# Overview

Claudio Delrieux

# Overview

- We will now present the candidate problems for your term projects.
- First, we will present a hypothetic, but realistic scenario in which there are specific requirements for a data science (DS) solution.
- The requirements encompass all the stages that we described in the prior lectures.
- The datasets are more complex and unclean than the ones we used in the midterm projects.

# Overview (cont.)

- In each of these scenarios, two or three specific problem statements will be presented.
- Each states clearly the problem, the dataset, and the goals. Please carefully read all the descriptions before you make your choice.
- Working in two-person groups is recommended, but individual projects are also welcome.

Overview

# The End

# The Kronos Incident Scenario

Claudio Delrieux

# The Kronos Incident Scenario

The Kronos incident

- In the years GAStech has been operating a natural gas production site in the island country of Kronos, it has produced remarkable profits and developed strong relationships with the government of Kronos. However, GAStech has not been as successful in demonstrating environmental stewardship. In January 2014, the leaders of GAStech are celebrating their newfound fortune as a result of the initial public offering of their very successful company.

# The Kronos Incident Scenario (cont.)

- In the midst of this celebration, several employees of GAStech go missing. An organization known as the Protectors of Kronos (POK) is suspected in the disappearance, but things may not be what they seem.
- Your task is to bring law enforcement up to date on the current organization of the POK and how that organization has changed over time, as well as to characterize the events surrounding the disappearance.

# The End

# Term Project 1

## The Kronos Incident: The Kidnapping

Claudio Delrieux

# Term Project 1: The Kronos Incident The Kidnapping, Part I

Historical analysis of the Protectors of Kronos

- You are provided with a set of current and historical news reports at your disposal, as well as resumes of numerous GAStech employees and email headers from two weeks of internal GAStech company email.
- You are being counted on to bring law enforcement up to date on the current organization of the POK and how that organization has changed over time, as well as to characterize the events surrounding the disappearance.

# Term Project 1: The Kronos Incident The Kidnapping, Part II

## Data

- Data provided consists of a collection of text-based files dealing with the kidnapping of the GAStech employees by members of the social movement group POK. As an analyst, you have the following data at your disposal.
  1. A map of Kronos
  2. A chart describing the local GAStech organization, in PDF format

# Term Project 1: The Kronos Incident The Kidnapping, Part III

## Data (cont.)

3. An Excel file of GAStech employee records: the primary worksheet contains the data; the index worksheet contains the data dictionary
4. Email headers from two weeks of internal GAStech company email, in comma-separated values (CSV) format
5. Resumes and short biographies of many, but not all, of the GAStech employees, in Microsoft Word format

# Term Project 1: The Kronos Incident The Kidnapping, Part IV

## Data (cont.)

6. Historical reports and descriptions of the countries involved, in Microsoft Word format
7. Relevant current and historical news reports from multiple domestic and translated foreign sources, in text file format: because these articles have come from multiple sources and original formats, some of them may contain corrupted characters, which is typical for this type of data; these corrupted characters should not interfere with your ability to analyze the data

# Term Project 1: The Kronos Incident The Kidnapping, Part V

**Goals**

1. Provide a clear analysis of the structure of the POK network, with supporting evidence.
   a. Who are the leaders?
   b. Who is part of the extended network?
   c. How has the group structure and organization changed over time?
   d. Where are the potential connections between the POK and GAStech?

# Term Project 1: The Kronos Incident The Kidnapping, Part VI

**Goals (cont.)**

2. Describe the events of January 20–21, 2014. What is the timeline of events?
3. Provide at least two possible explanations why the GAStech employees may be missing. What evidence do you have to support each of these explanations?

# Term Project 1: The Kronos Incident The Kidnapping, Part VII

In order to explore this data, you will need to perform several kinds of data integration activities and analyses to generate hypotheses about the **missing employees.** The main structure of the POK is described in the two historical reports dated five and ten years ago. Updates can be found in news articles and in some of the other datasets. Your investigation into this text data can be supported by text analytic tools.

# Term Project 1: The Kronos Incident The Kidnapping, Part VIII

When analyzing emails headers (containing *to, from, date,* and *subject* fields), try to reconstruct some sort of a communication network within the GASTech organization during the period for which data was provided. You should find clues to POK members or sympathizers, plus hints at socialization patterns among employees.

# The End

# Term Project 2

## The Kronos Incident: Geospatial-Temporal Patterns

Claudio Delrieux

# Term Project 2: The Kronos Incident
# Geospatial-Temporal Patterns, Part I

Geospatial-temporal patterns of life analysis

- In the years GAStech has been operating a natural gas production site in the island country of Kronos, it has produced remarkable profits and developed strong relationships with the government of Kronos. However, GAStech has not been as successful in demonstrating environmental stewardship. In January 2014, the leaders of GAStech are celebrating their newfound fortune as a result of the initial public offering of their very successful company.

# Term Project 2: The Kronos Incident
# Geospatial-Temporal Patterns, Part II

- In the midst of this celebration, several employees of GAStech go missing. An organization known as the Protectors of Kronos (POK) is suspected in the disappearance, but things may not be what they seem.
- Many of the Abila, Kronos-based employees of GAStech have company cars which are approved for both personal and business use. Those who do not have company cars have the ability to check out company trucks for business use, but these trucks cannot be used for personal business.

# Term Project 2: The Kronos Incident
# Geospatial-Temporal Patterns, Part III

- Employees with company cars are happy to have these vehicles, because the company cars are generally much higher quality than the cars they would be able to afford otherwise.
- However, GAStech does not trust their employees. Without the employees' knowledge, GAStech has installed geospatial tracking software in the company vehicles. The vehicles are tracked periodically as long as they are moving.

