Big Data Applications
IS71061A

- Dr. Kate (Katayoun) Farrahi

About Me

- Research
  - machine learning
  - data science
  - computational social science
  - mobile sensing (Reality Mining)
  - epidemiology
  - music recommendation

Today’s Agenda:

1. introduction to data science & big data
2. introduction to the course
3. introduction to python

What is a data scientist?

- “A data scientist is that unique blend of skills that can both unlock the insights of data and tell a fantastic story via the data.”–DJ Patil
- “A data scientist is someone who blends, math, algorithms, and an understanding of human behavior with the ability to hack systems together to get answers to interesting human questions from data.”–Hilary Mason
- “The four qualities of a great data scientist are creativity, tenacity, curiosity, and deep technical skills. They use skills in data gathering and data munging, visualization, machine learning, and computer programming to make data driven decisions and data driven products. They prefer to let the data do the talking.”–Jeremy Howard
- “By definition all scientists are data scientists. In my opinion, they are half hacker, half analyst, they use data to build products and find insights. It’s Columbus meet Columbo – starry eyed explorers and skeptical detectives.”–Monica Rogati

Data scientists

The New Rock Stars of the Tech World

DATA
is the new oil

Big data

data sets that are too large and complex to manipulate or interrogate with standard methods or tools

3 (4) V’s of Big Data - Dimensions of Data

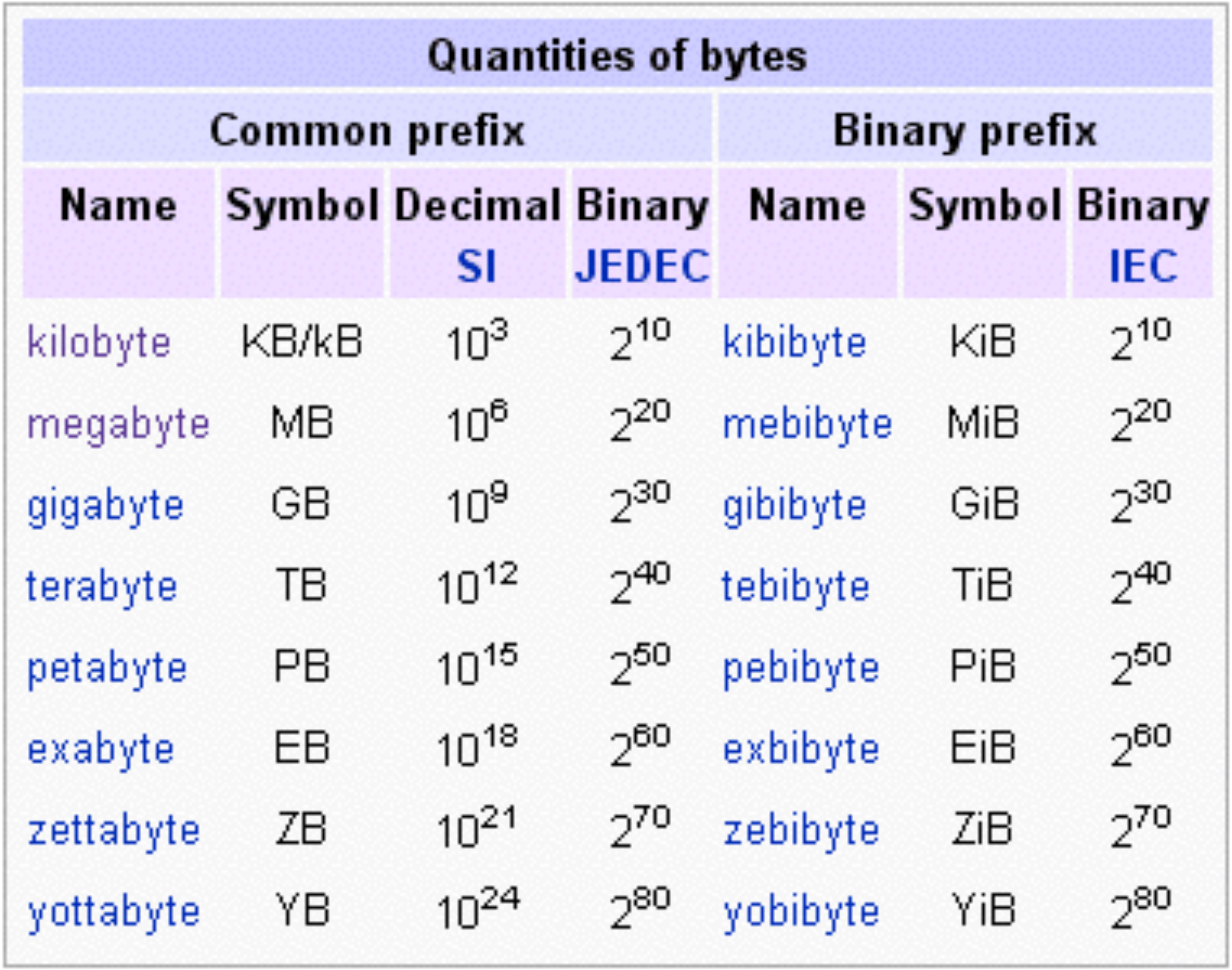
The challenges of “big data” result from the expansion of all these properties

