Opinionated Analysis Development

Hilary Parker
@hspter

STITCH FIX

Etsy

Not So Standard Deviations

Hilary Parker
@hspter

Been thinking on this -- we're missing a word for analysts who write careful/reproducible code for analysis, but aren't package developers

RETWEETS	LIKES
13	45

9:07 AM - 25 Nov 2015

Hilary Parker @hspter · 25 Nov 2015

Been thinking on this -- we're missing a word for analysts who write careful/reproducible code for analysis, but aren't package developers

22 13 45

Hilary Parker
@hspter

What about "analysis developer"? Any other ideas?

RETWEET	LIKES
1	4

9:07 AM - 25 Nov 2015

Peadar Coyle
@Springcoil

@hspter Is that different than data scientist?

LIKE
1

9:10 AM - 25 Nov 2015

MODERN DATA SCIENTIST

Data Scientist, the sexiest job of the 21th century, requires a mixture of multidisciplinary skills ranging from an intersection of mathematics, statistics, computer science, communication and business. Finding a data scientist is hard. Finding people who understand who a data scientist is, is equally hard. So here is a little cheat sheet on who the modern data scientist really is.

MATH & STATISTICS

- Machine learning
- Statistical modeling
- Experiment design
- Bayesian inference
- Supervised learning: decision trees, random forests, logistic regression
- Unsupervised learning: clustering, dimensionality reduction
- Optimization: gradient descent and variants

PROGRAMMING & DATABASE

- Computer science fundamentals
- Scripting language e.g. Python
- Statistical computing packages, e.g., R
- Databases: SQL and NoSQL
- Relational algebra
- Parallel databases and parallel query processing
- MapReduce concepts
- Hadoop and Hive/Pig
- Custom reducers
- Experience with xaaS like AWS

DOMAIN KNOWLEDGE & SOFT SKILLS

- Passionate about the business
- Curious about data
- Influence without authority
- Hacker mindset
- Problem solver
- Strategic, proactive, creative, innovative and collaborative

COMMUNICATION & VISUALIZATION

- Able to engage with senior management
- Story telling skills
- Translate data-driven insights into decisions and actions
- Visual art design
- R packages like ggplot or lattice
- Knowledge of any of visualization tools e.g. Flare, D3.js, Tableau

isomorphismes
@isomorphisms

@hspter reproducible-research er ?

3:28 PM - 25 Nov 2015




Reproducible

Accurate

Collaborative

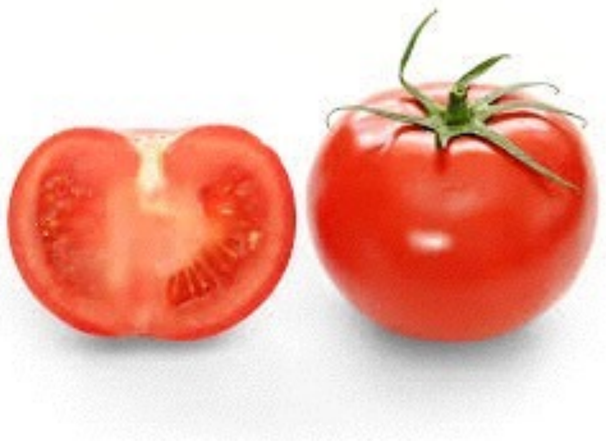
Hadley Wickham
@hadleywickham

@hspter data analysis engineer?

Translate from Dutch

LIKE
1

12:10 PM - 25 Nov 2015



Software developer

Software engineer

Development

Narrative

- What models are you going to use?
- What argument are you going to make?
- How are you going to convince your audience of something?

Technical Artifact

- What tools are you going to use to make the deliverable?
- What technical and coding choices will you make for producing the deliverable?

