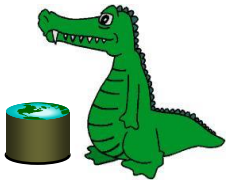
Challenges and Lessons from NIST Data Science Pre-pilot Evaluation in Introduction to Data Science Course Fall 2015

Dr. Daisy Zhe Wang
Director of Data Science Research Lab
University of Florida, CISE Department
March 17th, 2016

DSR @ UF
Data Science Research

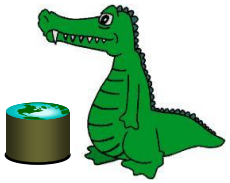


UF Data Science Research Lab

- Founded 2011: http://dsr.cise.ufl.edu/
- 2 Ph.D. Graduated and 7 Ph.D. in progress
- Archimedes: A Master Probabilistic Knowledge Base System (SML + Big data systems)
  - Multimodal extraction, in-Database text analytics, statistical relational model, probabilistic knowledge graphs, probabilistic integration and fusion, scalable inference, crowd-sourcing
- 3-course data science curriculum started 2011
  - Intro, Project, Advanced Topics in DS
  - undergrad + grad level mixed (enrollment: ~15→~100)

DSR @ UF
Data Science Research

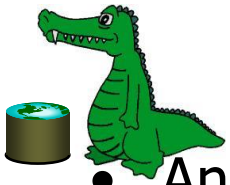
Archimedes



NIST Data Science Pre-Pilot Evaluation as Final Project, Fall 2015

- Introduced on Day 1 of the Intro to DS class
  - Upon invitation of participation to UF DSR Lab from Dr. Bonnie Dorr
  - Raised significant interest in students with ~100 enrollment (undergrad and grad), more on waiting list
  - with the prospect of working with real, practical and important problems and data
  - First course in DS to many and few has experience with Data Science technology
  - Question: Can they pull it off?

DSR @ UF
Data Science Research



Course Outline

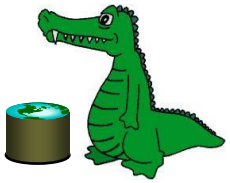
- An introduction to the basic data science techniques including programming in Python, SQL/SPARQL and Map-Reduce for small and big data manipulation and analytics.

- Teach basic techniques for data collection, data preparation, data querying, data analytics including pattern mining, classification, clustering, data visualization, and parallel computing platforms.

- Teach advanced data analytics techniques including NLP, knowledge extraction, graph analytics, graph querying, knowledge bases and crowd sourcing.

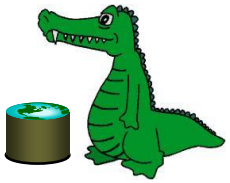
- Introduce key application areas of data science including business intelligence, social media, biomedicine, and e-discovery.

DSR @ UF
Data Science Research



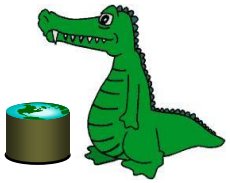
Course Formats

- Pre-requisites: Data Structure and Algorithms
  - Op: DBMS, Stats & Probabilities, Programming langs
- Lectures
  - 28 Lectures on 14 topics (data science pipeline & tech)
  - Guest Lectures (on related DS research topics)
- Labs and Homework 35% (1-2 person, 1-2 weeks)
  - 1 bootcamp on unix/python (a.k.a., lab0)
  - 4 Labs (Pandas, NLTK, map-reduce, scikit)
  - AWS credits (100$/person)
- Final Project 25% (~6 person, 1-2 months)
  - System and algorithm development
  - Presentations & write-ups
- In-class Midterm and pop Quizzes 40%
  - Review lectures



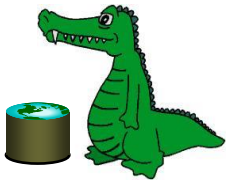
Overview of 2015 NIST Data Science Pre-pilot Participation (I)

- Submissions are *mainly* drawn from Fall 2015 Introduction to Data Science class
  - Cleaning Task: clean traffic lane detector measurements containing incorrect flow values, providing correct traffic flow values for the erroneous traffic flow measurements.
  - Alignment Task: analyze video from camera feeds to detect an event and match it to a separate inventory of traffic events (disabled car, accidents, etc).
  - Prediction Task: develop a system that can predict the number and types of traffic events by type for a given (geographical bounding, interval of time) pair.
  - Forecasting Task: leverage past traffic information and current conditions (weather, maps) to forecast vehicle flows on major roads.

