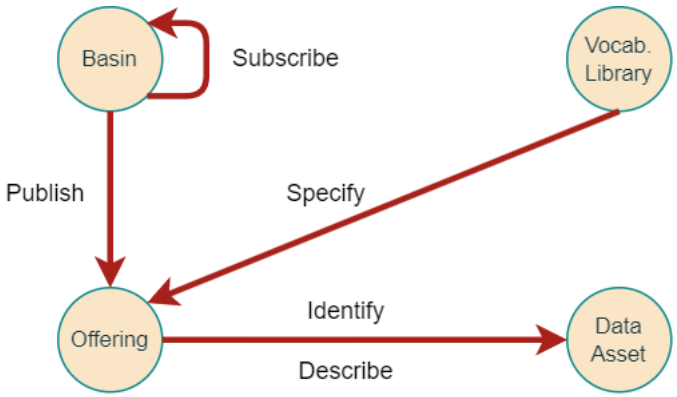
PhD Overview (Nasr Kasrin)

How can we facilitate the exchange of datasets and other data assets across heterogeneous technical and social environments? This is a classical problem that is more relevant today than ever before, relevant both on a global scale as well as the scale of organizations. We need to distinguish between two notions of data sharing and exchange: one based on the integration of data at the schema level, and the other, which is our focus here, is based on the description, publication, discovery, compilation of data as an asset (similar to how a book is managed in a library).

The state-of-the art of the data sharing and exchange problem is polarized. On one end we have emerging platforms which can be susceptible to become walled-gardens that do not interoperate with each other. On the other, we have recommendations such as FAIR guidelines, or architectural abstractions such as data lake and data catalog that help orchestrate vendors in developing interoperable technologies; taking us in the right direction but not too far, due to being too high-level and more normative/prescriptive than constructive. This work fills a gap by making a (1) 'medium-level' proposal, (2) in the form of a constructive recommendation (an architectural pattern), (3) which incorporates several major recommendations at the same time.

We propose the Basin Network, a distributed architectural pattern which revolves around two novel abstractions: the Offering (basic information unit) and the Basin (system to manage those). We adopt an indirection approach: a data asset is managed and exchanged via a surrogate (intermediate representation) which identifies, represents, describes, and effectively 'stands in' for it: this is what the Offering is for. The Basin is structure to author, catalog, and manage Offerings and exchange them with other Basins (publish/subscribe), resulting in a Basin Network (akin to a network of data lakes). See figure below for a conceptual overview.

We demonstrate the applicability of the proposal by applying it to three use-cases: (1) a computer-aided manufacturing project, (2) an IoT project in smart agriculture, and (3) crowd data management (See use-case Figures on the following page).



