Chapter 1 : Introduction to Data Fusion

John Klein
john-klein.github.io

Université de Lille 1 - CRIStAL UMR CNRS 9189

Chapter organization

1. Problem statement
2. Data Fusion FAQs
3. Unaddressed fusion problems

Data fusion is :

- a general problem related to information sciences and technologies.
- not really a branch of a given scientific domain.
- an active research field intersecting several scientific domains.

Data fusion :

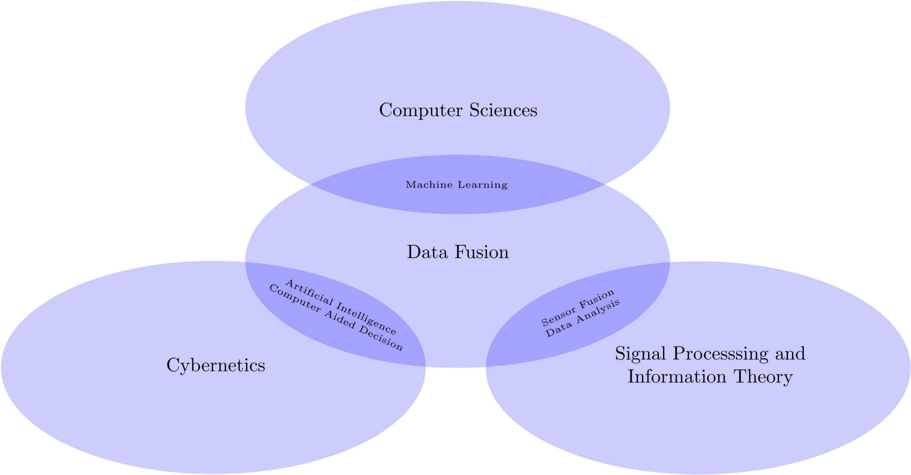


FIGURE – Data fusion related topics.

Despite the specificities of data fusion approaches, they can be somehow summarized by this diagram :.

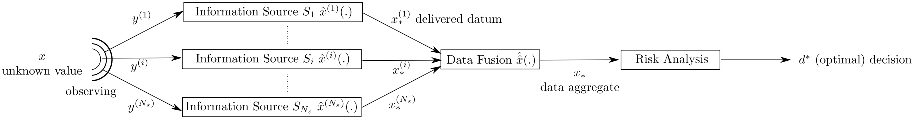

x
unknown value
observing
$y^{(1)}$
$y^{(i)}$
$y^{(N_s)}$
Information Source S_1 $\hat{x}^{(1)}(.)$
Information Source S_i $\hat{x}^{(i)}(.)$
Information Source S_{N_s} $\hat{x}^{(N_s)}(.)$
$x_*^{(1)}$ delivered datum
$x_*^{(i)}$
$x_*^{(N_s)}$
Data Fusion $\hat{\hat{x}}(.)$
x_*
data aggregate
Risk Analysis
d^* (optimal) decision

FIGURE – Data fusion process.

Let us define a little of bit of data fusion jargon :

- **Unknown value** : $x \in \mathcal{X}$
  - This is what we are trying to find.
  - The term *value* should be understood in a very wide sense, meaning that x can be the real value of a parameter, multi-dimensional, qualitative, a mix of those or a function.
  - The space $\mathcal{X}$ will be called the solution space.

Let us define a little of bit of data fusion jargon :

- **Unknown value** : Example
  suppose x is the color of an object. Then x could be alternatively :
  - the wavelength of light rays reflected by the object surface, so x would live in $\mathbb{R}^+$,
  - an RGB triplet as in computer sciences, so x would live in $[0;255]^3$,
  - the whole spectrum of light rays, so x would live in $\mathcal{L}_2$ where $\mathcal{L}_2$ is the set of square integrable functions (w.r.t. the Lebesgue measure),
  - a qualitative datum, so x would live in a given set $\{pink, white, purple, orange\}$,
  - a mix of the preceding items, so x would live in an abstract space without algebraic structure.

