Course Project for
Managing Machine Learning Projects

Early Detect, Alert and Locate Colony Collapse Disorder in Commercial Bee Hives

Ruth Shacterman
Sep 15, 2023

Opportunity Eval.

The Product
The Problem
The Task Analysis
AI/ML Opportunities

Table of content

TOC

What is Colony Collapse Disorder?

Healthy Hive

Colony Collapse Disorder (CCD)

Why is CCD a Problem worth solving?

- CCD is a significant concern for beekeepers and agricultural communities because honeybees play a crucial role in pollination. In the US, honeybee pollination accounts for over $15 billion in agricultural products each year.

- Early detection and monitoring can provide valuable data on the progression of CCD. This information can help Beekeepers with strategies to prevent and/or better manage the CCD.

Why is Early Detection Important

CCD may cause a rapid and complete collapse of a bee colony. Early detection allows beekeepers to prevent this loss.

The loss of colonies can have severe impact for crop pollination, leading to reduced crops yields and economic losses for farmers.

CCD is a complex phenomenon with many contributing environmental stressors. Early detection and removal of a stressor can prevent their spread to neighboring bee colonies

Early detection and monitoring can provide valuable data on the progression of CCD. This information can help Beekeepers with strategies to prevent and/or better manage the CCD.

How is CCD currently addressed?

CCD Detection:
Regular and frequent inspections of each hive:

- Visually monitoring bee activity
- If activity seems low, opening the hive to examine the combs and their content
- Assess and treat based on experience and knowlege

The cost
Regular hive inspections are resource-intensive and costly.

Scale and frequency
There could be thousands of hives. Inspecting each requires time and labor.

High level of experience
Assessing and “reading” a hive requires an experienced beekeeper

Opportunity Evaluation - Why ML?

Automation & Augmentation

- Eliminates repetitive tedious work of inspecting each hive manually.
- Reduced cost of labor.
- Maximize user of Expert Beekeeping skills (no longer wasted on repetitive inspections).

Prediction

- Given sufficient data and right features, ML should be able to predict CCD and flag early cases.

Trends and Insights

- Accumulated data pattern can shed light on the state and trends in the Beekeeping business.
- Data trends may add to the understanding about the specific location and context in which the monitored hives exist

A Commercial Beekeeping Business

The Product:

This product uses an ML- driven mobile app, and sensor data to

Detect and alert Commercial Beekeepers to the onset of CCD

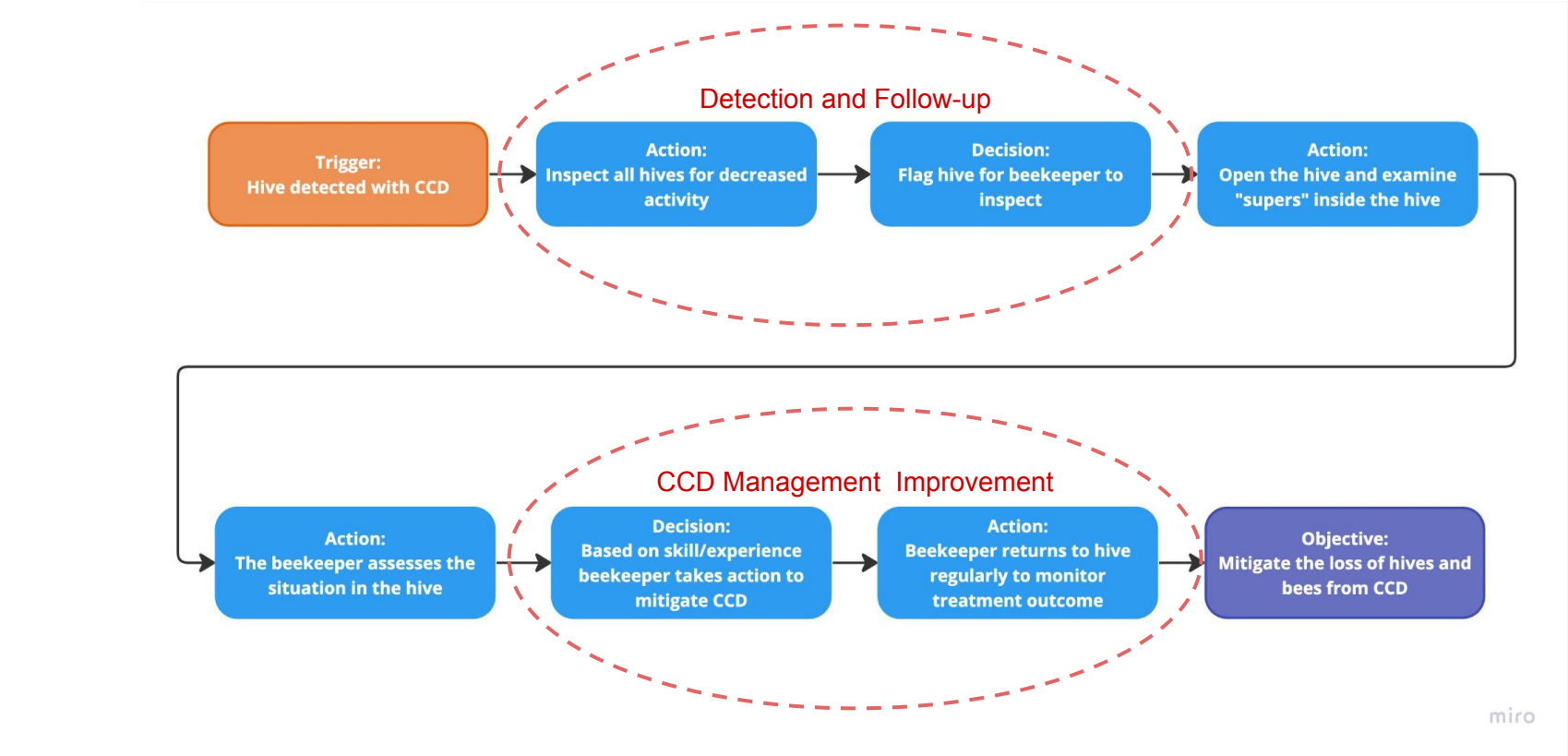
Assist the beekeepers with CCD management

