Data Science

Introduction

Concepts

- Data
  - What we have , is data.

Concepts

- Data
  - What we have , is data.
- Information
  - What of the data to be used , is information

Data Science -Introduction

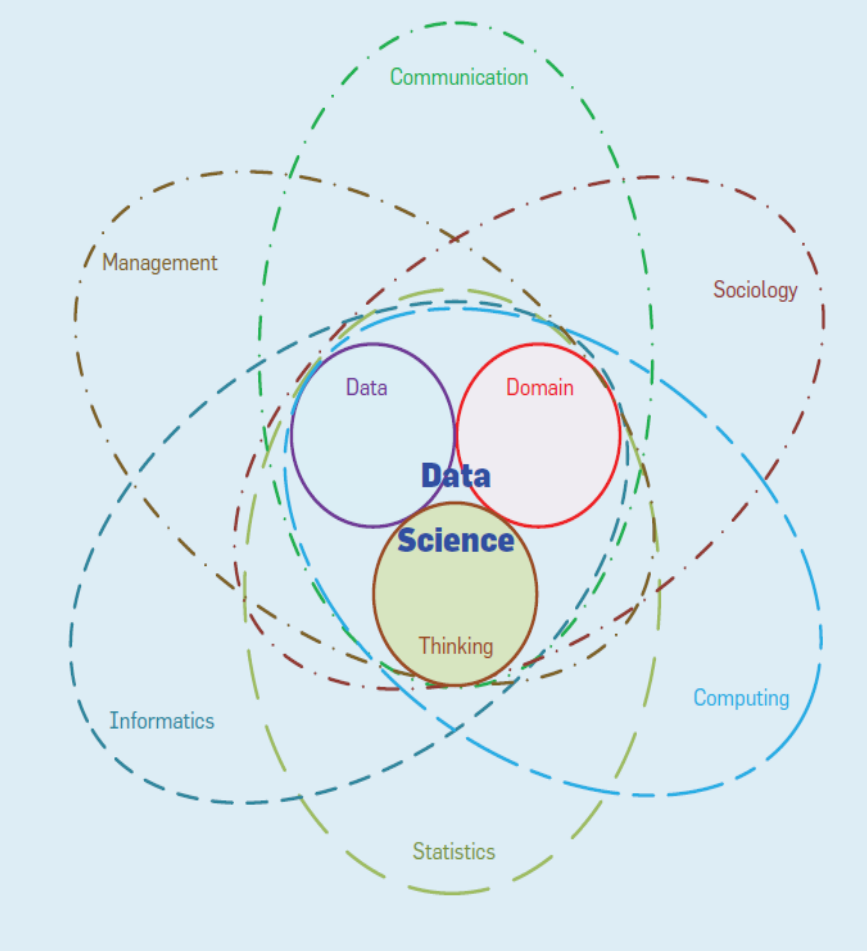
- Data Science
  - What is it?
  - Where can we find it?
  - How can we explore it?

Data Science - What is it?

- Data science is a new trans-disciplinary field that builds on and synthesizes a number of relevant disciplines and bodies of knowledge, including statistics, informatics, computing, communication, management, and sociology to study data following “data science thinking”.
- This data science formula can be specified as,

  Data science = {statistics ∩ informatics∩ computing ∩ communication ∩sociology ∩ management | data ∩domain ∩ thinking } where “|” means “conditional on.”

Transdisciplinary data science



COMMUNICATIONS OF THE ACM | AUGUST 2017 | VOL. 60 | NO. 8

- A core objective of data science is exploration of the complexities inherently trapped in data, business, and problem-solving systems
- Here, complexity refers to sophisticated characteristics, in terms of data (characteristics), behavior, domain, social factors, environment (context), learning (process and system), and deliverables.

