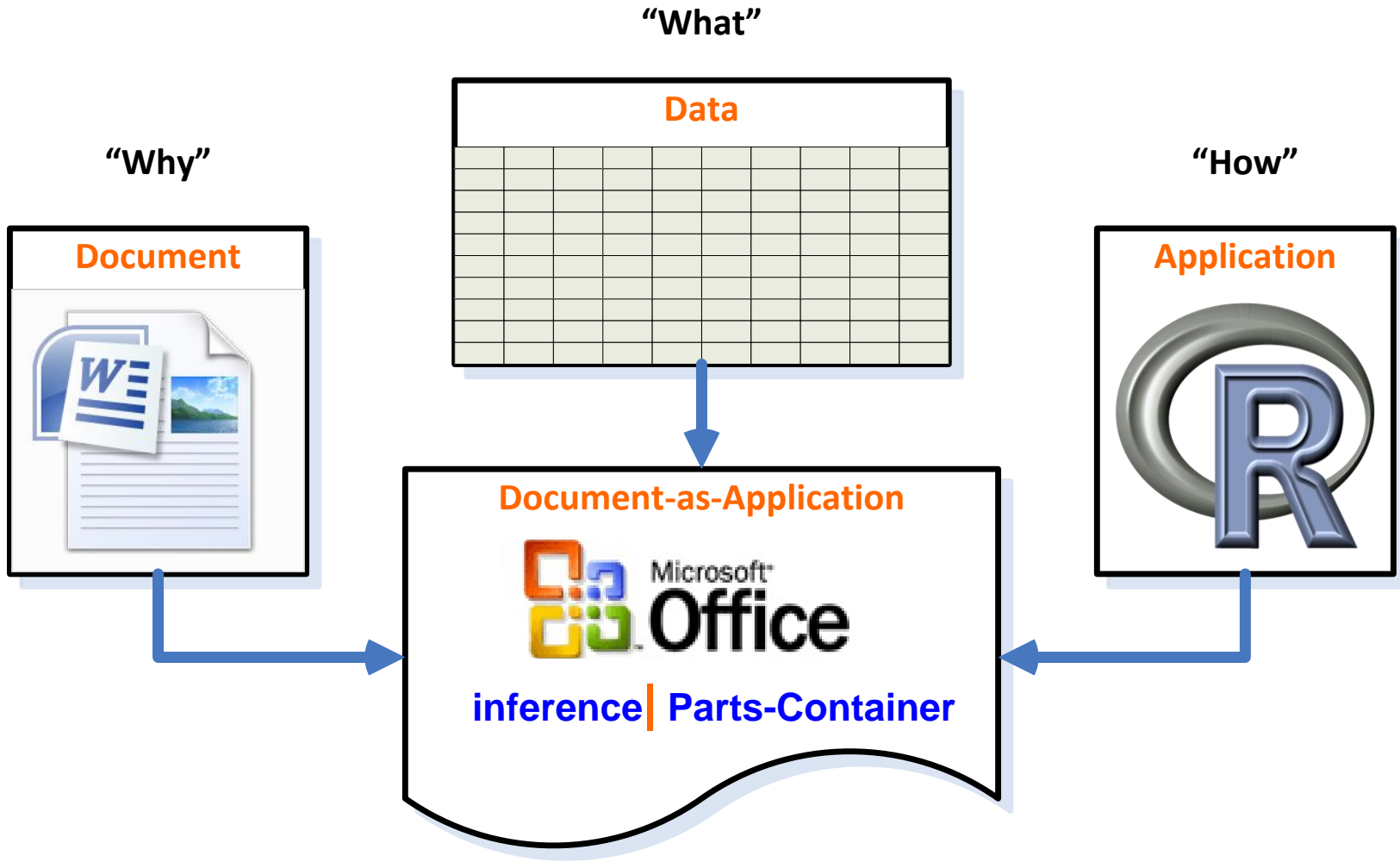


---

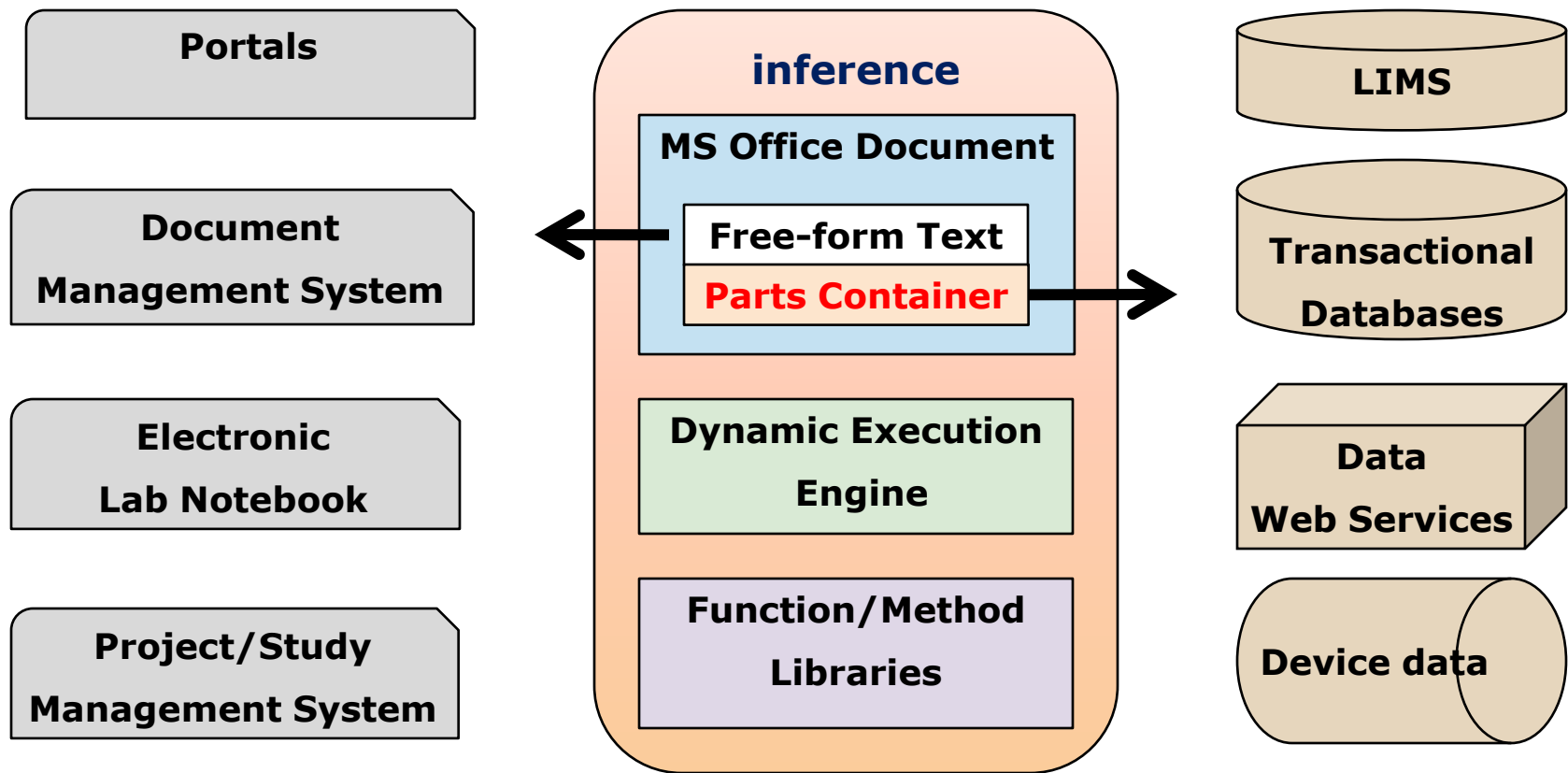
# Microsoft Office Documents as R-Applications

