8 Ways We Fail With Predictive Analytics in Business

Stephen Smith

Research Director, Data Science

Eckerson Group

Sponsored by:

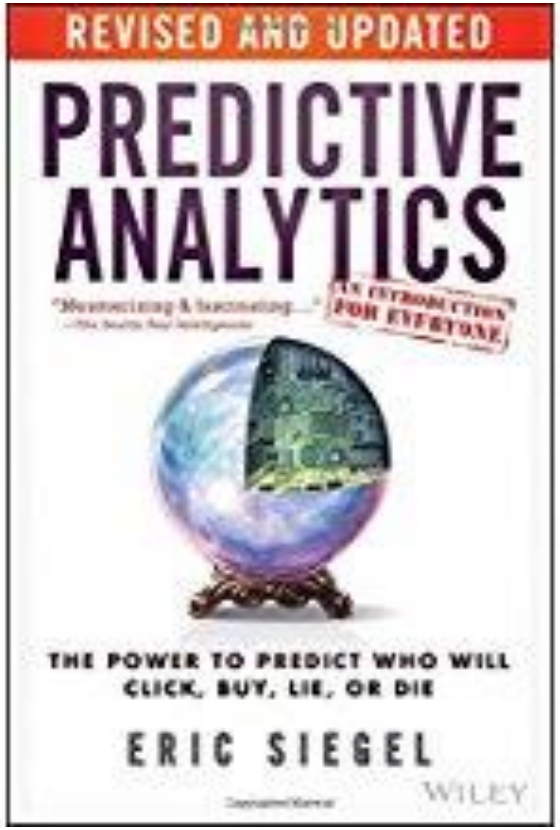
Sponsored by:

The Promise of Predictive Analytics

- **Best web ad selection**
  - *3.6% more revenue*
  - *$1 million every 19 months*

- **Decreased loss ratio in insurance**
  - *0.5% reduction is loss ratio*
  - *$50 million annually*

Reference: "Predictive Analytics" p.25



Great book!!!

Sponsored by: Mariner Innovations

Research Hypothesis

1. Every business **should be using** predictive analytics

2. Predictive analytics is **not being used** often enough

3. Because it is a bit **complex and dangerous**

Sponsored by:

We Interviewed Industry Experts

Research Findings

1. Every business **should be using** predictive analytics
2. Predictive analytics is **not being used** nearly enough
3. Because it is a bit **complex and dangerous**
4. It needs to be **automated and operationalized**

Sponsored by:

The Promise

Data

Predictive
Analytics

Predict
the future!

Sponsored by:

Reality

Where's My Data?

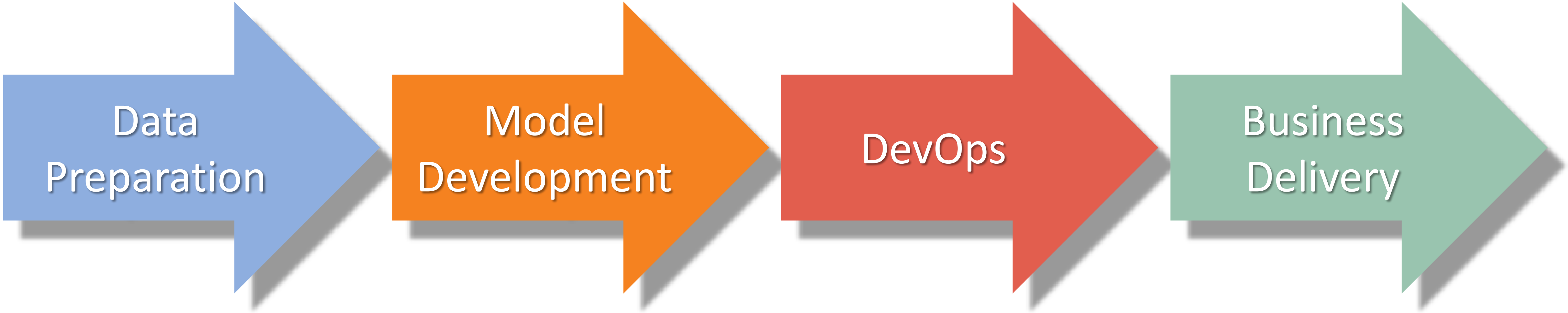
Predictive Analytics

What do I do with this model?

Business

Sponsored by: Mariner Innovations

Full Data Science Lifecycle



Sponsored by:

© Eckerson Group 2018

Cultural Evolution

Eckerson Group

Promise	Fear	Trust	Acceleration
• Expectations of reduced churn, increased cross sell etc. • Hires ‘data scientists’ • First models created	• Model doesn’t work in production • Models are late • Data scientist quits	• Focused projects begin to succeed • Business champions emerge requesting models • Best practices evolve and are enforced	• Hundreds to thousands of active models • Doubling data scientists quadruples number of models • Business users expect to use models

Technical Evolution

Experimentation
- Data discovery
- Non-actionable models
- Data scientist consultants

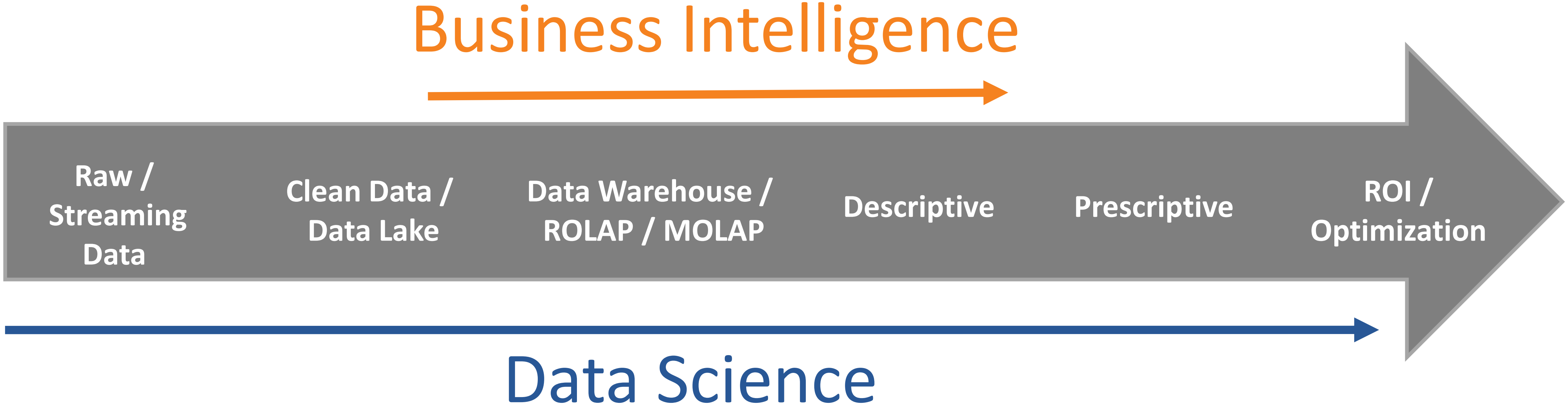
Artisanship
- Project-based
- Actionable models
- Irreplaceable data scientists

Automation
- Data lake
- Self-service
- Citizen data scientists

Operationalization
- Process-based
- Repeatable, auditable
- Agile teams

Sponsored by:
Mariner Innovations

Warning: This Time It's Different



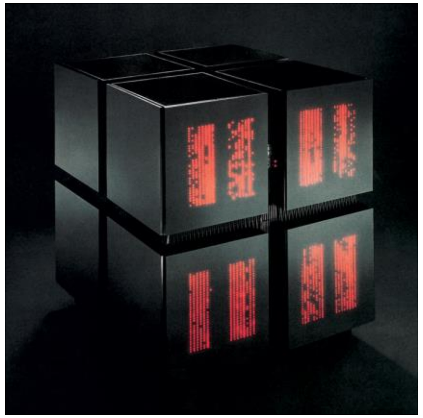
Sponsored by:

© Eckerson Group 2018

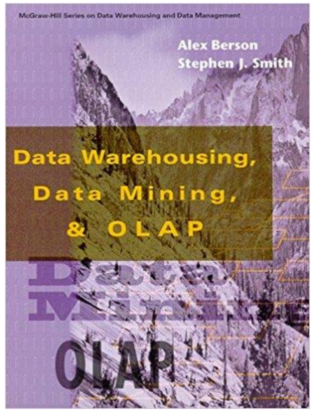
My Journey to Data Science

1990s

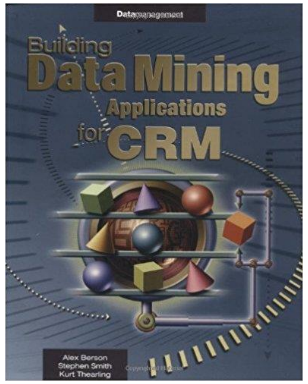
MPP



DW + OLAP



Data Mining



2000

Pharmaceuticals

OPTAS

2010

Education

JogNog

JogNog helps students remember!™

"Data Science"

