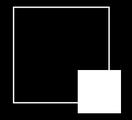


REPRODUCIBILITY

DISCOVERING AND COMMUNICATING INSIGHTS REPRODUCIBLY

EDUARDO ARIÑO DE LA RUBIA
CHIEF DATA SCIENTIST
DOMINO DATA LAB

EDUARDO@DOMINODATALAB.COM
TWITTER: @EARINO

WELCOME TO TDWI

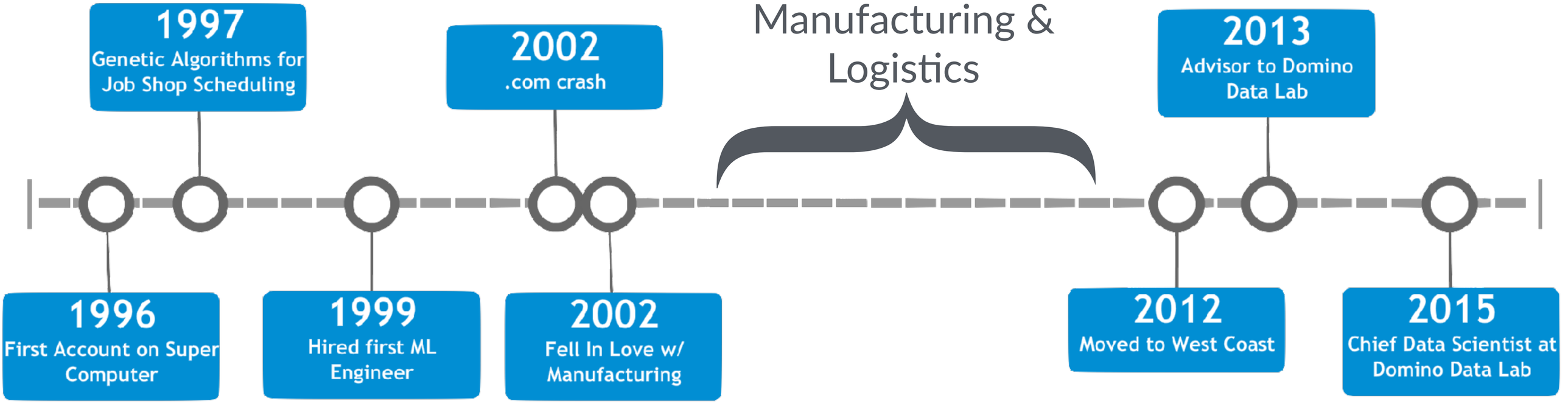
CONTENTS

A BIT ABOUT ME

DATA SCIENTIST

PICTURE SLIDE

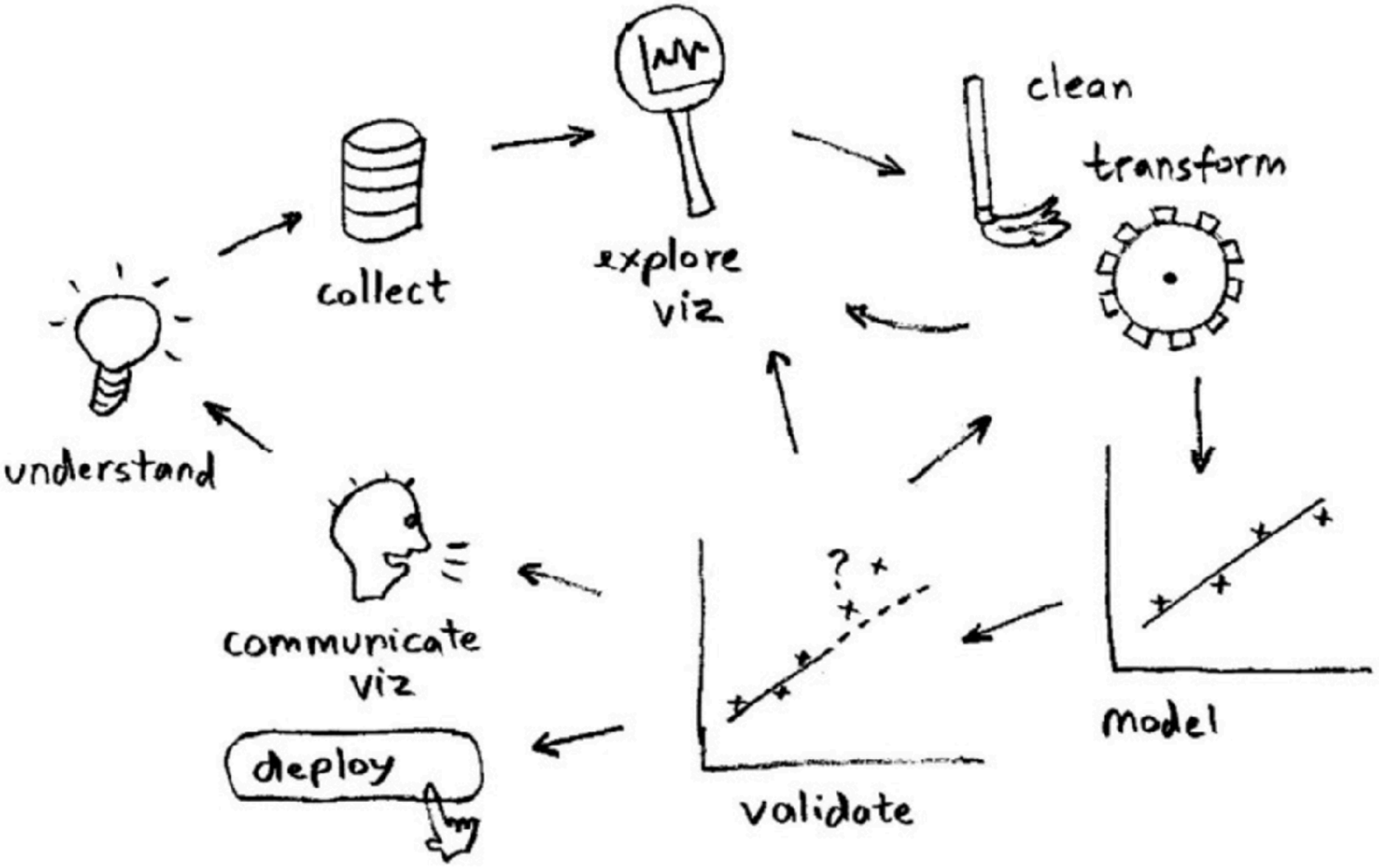
A QUICK TIMELINE



1

REPRODUCIBILITY

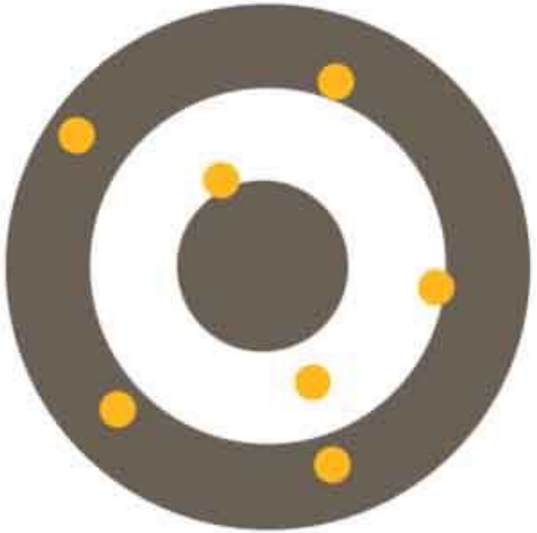
What is it, what isn’t it, why do you want it, and what will it enable you to do you can’t do otherwise.



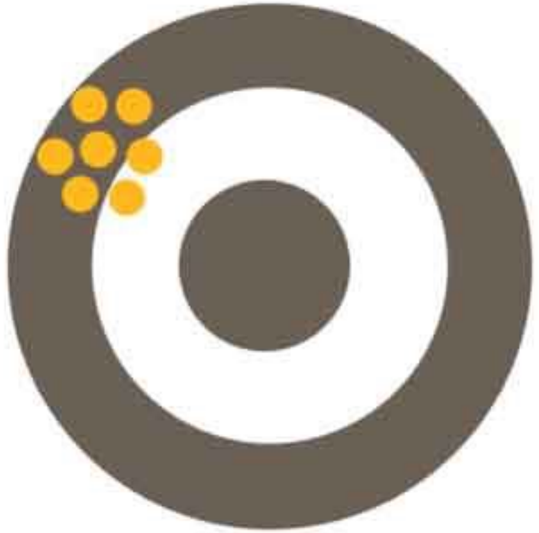
REPEATABILITY

Repeatability or test–retest reliability is the variation in measurements taken by a single person or instrument on the same item, under the same conditions, and in a short period of time. A less-than-perfect test–retest reliability causes test–retest variability.

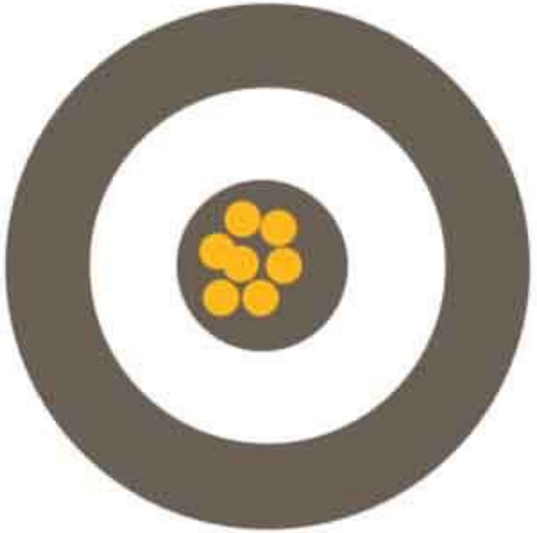
Your systems must be repeatable and reliable, the same query on the same data should return the same results. Otherwise, none of this applies.



