Big Data

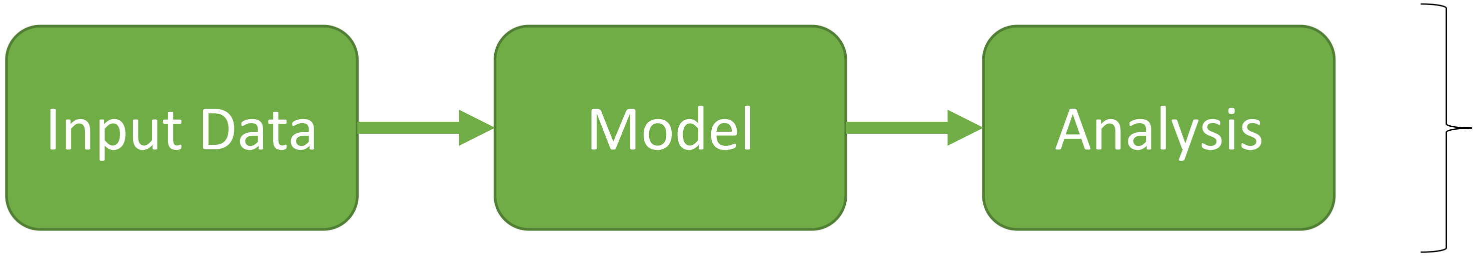
Introduction

Rachana B S, Usha Devi B G, Resma K S, Prafullata K A, Subramaniam K V
Department of Computer Science and Engineering
{rachanabs,ushadevibg, resma, prafullatak, subramaniamkv}@pes.edu
+91 80 6666 3333 Extn 877

What is Big Data?

There is no one standard single definition.

Big Data is data whose scale, diversity, and complexity require new architecture, techniques, algorithms, and analytics to manage it and extract value and hidden knowledge from it…



Modelling pipeline

<u>Model</u> – is a human construct that better helps us understand real-world systems/phenomena.

With Big Data, this means….

Big Data themes

- How to manage very large amounts of data (*data management)*
- and extract value and knowledge from them (*analytics)*

- Google: store index to WWW and search
- Amazon: store user purchases and make recommendations

Large-Scale Data Management

Big Data Analytics

Data Science and Analytics

Big Data: Motivating Example

Big Data themes

- How to manage very large amounts of data (*data management)*
- and extract value and knowledge from them (*analytics)*

- Google: store index to WWW and search
- Amazon: store user purchases and make recommendations

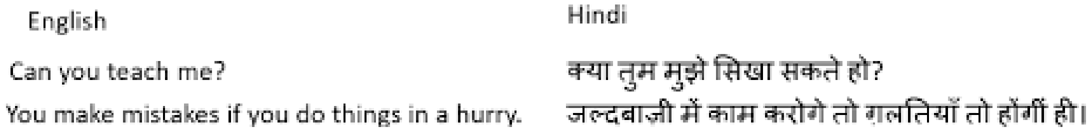
Large-Scale Data Management

Big Data Analytics

Data Science and Analytics

High level approach: motivating example

- Machine Translation
- Translating a sentence from English → Hindi
- What would be the traditional approach?
- How will it differ from the Big Data approach?

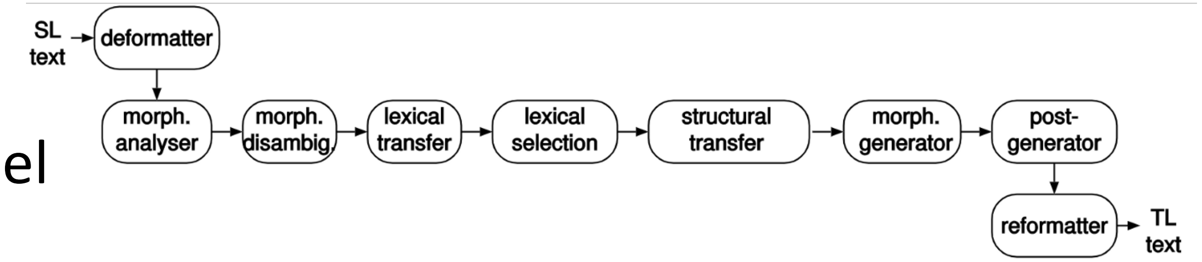


English	Hindi
Can you teach me?	क्या तुम मुझे सिखा सकते हो?
You make mistakes if you do things in a hurry.	जल्दबाज़ी में काम करोगे तो ग़लतियाँ तो होंगी ही।

https://towardsdatascience.com/intuitive-explanation-of-neural-machine-translation-129789e3c59f