Eckerson Group

Technology Evolved

• Massively Parallel Processing	=> Hadoop
• OLAP	=> Business Intelligence
• Data Mining	=> Predictive Analytics
• Neural Networks	=> Deep Learning
• Data Warehousing	=> Data Lake (+MDM,DG,ETL…)
• -	=> Data Science

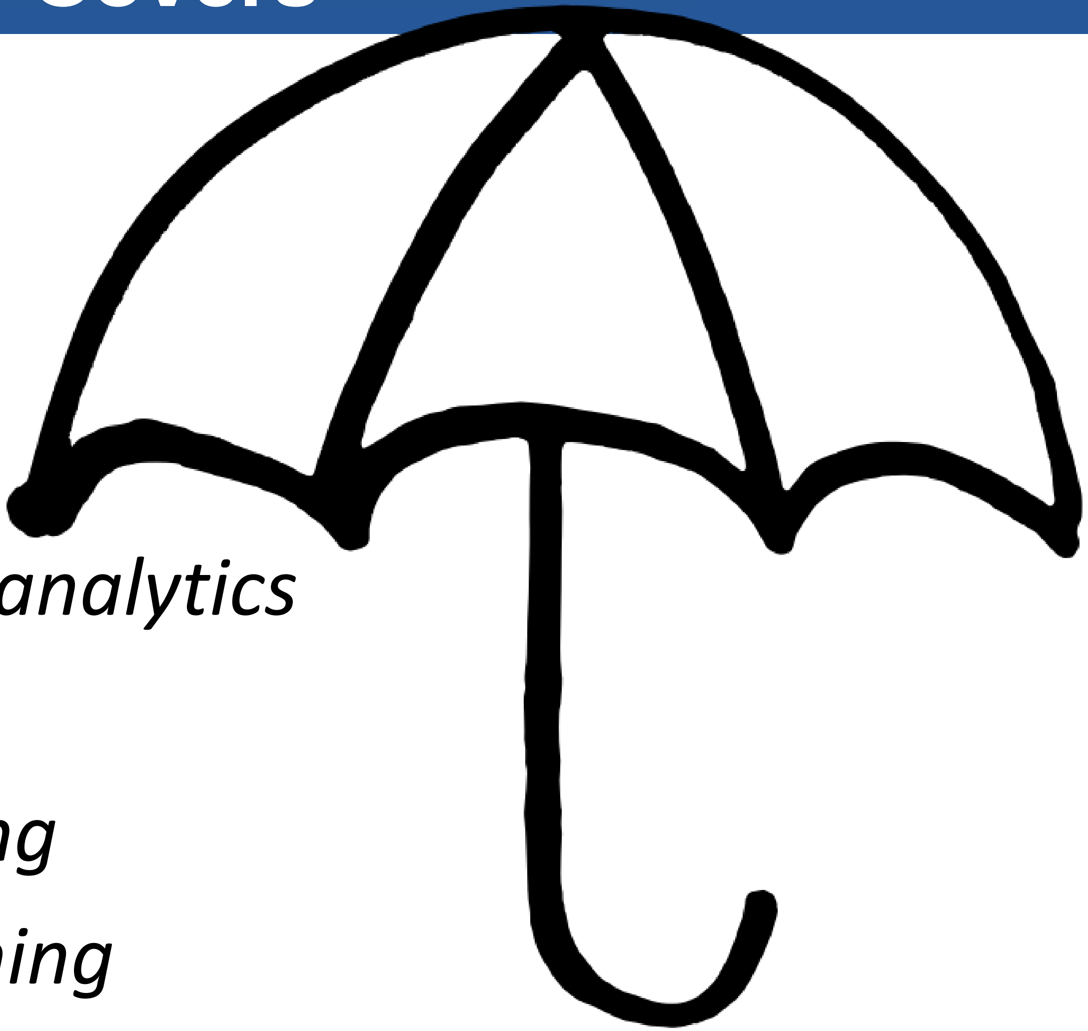
Sponsored by: Mariner Innovations

What we Need Now is…

- ‘Computer Science’ was Born in 1959
- A need to describe the practice of using computers
- Interestingly:
  - *The alternative name proposed for “Computer Science” was “Data Science”*

Sponsored by:

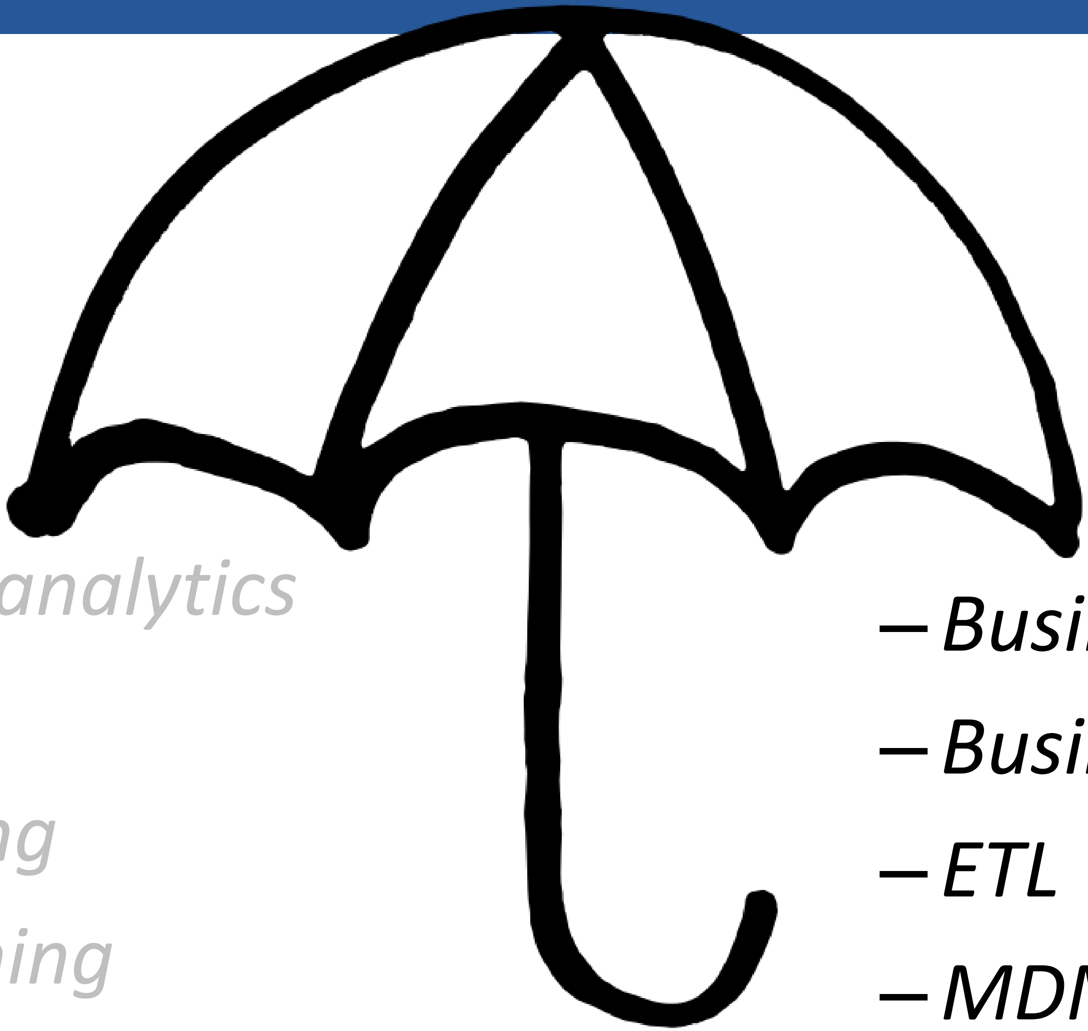
Data Science Covers



- *Predictive analytics*
- *Statistics*
- *Data mining*
- *Deep learning*
- *Artificial Intelligence*
- *Machine Learning*

Sponsored by: Mariner Innovations

Maybe also



- Predictive analytics
- Statistics
- Data mining
- Deep learning
- Artificial Intelligence
- Machine Learning

- *Business intelligence*
- *Business analytics*
- *ETL*
- *MDM*
- *Data engineering*

Sponsored by:

The Elephant in the Room

Why isn't data science used more?

