DATA SCIENCE SCENARIO : CRIMINAL ACTIVITIES

An international government contractor, working under contract for a US-based intelligence agency chartered with US national security, needs to conduct a study to help the agency focus resources on identifying and thwarting sources of crime in the world. There is a strong belief that correlative patterns in many locales could amplify and enable criminal syndicate activities in the United States, the United Kingdom, the EU, and western economies in general. In order to better plan their global investigations and investments over the next 10 years, the agency need to pinpoint areas to deploy undercover agents, recruit informants, and insert top-secret data gathering technologies. While many sources for data for criminal activities currently exist, the contractor requires an investigation that will identify relationship networks, potential causes, consequences, and correlates that enable and feed criminal activities. The study should answer the following questions:

Key Questions:

- In the aggregate, which areas or regions in the world are sources for the greatest amount of criminal activities? What is the basis for ranking?
- Is it possible, using public information, to identify and map crime-enabling relationship networks on a governmental, organizational, or individual level?
- Is it possible to find correlative patterns associated with criminal activites?
- Is it possible to predict crime levels or patterns in a given locale given correlates?
- To achieve optimal results (with focus on ROI), which nations, networks, or correlates should be addressed first?

In addition to the many data sets referenced on the Quality of Government site (http://qog.pol.gu.se/), the investigation should also identify and cite at least 3 additional data sources, and determine the current and projected costs, if any, for on-going data acquisitions costs.

OUTPUT (ARTIFACTS)

- Identify data sources
- Preliminary study of the data
- A plausible data architecture for ingestion and analysis
- POC architecture
- Ingestion data and wrangling report
- Iterative / interactive analysis results
- A set of models for projecting outcomes based on interactive assumptions
- Interactive dashboard
- Key findings
- Answers to the Key Questions

GOVERNMENT CONTRACTOR DATA

P.C. Labs

The government contractor in question does business globally under the name P. C. Labs (**Project Cadmus Labs**), a secretive organization that specializes in genetic engineering and high technology surveillance. Little is known of the secretive organization, other than it is private, allegedly has been involved with a laundry list of governmental agencies including the NSA, the CIA, the FBI, the Torchwood Institute, MI6, and Europol.

CEO: Professor Dabney Donovan

The CEO of P.C. Labs, **Professor Dabney Donovan**, is a proponent of genetic engineering without limits and sometimes uses P.C. Labs in tandem with intelligence agency projects to further his own goals. He is not a proponent of Data Science solutions and typically invests in human enhancement protocals and survellience assets and technologies.

There is a component of the contract in question that \ requires P.C.Labs to provide a preliminary study of the plausibility of using public data sources to identify correlates to crime and even create useful predictive models if possible. Donovan does not believe the study is useful and has only agreed to the subcontract in order to do other work under the contract which is more in line with his interests.

Head of Labs: Don Snow

Don Snow was seconded to Cadmus labs in April 2015, and since then works with Cadmus as an external consultant, now in the lead on P.C.Labs internal "Safety Everywhere" program. Don leads a small team of security analysts and works on matters of high priority for national security. His background is in computer science,

software engineering, information retrieval, database engineering, statistics, application management and software security.

A direct subordinate of Professor Donovan, Snow merged into the Professor's direct analyst team for the previous 6 months, even though Snow is still maintans direct relationships with many NSA elements outside of the Cadmus hierarchy. Don and the Professor do not have a good management relationship and do not communicate effectively.

Snow runs an internal team of data analysts operating from the P.C.Labs Cheyanne Mountain complex who have access to much more sensitive crime information and generate internal security products for other analysts, assets, and policy makers. His team needs more people to build these large-scale open-source projects. If successful Think Big will support Snow's team directly and perhaps open up other avenues for further sensitive contracts. It is likely that P.C.Labs will security vet the team and bring them into the organization for around 6 months before any project can begin. *According to Don*, Don's opinion matters in qualifying the technical capabilities of Think Big to begin the process of accepting them as a contractor.

