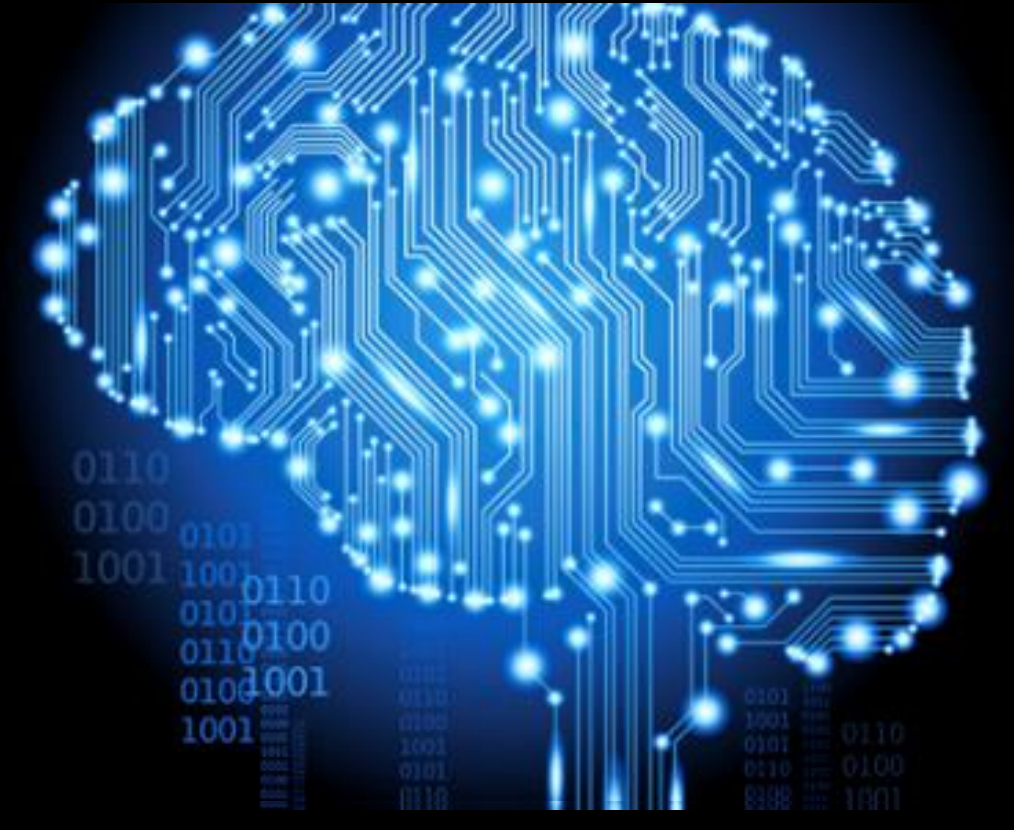
THE DOWNLOAD

TECH TALKS BY HPCC SYSTEMS



HPCC SYSTEMS®

The Download: Community Tech Talks Episode 3

March 30, 2017

RELX Group

Welcome!

- Please share: Let others know you are here with #HPCCTechTalks 
- Ask questions! We will answer as many questions as we can following each speaker.
- Look for polls at the bottom of your screen. Exit full-screen mode or refresh your screen if you don't see them.
- We welcome your feedback - please rate us before you leave today and visit our blog for information after the event.
- Want to be one of our featured speakers? Let us know! techtalks@hpccsystems.com

HPCC SYSTEMS®

Community Announcements

- Still time for students to submit a proposal for the HPCC Systems Summer Internship Program!
  - Deadline extended to April 22
  - Details at https://hpccsystems.com/intern2017

- Call for Presentations and Poster Abstracts now open for the 2017 HPCC Systems Community Day!
  - Community Day will be held in Atlanta on October 3, 2017
  - Submission deadline on June 30
  - NEW! Sponsorship opportunities available to allow our partners to have a bigger presence
  - Details at https://hpccsystems.com/hpccsummit2017

- Machine Learning Update
  - Two new bundles coming in 6.4.0 planned for mid-year release
  - Linear Regression
  - Logistic Regression
  - Check for the first release candidate in May.

Dr. Flavio Villanustre
VP Technology
LexisNexis® Risk Solutions
Flavio.Villanustre@lexisnexis.com

HPCC SYSTEMS®

Today's Speakers

Joselito (Joey) Chua , PhD

Manager Software Engineer
Optimal Decisions Group
LexisNexis Risk Solutions
Joey.chua@lexisnexisrisk.com

Joselito (Joey) Chua leads the software engineering team in the Optimal Decisions Group in LexisNexis Risk Solutions. He specializes in information-theoretic approaches to machine learning. He is a fan of anime.

Jill Luber

Senior Architect,
LexisNexis® Risk Solutions
jill.luber@lexisnexisrisk.com

Jill Luber is a Senior Architect for LexisNexis Risk Solutions with leadership responsibility for strategy, implementation, and stability of all US and international data linking products, including the LexID, Business LexID, Healthcare Provider ID, and UK LexID. With 17 engineers across multiple geographies, the Linking Team develops the core competences underpinning all products at Risk executing on the HPCC big data platform. Jill has presented patented, Risk linking concepts at the 2015 RELX leadership conference as well as the RELX Board meeting. She has been a member of the technology organization for over 13 years.

HPCC SYSTEMS®

Today's Speakers

Michael Gardner

Software Engineer II,
LexisNexis® Risk Solutions
michael.gardner@lexisnexisrisk.com

Michael Gardner is a HPCC Platform team member and developer. He is responsible for the HPCC Systems Platform init system, various build issues, administrative scripts, and HPCC Java projects. His most recent active work includes systemd integration for the HPCC Systems Platform, and an antlr3 (c) to antlr4 (cpp) migration for the wssql project.

Bob Foreman

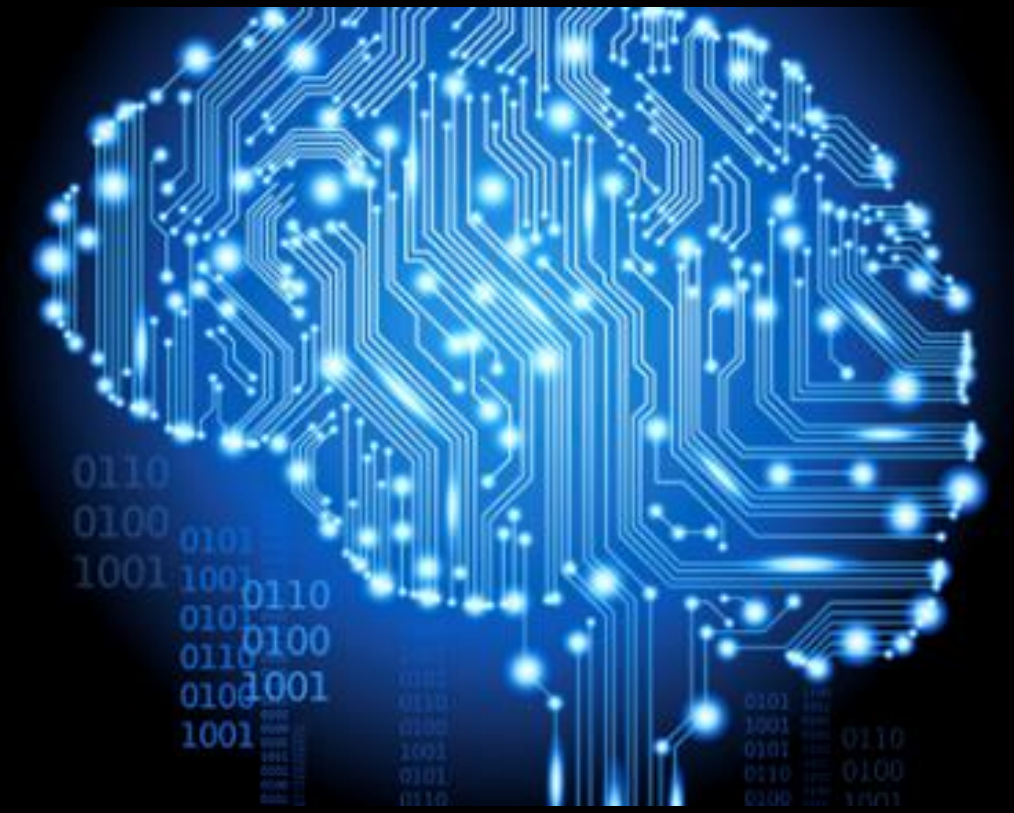
Senior Software Engineer
LexisNexis® Risk Solutions
bob.foreman@lexisnexisrisk.com

Bob Foreman has worked with the HPCC Systems technology platform and the ECL programming language for over 5 years, and has been a technical trainer for over 25 years. He is the developer and designer of the HPCC Systems Online Training Courses, and is the Senior Instructor for all classroom and Webex/Lync based training. This includes: Introduction to ECL - Concepts and Queries, Introduction to THOR - the Extract, Transform and Load (ETL) Process, Advanced ECL - Working with Relational Data, Advanced THOR - Super files, Working with XML, and Free-form Text Parsing, Introduction to ROXIE - Indexes and Queries, Advanced ROXIE - Complex Query Development, and Applied ECL – ECL Code Generation Tools.

HPCC SYSTEMS®

THE DOWNLOAD

TECH TALKS BY HPCC SYSTEMS



HPCC SYSTEMS®

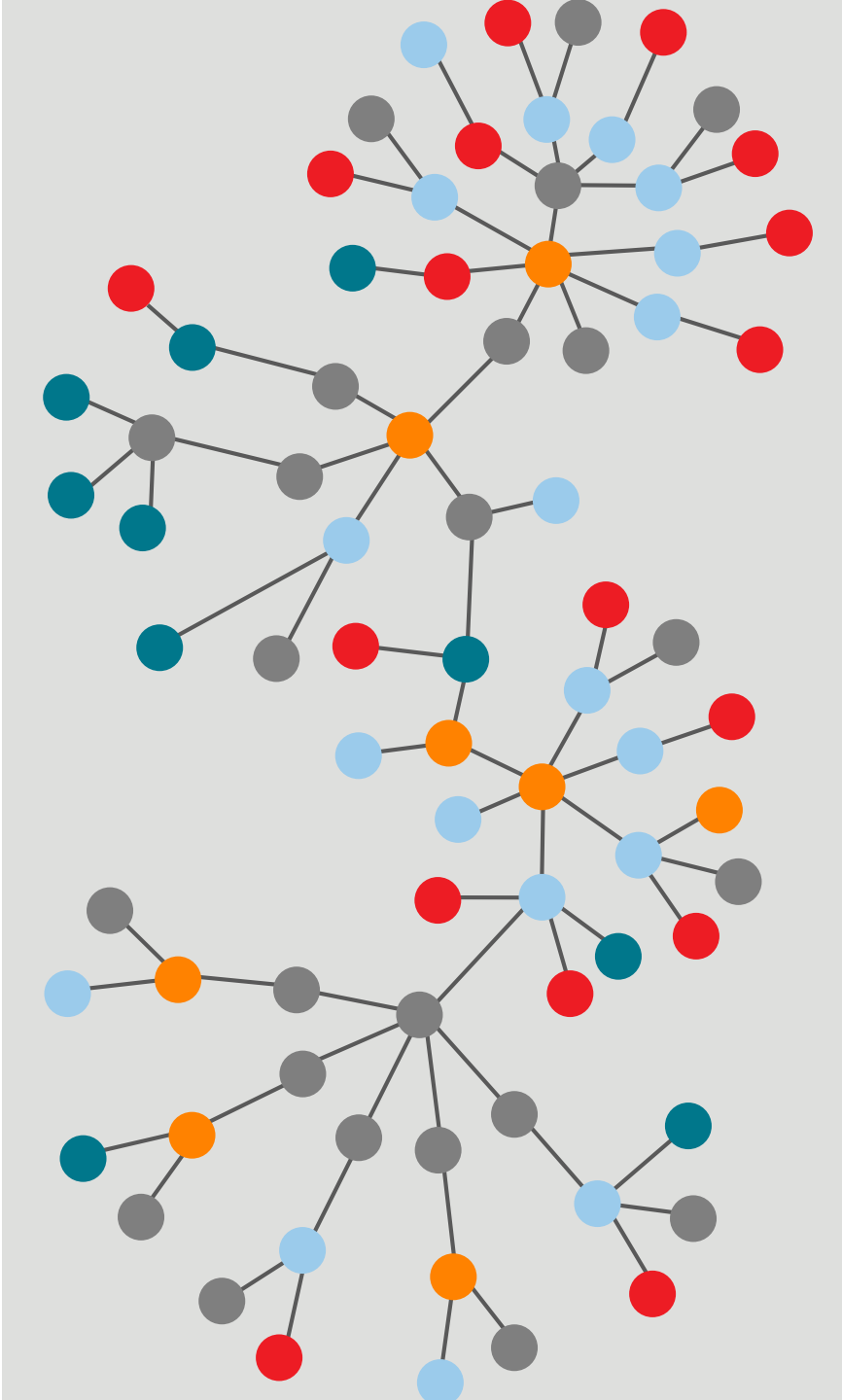
Prescriptive Analytics - a Software Engineering Perspective

Joselito (Joey) Chua , PhD
Manager Software Engineer
Optimal Decisions Group
LexisNexis Risk Solutions

LexisNexis® RISK SOLUTIONS

Quick poll: Would you use a driverless car?

See poll on bottom of presentation screen



HPCC SYSTEMS®

The ultimate goal of data analytics is to **improve outcomes** using insights from data.

