AUTOMATING HIV/AIDS GRANT ENCODING

Office of AIDS Research

Robert Cregg, Lead Data Scientist
Danny Murphy, Public Health Analyst

CIVIC DIGITAL FELLOWSHIP

NIH National Institutes of Health
Office of AIDS Research

WILLIAM HUANG
UCLA
Electrical Engineering

YUYANG ZHONG
UC Berkeley
Psychology &
Data Science

OBJECTIVE

Automate and cross-verify the manual encoding of SIC codes, objective codes, and areas of emphasis for HIV/AIDS-related projects.

CIVIC DIGITAL FELLOWSHIP

PROBLEM STATEMENT

Current Manual Coding:

- Grants receiving HIV/AIDS dollars – SIC codes
- Objective codes/areas of emphasis

This project:

- Natural Language Processing (NLP) models with titles, abstracts and specific aims
- Facilitate accurate & efficient grant coding (SIC/Obj/AoE)

CIVIC DIGITAL FELLOWSHIP

PROCESS ITERATIONS

- Classification on SIC Codes
- Classification on Objective Codes
- Term Frequency Approach (TF-IDF)
- Deep Learning Implementation
- Classification on Area of Emphasis
- Unsupervised Clustering

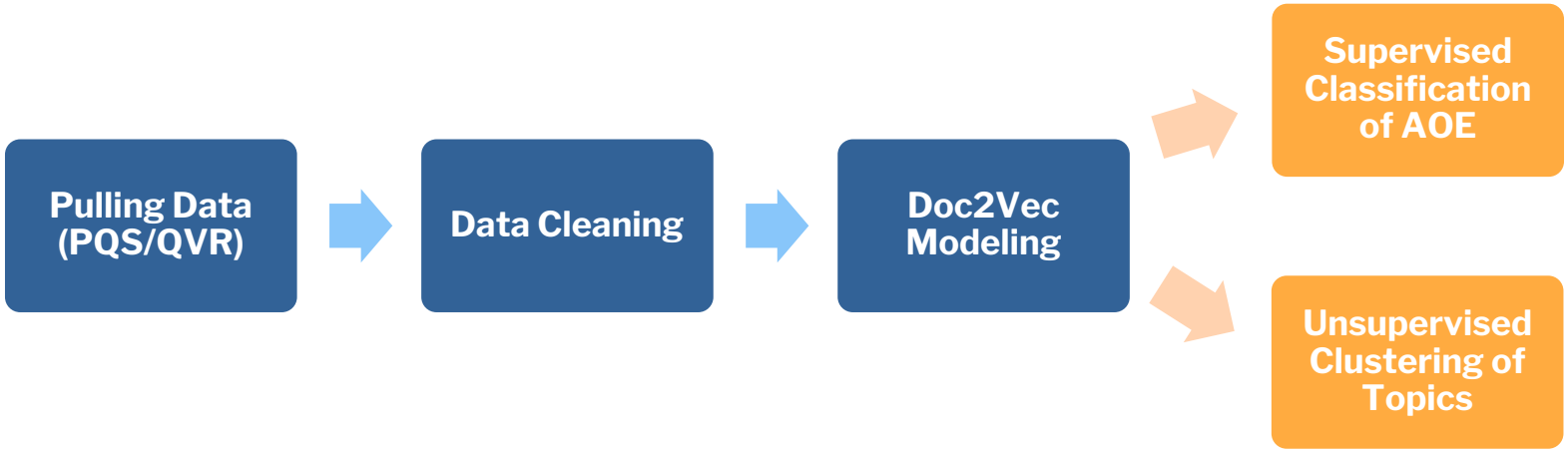
CIVIC DIGITAL FELLOWSHIP

DATASET

- **6,884 Grants**
  - FY 2011-2020
  - PQS data merged with text data from QVR
  - Focus on Areas of Emphasis (AOEs)
  - Research Project Grants (RPG)

CIVIC DIGITAL FELLOWSHIP

METHODS



CIVIC DIGITAL FELLOWSHIP

BASIC TEXT CLEANING

BEFORE CLEANING

Project summary/Abstract Accumulating evidence indicates that the incidence of HPV-associated neoplasia in HIV-positive individuals is substantially higher than in HIV-negative individuals despite effective antiretroviral therapy. These data strongly suggest that HIV may play a critical role in development of HPV-associated neoplasia of the anus, cervix and oropharyngeal cavity. However the mechanisms by which it does so are poorly understood. Our published work and preliminary data show that HIV may interact with oral and anal epithelia creating a tissue microenvironment where epithelial cells lose tight and adherence junctions.

