

Fundamental Methods of Data Science

Class 1

Tomer Libal

Data science

- ▶ Data science is the result of three factors:
 - ▶ The internet
 - ▶ Evolution of data processing methods
 - ▶ Advancements in Computers' processing power and storage space

Data size

How Much Data is Produced Every Day?



2.5 Exabytes are
are produced
every day

Which is equivalent to:

- 🎵 530,000,000 millions songs
- 📱 150,000,000 iPhones
- 💻 5 million laptops
- 📖 250,000 Libraries of Congress
- 📺 90 years of HD Video



<http://www.northeastern.edu/levelblog/2016/05/13/how-much-data-produced-every-day>

Data growth



<https://www.emc.com/leadership/digital-universe/2014iview/executive-summary.htm>

Sources of data

- ▶ Source: <http://expandedramblings.com/index.php/how-many-people-use-social-media/>
 - ▶ Facebook - 1,94 billion monthly active users
 - ▶ Gmail - 1 billion users
 - ▶ Skype - 300 million users
 - ▶ Tweeter - 328 million monthly active users
 - ▶ Amazon - 304 million users
 - ▶ EBay - 167 million active users

Sources of data

- ▶ Source: <http://expandedramblings.com/index.php/how-many-people-use-social-media/>
 - ▶ Facebook - 1,94 billion monthly active users
 - ▶ Gmail - 1 billion users
 - ▶ Skype - 300 million users
 - ▶ Tweeter - 328 million monthly active users
 - ▶ Amazon - 304 million users
 - ▶ EBay - 167 million active users
- ▶ Also,
 - ▶ Online payments, book reviews, etc.
 - ▶ Club membership: supermarket, etc.
 - ▶ Over 3.5 billion searches per day on Google

What can be done with all the Data

Consider an example from a *New York Times* story from 2004:

Hurricane Frances was on its way, barreling across the Caribbean, threatening a direct hit on Florida's Atlantic coast. Residents made for higher ground, but far away, in Bentonville, Ark., executives at Wal-Mart Stores decided that the situation offered a great opportunity for one of their newest data-driven weapons ... predictive technology.

A week ahead of the storm's landfall, Linda M. Dillman, Wal-Mart's chief information officer, pressed her staff to come up with forecasts based on what had happened when Hurricane Charley struck several weeks earlier. Backed by the trillions of bytes' worth of shopper history that is stored in Wal-Mart's data warehouse, she felt that the company could 'start predicting what's going to happen, instead of waiting for it to happen,' as she put it. (Hays, 2004)

Provost/Fawcett - Data Science for Business

What can be done with all the Data

Consider an example from a *New York Times* story from 2004:

Hurricane Frances was on its way, barreling across the Caribbean, threatening a direct hit on Florida's Atlantic coast. Residents made for higher ground, but far away, in Bentonville, Ark., executives at Wal-Mart Stores decided that the situation offered a great opportunity for one of their newest data-driven weapons ... predictive technology.

A week ahead of the storm's landfall, Linda M. Dillman, Wal-Mart's chief information officer, pressed her staff to come up with forecasts based on what had happened when Hurricane Charley struck several weeks earlier. Backed by the trillions of bytes' worth of shopper history that is stored in Wal-Mart's data warehouse, she felt that the company could 'start predicting what's going to happen, instead of waiting for it to happen,' as she put it. (Hays, 2004)

Provost/Fawcett - Data Science for Business

- How can you get a business opportunity from an hurricane and use data from previous events?

What can be done with all the Data

Consider an example from a *New York Times* story from 2004:

Hurricane Frances was on its way, barreling across the Caribbean, threatening a direct hit on Florida's Atlantic coast. Residents made for higher ground, but far away, in Bentonville, Ark., executives at Wal-Mart Stores decided that the situation offered a great opportunity for one of their newest data-driven weapons ... predictive technology.

