Data Science Ecosystem

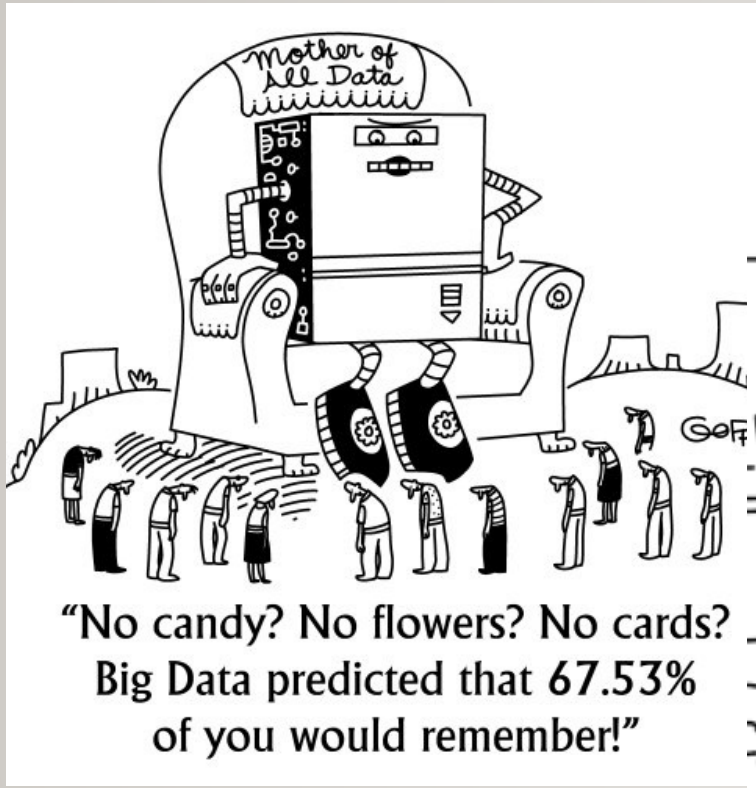
M. Tamer Özsu
U. Waterloo

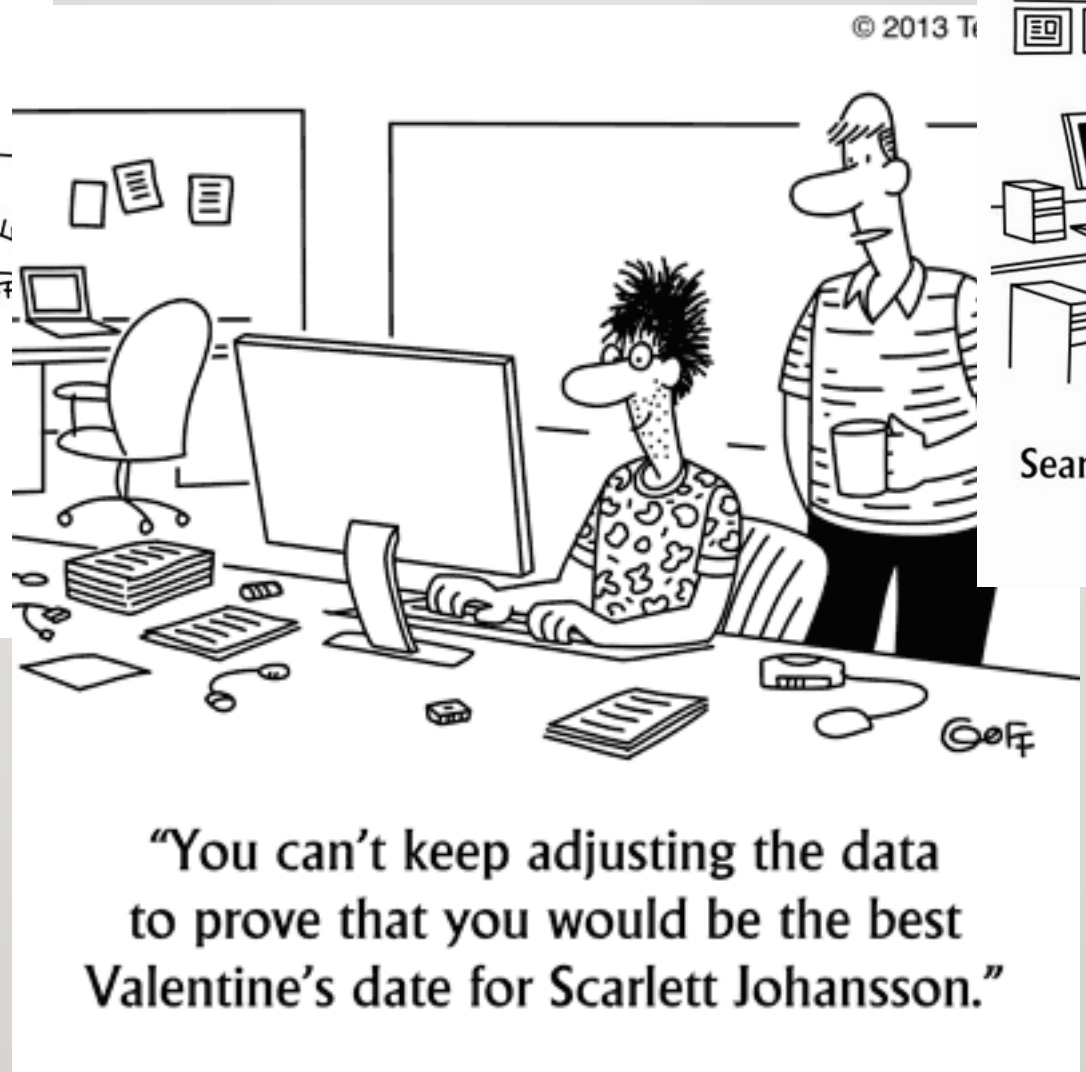
Nancy Reid
U. Toronto

Raymond Ng
UBC

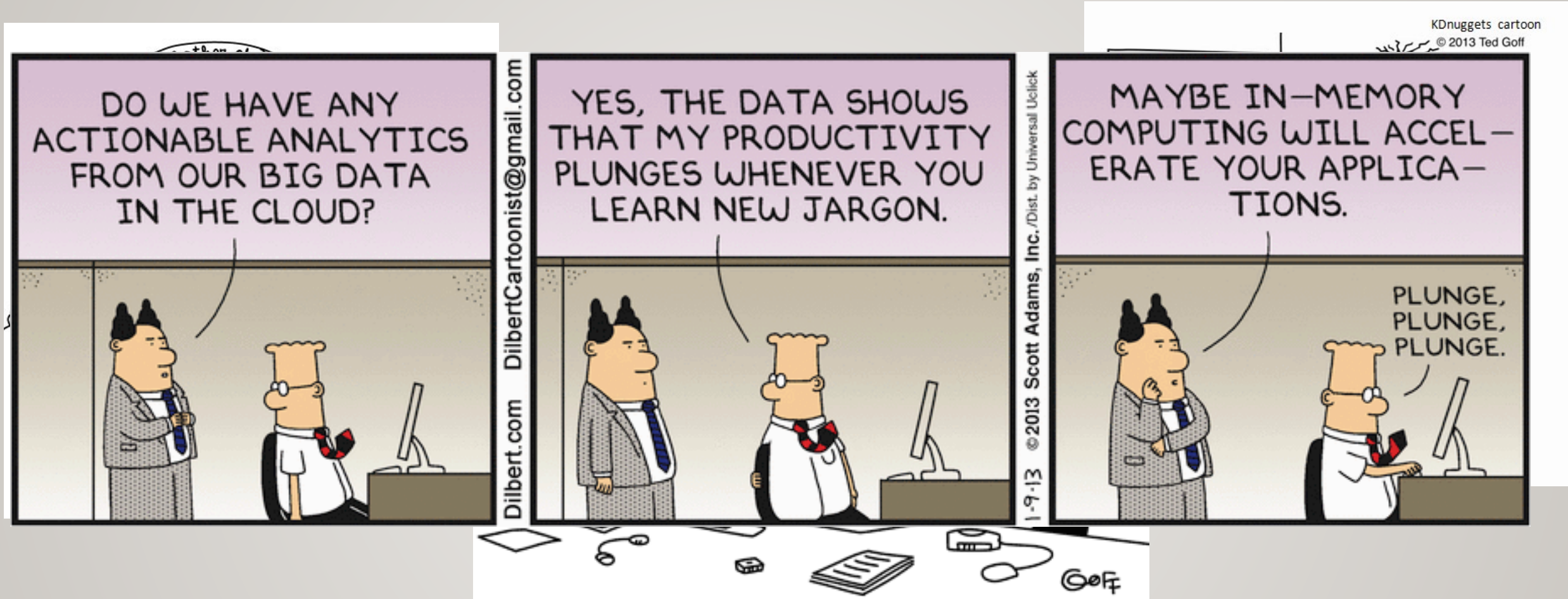
Data Science/Big Data in the News…

DATA SCIENCE EVERYWHERE!...



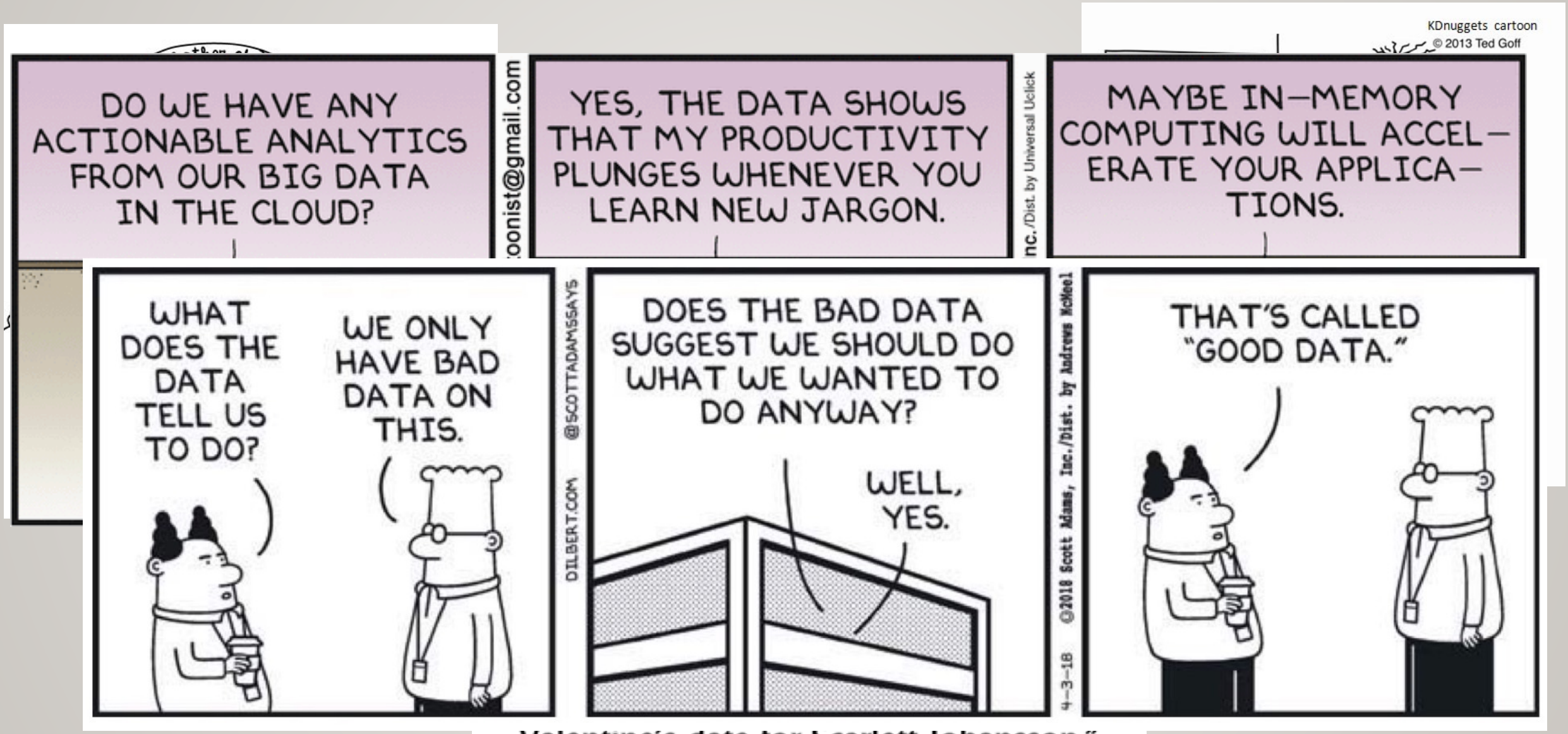


Data Science Everywhere!...



"You can't keep adjusting the data to prove that you would be the best Valentine's date for Scarlett Johansson."

DATA SCIENCE EVERYWHERE!...



Data Science Vocabulary

What is Data Science?

What is Data Science?

- "**Data science**, also known as **data-driven science**, is an interdisciplinary field of scientific methods, processes, algorithms and systems to extract knowledge or insights from data in various forms, either structured or unstructured, similar to data mining."

WIKIPEDIA
The Free Encyclopedia

What is Data Science?

- "**Data science**, also known as **data-driven science**, is an interdisciplinary field of scientific methods, processes, algorithms and systems to extract knowledge or insights from data in various forms, either structured or unstructured, similar to data mining."