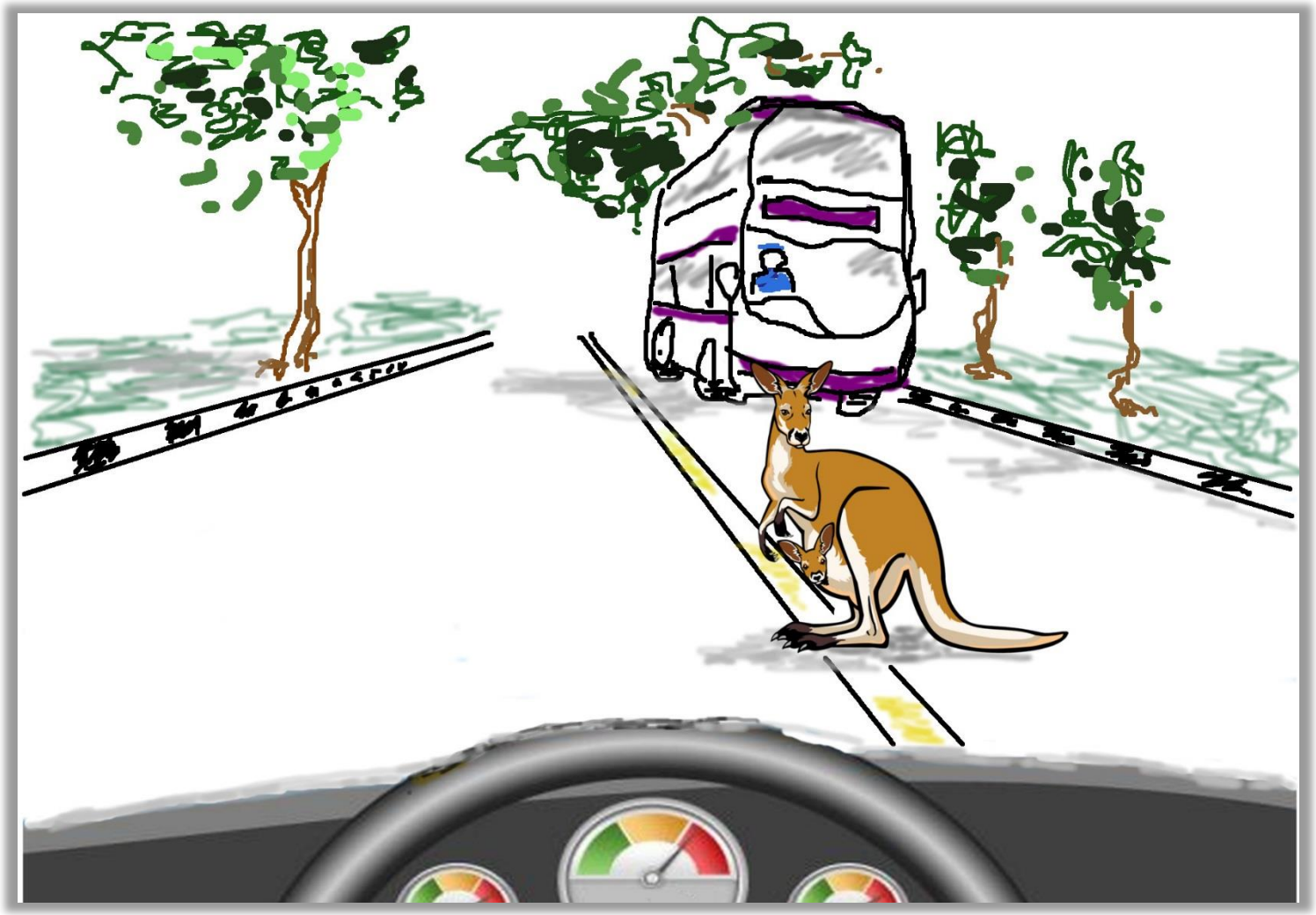
Curate	Collate	Describe	Predict	**Prescribe**
• Collect, manage, clean, share data	• Enrich, link, summarize	• What happened and why?	• What will happen and why?	• What should I do and why?

From Data to Insights

From Insights to **Decisions**

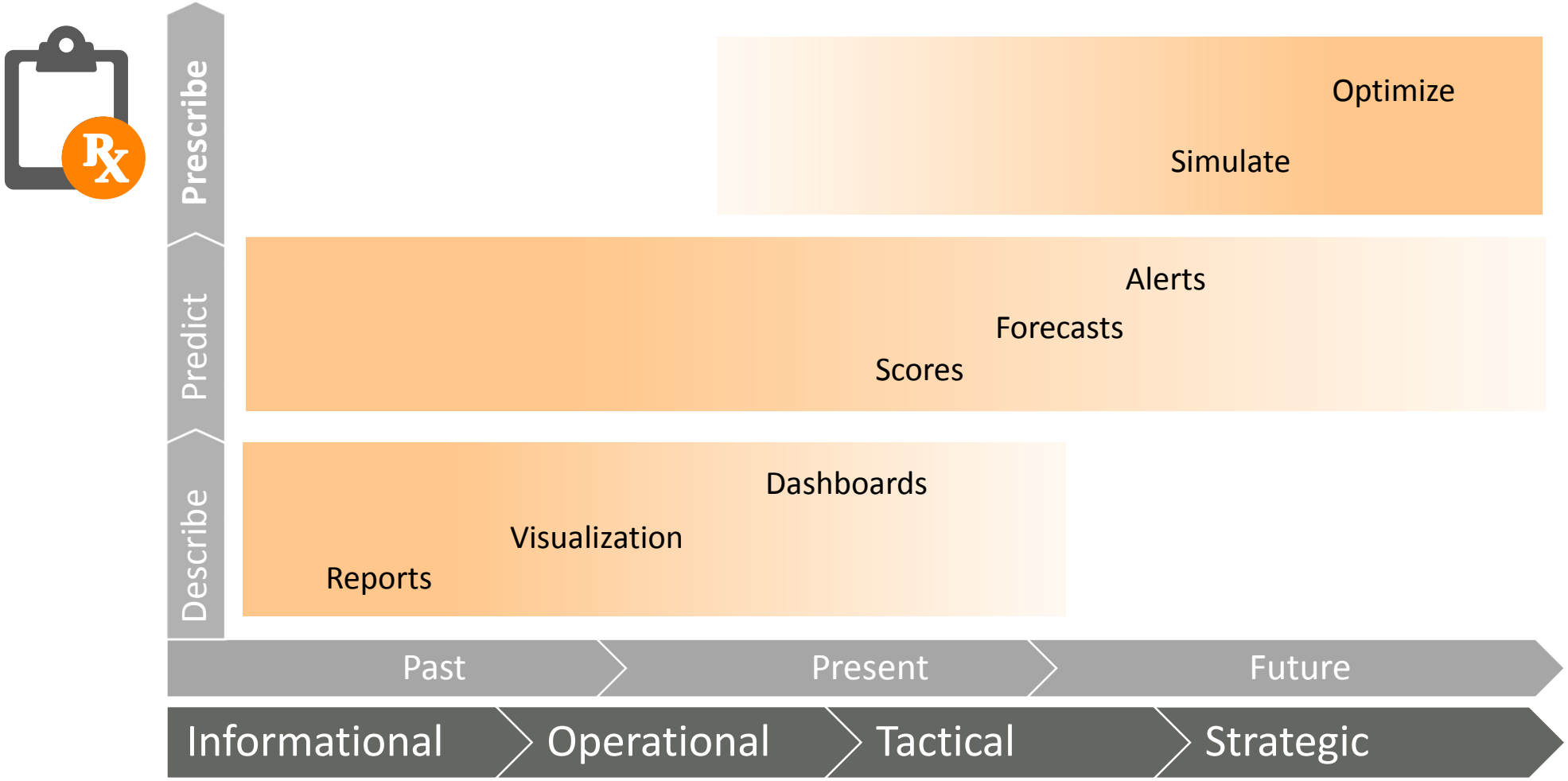
HPCC SYSTEMS®

Prescriptive tools are software solutions that turn insights into **decisions**.



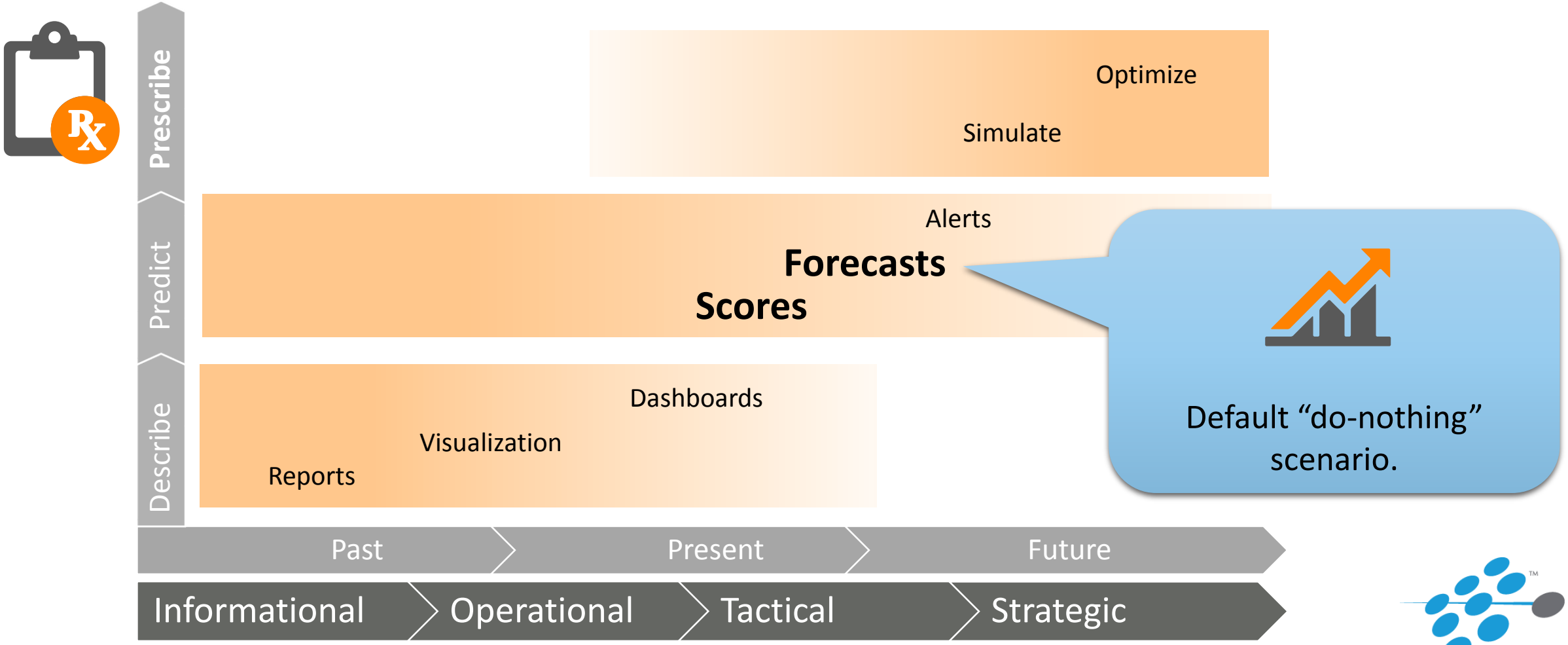
HPCC SYSTEMS®

Prescriptive tools are key components in realising the **value proposition** of data analytics and business intelligence solutions.



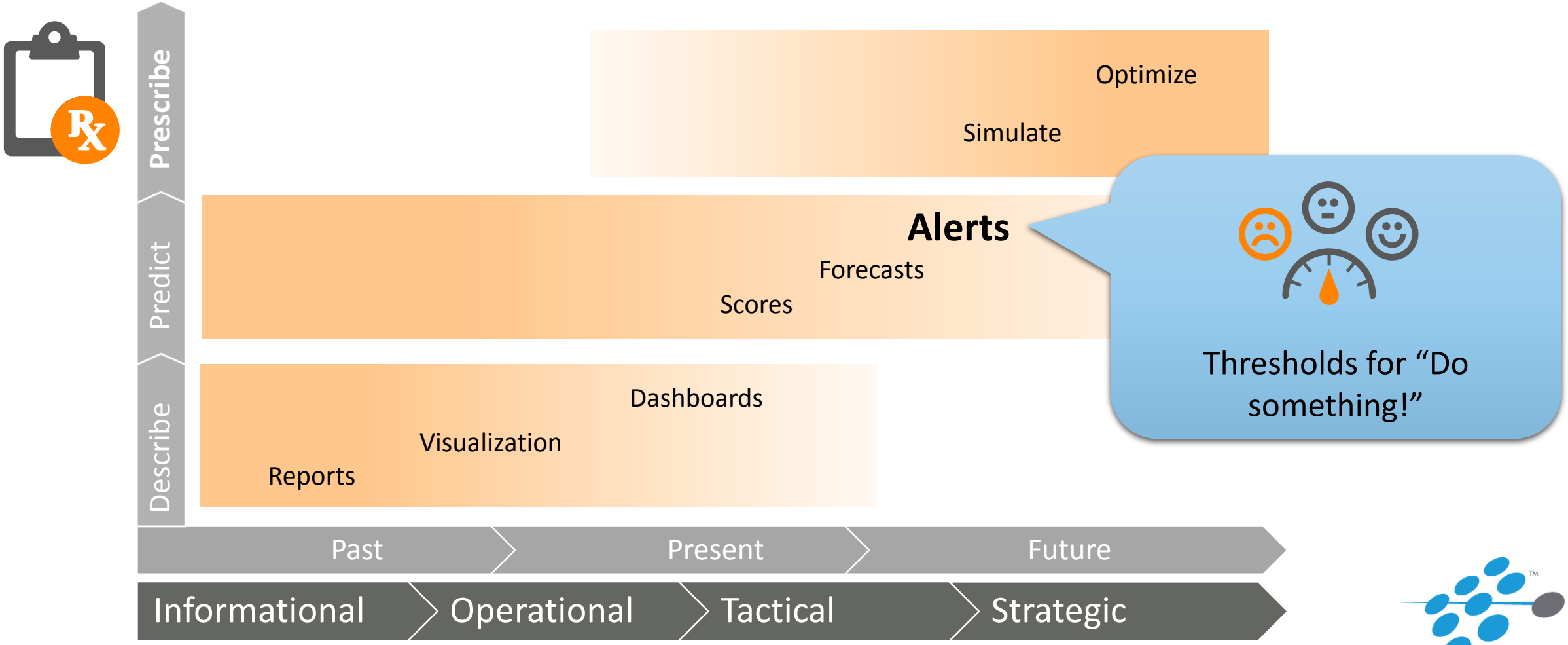
HPCC SYSTEMS®

Scores and forecasts relate descriptive attributes to future states or outcomes in a **business-as-usual** scenario.



