2018 Predictive Analytics Symposium

Session 08: M/S - Industry Best Practices for Data Protection

SOA Antitrust Compliance Guidelines
SOA Presentation Disclaimer

Industry Best Practices for Data Protection

Brad Lipic
Vice President, Head of Data Strategy
Global Research and Data Analytics (GRDA)

9.19.2019

Agenda

- Data Trends
- Data Protection Trends
- Supportive Frameworks and Tools for Data Protection

A Presentation of Two Halves

Establish a defense…

…and enable an offense (value creation)

2018 Trends to Follow

Big Data

- Cybersecurity and data protection will become major influencers of data strategy.

- Life insurers will be using cloud computing to access applicant health and medical data to expedite the underwriting process.

- Blockchain-based insurance contracts will allow for more efficient claims processing and combat insurance fraud.

Consumer Behavior

- Insurance gets gamified. Insurers are using apps and web-based games to engage and educate potential policyholders.

- Consumers are more willing to share their wearable data for incentivized premium discounts.

ge recog
s to use se
rmine age,
ealth habit

Insurance Distribution

- More insurers are using artificial intelligence to advise consumers. Robotic advisers can also assist with risk exposure assessment and policy administration.

- Customers demand multi-channel access for a smarter and more seamless way of accessing life insurance services.

- Accelerated underwriting programs will be expected to provide a more individualized experience to better fit consumer needs.

Insurance Services

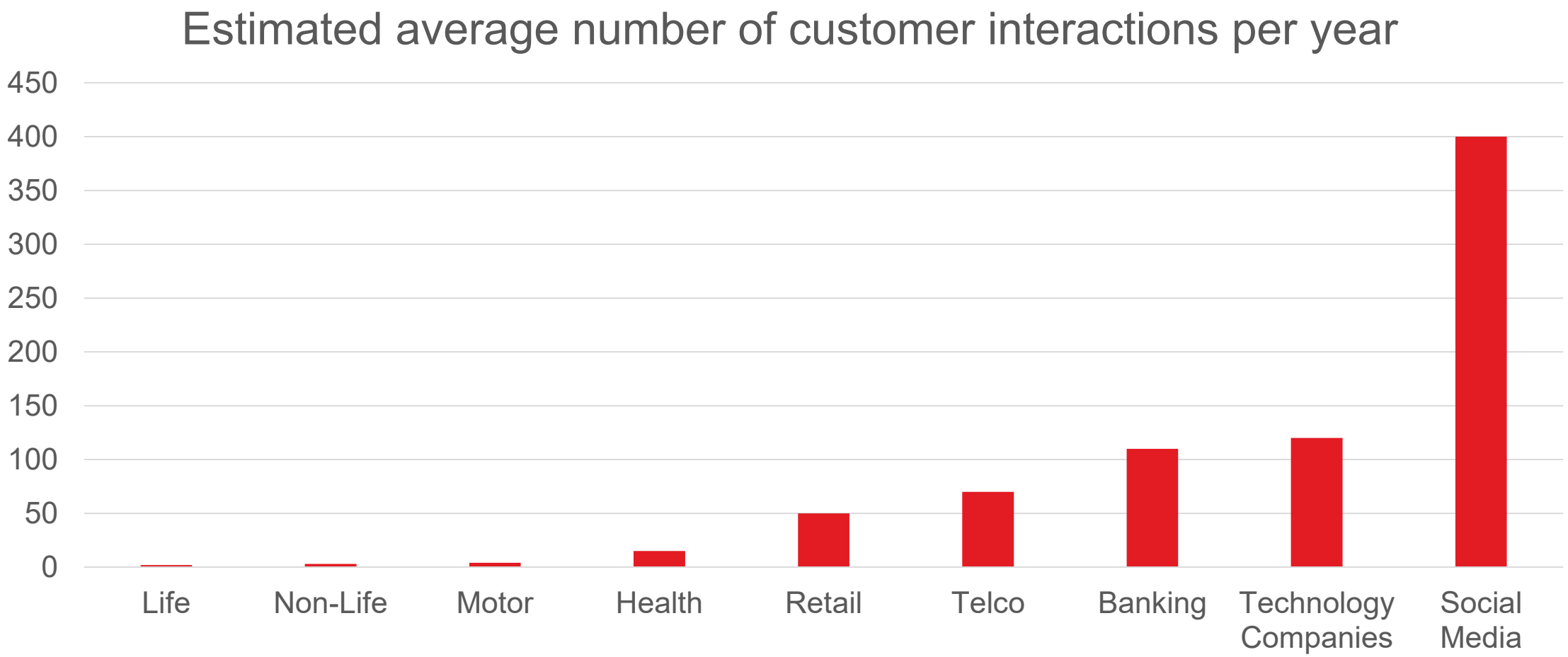
- Mobile-based micro-insurance policies offer life insurance coverage for shorter durations and convenient mobile premium payments.