# Term Project 2: The Kronos Incident
# Geospatial-Temporal Patterns, Part IV

- This vehicle tracking data has been made available to law enforcement to support their investigation. Unfortunately, data is not available for the day the GAStech employees went missing. Data is only available for the two weeks prior to the disappearance.
- In addition to the vehicle data, law enforcement has been given access to the personal and business credit and debit card transactions for the local GAStech employees for the two weeks preceding the kidnapping.

# Term Project 2: The Kronos Incident
# Geospatial-Temporal Patterns, Part V

- Many of the GAStech employees also use loyalty cards to gain discounts or extra benefits at the businesses they patronize, and law enforcement has been given access to two weeks of this loyalty card data as well.
- Your goal is to make sense of this data to identify suspicious patterns of behavior and to prioritize which of these may be related to the missing staff members. You must cope with uncertainties that result from missing, conflicting, and imperfect data to make recommendations for further investigation.

# Term Project 2: The Kronos Incident
# Geospatial-Temporal Patterns, Part VI

- **Data:** You have the following data at your disposal.
  1. A list of vehicle assignments by employee, in CSV format (car-assignments.csv)
  2. Employee last name
  3. Employee first name
  4. Car ID (integer)
  5. Current employment type (department; categorical)
  6. Current employment title (job title; categorical)
  7. ESRI shapefiles of Abila and Kronos
  8. A CSV file of vehicle tracking data (gps.csv, timestamp, car ID (integer), latitude, longitude)

# Term Project 2: The Kronos Incident
# Geospatial-Temporal Patterns, Part VII

- **Data:** You have the following data at your disposal. (cont.)
  9. A CSV file containing loyalty card transaction data (loyalty_data.csv, Timestamp, Location (name of the business), Price (real), FirstName (first name of the card holder), LastName (last name of the card holder))
  10. A CSV file containing credit and debit card transaction data (cc_data.csv, Timestamp, Location (name of the business), Price (real), FirstName (first name of the card holder), LastName (last name of the card holder))
  11. A tourist map of Abila with locations of interest identified, in JPEG format (map-tourist.jpg)

# Term Project 2: The Kronos Incident
# Geospatial-Temporal Patterns, Part VIII

- Map data, including shape files of Abila city streets and a visitor's map of the city sites and shops, are provided to you as an additional support to understand patterns of life. For example, even though employees' home addresses were not provided, patterns of life analysis may reveal where they spent their evening hours, typically indicating their home address. Regular gatherings at what appeared to be residences, but not corresponding to any employee's home address, could suggest different hypotheses.

# Term Project 2: The Kronos Incident
# Geospatial-Temporal Patterns, Part IX

**Goals**

1. Describe common daily routines for GAStech employees. What does a day in the life of a typical GAStech employee look like?
2. Identify up to twelve unusual events or patterns that you see in the data. If you identify more than twelve patterns during your analysis, focus your answer on the patterns you consider to be most important for further investigation to help find the missing staff members.

# Term Project 2: The Kronos Incident
# Geospatial-Temporal Patterns, Part X

**Goals (cont.)**

3. For each pattern or event you identify, describe it.
   - What is the pattern or event you observe?
   - Who is involved?
   - What locations are involved?
   - When does the pattern or event take place?
   - Why is this pattern or event significant?
   - What is your level of confidence about this pattern or event? Why?

# Term Project 2: The Kronos Incident
# Geospatial-Temporal Patterns, Part XI

- Like most datasets, the data you were provided is imperfect, with possible issues such as missing data, conflicting data, data of varying resolutions, outliers, or other kinds of confusing data.
- Considering data is primarily spatiotemporal, describe how you identified and addressed the uncertainties and conflicts inherent in this data to reach your conclusions in questions 1 and 2.

# The End

# Term Project 3

## Return to Kronos

Claudio Delrieux

# Term Project 3: Return to Kronos, Part I

- After the successful resolution of the kidnapping at GAStech's office in Abila (Kronos), GAStech officials determined that Abila offices needed a significant upgrade. At the end of 2015, the growing company moved into a new, state of the art, three-story building near their previous location.
- Even though the employee morale rose somewhat with the excitement of the new building, there are still a few disgruntled employees in the company.

# Term Project 3: Return to Kronos, Part II

- The new office is built to the highest energy efficiency standard, but as with any new building, there are still several HVAC (heating, ventilation, and air conditioning) issues to work out. The building is divided into several HVAC zones. Each zone is instrumented with sensors that report building temperatures, heating and cooling system status values, and concentration levels of various chemicals such as carbon dioxide ($CO^2$) and hazium (abbreviated Haz), a recently discovered and possibly dangerous chemical.

# Term Project 3: Return to Kronos, Part III

- CEO Sten Sanjorge Jr. has read about Haz and requested that these sensors be included. However, they are new and very expensive, so GAStech can afford only a small number of sensors.
- Also, with their move into the new building, GAStech also introduced new security procedures, which staff members are not necessarily adopting consistently. Staff members are now required to wear proximity (prox) cards while in the building.

# Term Project 3: Return to Kronos, Part IV

- The building is instrumented with passive prox card readers that cover individual building zones. The prox card zones do not generally correspond with the HVAC zones. When a prox card passes into a new zone, it is detected and recorded. Most, but not all, areas are still open to staff members even if they forget their prox cards.
- The builders included a free robotic mail delivery system. This robot, nicknamed Rosie, travels the halls periodically, moving between floors in a specially designed chute.

# Term Project 3: Return to Kronos, Part V

- Rosie is equipped with a mobile prox sensor, which identifies the prox cards in the areas she travels through. Data from Rosie, combined with other data, allowed contestants to hypothesize an insider threat employee, as well as how and when inappropriate activities were occurring.
- As an expert in data analytics, you have been hired to help GAStech understand its operations data. In this task, you are given two weeks of building and prox sensor data. Can you identify typical patterns and issues of concern?