- **Volume**: refers to the scale of data
  - By 2020, ~40 Zettabytes of data will be created
  - It is estimated 2.5 Quintillion bytes of data are created/day

- **Variety**: refers to the different forms of data

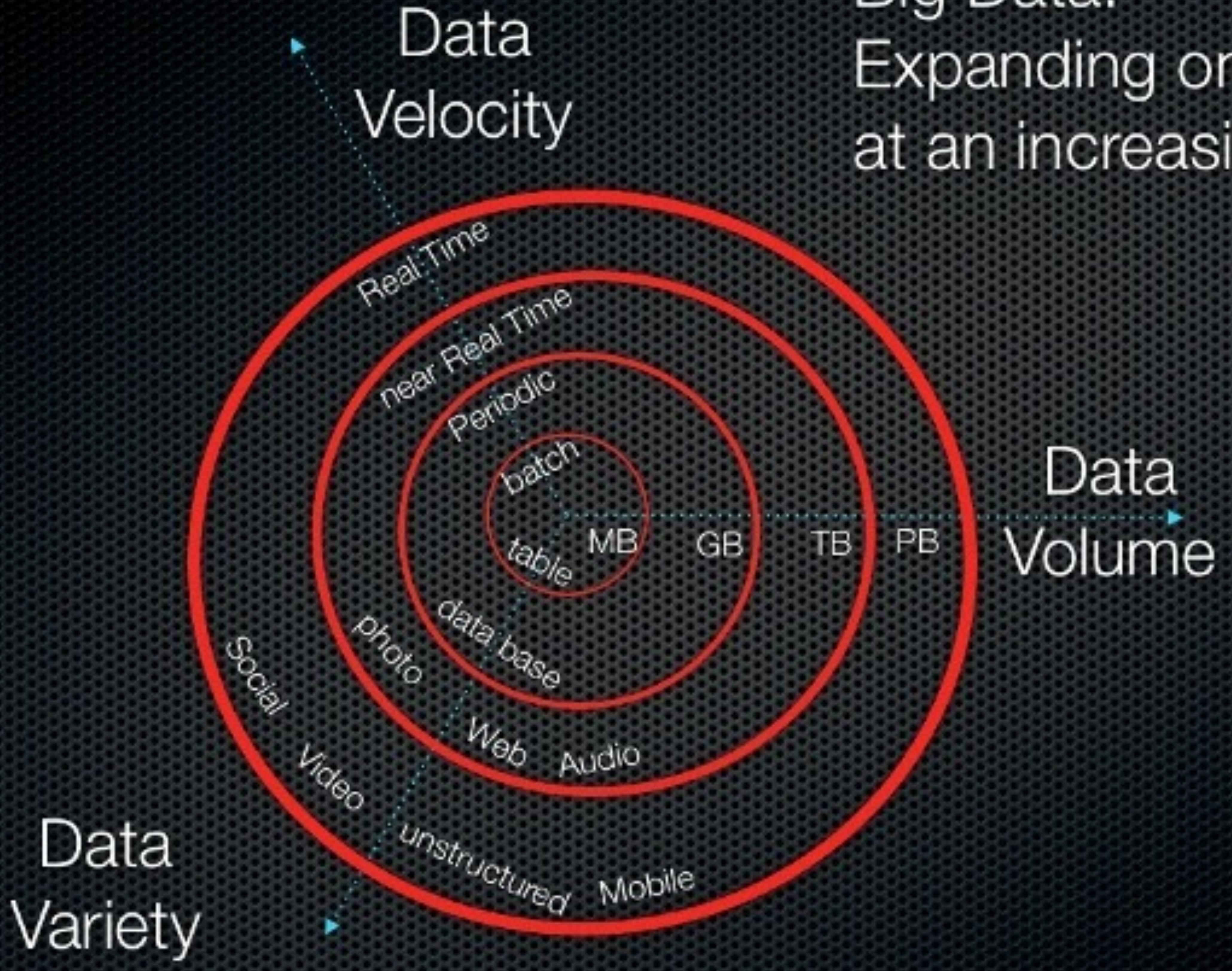
- **Velocity**: data is streaming into servers in real time, and the result of data analysis is often only useful if the delay is very short

