Tim Menzies
tim@menzies.us
LCSEE, WVU
Nov-31-10

Data Mining for Adaptive Business Intelligence

Summary

- We need to do more "data mining"
  - Not just on different projects
  - But again and again on the same project
- And by "data Mining" we really mean
  - Automated agents that implement
    - prediction
    - monitoring
    - diagnosis,
    - Planning
    - Adaptive business intelligence

This talk

* A plea for industrial partners to aid in this research

The new idea

A proposal

- Add roller skates to science and engineering
- Always use data mining on SE data

Two kinds of teams

- A few data collection teams
  - Collecting case study data
  - Publishing what they can
- Many many more data mining teams
  - Analyzing the data

Who benefits?

Researchers

- Data collection teams get
  - Comparisons of their data to data from other sites
  - Data analysis via crowd sourcing
- Data mining teams get
  - Fuel for their analysis

Industrial Standards Bodies

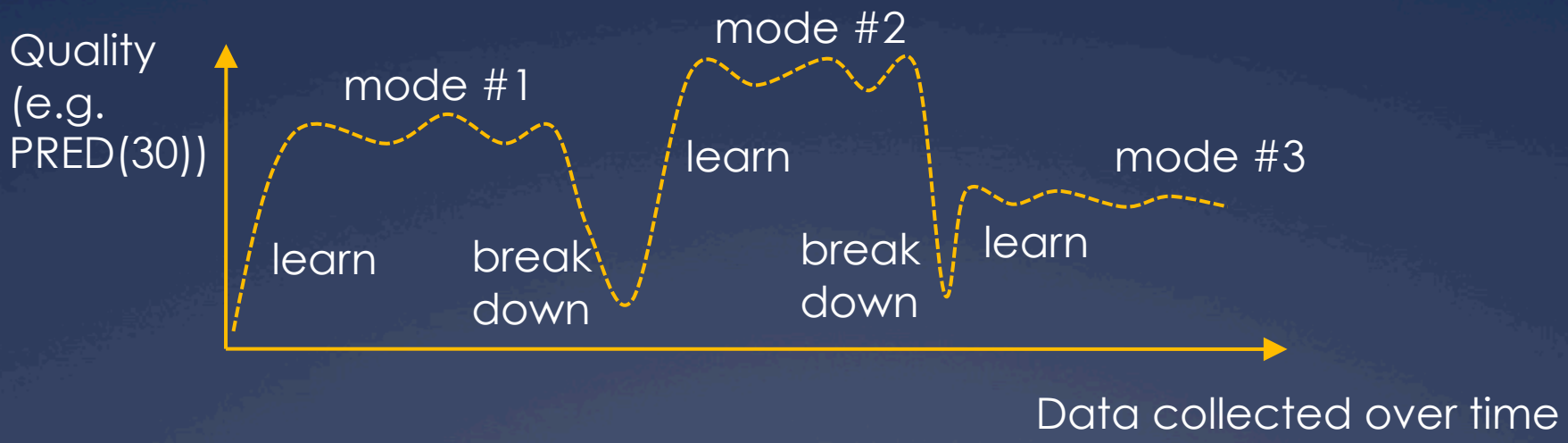
- 20th century SE
  - Prescriptions of how we think it should be
- 21st century SE
  - Descriptions of how it is
  - Massive data collection
  - Comparisons across different sites
  - Recognitions of common patterns
    - Good smells: patterns to be propagated
    - Bad smells: patterns to be avoided

