Introducing SCORE:

Sports Content for Outreach, Research, and Education

Why Sports?

Data availability

Lots of **public data**, freely available to anyone

Wide **variety** of data, problems, and methods

Popularity

Sports are widely **popular**. In 2019, 154.4 million U.S. viewers watched live sports at least once per month

Many students start as **subject-matter experts**.

Real-life **validation**

Transferability

Problems are **analogous** to those in non-sports applications.

Experience in sports **translates** to other fields

Sports are a **controlled** environment. Potentially easier to start with

Examples of Sports Analytics Problems

Teams

- Player personnel decisions, evaluate player performance (trades, FA signings)
  - How much is a player contributing to his/her team in terms of goals/points/runs?
  - How much is a player worth on the open market in terms of salary cap dollars?
  - How much is a player worth to us?
- Coaching decisions
  - Should we bunt? go for it on 4th down?
  - When should we pull our goalie?
  - Given locations of the players and ball/puck, what do we expect to happen? What is the best decision?
- Draft decisions
  - Who should we draft with our next pick?
  - Who will be available to draft?
  - How much is our draft pick worth?

Media - Talk about those decisions, team ratings and playoff probabilities
Betting - Many of the above problems, or variants of them
League - What realignment would minimize travel? What schedule would max. fairness and min. travel?
Business - Predicting demand for a game based on day, month, opponent, etc

Learning Objectives

Most of those can be answered reasonably with undergraduate level statistics and data science tools.

Students get experience with

- Solving real problems
- Joining data from multiple sources, working with several different types of data
- Data exploration/visualization
- Multivariable thinking, need for regression or something else
- Modeling
- Interpretation
- Etc

Most of the data is publicly available, or can be done with public available alternatives

Exception: player tracking data

Sports Analytics in Education

Dozens/hundreds of educators around the world have developed

- hundreds of examples,
- using multiple sports, and
- focusing on a variety of statistics and data science topics

Missing:

- Standardization
- Completeness
- Consolidation/centralization
  - Content creation is decentralized = good
  - Content is decentralized = less convenient
- Industry/media perspective

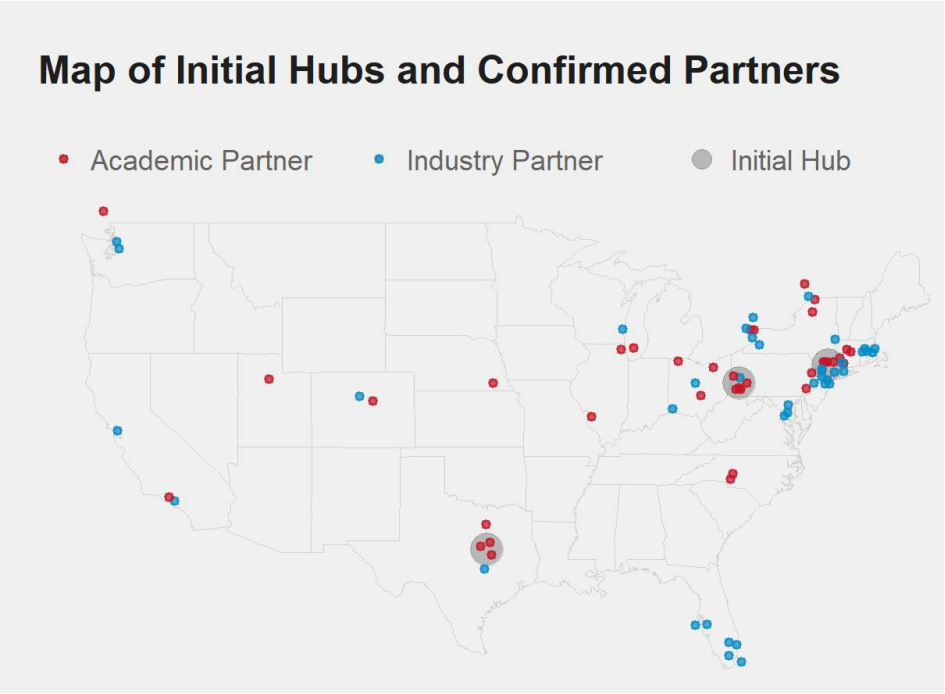
Enter SCORE!

SCORE with Data: building a **sustainable national network** for **developing and disseminating Sports Content** for Outreach, Research, Education in data science

Unique NSF-funded project combining academia, professional sports, and media to build a repository of educational materials for statistics & data science via sports applications/analytics