- More life insurers will be hiring and acquiring InsurTech companies to develop and offer more personalized policies.

- Peer-to-peer (P2P) insurance is disrupting the industry by allowing insureds to pool resources to develop a personalized insurance network with like-minded

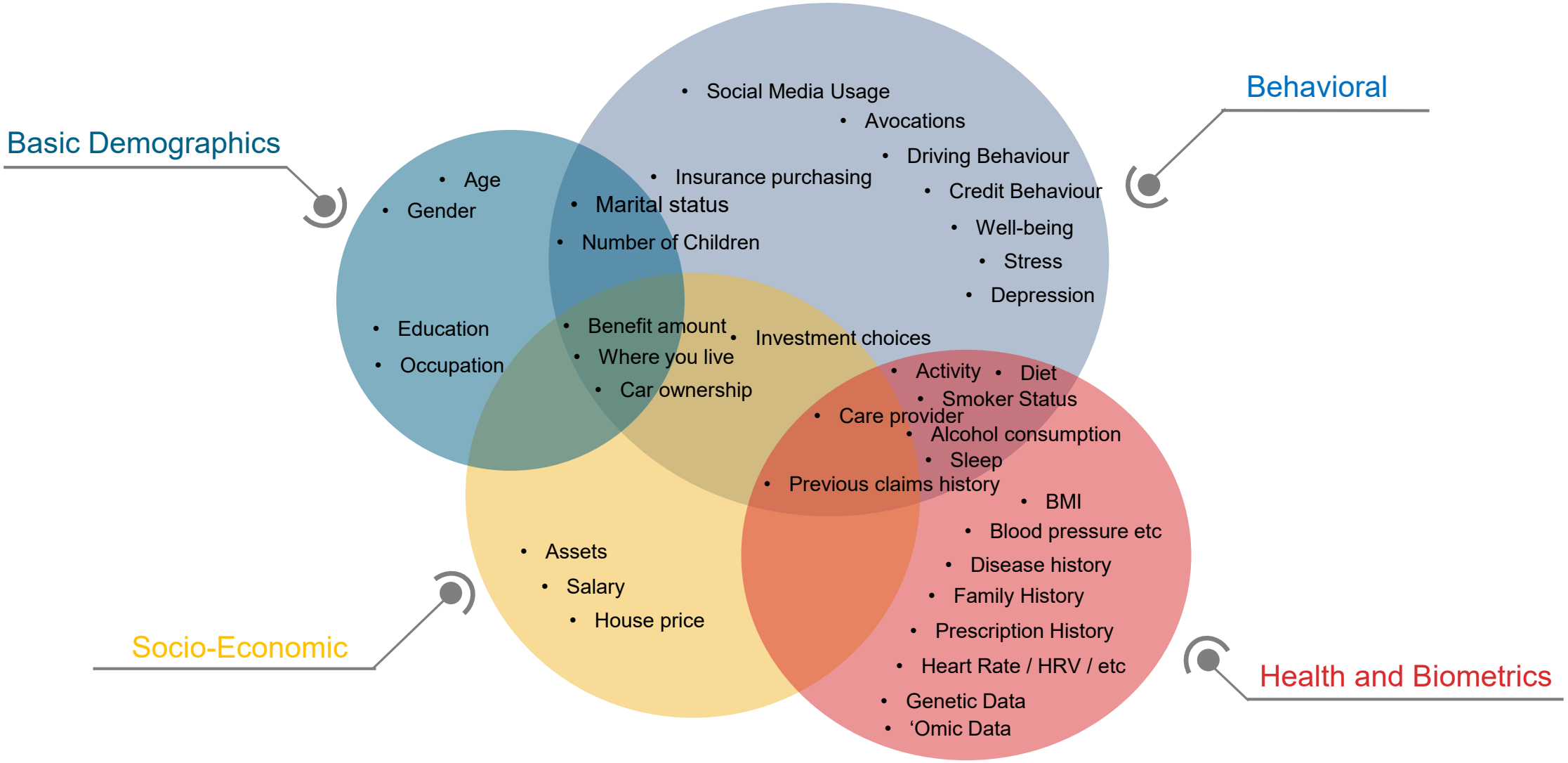
Personal Data Is Constantly Being Collected

What Data Could We Use?