Traditional Approach

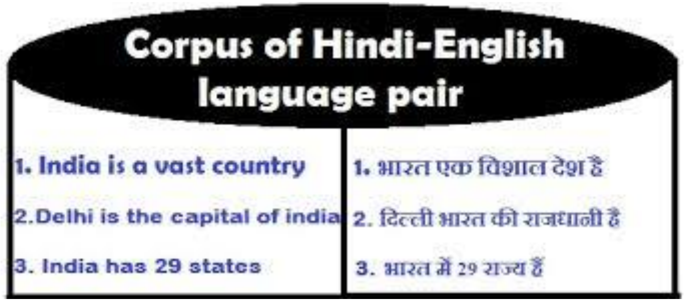
- Understand the system – linguistic approach – rule based
- Requires a linguistic expert to build a model
- Model should include
  - Language structure → morphology, grammar
  - Meaning of the words
  - Mapping words from one language to another



https://towardsdatascience.com/machine-translation-a-short-overview-91343ff39c9f

Big Data Approach

- No attempt to understand language
- Gather data about different sentences and translations
  - Requires a parallel corpus
  - Millions of sentences and their translations
- Build a statistical model
- For example:
  - Every time the word cat appears in the English sentence
  - The hindi equivalent has *billi*
  - So infer that cat can be translated as ***billi***

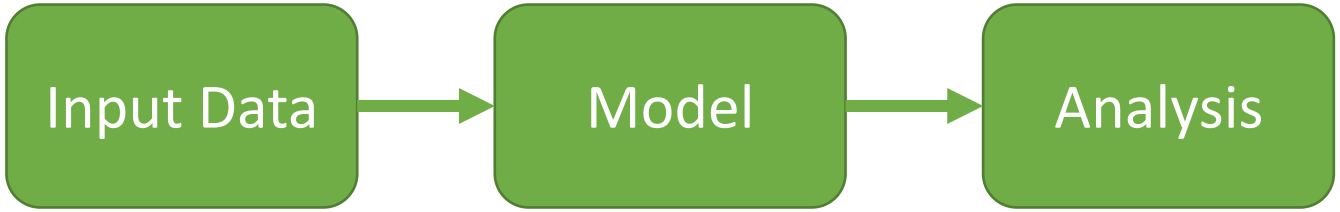


Corpus of Hindi-English language pair

1. India is a vast country	1. भारत एक विशाल देश है
2.Delhi is the capital of india	2. दिल्ली भारत की राजधानी है
3. India has 29 states	3. भारत में 29 राज्य हैं

https://techmediahub.com/machine-translation-complete-useful-guide/

Big Data and Analytics



<u>Traditional Approach</u>

The model is human generated

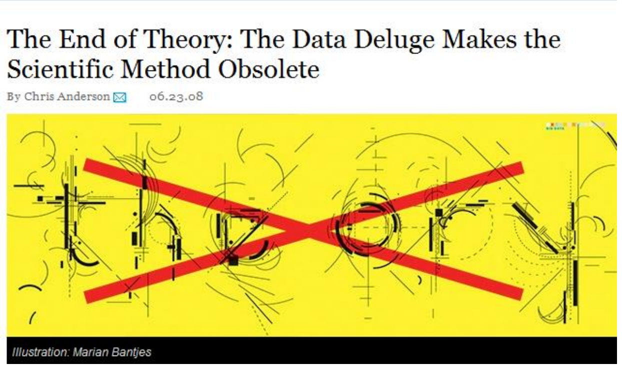
<u>Big Data Approach</u>

The model is machine generated

What about domain knowledge?

