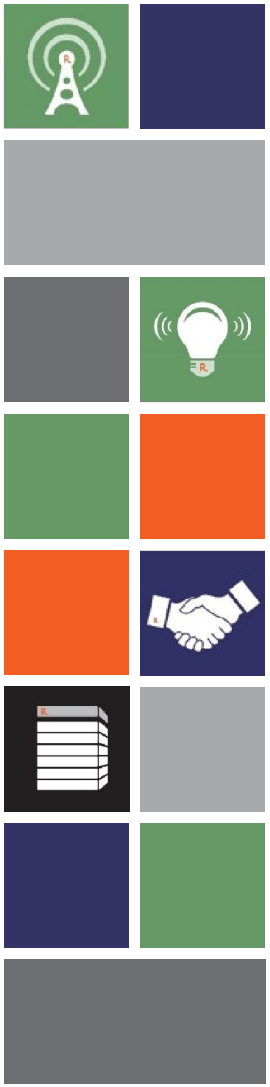
REVOLUTION ANALYTICS

How the growth of R helps data-driven organizations succeed

David Smith

Chief Community Officer

7th China R User Conference, May 2014



Agenda

- Introduction
- History of R
- Growth of R
- Applications of R
- The R Ecosystem
- Conclusion

Slides will be posted to:
blog.revolutionanalytics.com

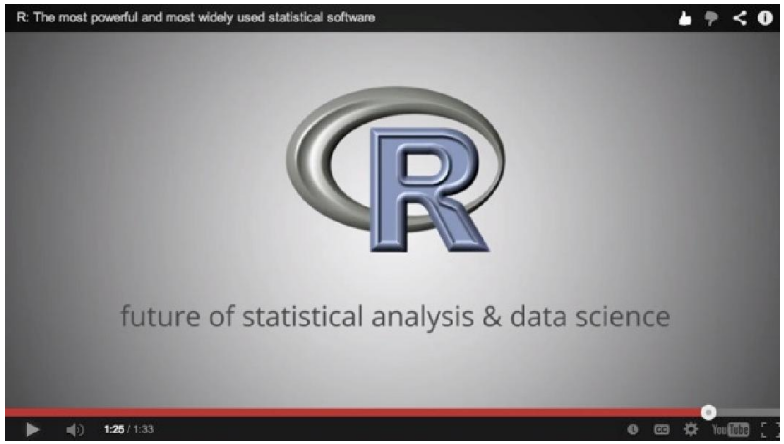
REVOLUTION ANALYTICS

What is R?

- Most widely used data analysis software
  - Used by 2M+ data scientists, statisticians and analysts
- Most powerful statistical programming language
  - Flexible, extensible and comprehensive for productivity
- Create beautiful and unique data visualizations
  - As seen in New York Times, Twitter and Flowing Data
- Thriving open-source community
  - Leading edge of analytics research
- Fills the talent gap
  - New graduates prefer R

R is Hot
bit.ly/r-is-hot

WHITE PAPER



REVOLUTION ANALYTICS

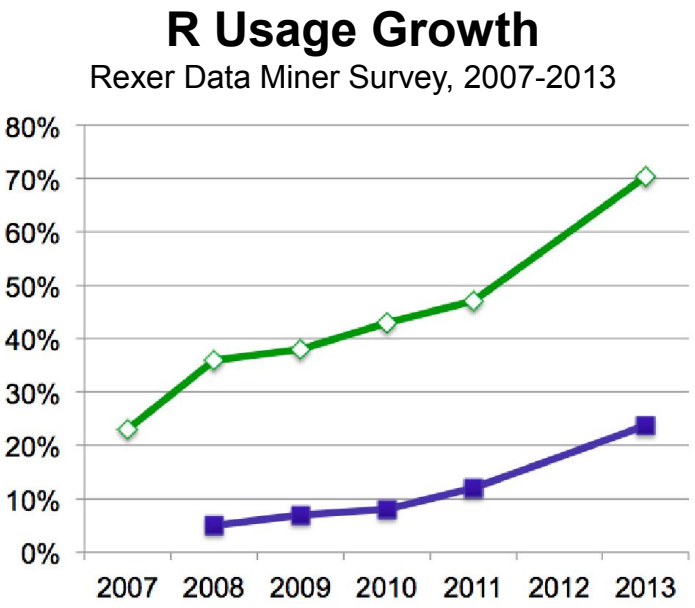
A brief history of R

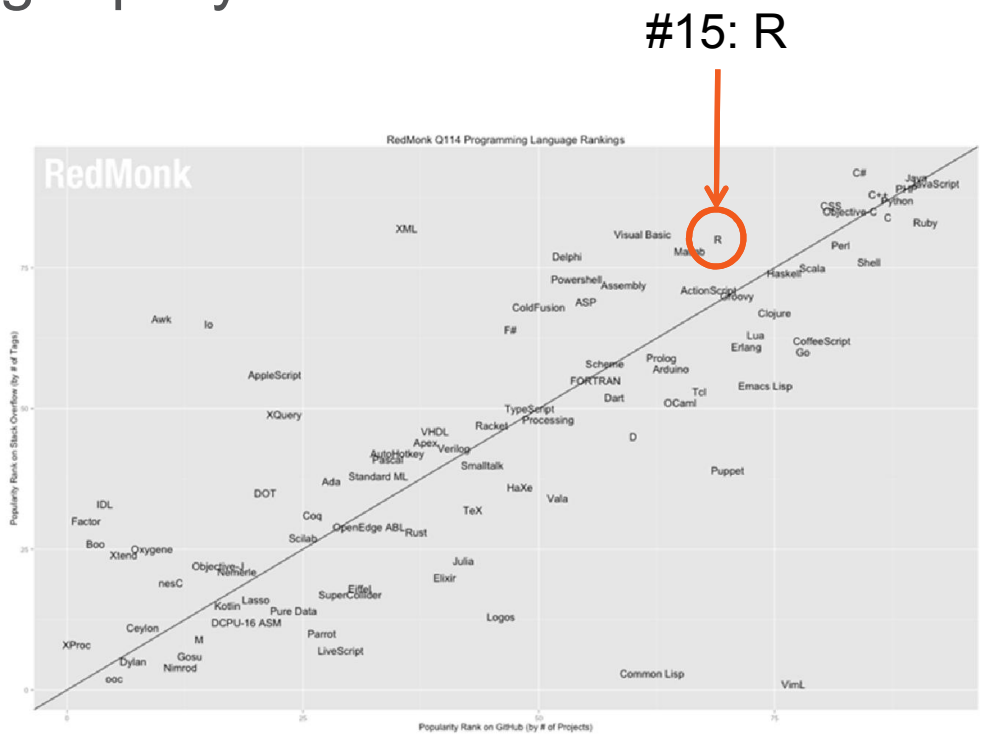
- **1993**: Research project in Auckland, NZ
  - Ross Ihaka and Robert Gentlemen
- **1995**: Released as open-source software
  - Generally compatible with the “S” language
- **1997**: R core group formed
- **2000**: R 1.0.0 released
- **2004**: R 2.0.0 released, first international user conference in Vienna
- **2009**: New York Times article on R
- **2013**: R 3.0.0 released