Data Characteristics

- Data complexity is reflected in terms of sophisticated data characteristics, which includes –
  - Large scale,
  - High dimensionality,
  - Extreme imbalance,
  - online and real-time interaction and processing,
  - cross-media applications,
  - mixed sources,
  - Strong dynamics,
  - high frequency,
  - uncertainty,
  - noise mixed with data,
  - unclear structures,
  - unclear hierarchy,
  - Heterogeneous or unclear distribution,
  - Strong sparsity, and
  - unclear availability of specific sometimes critical data.

Data Complexity Perspectives

1) **Behavior complexity -**
   - refers to the challenges involved in understanding what actually takes place in business activities by connecting to the semantics and processes and behavioral subjects and objects in the physical world which often are ignored or simplified in the data world.
   - Behavior complexities are individual and group behaviors, behavior in networking, collective behaviors, behavior divergence and convergence, "non-occurring" behaviors, behavior-network evolution, group-behavior reasoning, recovery of what actually happened, happens, or will happen in the physical world from the highly deformed information collected in the purely data world.
   - Behavior complexity gives insights, impact, utility, and effect of behaviors, and the emergence and management of behavior intelligence.

Data Complexity Perspectives

Behavior complexity -

- However, limited systematic research outcomes are available for comprehensively quantifying, representing, analyzing, reasoning about, and managing complex behaviors.

Data Complexity Perspectives

2) Domain complexity -

- It is critical aspect of data science for discovering intrinsic data characteristics, value, and actionable insight.
- Domain complexities are reflected in a problem domain as domain factors, domain processes, norms, policies, qualitative-versus-quantitative domain knowledge, expert knowledge, hypotheses, meta-knowledge, Behavior complexity gives insights, impact, utility, and effect of behaviors, and the emergence and management of behavior intelligence involvement of and interaction with domain experts and professionals, multiple and cross-domain interactions, experience acquisition, human-machine synthesis, and roles and leadership in the domain.

Data Complexity Perspectives

3) **Social complexity -**

- Social complexity is embedded in business activity and its related data.
- It is a key part of data and business understanding
- It may be embodied in business problems as social networking, community emergence, social dynamics, impact evolution, social conventions, social contexts, social cognition, social intelligence, social media, group formation and evolution, group interaction and collaboration, economic and cultural factors, social norms, emotion, sentiment and opinion influence processes, and social issues, including security, privacy, trust, risk, and accountability in social contexts.

Data Complexity Perspectives

4) Environment complexity -

- Environment complexity is another important factor in understanding complex data and business problems.
- It is reflected in environmental (contextual) factors, contexts of problems and data, context dynamics, adaptive engagement of contexts, complex contextual interactions between the business environment and data systems, significant changes in business environment and their effect on data systems, and variations and uncertainty in interactions between business data and the business environment.
- If ignored, a model suitable for one domain might produce misleading outcomes in another, as is often seen in recommender systems.

Data Complexity Perspectives

5) Learning complexity -

- Learning (process and system) complexity must be addressed to achieve the goal of data analytics.
- Challenges in analyzing data include developing methodologies, common task frameworks, and learning paradigms to handle data, domain, behavioral, social, and environmental complexity.
- Data scientists must be able to learn from heterogeneous sources and inputs, parallel and distributed inputs, and their infinite dynamics in real time;
- support on-the-fly, active and adaptive learning of large data volumes in computational resource-poor environments (such as embedded sensors), as well as multi-source learning, while considering the relations and interactions between sensors;

Data Complexity Perspectives

Learning complexity -contd

- enable combined learning across multiple learning objectives, sources, feature sets, analytical methods, frameworks, and outcomes.
- Other requirements for managing and exploiting data include appropriate design of experiments and mechanisms. Inappropriate learning could result in misleading or harmful outcomes, as in a classifier that works for balanced data but could mistakenly classify biased and sparse cases in anomaly detection.