- (**Veracity**): refers to the uncertainty of data
  - poor data quality
  - accuracy of data uncertain



Quantities of bytes						
Common prefix				Binary prefix		
Name	Symbol	Decimal	Binary	Name	Symbol	Binary
		SI	JEDEC			IEC
kilobyte	KB/kB	10^3	2^{10}	kibibyte	KiB	2^{10}
megabyte	MB	10^6	2^{20}	mebibyte	MiB	2^{20}
gigabyte	GB	10^9	2^{30}	gibibyte	GiB	2^{30}
terabyte	TB	10^{12}	2^{40}	tebibyte	TiB	2^{40}
petabyte	PB	10^{15}	2^{50}	pebibyte	PiB	2^{50}
exabyte	EB	10^{18}	2^{60}	exbibyte	EiB	2^{60}
zettabyte	ZB	10^{21}	2^{70}	zebibyte	ZiB	2^{70}
yottabyte	YB	10^{24}	2^{80}	yobibyte	YiB	2^{80}

Big Data: Expanding on 3 fronts at an increasing rate.



Google trends

- Google trends shows how often a particular search-term is searched for relative to the total number of searches, globally
- https://www.google.com/trends/explore#q=big%20data%2C%20data%20science%2C%20machine%20learning&cmpt=q

By 2018, the United States alone could face a shortage of 140,000 to 190,000 people with deep analytical skills as well as 1.5 million managers and analysts with the know-how to use the analysis of big data to make effective decisions [1].

[1] http://www.mckinsey.com/insights/business_technology/big_data_the_next_frontier_for_innovation

[E.g.] http://www.xcedesolutions.com/jobs/2516751/junior-data-scientist-london-retail.asp

introductions

- your name, background (education level, field of study)
- your personal interest in this program? why have you chosen this program?
- what are your future plans (grad school, industry, start your own company)?
- your experience with computer science?
  - have you used databases?
  - have you programmed before? (your level)
  - which programming languages are you familiar with?
- something fun about yourself

Topics

- Term 1 topics: Python, Relational databases, MapReduce, Hadoop, Recommender systems, (PageRank*, Hits*)

Reading group discussion topics:

Reality mining, computational social science, data in finance, social networks, wearable sensing

- Term 2 topics: NoSQL databases, MongoDB, Pig, Hive, social network analysis, web crawling, data visualization, (introduction to deep learning*), (Amazon web services*)

