Klaus Felsche,
Director Intent Management & Analytics

Department of Immigration and Citizenship

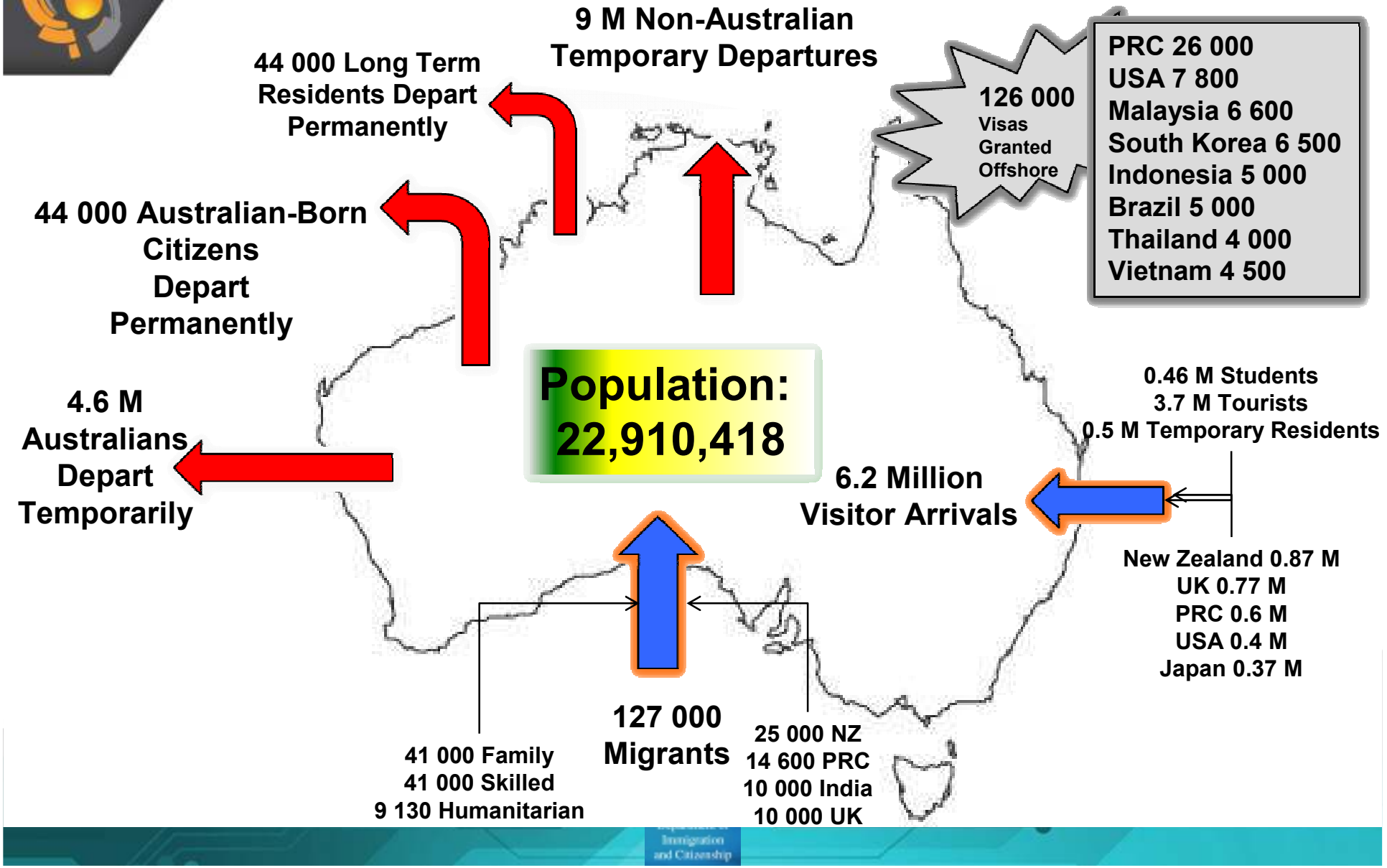
Scope: Real-Time Analytics Challenges for Immigration