Karl Broman
@kwbroman

@hspter How about “good analyst”. 😉

LIKES
7

10:02 AM - 25 Nov 2015

Karl Broman
@kwbroman

@hspter How about “good analyst”. 😉

LIKES
7

10:02 AM - 25 Nov 2015

disclaimer: Karl is extremely nice, helpful & non-judgmental

Bad analyst? :(

Hilary Parker @hspter · 25 Nov 2015

Been thinking on this -- we're missing a word for analysts who write careful/reproducible code for analysis, but aren't package developers

22 13 45

Hilary Parker
@hspter

What about "analysis developer"? Any other ideas?

RETWEET	LIKES
1	4

9:07 AM - 25 Nov 2015

Hilary Parker @hspter · 25 Nov 2015

Been thinking on this -- we're missing a word for analysts who write careful/reproducible code for analysis, but aren't package developers

22 13 45

Hilary Parker
@hspter

What about "analysis developer" ment? Any other ideas?

RETWEET	LIKES
1	4

9:07 AM - 25 Nov 2015

Why?

- Creating an analysis is a hard, error-ridden process that we yada-yada away right now

Why?

- Creating an analysis is a hard, error-ridden process that we yada-yada away right now
  - don’t want to “limit creativity”
  - embarrassed by personal process

Why?

- Creating an analysis is a hard, error-ridden process that we yada-yada away right now
  - don’t want to “limit creativity”
  - embarrassed by personal process

Why?

- There are known, common problems when creating an analysis (as well as solutions to these problems)

- Making these easy to avoid frees up time for creativity

You re-run the analysis and get different results.

Someone else can’t repeat the analysis.

You can’t re-run the analysis on different data.

An external library you’re using is updated, and you can’t recreate the results.

You change code but don’t re-run downstream code, so the results aren’t consistently updated.

You change code but don’t re-execute it, so the results are out of date.

You update your analysis and can’t compare it to previous results.

You can’t point to code changes that resulted in a different analysis results.

A second analyst can’t understand your code.

Can you re-use logic in different parts of the analysis?

You change the logic used in analysis, but only in some places where it’s implemented.

Your code is not performing as expected, but you don’t notice.

Your data becomes corrupted, but you don’t notice.

You use a new dataset that renders your statistical methods invalid, but you don’t notice.

You make a mistake in your code.

You use inefficient code.

A second analyst wants to contribute code to the analysis, but can’t do so.

Two analysts want to combine code but cannot.

You aren’t able to track and communicate known next steps in your analysis.

Your collaborators can only make requests in email or meetings, and they aren’t incorporated into the project.

Reproducible	You re-run the analysis and get different results.
	Someone else can't repeat the analysis.
	You can't re-run the analysis on different data.
	An external library you're using is updated, and you can't recreate the results.
	You change code but don't re-run downstream code, so the results aren't consistently updated.
	You change code but don't re-execute it, so the results are out of date.
	You update your analysis and can't compare it to previous results.
	You can't point to code changes that resulted in a different analysis results.
	A second analyst can't understand your code.
Accurate	Can you re-use logic in different parts of the analysis?
	You change the logic used in analysis, but only in some places where it's implemented.
	Your code is not performing as expected, but you don't notice.
	Your data becomes corrupted, but you don't notice.
	You use a new dataset that renders your statistical methods invalid, but you don't notice.
	You make a mistake in your code.
	You use inefficient code.
Collaborative	A second analyst wants to contribute code to the analysis, but can't do so.
	Two analysts want to combine code but cannot.
	You aren't able to track and communicate known next steps in your analysis.
	Your collaborators can only make requests in email or meetings, and they aren't incorporated into the project.

Why?

- We're doing a disservice to people by not defining this process, because it leaves a lot of space for "personal failure / human error"

Human Error

- Some helpful things we would borrow from operations

The Field Guide to Understanding 'Human Error'

Sidney Dekker

An Ashgate Book

THIRD EDITION

The Field Guide to Understanding Human Error

- Switch from “blaming” a person to blaming the process that they used

The Field Guide to Understanding Human Error

- The person did not want to make an error, and acted in a way that she thought wouldn’t create an error

- the current system failed the person with good intentions

Etsy

Etsy's Winning Secret: Don't Play The Blame Game!