Let us define a little of bit of data fusion jargon :

- Information source :
  - An information source S_i is an entity witnessing a phenomenon related to the unknown value.
  - It is fed by a datum $y^{(i)}$ that is not independent from x allowing it to produce a datum $x_*^{(i)}$ containing information about the value of x.
  - Input data are meant to allow us to retrieve the unknown value x therefore they are frequently called observations.
  - Output data will be called advocacies.
  - The action of an information source is embodied by a filtering function $\hat{x}^{(i)}$ mapping observations to advocacies : $\hat{x}^{(i)}\left(y^{(i)}\right) = x_*^{(i)}$.

Let us define a little of bit of data fusion jargon :

- Information source :
    - The acquired data $\left\{y^{(i)}\right\}_{i=1}^{N_s}$ are in general members of different observation spaces $\mathcal{Y}_i$ and are subject to many imperfections. Usually these imperfections are only partially mitigated by the filtering function $\hat{x}^{(i)}$ therefore $x_*^{(i)}$ is also an imperfect datum.
    - The advocacies $\left\{x_*^{(i)}\right\}_{i=1}^{N_s}$ are also in general member of differents spaces denoted by $\mathbb{X}_i$. Each of them is a superset of the solution space : $\forall i$, $\mathcal{X} \subseteq \mathbb{X}_i$. As a consequence, all advocacies are members of $\mathbb{X} = \sup_{1 \leq i \leq N_s} \mathbb{X}_i$ which we will call advocacy space.
    - When $\mathbb{X} = \mathcal{X}$, then the filtering functions are estimators.

Let us define a little of bit of data fusion jargon :

- Information source :
  - A source can be a sensor, an intelligent system or even a human being. In case of a sensor, then the estimator is the identity function and one has $y^{(i)} = x_*^{(i)}$. When a source of information is intelligent enough to make a decision or give an opinion, it is sometimes called an *agent*.
  - N_s stands for the number of sources involved in the fusion process. In this course, we will only discuss fusion problems with a constant source number.

Let us define a little of bit of data fusion jargon :

- Data fusion step and Data aggregate :
  - the data fusion step is fed by the source advocacies $\left\{ x_*^{(i)} \right\}_{i=1}^{N_s}$.
  - The action of this step is embodied by the fusion operator $\hat{\hat{x}}$.
  - The data aggregate x_* is the output of this operator and lives in the advocacy space : $\hat{\hat{x}} \left(x_*^{(1)}, .., x_*^{(N_s)} \right) = x_* \in \mathbb{X}$.
  - The aggregate can be viewed as the final (and best) representation of our knowledge on the value of x.

Let us define a little of bit of data fusion jargon :

- Decision :
    - the decision d^* is the value that one chooses for x given the features of the aggregate x_*.
    - For instance, in a probabilistic framework, the aggregate is a probability distribution related to the actual value of x. Instead of choosing the expectation or the most probable value, one can also take into account a loss function with respect to x. In this case, the risk analysis step is most often a minimization of some expected loss.
    - Suppose we are monitoring an air space and x is a binary variable with possible values *friend* or *foe*. The cost of deciding that an aircraft is foe while it is a friend is not the same as the cost of the opposite. This should be taken into account along with the probabilities of the events $\{friend\}$ and $\{foe\}$.

Data fusion as a formal problem :

- Data fusion is a process allowing to backtrack jointly a valuable piece of information along several estimation paths.
- Let us formalize a general definition of data fusion problems :

Definition

Data fusion is a subclass of inverse problems such that there are at least two information sources ($N_s \geq 2$) producing partial solutions living in a superset of the solution space.

Chapter organization

1. Problem statement
2. Data Fusion FAQs
3. Unaddressed fusion problems

- **What are the specificities of data fusion as compared to standard estimation problems ?**
  - Inputs for a fusion step are not any features. They are partial answers (advocacies) to our problems.
  - The pieces of information contained in advocacies are frequently overlapping.
  - Usually, advocacies are not point estimates and carry along some reliability / uncertainty meta-data which are calling for specific processing.

- **Is any multi-source problem a data fusion problem ?** No
  - Suppose you want to build a self-driving car. Your goal is thus to compute a command vector $\mathbf{u}_t$ at each time step. This vector has basically three components :
    - $u_{1,t}$ is the command signal to the engine.
    - $u_{2,t}$ is the command signal to the brakes.
    - $u_{3,t}$ is the command signal to the steering wheel.