Source: McKinsey & Company, presented in Transforming Life Insurance with Design Thinking

A Wide Variety of Information Is Being Accumulated and Used



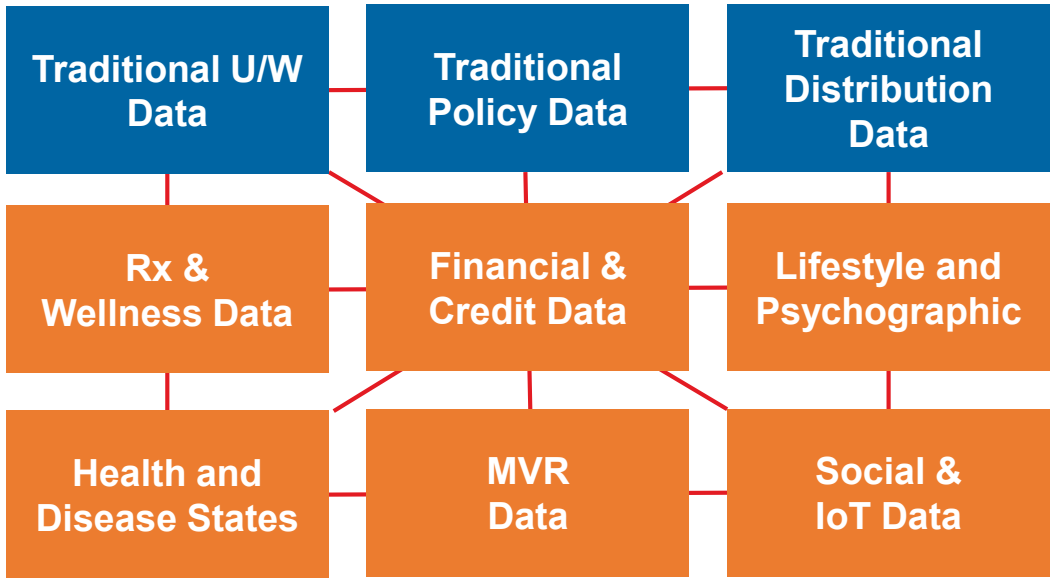
Linking Data Sources

Unique insights and robust solutions to business problems do not exist within a single data source…neither should our analyses and solutions