# Term Project 3: Return to Kronos, Part VI

**Data**

- The following data was provided for this task.
  1. A building layout for the GAStech offices, including the maps of the prox zones and the HVAC zones
  2. A current list of employees, roles, and office assignments
  3. A description of the data formats and fields provided
  4. Proximity sensor data for each of the prox zone regions

# Term Project 3: Return to Kronos, Part VII

**Data (cont.)**

- The following data was provided for this task.
  5. Proximity sensor data from Rosie the mobile robot
  6. HVAC sensor readings and status information from each of the building's HVAC zones
  7. Hazium readings from four sensors (Haz has become a recent concern on the island of Kronos; not much is known about its effects, but it is suspected that Haz is not good for people)

# Term Project 3: Return to Kronos, Part VIII

**Goals—**answer the following questions:

1. What are the typical patterns in the prox card data? What does a typical day look like for GAStech employees?
2. Describe up to ten of the most interesting patterns that appear in the building data. Describe what is notable about the pattern and explain its possible significance.

# Term Project 3: Return to Kronos, Part IX

**Goals—**answer the following questions: (cont.)

3. Describe up to ten notable anomalies or unusual events you see in the data. Prioritize those issues that are most likely to represent a danger or a serious issue for building operations.
4. Describe up to five observed relationships between the proximity card data and building data elements. If you find a causal relationship (for example, a building event or condition leading to personnel behavior changes or personnel activity leading to building operations changes), describe your discovered cause and effect, the evidence you found to support it, and your level of confidence in your assessment of the relationship.

# The End

# Mayhem at DinoFun World Scenario

Claudio Delrieux

# Mayhem at DinoFun World Scenario

- DinoFun World is a typical modest-sized amusement park, sitting on about 215 hectares and hosting thousands of visitors each day. It has a small-town feel, but it is well known for its exciting rides and events.
- One event last year was a weekend tribute to Scott Jones, internationally renowned football (“soccer,” in U.S. terminology) star. Scott Jones is from a town nearby DinoFun World. He was a classic hometown hero, with thousands of fans who cheered his success as if he were a beloved family member.

# Mayhem at DinoFun World Scenario (cont.)

- To celebrate his years of stardom in international play, DinoFun World declared “Scott Jones Weekend,” where Scott was scheduled to appear in two stage shows each on Friday, Saturday, and Sunday to talk about his life and career. In addition, a show of memorabilia related to his illustrious career would be displayed in the park’s pavilion.
- However, the event did not go as planned. Scott’s weekend was marred by crime and mayhem perpetrated by a poor, misguided, and disgruntled figure from Scott’s past.

# The End

# Term Project 4

## Mayhem at DinoFun World: Visitor Communication

Claudio Delrieux

# Term Project 4: Mayhem at DinoFun World (Visitor Communication), Part I

- While the crimes were rapidly solved, park officials and law enforcement figures are interested in understanding just what happened during that weekend to better prepare themselves for future events.
- They are interested in understanding how people move and communicate in the park, as well as how patterns changes and evolve over time, and what can be understood about motivations for changing patterns.

# Term Project 4: Mayhem at DinoFun World (Visitor Communication), Part II

**Data**

- The data for this task consists of three days' worth of communications from Friday through Sunday. This includes communications between the paying park visitors, as well as communications between the visitors and park services. In addition, the data also contains records indicating if and when the user sent a text to an external party.
- The data fields are timestamp, the originator's ID, the recipient's ID, and the park area from which the message was sent.

# Term Project 4: Mayhem at DinoFun World (Visitor Communication), Part III

- As can be seen in the accompanying map, the park is broken up into five themed areas: the Entry Corridor, Kiddie Land, Tundra Land, Wet Land, and Coaster Alley. So, while these locations were not precise, they indicate general geo-coordinate information for the analyses.
- There is also some auxiliary information that provides extra context (i.e., map, park website, news article). Use analytics to explore and analyze the available data and develop responses to the questions below.

# Term Project 4: Mayhem at DinoFun World (Visitor Communication), Part IV

**Goals**

- You are asked to characterize dominant communication IDs, interesting communication patterns, and suspicious patterns that could contribute to the analysis of the crime. Specifically:
  1. Identify groups that stand out for their large volumes of communication.
     I. Characterize the communication patterns you see.
     II. Based on these patterns, what do you hypothesize about these IDs?

# Term Project 4: Mayhem at DinoFun World (Visitor Communication), Part V

**Goals (cont.)**

2. Describe up to 10 communications patterns in the data. Characterize who is communicating, with whom, when, and where. If you have more than 10 patterns to report, please prioritize those patterns that are most likely to relate to the crime.
3. From this data, can you hypothesize when the vandalism was discovered? Describe your rationale.

# The End

# Term Project 5

## Mayhem at DinoFun World: Visitor Movement

Claudio Delrieux

# Term Project 5: Mayhem at DinoFun World (Visitor Movement), Part I

- While the crimes were rapidly solved, park officials and law enforcement figures are interested in understanding just what happened during that weekend to better prepare themselves for future events.
- They are interested in understanding how people move and communicate in the park, as well as how patterns change and evolve over time, and what can be understood about motivations for changing patterns.

# Term Project 5: Mayhem at DinoFun World (Visitor Movement), Part II

**Data**

- This task is focused on movement of people around the park. You have access to movement tracking information for all paying park visitors over the three days of the celebration. The datasets are .csv files for Friday, Saturday, and Sunday, containing a date-time stamp, a visitor ID, a tag as to whether the record referred to a movement within the park grid or a “check-in” to an amusement park ride, and a grid location (x,y coordinates).

# Term Project 5: Mayhem at DinoFun World (Visitor Movement), Part III

**Data (cont.)**

- There is also some auxiliary information that provides extra context (i.e., map, park website, news article).
- This data contains many patterns that are useful for planning park operations. On this particular weekend, a crime occurred and the data likely contains information pertinent to that crime. Analyze the available data and develop responses to the questions below.

