Finding the Unknown-Knowns within your data

You recognise it only when you see it

In the science of data analytics, *interesting* is frequently not the same as *correct*. A system may find anomalies that match my rules (and in that sense are 'correct') but very few of these matches may turn out to be significant and hence *interesting*. Similarly, whilst analysts are able to recognise if they have found what they are looking for, they are often unable to describe the rules that would enable it to be found; to separate the interesting wheat from the correct-but-dull chaff. This is not a failing of analysts, but a characteristic of the complex, chaotic, and fuzzy world in which we live.

Scenario Analytics and the Insight Engine

Technology must adapt to these uncomfortable facts, provide ways to describe what we would like to find, deducing *how* to find it for itself. We call this "scenario analytics", and in this paper Andrew Lea, our Technical Director in charge of Data Analytics, describes our solution to the business challenge of finding the lethal but hidden unknown-knowns deep in our data, which we will be blamed for missing.

data
Scenarios + Analytics
Insights

Understanding Scenario Analytics Capability

Scenario Analytics capability is best understood by way of examples:

Scenario Analytics answers these sorts of questions, which we call 'scenarios'

Field	Scenarios
Crime	Who is orchestrating, but not participating in, the sale of controlled drugs?
Fraud	What household benefit fraud is occurring that we don't know about?
	Which tax returns contradict publically filed data?
National Security	Is anyone conducting reconnaissance of critical national infrastructure, and to what end?
	Who is alienated and at risk of becoming radicalised? What are the key influencers? Where are the centres of radicalisation?
Cyber	Is this theoretically possible, but never-seen-before, cyber-attack happening?
	Our customer data may have been hacked. Is it being used 'out there'? If so, by whom, what for, and which of our customers does it impact?
Money laundering	What series of apparently independent financial transactions are there which, whilst each legitimate in their own right, together constitute money laundering or terrorist financing?
Staff safety	Is anyone trying to identify staff of a company carrying out legal but contentious research?
	Are any of my staff being bullied or discriminated against?
	Are any of our staff taking bribes?
Companies	Which protest groups are planning on disrupting my Annual General Meeting?
	Is anyone setting up a hostile take-over?
	What are my competitors researching, and planning to take to market?
Industry	What cascade failure might soon occur in my manufacturing plant?

Scenario Analytics does not need leads

Scenario Analytics analyses data in light of scenarios which would be interesting, and describes the instances it finds in natural language, with an explanation of supporting evidence.

Key Benefits

Scenario Analytics benefits differ in kind to those of traditional systems

- Finding things either:
  - according to an analysts description of *what would be interesting*, or
  - which are similar to a known existing scenario.
- The ability to continuously monitor large datasets against very rare but extremely serious threats, generating alerts when they or their progenitors are found.
- New leads from data, which it can do because it does not need a starting point.
- Explanations as to why a scenario instance is interesting, for human evaluation.
- Performs the same investigation as several staff, liberating resources or reducing costs.
- It does not get bored with tedious work, and so miss rare but critical events.

Business Impact

Finding the "unknown-knowns" in your data

Knowledge	Analytics
Known-knowns	Relational databases
Known-unknowns	Social network analysis
Unknown-knowns	Scenario Analytics
Unknown-unknowns	Unsupervised learning

Scenario Analytics is well placed to find the 'unknown-knowns' that catch us napping, with questions asked in hindsight as to why *it* was not known, even though *it* was found in the data (but after the event), and should have been obvious that *it* could happen. The unpalatable answer, of course, is nobody was looking because although indeed obvious *it* was (a) unlikely (b) hard to define the rules by which *it* could be found and (c) time-consuming, expensive, and tedious to find.

Insight output

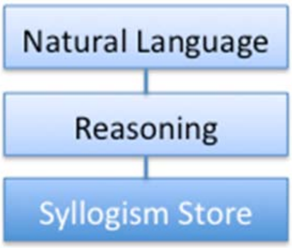
Scenario Analytics can raise alerts (for example by email) in response to scenarios arising from routine scanning of network data, or provide helpful visual explanations of those scenarios as web pages.

Technical Heritage

Scenario Analytics is based on the foundation layer of a 'deep' Artificial Intelligence architecture

The foundation is a highly flexible **syllogism store**, which holds:

- facts, facts about text or images, or extracted data
- facts-about-facts, such as confidence levels, permissions, scope, or even *another* syllogism store. (We could represent a book library with real-world borrowers, and for each book, the characters within it, without confusing the characters of one book with another, or with the real borrowers.)



