DATA 607
Fall 2021
Meetup 1

Andy Catlin

Experts in their fields tend to be motivated by criticism, and to see it as a sign of how well they're progressing toward their goals. Novices are more likely to seek praise, and to interpret it as a sign of whether to remain committed to the goals they've set.

Sue Shellenberg

If you don't
get feedback,
your confidence
grows much
faster than
your accuracy.

Philip Tetlock

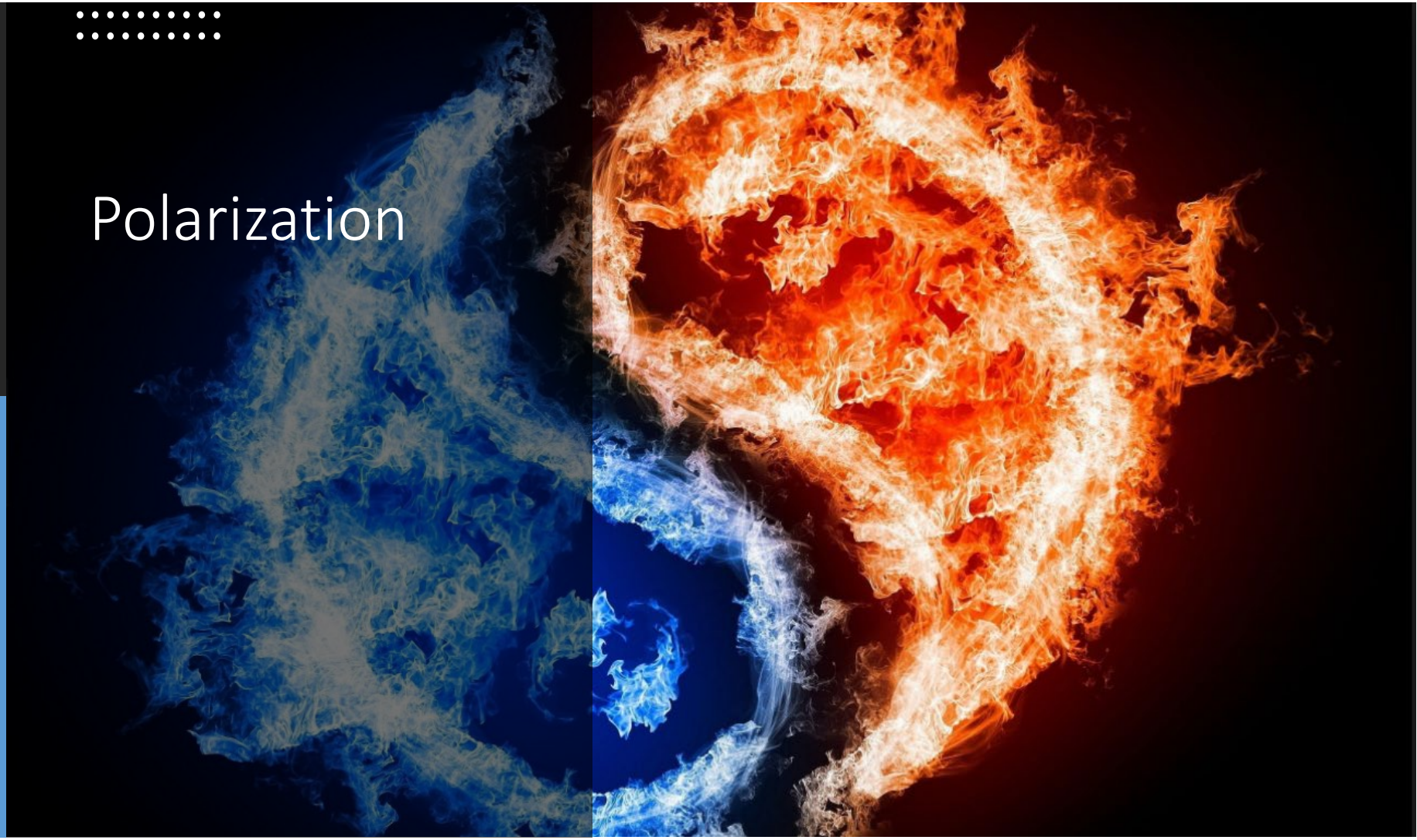
Topics

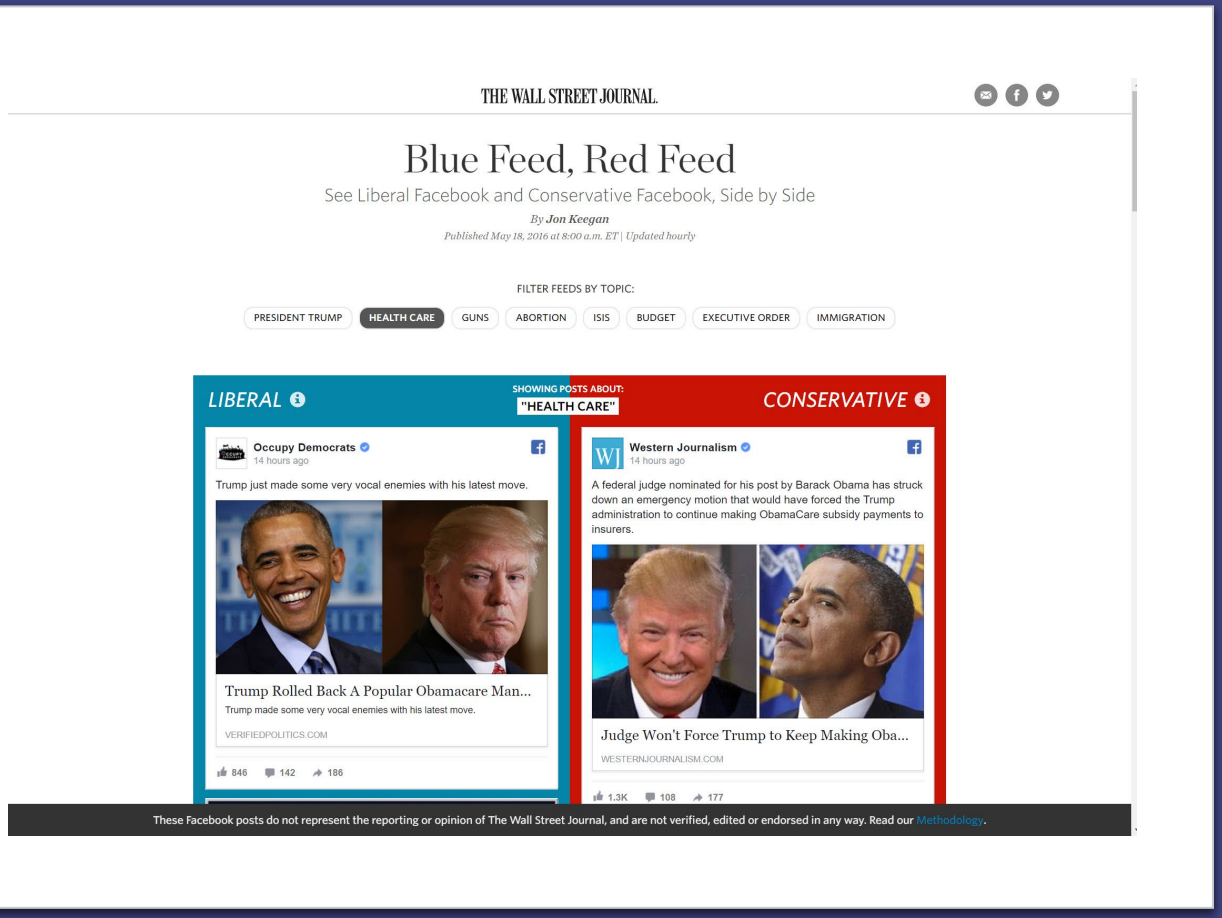
- **Course Overview – Course Site, assignments, resources**
- Data Ethics: Polarization and Privacy
- The Data Science Lifecycle
- Demonstration: GitHub