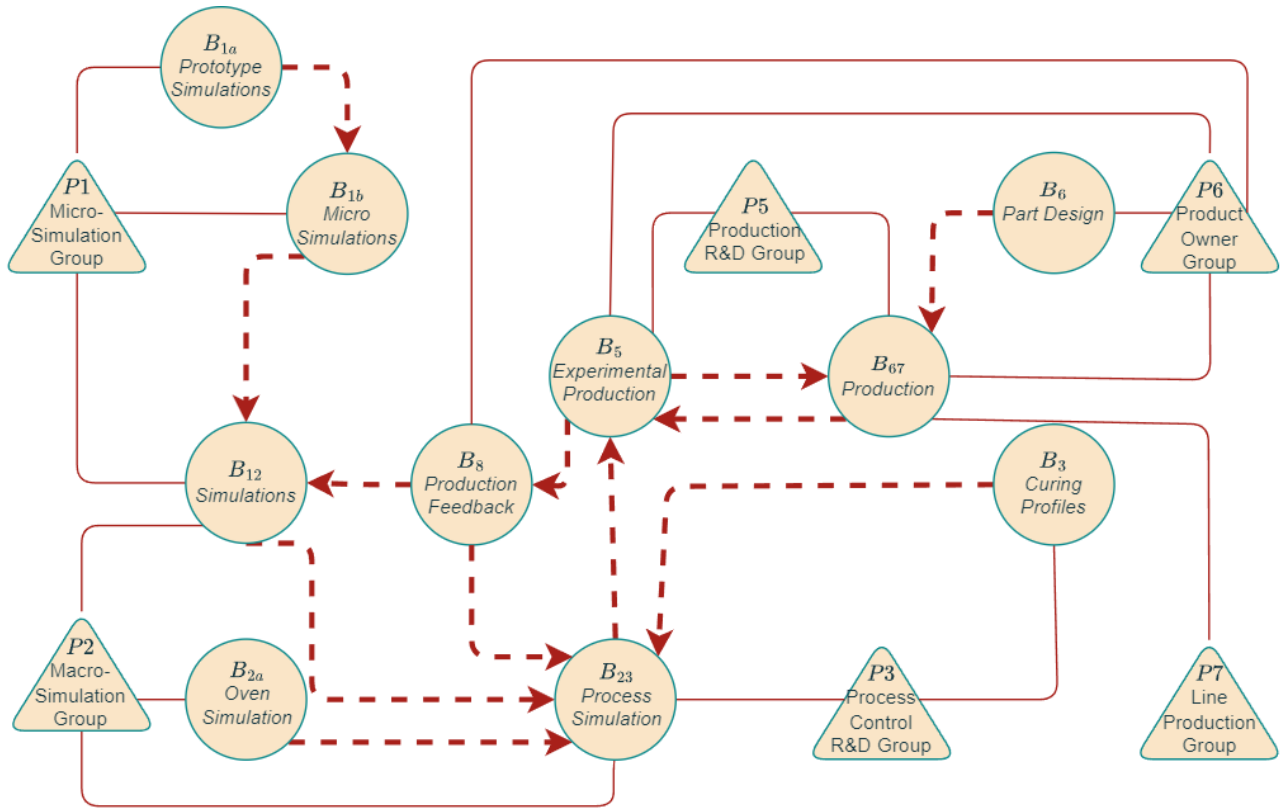


Figure 1: Manufacturing project. Triangle = team/group, circle = data space, dashed = data flow

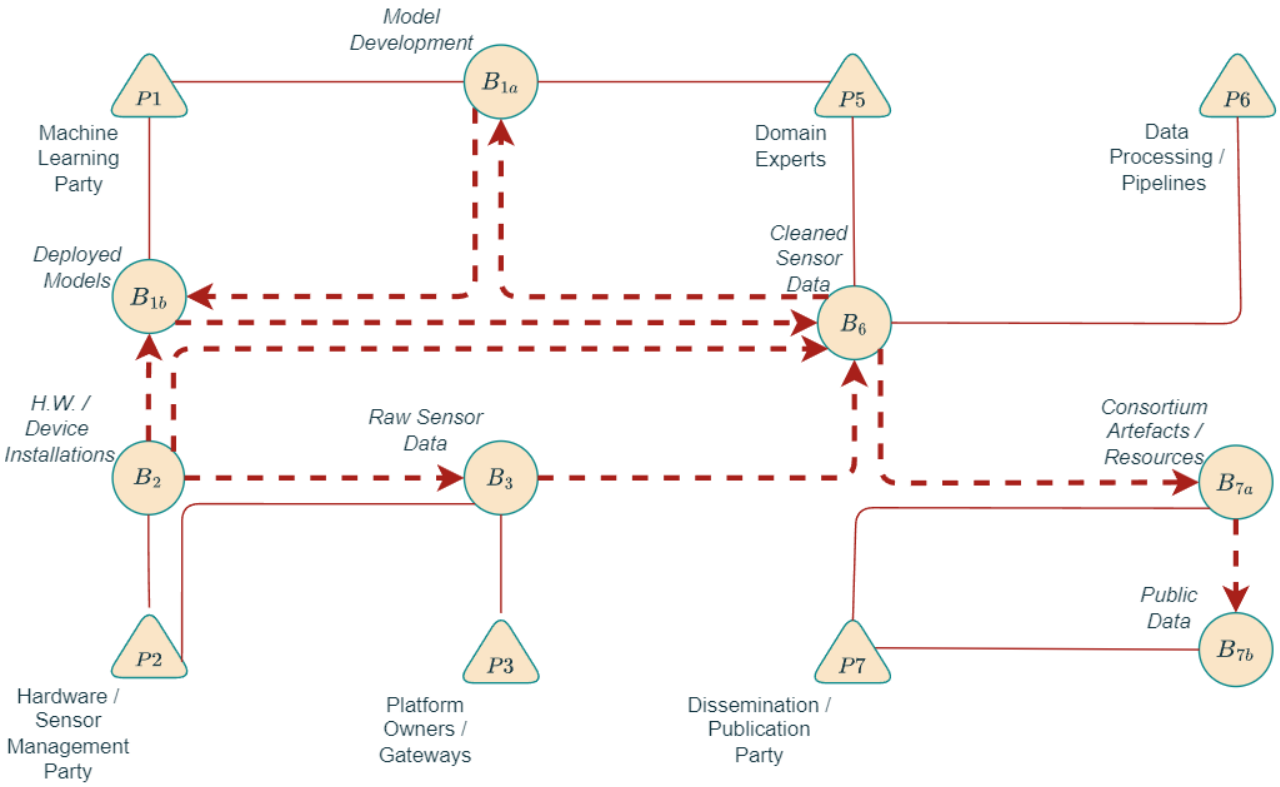


Figure 2: IoT / ML model management domain