# Term Project 5: Mayhem at DinoFun World (Visitor Movement), Part IV

**Goals**

- You are asked to characterize the movement of *"groups"* and individuals, with a special emphasis on what might be relevant to better understand the incident.

1. Characterize the attendance at the park on this weekend. Describe up to 12 different types of groups at the park on this weekend.
   - How big is the group type?
   - Where does this type of group like to go in the park?

# Term Project 5: Mayhem at DinoFun World (Visitor Movement), Part V

**Goals (cont.)**

- How common is this type of group?
- What are your other observations about this type of group?
- What can you infer about the group?
- If you were to make one improvement to the park to better meet this group's needs, what would it be?

2. Are there notable differences in the patterns of activity in the park across the three days? Describe the notable differences you found.

# Term Project 5: Mayhem at DinoFun World (Visitor Movement), Part VI

**Goals (cont.)**

3. What anomalies or unusual patterns do you see? Describe no more than 10 anomalies, and prioritize those unusual patterns that you think are most likely to be relevant to the crime.

- The definition of *group* is intentionally left to you to determine, so you can best formulate it within the context of your working hypotheses and evidence.

# The End

# Mystery at the Wildlife Preserve Scenario

Claudio Delrieux

# Mystery at the Wildlife Preserve Scenario

- Mistford is a mid-size city is located to the southwest of a large nature preserve. The city has a small industrial area with four light-manufacturing endeavors.
- Mitch Vogel is a post-doc student studying ornithology at Mistford College and has been discovering signs that the number of nesting pairs of the rose-crested blue pipit, a popular local bird due to its attractive plumage and pleasant songs, is decreasing!

# Mystery at the Wildlife Preserve Scenario (cont.)

- The decrease is sufficiently significant that the Pangera Ornithology Conservation Society is sponsoring Mitch to undertake additional studies to identify the possible reasons.
- Mitch is gaining access to several datasets that may help him in his work, and he has asked you to help him analyze these datasets.

# The End

# Term Project 6

## Mystery at the Wildlife Preserve: Vehicle Patterns of Life

Claudio Delrieux

# Term Project 6: Mystery at the Wildlife Preserve (Vehicle Patterns of Life), Part I

- As part of his investigation, Mitch needs to examine the movement of traffic through the Boonsong Lekagul Nature Preserve.
- His first working hypothesis is that there is some link between the traffic going through the preserve and the decline in the nesting: maybe the traffic noises are drowning out mating calls! Or perhaps some odd goings-on in the traffic patterns—perhaps campers are invading the bird's habitat areas.

# Term Project 6: Mystery at the Wildlife Preserve (Vehicle Patterns of Life), Part II

- There are park rangers working as caretakers of the nature preserve, and they have been collecting traffic data for their annual reporting to the local government. They have provided Mitch with some data, explanations about the data, and a map.

# Term Project 6: Mystery at the Wildlife Preserve (Vehicle Patterns of Life), Part III

**Data**

- You are provided with a description of how traffic through the preserve occurs and how traffic was measured through the sensors. You are also given some background information about the preserve and bitmapped files describing the gridded map against which the data is provided. The data contains a timestamp of when the vehicle passed a sensor location, a car-ID, a car type (as described in the background information), and a sensor identification.

# Term Project 6: Mystery at the Wildlife Preserve (Vehicle Patterns of Life), Part IV

**Goals**

1. “Patterns of life” analyses depend on recognizing repeating patterns of activities by individuals or groups. Describe up to six daily patterns of life by vehicles traveling through and within the park. Characterize the patterns by describing the kinds of vehicles participating, their spatial activities (where do they go?), their temporal activities (when does the pattern happen?), and provide a hypothesis of what the pattern represents.

# Term Project 6: Mystery at the Wildlife Preserve (Vehicle Patterns of Life), Part V

**Goals (cont.)**

2. Patterns of life analyses may also depend on understanding what patterns appear over longer periods of time (in this case, over multiple days). Describe up to six patterns of life that occur over multiple days (including across the entire data set) by vehicles traveling through and within the park. Characterize the patterns by describing the kinds of vehicles participating, their spatial activities (where do they go?), their temporal activities (when does the pattern happen?), and provide a hypothesis of what the pattern represents.

# Term Project 6: Mystery at the Wildlife Preserve (Vehicle Patterns of Life), Part VI

**Goals (cont.)**

3. Unusual patterns may be patterns of activity that change from an established pattern, or are just difficult to explain from what you know of a situation.
4. Describe up to six unusual patterns (either single day or multiple days) and highlight why you find them unusual. What are the top three patterns you discovered that you suspect could be most impactful to bird life in the nature preserve?

# The End

# Term Project 7

## Mystery at the Wildlife Preserve: Plume Analysis

Claudio Delrieux

# Term Project 7: Mystery at the Wildlife Preserve (Plume Analysis), Part I

- Mitch's hypothesis is that environmental contamination may be the cause of the problem.
- The primary job for Mitch is to determine which (if any) of the factories may be contributing to the problems of the rose-crested blue pipit. Often, air sampling analysis deals with a single chemical being emitted by a single factory. In this case, though, there are four factories, potentially each emitting four chemicals, being monitored by nine different sensors.

# Term Project 7: Mystery at the Wildlife Preserve (Plume Analysis), Part II

- Further, some chemicals being emitted are more hazardous than others. These factories were supposed to be compliant with recent years' environmental regulations, but Mitch had his doubts that the actual data has been closely reviewed. Substances of concern include *appluimonia, chlorodinine,* and *AGOC-3A*.
- Your task, as supported by data analytics that you apply, is to disentangle the data to help Mitch determine where problems may be.