Strong emphasis on outreach, inclusiveness, and building pipelines

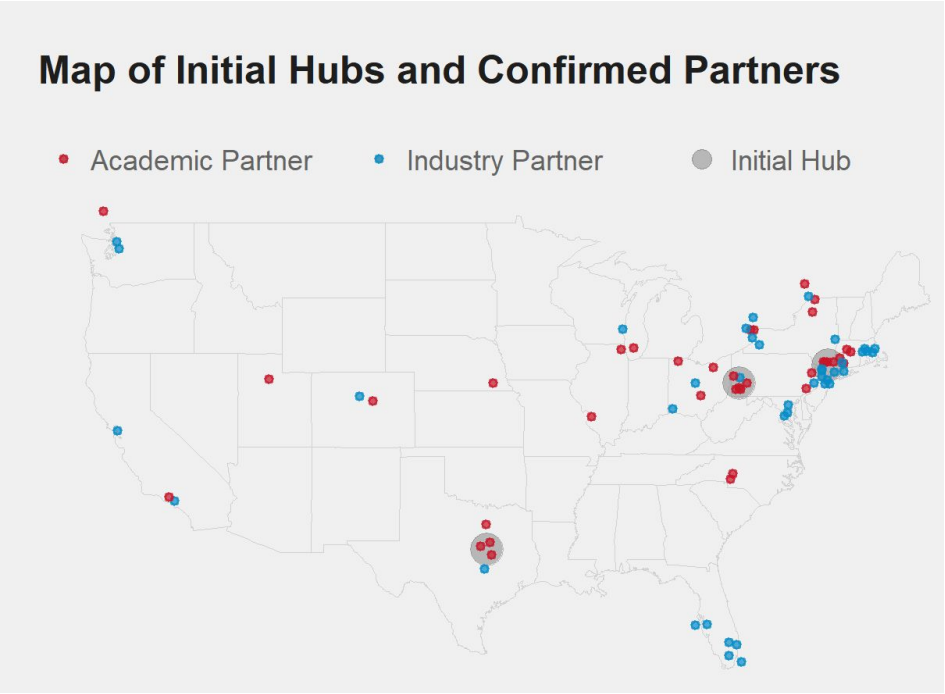
Sustainable National Network



Why a network?

- Decentralized/crowdsourced content creation
  - Ideally sustainable
- Diverse experience and skills
  - Academia, industry and media
  - Key part of review process
- Pre-existing informal network already

Sustainable National Network



Initial Network

- Over 80 initial confirmed partners in academia, industry and media
- 3 initial hubs
  - PIT - Carnegie Mellon, Pitt
  - NY - St Lawrence, West Point, Yale
  - TX - Baylor
- High level advisory board from academia and industry

SCORE with Data

Building a **sustainable national network** for **developing and disseminating Sports Content** for Outreach, Research, Education in data science

SCORE with Data

Building a **sustainable national network** for **developing and disseminating Sports Content** for Outreach, Research, Education in data science

Developing Content

Modules: described in detail later in the session

Goal: relevant, self-contained, plug and play

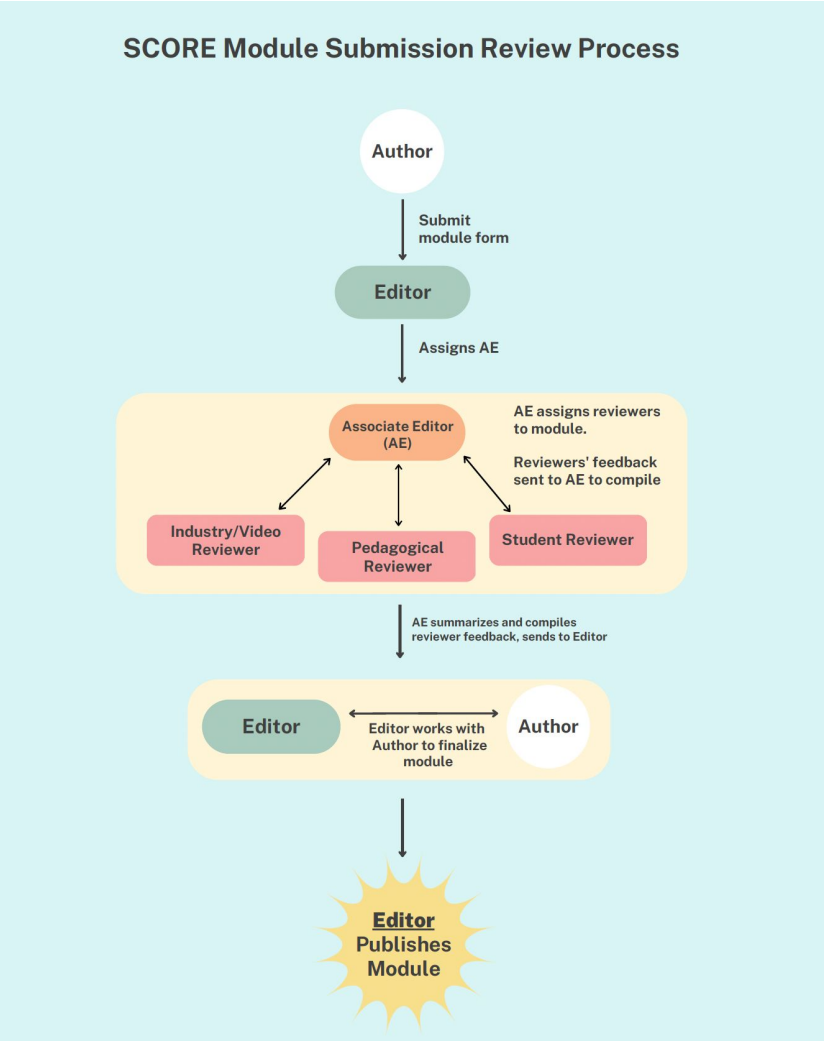
- Introductory motivation videos/content from sports professionals (athletes, analysts, management)
- Learning Objectives
- Data sets
- Lecture notes, handouts, activities, slides
- Multiple formats, languages: interactive/no code (ISLE, Excel, Minitab), R, Python, etc
- Modules can be downloaded or directly accessed through the SCORE website (e.g., can do data analysis online)
    - Support research on how people engage with and collaborate on data science
    - Use metrics available for educators/creators. Citations supported