AFTER CLEANING

accumulating evidence indicates incidence hpv-associated neoplasia hiv-positive individual substantially higher hiv-negative individual despite effective antiretroviral therapy data strongly suggest hiv may play critical role development hpv-associated neoplasia anus cervix oropharyngeal cavity however mechanism poorly understood published work preliminary data show hiv may interact oral anal epithelium creating tissue microenvironment epithelial cell lose tight adherence junction

CIVIC DIGITAL FELLOWSHIP

RCDC CONCEPT REPLACEMENT

BEFORE CLEANING

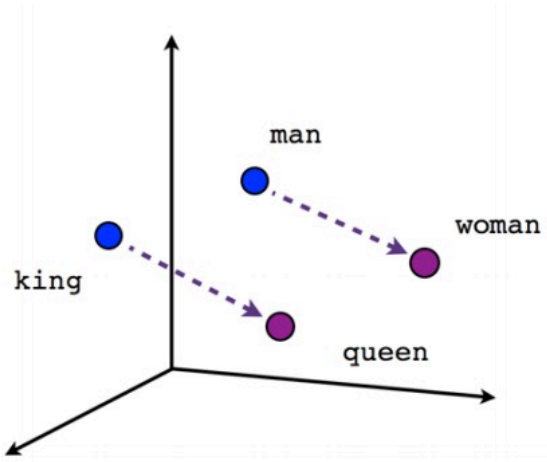
Project summary/Abstract Accumulating evidence indicates that the incidence of HPV-associated neoplasia in HIV-positive individuals is substantially higher than in HIV-negative individuals despite effective antiretroviral therapy. These data strongly suggest that HIV may play a critical role in development of HPV-associated neoplasia of the anus, cervix and oropharyngeal cavity. However the mechanisms by which it does so are poorly understood. Our published work and preliminary data show that HIV may interact with oral and anal epithelia creating a tissue microenvironment where epithelial cells lose tight and adherence junctions.

AFTER CLEANING

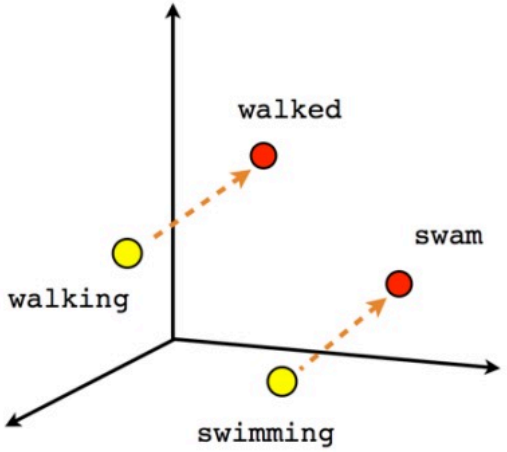
accumulating evidence concept_incidence concept_human_papillomavirus-associated concept_neoplasms concept_hiv-positive individual concept_hiv-negative individual effective concept_antiretroviral_therapy concept_data concept_hiv concept_play critical concept_role concept_development concept_human_papillomavirus-associated concept_neoplasms concept_anus concept_cervix_uteri concept_oropharyngeal cavity mechanism understood published concept_work preliminary concept_data concept_hiv interact concept_oral concept_anus epithelium creating tissue microenvironment concept_epithelial_cells lose tight concept_adherence junction

