26.02.19 14:52

Challenges and Opportunities with BIG Data

Jetic Gū

CMPT 843

#04-2019-1003-001_Präsentation_1

Overview

1. Introduction
2. Phases in the Processing Pipeline
   - Acquisition, Extraction, Integration, Analysis, Interpretation
3. Challenges in Big Data Analysis
   - Heterogeneity, Scale, Timeliness, Privacy, Human Collaboration
4. System Architecture
5. Conclusions

Coarse

1. "community white paper" "leading researchers across the US": not just for people from CS dept.
2. Big Data is everywhere
    1. Research, education, urban planning, transportation, environmental modelling, energy saving, smart materials, social science, business intel, defence, privacy
3. we know a lot about data
4. probably even seen it in action or are working with it
5. encourage investment and raise awareness: begging for money

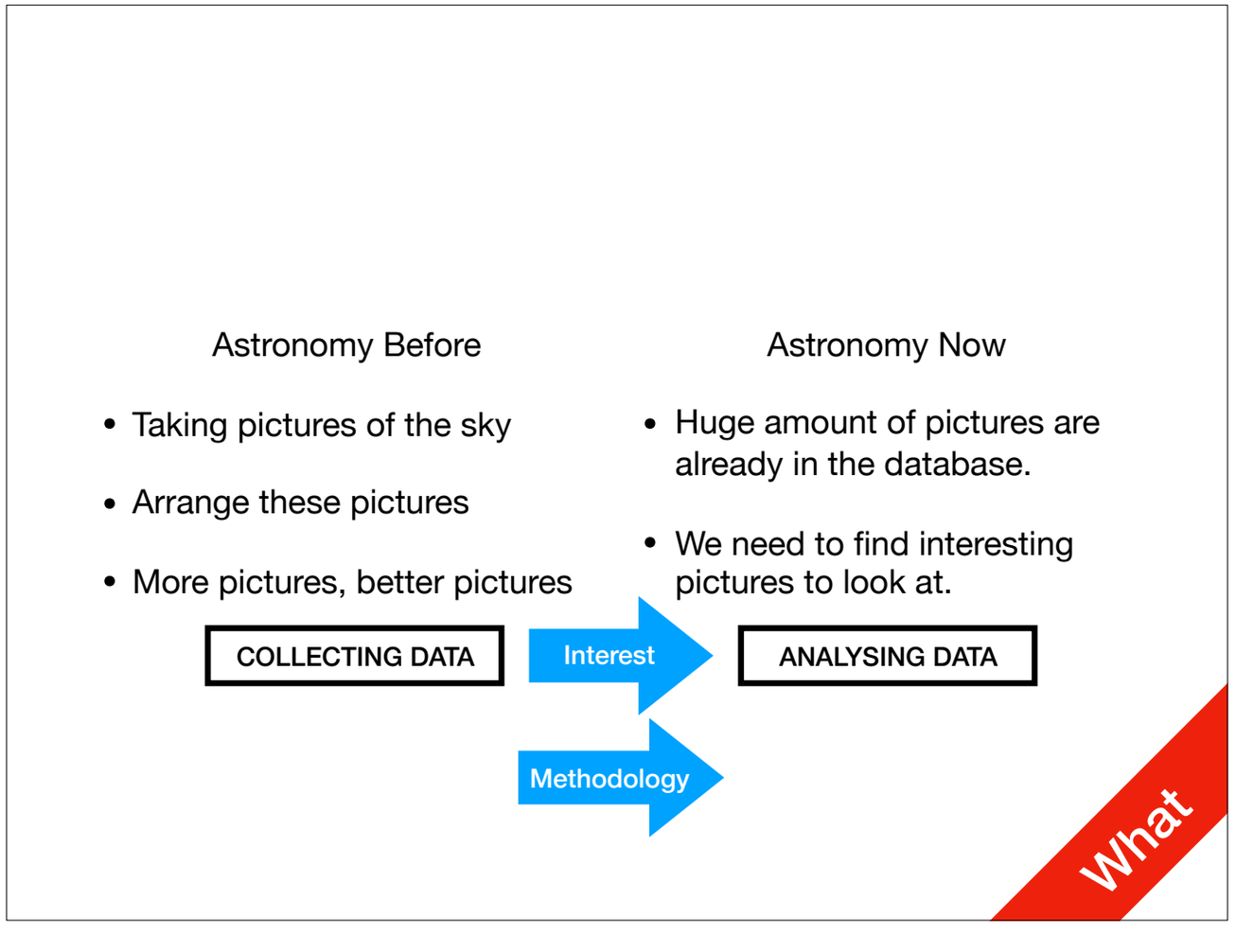
Overview

- Focus: BIG Data, What, How, and Why?
- Architecture: 5-Stage Pipeline
- Core Ideas:
    1. What has been in place
    2. Major difficulties
    3. Methodology shift

What is big data?
How to do big data?
Why big data

What is **BIG** Data?