Developing Content

Review process goal: use the diverse experiences and skills of the network to provide feedback from multiple perspectives to ensure

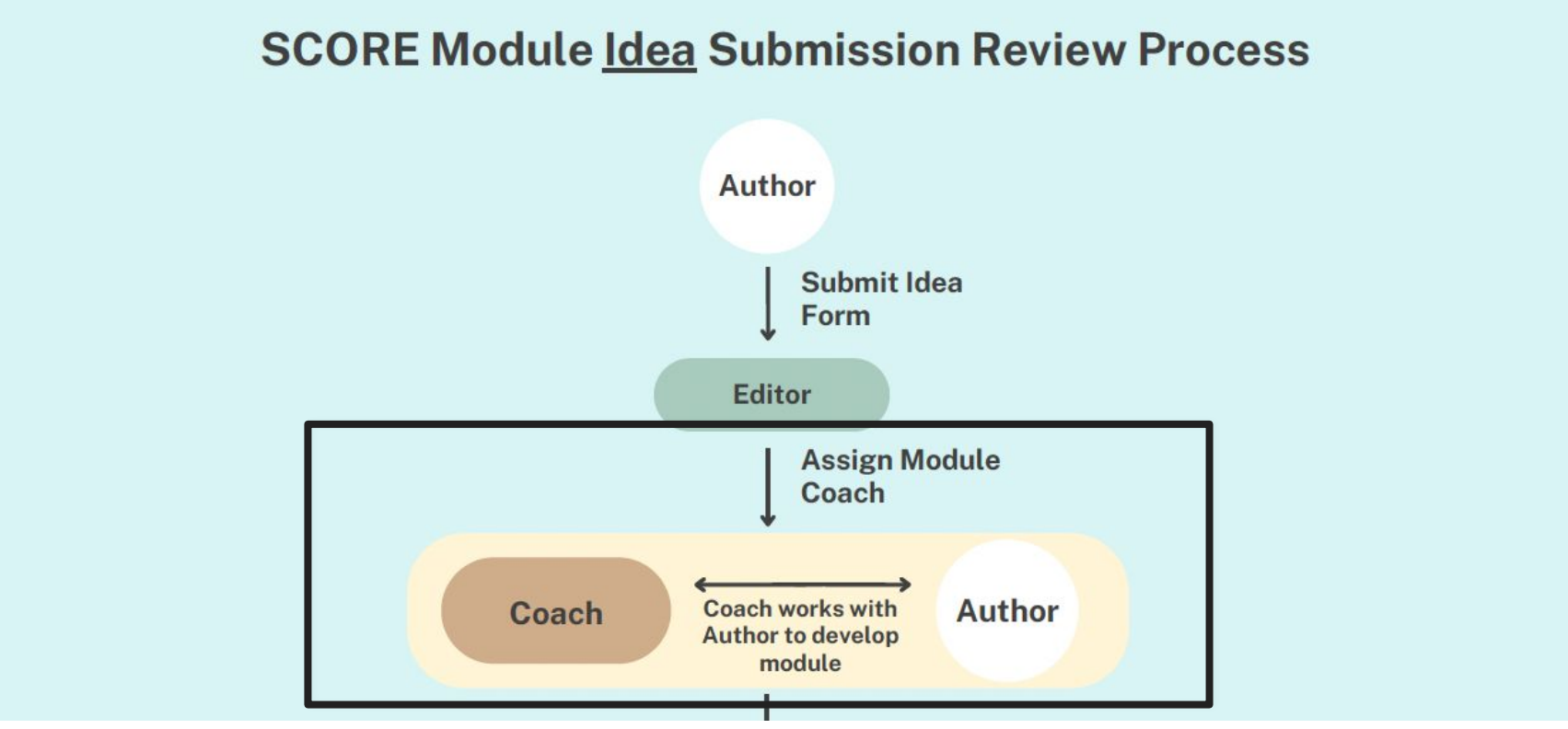
- Industry relevance
- Statistical and educational best practices
- Standardization

Non-adversarial



Developing Content

Can start with an idea, not a full module, and can be assigned a coach



SCORE with Data

Building a **sustainable national network** for **developing and disseminating Sports Content** for Outreach, Research, Education in data science

SCORE with Data

Building a **sustainable national network** for **developing and disseminating Sports Content** for Outreach, Research, Education in data science

Disseminating Content

ScoreNetwork.org

SCORE Module and Data Repository

Modules can be found at:

https://modules.scorenetwork.org/

Data can be found at:

https://data.scorenetwork.org/

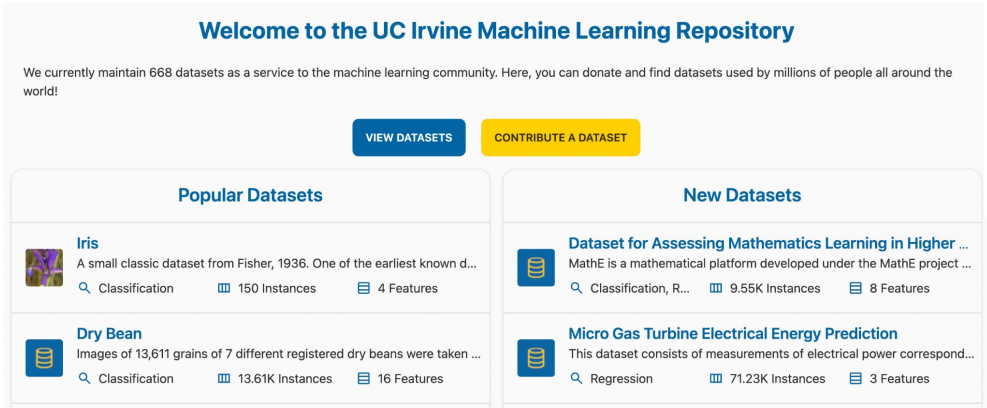
Module

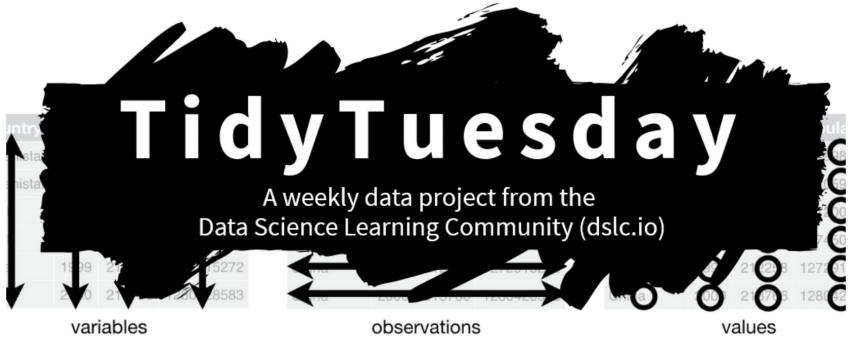
Educational materials for statistics and data science

SCORE Sports Data Repository

Ron Yurko
Assistant Teaching Professor, Department of Statistics & Data Science
Carnegie Mellon University

Datasets for class?





kaggle

CMU S&DS Data Repository

The Data Repository curates interesting datasets for use in statistics and data science education. Each dataset is supported by a *story* describing its origin and application, and a set of interesting *questions* that can be answered using the data. This means:

- Every dataset has context in a scientific field, pop culture, or daily life.
- Beyond context, datasets are interesting. They feature more than just a dozen observations from an antiquated scientific study — many feature thousands of observations of dozens of variables, and answer questions interesting to a wide audience.
- Just like in science, some datasets give null results.
- Instructors can easily build lessons and assignments from the suggested questions.

Datasets are organized by broad subject areas on the left, or you can browse a sortable list of all datasets.

Data Is Plural

... is a weekly newsletter (and seasonal podcast) of useful/curious datasets, published by Jeremy Singer-Vine. There have been 381 editions, dating from October 21, 2015 to July 31, 2024. To receive future editions, sign up here:

Check out: https://cmustatistics.github.io/data-repository/

What about sports datasets?

SCORE NETWORK

SPORTS REFERENCE

SportsDataverse

An open-source sports analytics and data organization.

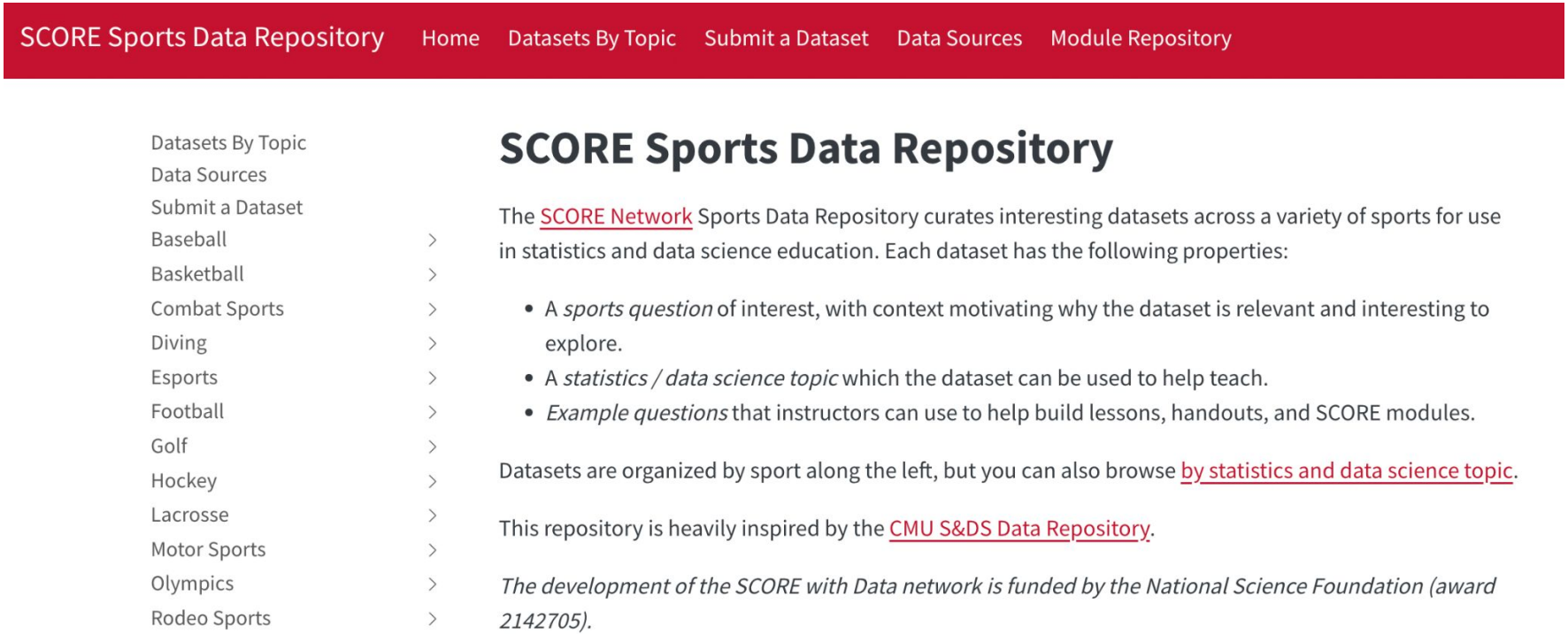
We provide utilities in Python, R, Node.js, etc.

