GENERAL ASSEMBLY

INTRO TO DATA SCIENCE
LECTURE 1: DATA EXPLORATION

Arun Ahuja – aahuja11@gmail.com

Gustavo Sandoval - gussand@gmail.com

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

COURSE MAILS: SCHOOLOGY (SIGN UP!)

WELCOME!

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

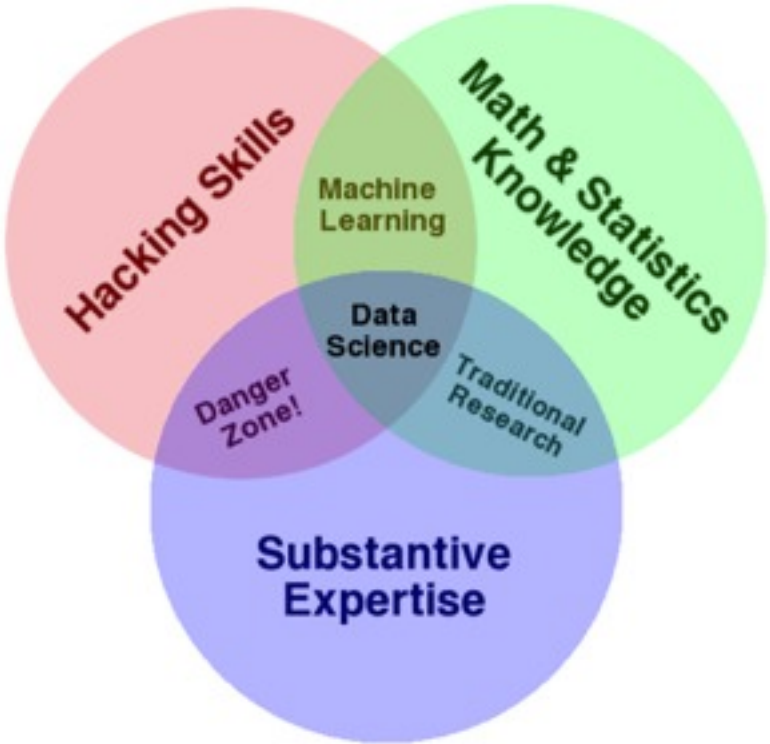
EXERCISES:
III. WORKING AT THE UNIX COMMAND LINE

I. WHAT IS DATA SCIENCE?

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

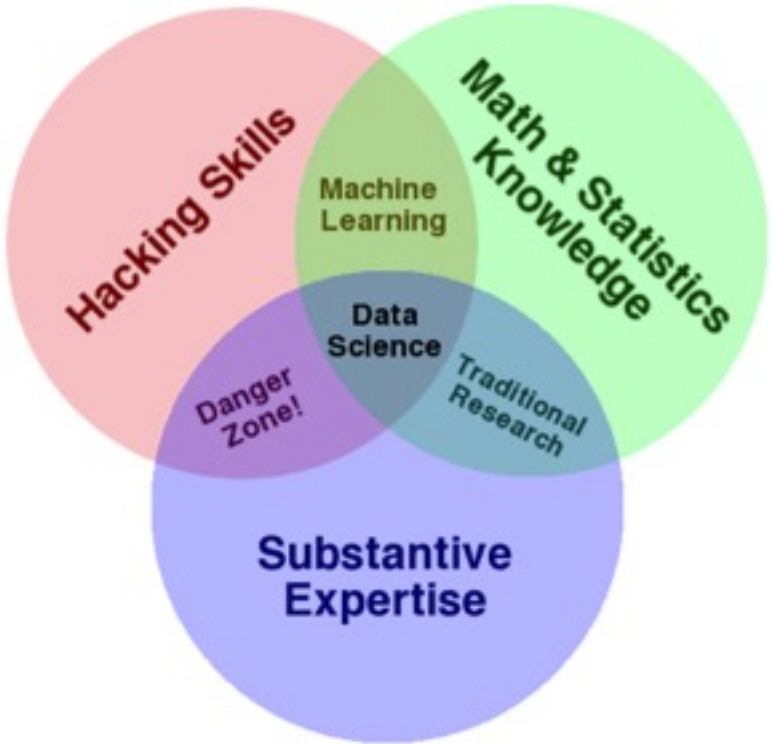
- A set of tools and techniques used to extract useful information from data.
- An interdisciplinary, problem-oriented subject.