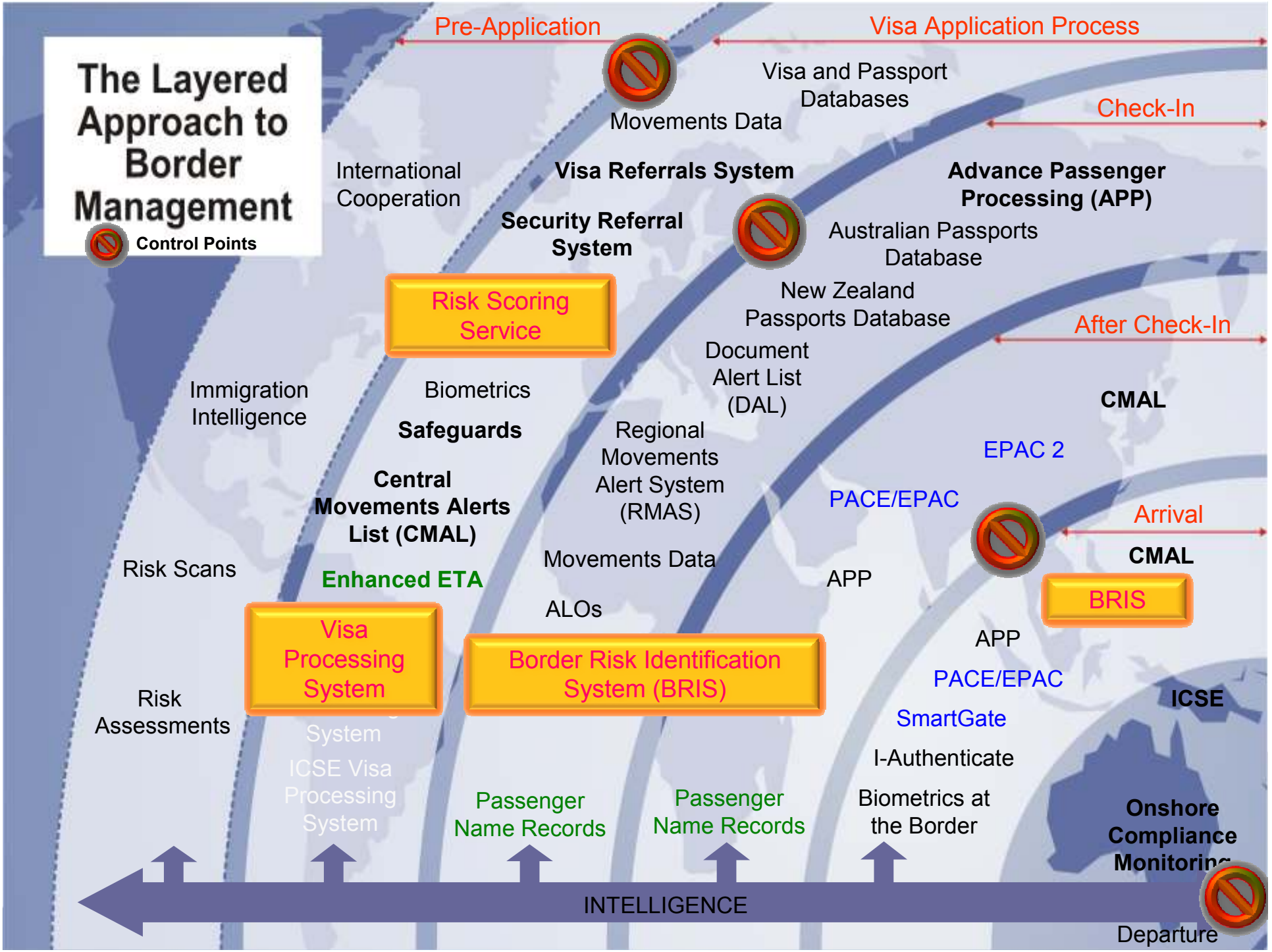
- Understanding the risk in real-time and creating a risk-scoring service for large volume transactions

- Mining your data – what are we looking for at the department when we analyse data?

- Identifying patterns in the data – how can you do this?

Australia – A Quick Overview




The Layered Approach to Border Management
Control Points
Pre-Application
Visa Application Process
Visa and Passport Databases
Movements Data
Check-In
International Cooperation
Visa Referrals System
Advance Passenger Processing (APP)
Security Referral System
Australian Passports Database
New Zealand Passports Database
Risk Scoring Service
After Check-In
Document Alert List (DAL)
Immigration Intelligence
Biometrics
CMAL
Safeguards
Regional Movements Alert System (RMAS)
EPAC 2
Central Movements Alerts List (CMAL)
PACE/EPAC
Arrival
Risk Scans
Movements Data
CMAL
Enhanced ETA
APP
BRIS
ALOs
Visa Processing System
Border Risk Identification System (BRIS)
APP
PACE/EPAC
Risk Assessments
ICSE
SmartGate
System
I-Authenticate
ICSE Visa Processing System
Passenger Name Records
Passenger Name Records
Biometrics at the Border
Onshore Compliance Monitorin
INTELLIGENCE
Departure

The Data – What we have

- The volume of data is so large that it is close to impossible to make sense of it.
  - Over 100 million visa applications
  - Over 500 million border crossings
  - Over 100 million passenger cards
  - Over 500 million alert list checks
  - Records for over a million overstayers and non-compliant visa holders
  - Gigabytes of unstructured intelligence
  - …..
- This is just the data collected and stored by DIAC!