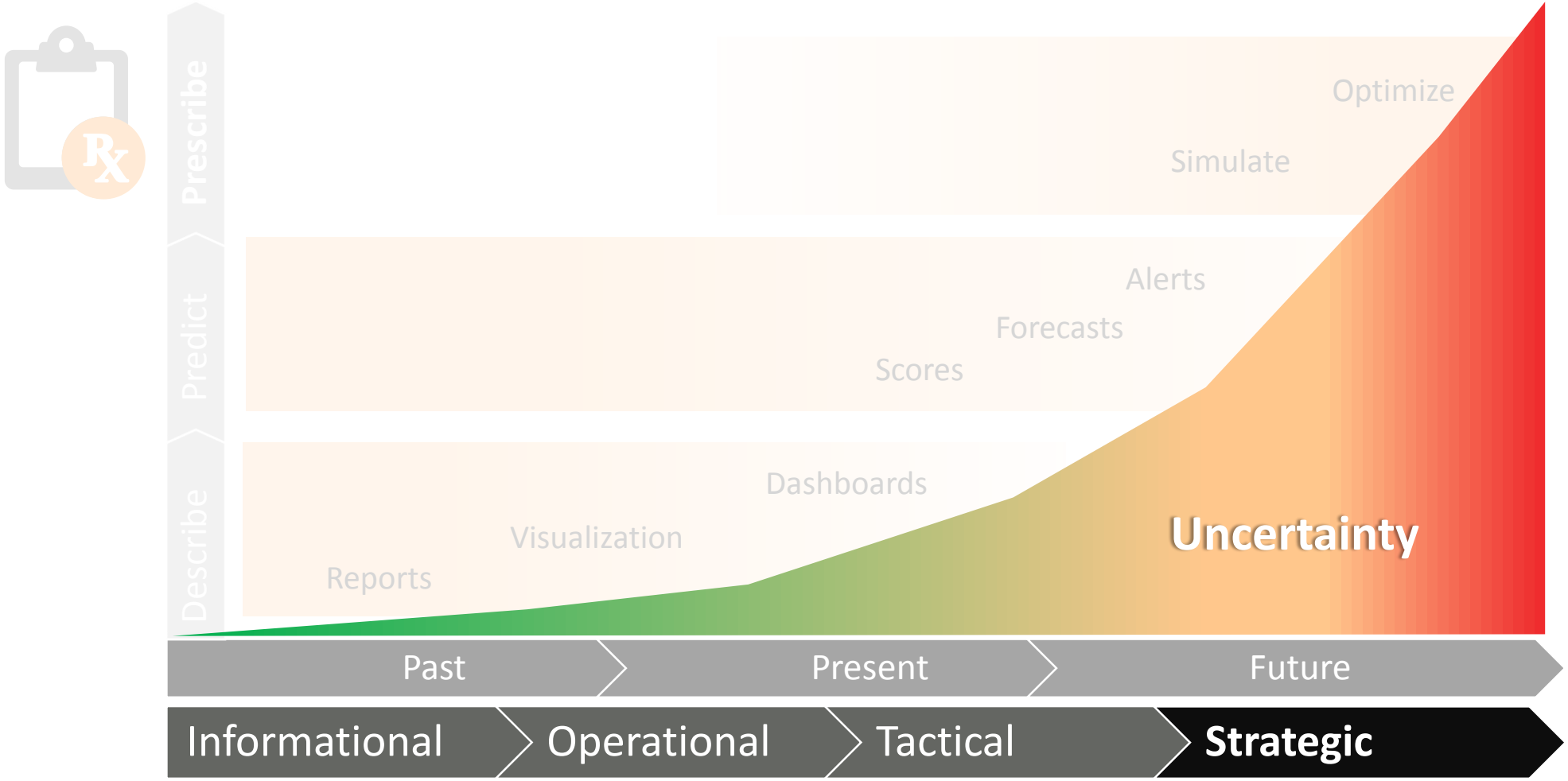
HPCC SYSTEMS®

Alerts indicate that actions are required but without prescribing what actions to take.



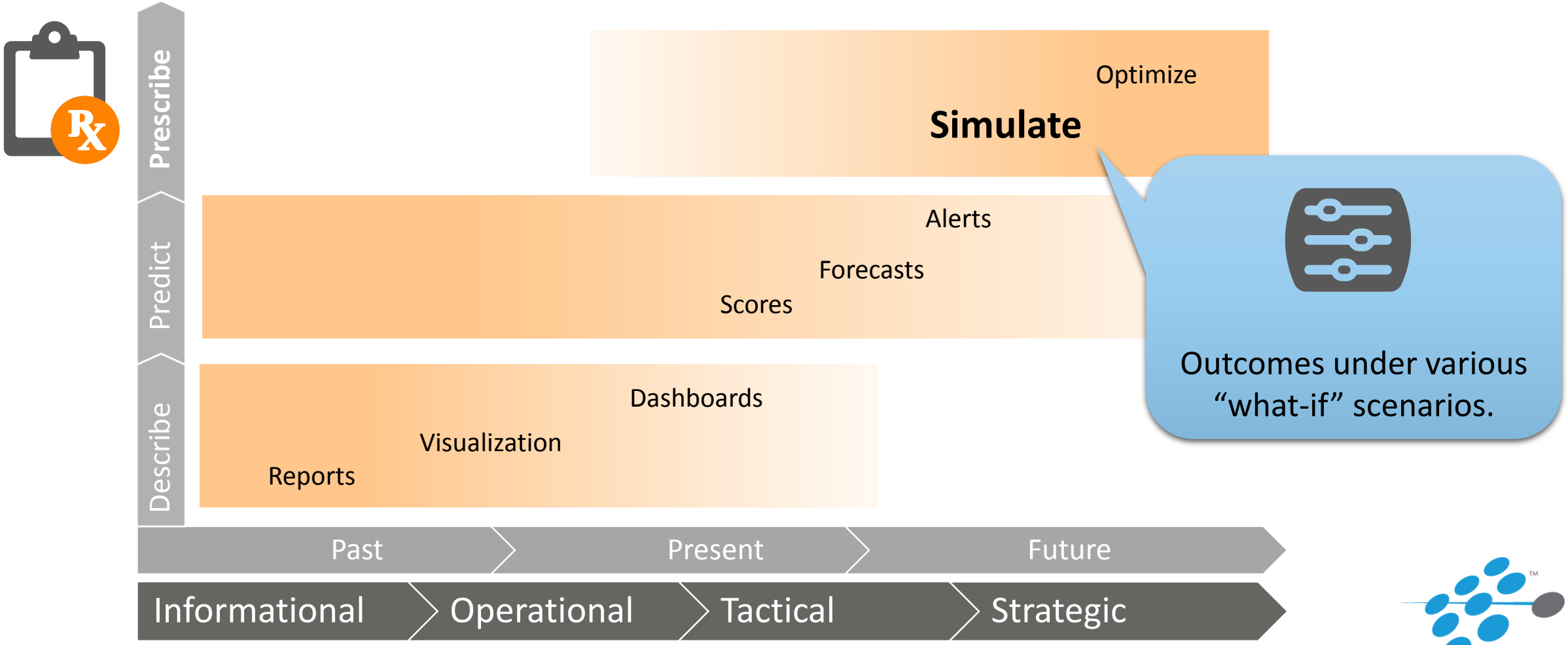
HPCC SYSTEMS®

Uncertainty increases in forward projections as options and constraints change, and more assumptions have to be made.



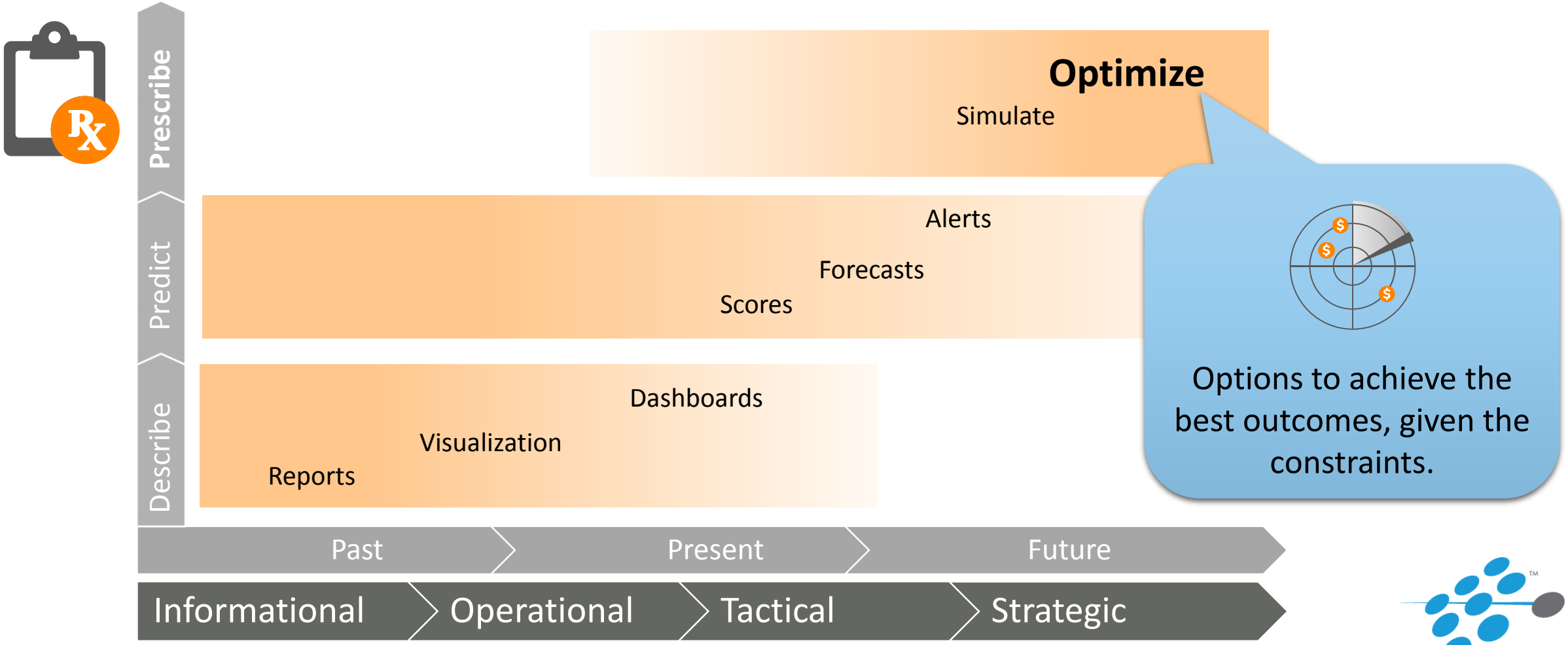
HPCC SYSTEMS®

Simulations predict outcomes under various combinations of **actions and assumptions** when uncertainty is high.