Topics

- Course Overview – Course Site, assignments, resources
- **Data Ethics: Polarization and Privacy**
- The Data Science Lifecycle
- Demonstration: GitHub

Polarization





News Recommenders

http://graphics.wsj.com/blue-feed-red-feed/

Video Recommenders

Google Brain's researchers wondered if they could keep YouTube users engaged for longer by steering them into different parts of YouTube, rather than feeding their existing interests. And they began testing a new algorithm that incorporated a different type of A.I., called reinforcement learning.

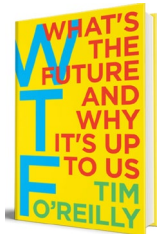
The new A.I., known as Reinforce, was a kind of long-term addiction machine. It was designed to maximize users' engagement over time by predicting which recommendations would expand their tastes and get them to watch not just one more video but many more.

Reinforce was a huge success. In a talk at an A.I. conference in February, Minmin Chen, a Google Brain researcher, said it was YouTube's most successful launch in two years. Sitewide views increased by nearly 1 percent, she said — a gain that, at YouTube's scale, could amount to millions more hours of daily watch time and millions more dollars in advertising revenue per year. She added that the new algorithm was already starting to alter users' behavior.

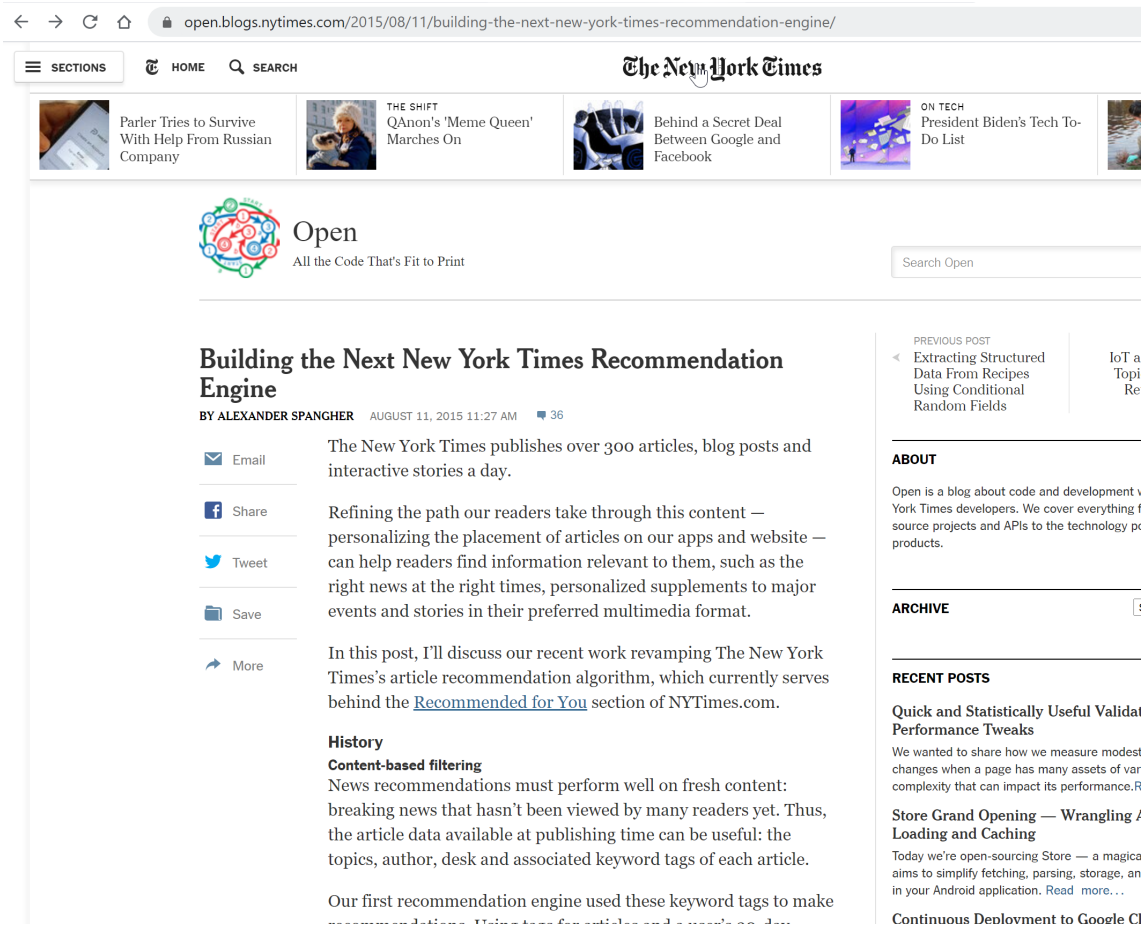
Kevin Rose, "The Making of a YouTube radical," New York Times, June 8, 2019.
https://www.nytimes.com/interactive/2019/06/08/technology/youtube-radical.html

humanity of the people it describes. Korzybski urged everyone to become viscerally aware of the process of abstraction, by which reality is transformed into a series of statements about reality—maps that can guide us but can also lead us astray.

