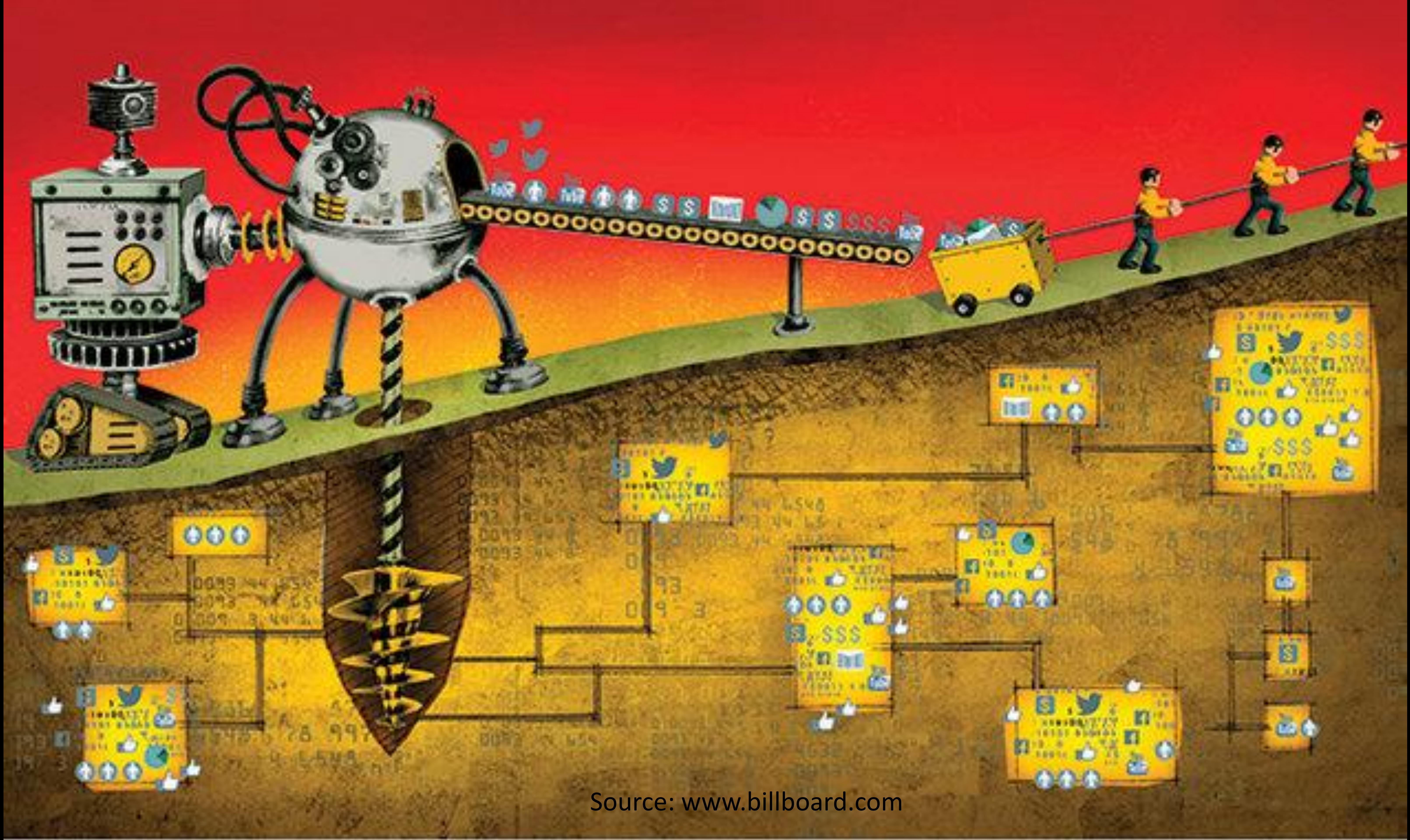
UNIVERSITY OF TWENTE.

Data analytics
Trends, Methods and Tools

PROF.DR. JOS VAN HILLEGERSBERG

INDUSTRIAL ENGINEERING AND
BUSINESS INFORMATION SYSTEMS

CHOIR



Source: www.billboard.com

Why?

Top objective of Big Data and AI investments	#1 Goal	Started	Success	Rate
Advanced analytics/better decisions	36.2%	84.1%	58.0%	69.0%
Improve customer service	23.2%	65.2%	34.8%	53.4%
Decrease expenses	13.0%	66.7%	40.6%	60.9%
Innovation/disruption	11.6%	46.4%	20.3%	43.8%
Speed to market	8.8%	53.6%	29.0%	54.1%
Monetization	7.2%	31.9%	8.7%	27.3%

Challenges?

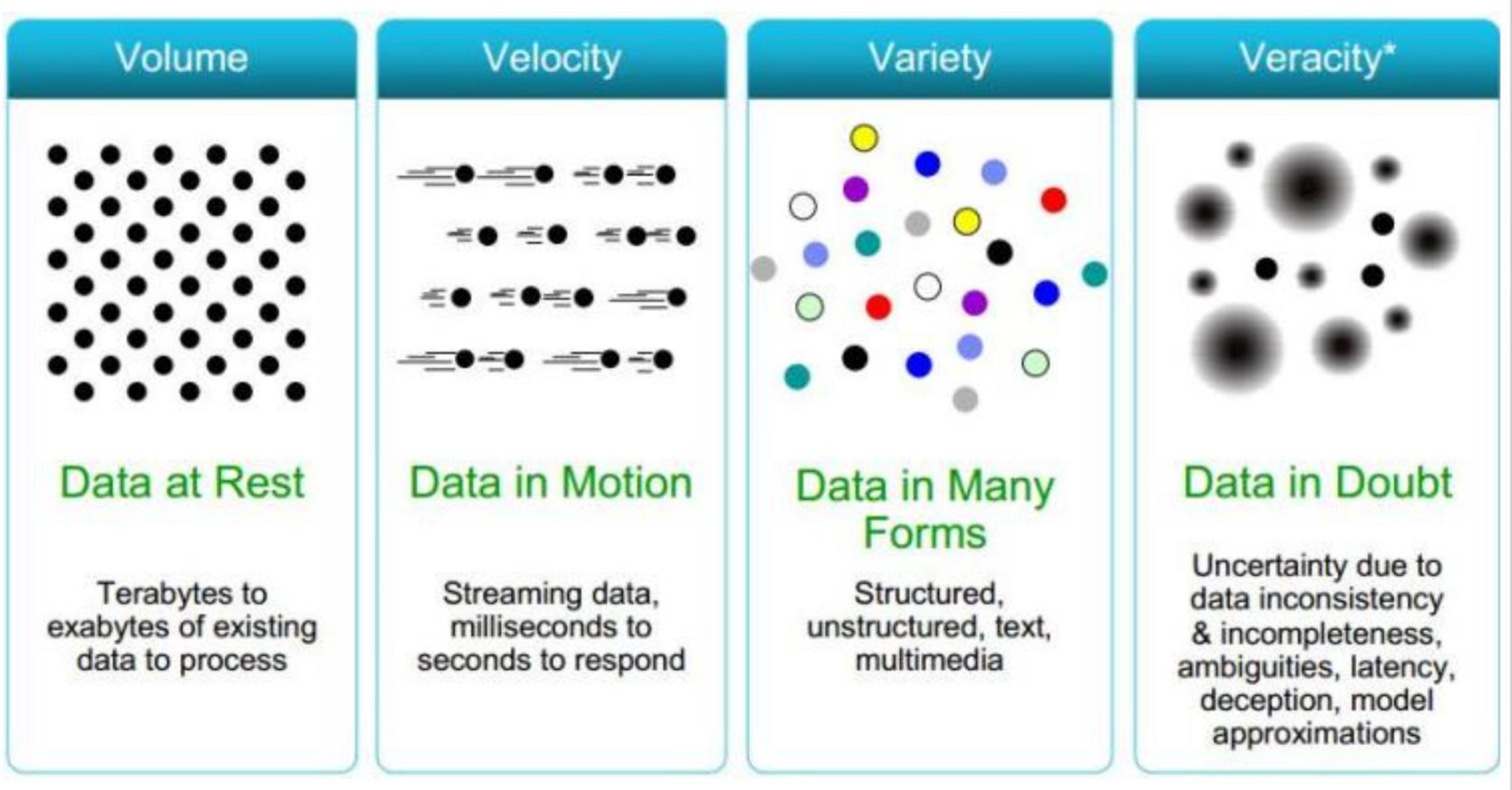
Business adoption of data initiatives remain a challenge	
- Yes	64.7%
No	35.3%

Challenges?

