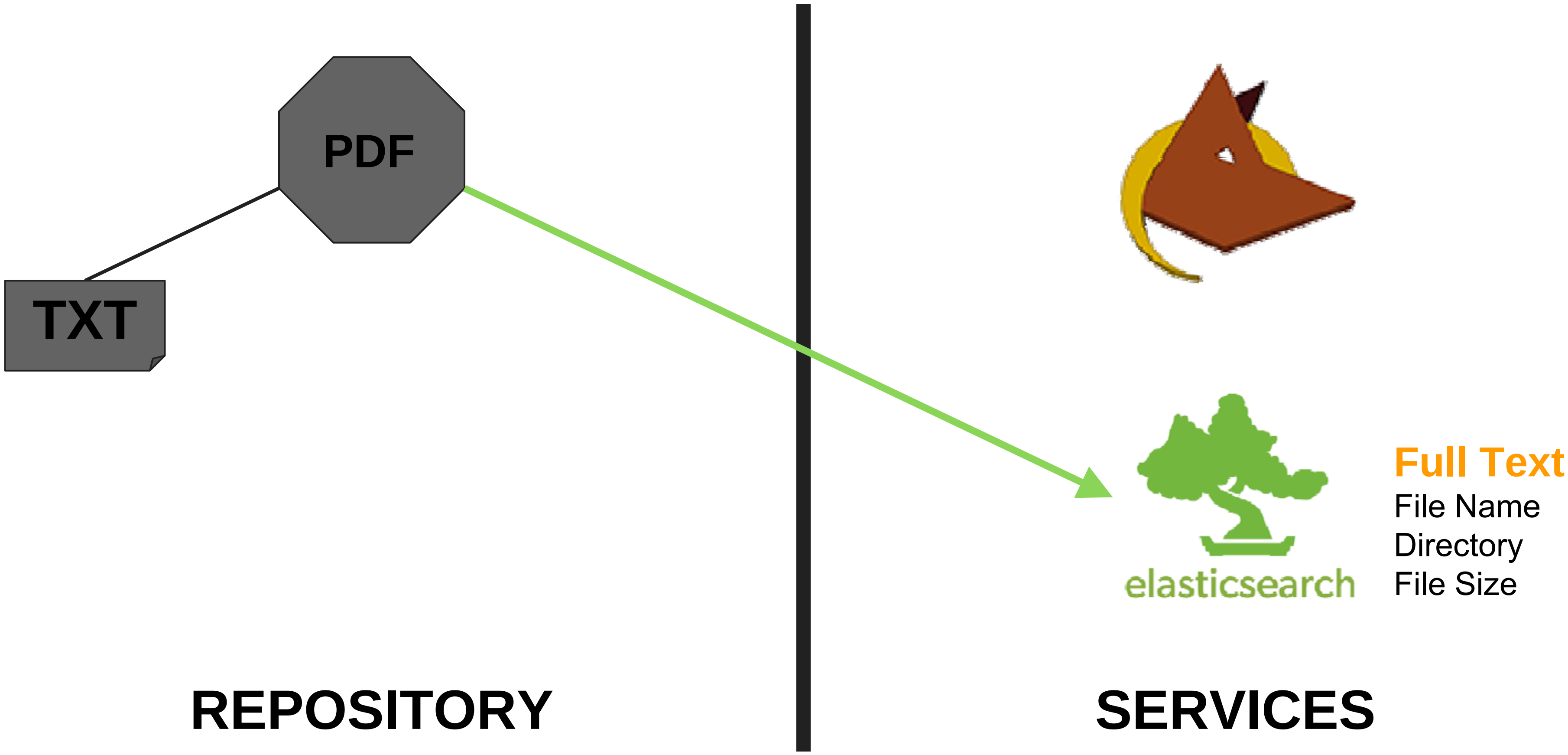
elasticsearch

File Name
Directory
File Size

REPOSITORY

SERVICES

Now we have a full text index..