Data Complexity Perspectives

6) **Deliverables complexity -**

- The complexity of a deliverable data product, or "deliverable complexity" becomes an obstruction when actionable insight is the focus of a data science application.
- Such complexity necessitates identification and evaluation of the outcomes that satisfy technical significance and have high business value from both an objective and a subjective perspective.
- The related challenges for data scientists also involve designing the appropriate evaluation, presentation, visualization, refinement, and prescription of learning outcomes and deliverables to satisfy diverse business needs, stakeholders, and decision support.
- In general, data deliverables to business users must be easy to understand and interpretable by nonprofessionals, revealing insights that directly inform and enable decision making and possibly having a transformative effect on business processes and problem solving.

Big Data-There is a lots and lot of data

Data Sources

- According to Wall Street Journal, the digital universe will reach 180 zetabytes by 2025.
- The new economy is more about analyzing rapid real-time flows of data, often unstructured.
  - The streams of photos and videos generated by users of social networks
  - The ream of information produced by commuters on their way to work
  - The flood of data from hundreds of sensors in a jet engine
  - Data from subway trains and wind turbines
  - Uber, known for cheap taxi rides, owns the biggest pool of data about supply and demand for personal transportation.
  - Tesla, maker of fancy electric cars collect mountains of data, which allow the firm to optimize its self-driving algorithms and then update the software accordingly.

Data Sources

- An Israeli startup, has devised a way to use drivers as data sources. Its app turns their smart phones into dashcams that tag footage of their travels via actions they normally perform . If many unexpectedly hit the brake at the same spot on the road, this signals a pothole or another obstacle. The firms goal is provide services that help drivers avoid accidents.

Data Sources

- GE, a non-technical firm, has developed an “operating system” called Predix to help customers control their machinery.
- Predix is a data collection system, it pools data from devices it is connected to, mixes these with other data and then trains algorithms that can help improve the operations of a power plant, when to maintain a jet engine before it breaks down and so on.

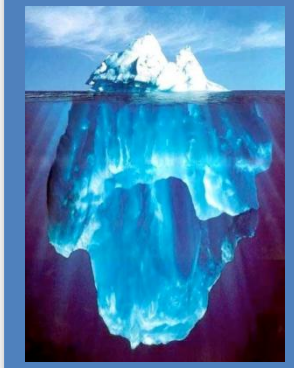
Data Sources

- More and more data generated through:

1. Facebook: data generated through photo sharing, text-photo messaging
2. Alphabet – mapping and navigation information
3. IBM – meterology data(weather data), health care data
4. INTEL –self-driving cars(Mobileye 2017- *Mobileye*, an *Intel* company, is a leader in automated technology and the world's largest supplier of cameras for advanced driver assistance systems (ADAS))
5. Microsoft – keyboard/ AI generated data , business networking data(Linkedln)
6. Oracle – Cloud data platform, marketing data

“Big Data” Sources

It’s All Happening On-line



Every:
Click
Ad impression
Billing event
Fast Forward, pause,…
Server request
Transaction
Network message
Fault
…

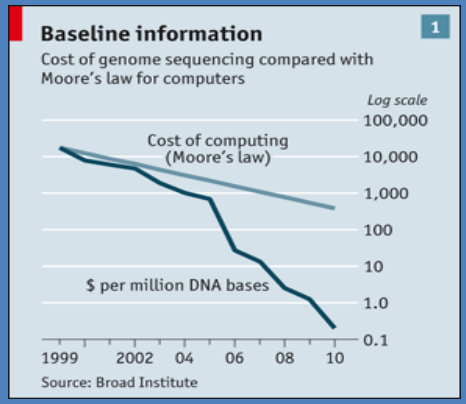
User Generated (Web & Mobile)




….
.