- Correlation is enough?
- Gene sequencing of DNA fragments found in ocean by J. Craig Venter
  - 1000s of new species
  - No idea of what species looks like or any other info

- All models are wrong, and increasingly you can succeed without them
  - Peter Norvig, Google's research director
  - “The unreasonable effectiveness of data”



The End of Theory: The Data Deluge Makes the Scientific Method Obsolete

By Chris Anderson 06.23.08

Illustration: Marian Bantjes

http://www.wired.com/science/discoveries/magazine/16-07/pb_theory

Conclusions from Peter Norvig's talk

- Algorithms are not important, data is
  - Domain knowledge (e.g., physics/grammar) is not important
- Demonstrates how images can be merged together using just data
- And translation of text giving examples of issues in segmentation

EXPERT OPINION

Contact Editor: **Brian Brannon,** bbrannon@computer.org

The Unreasonable Effectiveness of Data

Alon Halevy, Peter Norvig, and Fernando Pereira, *Google*

- Peter Norvig, Head Google Research, *The Unreasonable Effectiveness of Data* *https://www.youtube.com/watch?v=yvDCzhbjYWs*

What about domain knowledge?

- Can we rely only on data alone?
- Does this mean that domain knowledge is obsolete?

Big Data: Pitfalls in Analysis

Issues in machine translation

- What about *let the cat out of the bag*?
  - Naïve translation - *billi ko bag ke bahar chhod diya*
  - English meaning: reveal a secret

- To be able to solve this, we need information about the language → domain knowledge and some experimentation

EXPERT OPINION

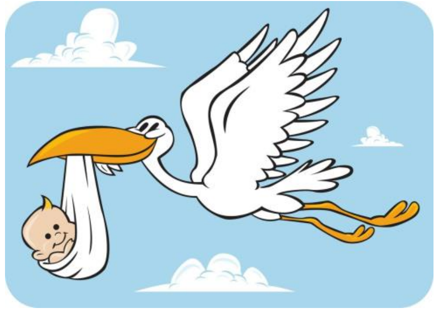
Contact Editor: **Brian Brannon**, bbrannon@computer.org

The Unreasonable Effectiveness of Data

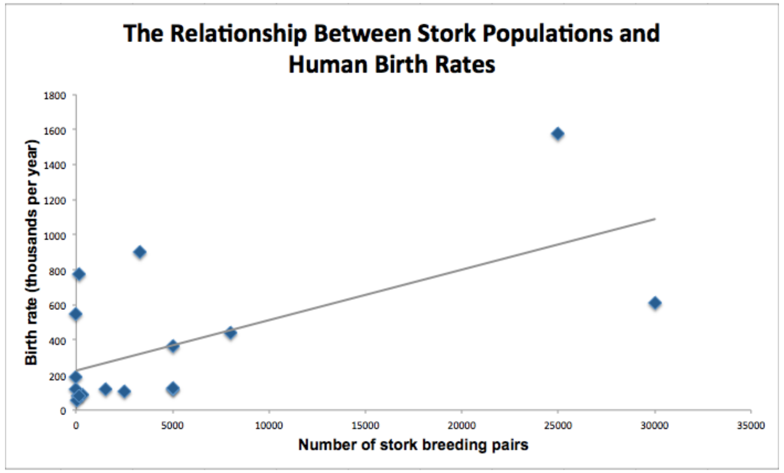
Alon Halevy, Peter Norvig, and Fernando Pereira, *Google*

- Peter Norvig, Head Google Research, *The Unreasonable Effectiveness of Data*
  https://www.youtube.com/watch?v=yvDCzhbjYWs

Pitfall : Spurious correlation



- C->A, C->B
  - *Does A->B?*
- Example:
  - Do storks deliver babies?
- Chart shows positive correlation between
  - Stork population and human birth rates in European countries
  - What it does not show is a hidden variable
    - Available nesting area?