This insight seems particularly important in the face of the fake news that bedeviled the 2016 US presidential election. It wasn't just the most outrageous examples, like the child slavery ring supposedly being run by the Clinton campaign out of a Washington, DC, pizza joint, but the systematic and increasingly algorithmic selection of news to fit and amplify people's preconceived views. Whole sectors of the population are now led by vastly divergent maps. How are we to solve the world's most pressing problems when we aren't even trying to create maps that reflect the actual road ahead, but instead drive toward political or business goals?



Privacy



Source: https://open.blogs.nytimes.com/2015/08/11/building-the-next-new-york-times-recommendation-engine/

Privacy

Snip Canada • August 12, 2015

I do not want you, NYT, to know what articles I'm reading. It's none of your business even though you think it is.

2 Recommend

Harold M Heft Princeton, NJ • August 12, 2015

Frankly, I don't want NYT tracking what I am viewing attached to my name or log in at all. If NYT wants to track views counts, please do it anonymously. I would really like an opt out choice to eliminate "Recommended for You" altogether on my NYT. It is none of your business what I am reading.

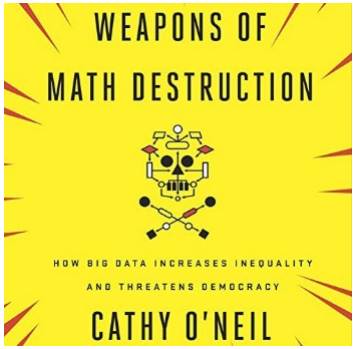
6 Recommend

Bob Foster Dublin, CA • August 12, 2015

Same old tired idea. You will assume you know what I want to read in the future based on what I have read in the past and I will never click on Recommended For You. If you force your assumptions on me by "customizing" the front page, I will stop reading the Times. What I want to read in the future are things I haven't read in the past.

Recommend

“Data Science doesn’t just predict the future; it causes the future.”
Cathy O’Neil

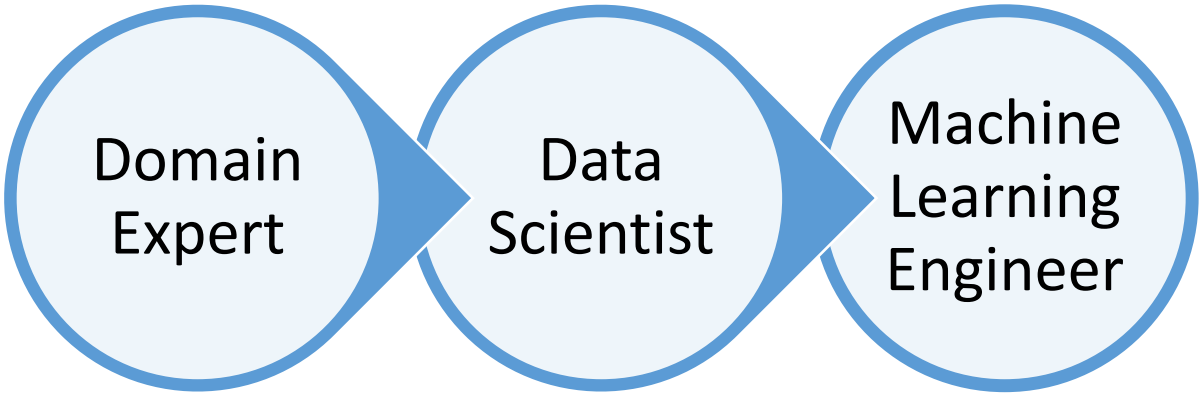



4А
5А
1П
2П
3П

Topics

- Course Overview – Course Site, assignments, resources
- Data Ethics: Polarization and Privacy
- **The Data Science Lifecycle**
- Demonstration: GitHub