Internet of Things / M2M

Health/Scientific Computing

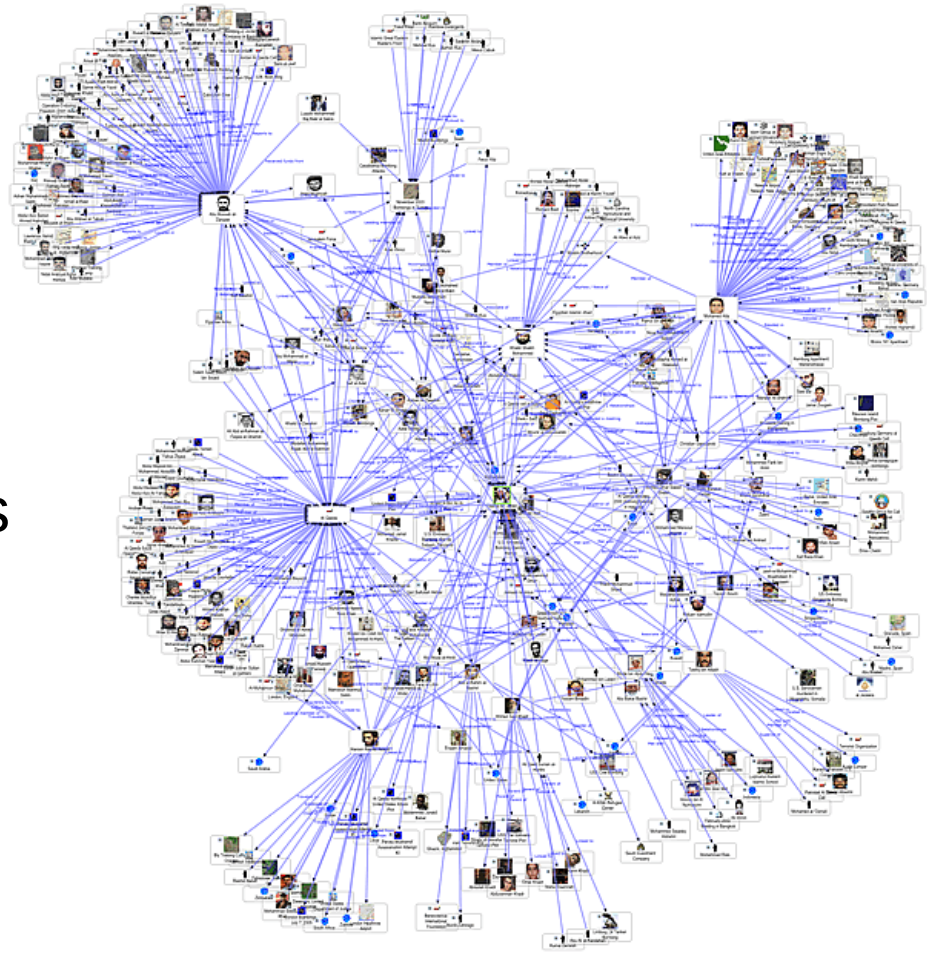


Graph Data

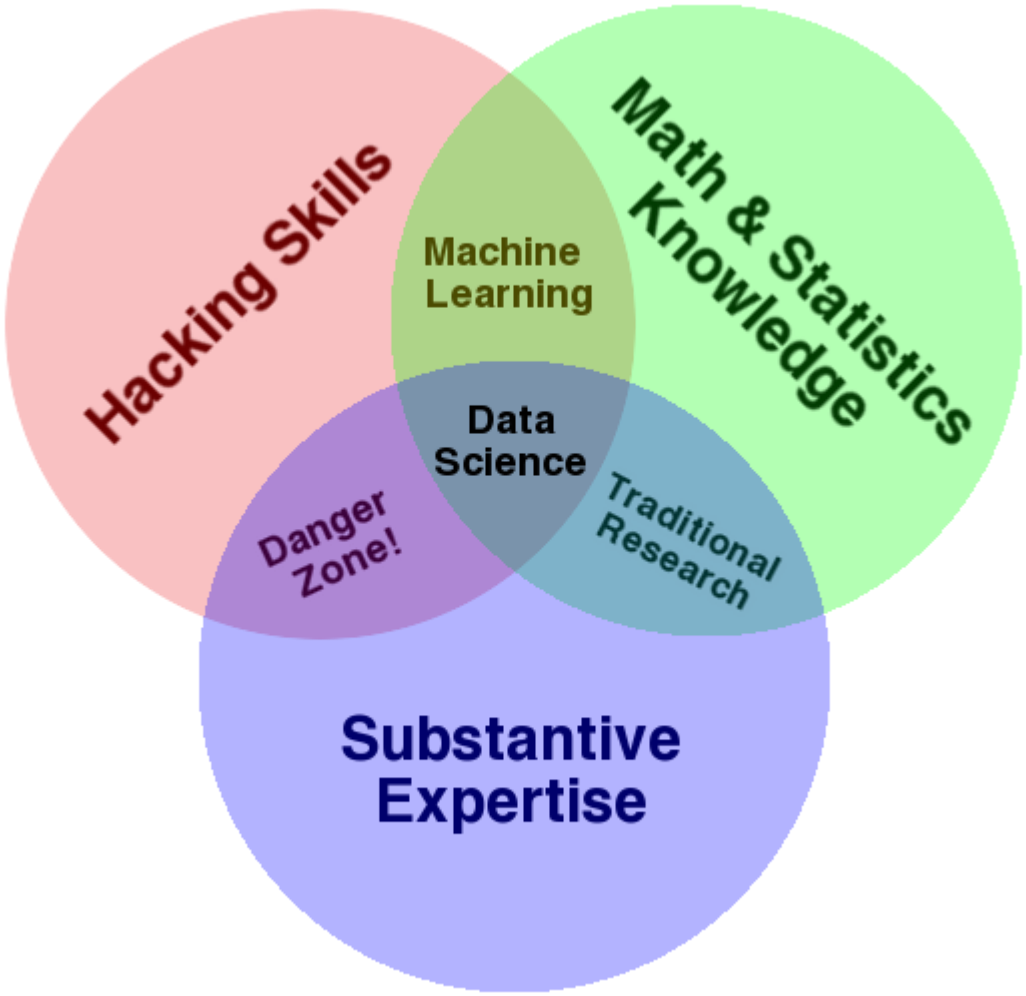
Lots of interesting data has a graph structure:

- Social networks
- Communication networks
- Computer Networks
- Road networks
- Citations
- Collaborations/Relationships
- …

Some of these graphs can get quite large (e.g., Facebook* user graph)



Data Science – One Definition



Contrast: Databases

	Databases	Data Science
Data Volume	Modest	Massive
Examples	Bank records, Personnel records, Census, Medical records	Online clicks, GPS logs, Tweets, Building sensor readings
Priorities	Consistency, Error recovery, Auditability	Speed, Availability, Query richness
Structured	Strongly (Schema)	Weakly or none (Text)
Properties	Transactions, ACID*	CAP* theorem (2/3), eventual consistency
Realizations	SQL	NoSQL: Riak, Memcached, Apache River, MongoDB, CouchDB, Hbase, Cassandra,…
Querying	Querying the past	Querying the future

ACID = Atomicity, Consistency, Isolation and Durability CAP = Consistency, Availability, Partition Tolerance

Learning

Definition:

- Learning in a broad sense can be defined as any computer program that improves its performance at some task through experience.

Well-defined learning problem

- A computer program is said to learn from experience E with respect to some class of tasks T and performance measure P, if its performance at tasks in T, as measured by P, improves with experience E.

Essential features of Well-defined learning problem

Three features:

1. the class of tasks
2. the measure of performance to be improved
3. the source of experience

Case study 1: Spam/Not spam emails

- Suppose your email program watches which emails you do or do not mark as spam, and based on that learns how to better filter spam. What is the task T in this setting?
  - **The task T** is classifying emails as spam or not spam
  - **The experience E** is watching you label emails as spam or not spam
  - **The performance P** is the number of emails correctly classified as spam or not spam

Case study 2: Checkers learning problem

A computer program that learns to play **checkers** might improve its performance as measured by its ability to win at the class of tasks involving playing checkers games, through experience obtained by playing games against itself.