Owen Thomas

May 15, 2012, 5:42 PM 11,788

FACEBOOK LINKEDIN TWITTER EMAIL PRINT

Etsy Code as Craft Speaker Series Events About Archive

Blameless PostMortems and a Just Culture

Posted by **John Allspaw** on May 22, 2012

Blameless post-mortem

- When an error happens, go through the technical aspects of how the error happened without assigning fault to the practitioner

Error Happens

Error Happens

Blameless postmortem:
discuss system changes to prevent error
weigh cost of implementing these changes

Error Happens

Blameless postmortem:
discuss system changes to prevent error
weigh cost of implementing these changes

Adopt new process

“Blameful culture”

“An engineer who thinks they’re going to be reprimanded is disincentivized to give the details necessary to get an understanding of the mechanism, pathology, and operation of the failure.”

https://codeascraft.com/2012/05/22/blameless-postmortems/

You re-run the analysis and get different results.

Someone else can’t repeat the analysis.

You can’t re-run the analysis on different data.

An external library you’re using is updated, and you can’t recreate the results.

You change code but don’t re-run downstream code, so the results aren’t consistently updated.

You change code but don’t re-execute it, so the results are out of date.

You update your analysis and can’t compare it to previous results.

You can’t point to code changes that resulted in a different analysis results.

A second analyst can’t understand your code.

Can you re-use logic in different parts of the analysis?

You change the logic used in analysis, but only in some places where it’s implemented.

Your code is not performing as expected, but you don’t notice.

Your data becomes corrupted, but you don’t notice.

You use a new dataset that renders your statistical methods invalid, but you don’t notice.

You make a mistake in your code.

You use inefficient code.

A second analyst wants to contribute code to the analysis, but can’t do so.

Two analysts want to combine code but cannot.

You aren’t able to track and communicate known next steps in your analysis.

Your collaborators can only make requests in email or meetings, and they aren’t incorporated into the project.

Error Happens

Blameless postmortem:
discuss system changes to prevent error
weigh cost of implementing these changes

Adopt new process

Executable analysis scripts	You re-run the analysis and get different results.
	Someone else can't repeat the analysis.
	You can't re-run the analysis on different data.
Defined dependencies	An external library you're using is updated, and you can't recreate the results.
	You change code but don't re-run downstream code, so the results aren't consistently updated.
Watchers for changed code and data	You change code but don't re-execute it, so the results are out of date.
Version control (individual)	You update your analysis and can't compare it to previous results.
	You can't point to code changes that resulted in a different analysis results.
Code review	A second analyst can't understand your code.
Modular, tested code	Can you re-use logic in different parts of the analysis?
	You change the logic used in analysis, but only in some places where it's implemented.
	Your code is not performing as expected, but you don't notice.
Assertive testing of data,	Your data becomes corrupted, but you don't notice.
assumptions and results	You use a new dataset that renders your statistical methods invalid, but you don't notice.
Code review	You make a mistake in your code.
	You use inefficient code.
Version Control (collaborative)	A second analyst wants to contribute code to the analysis, but can't do so.
	Two analysts want to combine code but cannot.
Issue tracking	You aren't able to track and communicate known next steps in your analysis.
	Your collaborators can only make requests in email or meetings, and they aren't incorporated into the project.

“Opinions”

Executable analysis scripts	You re-run the analysis and get different results.
	Someone else can’t repeat the analysis.
	You can’t re-run the analysis on different data.
Defined dependencies	An external library you’re using is updated, and you can’t recreate the results.
	You change code but don’t re-run downstream code, so the results aren’t consistently updated.
Watchers for changed code and data	You change code but don’t re-execute it, so the results are out of date.
Version control (individual)	You update your analysis and can’t compare it to previous results.
	You can’t point to code changes that resulted in a different analysis results.
Code review	A second analyst can’t understand your code.
Modular, tested code	Can you re-use logic in different parts of the analysis?
	You change the logic used in analysis, but only in some places where it’s implemented.
	Your code is not performing as expected, but you don’t notice.
Assertive testing of data,	Your data becomes corrupted, but you don’t notice.
assumptions and results	You use a new dataset that renders your statistical methods invalid, but you don’t notice.
Code review	You make a mistake in your code.
	You use inefficient code.
Version Control (collaborative)	A second analyst wants to contribute code to the analysis, but can’t do so.
	Two analysts want to combine code but cannot.
Issue tracking	You aren’t able to track and communicate known next steps in your analysis.
	Your collaborators can only make requests in email or meetings, and they aren’t incorporated into the project.