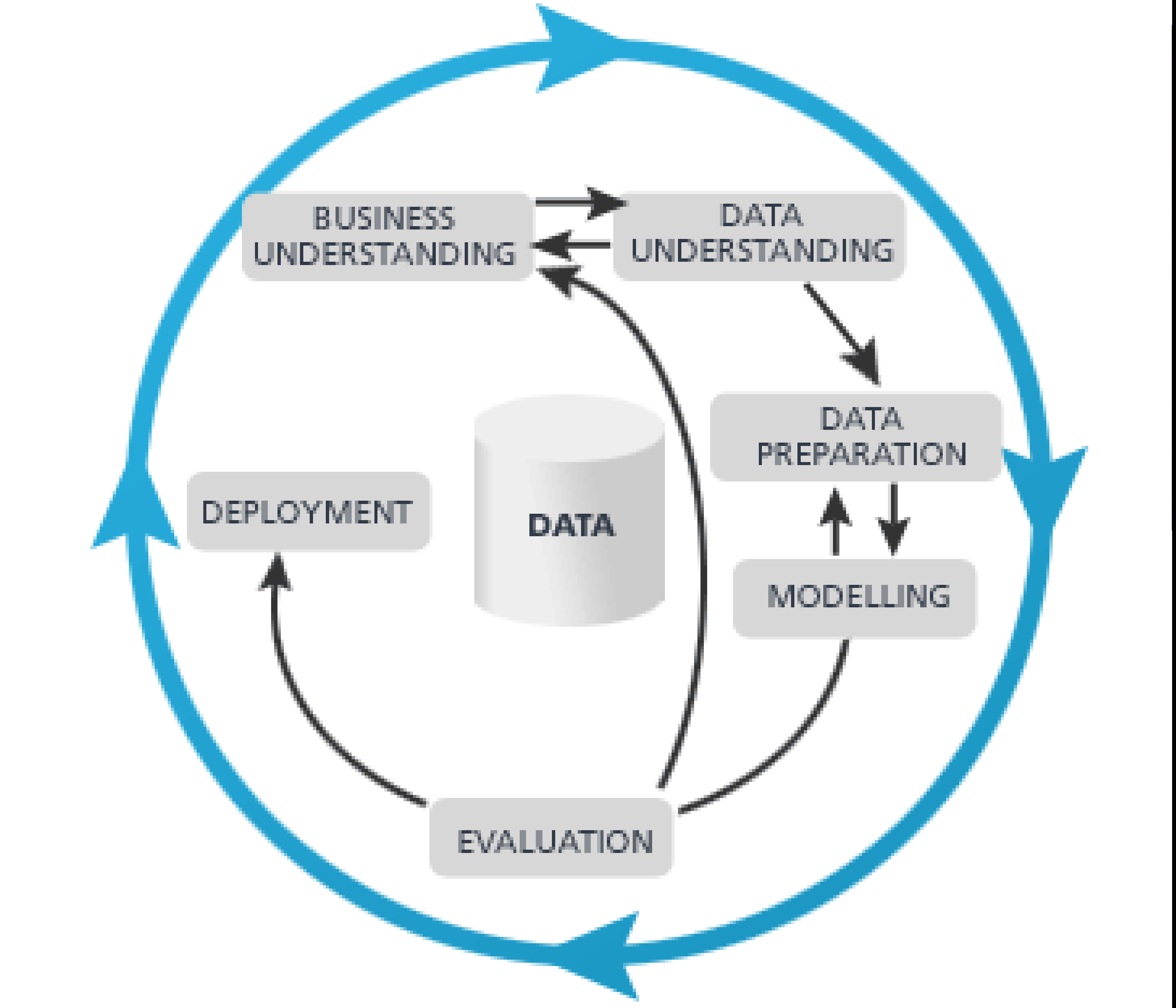
Biggest challenge to successful business adoption	
Cultural resistance to change	32.5%
Understanding of data as an asset	30.0%
Insufficient organizational alignment & business agility	25.0%
Lack of business diretion and executive leadership	7.5%
Lack of technology leadership and solutions	5.0%

Challenge?

Biggest challenge to becoming data-driven	
People	48.5%
Process	32.4%
Technology	19.1%

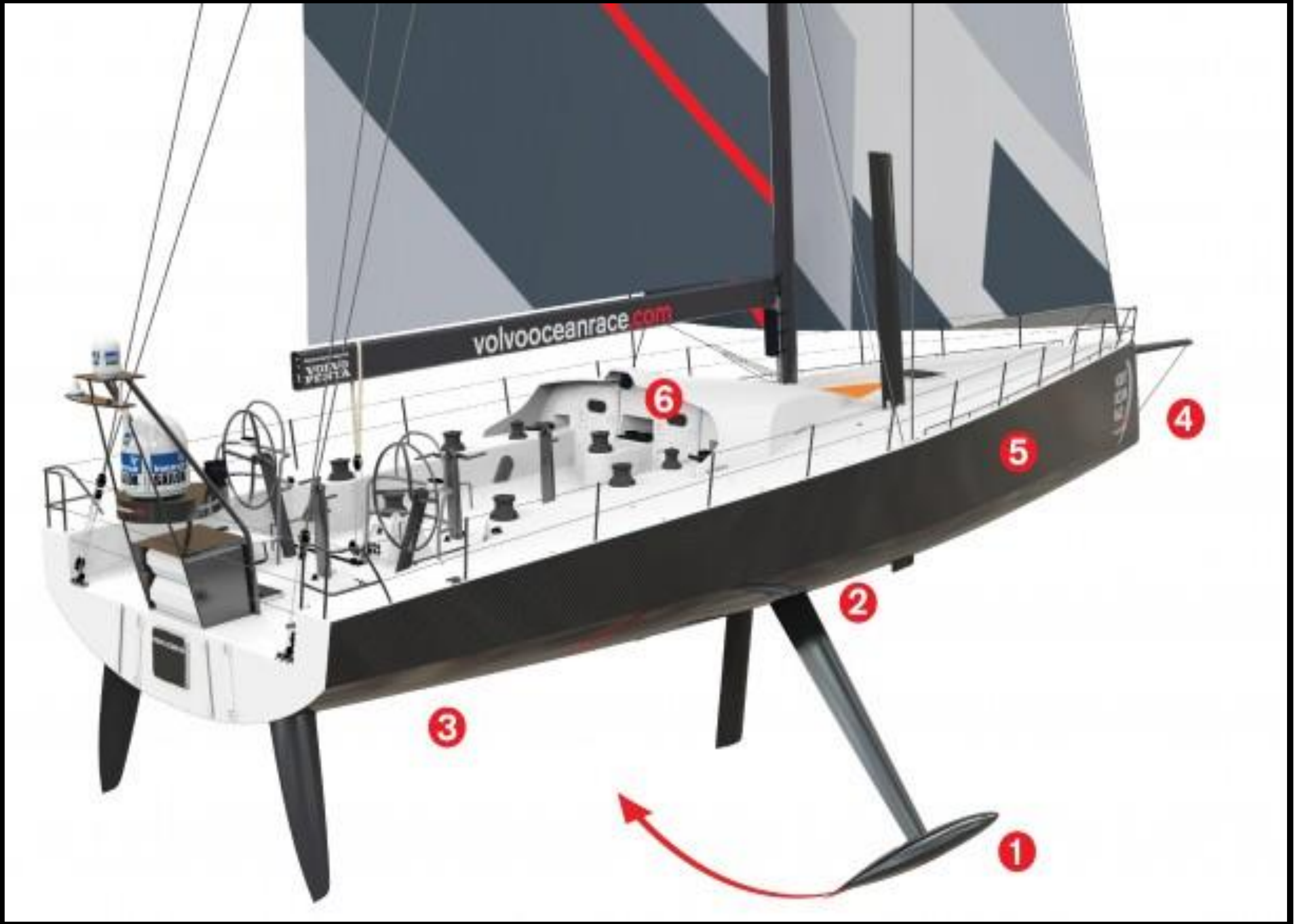

Volume
Velocity
Variety
Veracity*
Data at Rest
Data in Motion
Data in Many Forms
Data in Doubt
Terabytes to exabytes of existing data to process
Streaming data, milliseconds to seconds to respond
Structured, unstructured, text, multimedia
Uncertainty due to data inconsistency & incompleteness, ambiguities, latency, deception, model approximations