CRAN Task View: Sports Analytics

Maintainer:	Benjamin S. Baumer, Quang Nguyen, Gregory J. Matthews
Contact:	ben.baumer at gmail.com
Version:	2023-04-06
URL:	https://CRAN.R-project.org/view=SportsAnalytics
Source:	https://github.com/cran-task-views/SportsAnalytics/

SCORE Data Repository

https://data.scorenetwork.org/

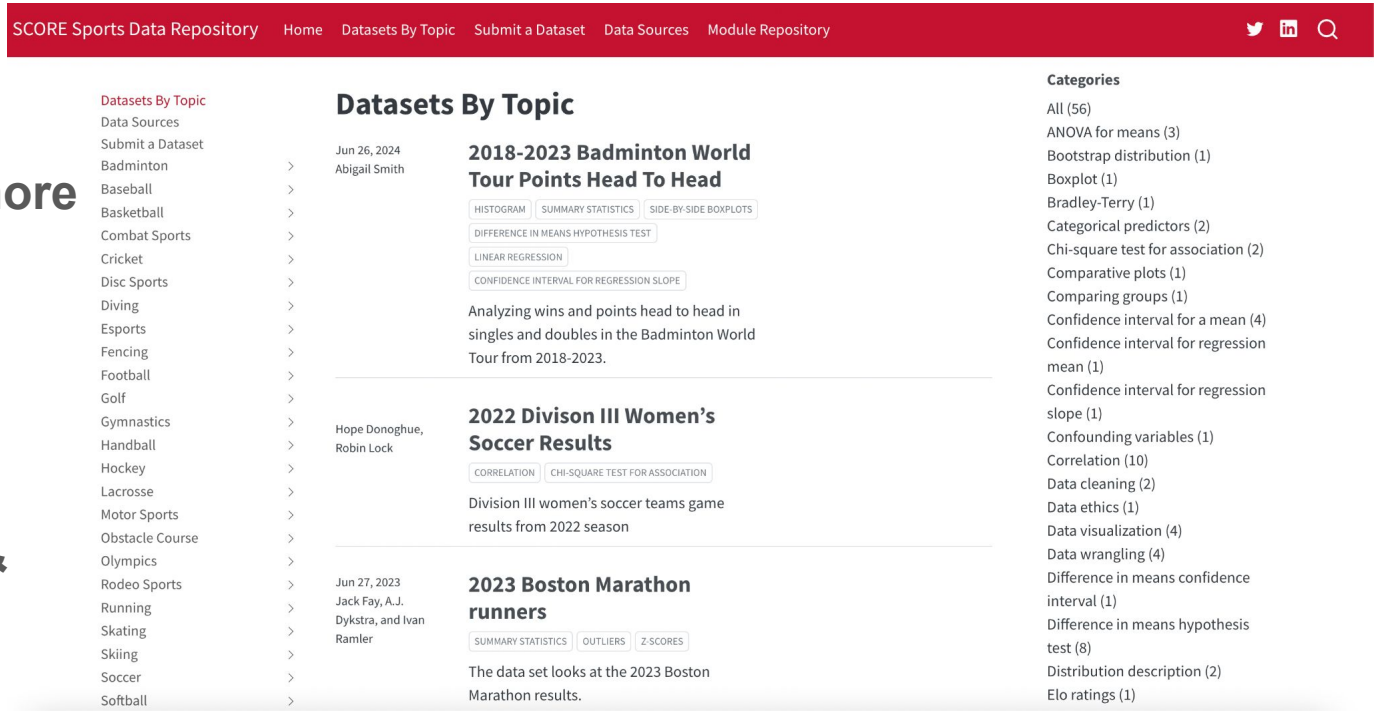


SCORE Data Repository

https://data.scorenetwork.org/

Features over 70 datasets across more than 30 different sports

Can search for datasets by sport and by statistics & data science topic



What does every dataset include?

- **Descriptive title**, e.g., Women's National Basketball Association Shots
- **Brief description the data background**, potentially include a brief summary of the sports problem and statistical situation
- List of **relevant statistics and data science topics/categories** for tags
- Information about **the source of the data motivating its usage**
- **Description of dataset**: what does one row represent? README file serving as the data dictionary for the columns
- **Example questions** associated with the data

Example dataset page information

Basketball > Women's National Basketball Association Shots

Women's National Basketball Association Shots

CLASSIFICATION | LOGISTIC REGRESSION | GENERALIZED ADDITIVE MODELS | MULTINOMIAL LOGISTIC REGRESSION | NAIVE BAYES CLASSIFIER | DENSITY ESTIMATION

Information about shots during the 2021-2022 WNBA season

AUTHOR
Ron Yurko

PUBLISHED
March 25, 2023

Motivation

The Women's National Basketball Association (WNBA) is the top professional women's basketball league in the world. The league records every shot players take along with contextual information about the shot such as its location, a description of the shot type, as well as the outcome. With this dataset, you can predict the success of each shot attempt to compute the expected value of shot types and compare team decision making.