- "Data science intends to analyze and understand actual phenomena with 'data'. In other words, the aim of data science is to reveal the features or the hidden structure of complicated natural, human, and social phenomena with data from a different point of view from the established or traditional theory and method."

WIKIPEDIA
The Free Encyclopedia


What is Data Science ?*
Fundamental Concepts and a Heuristic Example

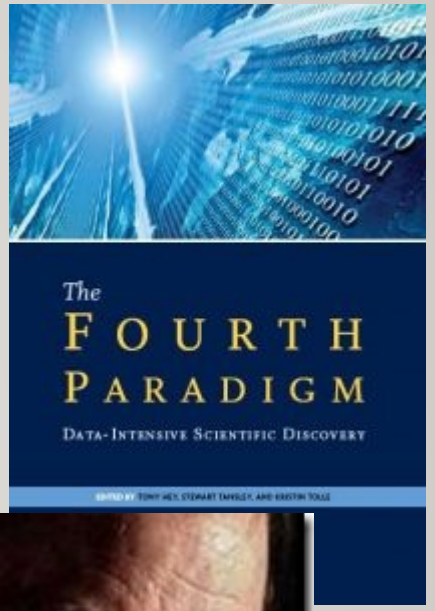
Chikio Hayashi
The Institute of Statistical Mathematics
Sakurugaoka, Bunjyan 304
15-8 Sakurugaoka, Shibuya-ku
Tokyo 150, Japan

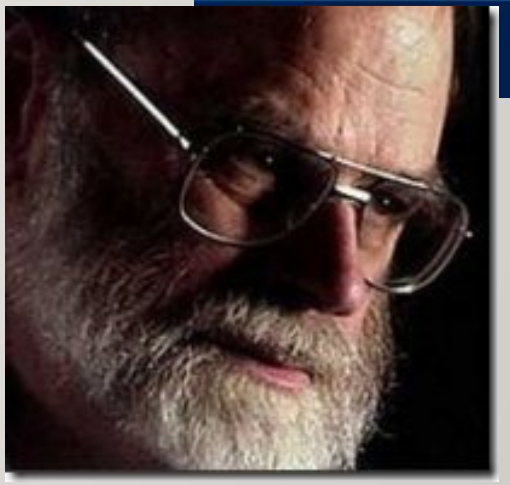
Summary: Data Science is not only a synthetic concept to unify statistics, data analysis and their related methods but also comprises its results. It includes three phases, design for data, collection of data, and analysis on data. Fundamental concepts and various methods based on it are discussed with a heuristic example.