One Challenge: Identify Risk in Incoming Traveller Stream

	This Year	2015 Forecast
Traveller Arrivals	14 685 923	18 000 000+
Per minute	28	34 +
Per day	40 000 per day	48 000 +
Airport Inspectors	60	60
Visas Granted	4 500 000	5 400 000 +
Overstayers	64 000	77 000 +

Our Data – Where From?

- In DIAC data is collected and stored in support of **our core business processes**:
  - Visa processing
  - Border processing
  - Citizenship
- Collected data primarily supports the legislative and regulatory requirements for the above processes:
  - Identify the person
  - Determine eligibility for visas/entry/stay/exit
  - Record decisions (and circumstances/reasons)
  - Record outcomes of processes (eg visa grant/refusal, cancellations)
  - Support process resourcing (volume, complexity, risk determines resourcing)

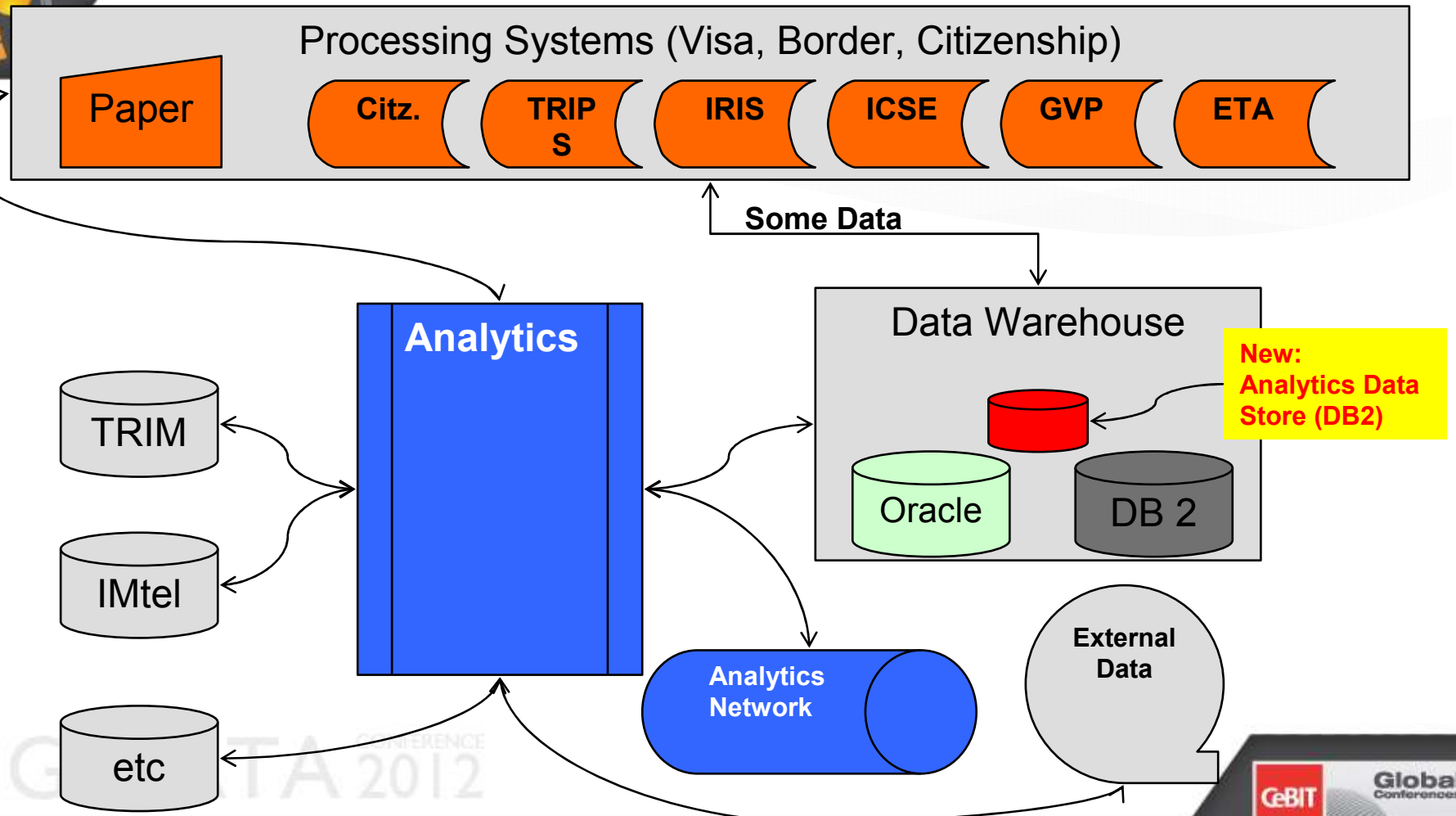
Core Data Characteristics

- Most visa data includes:
  - Identity of applicant
  - Reason for travel (Intent)
  - Sufficient information to allow assessment of above and:
    - Health status
    - Character assessment (criminal etc conduct)
    - Basic National Security
    - Meeting specific requirements for visa as specified by visa regulations.
- Border transaction data includes when, where, how and:
  - Identity of the traveller
  - Authority to enter/depart

Data – Where?

- Data may be held in many forms and in many locations:
  - On paper records (provided by clients and generated by our staff)
  - In processing systems (IRIS, ICSE, GVP, TRIPS, …)
  - In one of our two data warehouses (Oracle & DB2)
  - In transit ….
  - In stand-alone databases
  - Electronically in our records management system (TRIM)
