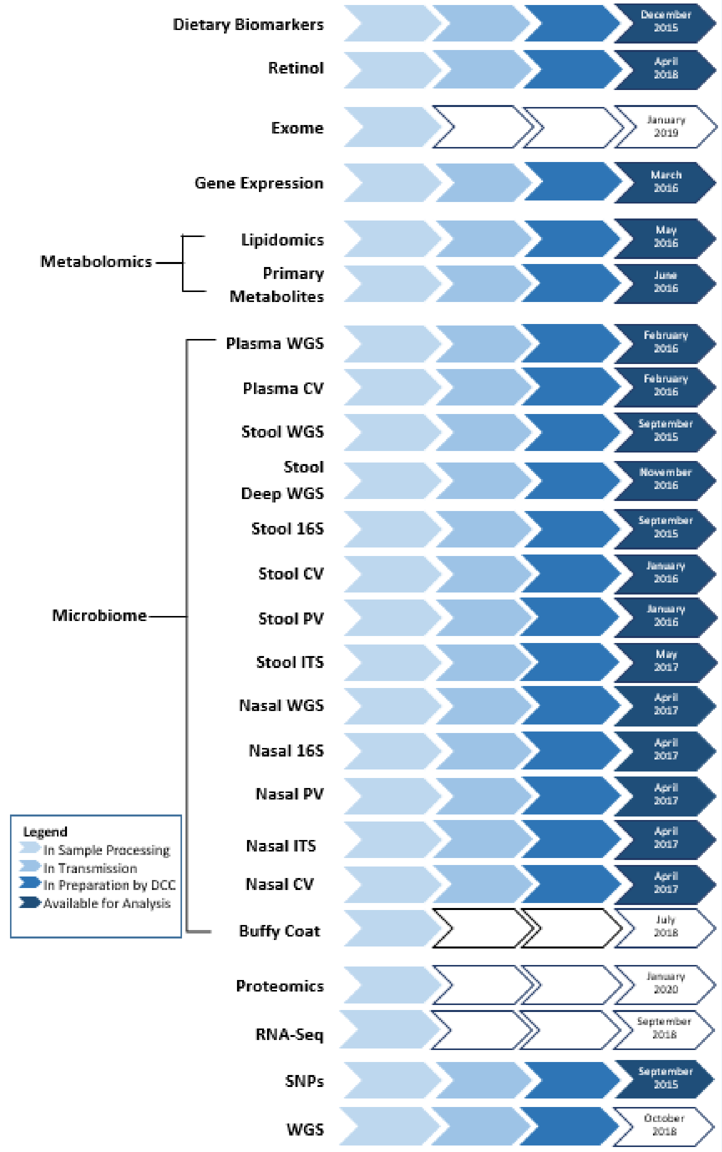
TEDDY

HII Technical Infrastructure

The Environmental Determinants of Diabetes in the Young

Data Management



	Dietary Biomarkers	December 2015
	Retinol	April 2018
	Exome	January 2019
	Gene Expression	March 2016
Metabolomics	Lipidomics	May 2016
	Primary Metabolites	June 2016
Microbiome	Plasma WGS	February 2016
	Plasma CV	February 2016
	Stool WGS	September 2015
	Stool Deep WGS	November 2016
	Stool 16S	September 2015
	Stool CV	January 2016
	Stool PV	January 2016
	Stool ITS	May 2017
	Nasal WGS	April 2017
	Nasal 16S	April 2017
	Nasal PV	April 2017
	Nasal ITS	April 2017
	Nasal CV	April 2017
	Buffy Coat	July 2018
	Proteomics	January 2020
	RNA-Seq	September 2018
	SNPs	September 2015
	WGS	October 2018

Legend
- In Sample Processing
- In Transmission
- In Preparation by DCC
- Available for Analysis

- TEDDY 'omics data
  - Quantity of data (~550 TB)
  - Diversity of data sources (9 labs, 28 analytes)
  - Number of analytical partners (9 EAP groups, 47 HPC users, 76 data sharing platform users)
  - Number of data releases (>65 releases)