# Term Project 7: Mystery at the Wildlife Preserve (Plume Analysis), Part III

**Data**

- Mitch discovered that the state government has been monitoring the gaseous effluents from the factories through a set of sensors, distributed around the factories, and set between the smokestacks, the city of Mistford, and the nature preserve. The state gave Mitch access to their air sampling data, meteorological data, and locations map. Data includes meteorological data that provided wind speed and direction for certain periods of time and sensor data for three months' worth of readings with the chemical detected, the sensor ID, the reading in parts per million, and the date/time.

# Term Project 7: Mystery at the Wildlife Preserve (Plume Analysis), Part IV

**Goals**

1. Characterize the sensors' performance and operation. Are they all working properly at all times? Can you detect any unexpected behaviors of the sensors through analyzing the readings they capture?
2. Now turn your attention to the chemicals themselves. Which chemicals are being detected by the sensor group? What patterns of chemical releases do you see, as being reported in the data?
3. Which factories are responsible for which chemical releases? Carefully describe how you determined this using all the data you have available. For the factories you identified, describe any observed patterns of operation revealed in the data.

# The End

# Term Project 8

## Mystery at the Wildlife Preserve: Multispectral Imagery

Claudio Delrieux

# Term Project 8: Mystery at the Wildlife Preserve (Multispectral Imagery), Part I

- As part of a visual exploration, Mitch is able to get around some of the nature preserve to study the bird, but because of the terrain and the vegetation, he is not able to cover the entire preserve with a car or on foot.
- Drones and air vehicles would scare the birds, making it difficult to assess the extent of the possible problem.

# Term Project 8: Mystery at the Wildlife Preserve (Multispectral Imagery), Part II

- While he puzzles out how to get a more complete picture, he talked with one of his professors at Mistford College, who suggested he look at satellite imagery over the area over the past few years. Perhaps there have been changes in the flora that are related to issues with fauna. Mitch thought this kind of associated study may be informative, so the professor provided him with some images of the preserve collected by the National Space Service.

# Term Project 8: Mystery at the Wildlife Preserve (Multispectral Imagery), Part III

**Data**

- A challenging component of this image dataset is that it is multispectral, that is, there are three other wavelength channels included beyond red, green, and blue. As mentioned, we provide a multispectral analysis primer to assist you with the analysis. We are specifically looking for you to go beyond simple displays of the image data that employ interesting data and visual analytics to support investigation into the changes in the area over time.

# Term Project 8: Mystery at the Wildlife Preserve (Multispectral Imagery), Part IV

**Goals**

1. Boonsong Lake resides within the preserve and has a length of about 3,000 feet (see the Boonsong Lake image file). The image of Boonsong Lake is oriented north-south and is an RGB image (not six channels as in the supplied satellite data). Using the Boonsong Lake image as your guide, analyze and report on the scale and orientation of the supplied satellite images. How much area is covered by a pixel in these images?

# Term Project 8: Mystery at the Wildlife Preserve (Multispectral Imagery), Part V

**Goals (cont.)**

2. Identify features you can discern in the preserve area as captured in the imagery. Focus on image features that you are reasonably confident that you can identify (e.g., a town full of houses may be identified with a high confidence level).
3. There are most likely many features in the images that you cannot identify without additional information about the geography, human activity, and so on. Mitch is interested in changes that are occurring that may provide him with clues to the problems with the pipit bird. Identify features that change over time in these images, using all channels of the images.

# The End

# Suspense at the Wildlife Preserve Scenario

Claudio Delrieux

# Suspense at the Wildlife Preserve Scenario

Mistford is a mid-size city located to the southwest of the Boonsong Lekagul Wildlife Preserve. The city has a small industrial area with four light-manufacturing endeavors. Mistford and the wildlife preserve are struggling with the possible endangerment of the rose-crested blue pipit, a locally loved bird. The bird's nesting pairs seem to have decreased alarmingly, prompting an investigation last year implicating a Mistford manufacturing firm.

# Suspense at the Wildlife Preserve Scenario (cont.)

Since the initial investigation, the situation has evolved: The firm insists that they have done nothing wrong! They assert that grad student Mitch Vogel and his professors are mere media-seekers trying to draw attention away from their lackadaisical research. The firm presents itself as an extremely eco-friendly organization. They have launched their own very public investigation into the issues raised last year and are reporting very different results! It's time to apply your data analytics expertise to help illuminate the path to good science.

# The End

# Term Project 9

## Suspense at the Wildlife Preserve: “Cheep” Shots

Claudio Delrieux

# Term Project 9: Suspense at the Wildlife Preserve ("Cheep" Shots), Part I

- The suspect firm may have been a primary contributor to the apparent reduction of the number of nesting pairs of the rose-crested blue pipit, a favorite bird of Mistford residents and Boonsong Lekagul Nature Preserve visitors. They supposedly used a banned substance in their manufacturing process. They surreptitiously dumped process waste in the northeast region of the preserve and the substance also was detected in their smokestack emissions.

# Term Project 9: Suspense at the Wildlife Preserve ("Cheep" Shots), Part II

- The firm now claims that the analysis was flawed and biased. To combat these conclusions, they launched their own "investigation" into the pipit situation, and they are now reporting that there are plenty of rose-crested blue pipits happily living and nesting in the preserve. To back up this claim, they have provided a set of pipit bird calls, recently recorded across the preserve, with locations of where they were recorded. Clearly, they claim, the pipits are a thriving population, and will provide even more supporting evidence as their investigation proceeds.

# Term Project 9: Suspense at the Wildlife Preserve ("Cheep" Shots), Part III

- The Pangera Ornithology Conservation Society is at their wit's end at what to do about this turn of events. The townsfolk and preserve rangers seem satisfied that the recordings back up the firm's claims. Mistford College does not have a pipit expert they can call upon for help. But they do have a collection of bird calls from the preserve that has been vetted by various ornithology groups as having accurate identifications. They have heard that new techniques from data science can be applied to situations like this.

# Term Project 9: Suspense at the Wildlife Preserve ("Cheep" Shots), Part IV

**Data**

- In addition to the recordings provided by the firm, you are provided with a collection of bird calls from the preserve that have been vetted by various ornithology groups as having accurate identifications.

