O'REILLY®

Building Intelligent Analytics through Time Series Data

Sanjian Chen, Ph.D.
Jian Chang, Ph.D.

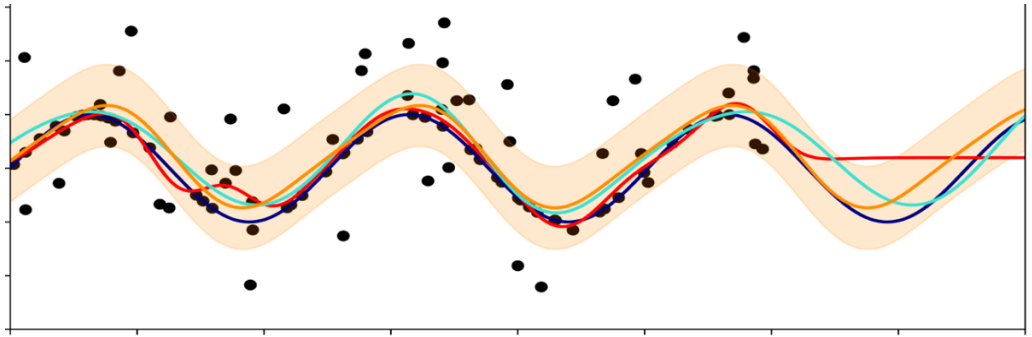
Jan / 2020

Agenda

- Introduction (PPT)
- Time-Series Forecasting Tutorial: Session 1 (Jupyter)
- Break (10min)
- Time-Series Forecasting Tutorial: Session 2 (Jupyter)
- Q & A

Scope of Discussion

- Time Series Data:
  - Data points ordered by timestamps



- Broad Applications
  - Retail sales data
  - Electricity consumption over time
  - Stock price trend

- Use cases
  - Anomaly detection: Is the time series trend abnormal?
  - **Forecasting**: How will the time series evolve in the future?

What can be forecasted?

- Some events are easier to forecast
  - Sunrise time vs weather
- Capture the pattern and ignore noise
  - Not always easy to tell the difference between the two
- Predictability depends on several factors
  - Correlation of past and future patterns
  - Availability of historical Data
  - Knowledge about what’s driving the change