Data Science Project Roles

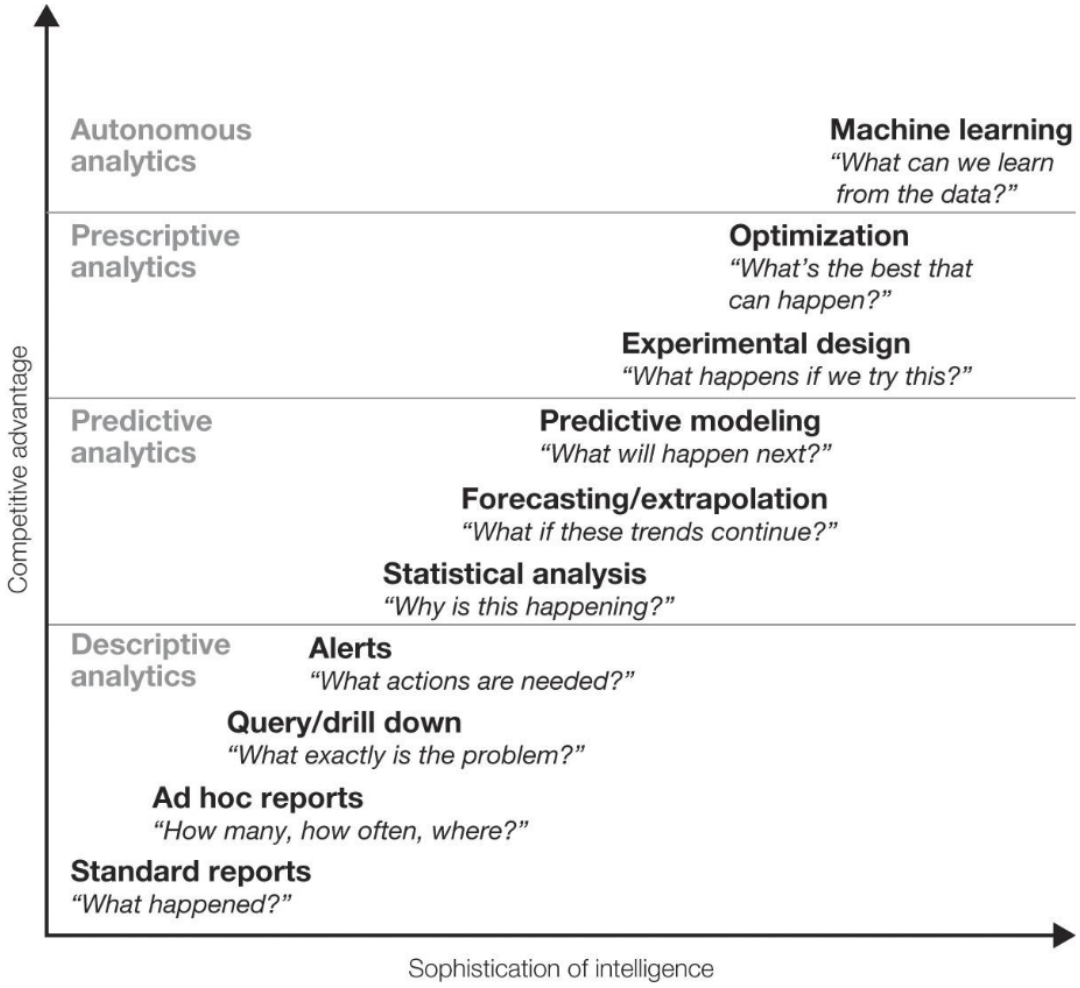


Is it better to be a generalist or a specialist?

Suppose you work on a team of hundreds of data scientists at American Express or UPS?

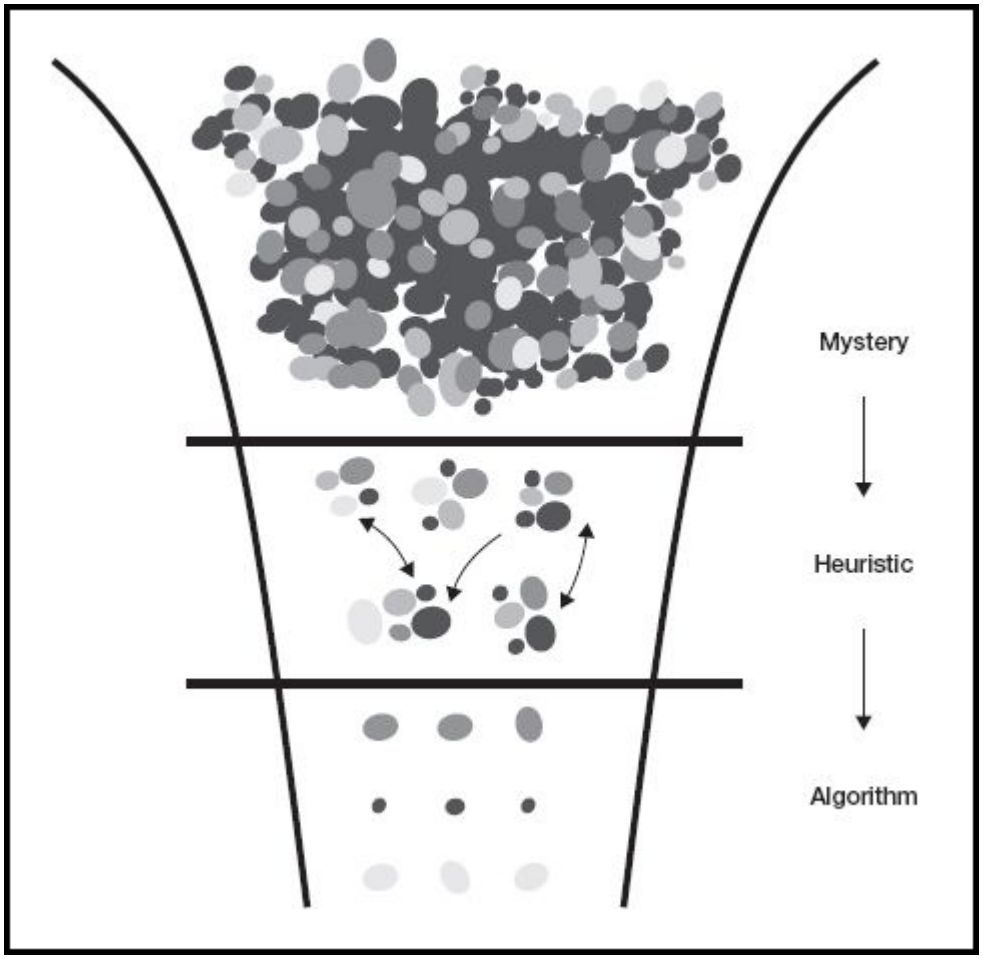
Suppose you work in a small start-up, where you need to assume several roles?

