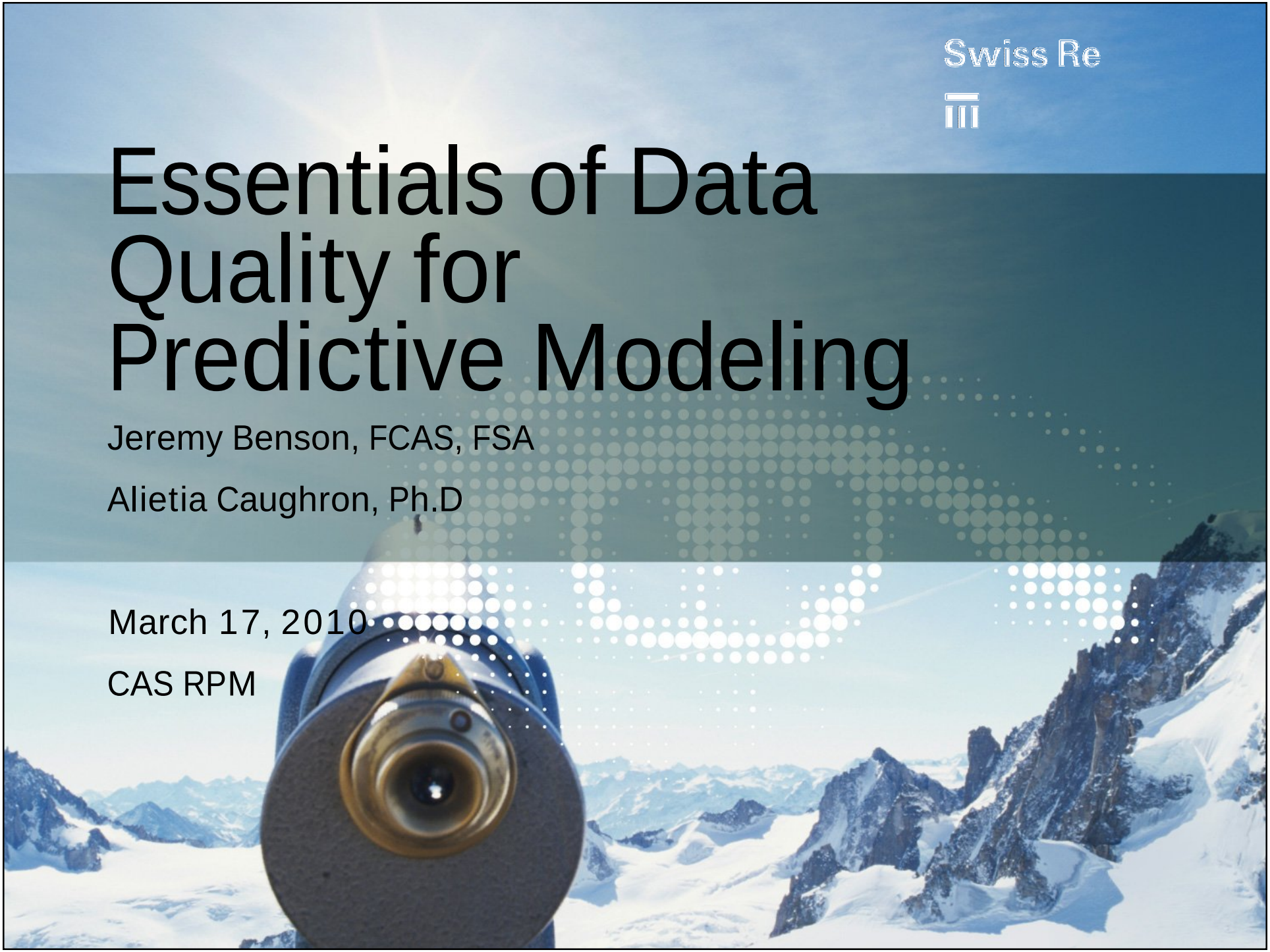


Swiss Re

Essentials of Data Quality for Predictive Modeling

Jeremy Benson, FCAS, FSA

Alietia Caughron, Ph.D

March 17, 2010

CAS RPM

Swiss Re

Antitrust Notice

- The Casualty Actuarial Society is committed to adhering strictly to the letter and spirit of the antitrust laws. Seminars conducted under the auspices of the CAS are designed solely to provide a forum for the expression of various points of view on topics described in the programs or agendas for such meetings.