It is (not) just a lot of data.



Methodology: how we view data, how we use data

The Major Difficulties Are What Defines Big Data

- Heterogeneity: as oppose to homogeneous
- Scale: as oppose to traditional DB
- Timeliness: processing speed, learning and inference
- Privacy: legal, trust issues, paranoia
- Human collaboration: understanding data

What

Heterogeneity

1. Natural language: powerful, expressful, but ambiguous
2. Noise, very noisy and coarse
3. Multimedia
4. Redundancy, missing information

Scale

1. Data volume is expanding faster than computing resources
2. Serialise computation -> parallel
3. Cluster managements
4. IO device revolution: changes in storage system affects general architecture.

Timeliness

Speed, training, inference, retraining, data gathering

Privacy

How do we protect data privacy? Easier to get data in China than Canada

Location data zB

Human collaboration

We don’t want machines to make all our decisions

crowd-sourcing?

What is Big Data

- Data is resource

Before

- The data is fine
- Know the data
- Know the question
- Use the data

Now

- The data is coarse
- Don’t know the data
- Don’t know the question
- Understand the data

What

1. Fine data: structured, formatted neatly
   - Noisy mostly useless
   - More formats that one person can know
   - Unstructured, missing elements
   - Multimedia
2. We have presumptions about the data, through humanly going over them and discover patterns
   - The patterns are learned automatically now
3. We are using the data to solve a very specific question, like given a geographic DB, we want to know what is the third highest mountain
   - We have a goal, like given a companies’ sales records and internal documents, we want to maximise the profit
   - We have the data about all professional Hockey players in the NHL, what do we do of it?
   - We have all twitter data, now what?
4. We use the data to achieve specific goals like in (3)
   - Knowledge acquisition: we want to be able to understand the hidden patterns and depending on which make decisions, or just knowing.

How to do **BIG** Data?

5 Stages

- Acquisition
- Extraction
- Integration
- Analysis
- Interpretation

How

Acquisition

- Getting the original raw data
- Filtering out useless data
  - Cannot be used directly now
- Metadata: what is recorded, how is it recorded and measured

How

Filters:

1. filters need to be designed to not discard useful information
2. online and offline

Extraction & Integration

- Extraction
  - Convert data into a format that can be used to perform analysis, pull out the useful information from the raw structure
  - Driven by the goal you have in mind
- Integration
  - Data structure design, hardware storage design, etc.
  - Eliminate basic errors

How

We haven’t started doing serious analysis yet.

Modelling & Interpretation

- Modelling: achieving the goal; Interpretation: understanding the results from modelling

Before	Now
• Model a priori	• Models need definition
• Rules/statistics a priori	• Discover patterns
• Interpretation a priori	• Discover interpretation
• Data(structure) is static	• Data(structure) is growing

How

Interactive
Life-time learning
Interpretation: how much do we know about the result generation phase?

Advantages

- Huge amount of noisy data > general statistics from tiny data
  - Noise control: data usually contain errors, limited noise can actually help the model generalise better
- Distributed representations of data: everything is stored as vectors -> easier comparison and general computation
- Realtime interactive analysis: instead of pre-training all models, information could be gathered on-the-go
  - recommendation systems

How

Why: Why is Big Data still so much trouble?

1. How we use data
2. Analytical models are not very good
3. Actually, we don’t have enough data
4. Actually, we don’t even have data

What now?

1. Future by Goldberg
2. Research trends, the bottleneck is real for good reason
3. Why big data?
    1. rethink the 5-stage pipeline, is it good enough?
    2. the gap between people who use data and people who produce data
    3. why it matters more and more

How we utilise data (Past)

- Rule-based systems: we know what's going on
- Statistics methods: we have some ideas on what's going on
- Neural methods: it is magic, we are all muggles :-(

Why

Why

1. How we use data
2. Analytical models are not very good
3. Actually, we don't have enough data
4. Actually, we don't even have data

What now?

