An Academic and Industry Partnership Training the Next Generation of Data Scientists

Nicholas J. Horton

Department of Mathematics and Statistics
Amherst College, Amherst, MA, USA

JSM2020, August 6, 2020

Slides at https://github.com/Amherst-Statistics/JSM2020

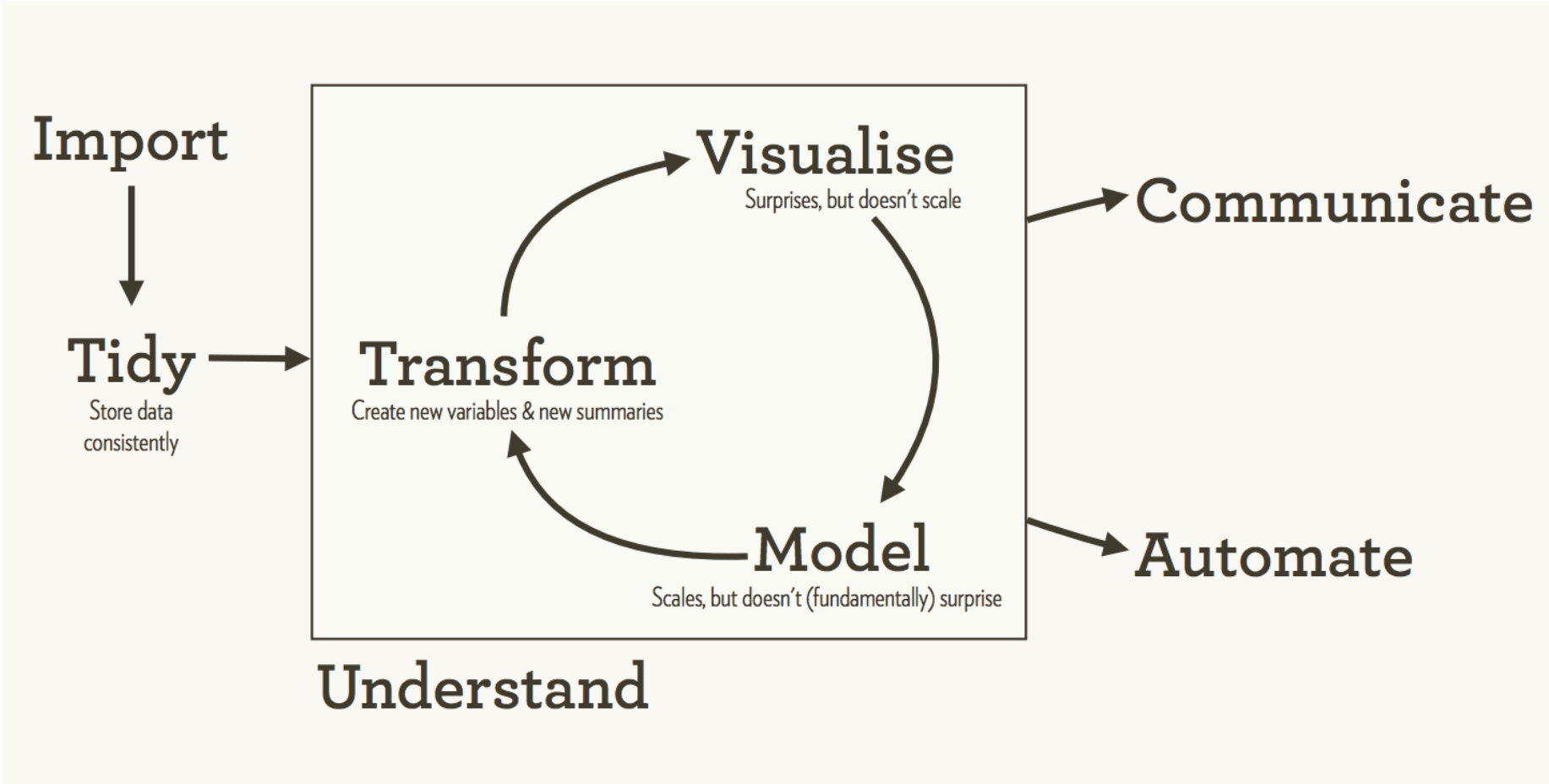
nhorton@amherst.edu

ASA Undergrad Guidelines for Statistics Programs, 2014

We are concerned that many of our graduates do not have sufficient skills to be effective in the modern workforce. Thomas Lumley (personal communication) has stated that our students know how to deal with $n \to \infty$, but cannot deal with a million observations.

If statistics is the science of learning from data, then our students need to be able to "think with data" (as Diane Lambert of Google has so elegantly described).
- Horton and Hardin (TAS, 2015)

Data analysis cycle (Wickham and Grolemund)



Motivation

Question: How can we ensure that our students have the skills they need to succeed in the real world?

Motivation

Question: How can we ensure that our students have the skills they need to succeed in the real world?

Answer: Develop partnerships between academia, industry, and government!

2019 SPAIG Award to the Five Colleges and MassMutual

This public-private partnership that has promoted the rapid growth of the statistics and data science community in Western Massachusetts (Pioneer Valley) in the last five years.

Overview of the partnership

Academic institutions

- Smith College (Ben Baumer)
- Mount Holyoke College (Andrea Foulkes)
- Amherst College (Nicholas Horton and Amy Wagaman)
- Hampshire College (Ethan Meyers)
- University of Massachusetts-Amherst (Brant Cheikes, Krista Gile, and Nick Reich)

Industry MassMutual Financial Group (Sears Merritt and Christine Pfeil)

Overview of the partnership (more later from Sears and Christine)

- Ten-year, $15 million grant from MassMutual to the UMass Center for Data Science to hire new faculty and launch a new graduate concentration in data science
- Four-year, $2 million women in data science grant from MassMutual to MHC and Smith colleges used to hire new faculty and launch majors in statistics and data science
- MassMutual Data Science Development Program (DSDP), an innovative three-year work-study training program that recruits recent Five College graduates, hires them as junior data scientists, and enrolls them in graduate programs at UMass
- Sponsorship of the ASA Five College DataFest by MassMutual

Where is the Pioneer Valley?