  - In our intelligence system (IMtel)

Data Everywhere



Our Process

- Predictive Models:
  - As we know previous adverse outcomes for travellers we can use the data collected for these cases.
  - Using sophisticated analytics software, we ‘feed’ the adverse cases into a large slice of our data.
  - Analysts then use various techniques to generate models (or patterns) that define the adverse outcomes.
  - Some of the simpler models can be deconstructed into a set of rules, others are so complex that they are best left in their ‘native’ form.
  - Analysts then test the reliability and validity of the models by taking other slices of historical data (for which we also know actual outcomes) to test the reliability of the models.
  - The most reliable model is then selected as the predictive model.

A Simple Model

Attributes	Values (based on 2006 data)
Previous entries made	Nil
Visa subclass at the time of movement	159, 416, 418, 422, 427, 428, 570, 572, 573, 773
Birth Country	Detail Removed
Airlines	BI (Royal Brunei Airline), CI (China Airlines), CX (Cathay Pacific), DJ (Pacific Blue), EK (Emirates), FJ (Air Pacific), GF (Gulf Air), MH (Malaysian Airlines), MK (Air Mauritius), OS (Austrian Airlines), TR (Tiger Airlines), UA (United Airlines), VN (Vietnam Airlines)
Citizenship at the time of movement	Detail removed
DIAC post granting the visa for the movement	Central Office, Darwin, Dubai, Hong Kong, London, Melbourne, Moscow, Southport, Washington
Port of Arrival	Adelaide, Eagle Farm (Brisbane), Perth
PID Percentile Score	< 67.5

64.3% chance that a passenger matching this profile will be the person of interest