Potential competitive advantage increases with more sophisticated analytics

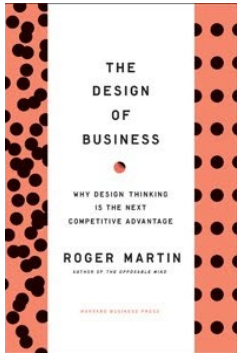


Source: Thomas Davenport, *Competing on Analytics*

The Design of Business

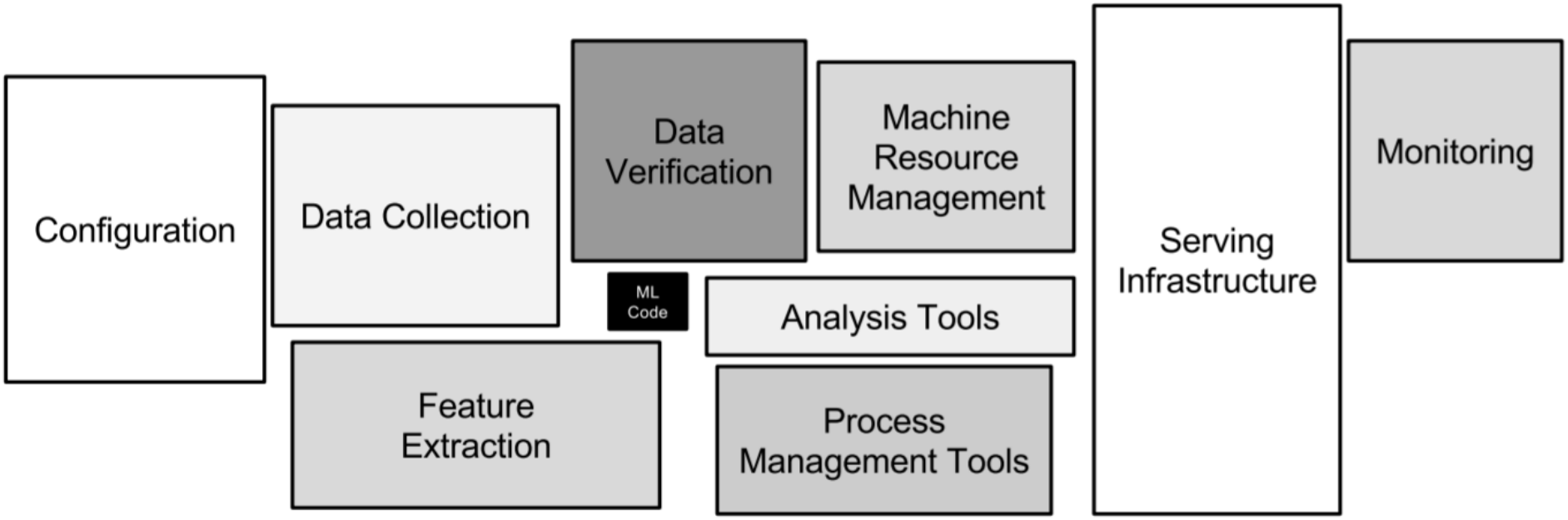


Source: Martin, Roger L. *The Design of Business: Why Design Thinking Is the next Competitive Advantage.* Boston, MA: Harvard Business, 2009. Print.



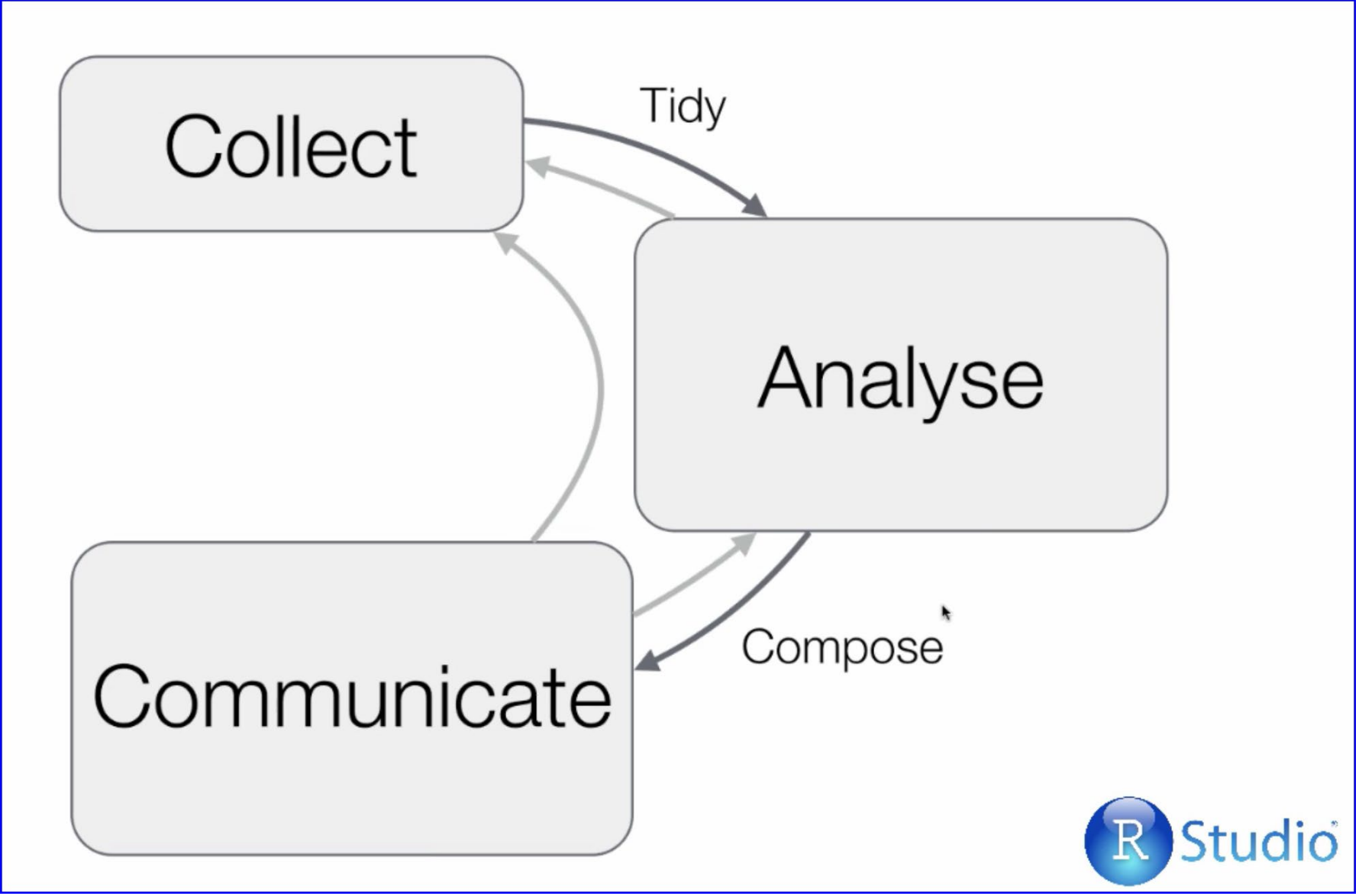
Data Science leads to…

- Process Improvement
- **Process Automation**
- Product Improvement
- **Data Products**