REVOLUTION ANALYTICS

New York Times:
Data Analysts Captivated by R's Power
6 Jan 2009

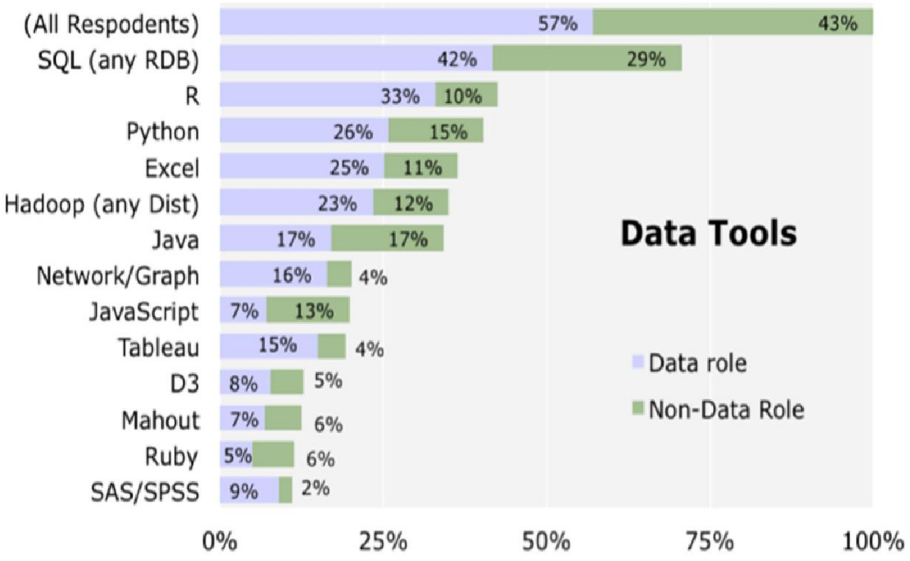
R's popularity is growing rapidly

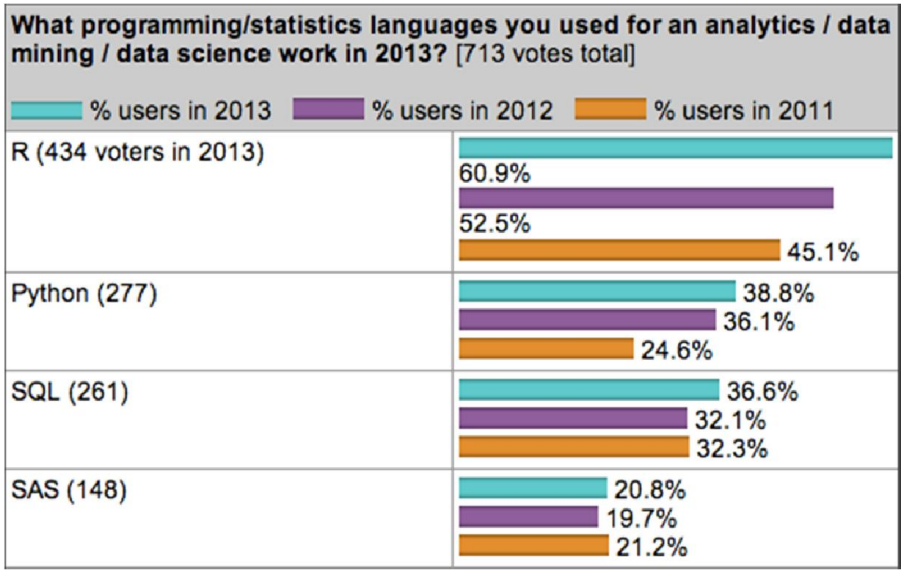




- [Rexer Data Miner Survey, 2013]
- [RedMonk Programming Language Rankings, 2013]

R is used more than other data science tools





- [O'Reilly Strata 2013 Data Science Salary Survey]
- [KDNuggets Poll: Top Languages for analytics, data mining, data science]

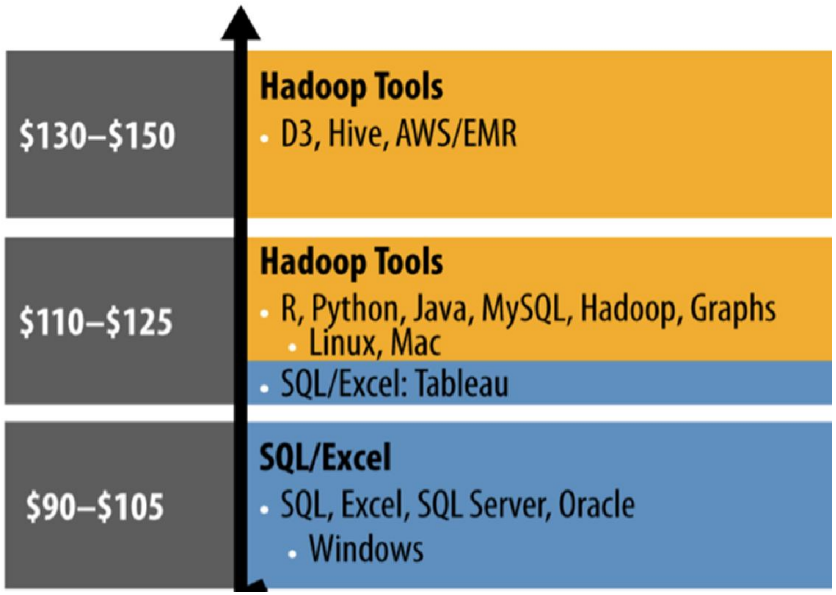
REVOLUTION ANALYTICS

R is among the highest-paid IT skills in the US

AVERAGE SALARY FOR High Paying Skills and Experience

SKILL	2013	YR/YR CHANGE
R	$ 115,531	n/a
NoSQL	$ 114,796	1.6%
MapReduce	$ 114,396	n/a
PMBok	$ 112,382	1.3%
Cassandra	$ 112,382	n/a
Omnigraffle	$ 111,039	0.3%
Pig	$ 109,561	n/a
SOA (Service Oriented Architecture)	$ 108,997	-0.5%
Hadoop	$ 108,669	-5.6%
Mongo DB	$ 107,825	-0.4%

