Atlas: Data Sharing in HIV Research

Sarah Ramsay
CHAVI/VISC/CDS Program Manager,
SCHARP
sramsay@scharp.org

Statistical Center for
HIV/AIDS Research & Prevention
SCHARP

FRED **HUTCHINSON**
CANCER RESEARCH **CENTER**
A LIFE OF SCIENCE

Agenda

- Background
- Atlas data sources
- Atlas usage and value
- Challenges
- Lessons and recommendations

Background: SCHARP

- Provides statistical collaboration to infectious disease researchers around the world
  - Includes statistical methodology and mathematical modeling research
- Collects, manages, and analyzes data from clinical trials and epidemiological studies of infectious disease
- Part of the Vaccine and Infectious Disease Division (VIDD) of the Fred Hutchinson Cancer Research Center
- Funded as the Data Management and Statistical Center for 3 large HIV research networks (MTN, HVTN, HPTN)
 - CHAVI, CHAVI ID, CAVD and more

Background: Atlas

- Goal of increasing transparency and improving operational efficiency in distributed collaborations
- Development started in July, 2005, launched in August, 2006
- Primary contributors:

Years	Network/Group	Funder	Focus
2005-2012	CHAVI	NIH (via Duke)	CRF sharing, specimen tracking
2006-2012	CAVD/VISC	BMGF	Assay tools: NAb, GPAT
2007-2008	SCHARP (Shared)	Multiple	Dev tools, APIs
2009-2012	HVTN	NIH	Admin features, study/specimen scalability
2009-2012	MTN	NIH	Full-text search
2011-2012	HPTN	NIH	Protocol-specific tools

Background: Atlas

- Originally began as part of the CHAVI and CAVD grants
  - Mandate for an online data sharing platform
- Large growth in the last 6 years of use
  - 2200 Active User Accounts
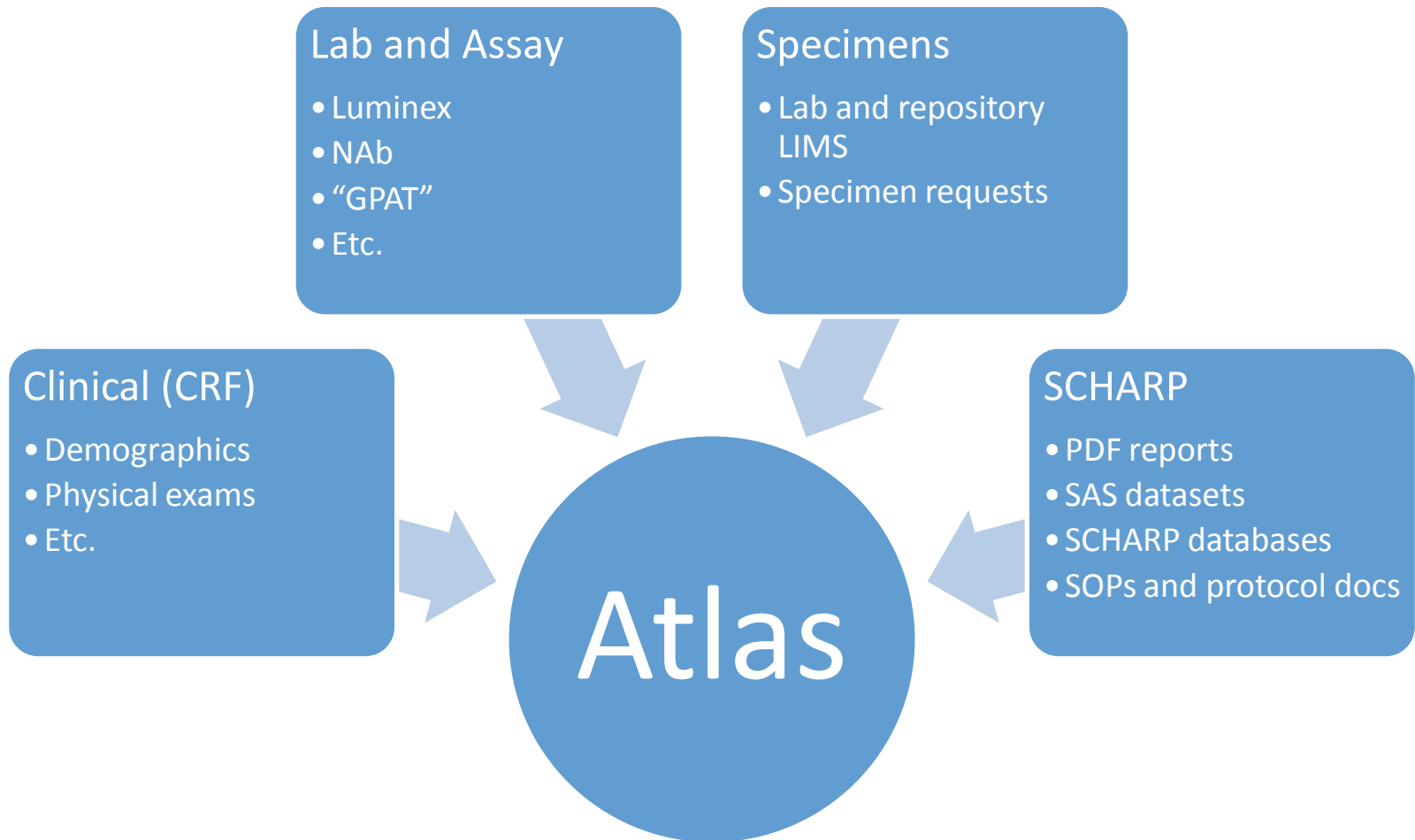 - 3800 Folders
  - 46 Projects

