GENERAL ASSEMBLY

INTRO TO DATA SCIENCE
LECTURE 1: DATA EXPLORATION

Arun Ahuja – aahuja11@gmail.com

Sandip Trivedi - trivedi.sandip@gmail.com

WELCOME!

COURSE MEETING:

T/TH 6:30 - 9:30

GA WEST 21ST STREET

COURSE NOTES:

HTTP://GADATASCIENCE.COM

COURSE MAILS: DAT-NYC-10@GA-GROUPS.COM

WELCOME!

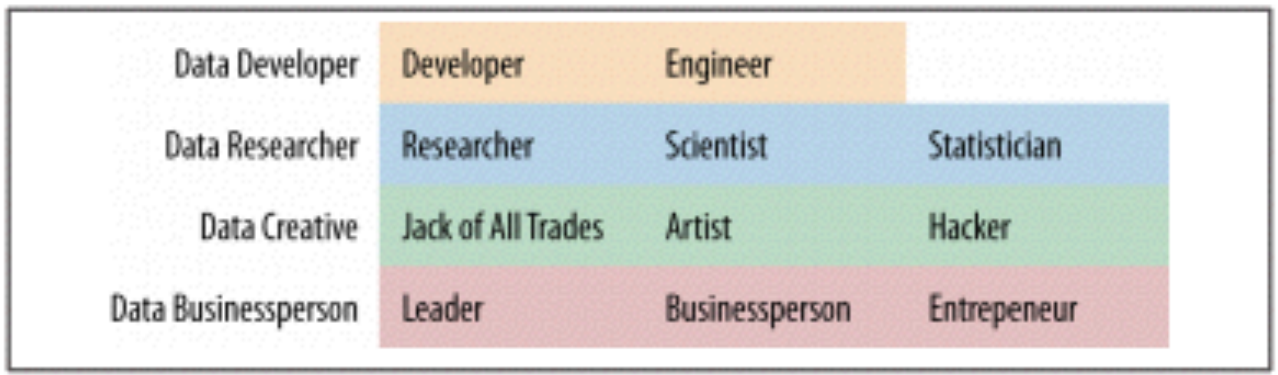
I. WHAT IS DATA SCIENCE?
II. THE DATA MINING WORKFLOW

EXERCISES:
III. WORKING AT THE UNIX COMMAND LINE

I. WHAT IS DATA SCIENCE?

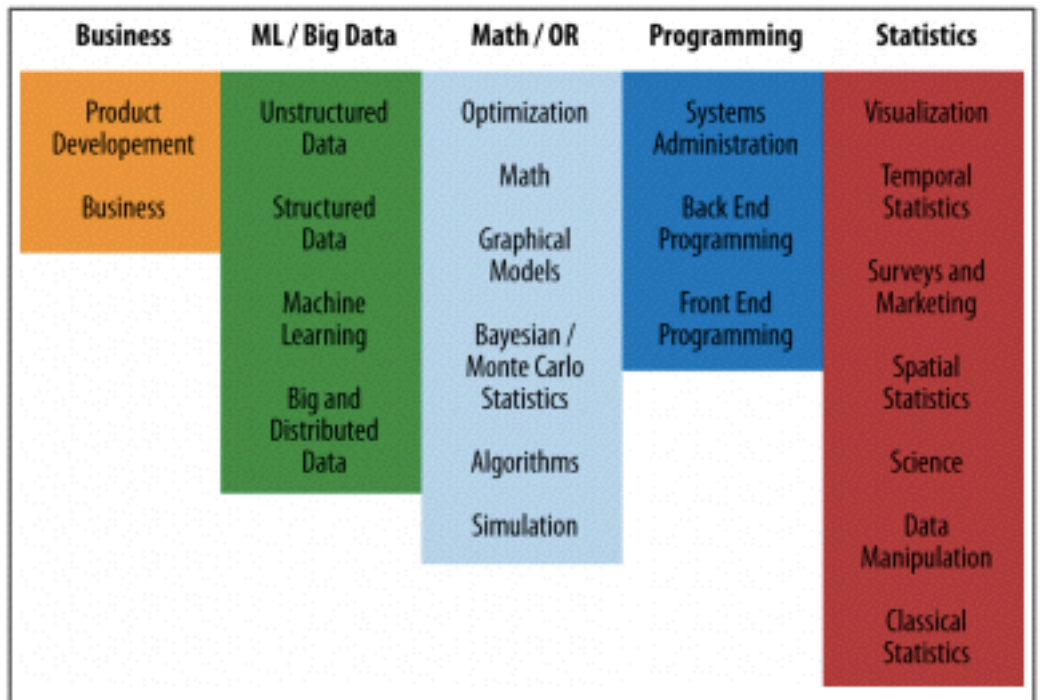
- A set of tools and techniques used to extract useful information from data.

- A set of tools and techniques used to extract useful information from data.
- An interdisciplinary, problem-oriented subject.



source: Analyzing the Analyzers

source: Analyzing the Analyzers



source: Analyzing the Analyzers

Michael E. Driscoll
@medriscoll

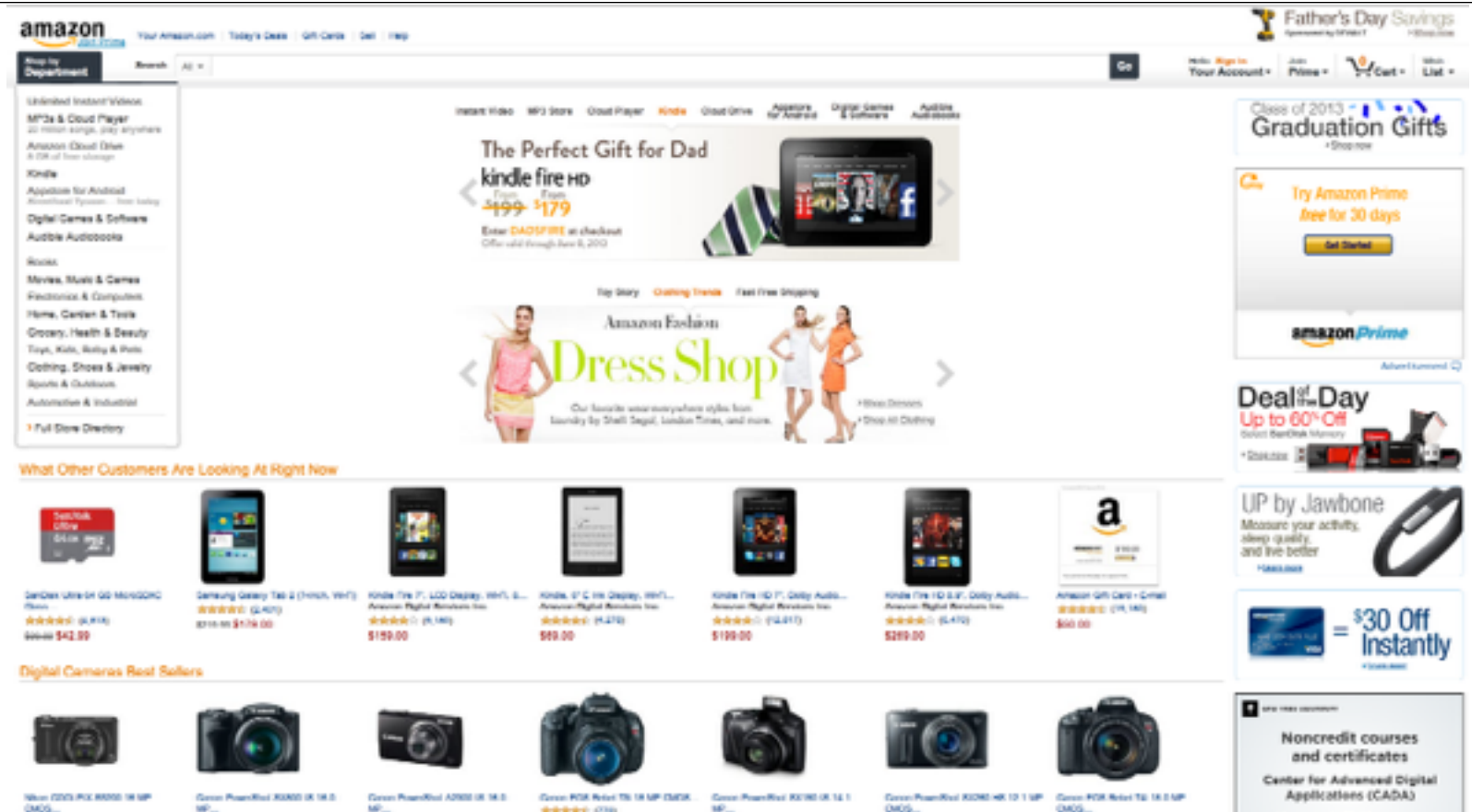
Following

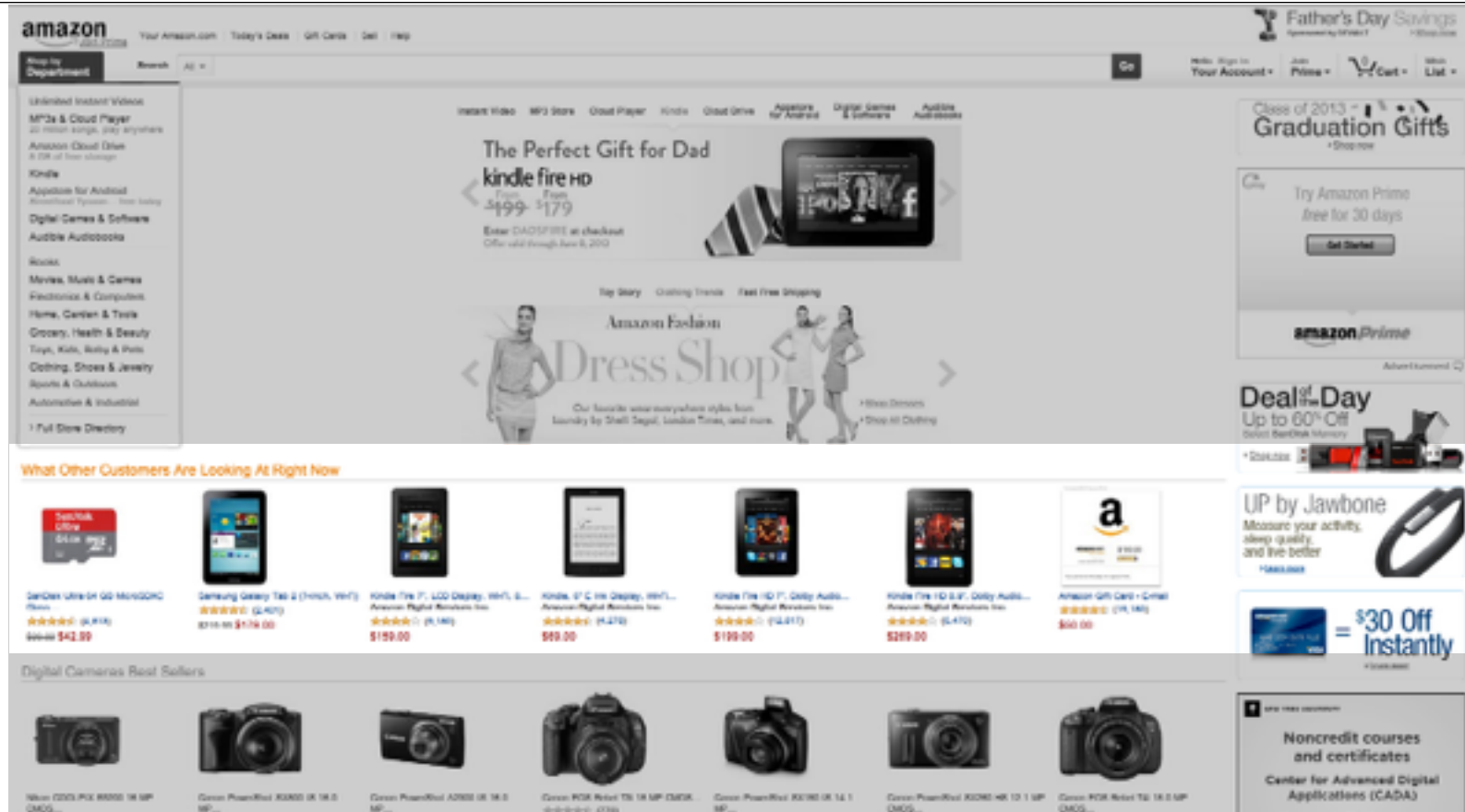
Data scientists: better statisticians than most programmers & better programmers than most statisticians bit.ly/NHmRqu @peteskomoroch

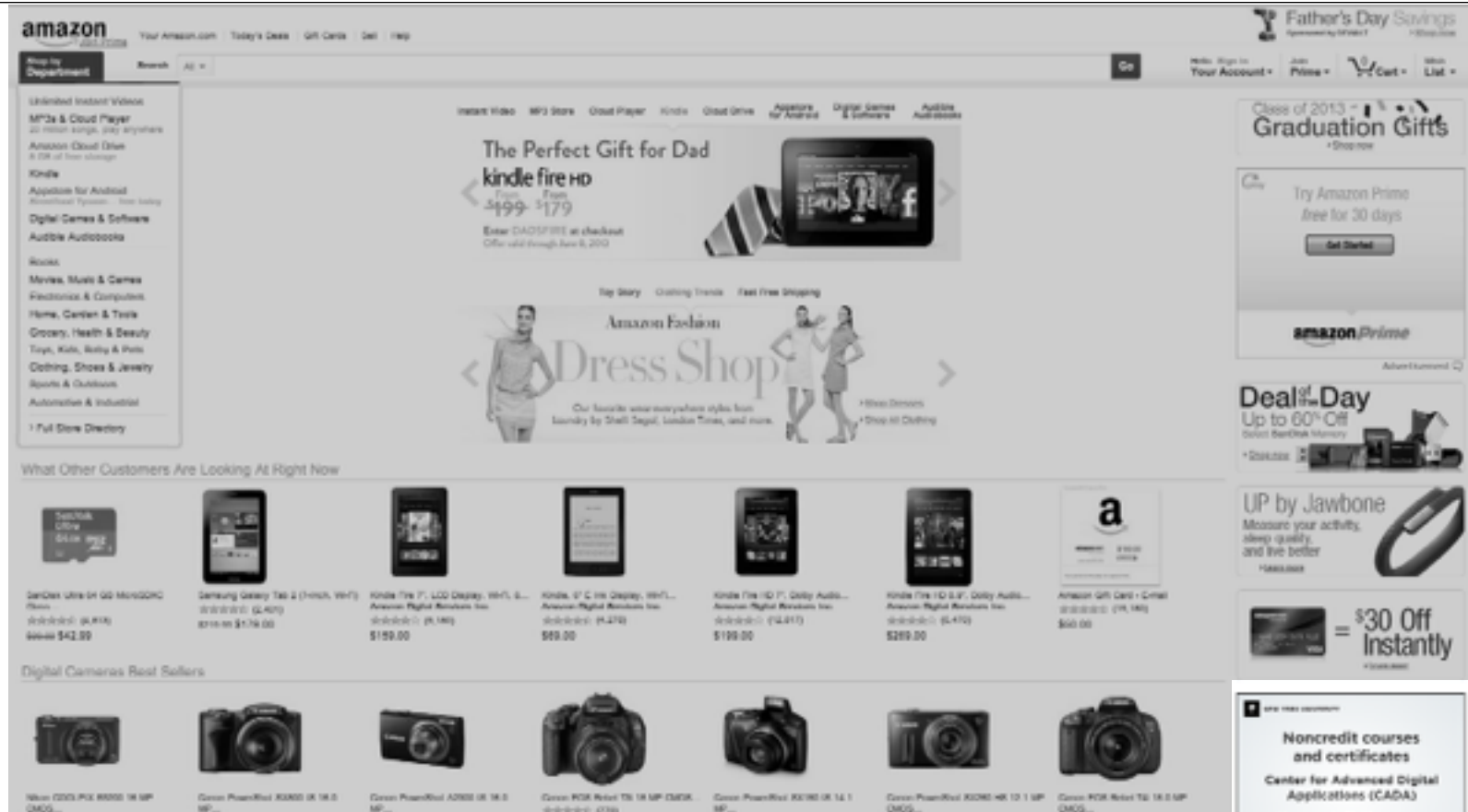
Reply Retweet Favorite More Pocket

- A set of tools and techniques used to extract useful information from data.
- An interdisciplinary, problem-solving oriented subject.
- The application of scientific techniques to practical problems.


NETFLIX
amazon
facebook
Google
LinkedIn
FiveThirtyEight
Nate Silver's Political Calculus
foursquare
Etsy
2012
BARACKOBAMA.COM
Microsoft


amazon
Father's Day Savings
Go
Your Account
Prime
Cart
List
Shop by Department
Search
Unlimited Instant Videos
MP3s & Cloud Player
Amazon Cloud Drive
Kindle
Appstore for Android
Digital Games & Software
Audible Audiobooks
Books
Movies, Music & Games
Electronics & Computers
Home, Garden & Tools
Grocery, Health & Beauty
Toys, Kids, Baby & Pets
Clothing, Shoes & Jewelry
Sports & Outdoors
Automotive & Industrial
Full Store Directory
Instant Video
MP3 Store
Cloud Player
Kindle
Cloud Drive
Appstore for Android
Digital Games & Software
Audible Audiobooks
The Perfect Gift for Dad
kindle fire HD
From $199 From $179
Enter DADSFIRE at checkout
Toy Story
Clothing Trends
Fast Free Shipping
Amazon Fashion
Dress Shop
Shop All Clothing
Class of 2013
Graduation Gifts
Shop now
Try Amazon Prime
free for 30 days
Get Started
amazonPrime
Deal of the Day
Up to 60% Off
UP by Jawbone
Measure your activity, sleep quality, and live better
$30 Off Instantly
What Other Customers Are Looking At Right Now
$42.99
$159.00
$69.00
$199.00
$269.00
Digital Cameras Best Sellers
Noncredit courses and certificates
Center for Advanced Digital Applications (CADA)


amazon
Father's Day Savings
Go
The Perfect Gift for Dad
kindle fire HD
Class of 2013
Graduation Gifts
Try Amazon Prime
free for 30 days
Get Started
amazon Prime
Deal of the Day
Up to 60% Off
Amazon Fashion
Dress Shop
What Other Customers Are Looking At Right Now
UP by Jawbone
Measure your activity, sleep quality, and live better
$30 Off Instantly
Digital Cameras Best Sellers
Noncredit courses and certificates
Center for Advanced Digital Applications (CADA)


amazon
Father's Day Savings
Go
Your Account
Prime
Cart
List
The Perfect Gift for Dad
kindle fire HD
Class of 2013
Graduation Gifts
Try Amazon Prime
free for 30 days
Get Started
amazon Prime
Amazon Fashion
Dress Shop
Deal of the Day
Up to 60% Off
UP by Jawbone
Measure your activity, sleep quality, and live better
$30 Off Instantly
What Other Customers Are Looking At Right Now
Digital Cameras Best Sellers
Noncredit courses and certificates
Center for Advanced Digital Applications (CADA)

Roll over image to zoom in

See 1 customer image

Share your own customer images

Star Trek [Blu-ray] (2009)

Chris Pine (Actor), Zachary Quinto (Actor), J.J. Abrams (Director) | Rated: PG-13 | Format: Blu-ray

(2,040 customer reviews)

List Price: ~~$22.98~~
Price: $9.99 & **FREE Shipping** on orders over $25. Details
You Save: $12.99 (57%)

In Stock.
Ships from and sold by **Amazon.com**. Gift-wrap available.

Want it Wednesday, June 5? Order within 20 hrs 11 mins and choose **One-Day Shipping** at checkout. Details

23 new from $9.98 19 used from $9.78 1 collectible from $49.99

Watch instantly with amazon instant video		Rent	Buy
Star Trek (2009)		$2.99	$9.99

Other Formats & Versions		Amazon Price	New from	Used from
Blu-ray	1-Disc Version	$9.99	$9.98	$9.78
DVD	Single-Disc Edition	$8.49	$5.65	$3.29

Quantity: 1

Yes, I want **FREE Two-Day Shipping** with Amazon Prime

Add to Cart

or

Sign in to turn on 1-Click ordering.

Add to Wish List

Sell Us Your Item
For up to a **$2.60** Gift Card
Trade in
Learn more

More Buying Choices
TechShowMe — Add to Cart
$19.99 & FREE Shipping on orders over $25. Details
43 used & new from $9.78
Have one to sell? Sell on Amazon

Share

This week only, save up to 62% on *Farscape: The Complete Series* in our Deal of the Week. Offer ends June 8, 2013. Learn more

Frequently Bought Together

Price for all three: $66.47

Add all three to Cart | Add all three to Wish List

Some of these items ship sooner than the others. Show details

- **This item:** Star Trek [Blu-ray] ~ Chris Pine Blu-ray $9.99
- Star Trek Into Darkness (Blu-ray 3D + Blu-ray + DVD + Digital Copy) ~ Chris Pine Blu-ray $24.99
- Iron Man 3 (Two-Disc Blu-ray / DVD + Digital Copy) ~ Robert Downey Jr. Blu-ray $31.49

What Other Items Do Customers Buy After Viewing This Item?

Star Trek Into Darkness (Blu-ray 3D + Blu-ray + DVD + Digital Copy) ~ Chris Pine Blu-ray
(199)
$24.99

Star Trek: Original Motion Picture Collection (Star Trek I, II, III, IV, V, VI + The Captain's Summit Bonus Disc) [Blu-ray] ~ William Shatner Blu-ray
(571)
$53.56

Sin City (Two-Disc Theatrical & Recut, Extended, and Unrated Versions) [Blu-ray] ~ Jessica Alba Blu-ray
(933)
$4.99

Roll over image to zoom in

See 1 customer image

Share your own customer images

Star Trek [Blu-ray] (2009)

Chris Pine (Actor), Zachary Quinto (Actor), J.J. Abrams (Director) | Rated: PG-13 | Format: Blu-ray

(2,040 customer reviews)

List Price: ~~$22.98~~
Price: $9.99 & **FREE Shipping** on orders over $25. Details
You Save: $12.99 (57%)

In Stock.
Ships from and sold by **Amazon.com**. Gift-wrap available.

Want it Wednesday, June 5? Order within 20 hrs 11 mins and choose **One-Day Shipping** at checkout. Details

23 new from $9.98 19 used from $9.78 1 collectible from $49.99

Watch instantly with amazon instant video		Rent	Buy
Star Trek (2009)		$2.99	$9.99

Other Formats & Versions		Amazon Price	New from	Used from
Blu-ray	1-Disc Version	$9.99	$9.98	$9.78
DVD	Single-Disc Edition	$8.49	$5.65	$3.29

Quantity: 1

Yes, I want **FREE Two-Day Shipping** with Amazon Prime

Add to Cart

or

Sign in to turn on 1-Click ordering.

Add to Wish List

Sell Us Your Item
For up to a **$2.60** Gift Card
Trade in
Learn more

More Buying Choices
TechShowMe — Add to Cart
$19.99 & FREE Shipping on orders over $25. Details
43 used & new from $9.78
Have one to sell? Sell on Amazon

Share

This week only, save up to 62% on *Farscape: The Complete Series* in our Deal of the Week. Offer ends June 8, 2013. Learn more

Frequently Bought Together

Price for all three: $66.47

Add all three to Cart | Add all three to Wish List

Some of these items ship sooner than the others. Show details

- **This item:** Star Trek [Blu-ray] ~ Chris Pine Blu-ray $9.99
- Star Trek Into Darkness (Blu-ray 3D + Blu-ray + DVD + Digital Copy) ~ Chris Pine Blu-ray $24.99
- Iron Man 3 (Two-Disc Blu-ray / DVD + Digital Copy) ~ Robert Downey Jr. Blu-ray $31.49

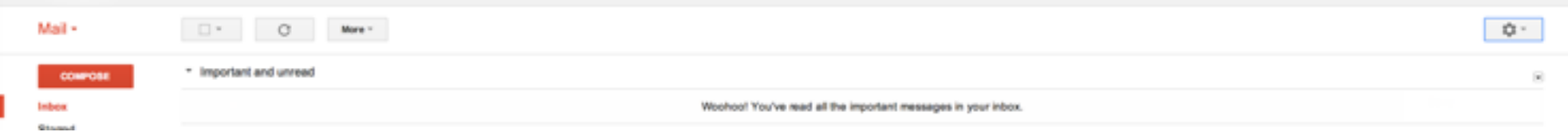
What Other Items Do Customers Buy After Viewing This Item?

Star Trek Into Darkness (Blu-ray 3D + Blu-ray + DVD + Digital Copy) ~ Chris Pine Blu-ray
(199)
$24.99

Star Trek: Original Motion Picture Collection (Star Trek I, II, III, IV, V, VI + The Captain's Summit Bonus Disc) [Blu-ray] ~ William Shatner Blu-ray
(571)
$53.56

Sin City (Two-Disc Theatrical & Recut, Extended, and Unrated Versions) [Blu-ray] ~ Jessica Alba Blu-ray
(933)
$4.99

WHO USES DATA SCIENCE?



II. THE DATA SCIENCE WORKFLOW

acquire parse filter mine represent refine interact

source: http://benfry.com/phd/dissertation-110323c.pdf

source: http://benfry.com/phd/dissertation-110323c.pdf

THE DATA SCIENCE WORKFLOW

source: http://benfry.com/phd/dissertation-110323c.pdf

source: http://benfry.com/phd/dissertation-110323c.pdf

THE DATA SCIENCE WORKFLOW

source: http://benfry.com/phd/dissertation-110323c.pdf

source: http://benfry.com/phd/dissertation-110323c.pdf

source: http://benfry.com/phd/dissertation-110323c.pdf

- From Jeff Hammerbacher:

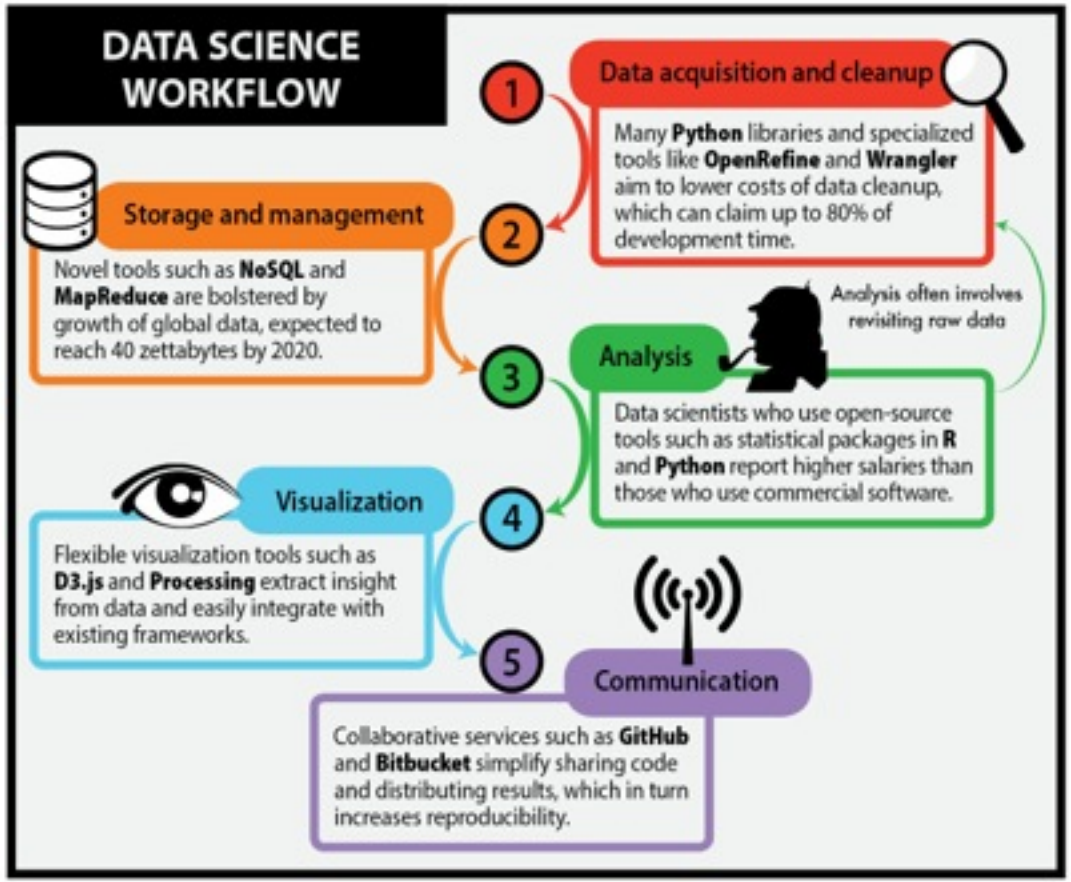
1. Identify problem
2. Instrument data sources
3. Collect data
4. Prepare data (integrate, transform, clean, impute, filter, aggregate)
5. Build model
6. Evaluate model
7. Communicate results

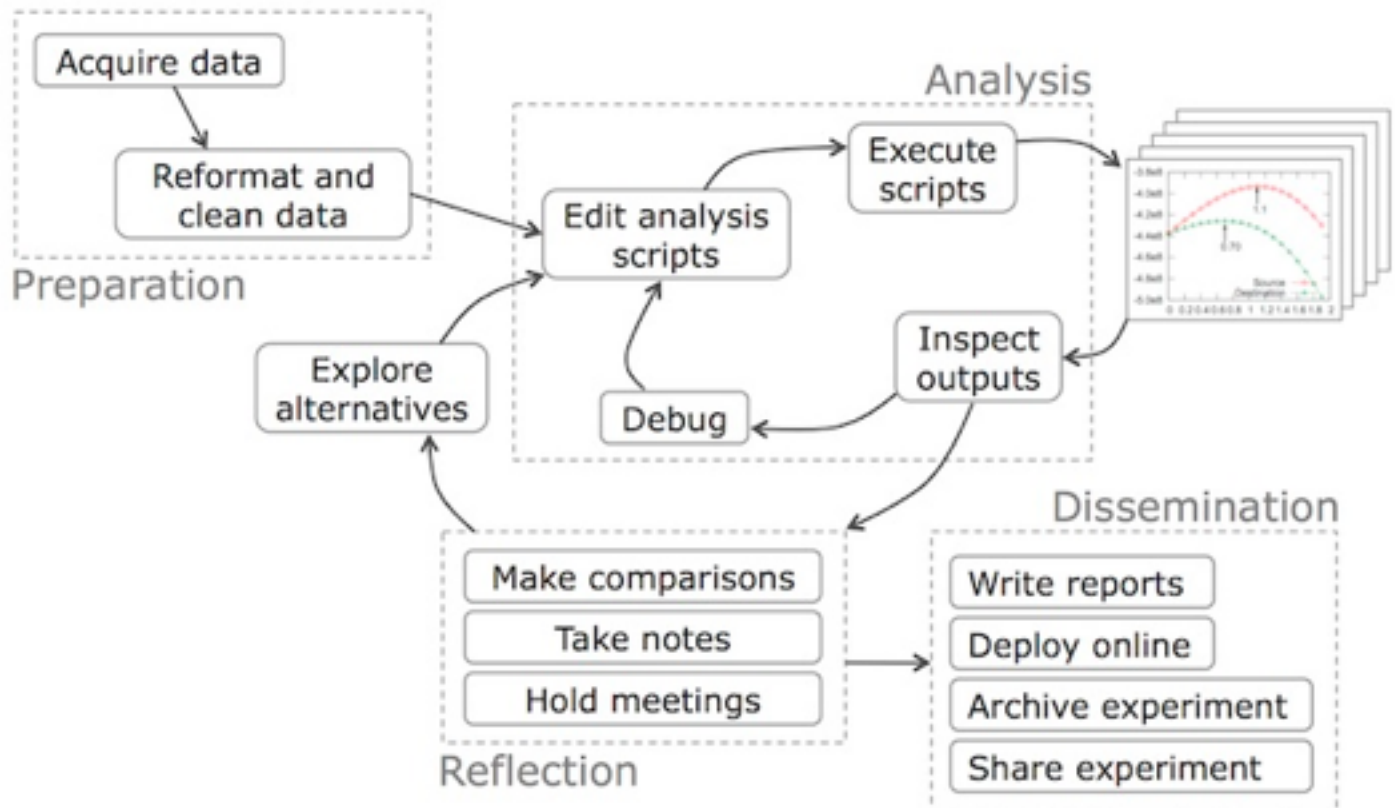
- From Jeff Hammerbacher:

1. Identify problem
2. Instrument data sources
3. Collect data
4. Prepare data (integrate, transform, clean, impute, filter, aggregate)
5. Build model
6. Evaluate model
7. Communicate results

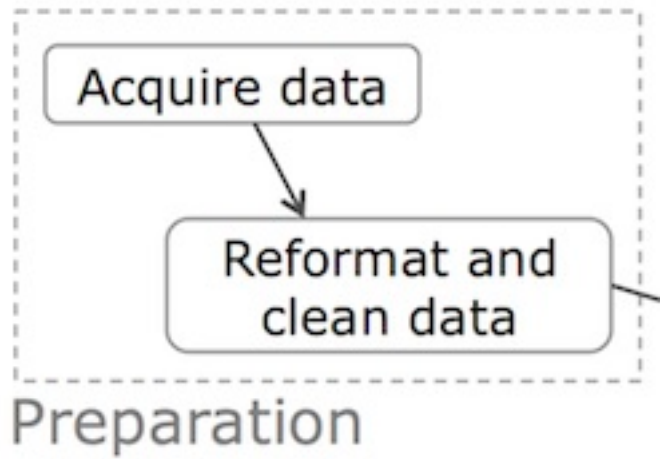
From Dataists Blog

1. Obtain
2. Scrub
3. Explore
4. Model
5. Interpret

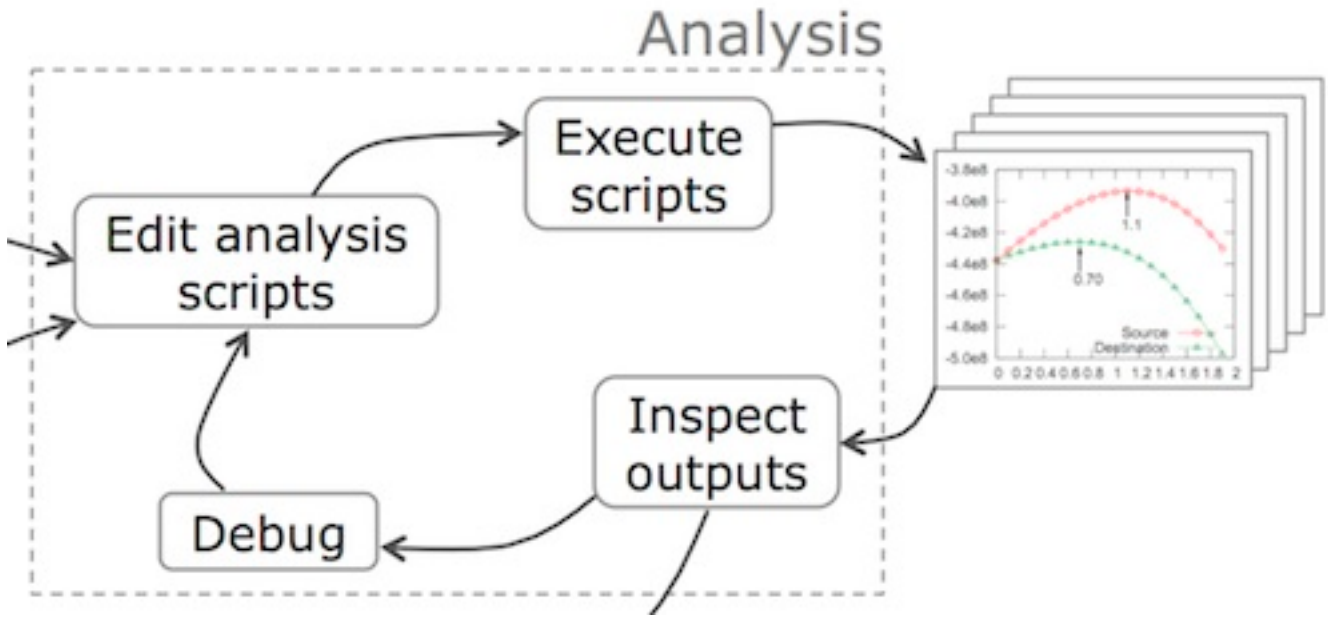

DATA SCIENCE WORKFLOW
1
Data acquisition and cleanup
Many Python libraries and specialized tools like OpenRefine and Wrangler aim to lower costs of data cleanup, which can claim up to 80% of development time.
2
Storage and management
Novel tools such as NoSQL and MapReduce are bolstered by growth of global data, expected to reach 40 zettabytes by 2020.
3
Analysis
Analysis often involves revisiting raw data
Data scientists who use open-source tools such as statistical packages in R and Python report higher salaries than those who use commercial software.
4
Visualization
Flexible visualization tools such as D3.js and Processing extract insight from data and easily integrate with existing frameworks.
5
Communication
Collaborative services such as GitHub and Bitbucket simplify sharing code and distributing results, which in turn increases reproducibility.



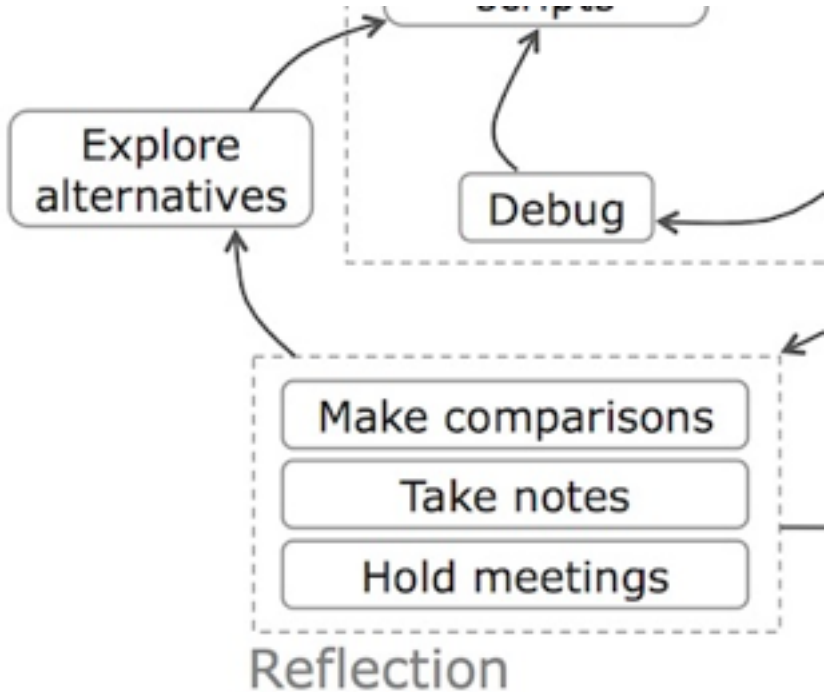
source: Phillip Guo - BLOG@CACM Data Science Workflow: Overview and Challenge



source: Phillip Guo - BLOG@CACM Data Science Workflow: Overview and Challenge



source: Phillip Guo - BLOG@CACM Data Science Workflow: Overview and Challenge



source: Phillip Guo - BLOG@CACM Data Science Workflow: Overview and Challenge

PROBLEM: WHAT ARE THE LEADING INDICATORS THAT A USER WILL MAKE A NEW PURCHASE?

PROBLEM: WHAT ARE THE LEADING INDICATORS THAT A USER WILL MAKE A NEW PURCHASE?

1. Collect data around user retention, user actions within the product, potentially find data outside of company

PROBLEM: WHAT ARE THE LEADING INDICATORS THAT A USER WILL MAKE A NEW PURCHASE?

1. Collect data around user retention, user actions within the product, potentially find data outside of company
2. Extract aggregated values from raw data
    1. How many times did a user share through Facebook within a week? A month?
    2. How often did they open up our emails?

PROBLEM: WHAT ARE THE LEADING INDICATORS THAT A USER WILL MAKE A NEW PURCHASE?

1. Collect data around user retention, user actions within the product, potentially find data outside of company
2. Extract aggregated values from raw data
    1. How many times did a user share through Facebook within a week? A month?
    2. How often did they open up our emails?
3. Examine data to find common distributions and correlations

PROBLEM: WHAT ARE THE LEADING INDICATORS THAT A USER WILL MAKE A NEW PURCHASE?

1. Collect data around user retention, user actions within the product, potentially find data outside of company
2. Extract aggregated values from raw data
    1. How many times did a user share through Facebook within a week? A month?
    2. How often did they open up our emails?
3. Examine data to find common distributions and correlations
4. Extract new meaning to predict if a user would purchase again or not

PROBLEM: WHAT ARE THE LEADING INDICATORS THAT A USER WILL MAKE A NEW PURCHASE?

1. Collect data around user retention, user actions within the product, potentially find data outside of company
2. Extract aggregated values from raw data
    1. How many times did a user share through Facebook within a week? A month?
    2. How often did they open up our emails?
3. Examine data to find common distributions and correlations
4. Extract new meaning to predict if user would purchase again
5. Share results (and probably also go back to the drawing board)

Let's build an analytics team:

You are on of the following organizations:

GA, Pinterest, Tesla, Pixar, Blue Bottle Coffee

Answer the following questions:

1. Define the top priorities of the organization
2. What products could you build or studies could you run?
3. What data would you need to collect?
4. What will your top challenges be?

III. WORKING AT THE UNIX COMMAND LINE

LET'S TAKE A LOOK AT THE 538 DATASET

DISCUSSION