Lecture 1

Data Science for Economics and Finance

What will you learn in this class?

Data Scientist:

The Sexiest Job of the 21st Century

Meet the people who can coax treasure out of messy, unstructured data.

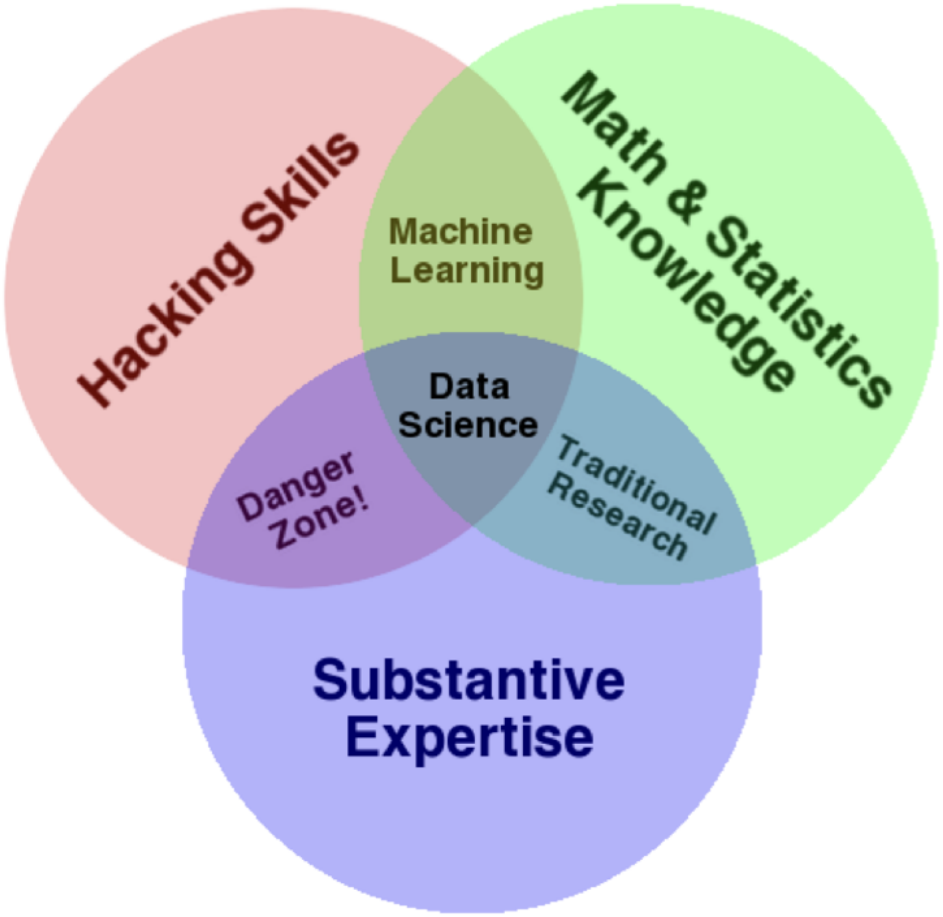
by Thomas H. Davenport and D.J. Patil

When Jonathan Goldman arrived for work in June 2006 at LinkedIn, the business networking site, the place still felt like a start-up. The company had just under 8 million accounts, and the number was growing quickly as existing members invited their friends and colleagues to join. But users weren't seeking out connections with the people who were already on the site at the rate executives had expected. Something was apparently missing in the social experience. As one LinkedIn manager put it, "It was like arriving at a conference reception and realizing you don't know anyone. So you just stand in the corner sipping your drink—and you probably leave early."

HBR Harvard Business Review October 2012


```
"I keep saying the sexy job in the next ten years will be statisticians.
People think I'm joking, but who would've guessed that computer
engineers would've been the sexy job of the 1990s?"
– Hal Varian (Chief Economist at Google, 2009).
```

What is Data Science?