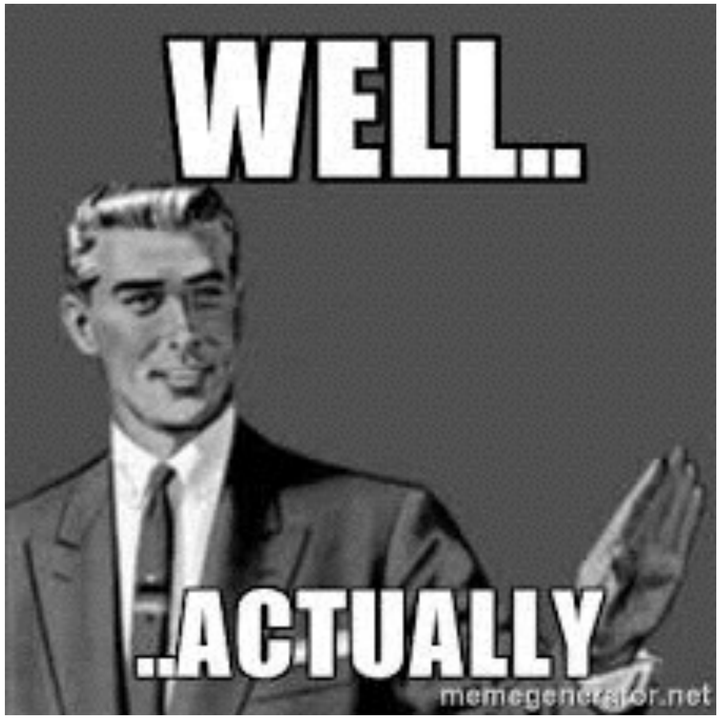
Opinionated

- 30+ years of experience at being opinionated

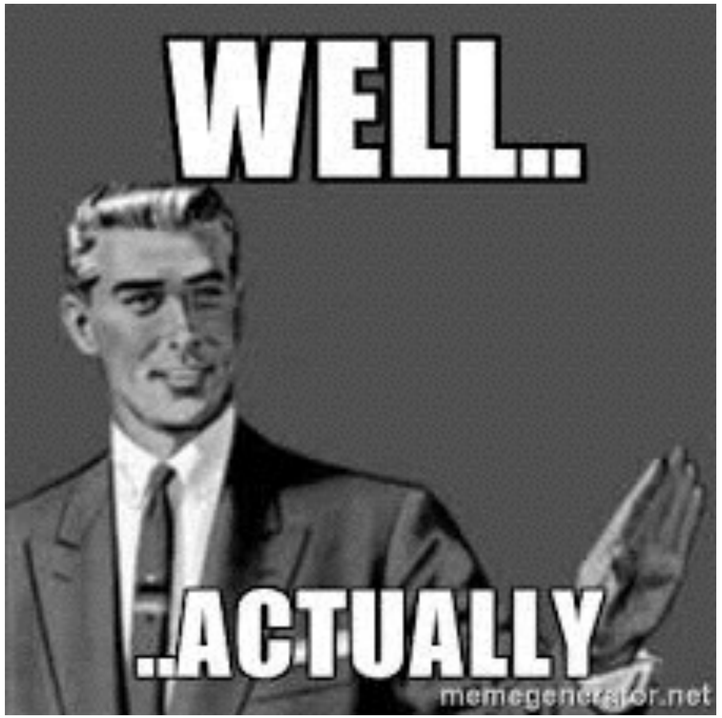
Opinionated

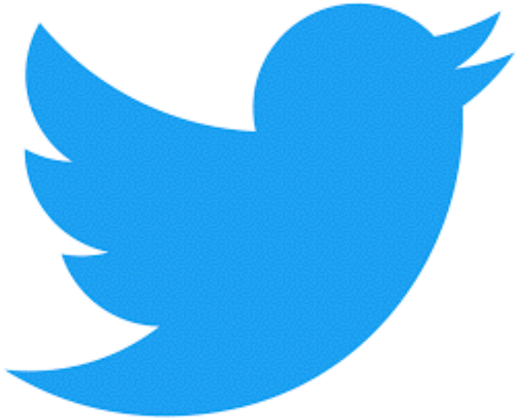
- “Hilary, I never fight with ***anyone