- Dice Tech Salary Survey, January 2014

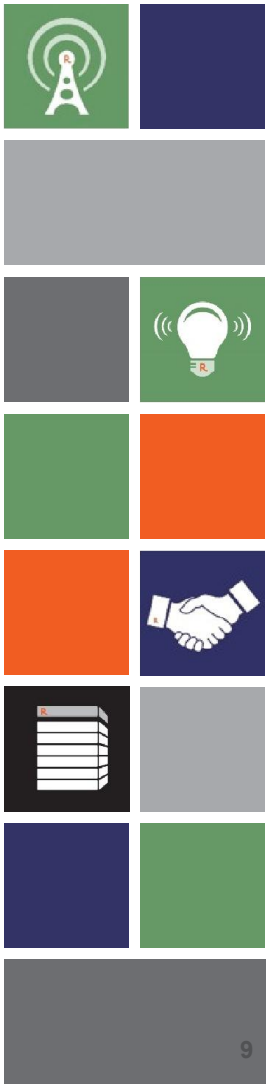


- O'Reilly Strata 2013 Data Science Salary Survey

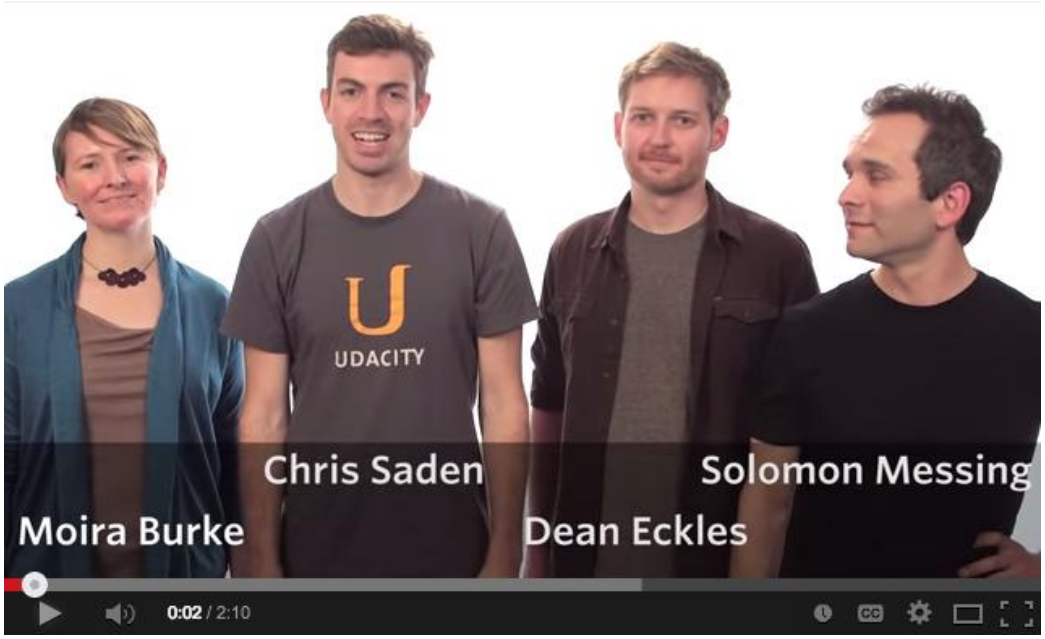
REVOLUTION ANALYTICS

REVOLUTION ANALYTICS

Applications of R



facebook



- Exploratory Data Analysis
- Experimental Analysis

"Generally, we use R to move fast when we get a new data set. With R, we don't need to develop custom tools or write a bunch of code. Instead, we can just go about cleaning and exploring the data."
— Solomon Messing, data scientist at Facebook

REVOLUTION ANALYTICS

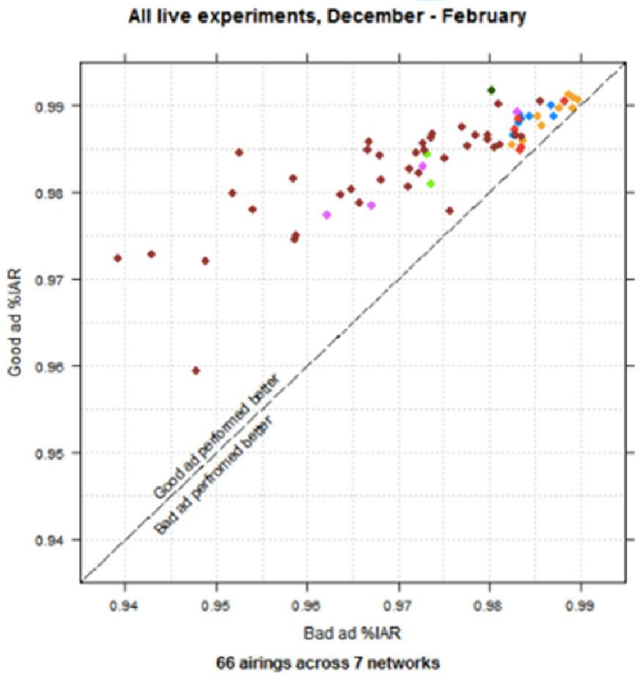
facebook

- Big-Data Visualization

REVOLUTION ANALYTICS

"It resonated with many people. It's not just a pretty picture, it's a reaffirmation of the impact we have in connecting people, even across oceans and borders." — Paul Butler, data scientist, Facebook

Google



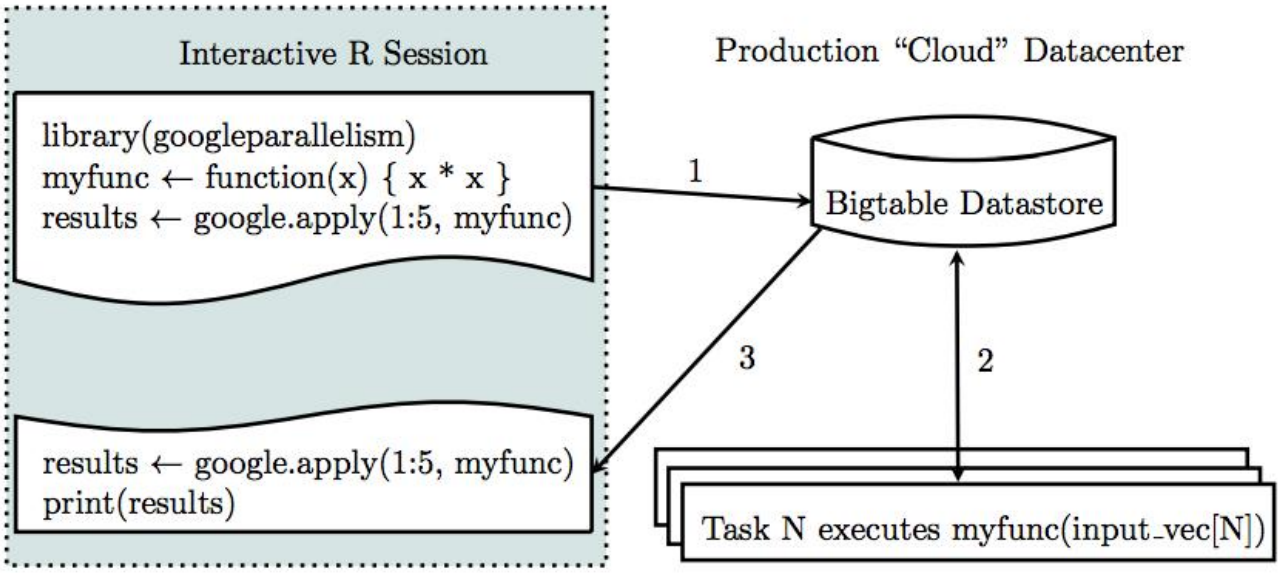
"R is really important to the point that it's hard to overvalue it."
— Daryl Pregibon Head of Statistics, Google

- Advertising Effectiveness

"The great beauty of R is that you can modify it to do all sorts of things."
— Hal Varian Chief Economist, Google

- Economic forecasting

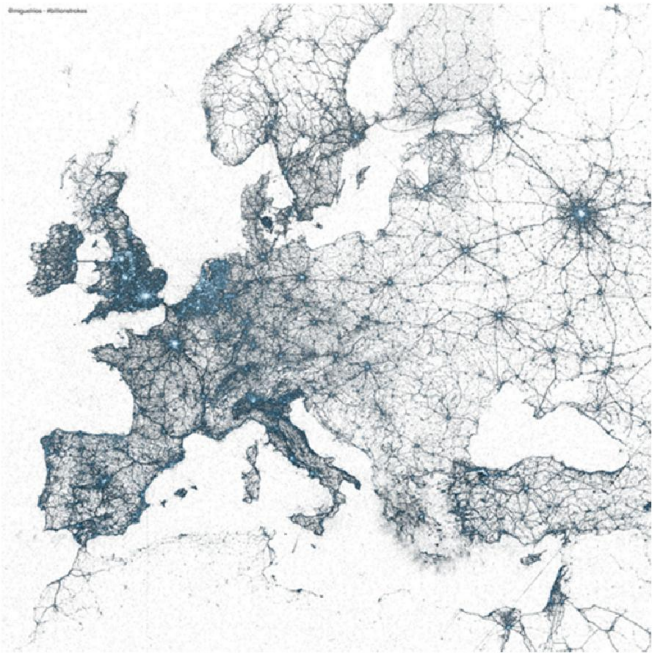
Google



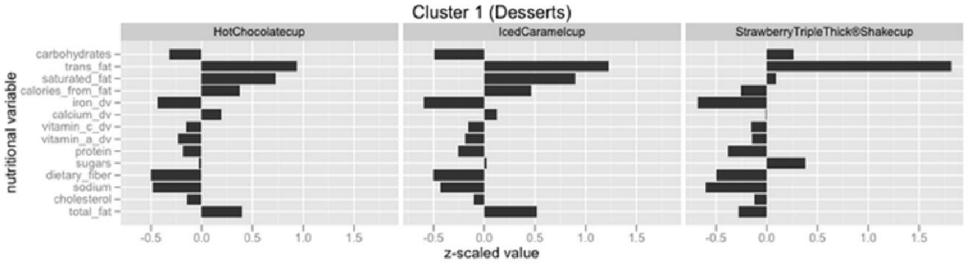
- Big-data statistical modeling (Large-Scale Parallel Statistical Forecasting Computations in R, JSM 2011)