THE QUALITIES OF A DATA SCIENTIST



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

THE QUALITIES OF A DATA SCIENTIST



ONE MORE THING!

Communication skills

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

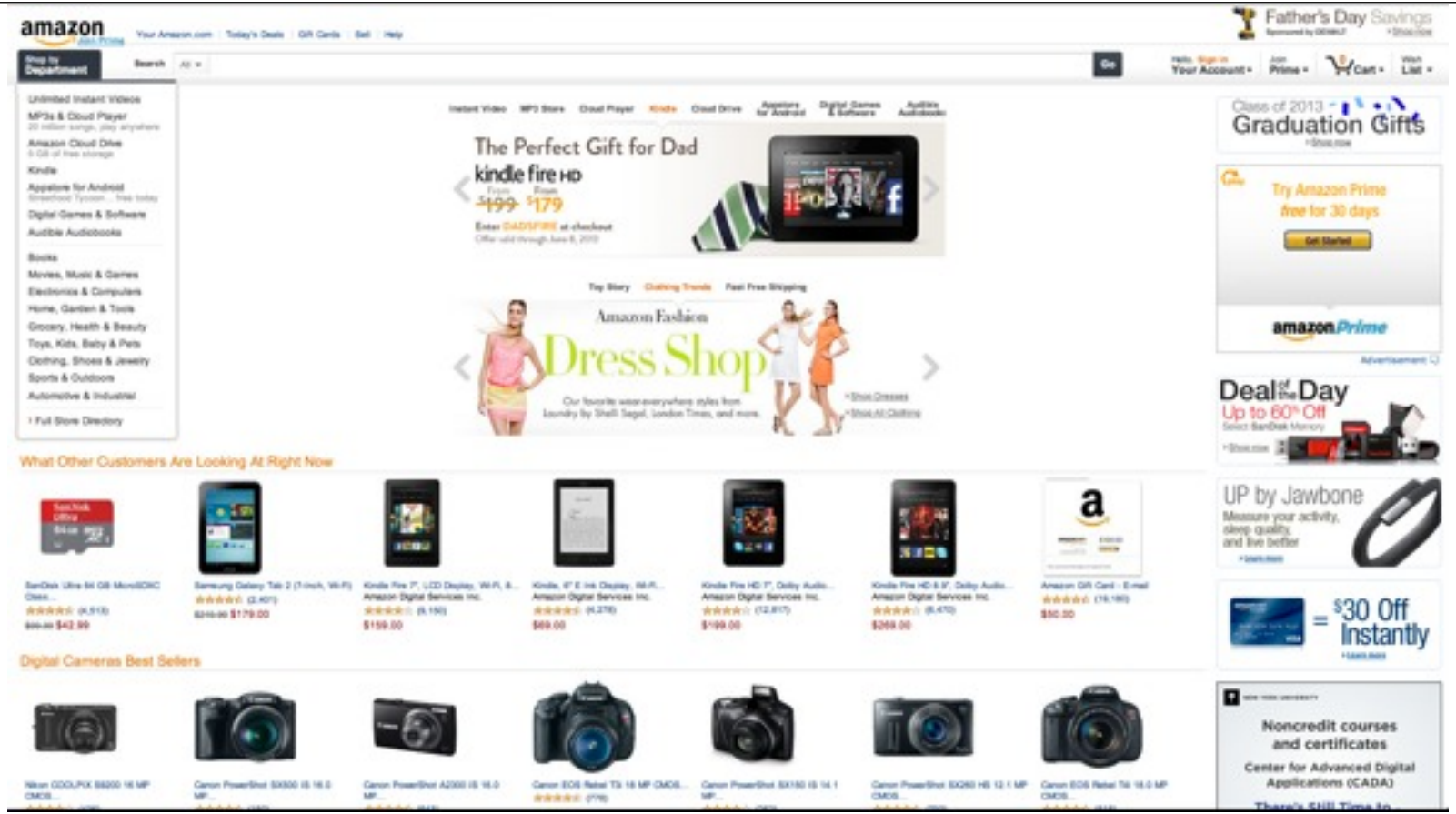
- A set of tools and techniques used to extract useful information from data.
- An interdisciplinary, problem-solving oriented subject.
- The application of scientific techniques to practical problems.