PDF

TXT

Full Text

File Name

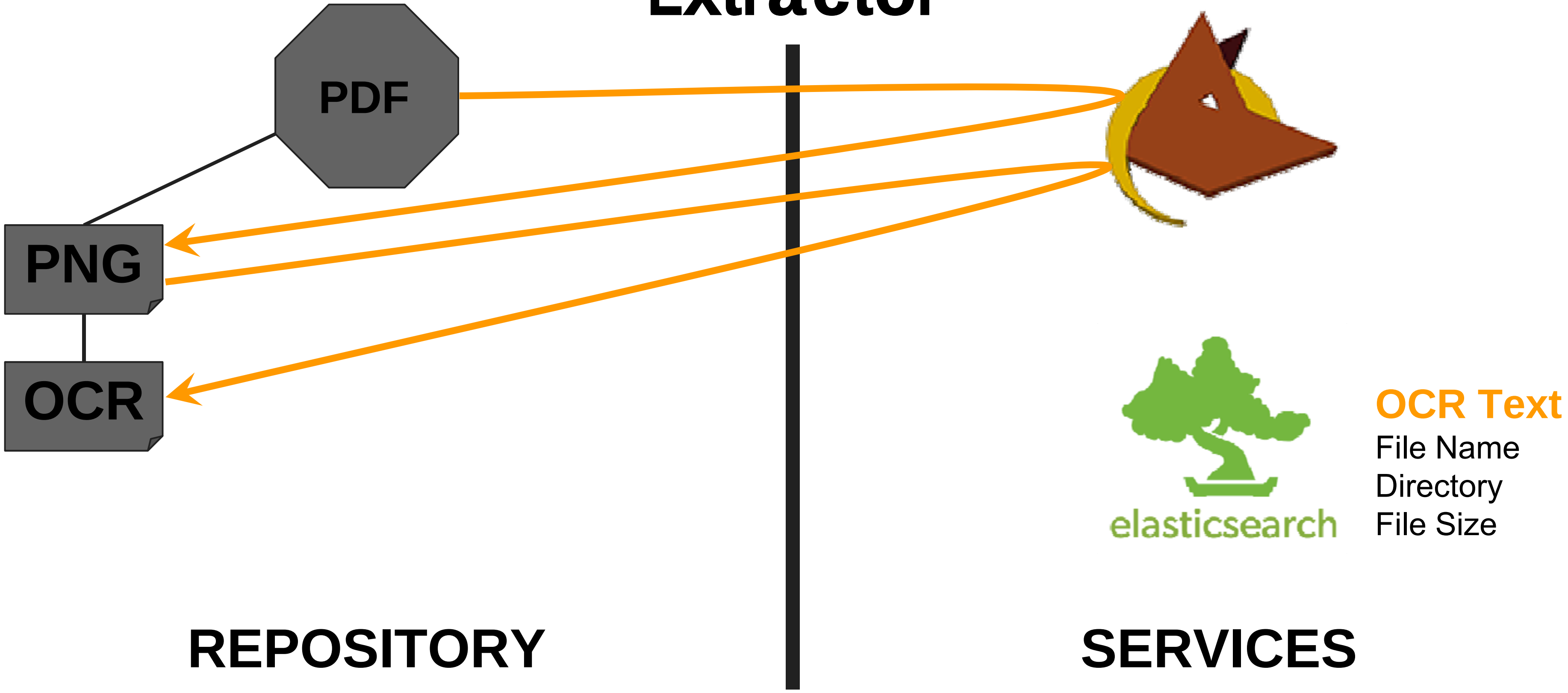
Directory

File Size

REPOSITORY

SERVICES

Optical Character Recognition (OCR) Extractor



PDF

PNG

OCR

OCR Text

File Name

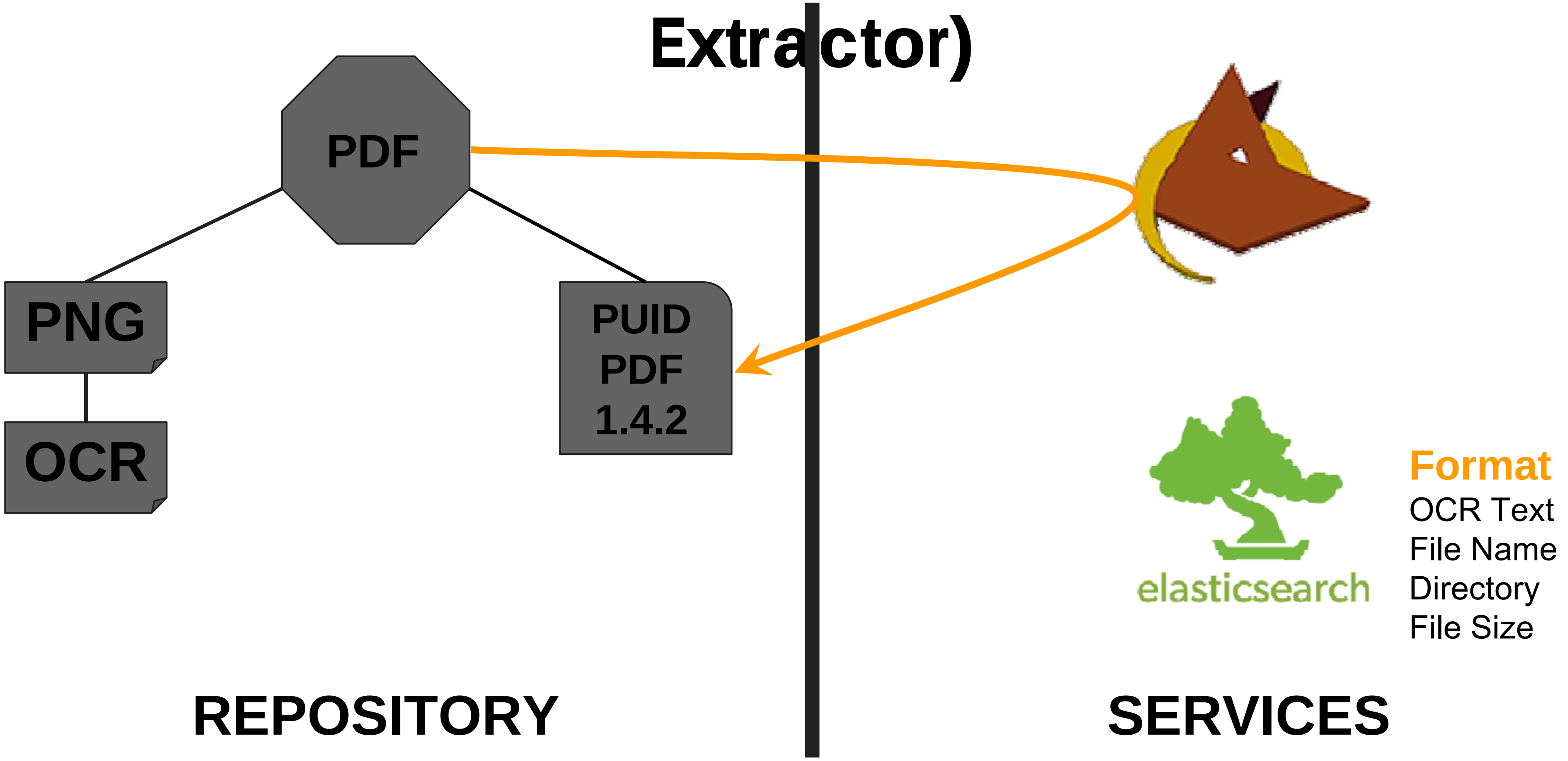
Directory

File Size

REPOSITORY

SERVICES

Format Recognition (Siegfried PRONOM Extractor)



