Why Statistics is Not Data Science

Participation: www.StatsClass.org

Christopher Malone
cmalone@winona.edu

Tisha Hooks
thooks@winona.edu

WINONA
STATE UNIVERSITY

USCOTS 2017

QUESTION #1

Suppose you are designing a postcard to advertise a statistics major to prospective students. You want this postcard to include an answer to the question "What is statistics?" What would you write?

Task: Submit your answer at www.StatsClass.org.

CURRICULUM GUIDELINES FOR STATISTICAL SCIENCE

- Endorsed by the ASA in 2014
- Working group consisted of members of the ASA community (from both academia and industry)

Statistics Guidelines

GUIDELINES FOR STATISTICAL SCIENCE – KEY SKILLS

- Statistical Methods and Theory

  Study design, EDA, model building, inference

- Data Manipulation and Computation

  Use statistical software for data exploration, cleaning, and analysis;
  Think algorithmically and program in a higher-level language;
  Manage and manipulate data

- Mathematical Foundations

  Apply ideas from linear algebra, calculus, and probability

- Statistical Practice/Communication

  Write clearly, speak fluently, and create effective visualizations; Collaborate

- Discipline-Specific Knowledge

  Encourage study in a substantive area of application

Statistics Guidelines

GUIDELINES FOR STATISTICAL SCIENCE – KEY SKILLS

- Statistical Methods and Theory

  Study design, EDA, model building, inference

- Data Manipulation and Computation

  Use statistical software for data exploration, cleaning, and analysis;
  Think algorithmically and program in a higher-level language;
  Manage and manipulate data

- Mathematical Foundations

  Apply ideas from linear algebra, calculus, and probability

- Statistical Practice/Communication

  Write clearly, speak fluently, and create effective visualizations; Collaborate

- Discipline-Specific Knowledge

  Encourage study in a substantive area of application

Statistics Guidelines

GUIDELINES FOR STATISTICAL SCIENCE – KEY SKILLS

- Statistical Methods and Theory
  - Study design, EDA, model building, inference
- Data Manipulation and Computation
  - Use statistical software for data exploration, cleaning, and analysis;
  - Think algorithmically and program in a higher-level language;
  - Manage and manipulate data
- Mathematical Foundations
  - Apply ideas from linear algebra, calculus, and probability
- Statistical Practice/Communication
  - Write clearly, speak fluently, and create effective visualizations; Collaborate
- Discipline-Specific Knowledge
  - Encourage study in a substantive area of application

Statistics Guidelines

GUIDELINES FOR STATISTICAL SCIENCE – KEY SKILLS

- Statistical Methods and Theory

  Study design, EDA, model building, inference

- Data Manipulation and Computation

  Use statistical software for data exploration, cleaning, and analysis;
  Think algorithmically and program in a higher-level language;
  Manage and manipulate data

- Mathematical Foundations

  Apply ideas from linear algebra, calculus, and probability

- Statistical Practice/Communication

  Write clearly, speak fluently, and create effective visualizations; Collaborate

- Discipline-Specific Knowledge

  Encourage study in a substantive area of application

Statistics Guidelines

GUIDELINES FOR STATISTICAL SCIENCE – KEY SKILLS

- Statistical Methods and Theory

  Study design, EDA, model building, inference

- Data Manipulation and Computation

  Use statistical software for data exploration, cleaning, and analysis;
  Think algorithmically and program in a higher-level language;
  Manage and manipulate data

- Mathematical Foundations

  Apply ideas from linear algebra, calculus, and probability

- Statistical Practice/Communication

  Write clearly, speak fluently, and create effective visualizations; Collaborate

- Discipline-Specific Knowledge

  Encourage study in a substantive area of application

Question 2

QUESTION #2

Suppose you are designing a postcard to advertise a data science major to prospective students. You want this postcard to include an answer to the question "What is data science?" What would you write?

Task: Submit your answer at www.StatsClass.org.