else*** except for you” — a friend, exasperated, at his birthday party

How Not To Win Friends and Influence People

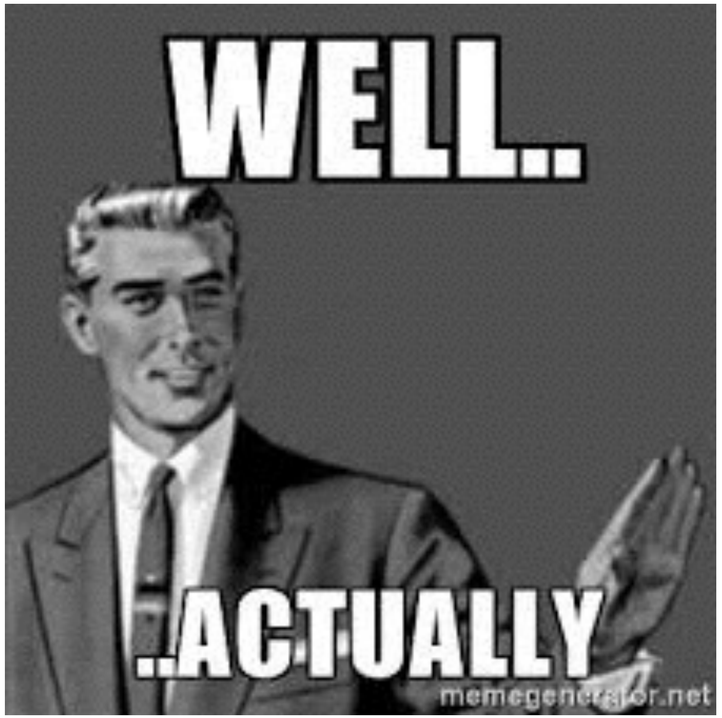


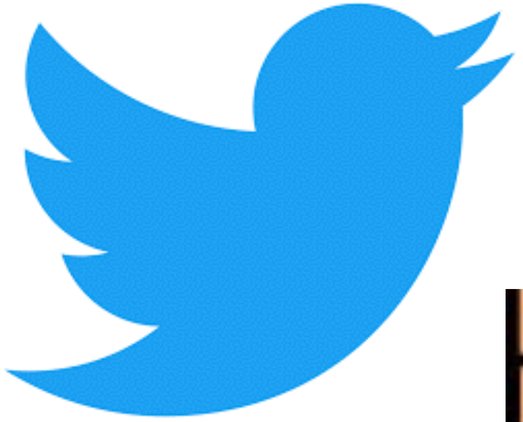
How Not To Win Friends and Influence People