PDF

PNG

OCR

PUID
PDF
1.4.2

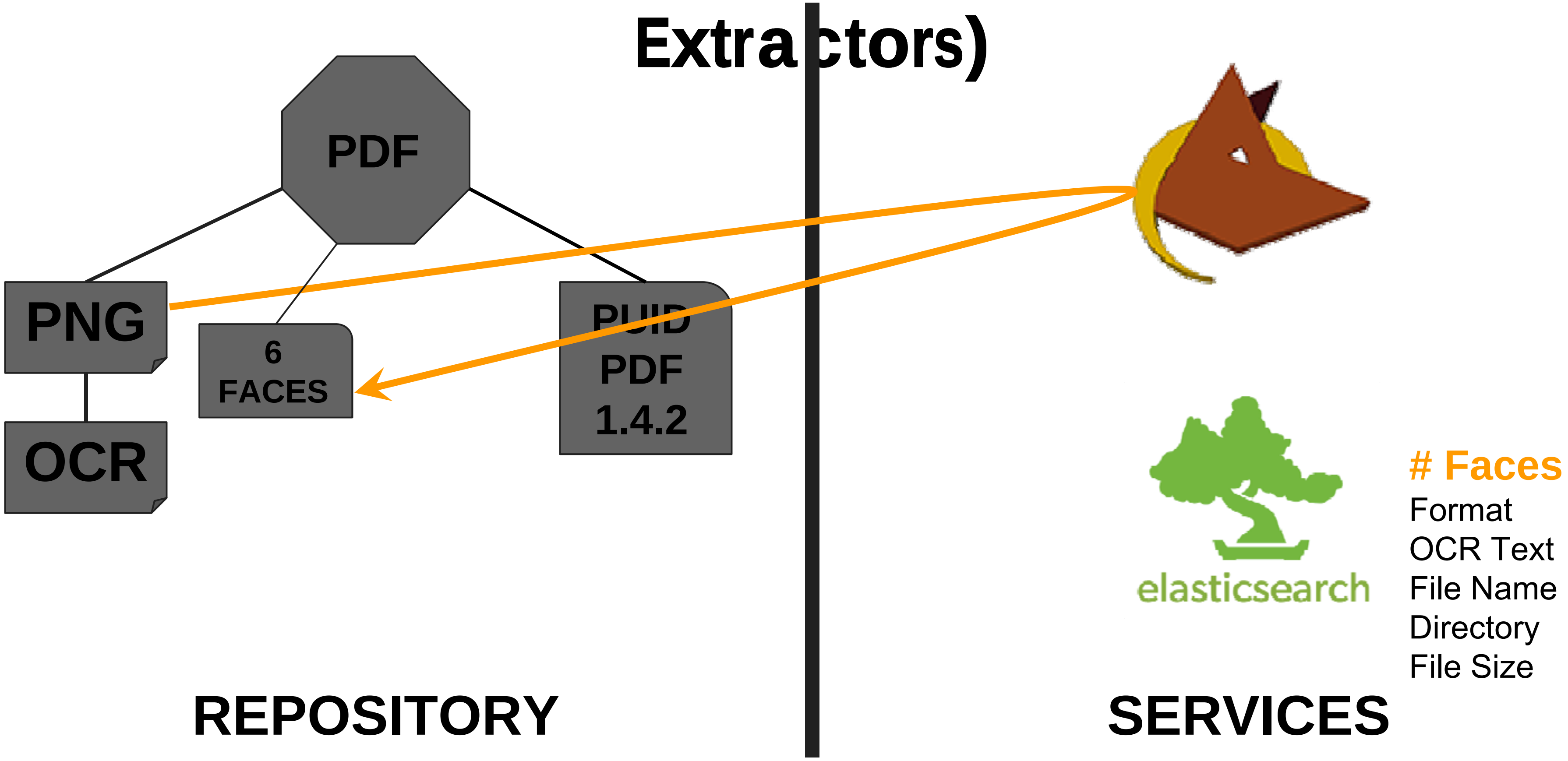
elasticsearch

Format
OCR Text
File Name
Directory
File Size

REPOSITORY

SERVICES

Facial Recognition (Computer Vision Extractors)

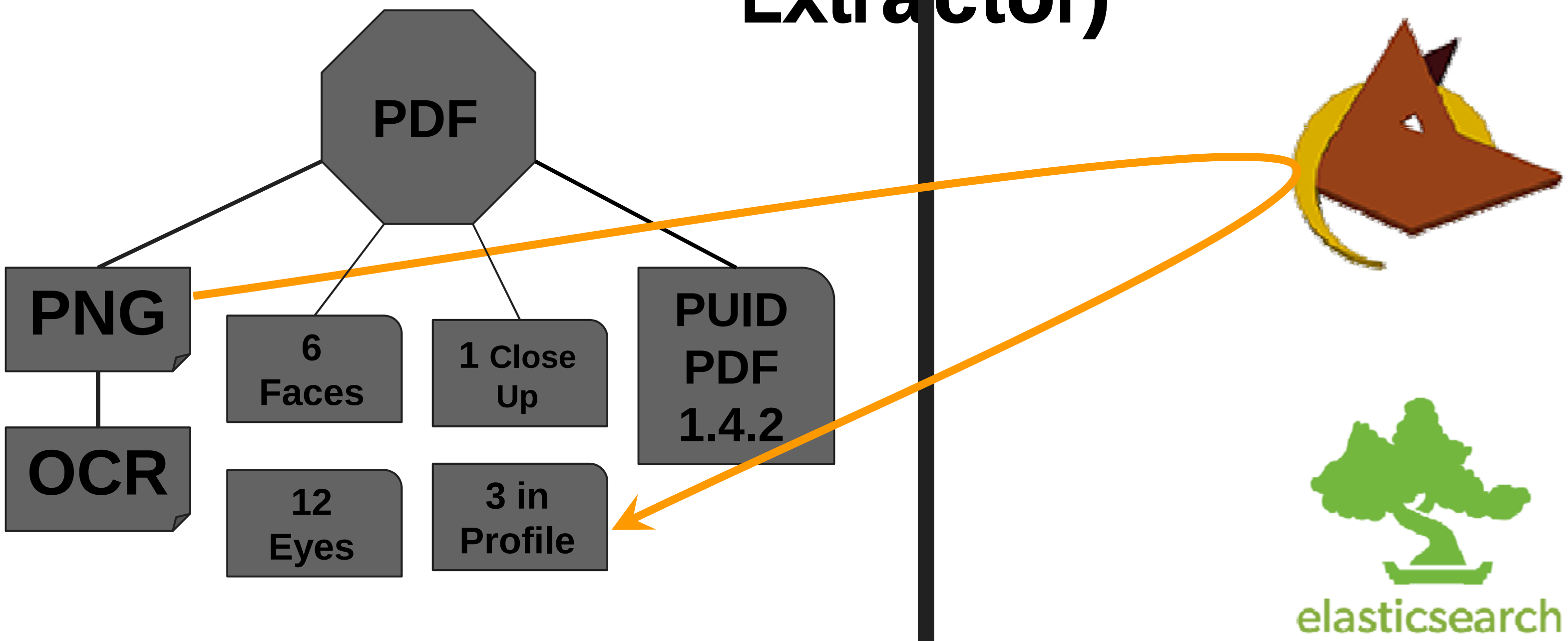


# Faces
Format
OCR Text
File Name
Directory
File Size

REPOSITORY

SERVICES

Facial Recognition (Computer Vision Extractor)