"We demonstrate the utility of massively parallel computational infrastructure for statistical computing using the MapReduce paradigm for R. This framework allows users to write computations in a high-level language that are then broken up and distributed to worker tasks in Google datacenters."

twitter



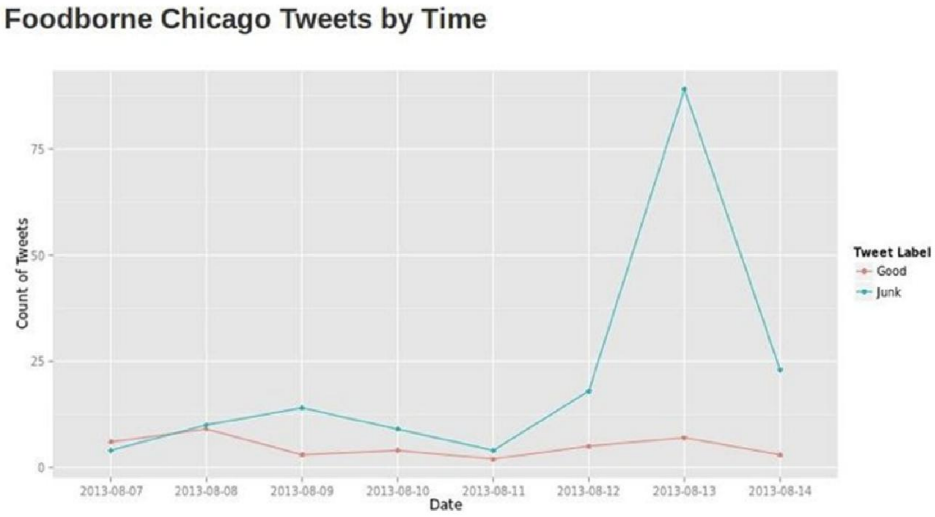
"A common pattern for me is that I'll code a MapReduce job in Scala, do some simple command-line munging on the results, pass the data into Python or R for further analysis, pull from a database to grab some extra fields, and so on, often integrating what I find into some machine learning models in the end" — Ed Chen, Data Scientist, Twitter



- Data Visualization
- Semantic clustering

REVOLUTION ANALYTICS

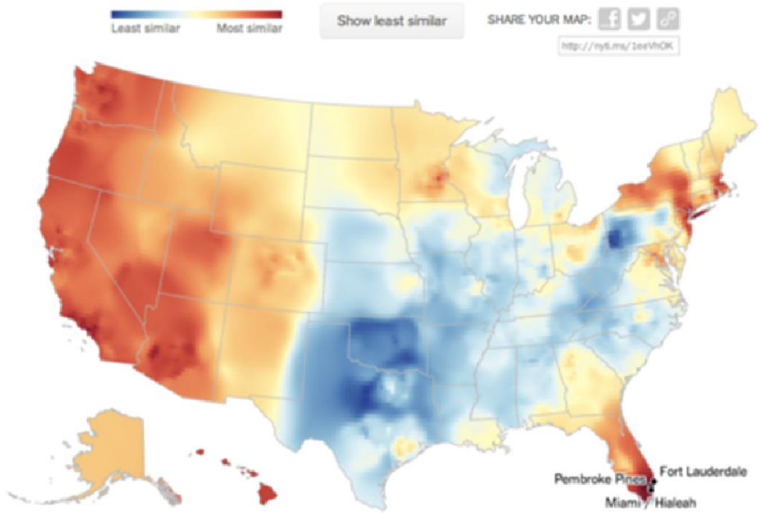
Public Health



- Food poisoning monitor

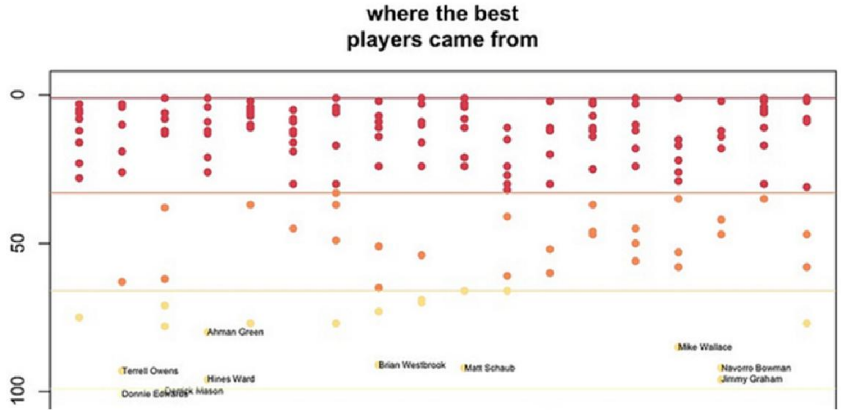
The New York Times

Interactive Features



- Election Forecast
- Dialect Quiz

Data Journalism

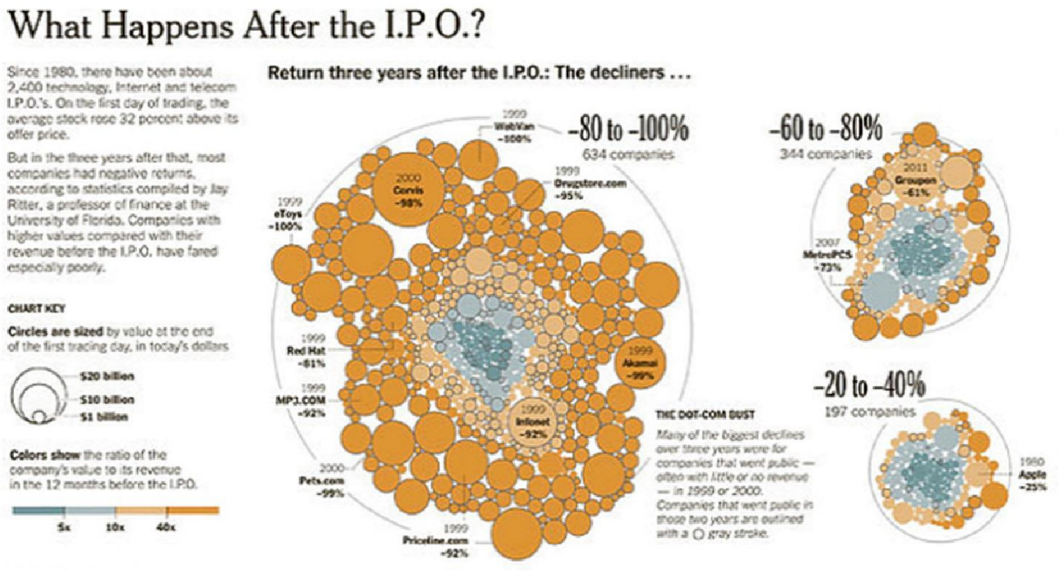


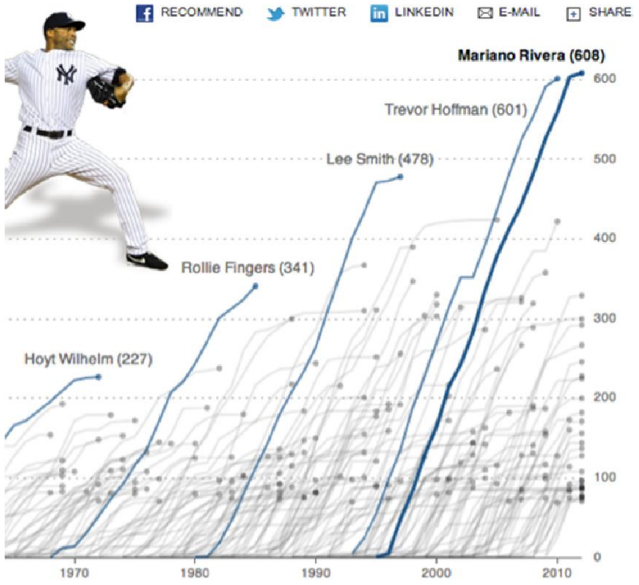
- NFL Draft Picks
- Wealth distribution in USA

REVOLUTION ANALYTICS

The New York Times

Data Visualization



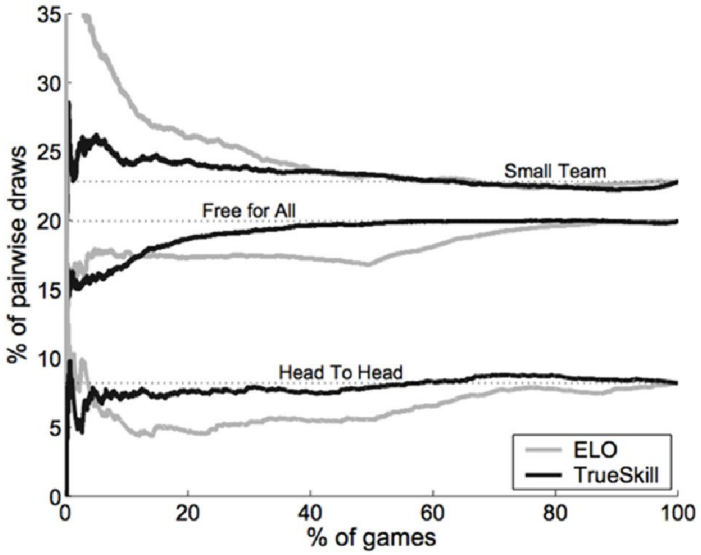


- Facebook IPO
- Baseball legends

REVOLUTION ANALYTICS

Video Games

XBOX ONE



- Multiplayer Matchmaking

scientific revenue

- Player Churn
- Game design
  - Difficulty curve
  - Level trouble-spots
- In-game purchase optimization
- Fraud detection
- Player communities

- Game Analysis

REVOLUTION ANALYTICS

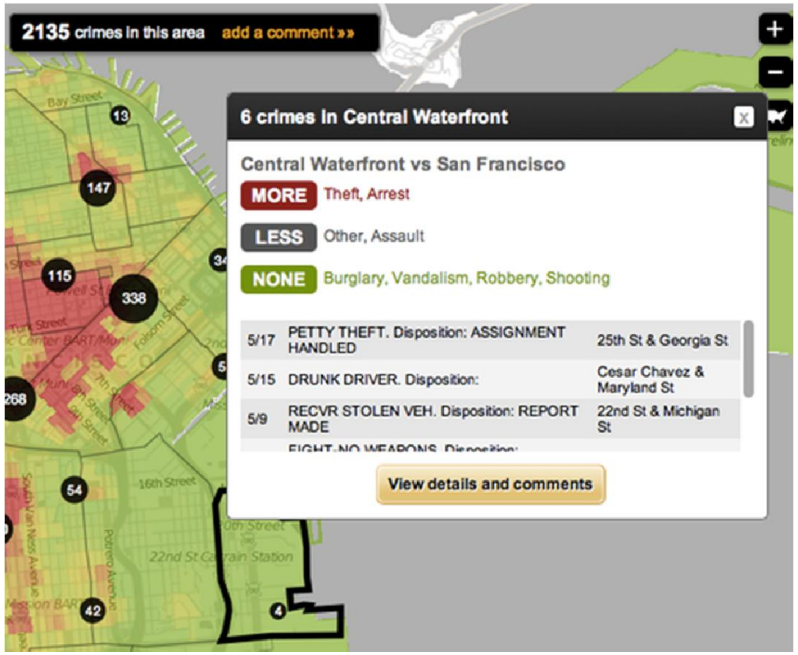
Real Estate

Zillow®
Your Edge in Real Estate

"The core innovation that Zillow offers are its advanced statistical predictive products, including the Zestimate®, the Rent Zestimate and the ZHVI® family of real estate indexes. By using R in production as well as research, Zillow maximizes flexibility and minimizes the latency in rolling out updates and new products."

- Statistical forecasting

trulia®



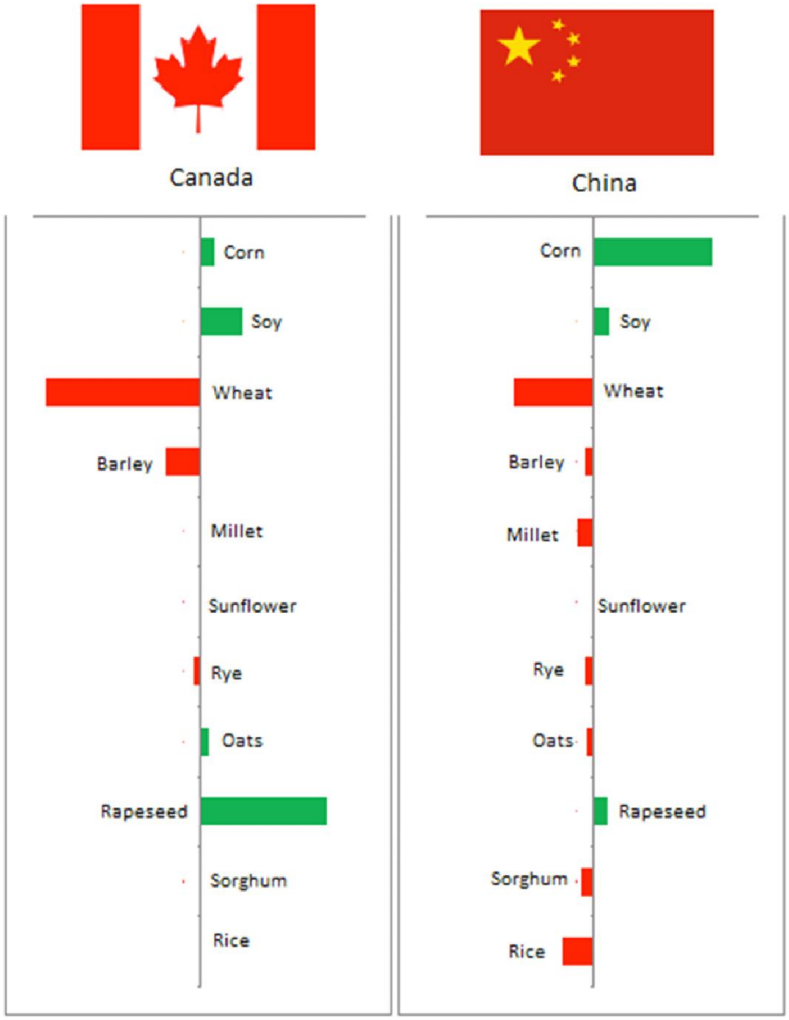
- Crime mapping

REVOLUTION ANALYTICS

JOHN DEERE

Statistical Analysis:

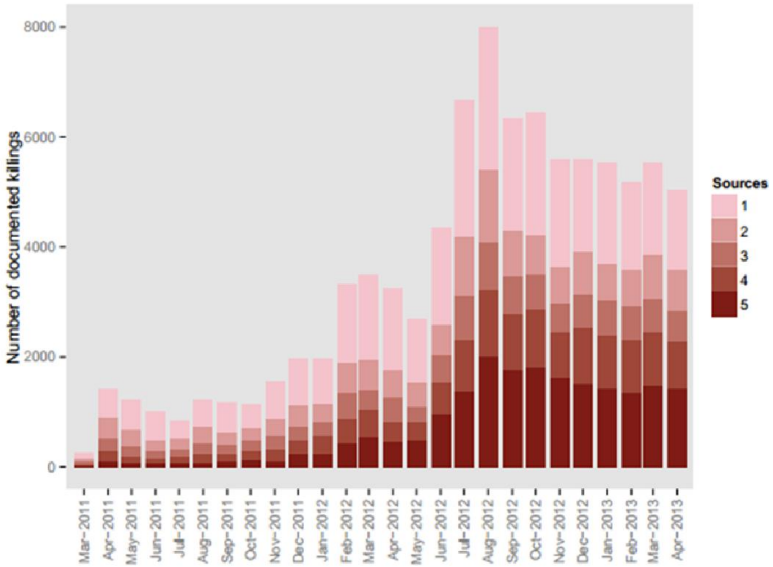
- Short Term Demand Forecasting
- Crop Forecasting
- Long Term Demand Forecasting
- Maintenance and Reliability
- Production Scheduling
- Data Coordination



REVOLUTIO ANALYTI

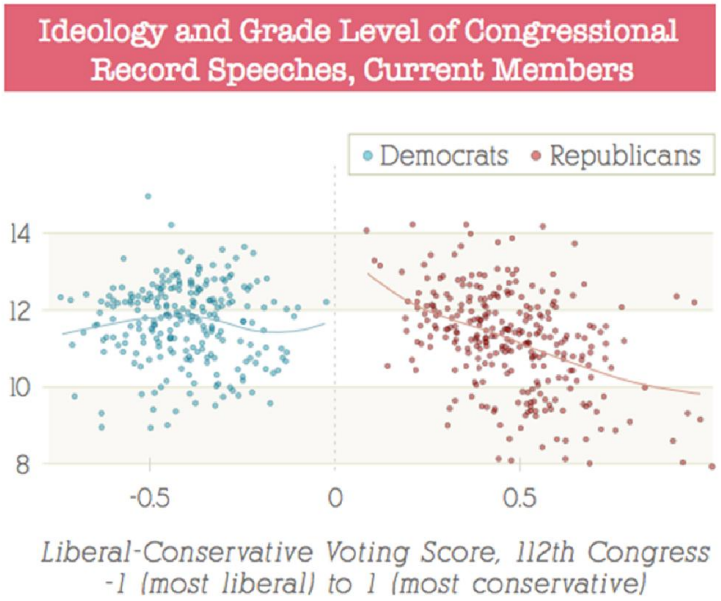
Public Affairs

HRDAG



- Casualty estimation in Warzones

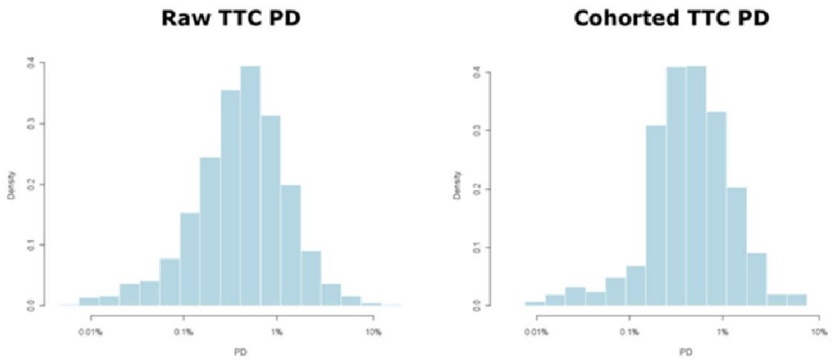
SUNLIGHT FOUNDATION



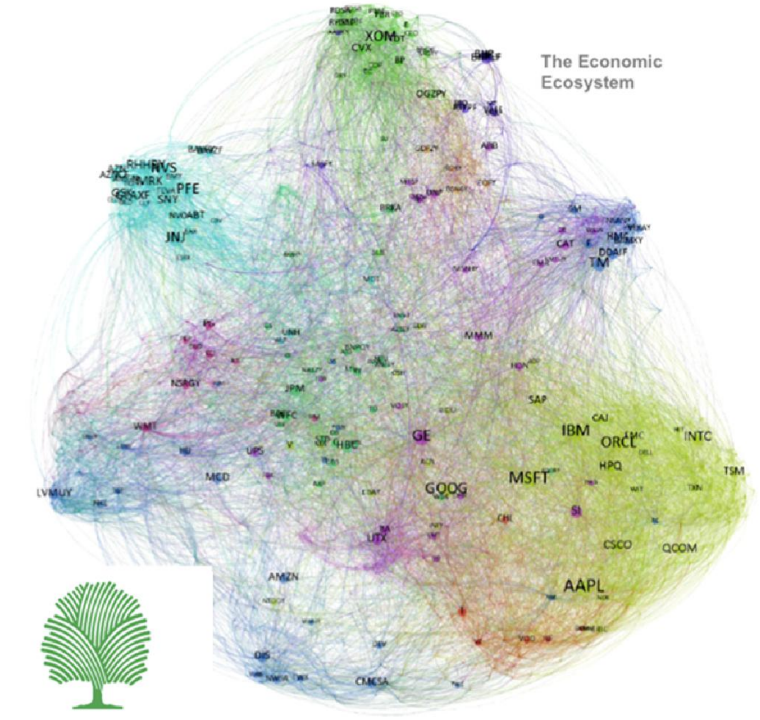
- Political Analysis

REVOLUTION ANALYTICS

Finance and Banking



- Credit Risk Analysis

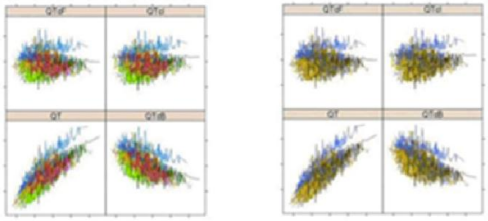


AMERICAN CENTURY

- Financial Networks

REVOLUTION ANALYTICS

Pharmaceuticals



"R use at the FDA is completely acceptable and has not caused any problems." — Dr Jae Brodsky, Office of Biostatistics, Food and Drug Administration

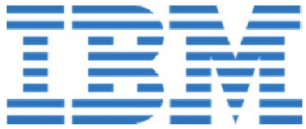
Example from an FDA report. Left: a plot made with the default rainbow color palette. This figure is not 508 compliant because some information is lost to colorblind people. Right: the red-green colorblind simulation. It is not possible to determine which lines correspond to individual subjects when the color palette is reduced.

Regulatory Drug Approvals

- Reproducible research
- Accurate, reliable and consistent statistical analysis
- Internal reporting (Section 508 compliance)

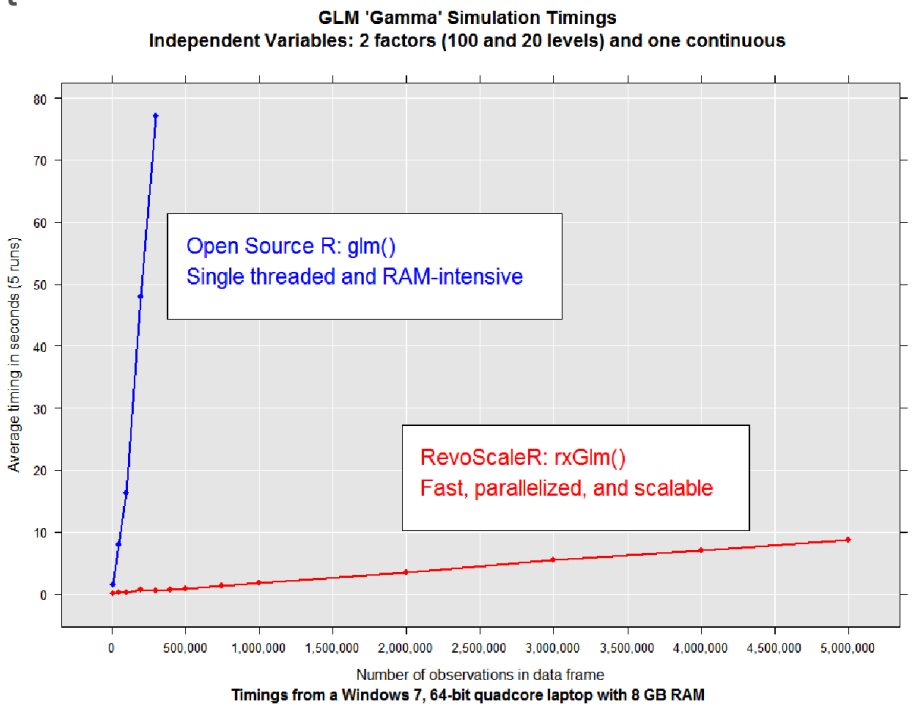
REVOLUTION ANALYTICS

Software Vendors and Service Providers



REVOLUTION ANALYTICS

- Open-Source R Technical Support
- Open Source development
  - RHadoop, ParallelR
- Community Support
  - User Group Sponsorship
  - Meetups, Events
  - Revolutions Blog
- Revolution R Enterprise
  - Big Data (ScaleR package)
  - Integration (Web Services API)
  - Enterprise Platforms
  - Cloud (Amazon AWS)



Companies Using R

Social media

Google
Facebook
Twitter
Foursquare
Kickstarter
eHarmony

Government

FDA
CPFB
City of Chicago
NOAA
NIST

Media

New York Times
Economist
New Scientist
Xbox

Public Affairs

HRDAG
Sunlight Foundation
Benetech
RealClimate

Software Vendors

Revolution Analytics
Rstudio
Zementis
Alteryx
SAP
IBM
SAS
Teradata
TIBCO
Oracle
OneTick
DataCamp

Services

Mango
Accenture
Deloitte
Scientific Revenue
OpenBI
Coursera

Analytics

Zillow
Trulia
DataSong
Exelate
X+1
PredictWise

Finance

American Century
ANZ
Credit Suisse
Nationwide
Lloyds

Manufacturing

Ford
John Deere
Monsanto
SZMF

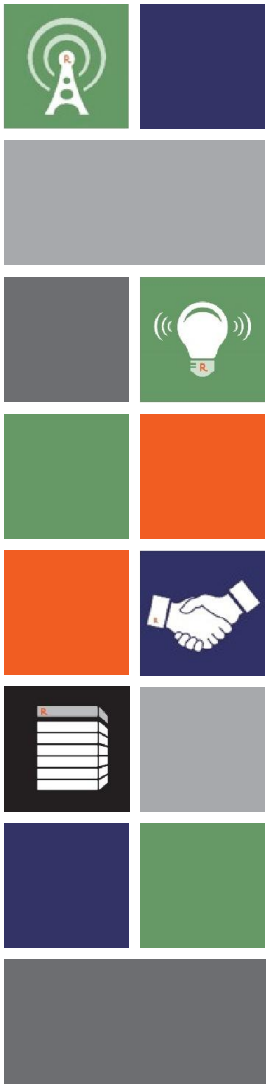
REVOLUTION ANALYTICS

Why are so many companies using R?

- Big Data
- Data Science
- Competition and Innovation
- Open Source
- Ecosystem
- People

REVOLUTION ANALYTICS

REVOLUTION ANALYTICS



Thank You

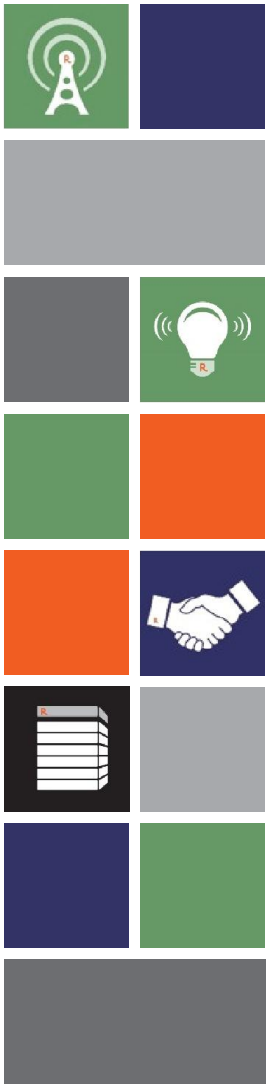
David Smith

david@revolutionanalytics.com, @revodavid

blog.revolutionanalytics.com

REVOLUTION ANALYTICS

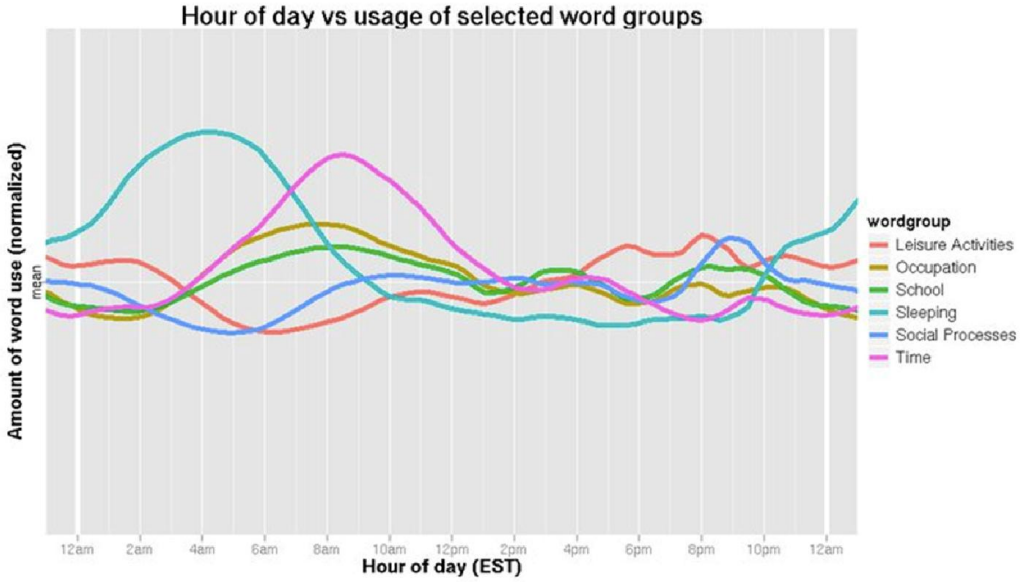
Bonus Applications



facebook

Eric Sun's Predicted Colleague Interactions

		User	Score
1		Alexander Strehl	9.3318008583172
2		Aaron Binbin Liao	8.1162385345038
3		Srinivas Narayanan	3.2938821239025
4		Jack Zhao	3.0129828038274
5		Austin Haugen	2.4928669082428



- Human Resources

- Status Update Trends
- User Profile Images

REVOLUTION ANALYTICS

MONSANTO

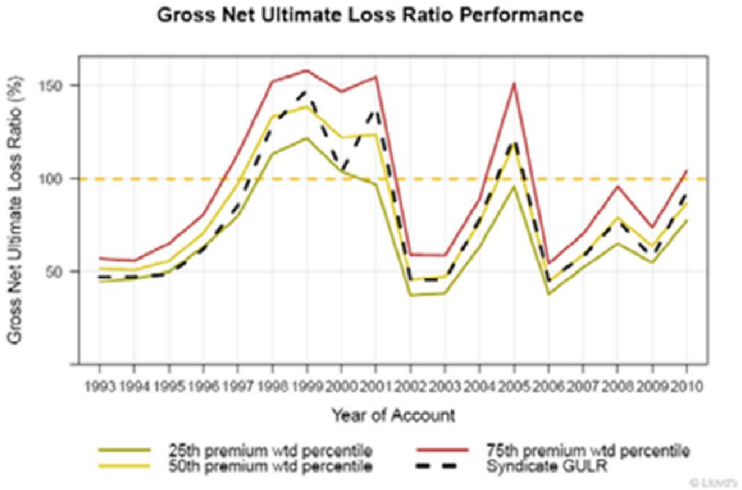
Statistical Analysis:

- Plant Breeding
- Fertility mapping
- Precision Seeding
- Disease Management
- Yield forecasting

REVOLUTION ANALYTICS

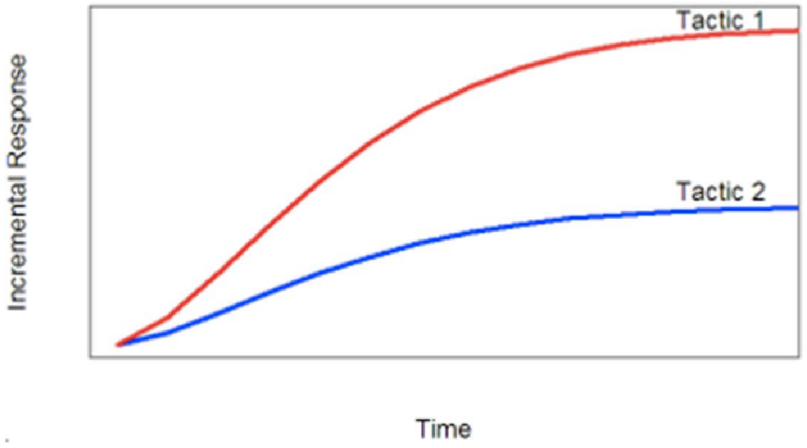
Insurance

LLOYD'S
LLOYD'S OF LONDON



- Risk Analysis
- Catastrophe Modeling

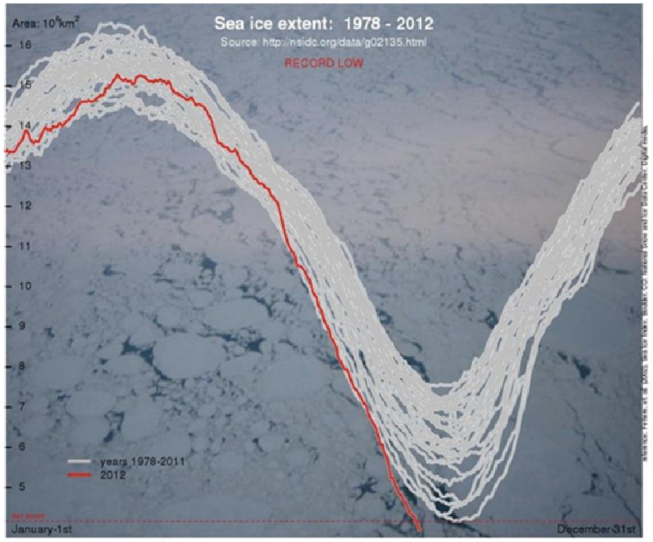
Nationwide®



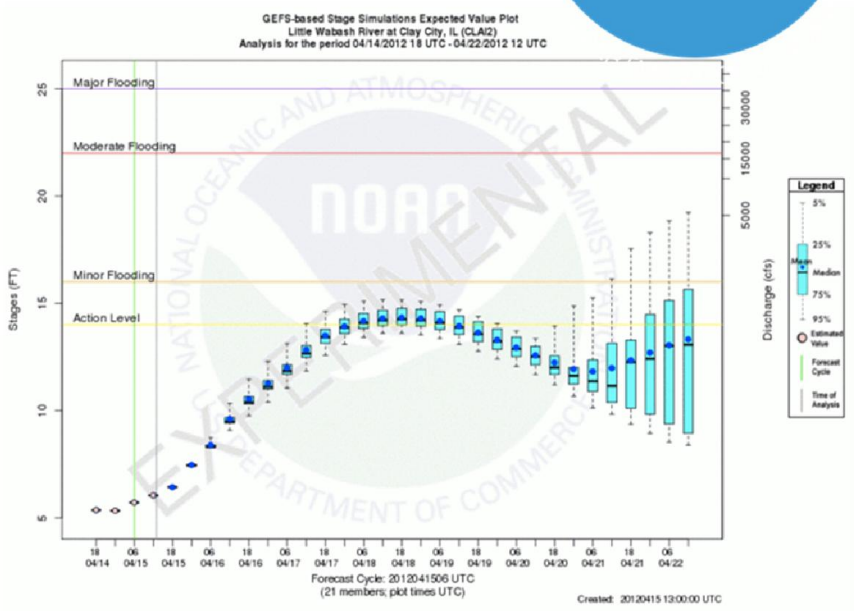
- Marketing Analytics

REVOLUTION ANALYTICS

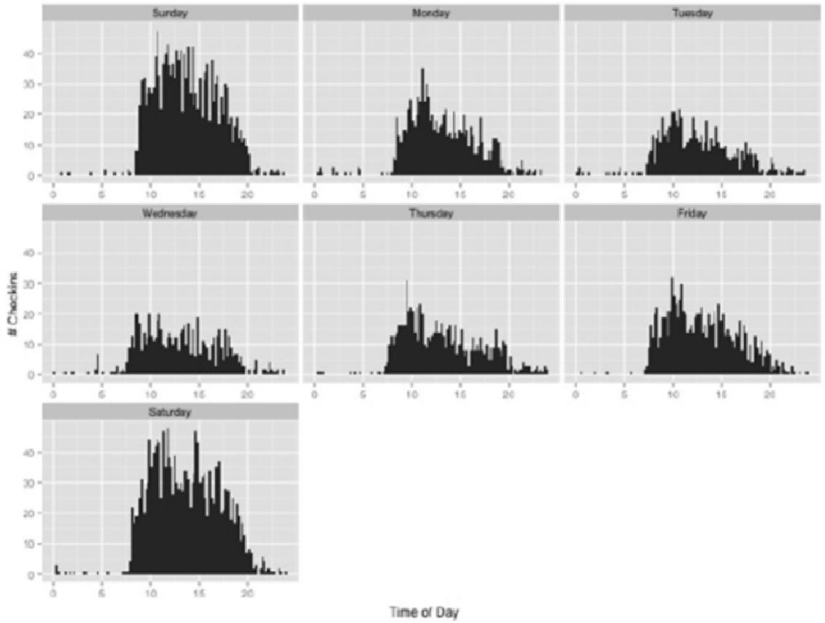
Weather and Climate



- Climate change forecasts



- Flood Warnings



- Recommendation Engine

Marketing Analytics

Power

4X performance
50M records scored daily

"We've combined Revolution R Enterprise and Hadoop to build and deploy customized exploratory data analysis and GAM survival models for our marketing performance management and attribution platform. Given that our data sets are already in the terabytes and are growing rapidly, we depend on Revolution R Enterprise's **scalability and *power*** – we saw about a 4x performance improvement on 50 million records. **It works brilliantly**."
- CEO, John Wallace, DataSong

Scalability

TB's data from 200+ data sources
10's thousands attributes
100's millions of scores daily

"We've been able to **scale** our solution to a problem that's so big that most companies could not address it. If we had to go with a different solution we wouldn't be as efficient as we are now."
- SVP Analytics, Kevin Lyons, eXelate

Performance

2X data
2X attributes
no impact on performance

"We need a ***high-performance*** analytics infrastructure because marketing optimization is a lot like a financial trading. By watching the market constantly for data or market condition updates, **we can now identify opportunities for our clients that would otherwise be lost.**"
- Chief Analytics Officer, Leon Zemel, [x+1]

REVOLUTION ANALYTICS