Data Science Guidelines

CURRICULUM GUIDELINES FOR DATA SCIENCE

- Published in 2016
- Composed by participants of the Park City Math Institute (PCMI) 2016 Summer Undergraduate Faculty Program
- The group consisted of undergraduate faculty primarily from the disciplines of mathematics, statistics, and computer science
- Endorsed by the ASA Board of Directors

GUIDELINES FOR DATA SCIENCE – KEY SKILLS

- Statistical Thinking

  Understand data analysis, data collection, modeling, inference;
  Formulate problems, identify and gather relevant data, provide insights

- Computational Thinking

  Access and organize data from databases, scrape data from web, process text data;
  Possess foundational software/programming skills, think algorithmically;
  Use statistical software and understand underlying principles of programming in these packages

GUIDELINES FOR DATA SCIENCE – KEY SKILLS

- Statistical Thinking

  Understand data analysis, data collection, modeling, inference;
  Formulate problems, identify and gather relevant data, provide insights

- Computational Thinking

  Access and organize data from databases, scrape data from web, process text data;
  Possess foundational software/programming skills, think algorithmically;
  Use statistical software and understand underlying principles of programming in these packages

Data Science Guidelines

GUIDELINES FOR DATA SCIENCE – KEY SKILLS

- Statistical Thinking

  Understand data analysis, data collection, modeling, inference;
  Formulate problems, identify and gather relevant data, provide insights

- Computational Thinking

  Access and organize data from databases, scrape data from web, process text data;
  Possess foundational software/programming skills, think algorithmically;
  Use statistical software and understand underlying principles of programming in these packages

Understand connections between statistical and computational thinking;
Wide variety of problem-solving approaches; work with a diverse collection of tools;
Ability to learn new tools and adapt to changes

GUIDELINES FOR DATA SCIENCE – KEY SKILLS

- Mathematical Foundations

  Choose, fit, and use mathematical models;
  Demonstrate structured mathematical problem-solving;
  Understand issues of optimization/convergence of algorithms used in model building

- Model Building and Assessment

  Informal modeling

  Identify sources of variation; Discern between stochastic and deterministic variation; Understand how these might be modeled mathematically andcomputationally

  Formal modeling

  Build/assess statistical and machine learning models; Employ variety of formal inference procedures, draw appropriate conclusions; Understand how issues such as data collection and sources of bias impact analysis and conclusions; Bring computational considerations to data analysis, including issues of scale

GUIDELINES FOR DATA SCIENCE – KEY SKILLS

- Mathematical Foundations

  Choose, fit, and use mathematical models;
  Demonstrate structured mathematical problem-solving;
  Understand issues of optimization/convergence of algorithms used in model building

- Model Building and Assessment

  Informal modeling

  Identify sources of variation; Discern between stochastic and deterministic variation; Understand how these might be modeled mathematically andcomputationally

  Formal modeling

  Build/assess statistical and machine learning models; Employ variety of formal inference procedures, draw appropriate conclusions; Understand how issues such as data collection and sources of bias impact analysis and conclusions; Bring computational considerations to data analysis, including issues of scale

GUIDELINES FOR DATA SCIENCE – KEY SKILLS

- Algorithms and Software Foundation

  Employ algorithmic problem-solving skills in a high-level language;
  Understand memory and execution performance;
  Utilize best practices in documentation and program structure;
  Leverage existing tools to solve problems

- Data Curation

  Data Preparation
  Work with data from a variety of sources/formats;
  Prepare data for use and recognize how quality of data and means of collection may affect conclusions

  Data Management
  Ensure integrity of data while it passes through all stages of analysis;
  Work with relational databases

GUIDELINES FOR DATA SCIENCE – KEY SKILLS

- Algorithms and Software Foundation

  Employ algorithmic problem-solving skills in a high-level language;
  Understand memory and execution performance;
  Utilize best practices in documentation and program structure;
  Leverage existing tools to solve problems

- Data Curation

  Data Preparation

  Work with data from a variety of sources/formats;
  Prepare data for use and recognize how quality of data and means of collection may affect conclusions