Industrial practitioners

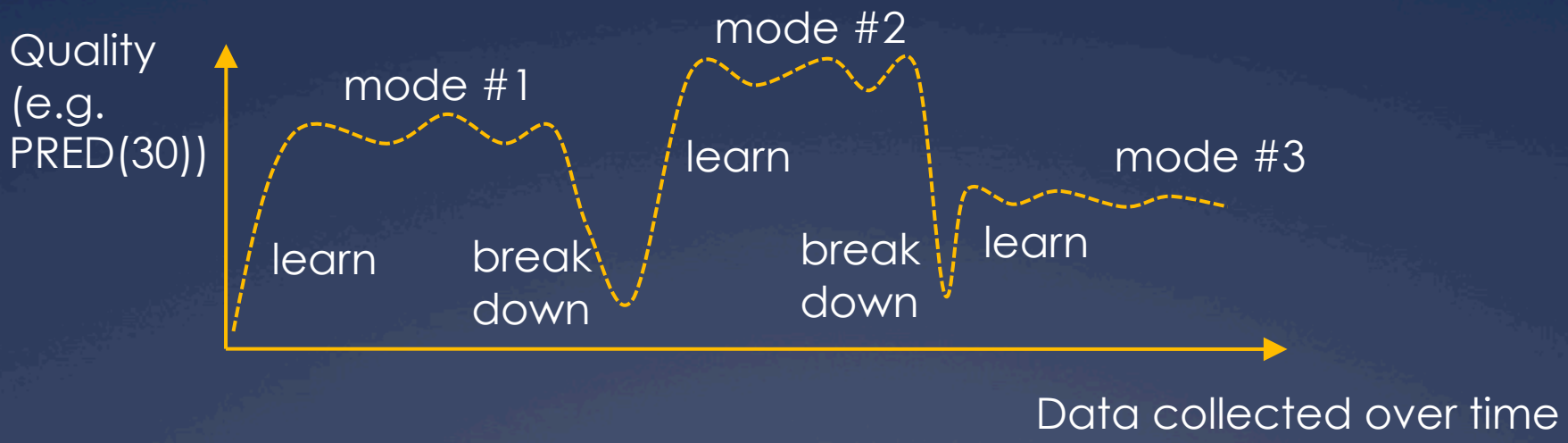
- Analysts-in-a-box
  - Rapid and automated analysis
- Learn best practices
  - Learn when other people's best practices do not apply to you
- Get more data from other sites
  - So, if using a technology that is new to you
  - But has been explored at other site
  - Then you can still determine (say)
    - Expected defect rates
    - Expected development effort

Details

Agents for adaptive business intelligence

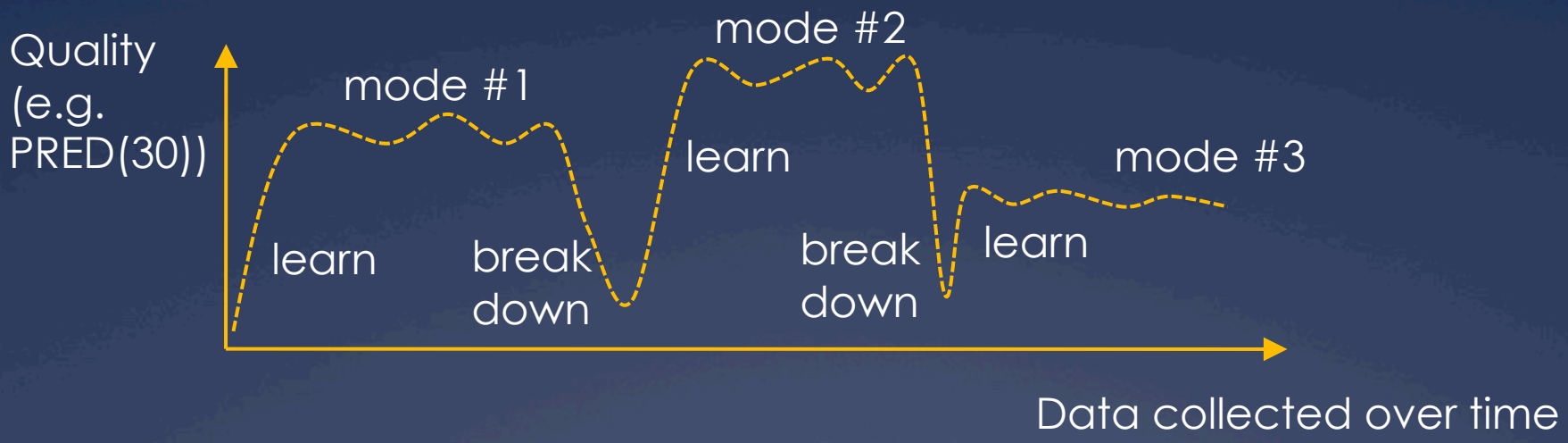


Agents for adaptive business intelligence



- Adaptive business intelligence
  - Via incremental data mining
  - Technology: incremental discretization + incremental clustering and classification