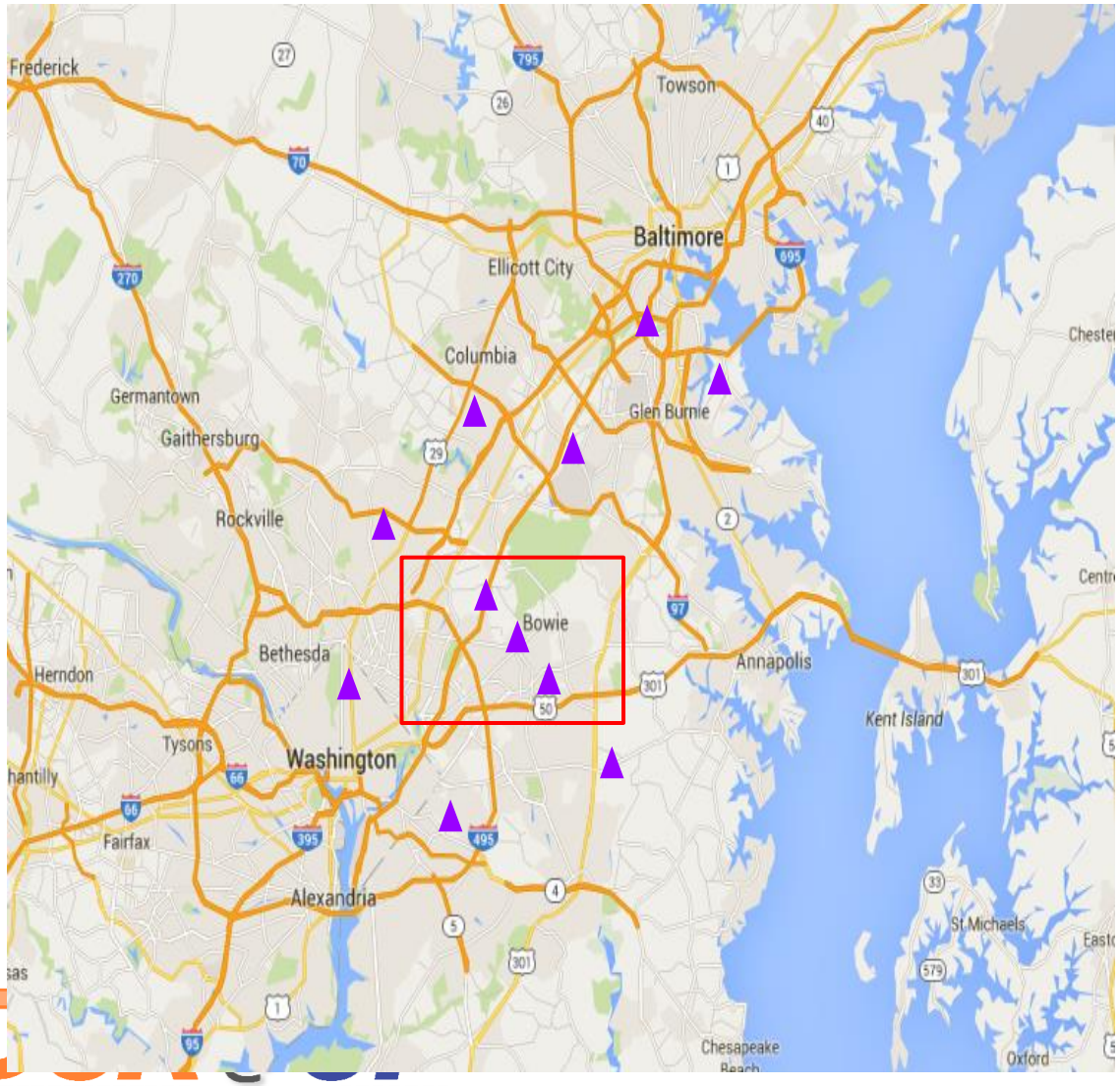


Overview of 2015 NIST Data Science Pre-pilot Participation (II)