Source: IBM

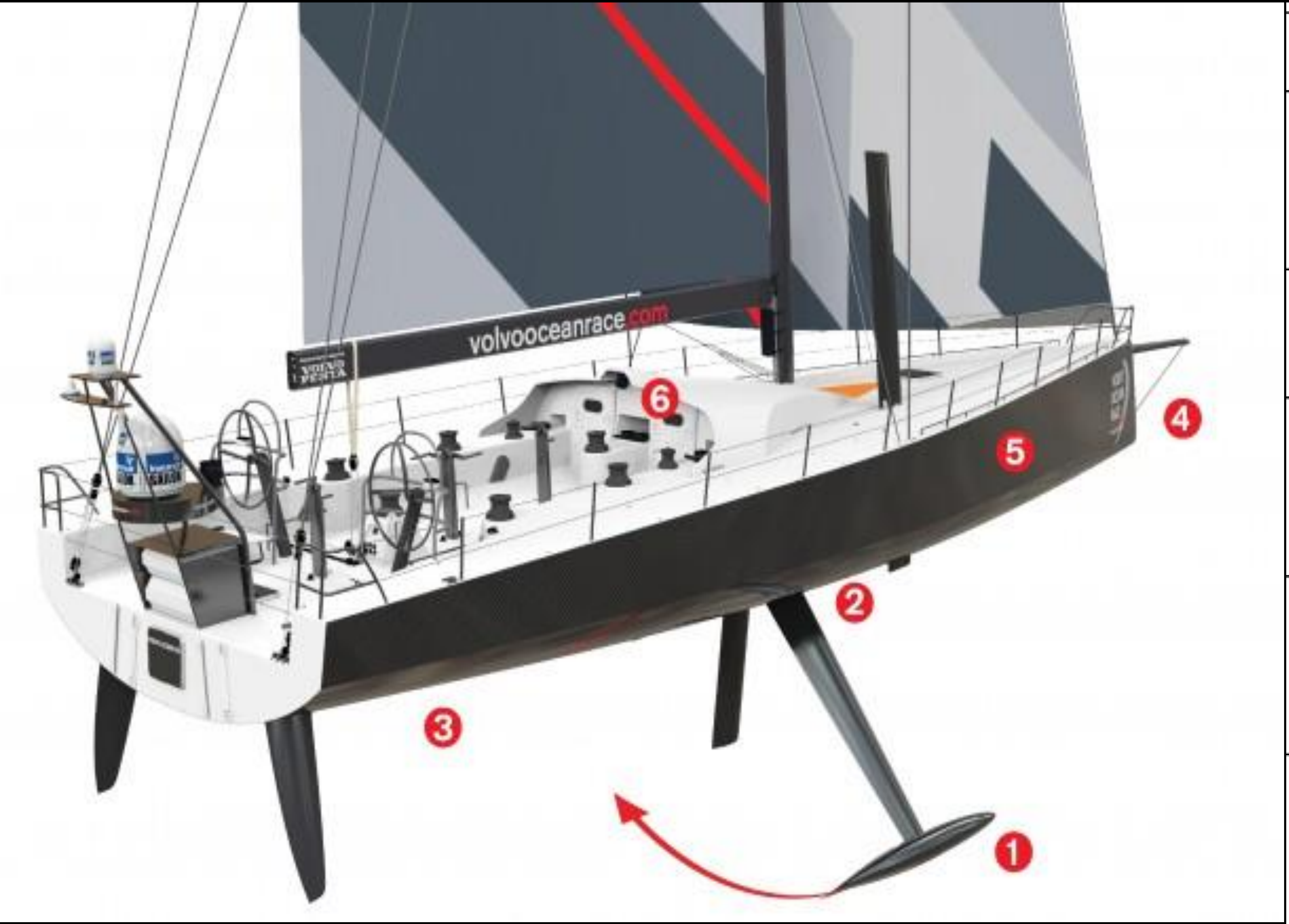

BUSINESS UNDERSTANDING
DATA UNDERSTANDING
DATA PREPARATION
MODELLING
EVALUATION
DEPLOYMENT
DATA


talk
to the
experts

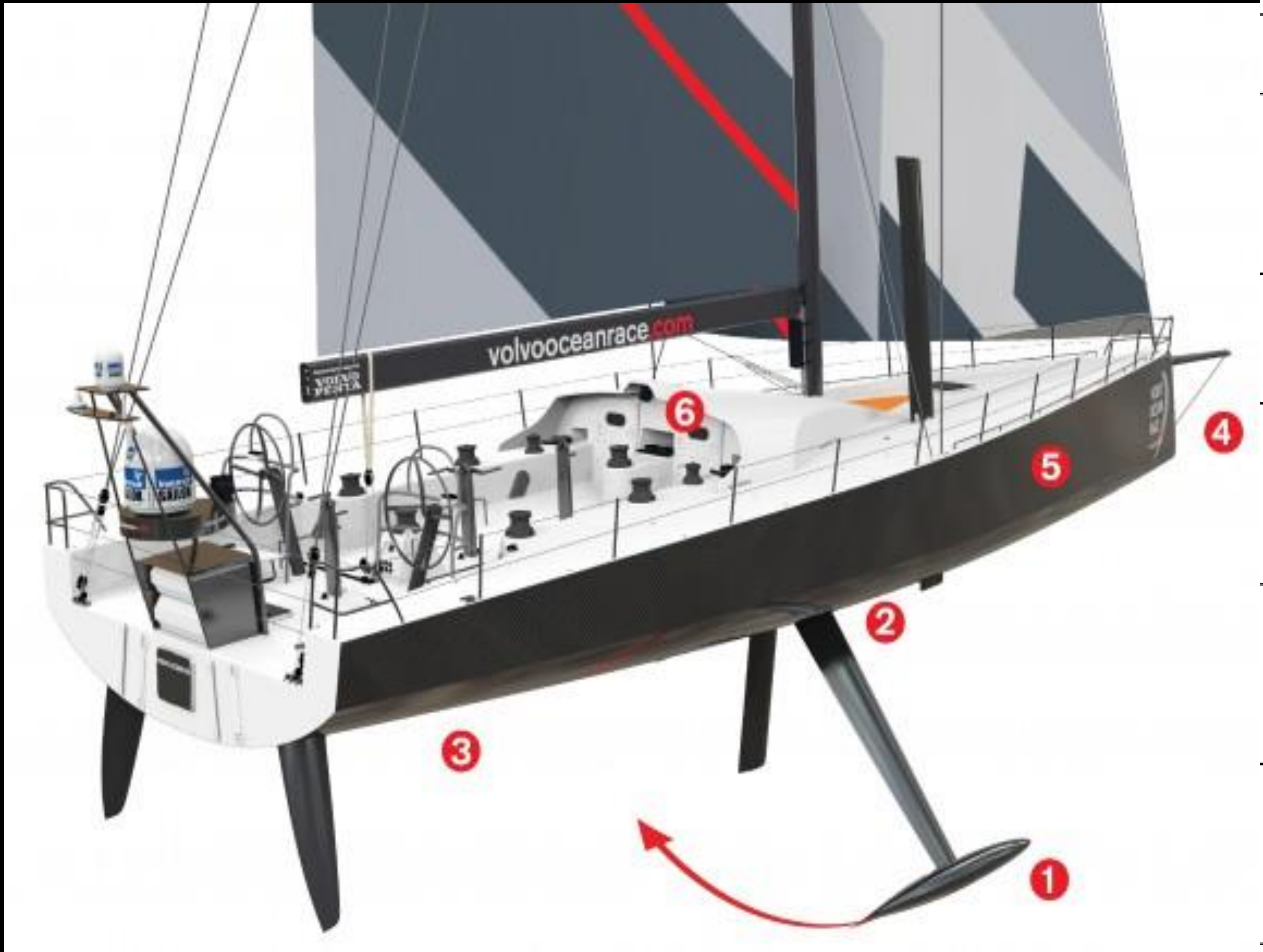
An example – "Help us to Sail Faster using Sensor Data" 12th Volvo Ocean Race



Wind Sensor Data	
True wind speed	Speed of the wind, without the motion of the boat.
True wind direction	Direction of the wind.
True wind angle	Direction of the wind relative to the ship.
Apparent wind speed	Speed of the wind as experienced on the moving boat.
Apparent wind angle	Angle of the wind as experienced on the moving boat



Navigation Sensor Data	
Boat speed	Speed of the boat including currents.
Compass heading	Course indicated by the compass.
Speed over ground	Speed of the boat excluding current.
Course over ground	Course relative to the land.
Position GPS	Coordinates of the boat's position.



Boat Sensor Data	
Heel	The angle that the boat heals to one side.
Trim	The angle that the boat heals forward/backward.
Rudder	The angle of the rudder.
Keel	The angle of the adjustable keel.

Data analytics support of Navigator
Andrew Cape and UT

Mark Vroling, Floris Smit
David Lamers (BIT), Tim Paauw (TBK/BIT),
Madelon Voets (Euros),
Peter Paul van der Wurff (BIT),
Monique van Leeuwen (BIT)


HARKEN
VOLVO OCEAN RACE
VOLVO OCEAN RACE

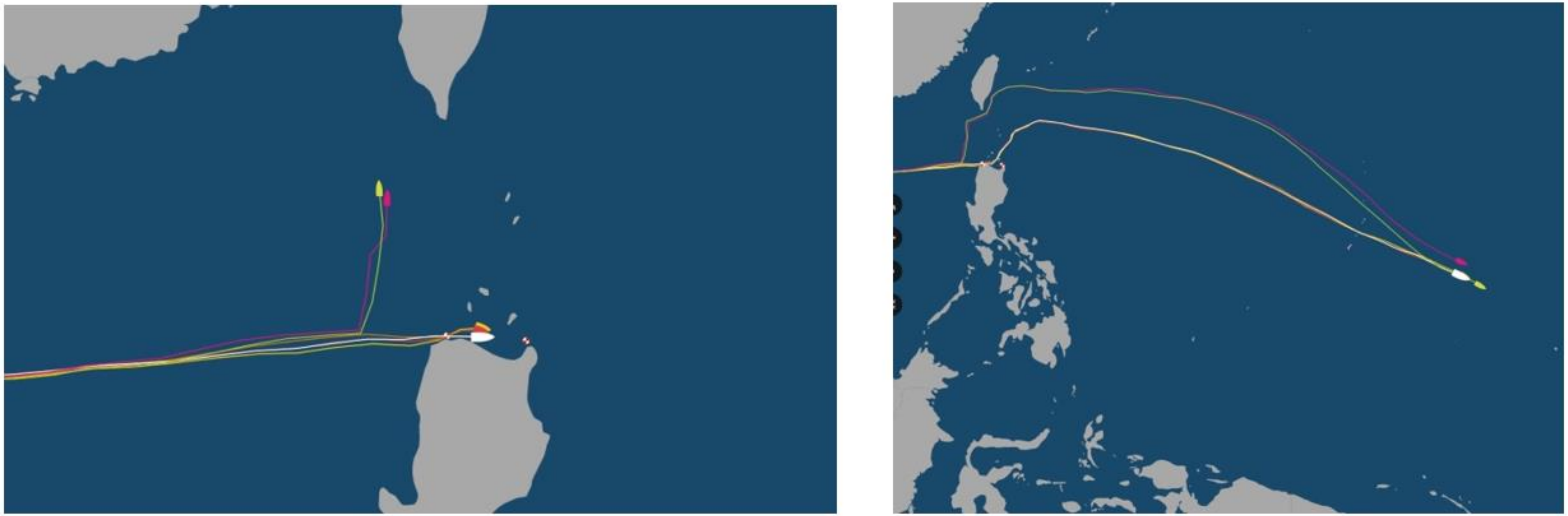


Figure 2: Example of a big decision by two teams to take the northern route that would later give a small but non-decisive advantage in the leg from Sanya to Auckland