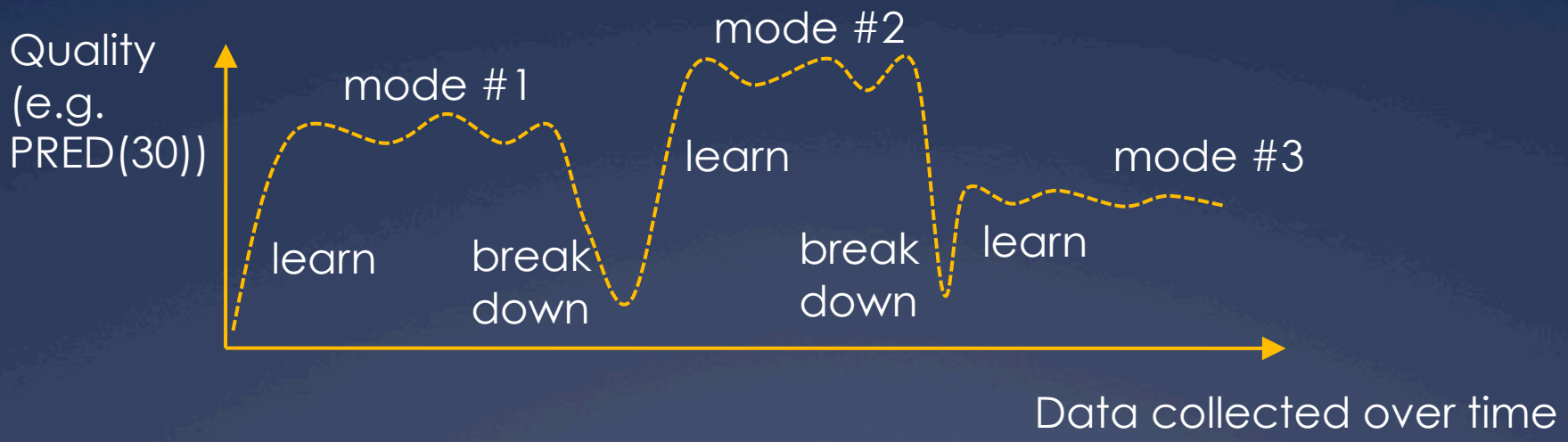
Agents for adaptive business intelligence



- What is different here?
  - Not "apply data mining to build a predictor"
  - But add monitor and repair tools to recognize and handle the breakdown of old predictors
  - Trust = data mining + monitor + repair

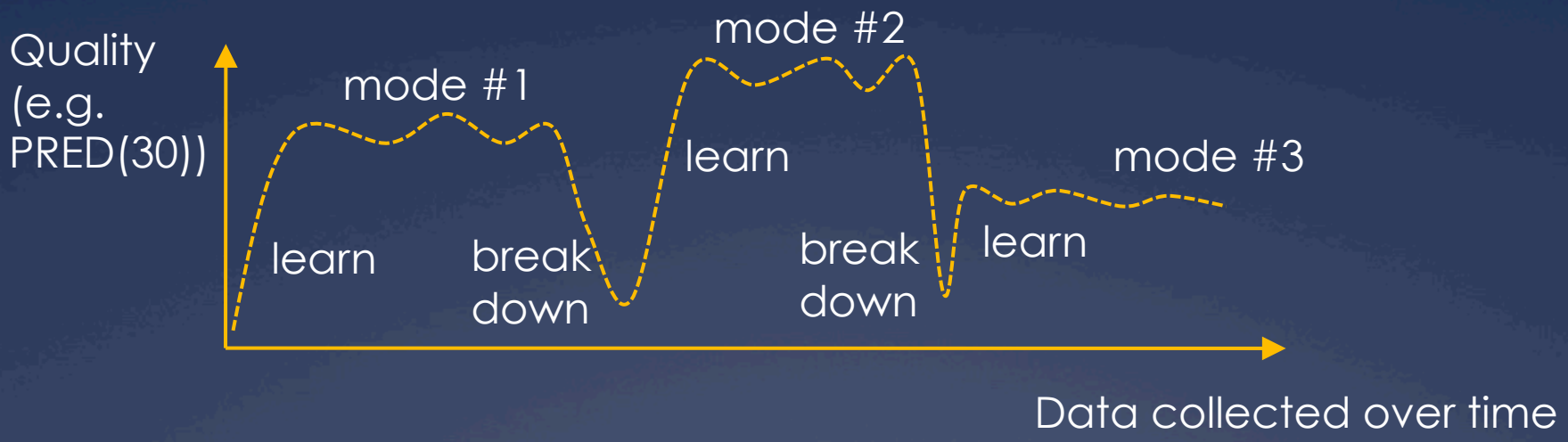
Most of the technology required for this approach can be implemented via data mining

