

INTRO To DATA SCIENCE

LECTURE 1: DATA EXPLORATION

Arun Ahuja – aahuja11@gmail.com

Gustavo Sandoval - gussand@gmail.com

INTRO TO DATA SCIENCE

WELCOME!

COURSE MEETING:**T/TH 6:30 - 9:30****TUESDAYS @ GA WEST 21ST STREET****THURSDAYS @ GA ANNEX 17TH STREET****COURSE NOTES:****[HTTP://GITHUB/ARAHUJA/GADS7](http://github.com/ARAHUJA/GADS7)****COURSE MAILS: SCHOOLOGY (SIGN UP!)**

INTRO TO DATA SCIENCE

WELCOME!

I. WHAT IS DATA SCIENCE?

II. THE DATA MINING WORKFLOW

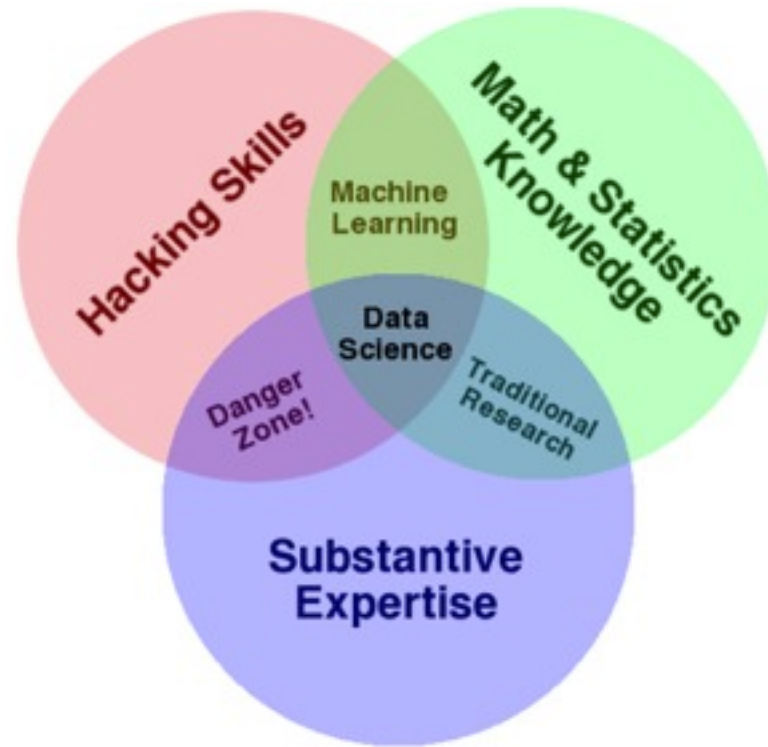
EXERCISES:

III. WORKING AT THE UNIX COMMAND LINE

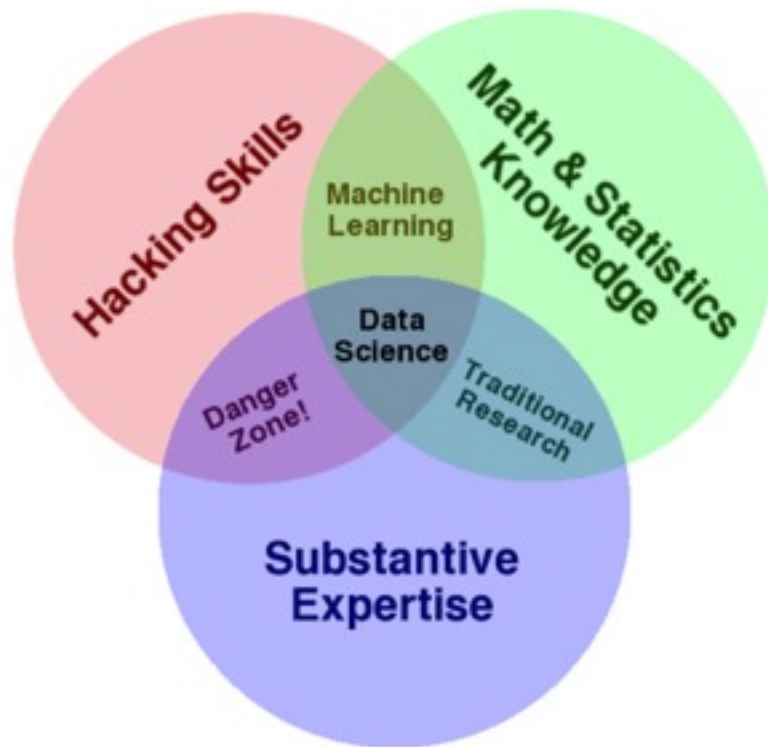
I. WHAT IS DATA SCIENCE?