WHAT IS DATA SCIENCE?

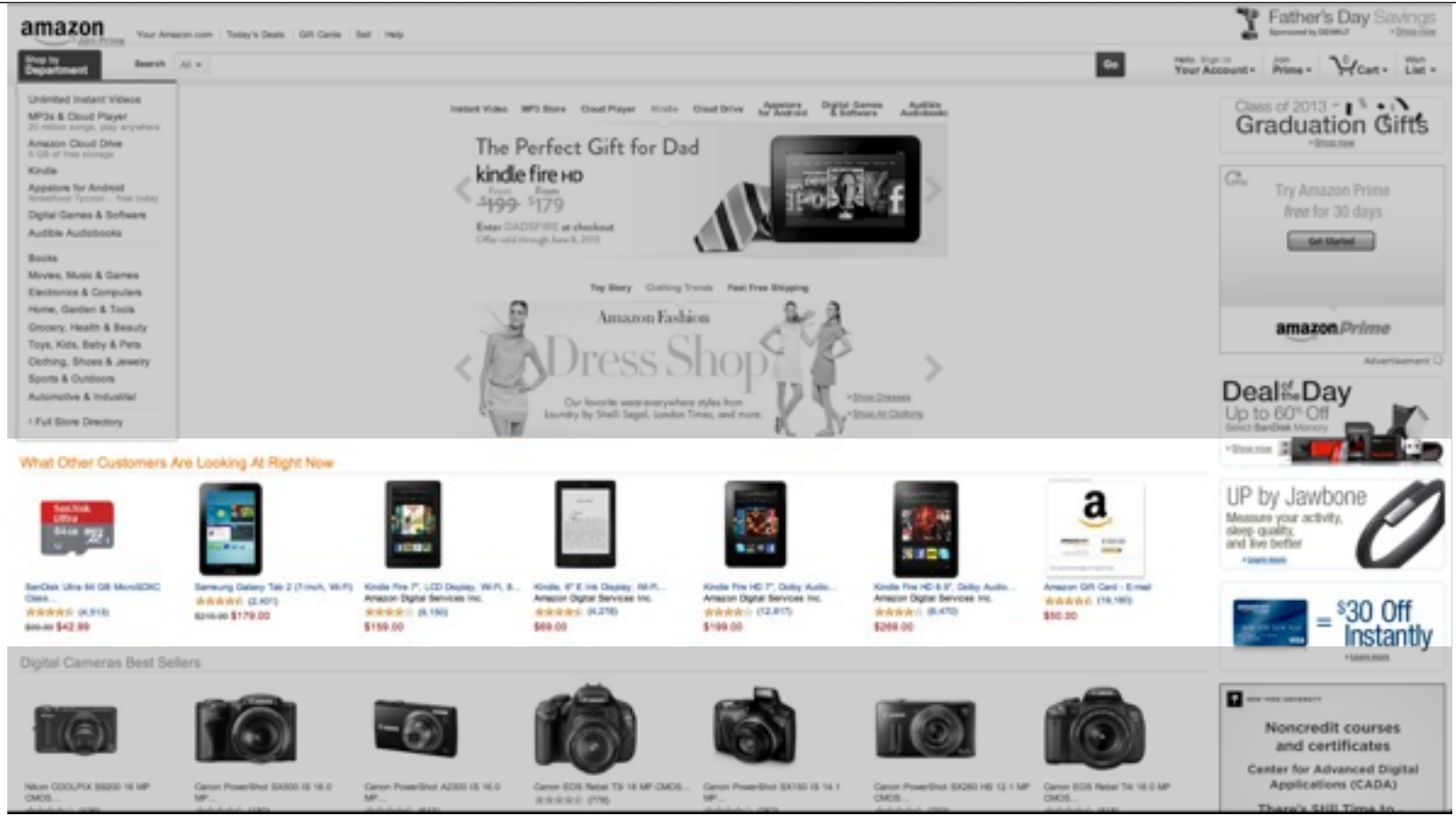
- A set of tools and techniques used to extract useful information from data.
- An interdisciplinary, problem-solving oriented subject.
- The application of scientific techniques to practical problems.
- A rapidly growing field.