TEDDY
The Environmental Determinants
of Diabetes in the Young

Data Management

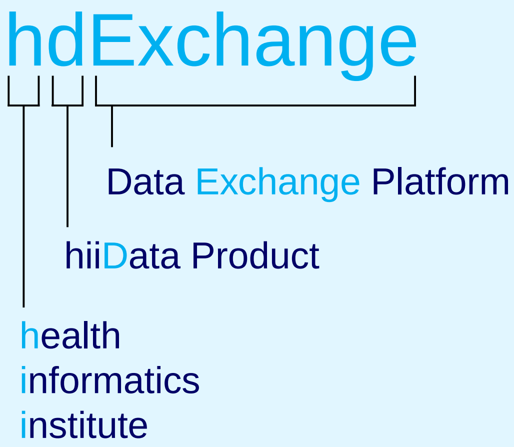
- Total raw data storage as of APR 2018: 2.0 PB
- Total expected Case-Control data: ~550 TB
    - Dietary Biomarkers - 4.1 MB
    - Exome – 100 GB
    - Gene Expression - 12 to 14 TB
    - Metabolomics - 16 to 24 TB
    - Microbiome & Metagenomics – 86 TB
    - Proteomics – 2 to 3 TB
    - SNPs – 60 GB
    - RNA Sequencing – 150 TB
    - Whole genome sequencing – 250 TB