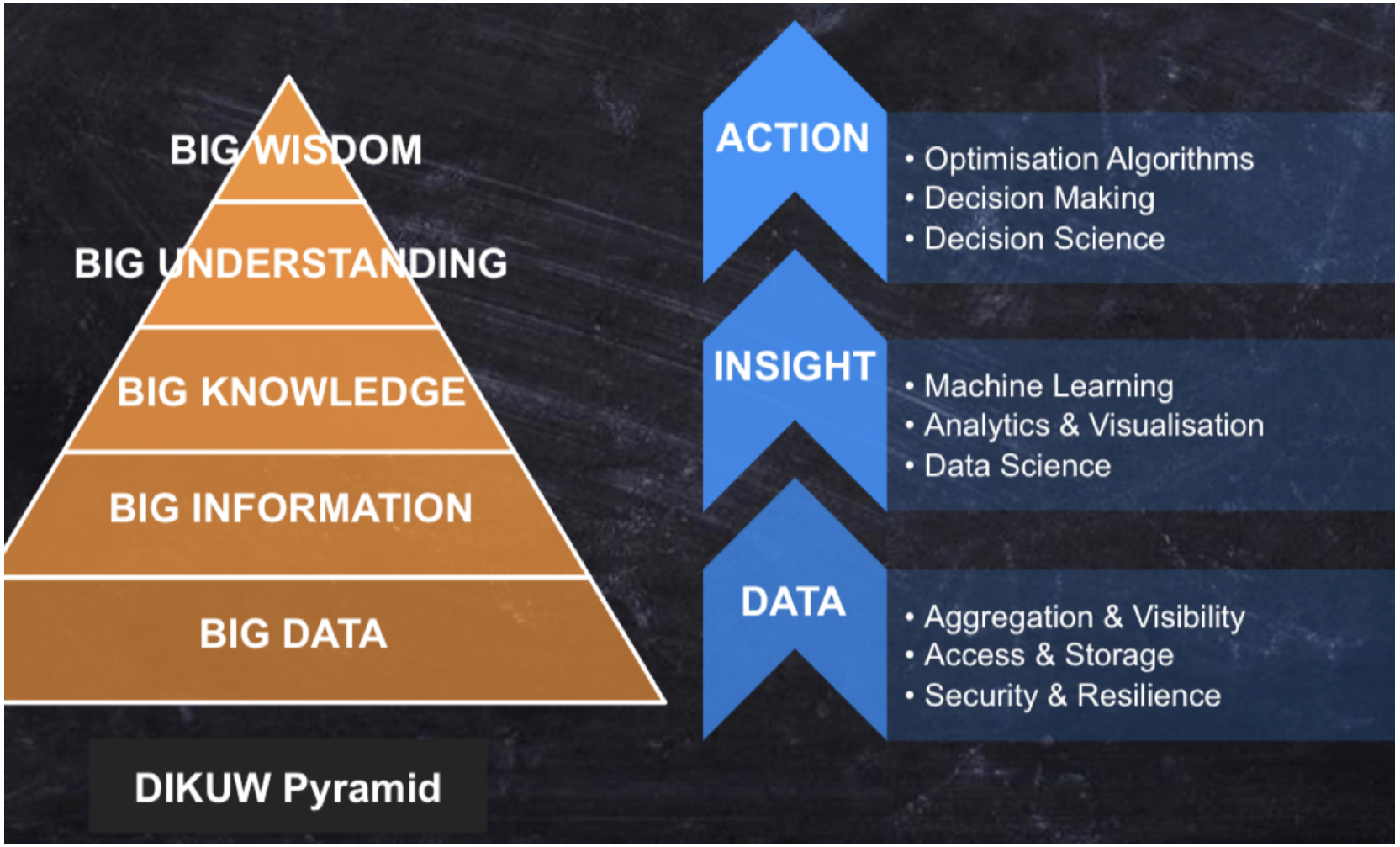
LOOKING BACKWARD AND FORWARD

FIRST THERE WAS BUSINESS INTELLIGENCE	NOW WE'VE ADDED DATA SCIENCE
Deductive Reasoning	Inductive and Deductive Reasoning
Backward Looking	Forward Looking
Slice and Dice Data	Interact with Data
Warehoused and Siloed Data	Distributed, Real Time Data
Analyze the Past, Guess the Future	Predict and Advise
Creates Reports	Creates Data Products
Analytic Output	Answer Questions and Create New Ones
	Actionable Answer

Inductive and deductive reasoning

- Data Science supports and encourages shifting between deductive (hypothesis-based) and inductive (pattern-based) reasoning
- This is a fundamental change from traditional analysis approaches.
- Inductive reasoning and exploratory data analysis provide a means to form or refine hypotheses and discover new analytic paths.
- Models of reality no longer need to be static.
- They are constantly tested, updated and improved until better models are found.