Agents for adaptive business intelligence



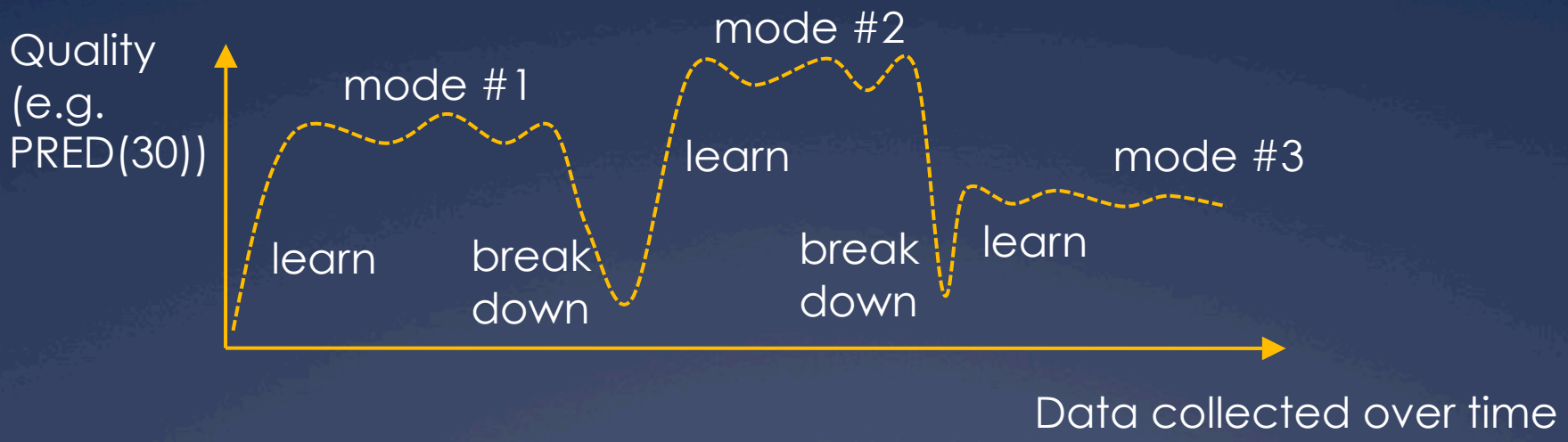
- Q1: How to learn faster?
  - Technology: active learning: reflect on examples to date to ask most informative next question
- Q2: How to recognize breakdown?
  - Technology: bayesian anomaly detection