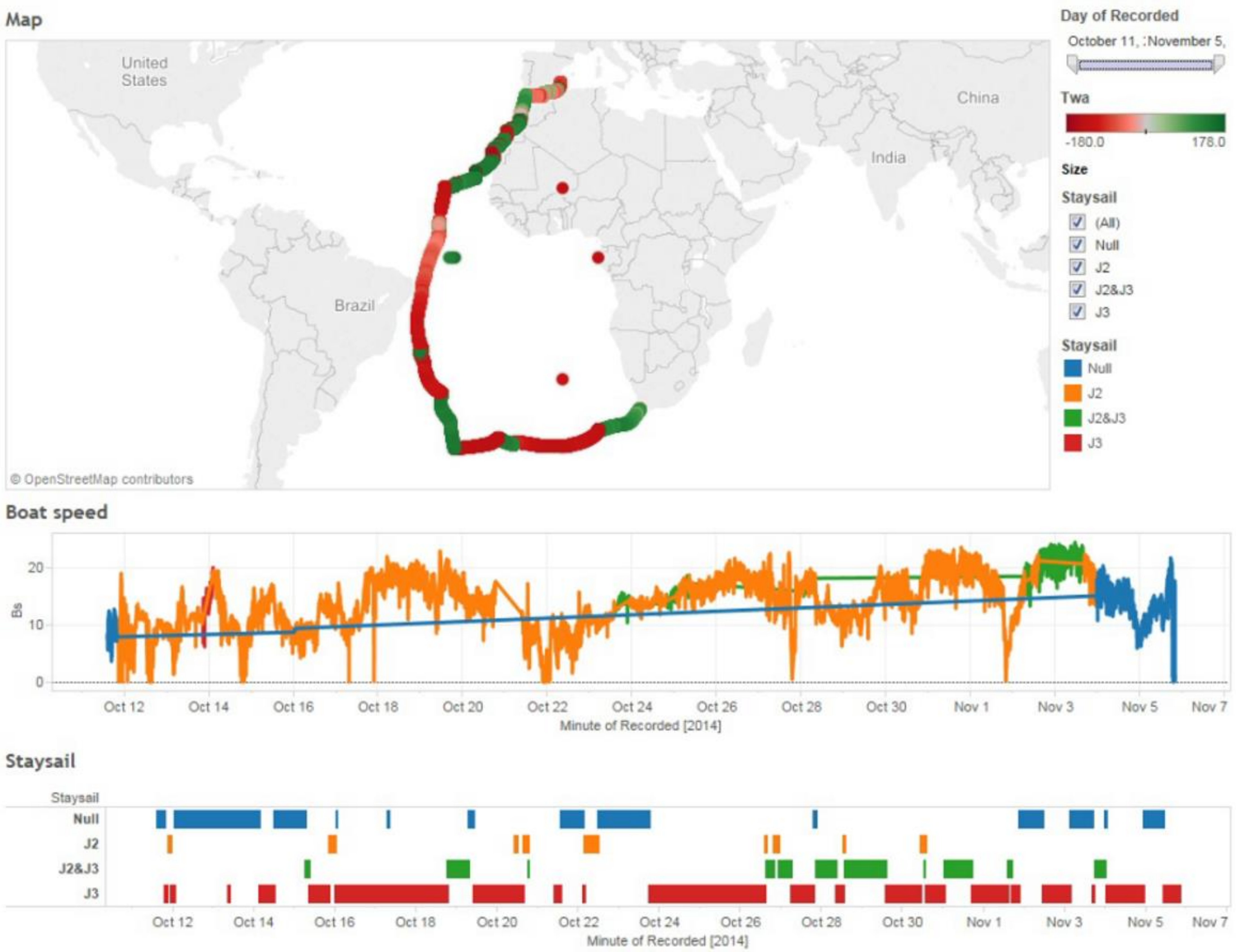


Figure 6: Map Speed Dasboard

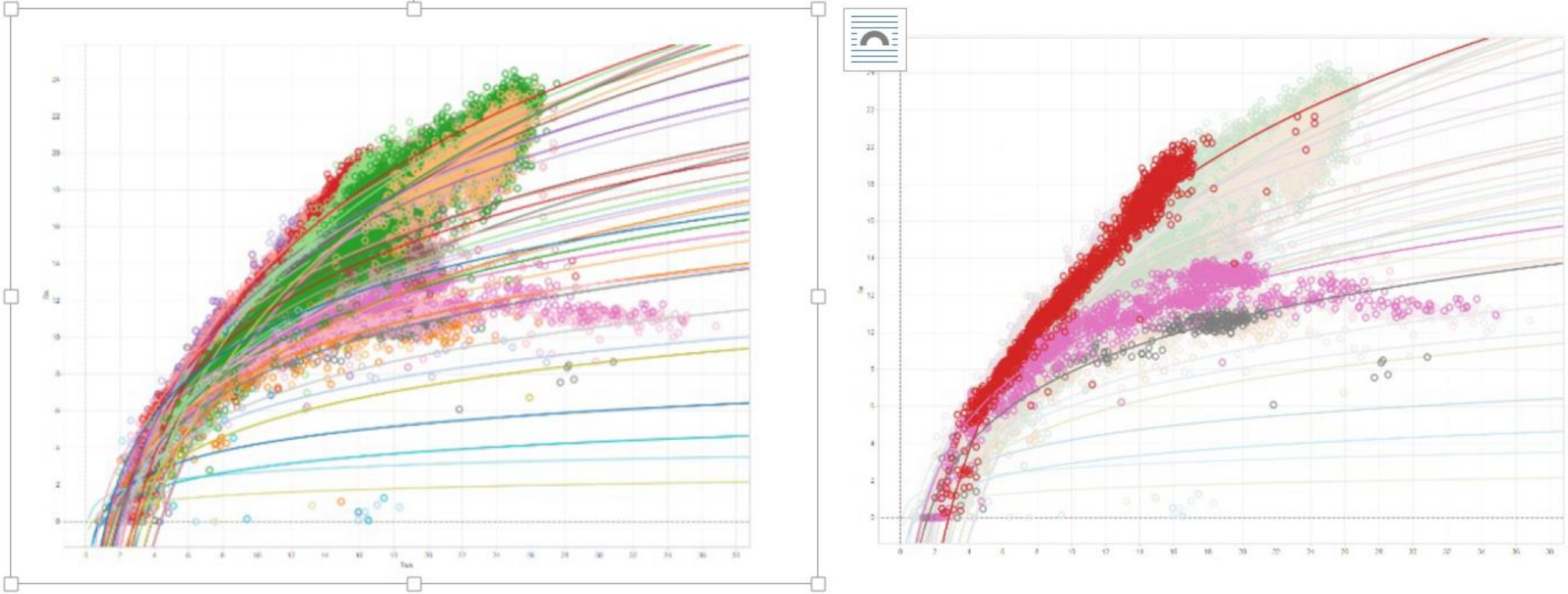
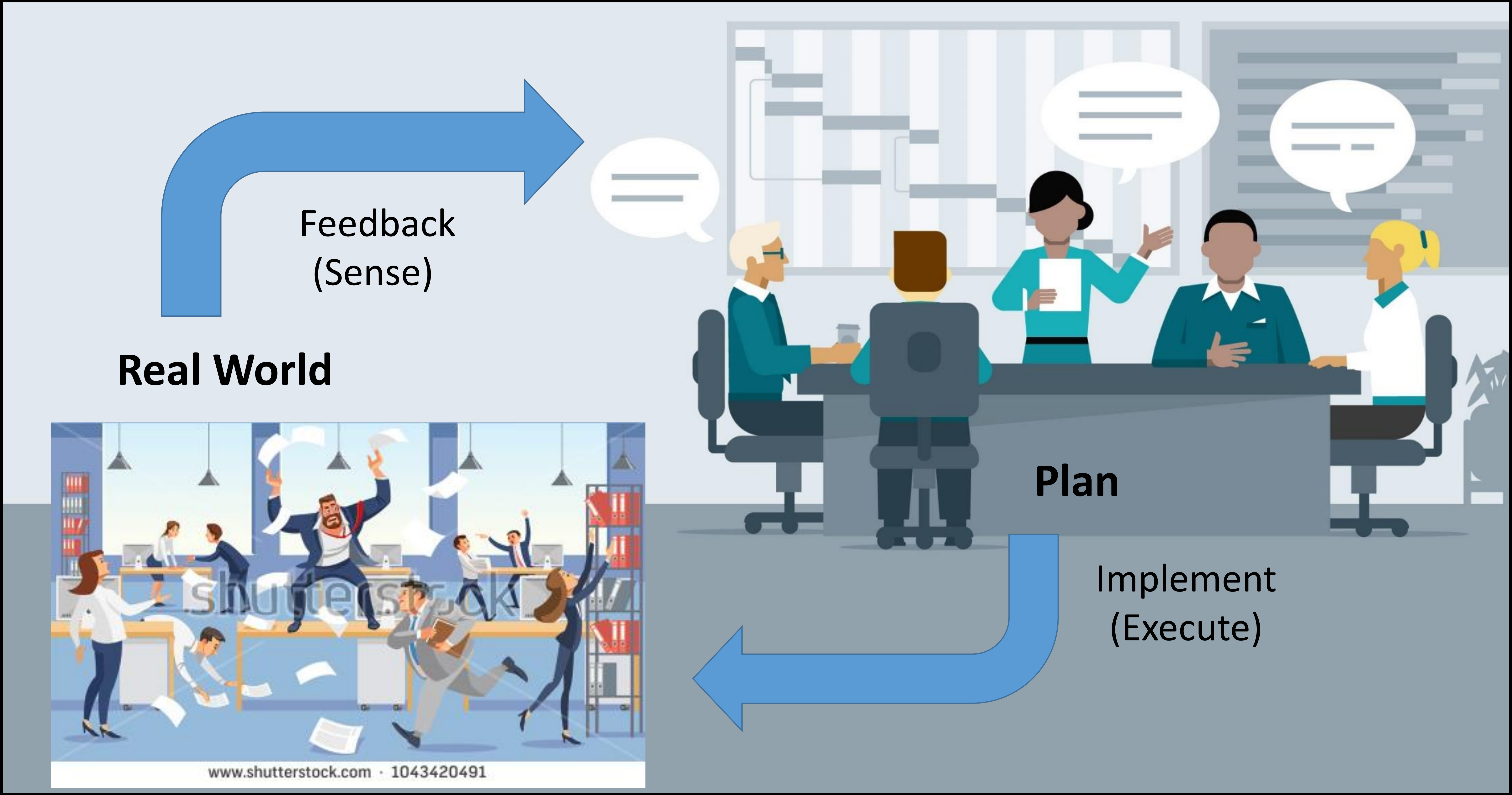