A week ahead of the storm's landfall, Linda M. Dillman, Wal-Mart's chief information officer, pressed her staff to come up with forecasts based on what had happened when Hurricane Charley struck several weeks earlier. Backed by the trillions of bytes' worth of shopper history that is stored in Wal-Mart's data warehouse, she felt that the company could 'start predicting what's going to happen, instead of waiting for it to happen,' as she put it. (Hays, 2004)

Provost/Fawcett - Data Science for Business

- ▶ How can you get a business opportunity from an hurricane and use data from previous events?
- ▶ Track unusual purchases due to hurricanes

What can be done with all the Data

Indeed, that is what happened. *The New York Times* (Hays, 2004) reported that: “... the experts mined the data and found that the stores would indeed need certain products—and not just the usual flashlights. ‘We didn’t know in the past that strawberry Pop-Tarts increase in sales, like seven times their normal sales rate, ahead of a hurricane,’ Ms. Dillman said in a recent interview. ‘And the pre-hurricane top-selling item was beer.’”¹

Provost/Fawcett - Data Science for Business

Another example

- ▶ TelCo, a major telecommunications firm, wants to investigate its problem with customer attrition, or “churn”
 - ▶ Lets consider this for now as a marketing problem only

Another example

- ▶ TelCo, a major telecommunications firm, wants to investigate its problem with customer attrition, or “churn”
 - ▶ Lets consider this for now as a marketing problem only
 - ▶ How would you go about targeting some customers with a special offer, prior to contract expiration?
 - ▶ Think about what data should be available for your use.

Another example


How Target Figured Out A Teen Girl Was Pregnant Before Her Father Did



Kashmir Hill, FORBES STAFF 

Welcome to The Not-So Private Parts where technology & privacy collide **FULL BIO** 

Every time you go shopping, you share intimate details about your consumption patterns with retailers. And many of those retailers are studying those details to figure out what you like, what you need, and which coupons are most likely to make you happy.

Target  **TGT +0.95%**, for example, has figured out how to data-mine its way into your womb, to figure out whether you have a baby on the way long before you need to start buying diapers.



Target has got you in its aim

Another example

The Obama 2012 campaign famously used [big data predictive analytics](#) to influence individual voters. They hired more than 50 analytics experts, including [data scientists](#), to predict which voters will be positively persuaded by political campaign contact such as a call, door knock, flyer, or TV ad. Uplift modeling (aka persuasion modeling) is one of the hottest forms of predictive analytics, for obvious reasons — most organizations wish to persuade people to do something such as buy! In this special episode of Forrester TechnoPolitics, [Mike](#) interviews Eric Siegel, Ph.D., author of [Predictive Analytics](#), to find out: 1) What exactly is uplift modeling? and 2) How did the Obama 2012 campaign use it to persuade voters? (< 4 minutes)



Applications of data science

- ▶ Data science is applied not only in marketing

Applications of data science

- ▶ Data science is applied not only in marketing
 - ▶ Fraud detection
 - ▶ Algorithmic trading
 - ▶ Spam filters
 - ▶ Bioinformatics
 - ▶ History, Literature, ...

The challenge

- ▶ To identify valid, novel, useful patterns within large scale data

Data analytics

- ▶ Analysing data is not only about science

Data analytics

- ▶ Analysing data is not only about science
- ▶ It also requires
 - ▶ Craft
 - ▶ Creativity
 - ▶ Common sense

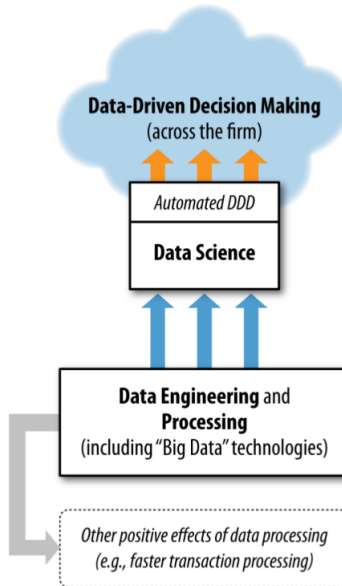
Data science and data mining

- ▶ **Data science** involves principles, processes, and techniques for understanding phenomena via the (automated) analysis of data
- ▶ **Data mining** (or “analytics”) includes the techniques used in data science
- ▶ **Data-Driven Decision-Making (DDD)** refers to the practice of basing decisions on the analysis of data, rather than purely on intuition
- ▶ We will not make a distinction between them

Data science and big data

- ▶ **Big Data** - data sets that are too large for traditional data processing systems (usually do not fit into RAM memory), and therefore require new processing technologies
- ▶ **Big Data Technologies** are used to process and handle big data, and include pre-processing prior to implementing data mining techniques
 - ▶ can also be used in conjunction with specific implementations of data mining methods (e.g., speed it up by parallelizing)

Data science for business



Roles in data mining

- ▶ Data scientist
 - ▶ Does data modeling
 - ▶ Can translate from business to execution
 - ▶ Knows enough statistics and computer science
- ▶ Collaborator in a data-centric project
 - ▶ Can translate from business to the execution
- ▶ Managing a data-mining project
 - ▶ Understanding the potential
 - ▶ Ability to evaluate a proposal and execution
 - ▶ Ability to interface with a broad variety of people
- ▶ Strategist, Investor, ...
 - ▶ envision opportunities, come up with novel ideas, evaluate the promise of new ideas, ...