CIVIC DIGITAL FELLOWSHIP

WORD2VEC OVERVIEW



Male-Female

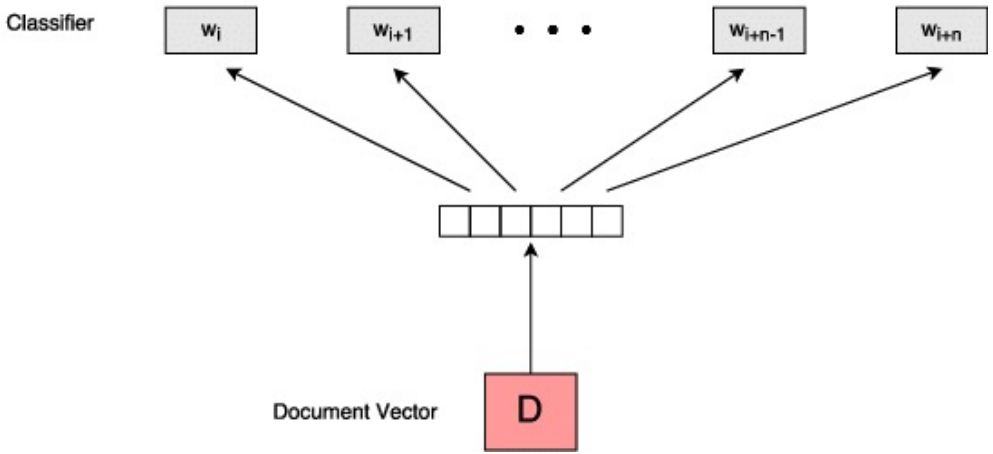


Verb tense

CIVIC DIGITAL FELLOWSHIP

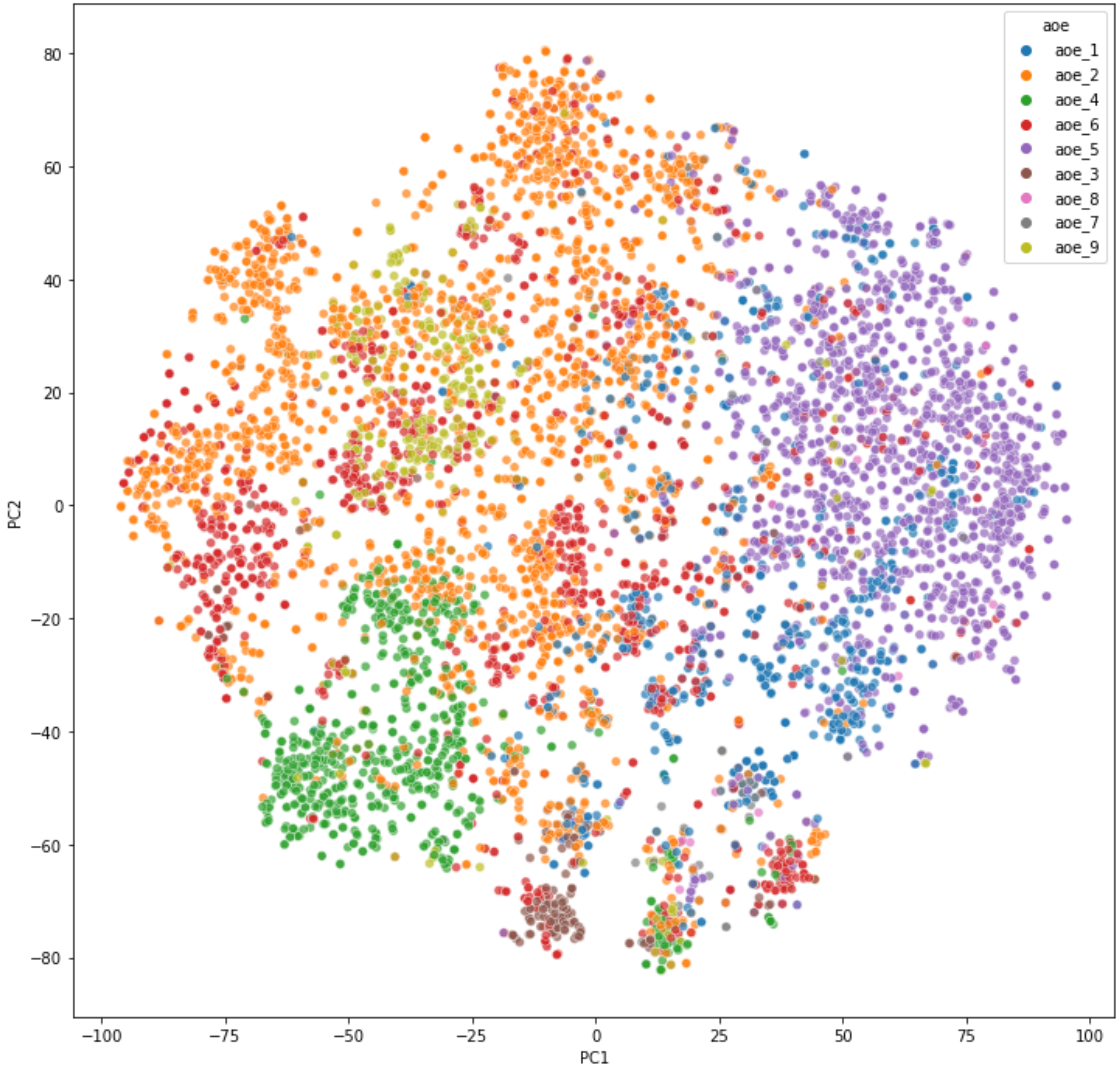
DOC2VEC OVERVIEW

- Distributed Bag of Words (DBOW)
  - Use the center word to predict words around it -> context-based