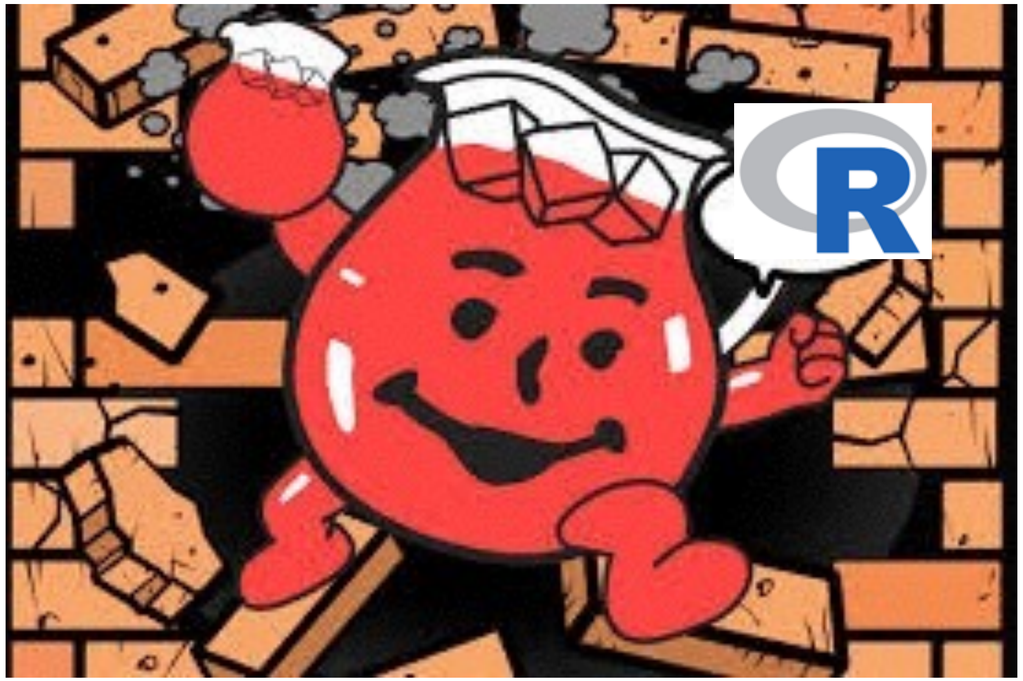




How Not To Win Friends and Influence People







Opinionated

- “A funny thing happens when engineers make mistakes and feel safe when giving details about it: they are not only willing to be held accountable, **they are also enthusiastic in helping the rest of the company avoid the same error in the future**. They are, after all, the most expert in their own error. They ought to be heavily involved in coming up with remediation items.”

https://codeascraft.com/2012/05/22/blameless-postmortems/

Opinionated Software

- Shift from “I know better than you” to “lots of people have run into this problem before, and here’s software that makes a solution easy”

Opinionated Software

“Opinionated Software is a software product that believes a certain way of approaching a business process is inherently better and provides software crafted around that approach.”

https://medium.com/@stueccles/the-rise-of-opinionated-software-ca1ba0140d5b#.by7aq2c3c

Opinionated Software

“It’s a strong disagreement with the conventional wisdom that everything should be configurable, that the framework should be impartial and objective. In my mind, that’s the same as saying that everything should be equally hard.”

— David Heinemeier Hansson

Non-opinionated software

This is bowling. There are rules.

This is ~~bowling.~~ There are rules.

analysis

This is ~~bowling.~~ There are rules.

analysis development

Opinionated Analysis Development

Define the process of technically creating the analysis

Define opinions based on common process errors /

- Shift blame away from individual practitioners using blameless postmortems

Push for software that makes it easy to implements opinions

Focus on creativity!

Opinionated Analysis Development

Define the process of technically creating the analysis

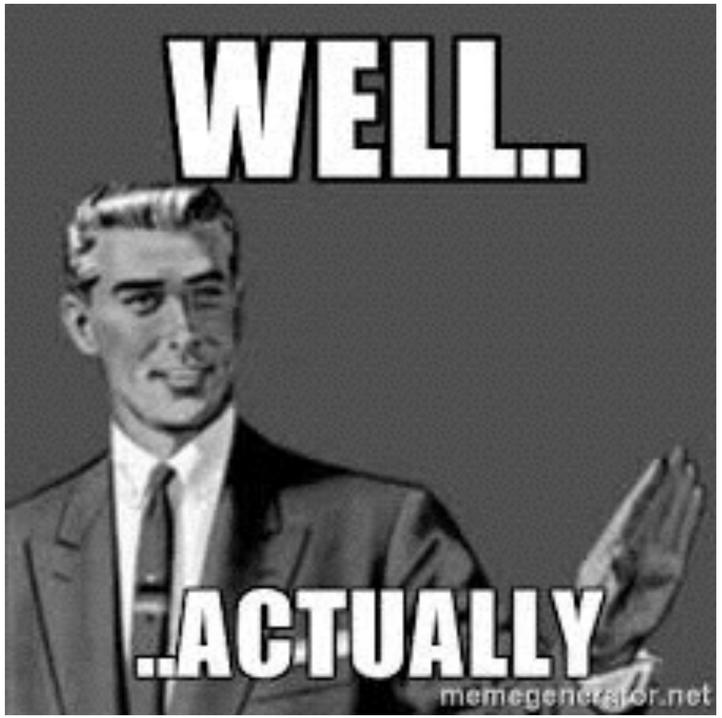
Define opinions based on common process errors /