Low repeatability, Low accuracy



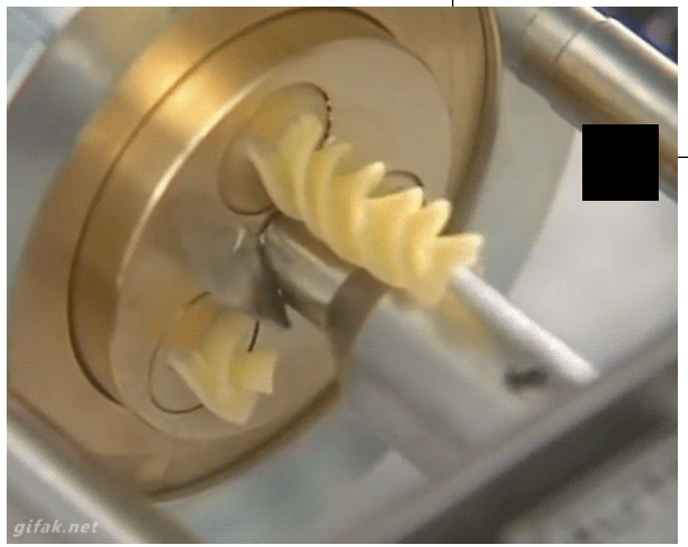
High repeatability, Low accuracy



High Repeatability, High accuracy

An analysis is **reproducible** if there is a specific set of computational functions/ analyses (**usually specified in terms of code**) that exactly reproduce all of the numbers in a published paper from raw data. It is now recognized that a critical component of the scientific process is that data analyses can be reproduced.

REPRODUCIBLE



REPLICABLE

But just because a study is **reproducible** does not mean that it is ***replicable***. **Replicability** is stronger than **reproducibility**. A study is only **replicable** if you perform the exact same experiment (at least) twice, collect data in the same way both times, perform the same data analysis, and arrive at the same conclusions.

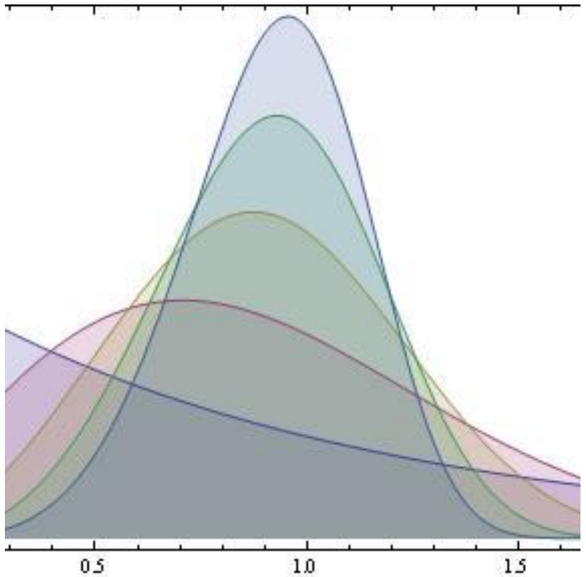
STODDEN'S TAXONOMY OF REPRODUCIBILITY

COMPUTATIONAL

when detailed information is provided about code, software, hardware and implementation details.