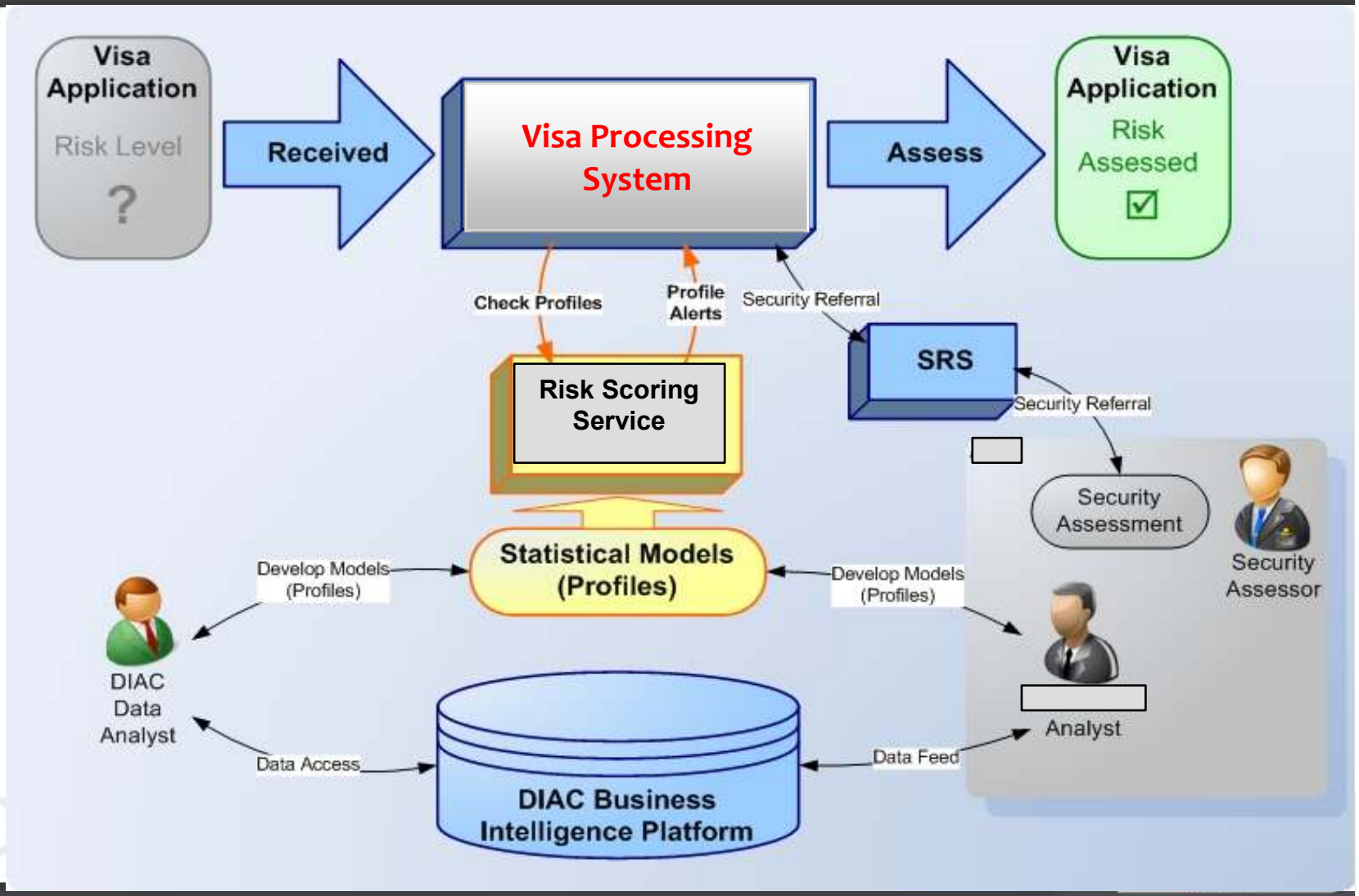
Predictive Analytics: Method

- **Pre-modelling**
  - Risk modelling in DIAC includes extensive consultation with business specialists before, during and after model construction

- **Model development**
  - We use the R environment for statistical computing
  - models are 'trained' using data available to DIAC
  - The developer may utilise any modelling technique that can be implemented as R source code
    - Examples:
      - decision trees,
      - random forests,
      - boosted models,
      - association rules,
      - clustering methods,
      - linear models,
      - time series and neighbourhood methods

Deploying Predictive Models

- **Model validation**
  - The preparation stage includes several automated code quality checks including:
    - ensuring that the model is able to:
      - correctly score transactional data
      - respond appropriately to noisy and missing data
    - ensuring error handling has been correctly implemented
    - the bundling of self-check regression test data
  - The deployed model consists of executable R code and meta-data.
- **Model deployment**
  - A custom Graphical User Interface is used to deploy the model file
  - Standard self tests on the model ensure correct operation
- **Model monitoring**
  - The RSS includes reporting and automated monitoring of deployed models.
  - In the event of a model performing outside expected parameters, appropriate treatment action can be automatically applied.
  - Models are updated on a period basis as business requirements dictate.


Visa Application
Risk Level
?
Received
Visa Processing System
Assess
Visa Application
Risk Assessed
Check Profiles
Profile Alerts
Security Referral
SRS
Risk Scoring Service
Security Referral
Security Assessment
Security Assessor
Statistical Models (Profiles)
Develop Models (Profiles)
Develop Models (Profiles)
DIAC Data Analyst
Analyst
Data Access
Data Feed
DIAC Business Intelligence Platform

Sample Projects Under-Way

1. Risk Tiering:
   - Workflow Management based on risk

2. Alerts Dashboard:
   - Automatically monitors business activity
   - Automatic alerts when activity falls outside expected parameters

3. Border Risk Identification System:
   - State-of-the-art risk scoring engine supporting our International airports
   - Sampling 100% of inbound passengers in real time, 24 hours a day

4. Identity Insight:
   - Enables real time search of linked DIAC data to determine connections between entities

Risk Tiering

Risk tiering enables DIAC to apply an **evidence based** automated risk assessment to visa processing to target high risk cases and reduce over processing of low risk cases.