The Relationship Between Stork Populations and Human Birth Rates

- http://en.wikipedia.org/wiki/Spurious_relationship
- http://www.cut-the-knot.org/do_you_know/misuse.shtml

Pitfall : Gaps in the data

- Selection bias
- Convenience
- Example
  - Rutgers University study
  - Examine decision-making process in emergency
  - Study tweets during Hurricane Sandy
  - Most tweets from Manhattan!
  - If studying impact of Sandy: *Manhattan most impacted!*

- *More Data, More Problems: Is Big Data Always Right?* ARI ZOLDAN http://www.wired.com/insights/2013/05/more-data-more-problems-is-big-data-always-right/

Pitfall : Gaps in the data

- Another example: medicine

- Missing data is always a challenge
    - but we also know that “negative results” are more likely to go missing.
    - This means we have a *__biased sample__*, overestimating the benefits of treatments.

O'REILLY

Strata CONFERENCE Making Data Work

Feb 28–Mar 1, 2012
Santa Clara, California

#strataconf

strataconf.com

Keynote by Ben Goldacre

Ben Goldacre, *Bad Science*

- *The Information Architecture of Medicine is Broken* Ben Goldacre http://strataconf.com/strata2012/public/schedule/detail/22941
- https://www.youtube.com/watch?v=AK_EUKJyusg

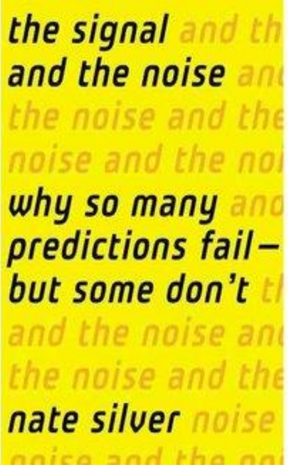
Big Data: How to address the issues?

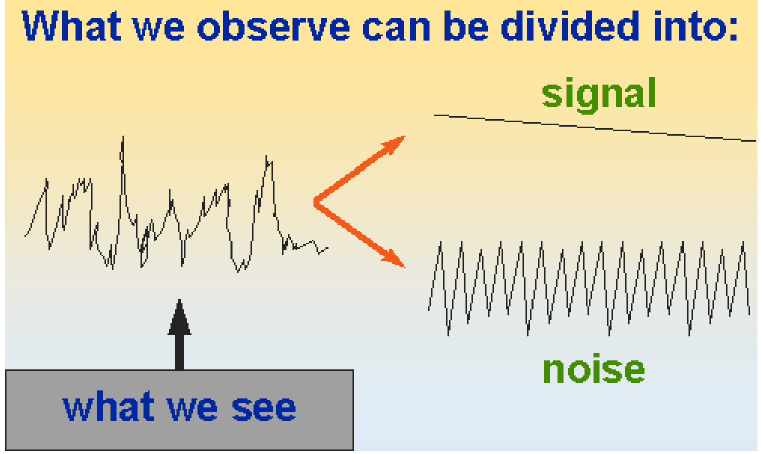
Summary of the methods

- Use domain knowledge to check model for validity
- Estimate errors

Let’s look to some experts

- Nate Silver book
  - The Signal and the noise
- On Time Magazine 2009 – 100 most influential people
- Correctly predict US 2008/2012 elections