1. Introduction

What is Data Science?

- Fourth paradigm
  - "… change of all sciences moving from observational, to theoretical, to computational and now to the 4th Paradigm – Data-Intensive Scientific Discovery"

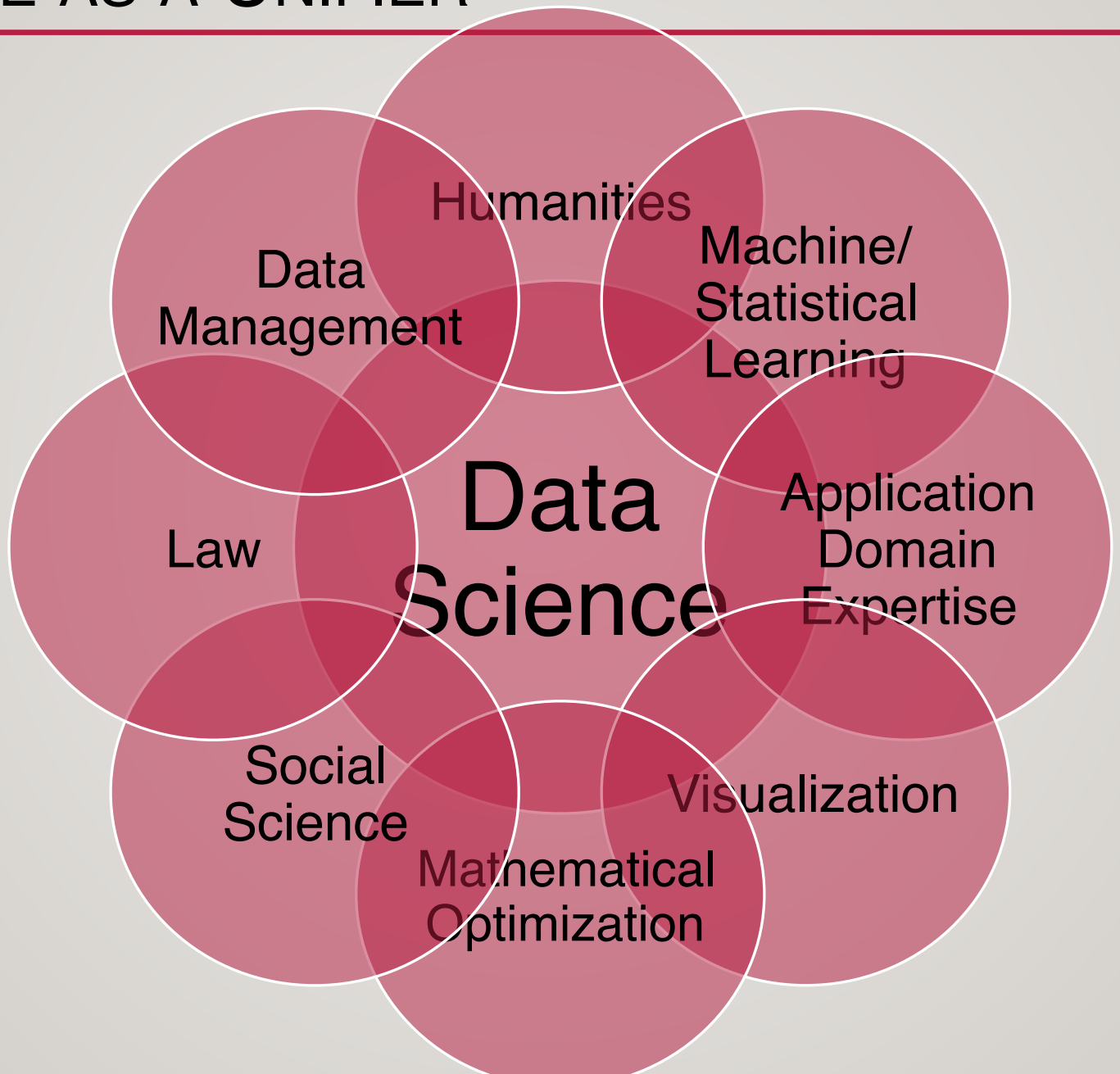




What is Important?

Need to solve a real problem using data…
No applications, no data science.