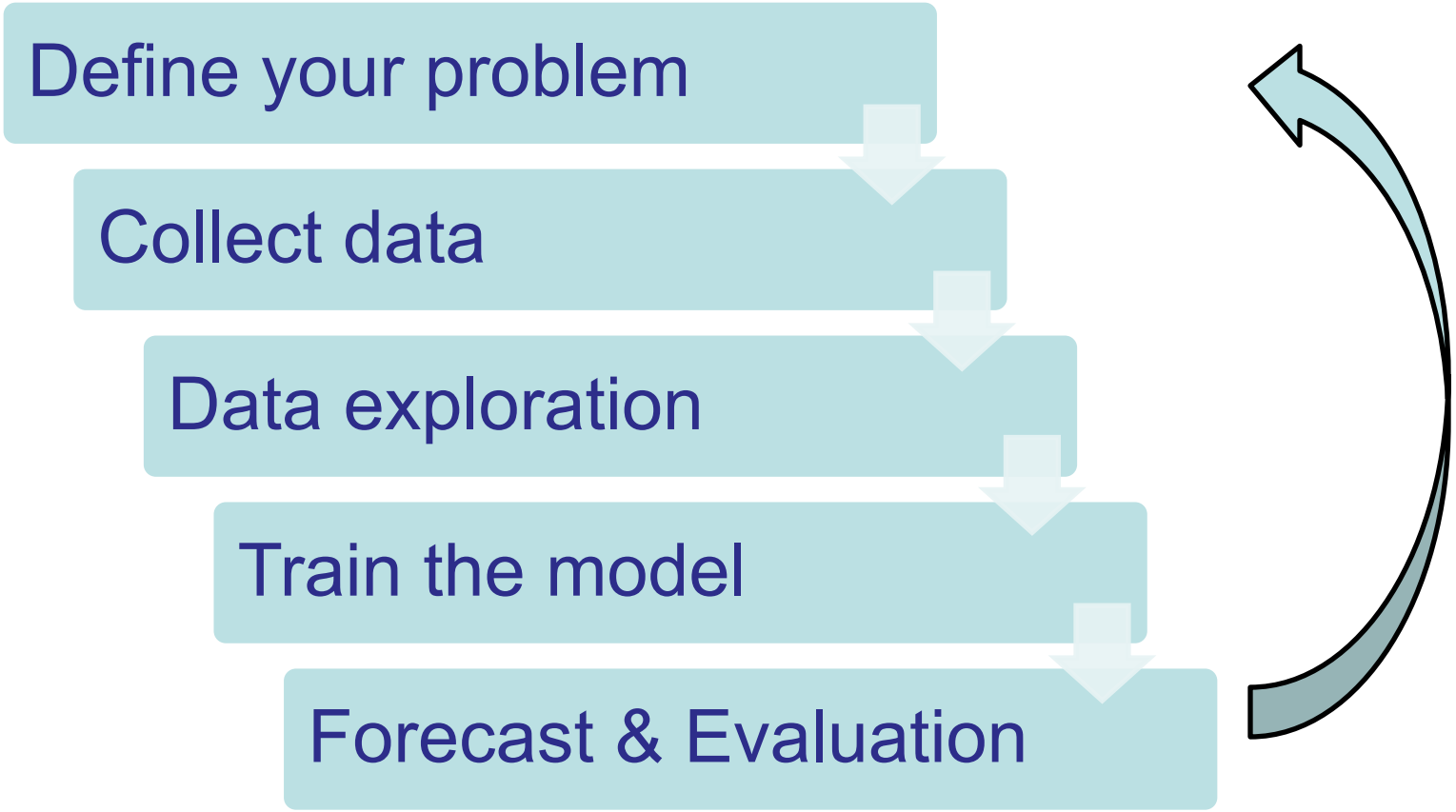
Formulate a Forecasting Problem

- What quantities need to be forecasted
  - Identify data
- Forecasting horizon: long-term vs short-term
- When to update the forecasts
  - Refresh results vs rebuild the model

Forecasting Models

- Explanatory models
  - Explicitly capture the effects of the driving factors
- Data-driven models
  - Only use past observations to predict future values
- Combine the two methods

Basic Forecasting Workflow



Session 1

- Jupyter/Python Quick Intro

- Time Series Forecasting
  - Data Exploration
  - Trend, Seasonality, Residual
  - STL Forecasting
  - Accuracy Measurement

- Prophet Forecasting

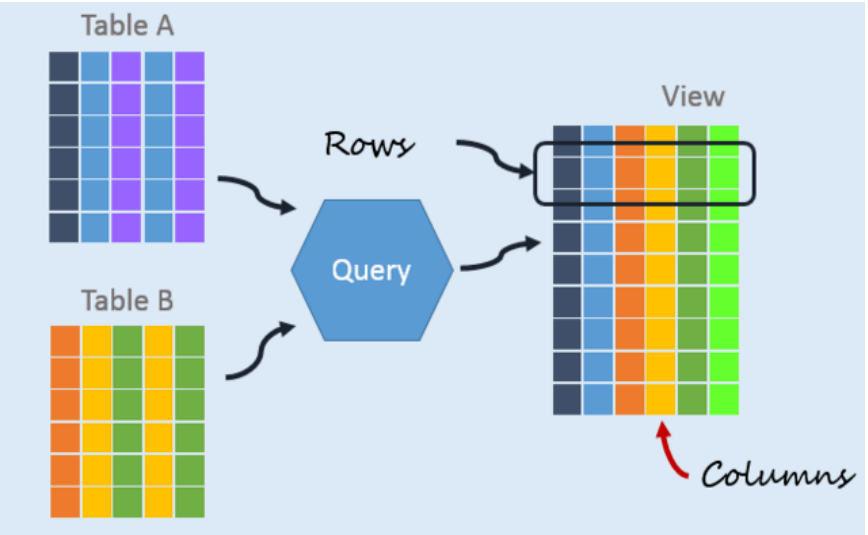
Session 2

- More Forecasting Models
  - Exponential Smoothing (ETS)
  - ARIMA

- Deep Learning Methods and Latest Toolkits
  - GluonTS
  - Deep Neural Network Models
  - “Hacking” the Performance
  - Other Advanced Deep Learning Models

Challenges of Time Series Data Management

- Patterns of Writing:
  - Rate of writing >> rate of query
  - Recent overwrites
  - Number of time series can explode
- Patterns of Query:
  - Requires low-latency response
  - Selecting data by time range
  - Complex aggregation operations (min, max, avg, *etc.*)



- Require high availability, often no needs for transaction semantics (ACID)

TSDB Landscape

☐ include secondary database models — 31 systems in ranking, July 2019

Rank Jul 2019	Rank Jun 2019	Rank Jul 2018	DBMS	Database Model	Score Jul 2019	Score Jun 2019	Score Jul 2018
1.	1.	1.	InfluxDB	Time Series	18.00	+0.02	+6.31
2.	2.	2.	Kdb+	Time Series, Multi-model	5.88	+0.07	+2.65
3.	↑ 4.	↑ 6.	Prometheus	Time Series	3.46	+0.14	+2.07
4.	↓ 3.	4.	Graphite	Time Series	3.44	+0.11	+0.89
5.	5.	↓ 3.	RRDtool	Time Series	2.78	+0.11	+0.13
6.	6.	↓ 5.	OpenTSDB	Time Series	2.30	+0.06	+0.56
7.	7.	7.	Druid	Multi-model	1.85	+0.07	+0.67
8.	8.	↑ 12.	TimescaleDB	Time Series, Multi-model	1.27	+0.16	+1.11
9.	9.	↓ 8.	KairosDB	Time Series	0.53	+0.03	+0.08
10.	10.	↓ 9.	eXtremeDB	Multi-model	0.43	+0.01	+0.10
11.	11.	↑ 21.	Heroic	Time Series	0.39	-0.01	+0.39
12.	12.	↑ 16.	GridDB	Time Series, Multi-model	0.37	+0.01	+0.29
13.	13.	13.	FaunaDB	Multi-model	0.34	-0.02	+0.22
14.	↑ 16.	↓ 11.	Axibase	Time Series	0.26	+0.05	+0.09
15.	↓ 14.		Amazon Timestream	Time Series	0.24	-0.01	