FREE DATA SOURCES

GENERAL

1. **UNData:** A statistical database of all United Nations data.

2. **Amazon Public Data Sets:** A repository of large datasets relating to biology, chemistry, economics, and physiology, including the Human Genome Project.

3. **Pew Research:** Public opinion polls, demographic research, content analysis, and other data-driven social science research.

4. **Google Scholar:** A wide array of information, including articles, theses, books, abstracts, white papers, and court opinions.

5. **Datasets Subreddit:** A dive into anything and everything, from English grain prices of the 14th Century to U.S. homelessness rates.

6. **FiveThirtyEight:** Statistical analysis that tells compelling stories about elections, politics, sports, science, economics, and more.

7. **Qlik DataMarket:** A place to check out data related to economics, healthcare, food, agriculture, and the automotive industry.

8. **The Upshot by New York Times:** News, analysis, and graphics about politics, policy, and everyday life.

CONTENT MARKETING

9. **Content Marketing Institute:** The latest news, studies, and research on content marketing.

10. **Buffer:**Data insights on digital marketing.

11. **Moz:** Insights on SEO.

12. **HubSpot:** A large repository of marketing data.

CRIME

13. **Bureau of Justice Statistics:** Information on anything related to U.S. justice system, including arrest-related deaths, census of jail inmates, national survey of DNA crime labs, surveys of law enforcement gang units, etc.

14. **Uniform Crime Reporting Statistics:** Statistics on violent crime, such as murder, rape, robbery, and assault; has decades of data at city, county, state, and national levels.

15. **FBI Crime Statistics:** Statistical crime reports and publications detailing specific offenses and outlining trends to understand crime threats at both local and national levels.

16. **National Archive of Criminal Justice Data:** Original research based on archived data concerning criminal justice and criminology.

DRUGS

17. **Drug Data and Database by First Databank:** Drug data and drug databases provided with the hope of drug knowledge inspiring change in the medication decision-making process.

18. **U.S. Food and Drug Administration:** Drug approvals and databases, including therapeutic equivalence evaluations for approved multi-source prescription drug products.

19. **National Institute on Drug Abuse:** Resources that cover a variety of drug-related issues, such as drug usage, emergency room data, and prevention and treatment programs.

20. **United Nations Office on Drugs and Crime:** Research, trend analysis, and forensics with global and regional data collections.

21. **Drug War Facts:** Thorough look at drugs and drug policy, applied to public health and criminal justice issues.

EDUCATION

22. **National Center for Education Statistics:** The primary federal entity for collecting and analyzing data related to education.

23. **Government Data About Education:** Education datasets, apps, resources for the classroom, and details about paying for college.

24. **Education Data by the World Bank:** Comprehensive data and analysis source for key topics in education, such as literacy rates and government expenditures.

25. **Education Data by Unicef:** Data related to sustainable development, school completion rates, net attendance rates, literacy rates, and more.

ENTERTAINMENT

26. **BLS: Arts, Entertainment, and Recreation:** Related industries at a glance, with statistics and datasets relevant to arts, entertainment, and recreation.

27. **Million Song Dataset:** A collection of 28 datasets containing audio features and metadata for a million contemporary popular music tracks.

28. **The Numbers:** Detailed movie financial analysis, including box office, DVD and Blu-ray sales reports, and release schedules.

29. **BFI Film Forever:** Research data and market intelligence focused on the UK film industry and film culture.

30. **IFPI:** Global statistics about the recording industry.

31. **Statista: Video Game Industry:** Statistics and facts about the video game industry, ranging from global gaming software expenditure to U.S. brand equity of Nintendo Wii.

32. **Statista: Film Industry:** Statistics and facts about the film industry, from the number of movie tickets sold in U.S. and Canada to the number of 3D cinema screens worldwide.

33. **Statista: Music Industry:** Statistics and facts about the music industry, ranging from concert revenue to record company market share.