# Term Project 9: Suspense at the Wildlife Preserve ("Cheep" Shots), Part V

**Goals**

1. Using the bird call collection and the included map of the wildlife preserve, characterize the patterns of all of the bird species in the preserve over the time of the collection. Please assume we have a reasonable distribution of sensors and human collectors providing the recordings, so that the patterns are reasonably representative of the bird locations across the area. Do you detect any trends or anomalies in the patterns?

# Term Project 9: Suspense at the Wildlife Preserve ("Cheep" Shots), Part VI

**Goals (cont.)**

2. Turn your attention to the set of bird calls supplied by the firm. Does this set support the claim of pipits being found across the preserve? A machine learning approach using the bird call library may help your investigation. What is the role of visualization?
3. Formulate a hypotheses concerning the state of the rose-crested blue pipit. What are your primary pieces of evidence to support your assertion? What next steps should be taken in the investigation to either support or refute the firm's claim that the pipits are actually thriving?

# The End

# Term Project 10

## Suspense at the Wildlife Preserve: Like a Duck to Water

Claudio Delrieux

# Term Project 10: Suspense at the Wildlife Preserve (Like a Duck to Water), Part I

The suspect firm was implicated in environmental damage to the Boonsong Lekagul Wildlife Preserve for both dumping toxic waste and polluting the air with chemicals from its manufacturing process. But they deny any accusation of industrial waste dumping and state that there isn’t any ground contamination near any remote ranger station, and they have inspected that area and found it as pristine as the rest of the preserve.

# Term Project 10: Suspense at the Wildlife Preserve (Like a Duck to Water), Part II

Outraged ornithology professors from Mistford College journeyed out to look over the dumping site themselves and perform soil analyses. They returned to report that the site looked like there had been recent excavation and building activities going on. Boonsong Preserve rangers later confirmed that a new ranger station was being built at that site! Soil samples taken from the site were inconclusive in detecting toxic manufacturing chemicals, as new topsoil had been trucked in.

# Term Project 10: Suspense at the Wildlife Preserve (Like a Duck to Water), Part III

With a primary piece of evidence now gone, investigators will need to take another approach. Professors in the Mistford College Hydrology Department have come forward with several years of water sensor readings from rivers and streams in the preserve. These samples were taken from different locations scattered throughout the area and contain measurements of several chemicals of possible interest. Could data analytics help reveal something in this data that could make up for the soil evidence that was destroyed?

# Term Project 10: Suspense at the Wildlife Preserve (Like a Duck to Water), Part IV

**Data**

- You are given a map of the preserve, with named sampling sites indicated on the map (the names have local significance, but are just mnemonics for your study). You are also provided with readings from each sampling station over time for several different chemicals and water properties.

# Term Project 10: Suspense at the Wildlife Preserve (Like a Duck to Water), Part V

**Goals**

1. Characterize the past and most recent situation with respect to chemical contamination in the Boonsong Lekagul waterways. Do you see any trends of possible interest in this investigation?
2. What anomalies do you find in the waterway samples dataset? How do these affect your analysis of potential problems to the environment?

# Term Project 10: Suspense at the Wildlife Preserve (Like a Duck to Water), Part VI

**Goals (cont.)**

3. Is the Hydrology Department collecting sufficient data to understand the comprehensive situation across the preserve? What changes would you propose to make in the sampling approach to best understand the situation?
4. After reviewing the data, do any of your findings cause particular concern for the pipit or other wildlife? Would you suggest any changes in the sampling strategy to better understand the waterways situation in the preserve?

# The End

# Term Project 11

## Suspense at the Wildlife Preserve: “Organized” Crime?

Claudio Delrieux

# Term Project 11: Suspense at the Wildlife Preserve ("Organized" Crime?), Part I

- Mitch Vogel left his work at Mistford College, but has not forgotten the rose-crested blue pipit. Soon after arriving in the small town of Sulev in Northern Europe, Mitch started to see the telltale signs of contamination damage at the nearby Panteleimon Aviary Sanctuary. Mitch hears from local bird watchers that populations of the greater Eurasian red-throated pipit have been affected. Mitch receives an anonymous letter from an insider in a local branch of the suspect firm who's willing to help. A fellow pipit lover has gathered up a variety of data and identified a suspicious group within the firm.

# Term Project 11: Suspense at the Wildlife Preserve ("Organized" Crime?), Part II

- Attached to the letter Mitch receives is a USB drive with phone, email, meeting, and procurement records for the firm over the past two and a half years. Mitch wonders if the fate of the Eurasian pipit lies somewhere in that data. Mitch intends to put this data together to see if the problems with the firm are much larger than he initially suspected.
- Your task, as supported by data analytics that you apply, is to help Mitch determine the organizational structure of the group within the firm that was referenced by the insider providing Mitch the data. How is it internally connected and is anyone else involved?

# Term Project 11: Suspense at the Wildlife Preserve ("Organized" Crime?), Part III

**Data**

- You will find phone, email, meeting, and procurement records for the firm over the past two and a half years. The data includes the source of each transaction, the recipient (destination), and the time of the transaction, but the contents of emails or phone calls are not available. There is also a company index that shows the name of all 642,631 individuals in the company and their associated ID numbers. Many of these transactions have been already flagged as suspicious.

# Term Project 11: Suspense at the Wildlife Preserve ("Organized" Crime?), Part IV

**Goals**

1. Using the four datasets, combine the different sources to create a single picture of the company. Characterize changes in the company over time. According to the company's communications and purchase habits, is the company growing?
2. Combine the data sources for the group that the insider has identified as being suspicious and locate the group in the larger dataset. Determine if anyone else appears to be closely associated with this group. Highlight which employees are making suspicious purchases.

