DATA HACKATHON

Feb 12th-14th, 2016

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 all participant has equal access to it.

Title	Uber and NYC Taxi & Limousine Commission (TLC) Data
Provided By	FiveThirtyEight
Questions	• How do car-sharing affect traffic of New York City? • Future of native Taxi industry?
Resources	Link: https://github.com/fivethirtyeight/uber-tlc-foil-response
Potential	You hack could be potentially used for consumer use for navigating New York and also design better policies regarding car-sharing.

Title	Patent Examination Research Dataset
Provided By	USPTO
Questions	• Understand intellectual property applications and patterns in innovation
Resources	Link: http://www.uspto.gov/learning-and-resources/electronic-data-products/patent-examination-research-dataset-public-pair
Potential	You hack could be potentially used by thousand of innovators and investment firms to understand trove of patent application datasets.

Title	AirBnb Listing Database
Provided By	Inside AirBnb
Questions	• What makes a Airbnb rental charge more money? (walkability score, or crime score?)
Resources	Link: http://insideairbnb.com/get-the-data.html
Potential	You hack would be potentially used by policy makers and consumers to understand disruption in the hotel market.

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	**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	**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.
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	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.