TEDDY

The Environmental Determinants
of Diabetes in the Young

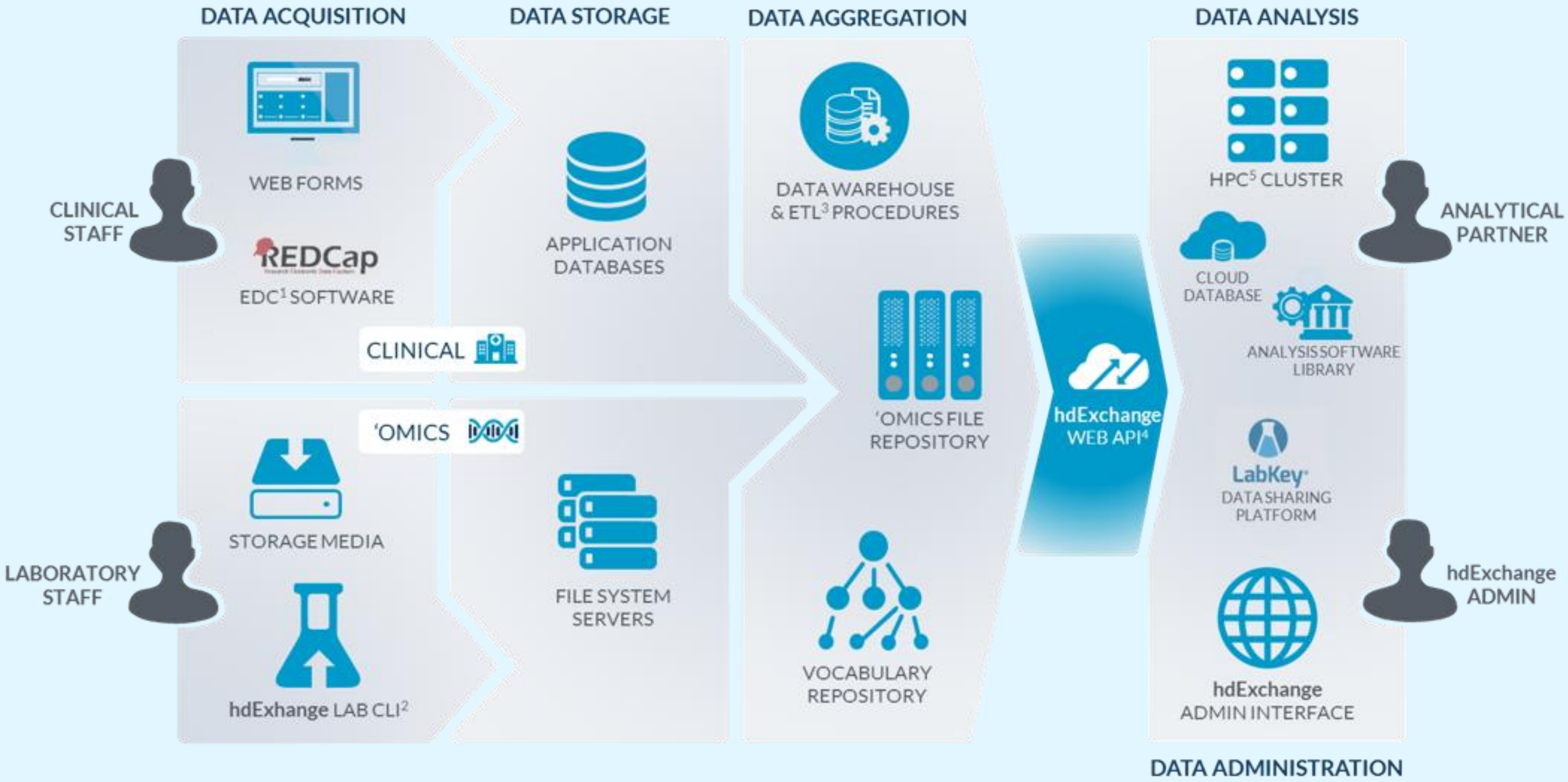
Technical Infrastructure

- Objective:
  - Comprehensively store, manage, and share HII Big Data assets in support of ‘omics analysis
  - Allow analytical partners to bring their analyses to the data
- Components:
  - Data Infrastructure
    - Clinical Data Warehouse
    - Big Data Repository
    - Controlled Vocabulary Repository
    - Laboratory Transfer CLI
    - Data Exchange API
  - Analytical Infrastructure
    - High Performance Computing (HPC) Cluster
    - Analysis Software Library



TEDDY
The Environmental Determinants
of Diabetes in the Young

Technical Infrastructure



[1]**EDC** – Electronic Data Capture | [2]**CLI** – Command Line Interface | [3]**ETL** – Extraction, Transformation, & Loading | [4]**API** – Application Programming Interface | [5]**HPC** – High Performance Computing

TEDDY
The Environmental Determinants
of Diabetes in the Young

Analytical Infrastructure: HPC Cluster

- Hardware
  - The HPC platform consists of two clusters:
  - HII (hii): 90+ nodes with ~ 1600 Cores / 8 TB Memory
  - RC (circe): 400+ nodes with ~ 5000 Cores / 12 TB Memory
- Nodes in the clusters are upgraded and expanded on a continual basis. Specs of latest 40 nodes provisioned:
  - Processor: 20-core E5-2650 v3 @ 2.30GHz (Haswell Microarchitecture)
  - Memory: 128 GB @ 2133 Mhz
  - MPI/Storage Interconnect: QDR Infiniband @ 32 Gb/s
- All nodes have access to the following storage arrays:
  - 1.7 PB DDN GPFS (Home Directories, Group Shares, and Scratch)
  - 300 TB Oracle ZFS (Genetic Sequence Files and Curated Results)

https://usf-hii.github.io/pages/hii-hpc.html

Data Infrastructure: hdExchange

- hdExchange API
  - Primary mechanism for programmatically accessing TEDDY clinical and ‘omics data from HPC environment
  - Hides the complexities of backend data management providing single point of contact for straightforward access to data assets
  - https://exchange.hiidata.org/documentation.htm