34. **Academic Rights Press:** A repository of historical and current music sales data with insight on how such numbers can be applied.

ENVIRONMENTAL/WEATHER DATA

35. **Environmental Protection Agency:** Information for more than 540 chemical substances, containing information on human health effects that may result from exposure to various substances in the environment.

36. **National Center for Environmental Health:** Nationally funded data systems that have a relationship to environmental public health.

37. **National Climatic Data Center:** Quick links from the National Oceanic and Atmospheric Administration, covering everything from storm data to climate indices.

38. National Weather Service**:** Climate data, including past weather conditions and long-term averages, from specific observing stations around the United States.

39. Weather Underground**:** Tracked weather by regional radar, regional severe weather, and global temperatures.

40. National Centers for Environmental Information**:** Weather record published since 1927, including monthly mean values of pressure, temperature, precipitation, and station metadata notes documenting observation practices and station configurations.

41. WeatherBase**:** Travel weather, climate averages, forecasts, current conditions, and normals for 41,997 cities worldwide.

42. International Energy Agency Atlas**:** A look at climate change that focuses on how each country produces and consumes energy.

FINANCIAL/ECONOMIC DATA

43. Google Finance**:** Real-time stock quotes and charts, financial news, currency conversions, or tracked portfolios.

44. Google Public Data Explorer**:** Searchable large datasets on economic development worldwide.

45. U.S. Bureau of Economic Analysis**:** U.S. economic statistics, including national income and gross domestic product.

46. National Bureau of Economic Research**:** Macro data, industry data, productivity data, trade data, international finance, data, and more.

47. U.S. Securities and Exchange Commission**:** Quarterly datasets of extracted information from exhibits to corporate financial reports filed with the Commission.

48. World Bank Open Data**:** Education statistics about everything from finances to service delivery indicators.

49. Financial Data Finder at OSU**:** Plentiful links to anything related to finance, no matter how obscure.

50. IMF Economic Data**:** Global financial stability reports, regional economic reports, international financial statistics, exchange rates, directions of trade, and more.

51. The Atlas of Economic Complexity**:** Analysis of trade flows and the sectoral composition of an economy with data visualizations.

52. World Bank Doing Business Database**:** An incredibly useful source of information that evaluates business environment indicators around the world, including trade capabilities and costs.

53. UN Comtrade Database**:** Raw data on high-level trade with visualizations.

54. Global Financial Data**:** Covers 60,000 companies across 300 years, analyzing the twists and turns of the global economy.

55. Visualizing Economics**:** Data visualizations about the economy.

GOVERNMENT/WORLD

56. The CIA World Factbook**:** Facts on every country, dependency, and geographic entity in the world; focuses on history, people, government, economy, energy, geography, communications, transportation, military, and transnational issues.

57. U.S. Census Bureau**:** Government-informed statistics on population, economy, education, geography, and more.

58. Data.gov**:** Open data of the U.S. government, focuses on everything from agriculture and ecosystems to manufacturing and science.

59. Unicef**:** Evidence on the situation of children and women around the world to inform national and global decision-making.

60. Data Catalogs**:** Comprehensive list of open data catalogs in the world, curated by a group of leading open-data experts.

61. European Union Open Data Portal**:** – Data pulled from European Union institutions.

62. Open Data Network**:** Government-related data with some visualizations tools built in.

63. Gapminder**:** Massive collection of data sources that cover everything from agriculture and employment to aid given and death.

64. Land Matrix (Transnational Land Database)**:** A meticulously developed database of international land transactions with plenty of visualization tools.

65. The World Bank's World Development Indicators**:** Huge collection of national data on hundreds of indicators, with data on every country.

66. UNDP's Human Development Index**:** A ranking of country progress under the lens of human development.

67. OECD Aid Database**:** Visualized data regarding aid collected from governments.

HEALTH

68. HealthData.gov**:** High-value health data for entrepreneurs, researchers, and policy makers; includes data on Medicaid, Medicare, clincial studies, and treatments.