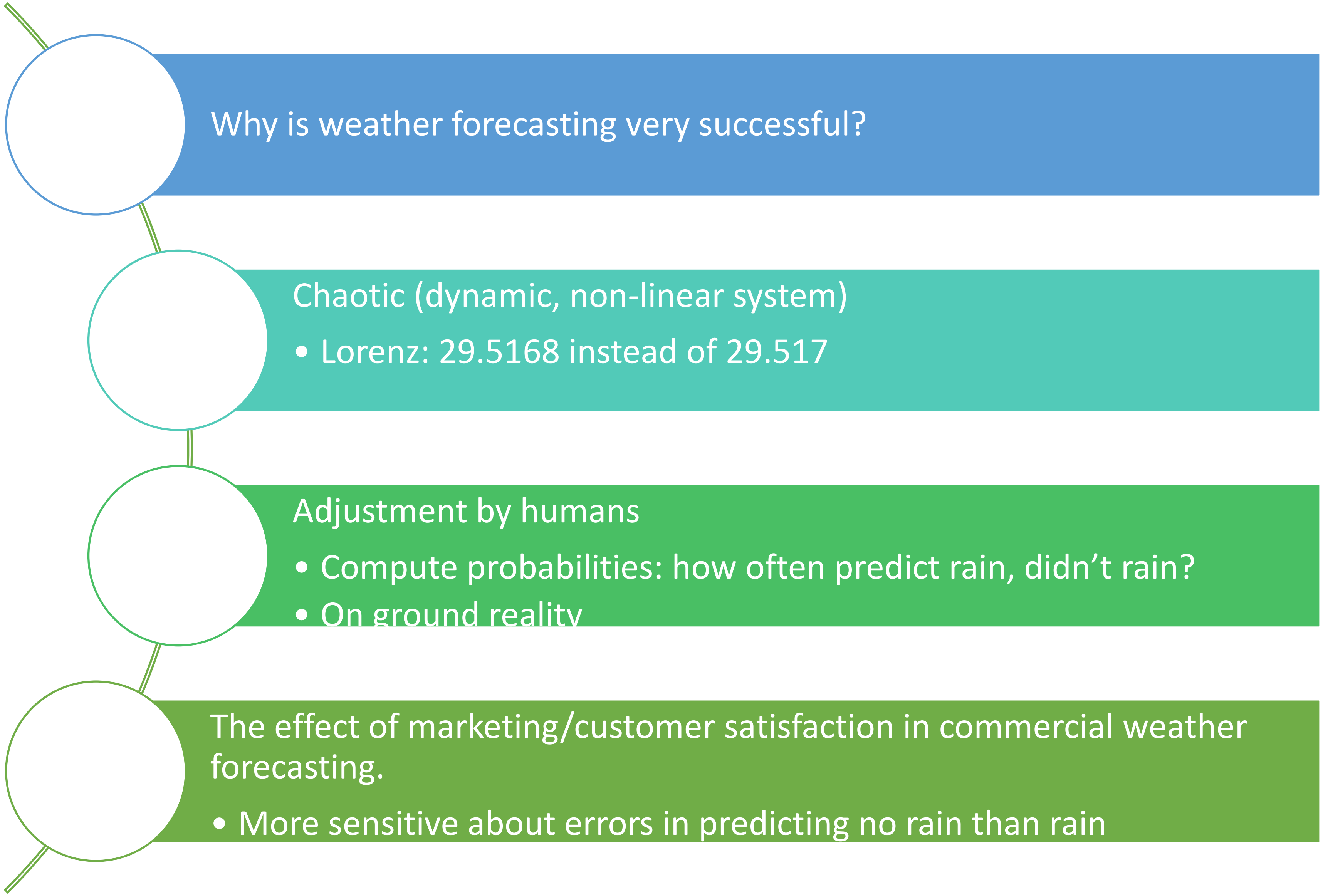

What we observe can be divided into:

signal

noise

what we see

Example: Weather Forecasting



Why is weather forecasting very successful?

Chaotic (dynamic, non-linear system)

- Lorenz: 29.5168 instead of 29.517

Adjustment by humans

- Compute probabilities: how often predict rain, didn’t rain?
- On ground reality

The effect of marketing/customer satisfaction in commercial weather forecasting.

- More sensitive about errors in predicting no rain than rain

Big Data Error Estimation

- Purely empirical: cannot be analysed by theory
- Divide data into *training set* and *testing set*
- Develop algorithm using training set; estimate error from testing set
  - Can be used to compare analytics algorithms
- Examples
  - Nate Silver: weather prediction: human adjustment
  - Amazon recommendations
    - Derive model using historical data; make recommendations
    - Get statistics on how many people look at or buy recommendations

Big Data: Summary and architecture

Big Data themes

- How to manage very large amounts of data (*data management)*
- and extract value and knowledge from them (*analytics)*

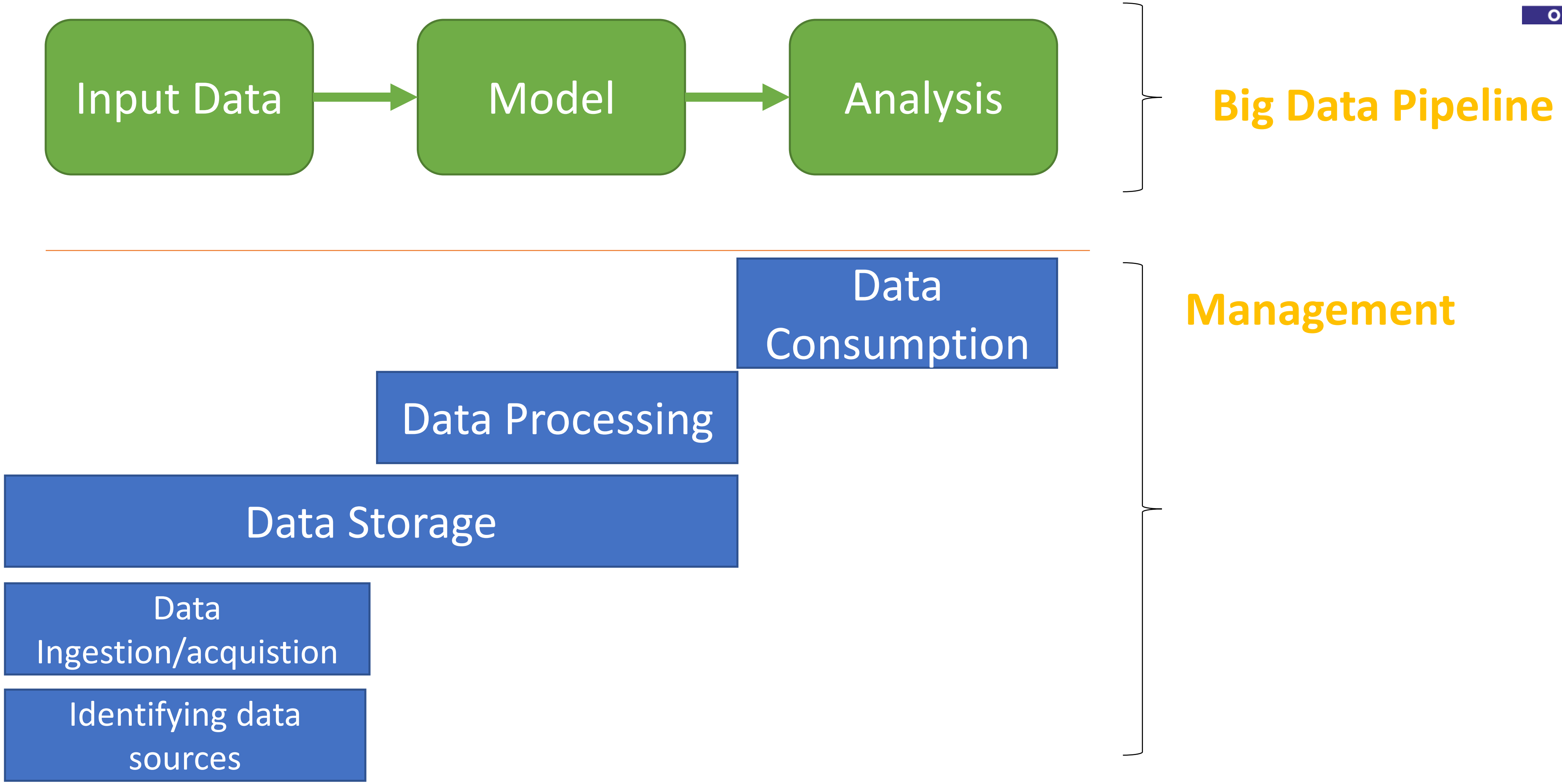
- Google: store index to WWW and search
- Amazon: store user purchases and make recommendations

Large-Scale Data Management

Big Data Analytics

Data Science and Analytics

Big Data and Analytics



THANK YOU

K V Subramaniam

Department of Computer Science and Engineering

subramaniamkv@pes.edu

+91 80 6666 3333 Extn 877

Content Creators

Prof Prafullata K A
Prof Usha Devi
Prof Rachana
Prof Resma