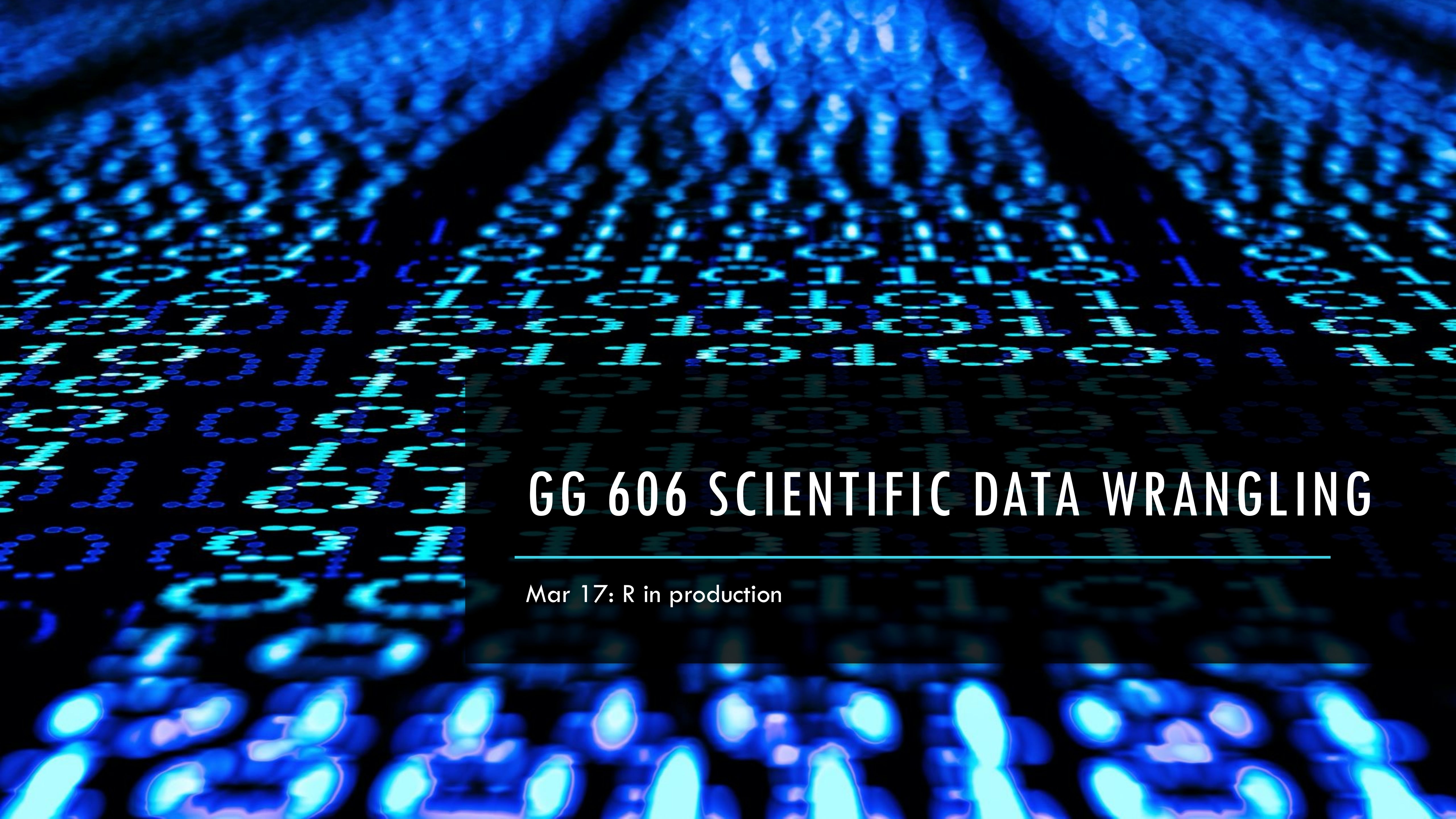