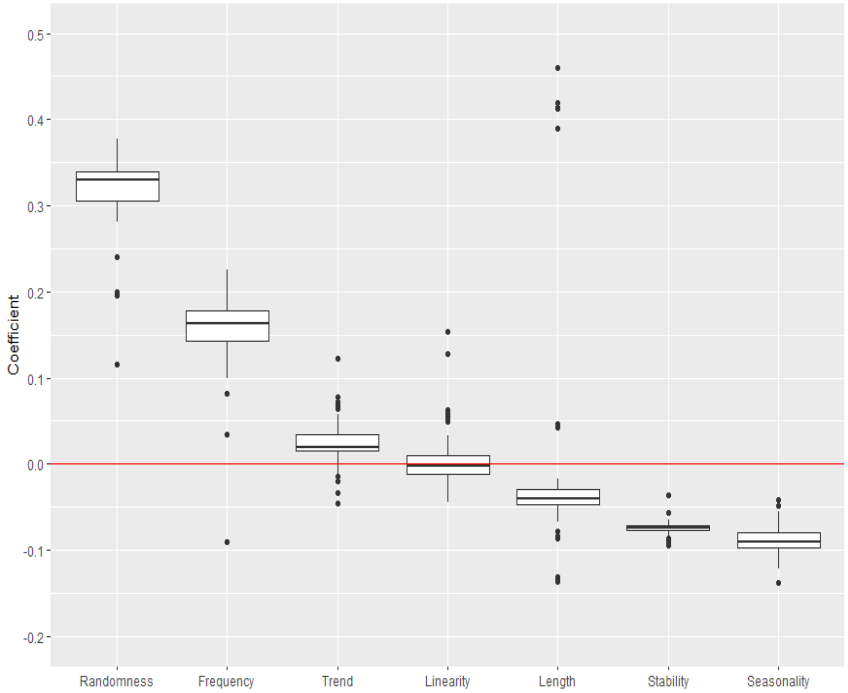
Source: https://db-engines.com/en/ranking/time+series+dbms

No One-Size-Fits-All Solution

- Choosing the right TSDB solution
  - Fixed Schema vs. Dynamic Schema (Further: SQL vs NoSQL)
  - Single Value vs Multiple Values Per Timestamp (*e.g.,* Structured Data)
  - Indexes (*e.g.,* B+ Tree, Hash Table)
  - Generic vs. Specially Designed Storage (*e.g.,* Hot, Warm, Cold Data)
  - Data Compression (*e.g.,* Timestamp & Value)

- Other considerations
  - User interface: Restful API, SQL querying capability
  - GPU acceleration

Visualizing Forecasting Algorithm Performance



- Hyndman, Wang and Laptev. “Large scale unusual time series detection” (2015).
- Kang, Hyndman & Smith-Miles. “Visualising forecasting algorithm performance using time series instance spaces” (2017).
- Talagala, Hyndman and Athanasopoulos. “Meta-learning how to forecast time series” (2018)

Understanding the Uncertainty in Forecasting

	KNOWN	UNKNOWN
KNOWN	I. Known/Knowns (Normal conditions, Law of large numbers, independent events, wisdom of the crowds) Forecasting: Accuracy measurable Uncertainty: Thin tailed and measurable Risks: Manageable (e.g. having inventories)	III. Unknown/Knowns (Cognitive biases, Strategic actions, self-fulfilling and self-defeating prophesies, game theory) Forecasting: Purely Judgmental Uncertainty: Extensive/hard to measure Risk: Depends on biases, strategic actions/reactions
UNKNOWN	II. Known/Unknowns (Special settings, effects of the next recession on economy/firms, madness of crowds) Forecasting: Inaccuracy can vary considerably Uncertainty: Fat tailed, hard to measure Risks: Can be substantial, tough to manage	IV. Unknown/Unknowns (Black Swans) (Black Swans: Low probability high impact events, e.g. implications of the total collapse of global trade) Forecasting: Impossible Uncertainty: Infinite Risks: Unmanageable, need for antifragile strategies

Reference: Spyros Makridakis, The Contributions of the M4 Competition to the Theory and Practice of Forecasting