Figure 7: Wind Graphs

True Wind Angle (2 degrees)

True W..	40	42	44	46	48	50	52	54	56	58	60	62	64	66	68	70	72	74	76	78	80	82	84	86	88	90	92	94	96
36				J2	J1		J2											J1					J1		J1	J1		J1	J1
34		J1	J1	J1	J1	J2	J2	J2															J1	J1					J1
32		J2	J1	J1	J2	J2	J2	J2	J2	J1			J2						J1	J1			J1		J1		J1		
30			J2	J1	J2	J2	J2	J2		J1	J2	J2	J2	J2	J2	J2	J2	J2	J2				J2	J1			J1		
28	J1	J2	J2	J2	J2	J2	J2	J2	J1	J2	J2	J2	J2	J2	J2	J2	J2	J2	J2	J2	J2	J1	J1	J1	J1		J1		J1
26	J1	J2	J2	J2	J2	J1	J2	J2	J1	J1	J2	J2	J2	J2	J2	J2	J2	J2	J2	J2	J2	J1	J2	J1 …	J1	J1	J1	J1	J1
24	J1	J1	J1	J1	J1	J1	J1	J1	J2	J1	J2	J2	J1	J2	J2	J2	J1	J2	J2	J2	J2	J2	J2	J1	J1	J1	J1	J1	J1
22	J1	J1	J1	J1 …	J1	J1	J2	J1	J2	J1	J1	J1	J1	J2	J1	J2	J2	J1	J1	J2	J1	J1	J1	J1	J1	J1	J1	J1	J1
20	J1	J1	J1	J1	J1	J1	J1	J1	J1	J1	J1	J1	J1	J1	J1	J1	J1	J1	J1	J1	J1	J1	J1	J1	J1	J1	J1	J1	J1
18	J1	J1	J1	J1	J1	J1	J1	J1	J1	J1	J1	J1	J1	J1	J1	J1	J1	J1	J1	J1	J1	J1	J1	J1	J1	J1	FR	J1	J1
16	J1	J1	J1	J1	J1	J1	J1	J1	J1	J1	J1	J1	J1	J1	J1	J1	J1	J1	J1	J1	J1	J1	J1	MH	J1	J1	J1	J1	J1
14	J1	FR	J1	J1	J1	J1	J1	J1	MH	MH	J1	J1	…	J1	J1	FR	FR	J1	J1	J1	J1	J1	J1	J1	J1	J1	J1	J1	J1
12	MH	J1	MH	FR	J1	J1	MH	J1	J1	FR	FR	FR	FR	FR	FR	FR	FR	FR	FR	J1	MH	MH	FR	MH	MH	FR	J1	J1	J1
10	J1	J1	J2	J1	FR	FR	FR	FR	FR	FR	J1	FR	J1	J1	J1	FR	FR	FR	FR	FR	FR	FR	FR	MH	FR	FR	FR	…	FR
8	MH	MH	MH	A3	FR	J1	MH	J1	MH	J1	J1	FR	FR	J1	FR	FR	MH	FR	FR	MH	FR	FR	FR	MH	MH	MH	MH	FR	FR
6	MH	MH	MH	MH	MH	MH	MH	MH	MH	MH	MH	MH	MH	MH	MH	MH	MH	MH	FR	MH	MH	MH	MH	MH	MH	MH	MH	MH	MH
4	MH		MH	MH	MH	MH	MH	MH	MH	MH	FR	MH	MH	MH	MH	MH	MH	MH	MH	MH	A3	MH	MH	MH	MH	MH	MH	MH	MH


Feedback
(Sense)
Real World
Plan
Implement
(Execute)
www.shutterstock.com · 1043420491

Data mining as solution to detect fraud

van Capelleveen, G., Poel, M., Mueller, R. M., Thornton, D., & van Hillegersberg, J. (2016). Outlier detection in healthcare fraud: A case study in the Medicaid dental domain. *International journal of accounting information systems*, *21*, 18-31.

Following the money using medical claim reimbursements

- Who are the gate keepers?
- Go for prevention or detection?
- How to deal with a complex changing nature of health care domain?
- How to deal with continuously changing type of fraud?
- What to do with multiple domains in health care?
- How to determine if identified fraud is really fraud?

- **Fraud:** Purposely billing for services that were never furnished and or supplies not provided, medically unnecessary services and altering claims to receive higher reimbursement than the service produced.

- **Abuse**: The billings of practices that, either directly or indirectly, are not consistent with the goals of providing patients with services that are medically necessary, meet professionally recognized standards, and are fairly priced.

Grey area

Data mining fraud: a multi-disciplinary problem

Who is doing what?

- **Medical subject matter expertise**
- **Fraud expertise**
- **Data science expertise**

Fraud approaches in healthcare

Who commits fraud?

- Providers, a group of providers, patients, insurance provider, combinations?

Different systems, different fraud!

- Pay and chase or a gate keeper?
- Fixed fee, variable costs, diagnosis related
- Negotiated procurement
- Explanation of benefits?

How to fraud?

- Hit and run
- Steal a little all the time

Fraud Schemes in healthcare

- **Billing for services not rendered** (identity theft & phantom billing)
- **Up-coding of services and items**
- **Duplicate billing**
- **Un-bundling of claims** (or creative billing)
- **Medically unnecessary services** (bill padding)
- **Excessive services** (bill padding)
- **Kickbacks**

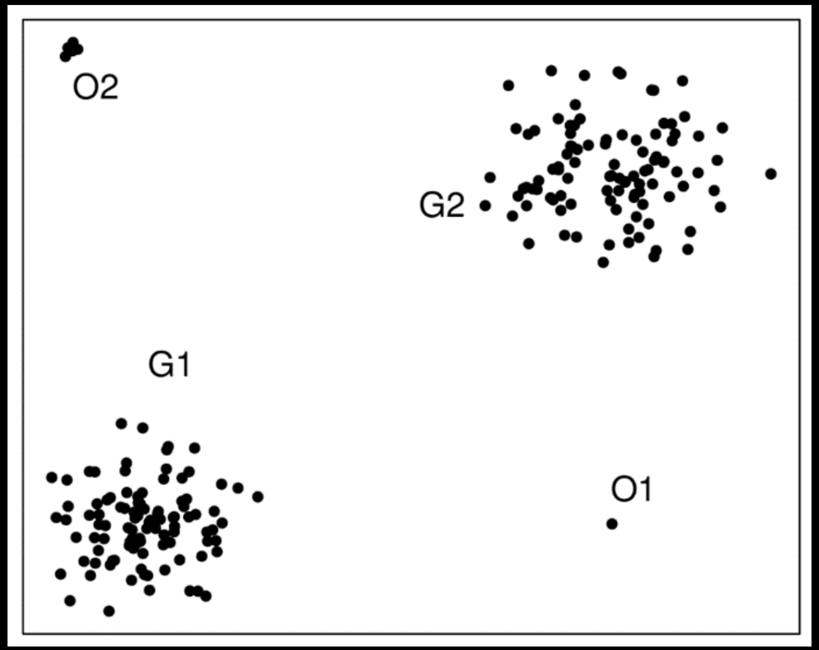
Sparrow, M. K. (2000). *License to steal: how fraud bleeds America's health care system*. Basic Books.

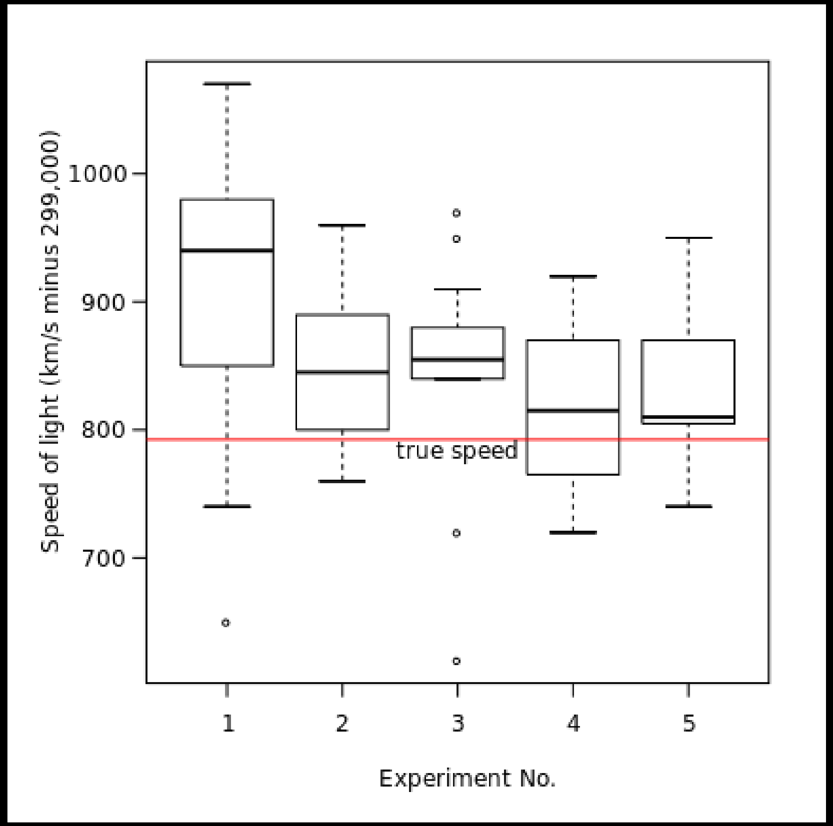
The healthcare data set

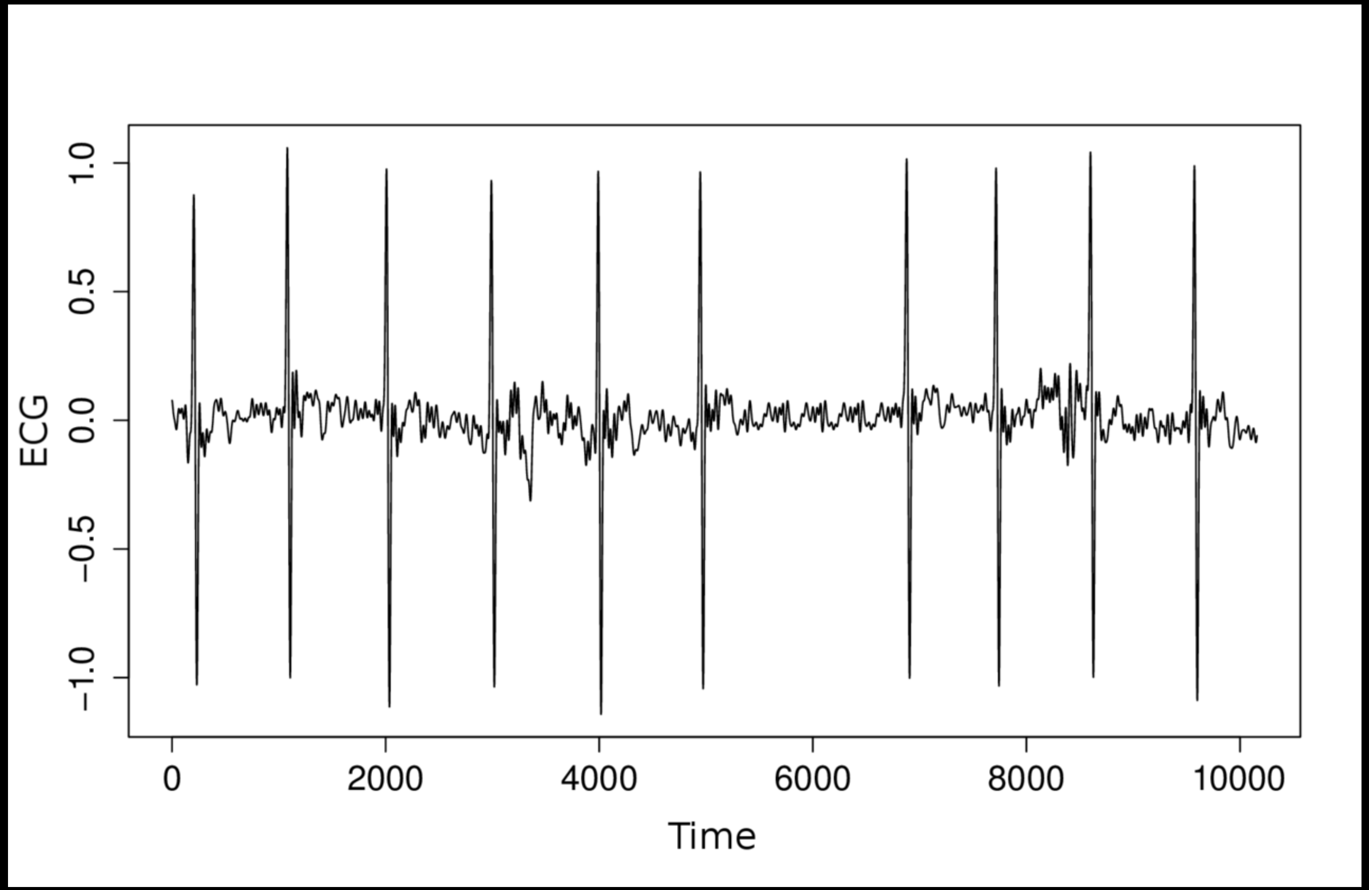
- Health insurance claim line dental data
- 30 providers from California (Medicaid insurance)
- Only data from 2016
- Cleaned (no null values, no messy data lines, values and reference are correct)
- Denormalized (no need for the reference tables)
- Adjusted (claim history already processed into the claim no., need to bind claims)
- Thus, minimal data wrangling

Outlier detection

- What is an outlier?
  - An outlier is an observation point that is distant from most of the other observations that group together or follow a pattern or distribution.







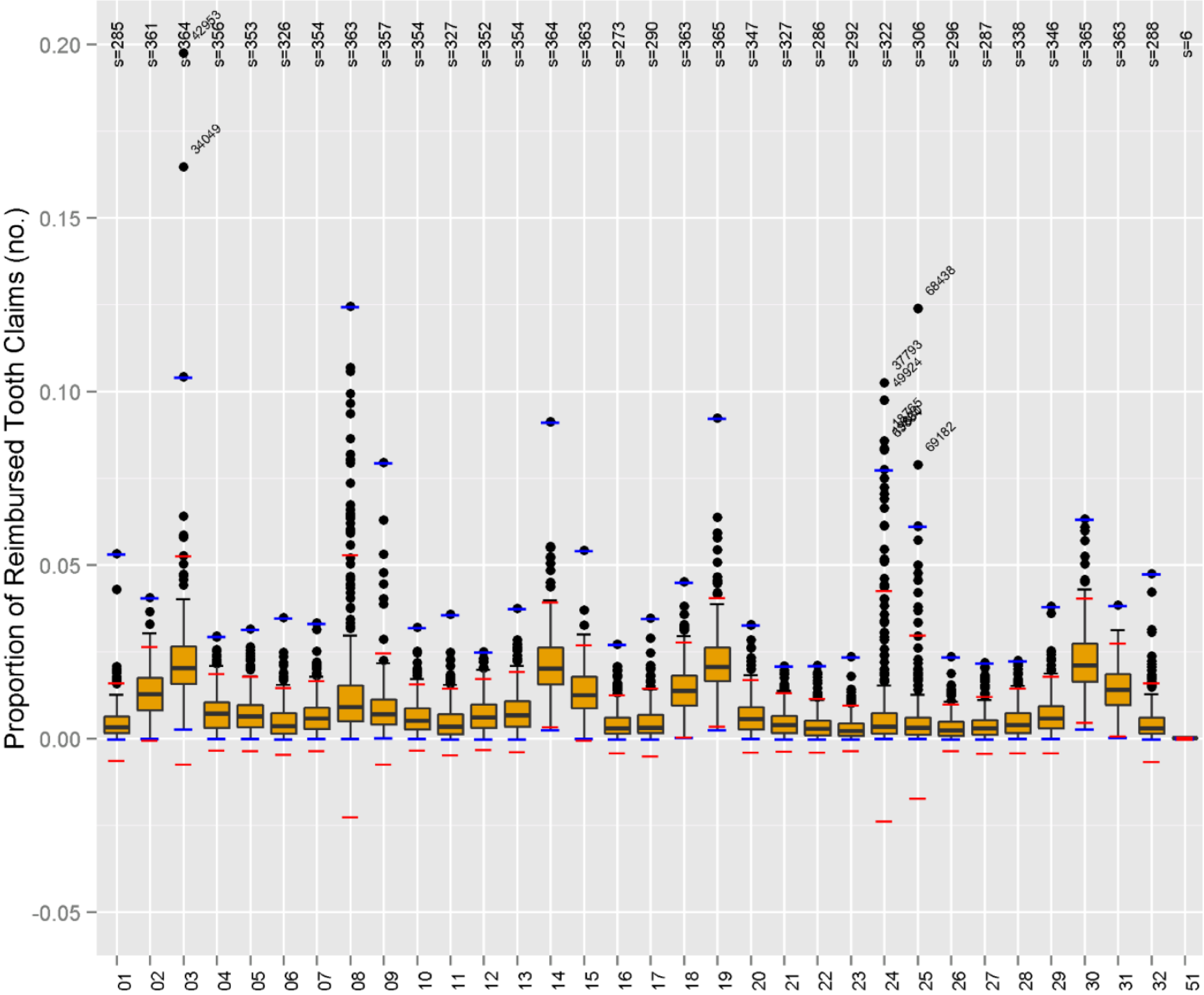
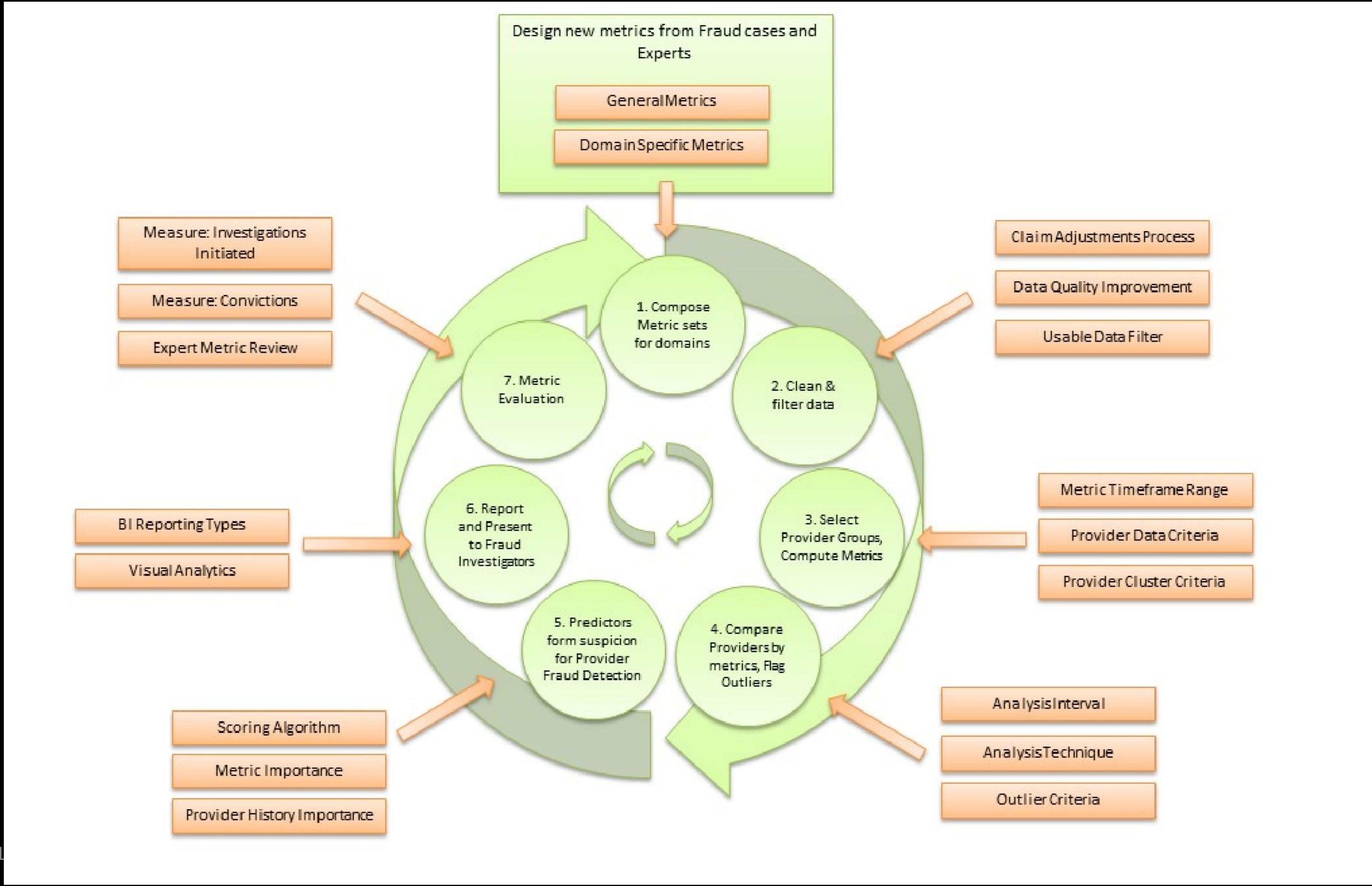
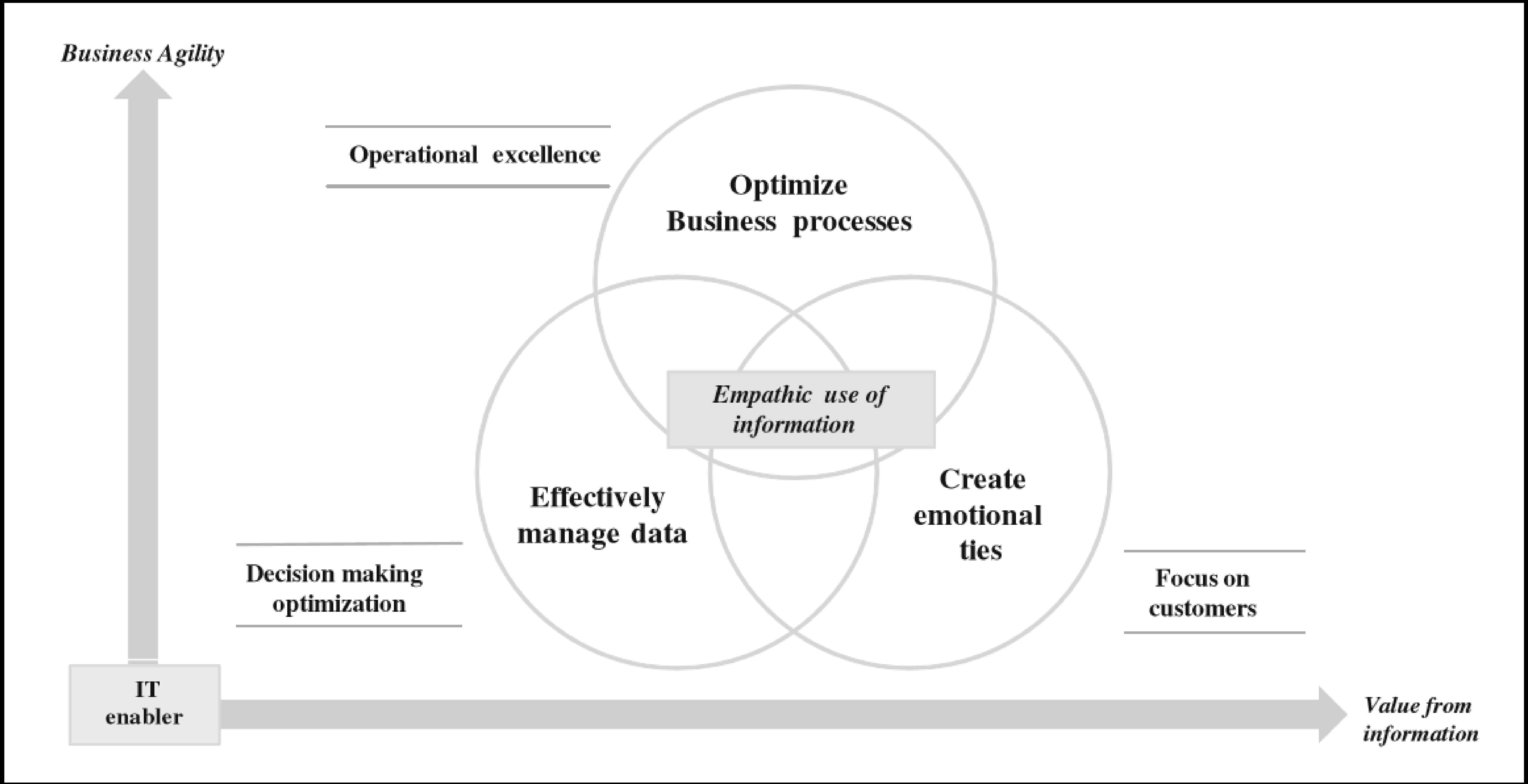


Fig. 3. Tooth Code Analysis


Design new metrics from Fraud cases and Experts
General Metrics
Domain Specific Metrics
Measure: Investigations Initiated
Measure: Convictions
Expert Metric Review
Claim Adjustments Process
Data Quality Improvement
Usable Data Filter
1. Compose Metric sets for domains
2. Clean & filter data
3. Select Provider Groups, Compute Metrics
4. Compare Providers by metrics, Flag Outliers
5. Predictors form suspicion for Provider Fraud Detection
6. Report and Present to Fraud Investigators
7. Metric Evaluation
Metric Timeframe Range
Provider Data Criteria
Provider Cluster Criteria
BI Reporting Types
Visual Analytics
Scoring Algorithm
Metric Importance
Provider History Importance
Analysis Interval
Analysis Technique
Outlier Criteria

How to use big data?

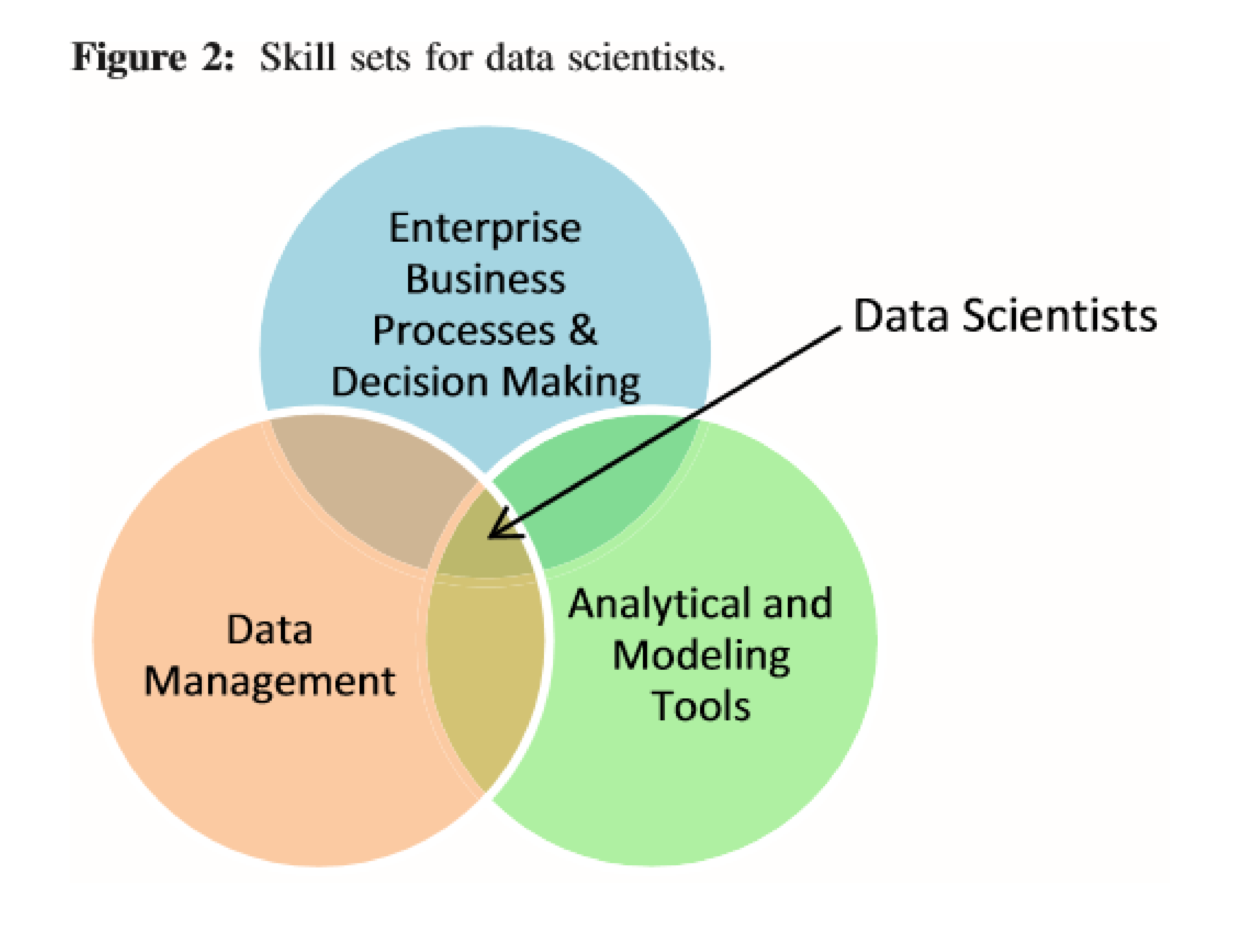


Source. Morabito, V. (2014). *Trends and challenges in digital business innovation*. New York: Springer International Publishing.

Tools for big data analytics
just examples by no means a ranking!

- Preparation, Cleaning, Transforming, Wrangling
  - Alteryx, Trifacta, Paxata,…
  - Openrefine.org (open source)
- Analytics, Inference, Mining, Visualization
  - R, SAS, SPSS, Tableau, Qlick, PowerBI, Statistica, Knowledge Studio, Google FusionTables,..
  - knime.org (open source)
- Mining, AI
  - IBM Watson analytics, Wolfram Alpha
  - Rapidminer.com, Weka (open source)
  - …..

THE WAY FORWARD: SKILLS



Source: Schoenherr, T., & Speier-Pero, C. (2015). Data science, predictive analytics, and big data in supply chain management: Current state and future potential. *Journal of Business Logistics*, *36*(1), 120-132.

kaggle

Host a competition

« Back to homepage

Overview | Business | Recruiting | Research | Get started

Kaggle makes it easy to connect your data with data scientists.

See what our community of 600,000+ data scientists can do for you.

Kaggle is a platform for data science competitions. We help you solve difficult problems, recruit strong teams, and amplify the power of your data science talent.

Interested in hosting a classroom competition? Visit **Kaggle in Class »**

For Business Problems

Bring in the world's largest community of data scientists to consult on your toughest business problems.

Learn more

3.Analytics Techniques
(Explore, Explain, Predict)

..
..
..
..
..
..
..

1.Questions
(no overview, no insight, no foresight)

..

2.Owner
(Who wants to know?)

..
..
..
..
..
..
..

4.Data Sources
(internal, external, open)

..
..
..
..
..
..
..

5.Changes needed
(people, process, organization)

..
..
..
..
..
..
..

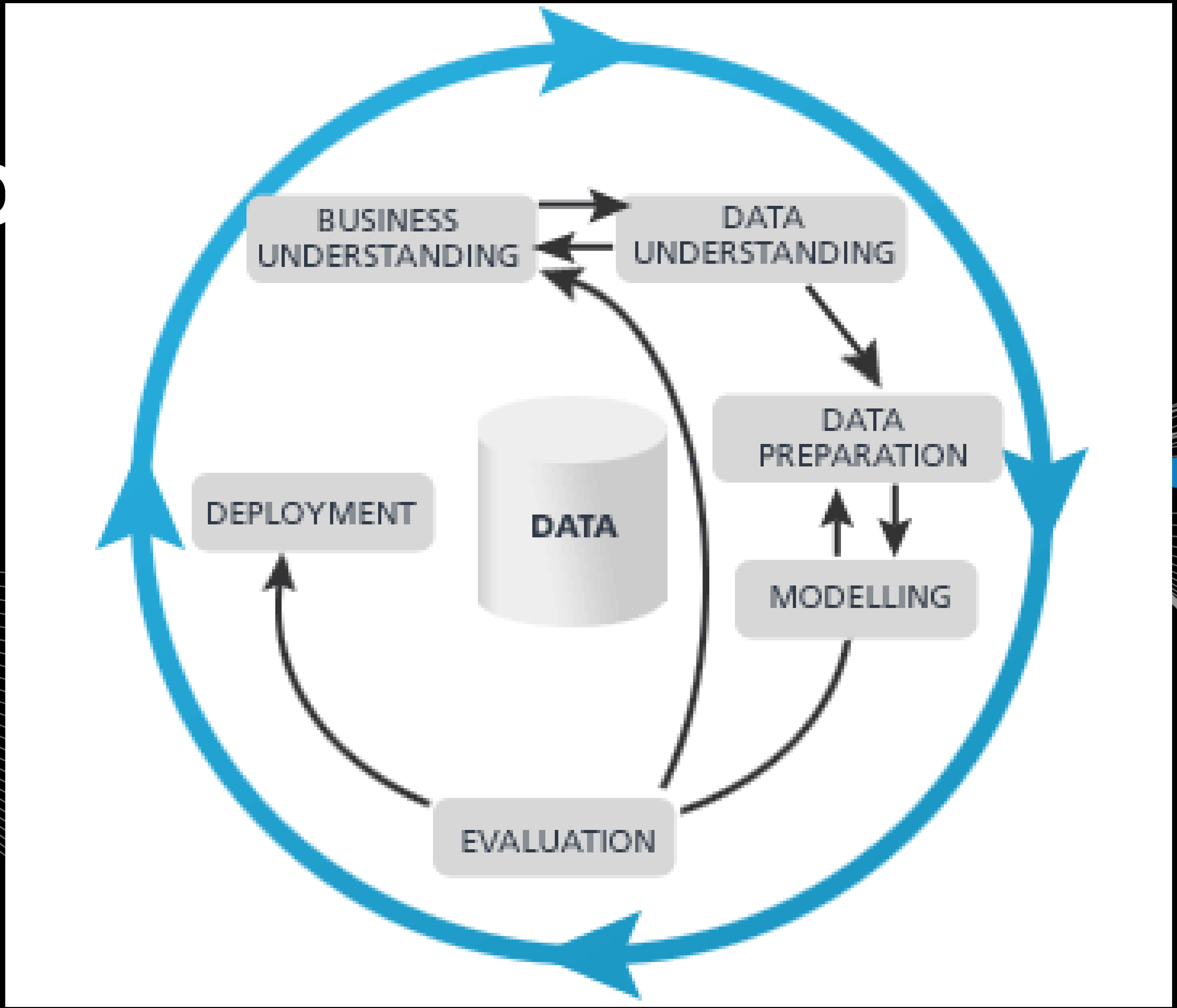
6.Challenges
(data availability and quality, skills, risks, regulation,,..)

..
..
..
..
..
..
..

Version:	Organization:	Name:

Source: (Big data canvas J van Hillegersberg ©)

Summary and Questio



Jos van Hillegersberg
J.VANHILLEGERSBERG@UTWENTE.NL