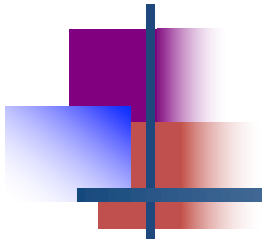


Predictive Analytics Using R & Python

Anusua Trivedi
Data Scientist, TACC
trivedi.anusua@gmail.com

A little about TACC

- One of the largest supercomputer center
- Non-profit, part of UT Austin
- HIPAA compliant
- How can TACC help you?

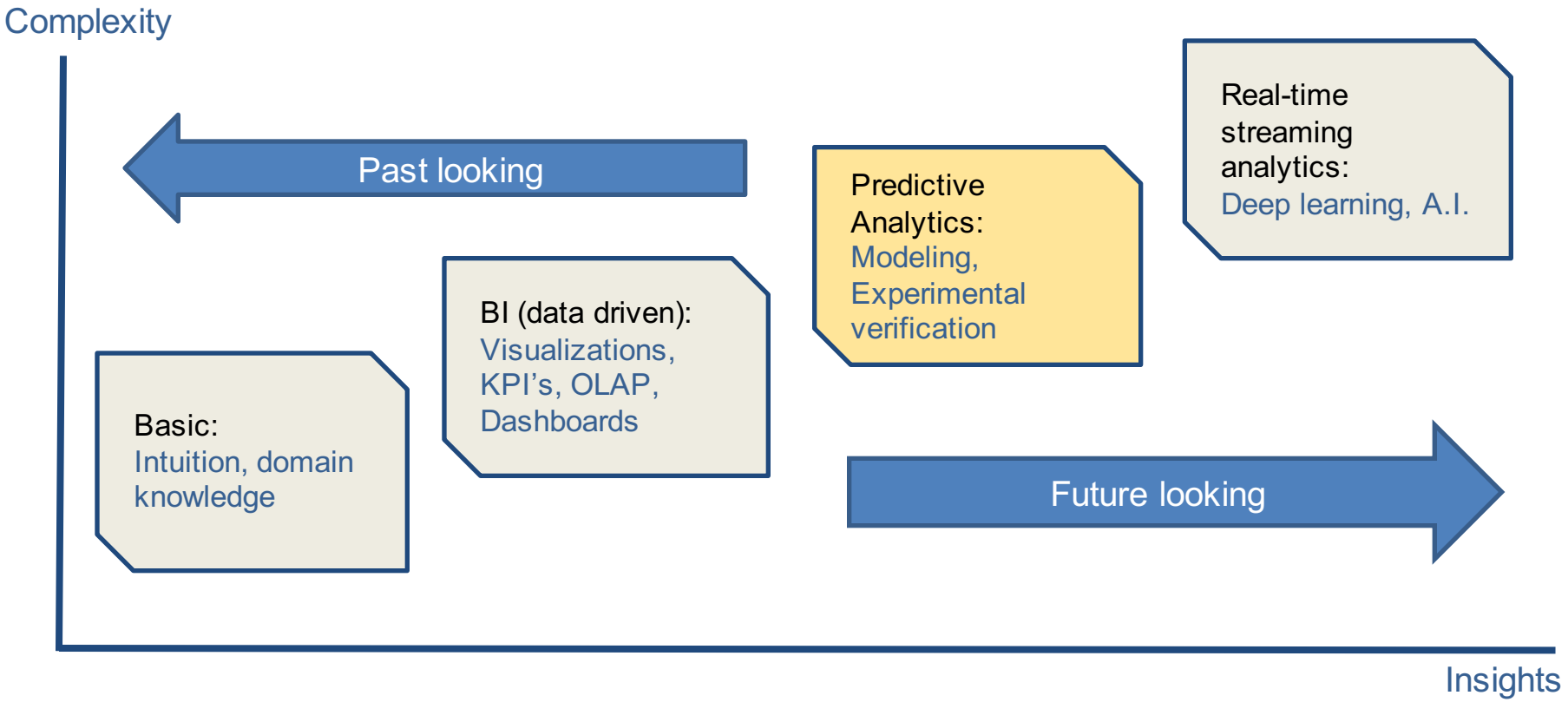
TACC

THE UNIVERSITY OF TEXAS AT AUSTIN

Talk Outline

- Evolution of Analytics
- Predictive Analytics: Definition
  - –Predictive Analytics: Process
  - –Traditional Predictive Analytics
- R
  - –Simple Graphics in R