Josh van Eikeren and Paul van Eikeren  
userR! 2009 (July 8-10, 2009)

# Problem to be Solved



# Industry: Bridge Two Worlds



**Free-form Text**  
**Document-centric**

**Structured Data**  
**Process-centric**

# Solution approach



# inference in Word

# inference in Word: rationale

## ■ **Leverage Systems You Already Have**

- Microsoft Office on every desktop (98% of corporate desktops)
- 500 Million users of Microsoft Office

## ■ **Focus on Supporting the Work Process**

- Reusable protocols as “best practices”

## ■ **Documents as Work Process Solutions**

- Documents at the core of industry workflow
- Technical workers comfortable with Microsoft Office (Word, Excel and PowerPoint)
- Documents leverage existing systems (document management, SOPs, regulatory practice)
- Technical workers still like paper!

# inference in Word: elements

Inference |  
Task Pane

Manage  
Parts-  
Container

Edit Parts  
Properties


The screenshot shows the Microsoft Word interface with the Inference Word Task Pane on the left and the Office Document Canvas on the right. The task pane is divided into two main sections: 'Manage Parts Container' and 'Edit CodeBlock Properties'. The 'Manage Parts Container' section shows a tree view of the document's parts, including 'Data Sets', 'Objects', 'Code Blocks', and 'Expressions'. The 'Edit CodeBlock Properties' section allows for editing the label, figure size, and execution options for a selected code block. The Office Document Canvas displays the document content, which includes a title 'Analysis of Michelson Data' and three sections: 'Background', 'The Data Set', and 'The Analysis'. The 'The Data Set' section contains a paragraph of text and a code block for a boxplot. The 'The Analysis' section contains a paragraph of text and another code block for a boxplot. The status bar at the bottom indicates 'Page: 1 of 1' and 'Words: 13/163'.

Office  
Document  
Canvas

Expressions

Code Blocks

## Source document



---

**Quiver Plots: Visualization of Vector Fields**

**The Data Set**

<p>Vector fields play an important role in science and engineering. They are used to describe a wide variety of phenomena including fluid flow. To illustrate vector fields, let us consider a collection of <math>x, y, z</math> values that describe a surface net. The <math>z</math> value for each node is calculated from the function</p> $f(x, y) = (3x^2 + y) \times e^{-x^2 - y^2}$ <p>The net is displayed in three dimensions in the figure on the right. A vector field would correspond to the magnitude and direction of the slope of the net at each node.</p>	<pre>displayData()</pre>
--	--------------------------

**The Analysis**


<p>Large vector fields often exhibit quite complex structures, which can be difficult to reveal, making efficient visualization of a vector field an important analysis. A preferred 2-D visualization, illustrated on the right, corresponds to a "quiver" plot. You may recall that a quiver is a carrying case for arrows.</p>	<pre>2: display quiver plot quiver2(     f2,     x,     y,     color.palette=terrain.colors )</pre>
---	---

**The Interpretation**

The analysis calculates a  $dx$  and  $dy$  component of the slope at each node in the net. It then uses this information to define a vector (an arrow) at each  $x, y$  position, whose length and orientation correspond to the magnitude and direction of the slope, respectively

http://www.inference.us
Page 1
7/7/2009

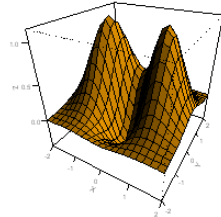
## Results document



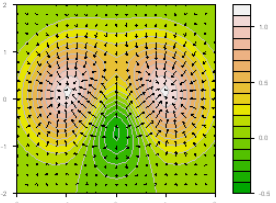
---

**Quiver Plots: Visualization of Vector Fields**

**The Data Set**

<p>Vector fields play an important role in science and engineering. They are used to describe a wide variety of phenomena including fluid flow. To illustrate vector fields, let us consider a collection of <math>x, y, z</math> values that describe a surface net. The <math>z</math> value for each node is calculated from the function</p> $f(x, y) = (3x^2 + y) \times e^{-x^2 - y^2}$ <p>The net is displayed in three dimensions in the figure on the right. A vector field would correspond to the magnitude and direction of the slope of the net at each node.</p>	
--	---

**The Analysis**

<p>Large vector fields often exhibit quite complex structures, which can be difficult to reveal, making efficient visualization of a vector field an important analysis. A preferred 2-D visualization, illustrated on the right, corresponds to a "quiver" plot. You may recall that a quiver is a carrying case for arrows.</p>	
---	--

**The Interpretation**

The analysis calculates a  $dx$  and  $dy$  component of the slope at each node in the net. It then uses this information to define a vector (an arrow) at each  $x, y$  position, whose length and orientation correspond to the magnitude and direction of the slope, respectively