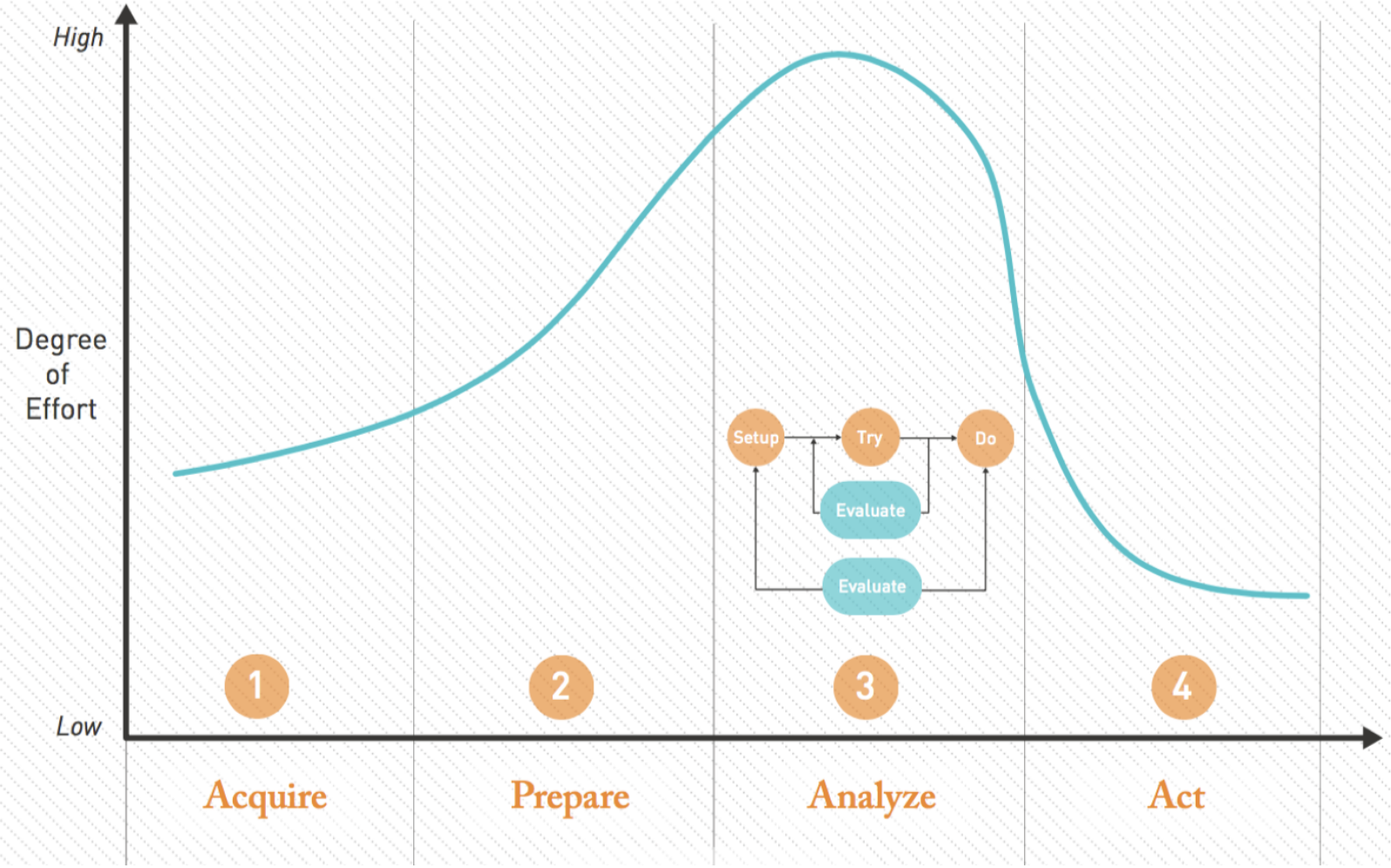
From data to wisdom



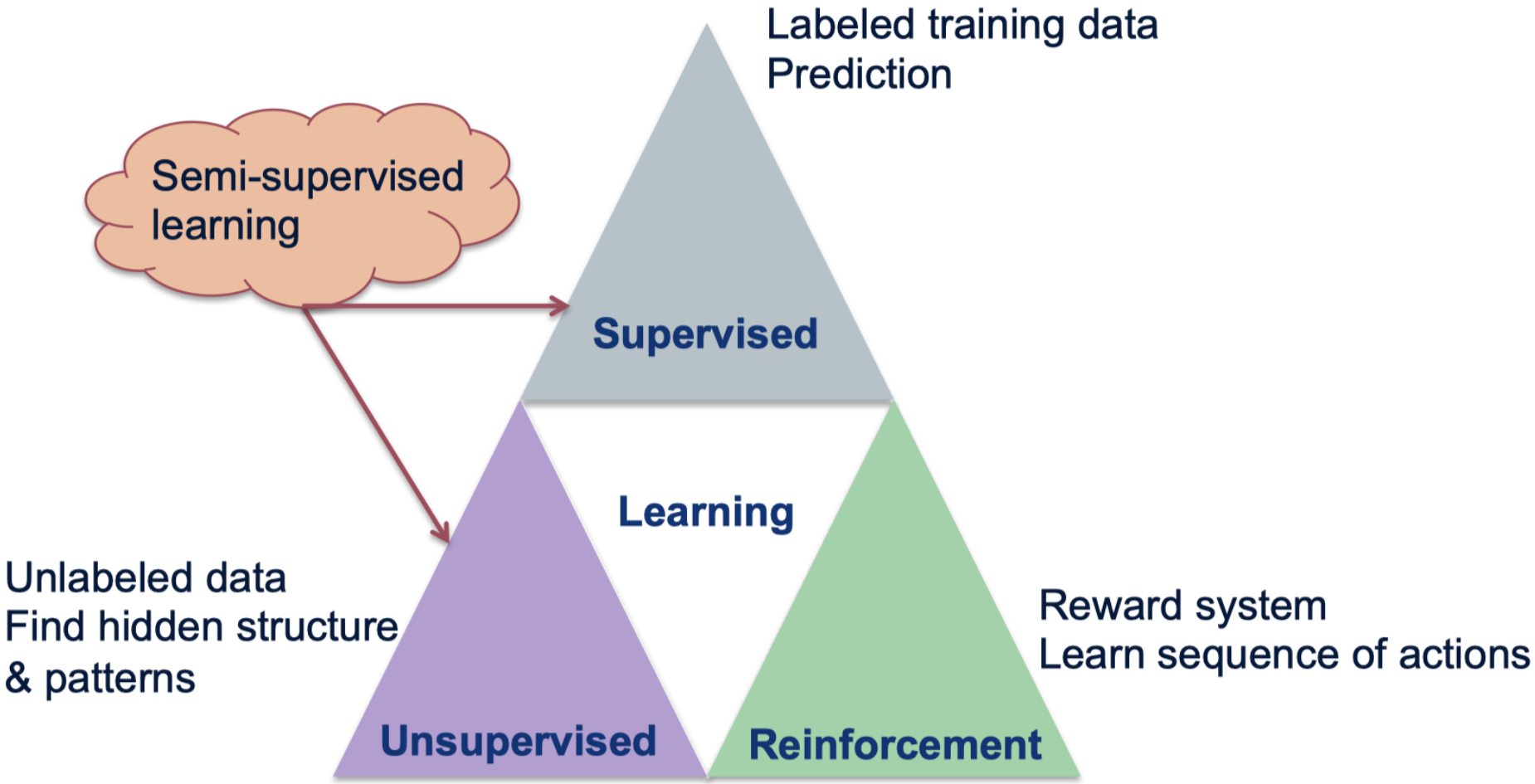
Data Science principles

- Be willing to fail.
- Fail often and learn quickly.
- Keep the goal in mind.
- Dedication and focus lead to success.