CIVIC DIGITAL FELLOWSHIP

Visualizing Doc2Vec via t-SNE



CIVIC DIGITAL FELLOWSHIP

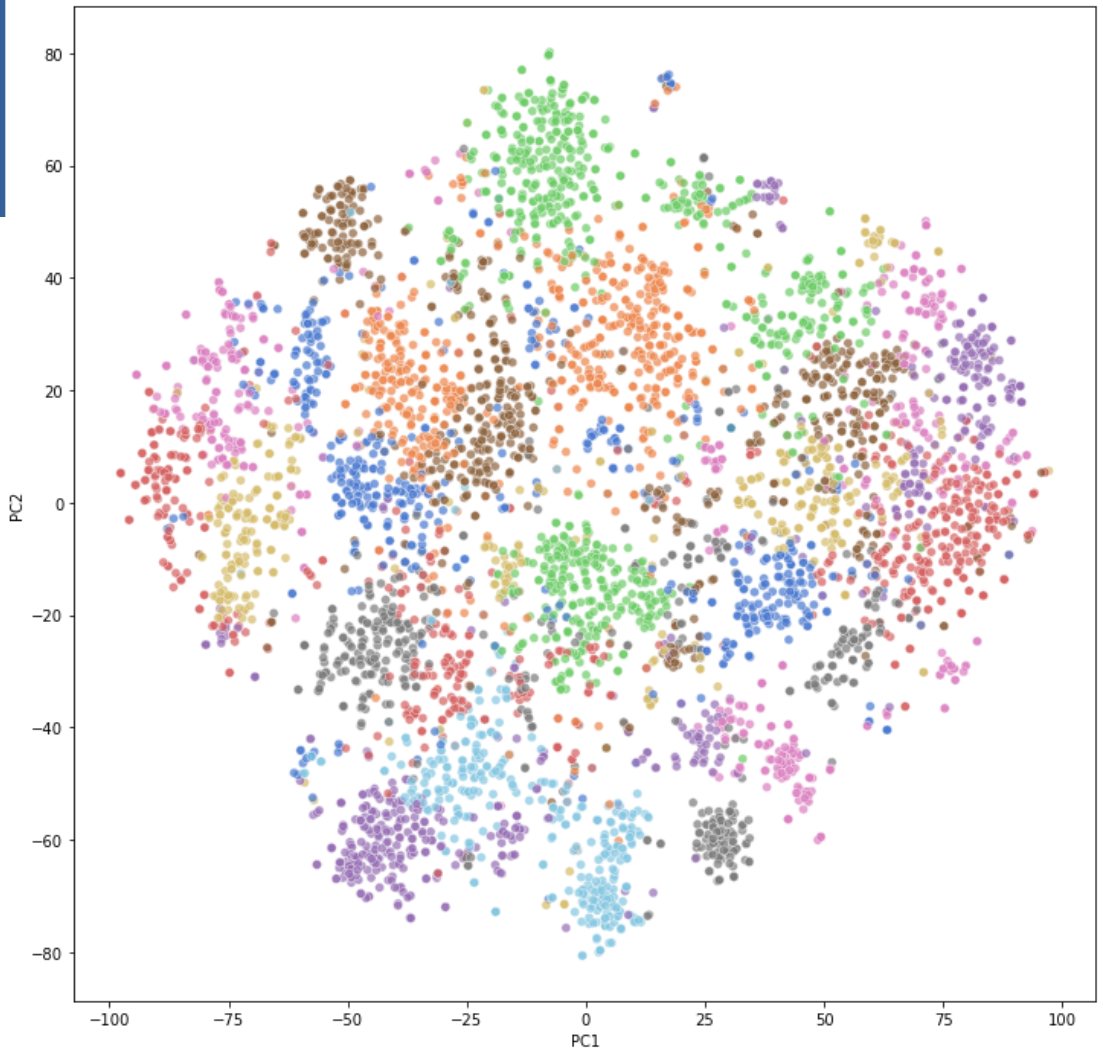
SUPERVISED CLASSIFICATION

Model	Accuracy
Logistic Regression	72.33%
Neural Network	74.22%
Random Forest	70.15%
Gradient Boosting	57.66%
Voting Classifier (Ensemble)	75.74%

- Partitioned 80% Training & 20% Testing
- The max allocated AOE is used for prediction
    - Check for multiple AOE situation: If predicted AOE is one of the lesser allocated AOEs
- Final Test Accuracy after Correction: **79.59%**

CIVIC DIGITAL FELLOWSHIP

Unsupervised Clustering By KMeans



CIVIC DIGITAL FELLOWSHIP

UNSUPERVISED CLUSTERING CONT.

Select Cluster Topics
Social Stigma, AIDS Prevention, Young Men who have Sex with Men, Counseling, Black Men who have Sex with Men
CD4, Effector, Control, Immunity, GC, CD8B1 Gene, Signal Transduction, B Lymphocyte
Alcohol or other Drugs Use, Imprisonment, Risk Behaviors, User, Services, offender, Community, Injecting Drug User, Drug

