Target Information Architecture (TIA) – to satisfy twin task demands:

1. Supply a working blueprint [architecture] & roadmap for Health Service Analytics Innovation

2. Set requirements for data de-identification framework and standard operating procedures that operationalize that framework

PART 1
TARGET INFORMATION ARCHITECTURE

Adapted from: presentation to the Pan Canadian Enterprise Architecture Community of Practice (June 18, 2019)

July 11, 2019

Kenneth A. Moselle, PhD, R.Psych.
Director, Applied Clinical Research Unit, Island Health
Adjunct Assoc. Professor, University of Victoria
Co-Founder – Data Science Studio

Andriy V. Koval, PhD
Assistant Professor
Health Management & Informatics
University of Central Florida
Co-Founder – Data Science Studio

This presentation represents Part 1 of a set of documents concerned with two related challenges:

1. Part 1 - Target Information Architecture - derivation of products from health service data that are sufficiently statistically/methodologically robust and targeted that they at least warrant consideration as candidates for translation back to a health service system.
2. Part 2 – Data De-Identification - disclosure/access management of source health service data to those parties/team who are likely to possess the requisite combinations of clinical content domain knowledge and statistical/analytically expertise required to generate useful/usable products.

In effect, Part 1 sets out the requirements for Part 2 – the methodology covered in Part 2 must scale out to the types of datasets required to generate the products covered in Part 1.

Organization of the two presentations

PART I – Target Information Architecture for Health Service/Service System Analytics

- Why might we want to promote the use of target information architectures (TIA) in supporting a health service analytics innovation agenda?
- Example of a TIA for health service analytics

PART 2 – Data-Requirements-Informed Data De-Identification Scheme [separate presentation]

- Why would the analytics innovation agenda be concerned with data de-identification?
- Target information architecture as the basis for systematically "stress testing" a data de-identification methodology
- Distinctive privacy challenges associated with transactional data extracted from clinical information systems
- Data disclosure privacy risk model that scales out to high-dimensional health datasets (e.g., datasets extracted from clinical information systems)
- Data de-identification workflow – high-level
- Critical role of shared understanding and consensus around data de-identification – a 'fractal' data de-identification model.

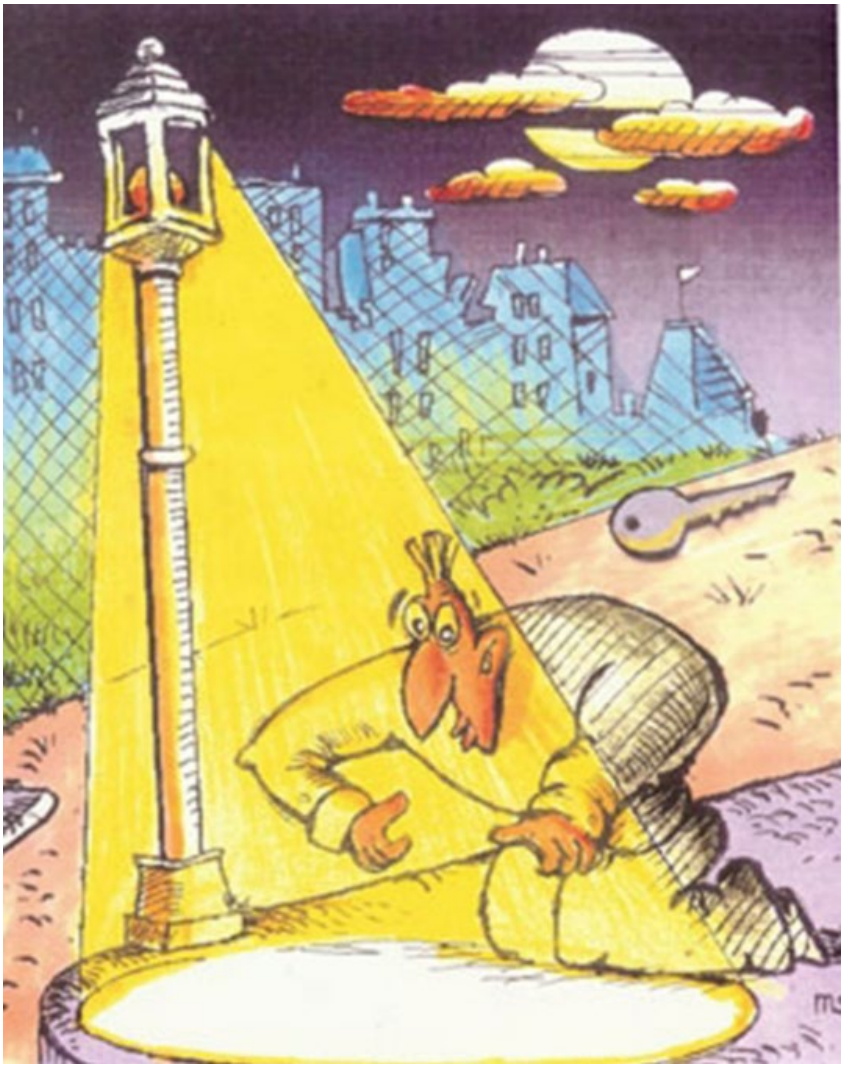
Part I
Target Information Architecture:

Why?

Why employ an information architectural approach to analytics innovation?

- So the necessary pieces (a) exist; and (b) fit together (into conceptual models; into statistical models).
- Relevance and impact - so the assembled analytically-derived 'objects' are fit for purpose and fit for context (i.e., ***targeting***)
- Analytical 'orphans' – e.g., process metrics (causes) not related to outcomes (effects); or effects not related to causes = diminished utility.
- Analytical gaps, e.g., essential risk-adjustments missing from models → ambiguous relevance.
- "Jumping to metrics" vs building to architectures – merits/demerits of different approachs
- Information dependencies dictating sequencing for analytics innovations
- Pushing off difficult-to-construct analytic entities to a time when we are not so busy – when might that be?
- Provide a reference model for analytics innovation strategy and tactics and plans – and resources and environments and partnerships.
- Provide a wire-frame within which data sources, information products and information-dependent functions can be catalogued and tracked.

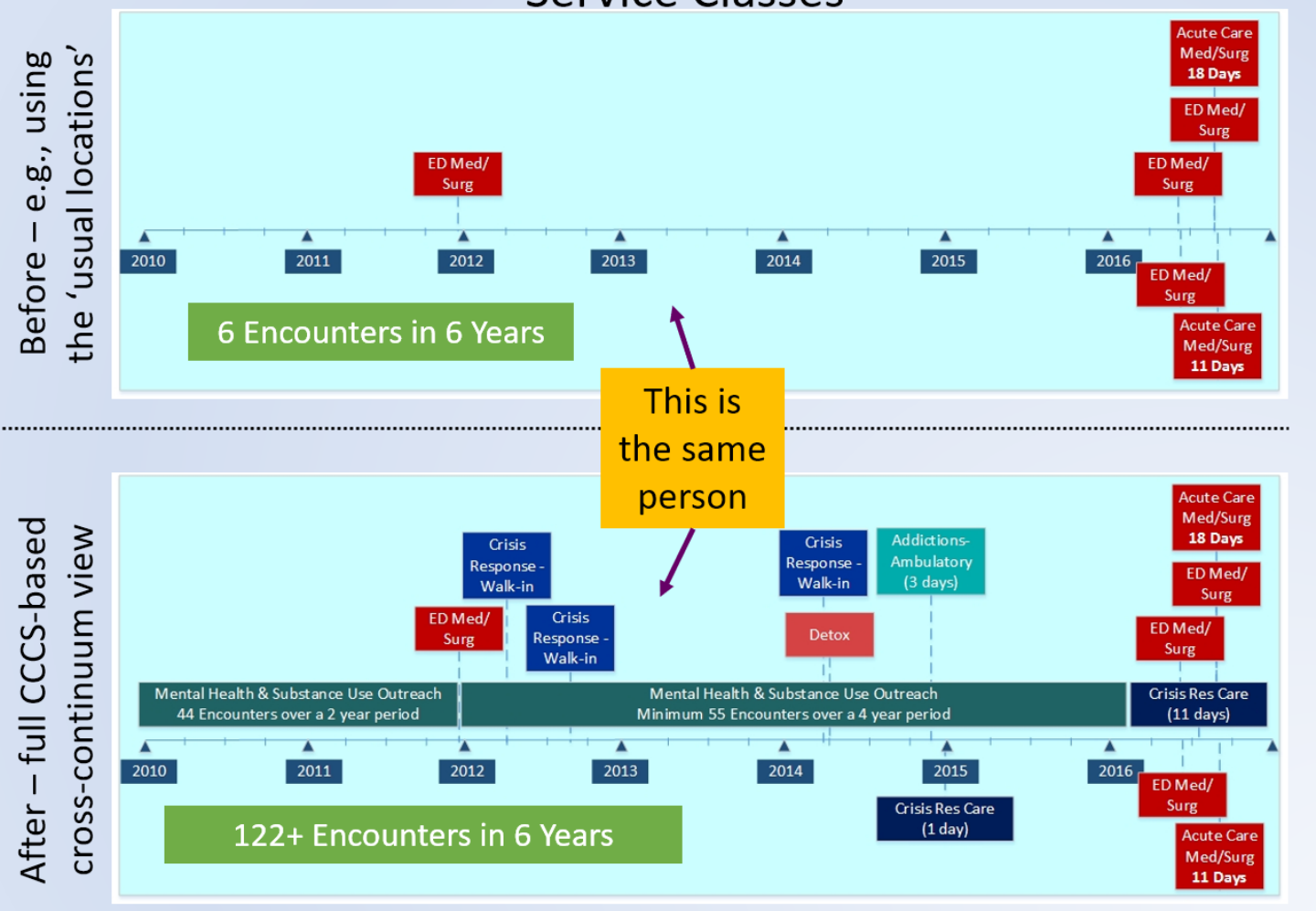
Illustrative example – working with data contents we have vs what we need