Data science workflow



Types of Learning



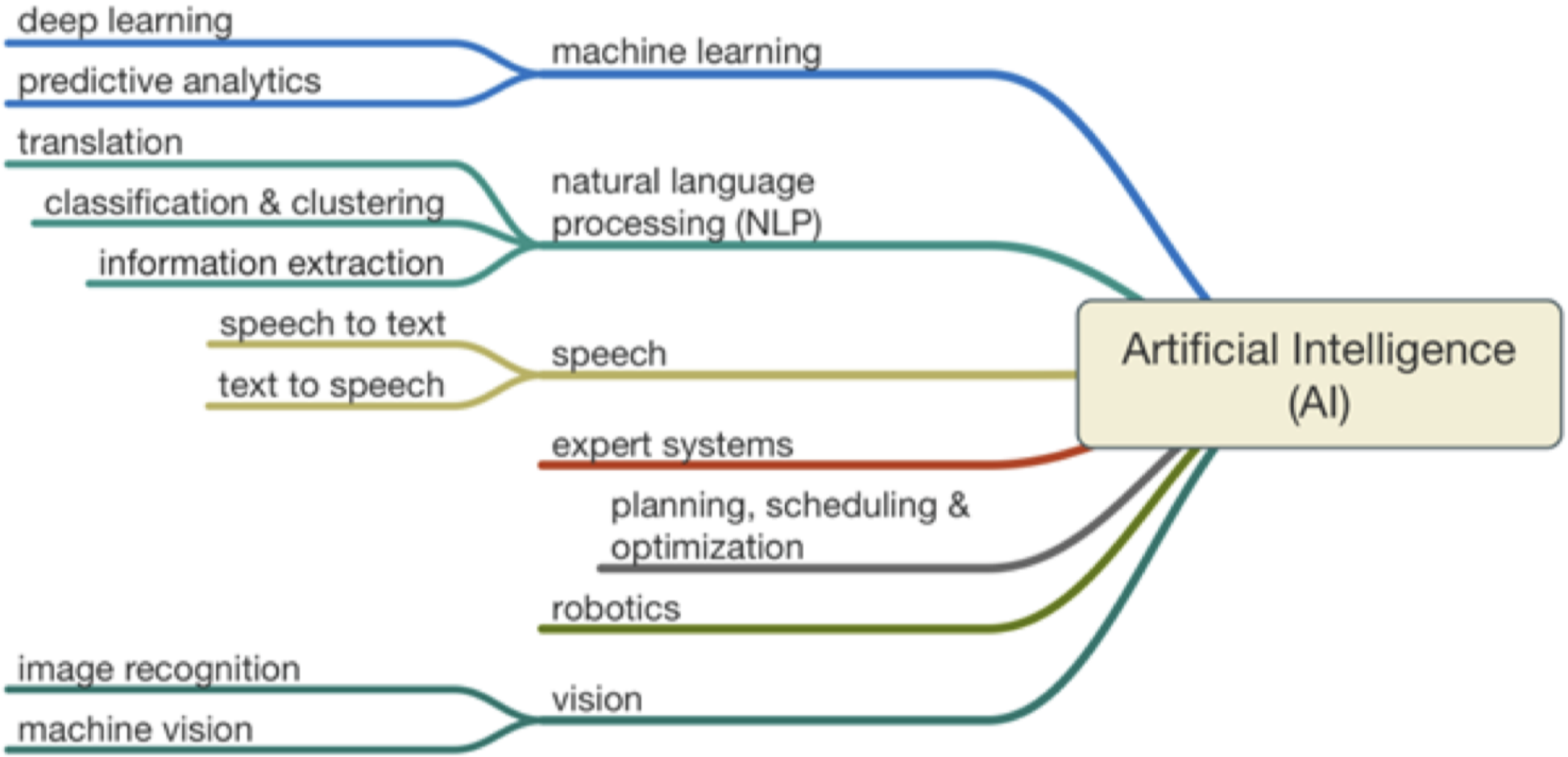
Why is Learning Important

- Impractical/impossible to specify systems correctly and completely at the time of design/implementation
- Implemented systems may not work as well as desired or expected when put in operation
- Knowledge about certain tasks may simply be too large to be explicitly encoded by humans
- The environment may change and hence the system’s goals need to be changed as well
- Hidden relationships and correlations among huge amounts of data

Why has ML become popular?

- Data explosion – Big Data!
  - Structured, unstructured, social media, labelled, unlabelled
  - Cost effective storage
- Computational power
- Faster processors, GPUs
- HPC, cloud computing, computing as a service
- Advances in algorithms and availability of toolkits