Agents for adaptive business intelligence



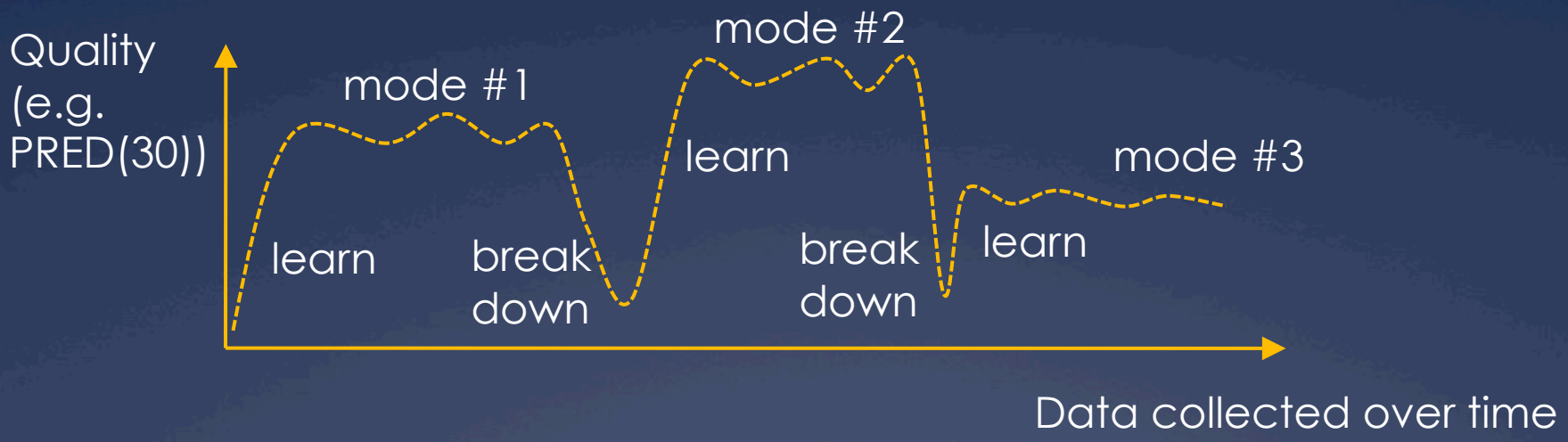
- Q3: How to classify mode?
  - Recognize if you've arrived at a mode seen before
  - Technology: Bayes classifier
- Q4: How to make predictions?
  - Using the norms of a mode, report expected behavior
  - Technology: table look-up of data inside Bayes classifier

Agents for adaptive business intelligence



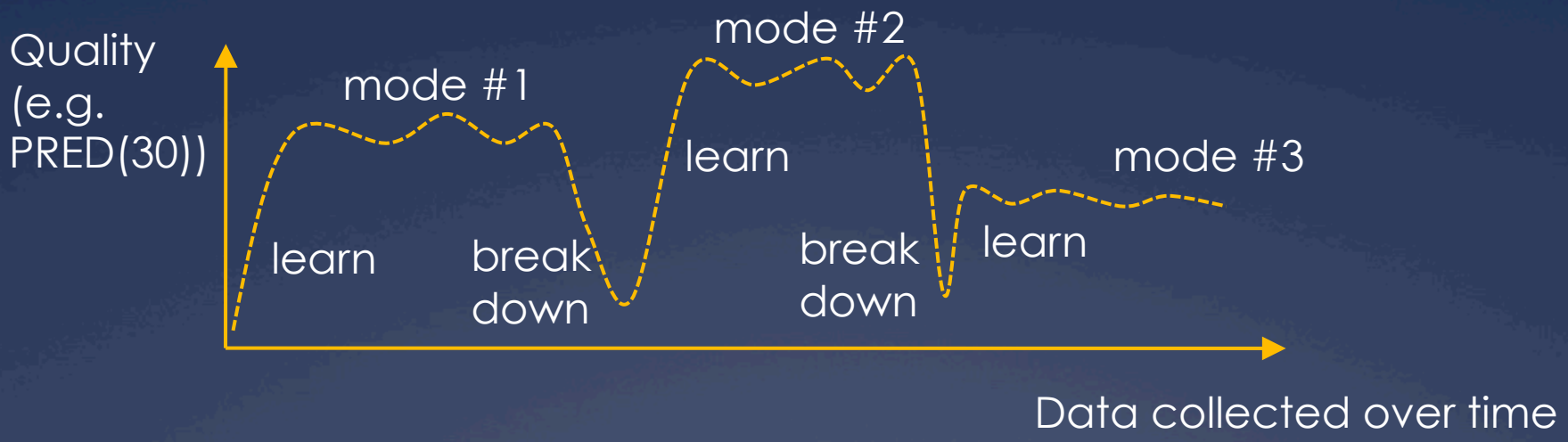
- Q5: What went wrong? (diagnosis)
  - Delta between current and prior, better, mode
- Q6: What to do? (planning)
  - Delta between current and other, better, mode
- Technology: contrast set learning