- **Is any multi-source problem a data fusion problem ?**
  - You will need a lot of sensors on your car to do achieve this goal. Let us imagine that these sensors are the following :
    - a radar to detect obstacles ahead,
    - an inertial navigation system to measure accelerations,
    - a GPS and a database to know where to go,
    - a camera and a computer vision software to read road signs.
  - There is no redundancy in the observations given by these sensors, they are focussing on different aspects of the problem. The collected data are combined to compute $\mathbf{u}_t$ but this is not a data fusion problem.

- **Is any multi-source problem a data fusion problem ?**
  - Now suppose the camera is also used to further analyze the scene, *e.g.* identify the road location. There is now a redundancy between the information given by the computer vision system with that of the GPS and the radar. The road localization is now a data fusion problem.

- **Is any multi-source problem a data fusion problem ?**
  - Something more *borderline* : Suppose one wants use an electro-cardiogram and a diaphragm myogram to check if a patient is nervous or happy. There is a redundancy in the information that these two signals carry. Both of them address the same inverse problem.
    - If I process these two signals separately using a given pair of estimators, then I will have to perform data fusion on the two estimated values. This falls in the definition of data fusion presented in a few slides ago.
    - If I extract some features in these two signals, concatenate these features, perform feature selection or dimensionality reduction and finally use a single estimator, then there is no data fusion step to perform. This falls in the definition of what is sometimes called **feature fusion**. In this course, we advocate that this not a data fusion problem but a feature selection problem.

- **In what kind of systems is a data fusion step necessary ?**
  - Data fusion steps are found in nearly all intelligent systems, especially when these systems need to be secure and robust.
  - For instance, aerospace systems always have redundant sensors, *i.e.* sensors measuring the same thing. This allows the system not only to keep being functional in case one of the sensor is faulty but also to reduce uncertainty in measures.
  - Redundant sensor fusion is a textbook case of data fusion.
  - Also, in machine learning, competition winning algorithms are often classifier ensembles which contain a data fusion step.

- **Is there a general solution to data fusion problems ?**
  - Fusion is a data driven problem !
  - There are as many fusion problems as advocacy types and configurations. The diversity in data quality is especially difficult to handle and requires tuned solutions.
  - Information delivered by each source individually can be tainted with many kinds of imperfection. Let us review them.

Data imperfections :

- Uncertainty :
  - Suppose your source is a sensor and the measure delivered by the sensor is subject to random variations.
  - The uncertain part of the datum is often called noise. This noise is generally modeled by a random variable.
  - The greater the expectation and variance of the noise, the poorer the measure.

Data imperfections :

- Imprecision :
  - A piece of information is imprecise when it involves at least two candidate values for x with no possibility to refine this information on each value individually.
  - For instance, suppose x is a real number then a piece of information like $x \in [0;1]$ is imprecise. Another example is quantized data.
  - A stronger case of imprecision is often referred to as vagueness. Suppose your information sources are human agents and one of them gave you this testimony : « this guy is about 30 years old ». What do you make of such a statement when the set of candidate values has a fuzzy frontier ?

Data imperfections :

- Ambiguity :
    - Ambiguity occurs more rarily except in human language related systems.
    - Let us for example analyze Bob Marley's song title « No woman no cry ». For someone who has never heard this song before, the title can be understood in pretty different ways :
        - If there's no woman then there is no cry.
        - No ! woman, please don't cry !
        - No, women do not cry.
    - This is a typical case of ambiguous statement.

Data imperfections :

- Untruthfulness :
  - Untruthfulness is a very difficult imperfection to circumvent because in this case the information delivered is wrong.
  - An untruthful source can lie purposely or accidentally.
  - Suppose your sources are computers in a network. If one of them has been hacked then it may lie purposely to make you take wrong decisions.