PDF

PNG

OCR

6 Faces

12 Eyes

1 Close Up

3 in Profile

PUID PDF 1.4.2

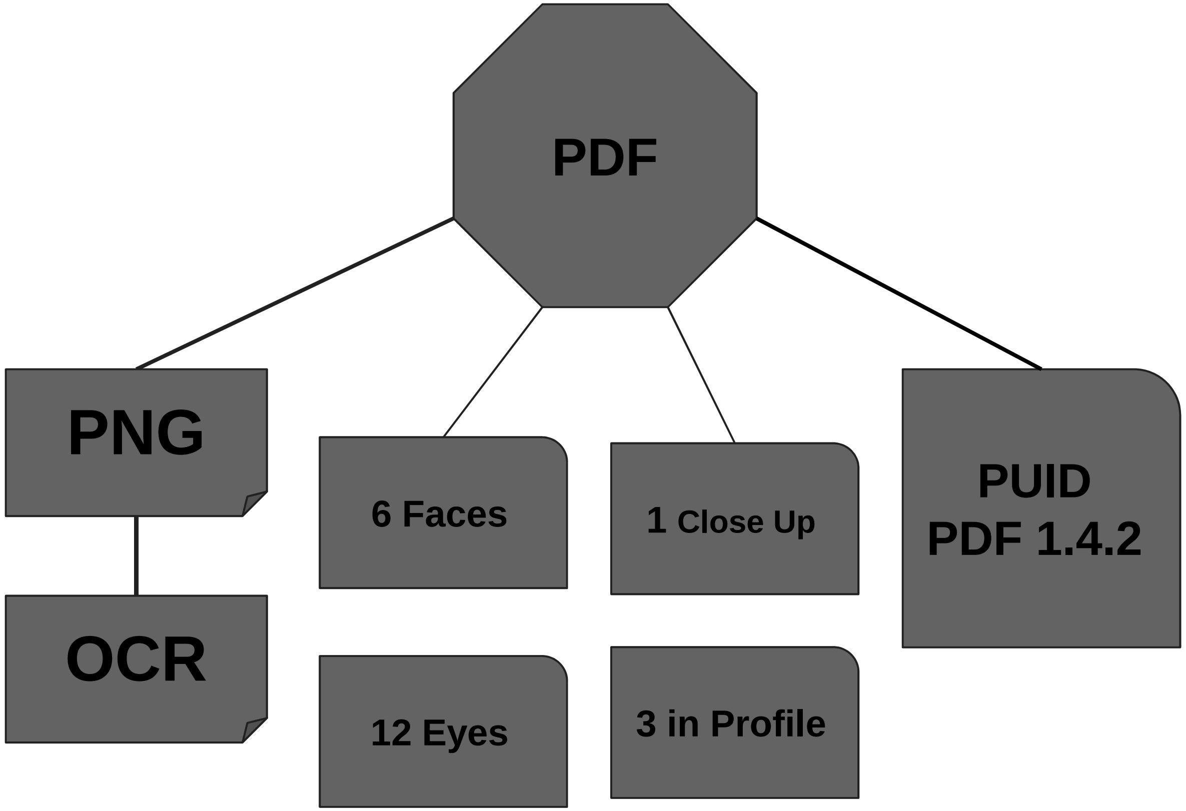
elasticsearch

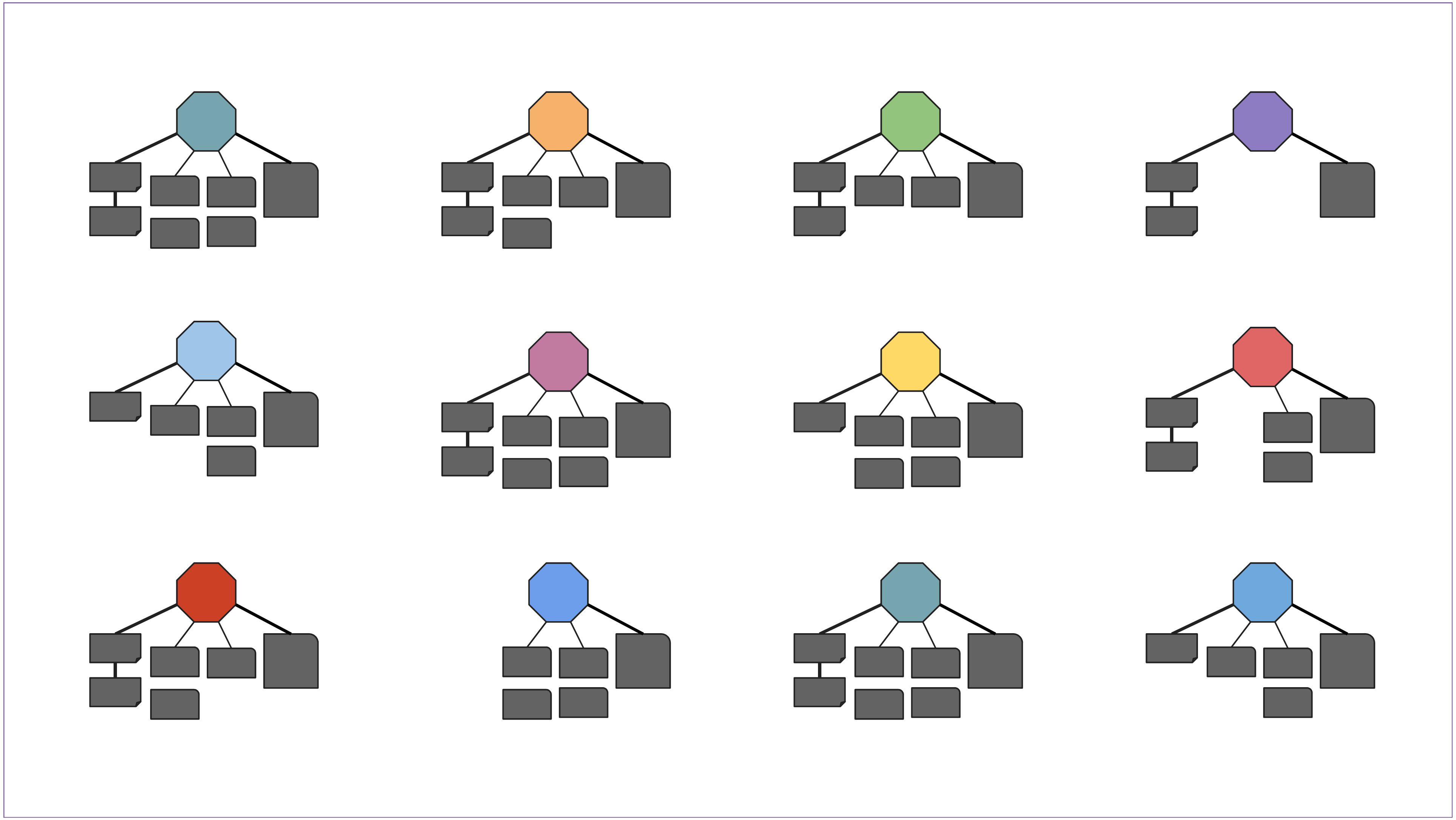
# Faces
# Eyes
# Close Ups
# Profiles
Format
OCR Text
File Name
Directory
File Size