Mary

There's something about Mary…

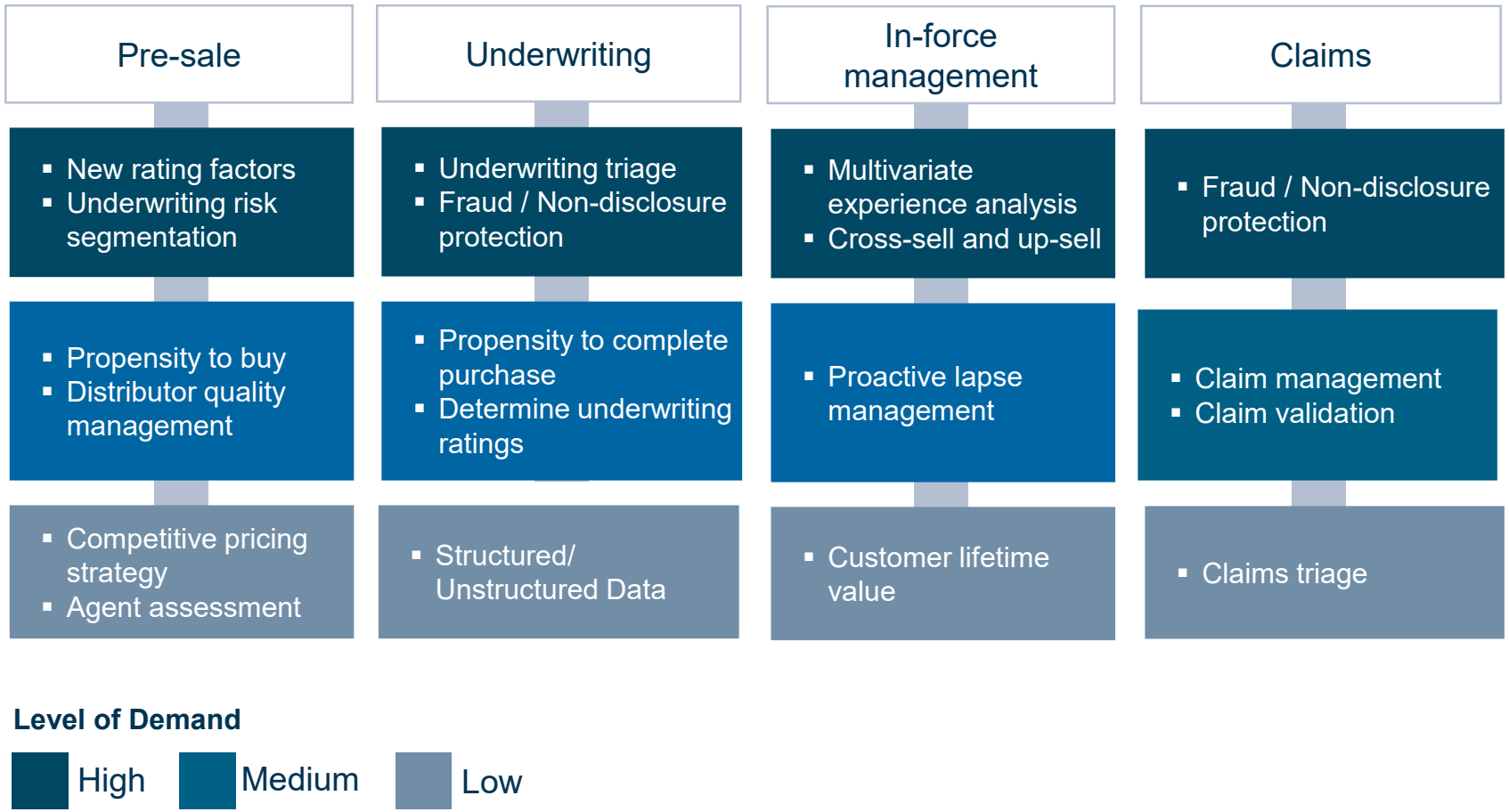
Datasets, in combination, can create unique insights about Mary



PII is required for linkages

Without PII, we become limited in our ability to better understand the multifaceted forces that predict future life events of Mary (and others)

Insurance Value Chain and Applications



Data Protection Trends and Regulatory Climates

In the News: Data Breaches

According to Forbes, in 2018, an average data breach incident costs the highest in U.S. as $7.91 million

AT&T Hit With Record-Breaking $25 Million Data Breach Fine

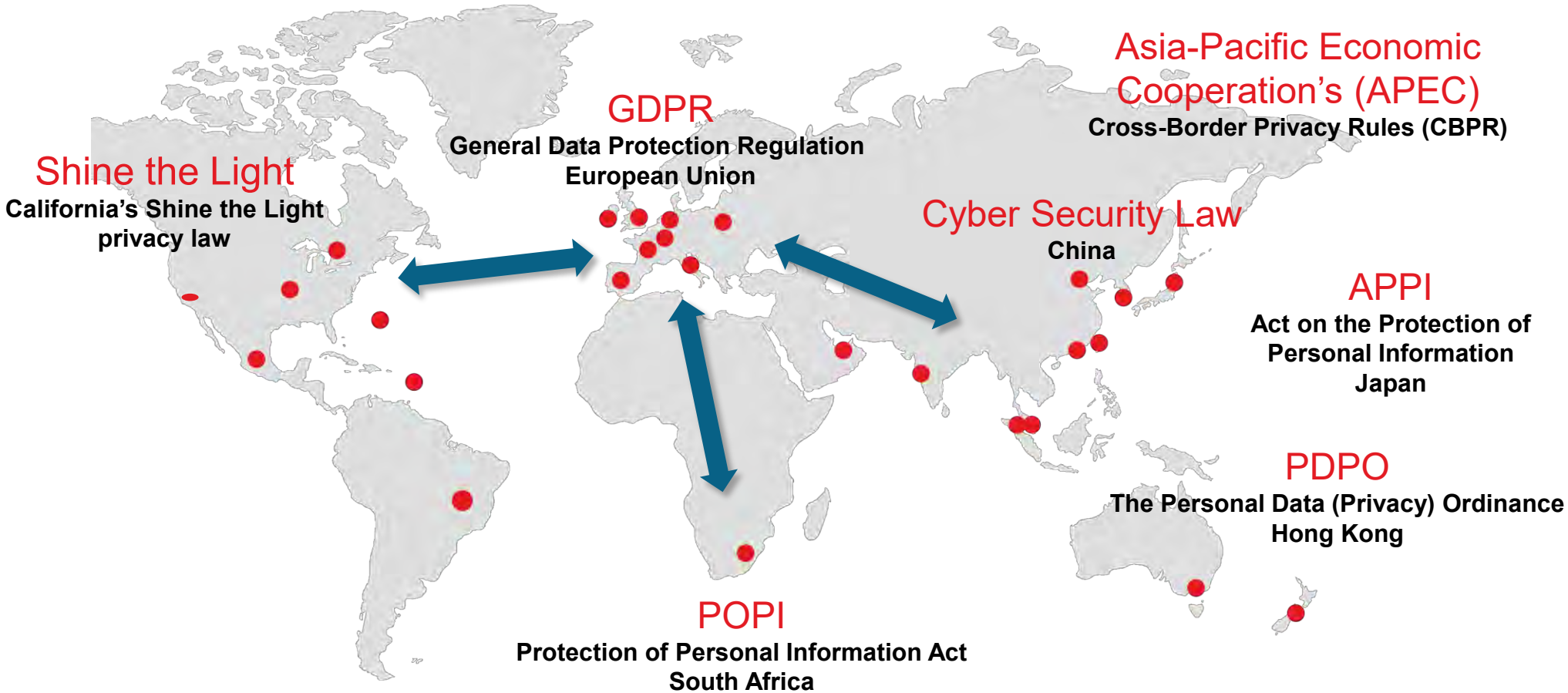
Yahoo fined £250,000 over 2014 data breach