  - –Predictive Analytics Use Cases using R
- Python
  - –Predictive Analytics Use Cases using Scikit-learn

Evolution of Analytics

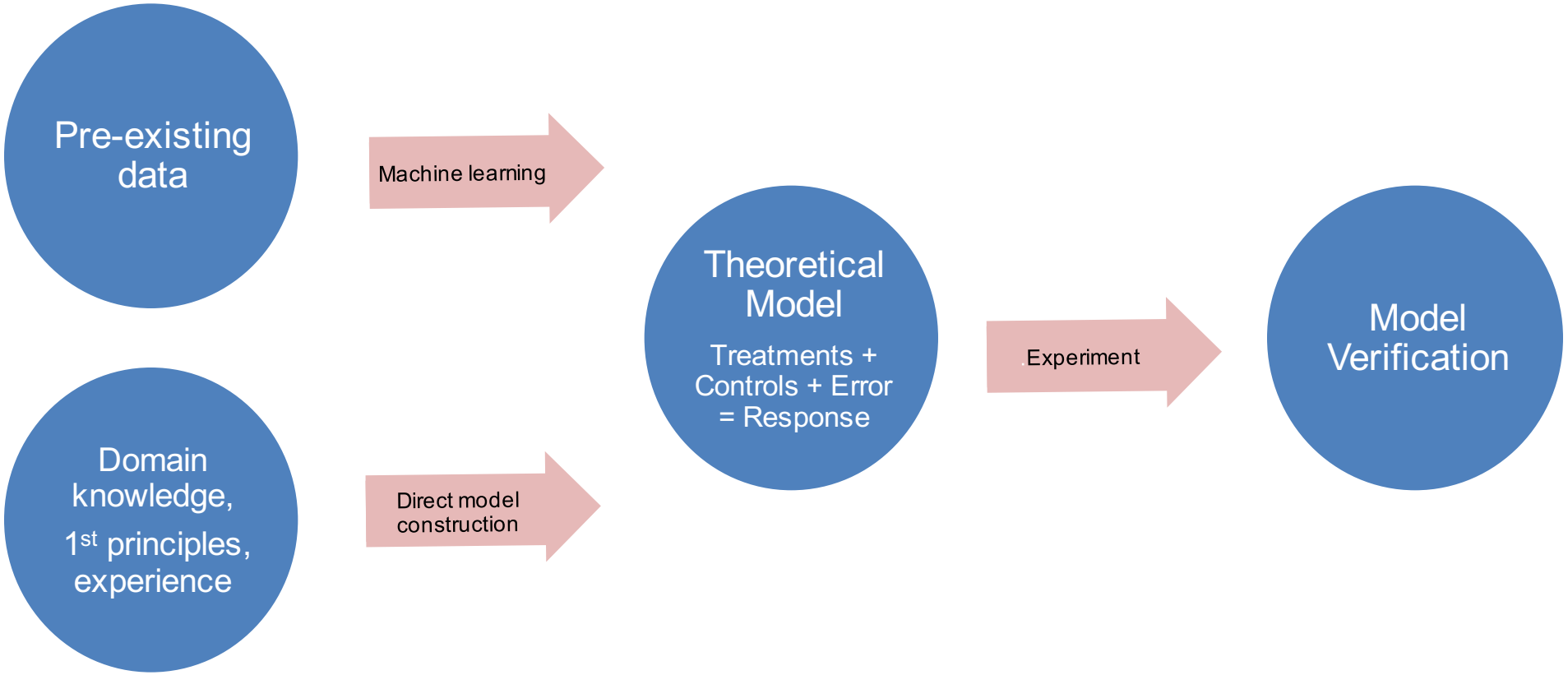


Predictive Analytics - Definition

Predictive analytics is an area of data mining that deals with extracting information from data and using it to predict trends and behavior patterns. Often the unknown event of interest is in the future, but predictive analytics can be applied to any type of unknown whether it be in the past, present or future.