http://www.inference.us
Page 1
7/7/2009

# inference in Word: tailored solution

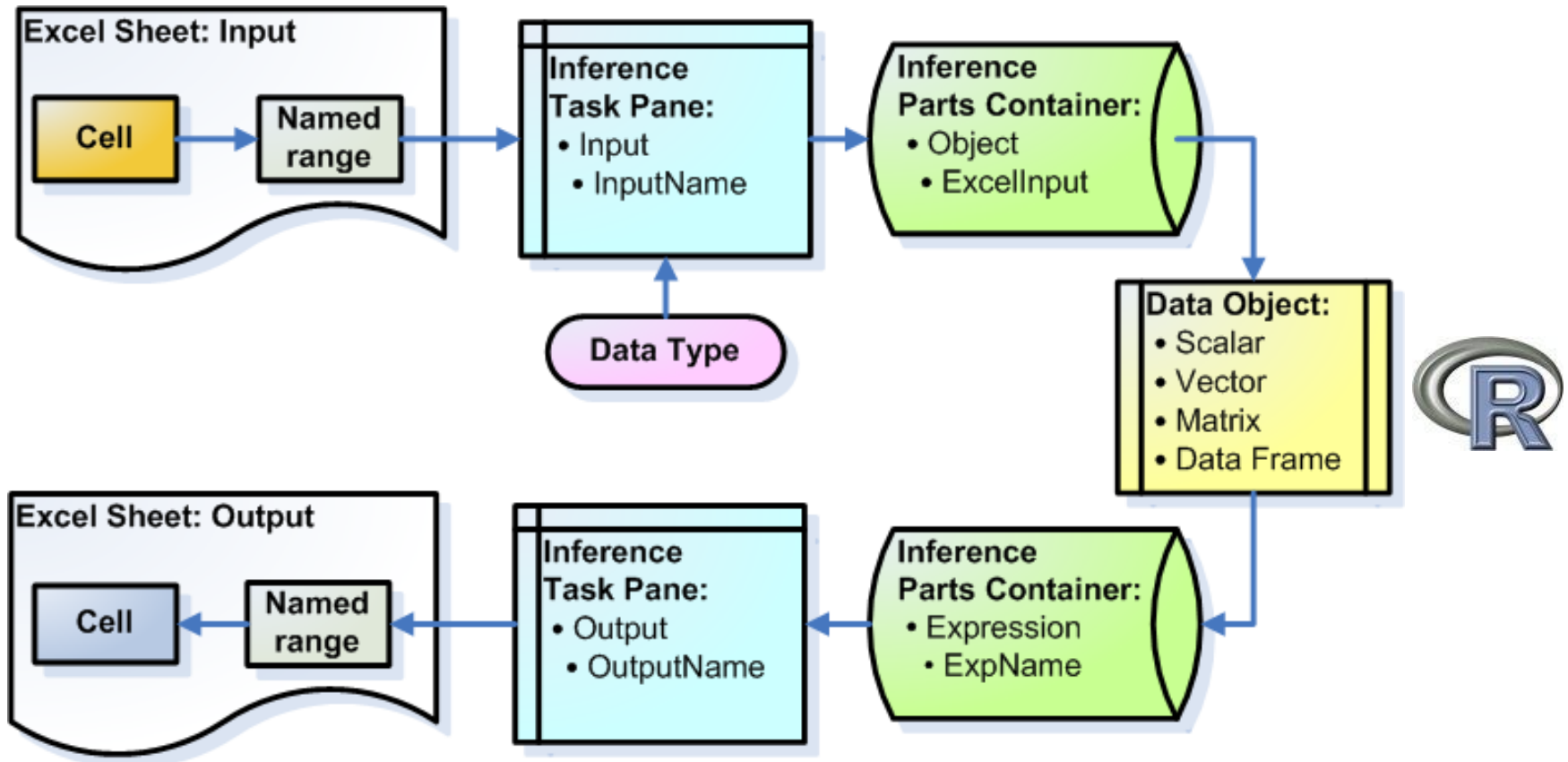
- ❑ Users only see the capability that they need
- ❑ Shallow learning curve using Office
- ❑ Data integration and management tightly integrated
- ❑ Analysis protocols provide audit trail for how results obtained
- ❑ Analysis and documentation occur concurrently
- ❑ Automated generation of Results Documents
- ❑ Adaptable and extensible using the Inference platform
- ❑ Inexpensive to set up and maintain because Office already deployed

# inference in Excel

# inference in Excel: elements

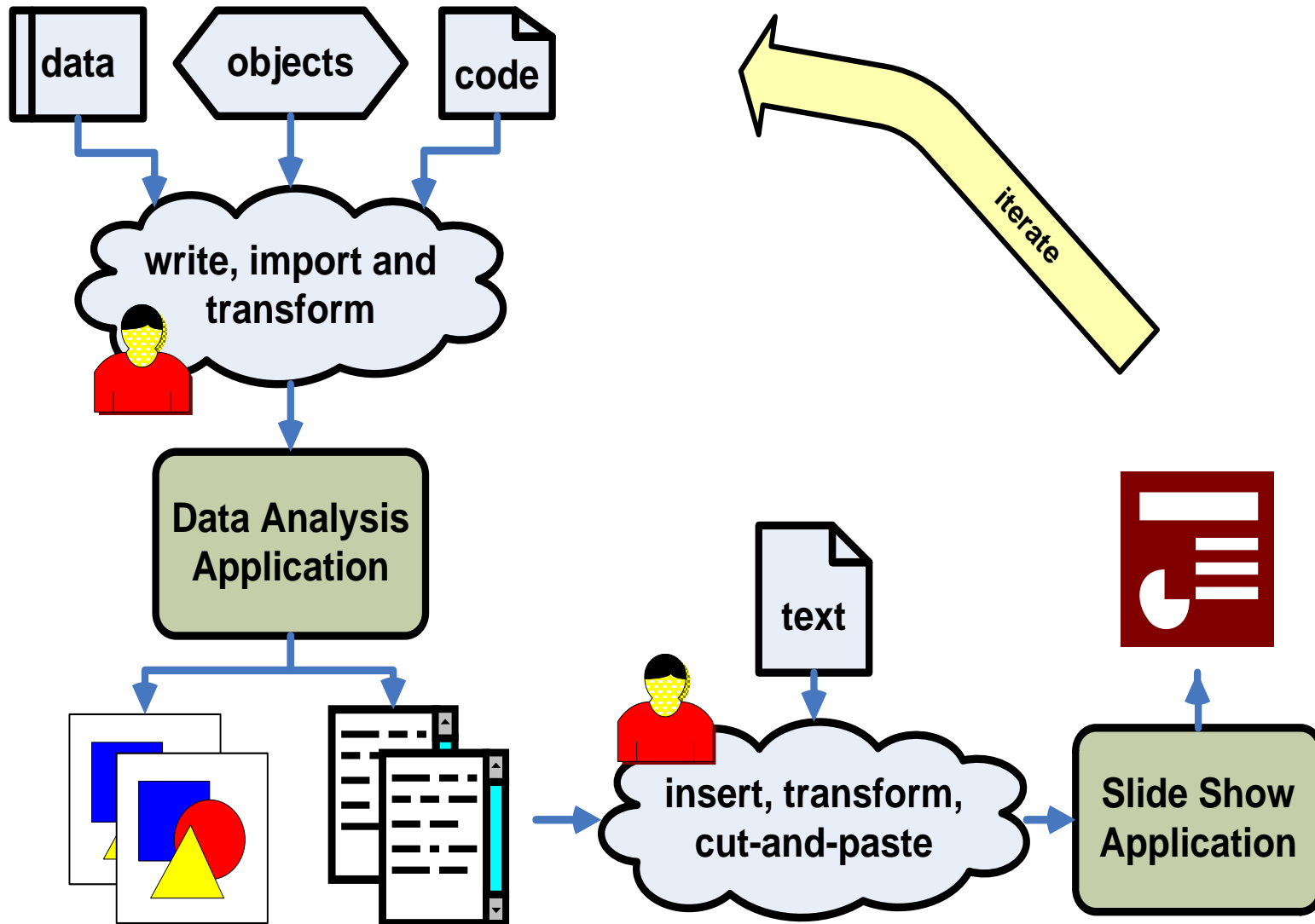
The screenshot shows the Microsoft Excel interface with the Inference add-in. The ribbon includes Home, Insert, Page Layout, Formulas, Data, Review, View, Developer, Add-Ins, Inference, and Acrobat. The Inference ribbon contains buttons for 'Add Container', 'Remove Container', 'Show Task Pane', 'Run Now', 'Clear Results', 'To Results Document View', 'To Results Document File', 'To Object File', 'To Object Service', 'Up-Convert to Office 2007', 'Down-Convert to Office 2003', 'Tools', 'Inference Getting Started', and 'Help and Support'. The task pane on the left is titled 'inference from Blue Reference' and shows a 'Parts Container' with a tree view of objects and code blocks. The main worksheet area displays a candlestick chart for MRK stock from 2009-01-02 to 2009-07-06. The chart shows a price range from approximately 20 to 32, with a volume bar chart below it showing trading volume in millions, with a peak of 23,135,700. The status bar at the bottom indicates 'Ready' and '100%' zoom.