Data imperfections :

- Untruthfulness :
  - A textbook case of accidentally untruthful source are uncalibrated sensors. Indeed, such a sensor will deliver biased data.
  - Untruthfulness itself can take various forms. For example, an untruthful source can say the opposite of what it knows to be true or it can tell you something partially wrong. Suppose the actual knowledge of the source is $x \in A$ but it delivers $x \in B$. If $B = A^c$ then it tells you the opposite of what it knows. If $B \cap A \neq \emptyset$ then it tells you something partially wrong.
  - Untruthfulness can only be solved if some contextual information on the quality of the source is available. These pieces of information are called meta-data. Meta-data can be used to update the estimator accordingly. Untruthfulness is also frequently termed unreliability.

Data imperfections :

- Incompleteness :
  - Incompleteness occurs when there are missing data in the dataset that a given source was supposed to deliver.
  - Imagine your source is a camera and that its photosensitive cell array has been damaged by an overvoltage for instance. Consequently, the image delivered by the source contains black pixels corresponding to malfunctionning cells.
  - The image is thus incomplete but there is perhaps still something that you can do about it to retrieve the information you seek.

Data imperfections :

- Incompleteness :

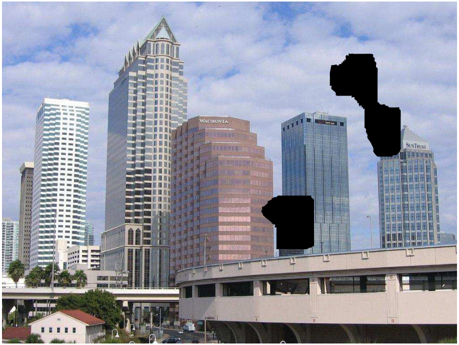


Figure – Example of incomplete image.

Data imperfections : remarks.

- The above imperfections are obviously somehow related to each others.
- For instance, incompleteness implies imprecision which itself implies uncertainty in some sense.
- The very goal of data fusion should be to allow the best representation of our knowledge about x by wiping out at least partially these imperfections. We should say exactly what we are able to deduce from the observations, no more no less.
- If one succeeds to do so, then the data fusion step is optimal. Note that the aggregate x_* may not be sufficiently precise or sure to make a decision but it is sometimes far more cautious to conclude that there is not enough data to make a decision rather than taking foolish risks.

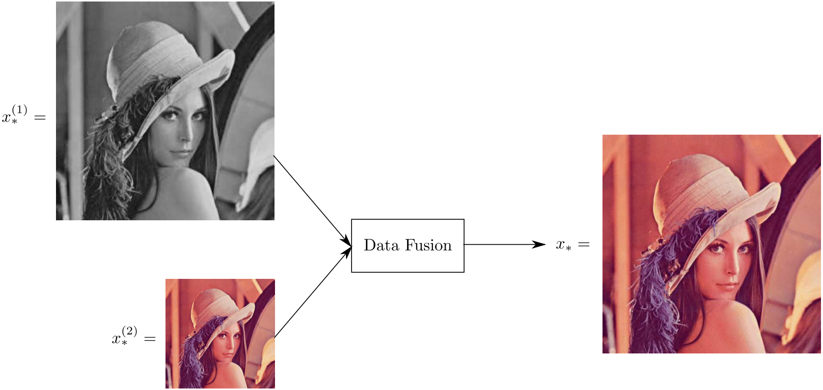
- **Is data fusion a really hard problem ?**
  - Yes, data fusion problems are in general rather hard to solve especially when the data are tainted with multiple imperfections.
  - To illustrate the challenges of data fusion, one just needs to take a group of individuals and ask them :

    Can you guess my height ?
  - Let's give it a try !

- **What concepts are central in data fusion ?**
  - conflict :
    - Suppose S_1 and S_2 are involved in the same inverse problem. Given its advocacy $x_*^{(1)}$, you are able to deduce that S_1 supports $x \in A$. Given its advocacy $x_*^{(2)}$, you are able to deduce that S_2 supports $x \in B$. Then what if we have $A \cap B = \emptyset$? Such a situation is said to be conflictual.
    - More precisely, this is a case of total conflict where S_1 and S_2 are incompatble. Conflict can also be partial like when S_1 and S_2 deliver incompatible data for x_1 and compatible data for x_2 with $\mathbf{x} = (x_1, x_2)^T$ a 2-dimensional vector as unknown value.
    - Conflict is treated in many different ways depending on the chosen formalism. Again, there is no best way to cope with conflict as this is application dependent.
    - You may also note that conflict allows you to detect cases of untruthfulness but neither does it tell you who is lying nor what to do about it.