- Under no circumstances shall CAS seminars be used as a means for competing companies or firms to reach any understanding – expressed or implied – that restricts competition or in any way impairs the ability of members to exercise independent business judgment regarding matters affecting competition.

- It is the responsibility of all seminar participants to be aware of antitrust regulations, to prevent any written or verbal discussions that appear to violate these laws, and to adhere in every respect to the CAS antitrust compliance policy.

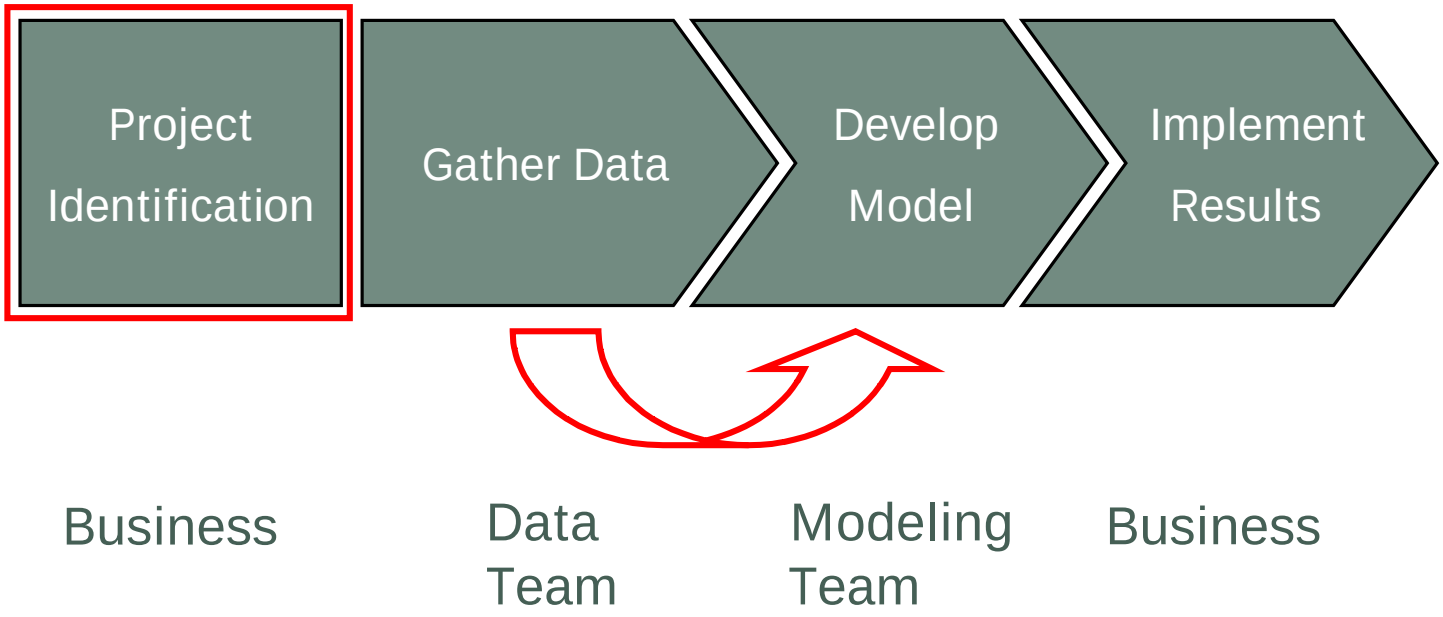
Swiss Re

Agenda

- Structure
- Goals of Data Quality
- Process
- Benefits of Data Quality
- Knowledge Transfer
- Lessons Learned