  Data Management

  Ensure integrity of data while it passes through all stages of analysis;
  Work with relational databases

GUIDELINES FOR DATA SCIENCE – KEY SKILLS

- Knowledge Transference

  Apply discipline outside core of statistics, computing, and mathematics

  Communication

  Ability to use oral, written, and visual modes to communicate effectively to a variety of audiences

  Ethics and Reproducibility

  Understand ethical issues such as data ownership, data security, privacy concerns, transparency, and reproducibility

Question 3

DISCUSSION ITEM

- Review the responses to Questions 1 and 2. Discuss the similarities/differences in the two definitions and whether the suggested definitions align with the curriculum guidelines.

Question 3

QUESTION #3

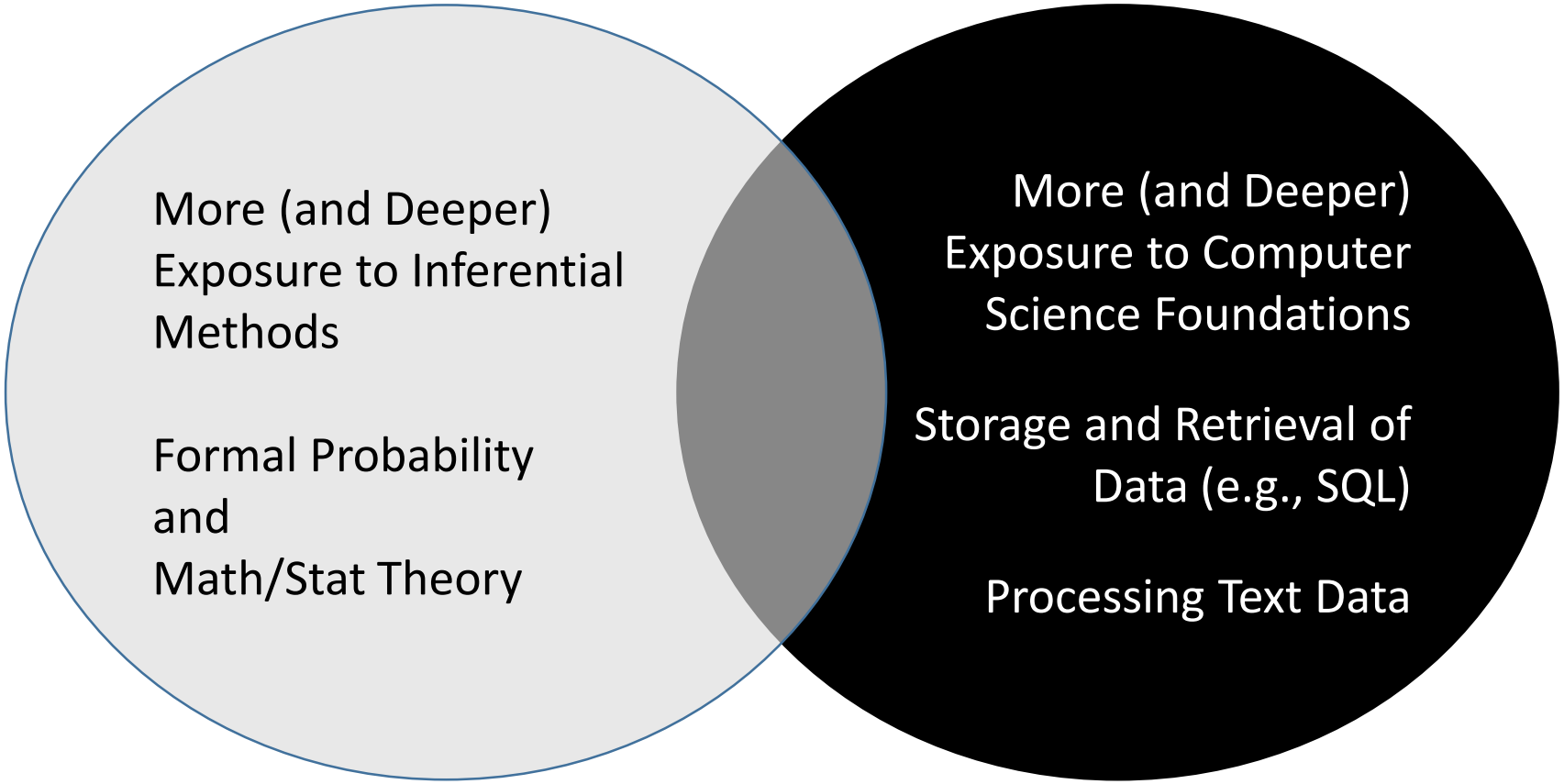
In your opinion, what percent of data science consists of statistics?