Free for commercial use – no attribution: https://pixabay.com/en/elephant-eye-tears-sad-animal-266791/

Question #1

"I have data?"

Sponsored by:

Free for commercial use – no attribution: https://pixabay.com/en/child-surprise-think-interactivity-2800835/

Question #2

"Where's my data?"

Sponsored by:

Free for commercial use – no attribution: https://pixabay.com/en/boy-teen-schoolboy-got-to-thinking-2736655/

Question #237

"What will happen next?"

Sponsored by:

Question #238

“What should I **do** next?”

Sponsored by:

"How do I start using predictive analytics?"

Sponsored by:

WARNING

Sponsored by:

“Today we can process Exabytes of data at lightning speed, and this gives us the **potential to make bad decisions** far more quickly, efficiently and with greater impact than we did in the past.”

- **Susan Etlinger**

Ted Talk: What do we do with all this big data?

TED Ideas worth spreading

Sponsored by:

Mariner Innovations

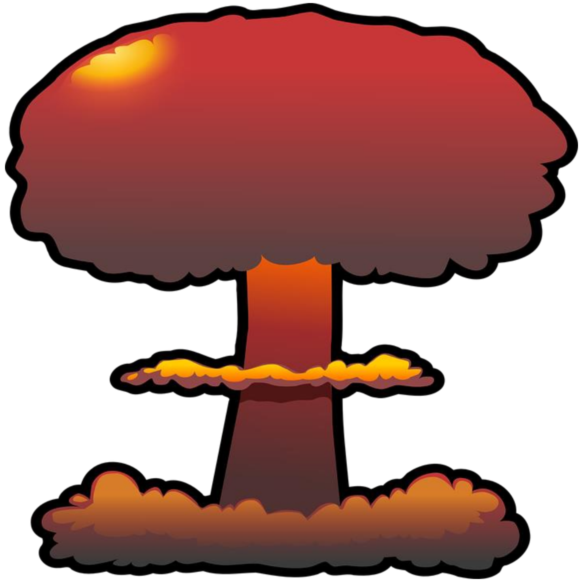
Uranium Ore

Uranium Ore = Perfectly Safe

Sponsored by:

You Make Plutonium from Uranium

Data = Uranium Ore



Data Science = Plutonium

Sponsored by:

Careful Process Control Required

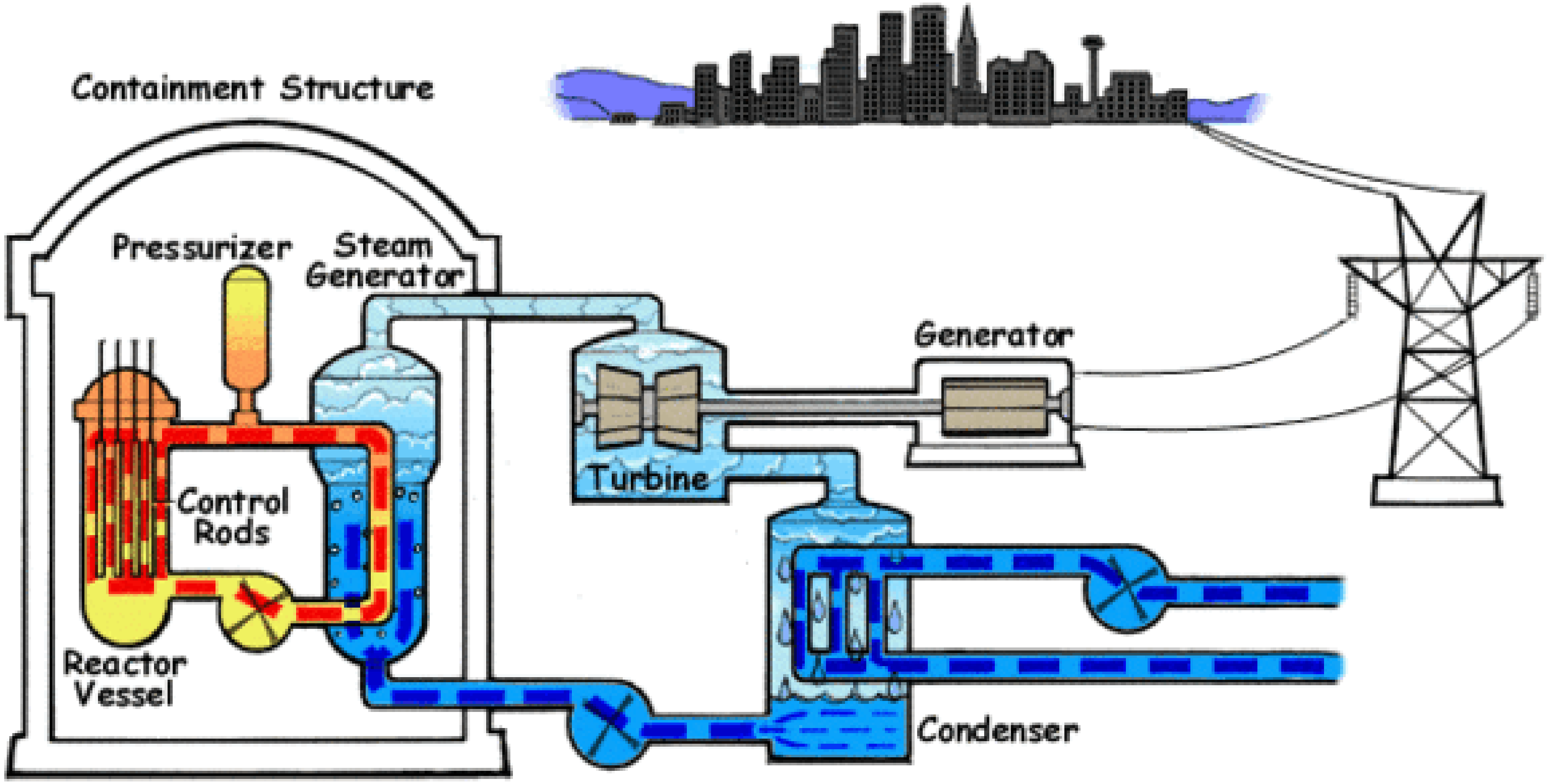


Image: Public domain: http://www.nrc.gov/reading-rm/basic-ref/students/animated-pwr.html

Sponsored by: Mariner Innovations

Predictive Analytics is Like Plutonium

	Plutonium	Predictive Analytics
Powerful	Generates electricity	Drives new revenue with little investment
Dangerous	Explosions	Mistakes can cost hundreds of millions of $$
Handle with care	Requires operationalized processes and tools	Requires operationalized processes and tools

Sponsored by:
MARINER INNOVATIONS

WHERE DATA SCIENCE FAILS

Sponsored by:

Free for commercial use – no attribution: https://pixabay.com/en/water-pipe-leaking-water-drip-880975/

How costly are data science fails?

Sponsored by:

A Bad Predictive Model Could:

1. Lose all the investment in a hedge fund.
2. Send expensive offers to the wrong customers.
3. Ruin a brand.

Sponsored by:

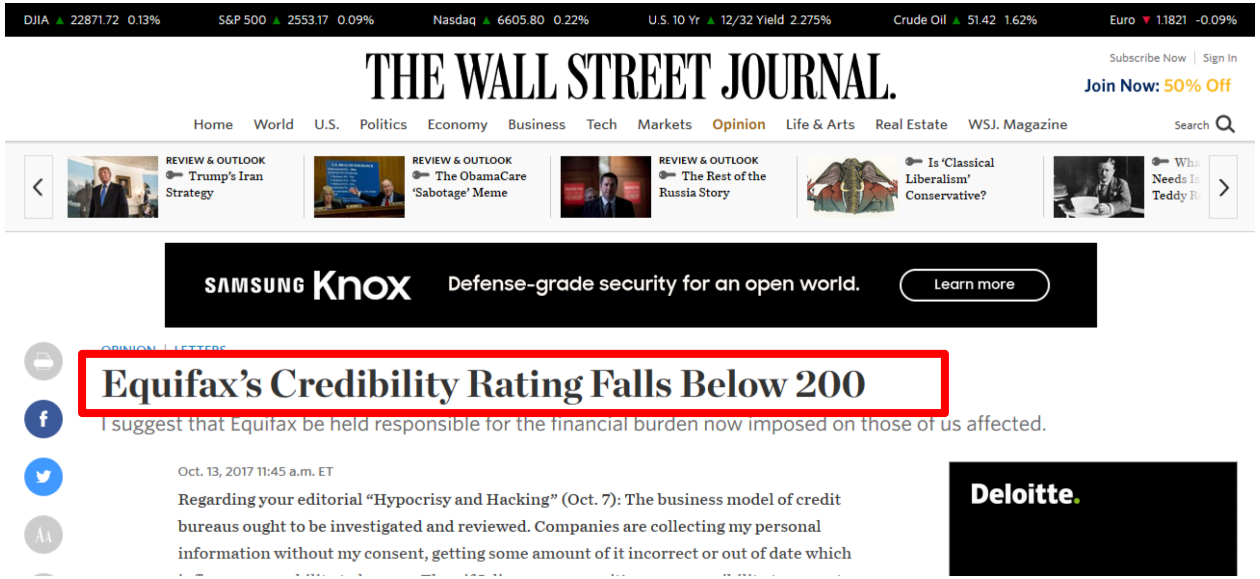
Impact on Brand from Breach of PII

$184-$334 million

TARGET

"Reputation Impact of a Data Breach"