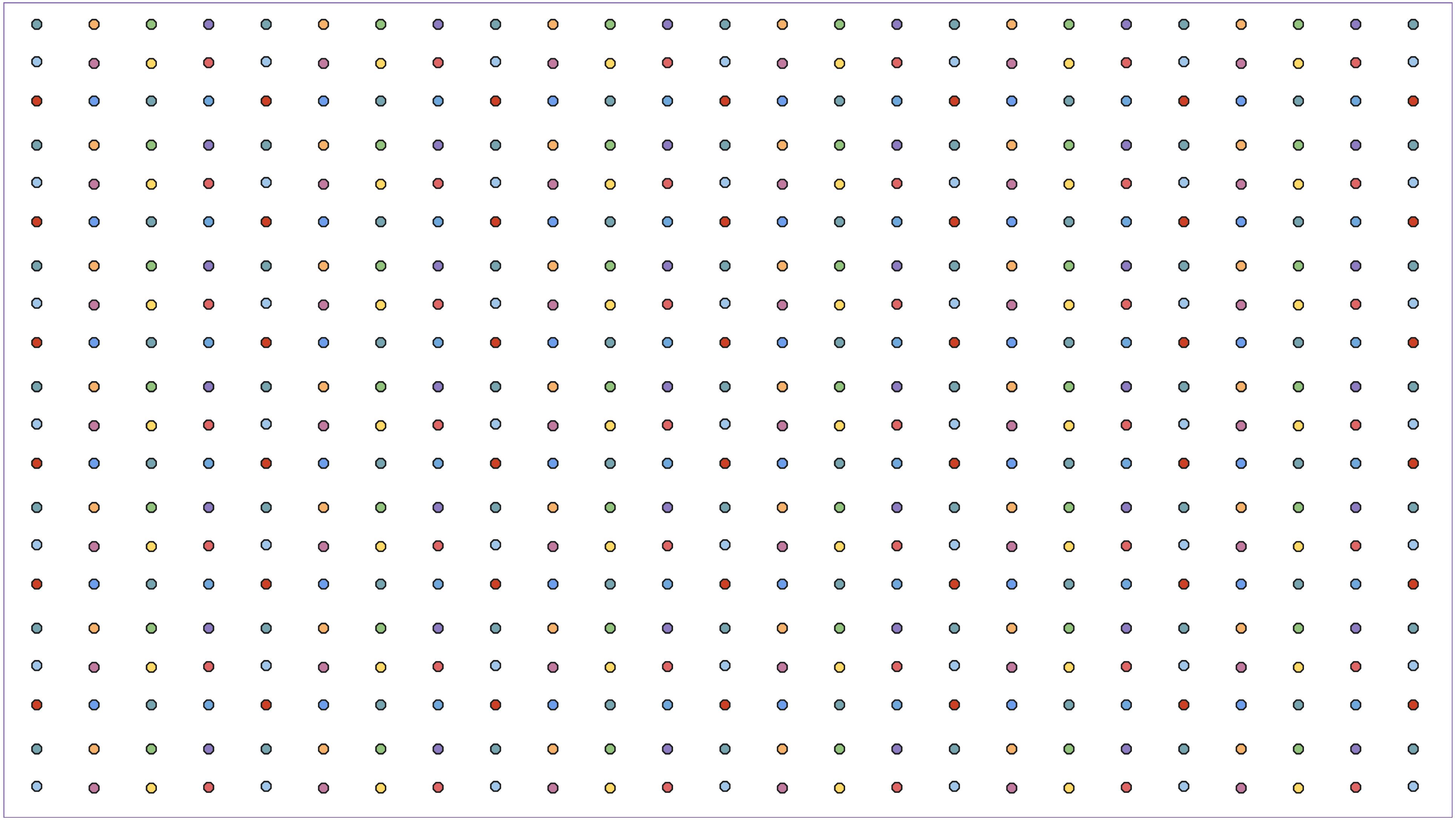
REPOSITORY

SERVICES

PDF Object Enhanced with Extracted Metadata





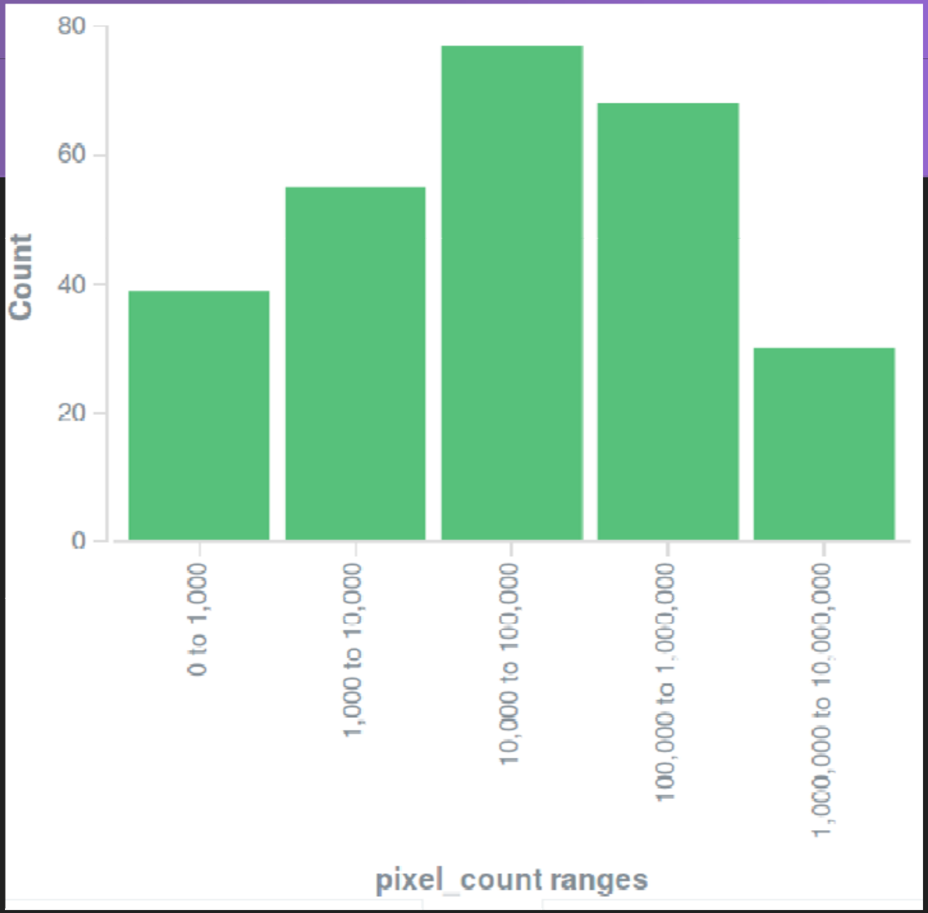


DON'T

PANIC

ElasticSearch + Kibana

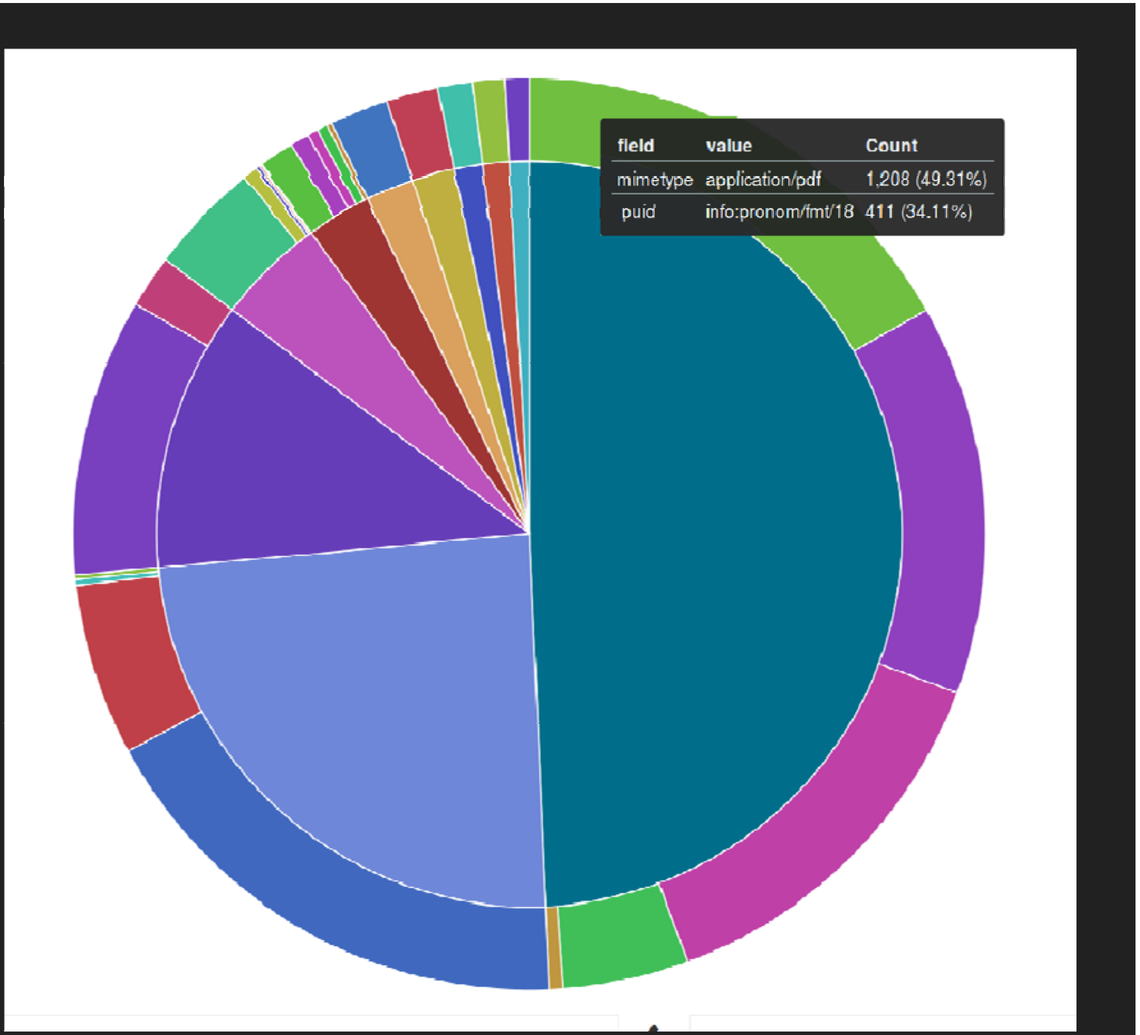
- Free plugin for Elasticsearch
- Gives shape to an Elasticsearch index