-- Wikipedia

Predictive Analytics Process



Traditional Predictive Analytics

- Applications
  - –R&D in agriculture and industry
  - –Verifying the effectiveness of healthcare treatments
    - •Clinical trials (randomized double blind are best)
  - –Predicting the weather
  - –Polls: opinions, election politics, Nielson ratings.
  - –Standardized tests: e.g. SAT's to predict college success
  - –Actuarial Science: life expectancy, etc.

- Limitations
  - –Each data point must be planned and collected intentionally
  - –Expensive to design, collect the data

How Predictive Analytics is done?

- In Machine Learning, Predictive Analytics is a kind of supervised learning.

- Predictive Analytics is about predicting future outcome based on analyzing data collected previously. It includes two phases:
    1. Training phase: Learn a model from training data
    2. Predicting phase: Use the model to predict the unknown or future outcome

- We can choose many models, each based on a set of different assumptions regarding the underlying distribution of data.

- We are interested in two general types of problems in this discussion:
    1. Classification—about predicting a category (outcome is discrete & finite, with no ordering implied).
    2. Regression—about predicting a numeric quantity (outcome is continuous and infinite, with ordering).

Use Cases

- Business / organizational goals drive initiatives:

1. Improve outcomes
    - Improve client outcomes
    - Improve products, customer satisfaction
2. Increase profits by increasing revenue
3. Increase profits by reducing costs

What is R?

- R implements a dialect of the language that was developed at AT&T Bell Laboratories.
- Versions of R are available, at no cost, for versions of Microsoft Windows, Unix/Linux, Mac OS.
- Supporting packages are available through the Comprehensive R Archive Network (CRAN).
- Neither R nor any statistical system will provide you the statistical expertise, it just acts as a analysis tool.

What is RStudio?

- RStudio is the convenient interface for R.
- Versions of RStudio are available, at no cost, for versions of Microsoft Windows, Unix/Linux, Mac OS.
- When we first launch RStudio:
  - The frame in the upper right contains your workspace
  - Any plots will show up in the lower right corner.
  - The left frame is the console where the action happens.
  - Below console is the prompt, which is really a request for a command.
- RStudio is a fairly efficient way to access and organize data, describe and invoke statistical computations.

What is Anaconda?

- **Anaconda**
  - Completely free Python distribution for large-scale data processing, predictive analytics, and scientific computing
  - 130+ of the most popular Python packages for science, math, engineering, data analysis
  - Cross platform on Linux, Windows, Mac
  - Miniconda available for small footprint installs -- contains conda and Python

What is IPython?

- "The IPython Notebook is a web-based interactive computational environment where you can combine code execution, text, mathematics, plots and rich media into a single document”
  - More than an IDE
  - Programmers and people who program
  - Integrated visualization and processing
  - Storying telling with Data

What I will talk about

- Machine Learning Methods using R & Python
  - –Simple methods
  - –Helpful Libraries
- Method Details
  - –Ideas
  - –Assumptions
  - –Implementations

What I won't Talk about

- Machine Learning Methods
  - –Classical, but complex methods (e.g., neural networks, deep learning)
  - –Methods not widely used

Acknowledgement

1. Ricky Ho tutorials/blogs
2. Joshua Reich blogs
3. Code adapted from R-bloggers: http://www.r-bloggers.com/
4. Quick-R: http://www.statmethods.net
5. Anaconda/IPython
6. Scikit Learn
7. Open blogs/tutorials

Hands-on Tutorials