- 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: <http://www.dataists.com/2010/09/the-data-science-venn-diagram/>



ONE MORE THING!

Communication skills

source: <http://www.dataists.com/2010/09/the-data-science-venn-diagram/>

- 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.

- 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.
- A rapidly growing field.



WHO USES DATA SCIENCE?

14

The screenshot displays the Amazon.com homepage. At the top, the Amazon logo is on the left, and navigation links for 'Your Amazon.com', 'Today's Deals', 'Gift Cards', 'Sell', and 'Help' are on the right. A search bar is centered below the logo. On the far right, there are links for 'Hello, Sign in', 'Your Account', 'Join Prime', 'Cart', and 'Wish List'. Below the search bar, a left sidebar lists categories: 'Shop by Department', 'Unlimited Instant Video', 'MP3s & Cloud Player', 'Amazon Cloud Drive', 'Kindle', 'Apps 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', and 'Full Store Directory'. The main content area features a large banner for 'The Perfect Gift for Dad' showcasing the Kindle Fire HD, with a price reduction from \$199 to \$179. Below this is an 'Amazon Fashion Dress Shop' banner. To the right, there are promotional boxes for 'Father's Day Savings', 'Class of 2013 Graduation Gifts', 'Try Amazon Prime free for 30 days', 'Deal of the Day' (up to 60% off on Sandisk memory), and 'UP by Jawbone'. The bottom section, titled 'What Other Customers Are Looking At Right Now', displays a row of product listings for various Kindle Fire HD models and an Amazon Gift Card. Below this, a section titled 'Digital Cameras Best Sellers' shows a row of camera listings, including Nikon COOLPIX S6200 and Canon PowerShot SX600. On the far right, there are additional promotional boxes for 'Amazon Prime' and 'Noncredit courses and certificates'.

amazon.com

Shop by Department

Unlimited Instant Video
MP3s & Cloud Player
Amazon Cloud Drive
Kindle
Apps 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

The Perfect Gift for Dad
Kindle Fire HD
From \$199 to \$179
Enter **OADSPRE** at checkout
Offer valid through June 8, 2013

Amazon Fashion
Dress Shop
Our favorite wear-everywhere styles from Laundry by Shelli Segal, London Times, and more

What Other Customers Are Looking At Right Now

SanDisk Ultra 64 GB MicroSDHC Class 10
\$49.99 \$42.99

Samsung Galaxy Tab 2 (7-inch, Wi-Fi)
\$249.99 \$179.00

Kindle Fire 7", LCD Display, Wi-Fi, 8-GB
Amazon Digital Services Inc.
\$159.00

Kindle, 8" E-Ink Display, Wi-Fi, 8-GB
Amazon Digital Services Inc.
\$69.00

Kindle Fire HD 7", Dolby Audio...
Amazon Digital Services Inc.
\$199.00

Kindle Fire HD 8.9", Dolby Audio...
Amazon Digital Services Inc.
\$269.00

Amazon Gift Card - E-mail
\$10.00

Digital Cameras Best Sellers

Nikon COOLPIX S6200 16 MP CMOS...
\$149.99

Canon PowerShot SX600 IS 16.0 MP...
\$149.99

Canon PowerShot A2300 IS 16.0 MP...
\$149.99

Canon EOS Rebel T3 18 MP CMOS...
\$799.99

Canon PowerShot SX130 IS 14.1 MP...
\$149.99

Canon PowerShot SX260 HS 12.1 MP CMOS...
\$149.99

Canon EOS Rebel T4i 18.0 MP CMOS...
\$1,199.99

Father's Day Savings
Sponsored by Saks Off 5th

Class of 2013 Graduation Gifts
Sponsored by Amazon.com

Try Amazon Prime free for 30 days
Get Started

Deal of the Day
Up to 60% Off
Select Sandisk Memory
\$10.00

UP by Jawbone
Measure your activity, sleep quality, and live better
\$49.99

Noncredit courses and certificates
Center for Advanced Digital Applications (CADA)
There's Still Time to

The screenshot displays the Amazon.com homepage with the following elements:

- Header:** Amazon logo, navigation links (Your Amazon.com, Today's Deals, Gift Cards, Sell, Help), a search bar, and links for sign-in, Prime, Cart, and Wish List.
- Left Sidebar:** A list of product categories including Unlimited Instant Videos, MP3s & Cloud Player, Amazon Cloud Drive, Kindle, Apps 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, and a Full Store Directory.
- Main Banners:**
 - Top Banner:** "The Perfect Gift for Dad" featuring the Kindle Fire HD, with a price reduction from \$499 to \$179 and a "DADSPRE" code.
 - Second Banner:** "Amazon Fashion Dress Shop" featuring women's clothing.
- Right Sidebar:**
 - Father's Day Savings:** Sponsored by Oldemark.
 - Class of 2013 Graduation Gifts:** Sponsored by Amazon.
 - Try Amazon Prime:** Free for 30 days.
 - Deal of the Day:** Up to 60% Off on Sandisk Memory.
 - UP by Jawbone:** Measure your activity, sleep quality, and live better.
 - \$30 Off Instantly:** Promotion on Visa gift cards.
- Product Recommendations:**
 - What Other Customers Are Looking At Right Now:** A row of products including SanDisk Ultra 64 GB, Samsung Galaxy Tab 2, Kindle Fire 7", Kindle 8" E Ink, Kindle Fire HD 7", Kindle Fire HD 8.9", and an Amazon Gift Card.
 - Digital Cameras Best Sellers:** A row of cameras including Nikon COOLPIX S6200, Canon PowerShot SX300, Canon PowerShot A2000, Canon EOS Rebel T3, Canon PowerShot SX150, Canon PowerShot SX260, and Canon EOS Rebel T4.

The screenshot displays the Amazon.com homepage with the following elements:

- Header:** Amazon logo, navigation links (Your Amazon.com, Today's Deals, Gift Cards, Sell, Help), a search bar, and account links (Hello, Sign in, Your Account, Join Prime, Cart, Wish List).
- Left Sidebar:** A list of product categories including Unlimited Instant Videos, MP3s & Cloud Player, Amazon Cloud Drive, Kindle, Apps 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, and a Full Store Directory.
- Main Banners:**
 - "The Perfect Gift for Dad" featuring the Kindle Fire HD with a price reduction from \$199 to \$179.
 - "Amazon Fashion Dress Shop" with the tagline "Our favorite wear-everywhere styles from Laundry by Shelli Segal, London Times, and more."
- Product Recommendations:** A section titled "What Other Customers Are Looking At Right Now" featuring a row of products including SanDisk Ultra 64 GB MicroSDHC Cards, Samsung Galaxy Tab 2, and various Kindle Fire models.
- Digital Cameras Best Sellers:** A row of digital cameras including Nikon COOLPIX S6200, Canon PowerShot SX300, Canon PowerShot A2000, Canon EOS Rebel T3, Canon PowerShot SX150, Canon PowerShot SX260, and Canon EOS Rebel T4.
- Right Sidebar:**
 - "Class of 2013 Graduation Gifts" with a link to Shop Now.
 - A promotion to "Try Amazon Prime free for 30 days" with a "Get Started" button.
 - "Deal of the Day" for a Sandisk Memory card.
 - "UP by Jawbone" advertisement.
 - "\$30 Off Instantly" promotion.
 - A New York University advertisement for noncredit courses and certificates.



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
[View details](#) [\(2,040 customer reviews\)](#)

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 **18 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			
Blu-ray	1-Disc Version	Amazon Price	New from Used from
DVD	Single-Disc Edition	\$8.49	\$5.65 \$3.29



This week only, save up to 62% on [Friscape: 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
[View details](#) (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
[View details](#) (571)
\$53.56

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

Quantity:

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 [Email](#) [Facebook](#) [Twitter](#) [Pinterest](#)



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
[View details](#) (2,040 customer reviews)

List Price: ~~\$22.66~~

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 27? Order within 20 hrs 11 mins and choose **One-Day Shipping** at checkout. [Details](#)

23 new from \$9.98 **18 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



This week only, save up to 62% on [Fingerboard: 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

Quantity:

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

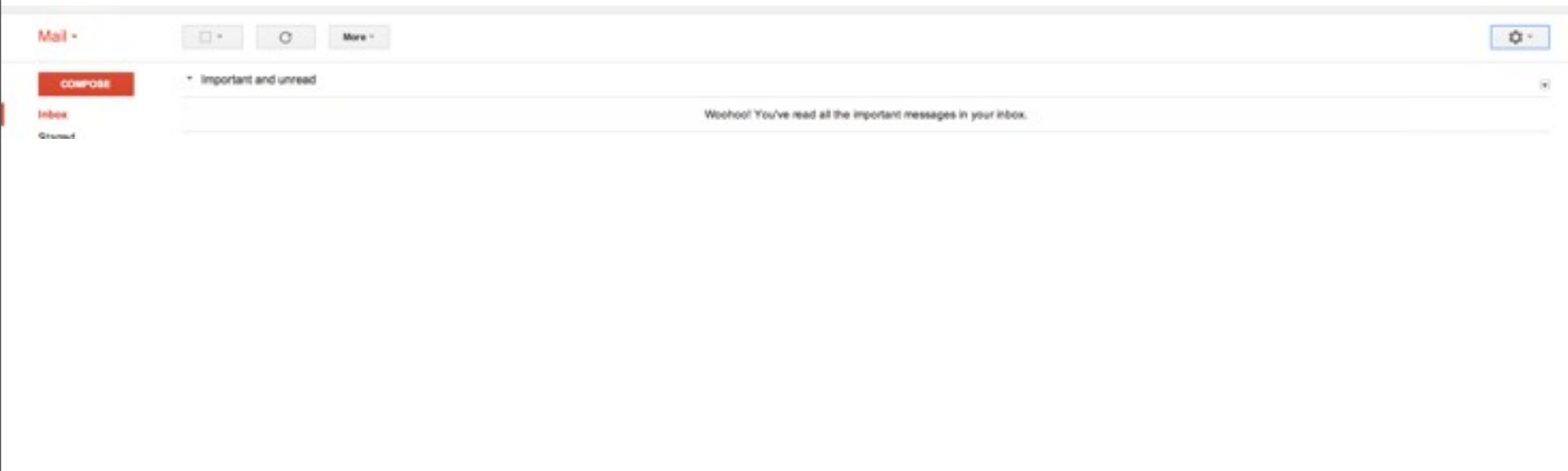
TechShowMe [Add to Cart](#)

\$19.99 & FREE Shipping on orders over \$25. [Details](#)

43 used & new from \$9.78

Have one to sell? [Sell on Amazon](#)

Share [Email](#) [Facebook](#) [Twitter](#) [Pinterest](#)



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>



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



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



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



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



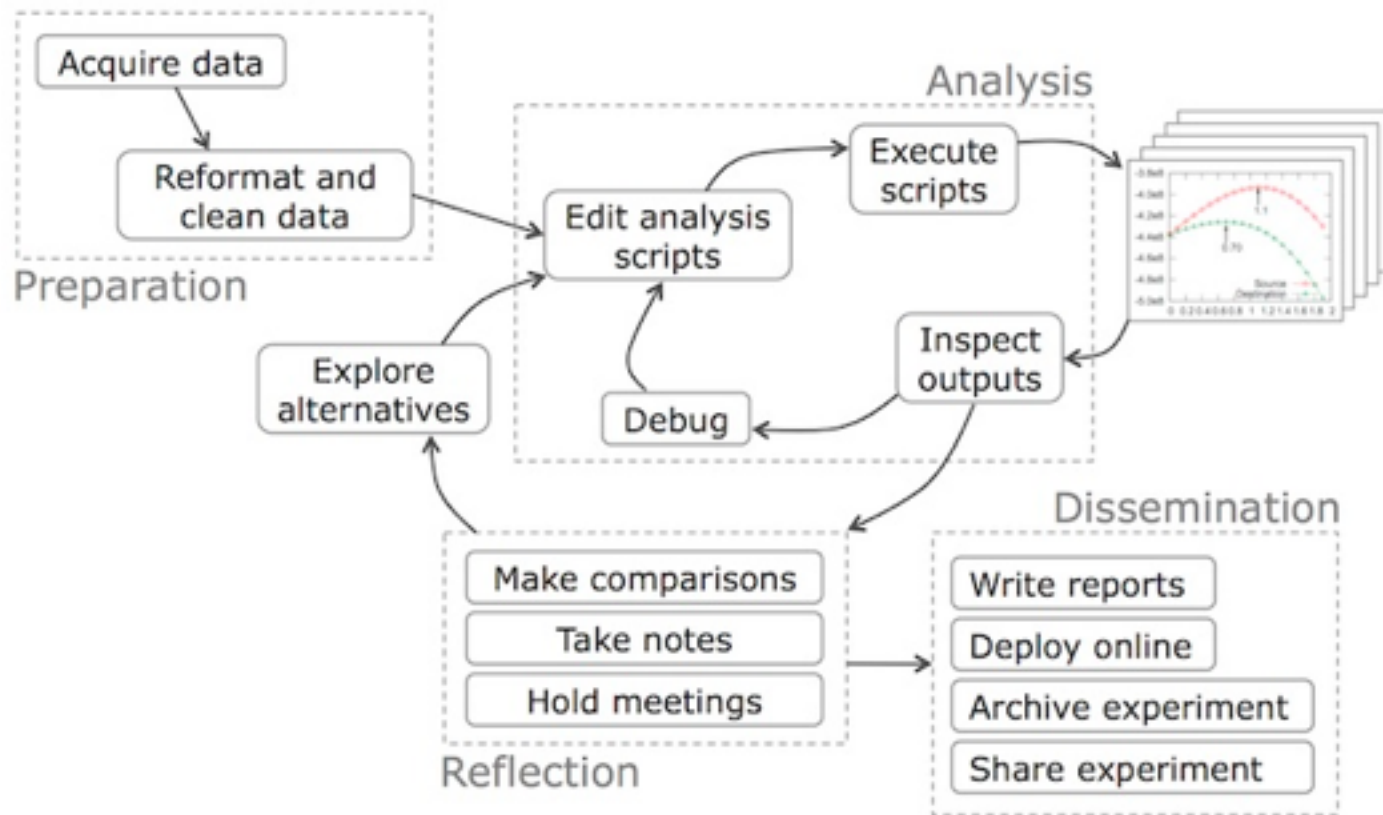
source: <http://benfry.com/phd/dissertation-110323c.pdf>