EMPIRICAL

when detailed information is provided about non-computational empirical scientific experiments and observations. In practise this is enabled by making data freely available, as well as details of how the data was collected.

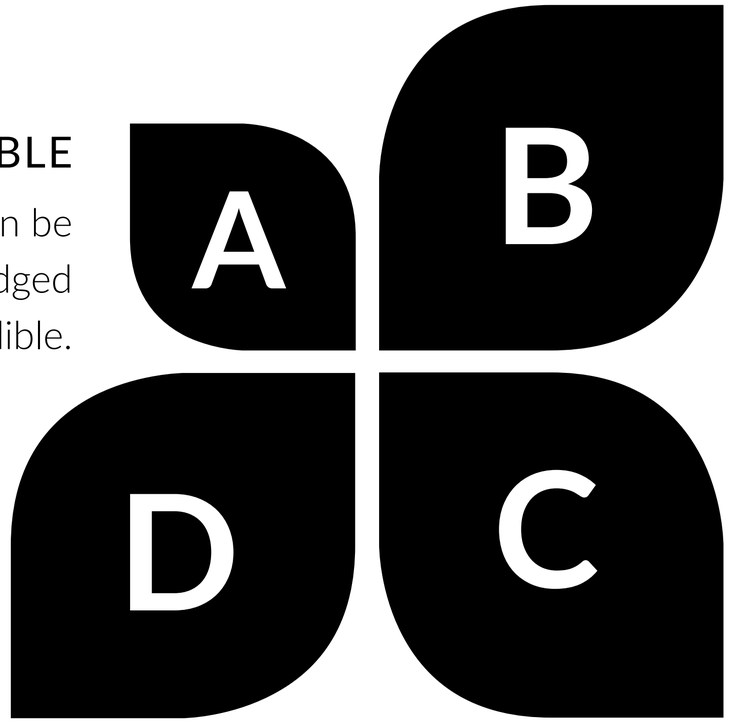


STATISTICAL

when detailed information is provided about the choice of statistical tests, model parameters, threshold values, etc. This mostly relates to pre-registration of study design to prevent p-value hacking and other manipulations

COMPUTATIONAL REPRODUCIBILITY

SHOULD BE



REVIEWABLE

The descriptions of the research methods can be independently assessed and the results judged credible.

REPLICABLE

Tools are made available that would allow one to duplicate the results of the research, for example by running the authors' code to produce the plots shown in the publication.

AUDITABLE

Sufficient records (including data and software) have been archived so that the research can be defended later if necessary or differences between independent confirmations resolved. The archive might be private, as with traditional laboratory notebooks.

CONFIRMABLE

The main conclusions of the research, or the model in production can be attained independently without the use of software provided by the author.

2

BEST PRACTICES

Nine Simple Rules for Reproducible Computational Research

TRACK RESULTS

Whenever a result may be of potential interest, keep track of how it was produced. As a minimum, you should at least record sufficient details on programs, parameters, and manual procedures to allow yourself, in a year or so, to approximately reproduce the results.

1

2

SCRIPT EVERYTHING

Whenever possible, rely on the execution of programs instead of manual procedures to modify data. If manual operations cannot be avoided, you should as a minimum note down which data files were modified or moved, and for what purpose.

STORE PROGRAM VERSIONS

In order to exactly reproduce a given result, it may be necessary to use programs in the exact versions used originally. As a minimum, you should note the exact names and versions of the main programs you use.

3

4

USE VERSION CONTROL

Even the slightest change to a computer program can have large intended or unintended consequences. As a minimum, you should archive copies of your scripts from time to time, so that you keep a rough record of the various states the code has taken during development.

5

STORE DATA & INTERMEDIATE RESULTS

In principle, as long as the full process used to produce a given result is tracked, all intermediate data can also be regenerated. In practice, having easily accessible intermediate results may be of great value. As a minimum, archive any intermediate result files that are produced when running an analysis .

6

SET A RANDOM NUMBER SEED

Many analyses and predictions include some element of randomness, meaning the same program will typically give slightly different results every time it is executed. As a minimum, note which analysis steps involve randomness, so that a level of discrepancy can be anticipated when reproducing the results.

STORE DATA VIZ INPUTS

From the time a figure is first generated to it being part of an analysis. It is critical to store the data and process that generated it. As a minimum, one should note which data formed the basis of a given plot and how this data could be reconstructed.

7

8

ALLOW LEVELS OF ANALYSIS

In order to validate and fully understand the main result, it is often useful to inspect the detailed values underlying the summaries. Make those fluid and explorable, as a minimum at least once generate, inspect, and validate the detailed values underlying the summaries.

GENERATE TEXT PROGRAMMATICALLY

There is nothing quite as embarrassing as having a disagreement between your analytical narrative in a writeup and having the tables and figures disagree. Connect them so that there is no change of disagreement.