The **streetlight effect**, or the **drunkard's search** principle, is a type of observational bias that occurs when people only search for something where it is easiest to look.
https://en.wikipedia.org/wiki/Streetlight_effect

Measurement based on what is readily accessible vs measuring based on a working model that describes essential features of what you are trying to measure

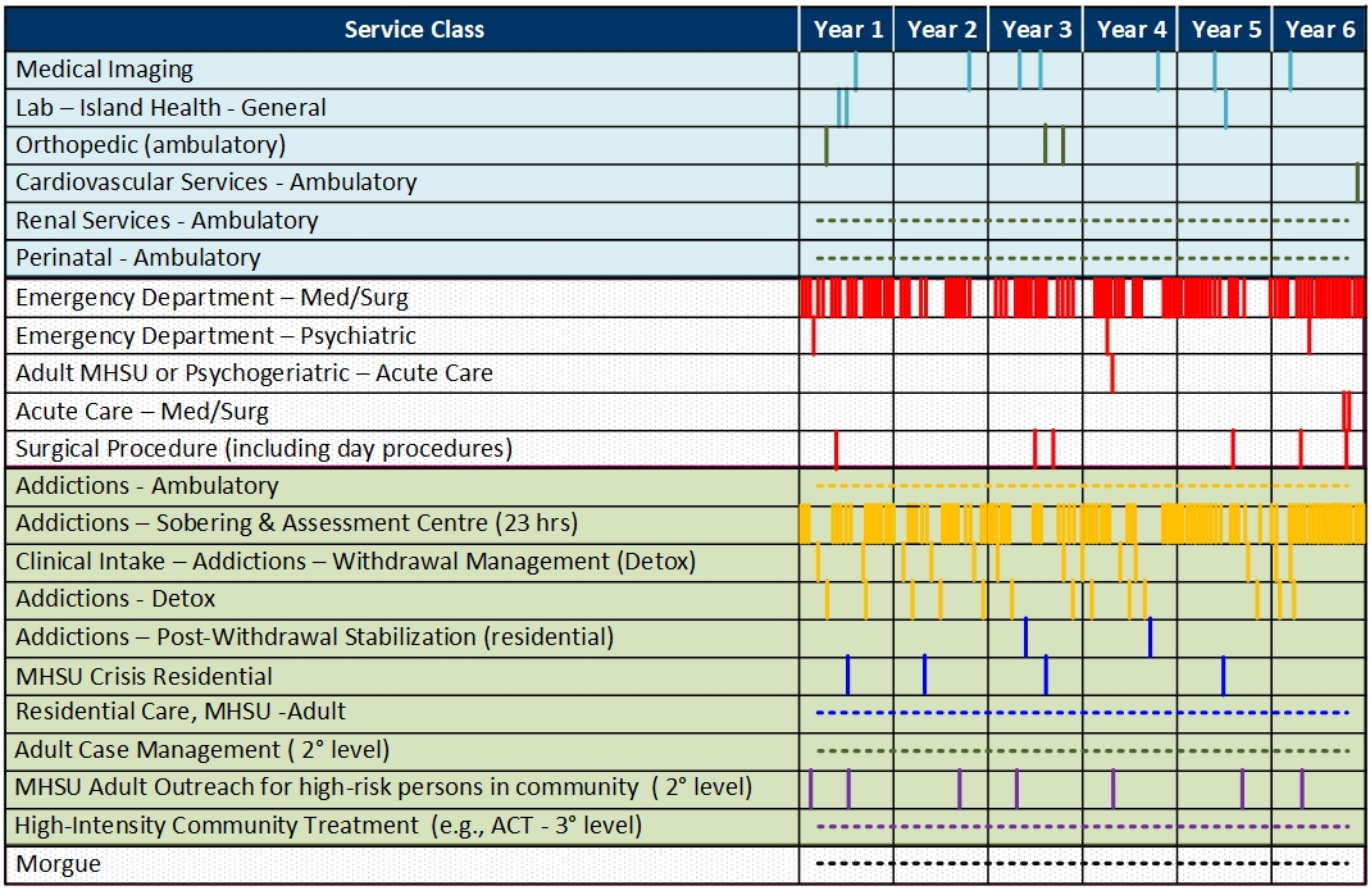
Working solution: Clinical Context Coding Scheme (CCCS) – transforming 1700+ Unit Names into 153 Meaningful Service Classes

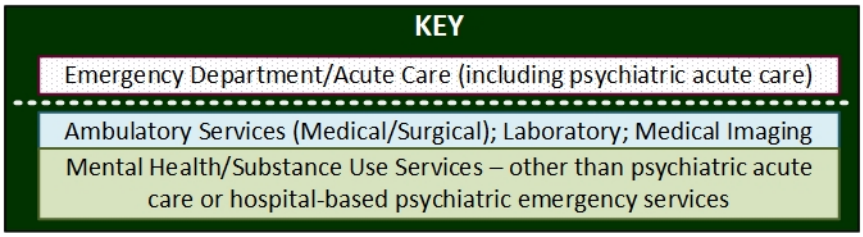


Prediction models based on the upper figure would be incorrect; estimates of demand for services to meet population need would not reflect the profiles of at-risk or affected populations.

Service Terrain Navigated by a Prototypical High Risk/High Needs Person Contending with Chronic/Recurring Mental Health & Substance Use Issues

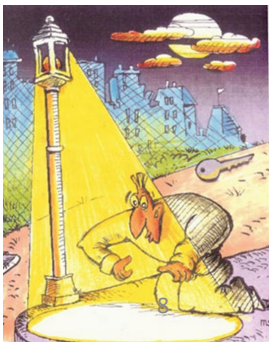
Within-person-over-time visualization of a single patient/client "journey"





Goal: generate useful information products from datasets that reflect the full "patient" journey.

Approach: We may ***initiate*** this analytical work with contents that are readily available (under the streetlamp) from most provider systems (e.g,. ED-plus-Acute-Care data). To meet the goal, our ***target*** information architecture must understand the full "journey". It must be built around the full suite of ***data 'traces'*** that are created as the person navigates the terrain.