https://news.sky.com/story/yahoo-fined-250000-over-2014-data-breach-11402490

https://www.esecurityplanet.com/network-security/att-hit-with-record-breaking-25-million-data-breach-fine.html

https://digitalguardian.com/blog/hilton-was-fined-700k-data-breach-under-gdpr-it-would-be-420m

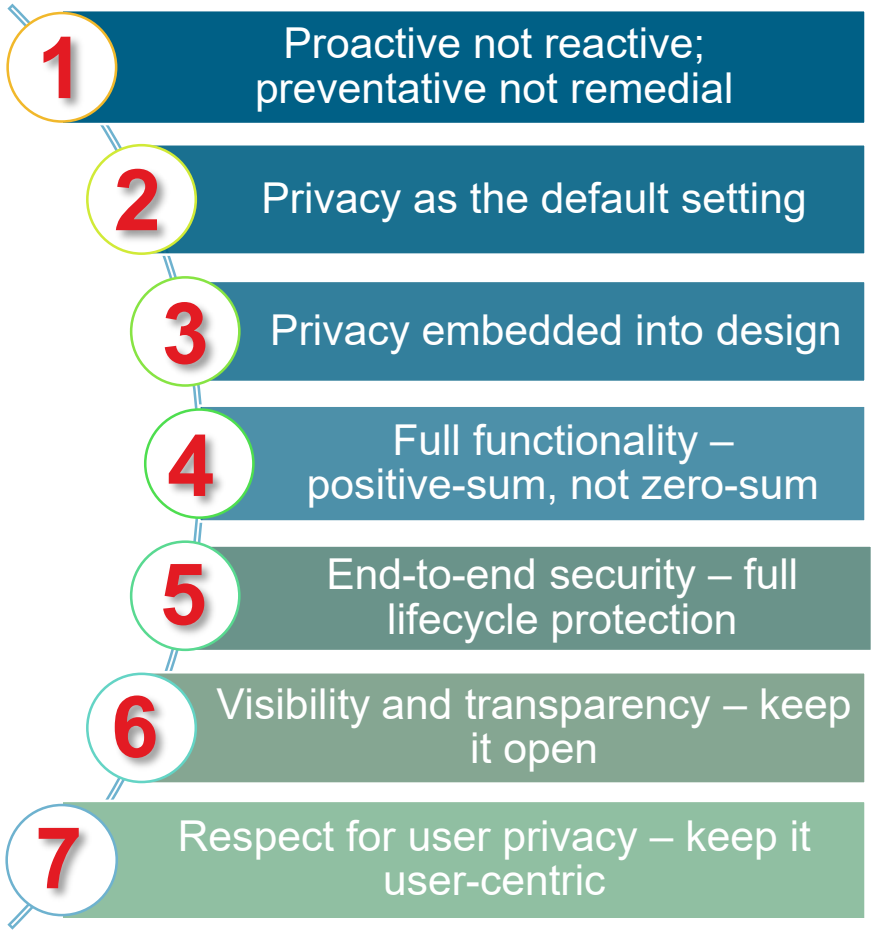
Regulations Driving Change Around the Globe