9

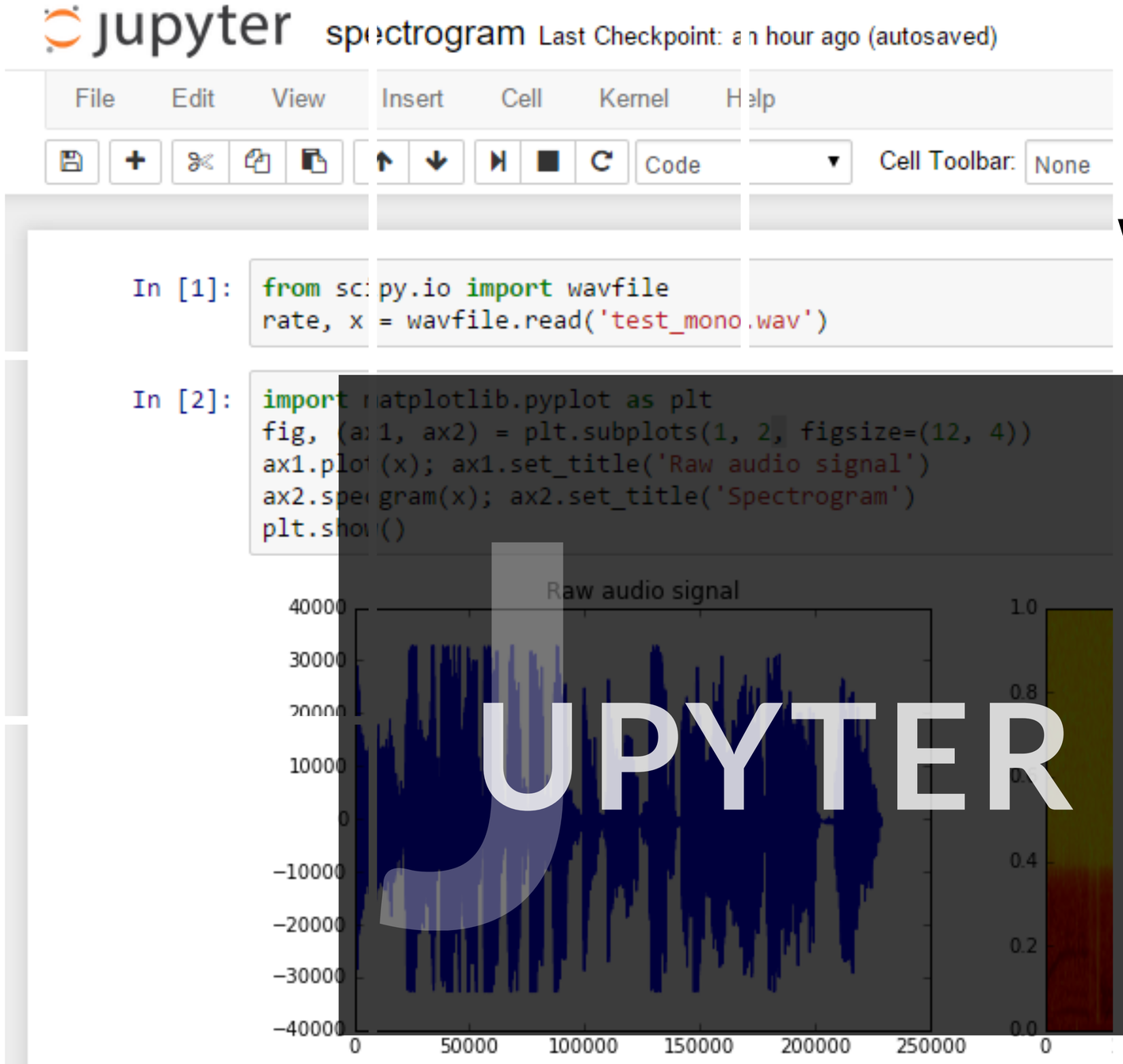
3

TOOLS AND INTERFACES

The right tool for the right job isn’t just an aphorism, it’s legitimately good advice.

I am convinced that we need better tools for scientific computing, and that our efforts to build them based on the Python language can make a significant difference to how scientific research is conducted and disseminated.

- Fernando Perez



WHAT ARE

JUPYTER NOTEBOOKS

Notebook documents (or “notebooks”, all lower case) are documents produced by the Jupyter Notebook App, which contain both computer code (e.g. python) and rich text elements (paragraph, equations, figures, links, etc…). Notebook documents are both human-readable documents containing the analysis description and the results (figures, tables, etc..) as well as executable documents which can be run to perform data analysis.