The Goal:

The product should save costs on repetitive labor and maximize the use of beekeeping time by:

- Accurately detecting of CCD
- Providing help with CCD outbreaks
- Improving hive management strategies

Task Analysis and Insights



Task Analysis - AI opportunity summary

The idea is to augment the Beekeeping workflow by automating tasks which require pattern detection from large data:

- Early detect of CCD outbreaks through prediction - **using a binary classification tree model and sensor input**
- Assist the beekeeper with assessments and strategies to mitigate outbreaks **most likely using a regression model for suggestions**

Opportunity Evaluation - Auto detect

Sample Numbers:

Hives in medium size commercial business:
5,000 hives

Average inspection time per hive about:
6 min

Average inspection per month:
2

CCD mortality rates per season:
X < 5% considered low (100 hives)
X > 20% considered high (1000 hives)

**Very rough estimates*

Regular Inspection:

Expert Beekeeper hours per month:
(6 min x 5,000 hives)/60min x 2
= **1000 Hr/Mon**

Note: Regardless of number of CCD cases, if number of hives remains more or less constant

With Detect and Alert App:

Given a "bad case" of 20% cases to inspect
(15 min x 1000) /60
= **250 Hr/Mon**

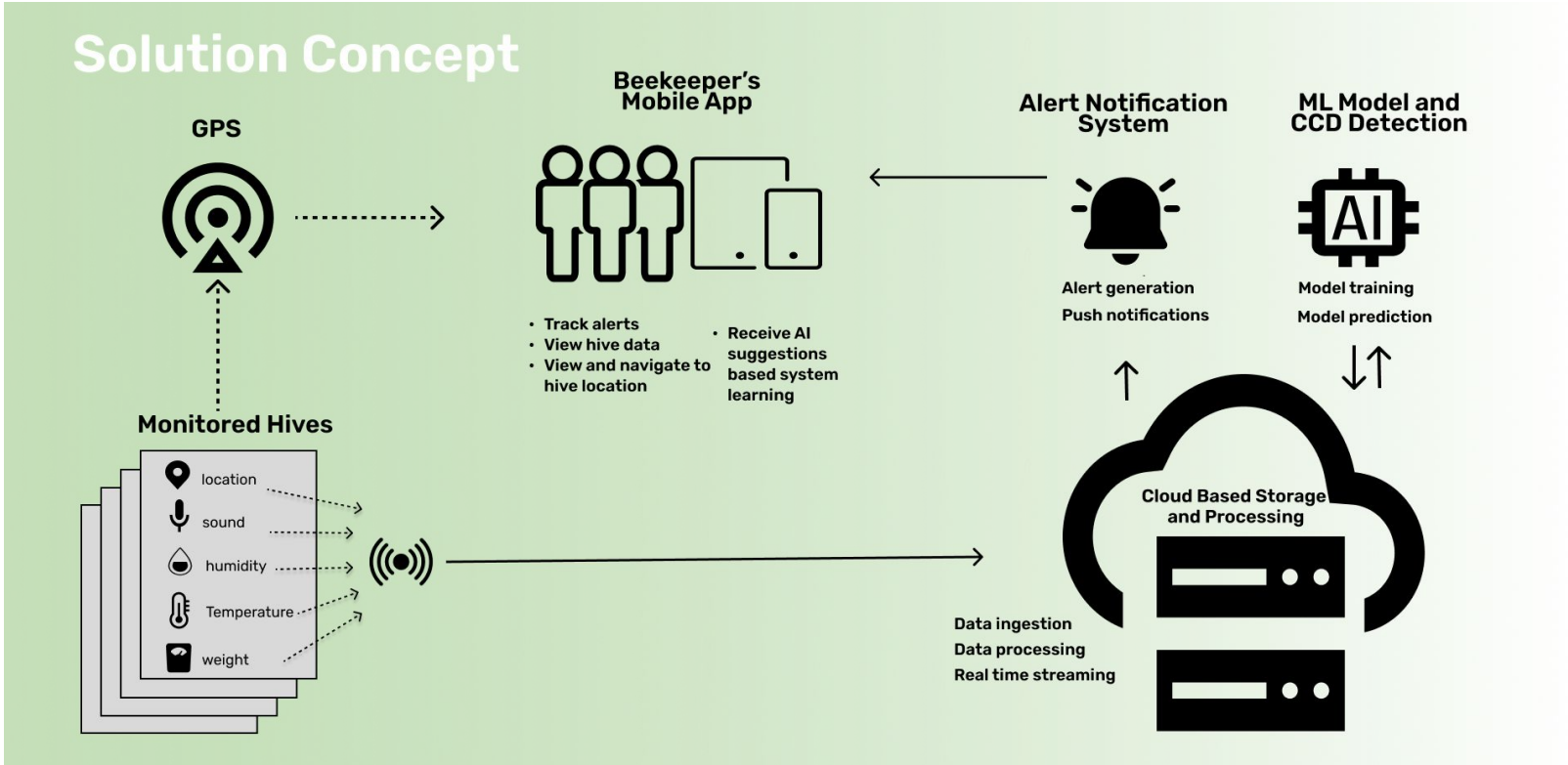
Notes:

- *On demand alerts, no need for recurring inspection*
- *Added 12 min for longer average inspection time*
- *Cases (< 20%) total hrs will decrease as well.*

Success Criteria

Outcomes	Metrics	Targets
Improve the beekeeper's CCD Detection Process Maximize the effectiveness of expert beekeeping skills. Offer diagnostic options Offer treatment suggestions Improve diagnostics and suggestion	**Outcome**: Increased detection stats and reduced CCD related labor and cost. Improved assessment and treatment results **Output**: The product uses a Binary Classification Tree model (0/1) for CCD detection. Output include prediction confidence level.	**Outcome:** Reduction in CCD detection labor/cost >= 60%. Colony mortality with ML <= mortality with the expert inspection. **Output:** Detection probability approaches 100% (with a possible bias towards false positive)

The Product - Solution Concept



What is Colony Collapse Disorder?

Healthy Hive

Colony Collapse Disorder (CCD)

Opportunity Evaluation: The Details

Sample Numbers:
Hives in medium size commercial business:
5,000 hives

Average inspection time per hive:
8 min

Average inspection per month:
2

CCD mortality rates per season:
X < 5% considered low (100 hives)
X > 20% considered high (1000 hives)

**These are very rough estimates*

Regular Inspection:
Expert Beekeeper hours per month:
(8min x 5,000 hives)/60min x 2
= 1333 Hr/Mon

Note: Regardless of number of CCD cases, if number of hives remains constant, the time wil

With Detect and Alert App:
Given a "bad case" of 20% cases to inspect
(20min x 1000) /60
= 333 Hr/Mon

Notes:

- *On demand alerts, so no need for recurring inspection*
- *Added 10 min per hive to adjust for longer journey between hives (might be much less)*
- *As cases decrease (<20%) total hrs decrease as well.*

Opportunity Evaluation: Assessment

Can ML Solve It?

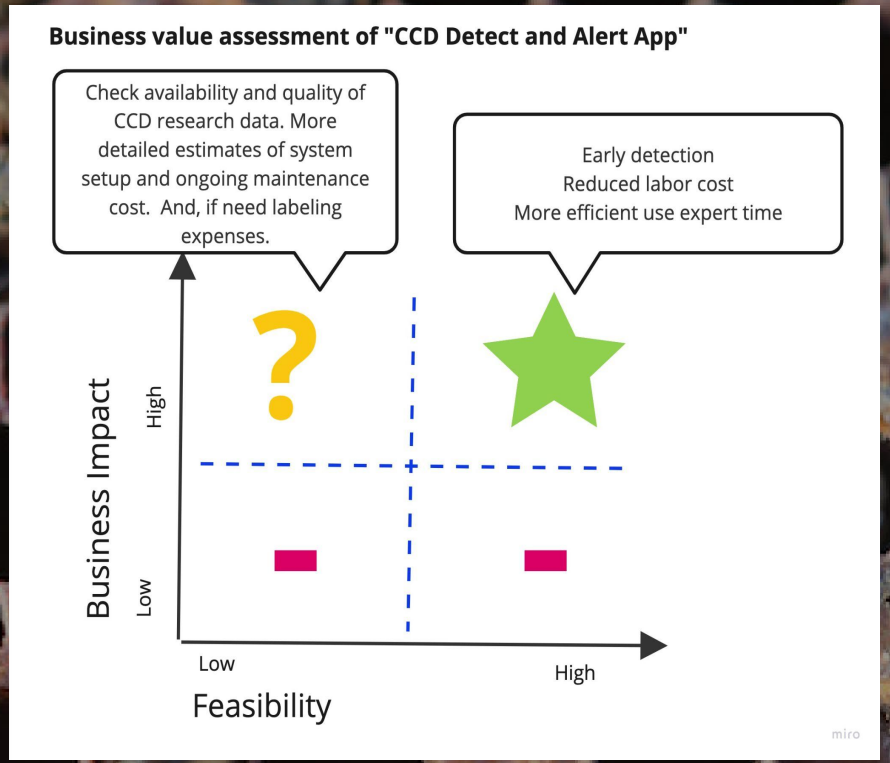
Yes. Easy, but requires:

- Initial investment in hardware/software
- More user research

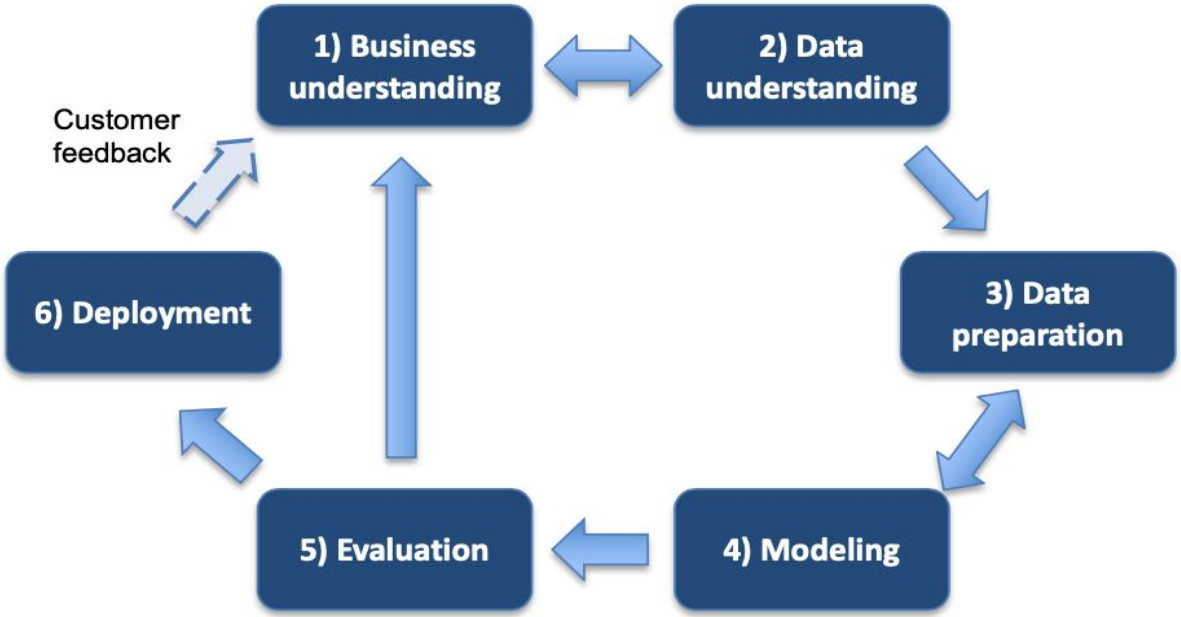
Can we get data?