# inference in Excel: data flow



# inference in PowerPoint

# PowerPoint: conventional approach



# inference in PowerPoint: elements

TurboCharge Stock Analysis.pptx - Microsoft PowerPoint

Home Insert Design Animations Slide Show Review View Developer Inference Acrobat

Add Parts Container Remove Parts Container Show Task Pane Run Now Clear Results To Results Document View To Results Document File To Object File To Service Up-Convert to Office 2007 Down-Convert to Office 2003 Tools Inference Getting Online Started Help and Support

**inference**  
from Blue Reference

## Summary & Candle Charts

■ Summary Statistics      ■ Candle chart

Stock Symbol: <stockSymbol>  
Date Range: <dateRange>  
Data Source: <dataSource>

Summary Statistics:  
<summaryStatistics>

<add Volume>

Platform: R  
References: + x Check  
quantmod  
TTR  
xts

Slide 4 of 6 "Presentation Template" 82%

# inference in PowerPoint: results

TurboCharge Stock Analysis.pptx - Microsoft PowerPoint

Home Insert Design Animations Slide Show Review View Developer **Inference** Acrobat

Add Parts Container Remove Parts Container Show Task Pane Run Now Clear Results To Results Document View To Results Document File To Object File To Service Up-Convert to Office 2007 Down-Convert to Office 2003 Tools Inference Getting Online Started Help and Support

**inference** from Blue Reference

**Summary & Candle Charts** **inference** from Blue Reference

■ Summary Statistics ■ Candle chart

Stock Symbol: GOOG  
Date Range: 2007-01-03 to 2009-07-07  
Data Source: yahoo

Summary Statistics:

dates	GOOG.Open	GOOG.High	GOOG.Low
Min. :2007-01-03	Min. :262.5	Min. :269.4	Min. :247.3
1st Qu.:2007-08-19	1st Qu.:402.8	1st Qu.:409.1	1st Qu.:394.9
Median:2008-04-05	Median:473.0	Median:479.1	Median:468.6
Mean :2008-04-04	Mean :475.1	Mean :481.7	Mean :467.9
3rd Qu.:2008-11-17	3rd Qu.:536.5	3rd Qu.:548.5	3rd Qu.:526.6
Max. :2009-07-07	Max. :741.1	Max. :747.2	Max. :725.0

GOOG.Close	GOOG.Volume	GOOG.Adjusted
Min. :257.4	Min. :1628400	Min. :257.4
1st Qu.:400.1	1st Qu.:3690625	1st Qu.:400.1
Median:472.8	Median:5022200	Median:472.8
Mean :474.6	Mean :5579164	Mean :474.6
3rd Qu.:535.4	3rd Qu.:6662025	3rd Qu.:535.4
Max. :741.8	Max. :23287300	Max. :741.8

GOOG [2009-01-02:2009-07-07]  
Last 396.63  
Volume (millions): 3,259,800

Jan 02 2009 Feb 17 2009 Mar 30 2009 May 11 2009 Jun 22 2009

©2009 Blue Reference, Inc. All rights reserved.

Stock Symbol: GOOG  
Date Range: 2007-01-03 to 2009-07-07  
Data Source: yahoo

Slide 4 of 6 "Presentation Template"

# inference in PowerPoint: message

- Lasting presentations
  - all information in one place
  - accessible raw data
  - symbolic and executable solution
  - automated results
  - reproducible and reusable
- Unified and integrated assembly process
  - better
  - faster
  - cheaper
- Presentation with lasting impact
  - no “walled gardens”
  - enables practice of “don’t tell me, show me”
  - learn from “view source” concept

Integrated Development Environment

**inference Studio**

The screenshot displays the Inference Studio interface for a file named "Michelson Data.infcontainer". The interface is divided into several sections:

- Toolbar:** Includes buttons for "New", "Delete", "Move Up", "Move Down", "Paste", "Copy", "Undo", "Redo", "Find", "Replace", "Insert Snippet", "Surround With", "Indent", "Outdent", "Comment", "Uncomment", "Start", "Stop", "Step", "Add", "Delete", "Delete All", "View", and "Help".
- Code Blocks:** A list on the left shows three blocks: "1: Display Code", "2: Display Graphic" (highlighted), and "3: Code Block".
- Code Editor:** Contains the following R code:

```
1 boxplot(  
2   Speed ~ Expt,  
3   data=MichelsonData,  
4   main="Speed of Light Data",  
5   xlab="Experiment No.",  
6   col="yellow"  
7 )
```
- References:** Shows a table with one entry: "xtable".
- Code Results:** Displays the output of the code block "2: Display Graphic". It includes a table of statistics and a boxplot.

	[,1]	[,2]	[,3]	[,4]	[,5]
[1,]	740	760	840	720	740
[2,]	850	800	840	765	805
[3,]	940	845	855	815	810
[4,]	980	890	880	870	870
[5,]	1070	960	910	920	950

attr(,"class")  
1  
"integer"  
\$n  
[1] 20 20 20 20 20  
\$conf  
[1,] 894.0712 813.2031 840.868 777.9036 787.0356  
[2,] 985.9288 876.7969 869.132 852.0964 832.9644
- Figure 1:** A boxplot titled "Speed of Light Data" showing the distribution of "Speed" for five "Experiment No."s (1 to 5). The boxes are yellow, and the plot includes whiskers and individual data points.

- Excel can be used to manage data frames
- Data frame object: R, .NET and SDML instances

## inference data integration

# Dataframe container w/ dictionary

The screenshot shows an Excel spreadsheet with the following data table:

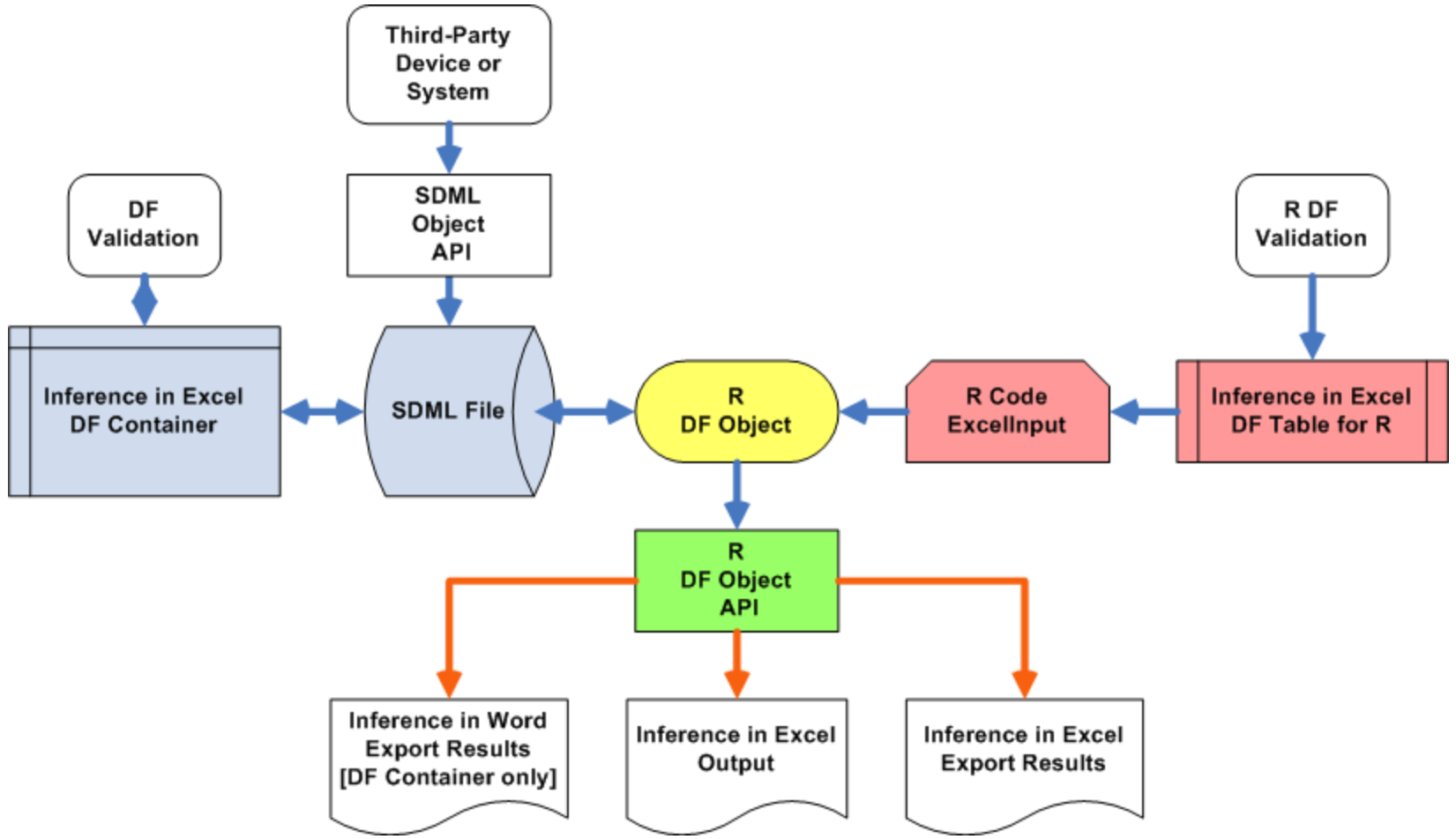
variable	index	runID	block	repID	pyrrolidone	teaHF	holdTime	holdTemp	yield
1	31	B4	23	5	1.335	25	10	75.82	
2	3	B1	1	4	1.17	22	15	82.93	
3	4	B1	3	4	1.17	28	15	88.07	
4	22	B1	5	4	1.5	22	15	88.68	
5	25	B1	7	4	1.5	28	15	93	
6	28	B1	9	6	1.17	22	15	77.21	
7	30	B1	11	6	1.17	28	15	83.6	
8	5	B1	13	6	1.5	22	15	84.86	
9	29	B1	15	6	1.5	28	15	88.71	
10	16	B2	17	3	1.335	25	20	94.13	
11	11	B2	18	7	1.335	25	20	89.94	
12	14	B2	19	5	1.005	25	20	88.21	
13	24	B2	20	5	1.665	25	20	93.11	
14	1	B2	21	5	1.335	19	20	89.78	
15	10	B2	22	5	1.335	31	20	94.61	
16	7	B2	25	5	1.335	25	20	93.32	
17	23	B2	25	5	1.335	25	20	92.32	
18	19	B2	25	5	1.335	25	20	93.68	
19	9	B2	25	5	1.335	25	20	93.27	
20	27	B2	25	5	1.335	25	20	92.87	
21	6	B2	25	5	1.335	25	20	92.96	
22	15	B2	25	5	1.335	25	20	93.07	
23	18	B3	2	4	1.17	22	25	94.04	
24	17	B3	4	4	1.17	28	25	93.97	
25	21	B3	6	4	1.5	22	25	94.3	
26	12	B3	8	4	1.5	28	25	93.42	
27	26	B3	10	6	1.17	22	25	92.99	
28	20	B3	12	6	1.17	28	25	94.38	
29	13	B3	14	6	1.5	22	25	94.26	
30	2	B3	16	6	1.5	28	25	94.66	
31	8	B5	24	5	1.335	25	30	93.25	

The 'inference' task pane on the left shows a 'DataFrame Container' with the following structure:

- Columns
  - runID (numeric-integer)
  - block (categorical) (5 Labels)
    - repID (numeric-integer)
      - min = 1
      - max = 25
    - pyrrolidone (numeric-real)
      - min = 3
      - max = 7
    - teaHF (numeric-real)
      - min = 1.005
      - max = 1.665
    - holdTime (numeric-real)
      - min = 19
      - max = 31
    - holdTemp (numeric-real)
      - min = 10
      - max = 30
    - yield (numeric-real)
      - min = 75.82
      - max = 94.66
    - startMat (numeric-real)
      - min = 0
      - max = 22.85
    - impurity (numeric-real)
      - min = 0.21
      - max = 5.29
  - Attributes
    - authors
    - title
    - source
    - year

The 'Selected Column Info' section at the bottom of the task pane indicates: "No column selected."