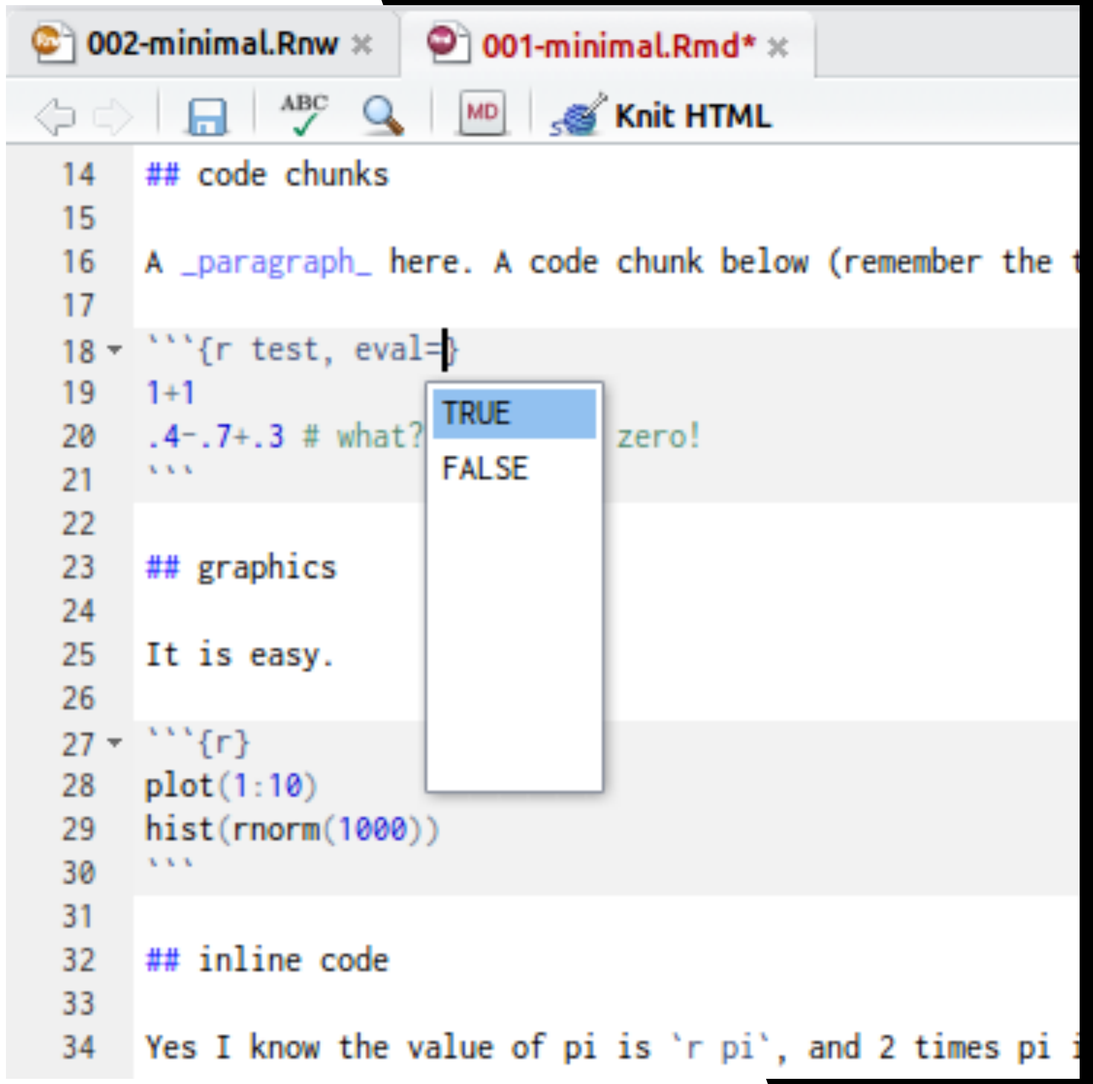
YIHUI XIE

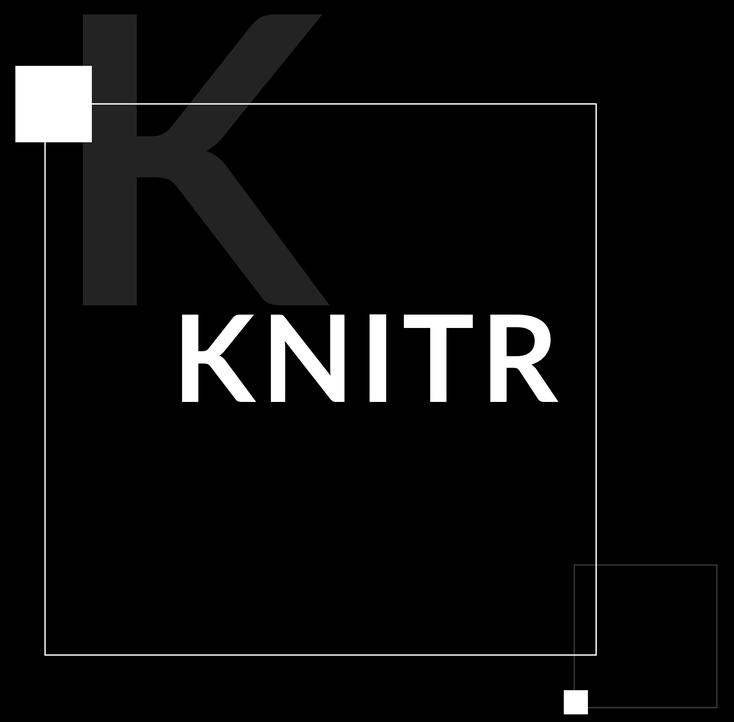
A software engineer working at RStudio, Inc. He earned his PhD from the Department of Statistics, Iowa State University. His thesis was *Dynamic Graphics and Reporting for Statistics*, advised by Di and Heike.

Creator of the knitR package

PIONEER IN REPRODUCIBILITY

knitr is an engine for dynamic report generation with R. It is a package in the statistical programming language R that enables integration of R code into LaTeX, LyX, HTML, Markdown, AsciiDoc, and reStructuredText documents. The purpose of knitr is to allow reproducible research in R through the means of Literate Programming.



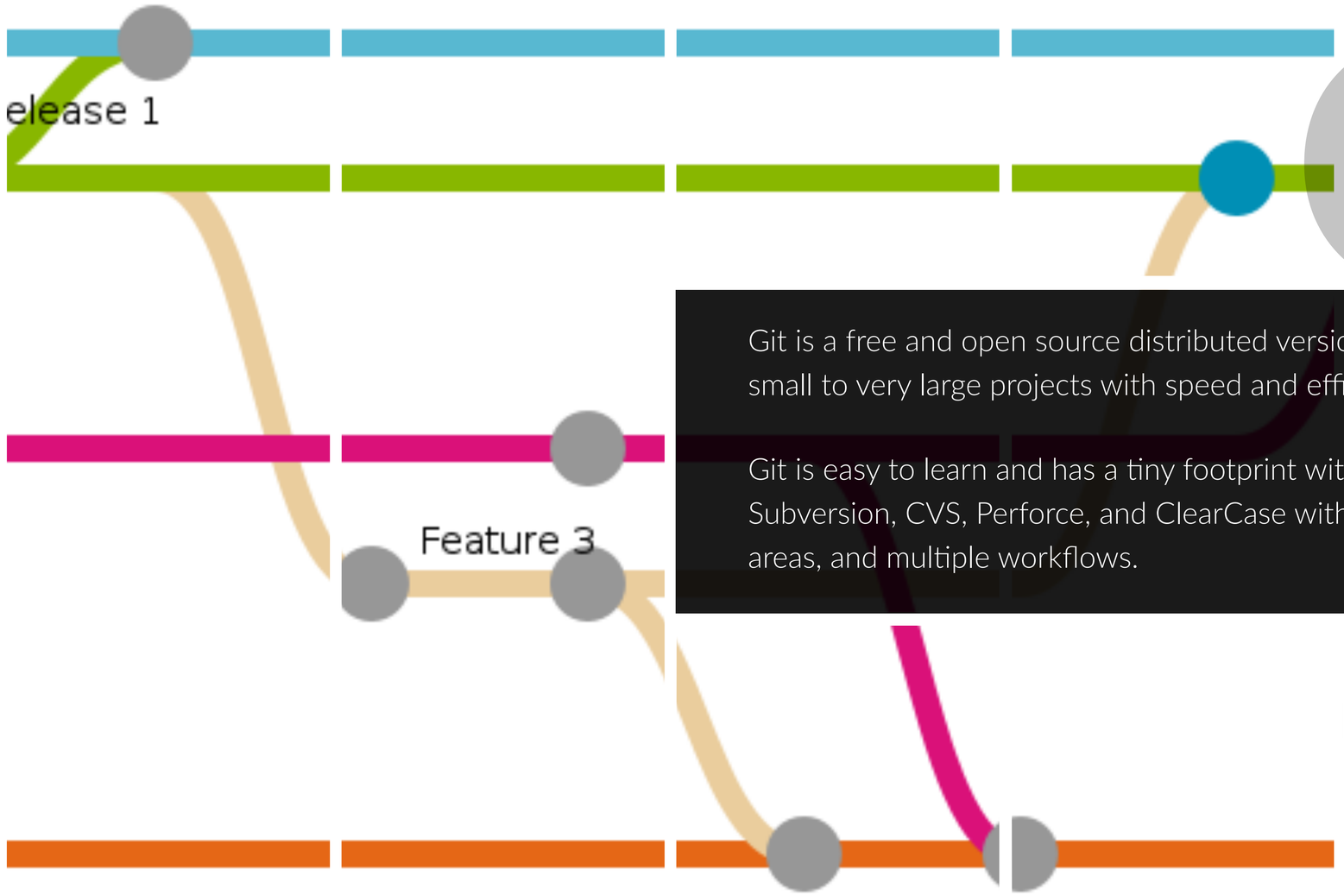


COMPARE JUPYTER & KNITR

SO, WHICH ONE?

There are a number of key differences between jupyter notebooks and knitr, but the most important is that Jupyter notebooks have a **LIVE** kernel, cells can be run **OUT OF ORDER**

knitR documents must be reproducible, they run once top to bottom.



GIT

Git is a free and open source distributed version control system designed to handle everything from small to very large projects with speed and efficiency.