Swiss Re

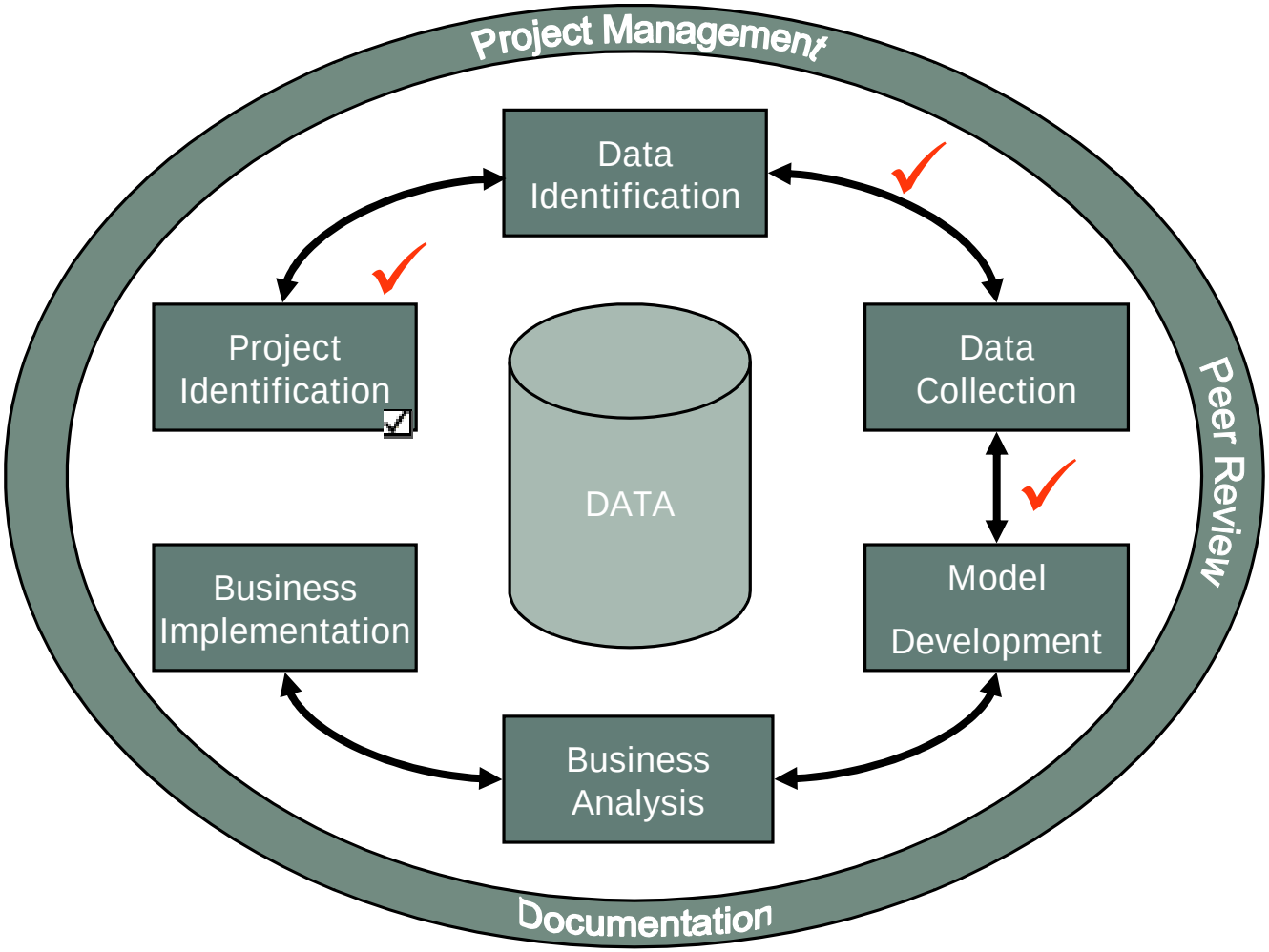
Project Structure, naïve version



Swiss Re

Project Structure, more realistic

- Iterative process
- Data stages are the most time consuming.
- Data Quality focus is part of the Data Collection stage.



Swiss Re

Separate Data and Modeling Teams

- Advantages
  - Increased focus on data quality
  - Different skills sets are needed for data management than for modeling
  - Data quality issues go beyond modeling
  - Data quality team can start next project earlier
- Disadvantages
  - Modeling team is not as knowledgable about the data, its strengths and weakness
  - Data team may have ‘cleaned’ the data introducing modeling bias
  - Knowledge transfer to the modeling team may be incomplete

Swiss Re

Overall Goals of Data Quality

- Accurate, Consistent, Complete Data
- The data should be appropriate for the purpose of the analysis
- Improved ability to explain and defend decisions
- Better decisions result from better data
- Actuary's time is freed up to focus on core professional responsibilities, decision and analysis

Goals of Data Quality for Predictive Modeling

Swiss Re

- Data – Fundamentally a statistical exercise that requires data

  Require there to be a Data Collection Plan.

  – "Clean" data – If data is missing, potential for lack of convergence, internal / external aliasing, biased results.

    Cleaning the data potentially introduces bias.

  – "Good" clean data – Quality of the data impacts models' predictive accuracy.

    Need to balance quality, materiality, and business needs.

- Documentation – Supports repeatability (model updates) & buy-in from others.
- Communication – Feedback loop improves Knowledge Transfer & assists in prioritizing efforts.