1. Future by Goldberg
2. Research trends, the bottleneck is real for good reason
3. Why big data?
    1. rethink the 5-stage pipeline, is it good enough?
    2. the gap between people who use data and people who produce data
    3. why it matters more and more

Problems?

- "Cloud computing is the future"
- "Blockchain is the future"
- "Deep learning is the future"
- Noah A. Smith: "machine translation: solved"
- "One model (DL) to solve them all?"

Why

Why

1. How we use data
2. Analytical models are not very good
3. Actually, we don’t have enough data
4. Actually, we don’t even have data

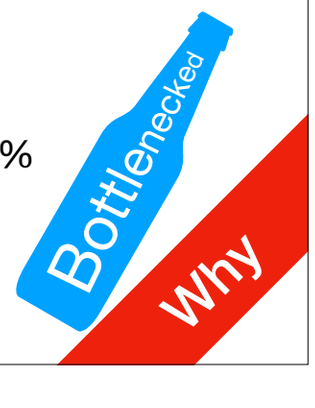
What now?

1. Future by Goldberg
2. Research trends, the bottleneck is real for good reason
3. Why big data?
    1. rethink the 5-stage pipeline, is it good enough?
    2. the gap between people who use data and people who produce data
    3. why it matters more and more

Problems?
Analytical models aren't good enough

Take machine translation in NLP: Not just from French to English, could be from German to Python, English to SQL, etc.

- Early 20th century, expert systems: doesn't work
- 1949 - 2014 statistical systems: remember Google translate back in the days?
- Now: 2014 onwards Neural Machine Translation
  - "Solved?" SoTA: ~40%; Google Translate: <22%
  - Inference issues! SlooooooooNULL

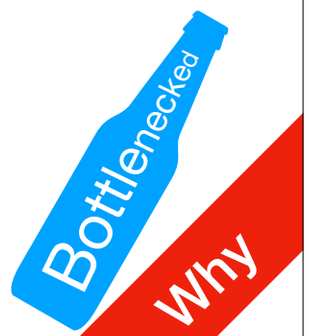


Why
1. How we use data
2. Analytical models are not very good
3. Actually, we don't have enough data
4. Actually, we don't even have data

What now?
1. Future by Goldberg
2. Research trends, the bottleneck is real for good reason
3. Why big data?
    1. rethink the 5-stage pipeline, is it good enough?
    2. the gap between people who use data and people who produce data
    3. why it matters more and more

Problems?
Analytical models aren't good enough

- Even better models?
  - Complex models provide better performance, but computational costs increase dramatically
  - Infrastructure support: complex distributed system architecture required
- NLP solved using deep learning:
  - English POS tagging (Noun, Verb, Adj, etc.)



Why

1. How we use data
2. Analytical models are not very good
3. Actually, we don't have enough data
4. Actually, we don't even have data

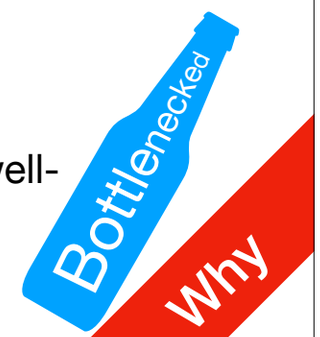
What now?

1. Future by Goldberg
2. Research trends, the bottleneck is real for good reason
3. Why big data?
    1. rethink the 5-stage pipeline, is it good enough?
    2. the gap between people who use data and people who produce data
    3. why it matters more and more

Problems?
Actually, we don't even have enough data

For the areas like biomedical, we are using models designed for Big Data, but we don't have nearly enough data

- Classification
  - LM: Hundreds of millions of training samples
  - Biomedical: thousands
- In average, the time it takes us to clean up a well-mined NLP dataset is usually weeks/months



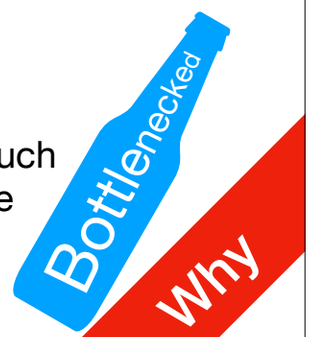
Why
1. How we use data
2. Analytical models are not very good
3. Actually, we don't have enough data
4. Actually, we don't even have data

What now?
1. Future by Goldberg
2. Research trends, the bottleneck is real for good reason
3. Why big data?
    1. rethink the 5-stage pipeline, is it good enough?
    2. the gap between people who use data and people who produce data
    3. why it matters more and more

Problems?
Actually, we don’t even have enough data

Huge amounts of data, but often unfeasible to fully leverage them

- Machine translation
  - Data available: FR-EN billions parallel sentences
  - Typical industry training data: <100 million
  - Academia: 200k—5 million
- Why? Cost of Big Data platform that can support such computation is way too costly, and unworthy for the academia. Costly Data = No Data



Why

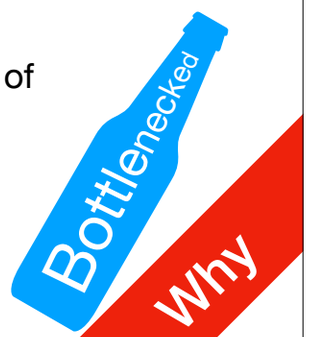
1. How we use data
2. Analytical models are not very good
3. Actually, we don’t have enough data
4. Actually, we don’t even have data

What now?

1. Future by Goldberg
2. Research trends, the bottleneck is real for good reason
3. Why big data?
    1. rethink the 5-stage pipeline, is it good enough?
    2. the gap between people who use data and people who produce data
    3. why it matters more and more

Problems?
Actually, sometimes we don't have data

- How much noise is acceptable?
  - <50% of samples contain errors in a typically mined dataset: this is unusable. Problematic data < No data
  - Typical precision system training data: <10% noise
- Lack of coordination between database systems
  - 'Today's analysts are impeded by a tedious process of exporting data from the database'
  - Mined dataset lack clear evaluation guidelines and metrics



Why

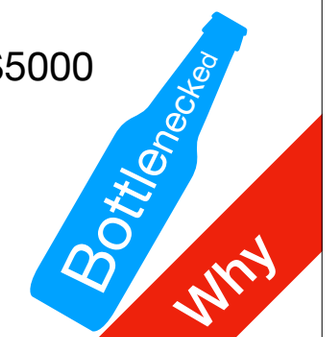
1. How we use data
2. Analytical models are not very good
3. Actually, we don't have enough data
4. Actually, we don't even have data

What now?

1. Future by Goldberg
2. Research trends, the bottleneck is real for good reason
3. Why big data?
    1. rethink the 5-stage pipeline, is it good enough?
    2. the gap between people who use data and people who produce data
    3. why it matters more and more

Problems?
Actually, sometimes we don't have data

- Low-resource languages? Try no-resource languages
- Linguistics data? Linguists have their own agendas for annotation!
- Proprietary data only: anybody wants to pay $5000 for 500 samples of training data?



Why

1. How we use data
2. Analytical models are not very good
3. Actually, we don't have enough data
4. Actually, we don't even have data

What now?

1. Future by Goldberg
2. Research trends, the bottleneck is real for good reason
3. Why big data?
    1. rethink the 5-stage pipeline, is it good enough?
    2. the gap between people who use data and people who produce data
    3. why it matters more and more

What now?

Why

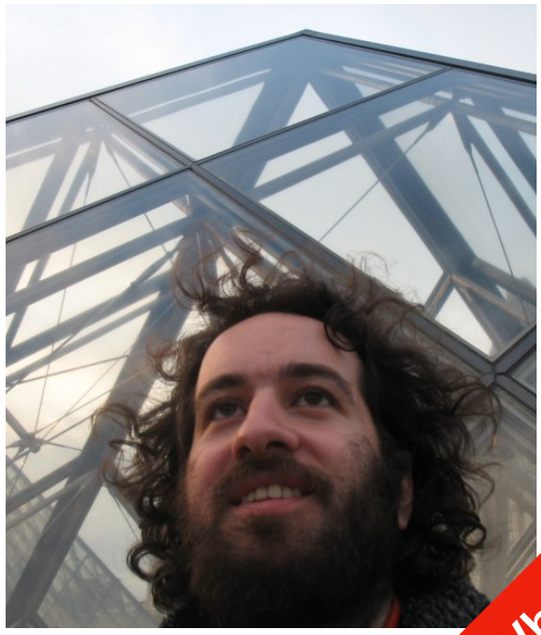
1. How we use data
2. Analytical models are not very good
3. Actually, we don’t have enough data
4. Actually, we don’t even have data

What now?

1. Future by Goldberg
2. Research trends, the bottleneck is real for good reason
3. Why big data?
    1. rethink the 5-stage pipeline, is it good enough?
    2. the gap between people who use data and people who produce data
    3. why it matters more and more

"We will take neural networks and blend in expert systems" (EMNLP2018)

- Yoav Goldberg
- Senior Lecturer at Bar Ilan University, working on NLP
- For non-experts: leading AI researcher
- He likes to say co stuff for attention
- People were very



Why

Why

1. How we use data
2. Analytical models are not very good
3. Actually, we don’t have enough data
4. Actually, we don’t even have data

What now?

1. Future by Goldberg
2. Research trends, the bottleneck is real for good reason
3. Why big data?
    1. rethink the 5-stage pipeline, is it good enough?
    2. the gap between people who use data and people who produce data
    3. why it matters more and more

Research Trends in Using Data

Industry	Academia
• Utilise 10TB of crowd-sourced/ mined data	• Utilise limited data (too poor to buy 10GPUs/AWS)
• Just stack layers of NN	• Better modelling
• ~99% PhDs optimising existing architectures	• ~5% PhDs working on new architectures
• NN is everything	• NN is not everything
• We need more ~~slaves~~ to annotate our data!	• We need more fine-grained clean data!
• We need more profit	• We need more knowledge
• 5-stage pipeline (kinda) works	• 5-stage pipeline enough?

Why

Why

1. How we use data
2. Analytical models are not very good
3. Actually, we don’t have enough data
4. Actually, we don’t even have data

What now?

1. Future by Goldberg
2. Research trends, the bottleneck is real for good reason
3. Why big data?
    1. rethink the 5-stage pipeline, is it good enough?
    2. the gap between people who use data and people who produce data
    3. why it matters more and more

What is Big Data

- Data is resource

Before

- The data is fine
- Know the data
- Know the question
- Use the data

Now

- The data is coarse
- Don’t know the data
- Don’t know the question
- Understand the data

What

What now?

1. Future by Goldberg
2. Research trends, the bottleneck is real for good reason
3. Why big data?
    1. rethink the 5-stage pipeline, is it good enough?
        1. The architecture is very very complicated and expensive! Even for big corporations! A lot of resources are wasted
    2. the gap between people who use data and people who produce data
    3. why it matters more and more

Why Big Data

- Data is resource

Solution		Now
• Beyond simple filters	⟷	The data is coarse
• Automated analysis	⟷	Don’t know the data
• Give hidden correlations	⟷	Don’t know the question
• Interpretable models	⟷	Understand the data

Why

1. We need better filtering strategies than simple filters: (NN?)
2. Automated analysis tools should be able to recognise basic data properties, give essential information regarding the dataset
3. Automated analysis, the analysis tool should establish basic correlations between different entries of data to provide causal insights
4. Not just data, the models should allow us to understand data.
    1. More difficult than just getting results

Why Big Data

- 5 stages:
  - acquisition, extraction, integration: preparing data
  - modelling, interpretation: performing analysis
- Future
  - Each stage itself needs a lot of work
  - Is this division good enough?

Why

What now?

1. Future by Goldberg
2. Research trends, the bottleneck is real for good reason
3. Why big data?
    1. rethink the 5-stage pipeline, is it good enough?
        1. The architecture is very very complicated and expensive! Even for big corporations! A lot of resources are wasted
    2. the gap between people who use data and people who produce data
    3. why it matters more and more

Why Big Data

- How often do researchers mine their own data?
  - Very very rare. The communities are quite different.
- Why should they work together more?
  - Some tasks require more, like lifelong learning (the future)
  - Some tasks make it ESSENTIAL

Why

What now?

1. Future by Goldberg
2. Research trends, the bottleneck is real for good reason
3. Why big data?
    1. rethink the 5-stage pipeline, is it good enough?
        1. The architecture is very very complicated and expensive! Even for big corporations! A lot of resources are wasted
    2. the gap between people who use data and people who produce data
    3. why it matters more and more

DB people should know how to do complex analysis!

- Database -> Knowledge-base: how do we represent general knowledge?
  - SoTA: tuples, **(Beijing, isCapitalOf, China)**
  - How do we store them?
- Incompleteness: there are always knowledge not in the KB
- Common-sense modelling: there are always knowledge not located anyway but in our intuition

Why

What now?

1. Future by Goldberg
2. Research trends, the bottleneck is real for good reason
3. Why big data?
    1. rethink the 5-stage pipeline, is it good enough?
        1. The architecture is very very complicated and expensive! Even for big corporations! A lot of resources are wasted
    2. the gap between people who use data and people who produce data
    3. why it matters more and more

There are still a lot to do with Big Data,

thank you.