Git is easy to learn and has a tiny footprint with lightning fast performance. It outclasses SCM tools like Subversion, CVS, Perforce, and ClearCase with features like cheap local branching, convenient staging areas, and multiple workflows.

Pinned Tweet

E. Ariño de la Rubia @earino · Mar 30

Please RT. Answer if a "data scientist" (or statistical programming language user). Do you agree with the following statement: "I use git."

82% Agree

18% Disagree

931 votes • Final results

10 85 14

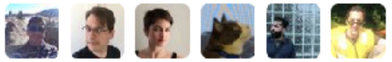
Leonard Kiefer
@lenkiefer

Following

Replying to @earino

Does knowing only a handful of commands and otherwise Googling furiously count as "using"? Asking for a friend.

LIKES
7



10:48 AM - 30 Mar 2017

3 7

Hadley Wickham @hadleywickham · Mar 30
Replying to @lenkiefer @earino
there's another way to use git?

2 7

@bistromath2013 · Mar 31
Replying to @earino @igelundfuchs
I use git but I hate it.

1

Leonard Kiefer

@lenkiefer FOLLOWS YOU

Deputy Chief Economist at Freddie Mac.
I help people understand the economy,
housing, and mortgage markets.

Hadley Wickham

@hadleywickham FOLLOWS YOU

R, data, visualisation.

DRAKE

DATA FRAMES IN R FOR MAKE

Drake is a workflow manager for R. When it runs a project, it automatically builds missing and outdated results while skipping over all the up-to-date output. This automation and reproducibility is important for data analysis workflows, especially large projects under heavy development.

DRAKE

MAKE FOR DATA

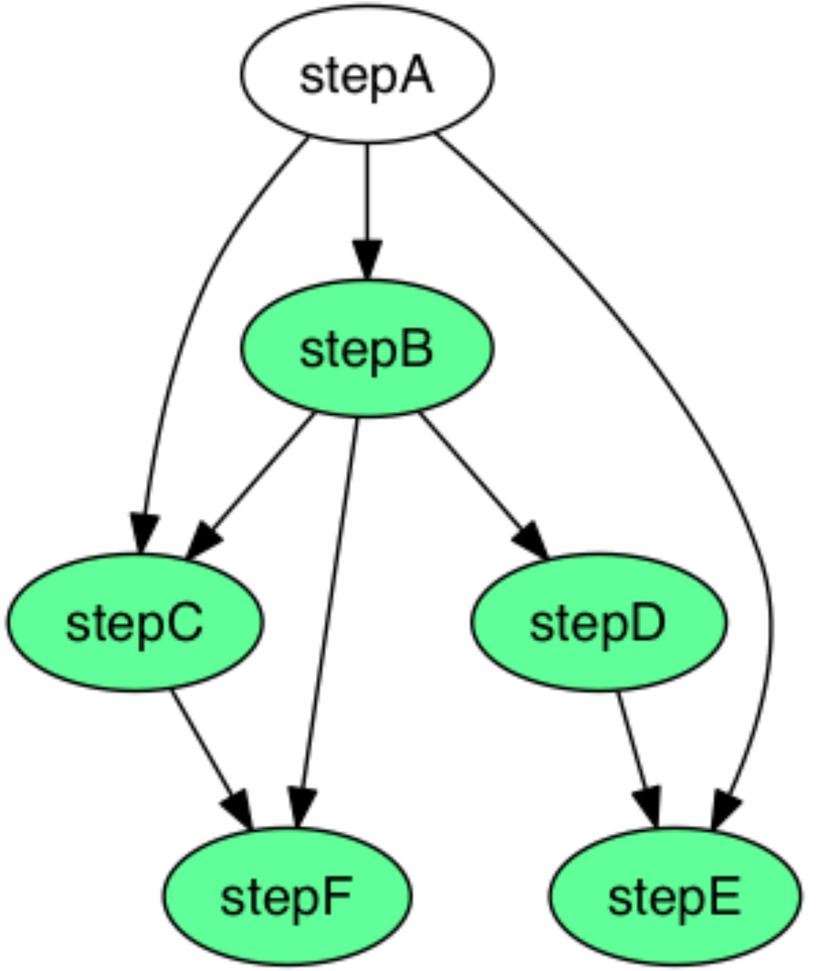
Yes, there are two separate projects that are makes for data named drake. This is not a great situation, I agree.

Drake is a text-based command line data workflow tool that organizes command execution around data and its dependencies. Data processing steps are defined along with their inputs and outputs. It automatically resolves dependencies and provides a rich set of options for controlling the workflow. It supports multiple inputs and outputs and has HDFS support built-in.

```
;
; Grabs us some data from the Internets
;
contracts.csv <-
  curl http://www.ferc.gov/docs-filing/eqr/soft-tools/sample-csv/contr

;
; Filters out all but the evergreen contracts
;
evergreens.csv <- contracts.csv
  grep Evergreen $INPUT > $OUTPUT

;
; Saves a super fancy report
;
report.txt <- evergreens.csv [python]
  linecount = len(file("$[INPUT]").readlines())
  with open("$[OUTPUT]", "w") as f:
    f.write("File $[INPUT] has {0} lines.\n".format(linecount))
```




drake


DRAKE

DOCKER IS THE FUTURE

Containerization is kind of a must...

Docker is an open-source project that automates the deployment of applications inside software containers.

> Docker containers wrap up a piece of software in a complete filesystem that contains everything it needs to run: code, runtime, system tools, system libraries – anything you can install on a server. **This guarantees that it will always run the same, regardless of the environment it is running in.**

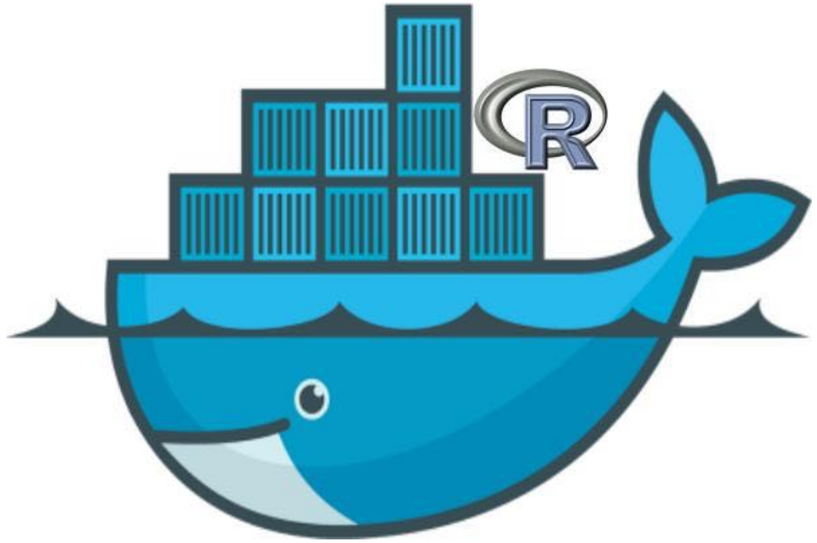
Docker provides an additional layer of abstraction and automation of operating-system-level virtualization on Windows and Linux. Docker uses the resource isolation features of the Linux kernel such as cgroups and kernel namespaces, and a union-capable file system such as OverlayFS and others to allow independent "containers" to run within a single Linux instance, avoiding the overhead of starting and maintaining virtual machines.