- **Task T :** playing checkers
- **Performance measure P :** percent of games won against opponents
- **Training experience E :** playing practice games against itself

Case study 3: Handwriting recognition learning

- **Task T :** recognizing and classifying handwritten words within images
- **Performance measure P :** percent of words correctly classified
- **Training experience E:** a database of handwritten words with given classifications

Learning Algorithms

- Supervised learning
- Un-supervised learning

Learning definition

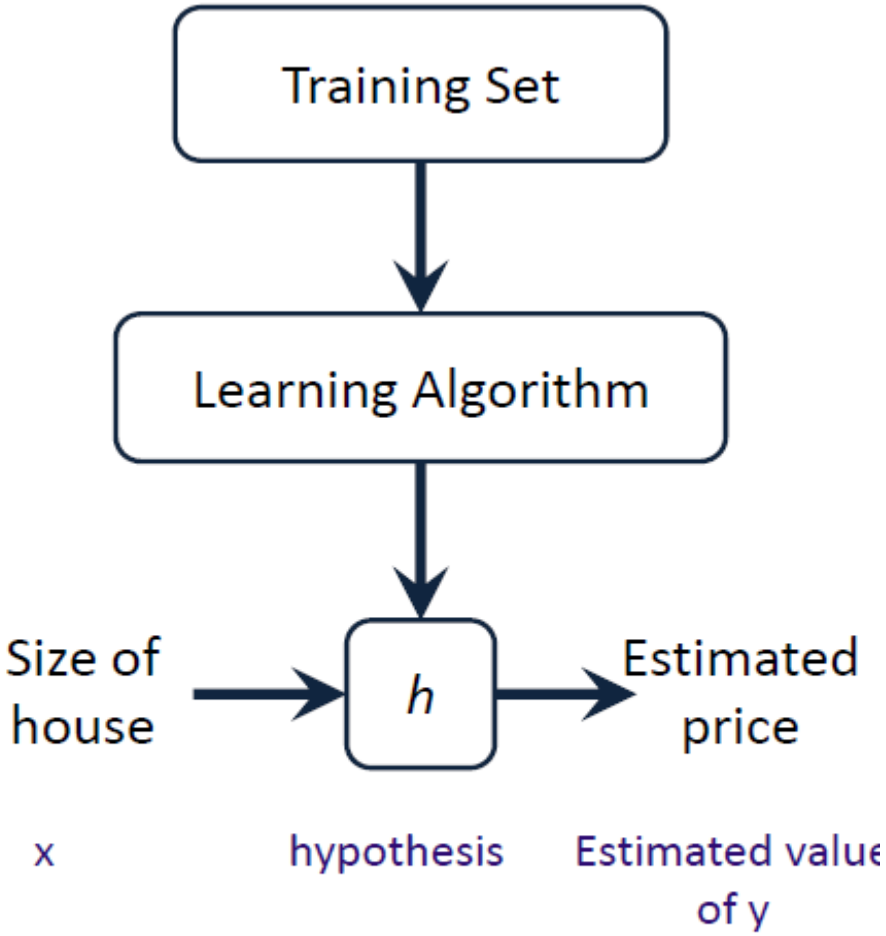
The Learner :

–Input : a set of m hand-labeled documents

(x1, y1),(x2, y2),------,(xm, ym)