- We (TAs did most of the hard work) provided detailed guidance, tutorial & QA
  - Data storage (AWS), access and description/stats/viz
    - Lane-measurement, traffic events, OSM, Weather
  - Task description, IO, Training/Dev data, API
  - Features, examples, baselines, models
    - Cleaning: detection + correction, constraints, smoothing
    - Prediction: road-independent, road-dependent model
  - Piazza forum for QA (~1 month)
    - More AWS credits, scalability, API, integration of data sources, map-reduce, accuracy of prediction
  - TA's prototypes and result post-processing (Ensemble)



Construct training data (road-independent model)



Randomly Selected bounding boxes

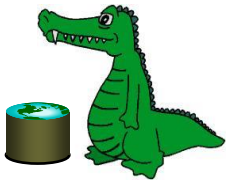
Precipitation events

Precipitation data construction

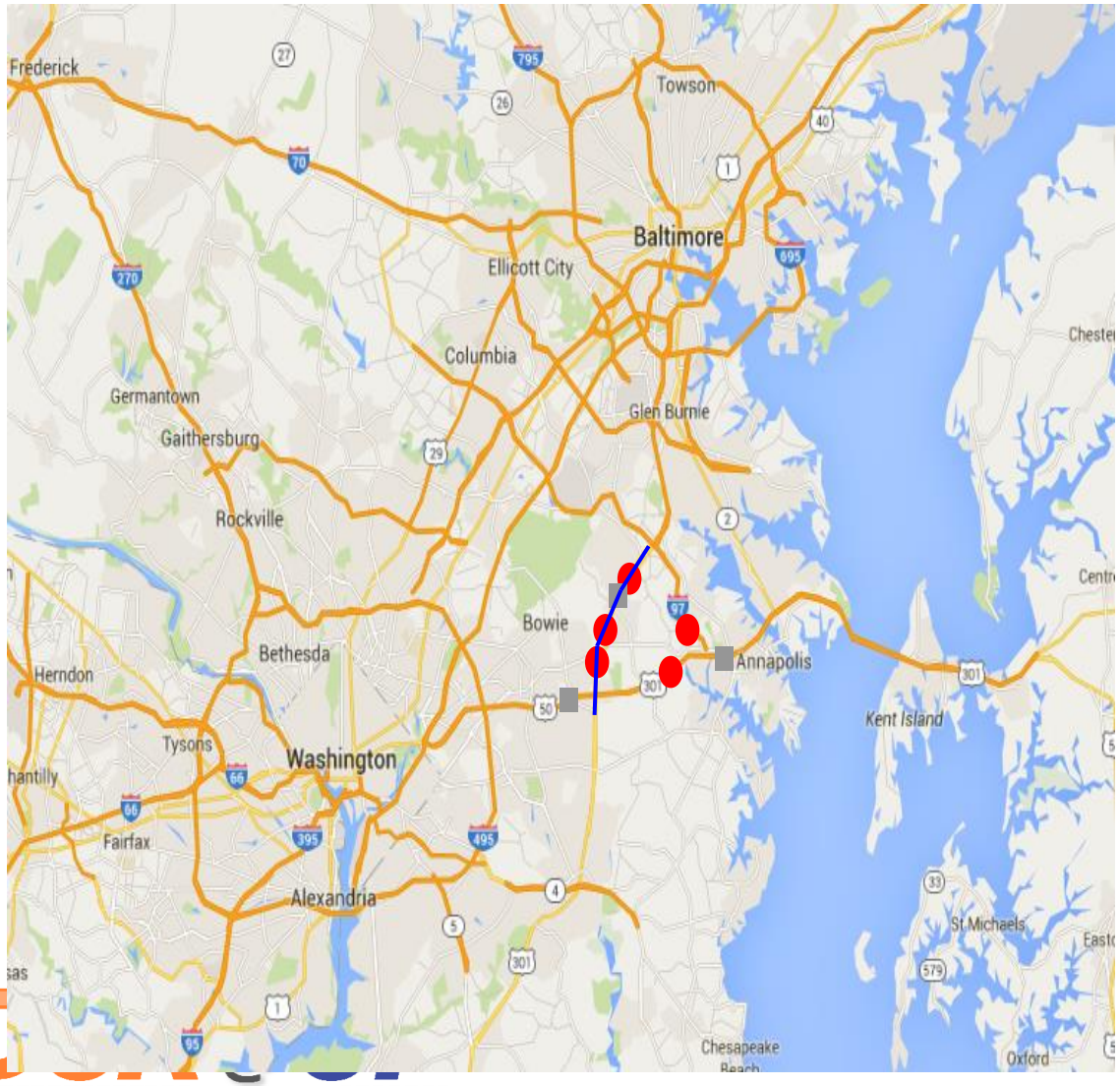
1. Randomly select a region.
2. Count #events of type precipitation occurs within the region in a random time interval of one month.
3. Obtain a single training entry: **(longitude,latitude,w,h,t) #events**
1. Repeat step 1-3 to generate more training entries.

Notes:

1. The shape of selected area should vary noticeably.
2. The method is for your reference only. You are encouraged to come up with other solutions.



Construct training data (road-dependent model)



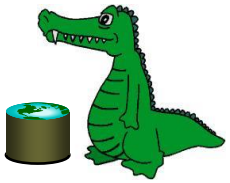
Accident And Incident Events (AAI)
Roadwork Events
Randomly Selected Road Segments

Accident And Incident Events data construction

1. Randomly select a road segment s.
2. Count #events of type AAI occurs within the segment in a random time interval of one month.
3. Obtain a single training entry: **(length,#lanes,time,…) #events**
1. Repeat step 1-3 to generate more training entries.

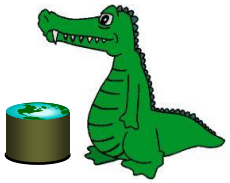
Notes:

1. Selected road segments should have records in the training data.
2. Length of selected road segments should vary noticeably.
3. Time of event counts should vary noticeably (e.g. Jan, Feb, …, Dec).
4. The method is for your reference only. You are encouraged to come up with other solutions.



Overview of 2015 NIST Data Science Pre-pilot Participation (II)

- Choice to participate in NIST DS pre-pilot as the final project -- ~40 students signed up
- 7 groups of 3-6 students participated
  - 4 groups are undergrad, some mixed
  - 3 groups are grad (master + Ph.D)
  - All completed prediction, half completed cleaning
- Question: How well did they do?
  - Notable Groups and Students:
    - 2 (3 masters, best prediction in class)
    - 3 (6 undergrads, second best prediction in class)
    - 7 (6 masters, best cleaning in class)
    - 8 (2 Ph.D. + 1 undergrad, most thorough in exploration)



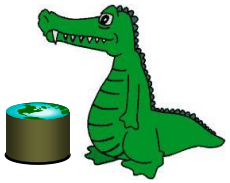
Summary of 2015 NIST Data Science Pre-pilot Results (I)

- Prediction Task – RMSE for 7 submissions
  - TA's Run: 5.17
  - Group 2 Run: 6.10
  - Group 3 Run: 6.52
  - Ensemble 1: 9.04
  - Ensemble 2: 10.10
  - Group 8 Run: 10.23
  - Group 7 Run: 33.44

Prediction Task Score Results

System	Score
ufdsrC	5.17
ufdsrD	6.10
ufdsrE	6.52
ufdsrB	9.04
ufdsrA	10.10
ufdsrG	10.23
ufdsrF	33.44

DSR @ UF
Data Science Research



Summary of 2015 NIST Data Science Pre-pilot Results (II)

- Cleaning Task – 3 sub's
  - Group 7 Run: 0.4007
  - Ensemble Run: 3.8007
  - Group 8 Run: 7.3066

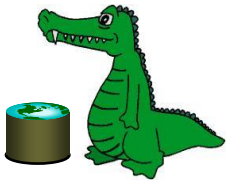
Cleaning task scores for all Systems.

system	score
baselinereferenceA	0.2857
baselinenoinfo	0.2872
ufdsrA	0.4007
ufdsrC	3.8007
ufdsrD	7.3066

Cleaning Task Scores of Systems

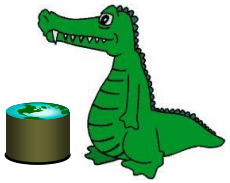
  - Group 7 cost for cleaning: 400$ AWS credits;
  - Total EMR instance time: 830 hours (multiple runs);
  - Time for cleaning (one pass): 22 hours;
  - Hardware: Three clusters consist of 5 * EMR m3.xlarge instances

DSR @ UF
Data Science Research

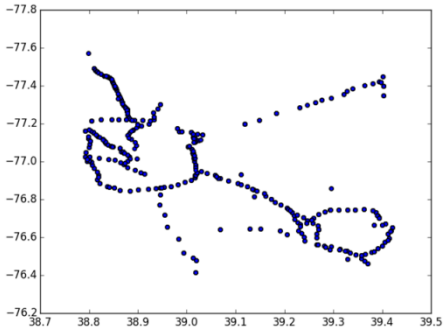


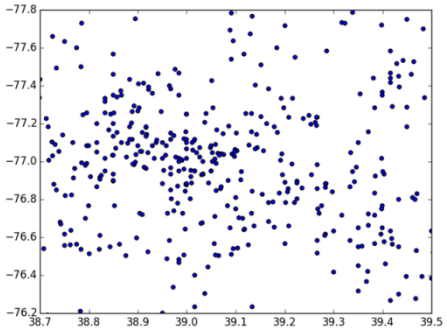
Challenges (I)

- Scalability (Volume)
  - Cleaning task (108 files of total 150G)
  - Python/Pandas does not scale (~400MB file >5 hours)
  - Solution0: run multiple processes in parallel → very expensive and does not scale (*r3.8xlarge* ec2 instances)
  - Solution1: AWS and Map-Reduce (<10min/file)
  - Solution2: JAVA custom implementation with fast I/O and efficient array representation (~13sec/file)
  - Other attempted solutions: GPU, SPARK
  - How long does one run takes? much does one run cost?
    - Days -> Hours -> Minutes. >1000$ -> hundreds -> <100$
    - Scaling to TBs? Scaling prediction?



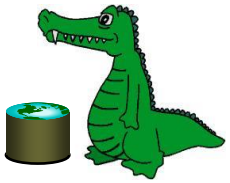
Challenges (II)



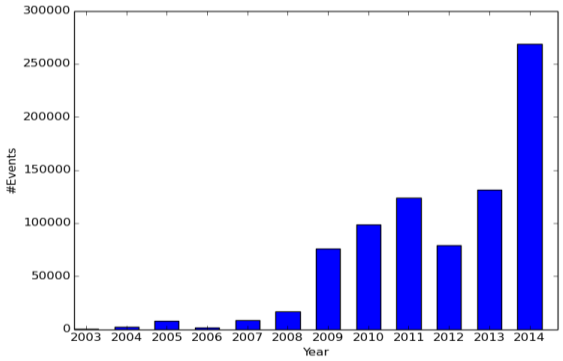


- Data Integration (Variety)
  - 5 groups integrated multiple data sources for prediction
    - Integrating open street map (OSM) data and events
    - Integrating weather and events data
  - Difficulty in learning data API, data format, parsing, understanding, cleaning
  - Difficulty in mapping (e.g. weather stations, events, road segments, bounding boxes)
  - Little correlative and explorative analysis for usefulness for an additional data source

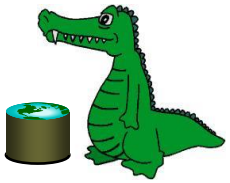
DSR @ UF
Data Science Research



Challenges (III)

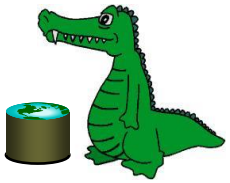


- Modeling (Veracity)
  - Data incompleteness from collection
  - Noisy Data from data collection, cleaning, parsing and integration
  - Data dependency
    - Weather is strongly dependent on time (especially month)
    - Road Features (OSM) are strongly dependent on location
    - Usefulness in incorporating new data source is also dependent on the prediction task (e.g., granularity)
  - Models applied: linear/polynomial regression, KNN, clustering, support vector/decision tree regression
  - Modeling Tools used: Scikit, Spark/MLlib (scaling), Weka



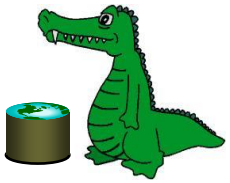
Lessons From the Trenches

- Devise Scalable & Efficient solutions from the start or prepare to throw away the prototype
  - Parallel data processing/machine learning frameworks, efficient data presentation, IO/SerDe
- Simple model first with less data types
  - Fit different models over different slice/dice data
  - Dependency analysis between data types and sources in explorative analysis
  - "The number of events are location, time dependent. ⇒ Three dimensions (x,y,t) work the best."
- Curse of dimensionality
  - Experience from Group 8: "Any feature/dimension we add, the results become worse!"



Feedback (I)

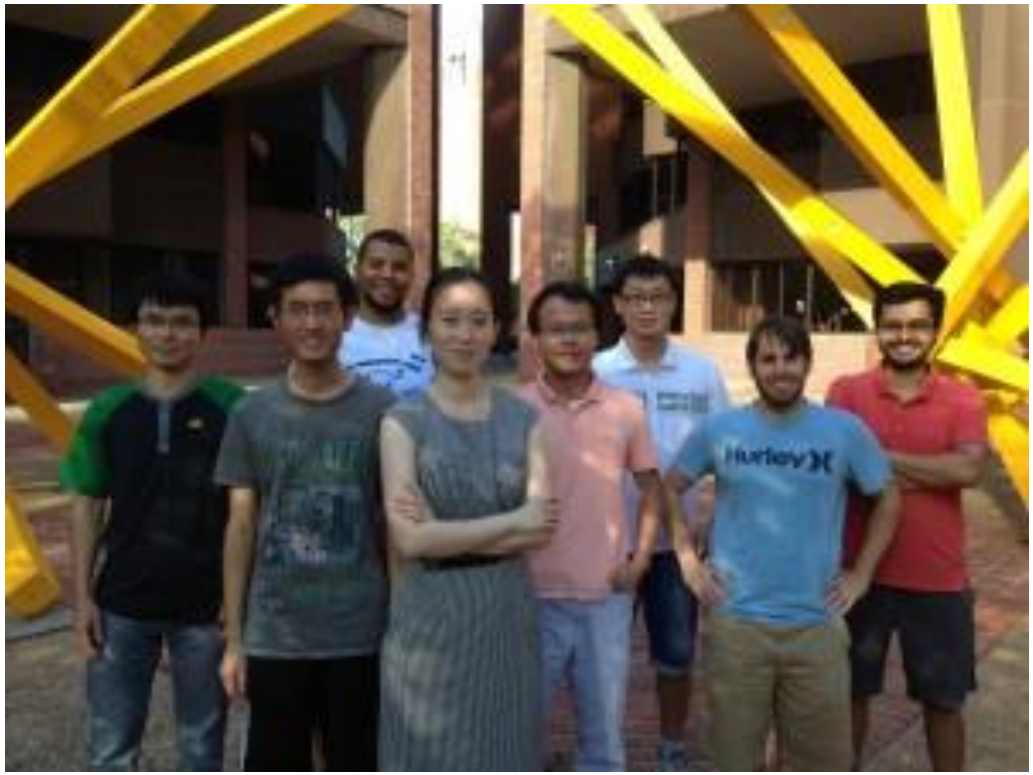
- The NIST pre-pilot DS Eval. Task exercises various muscles of a Data Science engine/pipeline
  - Cleaning, integration/alignment, prediction/forecast
  - Scalable data processing, analytics and statistical modeling
  - Tasks are well-structured/defined (description, test data, training data and performance metrics)
  - Tasks are independent, allow entry from groups with different expertise (e.g., video analysis vs. prediction)
  - Can be done by student groups at undergrad and grad level with different sophistication of models and tools
  - Difficult but Awesome experience → land internships!



Feedback (II)

- How much data collection bias affect modeling?
  - Incompleteness, noisy, dependency, uncertainty
  - Linear/Polynomial Regression over time is a side effect of data collection bias?
- Analysis of change over time and space
  - Plant Identification from Remote Sensing
- Adding computational cost as a new metric?
  - Cloudsort: Minimum cost for sorting a very large amount of data on a public cloud. (http://sortbenchmark.org/)
- Adding Confidence scores, Human computation

Thanks to the hardwork from the TAs and students! Thank You. Questions?



Data Science Lab @ UF CISE
http://dsr.cise.ufl.edu/

DSR @ UF
Data Science Research