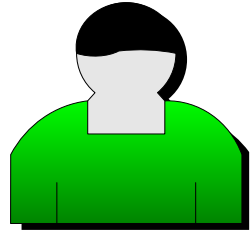
# Dataframe objects: R .NET SDML



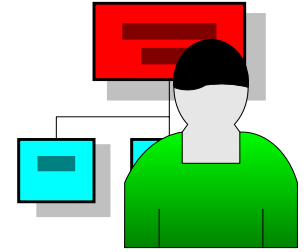
Test Driven Development

# inference applications

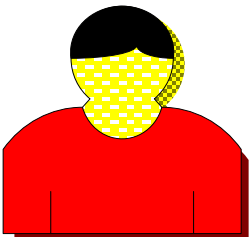
# Test-Driven Development: Workflow



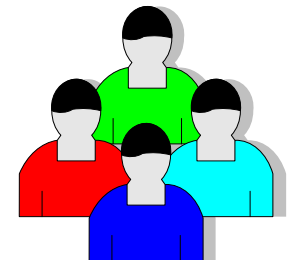
Developer/Analyst



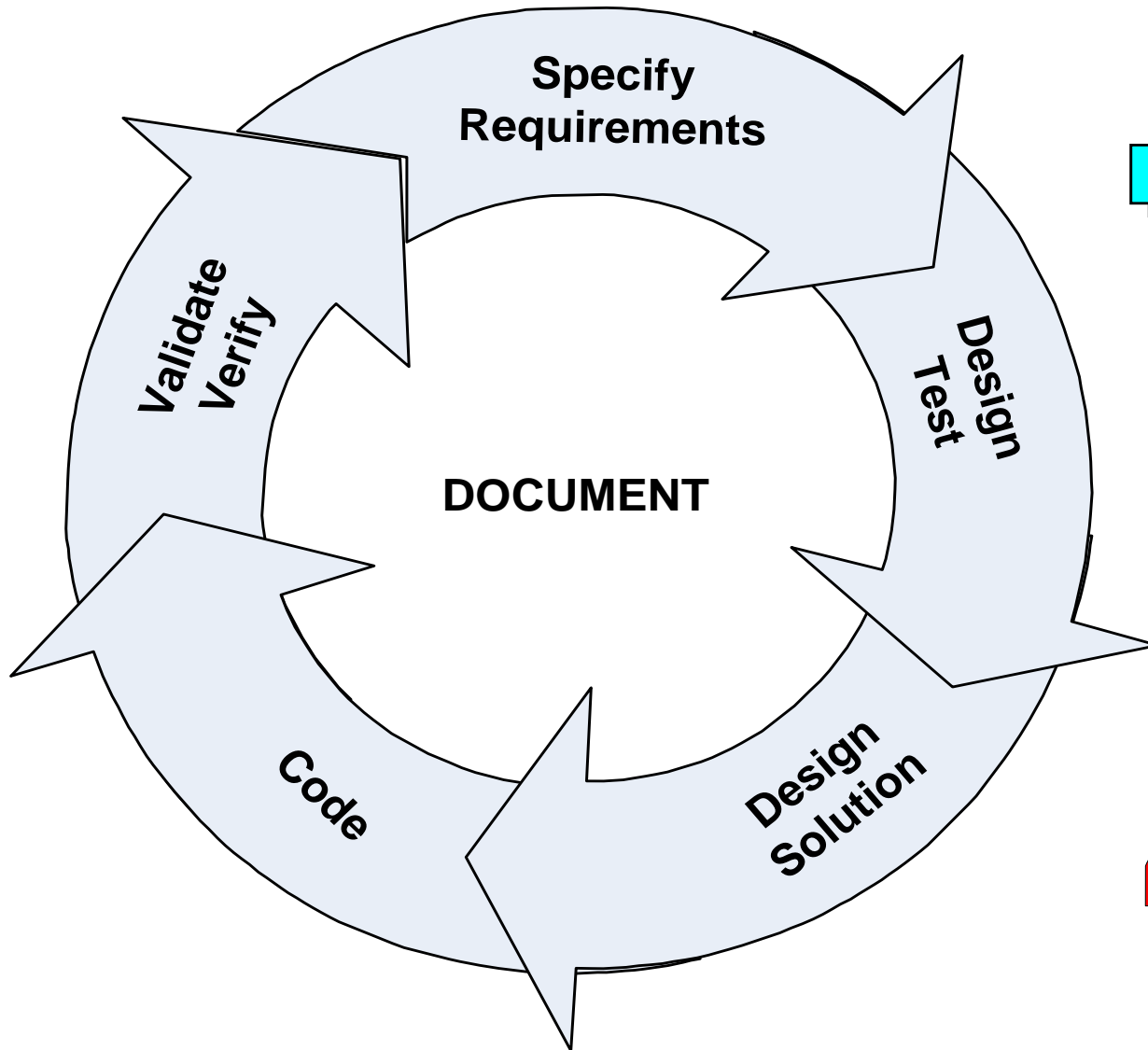
Business Systems



Domain Expert



Customers



# Test-Driven Development: benefits

- Benefits from integration
- Involves directed workflow and deliverables
  - Collect requirements and test data
  - Define test scripts
  - Specify code and run test scripts
  - Collect and report test verdicts (pass or fail)
  - Document entire process
  - Support collaboration among stakeholders
- Deliverables are basis for validation and verification

# 1. Requirements and Test Data

Microsoft Word window: INF03 Test-Driven Data-Analysis Development A.docx - Microsoft Word

**inference**  
from Blue Reference

## Data Analysis Method Development

Business User	Developer/Analyst	Business Systems
Bill Murray Biology Assays	John Smith Non-clinical Statistics	Paul van Eikeren Quality Assurance

### New Function Requirements

#### Application Requirements

Many biological targets are enzymes. Screening studies involve examining the effect of treatments on the kinetic behavior of such enzymes where the behavior is characterized by changes in the reaction velocity as a function of increasing substrate concentration. Changes in behavior are analyzed in term of the Michealis-Menten relationship reflected in the estimated value of the maximum velocity ( $V_m$ ) and Michealis-Menten constant ( $K$ ) for the untreated and treated cases.

$$V = \frac{V_{\max} S}{K_M + S} + \epsilon$$

In order to properly interpret changes in behavior it is valuable to obtain an outline of the approximate pairwise confidence region ( $V_m$  vs  $K$ ) for the nonlinear model fit, using nls in R, for the untreated and treated cases on the same plot.

#### Function and Interface Requirements

Need a function that outputs the a single plot showing the ellipsoidal confidence region for the  $V_m$  and  $K$  pair estimated for the nonlinear fit of enzyme initial rate versus substrate concentration for in the absence and presence of treatment. The function interface needs to consist of the following:

- `fit1` and `fit2` corresponding to nonlinear model objects corresponding to the untreated and treated cases, respectively.
- `level` corresponding to the confidence level of the region.

#### Functional Test Specifications

TestDataset contains the results of a typical treatment study. Fit the untreated and treated data to the Michaelis-Menten model using nonlinear least squares. Use the models in conjunction with the new function to display a plot of the confidence regions.

**Data Set Properties**

Label: TestDataset  
 Filename: INF03 Enzyme Kinetics galactosyltra  
 Format: Inference in Excel 2007 DataFrame

Page: 1 of 1 Words: 262

# 2. Define/Run Functional Tests

The screenshot shows a Microsoft Word document with the 'Inference' ribbon active. The document content is as follows:

**Data Analysis Method Development**

Business User	Developer/Analyst	Business Systems
Bill Murray Biology Assays	John Smith Non-clinical Statistics	Paul van Eikeren Quality Assurance

**New Function Requirements**

**Application Requirements**

Many biological targets are enzymes. Screening studies involve examining the effect of treatments on the kinetic behavior of such enzymes where the behavior is characterized by changes in the reaction velocity as a function of increasing substrate concentration. Changes in behavior are analyzed in term of the Michaelis-Menten relationship reflected in the estimated value of the maximum velocity ( $V_m$ ) and Michaelis-Menten constant (K) for the untreated and treated cases.

$$V = \frac{V_m S}{K_m + S} + \epsilon$$

In order to properly interpret changes in behavior it is valuable to obtain an outline of the approximate pairwise confidence region ( $V_m$  vs K) for the nonlinear model fit, using `nls` in R, for the untreated and treated cases on the same plot.