- Write queries visually and interactively



Lots of ways to explore the data

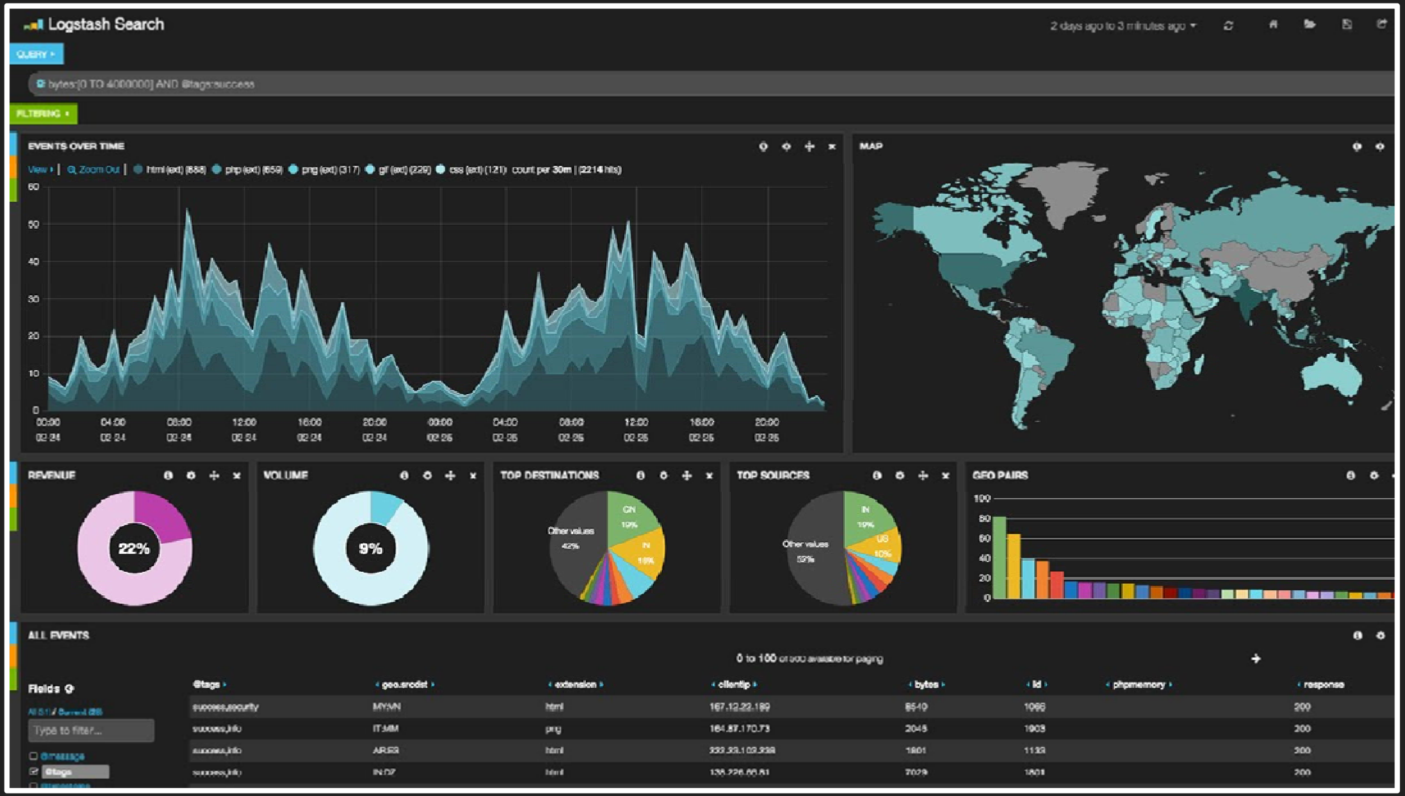
Files Formats

- ○ Concentric Pie Chart
- ○ Inner: Mimetype
- ○ Outer: PRONOM PUID



field	value	Count
mimetype	application/pdf	1,208 (49.31%)
puid	info:pronom/fmt/18	411 (34.11%)

