Data Science: Historical View

S. Lakshmivarahan
School of Computer Science
University of Oklahoma
Norman, OK- 73019, USA

Early beginnings in Astronomy

- Much of what we know in physical sciences had their origins in Astronomy - with observations of celestial objects
- Thanks to the Herculean efforts of pioneers:
- Copernicus (1473-1543)
- Galileo (1544-1642)
- Kepler (1571-1630)
- Newton (1643-1727) are note worthy among others

Early observations suggest existence of simple laws of nature

- Observations collected over decades were meticulously analyzed to formulate new laws of nature: Examples include
- Heliocentric system
- The laws of Kepler
- Law of gravitation by Newton
- Newton's laws

Note: Within the context of physical sciences these are some of the earliest examples of **data mining**

What is data mining (DM)?

- DM is the process **extracting the structure** or **patterns** that are inherent in the data/ observations
- These patterns give clues about the **data generating process**
- **Goal of DM is to understand the data generating process**
- Since the motion of celestial objects inherently followed certain laws, early pioneers with their hard work and ingenuity could discover the laws that laid the foundation of the physical sciences and engineering as we know today

Renewed interest in DM with the Abundance of data

- It is estimated that the volume of data collected **doubles in every three years** -thanks to computers, large scale storage device, communication and sensor technologies
- **Today interest in DM include** - Physical sciences, Biological sciences, Space exploration, All of Engineering, Environmental Sciences, Medical Sciences, Economics, Finance, Banking and Commerce, Sports and recreation, Governments at all levels, to name a few
- More about DM a bit later – back to early astronomy

Development of models

- The newly found Newton's laws along with the concurrent developments in **Calculus** by Newton (1643-1727) and Leibnitz(1646-1716)and others naturally lead to the development of **dynamic models** to describe the motion of planets around the sun
- With the availability of models, the **potential for forecast or prediction** became very clear

Beginnings of DA

- Gauss (1777-1855) (when he was only 24 years old) using the known models of his time, takes up the challenging problem of predicting when the celestial object called ***Ceres*** will reappear on the telescope
- By combining the model and the observations, he then created the "**first assimilated model**" to make an accurate prediction of the time and location of reappearance of the lost object
- Data assimilation involves estimation of the unknown parameters of the model in question

Gauss in 1801 laid the foundation for DA

- This work leads the development of the method of **least squares** which continues to be the work horse for the data assimilation industry today
- By this time, he has also invented the notion of distribution observational errors following the bell-shaped curve which we now call as the **normal or Gaussian distribution**
- Gauss made a successful prediction of the bearing and the time of reappearance of the last object with great accuracy

What is DA? – Fusion of model with data

- Models are general descriptions of the underlying physical processes in question.
- Data/observations reveal all the secrets of or the truth about the process that model tries to capture
- By combining models and data, we can get a specialized model called the **assimilated model**
- This assimilated model is a good tool for creating forecast or prediction
- One of the standard tools for the fusion of model and data is based on the **method of least squares.**

Goal of DA is to generate good forecast: Examples: Model + data

- Predict the path of a hurricane, tornado- use of several models + data collected via special planes
- From the crime scene data, reconstruct the case – CSI, Miami
- NTSB – estimate the causes of failure using the data from the debris
- Predict the potential tax revenues so that a Government can develop its budget for the next year
- Medical diagnosis – from symptoms to the cure

Early Indian Astronomy

The four yugas: Krita, Treta, Dvapara and Kali

Ramayana in Treta yuga – about 7,000 years ago

Mahabharata in Dvapara yuga – about 5000 years ago

Lord Rama's horoscope: 5 planets in exalted position – Valmiki Ramayana

- Sun is in Aries & exalted
- Moon is in Cancer – in his own home
- Jupiter is in Cancer & is exalted
- Saturn is in Libra & is exalted
- Mars is in Capricorn & is exalted
- Venus is in Pisces & is exalted
- Rahu is in Sagittarius
- Ketu is in Gemini

Great Astronomers of Kaliyuga

- Aryabhata (476-550) –Patna –Gupta Dynasty
  - Earth is spherical, revolves around the sun and 365 days in a year
- Varahamira (505-587) –Ujjain, MP
- Brahmagupta (598-668)-Bhinmal, Rajasthan
- Baskara (598-680) – Sourastra, Gujarat
- A whole host of scholars who have made lasting contributions to Astronomy and Mathematics

Aspects of DA: A Classification of models

- Static vs. Dynamic (ODE or PDE)
- Deterministic vs. stochastic
- Linear vs. nonlinear
- Discrete vs. continuous time
- Discrete vs. continuous space

Computational model

- Embed a grid over the domain(1,2,3-D space and time) of interest
- N denote the number of grid points
- At each point we may define L physical variables: u, v, w, p, T, q, etc
- n = N*L denoted the number of state variables in the model ~ 10*6 to 10*9
- Model state is denoted by an n vector, x

Relation between model state and observations

- Observations are denoted by an m vector, z
- Relation between z and x is critical to data assimilation
- The model state x may not be directly observable, but a function of x is
- This function is either linear: $z = Hx$ or
- Nonlinear: $z = h(x)$

Examples of the relation between state and observations

- **Doppler Radar**: z is reflectivity, x is rain – nonlinear empirical relation
- **Satellites**: z is energy radiated, x is temperature, nonlinear Stefan's or Planck's law
- **Balloons**: Directly measures the temperature, wind, pressure, etc
- **OK Mesonet**: Directly measures temperature, pressure, temperature, humidity soil moisture, etc
- **Dual polarization and Phased array radars**: Amount of data is going to an order of magnitude larger than the Doppler radars