BLACK AND WHITE WORLD

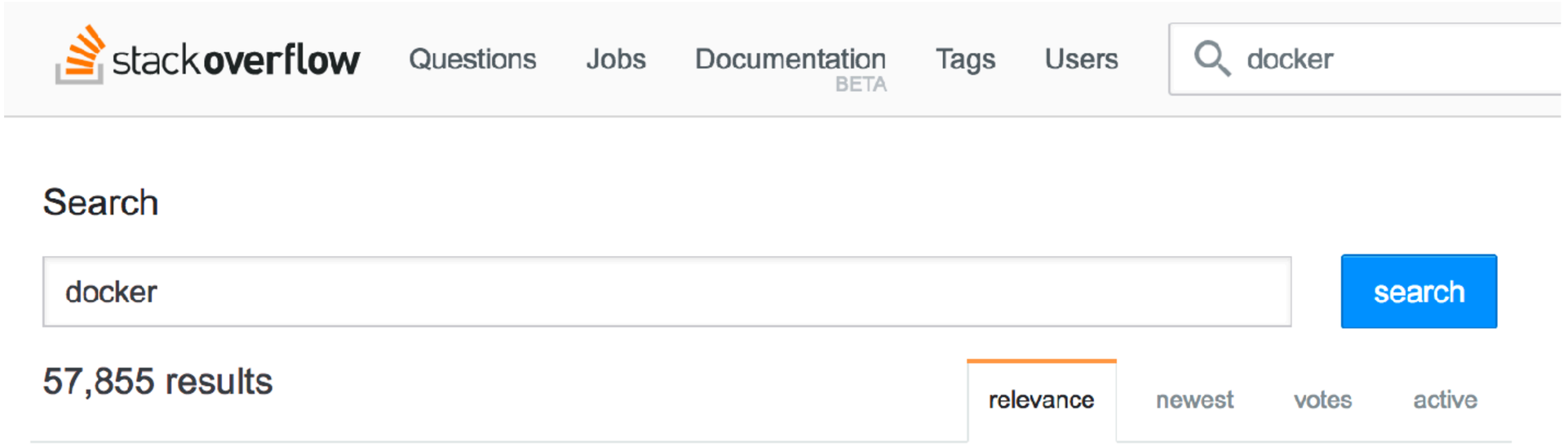
ANACONDA

Powered by Continuum Analytics®

docker



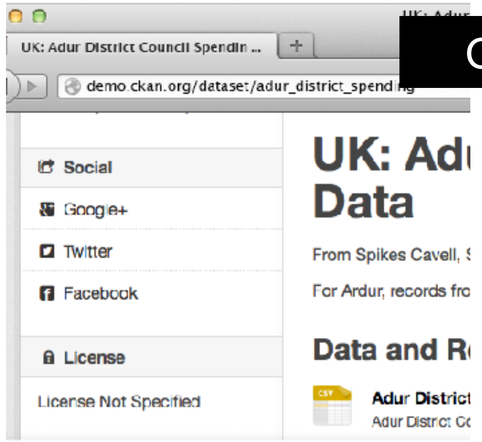
julia


stackoverflow
Questions
Jobs
Documentation
BETA
Tags
Users
docker
Search
docker
search
57,855 results
relevance
newest
votes
active

TIE IT ALL TOGETHER

JENKINS

Jenkins helps to automate the non-human part of the whole software development process, with now common things like continuous integration, but by further empowering teams to implement the technical part of a Continuous Delivery.



CKAN

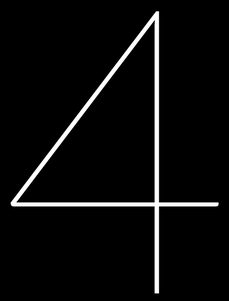
CKAN is a powerful data management system that makes data accessible – by providing tools to streamline publishing, sharing, finding and using data. CKAN is aimed at data publishers (national and regional governments, companies and organizations) wanting to make their data open and available.

ELASTIC SEARCH

Elasticsearch is a search engine based on Lucene. It provides a distributed, multitenant-capable full-text search engine with an HTTP web interface and schema-free JSON documents

GITHUB

Oh yeah git by itself definitely won't work, and it doesn't track experiments or data really well (large files suck.) You're gonna need GitHub enterprise.



LEVERAGING PLATFORMS

It would be nice if there was a way to make this all just work automatically, wouldn’t it?


