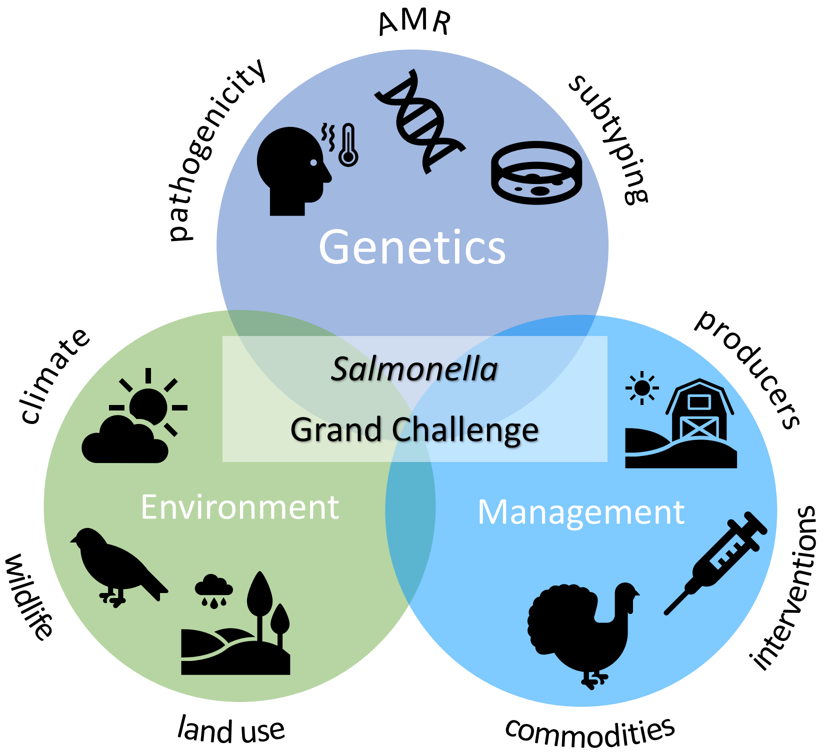
Use of data science to improve food safety:

The ARS *Salmonella* Grand Challenge and Data Analytics

Tatum Katz, Ph.D. M.A.Stat

ORISE SCINet Postdoctoral Fellow

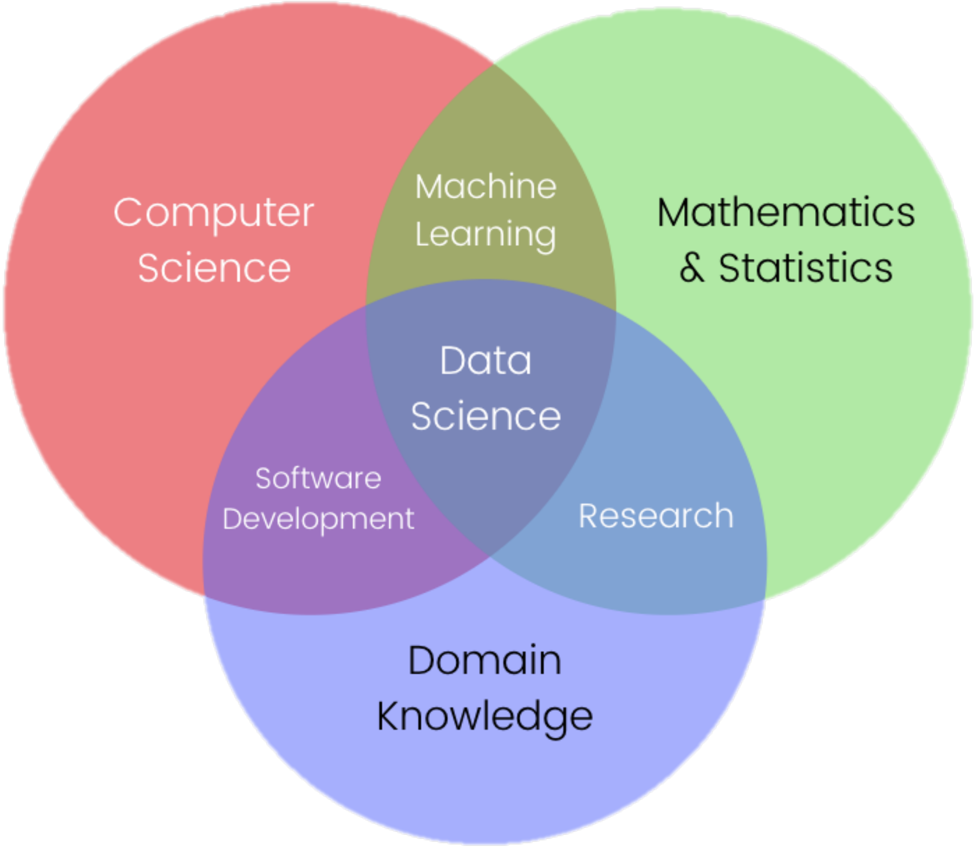
Meat Safety and Quality Research Unit, Clay Center, NE



My goals for the audience

- Learn about the ARS *Salmonella* Grand Challenge
- Understand what data science is
- Understand what data science can contribute to a research program
- Think about how you can leverage data science in your own research, and what challenges you may face

What is data science?

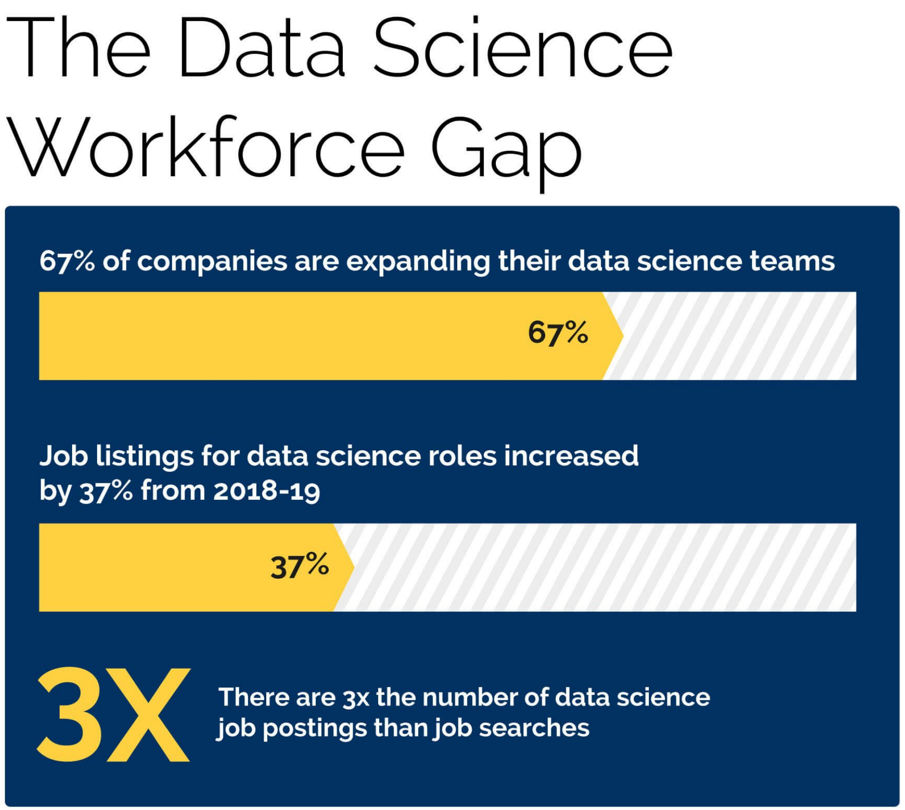


- Inter/transdisciplinary
- Combines math-stat, programming, and domain knowledge
  - Domain knowledge: expertise in the field of application
    - Ex, my domain is infectious disease ecology
- Goal:
  - Wrangle data from various sources
  - Identify novel insights to drive research forward
  - Hypothesis generation AND hypothesis testing
  - Create tools to help other researchers and stakeholders use data effectively

Why is everyone talking about data science all the time?