Data Science as a Unifier

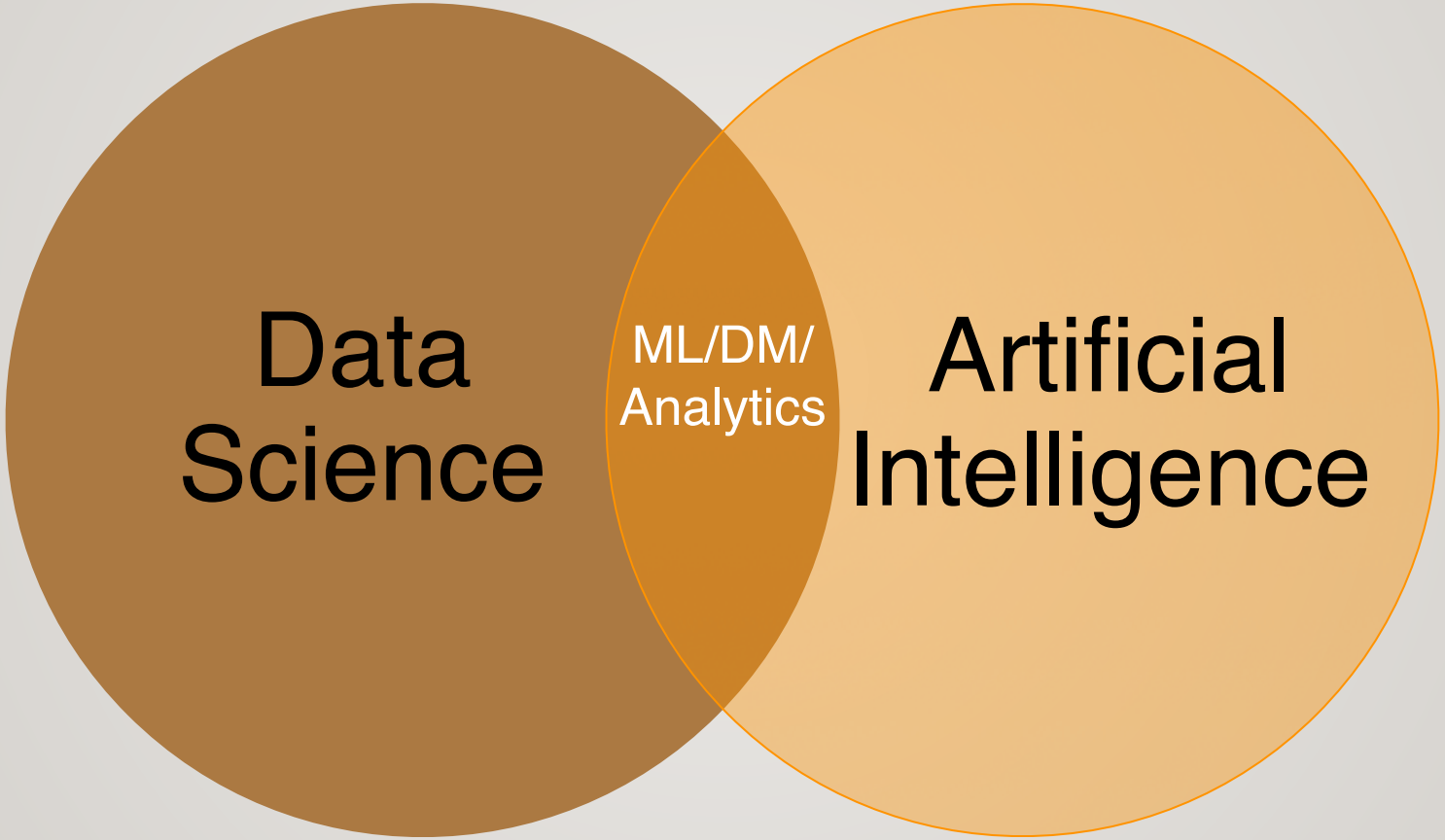


Data Science and Big Data

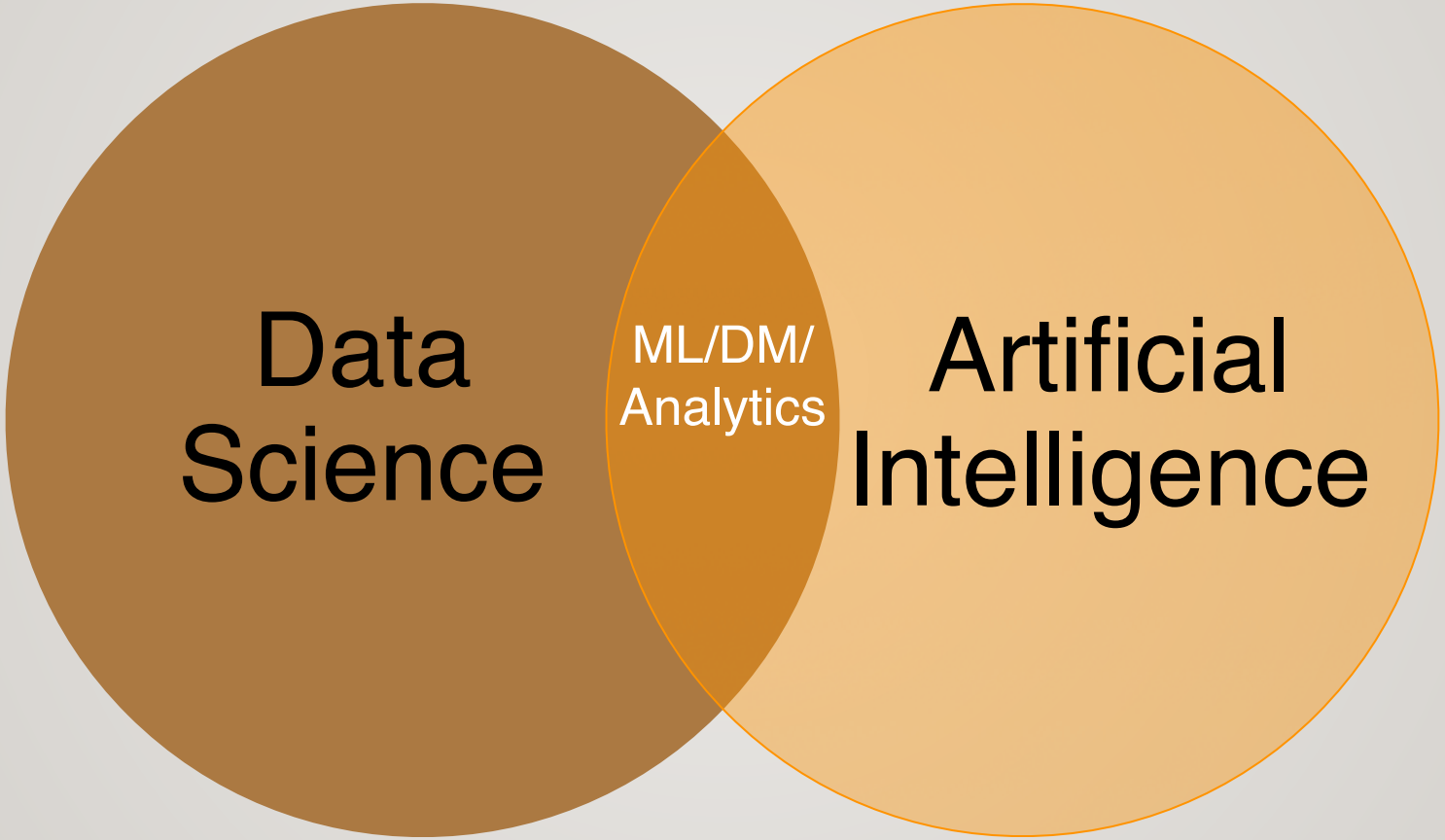
- They are not the “same thing”
- Big data = crude oil
  - Big data is about extracting “crude oil”, transporting it in “mega tankers”, siphoning it through “pipelines”, and storing it in “massive silos”
- Data science is about refining the “crude oil”

Carlos Samohano
Founder, Data Science London

Data Science and Artificial Intelligence



Data Science and Artificial Intelligence



"**Data science** produces **insights**.
Machine learning produces **predictions**"

Data Science Application Examples

- Fraud detection
  - Investigate fraud patterns in past data
  - Early detection is important
    - Before damage propagates
    - Harder than late detection
  - Precision is important
    - False positive and false negative are both bad
  - Real-time analytics

