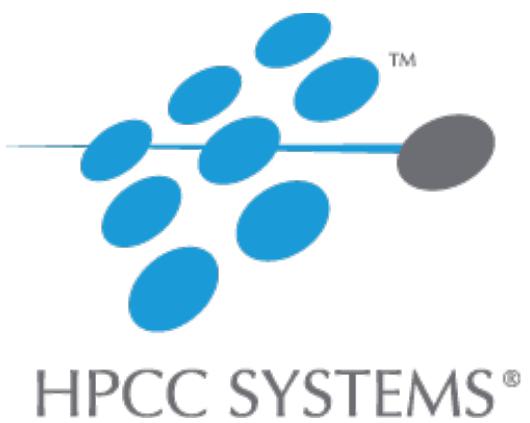




## HPCC Systems Intern Program



Lorraine Chapman  
Manager Business Analyst  
HPCC Systems Intern Program Manager  
LexisNexis Risk Solutions



# What is HPCC Systems?

- End to end big data analytics solution – but it doesn't have to be
- First created in 2000, went open source in 2011 – [hpccsystems.com](http://hpccsystems.com)
- Underpins a number of our own data driven business initiatives
- Used in a wide variety of ways – IoT, academic research, business
- Thor Cluster – Data Refinery
- ROXIE Cluster – Data Delivery
- Queries coded in ECL (Enterprise Control Language) training available
- Contribute to the platform / machine learning library /use case
- Cloud native platform - <https://wiki.hpccsystems.com/x/FYD0B>

# HPCC Systems Intern Program



- 12 weeks paid program
- High school through to PhD
- Global – Asia, US and Europe
- May through August
- Remote or LN office based (Alpharetta, GA or Boca Raton, FL)
- Proposal period deadline: 18<sup>th</sup> March 2022
- Review panel decides
- Mentoring
- Community involvement

Read my blog

[hpccsystems.com/intern](http://hpccsystems.com/intern)



Join the HPCC Systems team as an intern

The proposal application period for the 2022 HPCC Systems Intern Program is NOW OPEN

Final deadline date for proposal is Friday 18th March 2022

\*\*\*\*\*

Every year, HPCC Systems publishes a [list of projects](#) which are designed to be completed by students during our [summer intern program](#). These projects cover a wide range of areas from web interfaces, machine learning, JAVA programming, internet of things, compiler related projects and more. To apply, you need to submit a detailed project proposal showing how you plan to complete either one of our suggested [projects on our list](#) or submit a proposal outlining a project idea of your own that leverages HPCC Systems in some way.

We do offer places on our intern program in advance of the final deadline date to students who submit an excellent proposal we know we want to accept.

Students can work remotely or in one of our offices. In 2021, all students worked remotely due to the COVID-19 Pandemic. [Find out how remote working works in this blog](#) and learn about about

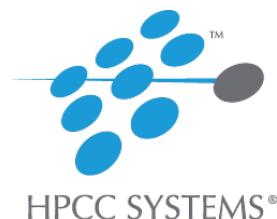
# Timeline - NCF Perspective

## 2022 internships

- 12 weeks
- Deadline - Friday 18<sup>th</sup> March
- Looks for projects now: [hpccsystems.com/ideas-list](http://hpccsystems.com/ideas-list)
- Suggest your own using HPCC Systems

## 2023 Internships as practicum

- 16 weeks
- Apply December 2022 for an early start (Jan/Feb)

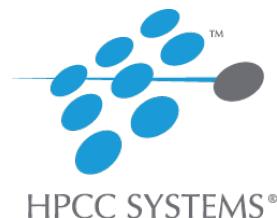


## Class of 2021 Projects - Use Cases

- Ingestion/Analysis of collegiate women's basketball GPS data using HPCC Systems and RealBI
- COVID-19 Tracker and Global Map Improvements
- Processing Robotics Data with and HPCC Systems Cluster on Kubernetes
- Improvements to the HPCC Systems Structured Query Language (HSQL)

## Class of 2021 Projects – Cloud Related

- Using Azure Spot Instances
- Ingress Configuration
- Apply Docker Image Build and Kubernetes Security Principles



# Class of 2021 Projects – Machine Learning

- Implement a PMML Processor
- Toxicity Detection Machine Learning Project
- Causality – Probabilities and Conditional Probabilities
- Causality– Independence, Conditional Independence and Directionality
- Causality – Counterfactual and Interventional Layers
- Causality 2021 Project Details  
<https://hpccsystems.com/blog/causality-2021>
- Learn more: <https://hpccsystems.com/blog/intern-intro-2021>

# Project Proposal

- Find out more:

[hpccsystems.com/student-wiki](http://hpccsystems.com/student-wiki)

- Available Projects List

[hpccsystems.com/ideas-list](http://hpccsystems.com/ideas-list)

- Highlight the main tasks
- Timeline of work for each week
- Challenges and solutions
- Liaise with the project mentor

Project Proposal	
<b>Title : Implement Latent Semantic Analysis in ECL-ML</b>	
<b>Deliverables : Will be implemented</b>	
1) FUNCTION to convert Document Corpus into term-document Matrix efficiently 2) FUNCTION to perform SVD on constructed term-document Matrix 3) FUNCTION to reduce Components of SVD by given Rank 4) Transform initial Document Term Vectors into Reduced Representation 5) Implementation of "Folding-In" method of LSA to make addition of new Documents in pre-computed LSA results efficiently. 6) Checks to determine when LSA needs to be re-performed due to repeated "Folding-In" 7) FUNCTION to compute query representation in reduced dimension 8) FUNCTION to calculate Document-Query Similarity and return best matched documents 9) Tests and Documentation	
<b>Wishlist :</b> 10) Implementation of SVD for Dense Matrix 11) Checks for Performance in both Sparse and Dense Matrix format. 12) Improving accuracy of LSA by implementing Locality Sensitive Hashing 13) Including other Information Retrieval Measures like Latent Dirichlet Model and Topic Modelling based on LSA	

Timeline :	
Design of Workflow from Data Input till result production Collection of Test Documents. Best Sources include datasets found in TREC IR competitions as well as from Wikipedia for benchmarking. Preprocessing and dataset-specific cleaning of above documents	25 <sup>th</sup> May – 5 <sup>th</sup> June
Convert text documents in RECORDs using Enumerate Function in Docs Module	6 <sup>th</sup> June – 15 <sup>th</sup> June
Improve methods for cleaning and splitting Text Documents into words. Specifically, : Include implementation for SnowBall Stemmer, which performs universally better than Porter Stemmer Include implementation for Lemmatization using WordNet	

# Mentoring and Support

- RELX Colleagues, School Teachers, Business Contacts, Professors
- Subject Matter Experts and Wellbeing
- Evaluations – Mid Term and Final
- Progress Reports - Weekly
- Daily stand-up call
- Email, Gitter, Teams/Zoom etc
- Intern Chat and Share

## Available Projects

- Additional Embedded Languages
- Additional External Datastores
  - Address Cleaner Plugin Optimizations
- Cloud specific projects
  - Develop an automated ECL Watch Test Suite
  - ECL Code Documentation Generator Improvements
  - Implement a Reverse activity
  - Implement reference dafilesrv in other languages
  - Incorporating self test code into a bundle
  - Investigate Test Frameworks and Best Practices for HPCC Systems Cloud
  - Investigate Third Party Environments Working with the HPCC Systems Cloud
  - Locking engine to replace DALI - Investigative project
- Machine Learning Algorithms on the HPCC Platform
- Marketing / Documentation Projects
- Natural Language Processing Projects
- Performance Testing - Bare Metal vs Cloud Native
- Provide test code for bundles with no self test

# Opportunities to Share Your Work

- Personal Blog Journal
- HPCC Systems Blog
- Presentations
- Poster Contest
- Community Day
- Social Media Channels
- HPCC Systems Website
- GitHub Repository



New College of Florida

Measuring the Geo-Social Distribution of Opioid Prescriptions

Nicole Navarro  
Mentor: Jo Richardson

**Project Goals**

- Broaden understanding of drug diversion behavior within organized social groups
- Rank and isolate communities with higher geo-social connectivity and high rates of opioid prescriptions who are better candidates for intervention through treatment and education programs

**High Risk Social Networks**

Choose specific social networks in order to view their prescription levels over time as well as how they are spread out geographically

HPCC > Nicole Navarro - Poster presentation entry 2018 > Nicole Navarro - HPCC Technical Poster Presentation.png

**Unique Perspective**

Instead of focusing on individual patients or prescribers we explored:

- 1) Social Networks: How are patients receiving opioid prescriptions connected to each other
- 1) Geography: How are patients receiving opioid prescriptions spread out geographically

**Approach**

- Generate Features: Create new data features to better measure diversion behavior
- Interactive Dashboards: Produce interactive visualizations and dashboards that allow for deep dives into the data

**Utilizing HPCC Systems Tools**

The HPCC Systems platform allowed for scaling the effort of computing attributes and creating visualizations using 5 separate files that ranged from 7 million records to 323 million records utilizing the following tools:

- ECL
- Data Science Portal (DSP)
- Knowledge Engineering Language (KEL)

Contact  
Nicole Navarro  
New College of Florida  
nicole.navarro@newcollege.edu

LexisNexis® RISK SOLUTIONS

HPCC SYSTEMS®

## Applying HPCC Systems TextVectors to SEC Filings

Lorraine Chapman on 07/30/2020

This blog features a student from our 2020 HPCC Systems intern program. Matthias Murray, joined us from New College of Florida, to work on a machine learning related project which will be a significant and popular contribution to our machine learning library. Read on to find out about Matthias's achievements as he introduces his project and demonstrates its potential.

# Personal Development Experience

- Office/teamworking experience
- Project management
- Communication skills
- Research and development real world experience
- Bleeding edge development
- Confidence in decision making
- Use of developer tools and processes
- Contribute to an active open-source community

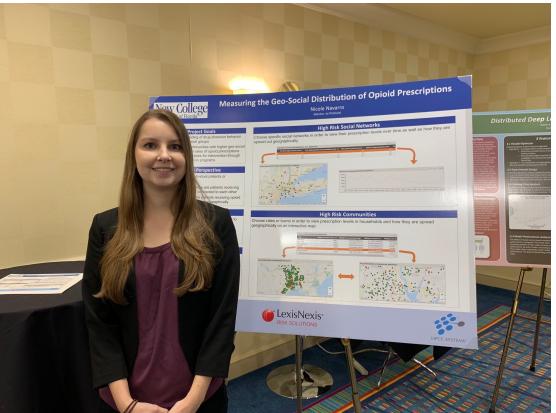
# Be Part of the Team and Make Your Mark



Nicola Navarro

2<sup>nd</sup> Place Poster Winner 2018

Measuring the geo-social distribution of opioid prescriptions



Matthias Murray

Applying HPCC Systems word vectors to SEC filings

The screenshot shows a GitHub repository page for 'hpcc-systems / EDGAR-SEC-Filings'. At the top, there are buttons for 'Watch 9' and 'Star 1'. Below the repository name, there are tabs for 'Code', 'Issues', 'Pull requests', 'Actions', 'Projects 1', 'Security', and 'Insights'. A dropdown menu shows 'master', '1 branch', and '0 tags'. Buttons for 'Go to file' and 'Code' are also present. The main area lists commits from 'MatthiasMurray' on 'Update README.md' on 7 Aug, with 56 commits. Other listed files include 'Data', 'EDGAR\_Extract', 'Internal', 'SEC\_2\_Vec', 'SEC\_Viz', 'pythonSEC', 'LICENSE', 'README.md', and 'Types.ecl'. On the right side, there are sections for 'About', 'Releases', 'Packages', 'Contributors', and 'Languages'. The 'About' section notes 'No description, website, or topics provided'. The 'Languages' section shows a bar chart with ECL at 91.5% and Python at 8.5%.

## Future Employment Opportunities

- Open positions at LexisNexis Risk Solutions Group:  
<https://risk.lexisnexis.com/group/careers>
- Previous interns have a head start and we can vouch for you
- Hired 2 interns from the 2021 program
- Not all interns are ready for immediate employment
- Intern for a second time or a third...

## Learn more...



- Student wiki - [hpccsystems.com/student-wiki](http://hpccsystems.com/student-wiki)
- Available Projects List - [hpccsystems.com/ideas-list](http://hpccsystems.com/ideas-list)
- Student journals and presentations  
<https://wiki.hpccsystems.com/x/GQFdB>
- Read about the program - [hpccsystems.com/interns/](http://hpccsystems.com/interns/)
- Student Testimonials <https://wiki.hpccsystems.com/x/KoNc>
- HPCC Systems GitHub Repository  
<https://github.com/hpcc-systems>
- Student Posters  
<https://wiki.hpccsystems.com/display/hpcc/HPCC+Systems+Technical+Presentations>



## Get in Touch

---

Lorraine.Chapman@lexisnexisrisk.com