- Closely related to **machine learning** and **artificial intelligence**
- Closely related to **big data**
- Closely related to **visualization** and **communication** of research findings
- Why now?
  - We are collecting more data than we can ever digest into useful results
  - Lack of expertise and training available
    - Historically, data scientists received training by real-life problem solving instead of formal classroom learning
    - Ex, first cohort at my graduate program to be able to specialize in data science

This is an excellent time to take advantage of the fresh crop of data scientists!



The Data Science Workforce Gap

67% of companies are expanding their data science teams

Job listings for data science roles increased by 37% from 2018-19

There are 3x the number of data science job postings than job searches

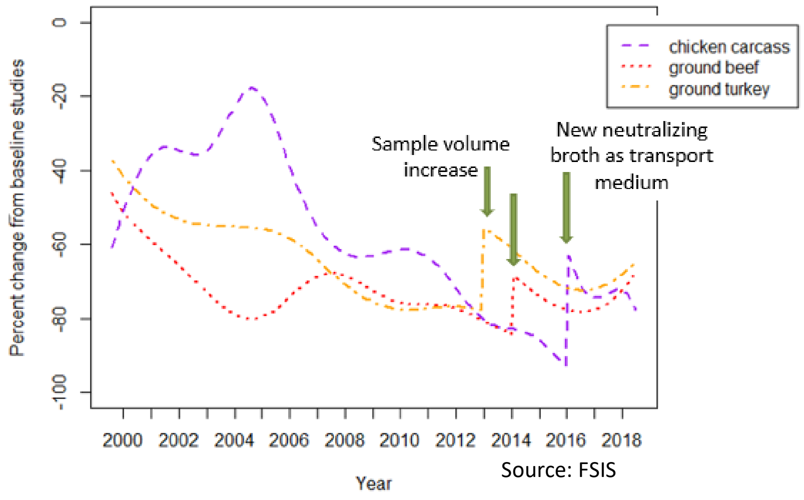
QuantHub, "The Data Scientist Shortage in 2020." (2020) https://quanthub.com/data-scientist-shortage-2020/

The *Salmonella* problem

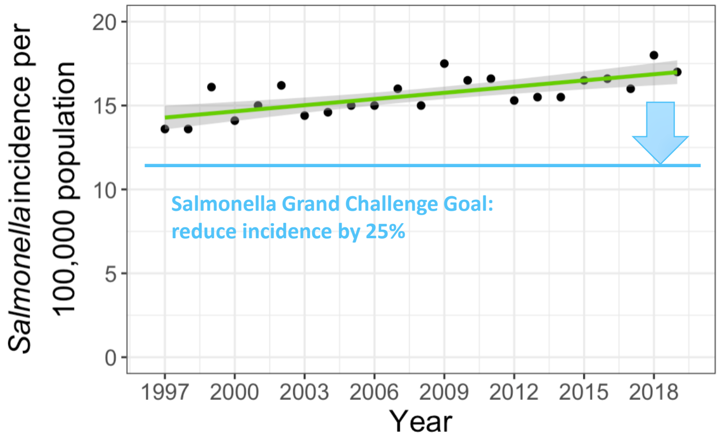
- Despite high investment and a downward trend in *Salmonella* contamination on product, illness rates have not decreased in 20 years

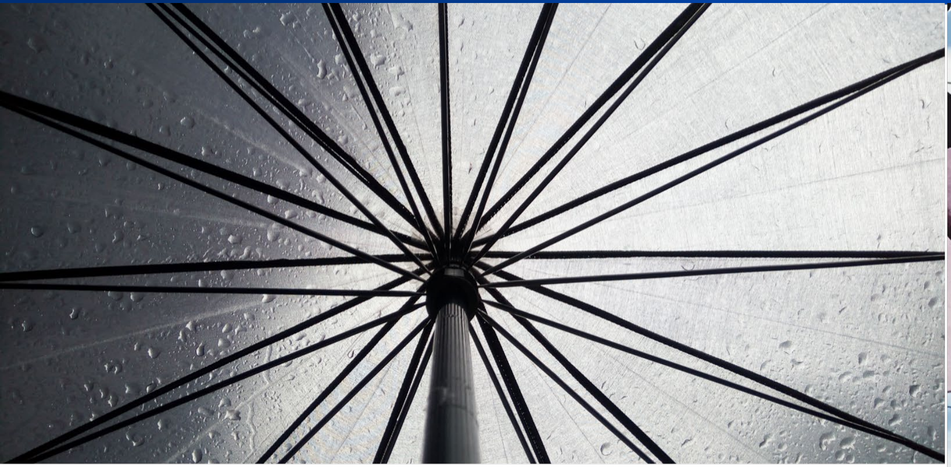
- New approaches and innovative thinking are needed to address the on-going challenge

Downward Trend In Salmonella Contamination



Source: FSIS



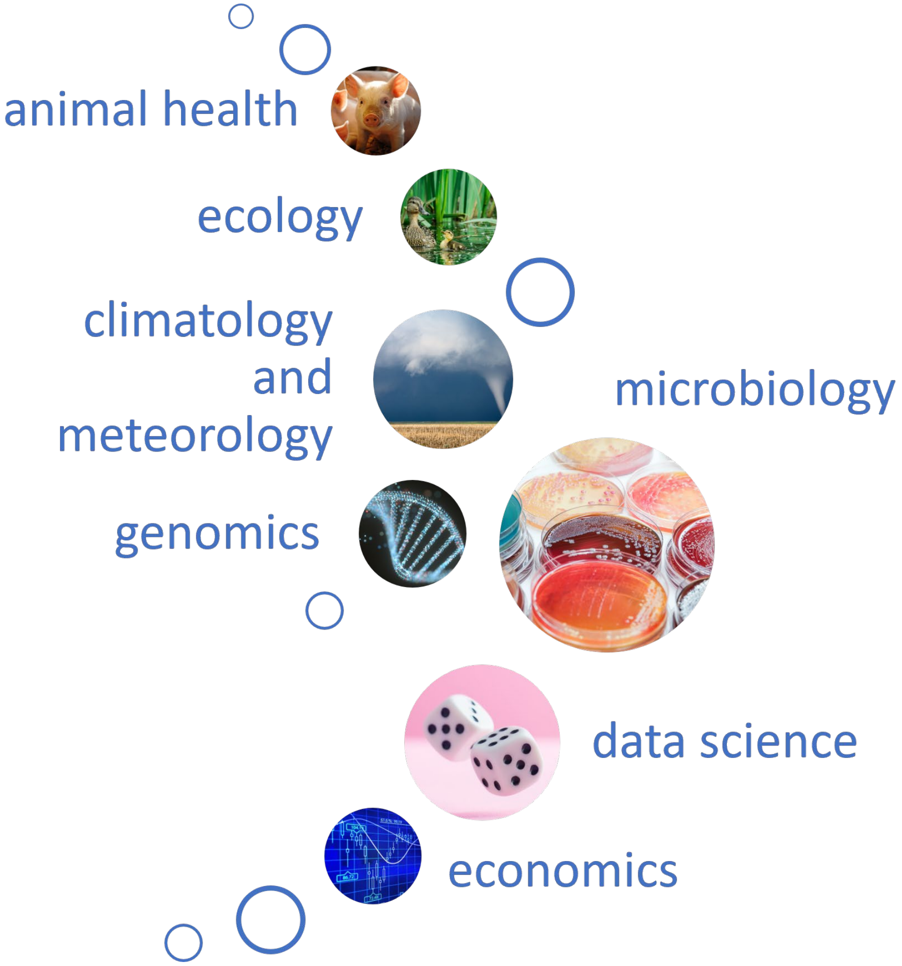


The ARS *Salmonella* Grand Challenge

An **umbrella** for ARS *Salmonella* research with the team goal of improving scientific impact, encouraging innovation and innovative thinking and enhancing collaborative, multidisciplinary work