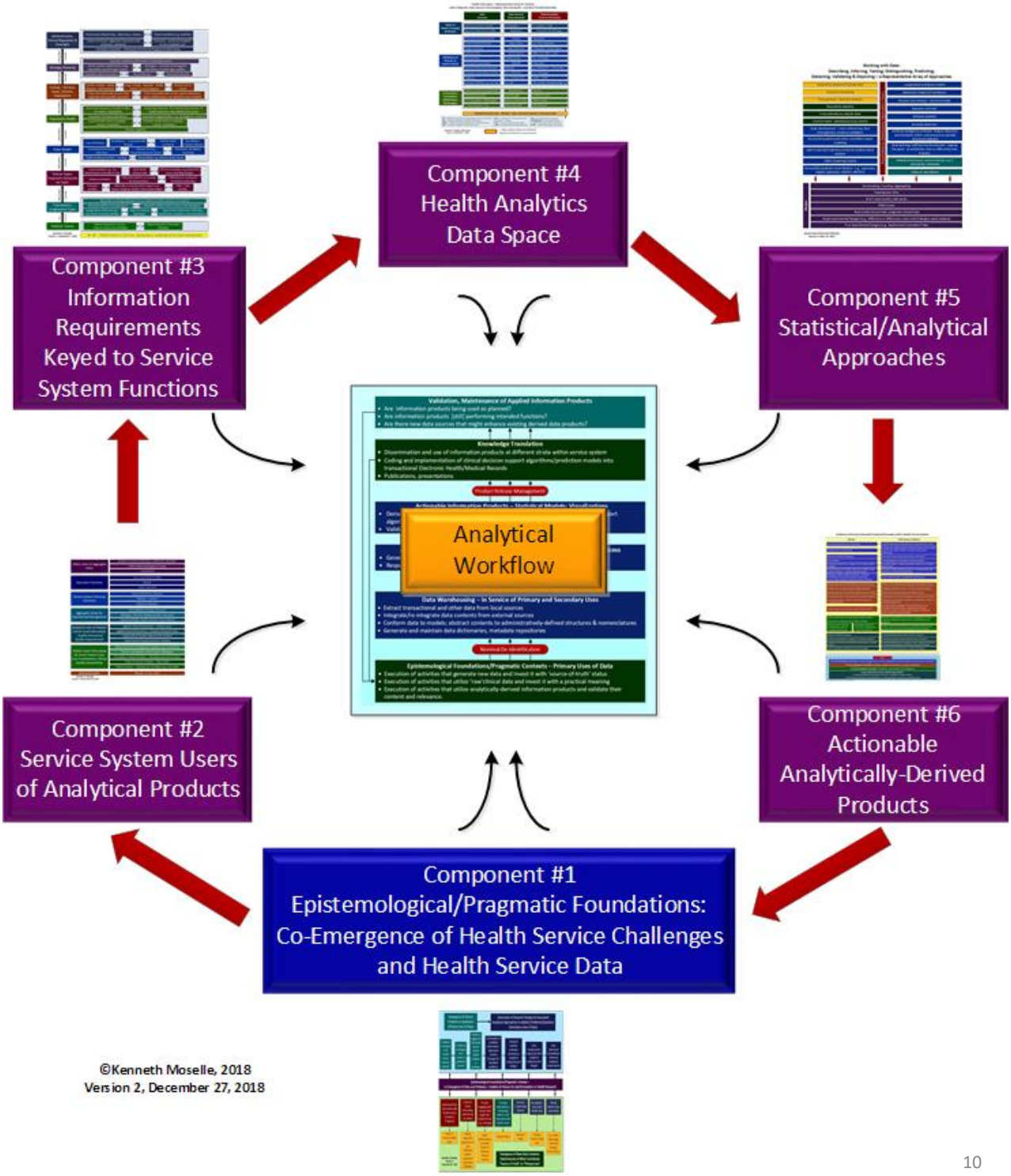


Target Information Architecture

a working example

Target Information Architecture

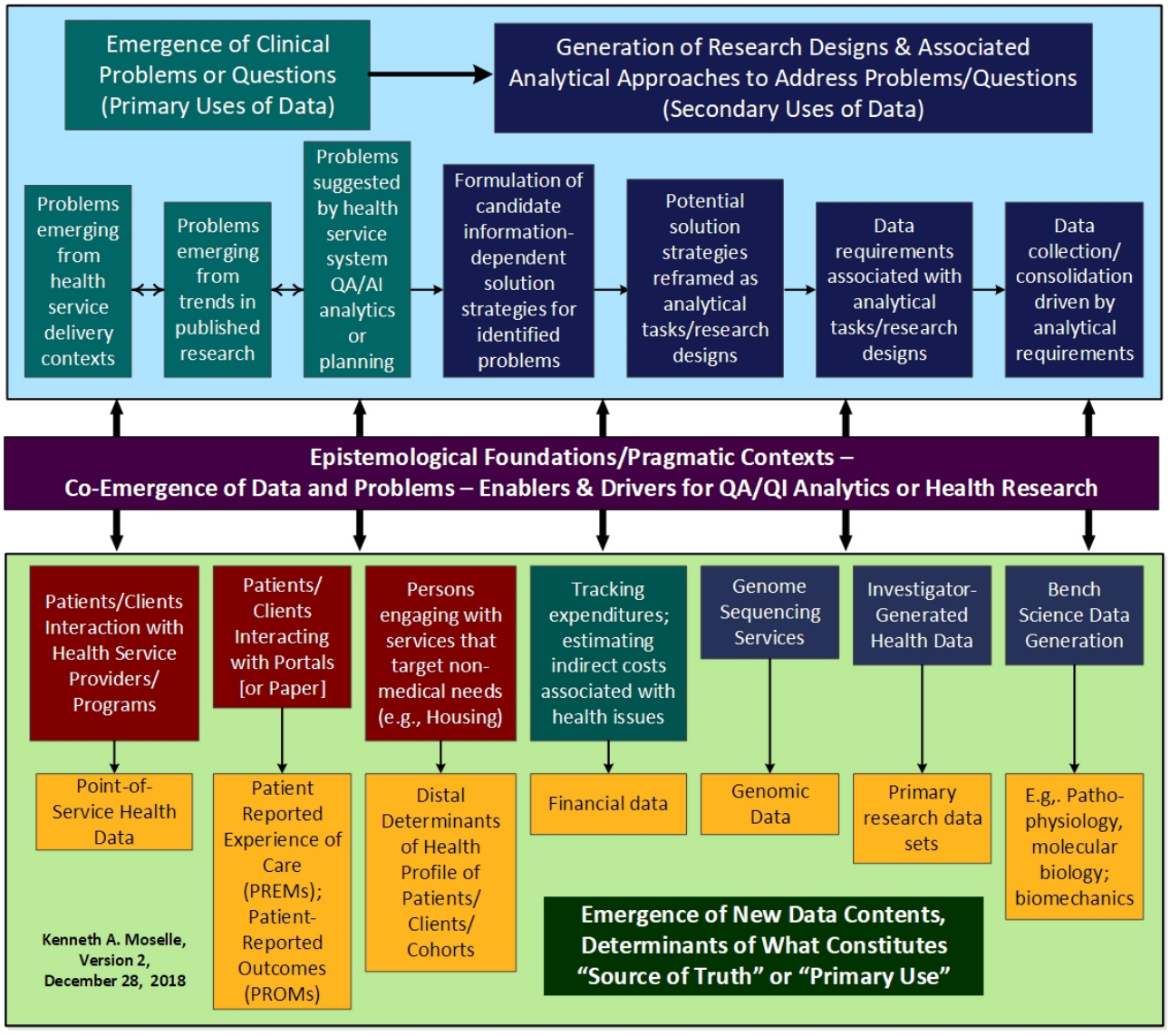
Enterprise Target Information Architecture for Health Service System Quality Assurance/Quality Improvement and Research-Based Analytics Innovation – Overarching View



Version 2, December 27, 2018

Component #1 - Epistemological Foundations – where does analytically useful health information originate?

- Epistemology – concerned with sources/emergence of knowledge.
- Where does knowledge of clinical/health risk, need and outcome originate?
- If we want our analytically-derived information products to interact constructively with processes at points of service – where MUST at least some of the knowledge originate??
- If we want to address issues using information, where must we target our analytically-derived products? And, what form should those products take??

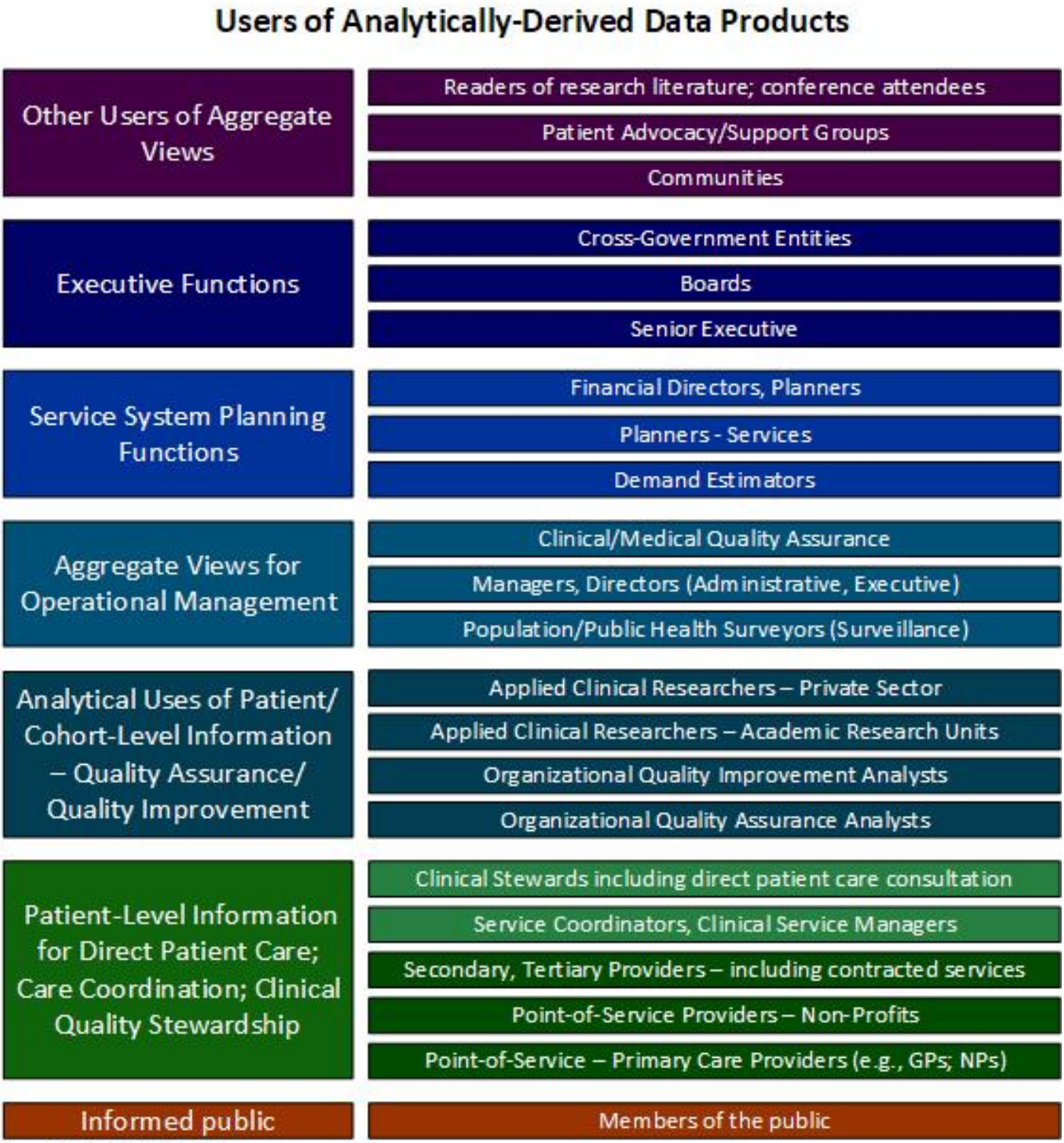


"Out of nothing shall not come something" – words allegedly spoken by Heinz Werner (Werner & Kaplan – ***Symbol Formation,*** 1984)