–Output: a learned model h:x → y

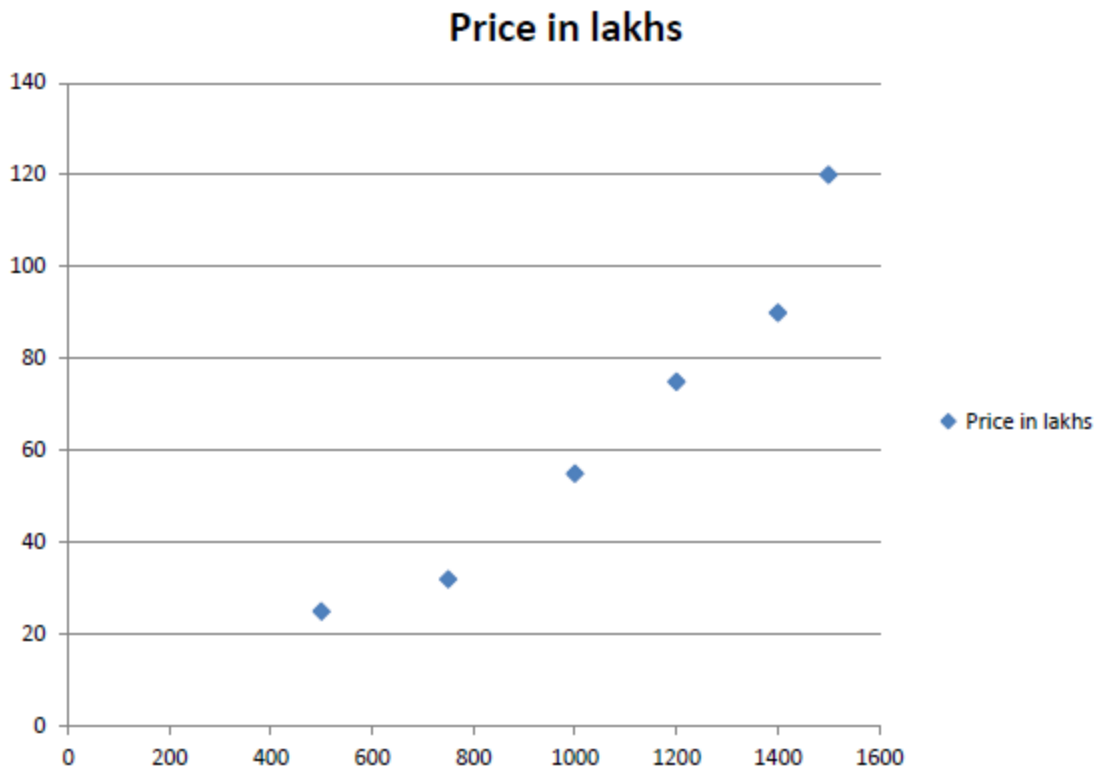
Housing Price Prediction



h maps from x's to y's

Example 1: Housing price prediction

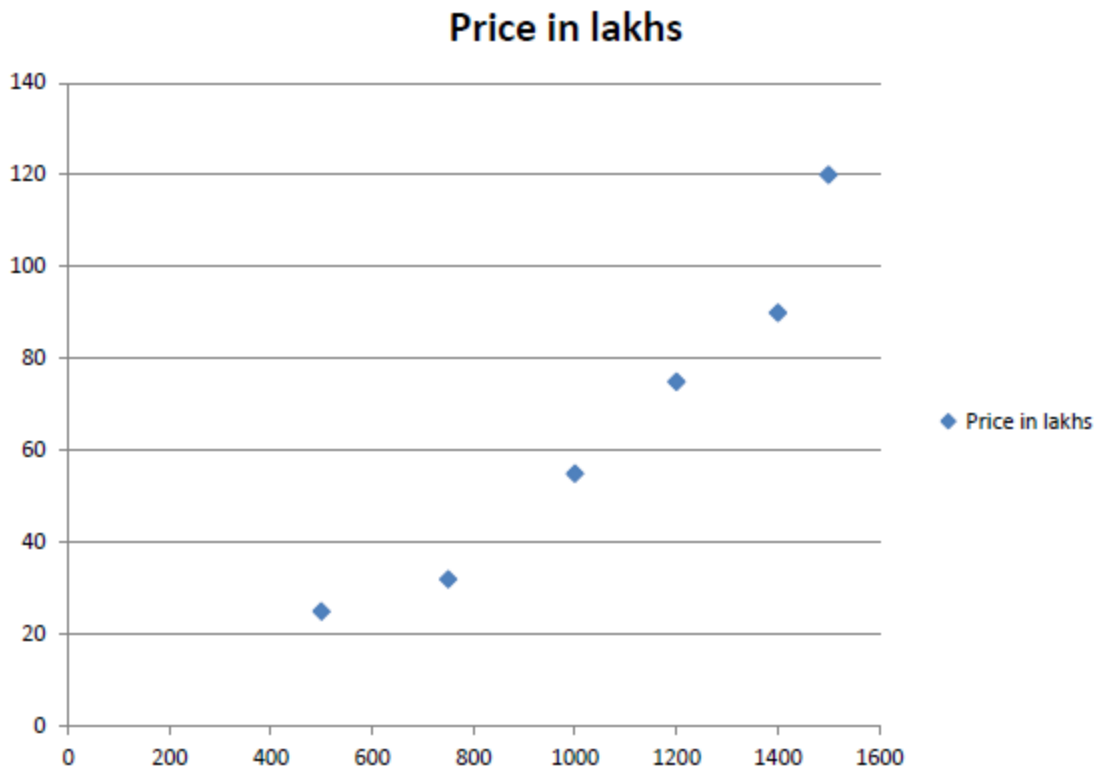
Size in Feet^2	Price in lakhs
500	25
750	32
1000	55
1200	75
1400	90
1500	120



Supervised Learning : Labeled data is given.

Example: Housing price prediction

Size in Feet^2	Price in lakhs
500	25
750	32
1000	55
1200	75
1400	90
1500	120



Supervised Learning : Labeled data is given.

Regression : Predict continuous valued output(price)

Example2 : Positive/Negative Sentiment Prediction

Doc ID	Sentiment
D1	+ve
D2	+ve
D3	-ve
…	
…	
D1000	-ve

Positive Sentiment features : good, extraordinary, cool, awesome, attractive, special, etc.,

Negative Sentiment features : not good, bad, worse, hate, sad, abused, awkward, dark, etc.,

You're running a company, and you want to develop learning algorithms to address each of two problems.

Problem 1: You have a large inventory of identical items. You want to predict how many of these items will sell over the next 3 months.
Problem 2: You'd like software to examine individual customer accounts, and for each account decide if it has been hacked/compromised.
Should you treat these as classification or as regression problems?

1. Treat both as classification problems.
2. Treat problem 1 as a classification problem, problem 2 as a regression problem.
3. Treat problem 1 as a regression problem, problem 2 as a classification problem.
4. Treat both as regression problems.

Unsupervised Learning

Unsupervised Learning

Example 1:Given a collection of text documents, organize them according to content similarity, to produce a topic hierarchy.

Unsupervised Learning

Example 1:Given a collection of text documents, organize them according to content similarity, to produce a topic hierarchy.

Example 2: In marketing, segment customers according to similarities, to do targeted marketing.

Unsupervised Learning

Example 1:Given a collection of text documents, organize them according to content similarity, to produce a topic hierarchy.

Example 2: In marketing, segment customers according to similarities, to do targeted marketing.

Example 3: On social networks, identifying research communities working on same problem.

Of the following examples, which would you address using an unsupervised learning algorithm? (Select all that apply.)

- Given email labeled as spam/not spam, learn a spam filter.
- Given a set of news articles found on the web, group them into set of articles about the same story.
- Given a database of customer data, automatically discover market segments and group customers into different market segments.
- Given a dataset of patients diagnosed as either having diabetes or not, learn to classify new patients as having diabetes or not.