experian™



Sponsored by: MARINER INNOVATIONS

Data Preparation → Model Development → DevOps → Business Delivery

#1 Data Drift, Anomalies, Errors

Sponsored by:
Mariner Innovations

Example: Hedge Fund

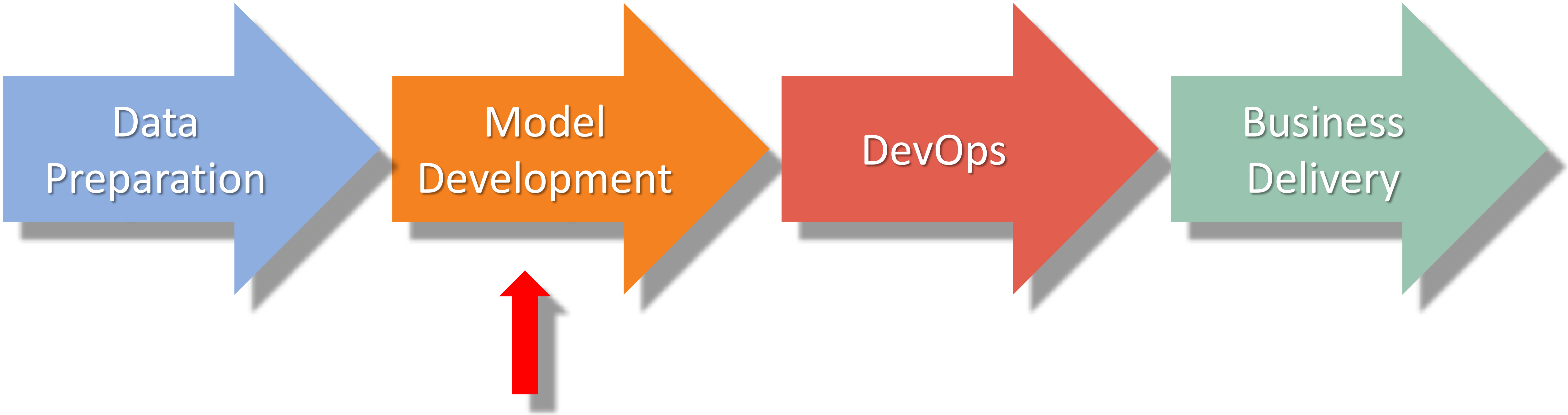
- **Solution**
  - *Proprietary algorithms: radial basis function, Black-Scholes, ANN*
- **Failure**
  - *Model prediction didn't match actual*
- **Why?**
  - *Errors in industry-wide data sets used by all hedge funds*

Sponsored by:
Mariner Innovations

- **Challenge**
  - *Input data values can change over time*
  - *Errors can creep into data*

- **Best Practices**
  - *Tools that have automated alerts for data changes*
  - *Algorithms that manage missing or corrupt values*

Sponsored by:

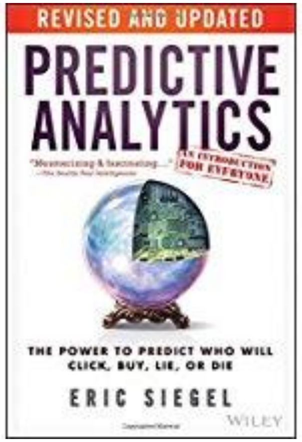


#2 Bad Model Validation

Sponsored by:
Mariner Innovations

Example: Predict the stock market

- **Solution**
  - *Proprietary time-series algorithms*
- **Failure**
  - *None – error detected before deployment*
- **Why?**
  - *Temporal leakage*
  - *Time window mistakenly shifted by one day*

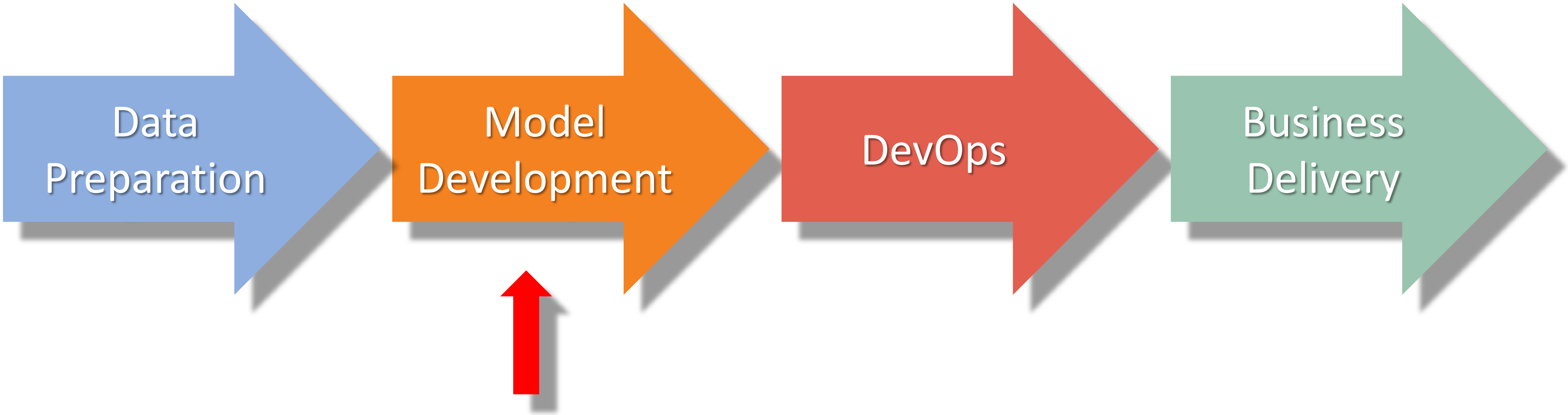


Sponsored by:

- **Challenge**
  - *Concepts like target leakage are too complex for citizen data scientists*

- **Best Practices**
  - *Trial model in non-actionable live environment*
  - *Have 3rd party reverse engineer model*
  - *Be careful!*

Sponsored by:



#3 Regulatory Compliance

Sponsored by: Mariner Innovations

Example: Predict pregnancy

- **Solution**
  - *Target looks at market basket analysis*
- **Failure**
  - *Father notices teenage daughter is receiving coupons for baby items from Target*
  - *Uncomfortable conversation with daughter ensues …*
- **Why?**
  - *Dealing with dangerous data*

Sponsored by: Mariner Innovations

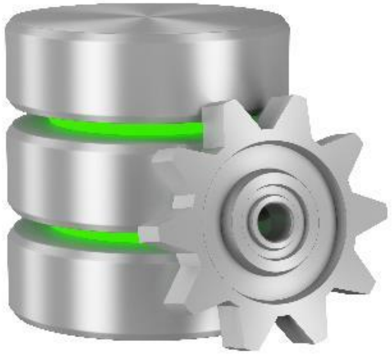
- **Challenge**
  - *Privacy violations are complex*

- **Best Practices**
  - *Agile teams with privacy advocate*
  - *Evaluation of risks vs. benefits*
  - *Build processes for CDO/CIO/CPO approval for things that could damage the corporate brand*
  - *Ability to rollback model creation environment*