Yes.

- We can access existing data used in CCD research
- We would also collect training data, but again, this will require initial investment in sensors and hardware



CRISP-DM Process



CRISP-Business Understanding

Problem Definition

Target User

Beekeepers (experts doing the beekeeping work)
Beekeeping business owner
Beekeeping Operations manager

Problem

Colony Collapse Disorder is destroying honey bee colonies, hurting the commercial beekeeping industry. Early detection can save colonies and prevent spread. The cost of not addressing CCD has an economic, environmental, and agricultural impact.

Why it Matters

Regular inspections of all hives to find, locate, and treat infected colonies are labor and cost intensive. Not inspecting the hive frequently enough, may result in the loss of bee colonies and the spread of infections which may in turn prevent the Beekeepers from meeting obligations.

Current state

The most common way CCD is addressed today is through regular inspections (weekly/monthly) of all hives. There are also ongoing efforts (similar to this proposal), to use ML and sensors to help manage hives.

CRISP-Business Understanding

Success Criteria

Expected Impact

- Decrease time and cost of detecting and locating distressed colonies.
- Maximize the effectiveness of expert beekeeping skills.
- Improving early detection of CCD.
- Decrease mortality of colonies due to CCD.

Metrics

Outcome: Increased detection and treatment of CCD, while reducing labor and cost.
Output: The product will use a Binary Classification Tree model. The output would be a 0/1 prediction of CCD in the hive. Output will also include a probability score for the prediction confidence level.

Targets

Outcome: Reduction in CCD detection labor/cost > 70%.
Colony mortality with ML <= mortality with the expert inspection.
Output: Detection probability approaches 100% (with a possible bias towards false positive)

Constraints

Alert latency should be less than 24hrs from detection.
We want to keep system setup and ongoing maintenance costs low because the goal is cost savings. Ideally, ongoing costs <= 30% of cost savings (more research needed)

CRISP-DM Data Understanding

Factors that might impact model data

Environment Conditions	**In hive conditions** such as temperature, humidity, activity, **External changes:** in temperature, humidity, rainfall, pollination **Proximity to technology:** “Wifi trees” and power lines.
Geographic Location	Local types of plants and flowers can affect the nutrition of the bees. Different regions may have varying pesticide usage, Pests and diseases (e.g., Varroa mites, Nosema) may vary by region.
Seasonal Changes	Seasonal variations in flora, bee activity, and hive conditions can influence data patterns.
Species and Genetics	Different bee species and genetic strains may exhibit variations in behavior and susceptibility to CCD.
Business Assignment	Moving the hives to different locations for the purpose of pollination, introduces many different stressors that may impact the health of the hive.

CRISP -DM Data Understanding

CDD Understanding Data

Sensor data

In hive sensors for temperature, humidity, sound, Weight (Images?)

Environmental data

Weather data
- temperature
- humidity
- precipitation
- wind
- pollen counts

Hive location data

GPS Geographic coordinates of the hives (all and each)
Wifi towers in the vicinity
Power lines

Existing Research data

Check for existing data and models related to CCD and bee population.