Highlighting data contents/deliverables within the architecture based on three key data sources and consumers of analytical products – community-derived (including primary care); health-authority-derived; Ministry of Health derived

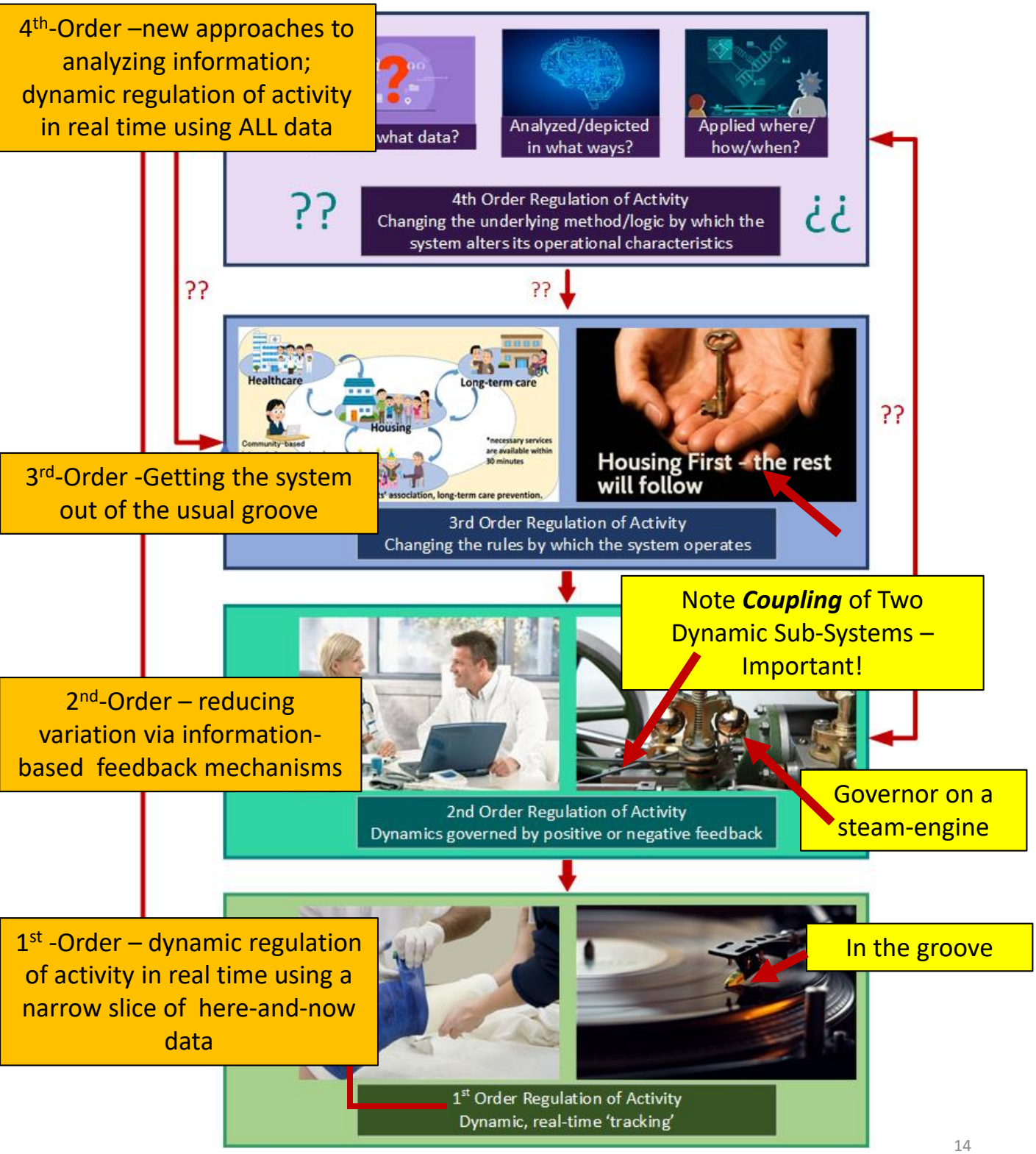
- Three key data sources and information consumers:
  - Community services - including primary care, and data generated directly by patients/clients
  - Health Authority – secondary, tertiary services
  - Ministry of Health – administrative data, with norm-references (e.g., Expected Length of Stay)
- In this section, some key components of the TIA are presented twice.
  - The first presentation of each component is intended to highlight the ***architecture*** of the entity in question. (e.g., slide #12 – "Service System Users of Information Products"). It also catalogues key contents associated with architectural elements.
  - The second presentation of a component is intended to highlight features of the component that relate to the three key data sources and consumers of information (community services; Health Authority; Ministry of Health) – colour-coded as indicated above.