Charts can be added to data dashboards..



bytes:[0 TO 4000000] AND @tags:success

0 to 100 of 500 available for paging

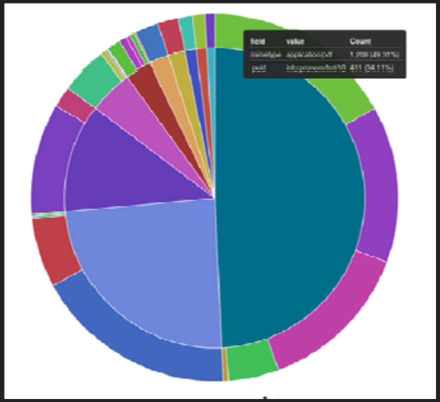
@tags	geo.srcdest	extension	clientip	bytes	id	phpmemory	response
success,security	MY:VN	html	167.12.22.189	8540	1066		200
success,info	IT:MM	png	164.87.170.73	2045	1903		200
success,info	AR:EG	html	233.33.103.238	1801	1133		200
success,info	IN:DZ	html	136.226.66.81	7029	1801		200

Arrangement can be used as a Facet

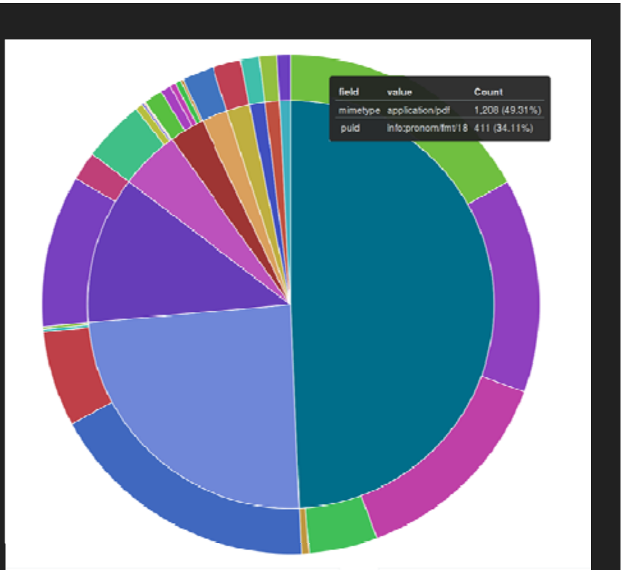
As you browse the hierarchy...

The entire dashboard is redrawn to reflect the particular record group, series or folder under study.

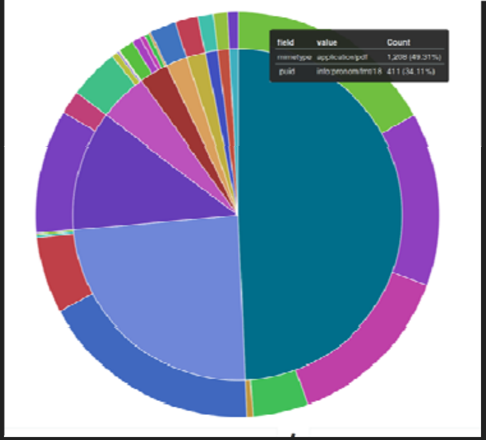
“Drill down” or zoom in and out of your collections.



field	value	Count
mimetype	application/pdf	1,208 (49.31%)
puid	info:pronom/fmt/18	411 (34.11%)



field	value	Count
mimetype	application/pdf	1,208 (49.31%)
puid	info:pronom/fmt/18	411 (34.11%)



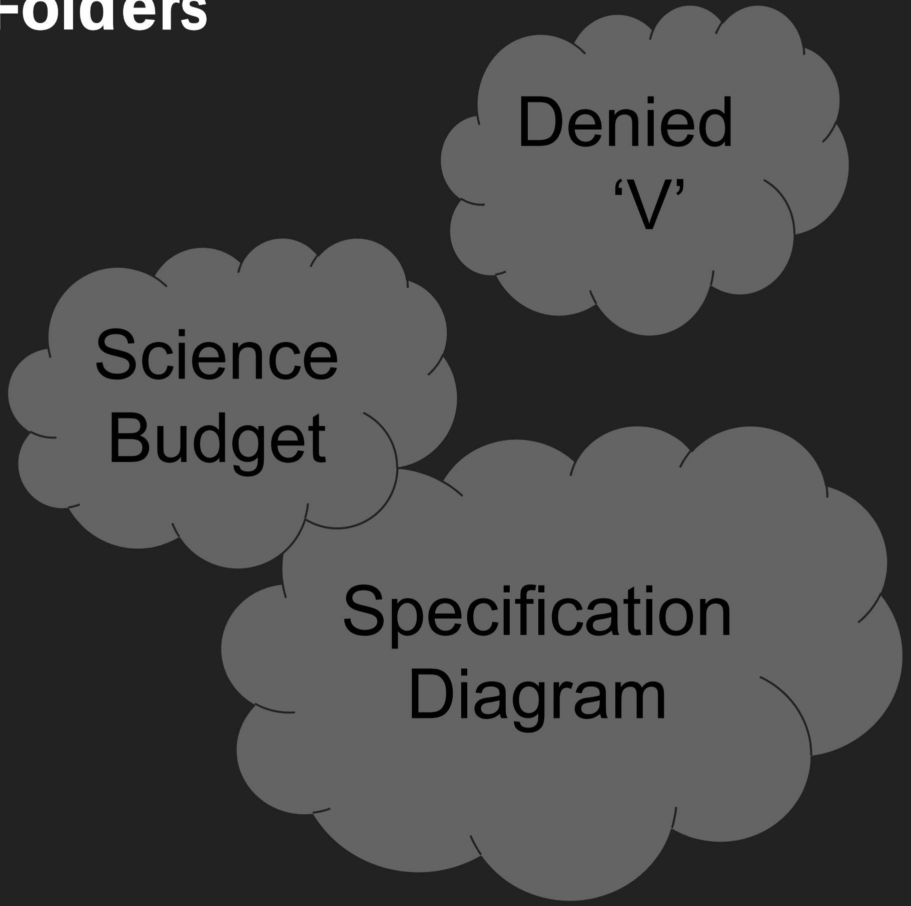
field	value	Count
mimetype	application/pdf	1,208 (49.31%)
puid	info:pronom/fmt/18	411 (34.11%)

Text Comparison between Folders

Significant Terms are based on full text.

They are significant within overall scope of query.

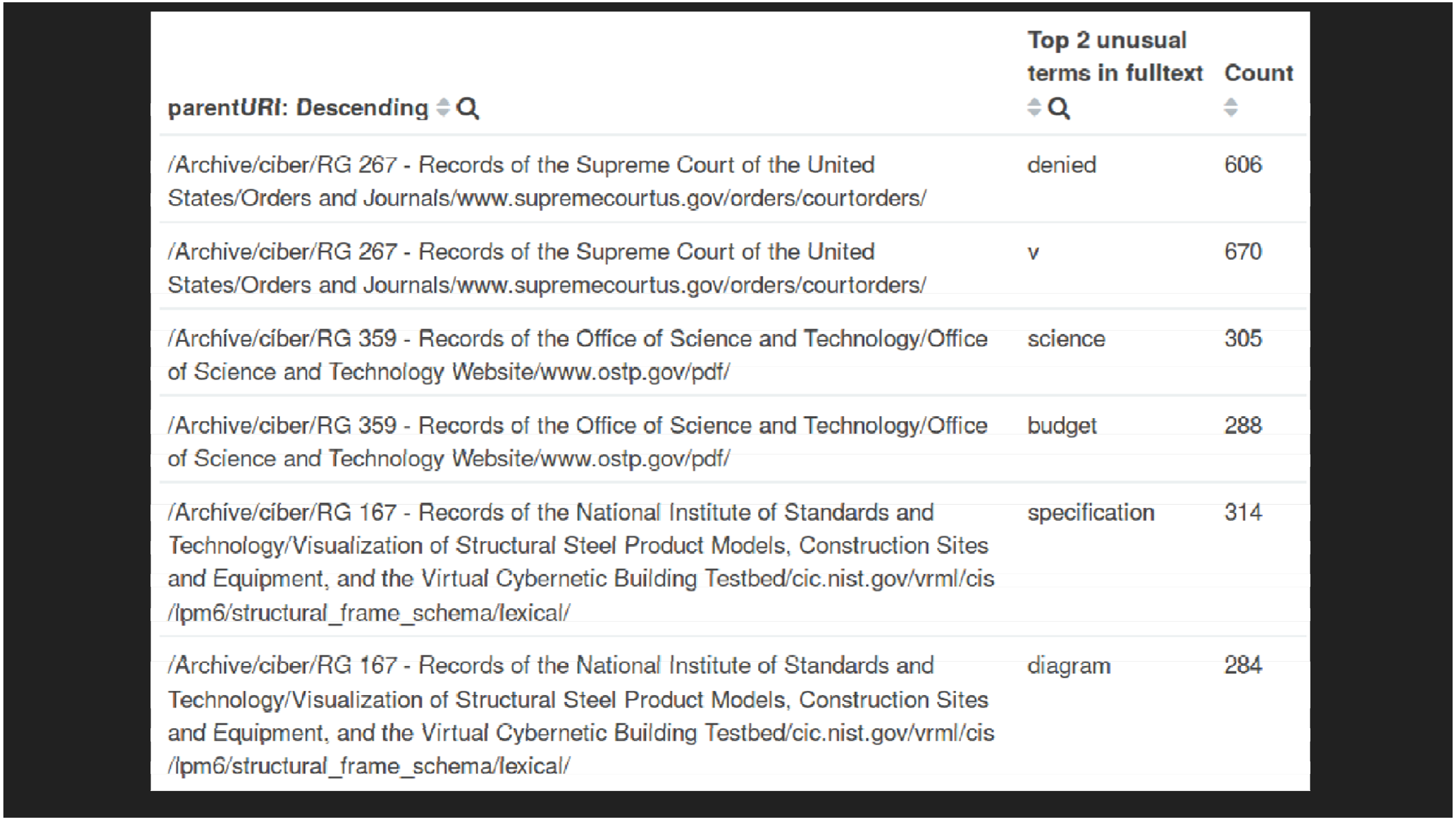
Significant Terms can be used to distinguish neighboring folders or documents.



Denied ‘V’

Science Budget

Specification Diagram



parentURI: Descending ⇕ 🔍	Top 2 unusual terms in fulltext ⇕ 🔍	Count ⇕
/Archive/ciber/RG 267 - Records of the Supreme Court of the United States/Orders and Journals/www.supremecourtus.gov/orders/courtorders/	denied	606
/Archive/ciber/RG 267 - Records of the Supreme Court of the United States/Orders and Journals/www.supremecourtus.gov/orders/courtorders/	v	670
/Archive/ciber/RG 359 - Records of the Office of Science and Technology/Office of Science and Technology Website/www.ostp.gov/pdf/	science	305
/Archive/ciber/RG 359 - Records of the Office of Science and Technology/Office of Science and Technology Website/www.ostp.gov/pdf/	budget	288
/Archive/ciber/RG 167 - Records of the National Institute of Standards and Technology/Visualization of Structural Steel Product Models, Construction Sites and Equipment, and the Virtual Cybernetic Building Testbed/cic.nist.gov/vrml/cis /lpm6/structural_frame_schema/lexical/	specification	314
/Archive/ciber/RG 167 - Records of the National Institute of Standards and Technology/Visualization of Structural Steel Product Models, Construction Sites and Equipment, and the Virtual Cybernetic Building Testbed/cic.nist.gov/vrml/cis /lpm6/structural_frame_schema/lexical/	diagram	284

JOIN FORCES

DRAS-TIC

Institutional R&D Partners
Use cases for Parallel Compute
Fedora Sprinters

Brown Dog

Try it on your Scientific Data
Become an Early Adopter of the API
Contribute Extractors & Converters

UMD iSchool

Partner with the DCIC on Projects
Digital Curation Certificate Program
Computational Archival Science

http://dcic.umd.edu
http://github.com/UMD-DRASTIC
http://browndog.ncsa.illinois.edu

marciano@umd.edu

Thank you ISGC 2017 Team:

Ludek, Simon, Stella, Vicky,

and many other staff