Questions

1. Build a classification model to predict the shot outcome based on the spatial x,y coordinates of the shot.
2. Create a visualization displaying the joint frequency of shot locations. Do there appear to be any clear modes of frequently taken shots? Create a conditional version of this display by shot outcome. Does the distribution shape vary by shot outcome? (You can also perform a similar analysis by team and shot type).

Data

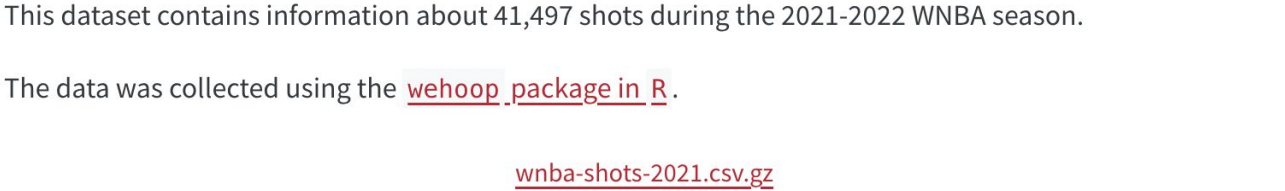
This dataset contains information about 41,497 shots during the 2021-2022 WNBA season.

The data was collected using the `wehoop` package in `R`.

Variable	Description
game_id	Unique integer ID for each WNBA game
game_play_number	Integer indicating the recorded play number for the shot attempt, where 1 indicates the first play of the game
desc	String detailed description of shot attempt
shot_type	String description of the shot type (e.g., dunk, layup, jump shot, etc.)
made_shot	Boolean denoting if the shot was made (TRUE) or not (FALSE)
shot_value	Numeric value of the shot outcome (0 for shots that were not made, and a positive value for made shots)
coordinate_x	Horizontal location in feet of shot attempt where the hoop would be located at 25 feet
coordinate_y	Vertical location in feet of shot attempt with respect to the target hoop (the hoop should be a little in front of 0 but the coordinate system is not exact)
shooting_team	String name of the team taking the shot

Dataset requirements

- **The data must be publicly shareable**
- **If you used code (such as an `R` package) to access the data then include your code with your submission.** The code will be publicly available on the repository's GitHub page for others to view.
- **There should be an interesting sports question coupled with a statistics and data science topic to motivate the use of the dataset in educational material.**
- **The dataset should be in a standard format such as a CSV file and be of reasonable size.** GitHub has a file size limit of 100 MB, and large files can be inconvenient for students. We recommend compressing files larger than a few megabytes. Note that gzip compression is a good choice since common tools such as `R` and `Python` feature ways to read `.csv.gz` files directly.
    - e.g., WNBA shot data was compressed and is available to download as .csv.gz



How to submit? (1) Google form

Please enter your name. As the dataset submitter, we will contact you at the email address provided above with any relevant questions. *

Your answer

Please provide a descriptive dataset title: *

Your answer

List all authors contributing the dataset, separated by commas (e.g., Author 1, Author 2): *

Your answer

Provide a brief one- or two-sentence description of the data. If possible, include a brief summary of the sports problem and statistical situation in this short description. *

Your answer

Upload a README file with a description of the columns in the dataset. This file should either be either an Excel file, CSV, or txt format with a row for each variable/column in the dataset and two columns explaining these variables/columns: (1) Variable - the column names, and (2) Description - a description of each variable, including units when possible. See other datasets on the website for example data descriptions. *

Add file

Subset the dataset file. The dataset should be in a standard format such as an Excel or CSV file and be of reasonable size. GitHub has a file size limit of 100 MB, and large files can be inconvenient for students. We recommend compressing files larger than a few megabytes. Note that gzip compression is a good choice since common tools such as R and Python feature ways to read .csv.gz files directly. *

Add file

(OPTIONAL) Submit code file used to prepare data that will be publicly available on GitHub.

Add file

How to submit? (2) Quarto template

This data repository is built using quarto and rendered into a website. You can get the template file in two ways:

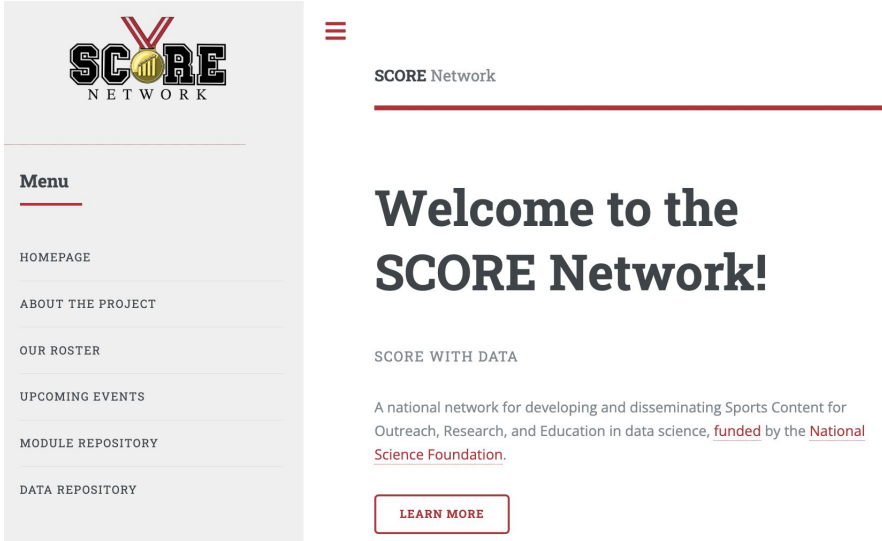
1. Copy the `_dataset-template.qmd` file from our GitHub repository and save it on your computer. Once you're done, you can email us the file and the data.