- Computationally generate topic labels for different KMeans clusters
    - Select high frequency words that do not appear frequently in other clusters
- Intended to be human readable and descriptive

CIVIC DIGITAL FELLOWSHIP

TAKEAWAY & NEXT STEPS

- **Automated Grant Categorization**
    - Automated method to help in the assignment of evaluators to grant proposals
- **Cross-Verification**
    - Check existing categorizations and compare them to what a computer might generate
    - Use D2V's clustering ability to find closely linked grants

CIVIC DIGITAL FELLOWSHIP

RCDC CATEGORY OVERLAP VISUALIZATION VIA TABLEAU

Office of AIDS Research

Robert Cregg, Lead Data Scientist
Danny Murphy, Public Health Analyst

CIVIC DIGITAL FELLOWSHIP

NIH National Institutes of Health
Office of AIDS Research

WILLIAM HUANG
UCLA
Electrical Engineering

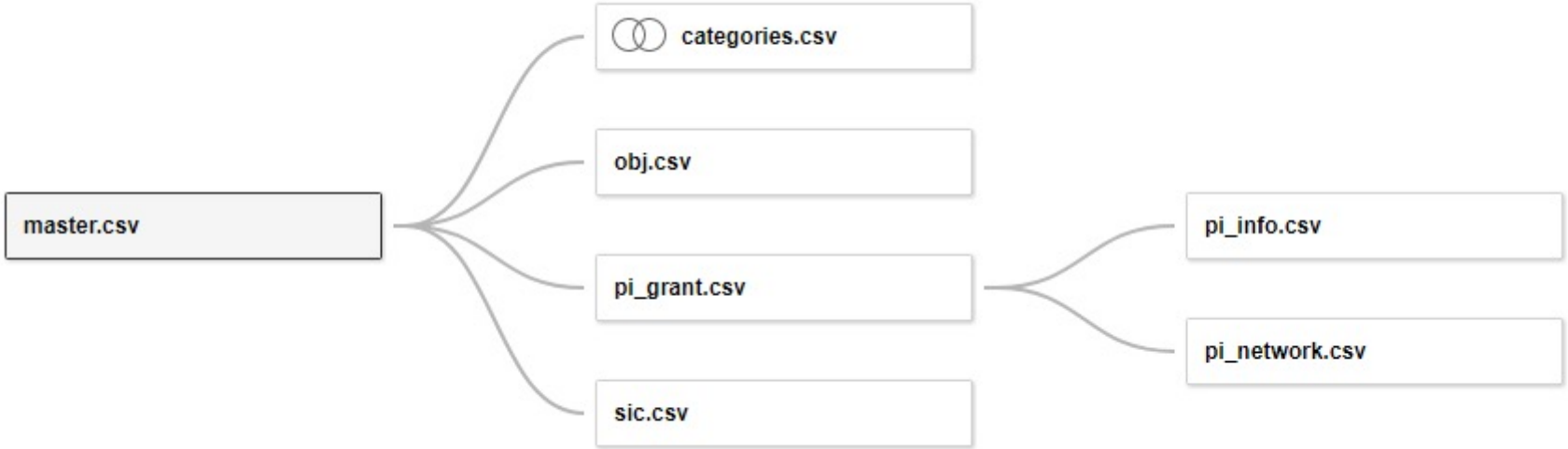
YUYANG ZHONG
UC Berkeley
Psychology &
Data Science