ANACONDA®


DATASCIENCE


KNIME


Pachyderm


rapidminer


data
iku

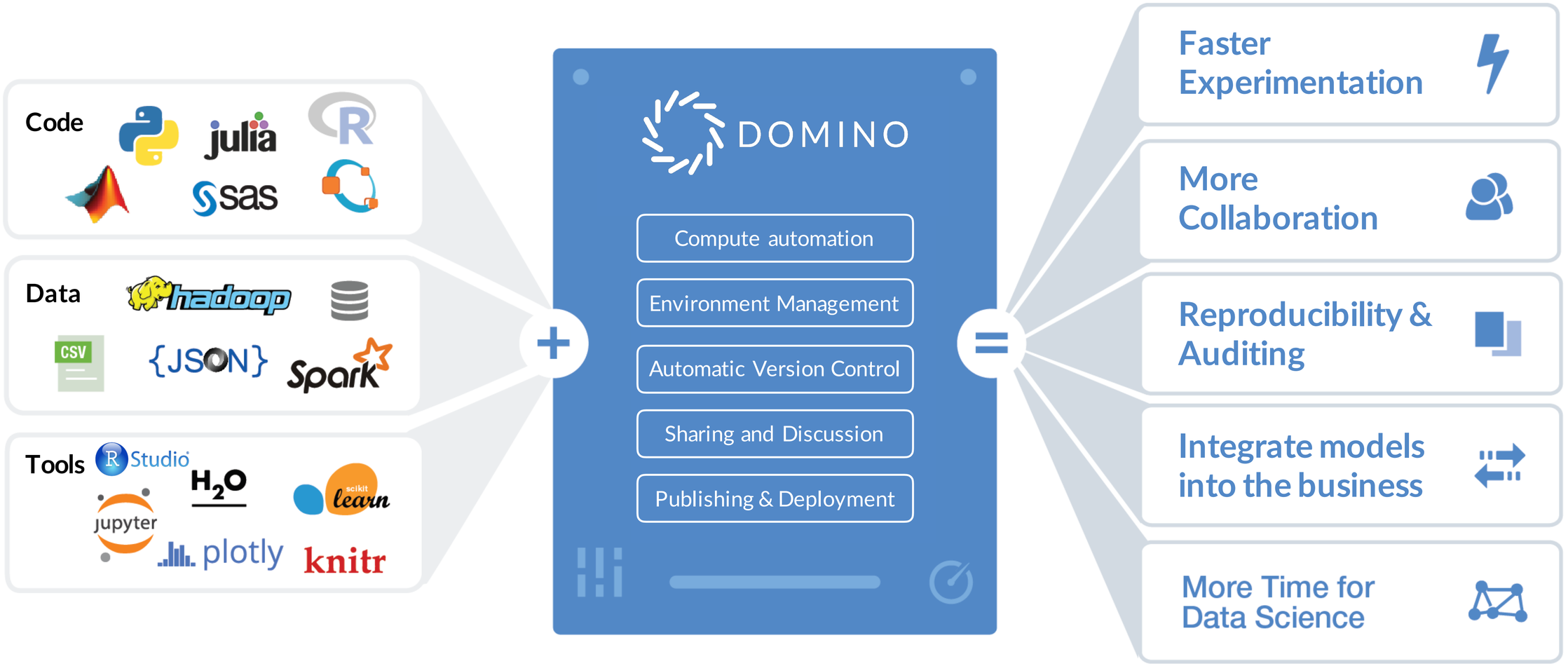

ALPINE DATA™


alteryx

DOMINO

CHECK US OUT AT
DOMINODATALAB.COM

Solution: An **open platform** to be a **central hub** for data science



More Time for Data Science


ZURICH


Carnival


CLOROX


MONSANTO


Allstate
You're in good hands.


UNIVERSITY of VIRGINIA


S&P
CAPITAL IQ


LIVE OAK BANK


bluestem
brands, inc.


DBRS

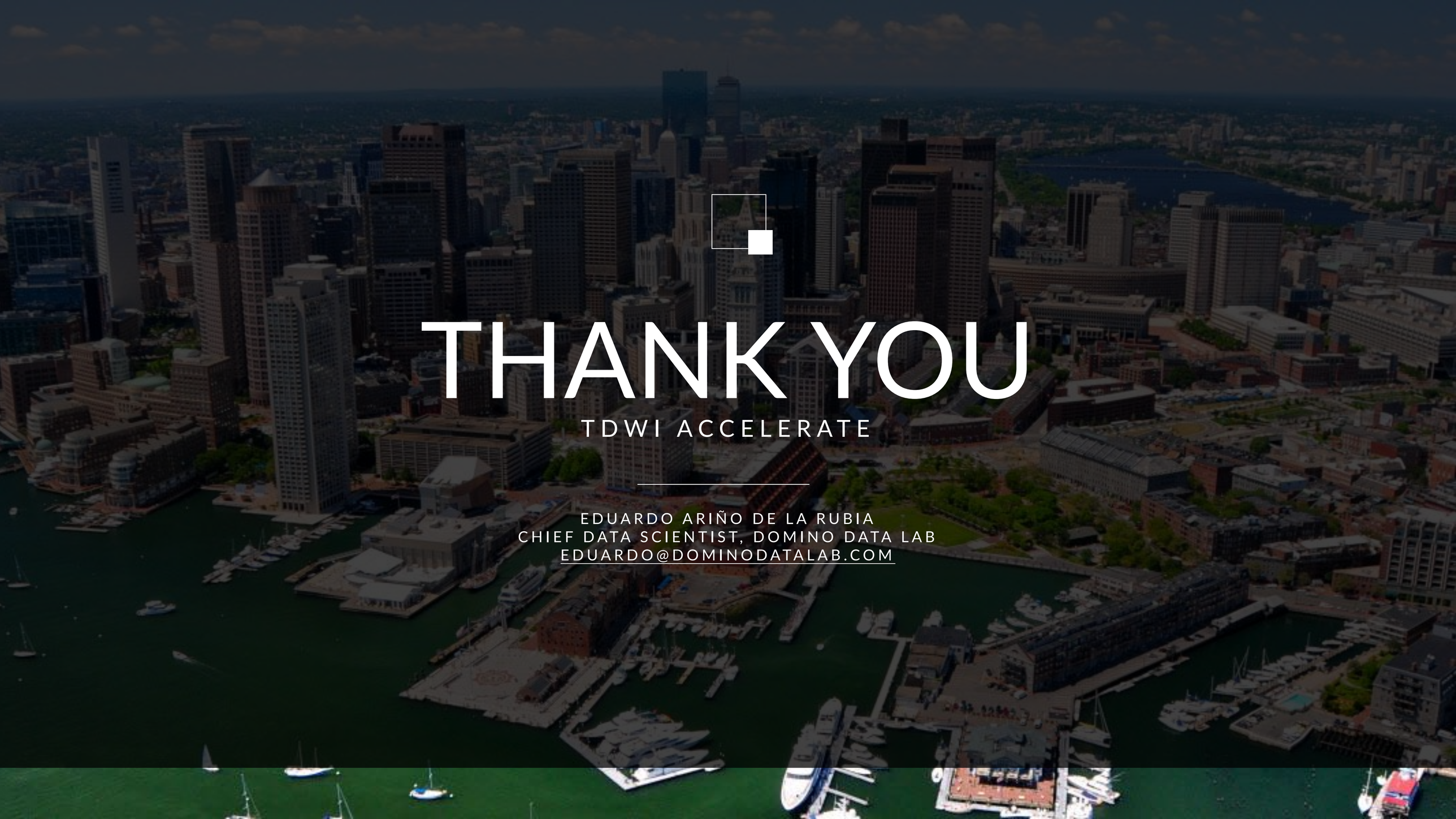

Eventbrite


NOV


Mashable


Instacart


KatRisk



THANK YOU

TDWI ACCELERATE

EDUARDO ARIÑO DE LA RUBIA
CHIEF DATA SCIENTIST, DOMINO DATA LAB
EDUARDO@DOMINODATALAB.COM