Course goals

- ▶ Help you view business problems from a data-analytic perspective:
 - ▶ Understand the basic concepts of data science, important considerations, opportunities and pitfalls
 - ▶ Become familiar with various business applications and uses of data science.

ADMINISTRATION



- ▶ **Instructor:** Tomer Libal (tlibal@aup.edu)
- ▶ **Office hours:** On calendly.com/tlibal
- ▶ **Blended classroom:** Some classes will be supported by Coursera
- ▶ **Grades:** 10% midterm, 20% participation and attendance, 40% exercises, 30% final project
- ▶ **Attendance:** Mandatory
- ▶ **Books:** Data Science for Business by Foster Provost and Tom Fawcett
- ▶ **Course homepage:** on Canvas

aula.education

- ▶ Main channel of communication
- ▶ Class material and discussions
- ▶ Questions and notifications
- ▶ Canvas will handle mini projects and grading as well as syllabus
- ▶ Weekly participation in Aula contributes to the grade

Exercises

- ▶ All information on the assignment tab in Canvas

Technology

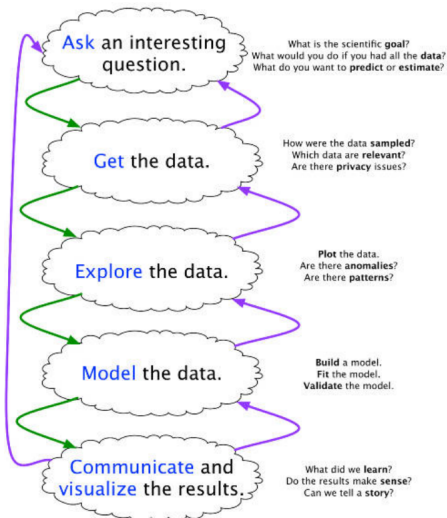
- ▶ Programming in Python
- ▶ Coding in Jupyter notebook



Getting familiar with Aula

- ▶ Instant messages
- ▶ Direct messages
- ▶ **Exercise:** Introduce each other - background, motivation, expectations, experience

The data science process



Joe Blitzstein and Hanspeter Pfister, created for the Harvard data science course <http://cs109.org/>.

Analyzing Hubway Data

- ▶ Introduction: Hubway is metro-Boston's public bike share program, with more than 1600 bikes at 160+ stations across the Greater Boston area. Hubway is owned by four municipalities in the area. By 2016, Hubway operated 185 stations and 1750 bicycles, with 5 million ride since launching in 2011
- ▶ The Data: In April 2017, Hubway held a Data Visualization Challenge at the Microsoft NERD Center in Cambridge, releasing 5 years of trip data
- ▶ The Question: What does the data tell us about the ride share program?

The Data Exploration/Question Refinement Cycle

- ▶ Our original question:
- ▶ What does the data tell us about the ride share program?
- ▶ Not good enough for driving a scientific exploration

The Data Exploration/Question Refinement Cycle

- ▶ Our original question:
- ▶ What does the data tell us about the ride share program?
- ▶ Not good enough for driving a scientific exploration

- ▶ Before we can refine the question, we have to look at the data!

	seq_id	hubway_id	status	duration	start_date	strt_statn	end_date	end_statn	bike_nr	subsc_type	zip_code	birth_date	gender
0	1	8	Closed	9	7/28/2011 10:12:00	23.0	7/28/2011 10:12:00	23.0	B00468	Registered	'97217	1976.0	Male
1	2	9	Closed	220	7/28/2011 10:21:00	23.0	7/28/2011 10:25:00	23.0	B00554	Registered	'02215	1966.0	Male
2	3	10	Closed	56	7/28/2011 10:33:00	23.0	7/28/2011 10:34:00	23.0	B00456	Registered	'02108	1943.0	Male
3	4	11	Closed	64	7/28/2011 10:35:00	23.0	7/28/2011 10:36:00	23.0	B00554	Registered	'02116	1981.0	Female
4	5	12	Closed	12	7/28/2011 10:37:00	23.0	7/28/2011 10:37:00	23.0	B00554	Registered	'97214	1983.0	Female