Data Science Application Examples

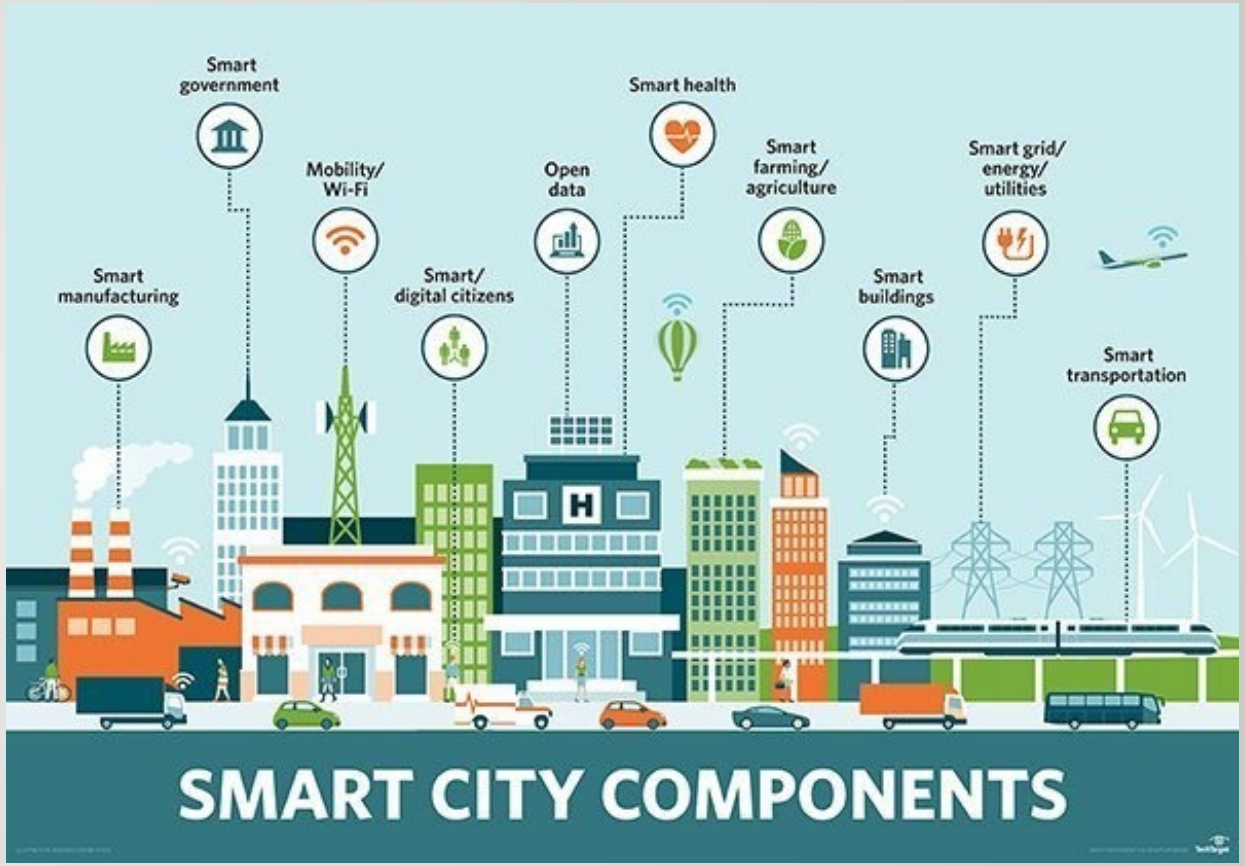
- Recommender systems
  - The ability to offer unique personalized service
  - Increase sales, click-through rates, conversions, …
    - Netflix recommender system valued at $1B per year
    - Amazon recommender system drives a 20-35% lift in sales annually
  - Collaborative filtering at scale

Data Science Application Examples

- Predicting why patients are being readmitted
  - Reduce costs
  - Improve population health
  - Find the “why” behind specific populations being readmitted
  - Data lakes of multiple data sources
  - Investigate ties between readmission an socioeconomic data points, patient history, genetics, …

Data Science Application Examples

- "Smart cities"
  - Not well-defined



Data Science Application Examples

- "Smart cities"
  - Not well-defined

lillisphotography / Getty / Emily Jan / The Atlantic

Stop Saying 'Smart Cities'

Digital stardust won't magically make future cities more affordable or resilient.

BRUCE STERLING | FEB 12, 2018 | TECHNOLOGY

Data Science Application Examples

- "Smart cities"
  - Not well-defined
  - Generally refers to using data ICT to
    - Better plan communities
    - Better manage assets
    - Reduce costs
  - Deploy open data to better en with community

lillisphotography / Getty / Emily Jan / The Atlantic

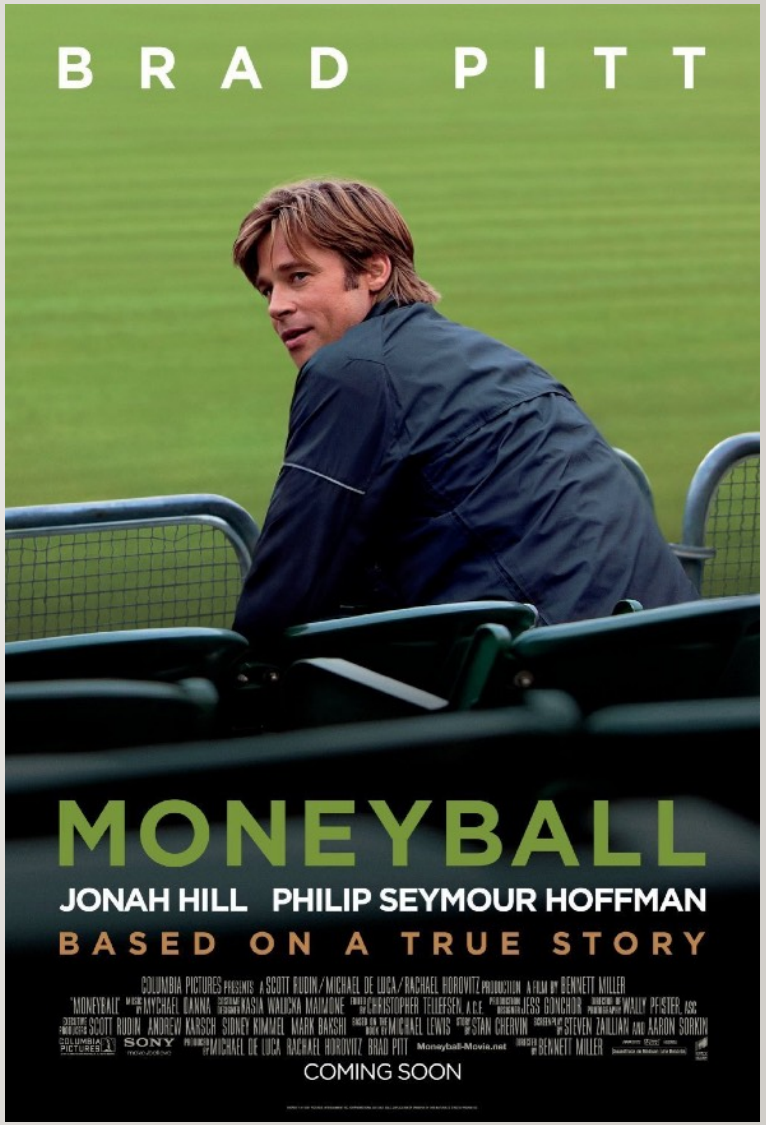
Stop Saying 'Smart Cities'

Digital stardust won't magically make future cities more affordable or resilient.