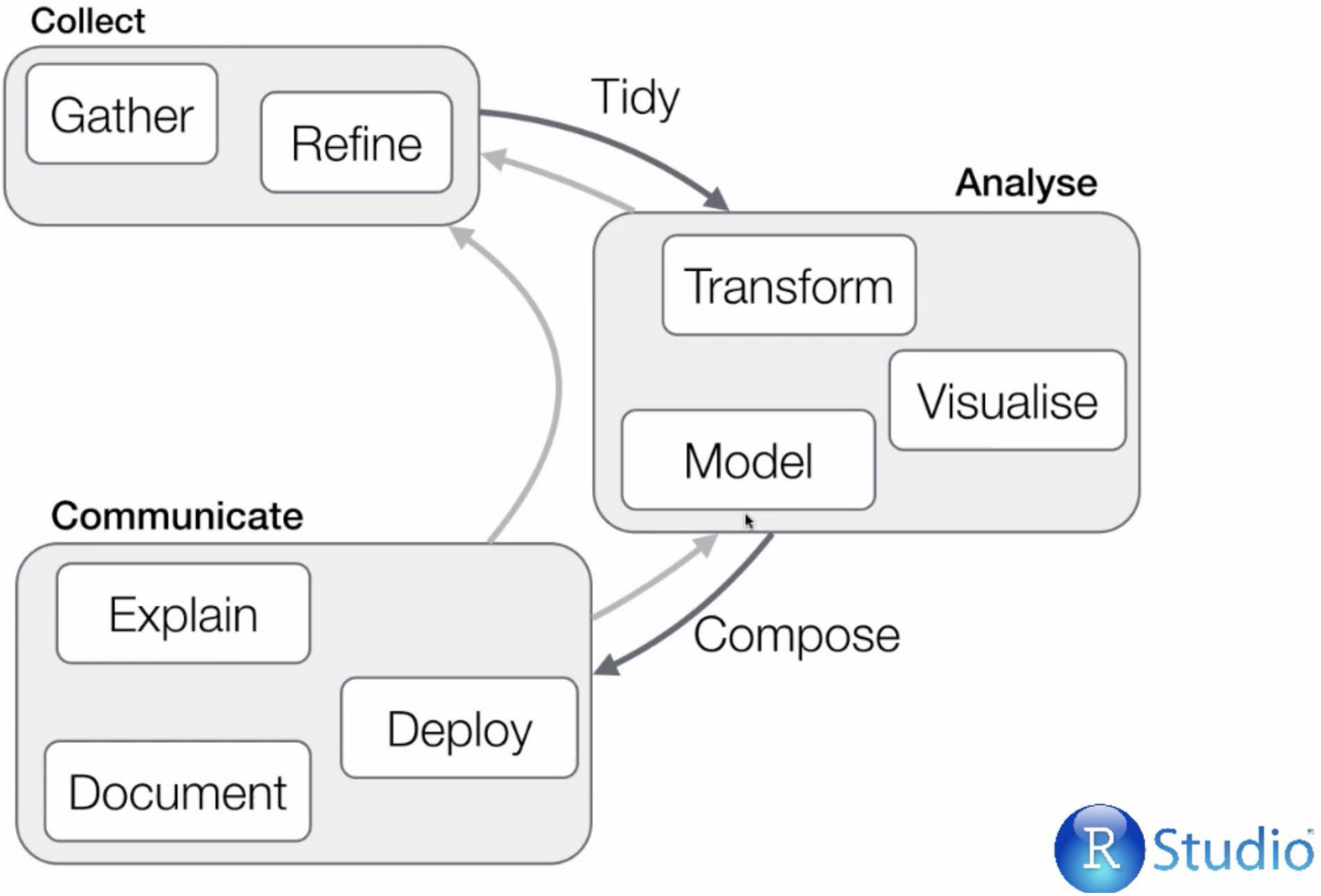


"Only a small fraction of real-world ML systems is composed of the ML code, as shown by the small black box in the middle. The required surrounding infrastructure is vast and complex."

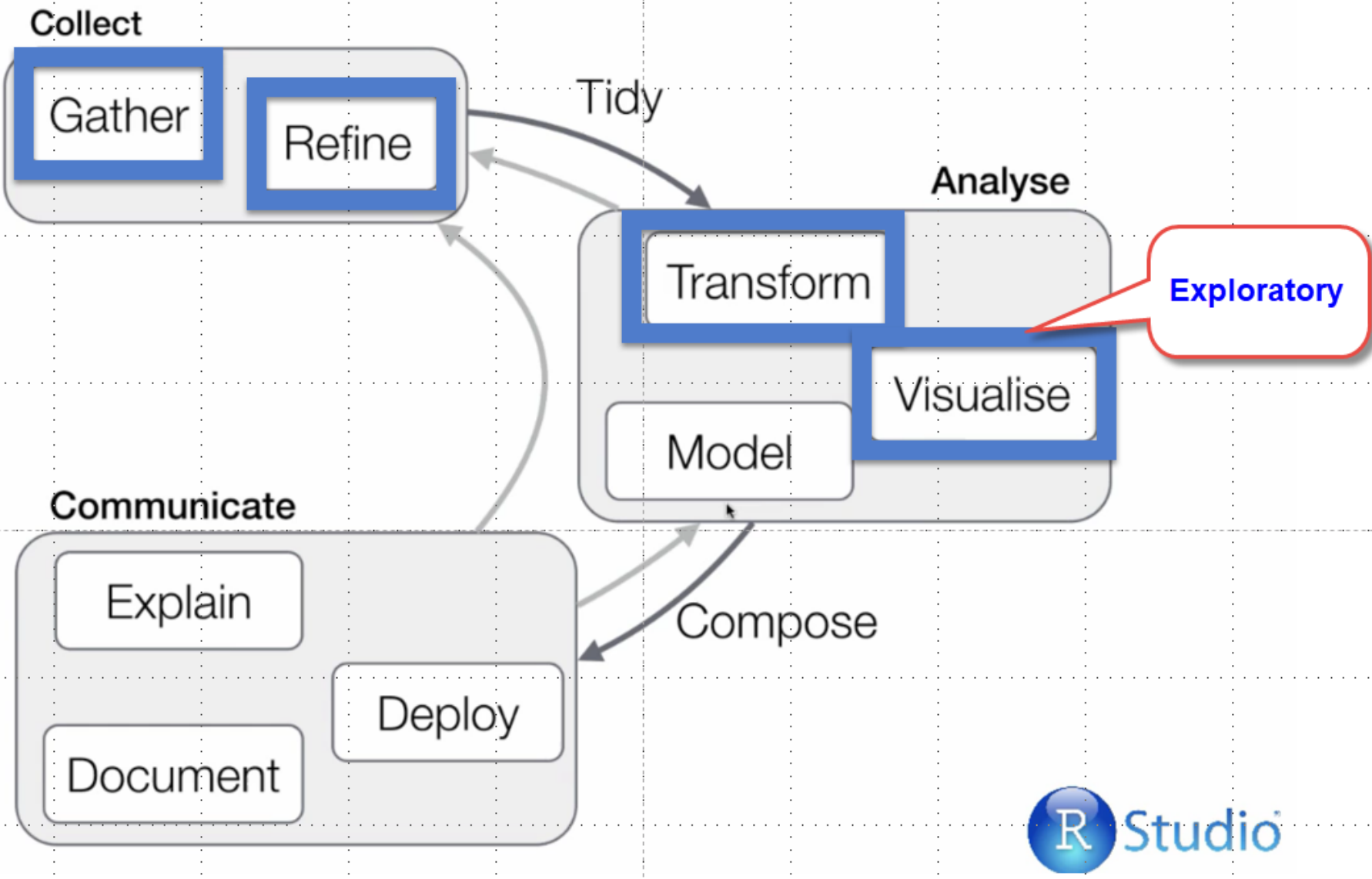
Source: D. Sculley et.al., "Hidden Technical Debt in Machine Learning Systems," NIPS'15 Proceedings of the 28th International Conference on Neural Information Processing Systems, Pages 2503-2511, Dec 7, 2015. https://papers.nips.cc/paper/5656-hidden-technical-debt-in-machine-learning-systems.pdf.