Methods of DA

- **3-D VAR** – static, deterministic or stochastic (linear or nonlinear) models, and linear or nonlinear observations – Related to Bayesian approach – similar to Kalman filtering
- **4-D VAR** – Dynamic, deterministic (linear or nonlinear) models and linear and nonlinear observations. This leads to the so called adjoint method (Lewis & Derber (1985), Le Dimet and Talagrand(1986)
- **Kalman filtering** – Dynamic stochastic (linear or nonlinear) models and linear and nonlinear observations – Kalman (1960), Kalman and Bucy (1961)

Methods for DA - continued

- To address the computational difficulties of large-scale nonlinear problems new ideas were introduced
- **Ensemble** (Kalman) **filtering** (Evensen(1994)
- **Unscented filtering** (Julier, Uhlmann and Durrant-Whyte (1995)
- **Particle filtering** (Metropolis and Wiener in 1940, but made feasible only in the 1980's. Refer to the book by Doucet, de Freitas and Gordon (2001)

Back to DM: Methods

- Largely relies on **empirical methods** including:
- **Time series analysis** (1950's) – Signal processing in EE, Medicine, Econometrics, Finance
- The goal is to build dynamic models in discrete time by exploiting the underlying correlation, seasonality properties of the data set
- Autoregressive, integrated, moving average (ARIMA) models
- This is one of the developed areas in empirical modeling

Ref: J. D. Hamilton (1994) *Time Series Analysis*, Princeton University Press, 799 pages

DM methods – ML/AI

- **Multivariate regression analysis** (1800's) – Statistics
- **Data reduction** using PCA (1940), ICA (1990) - Statistics
- **Classification** using **clustering** (1950's), **Neural Networks** (1950's), **Pattern Recognition** (1950's),
- **SVM** (1980's) **Association rules**, – Statistics, Image processing, voice recognition, etc.,
- **Decision trees** (1960) **probabilistic networks** (1990's) –AI, Machine learning
- **Genetic algorithms** – Discrete optimization
- **Random field** - Spatial data analysis

Representation of models in DM: explicit or implicit

- Time series provides an explicit ARIMA model
- Neural network, SVM, the classification rule are implicit representations of the model

Summary

- DM seeks to uncover the structure or patterns in the data so that we can summarize the intrinsic properties of the Data generating process as a model
- The goal of DA is to fuse or fit the model to data and the fitted model has better predictive power
- DM, DA and Prediction as parts of a continuum and are the pillars in the development of a Predictive Science/Data Science

Rose smells the same by whatever name you call

1) AI, Expert systems, Knowledge discovery,
2) M/L : Classification, Pattern Recognition- SVM, ANN/DNN/
3) Multivariate Data Analysis, Data Mining, Statistical Learning
4) Data Science, Big Data Analytics
5) Data centric AI

What is Data Science?

- The three basic building blocks of DS are:
  - Probabilistic, Statistical and Numerical Mathematics
  - Computer Science
  - Domain specific knowledge
- Growth in Computer Science was driven by advances in
  - Discrete mathematics, Theory of formal language, Algorithms and complexity theory and hard-ware technology
- Growth in Data Science will be propelled by advances
  - Statistics, Probability theory, numerical methods and large-scale simulation

Applications of AI/DNN

- Automatic speaker/speech recognition
- Image recognition-Handwritten
- Visual art processing – style, period,
- Natural Language processing
- Drug discovery/toxicology
- Customer relationship
- Recommender system
- Bioinformatics- discovery of gene functional relationship
- Image restoration

Algorithmic decision making-challenges

- DNN based approach lacks theory – why it works?
- Long way from building causality theory
- DNN is a step towards realizing strong AI, but not all-encompassing solution
- Transparency is needed to gain confidence
- Democratizing AI ? – Open access subject to malicious attacks
- DNN based decisions are sensitive perturbations in the input, malicious changes in the components of the architecture

Reading list

- B. G. Lindsay, J. Kettenring and D. O. Siegmund (2004) “A report on the future of Statistics”, Statistical Science, Vol 19, 387-413
- IMS Presidential Address: “Let us own data science” IMS Bulletin, Vol 43, Issue 7, 2014
- J. Fan, F. Han and H. Liu (2014) “Challenges of Big Data analysis”, National Science Review, Vol 1, 293-314
- D. Donoho (2017) 50 Years of Data Science, Journal of Computational and Graphical Statistics, 26:4, 745-766, DOI:10.1080/10618600.2017.1384734

References

- A. Blum, J. Hopcroft, and R. Kannan (2020) **Foundations of Data Sciences**, Cambridge University Press

- J. M. Lewis, S. Lakshmivarahan and S. K. Dhall (2006) **Dynamic Data Assimilation: a least squares approach**, Cambridge University Press

- V. Shikhman and D. Muller (2021) **Mathematical Foundations of Big Data Analytics,** Springer

References-continued

- I. Rish and G. Ya. Grabarnik (2015) **Sparse Modeling; Theory, Algorithms and Applications,** CRC Press

- **M Wainwright (2019) High-Dimensional Statistics: A non-asymptotic viewpoint, Cambridge University Press**

- J. Wang (2012) **Geometric Structure of High-Dimensional Data and Dimensionality** R**eduction**, Springer

- R. Vershynin (2020**) High Dimensional Probability Theory: an introduction with applications in Data Science**, Cambridge University Press
- B. Abu-Salih, et. al. (2021) **Social Big Data Analytics**, Springer