TEDDY
The Environmental Determinants of Diabetes in the Young

Data Infrastructure: hdExchange

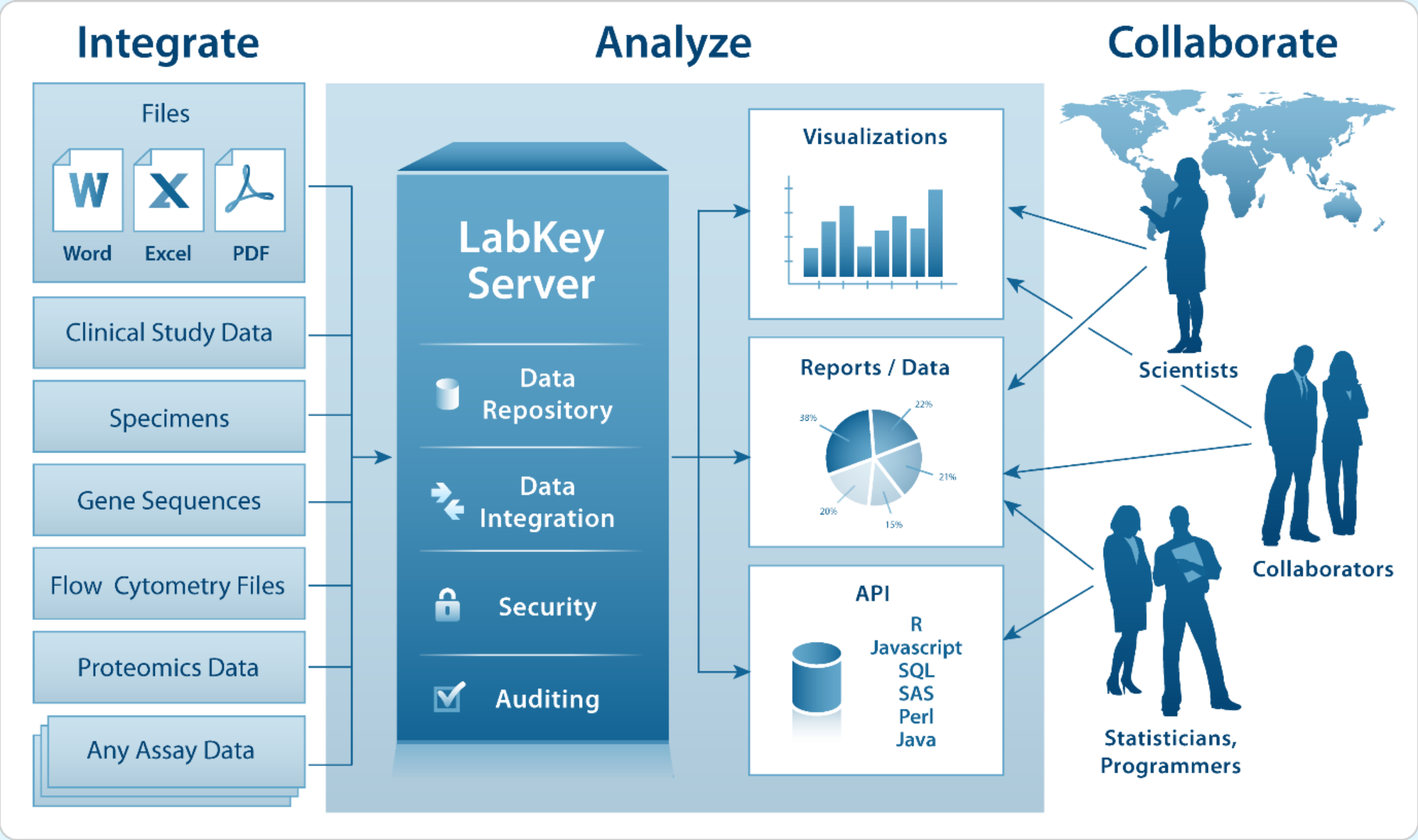
- Specifications
  - RestFUL Web API
    - W3C Standards for Rest Architecture
  - Token-based API Authentication
  - Synchronous Delivery of Tabular Data
    - Clinical Metadata & Data Dictionary
  - Asynchronous Processing of Data File Requests
    - Background Process and Message Queue for Scalability and Big Data

TEDDY
The Environmental Determinants
of Diabetes in the Young

Data Infrastructure: hdExplore

- Data Sharing Platform
  - Web application consisting of an interactive user interface for accessing TEDDY clinical metadata and associated documentation along with a suite of data manipulation and visualization tools
  - The platform is accessible only internally to authorized investigators.

Data Infrastructure: hdExplore



Integrate

Files

Word

Excel

PDF

Clinical Study Data

Specimens

Gene Sequences

Flow Cytometry Files

Proteomics Data

Any Assay Data

Analyze

LabKey Server

Data Repository

Data Integration

Security

Auditing

Visualizations

Reports / Data

API

R
Javascript
SQL
SAS
Perl
Java

Collaborate

Scientists

Collaborators

Statisticians, Programmers

TEDDY

The Environmental Determinants
of Diabetes in the Young

Funded by:

- National Institute of Diabetes and Digestive and Kidney Diseases (NIDDK)
- National Institute of Allergy and Infectious Diseases (NIAID)
- National Institute of Child Health and Human Development (NICHD)
- National Institute of Environmental Health Sciences (NIEHS)
- Juvenile Diabetes Research Foundation (JDRF)
- Centers for Disease Control and Prevention (CDC)

- Supported in part by the NIH/NCATS Clinical and Translational Science Awards to the University of Florida and the University of Colorado