Presentation by Hadley Wickham on “The Grammar and Graphics of Data Science”, https://www.rstudio.com/resources/webinars/archives/

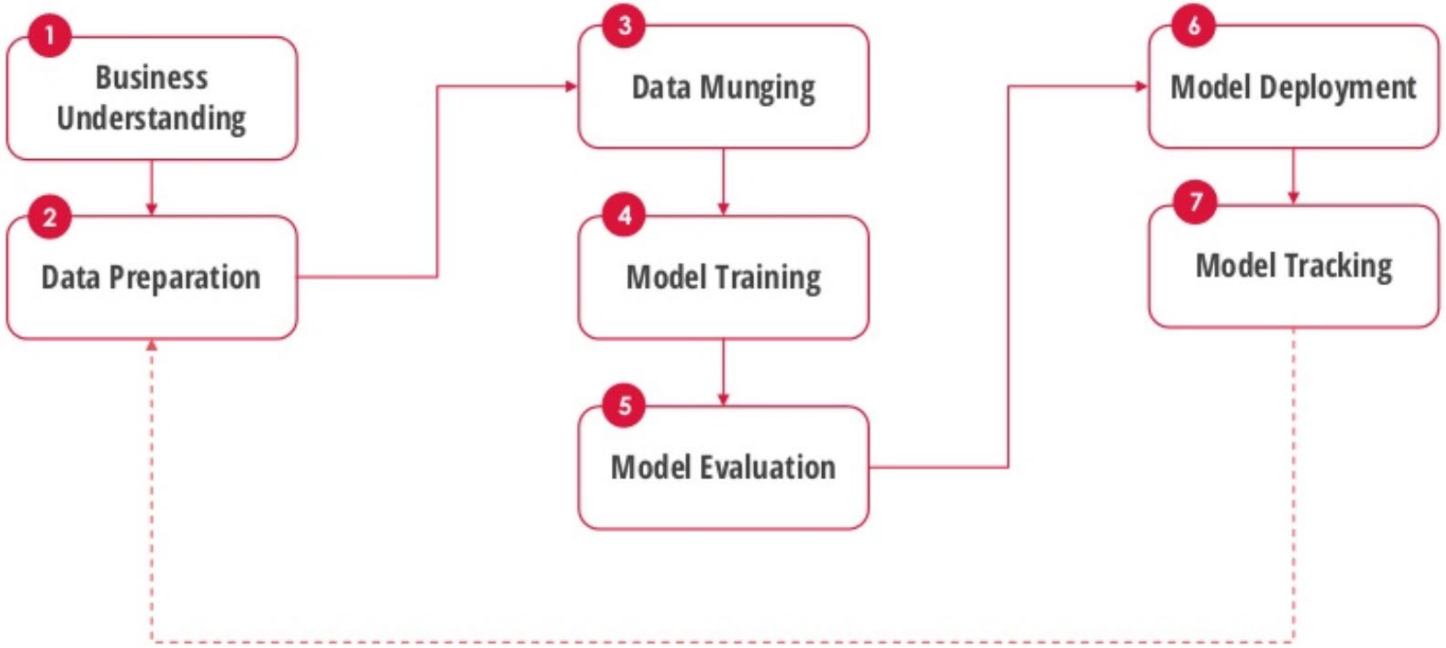


Presentation by Hadley Wickham on “The Grammar and Graphics of Data Science”, https://www.rstudio.com/resources/webinars/archives/