Supportive Frameworks and Tools

Privacy by Design:
A Framework Developed by Ann Cavoukian

Privacy is to be taken into account as it relates to design and operation throughout the whole of IT systems, networked infrastructure, and business practices



Take action in advance of the event, not after-the-fact

Responsibility is on IT system design; not the individual

Core and integral; not an add-on with potential diminished utility

Win-win in for all legitimate business purposes; not a trade-off between privacy and security

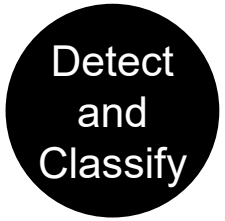
Data onboarding through destruction and all steps in-between are to be included

Ensure the system is subject to ongoing compliance and validation to stakeholders

Empower the user with strong privacy defaults and end-user options

Data Protection Capabilities

Provides visibility into the data available in the an ecosystem sourced from internal and external sources. This may include origin or storage source, data content, usage, or other elements the organization deems important to see and track



Analyzing the data set to identify fields that could contain PII, as well as combinations of fields that could be used to identify an individual. Data is then classified according to defined sensitive fields

Evaluates two or more data records containing the same, or similar data elements to make a determination if they are for the same individual

Protecting sensitive data through a variety of techniques, including anonymization, pseudonymization or redaction, and enabling only authorized users to see specific data elements in the clear

Capability: Catalog

Ability to register and track data sources

Benefits	Risks
▪ Enhances the data onboarding process ▪ Provides visibility and single source of truth for data, sources and projects using them ▪ Opportunity to cross-leverage sources ▪ Identifies gaps for future data acquisitions ▪ Enables regulatory compliance efforts	▪ Potential security concern given catalog could be exposed ▪ Potential financial impact at different capability levels

Measures of Success

- # of sources
- % of sources in catalog
- Date range of data available
- Geographies available
- Attributes available by geo or demographic
- Where PII is located
- Time spent cataloging

Capability: Detect & Classify

Ability to identify fields or combinations of fields that contain or constitute PII / SPII

Benefits	Risks
▪ Enables regulatory compliance (e.g. GDPR) ▪ Provides visibility into location of PII in applications and data sets ▪ Increases auditability ▪ Creates organizational consensus and clarity on PII definitions	▪ Mishandling of PII, due to vague definitions ▪ Hefty regulatory fines due to regulatory noncompliance ▪ Adoption resistance depending on level of maturity ▪ Decreased productivity due to additional process steps

Measures of Success

- # of data sets classified
- # of classified fields/columns
- Adoption by business units
- Dollars fined
- # of classifications based on regulatory entity
- # of instances estimated vs. observed
- Location of PII

Capability: Master Person Index

Evaluates two or more data records containing the same, or similar data elements to make a determination if they are for the same individual

Benefits
▪ Reduces time spent on manual linkage ▪ Manages jumbo risks / retention limits ▪ Builds trust in assigning aliases, across different partners and different data sets ▪ Increases data quality

Risks
▪ Could lead to inefficient use of partner data sets if incorrectly linked ▪ Potential data redundancy ▪ Potential flaws in procedures, selection bias

Measures of Success

- Reduction in record linkage time
- % of false positives or false negatives
- % accuracy based on samples taken
- # of new partnerships acquired over time
- % of enterprise adoption
- # of linkable datasets

Capability: Protect

Ability to protect PII from unnecessary exposure in the enterprise pipeline

Benefits	Risks
▪ Minimizes risk and engenders trust ▪ PII/SPII unintended disclosure risk is reduced ▪ Enterprise-wide implementation would ensure GDPR compliance	▪ Removal of fields could impede analysis ▪ Adoption resistance ▪ If policy definitions are too lenient the tool may not be leveraged efficiently ▪ Enterprise, integration risk due to potential number of touchpoints

Measures of Success

- Relies on the detection capability to quantify protected data
- Data risk assessment reduction
- # of supported applications
- # of fields
- # of users
- % of enterprise adoption
- # of new external partners acquired over time

A Few Protection Methods

Data Protection	Anonymized		Alternatives	
	Anonymization	Differential Privacy	Tokenization	Psuedonymization
Webster	Remove identifying information from something so that the original cannot be known	(None)	Symbol representation, or distinguishing feature	Use of a fictitious name
In-Data Practice		Method that seeks to maximize accuracy while minimizing the ability to identify the identity of data subjects	▪ Processing of data that can no longer be attributed to a specific data subject without the use of additional information ▪ Random generation of a value for plain text, which is then stored in a mapping database	
Distinguishing Features	One-way; permanently removes the personally identifiable information with no way of getting back to it	Adds random noise to the data while retaining meaingful aggregate statistics; can be applied to the "ride-along" data attributes	Can switch the data between the "masked token" and "in the clear" as it moves through workflows	Adds a field with a pseudonym associated with the identity; preferred method when needing to link data sources on the same individual identity

Anonymization

Raw Input

National ID	Name	DOB	HIV Indicator	Smoker Indicator	BMI
123456789	John Doe	10/12/1983	Y	Y	24.5

Protected Output

National ID	Name	DOB	HIV Indicator	Smoker Indicator	BMI
			Y	Y	24.5

Advantages

- Low risk of re-identification
- Theoretically impossible to get back to the identify of an individual (if anonymized properly)

Disadvantages

- Static
- Cannot merge with other data sources

Differential Privacy – The Issue

National ID	Name	HIV Indicator
123	John	Y
456	Jane	Y
789	Bob	Y
111	Betsy	N
333	Zach	N

Public Aggregate Analysis:
3 out of 5 people have HIV

Differential Privacy – The Issue

National ID	Name	HIV Indicator
123	John	Y
456	Jane	Y
789	Bob	Y
111	Betsy	N
333	Zach	N

Public Aggregate Analysis:
3 out of 5 people have HIV

What about Bob?
Suppose an adversary gets aggregate analysis, and all data records except for Bob

National ID	Name	HIV Indicator
123	John	Y
456	Jane	Y
789	Bob	
111	Betsy	N
333	Zach	N

Adversary can determine that Bob has HIV

Differential Privacy – A Solution

National ID	Name	HIV Indicator
123	John	Y
456	Jane	Y
789	Bob	Y
111	Betsy	N
333	Zach	N
…	…	…
...additional people…		
…	…	…

Before Differential Privacy

Aggregate Analysis:
600 out 1000 people have HIV

Differential Privacy Applied

Adds random noise to a returned query via a mathematical function with a specific privacy parameter, ε

After Differential Privacy

Agg. Analysis #1:
574 out of 960 have HIV (59.8%)

Agg. Analysis #2:
589 out of 980 have HIV (60.1%)

Differential Privacy ensures the above two analyses are differentially private

Advantages

- Minimizes risk of identifying individuals upon stitching together disparate pieces of information

Disadvantages

- Need to take into account the number of queries/analyses that will be conducted to ensure differential privacy is retained

Tokenization

In-force Data

National ID	Name	DOB	HIV Indicator	Smoker Indicator	BMI
123456789	John Doe	10/12/1983	Y	Y	24.5

User: Analyst

National ID	Name	DOB	HIV Indicator	Smoker Indicator	BMI
128*$ksdk	893fdz*&	90djg24	Y	Y	24.5

User: Claims

National ID	Name	DOB	HIV Indicator	Smoker Indicator	BMI
123456789	John Doe	10/12/1983	Y	Y	24.5

Advantages

- Protects the identity of the individual when it is not necessary
- Software can token/de-token as data flows through different systems and users

Disadvantages

okens by themselves are not consistent
or the same data subject coming from
ifferent data lineages

Psuedonymization

Raw Input

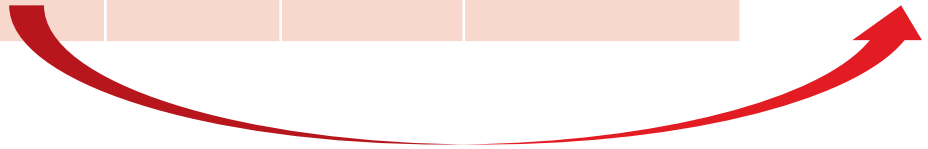
Name	DOB	HIV Indicator
John Doe	10/12/1983	Y