**Function and Interface Requirements**

Need a function that outputs the a single plot showing the the ellipsoidal confidence region for the  $V_m$  and K pair estimated for the nonlinear fit of enzyme initial rate versus substrate concentration for in the absence and presence of treatment. The function interface needs to consist of the following:

- `fit1` and `fit2` corresponding to nonlinear model objects corresponding to the untreated and treated cases, respectively.
- `level` corresponding to the confidence level of the region.

**Functional Test Specifications**

`TestDataset` contains the results of a typical treatment study. Fit the untreated and treated data to the Michaelis-Menten model using nonlinear least squares. Use the models in conjunction with the new function to display a plot of the confidence regions.

**Functional Test**

**Step 1: Fit Test Data to Michaelis-Menten Models**

Calculate model for the untreated data:

```
model1 <- nls(
  rate ~ Vm*conc/(K + conc),
  data = TestDataset,
  subset = state=="untreated",
  start = list(K = 0.05, Vm = 200)
)
summary(model1)

Formula: rate ~ Vm * conc/(K + conc)

Parameters:
  Estimate Std. Error t value Pr(>|t|)
K  6.412e-02  8.281e-03  7.743 1.57e-05 ***
Vm  2.127e+02  6.947e+00  30.615 3.24e-11 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 10.03 on 10 degrees of freedom

Number of iterations to convergence: 5
Achieved convergence tolerance: 8.824e-06
```

**Calculate model for the treated data:**

```
model2 <- nls(
  rate ~ Vm*conc/(K + conc),
  data = TestDataset,
  subset = state=="treated",
  start = list(K = 0.05, Vm = 200)
)
summary(model2)

Formula: rate ~ Vm * conc/(K + conc)

Parameters:
  Estimate Std. Error t value Pr(>|t|)
K  4.771e-02  7.782e-03  6.131 0.00173 ***
Vm  1.603e+02  6.480e+00  24.734 1.38e-09 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 9.773 on 9 degrees of freedom

Number of iterations to convergence: 6
Achieved convergence tolerance: 1.566e-06
```

**Step 2: Create Plot Comparing Confidence Regions of Models**

```
CR.compare(fit=model1, fit=model2, level=0.99)
```

Page 1 / 7/7/2009

Page 2 / 7/7/2009

# 3. Code, Test & Report Results

The screenshot displays a Microsoft Word document titled "INF03 Test-Driven Data-Analysis Development C.docx". The ribbon includes tabs for Home, Insert, Page Layout, References, Mailings, Review, View, Developer, Add-Ins, and Inference. The Inference tab contains various tools like "Add Parts Container", "Remove Parts Container", "Show Task Pane", "Run Now", "Clear Results", "To Results Document View", "To Results Document File", "To Object File", "To Service", "Up-Convert to Office 2007", "Down-Convert to Office 2003", "Tools", "Inference Online", "Getting Started", and "Help and Support".

The document is split into two pages. The left page (Page 3) contains the following content:

**New Function Development and Documentation**

**Approach**

The approach is based on the ellipse R-package comprised of various routines for drawing ellipses and ellipse-like confidence regions. Implementation based on plots described in Murdoch and Chow (1996), "A graphical display of large correlation matrices," *The American Statistician* 50, 178-180.

**CR.compare Function**

The function CR.compare produces ellipsoidal outlines of the approximate pairwise confidence regions for the two nonlinear fits. It takes three arguments:

- fit1 corresponding to the nonlinear least squares object for the Michaelis-Menten model of the enzyme kinetics without treatment
- fit2 corresponding to the nonlinear least squares object for the Michaelis-Menten model of the enzyme kinetics in the presence of treatment
- level corresponding to the confidence level for the two regions expressed as a fraction (default equals 95%)

```

CR.compare <- function (fit1,fit2,level)
{
  # get Vm and K from fits
  params1 <- fit1$m$getParams()
  params2 <- fit2$m$getParams()
  # plot fit1 ellipse
  plot(
    ellipse(fit1,whichec("Vm","K"),level=level),
    xlim=c(140,240),
    ylim=c(0,0.1),
    type="l",
    col="darkgreen"
  )
  # add Vm and K of fit1 to plot
  points(params1["Vm"],params1["K"],pch=15, cex=2, col="darkgreen")
  # add plot of fit2 ellipse
  lines(
    ellipse(fit2,whichec("Vm","K"), level=level),
    type="l",
    col="red"
  )
  # add Vm and K of fit2 to plot
  points(params2["Vm"],params2["K"],pch=15, cex=2, col="red")
}

```

**Execute Functional Test**

```

CR.compare(fit1=model1,fit2=model2,level=0.95)

```

Page 3 footer: <http://www.inference.us> Page 3 7/7/2009

The right page (Page 4) contains a plot of two confidence ellipses. The x-axis is labeled "Vm" and ranges from 140 to 240. The y-axis is labeled "K" and ranges from 0.00 to 0.10. A red ellipse is centered around Vm ≈ 160 and K ≈ 0.05. A green ellipse is centered around Vm ≈ 210 and K ≈ 0.07. A red square and a green square mark the center of each ellipse, respectively.

Page 4 footer: <http://www.inference.us> Page 4 7/7/2009

Study Management

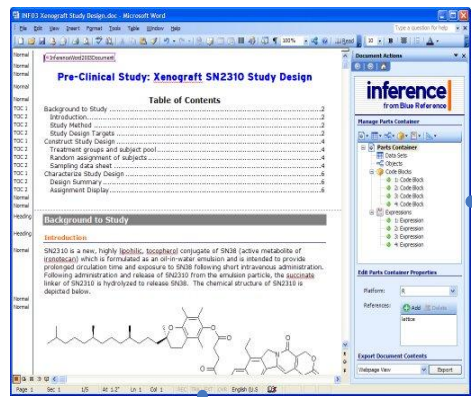
**inference application**

# Preclinical Study Management

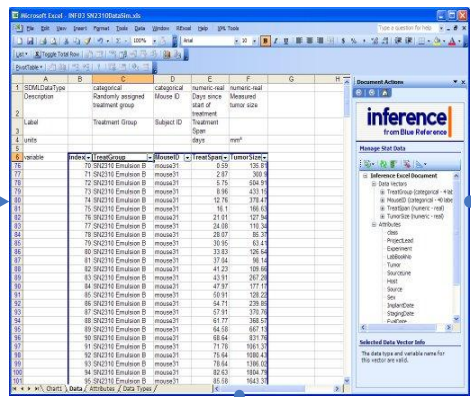
- Test safety and efficacy of new drug candidates in animals before testing in humans
- Scientists rely on paper and Excel
- Paper and Excel inefficiencies:
  - lack of easy-to-use tools for study design
  - lack of collaboration support
  - need to transcribe and merge data
  - data not accessible in real time
  - manual creation of reports

# inference in Study Management

## Study Design

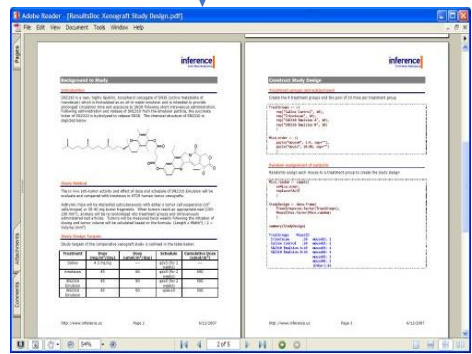


## Design Data

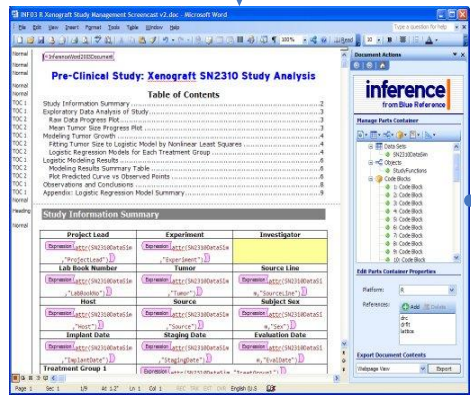


## Data Export

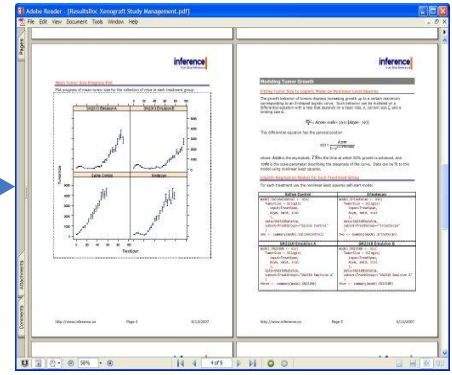
StatDataML  
XML  
Spreadsheet  
Text File  
Data Base



## Design Report



## Design Analysis



## Analysis Report

# inference for QbD study management

Inference  
Excel

Material	Process	Supplier	Material	Material	Material	Material	Material	Material	Material	Material	Material	Material	Material	Material	Material	Material	Material	Material	Material	Material
APR0004	Small	17%	James Inc.	James Inc.	James Inc.	James Inc.	James Inc.	James Inc.	James Inc.	James Inc.	James Inc.	James Inc.	James Inc.	James Inc.	James Inc.	James Inc.	James Inc.	James Inc.	James Inc.	James Inc.
APR0005	Small	15%	James Inc.	James Inc.	James Inc.	James Inc.	James Inc.	James Inc.	James Inc.	James Inc.	James Inc.	James Inc.	James Inc.	James Inc.	James Inc.	James Inc.	James Inc.	James Inc.	James Inc.	James Inc.

Manufacturing  
Data 1

2

3

Inference Word Templates

4

5

**Quality by Design Template Investigate Data Quality by Profiling**

Table of Contents

- Data Set Provenance and Scope
- Listing of Variables and Their Attributes
- Statistical Profile of Data Set
- Healthcare Display of Data Frame Data for Numeric Variables
- Heatmaps and Frequency Displays of Data Frame Data

**Data Set Provenance and Scope**

Company	Title	Record Create Date
Company	Company	Company
Drug Name	Drug Name	Drug Name
ATC Code	ATC Code	ATC Code
Drug Class	Drug Class	Drug Class
Formulation Code	Formulation Code	Formulation Code

**Quality by Design Template Statistical Quality Control Analysis**

Table of Contents

- Objective and Approach
- Drug Product Manufacturing Status
- Statistical Quality Control in Manufacturing Process
- Manufacturing Process Under Statistical Control?
- Control Chart of the Manufacturing Process
- Process Capability Analysis

**Objective and Approach**

The objective is to investigate the process for manufacturing tablets at a single dose. The key performance metric is process mean distribution. The objective of the manufacturing team is to investigate the process and improve quality capability.

**Quality by Design Template Identification of Critical Process Parameters**

Table of Contents

- Objectives and Approach
- What is the distribution between accepted and rejected lots?
- Quality Performance Analysis by Random Forests Classification
- How should the Random Forests Classifier on Manufacturing Data?
- Are there major outliers in the data set?
- How should the Random Forests Classifier on Manufacturing Data?
- Do Accepted/Rejected Lots cluster into separate groups?
- How is the Critical Process Parameter identified?
- How is setting values of CPP affect probability of lot rejection?
- Critical Process Parameter Significance

**Objectives and Approach**

The objective is to identify a critical Quality Attribute (CQA, e.g., Dissolution at 60 minutes) of the drug product (e.g., formulated tablet) by the collection of processing parameters (e.g., process settings and raw material attributes) that have the largest impact on the Critical Quality Attribute. This collection of processing parameters is termed Critical Process Parameters (CPPs).

**Quality by Design Template Identify Design Space**

Table of Contents

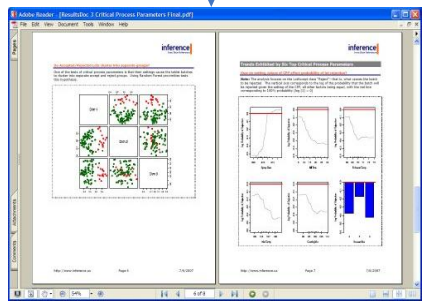
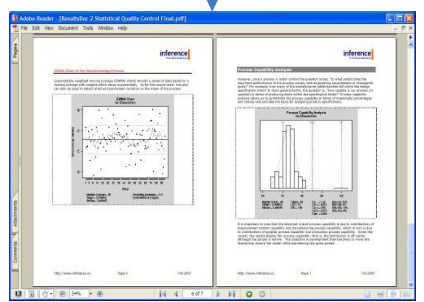
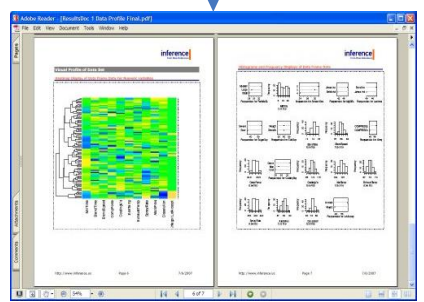
- Quality Attribute Process Model for Experimental Space
- Refine the Process Model for Experimental Space
- Define the Design Space
- Identify Critical to Process Parameters for Experimental Space
- Create a Multidimensional Contour Grid of Critical Process Parameters
- Evaluate Grid of Experimental Space
- Colour Map Visualization of Design Space in Experimental Space
- Accept/Reject Visualization of Design Space in Experimental Space

**Objective and Approach**

Definition of terms:

- Experimental Space:** the multidimensional region within the design space where the process is intended to achieve target quality goals
- Design Space:** the multidimensional region of experimental space where the process achieves target quality as determined by a science-based QbD approach
- Experimental Space:** the multidimensional region of experimental data where

Inference Results Documents



take-home message

## inference for R value

- Obtain increased returns on your investments in R training
- Extend R data-analysis capability to non-expert users
- Manage R data-analysis development, execution, collaboration and publication as a document-based business process
- Provide R data-analysis service as software
- Deploy “one-button” R data-analysis solutions
- Reuse R data-analysis solutions

For additional information contact:  
Paul van Eikeren, Ph.D.  
Paul.van.Eikeren@BlueReference.com

[www.InferenceforR.com](http://www.InferenceforR.com)