‣ From Jeff Hammerbacher:

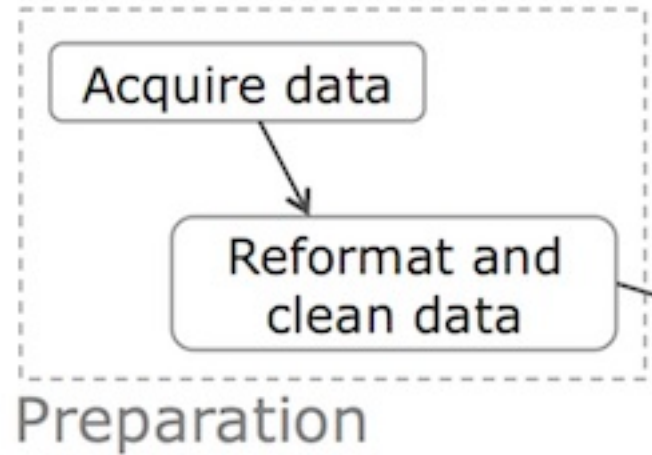
1. Identify problem
2. Instrument data sources
3. Collect data
4. Prepare data (integrate, transform, clean, impute, filter, aggregate)
5. Build model
6. Evaluate model
7. Communicate results

From Dataists Blog

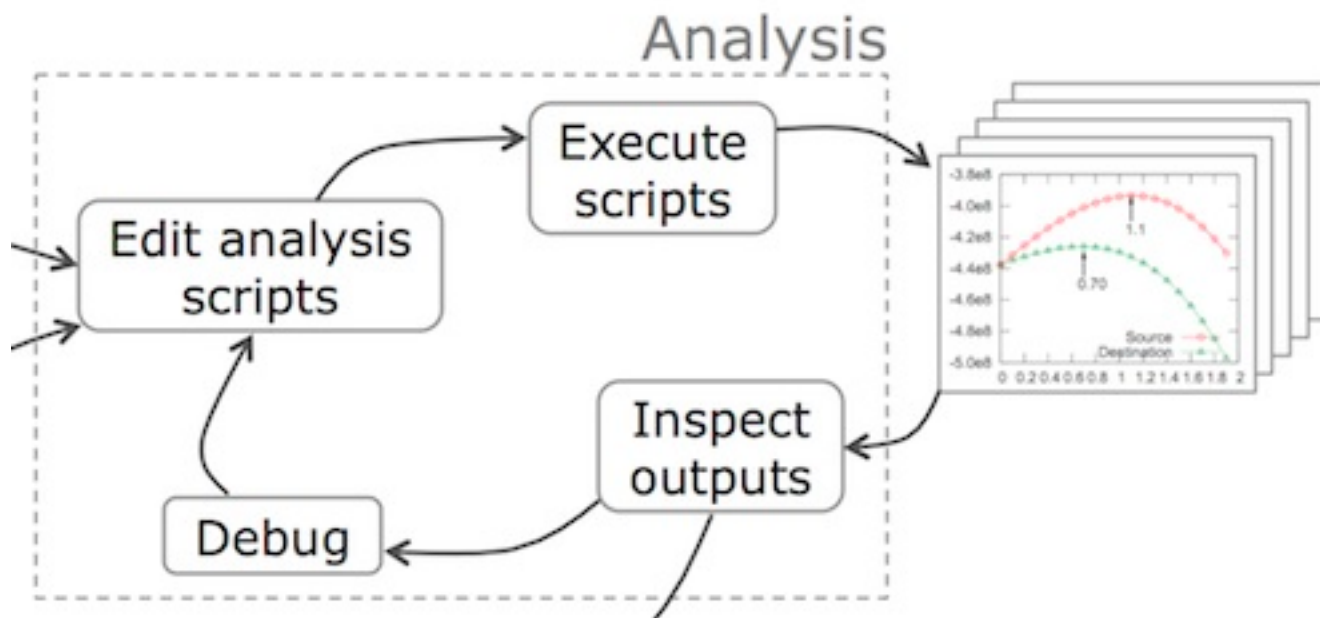
1. Obtain
2. Scrub
3. Explore
4. Model
5. Interpret