Swiss Re

Data Collection Process

- Data Profiling
- Data Preparation and Integration
- Data Quality Testing
- Data Scrubbing
- Documentation
- Communication

Swiss Re

Steps in Data Profiling

- Determine the Context of Data (Scoping)
- Define the Data (Operational Definitions)
- Determine the Types of Statistics
- Get the Data
- Profile and Document
- Interpret the Results
- Discuss with Functional/Business Experts
- Repeat
- Adjust Scope, Timelines or Resources

Swiss Re

Data Profiling

- Types of Data Profiling Statistics
  - Fill Rates, Min/Max, Frequency Distributions, Uniqueness
- Purpose
  - Understand data challenges early in the project so late surprises are minimized.
  - Use to prioritize data to use for analysis.
  - Assess the risk involved in integrating the data.
  - Help to prepare meaningful questions to ask subject matter experts.
  - Identify potential mapping issues between or within fields.

Swiss Re

Data Preparation and Integration

- Data Preparation
  - Get the Data
  - Primary Key Analysis
  - Aggregation/Duplication
  - Initial Data Scrubbing
  - Data Manipulation
  - Mapping
- Data Integration
  - Merge Data
  - Data Integration Tests
  - Reconciliation

Swiss Re

Data Quality Testing

- Data Profiling on Integrated Data
  - Fill Rates
  - Frequency Tests
  - Min/Max
- Business Rules
  - Loss Date should be after Effective Date
  - Claims should have a matching Policy
  - Tests for Negative Premium/Loss

Swiss Re

Data Scrubbing

- Normalizing Data
  - If fields are required to add to a certain value, adjusting the data so that they do.
- Imputations
  - Filling in of Missing Data
- Translations
  - Changing the value of a Data point to make it Consistent with other values that have the same meaning
- Cleaning
  - Correction of Erroneous Data
- Mapping
  - Process in which similar data is merged together and the values in the datasets are translated to become consistent with each other
  - Redefining the segmentation of data

Swiss Re

Documentation

- Clear and concise so that someone else can re-create the process
- Modeler should be able to understand the what each data element represents and any data scrubbing that took place for that element
- Any justification for changes to the dataset should be clearly documented
- Helps provide the modeler with a comfort level about the data.

Swiss Re

Communication Process

- IT – Determine how to access the data and obtain permission to the data
- Actuarial, Underwriting, Claims, Operations
  - Understand the intended uses of the data
  - Determination of the Scope of Records and Fields to use for Modeling
  - Input and Feedback on Results of Data Quality Testing
  - Drilldown into Root Cause Analysis of Data Quality Issues
- Project Sponsor
  - Set Overall Direction for Predictive Modeling Project

Swiss Re

Communication Process, predictive modeling team

- Provides direction for data collection & quality review, e.g. identifying “must-haves”, “nice-to-have”, “wish list”, and “not needed”
- Timeline management
- Facilitates Knowledge Transfer and mitigates concerns about separating the two functions (data collection/quality & modeling)

Swiss Re

Knowledge Transfer

- Identify issues that need to shared with the modeling team
- Know data issues and how they have been resolved
- Deepen understanding of what is meant by “data quality”
- Able to articulate specific requirements needed for data quality with respect to modeling projects
- Have a general sense of the data’s quality in order to identify strengths and weaknesses (of the model)

Swiss Re

Benefits of Data Quality for Predictive Modeling

- See the impact of Pre and Post Data Quality
- More time to focus on building the models
- Improvements in data that measurably impact the business can be taken care of because there are resources focused on the data
- Obtain early buy-in from the business

Swiss Re

Benefits of Data Quality

- Retrospective
  - Fixing Data Errors in Systems
  - Documentation of cleanup for other Actuarial Analysis
- Prospective
  - Involvement in System Architecture when setting up new system
  - Fix processes that caused data errors
  - Find errors before they adversely affect results
- Communication
  - Business Awareness of Data Quality Issues

Swiss Re

Lessons Learned

- Data Quality affects everyone – it is not just a business or IT issue
- Communication with Business experts is essential to understanding why there are errors in the data
- Time spent up-front on Data Profiling will save you time on the rest of the Data Collection
- Formalized Process for Data will produce both better data and data sooner
- It is important to have operational definitions
- If pulling from multiple datasets, mapping the datasets is HUGE
- Process and Communication are just as, if not more, important than the Data Quality Tests and Results.