Sponsored by: Mariner Innovations

Compliance Architecture

Data Lake / Data Warehouse	Metadata	Predictive Analytics Engine	Business User

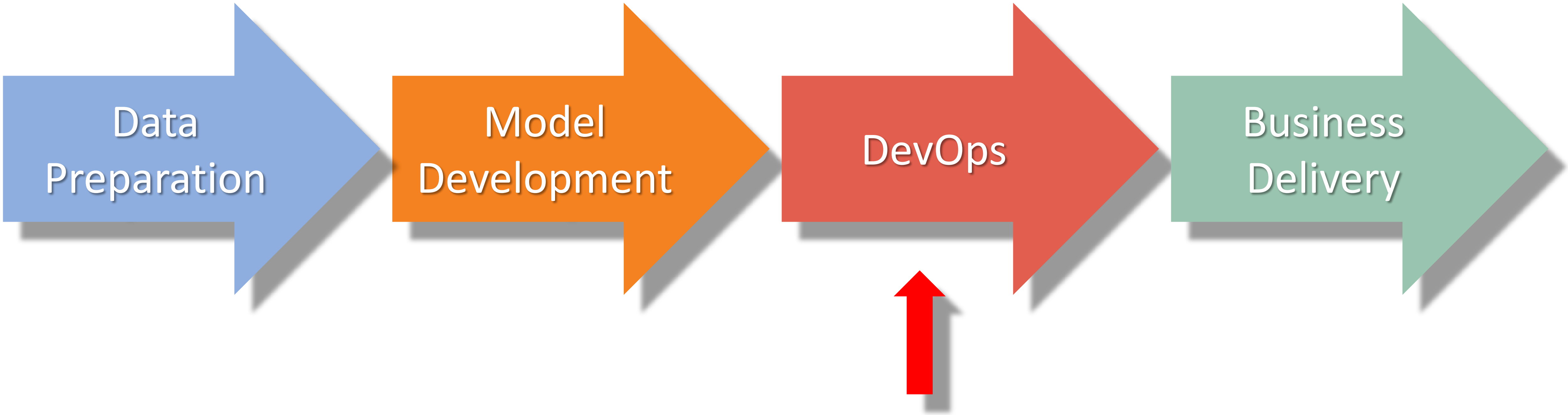


- Rollback of data
- Random sampling
- Flagging dangerous data
- Data lineage
- Data governance

- Test and control creation
- Detect highly correlated variables
- Alerts on variable variation
- Model lineage
- Model governance
- Rollback to models
- Model audit trails
- Temporal leakage detection
- Alert on model aging

- Business modeling / ROI
- Workflow
- Collaboration
- Model marketplace
- Model ranking

Sponsored by:

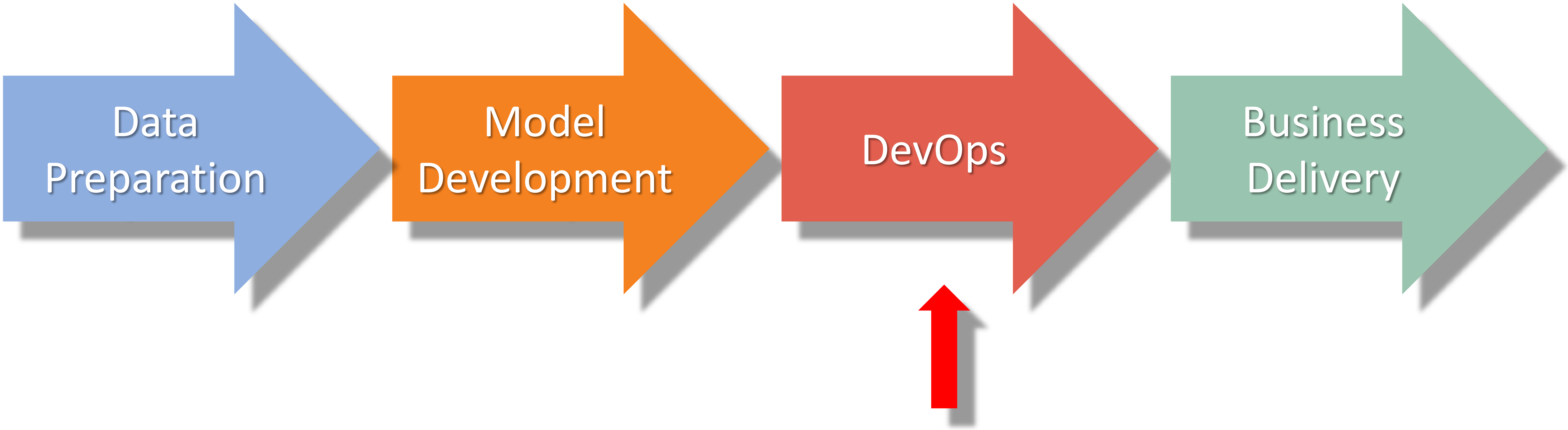


#4 Model Degradation

Sponsored by:

- **Challenge**
  - *Models decay in performance*
  - *Data scientists are overwhelmed*
- **Best Practices**
  - *Automate retraining or refreshing a model*
  - *Utilize tools that have alerting systems*
  - *Empower operations to offload data scientists*

Sponsored by:
Mariner Innovations

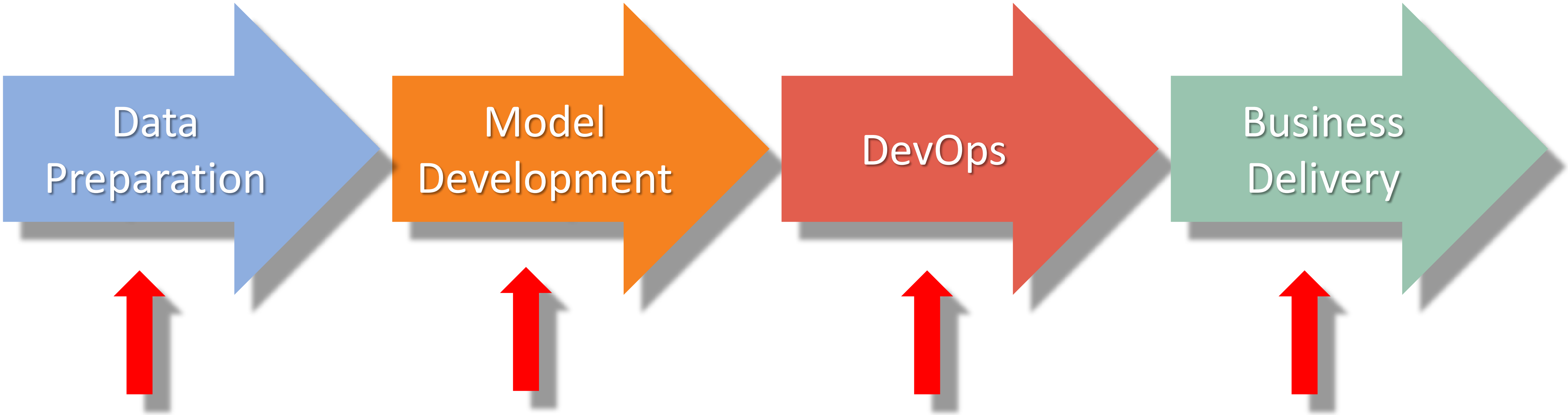


#5 Deployment Disconnect

Sponsored by:
Mariner Innovations

- **Challenge**
  - *Model is created in one database and deployed elsewhere*
  - *Must be recoded and tested*
  - *Can delay deployment by 3-6 months*
- **Best Practices**
  - *Use tools that automatically generate code*
  - *Build and deploy in the same database*

Sponsored by:

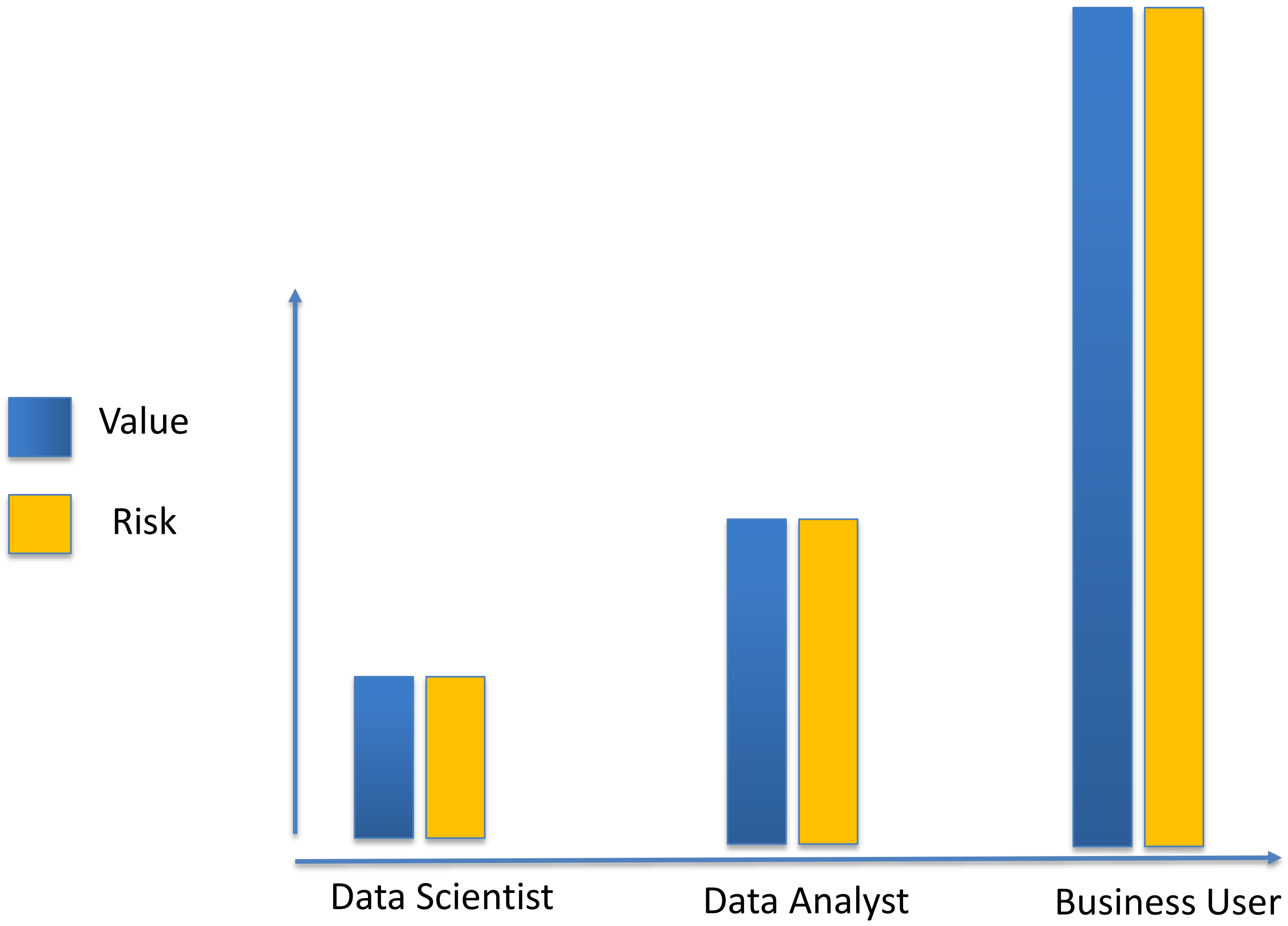


#6 Finding Data Scientists

Sponsored by: Mariner Innovations

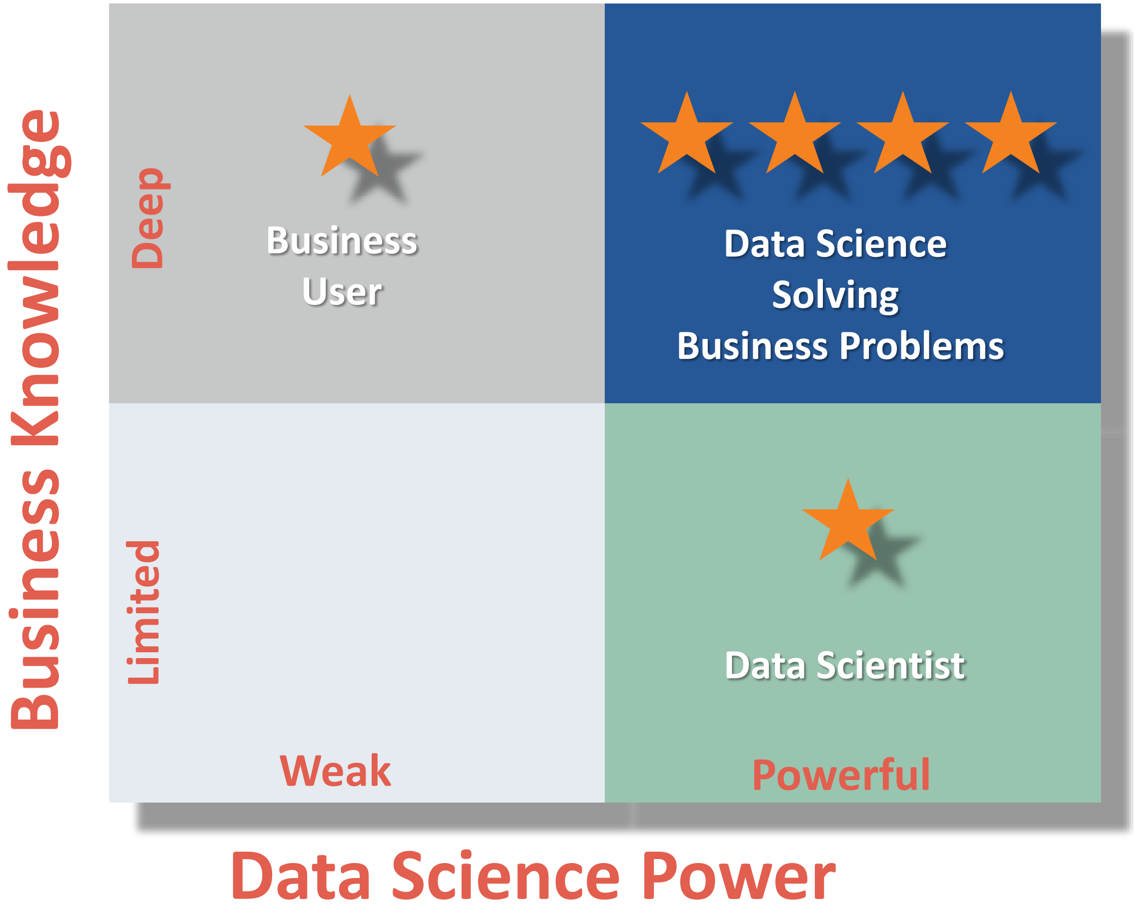
© Eckerson Group 2018

Value and Risk Increase Together



Sponsored by: Mariner Innovations

Acceleration = Business + DS



Sponsored by:

© Eckerson Group 2018

Data Scientist - Defined

Eckerson Group

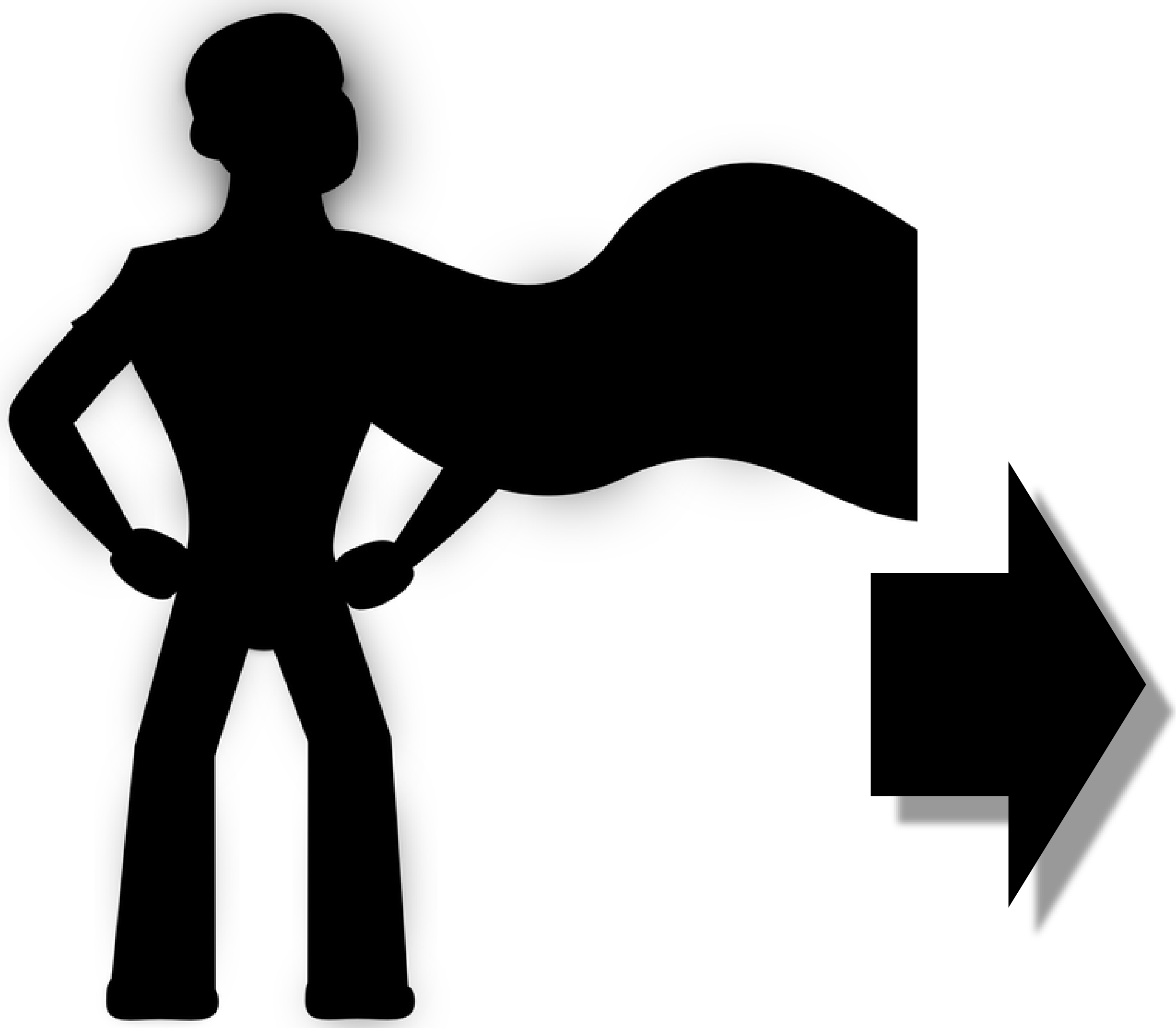
Mythical Data Scientist

- Data expert
- Data engineer
- Deep learning expert
- Statistician
- Coder
- Cloud expert
- Operations expert
- Parallel processing developer
- Deep thinker
- Good communicator
- Understands business problem
- Business rules expert

Required Skills

Sponsored by: Mariner Innovations

Data Scientist - Deconstructed



Mythical Superman Data Scientist

Multidisciplinary Agile Team

Sponsored by: Mariner Innovations

Data Scientist Has Many Agile User Roles

1. **Data Steward**
   - Knows what data is available, lineage, and governance
2. **Data Engineer**
   - Can access and transform data and optimize database performance
3. **Privacy Advocate**
4. **Data Scientist**
   - Understands predictive analytics, statistics, machine learning
   - Provides high accuracy models and scores
5. **Operational Engineer**
6. **Data Analyst**
   - Can translate between business requests and data realities
   - Provides ad hoc query and report support
7. **Business User**
   - Understands the business problem and responsible for ROI
   - Provides business case and logical validation of causality

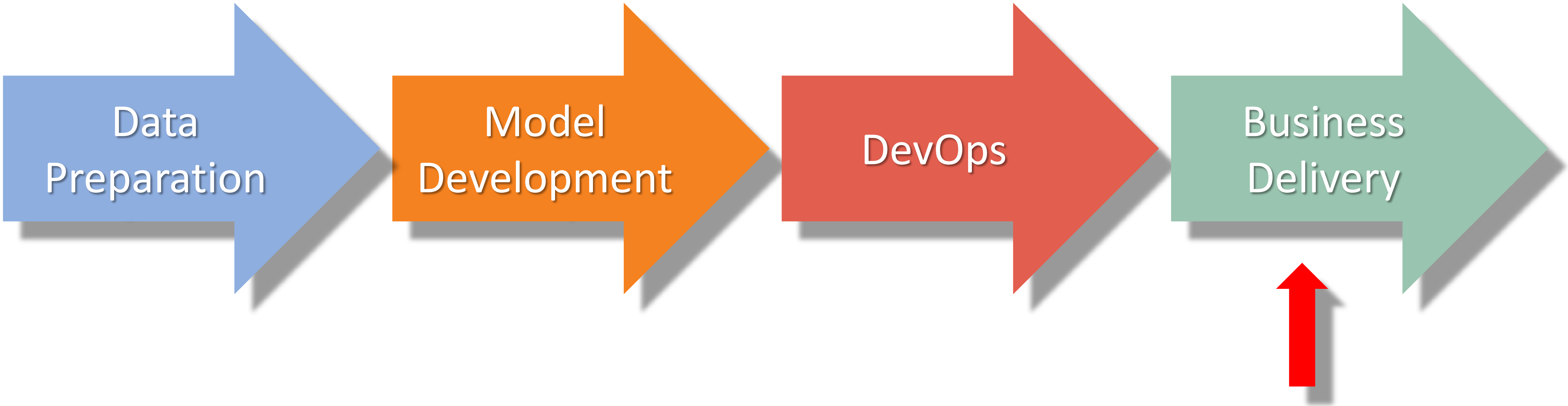
Sponsored by:

Data Scientist

- **Challenge**
  - *Heroic project-based model creation*
  - *High error rates*
  - *Can't scale*

- **Best Practices**
  - *Agile teams*
  - *Embed DS representative into business*

Sponsored by:



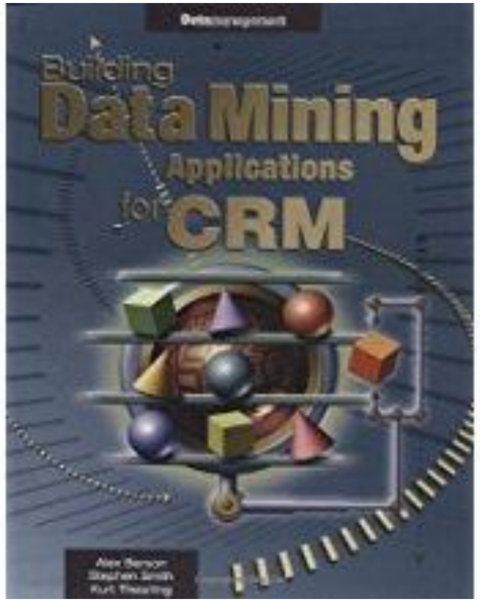
#7 Business Disconnect

Sponsored by:

Free for commercial use – no attribution: https://pixabay.com/en/water-pipe-leaking-water-drip-880975/

Example: Predict mobile phone churn

- **Solution**
  - *CART model in specialized PA tool*
- **Failure**
  - *Created higher churn than control group*
- **Why?**
  - *Predictive model excellent*
  - *BUT - Offer reminded customers of their anniversary*

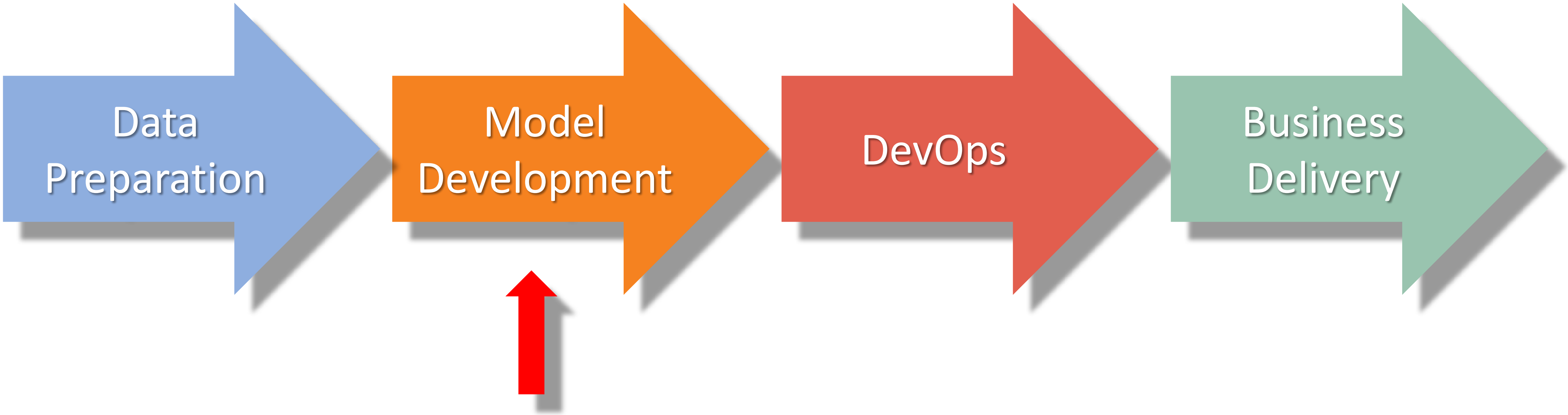


Great book!!!
(completely biased)

Sponsored by:

- **Challenge**
  - *Possibility of large errors*
  - *Gun shy business users*
  - *Frustrated data scientists*
- **Best Practices**
  - *Predict offer effect with small trials*
  - *Include business user in top down design*

Sponsored by:



#8 Lack of a Data Science Platform

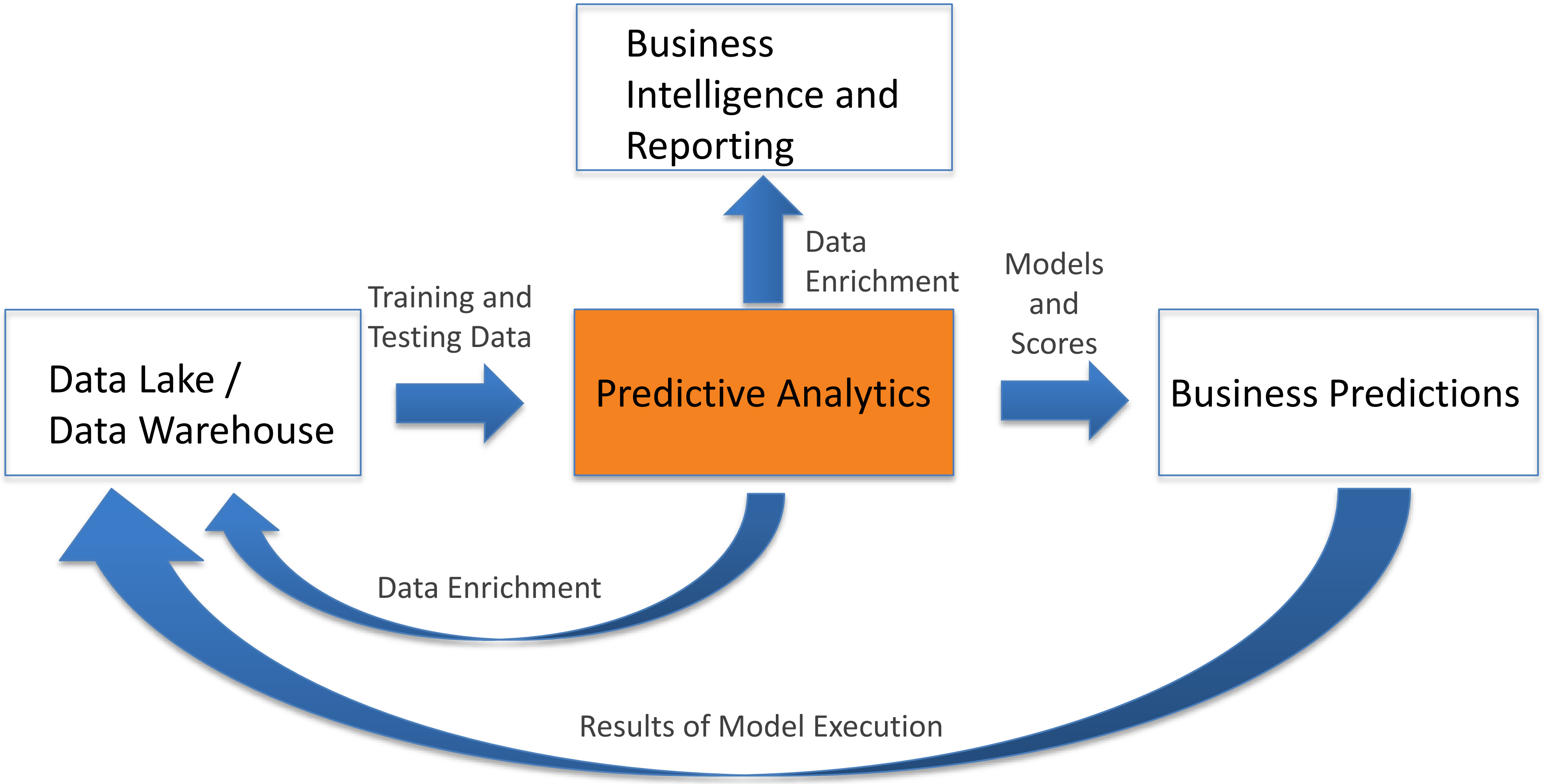
Sponsored by:

Data Science Platform

- **Challenge**
  - *High error rates*
  - *Long delivery cycles*
  - *Can’t scale*
- **Best Practices**
  - *Centralize and focus on building a ‘platform’*
  - *Deliver like an internal software company*
  - *Plan for scale but execute for success today*

Sponsored by:

Data Science Platform



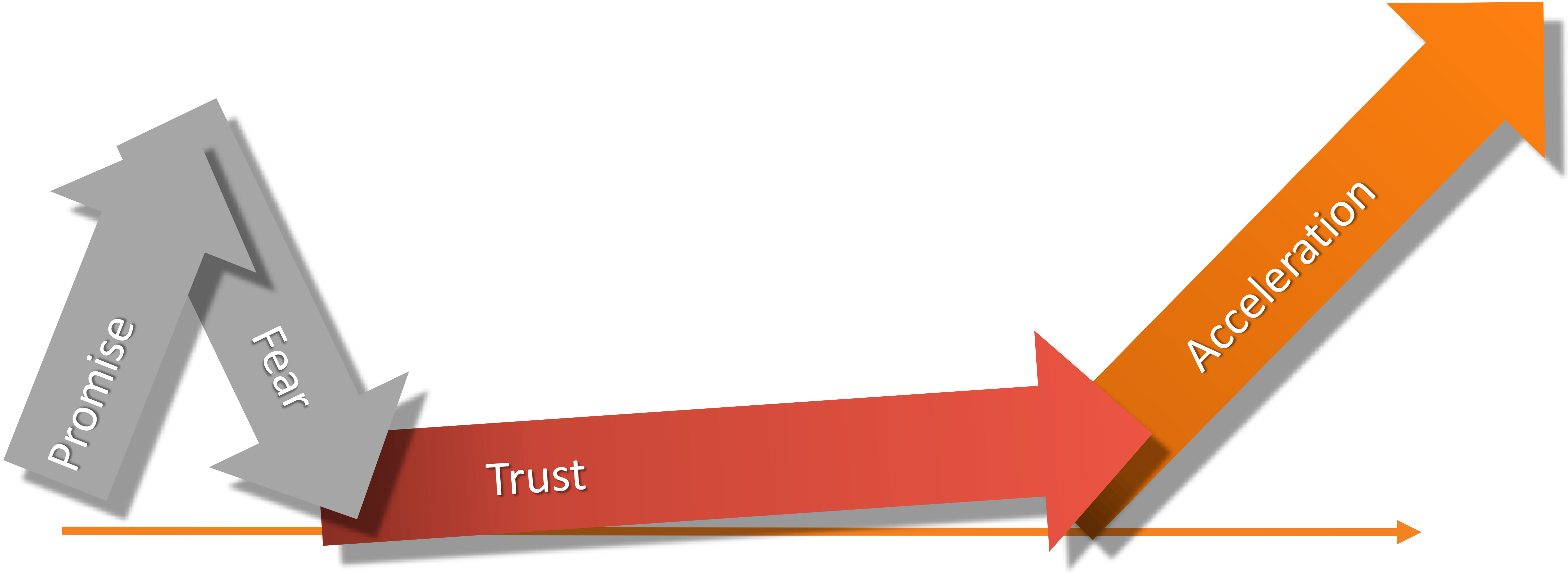
Sponsored by:

© Eckerson Group 2018

Top 8 Challenges to Data Science

1. Bad Data
2. Bad Validation
3. Regulatory Compliance
4. Model Degradation
5. Deployment Disconnect
6. Finding Data Scientists
7. Business Disconnect
8. No Platform

Sponsored by:



Path to Digital Transformation

Sponsored by:
Mariner Innovations

Four Years from Now

Eckerson Group

1. **Self-service data science** will be the dominant lifeform

2. **Citizen data scientists** will be effective

3. Data scientists will be **more effective**

4. Data science best practices will look a lot like **computer science best practices**

5. Operationalized data science will provide a **strategic competitive advantage**

Sponsored by:
MARINER INNOVATIONS

Questions?

Stephen Smith
Research Director, Data Science
Eckerson Group
Email: ssmith@Eckerson.com
Twitter: @steve4years

Karen Fegarty
Regional VP, Mariner Innovations
T/ 902-499-4983
Karen.fegarty@marinerinnovations.com
www.marinerinnovations.com

Resources

Papers on www.Eckerson.com

- **Best Practices in Data Science : Ten Keys**
- **The Demise of the Data Warehouse**
- **Eckerson Eight Innovations in Data Science**
- **Data Science is Plutonium**

Books

- **"Building Data Mining Applications for CRM"** – Stephen Smith, Alex Berson, Kurt Thearling
- **"Data Warehousing, Data Mining, & OLAP"** – Stephen Smith, Alex Berson
- **"Predictive Analytics"** - Eric Siegel
- **"Data Science for Business"** – Foster Provost, Tom Fawcett

Sponsored by: