DATA HACKATHON

March 6th-8th, 2015

NYC

Data Problems

Please adhere to rules and guidelines, as you work upon deciding your hack and what dataset you are going to use. Please feel free to reach out any of the mentors or organizers if you have any questions or seek clarifications about any of the following datasets or just want feedback on any stage of your project.

Once, again you are not limited to just following problem statements. Feel free to work on a dataset of your own interest. Just make sure, the dataset is publicly available and every participant has equal access to it.

Title	Open data resources on various economic categories
Provided By	RAND State Statistics (a service of Rand Corporation)
Questions	RAND is exclusively opening up their API to our Data Hackathon! Hundreds of economic datasets curated from multiple sources were collected from scanned documents and converted to machine-readable format. This data is not open to the public, so take advantage. RAND data will empower us to make applications geared towards making us a better and smarter society. Feel free to brainstorm with mentors on potentials ideas and questions if you have any. Look into one of the sample datasets about business dynamics: https://www.census.gov/ces/dataproducts/bds/ . A sample problem on this could be: Make a visualization tool to provide insights into the age of establishments and average size of core sectors vis-à-vis financial and service firms. Combining it with Pluto dataset from NYC would bring out interesting insights.
Resources	For the duration of the Hackathon, participating teams will be provided with the login credentials to the portal (http://www.randstatestats.org/statistics.php). Feel free to browse and explore in the meantime. Categories are: Population & Demographics Environment, Resources & Weather Health & Health Care K-12 Education Business & Economics Higher Education Crimes, Prisons & Courts Income, Expenditure, Wealth & Poverty Labor Force, Employment & Earnings Social Insurance & Human Services Energy Transportation & Travel Federal Government

	State & Local Government
Potential	With the new-age data economy, easy availability, cheap computation and much larger problems at civic and personal levels, the frequently updated government dataset opens up a huge space of opportunities to affect our everyday lives with empowered decision-making and vigilant citizenship.

Title	**Promote mixed-used, mixed-income communities anchored by affordable housing?**
Provided By	Accenture
Questions	• Correlate different types of strategic investments to support new housing and neighborhood revitalization and their ability to energize the affordable housing market • Track the ability of mixed-use, mixed-income communities to generate more affordable housing units • Correlate affordable housing to employment rates and workforce development rates • Evaluate the safety and habitability of affordable housing as compared to the rest of the housing inventory • Determine impact and availability of affordable housing to the citizen population of NYC • Quantify delays to affordable housing development versus all other housing inventory • Provide accurate views of available affordable housing to all demographics, including seniors, disabled, and minorities
Resources	Multiple data sources include problems from the NYC Housing Authority (https://data.cityofnewyork.us/data?cat=housing%20%26%20development), Zonal Maps (https://data.cityofnewyork.us/City-Government/Inclusionary-Housing-Designated-Areas/w83z-2kf9), ACRIS datasets on ownership, transfers and mortgages (19 of them: https://data.cityofnewyork.us/data?browseSearch=ACRIS&type=&cat=&scope=). Few datasets are also available on the repo (http://zillowhack.hud.opendata.arcgis.com/, filter to NYC). An example a similar data project was done by Seattle (link)
Potential	Make NYC a better place to live for first-time homebuyers, senior citizens, and low-income renters. This project is for city government and sponsored problems by Accenture. One of the major challenges for the citizens is to find a house considering their mobility patterns, accessibility and transportation. Your hack is going to affect thousand of lives and contribute a great tool, empowering citizens to make robust housing decisions.

Title	Fights against child sexual exploitation
Provided By	Thorn – Defenders of Children, problems shared by Bayes Impact
Questions	Development a visualization tool to gain insights about demand and supply into escort markets? Extract pricing information from unstructured texts and develop tools for accessing this information in a dashboard.
Resources	https://s3-us-west-1.amazonaws.com/thorn-hackathon-escort/escort_all.tar.gz 7+ M rows extracted from 3 escort services.
Potential	Online sex advertising market is huge and growing every year. Pricing information hidden within the texts of these ads can help better understand the demand and explore size-type granularities and pricing variations over cities, a develop a macro level perspective of this business. Over millions of children are exploited in sex trade globally, yet alone the number of potentially risked children US varies from 100k to 300k.

Title	Which hospital to visit for the given condition?
Provided By	NY state Department of Health
Questions	Are you restricted by your Health Insurance provider to visit only a specific hospital? Using the new data provided by NY state about in-patients visits across all hospitals in the state, how can you make smarter decision making when choosing hospital for a selective surgery or a medical condition? Features include cost, number of procedures, distance etc. combined with other metrics such as re-admissions rate and hospital quality.
Resources	https://health.data.ny.gov/Health/Hospital-Inpatient-Discharges-SPARCS-De-Identified/u4ud-w55t
Potential	Health-care is undergoing a revolution. Be part of the solution to crush the clunky and fragile health-care system. Finding the best hospital for a given condition can be challenging to find for selective procedures. Making this task easier would affect thousands of lives every year.

Title	How can we fight Ebola better?
Provided By	EbolaData.org and NDSSL Lab at VirginiaTech
Questions	How to build a disease spread model and related with casualties in the area. Real time map of disease outbreak, mapping epidemiological analysis of data and prevention and cure. Some sample problems people have worked in the past: http://hackebolawithdata.challengepost.com/submissions
Resources	Some of the Data files are available through this Github repo: https://github.com/cmrivers/ebola https://datamarket.com/data/list/?q=ebola&ref=search Suggested reads would be http://nyti.ms/1m3pYLs http://projects.iq.harvard.edu/files/hack/files/hackebolawithdatasummary.pdf Also check: http://www.vbi.vt.edu/ndssl/ebola/resources-overview A few links from Bayes Impact: https://docs.google.com/spreadsheets/d/19hfIVeOlkaBTzTEzgpGuSCDypYgxB0VRhTdgflmareA/edit?pli=1#gid=429291897
Potential	Ebola is one of the most dreaded diseases that sprawled an epidemic across several countries in Africa and killed several thousands. Your hack could potentially be used by the CDC and department of health to combat the disease.

Title	Improve Open source Collaborations
Provided By	Ghtorrent.org
Questions	Find pattern of contribution across open source communities and develop a model to incentivize or boost growth of the contributions. Example of such a tool is visualization of language popularity: http://ghtorrent.org/netviz/ A presentation detailing the dataset release: https://speakerdeck.com/gousiosg/the-ghtorrent-dataset-and-toolsuite
Resources	Trimmed down datasets: http://ghtorrent.org/msr14.html http://ghtorrent.org/vissoft14.html
Potential	The software industry is going through revolution and open source is sweeping across all sets of industries, even companies known for keeping code private are opening up certain projects. However, a robust community that cultivates a healthy environment of growth and effective volunteerism is crucial. You hack would help us understand features of such a community and how improvements can be made to boost

	contributions.

Title	Visualize and understand geo-spatial profile of NYC
Provided By	NYC Gov.
Questions	Mapping and understanding the largest dataset of human settlement ever-performed. It can be used to identify several questions in business, investments and pure research. Several questions can be asked such as- How zoning policies affect affordable housing?
Resources	Available here-- https://data.cityofnewyork.us/City-Government/Primary-Land-Use-Tax-Lot-Output-PLUTO-/xuk2-nczf. An interesting dataset that can be combined with this is Historical Crime Data of NYC: https://nycopendata.socrata.com/Public-Safety/Historical-New-York-City-Crime-Data/hqhv-9zeg Also check CartoDB's mapping of the given dataset-- http://cartodb.com/gallery/pluto/
Potential	Pluto is (short for Property Land Use Tax lot Output), which is a detailed tract of every piece of property in the city. The rows are unique tax lots, and each column describes an attribute — things like property value, number of buildings on the lot, square footage. Understanding and mapping this dataset is a huge challenge and potential business problems would affect several thousand lives.

Title	Risk of Flooding
Provided By	National Oceanic and Atmospheric Administration
Questions	What areas and regions are most exposed to flooding from sea level rise, storm or river-flooding?
Resources	Climate datasets- http://tidesandcurrents.noaa.gov/sltrends/sltrends.html http://tidesandcurrents.noaa.gov/est/ http://water.weather.gov/ahps/forecasts.php http://nationalmap.gov/ http://www.eia.gov/tools/faqs/faq.cfm?id=767&t=3 http://www.census.gov/geo/maps-data/data/tiger.html
Potential	Your hack would help researchers and policy makers reduce the vulnerability of flood hazards, and make reasonably accurate assessments to mitigate risk due to flooding on various infrastructure projects.

Title	Stopping Blooming of Toxic Algae
Provided By	Environmental Protection Agency, problems shared by Bayes Impact
Questions	What can we learn from MERIS satellite imagery and EPA's EnviroAtlas about harmful toxic algae blooms? Compare cyanobacteria algal bloom cell counts against different ecosystem services information. Focussing on a GIS spatial/temporal exercise. More on the problem: http://oceanservice.noaa.gov/hazards/hab/
Resources	Link to visualization and dataset download is available here: http://service.ncddc.noaa.gov/website/AGSViewers/HABSOS/maps.htm https://docs.google.com/document/d/1mBIF9BeFJlWXWM_1JFxQPJBIFCRAorOtkd4JE_3JKSM/edit?usp=sharing EnviroAtlas: http://enviroatlas.epa.gov/enviroatlas/atlas.html
Potential	Algae blooms gave caused poisoning of several water bodies and is a big threat to aquatic life, also putting risk to humans. Satellite instruments can be used to detect and quantify cyanobacteria blooms in lakes and estuaries. The current challenge is to refine a predictive model based on the best available science to forecast the location of cyanobacteria cell counts up to 7-days in advance of the previous available satellite observation. The model would be based on the best and most appropriate scientific information available to predict the growth and movement of cyanobacteria blooms. It's anticipated that the model would incorporate ancillary data sets in near-real time to provide a robust predictive capability (e.g., temperature data, meteorological

	predictions, etc.).

Title	Map search for small business planning
Provided By	Yelp, problems by Accenture
Questions	Predict star ratings / review counts / review sentiments / (....) by geography. Start by mapping / visualizing neighborhoods or geobins by cumulative business stats-- review count, stars, avg. review, etc. Refine the dataset by useful attributes Feature extraction from text is a major task here-- go nuts! Topic models, tf-idf, sentiment all fair game. Train a predictive model based on resulting feature vectors: location, categories, attributes, textual features. (Hint: start with regression.) Wrap it in an interface that takes some business attributes as input, and outputs the predicted rating. **Or**, input the features, and the application outputs the location that maximizes ratings according to some criteria.
Resources	Data is from: http://www.yelp.com/dataset_challenge *Some questions to consider:* • Make sure the strategy you use to segment the data for train / test makes sense-- reviewers from different populations might have different (latent) characteristics. • There are a number of ways you can organize the data by location-- geobins (as above) are high-resolution; neighborhood tags might be simpler but harder to visualize. How will you evaluate your model?
Potential	How yelp reviews data can be used to monitor business activity and aggregate it with other datasets such as Pluto to find affordable planning of small business. Your hack could potentially be used by several thousand small business owners.

Title	Restaurant recommender
Provided By	Yelp, problems by Accenture
Questions	• Train a model on the review dataset, and, given a user id, recommend a restaurant they'd like. • This is something of a canonical project for this sort of data, and there are many ways you could go about itt: Train a topic model on review text Extract sentiment from the the review text Weight reviews based on funny, cool, useful votes -- something we haven't touched here etc. There are a lot of recommender algorithms out there-- find one that suits this best.

Resources	Data is from: http://www.yelp.com/dataset_challenge *Questions to consider:* • The review dataset is not grouped by user_id. You will need to JOIN the records by user. It's a fairly large file, so consider your strategy carefully. (Hint: check out EMR or BigQuery.) • Batch loading? Compressed representations? • What evaluation metrics will you use? Do they make sense? There's no need to restrict this solely to restaurants, either, but note that the file is large. Do reviewer populations for different business categories differ in nontrivial ways?
Potential	Recommendations currently deployed aren't working to the best of our needs. Your hack will potentially affect every meal you decide to eat outside in NYC.