HIGH RISK PROCESS

Full assessment of all documents

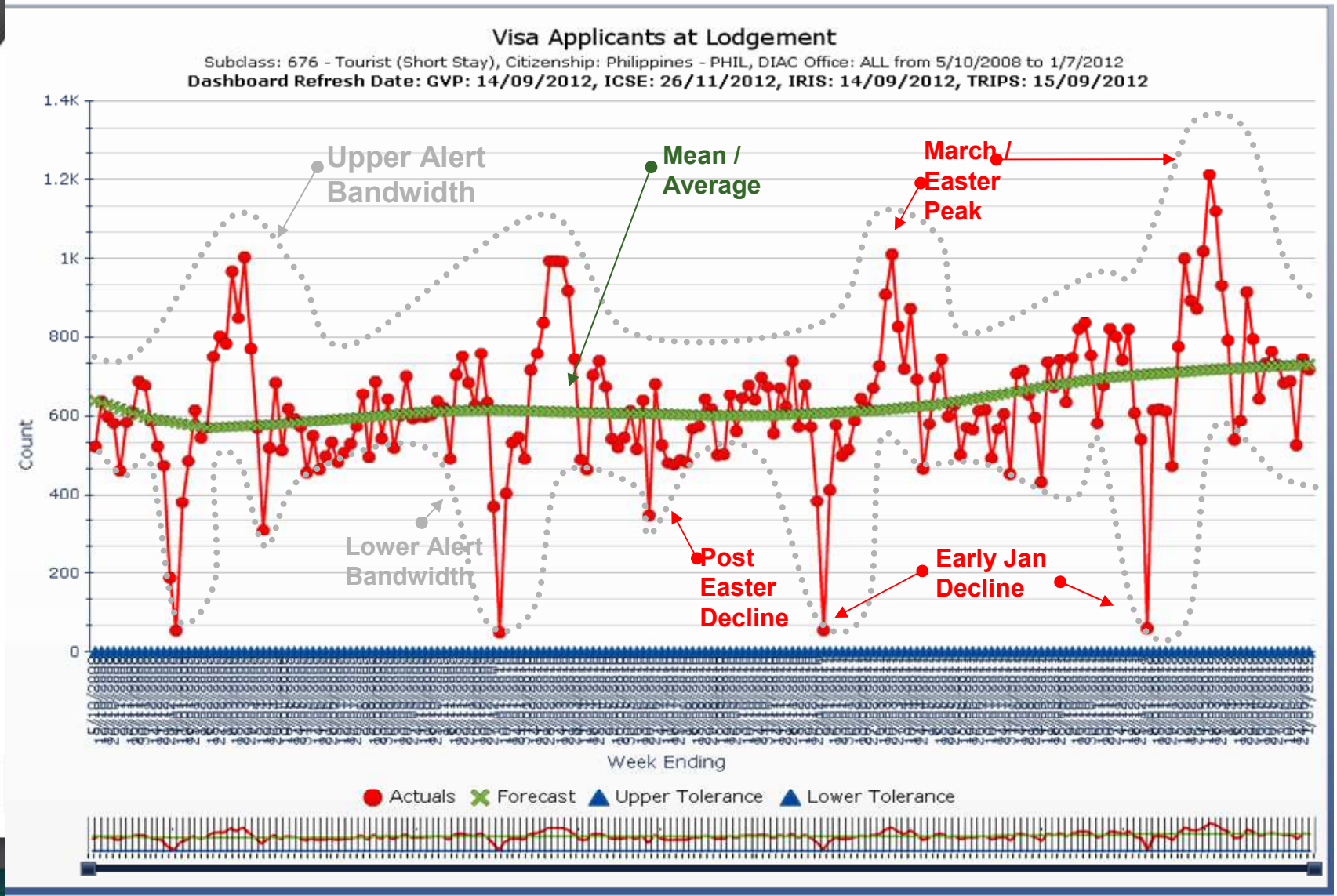
STANDARD RISK PROCESS

Partial assessment of documents

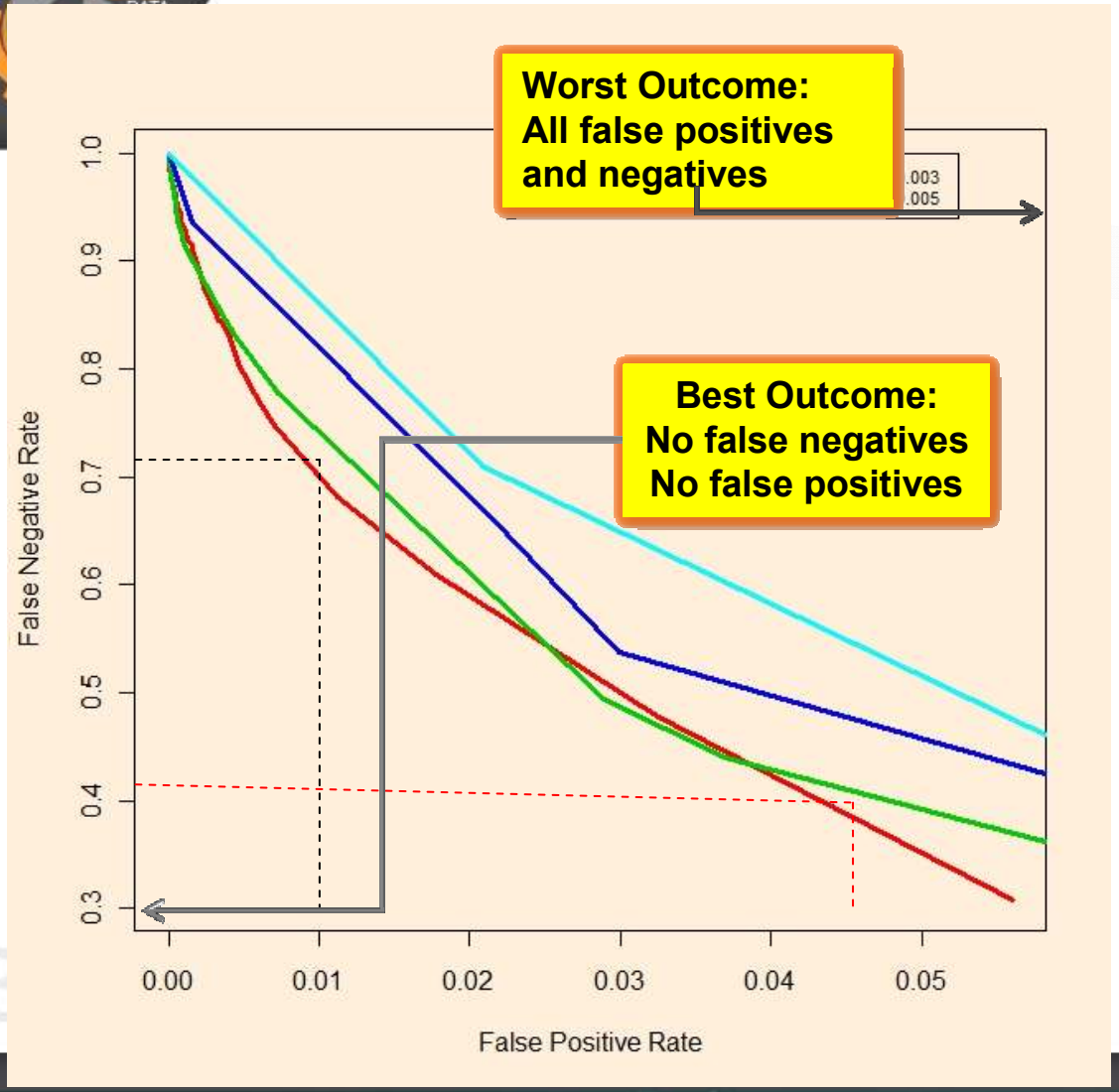
STREAMLINED PROCESS

Minor assessment of documents

Alert Dashboard



Predictive Models – Border Risk

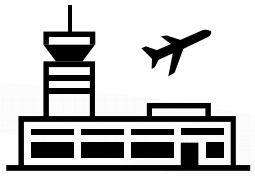


Using Historic data from December 2006 as a base:
- 2400 targets
- 678000 non-targets

Random (50:50)
30% target = 720
30% non-target = 203400
Using Red curve (our model):
Blue dotted line
30% target = 720
1% non-target = 6780
Red dotted line
60% target = 1440
4.5% non-target = 30510