# Term Project 11: Suspense at the Wildlife Preserve ("Organized" Crime?), Part V

**Goals (cont.)**

3. Using the combined group of suspected bad actors you created in question two, show the interactions within the group over time.
   - Characterize the group's organizational structure and show a full picture of communications within the group.
   - Does the group composition change during the course of their activities?
   - How do the group's interactions change over time?

# Term Project 11: Suspense at the Wildlife Preserve ("Organized" Crime?), Part VI

**Goals (cont.)**

4. A list of purchases might indicate illicit activity elsewhere in the company. Using the structure of the first group as a model, can you find any other instances of suspicious activities in the company? Are there other groups that have structure and activity similar to this one? Who are they? Each of the suspicious purchases could be a starting point for your search. Provide examples of up to two other groups you find that appear suspicious and compare their structure with the structure of the first group.

# The End

# Disaster at St. Himark! Scenario

Claudio Delrieux

# Disaster at St. Himark! Scenario

St. Himark is a vibrant community located in the Oceanus Sea. Home to the world-renowned St. Himark Museum, beautiful beaches, and the Wilson Forest Nature Preserve, St. Himark is one of the region's best cities for raising a family and provides employment across a number of industries including the Always Safe Nuclear Power Plant. All that was true before the disastrous earthquake that hit the area during the course of this year.

# Disaster at St. Himark! Scenario (cont.)

Major Jordan, city officials, and emergency services are overwhelmed and desperate for assistance in understanding the true situation on the ground and how best to deploy the limited resources available to this relatively small community. Officials are scrambling to determine the extent of the damage and dispatch limited resources to the areas in most need. They quickly receive seismic readings and use those for an initial deployment, but realize they need more information to make sure they have a realistic understanding of the true conditions throughout the city.

# The End

# Term Project 12

## Disaster at St. Himark!
## Crowdsourcing for Situational Awareness

Claudio Delrieux

# Term Project 12: Disaster at St. Himark!
# Crowdsourcing for Situational Awareness, Part I

- Prior to the earthquake, the city had released a damage reporting mobile application shortly before the earthquake. This app allows citizens to provide more timely information to the city to help them understand damage and prioritize their response.
- With emergency services stretched thin, officials are relying on citizens to provide them with much needed information about the effects of the quake to help focus recovery efforts.

# Term Project 12: Disaster at St. Himark!
# Crowdsourcing for Situational Awareness, Part II

- Use app responses in conjunction with shake maps of the earthquake strength to identify areas of concern and advise emergency planners (note: the ShakeMaps are from April 6 and April 8, respectively).
- By combining seismic readings of the quake, responses from the app, and background knowledge of the city, help the city triage their efforts for rescue and recovery.

# Term Project 12: Disaster at St. Himark!
# Crowdsourcing for Situational Awareness, Part III

**Data**

- The data includes one (CSV) file spanning the entire length of the event, containing (categorical) individual reports of shaking/damage by neighborhood over time. Reports are made by citizens at any time, however, they are only recorded in five-minute batches/increments due to the server configuration. Furthermore, delays in the receipt of reports may occur during power outages. Also included are two ShakeMap (PNG) files which indicate where the corresponding earthquake' epicenters originated, as well as how much shaking can be felt across the city.

# Term Project 12: Disaster at St. Himark!
# Crowdsourcing for Situational Awareness, Part IV

**Goals**

1. Emergency responders will base their initial response on the earthquake ShakeMap. Use data analytics to determine how their response should change based on damage reports from citizens on the ground. How would you prioritize neighborhoods for response? Which parts of the city are hardest hit?

# Term Project 12: Disaster at St. Himark!
# Crowdsourcing for Situational Awareness, Part V

**Goals (cont.)**

2. Use data and visual analytics to show uncertainty in the data. Compare the reliability of neighborhood reports. Which neighborhoods are providing reliable reports? Provide a rationale for your response.
3. How do conditions change over time? How does uncertainty in change over time?

# The End

# Term Project 13

## Disaster at St. Himark! Citizen Science to the Rescue

Claudio Delrieux

# Term Project 13: Disaster at St. Himark! Citizen Science to the Rescue, Part I

- One of St. Himark's largest employers is the Always Safe Nuclear Power Plant. The pride of the city, it produces power for St. Himark's needs and exports the excess to the mainland, providing a steady revenue stream.
- However, the plant was not compliant with international standards when it was constructed and is now aging. As part of its outreach to the broader community, Always Safe agreed to provide funding for a set of carefully calibrated professional radiation monitors at fixed locations throughout the city.

# Term Project 13: Disaster at St. Himark! Citizen Science to the Rescue, Part II

- Additionally, a group of citizen scientists led by the members of the Himark Science Society started an education initiative to build and deploy lower cost homemade sensors, which people can attach to their cars.
- The sensors upload data to the web by connecting through the user's cell phone. The goal of the project was to engage the community and demonstrate that the nuclear plant's operations were not significantly changing the region's natural background levels of radiation.

# Term Project 13: Disaster at St. Himark! Citizen Science to the Rescue, Part III

- When the earthquake struck St. Himark, the nuclear power plant suffered damage resulting in a leak of radioactive contamination. Further, a coolant leak sprayed employees' cars and contaminated them at varying levels. Now, the city's emergency management officials are trying to understand if there is a risk to the public while also responding to other emerging crises related to the earthquake, as well as satisfying the public's concern over radiation (note: reviewing the city description document may be helpful to understanding the landscape and character of the city).

# Term Project 13: Disaster at St. Himark! Citizen Science to the Rescue, Part IV

**Data**

- You will find two data files spanning the entire length of the events (12 a.m. on April 6, 2020 to 11:59 p.m. on April 10, 2020), containing radiation measurements from mobile and static radiation sensors.
- Also provided are a set of supporting files. Be prepared for missing and corrupted data, skipped timesteps, and other issues. Both radiation measurements and movements may be affected by conditions in the city.

# Term Project 13: Disaster at St. Himark! Citizen Science to the Rescue, Part V

**Goals**

- Help St. Himark's emergency management team combine data from the government-operated stationary monitors with data from citizen-operated mobile sensors to help them better understand conditions in the city and identify likely locations that will require monitoring, cleanup, or even evacuation. Will data from citizen scientists clarify the situation or make it more uncertain?

1. Visualize radiation measurements over time from both static and mobile sensors to identify areas where radiation over background is detected. Characterize changes over time.

# Term Project 13: Disaster at St. Himark! Citizen Science to the Rescue, Part VI

**Goals (cont.)**

2. Use data analytics to represent and analyze uncertainty in the measurement of radiation across the city.
   - Compare uncertainty of the static sensors to the mobile sensors. What anomalies can you see? Are there sensors that are too uncertain to trust?
   - Which regions of the city have greater uncertainty of radiation measurement? Use data analytics to explain your rationale.
   - What effects do you see in the sensor readings after the earthquake and other major events? What effect do these events have on uncertainty?

# Term Project 13: Disaster at St. Himark! Citizen Science to the Rescue, Part VII

**Goals (cont.)**

3. Given the uncertainty you observed, are the radiation measurements reliable to locate areas of concern?
   - Highlight potential locations of contamination, including the locations of contaminated cars. Should St. Himark officials be worried about contaminated cars moving around the city?
   - Estimate how many cars may have been contaminated when coolant leaked from the Always Safe plant. Use data analysis of measurements to determine if any have left the area.
   - Indicate where you would deploy more sensors to improve monitoring in the city. Would you recommend more static sensors, more mobile sensors, or both? Use your analysis of uncertainty to justify your recommendation.

# Term Project 13: Disaster at St. Himark! Citizen Science to the Rescue, Part VIII

**Goals (cont.)**

4. Summarize the state of radiation measurements at the end of the available period. Use your data analysis approaches to suggest a course of action for the city. Use analytics to compare the static sensor network to the mobile sensor network. What are the strengths and weaknesses of each approach? How do they support each other?

# The End

# Term Project 14

## Disaster at St. Himark! Voice from the People

Claudio Delrieux

# Term Project 14: Disaster at St. Himark! Voice from the People, Part I

- Seismic and survey data are useful for capturing the objective damage that the earthquake has caused St. Himark. However, this data has limitations.
- First, official surveys are time-consuming and do not stay current in a rapidly changing situation.
- Second, they don’t establish how citizens are reacting to the current crisis.
- Third, they are often insufficiently granular, providing little insight into differences between neighborhoods and emergency resources.

# Term Project 14: Disaster at St. Himark! Voice from the People, Part II

- In other words, the seismic and survey data do not provide an up-to-date view of the structural and humanitarian impact caused by the earthquake on a neighborhood-by-neighborhood basis.
- The city has concluded that this knowledge is necessary to determine where to allocate emergency resources.

# Term Project 14: Disaster at St. Himark! Voice from the People, Part III

- City officials have identified a subset of Y*INT, a community-based social media platform, as a potential source for revealing the current state of St. Himark's neighborhoods and people.
- Knowing that you are skilled in data analytics, the city has asked you to analyze Y*INT messages in order to determine the appropriate actions it must take in order to assist the community in this disaster.

# Term Project 14: Disaster at St. Himark! Voice from the People, Part IV

**Data**

- Data contains one CSV file, spanning from 04/06/2020 to 04/12/2020, with the following fields.
  - Time (date/time the message was posted)
  - Location (St. Himark neighborhood message was posted from)
  - Account (user handle of the person who posted the message)
  - Message (text of the message itself)
- Be prepared to discern between reliable and unreliable messages.

# Term Project 14: Disaster at St. Himark! Voice from the People, Part V

**Goals**

- The city has been using Y*INT to communicate with its citizens, even post-earthquake. However, city officials need additional information to determine the best way to allocate emergency resources across all neighborhoods of St. Himark. Your task is to determine the types of problems that are occurring across St. Himark. Then, advise the city on how to prioritize the distribution of resources. Keep in mind that not all sources on Y*INT are reliable, and that priorities may change over time as the state of neighborhoods also changes.

# Term Project 14: Disaster at St. Himark! Voice from the People, Part VI

**Goals (cont.)**

1. Using data analytics, characterize conditions across the city and recommend how resources should be allocated at 5 hours and 30 hours after the earthquake. Include evidence from the data to support these recommendations. Consider how to allocate resources such as road crews, sewer repair crews, power, and rescue teams.

# Term Project 14: Disaster at St. Himark! Voice from the People, Part VII

**Goals (cont.)**

2. Identify at least three times when conditions change in a way that warrants a reallocation of city resources. What were the conditions before and after the inflection point? What locations were affected? Which resources are involved?
3. Take the pulse of the community. How has the earthquake affected life in St. Himark? What is the community experiencing outside the realm of the first two questions? Show decision makers summary information and relevant/characteristic examples.

# The End

# Final Considerations

Claudio Delrieux

# Final Considerations, Part I

- During the next weeks, you will be working on the term project of your choice among the ones presented here.
- In the final week you will present your term projects according to guidelines you will find in the LMS.

# Final Considerations, Part II

- Like most real-world datasets, the data you are provided is imperfect, with possible issues such as missing data, conflicting data, data of varying resolutions, outliers, or other kinds of confusing data.
- Therefore, previous data wrangling for data curation, and also exploratory analysis, will likely be beneficial for an adequate task completion.

# Final Considerations, Part III

- Please prepare a single Python notebook with the answers. However, feel free to use more than a single Python file to solve tasks. If necessary, some results may be pre-computed and stored in separate files.
- Notebooks must clearly explain the answers to each question, supported by evidence derived from the data analysis, and visualized using the tools presented in class.

Final Considerations

# The End