Component #2 - Service System Users of Information Products

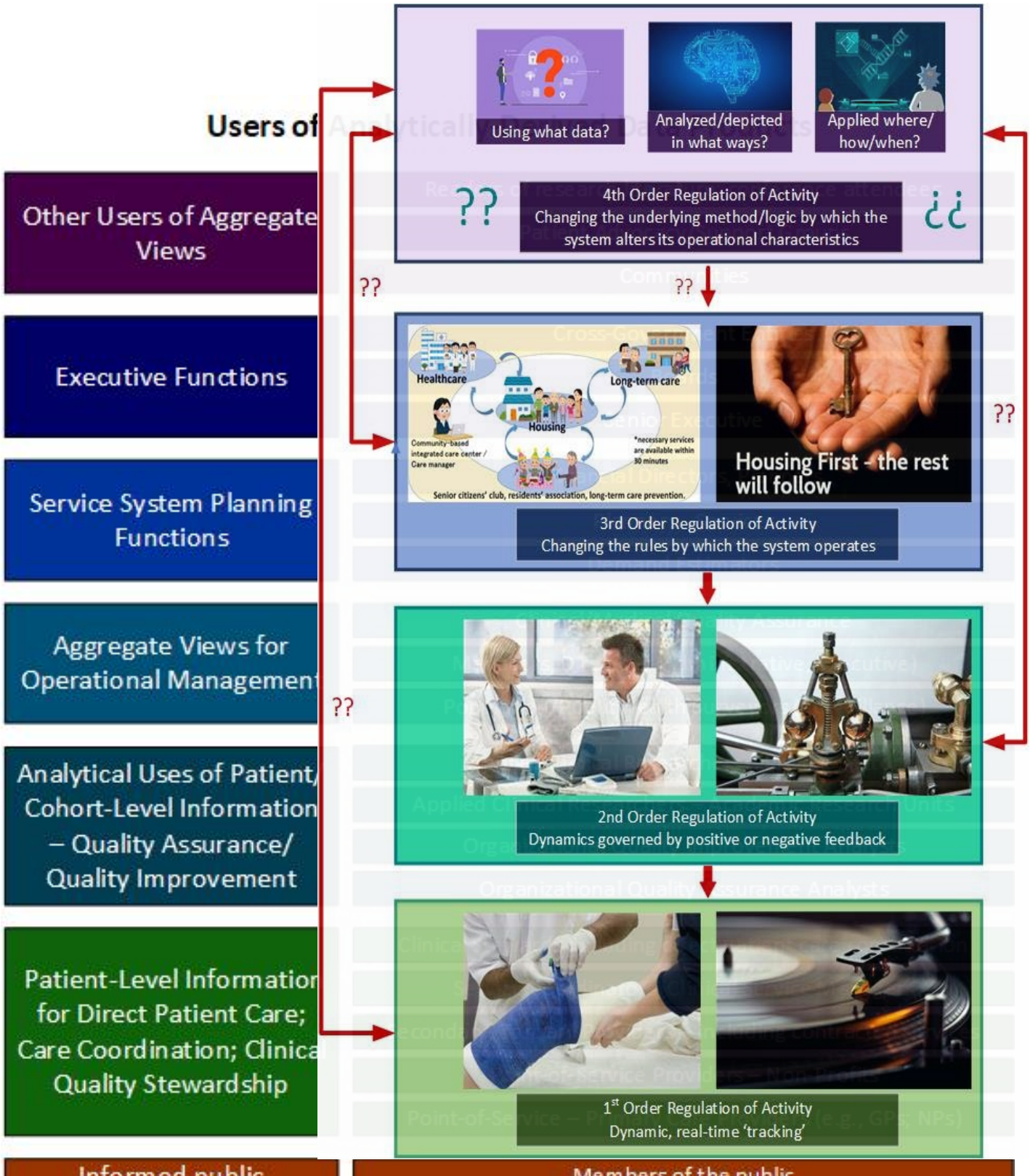


Kenneth A. Moselle,
Version 2, December 28, 2018

Deeper structure to Component #2 – a basic General Systems Theory framework



Service Systems – 1st, 2nd, 3rd Order Users of Information Products



Kenneth A. Moselle,
Version 2, December 28, 2018

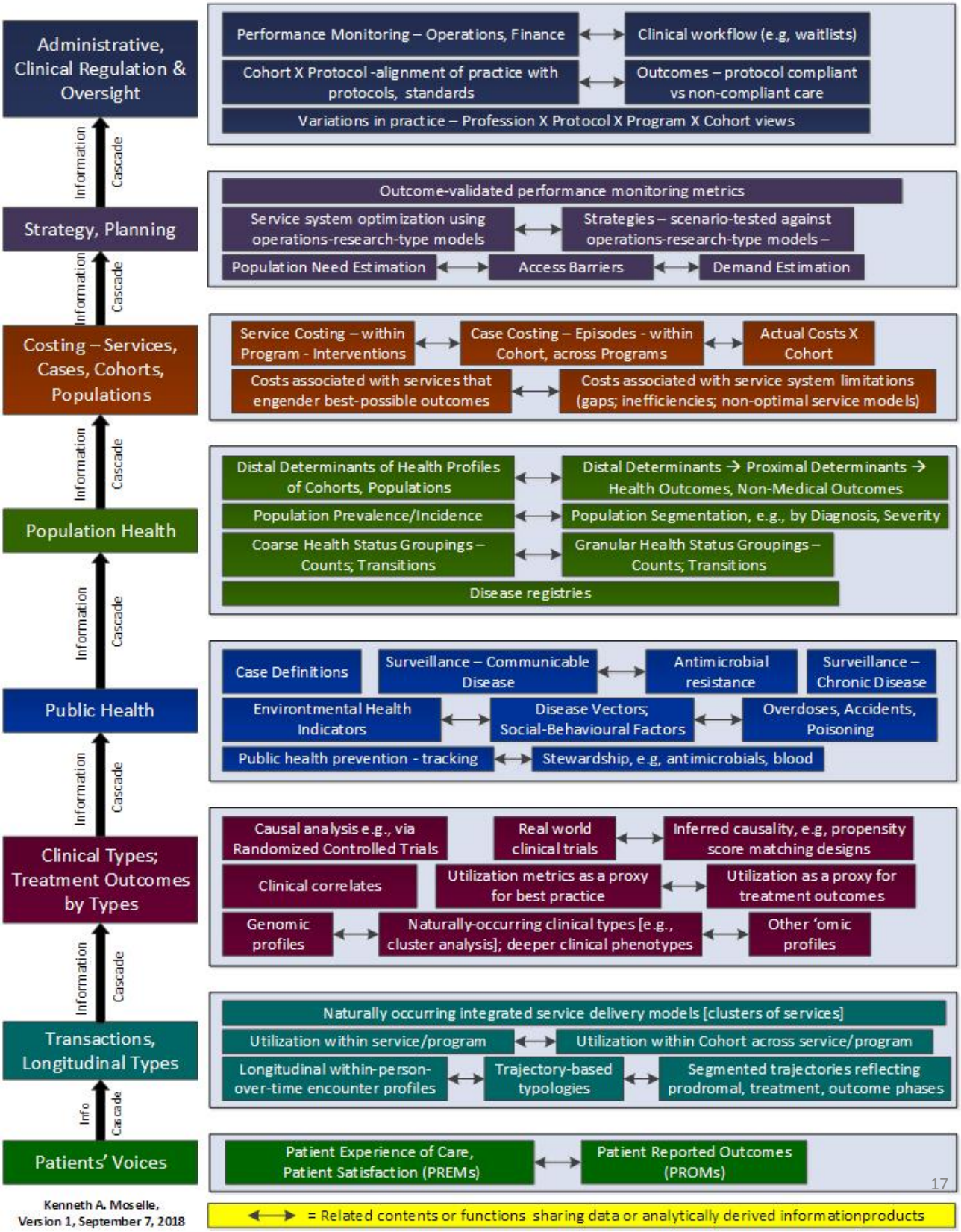
Service System Users of Information Products

Community

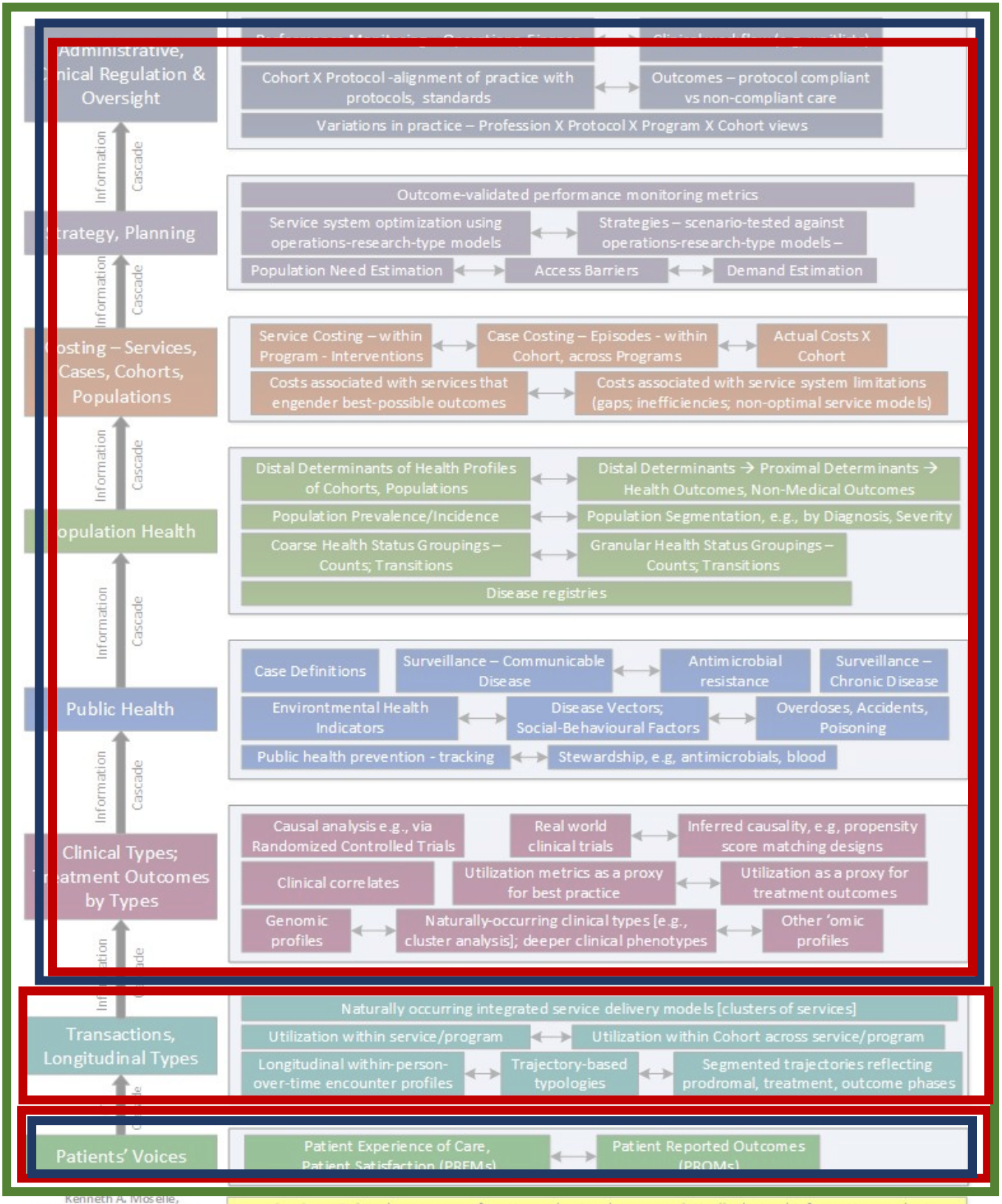
Health Authority

Ministry of Health

Component #3 - Information Products Positioned within a layered Health Service System – which information-dependent functions REQUIRE which products/analytical tools?