OBJECTIVE

Visualize funding overlap and other relationships between different major research topics related to HIV/AIDS.

CIVIC DIGITAL FELLOWSHIP

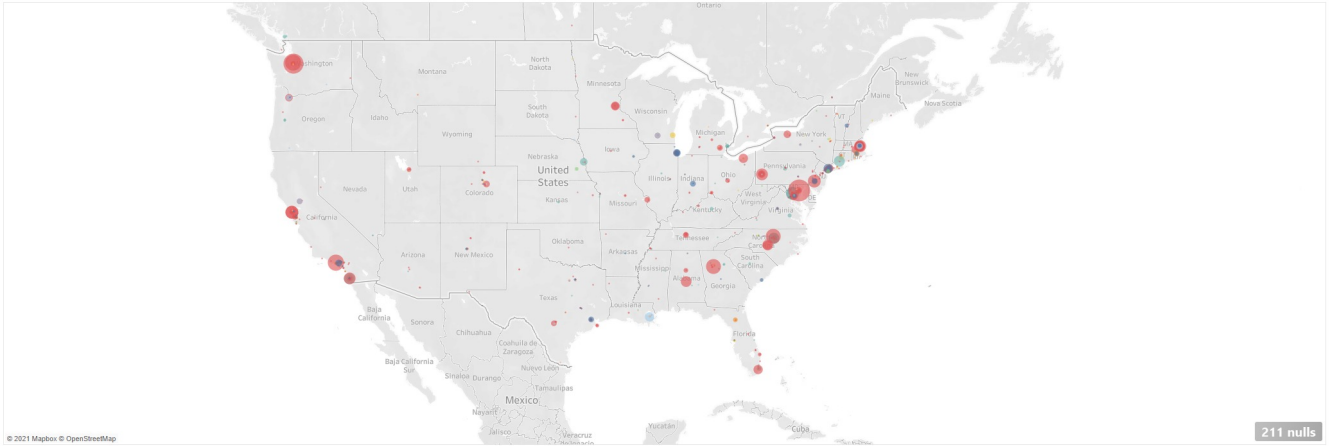
DATABASE SETUP



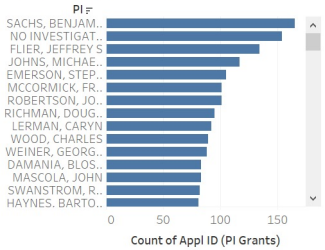
Current database organization from Tableau.