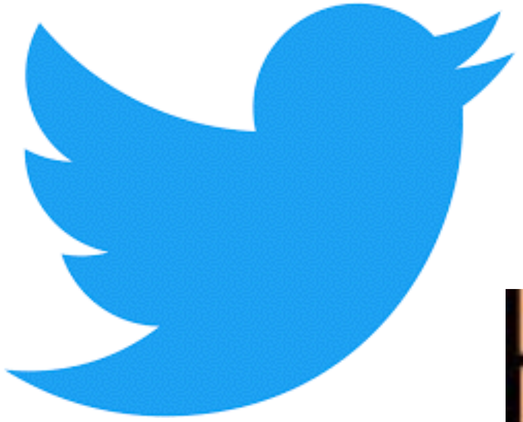
- Shift blame away from individual practitioners using blameless postmortems

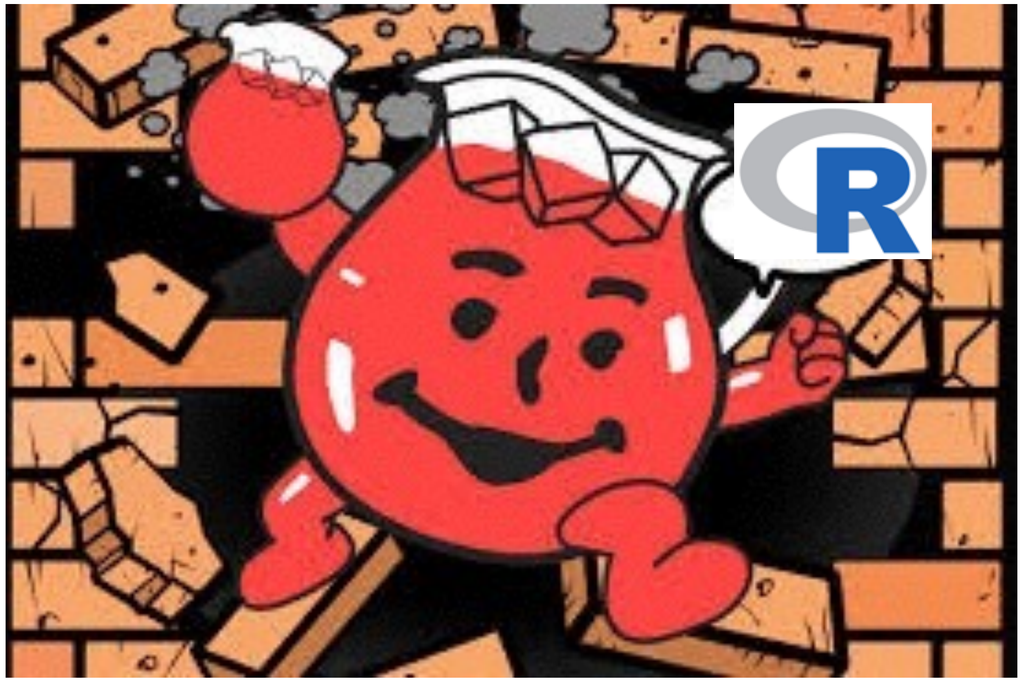
Push for software that makes it easy to implements opinions

Focus on creativity!

How Not To Win Friends and Influence People







Thanks!

Hilary Parker

@hspter