69. Centers for Disease Control and Prevention**:** Public health data and statistics by topic, from alcohol use to viral hepatitis.

70. **World Health Organization:** Information, data, statistics, and reports concerning international public health.

71. **President's Council on Fitness, Sports & Nutrition:** Information aimed to promote, encourage, and motivate Americans of all ages to become physically active and participate in sport.

72. **Partners in Information Access for the Public Health Workforce:** A collaboration of U.S. government agencies, public health organizations, and health sciences libraries.

73. **Health Services Research Information Central:** Selective links aimed at providing information and data regarding health services resources.

74. **MedicinePlus:** Health statistics ranging from percentage of obese citizens to rates at which people are catching the flu.

75. **National Center for Health Statistics:** Datasets, documentation, data access tools, growth charts, and resources for further vital records.

76. **America's Health Rankings:** Health reports that view the nation holistically, with in-depth data and analysis.

77. **Health & Social Care Information Centre:** National provider of information, data, and IT systems for health and social care.

78. **Medicare Hospital Quality:** A database on complication rates by hospital for interesting comparisons.

79. **SEER Cancer Incidence:** Cancer-related statistical summaries, interactive tools, and publications.

80. **The BROAD Institute:** Cancer program legacy publication resources and cancer-related datasets.

HUMAN RIGHTS

81. **Amnesty International:** Human rights information, run independent of any political ideology, economic interest, or religion.

82. **Human Rights Data Analysis Group:** Nonprofit, nonpartisan group applying rigorous science to the analysis of human rights violations around the world.

83. **Harvard Law School:** A collection of links that cover a variety of topics, including everything from international relations and human rights data, from political institution databases.

84. **The Armed Conflict Database by Uppsala University:** A look at fragile and conflict-affected states that dives into minor and major violent conflicts around the world.

LABOR/EMPLOYMENT DATA

85. **Bureau of Labor Statistics:** U.S. government's data collection of employment-related stats across regions, states, and local areas.

86. **Department of Labor:** Closely watched measures of employment and unemployment.

87. **U.S. Small Business Administration:** Employment data from business owners' perspective, including economic indicators and projections.

88. **Employment by U.S. Census:** Data that measures the state of the nation's workforce, including employment and unemployment levels, as well as weeks and hours worked.

POLITICS

89. **Open Secrets:** Nonpartisan, independent, and nonprofit; nation's premier research group tracking money in U.S. politics and its effect on elections and public policy.

90. **Crowdpac:** Calculates objective scores for political candidates showing their overall political position and their position on specific issues.

91. **Gallup:** Data-driven news based on U.S. and world polls.

92. **Real Clear Politics:** A look at everything from policy support to election polling data.

93. **Intro to Political Science Research by UC Berkeley:** Statistics and data for those interested in political science; an ideal starting place.

94. **California Field Poll:** Independent, nonpartisan, media-sponsored public opinion news service that examines California public opinion.

95. **Rand State Statistics:** Social science data for the U.S. at the national, state, and local levels.

96. **Roper Center for Public Opinion Research:** U.S. and international polling and public opinion survey data.

SOCIAL

97. **SocialMention:** Real-time social media search and analysis.

98. **Google Trends:** Data and trends by search engine engagement.

99. **Facebook Graph:** API that pulls data about Facebook engagement.

TRAVEL/TRANSPORTATION

100. **Bureau of Transportation Statistics:** Transportation statistical data, research activities, and budgetary resources.

101. **Monthly Tourism Statistics – U.S. Travelers Overseas:** A look at U.S. international air passenger statistics.

102. **SkiftStats:** Latest statistics, research, and data about the travel industry.

103. **Search the World:** Statistics, population, weather, webcams, and travel information for millions of locations worldwide.

104. **U.S. Travel Association:** Covers a wide variety of travel-related topics, from impacts of travel on state economies to analysis of what a stronger dollar means for the travel industry