After making a submission, review the responses of others.

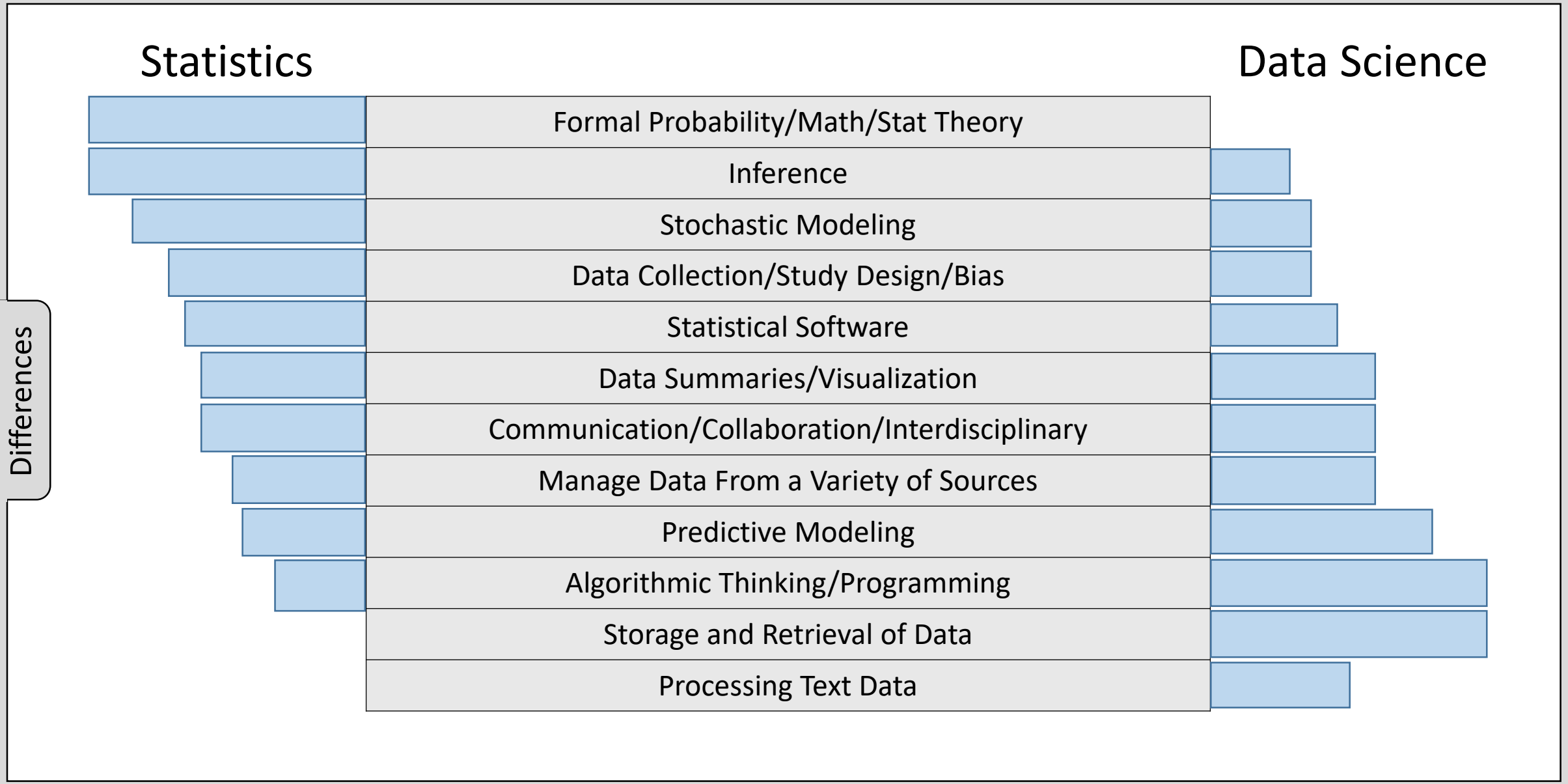
Why Statistics is Not Data Science

KEY DIFFERENCES IDENTIFIED FROM THE GUIDELINES

Differences



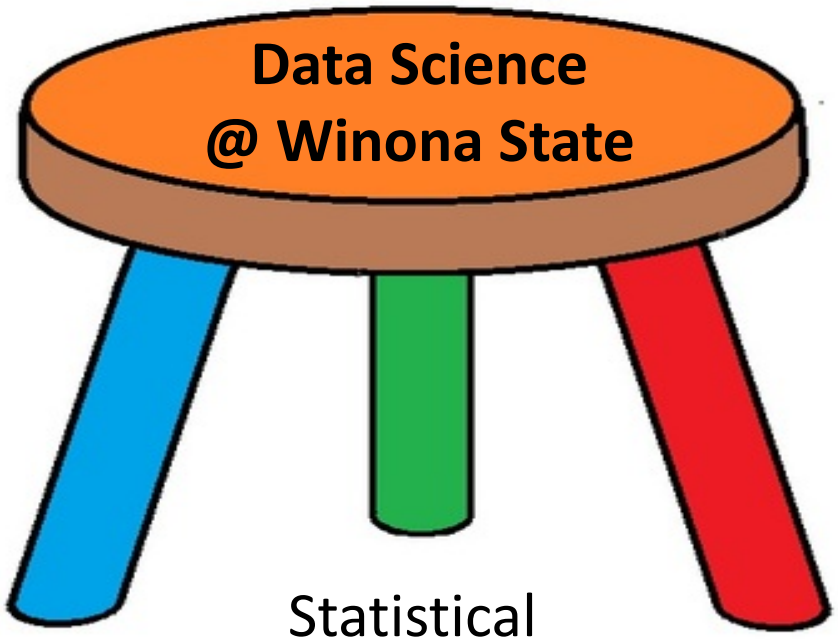
Why Statistics is Not Data Science



WSU Perspective

WHERE WE ARE AT WINONA STATE

- Data science is much more than statistics, but statistics has much to offer



Computational

Statistical

Interdisciplinary &
Professional Skill Development

MOVING FORWARD

- Data science curriculum guidelines are still evolving
- Committee on Applied and Theoretical Statistics (CATS) Roundtable Discussions are happening
- Statisticians, computer scientists, and members of other disciplines need to be involved
- Need to recognize that statistics and data science programs are distinct and require separate skills

Task #1 – Stop and Frisk

- Goal: Identify which New York City Police precincts are most unbiased (or biased) regarding the rate in which police "Stop/Question/Frisk" people.

QUESTION #4

What elements of Task #1 are statistical in nature?
Data science in nature?

After making a submission, review the responses of others.

Task #2 – Yelp Reviews

- Goal: Quantify the positivity/negativity of Yelp reviews for a restaurant in State College. Correlate these measures to the star rating.

QUESTION #5

What elements of Task #2 are statistical in nature?
Data science in nature?

After making a submission, review the responses of others.

FINAL THOUGHTS

FINAL THOUGHTS

FINAL THOUGHTS

FINAL THOUGHTS

Why Statistics is Not Data Science

Christopher Malone
cmalone@winona.edu

Tisha Hooks
thooks@winona.edu

The End