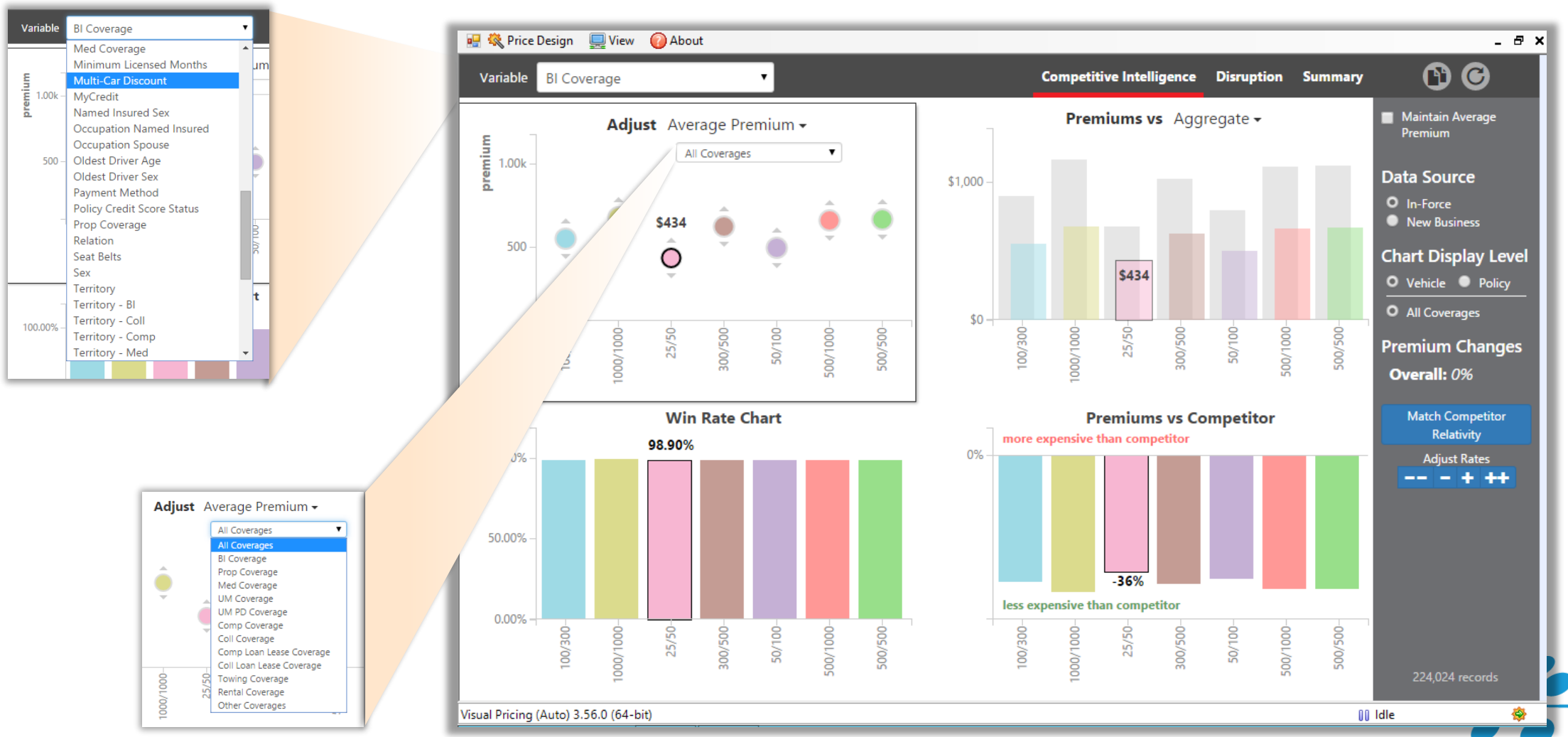
HPCC SYSTEMS®

Optimization relates the actions to the **business objectives and constraints**, sifting through the options towards the best outcomes.



HPCC SYSTEMS®

Example: Price designers need to assess the impacts of many possible combinations of adjustments to “pricing levers”.

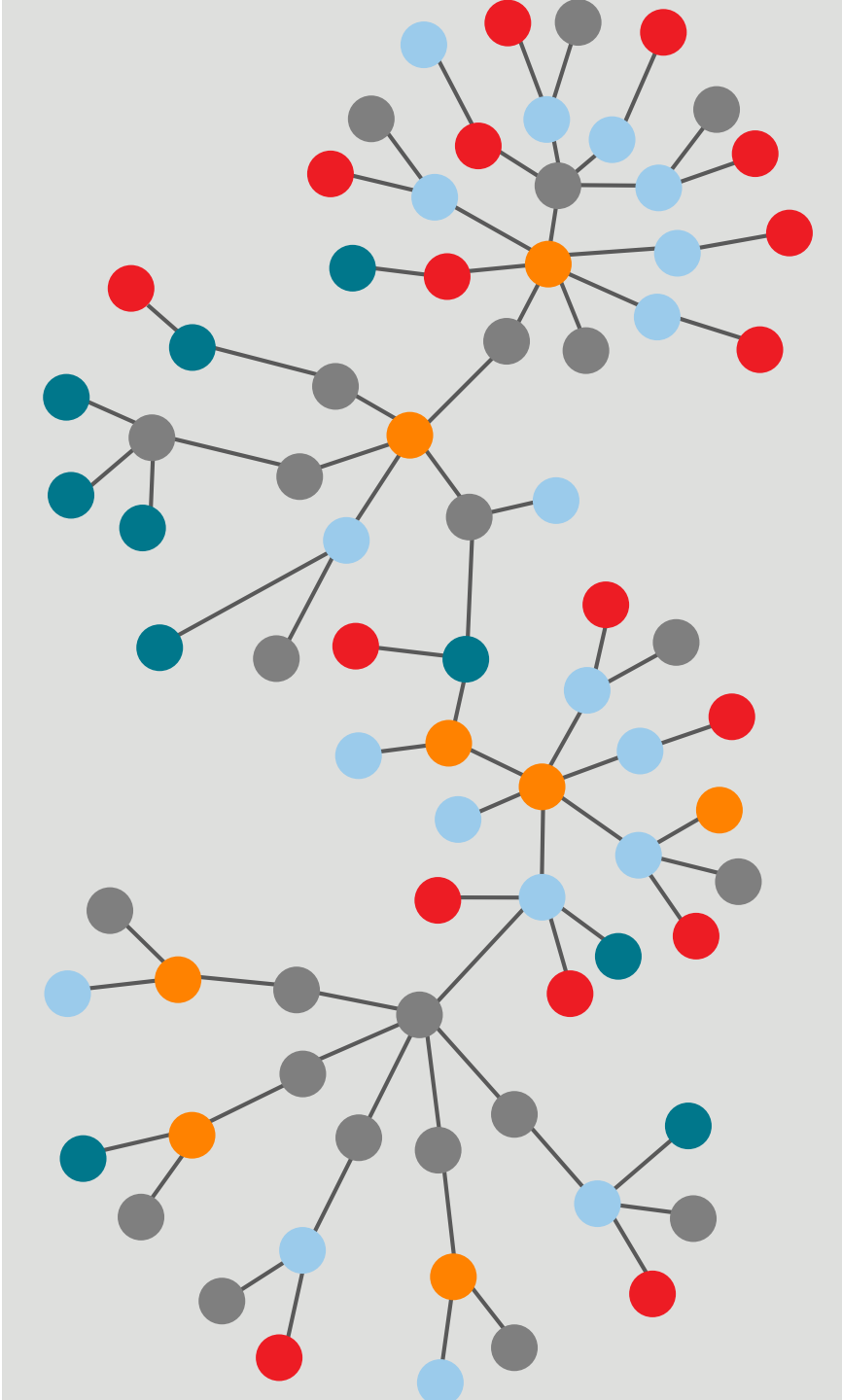


HPCC SYSTEMS®

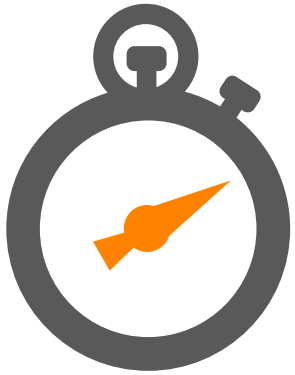
Quick poll: Do you work with prescriptive tools in your current role?

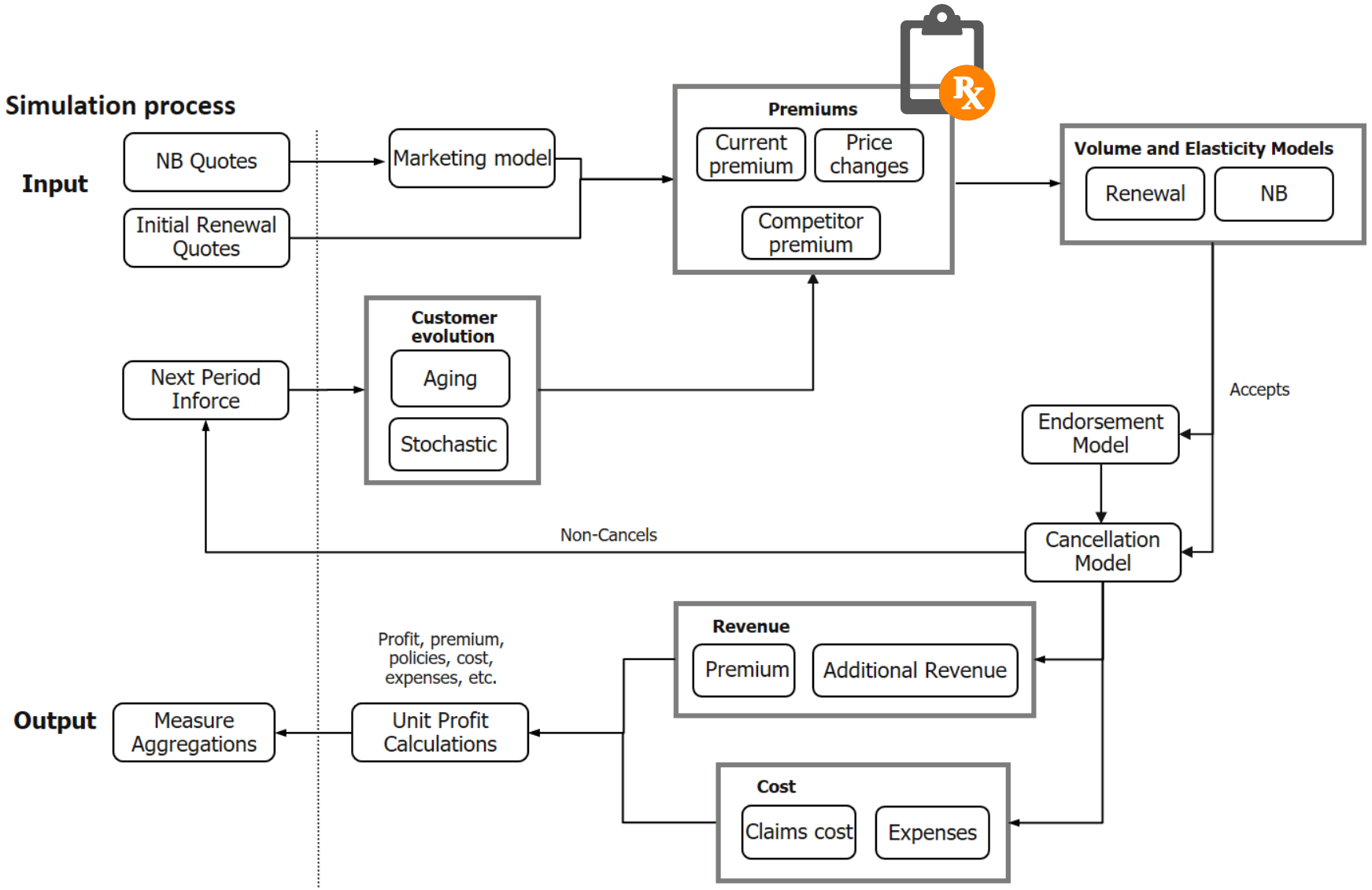
See poll on bottom of presentation screen

HPCC SYSTEMS®



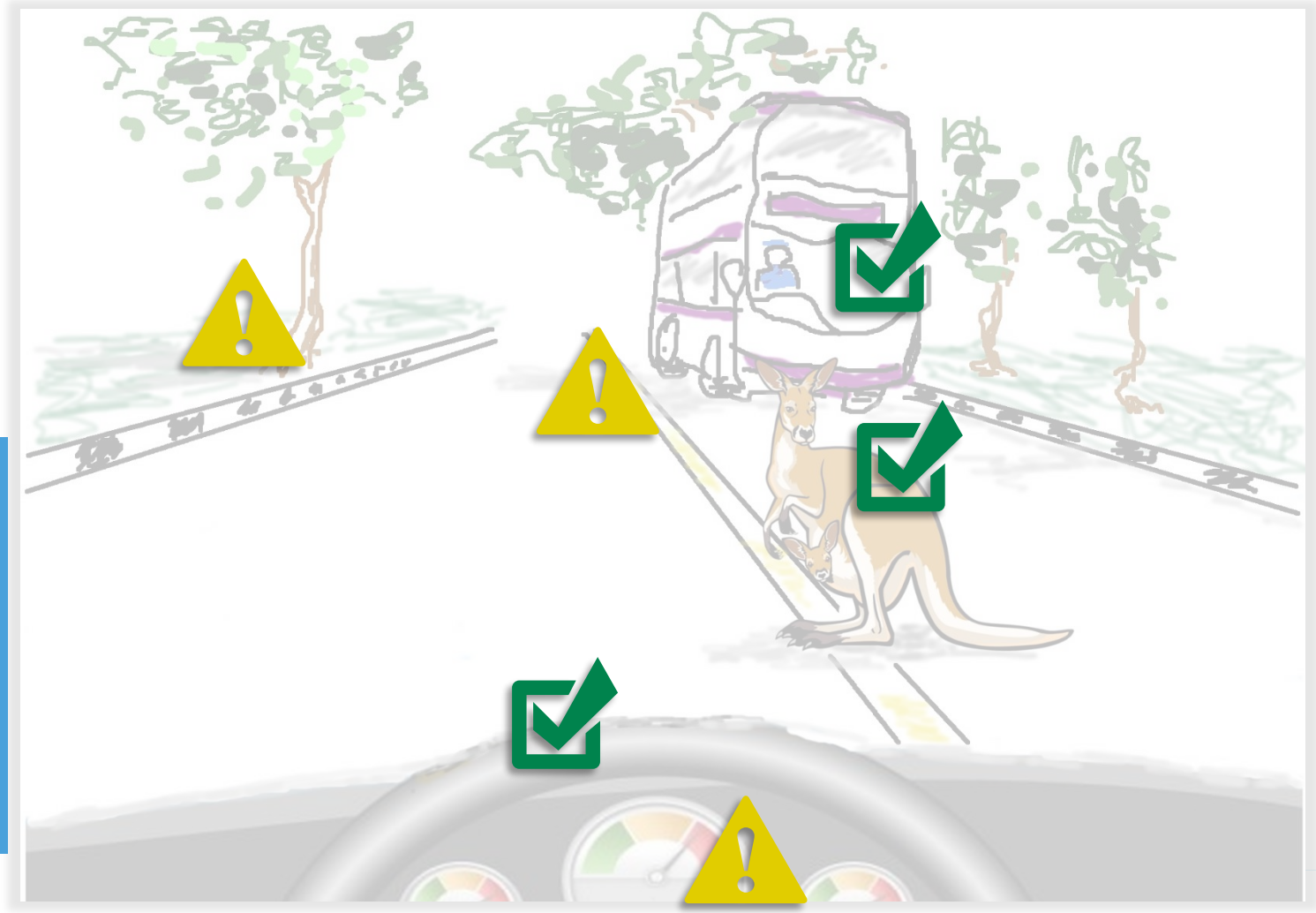
Prescriptive tools have several characteristics and elements that present software engineering **challenges**.





The desired outcomes often require fast decisions involving **multiple competing** objectives and constraints.

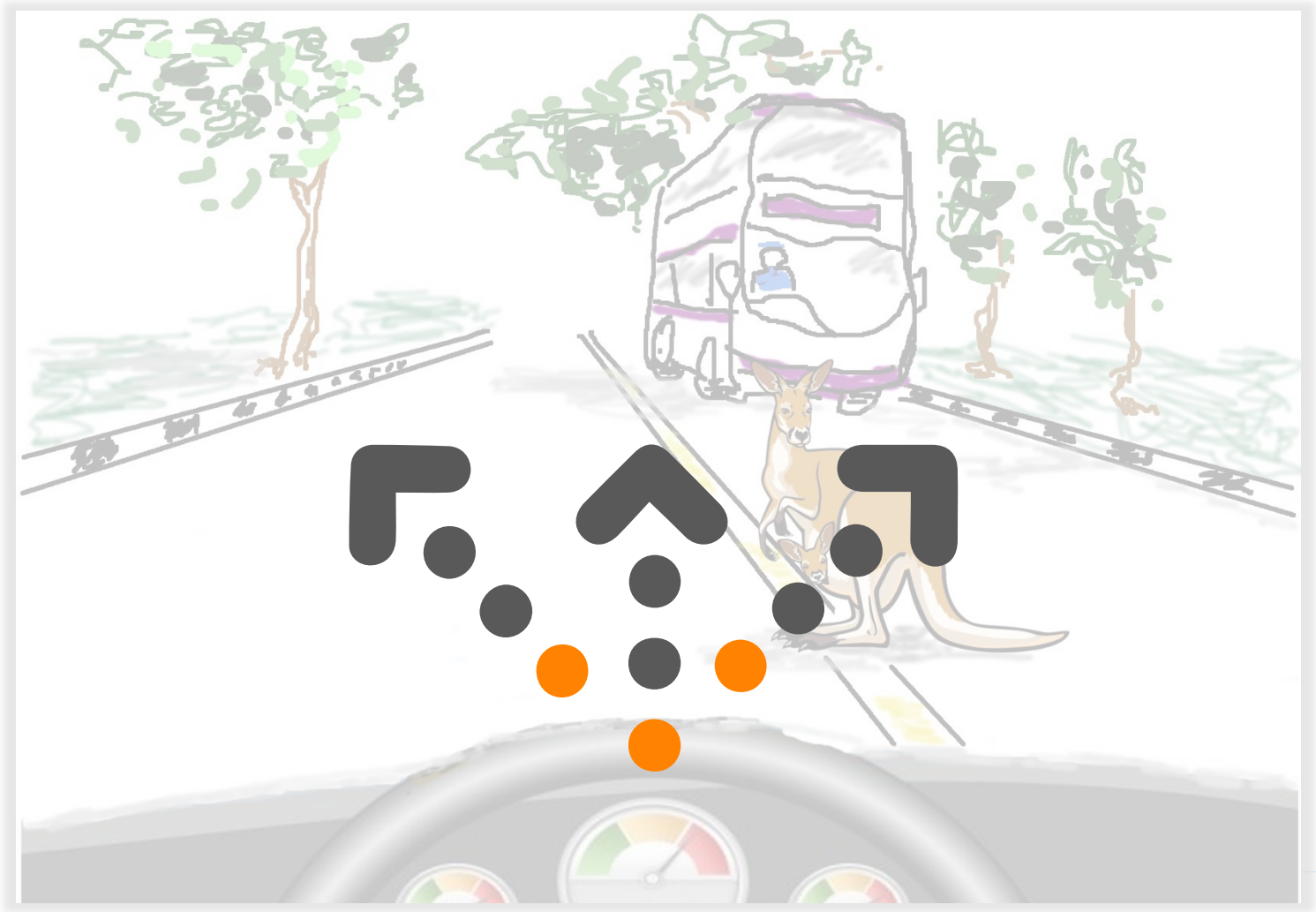
- ✓ Time-sensitive
- ✓ Non-unique solutions
- ✓ UI for representing objectives and constraints
- ✓ “Black-box” reputation
- ✓ Interrogate intermediate calculations



HPCC SYSTEMS®

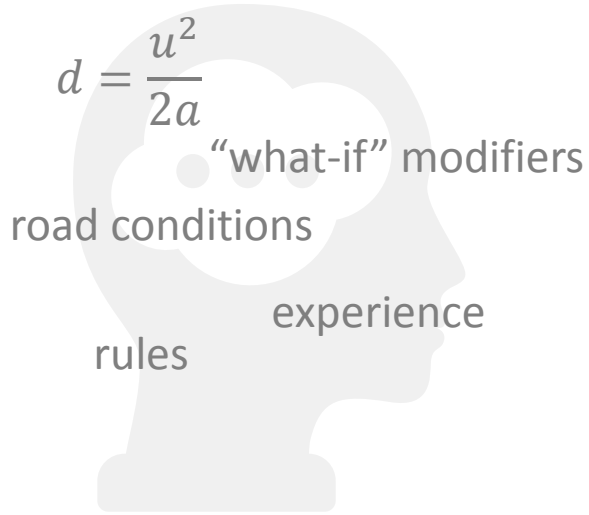
The number of **combinations** of possible actions are very large, where small perturbations can lead to very different outcomes.

- ✓ Fast algorithms
- ✓ Numerical stability
- ✓ Parallel calculations
- ✓ Local vs. global optima
- ✓ Visualization

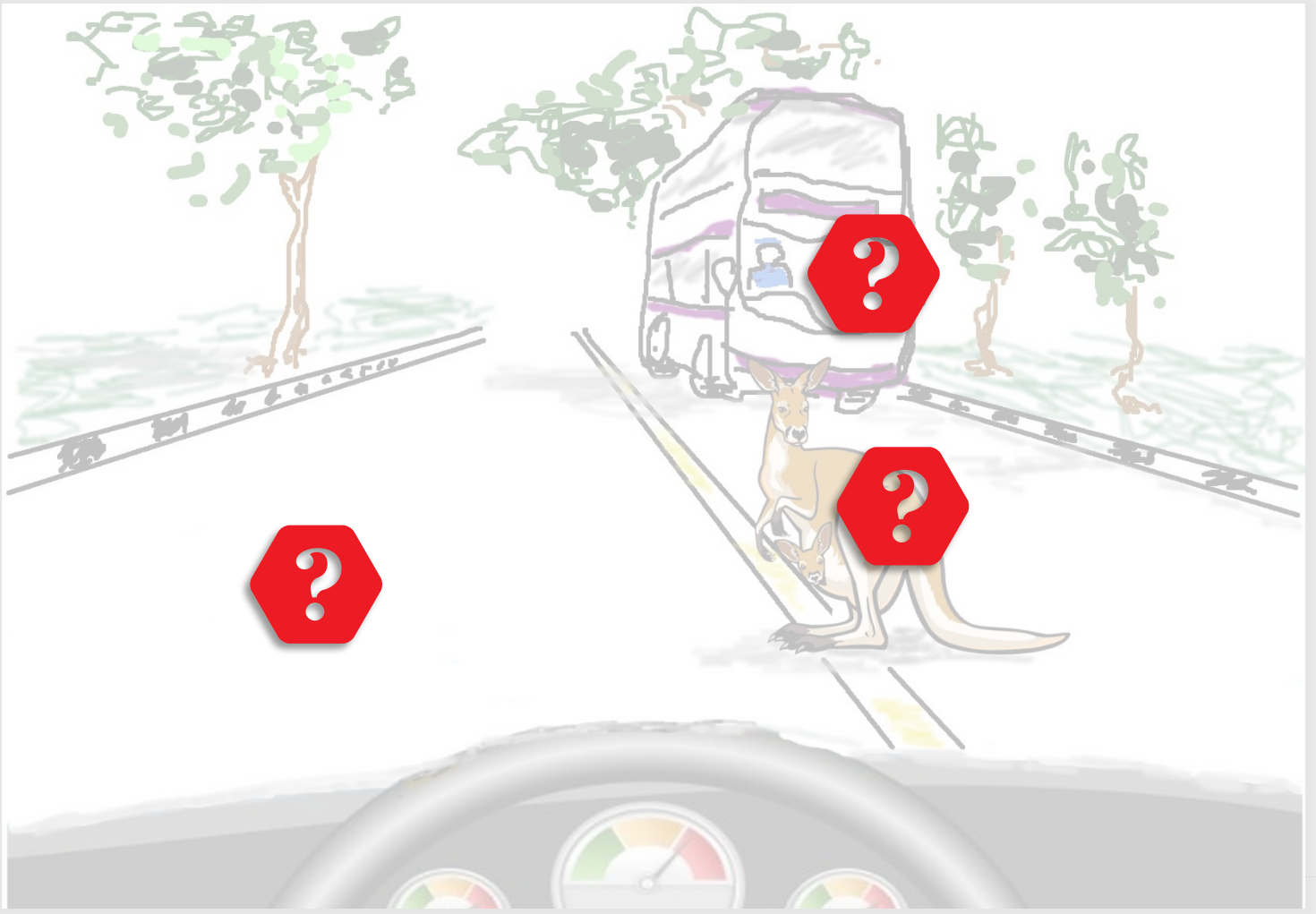


HPCC SYSTEMS®

The predicted impacts of various actions involve **causal and response models** with complex dependencies that are sensitive to assumptions and initial parameters.

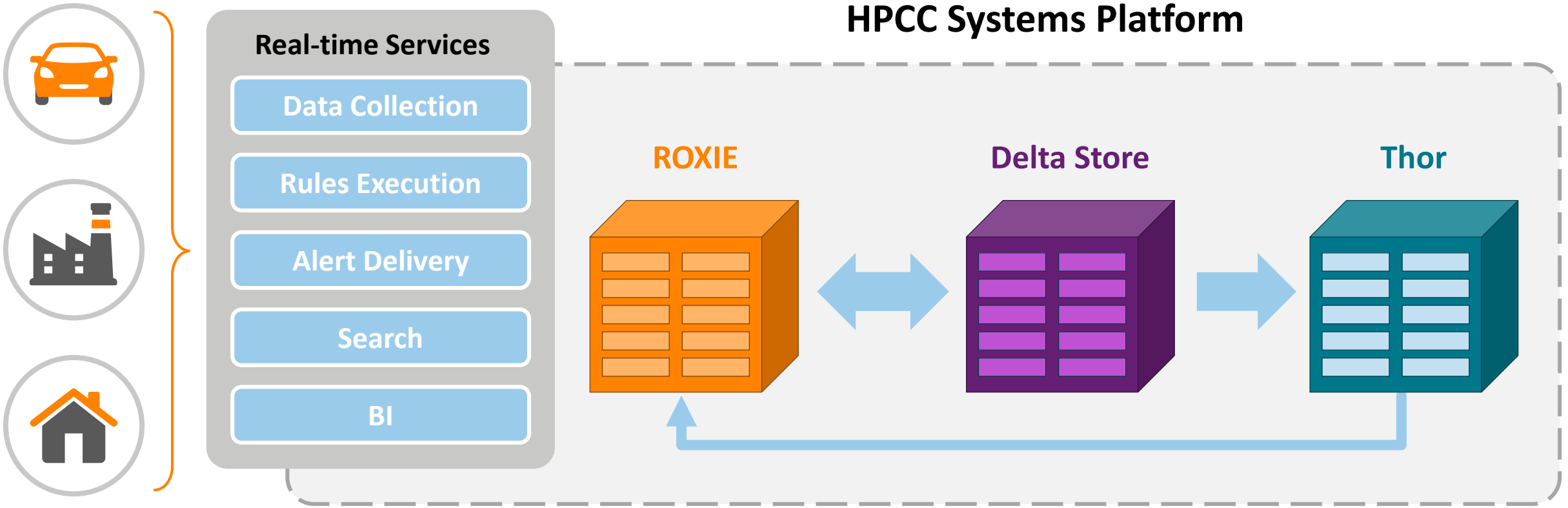


- ✓ **Behavioural** rather than descriptive attributes
- ✓ Large datasets
- ✓ Data hygiene and enrichment
- ✓ Batch processing
- ✓ Fast response to queries



HPCC SYSTEMS®

A robust and proven platform helps address the engineering challenges.



Distributed Massively Parallel Architecture

- Real-time indexed based search
- Real-time rules execution
- Alert call back
- Real-time store
- Real-time analytics on real-time data
- Long term store
- Batch analytics

HPCC SYSTEMS®

Prescriptive tools improve how insights can deliver better outcomes.

- Realise the value proposition of data analytics and business intelligence solutions.
- Simulate many possible scenarios, and select options that achieve the desired outcomes and satisfy constraints.
- Several characteristics and elements that present software engineering challenges
  - Multiple competing objectives and constraints
  - Large number of options
  - Causal and response models that require behavioural attributes
- High-performance cluster computing platform helps meet the engineering challenges.

Questions?

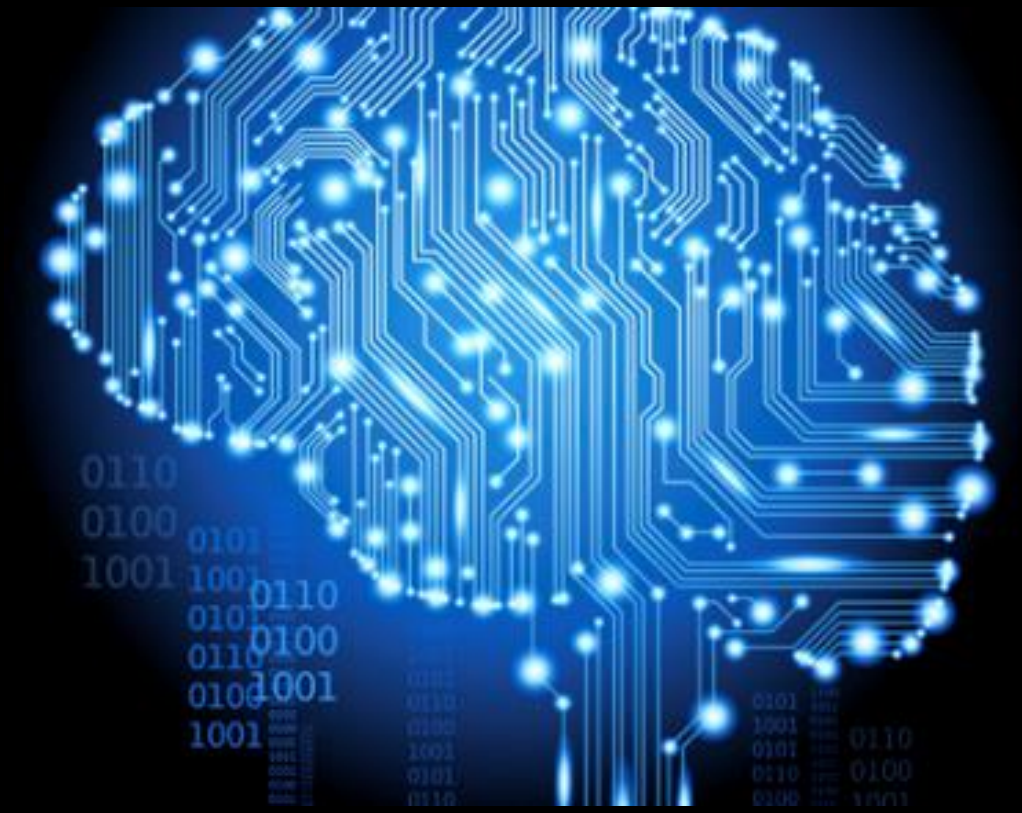
Joselito (Joey) Chua , PhD
Manager Software Engineer
Optimal Decisions Group
Joey.Chua@lexisnexisrisk.com

No animal was harmed in the making of this presentation

HPCC SYSTEMS®

THE DOWNLOAD

TECH TALKS BY HPCC SYSTEMS



Migrating an ECL Code Repository into Git

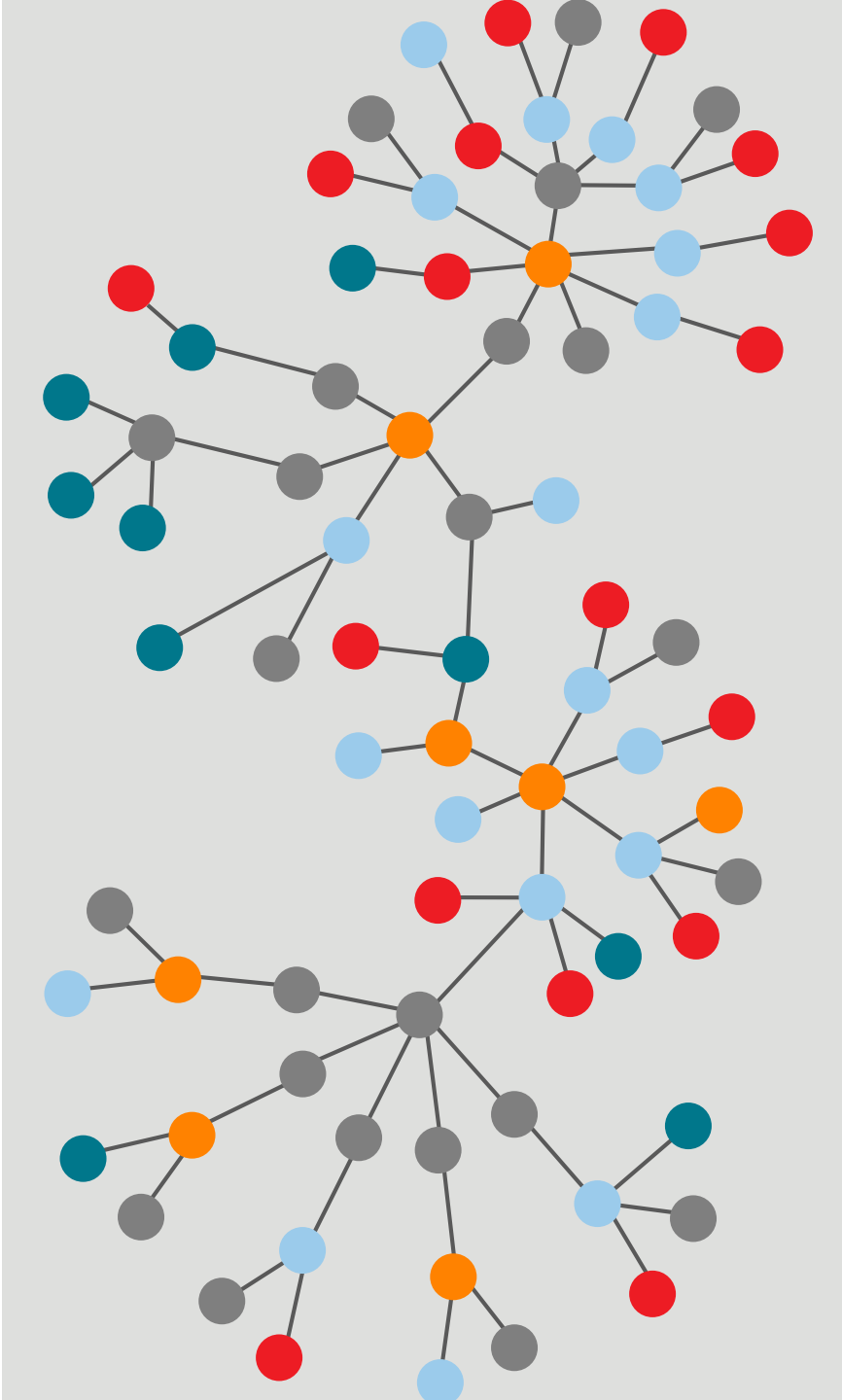
HPCC SYSTEMS®

Jill Luber
Senior Architect

LexisNexis® RISK SOLUTIONS

Quick poll: How are you currently managing your ECL code?

See poll on bottom of presentation screen



HPCC SYSTEMS®

Migration to Git – What and Why?

What is this migration about?

- LexisNexis Risk Solutions is currently undergoing a migration process to move all ECL code from MySQL to a Git source control system.
- We are using an enterprise version of GitLab.
- "**GitLab** is an application to code, test, and deploy code together. It provides Git repository management with fine grained access controls, code reviews, issue tracking, activity feeds, wikis, and continuous integration." - *https://about.gitlab.com/about*

Why are you migrating?

- Branching
- Release management
- Distributed code base

Migration to Git – How?

How is the code structured in Git vs. MySQL?

The MySQL existing code has this migration path:

- Dev -> QA -> Production Thor -> Production Roxie

A simplified new Git structure:

- Thor master – contains all code used to execute production Thor builds
- Roxie master – contains all code ready to be used in a Roxie deployment
- Multiple release branches: Roxie_Release_20070323, etc.
- Lessons Learned: If starting from scratch, I would maintain all file references and layouts in a separate project to remove the artificial dependences between Thor and Roxie code.

Migration to Git – How?

How are you Managing the Migration?

We decided to migrate Roxie first.

- Fewer touchpoints
- Fewest users
- Biggest operational improvement opportunity. An Example:

HPCC SYSTEMS®

Migration to Git – How?

How are you Managing the Migration?

We decided to migrate Roxie first.

- Fewer touchpoints
- Fewest users
- Biggest operational improvement opportunity. An Example:

HPCC SYSTEMS®

Migration to Git – How?

How are you Mana

We decided to migrate

- Fewer touchpoints
- Fewest users
- Biggest operational

xample:

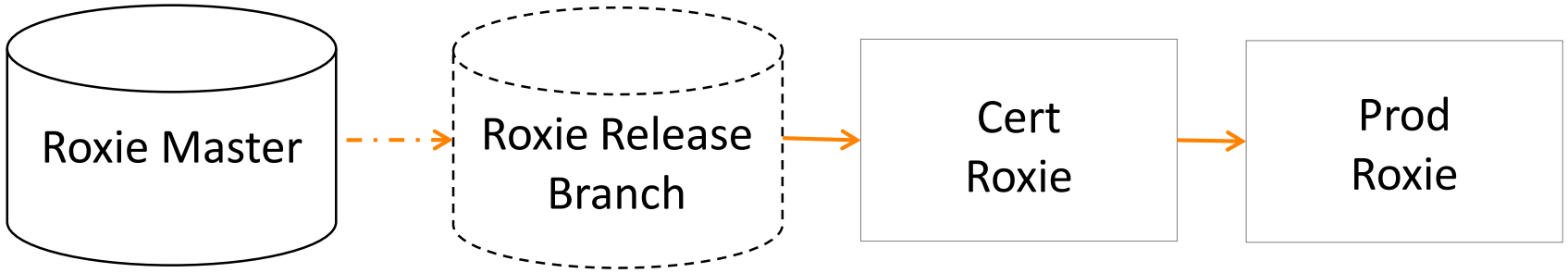
HPCC SYSTEMS®

Migration to Git – How?

How are you Managing the Migration?

We decided to migrate Roxie first.

- Fewer touchpoints
- Fewest users
- Biggest operational improvement opportunity. An Example:



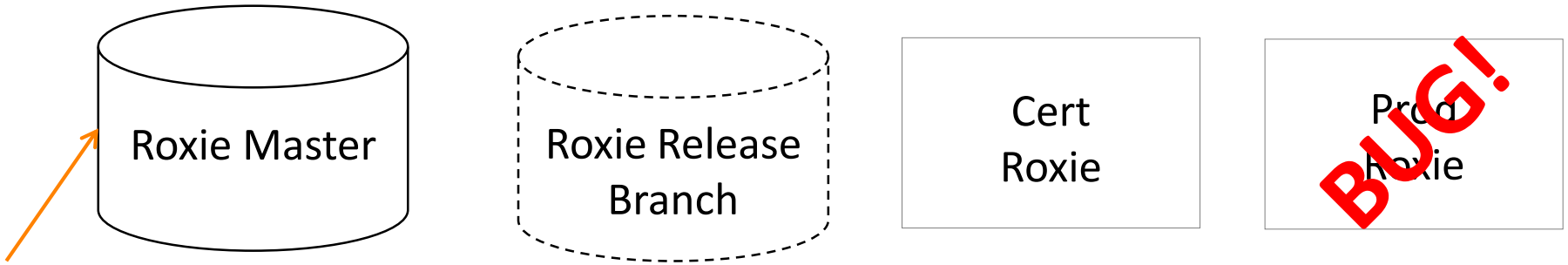
HPCC SYSTEMS®

Migration to Git – How?

How are you Managing the Migration?

We decided to migrate Roxie first.

- Fewer touchpoints
- Fewest users
- Biggest operational improvement opportunity. An Example:



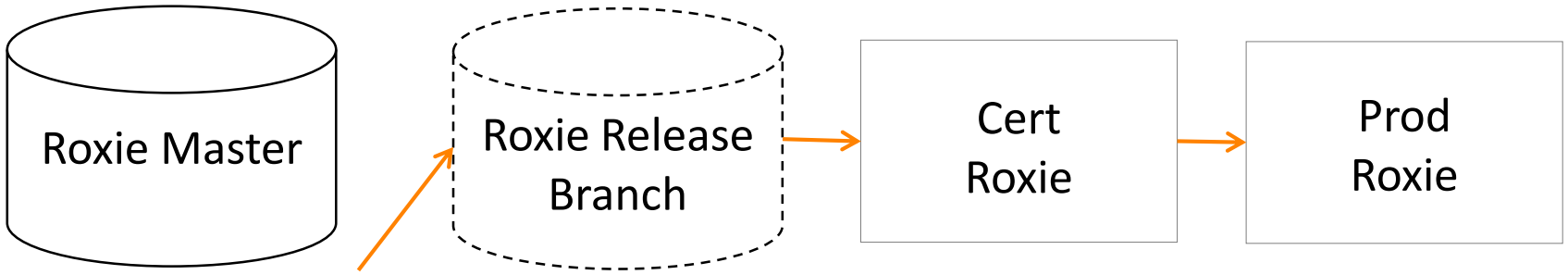
HPCC SYSTEMS®

Migration to Git – How?

How are you Managing the Migration?

We decided to migrate Roxie first.

- Fewer touchpoints
- Fewest users
- Biggest operational improvement opportunity. An Example:



HPCC SYSTEMS®

Migration to Git – How?

Where do you begin?