A syllogism store can be persistent or transitory (for 'on-the-fly' deductions). They contain complex recursive mechanisms (used by Scenario Analytics) for powerful optimised queries to be quickly executed on large, possibly big data, datasets.

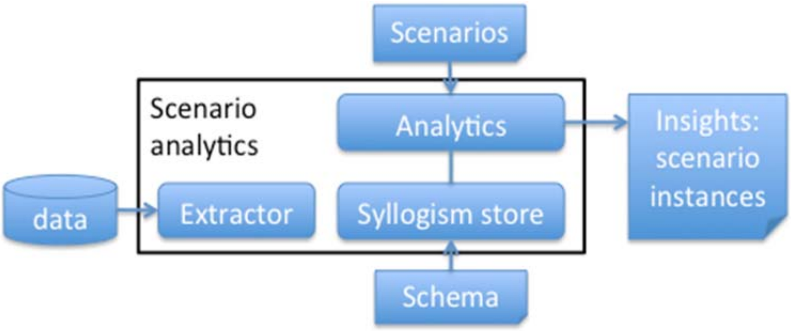
It can simulate some facets of human thought And language

The **reasoning layer** can make deductions, or request further facts be obtained. Being a type of Artificial Intelligence, it can even use a form of *artificial intuition* in which logical data gaps can be leaped: it can find scenarios instances which would be true but for one or two pieces of missing but likely evidence, describe those scenario instances, and explain the evidence which, if found, would complete the logical chain.

The top layer is responsible for **natural language understanding** and synthesis. Unusually for natural language systems, it uses the reasoning level to understand language. This layer could be added to Scenario Analytics if an interactive natural language interface is required.

Architecture Overview

Scenario analytics imports and fuses data into its syllogism store. Its analytics uses pre-defined scenarios, and a description of the data relationships in the scheme, to identify new scenario instances.

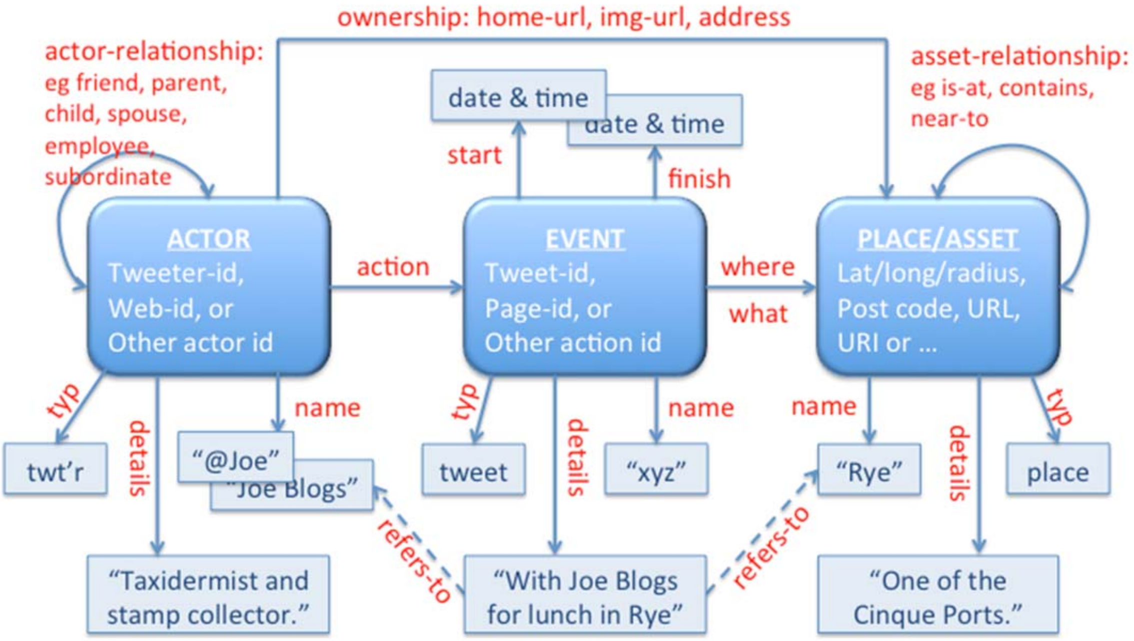


Real World Data Challenges

Complexity – using the right *schema*

Taming complexity requires an appropriate model or world-view. Since data is generally transformed from its original structure to the current schema during load, changing the schema does not need the system to be re-architected, as it might with a relational database.