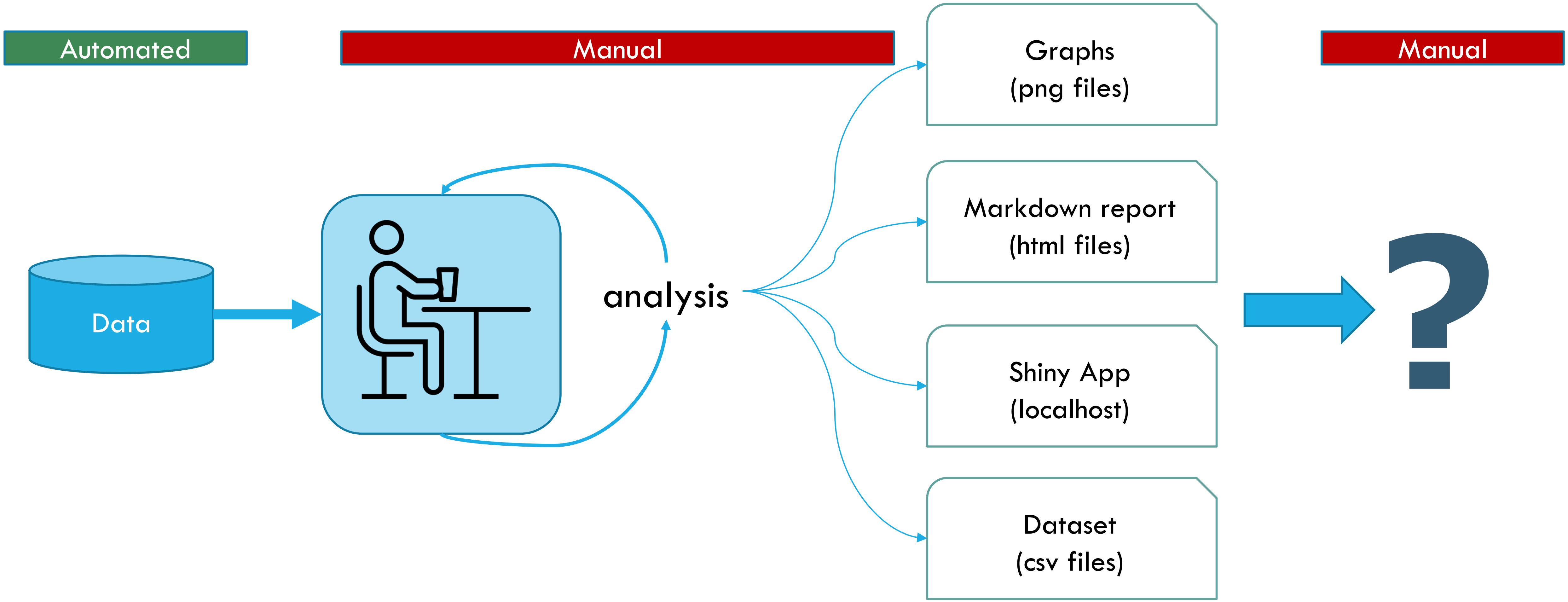