The Data Exploration/Question Refinement Cycle

- ▶ Our original question:
- ▶ What does the data tell us about the ride share program?
- ▶ Not good enough for driving a scientific exploration
- ▶ Before we can refine the question, we have to look at the data!

	seq_id	hubway_id	status	duration	start_date	strt_statn	end_date	end_statn	bike_nr	subsc_type	zip_code	birth_date	gender
0	1	8	Closed	9	7/28/2011 10:12:00	23.0	7/28/2011 10:12:00	23.0	B00468	Registered	'97217	1976.0	Male
1	2	9	Closed	220	7/28/2011 10:21:00	23.0	7/28/2011 10:25:00	23.0	B00554	Registered	'02215	1966.0	Male
2	3	10	Closed	56	7/28/2011 10:33:00	23.0	7/28/2011 10:34:00	23.0	B00456	Registered	'02108	1943.0	Male
3	4	11	Closed	64	7/28/2011 10:35:00	23.0	7/28/2011 10:36:00	23.0	B00554	Registered	'02116	1981.0	Female
4	5	12	Closed	12	7/28/2011 10:37:00	23.0	7/28/2011 10:37:00	23.0	B00554	Registered	'97214	1983.0	Female

- ▶ Based on the data, what kind of questions can we ask?

The Data Exploration/Question Refinement Cycle

- ▶ **Who?** Who's using the bikes?
 - ▶ Refine into specific hypotheses:

The Data Exploration/Question Refinement Cycle

- ▶ **Who?** Who's using the bikes?
 - ▶ Refine into specific hypotheses:
 - ▶ More men or more women?

The Data Exploration/Question Refinement Cycle

- ▶ **Who?** Who's using the bikes?
 - ▶ Refine into specific hypotheses:
 - ▶ More men or more women?
 - ▶ Older or younger people?

The Data Exploration/Question Refinement Cycle

- ▶ **Who?** Who's using the bikes?
 - ▶ Refine into specific hypotheses:
 - ▶ More men or more women?
 - ▶ Older or younger people?
 - ▶ Subscribers or one time users?

The Data Exploration/Question Refinement Cycle

- ▶ **Where?** Where are bikes being checked out?
 - ▶ Refine into specific hypotheses:

The Data Exploration/Question Refinement Cycle

- ▶ **Where?** Where are bikes being checked out?
 - ▶ Refine into specific hypotheses:
 - ▶ More in Boston than Cambridge?

The Data Exploration/Question Refinement Cycle

- ▶ **Where?** Where are bikes being checked out?
 - ▶ Refine into specific hypotheses:
 - ▶ More in Boston than Cambridge?
 - ▶ More in commercial or residential?

The Data Exploration/Question Refinement Cycle

- ▶ **Where?** Where are bikes being checked out?
 - ▶ Refine into specific hypotheses:
 - ▶ More in Boston than Cambridge?
 - ▶ More in commercial or residential?
 - ▶ More around tourist attractions?

The Data Exploration/Question Refinement Cycle

- ▶ **Where?** Where are bikes being checked out?
 - ▶ Refine into specific hypotheses:
 - ▶ More in Boston than Cambridge?
 - ▶ More in commercial or residential?
 - ▶ More around tourist attractions?
- ▶ Sometimes the data is given to you in pieces and must be merged!

The Data Exploration/Question Refinement Cycle

- ▶ **When?** When are the bikes being checked out?
 - ▶ Refine into specific hypotheses:

The Data Exploration/Question Refinement Cycle

- ▶ **When?** When are the bikes being checked out?
 - ▶ Refine into specific hypotheses:
 - ▶ More during the weekend than on the weekdays?

The Data Exploration/Question Refinement Cycle

- ▶ **When?** When are the bikes being checked out?
 - ▶ Refine into specific hypotheses:
 - ▶ More during the weekend than on the weekdays?
 - ▶ More during rush hour?

The Data Exploration/Question Refinement Cycle

- ▶ **When?** When are the bikes being checked out?
 - ▶ Refine into specific hypotheses:
 - ▶ More during the weekend than on the weekdays?
 - ▶ More during rush hour?
 - ▶ More during the summer than the fall?

The Data Exploration/Question Refinement Cycle

- ▶ **When?** When are the bikes being checked out?
 - ▶ Refine into specific hypotheses:
 - ▶ More during the weekend than on the weekdays?
 - ▶ More during rush hour?
 - ▶ More during the summer than the fall?
- ▶ Sometimes the feature you want to explore doesn't exist in the data, and must be engineered!

The Data Exploration/Question Refinement Cycle

- ▶ **Why?** For what reasons/activities are people checking out bikes?
 - ▶ Refine into specific hypotheses:

The Data Exploration/Question Refinement Cycle

- ▶ **Why?** For what reasons/activities are people checking out bikes?
 - ▶ Refine into specific hypotheses:
 - ▶ More bikes are used for recreation than commute?

The Data Exploration/Question Refinement Cycle

- ▶ **Why?** For what reasons/activities are people checking out bikes?
 - ▶ Refine into specific hypotheses:
 - ▶ More bikes are used for recreation than commute?
 - ▶ More bikes are used for touristic purposes?

The Data Exploration/Question Refinement Cycle

- ▶ **Why?** For what reasons/activities are people checking out bikes?
 - ▶ Refine into specific hypotheses:
 - ▶ More bikes are used for recreation than commute?
 - ▶ More bikes are used for touristic purposes?
 - ▶ Bikes are use to bypass traffic?

The Data Exploration/Question Refinement Cycle

- ▶ **Why?** For what reasons/activities are people checking out bikes?
 - ▶ Refine into specific hypotheses:
 - ▶ More bikes are used for recreation than commute?
 - ▶ More bikes are used for touristic purposes?
 - ▶ Bikes are use to bypass traffic?
- ▶ Do we have the data to answer these questions with reasonable certainty?
- ▶ What data do we need to collect in order to answer these questions?

The Data Exploration/Question Refinement Cycle

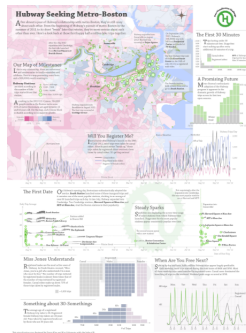
- ▶ **How?** Questions that combine variables
 - ▶ How does user demographics impact the duration the bikes are being used? Or where they are being checked out?

The Data Exploration/Question Refinement Cycle

- ▶ **How?** Questions that combine variables
 - ▶ How does user demographics impact the duration the bikes are being used? Or where they are being checked out?
 - ▶ How does weather or traffic conditions impact bike usage?
 - ▶ How do the characteristics of the station location affect the number of bikes being checked out?

Inspirations for Data Viz/Exploration

- ▶ Check out the winners of the Hubway Challenge
 - ▶ <http://hubwaydatachallenge.org/>



This is a visualization of the Highway shared-drive program and its adoption by the five people of the Boston water area. Viewed through the lenses of a romantic relationship, this visual diary displays an inventory of shared drives (2004-2006) across geographic, job/business, and personal dimensions.

Data science and creativity

- ▶ 70's - a famous university was sued for sex discrimination
 - ▶ Acceptance rate was significantly higher for men than for women

Data science and creativity

- ▶ 70's - a famous university was sued for sex discrimination
 - ▶ Acceptance rate was significantly higher for men than for women
- ▶ Data of six most popular departments:

Gender	Department	Admitted	Rejected
M	A	512	313
F	A	89	19
M	B	353	207
F	B	17	8
M	C	120	205
F	C	202	391
M	D	138	279
F	D	131	244
M	E	53	138
F	E	94	299
M	F	22	351
F	F	24	317

Data science and creativity

- ▶ 70's - a famous university was sued for sex discrimination
 - ▶ Acceptance rate was significantly higher for men than for women
- ▶ Data of six most popular departments:

Gender	Department	Admitted	Rejected
M	A	512	313
F	A	89	19
M	B	353	207
F	B	17	8
M	C	120	205
F	C	202	391
M	D	138	279
F	D	131	244
M	E	53	138
F	E	94	299
M	F	22	351
F	F	24	317

- ▶ What do you think?