Agents for adaptive business intelligence



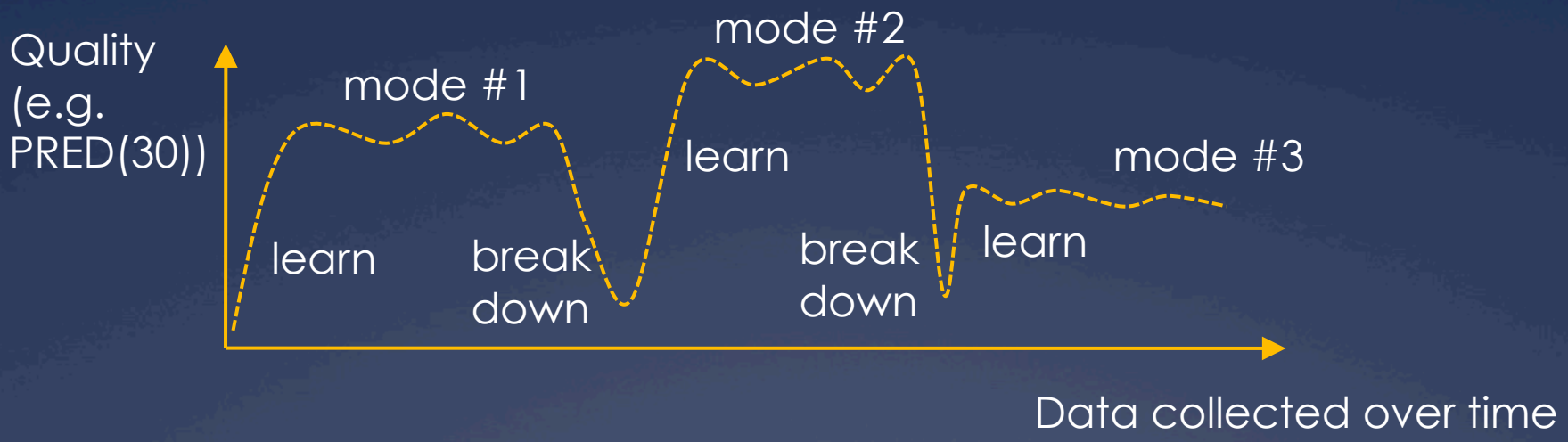
- Q7: How to understand a mode (explanation)
  - Presentation of essential features of a mode
- Technology: Dimensionality reduction, feature selection

Agents for adaptive business intelligence



- Q8: How to start?
- All the above can be based on a matrix P*T:
  - Products (P) described using a set of terms (T)
  - Probably a sparse matrix

Agents for adaptive business intelligence



- Q9: What terms to collect?
- A: Everything
  - Assign all project artifacts and unique ID
  - Place all project artifacts in (say) a wiki
  - Use text mining to generate the P*T matrix
  - Add a join key that maps artifacts to some quality measure (defects, effort, whatever is important and monitored at your site)

Finding the controllables

What are the control points?

Problem

- Most certainly,
  - Our initial data collection will be incomplete
- If we add all products to a wiki
  - We still might miss the process options that can change a project

Solution

- Domain interviews
  - When modes identified
  - Conduct structured interviews with managers …
  - … on the delta of this mode to others…
  - … to indentify the process actions or the market forces that resulted in a mode
- Some of these actions/ forces will be controllable
  - Augment T with these new actions/forces

And so…

We seek industrial partners

1. That will place textual versions of their products in a wiki
2. That will offer joins of those products to quality measures
3. That will suffer us interviewing their managers, from time to time, to learn the control points.

(Note: 1,2 can be behind your firewalls.)

In return, we offer

- Agents for
  - automatic, adaptive, business intelligence
  - that tunes itself to your local domain

Questions?
Comments?