The ARS *Salmonella* Grand Challenge

- 24 core members across the Agricultural Research Service and Economic Research Service, representing 8 locations
- A Computational Postdoctoral Community of Practice
- An Industry Advisory Board representing the four major meat and poultry commodities and four companies
- Additional collaborators across universities, industries, and associations



animal health

ecology

climatology and meteorology

microbiology

genomics

data science

economics

The ARS *Salmonella* Grand Challenge

Vision: Support stakeholders to implement affordable, effective, data-driven strategies to address Salmonella food safety goals

Gold Standard Protocols across projects and publicly available

Cutting-edge data management platforms and ontologies

Pilot studies to ground-truth findings for application in real-world systems

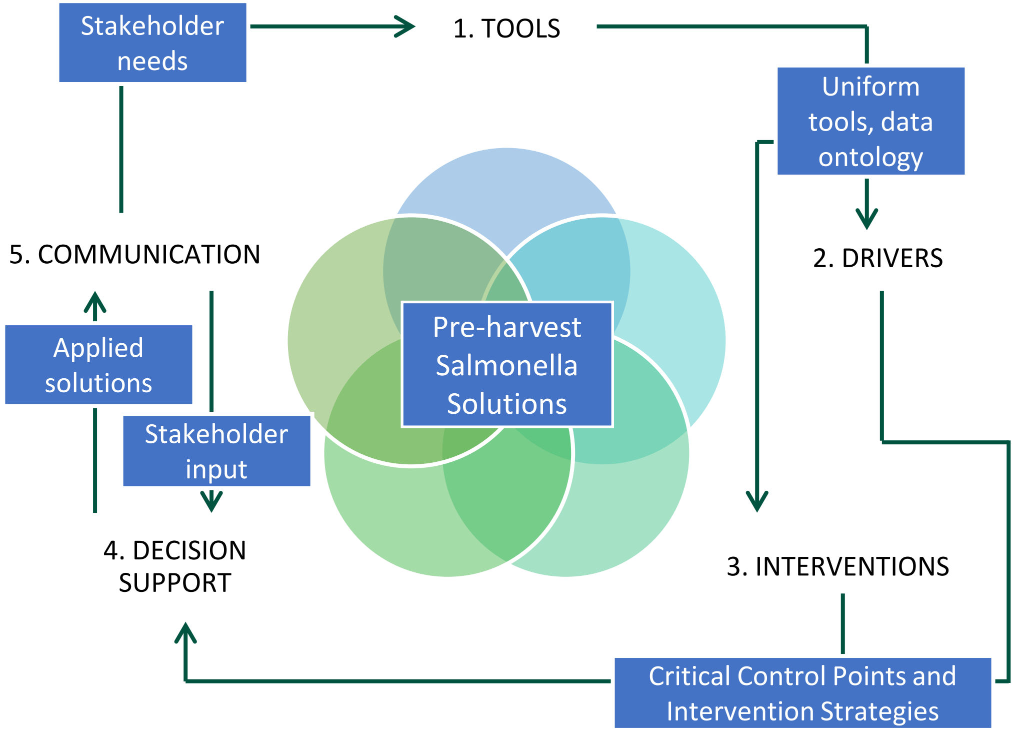
Cost-efficacy models for industry buy-in of interventions as economic viable

Decision support tools for stakeholders that are easy to interpret and use

Demonstration projects in collaboration with stakeholders to show value

The ARS *Salmonella* Grand Challenge

- Tools
  - Identify novel tools and standardized protocols
- Drivers
  - Define predictive reservoirs and drivers
- Interventions
  - Implement mitigation approaches that address production complexity
- Decision support
  - Apply AI and ML to develop easy to use decision support tools
- Communication
  - Implement a Communication, Outreach, and Data Management plan



Stakeholder needs

1. TOOLS

Uniform tools, data ontology

2. DRIVERS

3. INTERVENTIONS

Critical Control Points and Intervention Strategies

4. DECISION SUPPORT

Stakeholder input

Applied solutions

5. COMMUNICATION

Pre-harvest Salmonella Solutions

Using data science to improve food safety: A case study

with John W. Schmidt, Terrance Arthur, and Tommy Wheeler at USMARC

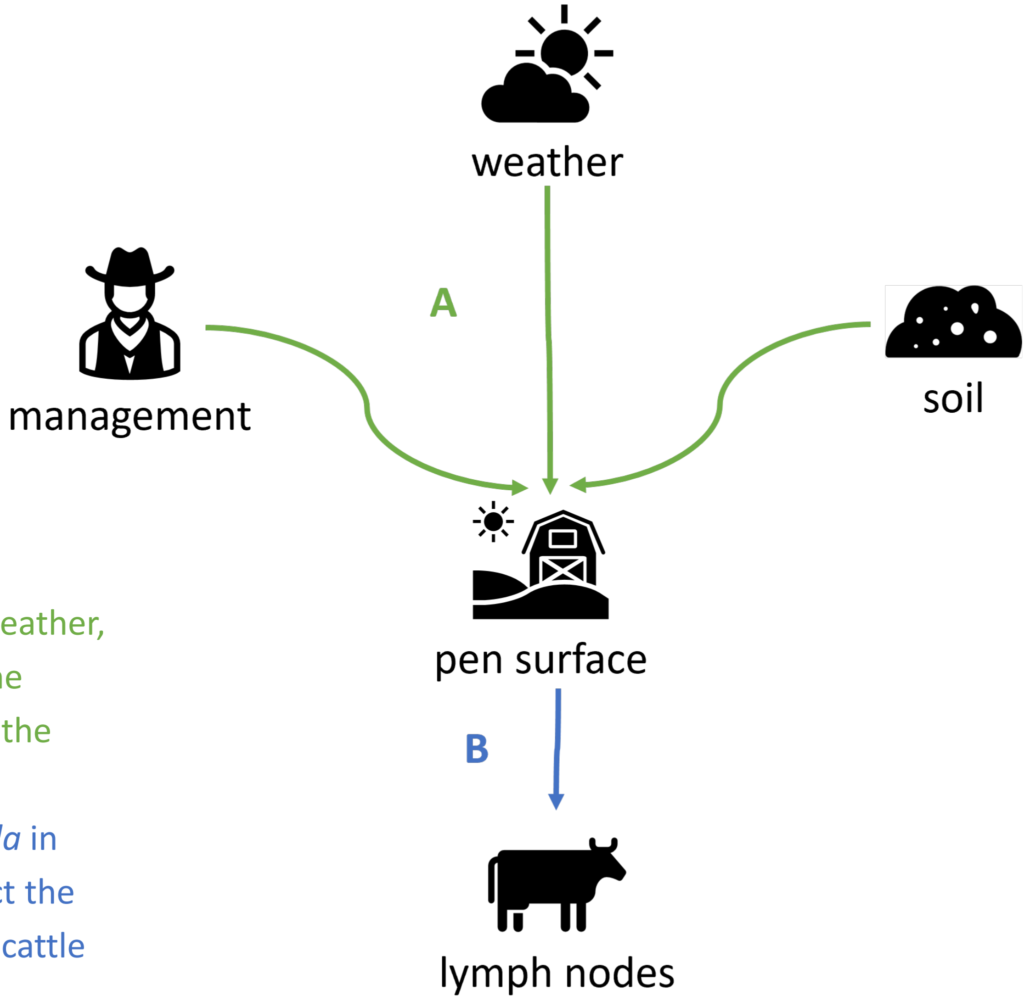
Goal:

Identify pre-harvest predictors of *Salmonella*-contaminated lymph nodes at harvest, for cattle

Working hypotheses:

A. Management decisions, weather, and pen soil can predict the presence of *Salmonella* in the pen surface

B. The presence of *Salmonella* in the pen surface can predict the presence of *Salmonella* in cattle lymph nodes at harvest

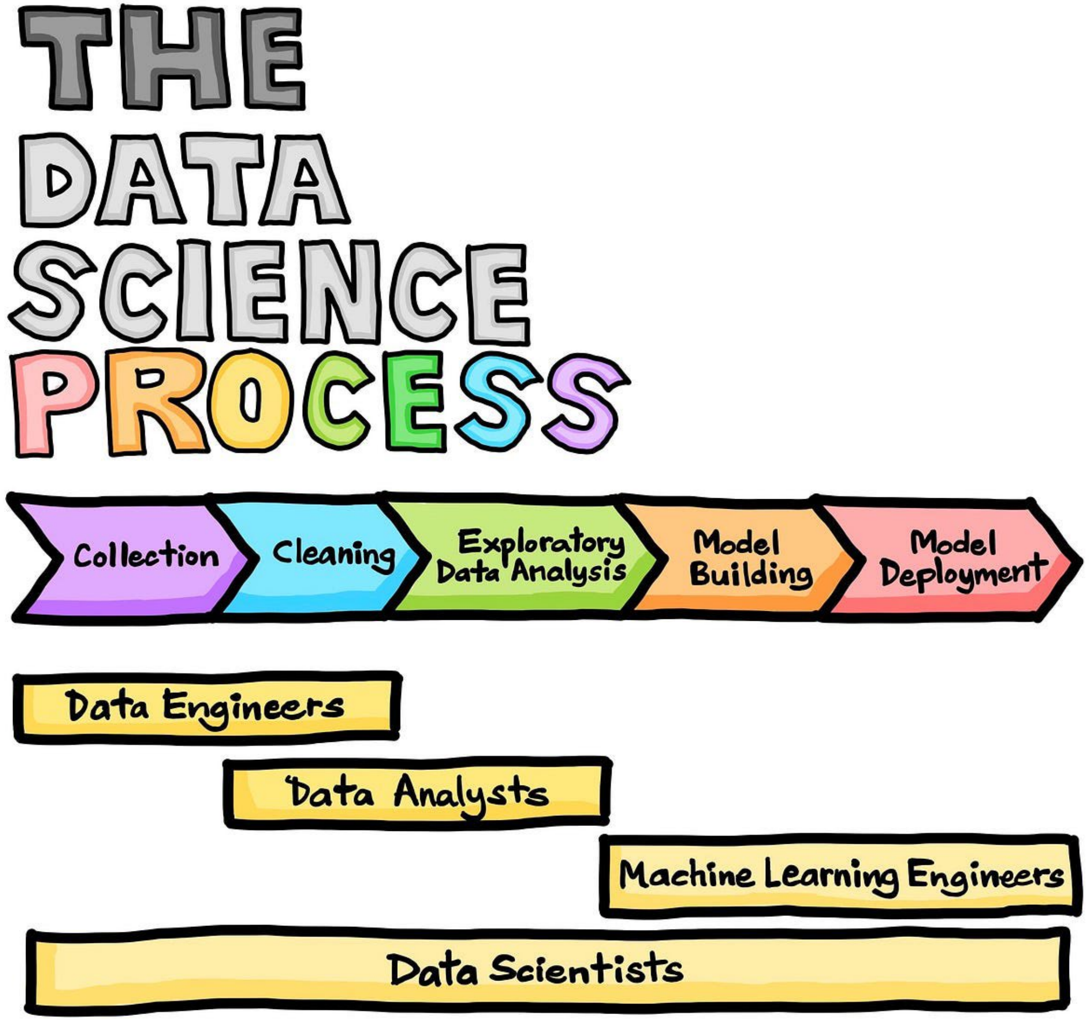


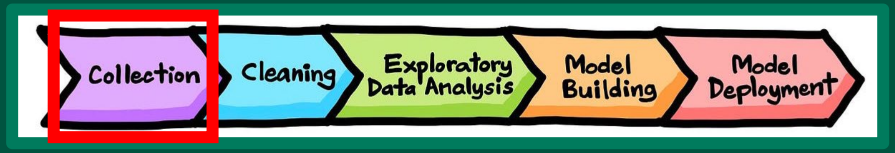
Using data science to improve food safety: A case study

with John W. Schmidt, Terrance Arthur, and Tommy Wheeler at USMARC

Goal:

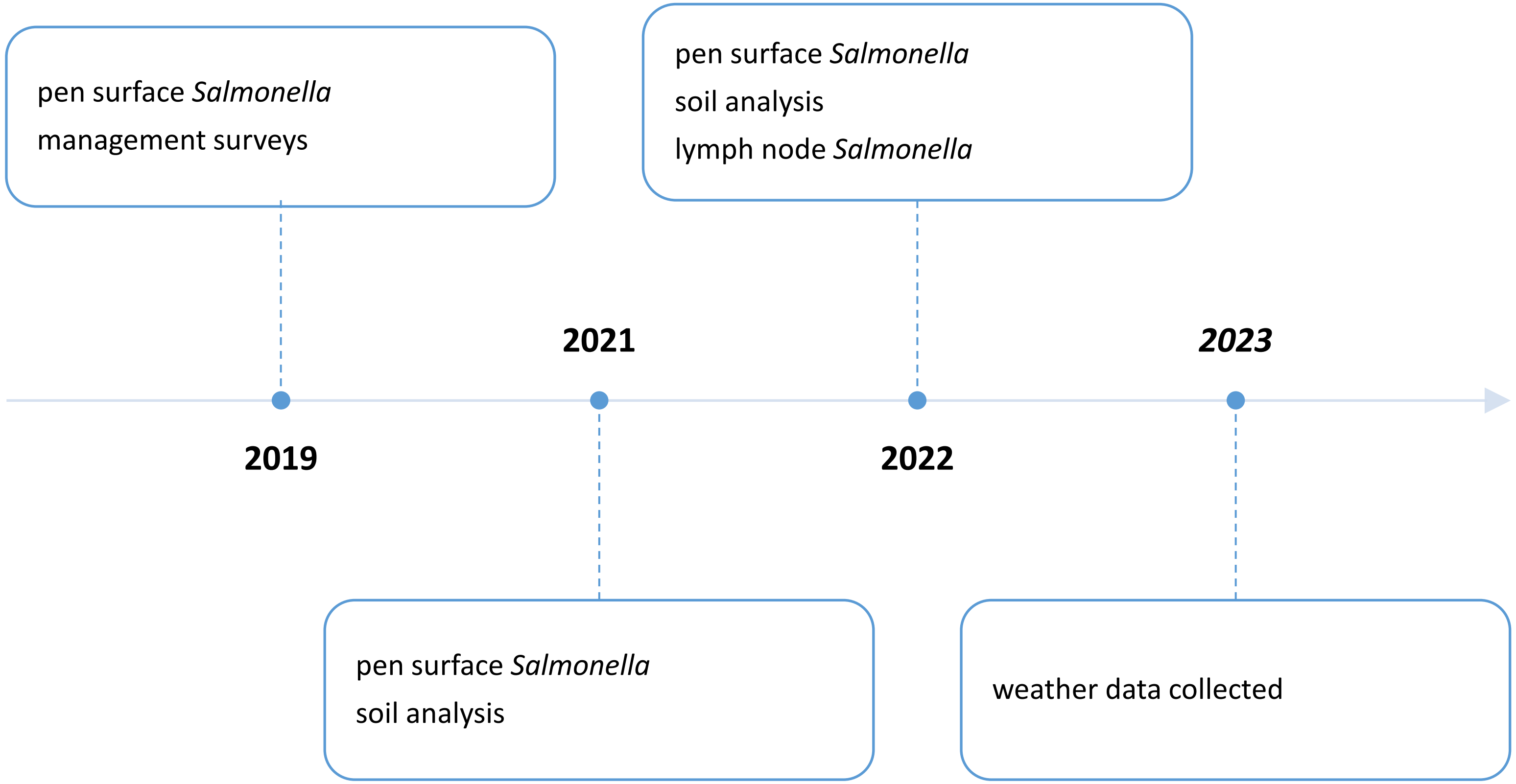
Identify pre-harvest predictors of *Salmonella*-contaminated lymph nodes at harvest, for cattle





Pre-harvest predictors of post-harvest *Salmonella* contamination

Data Collection



2019

pen surface *Salmonella*
management surveys

2021

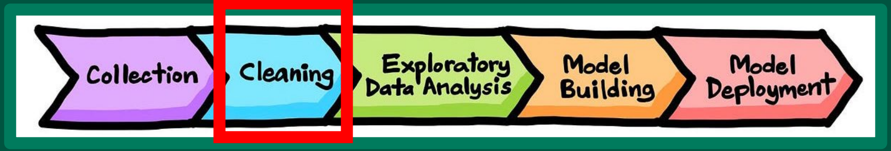
pen surface *Salmonella*
soil analysis

2022

pen surface *Salmonella*
soil analysis
lymph node *Salmonella*

2023

weather data collected



Pre-harvest predictions of post-harvest *Salmonella* contamination

Data cleaning

Pen surface *Salmonella*

- 4 samples per pen
- *Salmonella* detection
- *Salmonella* index
- *Salmonella* pathogenicity

Lymph node *Salmonella*

- One value per carcass, 25 carcasses per pen
- Three peripheral lymph nodes
  - superficial cervical
  - popliteal
  - subiliac
- *Salmonella* index

Soil analysis

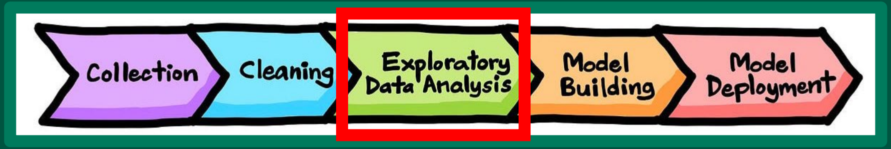
- One value per pen
- 100+ variables describing moisture, pH, minerals, metals, and salts in the soil

Management surveys

- Limited data
- One value per feedlot
- Where the cattle came from
- Pen density
- Probiotic information
- Tylosin use

Weather

- K State Mesonet data from local weather stations (matched by closest feedlot)
- One value per feedlot (some feedlots share a weather station)
- Air temperature
- Soil temperature at 2 and 4 inches depth
- Solar radiation
- Wind
- Precipitation
- Evapotranspiration of grass and alfalfa

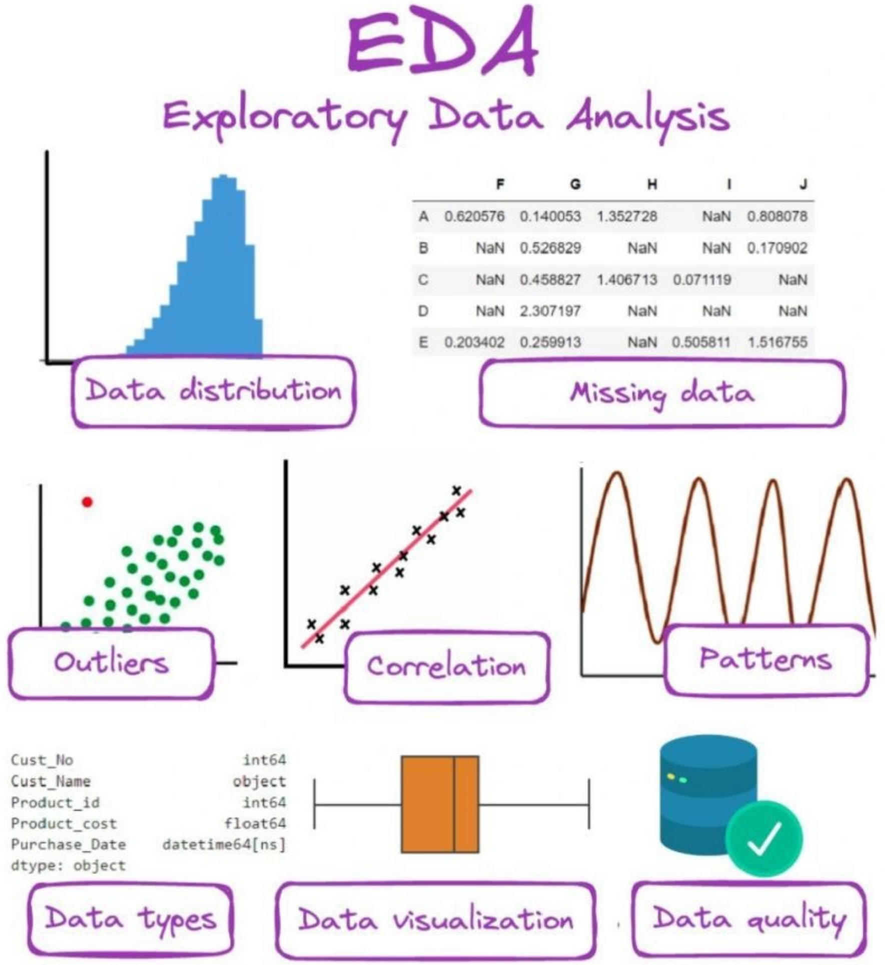


Pre-harvest predictors of post-harvest *Salmonella* contamination

EDA

Exploratory Data Analysis (EDA)

- Use visualization tools to learn about data
- Identify patterns
- Identify errors/outliers/issues
- Hypothesis generating