INSTITUTION VIEW

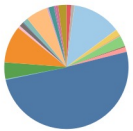
Institution Map



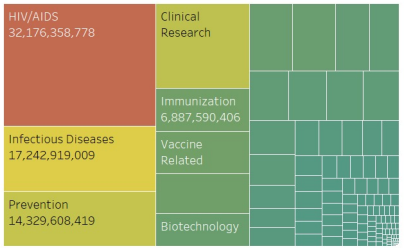
Total Grants Written



IC Funding Distribution

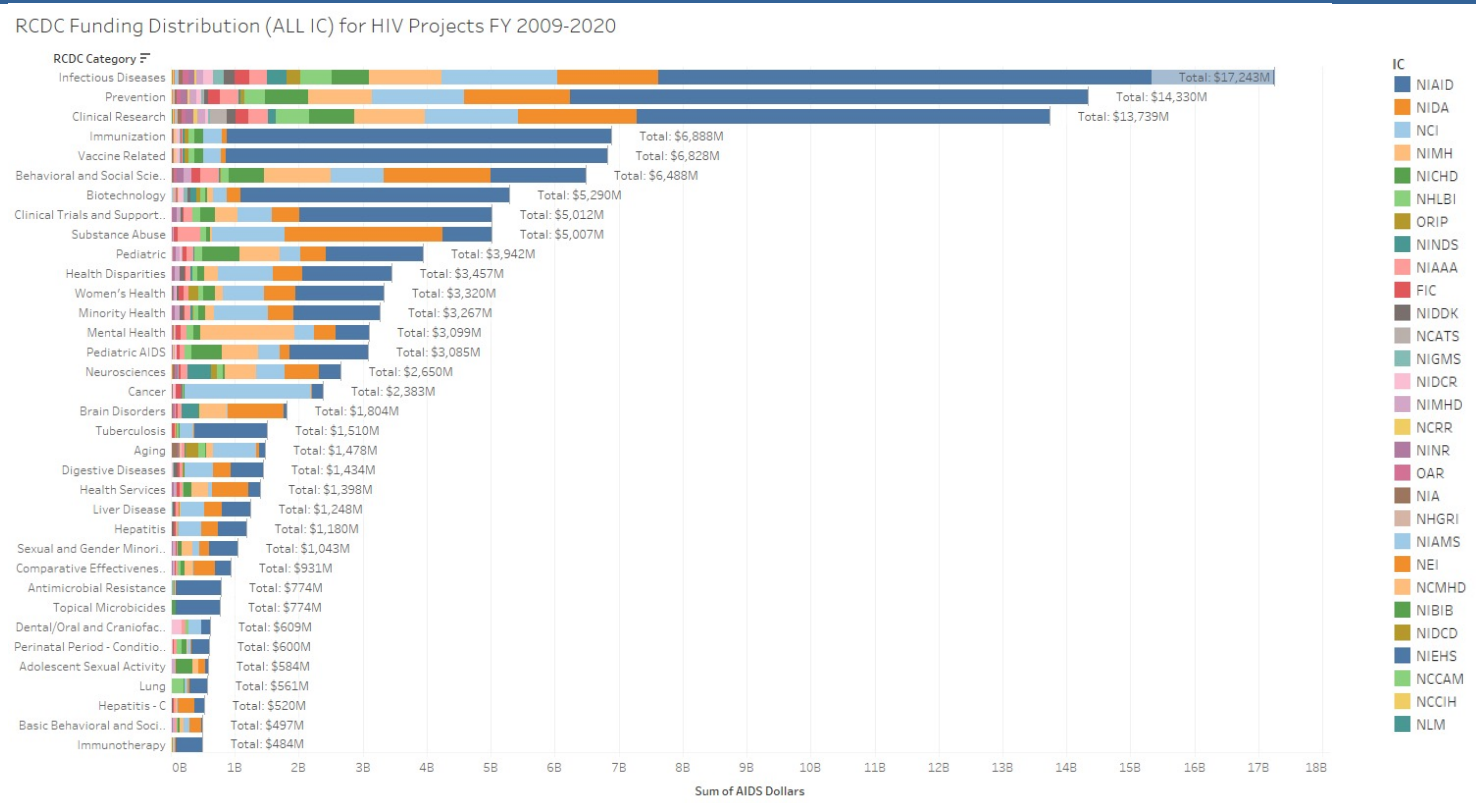


RCDC Funding Distribution (Filter by INST)



CIVIC DIGITAL FELLOWSHIP

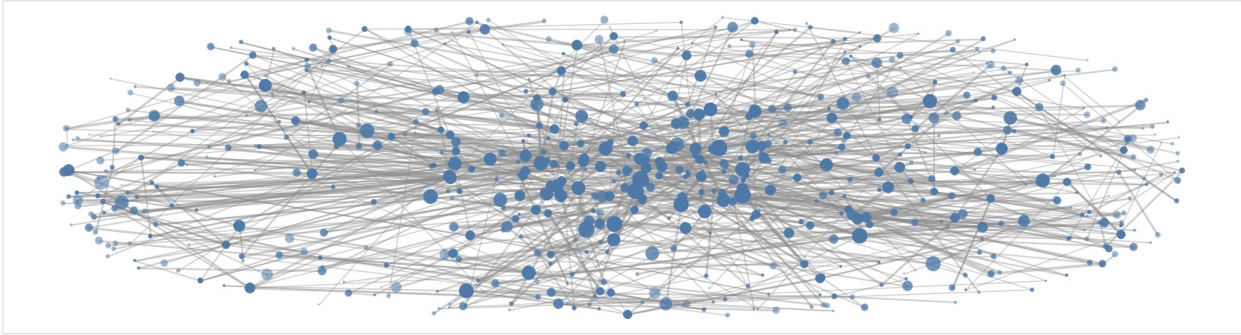
RCDC FUNDING DISTRIBUTION BY IC



CIVIC DIGITAL FELLOWSHIP

PI COLLABORATION NETWORK

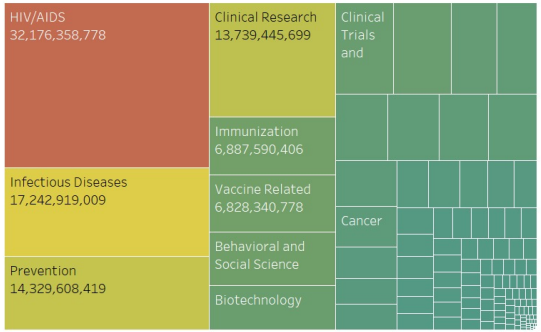
PI Network



PI Grants

Appl ID	Fiscal Y..	Project Title
74911..	2009	A MULTILEVEL HIV-PREVENTION STRATEGY FOR HIGH-RISK YOUTH
74932..	2009	DRUG ABUSE, SUSTANCE P AND HIV
74953..	2009	PLGA/ANTISENSE IL-10: GENE THERAPY FOR COCAINE ABUSERS WITH HAD
74962..	2009	ANALYSIS OF RESTRICTION FACTOR FUNCTION IN SITU
75327..	2009	IMMUNE RESPONSE TO PNEUMOCYSTIS IN SIMIAN MODEL OF AIDS
75334..	2009	MONOCLONAL ANTIBODIES FOR THE STUDY OF P. CARINII
75343..	2009	HIV/STD RISK REDUCTION FOR AFRICAN AMERICAN COUPLES
75343..	2009	ROLE OF ESTROGEN IN CERVICAL CANCER
75347..	2009	INVESTIGATION OF APOBEC3G/VIF INTERACTIONS BY SAXS AND X-RAY CRYSTAL
75348..	2009	MOLECULAR MECHANISMS OF CELLULAR GENE EXPRESSION REPROGRAMMING
75350..	2009	ADHERENCE TO HIV CARE: ADVANCING THE SCIENCE
75351..	2009	STUDIES OF NATURAL SIV INFECTION OF SOOTY MANGABEYS
75352..	2009	HIV-1 INTEGRASE STRUCTURE AND FUNCTION AS A THERAPEUTIC TARGET
75352..	2009	STRUCTURAL STUDIES OF HIV-1 INTEGRATION
75372..	2009	STRUCTURE AND FUNCTION OF THE HEPATITIS C VIRUS GENOME
75383..	2009	NATURAL HISTORY OF HIV INFECTION IN INJECTION DRUG USERS
75399..	2009	INTERFACE OF INNATE AND ADAPTIVE IMMUNITY IN HIV-1 INFECTION
75399..	2009	INTERVENTION FOR HIV+ ADULTS WITH CHILDHOOD SEXUAL ABUSE
75399..	2009	CONDITIONAL CASH TRANSFERS TO PREVENT SEXUALLY TRANSMITTED INFECTI
75404..	2009	KILLER CELLS & VIRAL LOAD IN VERTICAL HIV INFECTION
75409..	2009	MUCOSAL REOVIRUS-ADENOVIRUS VACCINES AGAINST HIV-1

RCDC Funding Distribution (Filter by PI)



CIVIC DIGITAL FELLOWSHIP

TAKEAWAY & NEXT STEPS

- **Easily Extendable**
    - Standardized database setup allows future data scientists to build more workbooks and visualizations
- **Accessible**
    - Leverage Tableau's different features to develop easily readable visualizations that can be shown to internal OAR members, Congress, or the general public.

CIVIC DIGITAL FELLOWSHIP