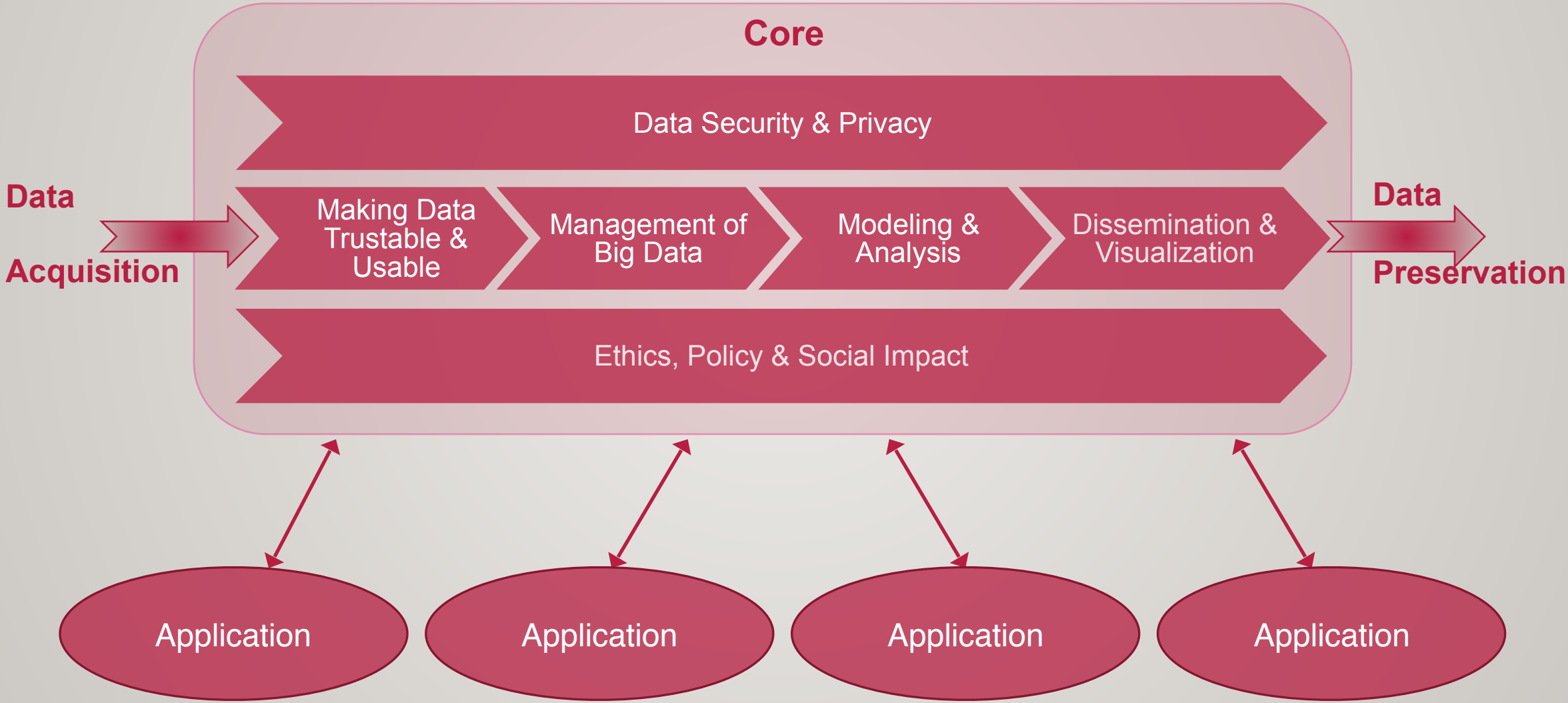
BRUCE STERLING | FEB 12, 2018 | TECHNOLOGY

Data Science Application Examples

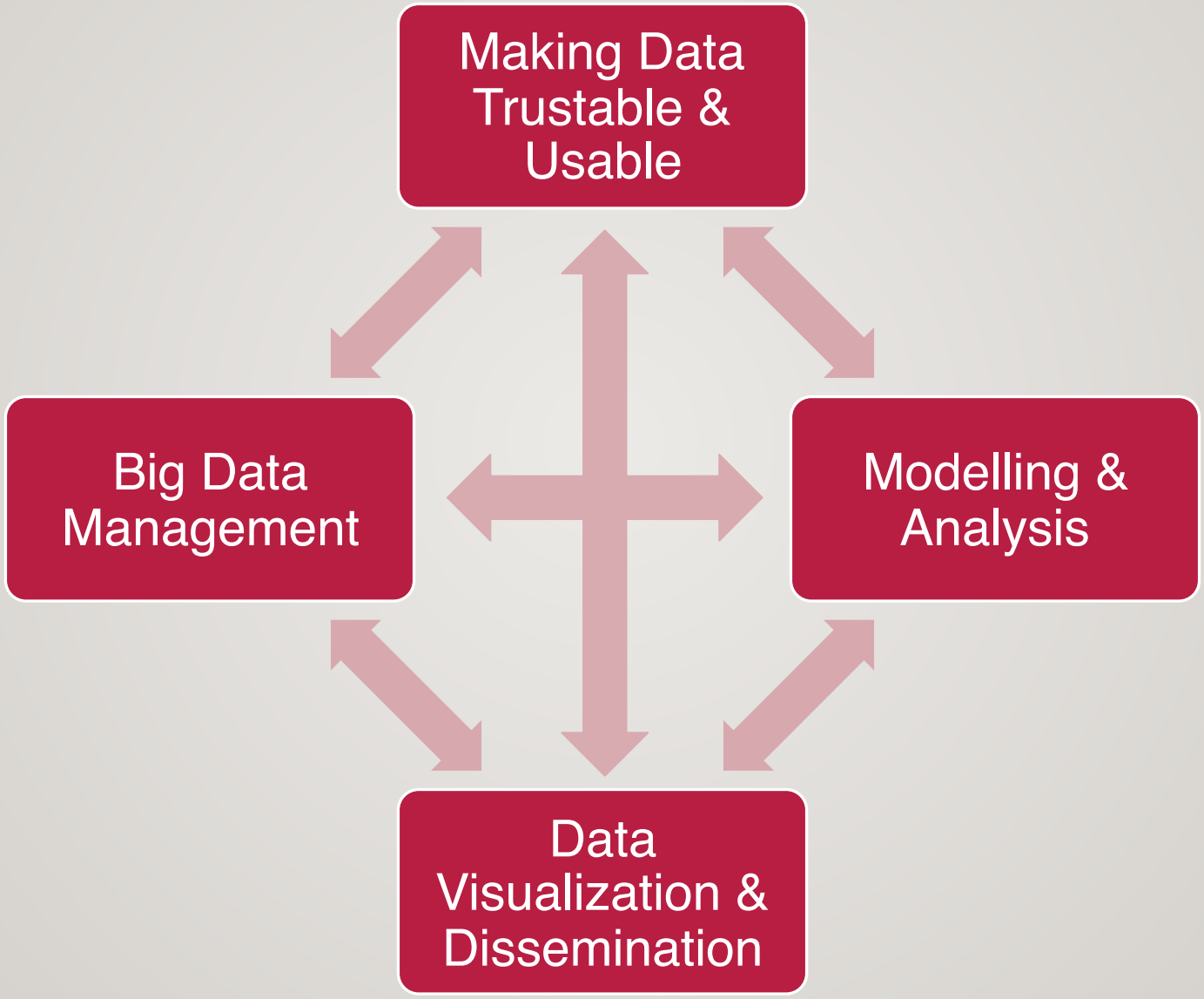
- Moneyball
  - How to build a baseball team on a very low budget by relying on data
  - *Sabermetrics*: the statistical analysis of baseball data to objectively evaluate performance
  - 2002 record of 103-59 was joint best in MLB
    - Team salary budget: $40 million
  - Other team: Yankees
    - Team salary budget: $120 million