	F	G	H	I	J
A	0.620576	0.140053	1.352728	NaN	0.808078
B	NaN	0.526829	NaN	NaN	0.170902
C	NaN	0.458827	1.406713	0.071119	NaN
D	NaN	2.307197	NaN	NaN	NaN
E	0.203402	0.259913	NaN	0.505811	1.516755

```
Cust_No                    int64
Cust_Name                 object
Product_id                 int64
Product_cost             float64
Purchase_Date     datetime64[ns]
dtype: object
```

Collection
Cleaning
Exploratory Data Analysis
Model Building
Model Deployment

Pre-harvest predictors of post-harvest *Salmonella* contamination

EDA

Problems:

- After cleaning and combining all data, we had 95 variables describing 24 observations (complete cases)
  - n vs p problem
- Visualizing 95 individual variables is a lot

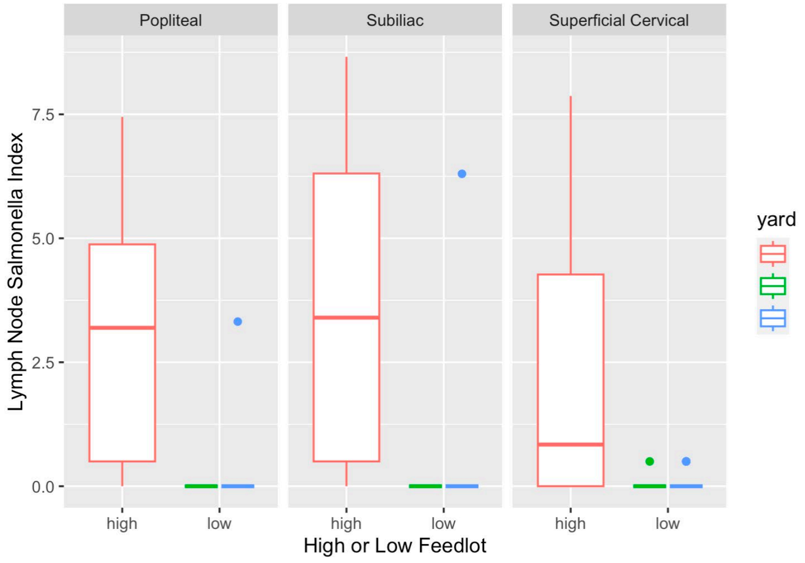
Solutions:

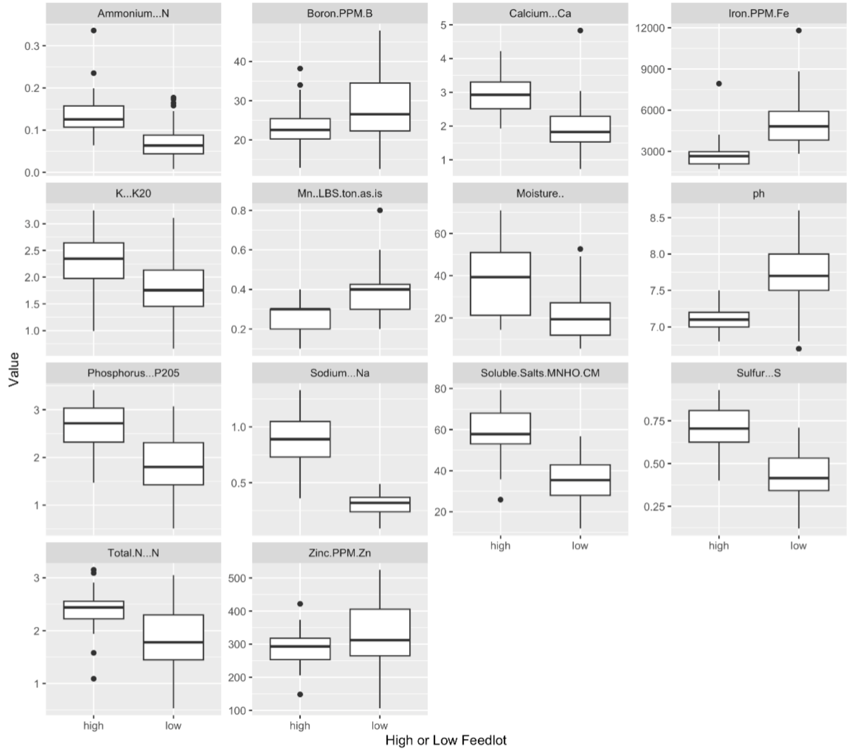
- Analyze each type of data (weather, management, soil) separately and together to identify trends (increase n, decrease p)
- Dimensionality reduction (PCA, CUR)
- Reclassify outcome variable from numeric continuous to categorical (*Salmonella* index to high/low *Salmonella*) to increase signal and reduce noise (increase power)
- Write programs to automate visualization of variables (still have to look at every single plot, though!)