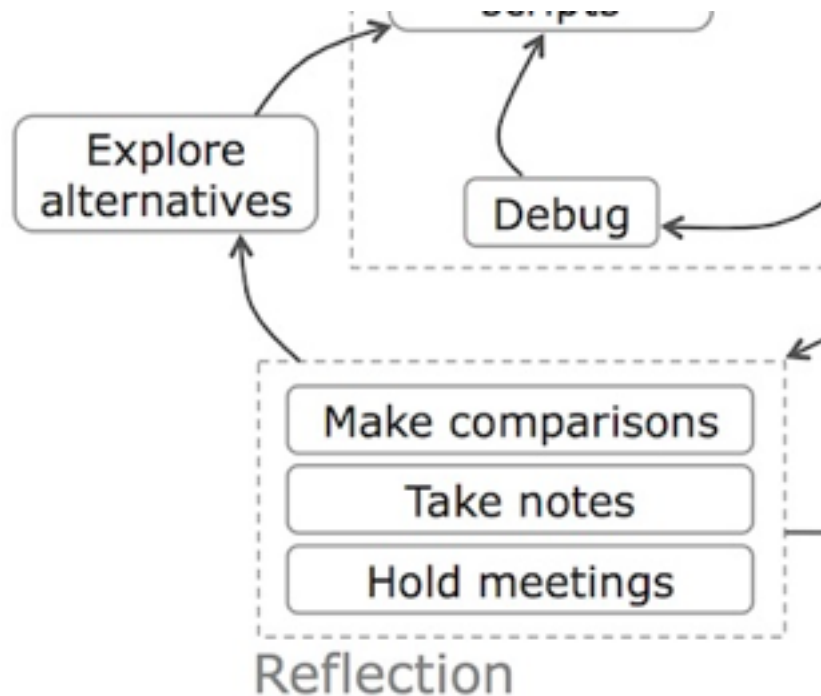
source: Phillip Guo - *BLOG@CACM Data Science Workflow: Overview and Challenge*



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*

- Build an analytics team:
 1. Define the top priorities of the organization
 2. Determine the data you'd like to collect
- What will your greatest challenges be?
- What products could you build?
- What studies could you run? How would these influence the organization?

III. WORKING AT THE UNIX COMMAND LINE

LET'S TAKE A LOOK AT THE 538 DATASET

KEY OBJECTIVES

- NAVIGATE THE FILESYSTEM
- CREATE, MOVE, COPY, AND DELETE FILES & DIRECTORIES
- VIEW & SEARCH FILES
- EDIT & INTERACT WITH FILES
- COMBINE STEPS
- LEARN MORE

TOOLS

- LS, CD
- CAT, TOUCH, MV, CP, MKDIR, RM, RMDIR
- HEAD, TAIL, LESS, CAT, GREP
- VIM, AWK, SED, TR, SORT, UNIQU, WC
- PIPE (|)
- MAN, APROPOS

NOTE

Being comfortable at the command line makes your life much easier!

INTRO TO DATA SCIENCE

DISCUSSION