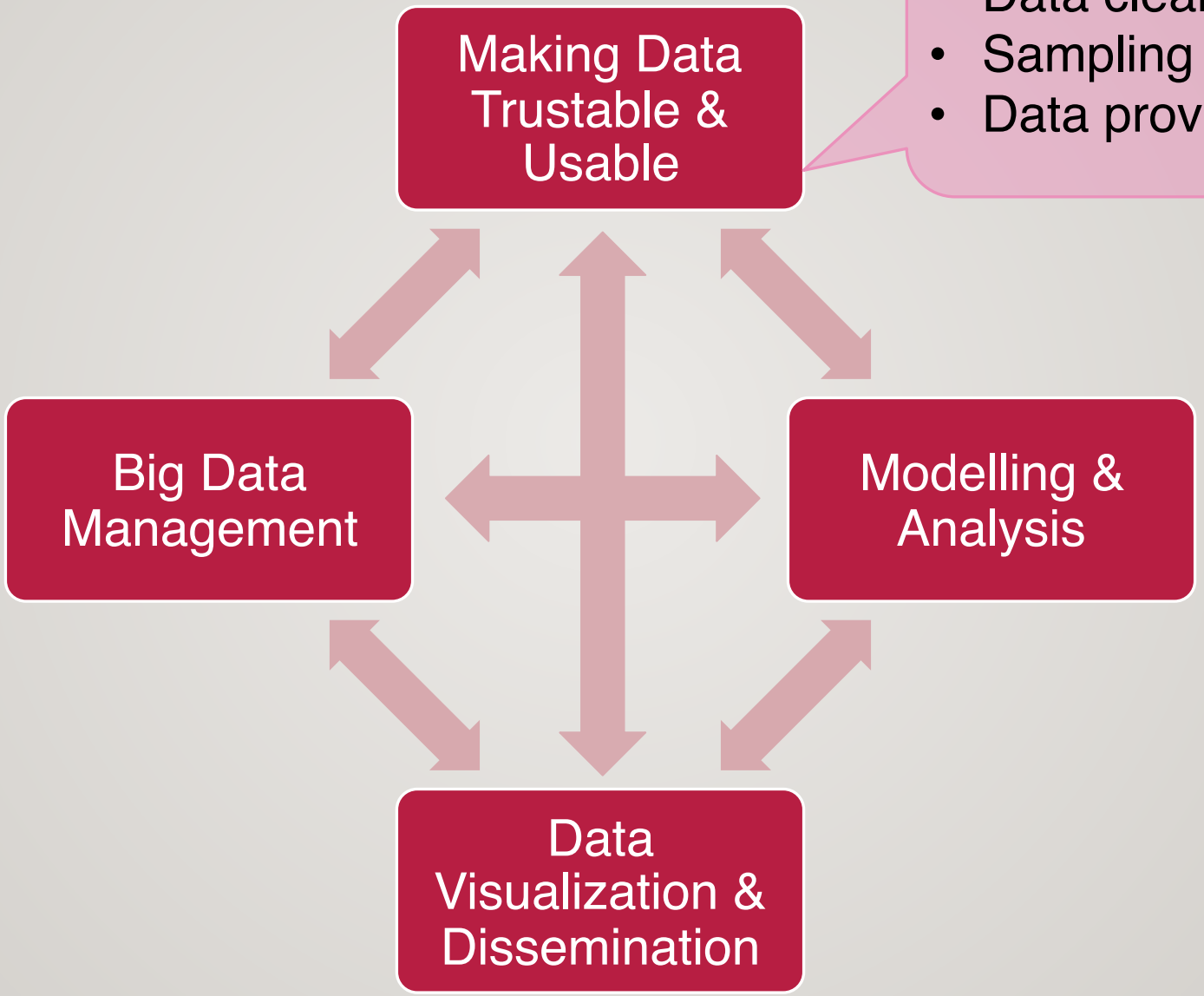
Holistic Approach to Data Science



Core Research Issues & Interactions



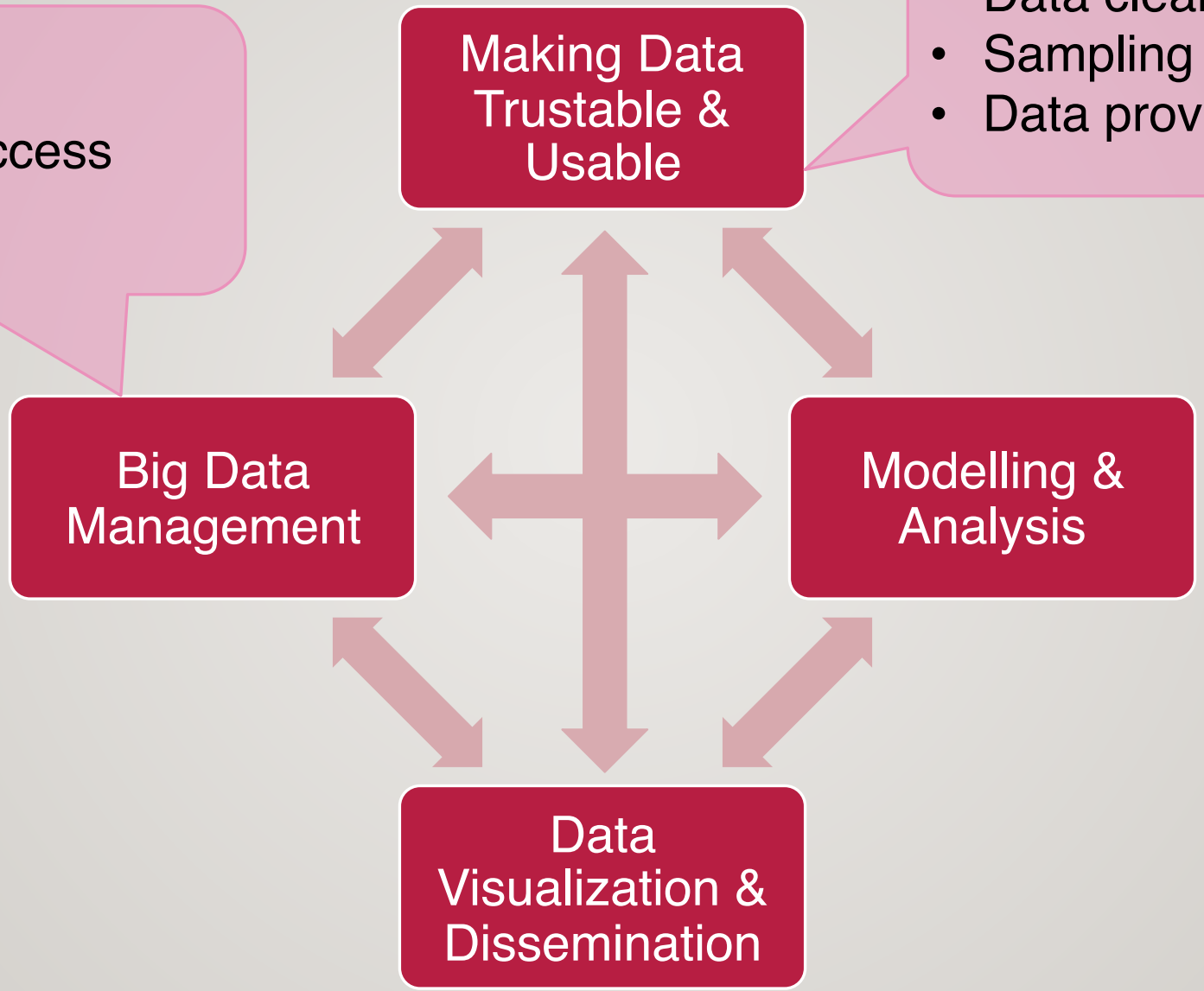
Core Research Issues & Interactions



- Data cleaning
- Sampling
- Data provenance

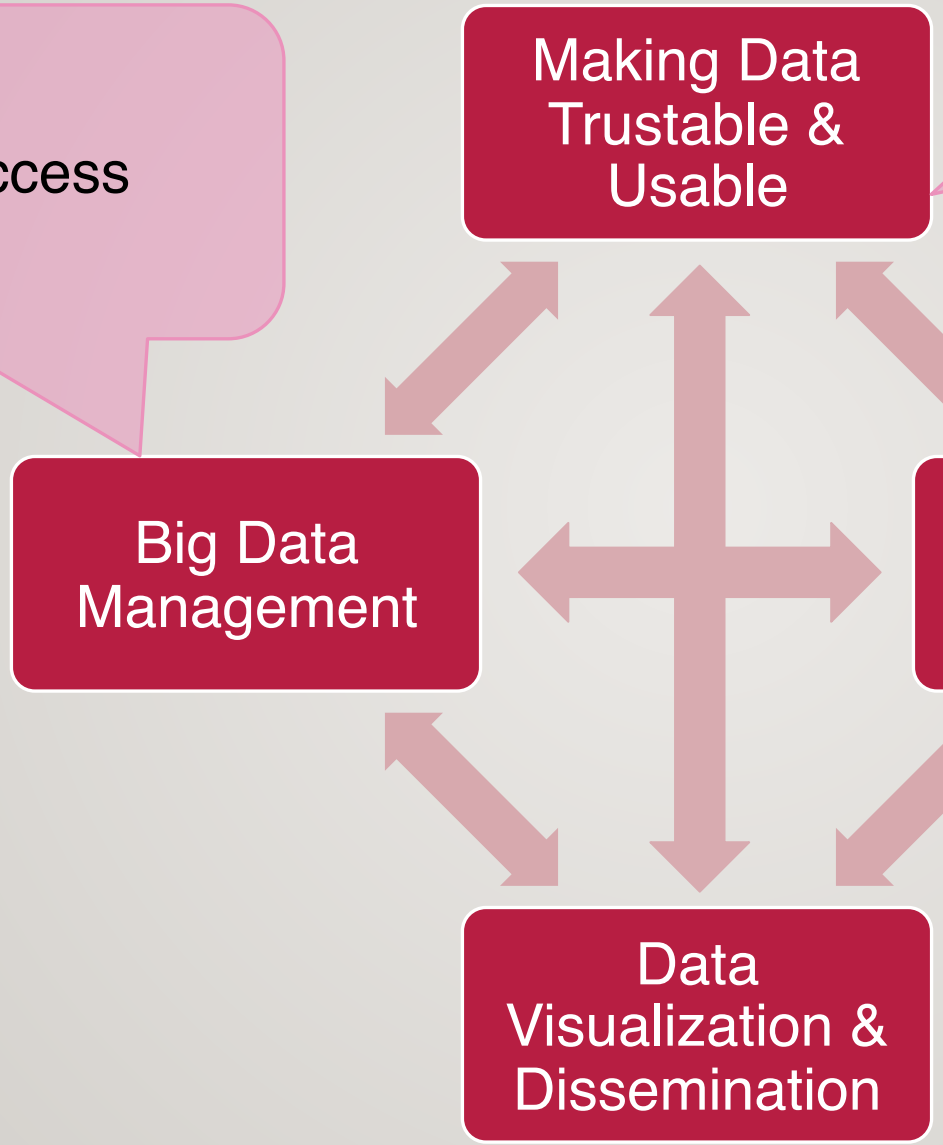
Core Research Issues & Interactions

- Data lakes
- Batch & online access
- Platforms