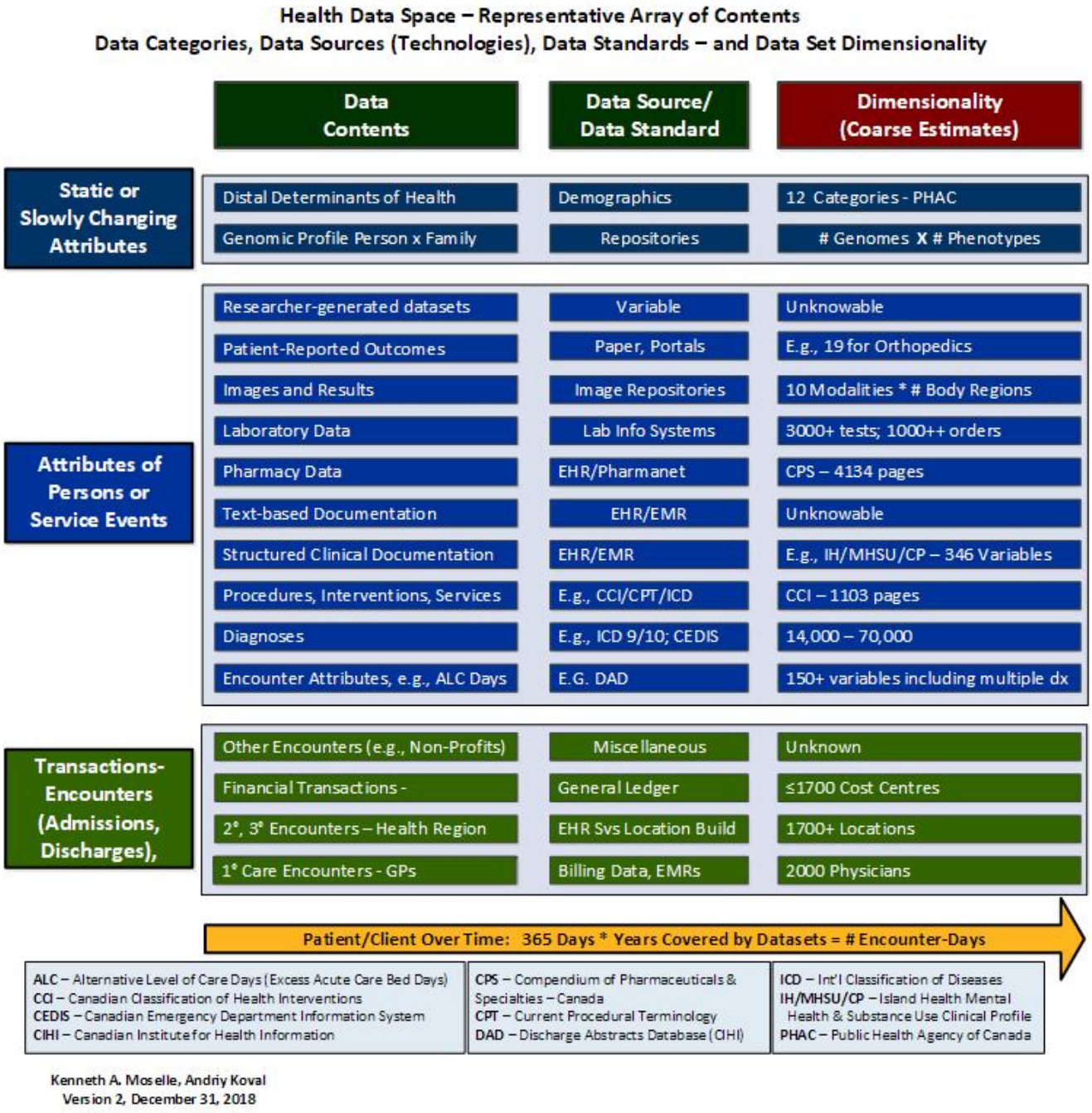
Information Products/Tools Positioned within a Layered Health Service System



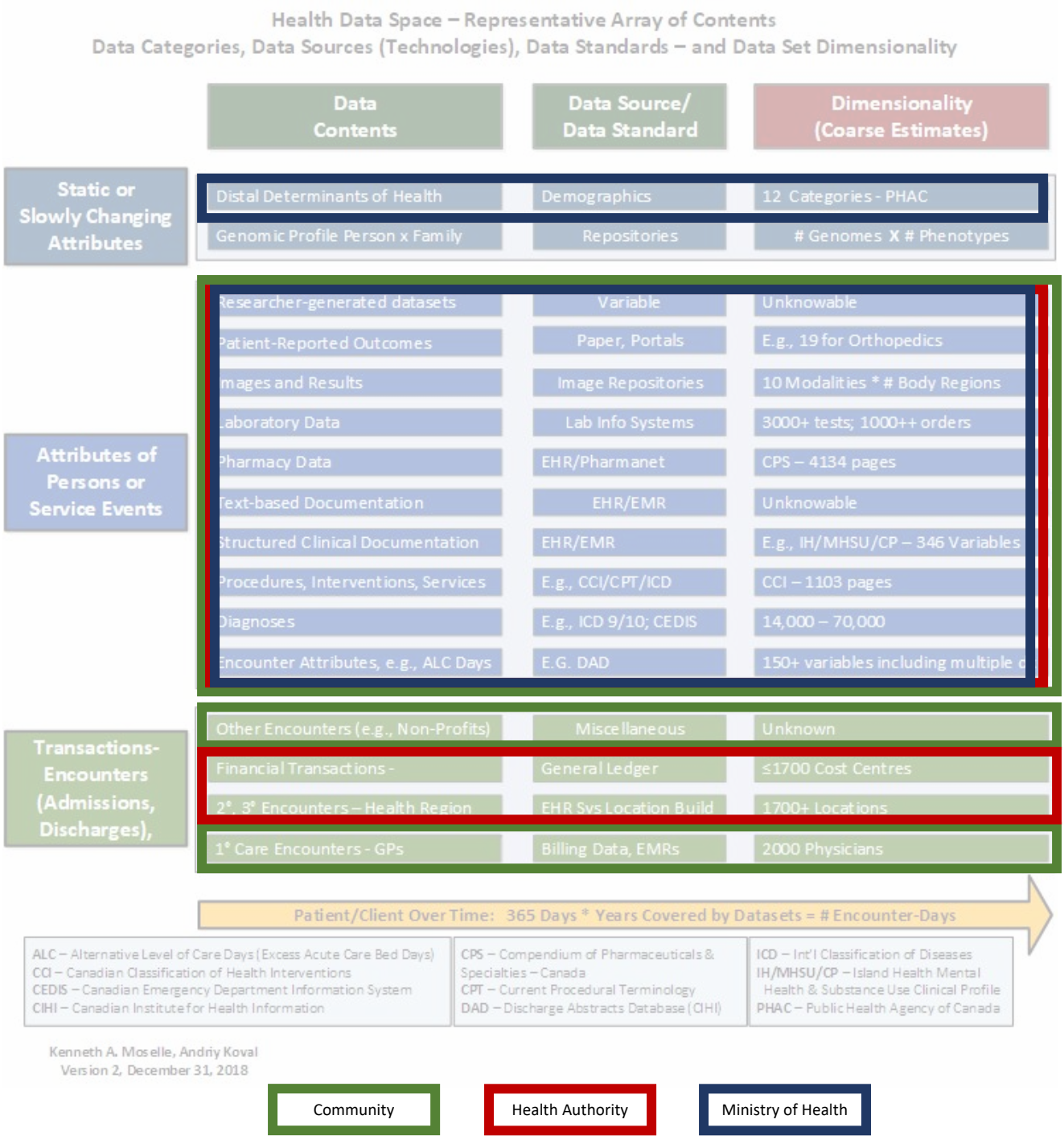
Kenneth A. Moselle,
Version 1, September 7, 2018

Community | Health Authority | Ministry of Health

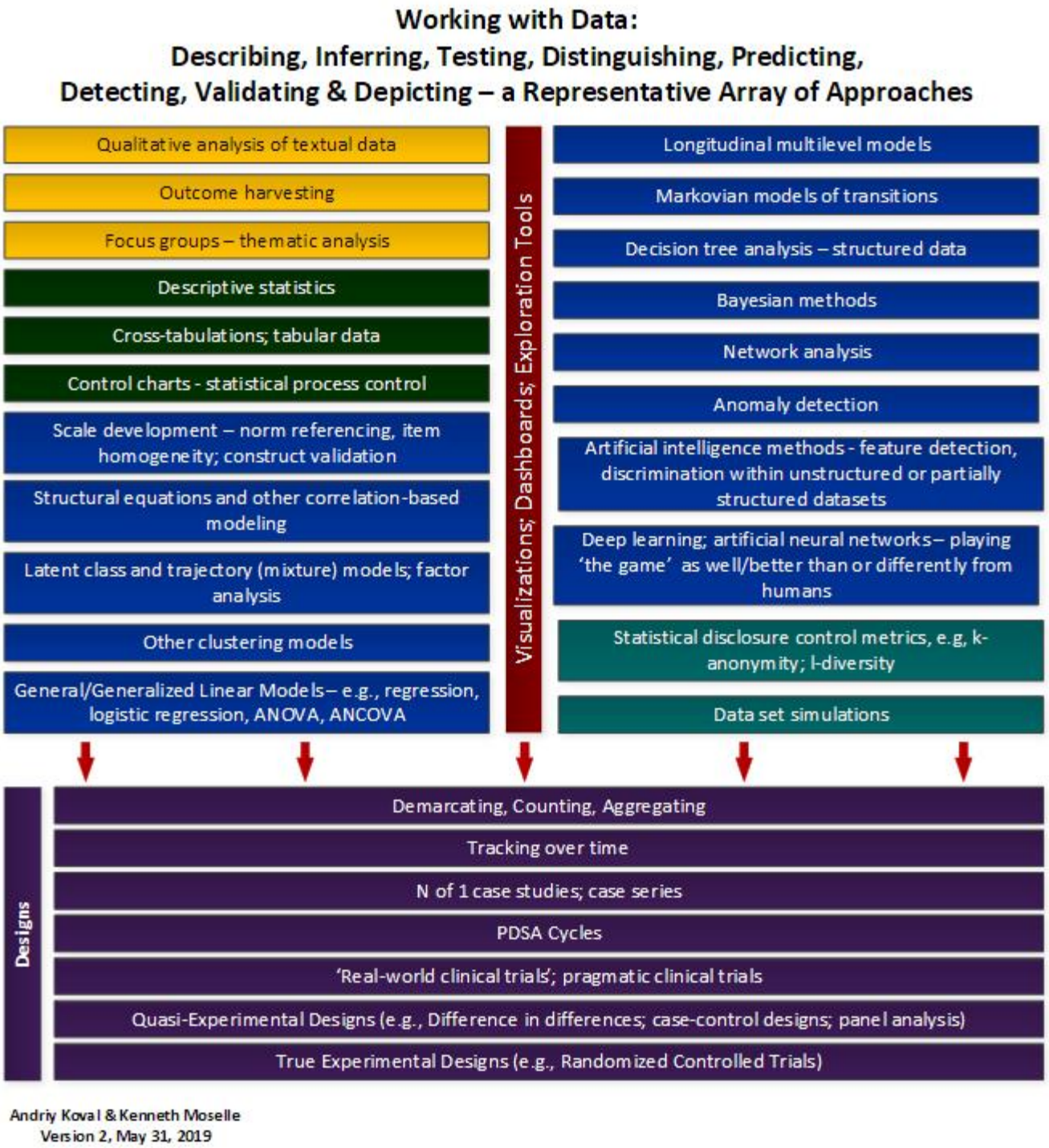
Component #4 - What data?



What data – broken out by community/health authority/ Ministry of Health



Component #5 - Statistical/Analytical Approaches



These are not clearly "owned" by any sector or strata within the full array of entities we may call the health service system – so unlike the other components, they are not marked according to "owner" or "stakeholder".

Component #6 – Actionable, Analytically-Derived Products

Activity	Information, Evidence
Executive-level performance monitoring	Useful & valid metrics linking process to outcomes & vice-versa.
Strategic Plans; Clinical Policies	Strategies pre-validated against operations-research-type models; sequences & priorities based on measured cause-effect relationships
New Service Delivery Paradigms	New overarching clinical service system 'orientations'
New programs	Clinically-targeted, functionally internally integrated
Evidentiary supports for cross-government health-relevant initiatives	Population health – incidence/prevalence linked to non-medical determinants
Proportionate service system response to changes in rates of chronic diseases – reflecting risks and impacts	Population health prevalence/incidence – chronic disease with appropriate risk/severity stratification
Population health prevalence/incidence – outbreak identification and response outcome monitoring – communicable disease; behaviourally-linked conditions	Public health functions – identification and response
Capacity planning at a clustered program level for cohorts	Patterns of svs utilization that engender best possible outcomes
Evaluate service system operations against standards	Validated performance metrics – for every level of svs system
Capacity planning at an individual program level for cohorts	Demand for services that have been shown to work.
Retrospective point-in-time supports for new/best practices	Conformance of practice to standards; outcomes of conformance
Clinical stewardship – real-time, persisting supports for new/best practices	Measures of conformance of practice to standards for cohorts – at provider/program level
Evidence-based clinical decision support within the EHR	Clinically-targeted 'packaged' views of EHR contents, e.g., Clinical/functional/behaviour characteristics of persons Procedures, orders, encounters with programs, professions Alerts, treatment recommendations
Clusters of Risk Factors and associated Health Problems 'Personalized' Integrated Service Delivery Models	Clinically contextualized cross-continuum packages of services targeting clinically over-determined or multiply determined problems.
Protocols, 'Best' Practices: • Do X because of A, B, C • Problem-focused care Personalized Healthcare: • Don't do X despite A,B, C • Person-focused care	Evidence-based best practices for conditions/problems and for people in the 'real world' contending with those problems
Measure treatment response	Clinically/operationally relevant outcome measures
Profile persons based on mediators of treatment response	Distal determinants, proximal determinants, risk factors
Collection of treatments collectively constituting an "episode of care"	Operational definitions of "episodes of care" (single service/point-in-time or longitudinal/cross-continuum), keyed to cohorts
Clinical 'targeting' of analytical derived knowledge	Diagnostic/Clinical Taxonomies – Deeper Phenotypes

KEY

3rd-Order Regulation of Activity: Unplanned/triggered or pre-planned [regularly schedule] changes to service system operations

2nd-Order Regulation of Activity: Guidelines, supports for standards, oversight, stewardship

1st-Order Regulation of Activity: Treatments, interventions, care management

Inputs, Information Drivers, Clinical Target Definers

- Causal foundations (taxonomies, metrics) – risk factors, causes, effects /
- Definitions of clinically relevant entities, e.g., case definitions, encounters, episodes of care

Moselle & Koval, verson 2, June 4, 2019

Component #6 – Actionable, Analytically-Derived Products

KEY

3rd-Order Regulation of Activity: Unplanned/triggered or pre-planned [regularly schedule] changes to service system operations

2nd-Order Regulation of Activity: Guidelines, supports for standards, oversight, stewardship

1st-Order Regulation of Activity: Treatments, interventions, care management

Inputs, Information Drivers, Clinical Target Definers

- Causal foundations (taxonomies, metrics) – risk factors, causes, effects /
- Definitions of clinically relevant entities, e.g., case definitions, encounters, episodes of care

Putting the TIA to use:

Using the TIA as a framework for characterizing and cataloguing the deliverables associated with a program of research focused on MoH Minimum Reporting Requirements (MRR) for Mental Health & Substance Use

Data Space – program of research concerned with high risk/high needs Mental Health & Substance Use Clients

Health Data Space – Representative Array of Contents
Data Categories, Data Sources (Technologies), Data Standards – and Data Set Dimensionality

	Data Contents	Data Source/ Data Standard	Dimensionality (Coarse Estimates)
Static or Slowly Changing Attributes	Distal Determinants of Health	Demographics	12 Categories - PHAC
	Genomic Profile Person x Family	Repositories	# Genomes X # Phenotypes
Attributes of Persons or Service Events	Researcher-generated datasets	Variable	Unknowable
	Patient-Reported Outcomes	Paper, Portals	E.g., 19 for Orthopedics
	Images and Results	Image Repositories	10 Modalities * # Body Regions
	Laboratory Data	Lab Info Systems	3000+ tests; 1000++ orders
	Pharmacy Data	EHR/Pharmanet	CPS – 4134 pages
	Text-based Documentation	EHR/EMR	Unknowable
	Structured Clinical Documentation	EHR/EMR	E.g., IH/MHSU/CP – 346 Variables
	Procedures, Interventions, Services	E.g., CCI/CPT/ICD	CCI – 1103 pages
	Diagnoses	E.g., ICD 9/10; CEDIS	14,000 – 70,000
	Encounter Attributes, e.g., ALC Days	E.G. DAD	150+ variables including multiple dx
Transactions-Encounters (Admissions, Discharges),	Other Encounters (e.g., Non-Profits)	Miscellaneous	Unknown
	Financial Transactions -	General Ledger	≤1700 Cost Centres
	2°, 3° Encounters – Health Region	EHR Svs Location Build	1700+ Locations
	1° Care Encounters - GPs	Billing Data, EMRs	2000 Physicians

Patient/Client Over Time: 365 Days * Years Covered by Datasets = # Encounter-Days

ALC – Alternative Level of Care Days (Excess Acute Care Bed Days)
CCI – Canadian Classification of Health Interventions
CEDIS – Canadian Emergency Department Information System
CIHI – Canadian Institute for Health Information

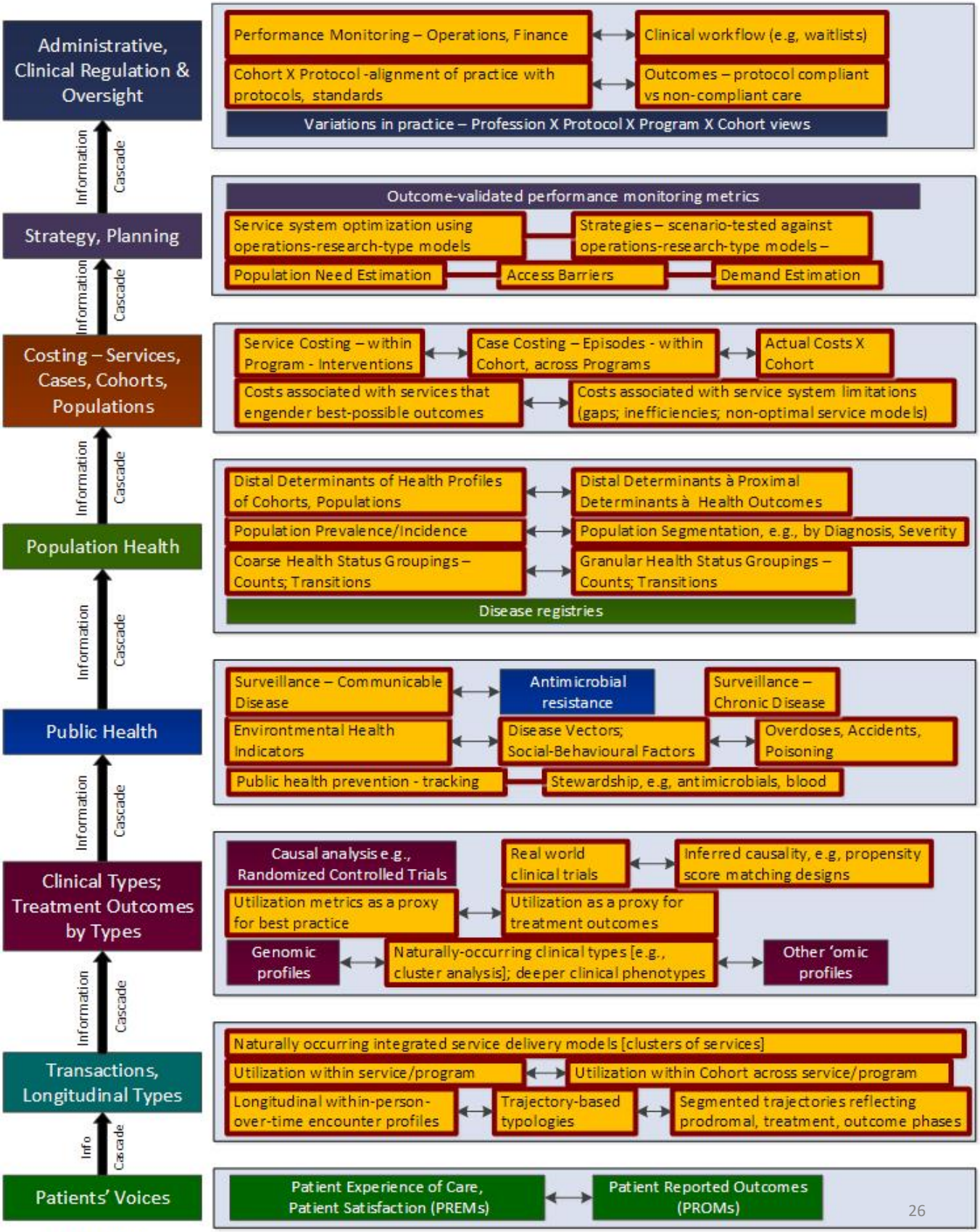
CPS – Compendium of Pharmaceuticals & Specialties – Canada
CPT – Current Procedural Terminology
DAD – Discharge Abstracts Database (CIHI)