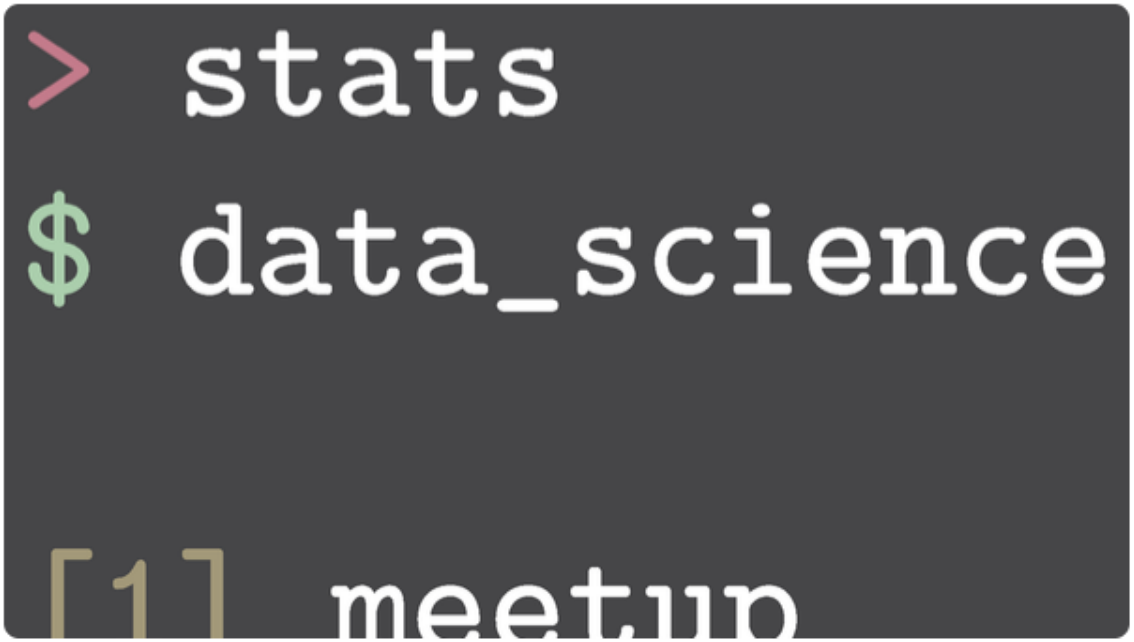
Overview of the Five Colleges

- Four liberal arts colleges (Amherst, Hampshire, Mount Holyoke, and Smith Colleges) and large public R-1 university (University of Massachusetts/Amherst)
- Founded in 1965 (as the Four Colleges Consortium, Hampshire founded in 1968)
- Facilitates cross-registration, free buses, shared libraries, faculty development

ASA Five College DataFest

ASA Five College DataFest

- Undergraduate students work in teams over a weekend to extract meaning from a complex dataset
- Judges and consultants from government and industry
- Co-sponsored by MassMutual



Western Mass Statist and Data Science

Northampton, MA
529 members · Public group