- Data cleaning
- Sampling
- Data provenance

Core Research Issues & Interactions

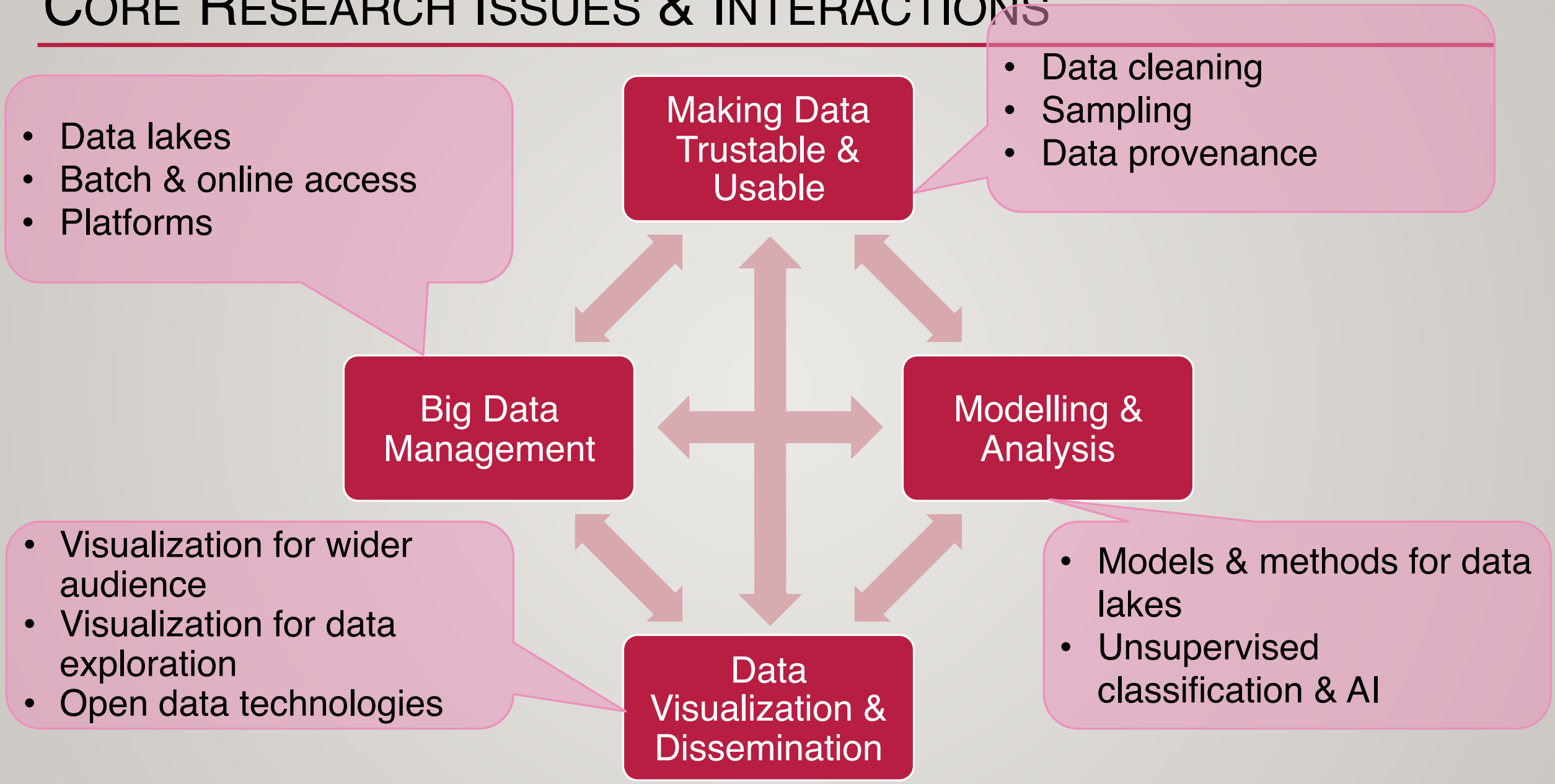


- Data lakes
- Batch & online access
- Platforms

- Data cleaning
- Sampling
- Data provenance

- Models & methods for data lakes
- Unsupervised classification & AI

Core Research Issues & Interactions



Core Research Issues & Interactions