We often use the Primary Key Three Entity model, as it is sufficiently flexible to encapsulate the ever-changing 'social' world.


ownership: home-url, img-url, address
actor-relationship: eg friend, parent, child, spouse, employee, subordinate
asset-relationship: eg is-at, contains, near-to
date & time
date & time
start
finish
ACTOR
Tweeter-id, Web-id, or Other actor id
action
EVENT
Tweet-id, Page-id, or Other action id
where
what
PLACE/ASSET
Lat/long/radius, Post code, URL, URI or ...
typ
details
name
typ
details
name
name
details
typ
twt'r
"@Joe"
"Joe Blogs"
tweet
"xyz"
"Rye"
place
refers-to
refers-to
"Taxidermist and stamp collector."
"With Joe Blogs for lunch in Rye"
"One of the Cinque Ports."

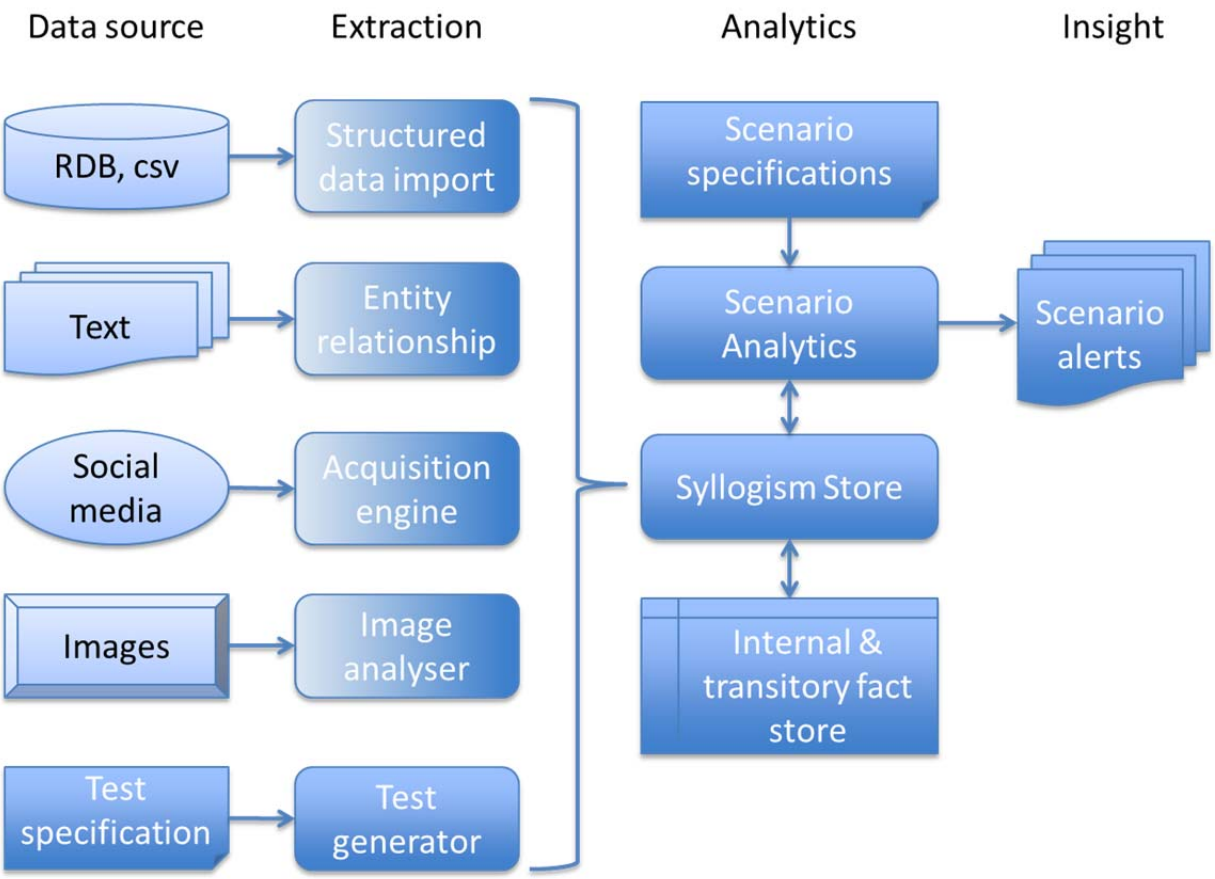
Heterogeneous Data

The Syllogism Store addresses many of the challenges and defects of real-world data.