GG 606 SCIENTIFIC DATA WRANGLING

Mar 17: R in production

TYPICAL R-BASED ANALYSIS WORKFLOW



Automated

Manual

Manual

Data

analysis

Graphs
(png files)

Markdown report
(html files)

Shiny App
(localhost)

Dataset
(csv files)

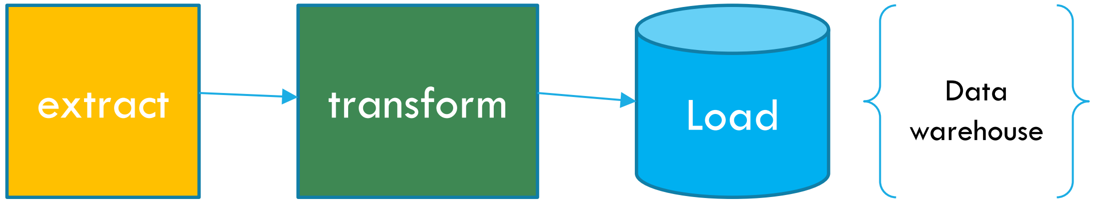
?

MOVING FROM ADHOC ANALYSIS TO SYSTEMS

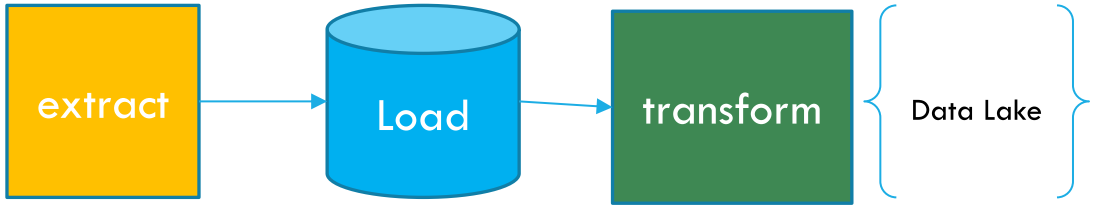
- Many organizations are increasingly aiming to become *data-driven* or aim to adopt *evidence-based* decision making processes
- How can we facilitate the use of data in business or decision-making processes
  - data → analyst → model → insights → decisions *does not scale well*
- How to built integrated systems that automate use of data/models
  - need to think about the technology environment within which an organization operates

“PRODUCTION” SOFTWARE SYSTEMS

- When we talk about ‘production’ vs ‘development’ we differentiate between those software components or systems that are vital to an organization’s functioning
- Databases, websites, transaction, sensors, data capture systems etc.
- custom scripts for moving data between systems
  - **ETL** – extract, transform, load
    - extract data from source systems
    - transform it by building variables, preparing final require datasets
    - load data into database for use by applications / models actually used by the organization



  - **ELT** – extract, load, transform



WHAT DO WE NEED TO THINK ABOUT WHEN PUTTING MODELS/ANALYTICS IN PRODUCTION?

- ANSWER: a whole lot
  - Security – data security, access controls, permissions (edit, read-only, execute, etc.), passwords, users, groups, access to external resources (e.g., APIs), etc.
  - Software environment – operating system, dependencies, upgrade cycle, what other software systems are used by users in the organization, etc.
  - Server environment – if we want website output need to consider where web applications are hosted and how they interact with other system components
  - workflow for updating software – CI/CD pipelines
  - workflow for version control – github repos

A lot of IT-related things you really don't want to deal with

It is HARD to put models and analytics in production systems, often it is not done except in specialized / tech-based organizations

YOUR R CODE

https://www.travelchannel.com/

PRODUCTION ENVIRONMENT

https://visualretailing.com/

R IN PRODUCTION – R SHINY

- Web-application development in R
- easy to use and quickly create a dashboard **locally** (localhost – only viewable on your own computer)
- more difficult to deploy, depends on
  - R-shiny Server – Windows
    - easy to set up, but not free
    - still requires configuring a web server software
  - R-Shiny Server – Linux
    - difficult to set up, free
- publish to an existing cloud service
  - easy but costs money
  - difficult to integrate within an existing website or organizational system

R IN PRODUCTION – R SHINY

- demo and examples
  - tutorials https://shiny.rstudio.com/tutorial/written-tutorial/lesson1/
  - demo - https://spatial.wlu.ca/rwe/
  - https://geographic.shinyapps.io/report-dashboard/

R IN PRODUCTION – R MARKDOWN

- Database reporting in R
  - may have data being collected continuously and want to do periodic analysis with some reports going out to stakeholders:
    - example architecture: https://wwwnc.cdc.gov/eid/article/16/10/10-0249-f2
  - Example

DEALING WITH ENVIRONMENT

1. Virtual Machines

- this is a *virtual* new computer or server which you can create on a disk partition
  - either on a server on the network or locally
- designate the OS and software dependencies
- works well for very computationally demanding applications with lots of specialized dependencies (e.g., ArcGIS Portal, R Studio Server)
- Problems:
  - lots of IT support needed
  - difficult to interact with (e.g. moving data in and out)
  - not great for sharing among different types of end-users or collaborators

DEALING WITH ENVIRONMENT

2. Containers

- like a mini-version of a virtual machine, deployed often for specific sets of tools or even individual applications
- a container encapsulates all OS requiremets in a self contained environment which an application needs to operate
- good use case for workflows with specific dependencies (e.g., package requires sf version 0.9.7 and sp version 2.3.1 etc.) which can cause issues when deployed on a server of regular operating systems
- requires extra overhead to configure and maintain
- good for reproducibility –
- https://colinfay.me/docker-r-reproducibility/

SUMMARY

- Data analysis tends to be disconnected from broader business and organizational processes and heavily dependent on manual steps
- This is starting to change with more incorporation of data-driven frameworks and machine learning models in particular
  - there is a LOT of complexity involved in moving from ad hoc analysis to a production-deployed model
  - requires consideration of broader ICT environment and coordination between IT and data science/analytics teams

GG 606 SCIENTIFIC DATA WRANGLING

Mar 17: R in production