- **What concepts are central in data fusion ?**
  - Heterogeneity in advocacies :
    When at least two sources produce advocacies such that $\mathbb{X}_1 \not\subseteq \mathbb{X}_2$ and $\mathbb{X}_2 \not\subseteq \mathbb{X}_1$ then the data fusion problem is said to be heterogeneous.
  - For instance, suppose two cameras filmed the same scene :
    - Cam n°1 produced an image with poor spatial resolution but high color definition.
    - Cam n°2 produced an image with high spatial resolution but only in grey level.
    - Our goal is to retrieve an image with high spatial resolution and color definition.
  - In this case, data is partially redundant but also complementary. This is a heterogeneous data fusion problem.

- **What concepts are central in data fusion ?**
  - Heterogeneity in advocacies :



$x_*^{(1)} =$

$x_*^{(2)} =$

Data Fusion

$x_* =$

- It mainly implies increased computation load because the fusion operator will be defined on a very large domain.

- **What concepts are central in data fusion ?**
  - Heterogeneity in information :
    - Heterogeneity in the observations are quite frequent in data fusion problems.
    - It is an advantage because such observations are usually independent and complementary.
    - Going back to the self-driving car example, suppose we use both a camera and radar to detect obstacles on the road. This a non-heterogeneous data fusion problem with heterogeneous observations. The camera produces an image while the radar produces a scalar distance measure.

The following types of fusion approaches will not be studied :

- conditional fusion : it is possible to design several fusion steps and select one of them depending on justified conditions. For instance, meta-data can be used to implement such conditional tests.

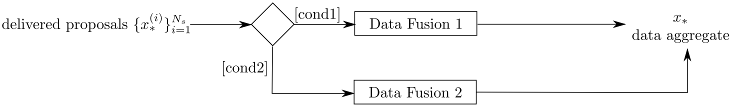


Figure – Flowchart of conditional fusion.

The following types of fusion approaches will not be studied :

- iterated fusion : some fusion techniques allow to compute iteratively the estimate $\hat{x}$ until some quality criterion is reached. Such a criterion is often related to the imprecision of the estimate or to its uncertainty. The improved estimate is usually obtained thanks to additionnal (meta-)data.

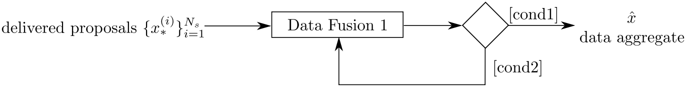


FIGURE – Flowchart of iterated fusion.

The following types of fusion approaches will not be studied :

- hierarchical fusion : the set of advocacies is dispatched into two subsets Z_1 and Z_2 such that $Z_1 \cup Z_2 = \left\{ x_*^{(i)} \right\}_{i=1}^{N_s}$. However, Z_1 and Z_2 may have non-empty intersections. Each set are processed separately by two different fusion steps which consitute the first fusion level. Two data aggregates $x_*^{(Z_1)}$ and $x_*^{(Z_2)}$ are thus given by the first fusion level. Afterward, these two aggregares are themselves merged by a third fusion step which constitutes the second fusion level. Finally, a unique estimate x_* is obtained.

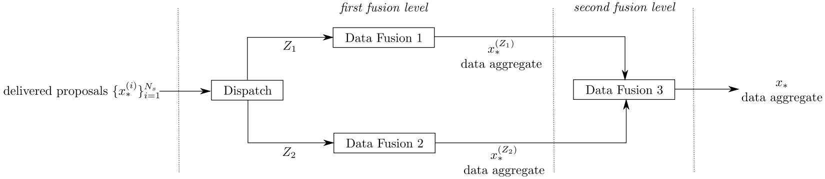


FIGURE – Flowchart of iterated fusion.

The following types of fusion approaches will not be studied :

- Sophisticated fusion systems that can be conditional, iterative and hiearchical at the same time.
- Quality criteria for data fusion : many such criteria can be defined and formalized mathematically in order to garantee that the fusion behaves in given way with respect to information sources, unknown value estimations or meta-data.