Main approaches in AI

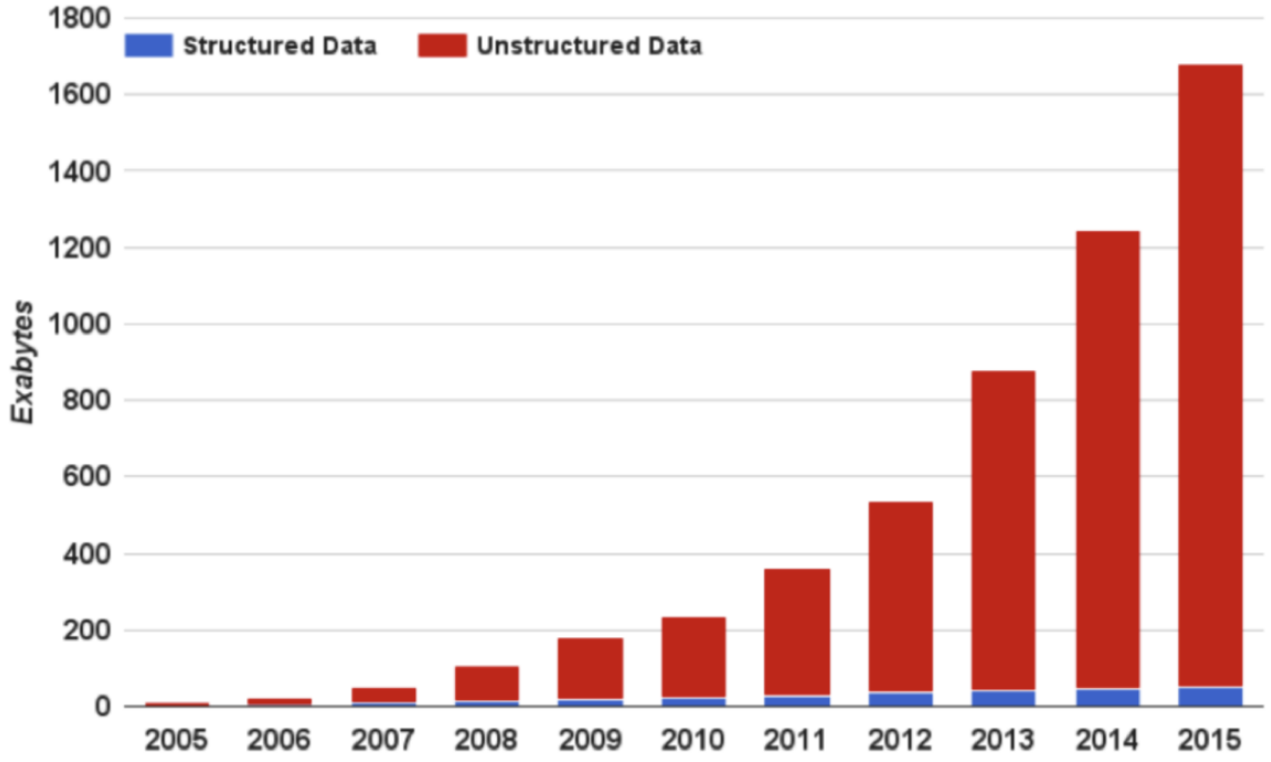


Data Science in the Wild

Personalisation

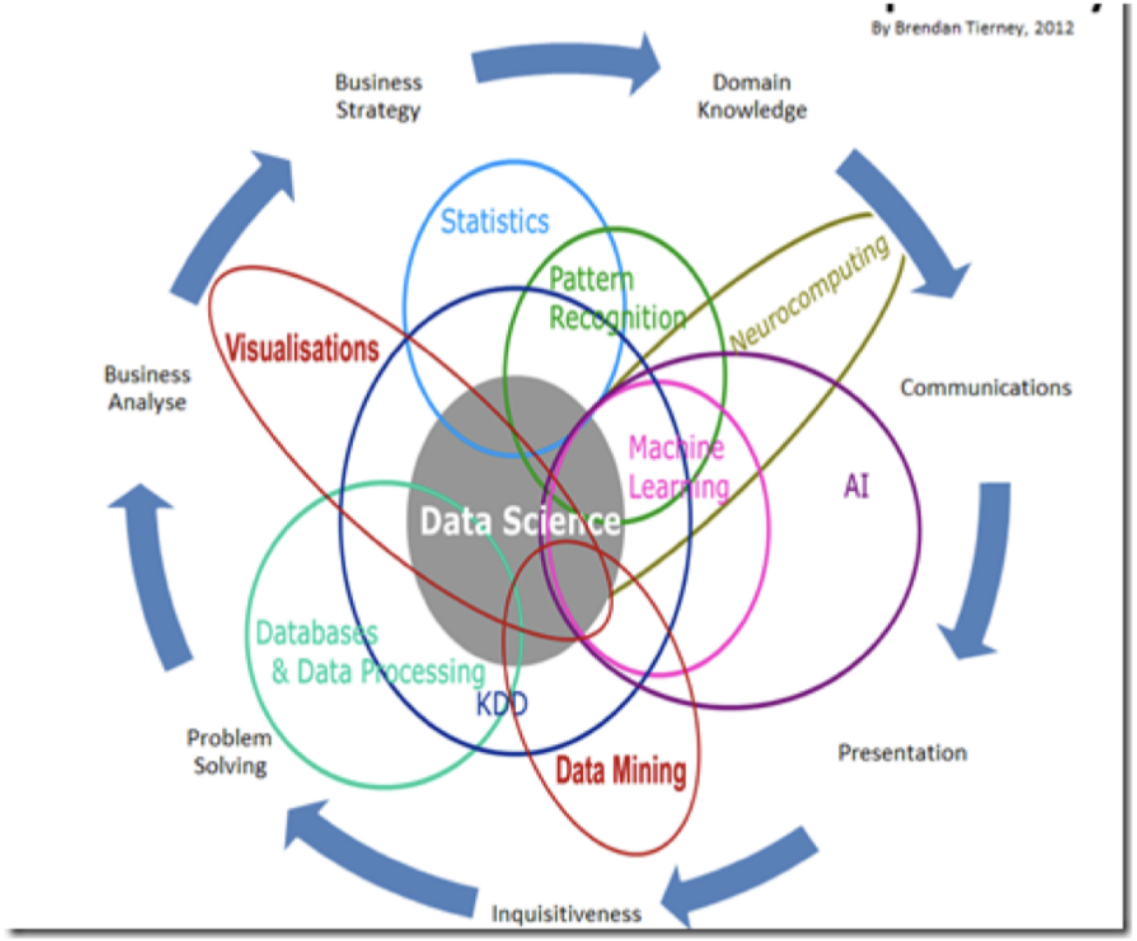
- What articles should be shown on the homepage of an online newspaper?
- What titles and images would attract the most clicks?
- Which product order would yield the highest profit?
- What is the best combination of drugs for patient?

Unstructured Data



A.Nadkarni, N.Yezhkova, “Structured versus unstructured data: The balance of power continues to shift.” IDC (Industry Development and Models), March 2014.

Ideal Data Scientist



What we will do

1. Read text documents as data.
2. Unsupervised learning techniques for interpreting corpora.
3. Supervised learning for regression and classification.
4. Word/document embedding for identifying dimensions of language.
5. Discourse analytics – summarization, question answering.

Big Data, Big Analytics

- Massive increase in availability of unstructured text datasets:
  - new social structures (the internet, email)
  - digitization efforts (govt documents, Google)
- Parallel increase in computational resources:
  - cheap disk space
  - efficient database solutions
  - compute: CPUs → GPUs → TPUs
- Parallel development of tools for natural language analysis
  - text by itself is not very useful
  - machine learning, natural language processing, causal inference

Corpora

- Text data is a sequence of characters called documents.
- The set of documents is the corpus, which we will call D.
- Text data is unstructured:
  - the information we want is mixed together with (lots of) information we don't.
- All text data approaches will throw away some information:
  - The trick is figuring out how to retain valuable information.
- The tools for `Tokenization` and `Dimension Reduction` are focused on this step:
  - transforming an unstructured corpus D to a usable matrix X.

Relating documents to metadata

- This course is on applied NLP:
  - the documents are not that meaningful by themselves.
  - we want to relate text data to metadata.
- e.g., measuring positive-negative sentiment Y in judicial opinions.
  - not that meaningful by itself.
- but how about sentiment Y_{ijt} in opinion i by govenor j at time t:
  - how does sentiment vary over time t?
  - does govenor from party p_j express more hawkish toward the monetary policy?

What counts as a document?

- The unit of analysis (the “document”) will vary depending on your question.
  - needs to be fine enough to fit the relevant metadata variation
  - should not be finer – would make dataset more high-dimensional without relevant empirical variation.
- What should we use as the document in these contexts?
  i. predicting whether a new agency is right-wing or left-wing in partisan ideology, from their written opinions.
  ii. predicting whether parliamentary speeches become more emotive in the run-up to an election
  iii. measuring whether newspapers use higher or lower sentiment toward different groups.

Handling Corpora

- There is already a vast amount of data out there that has already been compiled (e.g. CourtListener, Twitter, New York Times, Google Trends, Wikipedia).
- Everyone in this class should learn how to:
    i. query REST API's
    ii. run a web scraper in selenium
    iii. do pre-processing on corpora, e.g. to remove HTML markup, fix errors associated with OCR.
- I also recommend everyone to become familiar with hugginface datasets (https://huggingface.co/docs/datasets/)