BIG DATA

Check-In



Departure

DIAC Automated Risk Scoring Engine

Traveller Risk Scores

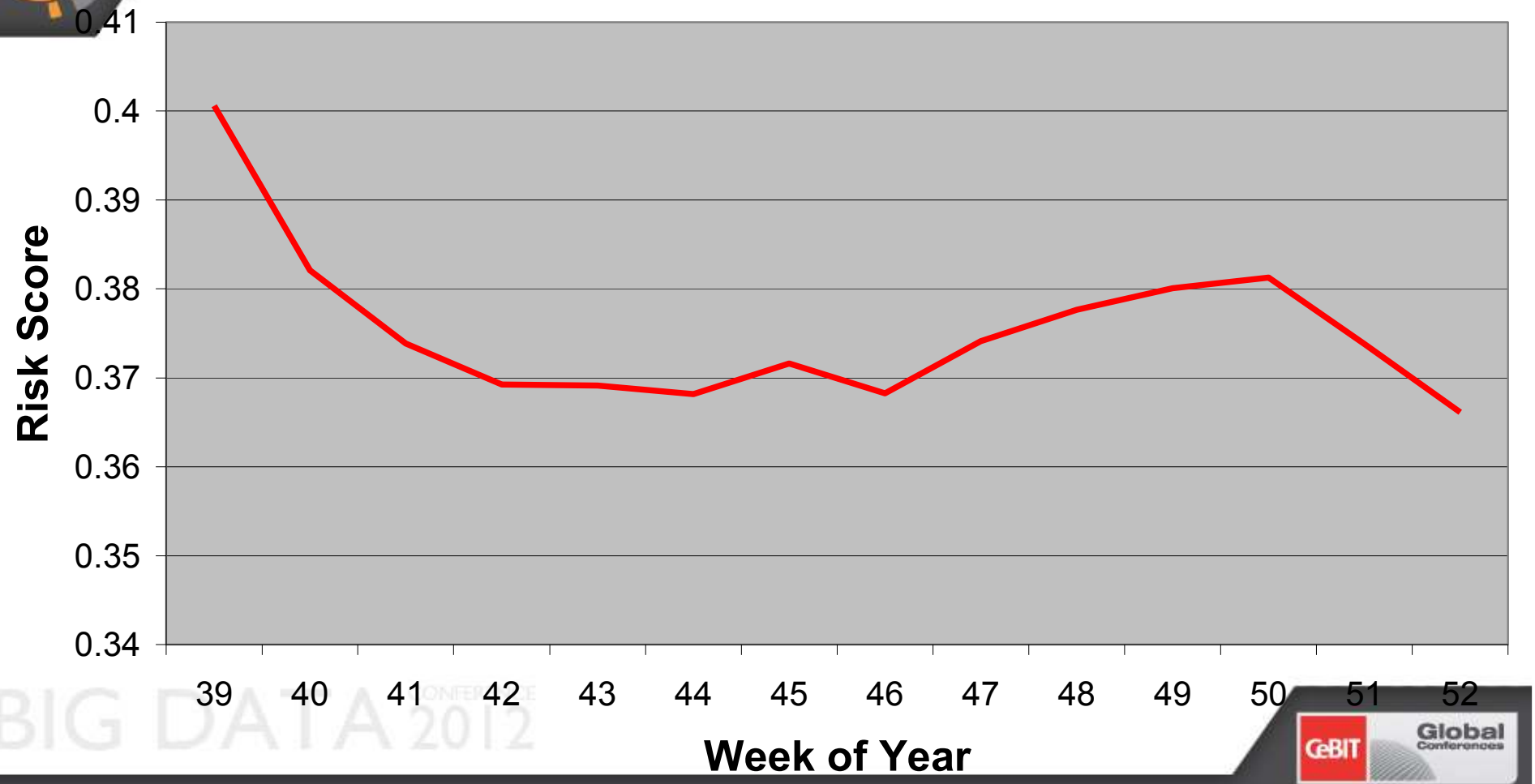
10 minutes after check-in

High Risk

Arrival

Low Risk

Average bona fide risk - 4th Quarter 2011



Not The Only Method

- The above process is not the only tool we use.
- Outlier detection. This looks for the ‘unusual’ pattern in the data.
- Flight Analysis:
  - looks at potential adverse patterns,
  - then looks for all the people connected to that adverse pattern on the same or related flights.
- Abnormal trend alerts:
  - Use historic data to forecast trends and then generate alerts when agreed tolerances are are breached.