Organized by **Nicholas H.** and **10 others**

Share:

About Events Members Photos Discussions More

Join this group

What we're about

This is a Statistical and Data Science user group for the Pioneer Valley and Five College Area.

We were founded in January, 2013 originally as an R users group. However, we have become more of an all-purpose Statistics and Data Science meetup group, hosting and cross-listing public talks and events related to these fields.

Organizers

Nicholas H. and **10 others**
Message

Members (529)

Monday, January 6, 2020

Research Bytes at MassMutual: "I Social Good"

Hosted by
Nicholas H. and Adam F.

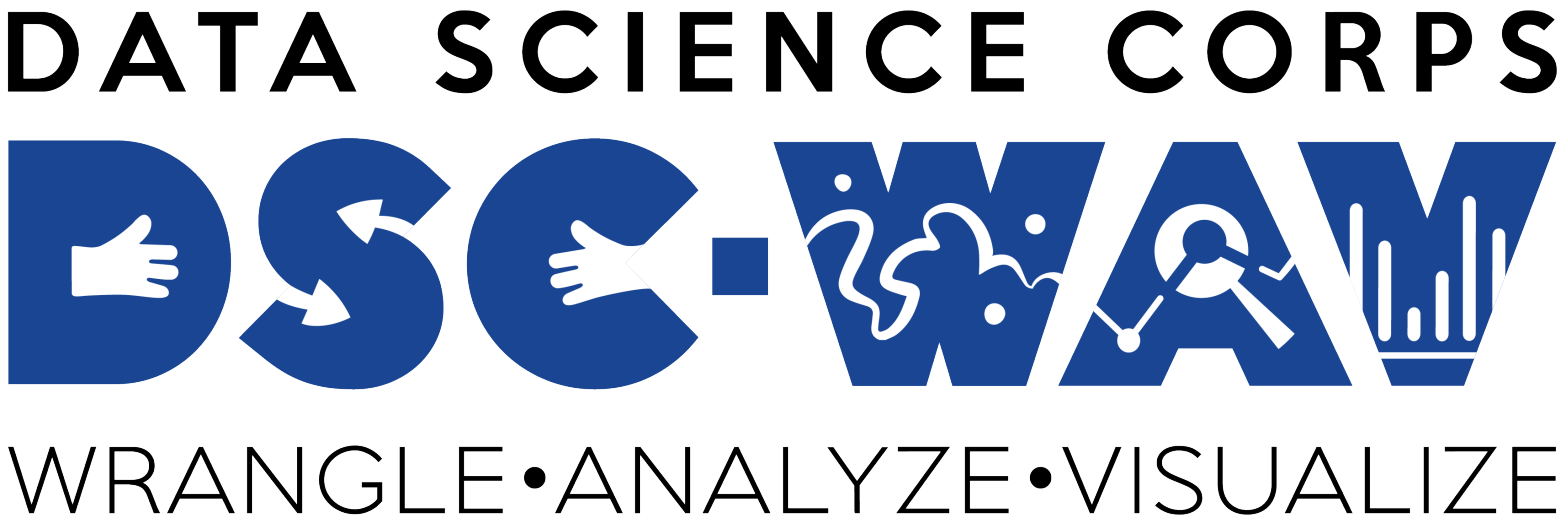
Western Mass Statistics and Data Science