ICD – Int'l Classification of Diseases
IH/MHSU/CP – Island Health Mental Health & Substance Use Clinical Profile
PHAC – Public Health Agency of Canada

Kenneth A. Moselle, Andriy Koval
Version 1, September 7, 2018

= Data requirements for Mental Health & Substance Use Scenario

Information Products, Information Dependencies/Cascades, Information-Dependent Activities – and Information Positioning within a Layered Health Service System



Kenneth A. Moselle,
Version 1, September 7, 2018

↔ = Related contents or functions sharing data or analytically derived information products

Discussion, Summary: Some TIA framework principles; some TIA facts of life

Where are some key limitations of the existing model

- It treats “executive level functions” in an undifferentiated way – but this ‘level’ is the core or foundational ‘business’ for a Ministry that does not deliver services directly.
- Divisions (interacting laterally) within a layer in a hierarchical structure, e.g., 1700+ programs that collectively constitute the ‘clinical business end’ of island Health. These cluster into a smaller set of entities which are not homogeneous with respect to information generated or used.
- Risk management functions and associated information requirements.
- Primary care – the model does not go very far into the primary care end of the service continuum. That is a problem!
- The model does not envision or spell out uses of derived information products by patients/clients (e.g,. via portals).

Framework, Principles

#1: General Systems Theory (GST – von Bertalanffy) – complex system = hierarchically-ordered array of interacting regulatory mechanisms – exchanging information with an environment

#2a: Law of requisite variety (see Ashby) – why information MUST roll up from the points-of-service if adequately informed executive-level system regulations are going to roll back down

#2b: Law of minimal granularity – roll up no more detail than is necessary for intelligent control to roll back down (converse of #2a)

#3: "Structural engineering" building code for load-bearing "multi-story" analytics that maps onto layered structure of an organization

- Epistemological foundations – source of clinical meaning and utility
- Data dependencies
- Structural properties of useful information
- Validity – ready for clinical prime-time – the concepts of "research-grade" and "clinical-grade" information

#4: Architectural dependencies – ***transactional*** foundations of everything; administrative data are derived entities

#5: Coupling (reciprocally-reinforcing transactions): exposure to information does not necessarily accomplish work; coupling *via* the intermediary of information puts information to work.

- Between information and a single recipient of care – one key, one lock
- Between providers, programs and cohorts – several doors, multiple keys
- Between strata within an organization – different floors, different access between levels
- Between organizations – different buildings

#6: Information 'highways' are only a part of the information road system

- Local roads vs logging roads vs superhighways – each have their place and time and function
- Roads to nowhere – untargeted analytics

Facts of Life

1. Transactional data and “administrative data” are not equivalent (even when you have a lot of longitudinal administrative data).
2. Time matters – between outcome and events that engender those outcomes or dynamics that regulate those events – why the eyes are in the head!
3. Persistence – sustained streams of information (driving requirements around reproducible paradigm, information streams, vs one-off injections of insights) – if you want your analytics to do constructive work.
4. Single cause, single treatment → simpler statistical models (though not THAT simple if we want to look at people over time)
5. Multiple causes, risk factors → multivariate models, undeniably complex
6. Unknown causes, unknown effects → iterative approaches
   - Voyage of statistical discovery – using old and new tools
   - Validation – using classic approaches
7. Therefore – partnerships between holders of data and holders of analytical expertise (and environments that enable those twains to meet).
8. This means we have to have processes that expose data (but not data subjects) to people who can analyze the data and generate useful products.

Part II

Data Disclosure Privacy Risk Model – Creating Relationships Between Health Data and Parties with Analytical Expertise (see Component #5 of Target Information Architecture)

Basis for a ‘Real World’ Contextualized Data Disclosure/Data De-Identification Methodology

Introductory Material

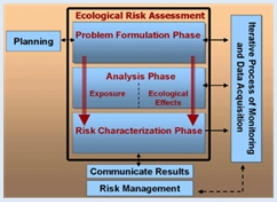
Data Disclosure Privacy Risk Model

The Four Components of Data Disclosure Privacy Risk

Components

Existential Identifiability of People and their Attributes	Mathematical Distinguishability of Cases in a Dataset	Theoretical Re-identifiability of Cases in the World	Pragmatic/ Contextualized Risk for Re-Identification of People's Data
The Natural Person(s) (the "Person", the "Population", the "Cohort") in the World →	The Data in the Dataset (the "Data") →	The Dataset in the World (the "World") →	Data and Data Users in Data Disclosure Environments (the "Context")

	A	B	C	D	E
1	ID	Var1	Var2	Var3	Var4
2	4734	1	2	3	4
3	6835	1	2	3	4
4	8266	4	3	2	1
5	4849	1	2	3	4



The Ecology of Data Re-Identification

Distinguishability of Persons in the World
- Unique Identifiers associated with the Person
- Between/beyond the Person and Unique Identifiers – knowable & distinguishing attributes of Person/Population
- Affiliation with what Communities of Interest

Distinguishing characteristics of data in the dataset
- Ignoring everything but the dataset – what is the ***mathematically quantifiable*** data re-identification worst-case scenario?
- 'Baked in' assumptions about the World.
- Data contents identify Communities of Interest.

Existence of distinguishing *plus* identifying data in the World
- Is it ***theoretically possible*** to reconnect a distinct Person in the World to a distinguished case in the Dataset?
- Is it possible to make a statistically "educated" guess about level of theoretical risk?

The Data User and Data in an Ecological Context:
- What actions are feasible?
- Is risk-actualization ***reasonbly likely***?
- What if assumptions about the person or the context are wrong?
- Differential sensitivity of different data contents – potential harm.

Quantifiables

Population prevalence
Prevalence of distinguishing characteristics – how many People share the same profile of distinguishing characteristics?

Distinguishable cases
- Range of metrics based on numbers of cases with the same profiles on multiple linked variables in Dataset.
- "Zero risk" and "disinguishable cases" do not go together.

Computed estimates of re-identifiability probabilities
- Statistical estimates of *theoretical* risk for re-identification
- Probabilities conditioned on assumptions about ***existence/linkability*** of Data in the World

Game-theoretic estimates of risk
- How many plausible scenarios can be generated where benefits outweigh the risks?
- **"Zero risk" is meaningful** in this measurement paradigm.

Caveats – 'Hard' vs 'Soft' Metrics

Coarse Estimates
- Sampling will effect estimates; 'poor' sampling yields 'poor' estimates
- Case definitions for population estimates may not match cohort definitions in dataset.
- Out-of-date estimates
- Estimates may not exist at all

'Hard' Computed Values
- 'Hard' values under a set of worst-case assumptions that may or may not be reasonable
- Greatest risk for inflated estimates of risk – greatest potential for putting data integrity at risk

'Soft' 'but seemingly 'hard' estimates of re-identifiability
- Probabilities conditioned on difficult-to-validate assumptions about Data in the World
- Quantifiables require assumptions about what Data User could possibly know.

Taking the measure of the World
- Metrics conditioned on assumptions about predictability of behaviour and robustness of technical controls.
- Risk/cost vs benefit not yet well-recognized in world of 'hard' data de-identification (statistical disclosure control)

Four components that collectively specify the risk profile of a candidate data disclosure – and provide an anchoring-point for ***operational*** definitions of key constructs (e.g., “de-identification”; “limiting disclosure”; “risk”)

- **Component #1 - People in the world** – with attributes that need to be preserved (e.g., response to treatment) in the data as disclosed, while preserving the privacy of the people associated with those attributes.
- **Component #2 – Mathematical distinguishability of cases** in the Dataset – without which there is no privacy risk associated with the disclosure – the data “”space”.
- **Component #3 – Dataset in “data space” meets data in the “real-world”** – from “distinguishability” to “theoretical re-identifiability”.
- **Component #4 – Logistical/pragmatic features of the disclosure** – how feasible and likely is it that someone will perform the actions required to transform theoretically re-identifiable contents into re-identified contents?

If only it were as simple as finding one key – anywhere!

For more information, please contact:

Kenneth A. Moselle, PhD, R.Psych.
Director, Applied Clinical Research Unit
Island Health
British Columbia
kenneth.moselle@viha.ca
250-812-3925