1. Move all ECL code into Git; we created a tool to automate this using C++.
2. Change migration process to include a push step into Git.
3. Create a release branch off of the Master before each release compile.
4. Deploy to Roxie using Git branches; this requires an ECLCC server on version 5.6.8 or 6.2.8 (or later).
5. Migrate automated Thor production builds; change any scripts to pull Git code before submitting a workunit.
6. Migrate developers; this is the hard part.
    1. Set up permissions via groups in Git.
    2. Establish migration roles and best practices (don't forget to include code reviews!)
    3. Training! Make sure you have educated your developers on Git. Note: There will be a slow down in productivity – at least initially – due to the learning curve.

Migration to Git – Tips

What to Consider Before Migrating

1. Code structure: How many projects do you need? Do they intersect? – Tip: file references and layouts in separate project if you can.
2. Groups for permissions
3. Integrate as much as you can with your issue tracking system – we use Jira.
4. Prepare the team.

Questions?

Jill Luber

Senior Architect,

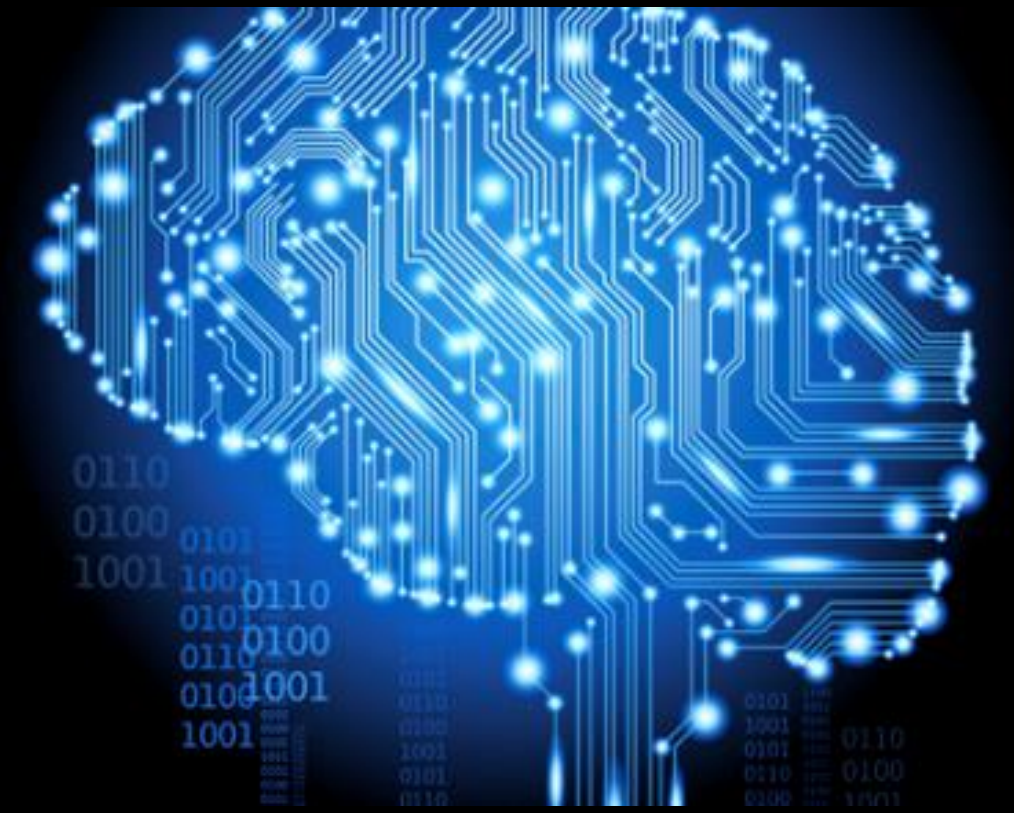
LexisNexis® Risk Solutions

jill.luber@lexisnexisrisk.com

HPCC SYSTEMS®

THE DOWNLOAD

TECH TALKS BY HPCC SYSTEMS



HPCC Systems Platform: Java APIs and Tools

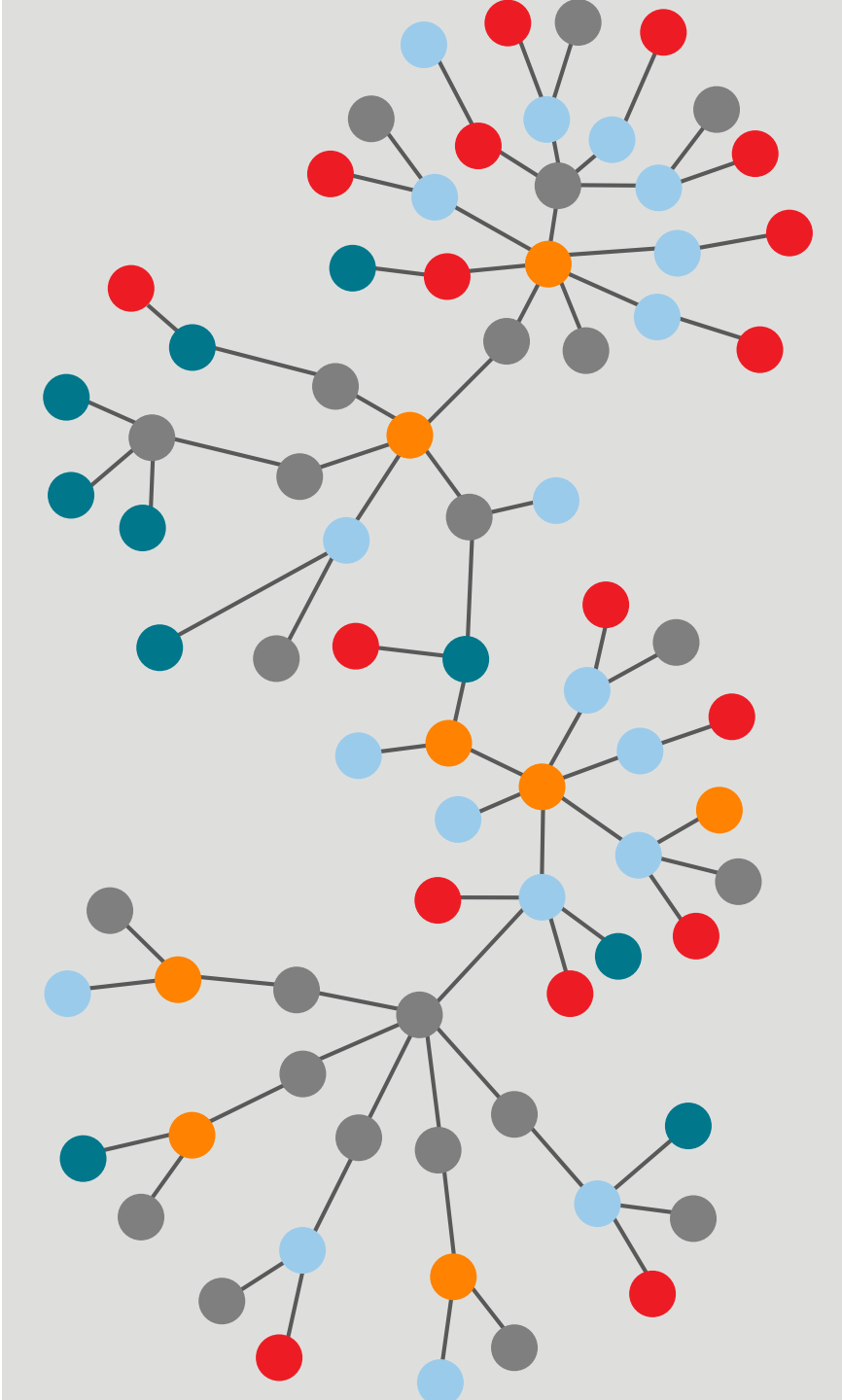
HPCC SYSTEMS®

Michael Gardner
Software Engineer II
LexisNexis® Risk Solutions

LexisNexis® RISK SOLUTIONS

Quick poll: Have you ever used a JAVI API, or a product which uses these tools/API in your daily work?

See poll on bottom of presentation screen



HPCC SYSTEMS®

HPCC Systems Java APIs and tools

- JAPI
  - wsclient
  - rdf2hpcc
  - clienttools
- JDBC

JAPI - wsclient

- What?
  - Java API to HPCC-Platform esp services
  - Built off of generated wsdls for consistent and accurate interface to services.
- Why?
  - Users were creating ad hoc SOAP calls and esp connections that were inconsistent and error prone.
  - Downstream developers were unaware of the proper workunit workflow

JAPI - rdf2hpcc

- Graph based data set.
- Many domains use RDF as standard data model
- Maps RDF model to a HPCC ingestible format

JAPI - clienttools

- Framework structured in a manner to allow easy addition of other client tool logic
- Local compilation of ECL through a Java class.
- Contributed from the RAMPS project by Hari Chittaluru

JDBC

- Java Database Connectivity (JDBC) is a standard Java API that enables Java applications or client tools that support JDBC to access data from a presumably SQL-compliant data source via the SQL language.
- JDBC makes it possible to write a single database application that can run on different platforms and interact with different database management systems.
- Allows the end user to interact with the HPCC Systems Platform as a data source. This is achieved by exposing HPCC logical files as RDB tables, and HPCC published queries as RDB stored procedures.

HPCC SYSTEMS®

Build Changes

- Projects were originally designed to be built with Cmake
- Moved to Maven, which is an industry standard tool for building Java projects
- Projects are now published to maven central repository. Allowing downstream developers to use at will.

```
<dependency>
   <groupId>org.hpccsystems</groupId>
   <artifactId>wsclient</artifactId>
   <version>1.0.0</version>
</dependency>
```

JAPI Project Integration

- Circuits/ECL Pentaho Plugin
- Logi Analytics
- ECL Builder
- Zoomdata
- HIPIE
- wssql
- Flow
- DSP

Community Contributors

- David Bayliss
- Luke Renn
- Drea Leed
- Mike Krumlauf
- Brian O'Neill
- Dhanasiddharth Selvam
- Andy Michel
- Vivek Nair
- Hari Chittaluru
- Joe Chambers

Potential Future

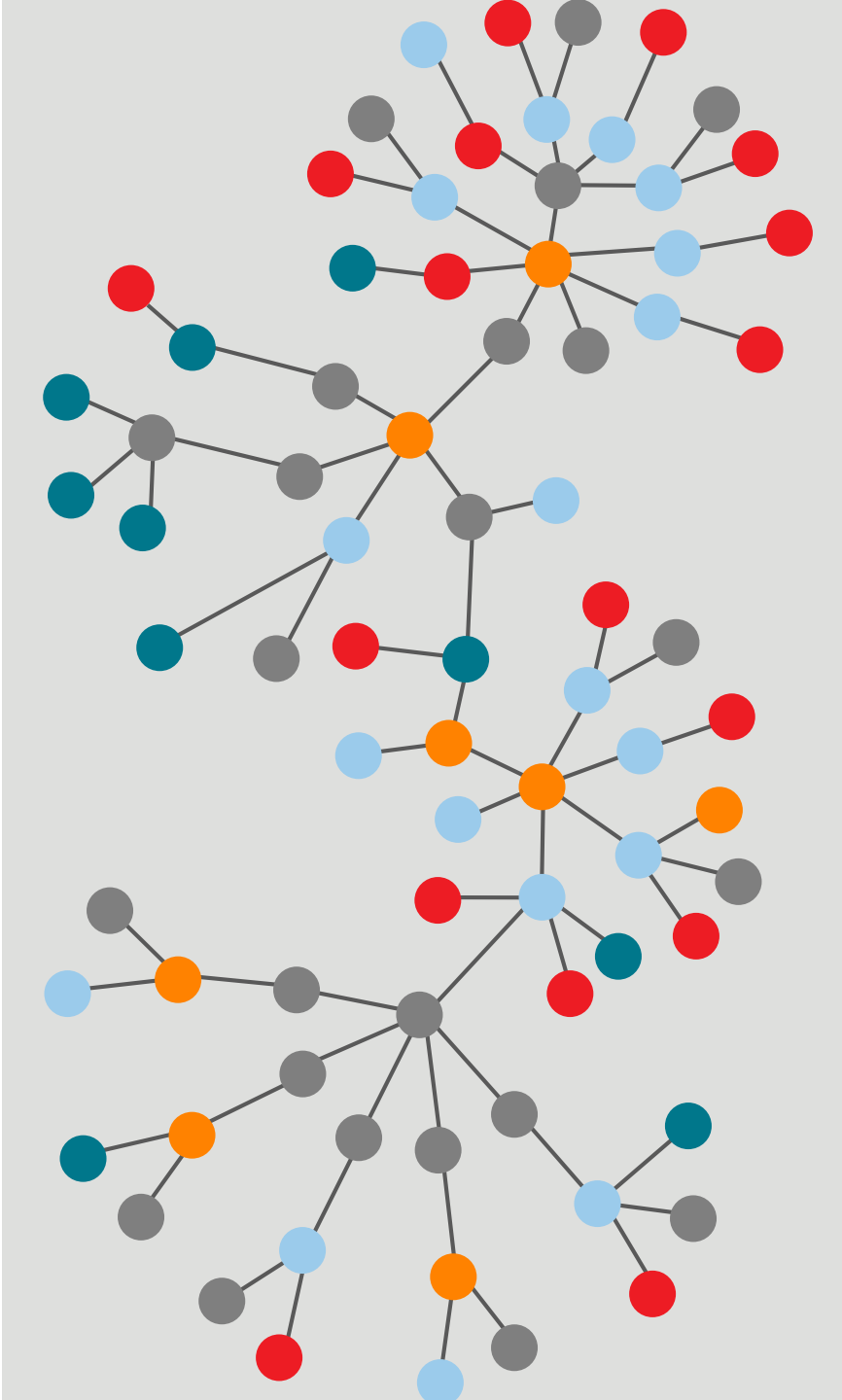
- Migration of other clienttool tools to Java libraries
- Automated validation of esp services through wsclient test suite

Resources

- https://github.com/hpcc-systems/HPCC-JAPIs
- https://github.com/hpcc-systems/hpcc-jdbc
- https://search.maven.org/#search|ga|1|org.hpccsystems

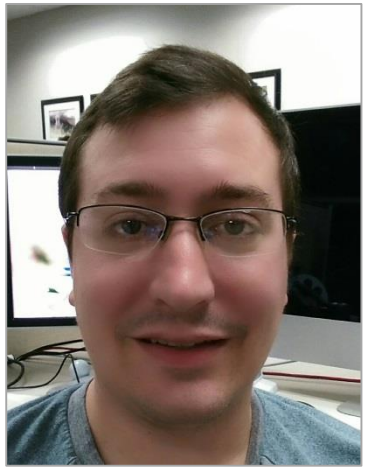
Quick poll: Are you familiar with the maven central repository?

See poll on bottom of presentation screen



HPCC SYSTEMS®

Questions?

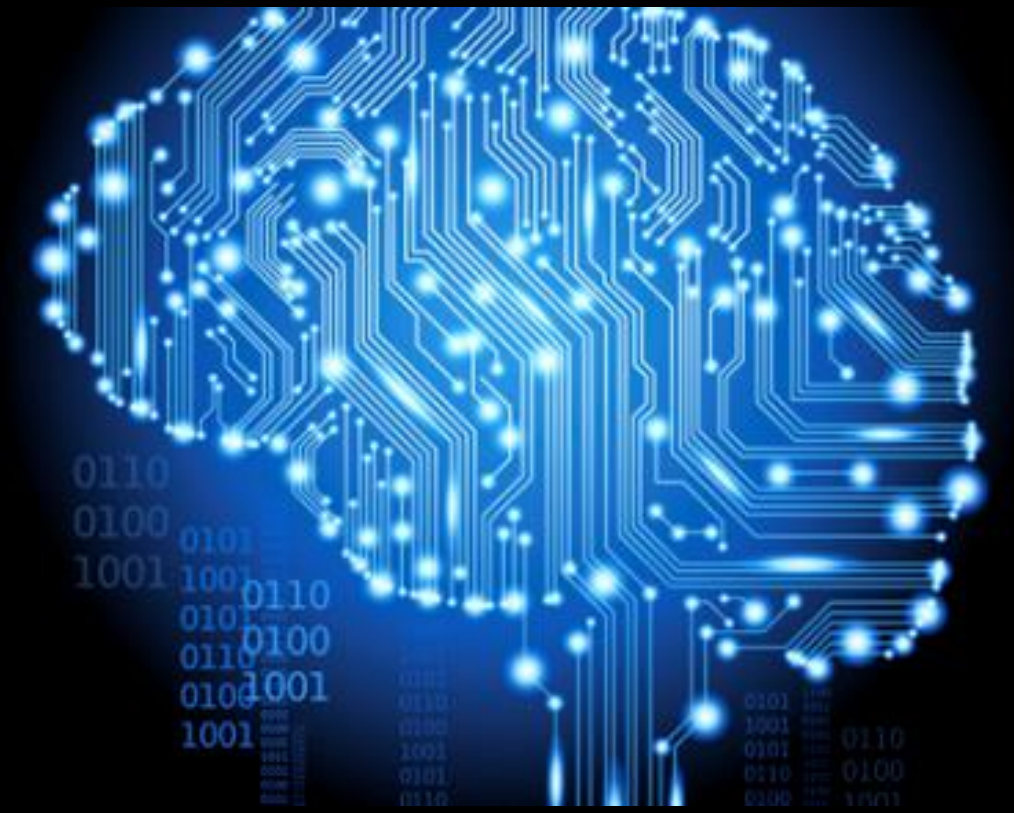


Michael Gardner
Software Engineer II,
LexisNexis® Risk Solutions
michael.gardner@lexisnexisrisk.com

HPCC SYSTEMS®

THE DOWNLOAD

TECH TALKS BY HPCC SYSTEMS



In Search of the Lost ECL Tutorial

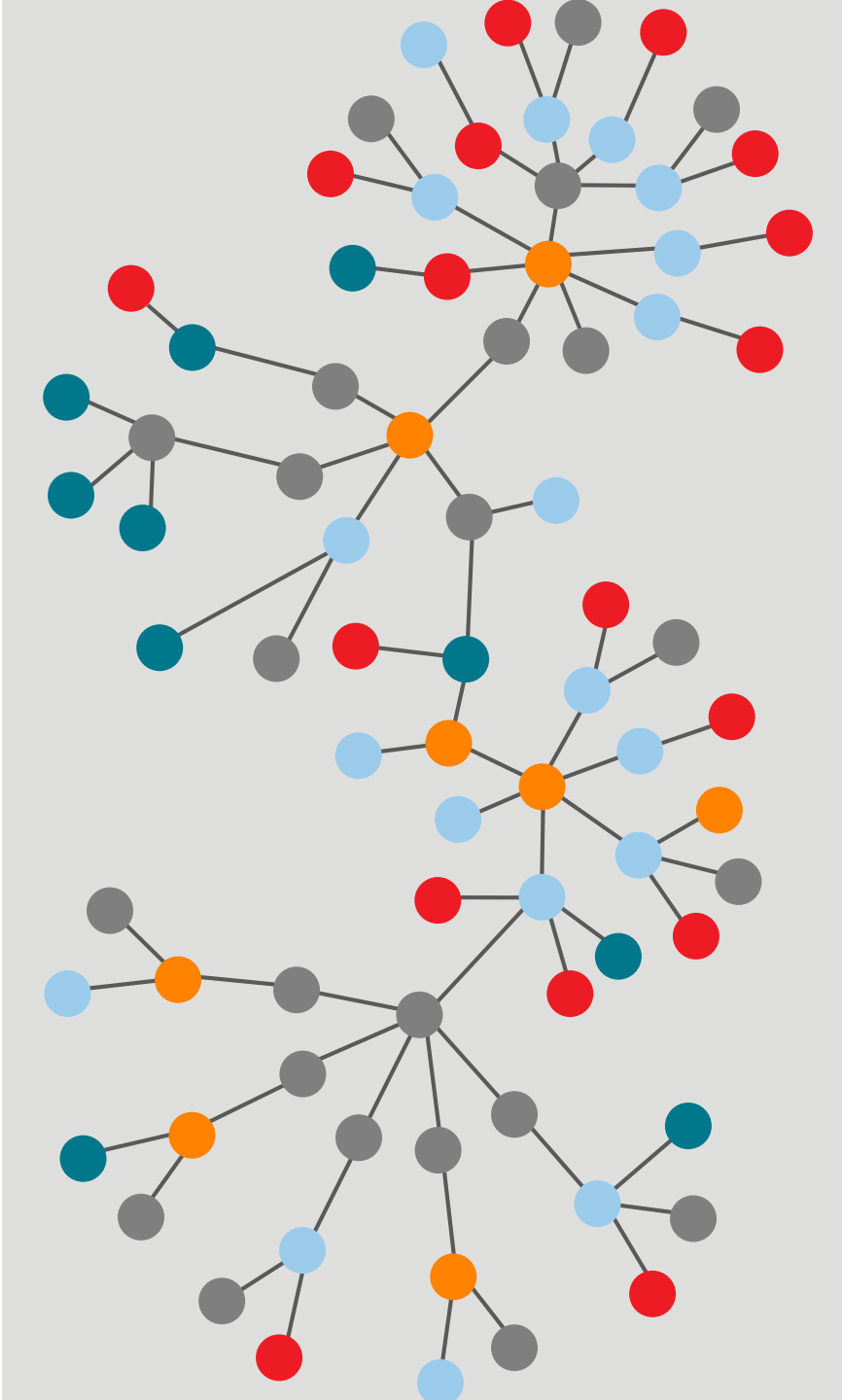
HPCC SYSTEMS®

Bob Foreman
Senior Software Engineer
LexisNexis Risk Solutions

LexisNexis® RISK SOLUTIONS

Quick poll: Have you ever visited David Bayliss' personal web site?

See poll on bottom of presentation screen



HPCC SYSTEMS®

The Best Online ECL Tutorial is the one that many ECL Developers have never seen!

Start at:

http://www.dabhand.org/Technical%20Documents.htm

Bible Search in ECL

Divided into 3 parts:

Part I

Getting, Smacking, Organizing, Structuring, and Preparing the Data

Part II

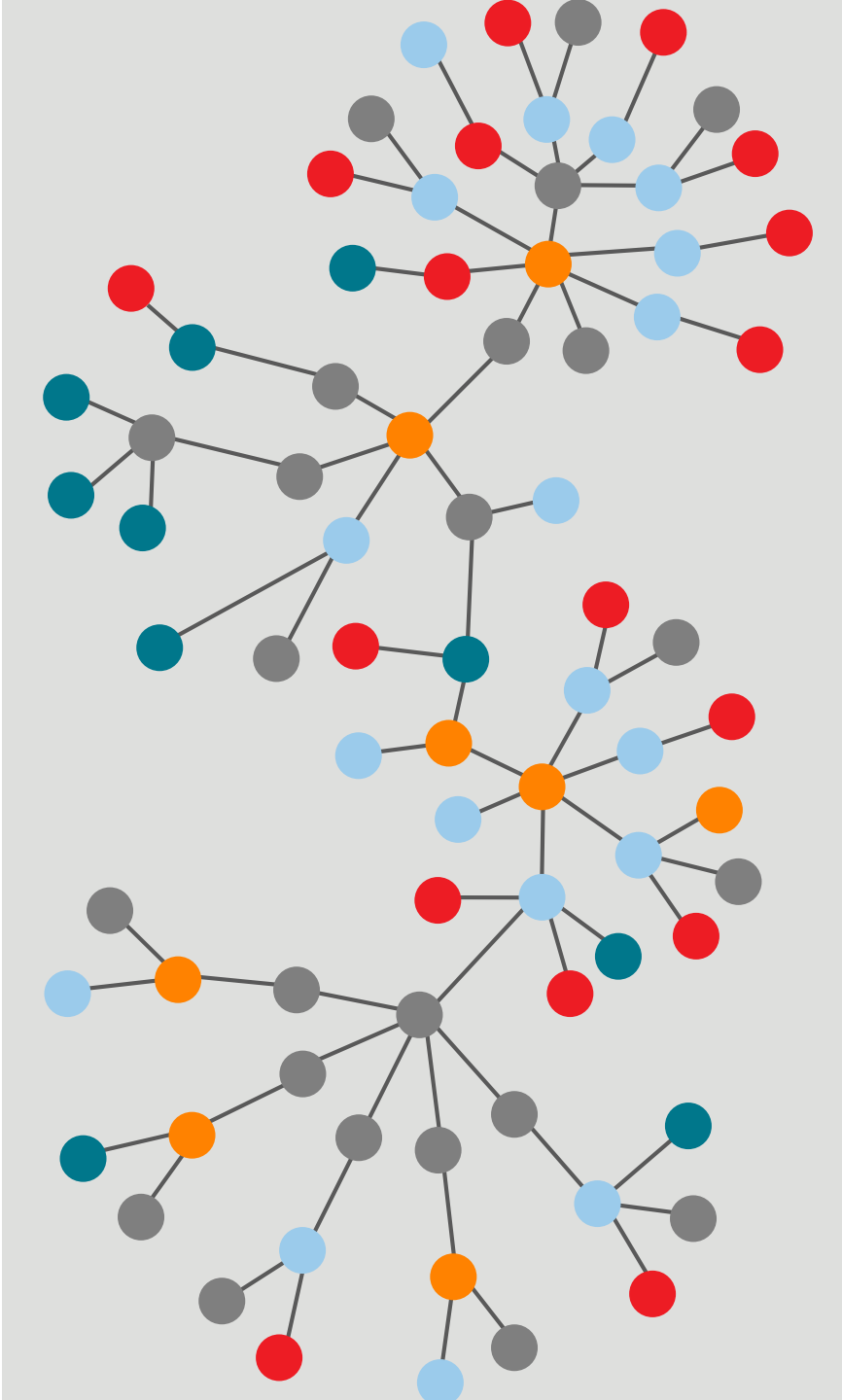
The GRAPH Function, Heart of the Search Engine

Part III

Build a ROXIE Query

Quick poll: Have you ever used the GRAPH function in your own ECL projects?

See poll on bottom of presentation screen



HPCC SYSTEMS®

Questions?

Bob Foreman
Senior Software Engineer
LexisNexis® Risk Solutions
robert.foreman@lexisnexisrisk.com

Download ECL Tutorial Examples from today's talk

HPCC SYSTEMS®

Submit a Talk for an Upcoming Episode!

- Have a new success story to share?
- Want to pitch a new use case?
- Have a new HPCC Systems application you want to demo?
- Want to share some helpful ECL tips and sample code?
- Have a new suggestion for the roadmap?
- Be a featured speaker for an upcoming episode! Email your idea to Techtalks@hpccsystems.com

Save the Date for our next Tech Talk on April 20!

Visit The Download Tech Talks wiki for more information:
https://wiki.hpccsystems.com/display/hpcc/HPCC+Systems+Tech+Talks

HPCC SYSTEMS®

Thank You!

A copy of this presentation will be made available soon on our blog:
hpccsystems.com/blog