Collection
Cleaning
Exploratory Data Analysis
Model Building
Model Deployment

Pre-harvest predictors of post-harvest *Salmonella* contamination

EDA

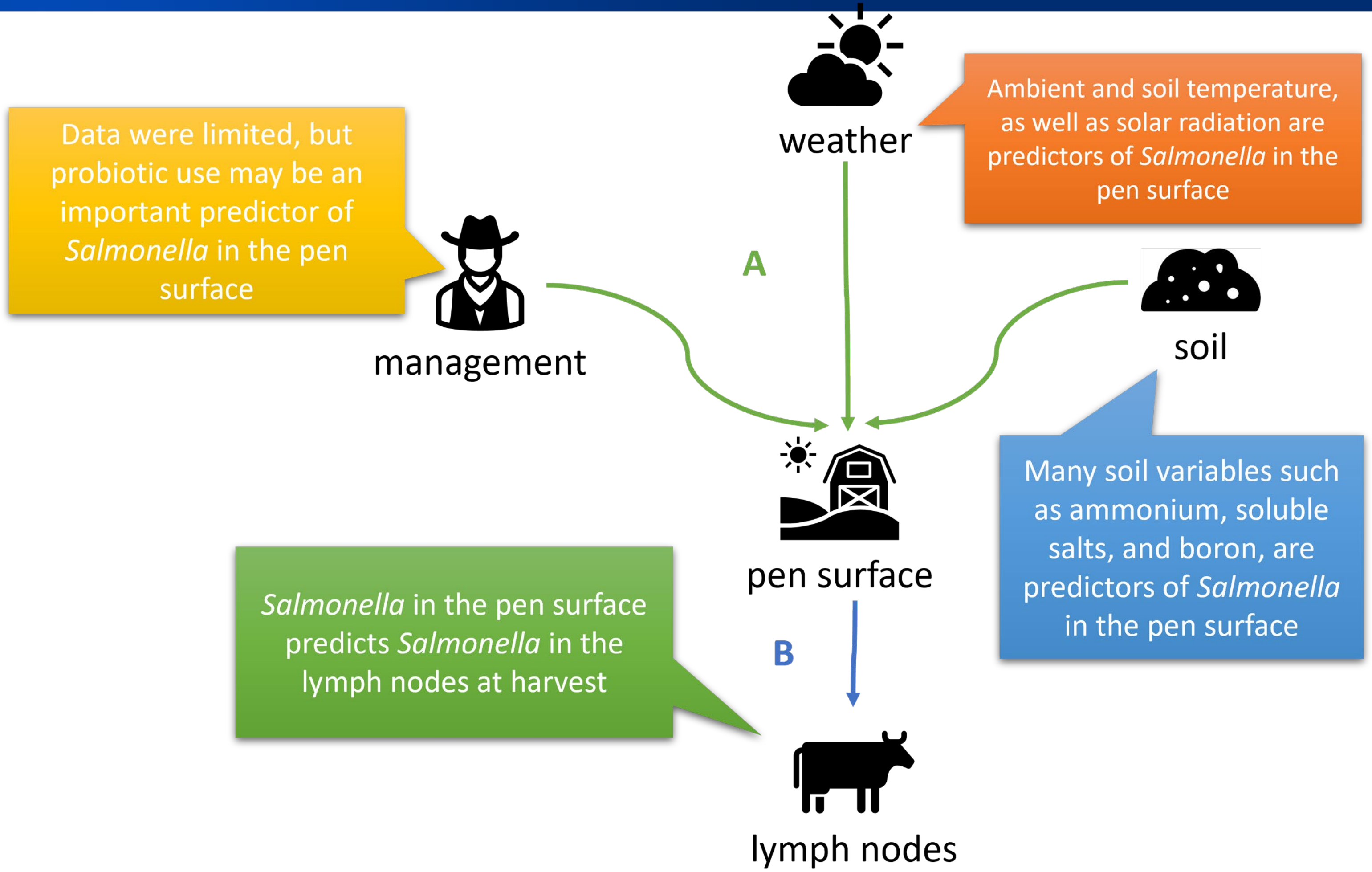




Collection
Cleaning
Exploratory Data Analysis
Model Building
Model Deployment

Pre-harvest predictors of post-harvest *Salmonella* contamination

Model Building



Data were limited, but probiotic use may be an important predictor of *Salmonella* in the pen surface

Ambient and soil temperature, as well as solar radiation are predictors of *Salmonella* in the pen surface

Many soil variables such as ammonium, soluble salts, and boron, are predictors of *Salmonella* in the pen surface

Salmonella in the pen surface predicts *Salmonella* in the lymph nodes at harvest

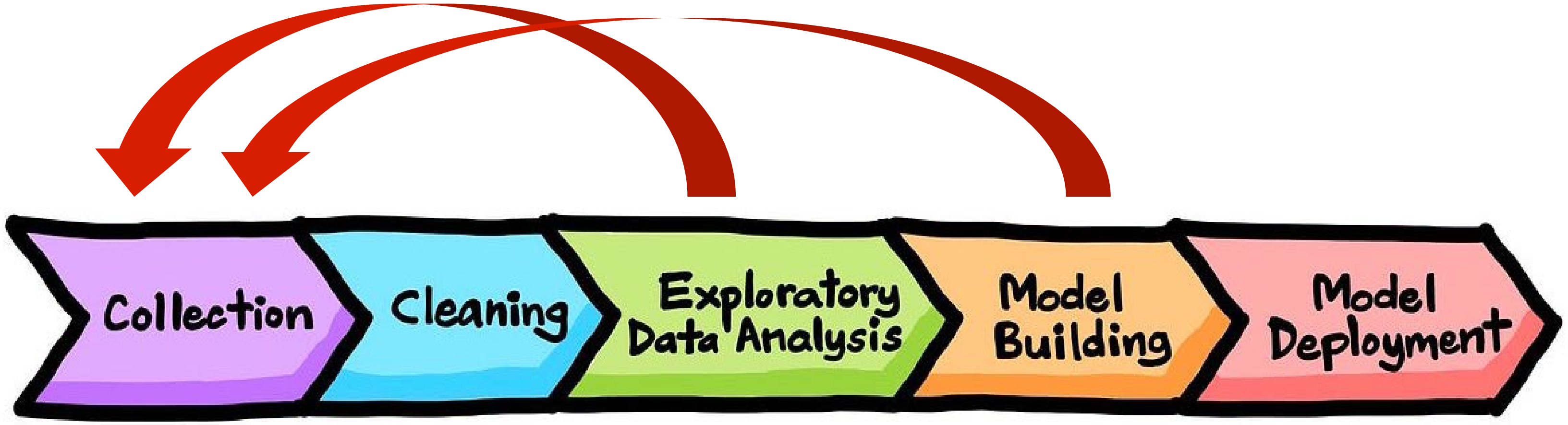
Pre-harvest predictors of post-harvest *Salmonella* contamination

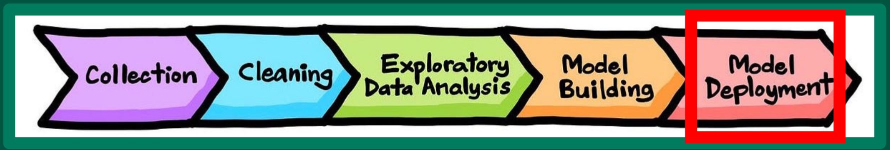
the process is not so linear!

Issues:

- Not enough data to model the entire system together
- Too few data to control for all confounding factors

Sometimes, we have to go back

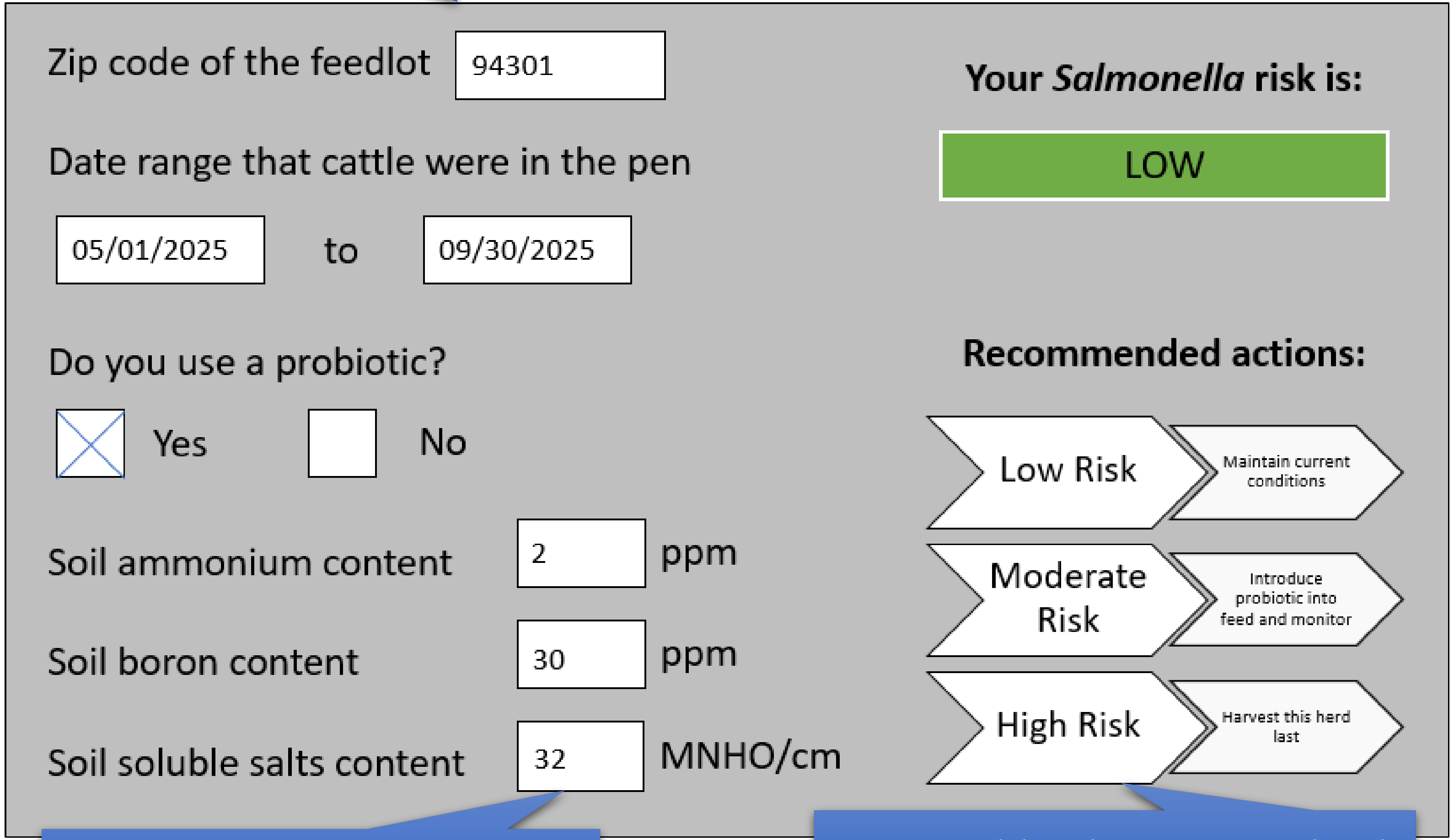




Pre-harvest predictors of post-harvest *Salmonella* contamination

Model deployment

Call data from weather database API based on user input



Zip code of the feedlot: 94301

Date range that cattle were in the pen

05/01/2025 to 09/30/2025

Do you use a probiotic?

☒ Yes □ No

Soil ammonium content: 2 ppm

Soil boron content: 30 ppm

Soil soluble salts content: 32 MNHO/cm

Your *Salmonella* risk is:

LOW

Recommended actions:

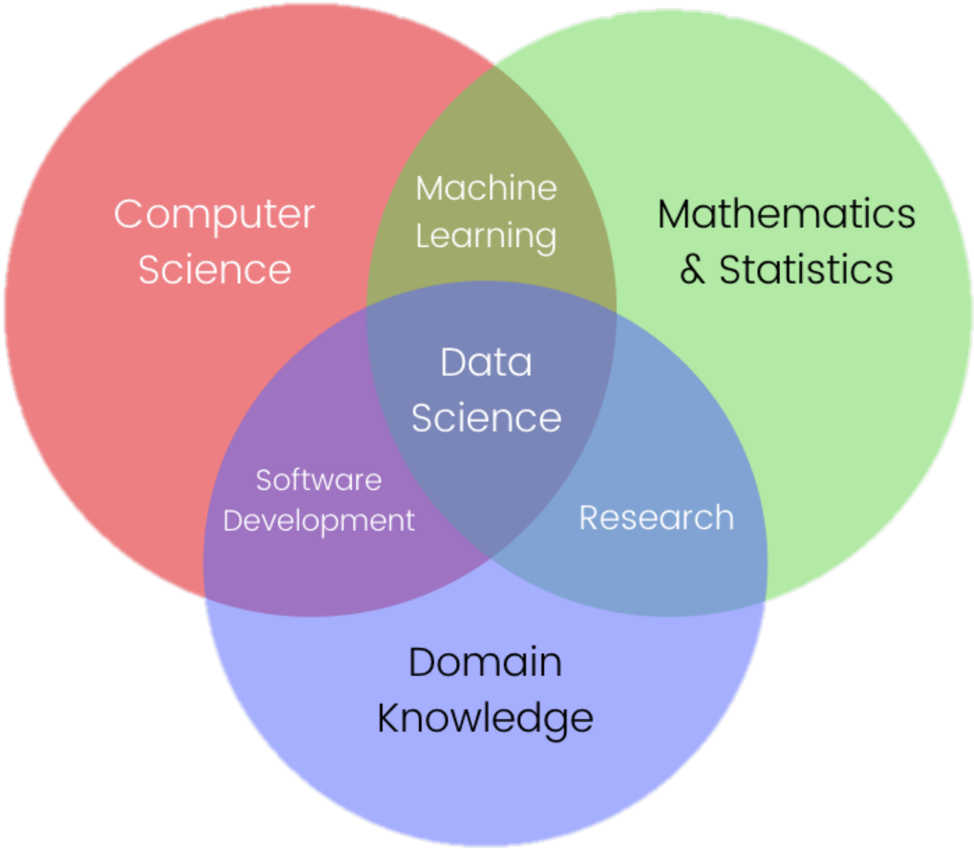
Low Risk → Maintain current conditions

Moderate Risk → Introduce probiotic into feed and monitor

High Risk → Harvest this herd last

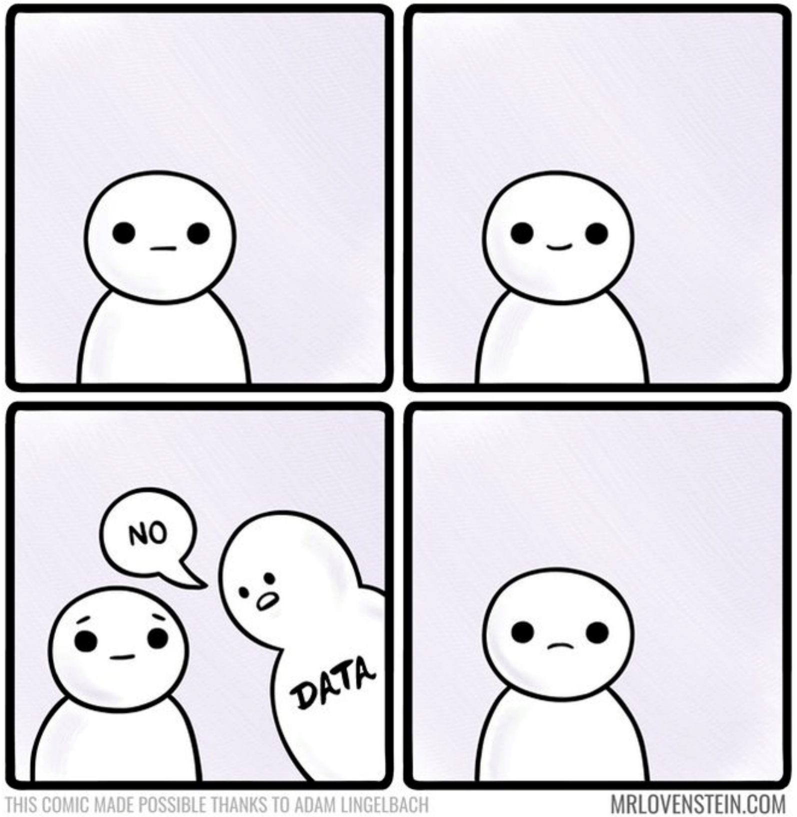
Calculate model predictions from user input

Convert model predictions to a simple red-yellow-green recommender system and provide insights to the user



The value add of data science

- **Domain knowledge** allows ease of communication between domain scientists and data scientists
- Traditional **math/stat background** ensures methods are quantitatively sound
- **Programming** creates reproducible results, development of software for non-programmers
- Ease of handling of ”big data”:
  - High volume
  - High diversity
  - High speed



Takeaways

- Data science combines mathematics, statistics, programming, and domain expertise to wrangle diverse data and produce novel insights with a focus on visualization and communication
- We used data science to identify pre-harvest predictors of post-harvest lymph node *Salmonella* contamination in cattle
- These findings will be used to collect more specific data to build and deploy a decision support tool to assist stakeholders
- The challenges we faced are not unique to our system
    - Low data
    - Diverse data
    - n vs p problem