Because of the flexible syllogism store representation, a wide range of heterogeneous data sources can be fused, given a suitable import routine. Sources may be:-

- Structured data in 'ordinary' relational databases, spread-sheets, or 'flat' csv files
- Meta-data
- Social media
- Textual documents from which entities and their relationships, are extracted
- Images, using information from specialist image extraction techniques, such as Primary Key Associates staff have used in the space industry
- Simulators for testing

Scenario Analytics fuses data from multiple sources.



Uncertainty and Contradictions – the uncertain world in which we live

Scenario Analytics can know how sure it is of a conclusion

The internal representation supports the concept of uncertainty or (equivalently) confidence, so explanations can explain the degree of confidence that is implied by our confidence in the underlying data items.

These same certainties allow the system to pursue independent and contradictory reasoning chains, and to select the one with the greatest credence or weight of evidence. Contradictions can be exploited: for example a declared income and expenditure discrepancy might indicate a tax fraud.

Data Integrity

Scenario Analytics never modifies source data, so it cannot corrupt your corporate record.

Data Protection and Proportionality

In many firms some groups, such as fraud investigation, may see Personally Identifiable Data, but not others, such as marketing. As the Syllogism Store can store entitlements with data and carry them through analysis, results can be limited to those with sufficient authorisation.

Data Size

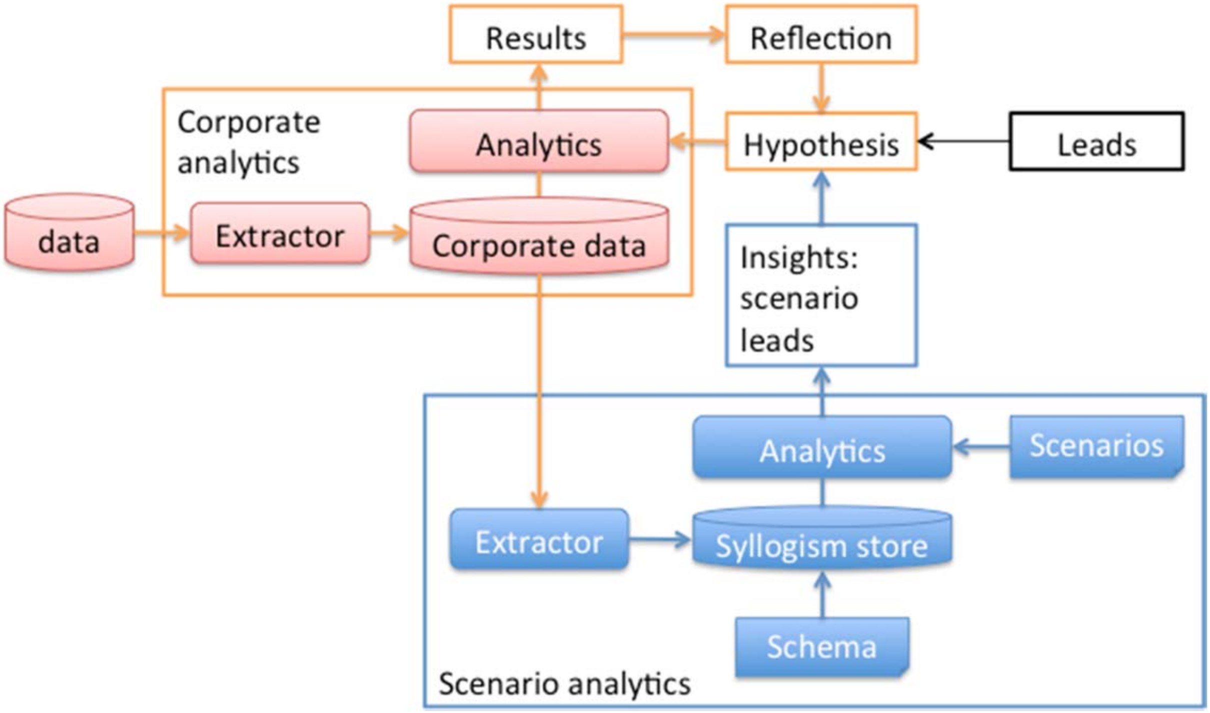
We have optimised Scenario Analytics to run quickly even on large datasets. We routinely process tens of millions of nodes with hundreds of millions of links and generate sub-second scenario results using off-the-shelf computing hardware.

Using Scenario Analytics and the Insight Engine

Integrating into the Work Flow

Scenario Analytics enhances your existing work-flow by generating new leads for investigation using your normal processes.

Scenario analytics enhances existing analytics



Deploying Scenario Analytics

Scenario Analytics is embodied in the Primary Key Insight Engine

To deploy Scenario Analytics, we recommend engaging Primary Key Associates to:

- work with your analysts to identify the scenarios that would be of interest to you were they to occur, so we can then write the scenarios in the specialised language.
- integrate scenario analytics to run on a timely basis, depending on data update rates.
- help you (1) pipe data into the Scenario Analytics System and (2) vector output – identified scenarios – to interested consumers.
- provide one day a week support and assistance during use, adding new scenarios, helping with urgent investigations, or dealing with broken data.

Testing Scenario Analytics

How do we know Scenario Analytics works?

We have a proven approach to testing Scenario Analytics, typically we use sophisticated models to generate test data, into which those scenarios are injected, and then verify that Scenario Analytics finds them. By way of example, recently generated a test case of 67,000,000 entities and checked the four significant injected cases were found.

Where we have access to real world data we can inject known scenario instances into that data for the same purpose or can modify that real world data (without addition) to contain scenario instances.

Procuring Scenario Analytics and the Primary Key Insight Engine

Copyright in the output lies with the client

We provide three procurement routes:

- As an **on-site service**. We charge a flat monthly fee for use of Scenario Analytics by each client company, regardless of the number of CPUs or users.
- As on **off-site service**. We hook up Scenario Analytics to our open source acquisition engines or take a data feed provided by you, and provide you with scenario alerts, via secure email. The monthly fee is slightly higher than that of an on-site service.
- For integration into a rapid application development platform as a re-usable block, using whatever pricing model you **license** your platform to your clients.

We charge integration, set-up, scenario development, and support at our standard rates.

To arrange a demonstration, or procure a value-proving experiment on your live data (for which the flat monthly fee, but not our time, is waived), please contact us.

About us

Since 2010 Primary Key Associates has delivered world-class consultancy, developing and exploiting cutting-edge analytics, artificial intelligence and digital investigation technologies.

Answer-Focused Technologies

We re-invest much of our revenue to develop cutting-edge data analytics, investigation and artificial intelligence technology to address the business problems we see our clients facing. Our IPR portfolio includes:

- Primary Key **Scenario Analytics** – a fast, modern and flexible data analytics technology to find, explain and illustrate the 'unknown knowns' in your large datasets.
- Primary Key **Insight Engine** – Scenario Analytics combined with powerful entity and relationship extraction, applied to digital forensics that reveals previously unknown connections in evidence.
- Primary Key **Illuminate** – a social media analysis and intelligence technology and service to find and analyse open source data to address particular business problems.
- Primary Key **Incipients** – predictive analytics that identify what is going to trend.
- Primary Key **Distil** – a technology toolkit we deploy on client sites to both find frauds and identify the business practices that make you vulnerable to fraud.
- **Fast analytic heuristics** to overcome the bottleneck of solving intrinsically non-parallel, yet vital, problems.

Expert consultancy and services

Our team are experienced professionals in IT and related areas:

- As experts in enterprise and security architectures (including SABSA), information security and cryptography we help design, validate, and test systems which are secure both physically and in cyber-space.
- We build software systems from the smallest (in C and assembler coded microcontrollers) through the mobile (Java or Swift coded Android or IoS apps), to the ubiquitous (Python or C+ on PCs, web technologies and virtualised servers) and even the esoteric (real-time and spacecraft).
- We understand data (SQL and graph databases) and artificial intelligence (computational linguistics, image analysis and machine learning) and have fast algorithms for timely and secure analysis of big datasets and evidential analysis of social media.
- We provide cyber threat intelligence and competitive intelligence.
- We support civil and criminal digital investigations from open source research through digital forensic analysis and providing expert witness reports and testimony.
- We undertake both technical and business programme and project management.

e: feedback@primarykey.co.uk
w: www.primarykey.co.uk
t: +44 1403 599900

@pkaluk

SWEET & MAXWELL CHECKED 2017 EXPERT WITNESS