Background: Atlas Usage
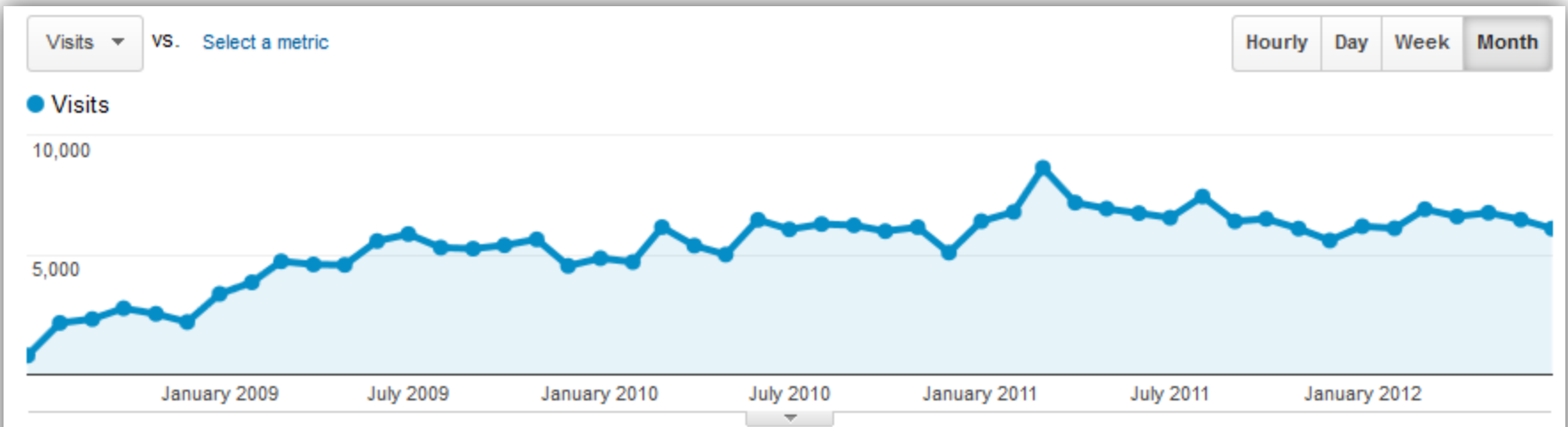
Agenda

- Background
- Atlas data sources
- Atlas usage and value
- Challenges
- Lessons and recommendations

Data sources: overview

Lab and Assay
- Luminex
- NAb
- "GPAT"
- Etc.

Specimens
- Lab and repository LIMS
- Specimen requests

Clinical (CRF)
- Demographics
- Physical exams
- Etc.

Atlas

SCHARP
- PDF reports
- SAS datasets
- SCHARP databases
- SOPs and protocol docs

Data sources: clinical

- DataFax
  - SCHARP-run and maintained
  - 42 protocols, 53,781 forms in August
  - 118,394 forms to data processed and imported for the 7 CHAVI protocols
- 3 Pipelines to Atlas
  - Datafax to Atlas
    - SCHARP-authored
    - Protocol-specific
    - Nightly import into Atlas Study Folders
    - Currently outputs a mix of TSV- and XML-based study formats
  - SAS to Atlas
    - SCHARP-authored
    - Nightly import into Atlas Study Folders
  - SAS Share
   - Direct external data source exposure in Atlas data grids

Data sources: clinical
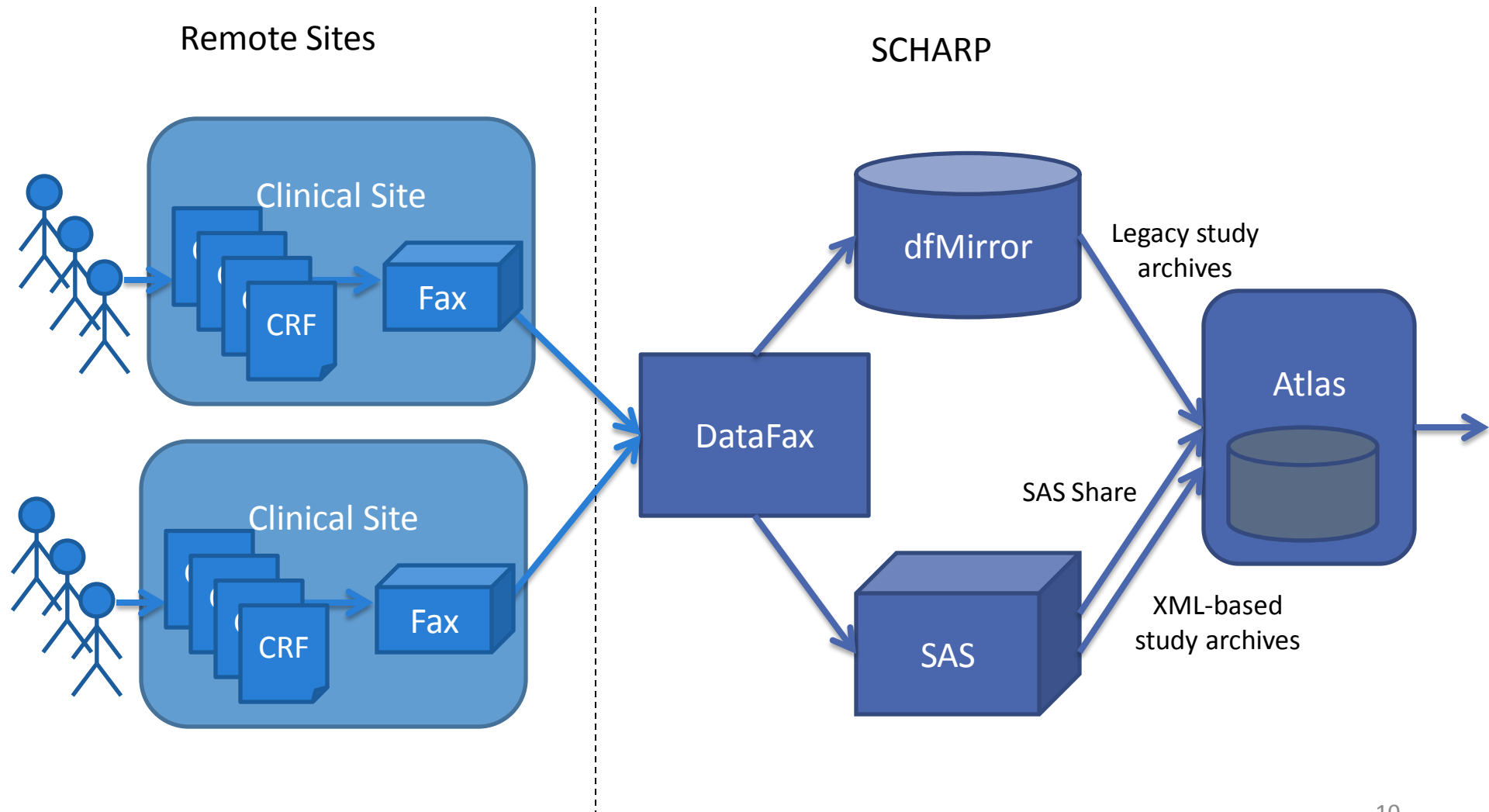
Data sources: Specimen

- Most Atlas specimen data starts with FSTRF
  - Data from > 132 location-specific LDMS installations is exported to FSTRF
  - FSTRF compiles and sends to SCHARP
- SCHARP-side pre-processing pipeline
  - Quality control checks
  - Data normalization
  - Exports to per-protocol or per-network LabKey Server specimen archives
- Data reloaded nightly into Atlas study folders

Data sources: specimen loading
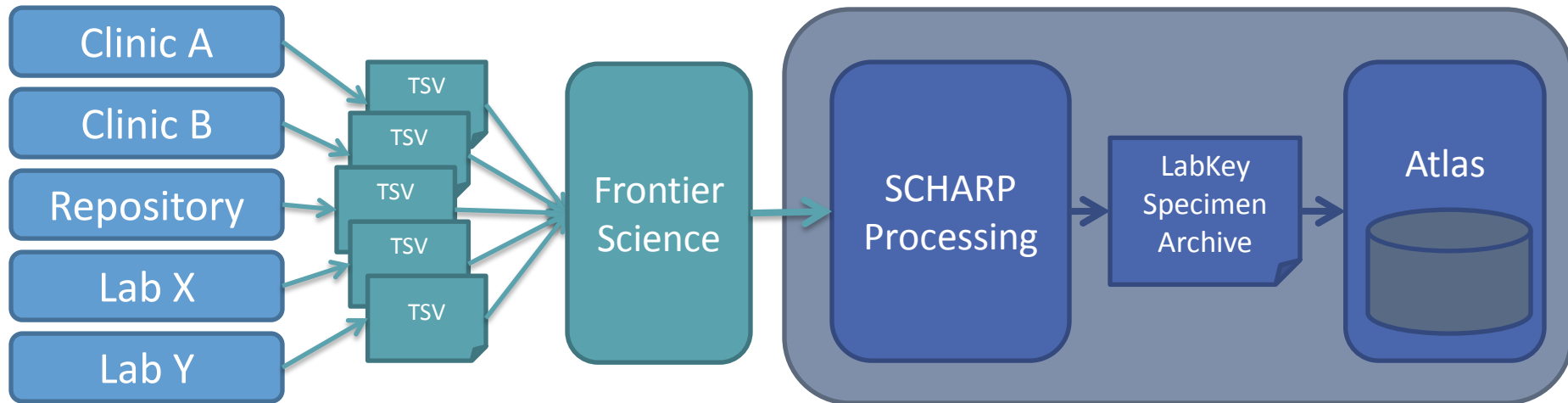
1. Data is exported from each location's LDMS system to FSTRF and uploaded to SCHARP nightly

2. Data is normalized and combined into a specimen archive. SCHARP has an extensive internally developed system for evaluating and integrating these data into a LabKey specimen archive

3. The specimen archive is loaded into Atlas via a nightly job which is started automatically by the SCHARP-side processing pipeline

Data sources: Specimen

| Summary (Vial Count)		Acute Cohort Enrollment	Int Vist	Acute Cohort Week 1	Int Vist	Acute Cohort Week 2	Int Vist	Acute Cohort Week 3	Acute Cohort Week 4	Int Vist	Acute Cohort Week 8	Int Vist	Acute Cohort Week 12
Blood (Whole)	[unspecified]												1
	Blood (Whole)	208		186		180		14	212		6		218
	Extracted DNA												
	Extracted RNA												
	PBMC Cells, Viable	2213		1330		1354		1227	1445		1459	15	1703
	Plasma	8878		5511		5616		5742	6265		6657	34	7546
	Serum	803		618		627		621	766		723	3	736
	Whole Blood Pellet, Specify Methodology	325		118		122		144	125		164		149
Breast Milk	Breast Milk - Spun	18		22		8		4	2				
	Breast Milk - Whole	8		13		8		15	1				
	Cryopreserved primary cells from Non-Blood Spec Type, Viable	1		5		3		3					
Cervical Vaginal Lavage	[unspecified]					1							
	Cryopreserved primary cells from Non-Blood Spec Type, Viable	30	1	25	1	33		30	36	2	28		34
	Fluid Portion from a Non-Blood Specimen Type	104	6	85	12	120		109	128	5	108		123
Cervix	[unspecified]	2			2	2		4	5		4		6
	Cryopreserved primary cells from Non-Blood Spec Type, Viable	2		1		1							
	Fluid Portion from a Non-Blood Specimen Type	2		3	1	1		2	1		5		3
	Tear Flo Strips	25	1	19	2	27		22	25	1	19	1	25
Saliva	Saliva	73		63		71		70	87		57	4	65
Semen	Cryopreserved primary cells from Non-Blood Spec Type, Viable	54		80		90	2	84	85	2	92		99
	Fluid Portion from a Non-Blood Specimen Type	197		270		322	10	323	367	5	351		392
Urine	Urine	1498		1200		1240		1165	1390		253	16	262
[unspecified]	[unspecified]	2											
	Cryopreserved primary cells from Non-Blood Spec Type, Viable	1			2			3	1		2		3
	Fluid Portion from a Non-Blood Specimen Type												

Data sources: Assay

- Assay data is the most diverse in format and origin
  - >7 labs, >6 assay types
  - Many more labs deliver data via file drop (30+)
  - New data formats appear monthly
- Direct lab import
  - NAb, Luminex, GPAT data uploaded directly to lab folders
  - "Copy to study" pushes QC'd data into shared folders
- Indirect/assisted import (via LDO)
  - Data file drops (FTP, Atlas file management tools)
  - Email
  - Other custom tools external to Atlas

Data sources: Assay
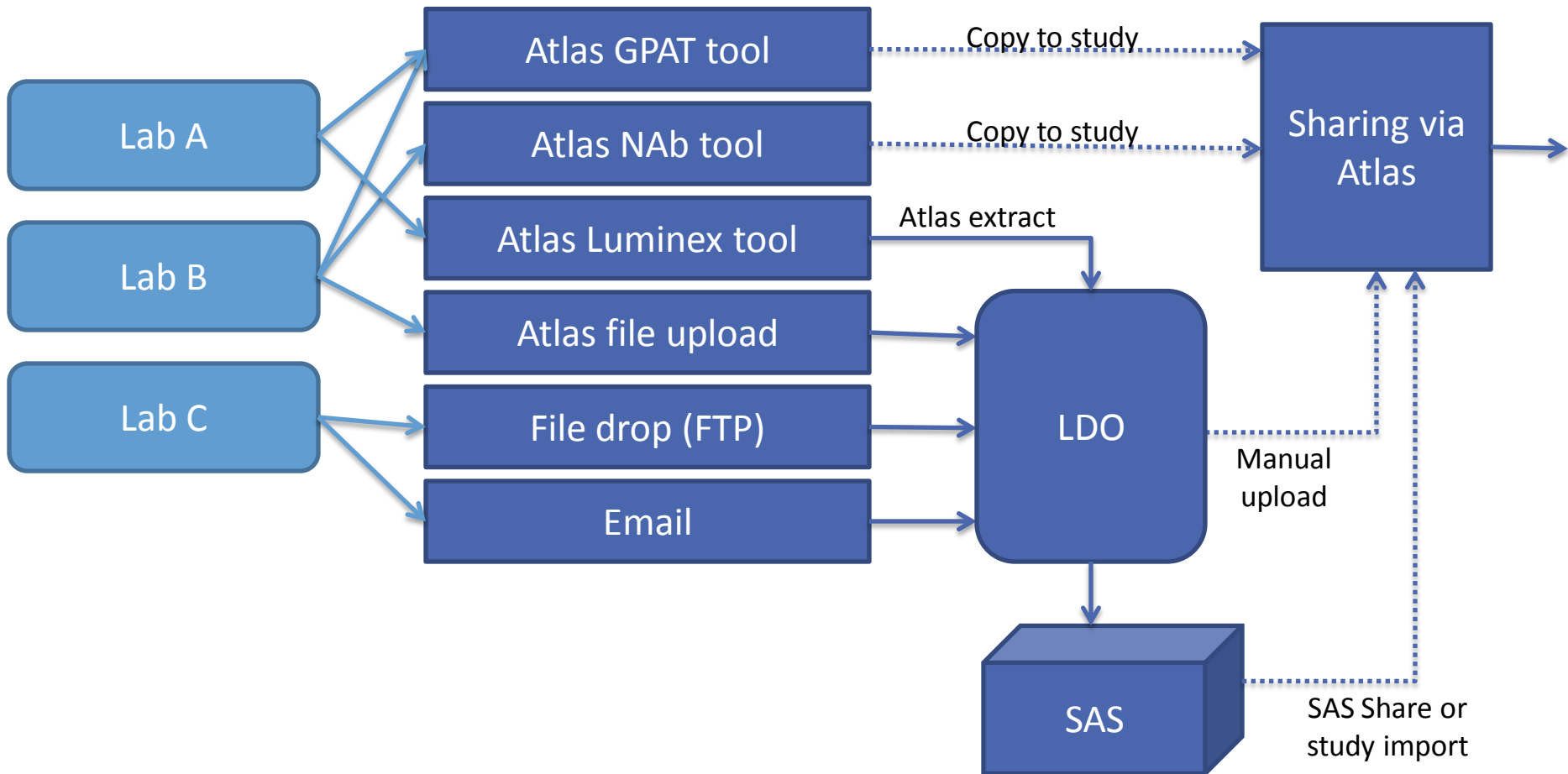
Data sources: Assay

- 3 Tiers of Assay Data
    - Standardized (and some times validated) has a well defined structure and analysis plan
    - Non-standardized usually has a good structure, but may change depending on developing analysis trends
    - R&D, highly unstable data structures with quick changes to layout and analysis

Standardized

Non-Standardized

R&D
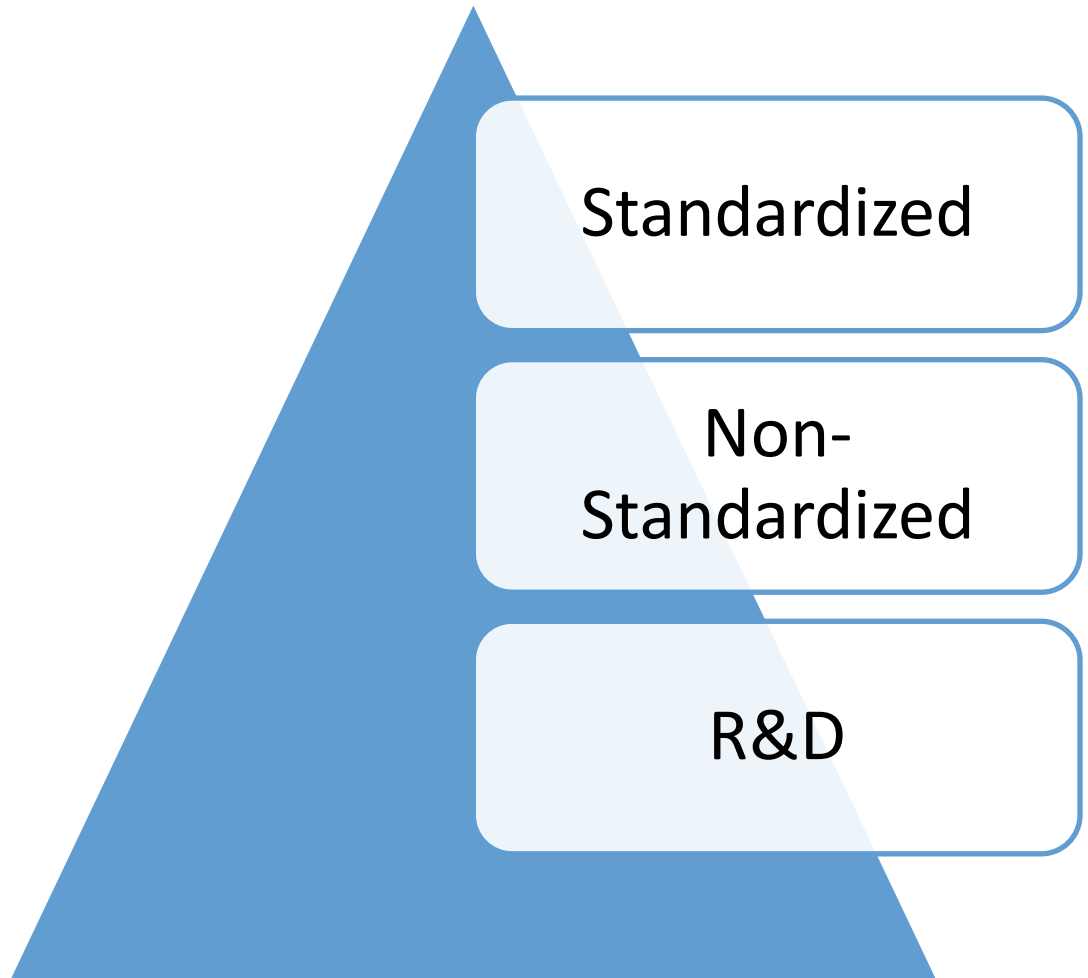
Data sources: NAb example

Run Summary: NAb1

Assay Id	NAb1	**Experiment ID**	Nab32
Created	2009-09-03 15:06:30.354	**Incubation Time**	30
Created By	eknelson	**Plate Number**	1
Virus Name	HIV-1	**File ID**	NAbresults2
Virus ID	P392	**Lock Graph Y-Axis**	false
Host Cell	T	**Curve Fit Method**	Five Parameter
Study Name	Demo	**Batch**	2009-09-03 batch 2
Experiment Performer	Elizabeth		
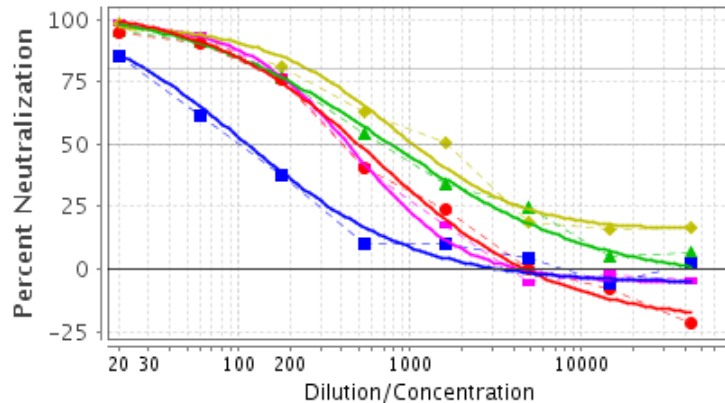
Range 54871
Virus Control 55931 ± 6%
Cell Control 1060 ± 6%

	Cutoff Dilutions			
	Curve Based		Point Based	
	50%	80%	50%	80%
526455390.2504.346	117	28	103	26
249325717.2404.493	504	134	400	132
249320619.2604.640	832	147	688	142
249328595.2604.530	979	243	1644	195
526455350.4404.456	425	157	408	149

- ■ 526455390.2504.346   ● 249325717.2404.493
- ▲ 249320619.2604.640   ◆ 249328595.2604.530
- ■ 526455350.4404.456

Specimen ID	Participant ID	Visit ID	Date	Dilution Factor	Initial Dilution	Method	Fit Error	AUC	PositiveAUC
526455390.2504.346	249318596	2804.0	2008-09-02	3.0	20.0	Dilution	4.7	0.224	0.238
249325717.2404.493	249320107	2804.0	2008-07-30	3.0	20.0	Dilution	4.1	0.372	0.405
249320619.2604.640	249320127	2804.0	2009-01-06	3.0	20.0	Dilution	3.4	0.477	0.477
249328595.2604.530	249320489	2804.0	2008-12-03	3.0	20.0	Dilution	4.5	0.547	0.547
526455350.4404.456	249320897	2804.0	2008-07-20	3.0	20.0	Dilution	2.9	0.378	0.390

Data sources: SCHARP-produced

- Atlas is also used to share data from internal systems
  - SAS datasets (via SAS Share)
  - SOPs, Protocol documents
  - Analysis results in various formats
  - Lists of antibodies, virus isolates, isotypes, etc.
  - Other postgres databases
- Permissions vary
  - Facilitating both internal and external workflows

Data sources: SCHARP-produced

- USMHRP RV144 Correlates Analysis
  - Study Management system
- Adjudication Tools
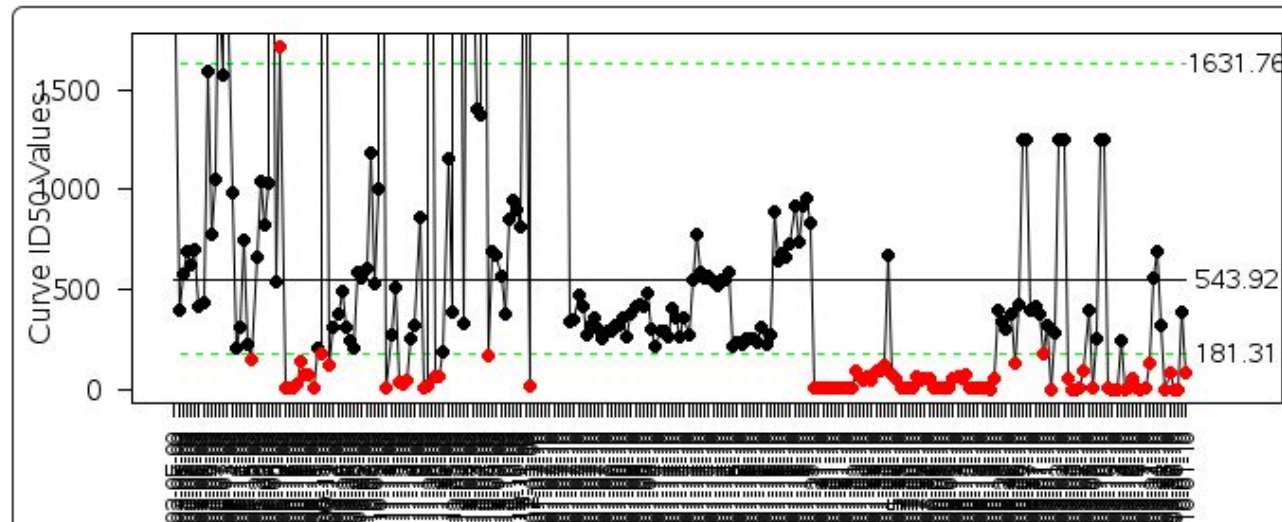- Assay Monitoring
- Specimen DB

NAb Trendline Report ▾ 🖉

Control:	HIVIG ▾	Virus:		▾
Performed by:	<choose filter> ▾	Study:	<choose filter> ▾
Start date:	📅	End date:	📅

Count	Mean	Std Dev	Acceptable Range (+/- 3x of mean)
249	543.92	697.83	181.31 - 1631.75

HIVIG, all viruses, performed by everyone, in all studies, all dates

PDF output file (click to download)

All Matching Results [Export to Excel]					
	Study	Virus Name	Experiment Date	Performed By	Curve ID50
[run details]	Haynes_McCormik_Zambia	SVA-MLV	2009-02-05	Forst, Kathleen	2500.00
[run details]	Haynes_McCormik_Zambia	6535.3	2009-02-05	Forst, Kathleen	393.77
[run details]	Haynes_McCormik_Zambia	QH0692.42	2009-02-05	Forst, Kathleen	572.28
[run details]	Haynes_McCormik_Zambia	SC422661.8	2009-02-05	Forst, Kathleen	686.83

Most Recent Pipeline Job Status

↑ ↓ ✕

[show recent pipeline step log]

Job Initiated At: 2012/09/16 19:00:02

19:00	19:50	19:54	20:50	21:14	22:40	22:40	22:56	22:59
Parse LDMS-CHAVI ✓	Parse Labware-CHAVI ✓	Parse LDMS-HVTN ✓	Parse LDMS-MTN ✓	Parse LDMS-HPTN ✓	Import Manifest Specimens ✓	Post Import Processing ✓	Export Data ✓	Trigger Atlas Import ✓
50m 29s	3m 55s	56m 21s	23m 5s	1h 25m 59s	0s	15m 56s	3s	2s
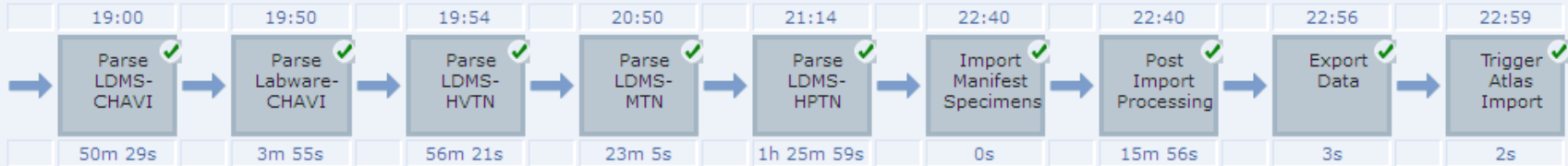
Alerts

❗ **Parse LDMS-CHAVI:** 31 records were excluded from import.

❗ **Parse Labware-CHAVI:** 1260 records were excluded from import.

❗ **Parse LDMS-MTN:** 1785 records were excluded from import.

❗ **Parse LDMS-HPTN:** 1661 records were excluded from import.

❗ **Import Manifest Specimens:** 24 records were excluded from import.

❗ **Post Import Processing:** The number of vials imported for lab 121 is significantly larger than usual.

❗ **Export Data:** The number of vials exported for lab 121 is significantly larger than usual.

		LDMS-CHAVI	LDMS-HVTN	LDMS-MTN	LDMS-HPTN	Labware-CHAVI
Job Initiated At	2012/09/16 19:00:02	**File Size (kB)**	336,966	337,859	152,271	556,937	11,384
Job Completed At	2012/09/16 22:59:39	**Vials Imported**	577,452	963,513	399,391	1,338,886	84,428
SCHARP Pipeline Time	3h 59m 37s	**Vials Exported**	575,705	963,513	399,391	1,338,886	84,369
		File Date/Time	2012/09/16 07:01:42	2012/09/16 07:16:16	2012/09/16 07:01:25	2012/09/16 07:16:20	2011/07/26 07:50:03
		Upload Time	50m 30s	56m 22s	23m 5s	1h 25m 59s	3m 56s

Agenda

- Background
- Atlas data sources
- **Atlas usage and value**
- Challenges
- Lessons and recommendations

Value to external collaborators

- Funders
  - Real-time reports available to DAIDS (e.g., HVTN IQC/EQC)
- Public
  - Data from ~70 CAVD studies available for public access
  - Free download of documents for 16 HPTN and 15 MTN protocols

Value to external collaborators

- Network core
  - Transparency into data management and analysis
- Labs
  - Access to CRF (clinical) data
  - Rapid access to integrated reports (e.g., Borrow queries)
  - Operational efficiencies when Atlas is workflow-integrated (e.g., NAb)
  - Example: cross-comparison of CRF and specimen data allowed automatic identification of mislabeled CHAVI vials

Atlas value to SCHARP staff

- Primary Investigators
  - Atlas is a selling point in grants/proposals
- SCHARP Operations
  - Secured delivery of reports (DSMB)
  - structured data delivery
  - Embedded quality control during data upload
    - Ptid, visit, specimen checks
- Statisticians
  - Flexible access (R, SAS, Excel) to data

Agenda

- Background
- Atlas data sources
- Atlas usage and value
- Challenges
- Lessons and recommendations

Challenges: people, politics and culture

- New drive towards collaborative research
  - HIV field historically competitive
  - Concerns that data will be misinterpreted
- Those paying the costs may not see the benefits
  - Technicians and project managers work to import data
  - PIs and statisticians benefit from growing data asset
- Change is difficult
  - Ownership of current methodologies hinders progress
  - Resistance to change is independent of need for change

Challenges: technological

- Flexibility is a double-edged sword
  - Rapid tool development is possible and desirable
  - Long-term support of vast tool library is very expensive
- In-house expertise in the LabKey platform
  - Difficult to hire staff with LabKey experience
- Hard to gauge technical path forward given the variety of available resources

Agenda

- Background
- Atlas data sources
- Atlas usage and value
- Challenges
- **Lessons and recommendations**

Lessons learned: platform development

- Interaction with LabKey
  - early on we were too hands off: features can miss the mark
  - Over correction to intense oversight: expensive, slow iteration
  - Work to achieve balance in team based approach
- Don't underestimate support costs
  - Allowed for organic growth of the system with downstream effects
  - "Quick" or "small" tools can be expensive
  - Documentation, standardization, planning are needed
  - Acceptance/regression testing
- Invest in self-empowerment
  - Developer tools/APIs and administrative tools have paid for themselves many times over

Lessons learned: platform adoption

- Lab buy-in requires covering their full workflow
  - Show immediate value after upload
- Adoption takes time
  - Much faster if system is integral to users' daily work

Any questions?

Sarah Ramsay
sramsay@scharp.org
(206) 667-6992