2. Fork our GitHub repository into your own GitHub account and edit it like any other Git repository. Once you're done, you can submit a pull request.

```
---
title: A descriptive dataset title
author: Your Name
date: Today's Date (e.g., April 17, 2023)
description: A one- or two-sentence description of the data. If possible, give a brief summary of
categories:
  - list the relevant
  - statistical methods
  - that can be used
  - with this dataset
  - one per line
  - with two spaces and a hyphen in front
---

## Motivation

The categories above determine how this dataset is listed on the [datasets by
methods](https://data.scorenetwork.org/by-statsds-topic.html) page.
Consult that page for a list of statistical categories already used by other
datasets.

In this first section, describe the source of the dataset and what it's about.
Give any necessary background about it and the sports research question of interest. See other
datasets on the website for examples.

This file is Markdown, so you *can* use formatting; [here is a guide to the
basics](https://quarto.org/docs/authoring/markdown-basics.html).

## Data
```

Thanks for your attention, any questions?

Building a sustainable national network for developing and disseminating Sports Content for Outreach, Research, Education in data science



- Email scorenetworkorg@gmail.com to join d-list, get more info, etc
- Workshops, trainings, network events for educators
- Data sets and educational materials available
- Connecting students with statistics & data science through sports and sports analytics
- Researching how students engage with data science

https://scorenetwork.org/

Module Submission & Review Process

- Module Components and Choices
- Submission Types
  - Idea vs Completed Module
- Role of Associate Editors and Editors
- Reviews and Review Types
- Miscellaneous Comments and Support

Module Components

Learning Goals

Introduction/Motivational Video

Methods (Statistical/Data Science)

Exercises/Activities

Conclusion

Dataset(s)

Data Glossary

ReadMe file

Module Choices

Sport(s)

Topic = Stat or Data Sci Content

Level (including Pre reqs)