Past records

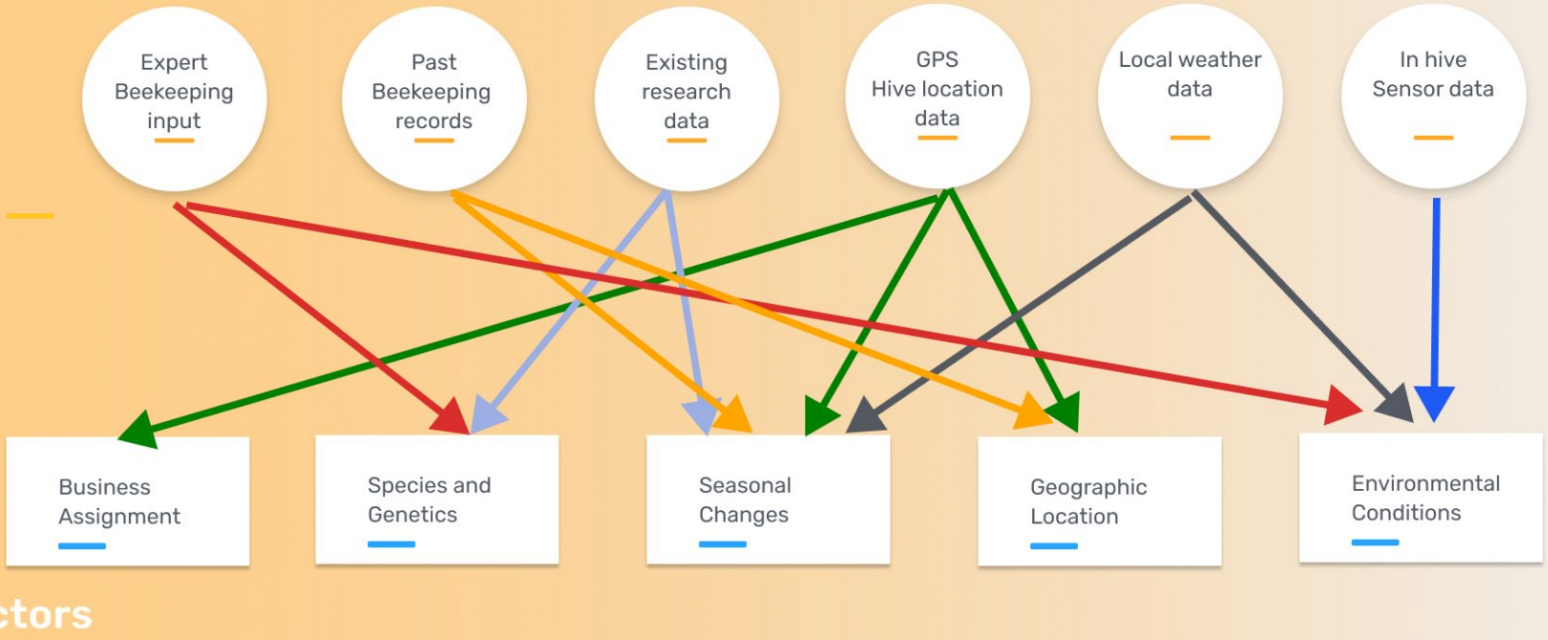
Past records on hive inspection and treatment can help with correlation.

Expert Input

Expert input can provide insight into feature selection and will likely be needed for labeling data.

CRISP -DM Data Understanding

Data Sources for different Factors



CRISP -DM Data Understanding

Data Considerations

- How much data is needed and what types?
  - Form sensor
  - On climate
- From GPS (for the hive cluster, for each hive, etc)
  - For each hive
  - For the target location of the hives
- How reliable is the data source?
- What is the cost of the data
  - From research
  - From weather sources
  - From GPS
  - From experts
- What is the cost of importing and integrating the data (if from external sources)

Note on data

- My hope here, which would need to be tested, is that most of our feature data will come from the hive sensors and weather predictions. If this is the case, we will have some measure of control over the data quality.

CRISP -DM Data Understanding

CCD Detect, Alert, Locate: Features

Sensor Data Features

Temperature:
- The mean temperature within the hive.
- Variability or fluctuations in temperature.
- Seasonal patterns in temperature.

Humidity:
- Mean humidity levels within the hive.
- Changes in humidity over time.
- Relative humidity variations.

Sound:
- Frequency and amplitude of hive sounds.
- Buzzing patterns and their changes.
- Sound variance or anomalies.

Weight:
- Hive weight changes over time.
- Weight fluctuations during specific periods (e.g., winter or foraging seasons).

Environmental Features

Weather Conditions:
- Temperature, humidity, and rainfall data from external weather sources.
- Correlations between hive conditions and weather patterns.

Forage Availability (if available):
- Information on the availability of floral resources and nectar-producing plants in the vicinity of the hive.
- Pollen counts and diversity of pollen sources.

Hive Location

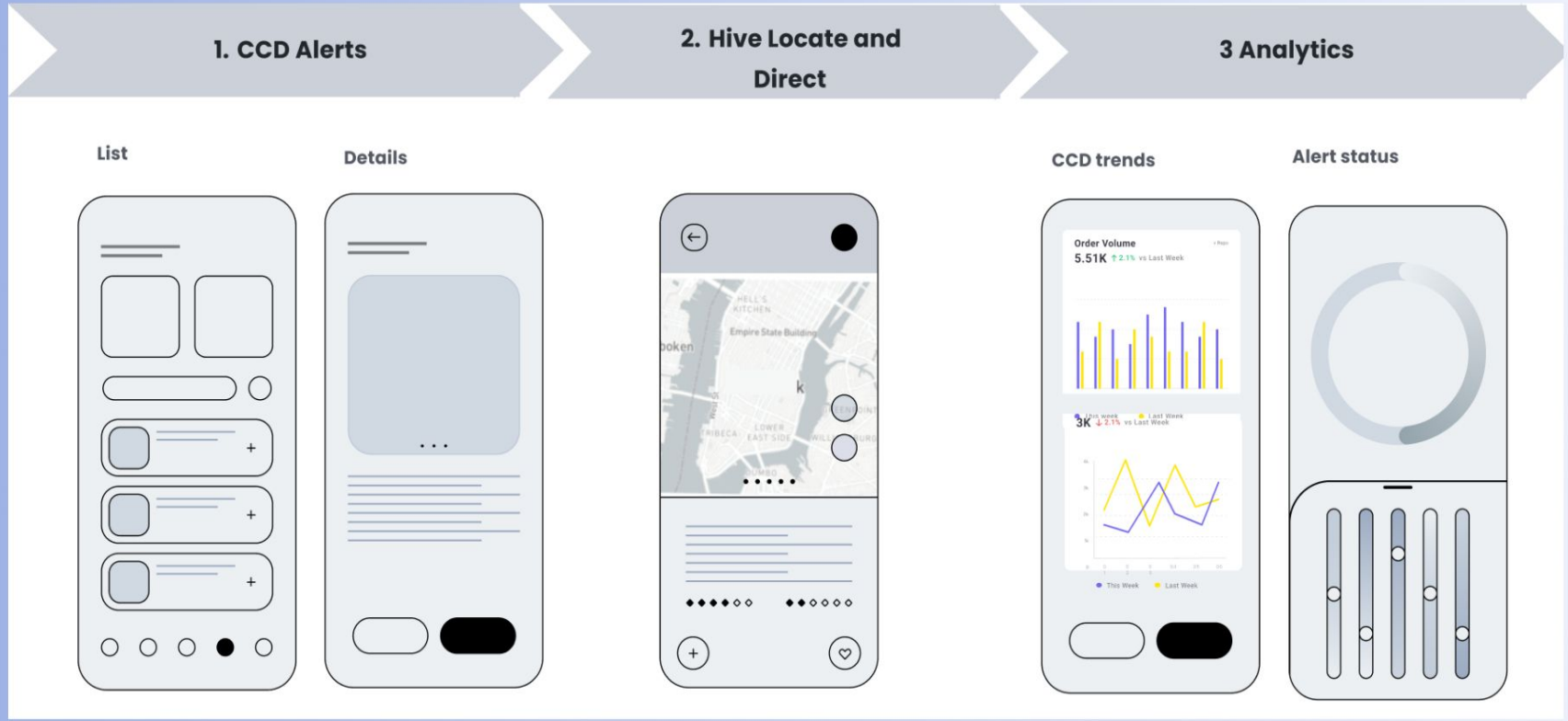
- Geographic coordinates (latitude and longitude) of the hive.
- Proximity to specific land use types (e.g., agricultural areas, urban areas).

Time-Related Features

Seasonal Indicators:
- Seasonal indicators to capture variations related to specific times of the year.

Time Series Features:
- Rolling averages or moving statistics for sensor data.
- Trends and patterns in time series data.

Validation - User Facing Functionality



CRISP- DM Solution Concept

Concept narrative

Data Collection

- **Sensor Data Collection**: Implement sensors in beehives to collect temperature, humidity, sound, weight, and other relevant data. This data is transmitted to a central cloud-based platform.
- **Location Tracking:** Use GPS or geolocation services to track the location of each hive, and send this data to the cloud.

Cloud-based Storage and Processing

- **Data storage**: Store sensor and location data in a cloud-based database or data lake.
- **Data Preprocessing**: Implement data preprocessing pipelines to clean, validate, and transform the incoming data. Handle missing values and outliers.
- **Real-time Streaming:** Use real-time data streaming capabilities to process and analyze incoming data, enabling timely detection of CCD-related patterns

ML and CCD Detection

- **Machine Learning Models**: Train and deploy machine learning models for CCD detection based on the preprocessed data. These models analyze sensor readings and other relevant features to predict CCD likelihood.
- **Model Deployment:** Host the trained models on the cloud platform for real-time predictions. Implement mechanisms for model updates and maintenance.

Alerting and Notification System

- **Alert Generation:** When the CCD model detects a potential issue or abnormal hive condition, generate alerts and notifications within the cloud platform.
- **Alert Delivery:** Send alerts to beekeepers and users via various channels, such as push notifications, email, SMS, or within the mobile app.

Mobile Application

- **User Interface:** Develop a user-friendly mobile app that allows beekeepers to access hive data, view CCD alerts, and track hive locations on a map.
- **Data Visualization**: Implement interactive data visualizations to display data trends and historical information.
- **User Authentication:** Incorporate user authentication and authorization to ensure secure access to hive data.

GPS services

- **Geospatial Integration**: Integrate geospatial services or mapping APIs to display hive locations and provide directions to beekeepers.

Solution 1

UX Validation Steps

New ta

Iteration	Itr 1 - Conpect General	Itr 2 - Alerting flow	Itr 3 - Location and Navigation	Itr 4 - Reporting	Itr5 - Putting it all together
Learning and Validation	Hive level Concept Mockups: Application flow and features • Early detection alerts and event management. • Hive locator and navigation to hive • Reporting and trends • Anything else?	Interactive wireframes to explore and validate the behavior of the CCD alerting system: 1. Alert list most recent on top 2. Alert Details 3. Alert status/workflow/filters and sorting	Interactive mapping app - re-used familiar mapping app display and controls • Show the affected hive on the map. • Show other affected hives in vicinity. • Show navigation controls • Show directions options for navigating to the hive.	Dashboard for tracking Outcome trends Mockup an interactive dashboard that allows user to explore outcome trends	Interactive Demo using high Fidelity Mockups that simulate real time interaction with sample taken from hive location, historical data and current research data
Details	Concept Validation • Validate concepts with users. • Validate core features and flow • Discuss feature priority • Learn more about what the app should do	Validate content and flow with users: • What content is most important to display at list level? • What detail content would be of interest to the Keeper? • What would be the most useful way to track alerts and followup? • What is the importance of workflow? • What filtering and sorting options would be useful?	Validate the mapping feature with Users: • How would they use it? (would they head out for each hive or would they batch hive visits?) • What type of information would be most useful? • If a CCD was detected, and the hive visited and treated, what would be the best way to integrate the hive status into the map?	Validate and explore the type of outcome trends that might be of interest to the Beekeeper for example • Number of CCD incidents detected in hives • Number of hives treated? • Number of hives saved? • Time taken to respond? • Geo-distribution of CCD incidences • Compare current mortality rates to historical stats • Explore possible ways to track expert time spent (for better outcome tracking)	Validate the product flow with users.