Public group

Monday, October 7, 2019

Research Bytes at Mass Mutual: Data provenance + monitoring of production models

Building on the partnership



NSF grant #1924017 Harnessing the Data Revolution (Data Science Corps) The Data Science WAV: Experiential Learning with Local Community Organizations

DATA SCIENCE FOR UNDERGRADUATES

Opportunities and Options

The National Academies of | SCIENCES ENGINEERING MEDICINE

nas.edu/EnvisioningDS

NASEM (2018) Components of data acumen

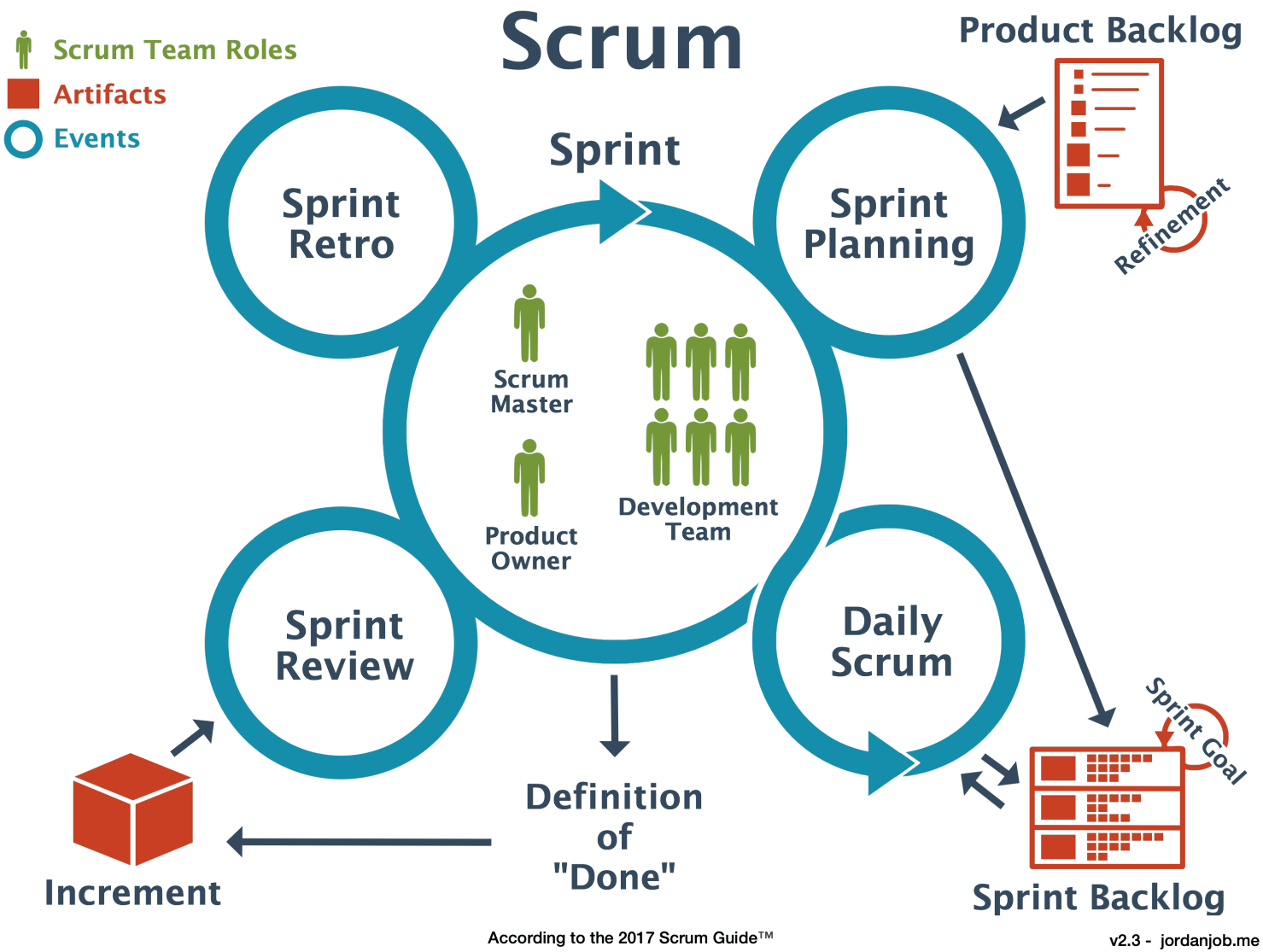
Finding 2.3: A critical task in the education of future data scientists is to
requires exposure to key concepts in data science, real-world data and pr
the limitations of tools, and ethical considerations that permeate many ap
involved in developing data acumen include the following:

- Mathematical foundations,
- Computational foundations,
- Statistical foundations,
- Data management and curation,
- Data description and visualization,
- Data modeling and assessment,
- Workflow and reproducibility,
- Communication and teamwork,
- Domain-specific considerations, and
- Ethical problem solving.

DSC-WAV Goal one: build workplace-ready skills

- teams of four specially-trained undergraduate students deployed to community-based organizations to **Wrangle, Analyze**, and **Visualize** their data
- build data science capacity at community organizations
- provide real-world experience that complements what students learn in the classroom

Better team data science using scrum and agile analysis



Girls Inc of the Valley

girls inc.

of the Valley

Students are working on a map for the organization to share with potential donors to demonstrate the accumulation of risk factors in the physical environment (air quality, water pollution, etc.) that face their target population on a daily basis

Western Massachusetts
Health Equity
Network

Students have researched and used publicly available date to build a dashboard interface, facilitating the retrieval of multi-year data on health characteristics by neighborhood in Springfield.

ValleyBike

Students are building visualizations to assist planners and inform decisions about how to expand the program most appropriately and best serve under-resources neighborhoods and communities.

VentureWell

Students are working to make data accessible by converting an archive of pdf grant applications to text, automating the loading of the files to a database, and making them searchable.

Nature Conservancy

Students are working to automate the import and analysis of patterns in images from wildlife cameras to assess whether a particular image includes an animal or not. The metadata for the images are read from a file and added to a growing database of available images.

DSC-WAV Goal two: Building capacity at community colleges

- help institutions create and improve data science programs
- foster faculty development
- create pathways for students to transfer and engage with DSC-WAV

Changing landscape of two year college data science education

Two Year College Data Science Summit (NSF funded, `https://www.amstat.org/ASA/Education/Two-Year-College-Data-Science-Summit.aspx`)

- certificate programs
- associates to workforce
- associates to transfer

Challenges in fostering partnerships

- No surprise: **time** is the major constraint (in government, in industry, and in academia)
- Need to protect junior faculty
- Conflicting need to build connections and engagement from day one
- Lack of incentives for building partnerships (How to credit and incentivize engagement across institutions?
- (MassMutual showering money on data science in the Pioneer Valley helps though!)

An Academic and Industry Partnership Training the Next Generation of Data Scientists

Nicholas J. Horton

Department of Mathematics and Statistics
Amherst College, Amherst, MA, USA

JSM2020, August 6, 2020

Slides at https://github.com/Amherst-Statistics/JSM2020

nhorton@amherst.edu