Language

Learning Goals

Data and Data Glossary

Methods

Activities

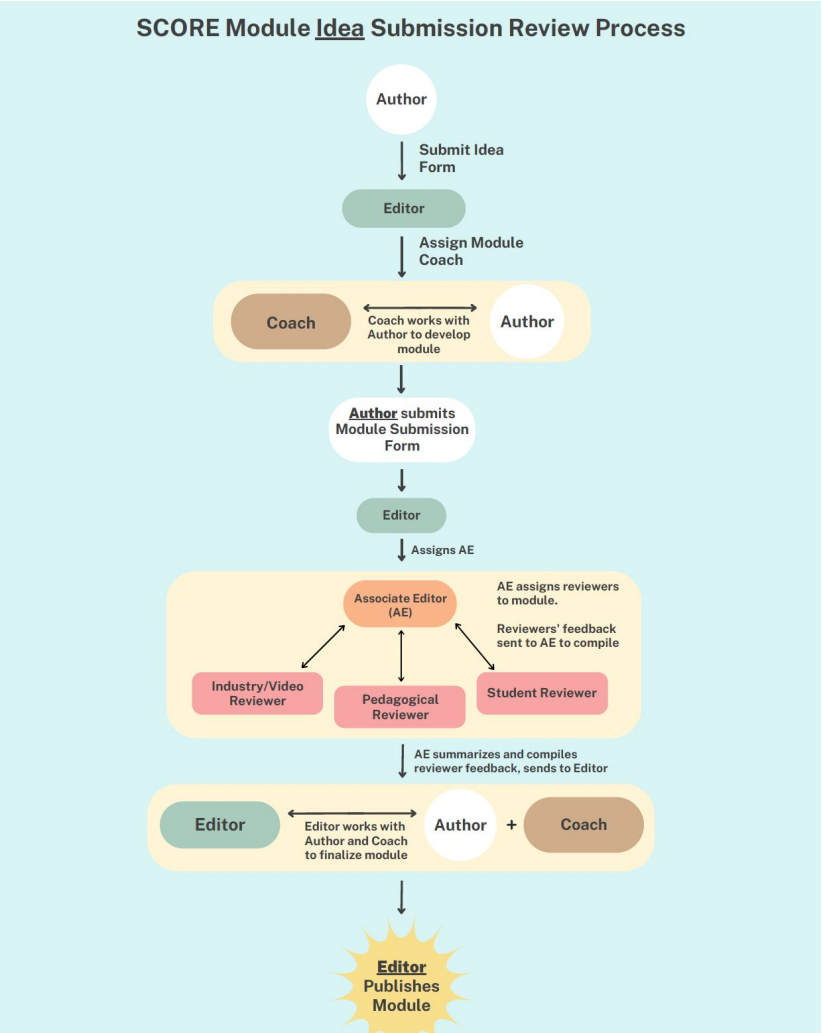
Conclusions

Submission Types

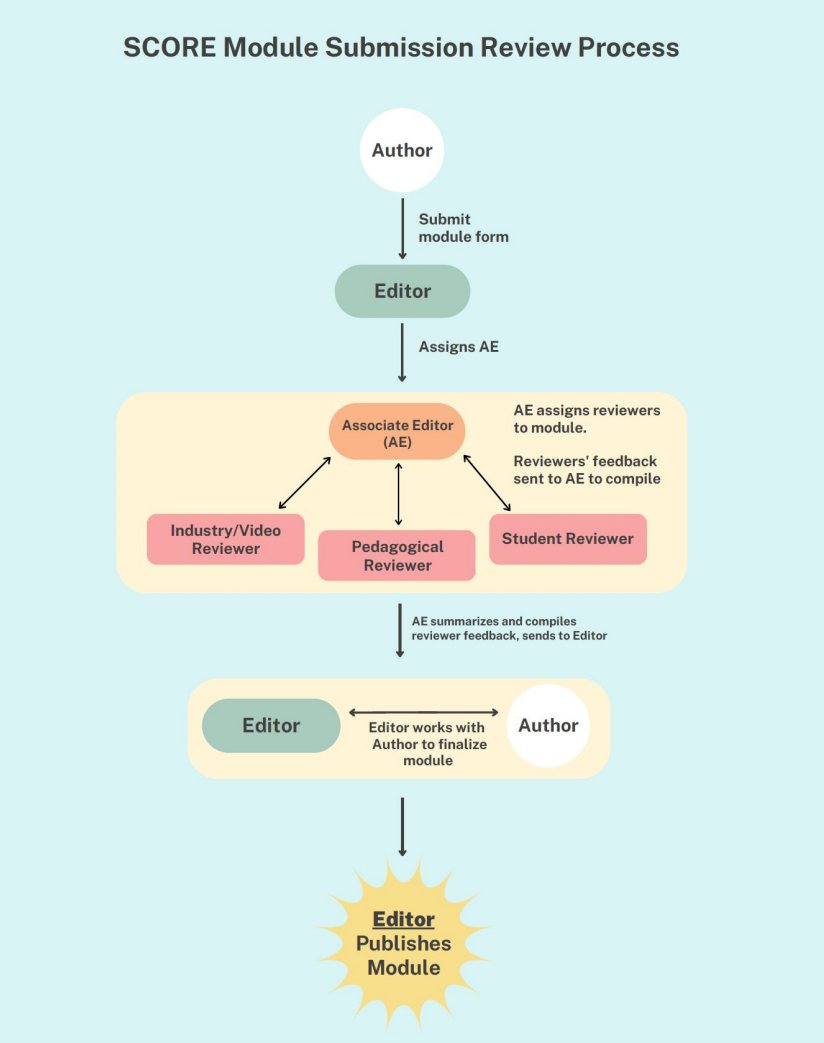
Idea

Complete or Nearly Complete Module

Module Idea


SCORE Module Idea Submission Review Process
Author
Submit Idea Form
Editor
Assign Module Coach
Coach
Coach works with Author to develop module
Author
Author submits Module Submission Form
Editor
Assigns AE
Associate Editor (AE)
AE assigns reviewers to module.
Reviewers' feedback sent to AE to compile
Industry/Video Reviewer
Pedagogical Reviewer
Student Reviewer
AE summarizes and compiles reviewer feedback, sends to Editor
Editor
Editor works with Author and Coach to finalize module
Author
+
Coach
Editor Publishes Module

Completed Module



Editor Role

Evaluate submission initially for

- Completeness
- Appropriateness

Find an AE

…

Assess AE & Reviewer recommendations

Make Decision on Publication

Work with Authors to get to publishable

GOAL: Publish as much as we can

Associate Editor Role

Evaluate module submission for appropriateness

Find Reviewers

- Pedagogical
- Industry
- Student (optional)

…

Make recommendation to Editor based upon Reviews

Reviewer Types

- Pedagogical
  - Learning Objectives, Alignment with Content, Motivation, Best Practices, Data and Documentation, Terminology, Inclusivity
- Industry
  - Connection and Application to Sports, Sports Question and/or Motivation, Video Review, Terminology/Clarity
- Student *(optional)*
  - Knowledge needed, Clarity, Time Taken, Connection to Sports, Tone, Interest, Engagement

Published Modules

- Video release form for any third parties (e.g. coach/player)
- Data sets must be able to be posted publicly
  - Author must attest during the submission process
- Author(s) will be given public attribution
  - Citations will be promoted and encouraged
  - Metrics will be available regarding use, downloads, etc

Credits

Editors: Rebecca Nugent, Michael Schuckers

Assoc Editors:

Nick Clark, Peter Freeman,

Andy Lee, Robin Lock,

Brian Macdonald, Josh Patrick,

Kostas Pelechrinis, Ivan Ramler,

Michael Schuckers, Rod Sturdivant,

Ron Yurko,

Admin: Sam Neilsen, Philipp Burckhardt *

Conclusions

- SCORE Network is awesome (and growing)
- Roles of Reviewers, Assoc Editors, Editors
- Foci for modules content and expecations
- Process of Peer Vetting

SCORE Network

Please Join Us: