OPEN SOURCE
LEADERSHIP SUMMIT

Introduction to the Community Data License Agreement

Nick Acosta, IBM

THE LINUX FOUNDATION

Introduction

Agenda

Why Open Source Data

Data sharing uncovers five ne

https://www.nih.gov/news-events/news-releases/data-sharing-uncovers-five-new-risk-genes-alzheimers-disease

NIH National Institutes of Health
Turning Discovery Into Health

Search NIH

NIH Employee Intranet | Staff Directory | En Español

Health Information | Grants & Funding | News & Events | Research & Training | Institutes at NIH | About NIH

Home » News & Events » News Releases

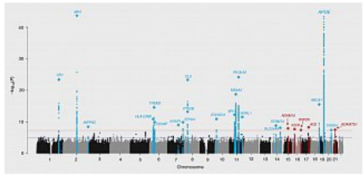
NEWS RELEASES

Thursday, February 28, 2019

Data sharing uncovers five new risk genes for Alzheimer's disease

NIH-funded project includes largest sample to date for Alzheimer's gene association.

Analysis of genetic data from more than 94,000 individuals has revealed five new risk genes for Alzheimer's disease, and confirmed 20 known others. An international team of researchers also reports for the first time that mutations in genes specific to tau, a hallmark protein of Alzheimer's disease, may play an earlier role in the development of the disease than originally thought. These new findings support developing evidence that groups of genes associated with specific biological processes, such as cell trafficking, lipid transport, inflammation and the immune response, are "genetic hubs" that are an important part of the disease process. The study, which was funded in part by the National Institute on Aging (NIA) and other components of the National Institutes of Health, follows results from 2013. It will be published online February 28, 2019 in the journal *Nature Genetics*.



Newly identified (red) and known (blue) genes linked to Alzheimer's disease spike in this table plotting results from genome-wide association analysis of 94,437 individuals with late onset Alzheimer's. *Kunkle et al and Nature Genetics.*

"This continuing collaborative research into the genetic underpinnings of Alzheimer's is allowing us to dig deeper into the complexities of this devastating disease," said Richard J. Hodes, M.D., director of the NIA. "The size of this study provides additional clarity on the genes to prioritize as we continue to better understand and target ways to treat and prevent Alzheimer's."

The researchers, members of the International Genomic Alzheimer's Project (IGAP), analyzed both rare and common gene variants in 94,437 individuals with late onset Alzheimer's disease, the most common form of dementia in older adults. IGAP is made up of four consortia in the United States and Europe that have been working together since 2011 on genome-wide association studies (GWAS).

Institute/Center

National Institute on Aging (NIA)

Contact

Joe Balintfy
301-496-1752

Connect with Us

Subscribe to news releases

RSS Feed

The National Institute on Aging Genetics of Alzheimer's Disease Data Storage

NIAGADS DATA REQUEST DOCUMENTS

DOCUMENT NAME	UPLOAD DATE
Data Distribution Agreement	2018-10-29
NIAGADS Application Instructions	2016-09-09
NIAGADS Renewal Instructions	2016-09-09
NIAGADS Data Use Certification	2015-04-16
NIH Biosketch Sample	2015-04-16
Research Use Statement Template	2016-09-09
Supplemental Information	2016-09-09
NIA AD Genomic Data Sharing Policy	2015-03-02
Sample Language Data Return to NIAGADS	2015-03-02

Data != Code

Data

Not protected by copyright

Data != Code

Data

Not protected by copyright

Patents do not apply to data

Data != Code

Data

Not protected by copyright

Patents do not apply to data

Value in analysis

Data != Code

Data

Not protected by copyright

Patents do not apply to data

Value in analysis

Code

Protected by copyright

Standard patent process

Value is intrinsic

CDLA

CDLA

Two model agreements, introduced and sponsored by the Linux Foundation

CDLA

Two model agreements, introduced and sponsored by the Linux Foundation

October 23, 2017
Modeled after leading open source agreements
Licenses that reflect the nuances of data
Designed for use by independent data communities

CDLA

CDLA promotes free exchange of data

CDLA promotes free exchange of data

Permits data to be freely used, modified, and republished
Authorship and source attribution statements must be preserved
No use restrictions permitted
Broader license coverage than mere copyright
Explicit permissions to create separate works and analyses of licensed data

Computational Use

Both CDLA licenses provide the right to Computational Use

Computational Use

Both CDLA licenses provide the right to Computational Use

Analyze data and create analytical works based upon it
Analytical works do not need to be relicensed under the terms of the CDLA
Minimal reps and no warranties
Broad Limitation of Liability
No prohibition on commercial use of data

CDLA Versions

Versions

Sharing

Permissive

Versions

Sharing

Republished data must be licensed under CDLA Sharing

Permissive

Data may be republished under any terms not inconsistent with the License

Versions

Sharing

Republished data must be licensed under CDLA Sharing

Includes:
Data Modifications to data received
Data Additions to data received

Excludes:
The results of any analysis

Permissive

Data may be republished under any terms not inconsistent with the License

Example

Example

CDLA

Example

Example

CDLA - Sharing

CDLA - Sharing

Example

CDLA - Permissive

Example, pt. 1

Deep Learning Developers Dataset

Deep Learning Frameworks

Allow for “easy” construction of neural networks

Deep Learning Frameworks

Allow for “easy” construction of neural networks

Abstract away lowest level detail
Standardize commonly used neural network concepts, from components to models
Provide tooling for deployment on various hardware configurations, including GPUs

Deep Learning Frameworks

TensorFlow

PYTORCH

Python Notebooks

DLClassifier

localhost:8888/notebooks/DLClassifier.ipynb

jupyter **DLClassifier** Last Checkpoint: 12 minutes ago (autosaved) Logout

File Edit View Insert Cell Kernel Widgets Help Trusted Python 3

Run Code

In [1]:
```
!pip install pygithub
```

```
Requirement already satisfied: pygithub in /anaconda3/lib/python3.6/site-packages (1.39)
Requirement already satisfied: pyjwt in /anaconda3/lib/python3.6/site-packages (from pygithub) (1.6.1)
You are using pip version 18.1, however version 19.0.3 is available.
You should consider upgrading via the 'pip install --upgrade pip' command.
```

In [2]:
```
import re
import sys
import json
import base64
import github
import operator
import numpy as np
from collections import *
from github import GithubException
```

In [3]:
```
login = '***'
password = '***'

g = github.Github(login,password)
```

In [4]:
```
targetfiles = ['.ipynb']
```

In [5]:
```
def download_directory(repository, path, framework):
    global dataset
    try:
        contents = repository.get_contents(path)
        for content in contents:
            if content.type == 'dir':
                download_directory(repository, content.path, framework)
            else:
                if content.content:
                    if len(str(content.name).split(".")) == 2:
                        if any(substring == ("." + str(content.name).split(".")[1]) for substring in targetlangs):
                            try:
                                dataset.append([repository,str(base64.b64decode(content.content),'utf-8'),csp])
                            except (GithubException, IOError) as exc:
                                print('Error processing %s: %s', content.path, exc)
    except (GithubException, IOError) as exc:
        print("error in dir ")
```

In [6]:
```
def getrepos():
    repos = []
    repos.append(list(g.search_repositories('TensorFlow')))
    repos.append(list(g.search_repositories('PyTorch')))
```

Notebook Licensing

Licenses vary widely among notebooks posted to GitHub

located in a separate file or repo

referenced in the notebook

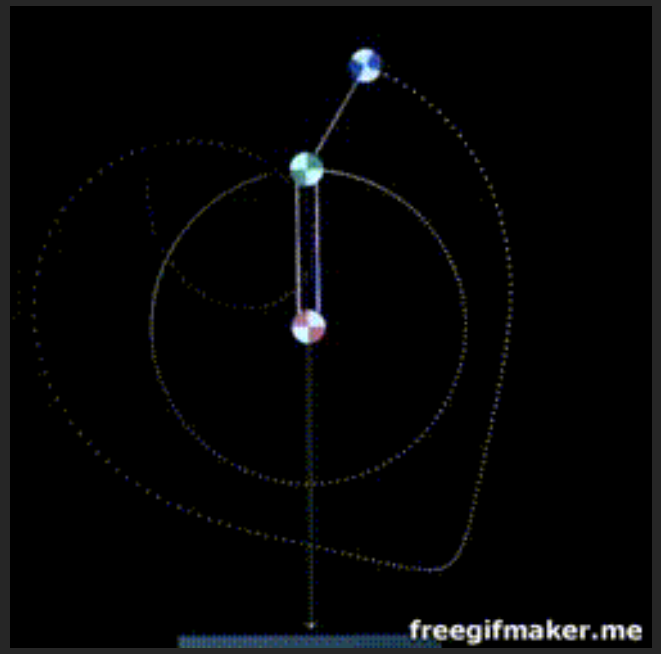
no, ambiguous or default license (Cloud Platform)

notebooks with one license referencing information with differing license

Example, pt. 2

Double Pendulum Chaotic Dataset

Double Pendulum Chaotic Dataset



Learning beyond simulated physics

NIPS 2018 Spatiotemporal Workshop

Used to showcase MI – Prometheus Framework

Promotes reproducibility in Machine Learning Research

The End

Interested in data sharing?

Thank You

Nick Acosta
Developer Advocate
—
nacosta@us.ibm.com
cdla.io


OPEN SOURCE
LEADERSHIP SUMMIT
THE LINUX FOUNDATION