Solution Validation Plan

Solution Concept Narrative - More

Cloud Hosting and Scalability

Cloud Platform

Utilize a cloud service provider (e.g., AWS, Azure, Google Cloud, OCI) for hosting your application, databases, and machine learning models. Leverage cloud resources for scalability and reliability.

Data Security and Compliance

Data Security

Implement strong security measures to protect sensitive data and ensure compliance with data privacy regulations.

Monitoring and Maintenance

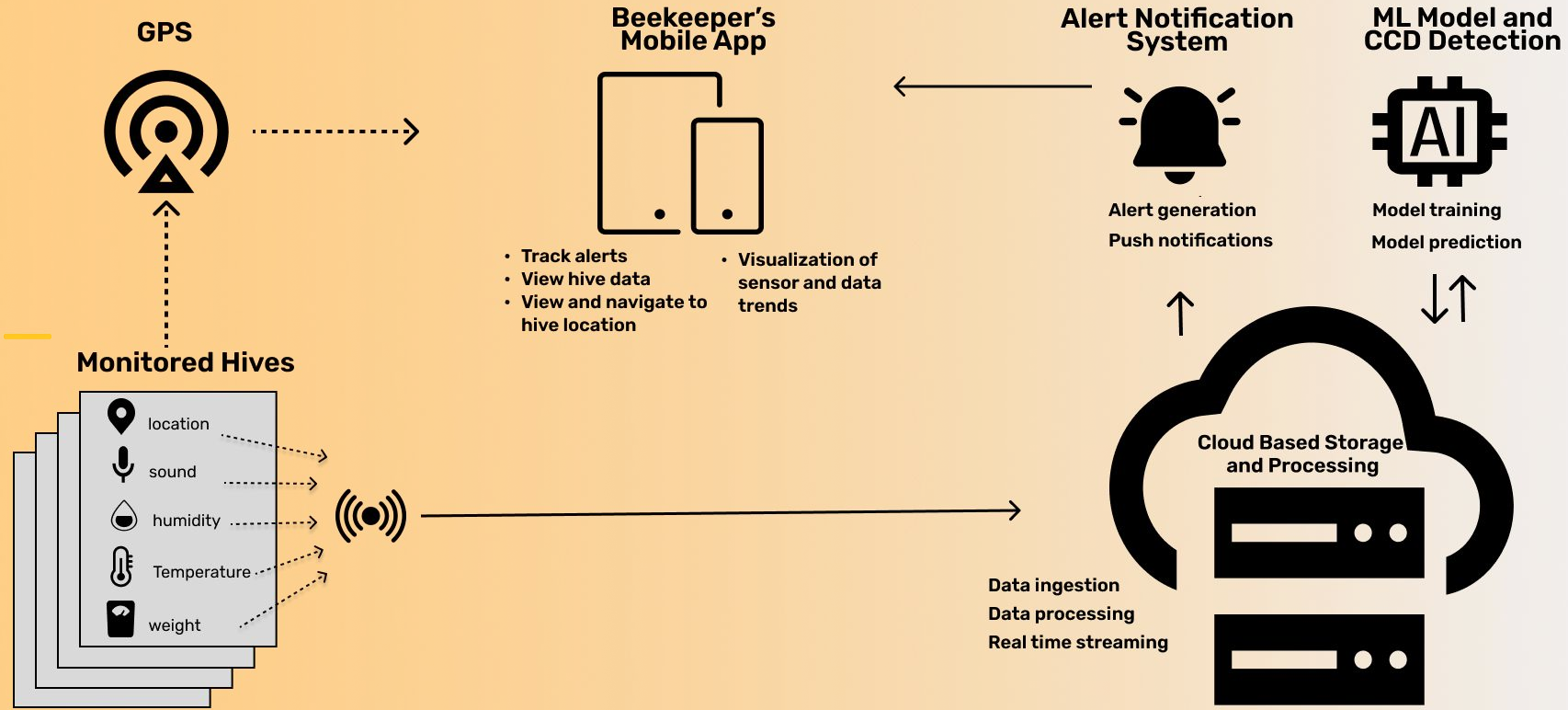
- **Monitoring Tools:** Set up monitoring and logging tools to track the performance of your cloud-based services and applications.
- **Model Maintenance:** Implement regular model updates and retraining to adapt to changing conditions and data drift.

Change management support

- **User Support**: Provide user support and training to beekeepers and app users.
- **Documentation**: Maintain documentation for the entire system, including data sources, data processing pipelines, model details, and deployment procedures.

CRISP - DM

Solution Concept



Solution

Solution Validation Steps

Iteration	Itr 1	Itr 2	Itr 3	Itr 4
Setup	**Interactive Wireframes - application flow** High level Concept Mockups highlighting value added to current workflow: Hard-coded data based on historical records and research data (if exists).	**Simulate DemoL local data alerts, lotion and data updates)** Interactive mockups to validate system behavior (Jobs To Be Done): • Sample **data taken live from customer sample hives**. • Simulate **behaviors (alerts, locations, data updates) are manually triggered**. • No storage, system setup or automation yet.	**Prototype - product, system and output validation System Implemented** • Implement the ML system design (see diagram) • Setup early adopters live sensor data capture • Clean and prepare the data	**Version 1.0 for Outcome validation** Setup a way to capture values that are critical for reporting on outcomes • Number of cases detected • Time to detection • Time spent on detection • Time spent on treatment
Purpose	**Concept Validation** • Validate concepts with users. • Feedback/Analysis.	Validate Jobs to be Done with group of **early adopter customers**. Use an interactive mockup (a live demo). • Validate understanding from Itr1 feedback. • Validate Product's expected behavior • Validate the time constraints of system an human responses	**Building and testing a minimal working prototype** • User testing and feedback • System testing • Data testing • Model validation • Output metrics evaluation	**CCD alert, detect direct app V 1.0: Validate** • UX • Outcome metrics evaluation • Model performance • System performance feedback • Cost validation
Data	Mock data based on historic records, Research data if exists.	With early adopter customers setup a demo with "real" data: • Define the data needs for a "live simulation" (Hive weight, sound, humidity, temperature). • Schedule date/time and locations for the simulated demo • During this time, collect data from the sample hives and use it to update the demo.	Full data set from at least 3 trial customers: • Sensor data • Environmental data • Location data	Combined dataset from all available sources • Sensor data • Environmental data • Location data • Historical baseline • Existing research data • Expert input
Model	Model not implemented yet. Mock data assumes: • Ability to predict CCD for a hive based on sensor input and environmental data. • Assumes and ability to locate a hive navigate user to it.	• Apply heuristic analysis to predict hive health based on historic data and in-coming sensor data from sample hives. • During the simulation, when data is updated, manually calculate to check when a hive reaches the CCD thresh-hold.	CCD detection is a classification prediction. This product will use Simple **Binary Classification Tree model**, with a with a bias to false positive. • **Cloud processing** (real time detection no needed) • **Offline learning** (immediate update not needed) • **Online prediction** (predictions should happen promptly)	I will add a probability threshold to my Binary Classification Tree.
App	Concept wireframes (Figma) • A CCD early Alerts "Inbox". • Hive alert details. • Hive geo-locator. • Dashboard with trend analysis.	Using "real" sample sensor data simulate the application behavior. • Define the behaviors that need to be simulated and how to trigger them in the demo. • Use the spreadsheet and heuristic calculations to update the"interactive" mockups in the demo and "manually" trigger the planned behaviors.	First real prototype with data flow from edge input (hive) to edge output (mobile app) through Cloud (and GPS)	Fully functional • UI/UX • Fully functional alerting system • Realtime hive location and navigation • Trends analysis (dashboard)

ML System Level Architecture

Edge versus Cloud

→ Cloud computation

- Unlike fraud detection, CCD detection does not need to happen in real-time.
- My assumption is that a latency of up to 24 hours from input can be tolerated.
- My assumption is that this latency can be achieved using cloud computation.
- Privacy is not a main consideration

Pros: Cloud would be a better choice for this application because:

- Privacy and latency are not a consideration
- It is expected to process data from a large number of hives and integrate sensor and environmental inputs
- It will integrate with GPS,
- It will benefit from centralized data management.

Note: Both assumptions (latency and processing performance need to be validated)

Offline versus Online Learning

→ Offline learning

- Online learning is most useful when the input may have an immediate impact on the prediction algorithm, and therefore the model needs to be retrained continuously.
- In the case of CCD, the input can change (on a dime) yet it is unlikely that this change will have an immediate effect on the prediction algorithm.
- Therefore, learning can take place "after hours", offline.

Pros: This will keep the model simpler and easier to evaluate.

Batch versus Online Predictions

→ Online Prediction

- The sooner the CCD case is addressed, the better. The aim therefore would be to minimize the time between CCD detection and alert generation.
- While real-time detection is not required in the case of CCD, it is best if the prediction is made and an alert goes out as soon as the input is available in the system.

Pros: Online (real-time) predictions will allow for faster response and intervention which is what the product hopes to offer.

Concept and Data DRIFT

Possible causes

Environmental changes (such as changes to land use and the introduction of new plant species and crops) can affect bee behavior and alter the factors that contribute to CCD.

Environmental Changes

New Genetic Bee Strains

New Bee strains can be introduced into the beekeeping business and these new trends may respond differently to CCD.

Pesticides are known to have a significant impact on bee health. The use of different types or formulations of pesticides can introduce changes in hive conditions and bee behavior that may influence sensor data patterns.

New Pesticides

Polination transportation

The transportation of beehives for pollination can add stressors such as vibrations, temperature fluctuations, movement disorientation, and limited foraging, all of which might impact hive conditions and data patterns.

Hive management practices, such as supplemented nutrition and added ventilation, can influence hive conditions and alter the patterns observed in sensor data.

Changes to hive management

Sensor degradation

CCD detection relies heavily on **Sensor input**. Over time, the sensors may degrade due to "wear and tear", environmental exposure, and calibration shifts. This may affect the input quality and contribute to data drift.

Drift

Concept and Data DRIFT

Managing Concept and Data Drift

1. Be on the lookout for new/unexpected conditions, and add new features when needed.
2. Adapt the model with regular updates and a retraining schedule.
3. Follow up on research and expert publications.

Environmental Changes

1. Actively monitor pesticide exposure. Train and update the model accordingly.
2. Followup on publication on new pesticides.
3. If needed, modify to capture new indicators of potential affects. Train and update the model accordingly.

New Pesticideds

1. Carefully track changes in hive management practice.
2. Train and update model regularly to account for such changes.
3. Segment data to closely track different phases of management practices.

Changes to hive management

New Genetic Bee Strains

1. Be on the lookout for new strains,
2. Collect data from hives we new strains,
3. Retrain the model to account for new behavior,
4. If needed, review features to include strain- specific behaviors that may affect CCD.

Transportation for Pollination

1. Establish baseline data for hives before transport, as a strong reference point.
2. Establish consistent transport protocols.
3. Detect and filter out data patterns related to transport stressors.

Sensor degradation

1. Regular Sensor Maintenance.
2. Data Quality Control measures to detect and filter out unreliable data.
3. Analyze historical data to detect trends.
4. Monitor sensor health.
5. Consider retraining and updating the model to changes in sensor input.