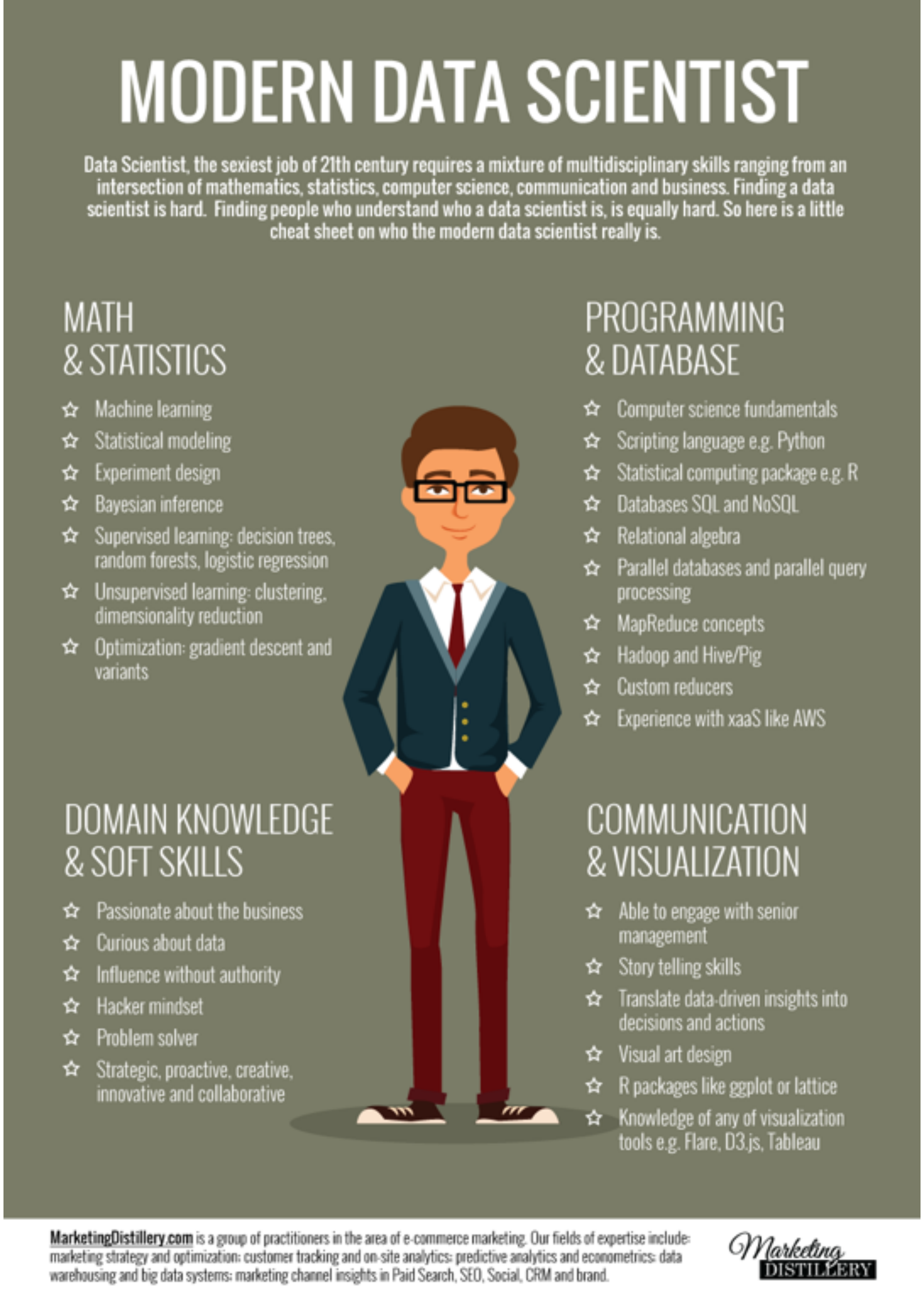
* advanced topics will depend on time, progress and interest

Assessment

- All information on the course page: learn.gold.ac.uk
- Add IS71061A to your course list
- Three graded assignments (30%, 30%, 40%)
- Deadlines: end of this term, early next term, end of next term

Reading Group Discussion

- List of reading group discussion papers will be posted on learn.gold.
  - Read through the paper abstracts and select the ones you are interested in
  - You are free to choose papers of interest which are NOT on the list as well (they need my approval)
  - Email me a list of your top 3 paper choices
  - Deadline to email paper choices: Oct 15th



MODERN DATA SCIENTIST

Data Scientist, the sexiest job of 21th century requires a mixture of multidisciplinary skills ranging from an intersection of mathematics, statistics, computer science, communication and business. Finding a data scientist is hard. Finding people who understand who a data scientist is, is equally hard. So here is a little cheat sheet on who the modern data scientist really is.

MATH & STATISTICS

- ☆ Machine learning
- ☆ Statistical modeling
- ☆ Experiment design
- ☆ Bayesian inference
- ☆ Supervised learning: decision trees, random forests, logistic regression
- ☆ Unsupervised learning: clustering, dimensionality reduction
- ☆ Optimization: gradient descent and variants

PROGRAMMING & DATABASE

- ☆ Computer science fundamentals
- ☆ Scripting language e.g. Python
- ☆ Statistical computing package e.g. R
- ☆ Databases SQL and NoSQL
- ☆ Relational algebra
- ☆ Parallel databases and parallel query processing
- ☆ MapReduce concepts
- ☆ Hadoop and Hive/Pig
- ☆ Custom reducers
- ☆ Experience with xaaS like AWS

DOMAIN KNOWLEDGE & SOFT SKILLS

- ☆ Passionate about the business
- ☆ Curious about data
- ☆ Influence without authority
- ☆ Hacker mindset
- ☆ Problem solver
- ☆ Strategic, proactive, creative, innovative and collaborative

COMMUNICATION & VISUALIZATION

- ☆ Able to engage with senior management
- ☆ Story telling skills
- ☆ Translate data-driven insights into decisions and actions
- ☆ Visual art design
- ☆ R packages like ggplot or lattice
- ☆ Knowledge of any of visualization tools e.g. Flare, D3.js, Tableau

MarketingDistillery.com is a group of practitioners in the area of e-commerce marketing. Our fields of expertise include: marketing strategy and optimization; customer tracking and on-site analytics; predictive analytics and econometrics; data warehousing and big data systems; marketing channel insights in Paid Search, SEO, Social, CRM and brand.