WHO USES DATA SCIENCE?

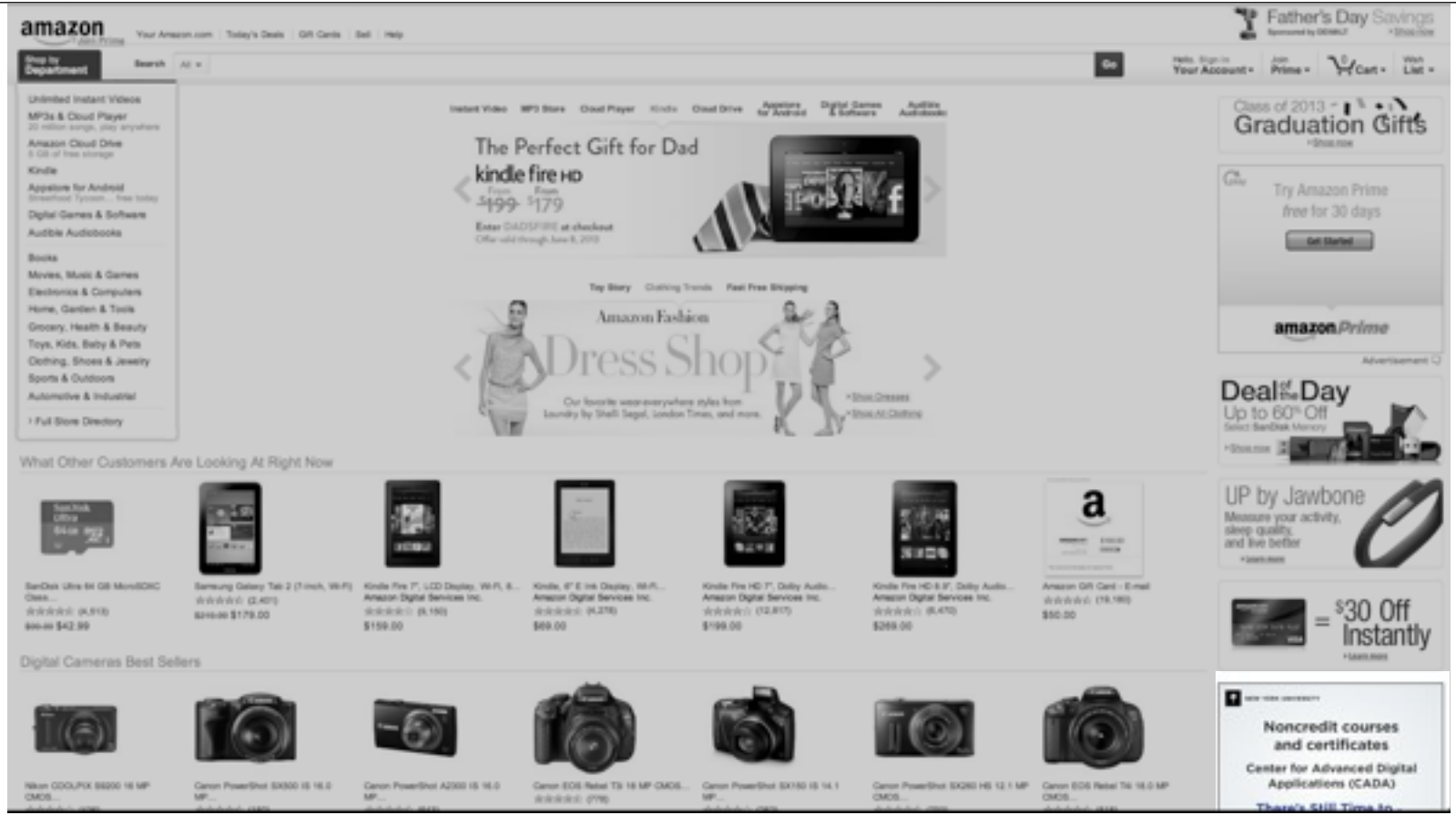
WHO USES DATA SCIENCE?



WHO USES DATA SCIENCE?



WHO USES DATA SCIENCE?