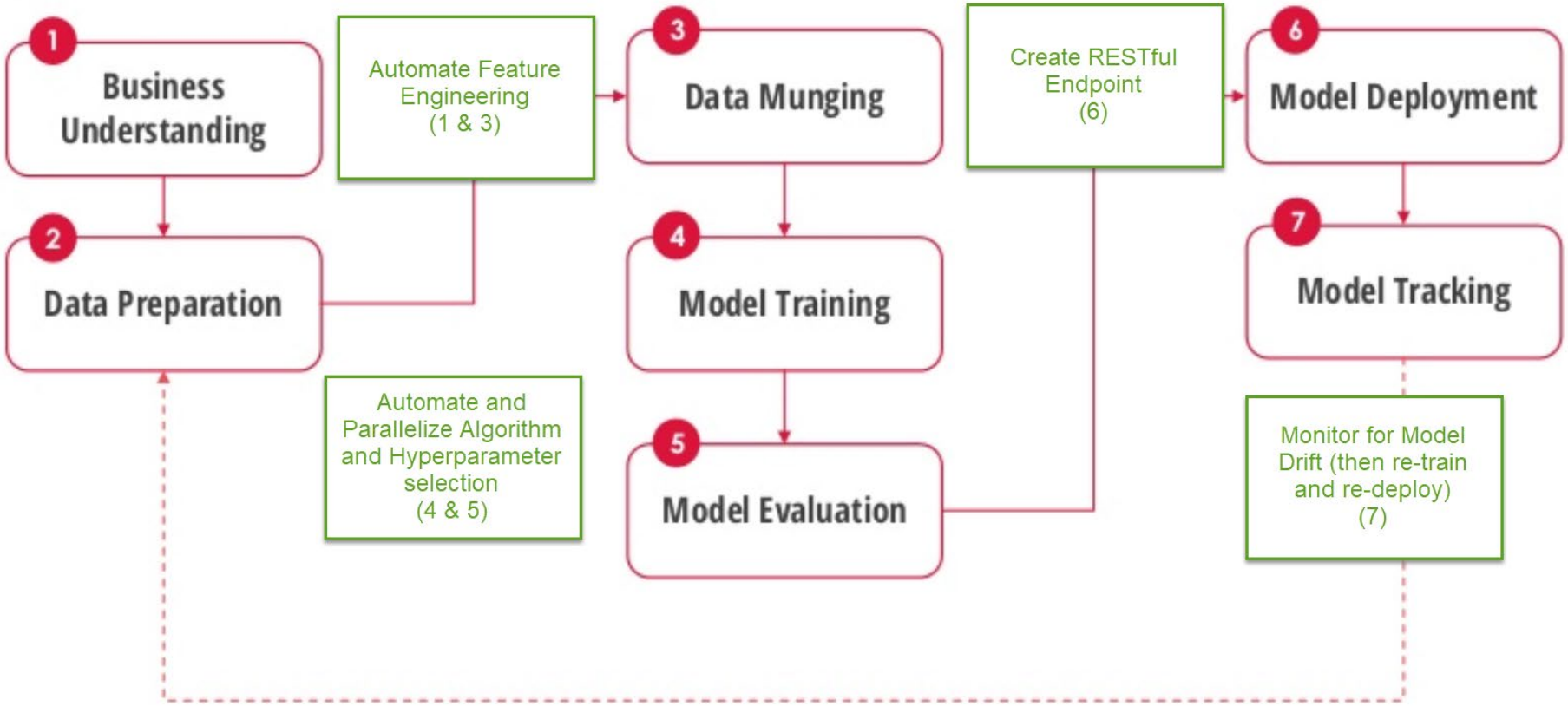


Presentation by Hadley Wickham on "The Grammar and Graphics of Data Science", https://www.rstudio.com/resources/webinars/archives/

Data Science Lifecycle: Processes



Source: Patel [2018]


1
Business Understanding
2
Data Preparation
Automate Feature Engineering (1 & 3)
Automate and Parallelize Algorithm and Hyperparameter selection (4 & 5)
3
Data Munging
4
Model Training
5
Model Evaluation
Create RESTful Endpoint (6)
6
Model Deployment
7
Model Tracking
Monitor for Model Drift (then re-train and re-deploy) (7)

The New York Times

TECHNOLOGY

For Big-Data Scientists, 'Janitor Work' Is Key Hurdle to Insight

By STEVE LOHR AUG. 17, 2014

Monica Rogati, Jawbone's vice president for data science, with Brian Wilt, a senior data scientist.
Peter DaSilva for The New York Times

Statistical Science
2001, Vol. 16, No. 3, 199–231

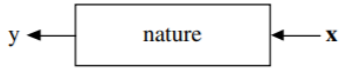
Statistical Modeling: The Two Cultures

Leo Breiman

Abstract. There are two cultures in the use of statistical modeling to reach conclusions from data. One assumes that the data are generated by a given stochastic data model. The other uses algorithmic models and treats the data mechanism as unknown. The statistical community has been committed to the almost exclusive use of data models. This commitment has led to irrelevant theory, questionable conclusions, and has kept statisticians from working on a large range of interesting current problems. Algorithmic modeling, both in theory and practice, has developed rapidly in fields outside statistics. It can be used both on large complex data sets and as a more accurate and informative alternative to data modeling on smaller data sets. If our goal as a field is to use data to solve problems, then we need to move away from exclusive dependence on data models and adopt a more diverse set of tools.

1. INTRODUCTION

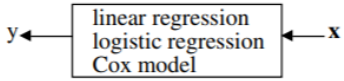
Statistics starts with data. Think of the data as being generated by a black box in which a vector of input variables **x** (independent variables) go in one side, and on the other side the response variables **y** come out. Inside the black box, nature functions to associate the predictor variables with the response variables, so the picture is like this:



There are two goals in analyzing the data:

Prediction. To be able to predict what the responses are going to be to future input variables;
Information. To extract some information about

The values of the parameters are estimated from the data and the model then used for information and/or prediction. Thus the black box is filled in like this:



Model validation. Yes–no using goodness-of-fit tests and residual examination.
Estimated culture population. 98% of all statisticians.

The Algorithmic Modeling Culture

The analysis in this culture considers the inside of the box complex and unknown. Their approach is to

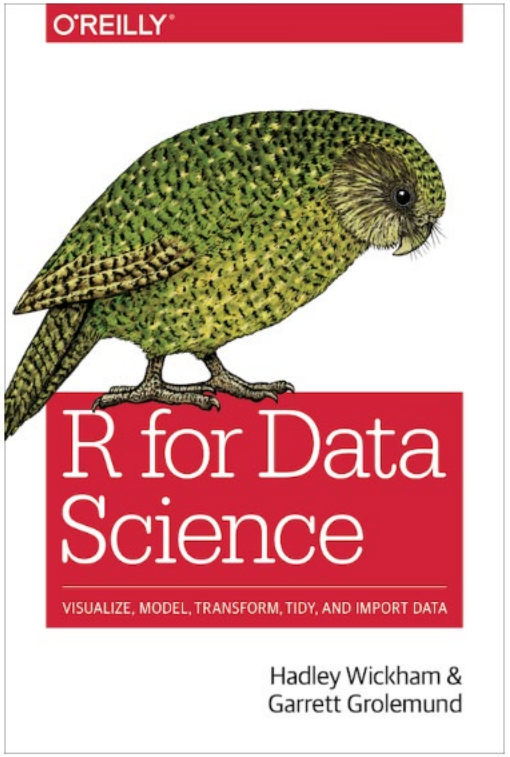
What are the different “flavors” of R?

Base R
Tidyverse
Other good choices:

Examples

- apply verbs vs. dplyr (and plyr) vs. data.table
- Base R graphics vs. ggplot2 vs. lattice

Which do you need to know?



Topics

- Course Overview – Course Site, assignments, resources
- Data Ethics: Polarization and Privacy
- The Data Science Lifecycle
- **Demonstration: GitHub**

Be scientists!

"Machine Learning is [a] mix of hard science and cooking recipes."
-- Julien Simon, AWS

Sources: Priyadarshini [2019] and Simon [2019]