

INTRO TO DATA SCIENCE

LECTURE 17: TIME SERIES AND DATA STREAMS

Paul Burkard

08/21/2015

LAST TIME:

- NETWORK ANALYSIS**
- NETWORK STATICS**
- NETWORK DYNAMICS**

QUESTIONS?

I. TIME SERIES MODELING

HANDS-ON: TIME SERIES

II. DATA STREAM MINING

HANDS-ON: DATA STREAMS

- ▶ What is time series modeling?
 - ▶ How do we do it?
 - ▶ When do we need it?
- ▶ What is data stream mining?
 - ▶ How do we do it?
 - ▶ When do we need it?

I. TIME SERIES MODELING

*Q: What is a **time series**?*

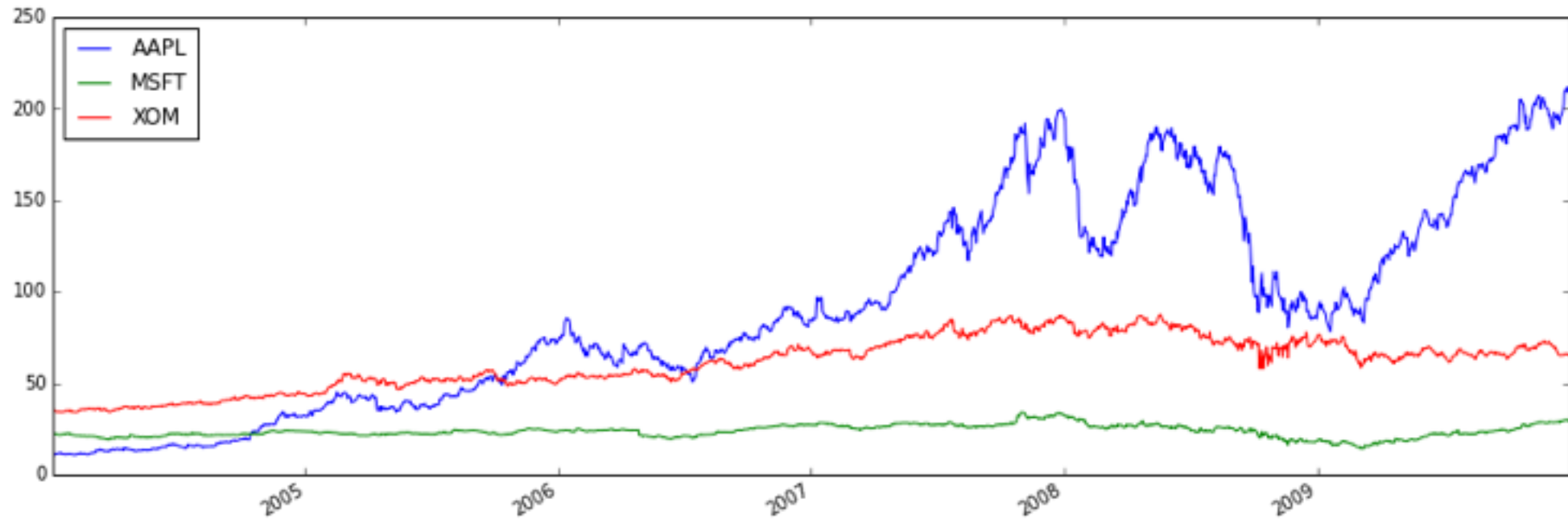
A: A sequence of datapoints where each has an associated timestamp.

*Q: What is can we do with **time series analysis**?*

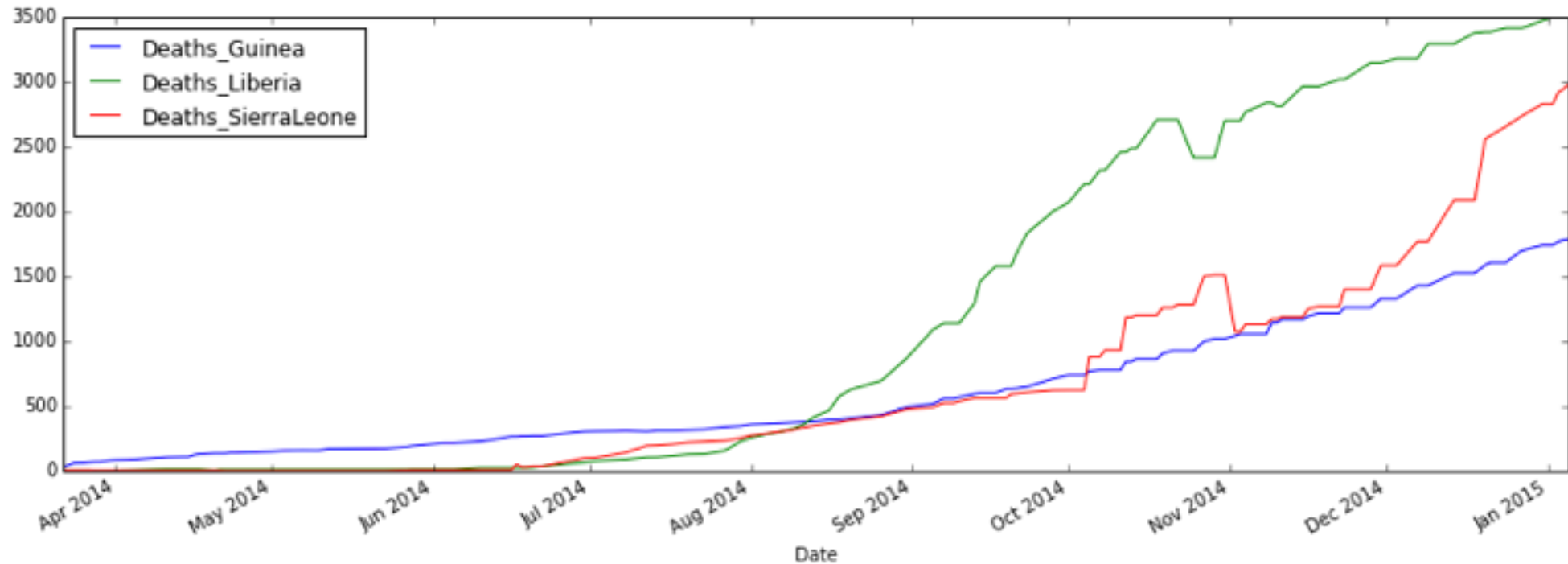
A:

- **Describe:** *extract meaningful statistics and characteristics from the time series*
- **Forecast:** *use a predictive model to predict future values of the time series from previous values (previous values are features!)*
- **Classify:** *use time series characteristics to predict an underlying state of affairs*

Finance



Disease Spread



What these plots have in common is that the x -axis is time.

Things to observe in a time series:

- *Trends – increasing or decreasing over time?*
- *Periodicity – are there observable cycles?*
- *Relationships – correlations to other time series?*

*Q: How might we perform time series **forecasting**?*

*A: **Autoregressive modeling***

Q: What are autoregressive models?

A: Regression models that use previous observations of the target variable as features to predict the target variable at the current time.

The Autoregressive Model:

$$X_t = \sum_{i=1}^p \theta_i X_{t-i}$$

We solve for the values of the thetas just like we would for any normal regression model.

INTRO TO DATA SCIENCE

HANDS-ON: TIME SERIES

II. DATA STREAM MINING

*Q: What is **data stream mining**?*

*A: Extracting knowledge from **continuous, rapidly streaming** data sources.*

Can you think of any such questions that might be interesting?

Q: When might we need data stream mining?

A: When we need incoming data to be incorporated into a ML model in (near) real-time.

*This is called the **data horizon**, aka how real-time does our training need to be?*

Q: When might we need data stream mining?

A: When data is evolving rapidly aka previous training data becomes quickly obsolete

*This is called **data obsolescence***

Q: When might we need data stream mining?

A: When we can only retain a limited set of the data in computational memory.

Q: How can we accomplish data stream mining?

A: 2 general approaches:

- ***Incremental Algorithms***
- ***Periodic Batch Retraining***

Q: What are incremental algorithms?

A: ML algorithms that don't need to retrain on all of the data at once to maintain the model.

*They can simply update themselves **incrementally** based on the delta data.*

Pros:

- *Simplicity (in terms of data management)*
- *Speed*

Cons:

- *Often sacrifice power and flexibility in your model in exchange for incremental procedure*

Our last option is to periodically fully retrain our model once a given batch size of data has come in.

Example: LSI Index with documents streaming in

As documents come in, they are simply folded into the index without retraining the whole index. Once a week, the entire index is retrained over all documents.

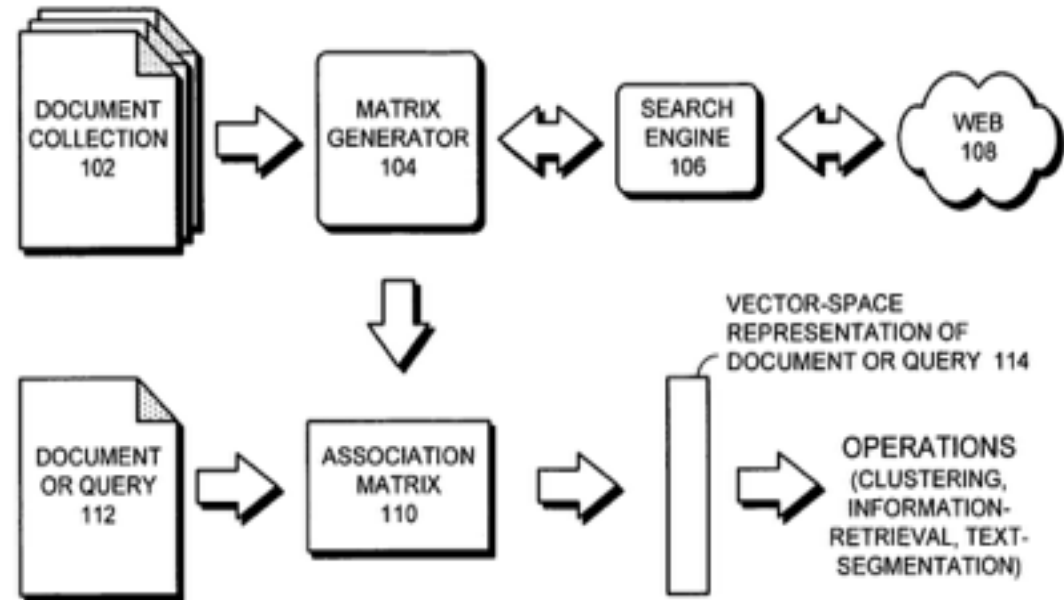


FIG. 1

Pros:

- *Given all the power/flexibility of general ML algorithms*
- *Often able to weight data based on its age in your models*

Cons:

- *More intensive to generate and maintain your model*

Batch methods are probably still preferred from a ML (accuracy/generalizability) perspective if your resources can handle it.

An ongoing research area are systems that automatically detect when a model should be batch retrained (vs. fixed schedule).

INTRO TO DATA SCIENCE

HANDS-ON: DATA STREAM MINING