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: ~~$22.99~~

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)

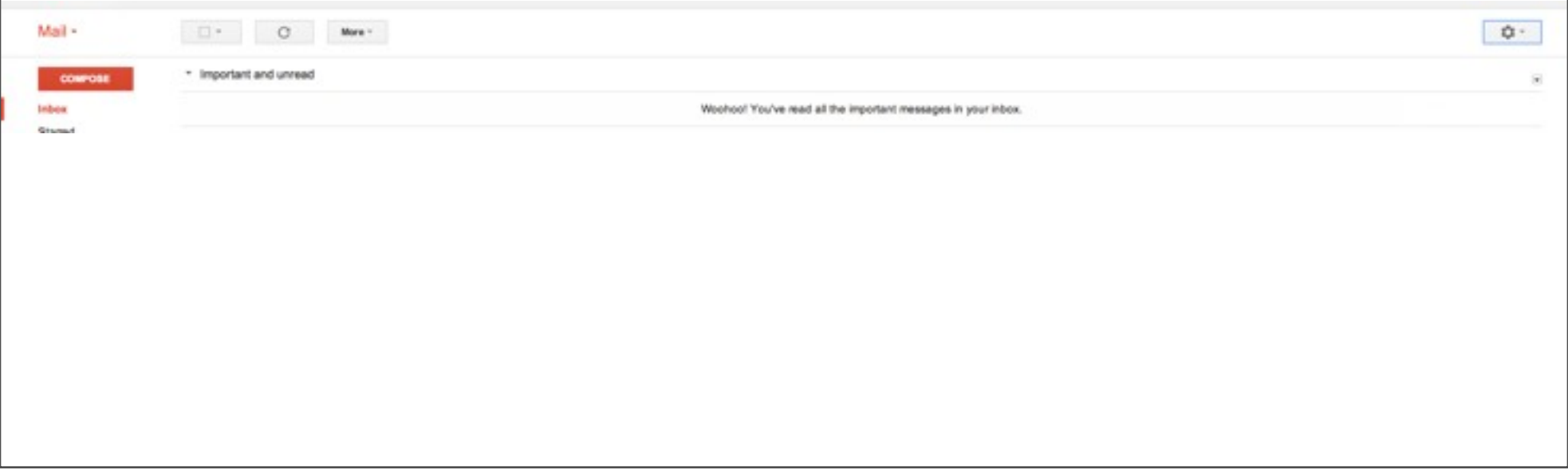
$53.56

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

THE DATA SCIENCE WORKFLOW

acquire parse filter mine represent refine interact

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

THE DATA SCIENCE WORKFLOW

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

THE DATA SCIENCE WORKFLOW

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

THE DATA SCIENCE WORKFLOW

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

THE DATA SCIENCE WORKFLOW

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

THE DATA SCIENCE WORKFLOW

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

THE DATA SCIENCE WORKFLOW

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

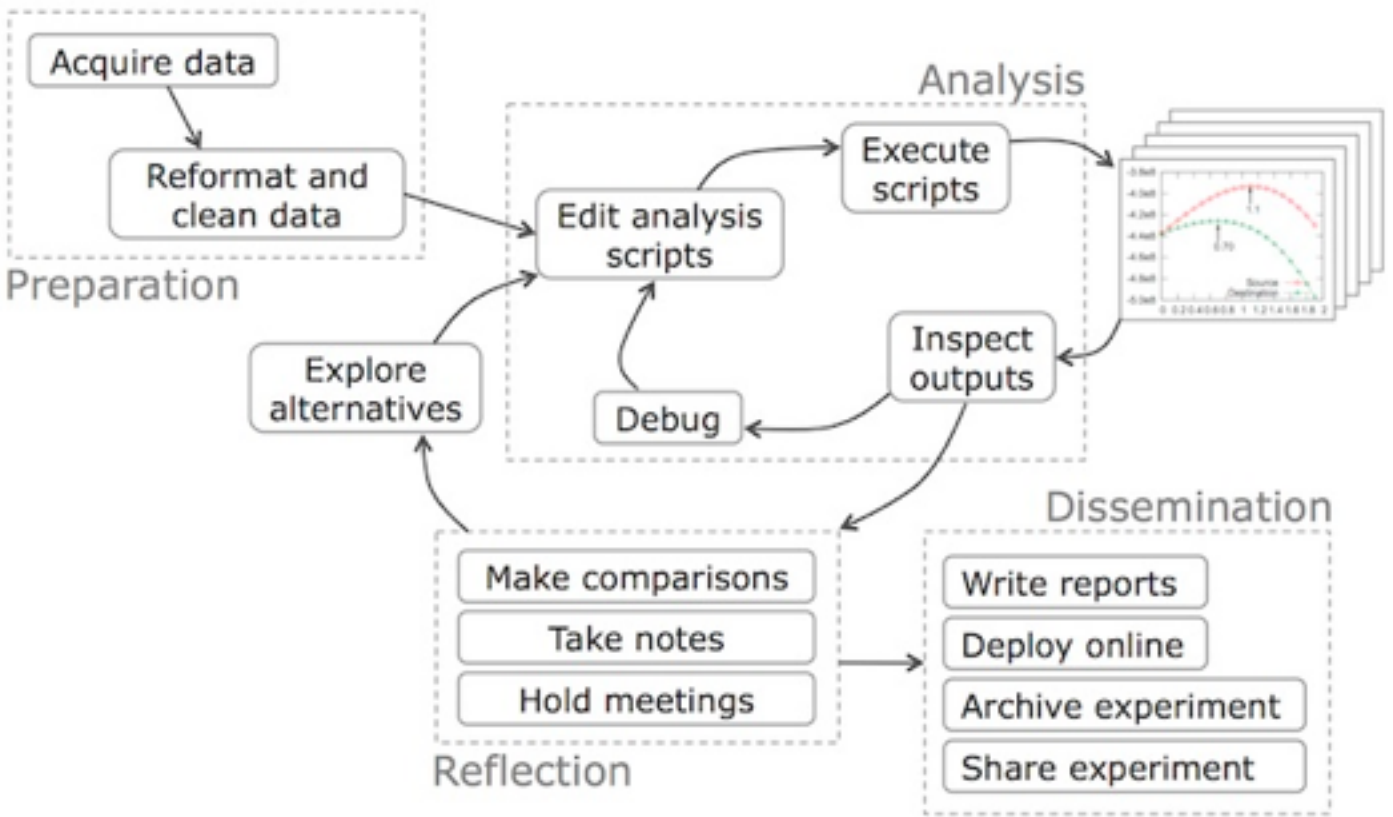
- From Jeff Hammerbacher:

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

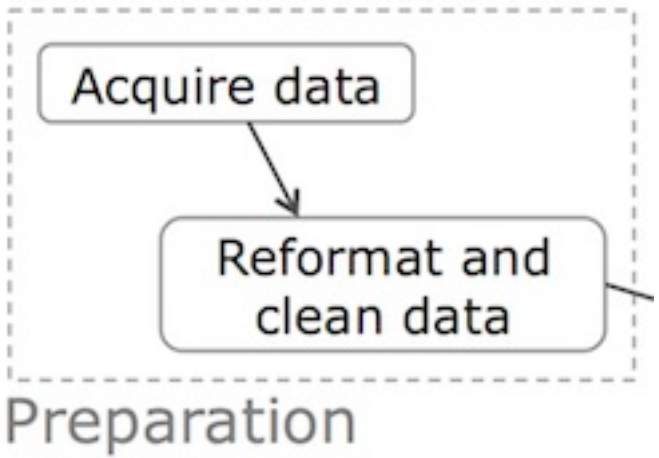
From Dataists Blog

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

THE DATA SCIENCE WORKFLOW

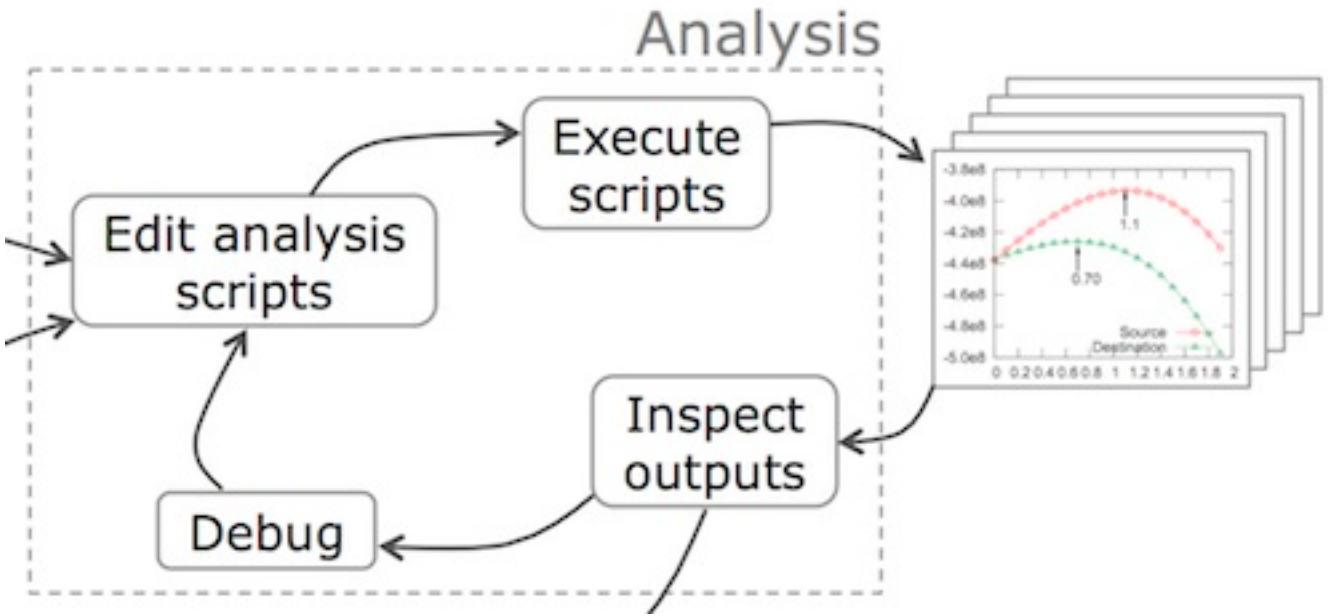


source: Phillip Guo - BLOG@CACM Data Science Workflow: Overview and Challenge



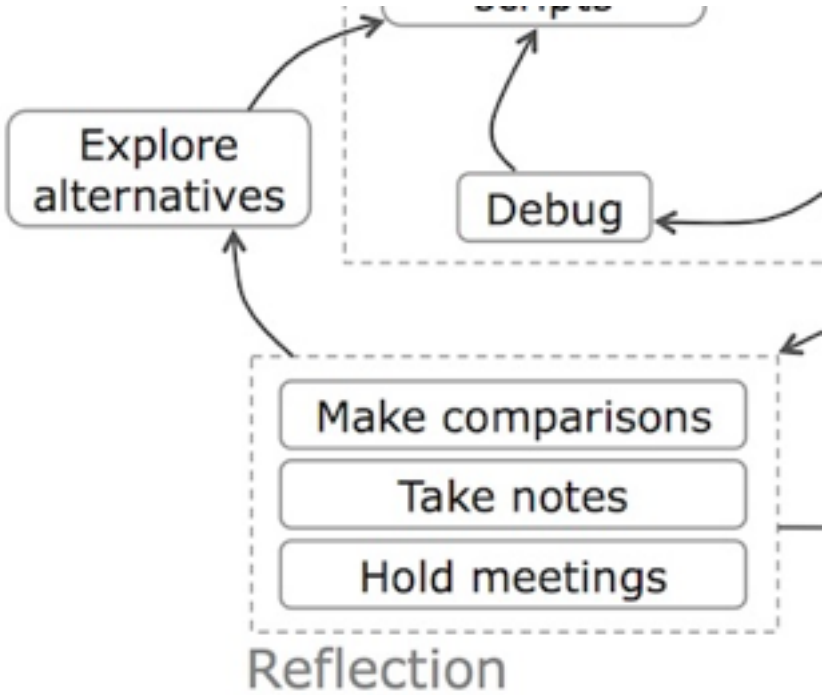
source: Phillip Guo - BLOG@CACM Data Science Workflow: Overview and Challenge

THE DATA SCIENCE WORKFLOW



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

THE DATA SCIENCE WORKFLOW



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

EXERCISE – WORKING AT THE UNIX COMMAND LINE

LET'S TAKE A LOOK AT THE 538 DATASET

EXERCISE – WORKING AT THE UNIX COMMAND LINE

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, UNIQ, WC
- PIPE (|)
- MAN, APROPOS

NOTE

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

DISCUSSION