Name	YOB	Smoker Indicator
Johnathon Doe	1983	N

Protected Output

Psuedo	Name	DOB	HIV Indicator
123ABC			Y

Pseudo	Name	YOB	Smoker Indicator
123ABC			N



Advantages

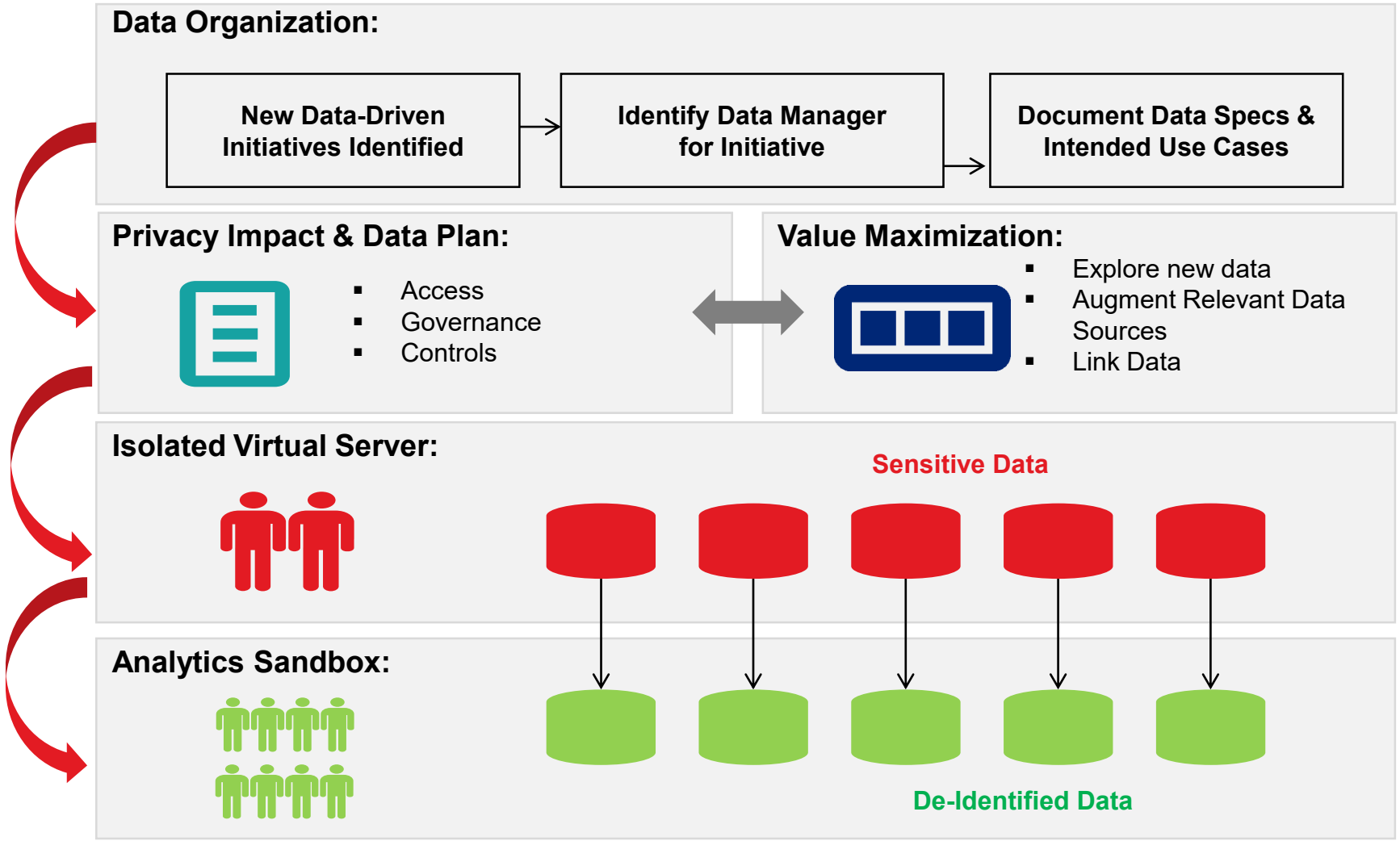
- Enables merging of de-identified datasets via common pseudonyms
- Dynamic ability to merge new sources to existing sources over time

Disadvantages

- Requires significant investment in an identify management service to link data subjects
- Requires IT and role-based separation of duties to minimize re-identification

A Structure for Focus on Data

Two key components to ensure data (i) governance/control and (ii) value maximization



In Closing…

Why does data protection matter to a data science capability?

- Prevent inappropriate use
  - Damage to reputation
  - Loss of business (current and future)
  - Substantial fines, damages, and costs

- Engender trust from data contributors
  - Keep the data flowing and therefore ‘things to do’