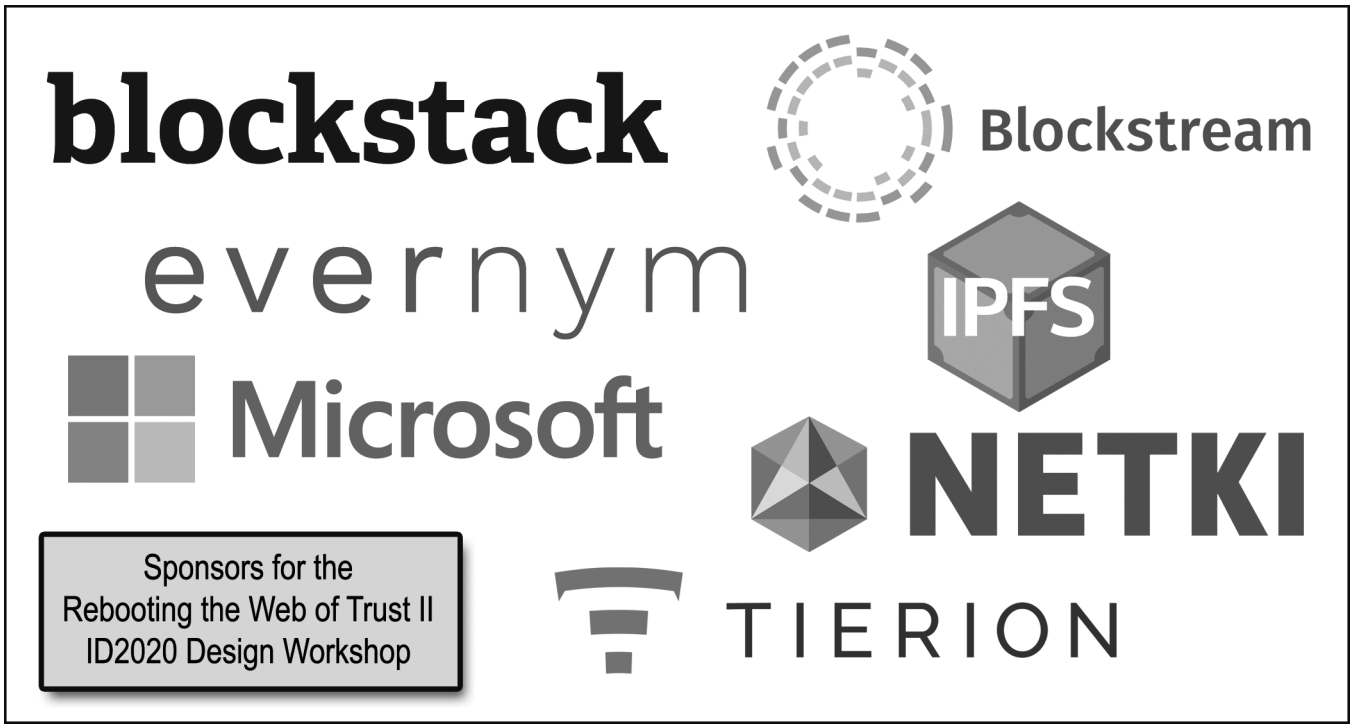
Identity Crisis: Clearer Identity through Correlation

A White Paper catalyzed by the Web of Trust II: ID2020 Design Workshop

By Joe Andrieu, Editor <*joe@joeandrieu.com*>, Kevin Gannon <*kevin.gannon@gmail.com*>, Igor Kruiper <*ikruiper@ymail.com*>, Ajit Tripathi <*tripathi.ajit@gmail.com*>, and Gary Zimmerman <*gzimmerman1024@gmail.com*>



1. INTRODUCTION

The term "identity" is a challenge.

Both laypeople and experts struggle to communicate clearly about it. The term has numerous rich and useful meanings. That same flexibility and expressivity also makes it easy to misunderstand subtle nuances and often leads to ideological debate rather than understanding and applications. We compensate with adjectives, creating new phrases like "digital identity" or "legal identity", but we often still speak past each other. We regularly refer to "identities" as things that are assigned to us or that we own, things we control or present, instead of using more rigorous terms such as "identifiers" or "credentials". This fluidity often confuses because, at its core, identity is an emergent phenomenon that doesn't have an existence independent of the observer.

We propose using "correlation" instead of "identity" when discussing concrete identities in identity systems. It isn't a word-for-word replacement, but using it *will* improve the conversation. We argue that "correlation" provides a more concise and clear understanding of how identity is created and applied in both digital and real-world systems, and that using it as an alternative to "identity" will improve communication and understanding.

For the length of this paper, we ask you to consider, "What if there were a different way to talk about *Identity*?" Take a moment to step outside your current use of that term and look through a different lens. Allow a different take on a well-worn term to illuminate your work in a new light.

2. IDENTITY FUNDAMENTALLY DEPENDS ON CORRELATION

Without an observer to recognize a subject, identity doesn't exist.

In simpler terms, any notion of identity is not particularly useful without the existence of a person or entity performing identification.

Using our alternative lens, we challenge the appropriateness of focusing on an "identity" as a property of a thing (or person), rather than as a phenomenon that emerges between an observer and a subject. We think that using the word "identity" as a concrete, own-able, controllable asset obfuscates more than it communicates.

We have personally experienced thousands of hours of discussion, debate, and disagreement about just what "identity" means. As identity professionals, we understand the need to clarify the lexicon. It's important. Unfortunately, in every new community that works toward a common understanding of the term, we see the same conversations repeated with different highlights and different influences.

Even with these potentially confusing uses of "identity", all of the varied understandings of the word depend on correlation. Consider three excerpts from dictionary definitions of "identity".

First, from the **Collins English Dictionary**[1]:

1. the state of having unique identifying characteristics held by no other person or thing

2. the individual characteristics by which a person or thing is recognized

Second, from the **Unabridged Random House Dictionary**[2]:

1. the state or fact of remaining the same one or ones, as under varying aspects or conditions: > *The identity of the fingerprints on the gun with those on file provided evidence that he was the killer.*

2. the condition of being oneself or itself, and not another: > *He began to doubt his own identity.*

Third, from **Merriam Webster**[3]:

1. a : sameness of essential or generic character in different instances

   b : sameness in all that constitutes the objective reality of a thing : oneness

2. a : the distinguishing character or personality of an individual : individuality

   b : the relation established by psychological identification

Collins favors identity as a collection of characteristics; Random House focuses on a state of unique continuity; and Merriam Webster suggests both. All eight definitions share the notion that identity addresses continuity across contexts. Identity means that a subject can somehow be

recognized in a later context as the same subject known from an earlier context:

- Three definitions concentrate on the characteristics that allow this recognition (Collins 1, 2, and Merriam 2a);
- four definitions highlight sameness (Random House 1, 2 and Merriam 1a, b); and
- three definitions feature the mental act of recognition (Collins 1, 2 and Merriam 2b).

These three different focuses (on characteristics, on sameness, and on recognition) all relate what is known about a subject in one context to something else known about the same subject in another context. In other words, "identity" means correlating information about the **same** subject in **different** situations. If we can identify a subject, we can know something about him or her that isn't based on observation.

In certain situations, information for correlating individuals with particular market segments or purchase intentions has direct value to marketers and manufacturers. We sometimes refer to this as an opportunity to sell an 'identity' for a price, discount, or other consideration. However, when we treat the entire notion of identity as a concrete asset (e.g., as "a digital identity") rather than as the emergent phenomenon of identity, we sometimes confuse the conversation. We often refer to the characteristics and credentials we use for recognition as if they constitute identity *independent* of recognition by an observer.

This is seen in discussions of modern identity systems, when professionals and engineers say things like:

- "you select your identity," or
- "we store the identities in the blockchain," or
- "users own and assert their own identity."

These statements ignore the role of the observer and often confuse the listener about what is actually selected, stored, owned, or asserted. We believe this conflation of identifiers, attributes, credentials, and identity is a primary driver of miscommunication and misunderstanding in identity discussions.

3. IDENTITY IS MORE THAN JUST BITS

Identity manifests when we see a face and recall a name. It is in play when we see a badge and acknowledge someone's authority. It emerges when we see an individual and treat them as white or black, gay or straight, male or female.

It doesn't exist without that correlation between an identifier or attribute and a subject. If you can't identify a thing, it means you can't place it. You can't relate it to something else you already know. It has no identity precisely because you can't correlate it with anything else.

Take an arbitrary string of hexadecimal digits:

190B95B104BD41ACA53D9C1486B26C71

Without further information, we can't tell if that is an identifier, an attribute, a credential, or just entropy. It may eventually become an identifier, but it isn't yet — not until someone associates it with the thing it identifies[4]. It's just an example string of digits. It is certainly not an identity.

Yet, if we were to assign that GUID as an identifier for some object, **everything** changes and it **becomes** an identifier. This is the semiotic nature of the signifier and the signified[5]: until the string is known as a signifier referring to some (potentially unknown) signified, it isn't an identifier. It's just a string of hexadecimal digits.

Similarly, any concrete "identity", such as a collection of attributes in a verified claim, isn't actually an identity until and unless an observer correlates it with a subject. We may have a username or an email address or a string of UTF8 characters with the label "name". These bits of information could exist anywhere: in a database, in a form submission, or printed on a piece of paper. They may even occur when millions of monkeys type on a typewriter. We might even know that the data is identifiers, rather than arbitrarily ordered bits, but until specific data is associated with a subject, it isn't an identity. It is, at best, identifiers and, at worst, arbitrary bits.

Consider the password cracking tool Medusa[6]. Like many password crackers, Medusa will accept a file containing strings to use as user names in a brute force attack. Unless the target system actually has a matching login for a given string, that string isn't an identity in that system (or potentially anywhere, as

the strings could be randomly generated). Until the target system correlates that string with the username of an account on the system, it is incorrect to describe that string as an identity. Yet we often refer to usernames as "identities".

We make that mistake every time we refer to "identities" in isolation from that fundamental act of correlation.

Consider this...

> DNA doesn't identify a person until it is used to correlate him or her with evidence at a scene of a crime, or links her or him to an ethnic group or ancestral lineage. Without correlation, DNA is simply a physical encoding that drives protein generation in the body.
>
> The social security number for a child, printed on a card and stored in a filing cabinet, isn't an identity until and unless he, or someone else, uses it. The number itself is just a number.
>
> An "identity" is not the sum collection of all the ways that one might be identified… we can watch any good detective show and see the trail of clues that lead to tracking down a suspect; yet we don't think of each clue as an "identity".
>
> Identity isn't the sum of all of the attributes and actions that might be wrongly or rightly ascribed to you, and may, in one way or another, be used to figure out "who you are". These are at best a digital profile, at worst the data bloat of a runaway surveillance network.

These are examples where common notions of "identity" lead to incomplete, inaccurate, and confusing interpretations.

4. THE PHILOSOPHY AND POLITICS OF IDENTITY

Typical digital systems use identifiers, attributes, and transaction logs to correlate individuals across contexts, such as across multiple visits to a website. Typical real-world systems issue credentials that bind identifiers to long-term, observable, measurable attributes and facts — such as name, race, height, weight, hair color, a picture, and birthdate — in order to correlate a human with certain privileges or responsibilities, like a license to drive.

In discussing both digital and real-world cases, we sometimes confuse the notion of identifying a singular "self" — a specific human person — with the mechanisms by which we do so. This leads to philosophical and civil debates about such things as the right to be forgotten and the innate value of anonymous and pseudonymous speech in a functioning democracy. When we think about "identity" in terms of "who we ***are***", we get caught up in the consequences and ramifications of policy and privacy and human rights. These are important debates, but they often slip into abstractions, miscommunication, and political disagreements that undermine our efforts to build functioning identity systems. On the other hand, when we think about "identity" as a mere collection of attributes or identifiers, we ignore and sometimes dismiss the deeper meanings others interpret in the word.

Even in the abstract case from psychology and literature, of a person searching for their own 'identity', the notion of identity generally derives meaning from the context of an observer associating themselves with groups of individuals or other entities that they 'identify' based on matching, or correlated observable attributes such as race, nationality, economic background, interests and so on.

We shouldn't need to resolve our political and philosophical differences to describe an identity system. Yes, the more human side of "identity" is absolutely relevant when we champion a particular feature or capability. But to describe and discuss the functional mechanisms of a given identity system, we are better served using concrete, non-political, non-philosophical terms.

In the digital and real-world systems described above, we were able to quickly describe the mechanisms of identity using the term "correlation" without getting sidetracked into more abstract discussions like "What is Legal Identity?"[7] or "Is anonymity possible?"

That is the point of this paper.

We argue that, when discussing identity systems, "correlation" enables a more concise discussion and clearer understanding of how identity is created and used[8]. It's not that "identity" is incorrect, it's that the mechanisms of identity are inherently mechanisms of correlation and, *therefore*, we can be clearer by focusing the discussion how correlation is managed.

Everyone, layperson and expert alike, can be more concise, more rigorous, and better understood by

using correlation (*and* anti-correlation) when discussing *the exact same identity systems.*

5. ILLUSTRATIONS OF IDENTITY USING CORRELATION

Following are several examples of identity in the modern world. We discuss each using the terminology of correlation. These situations represent a broad range of identity and identification, demonstrating that our proposed focus on correlation applies beyond specific technical implementations or use cases. It can be used to describe identity in any context.

- **More than a Piggy Bank** *Transitive Correlation*
- **Beverage Bracelet** *Temporary Correlation With Limited Disclosure*
- ***He* Did It!** *Correlation Using Neither Identifiers Nor Consent*
- **I Know Where You Live** *Unwanted Correlation*
- **Pinkeye Guy** *Spontaneous Correlation*

5.1 More than a Piggy Bank

Transitive Correlation

In the U.S, when we go to the bank to open an account we provide our social security number, in part so that banks can comply with federal regulations like reporting cash transactions over $10,000. The social security number is, generally, only used by the bank for regulatory filing (whereas they use an account number and a recorded signature to correlate our deposits and withdrawals with our accounts). In turn, the government uses our social security numbers to correlate our taxable and fiscally regulated transactions throughout our lifetime. This is transitive correlation, where an identifier is used not by the immediate recipient (the bank) for correlating our direct interactions with them, but by a third party (the government) when the recipient needs to refer to us in communications with that entity.

Because of the ready availability of social security numbers and their innate role in reporting personal finances to public agencies, they are *also* often used by financial intermediaries to query and report private financial interactions. Credit bureaus and creditors use social security numbers as a primary identifier to correlate individuals across credit transactions. This unintended use has made the social security number both more valuable and, unfortunately, more accessible, as a target for "identity theft".

Correlation by the US and state governments is the intended correlation. Correlation by creditors and credit bureaus is unintended, and the correlation by identity thieves is undesired.

5.2 Beverage Bracelet

Temporary Correlation with Limited Disclosure

When we attend a music festival, we sometimes receive a disposable, colored bracelet that allows us to purchase alcoholic drinks. To get the bracelet, we provide proof that we are at least the minimum legal drinking age to a single, designated agent at the event. Then the bracelet allows us to purchase drinks from bartenders throughout the grounds without further use of legal credentials. At the point of sale, the bartender can verify that the person ordering a drink has been vetted for the legal age limit using the presence of the bracelet, which can't be removed without destroying it.

This is an example of temporary correlation. These bracelets are durable enough to last for as long as a few days and are generally discarded afterward rather than reused; different events use different colors and patterns so it is a challenge for underage drinkers to know before hand what type of bracelet would let them sneak past the age restriction.

It's also an example of limited disclosure. The information contained in the bracelet is minimal: "the wearer has demonstrated proof of age." This restricts the disclosure of potentially risky personally identifiable information, like birth date or address, to the initial point of issuance[9]. The bartender gets just what they need, just when they need it.

Bracelets are an inexpensive privacy-enhancing technology that also reduces the time bartenders spend checking IDs — which increases sales and profit and reduces the compliance costs for the venue. Not only is it easier to manage than alternative systems, like isolated "beer gardens", but the bracelets themselves also provide evidence of due care to authorities who regularly punish non-compliant vendors with penalties from $10,000 up to revoking the liquor license.

5.3 *He* Did It!

Correlation Using Neither Identifiers nor Consent

It is a staple of crime dramas and real-world courtrooms to call on eyewitnesses to literally point out the alleged offender so the jury can see whom they are accusing. Prosecuting attorneys love eyewitnesses because they provide a human face for corroborating the physical evidence. At the same time, defending attorneys will go to great lengths to question the veracity and the character of the witness to undermine their claims. The success of one side or another can literally be a matter of life or death in cases of capital punishment.

Victory in the battle before the jury depends on whether or not they believe the asserted correlation. That identification does not depend on the eyewitness knowing the name of the accused, their address, their birthdate, or their social security number. The eyewitness simply needs to demonstrate that they reliably recognize the accused as the party they saw committing the crime[10]. Yes, eyewitnesses are known to be wrong sometimes, just as forensic evidence is never 100% accurate. The battle is between the efforts of the prosecutor to correlate the accused with the crime and the efforts of both the defense attorney and the accused to prevent that correlation. Criminals often go to great lengths to lay false trails and hide or destroy evidence, and may even lie or commit further crimes in their attempt to prevent such correlation. The "identity" of the killer ultimately depends on the court's ability to fairly and accurately resolve this battle of correlation.

While many valid digital identity use cases base their architecture on consent and control by the subject, that doesn't apply to all situations. In this example of identity, the subject (i.e., the alleged criminal) does not consent to the correlation. This lack of consent is especially true for law enforcement, border patrol, and the military[11].

In order to allow regulators, lawmakers, ambassadors, and heads of state to make decisions about digital identity systems, it will be vital to understand how such systems correlate people and how they prevent undesired correlations. The goal is a system that is flexible enough — and *understandable enough* — to allow organizations, companies, and sovereign states to choose the best tradeoffs for their needs.

5.4 I Know Where You Live!

Unwanted Correlation

In the Jungle of Calais, seven thousand refugees fleeing political strife and violence have forged a temporary home[12]. In nine months it went from virgin ground to the largest slum in Europe. In these harsh conditions, many fear any form of identification, knowing that their families back home could be punished or killed by the regime they fled if the link is made between them and those they left behind. The lack of identity credentials makes it hard to access justice and health services and to integrate into society. Their fear of persecution keeps many on the fringe. Some have destroyed identity documentation while others avoid even being recorded for unofficial documentaries.

This is the fear some refugees live with every day. For the regimes, it is identity weaponization; for the refugees, it is fear of unwanted correlation. In this case, the consequences isn't the harm done directly to the subject — the traditional focus of privacy efforts — but rather the harm that might be done to friends and family back home. Unfortunately, this directly conflicts with the approaches of several identity solutions presented at the recent ID2020 Summit and the related ID2020 Design Workshop. One in particular proposed a DNA registry designed to help refugees reconnect with family back home. It will be hard to get refugees to participate in such a program when being connected with family is exactly what they fear.

The challenge is to build a system that allows just-enough correlation, just-in-time, to enable the services necessary for human dignity and freedom, without facilitating unwanted correlation that can and does enable further violence and even genocide. Perhaps the trickiest part will be finding a solution that is so clear and obvious that the typical refugee, despite distrust of formal authority and despite speaking a second language, can understand it and believe it won't put their loved ones at risk.

5.5 Pinkeye Guy

Spontaneous Correlation

Consider a dinner party where, a guest who happens to have conjunctivitis (aka pinkeye) tragically trips and breaks the host's favorite vase. Later, we might not recall his name, or maybe we never knew it. Yet,

there's a good chance we'll remember that guy who had pinkeye:

> "Remember that guy who smashed Elly's vase?"
>
> "Who? Oh, you mean Pinkeye Guy?"
>
> "Yeah! I bumped into him the other day at the Royal Oak. Turns out he just published a book."

In social contexts, we sometimes use a unique feature or distinguishing moment to refer to shared or indirect acquaintances. These spontaneous monikers allow us to refer to the party in question in conversation with others who observed — or sometimes just heard about — the memorable distinction. The nickname may never have been used before and may or may not establish a long-term reference. Yet when understood by the people we're speaking with, it allows correlation between the current subject and the referenced individual.

In this spontaneous correlation, the subject has no control and often never even knows about the nickname. The identity is real, useful, and completely emergent.

6. CONCLUSION

Using "correlation" to describe identity systems provides a simpler, more coherent view of mechanisms, capabilities, and risks. The term doesn't change the nature of the system. It is simply more concise and more accurate. It results in discussions that are more rigorous and easier to understand. [13].

Even if you disagree with our arguments about *why* "identity" as a concrete property is problematic, you can still use "correlation" to be clearer.

When you find yourself in a project where the definition of "identity" seems to be a repeating source of challenging conversations, try describing the role of identity in terms of correlation. Shifting to alternative language may help you and your colleagues to see the commonalities in your perspectives rather than the differences. It may allow perspectives to be heard that were getting lost in the debate. It may highlight that the issue at hand is more political than technical and allow the group to steer the conversation in the most productive direction, whichever way that is.

We believe that *every* identity system can be fully characterized by how it manages correlation across contexts.

So, please, use "correlation" when you describe identity systems. Use it when you discuss identity systems with both laypeople and experts. Use it when you build, validate, and improve identity systems.

We believe that doing so will make you more effective and more productive, and your resulting systems more successful and more appreciated.

FOOTNOTES

[1] "Identity." Dictionary.com. Collins English Dictionary - Complete & Unabridged 10th Edition. HarperCollins Publishers. Online. Accessed June 09, 2016. *http://www.dictionary.com/browse/identity*

[2] "Identity." Dictionary.com. Dictionary.com Unabridged. Random House, Inc. Online. Accessed June 09, 2016. *http://www.dictionary.com/browse/identity*

[3] "Identity." Merriam-Webster.com. Online. Accessed June 09, 2016. *http://www.merriam-webster.com/dictionary/identity*

[4] In this case, it isn't at the time of this writing.

[5] Review any text on semiotics for further details.

[6] JoMo-Kun, "Medusa Parallel Network Login Auditor", *foofus.net*. Online. Accessed Jun 14, 2016. *http://foofus.net/goons/jmk/medusa/medusa.html*

[7] The topic of one of the panels at the ID2020 Summit, where, unfortunately, none of the panelists could offer a concise answer.

[8] Note, we did not say "how *identities* are created and used". That's the terminology that got us into this mess.

[9] Data typically found on credentials accepted as proof of age, such as a driver's license or passport

[10] Sometimes the witness didn't see the actual crime, but saw some other correlating activity, such as leaving the building with a suspicious weapon, etc.

[11] In particular, the 2014 Ottowa Treaty governing the use of indiscriminate antipersonnel land mines requires militaries to "positively identify" all targets. Obviously, this is not a matter where consent is appropriate.

[12] O'hara, Finlay. *Jangala.* May 31, 2016. Video. Online. Accessed June 7, 2016. *http://theworldwidetribe.com/2016/05/jangala/*

[13] While in literature and philosophy, identity can be and is often shared based on one or more observable attributes (e.g. gender = male), in discussions of 'digital identity' there is often an unstated assumption of uniqueness. Such uniqueness derives solely from the distribution of observable attributes in the context. For example, three of the authors of this paper are professional consultants and one of the authors is of Indian origin. If the context is set to the set {authors of this paper}, then it so happens that the attribute (country of Origin = India) identifies one of the authors uniquely whereas (occupation = consultant) does not. If the context is set to authors of a different paper written by three Indians and a consultant, then the reverse may be true.

This search for uniqueness, or the need for reduction in uncertainty in identification' may also exist in the social context in the 'immigrants searching for roots' such as 'Americans of Irish origin sometimes identifying themselves as Irish' and so on.

We will discuss this aspect of identity, or identity in the context of conditional probabilities in a subsequent paper.

Additional Credits

Lead Paper Editor: Joe Andrieu

Authors: Joe Andrieu, Kevin Gannon, Igor Kruiper, Ajit Tripathi, and Gary Zimmerman

About Rebooting the Web of Trust

This paper was produced as part of the **Rebooting the Web of Trust II** design workshop. On May 21st and May 22nd, 2016, over 40 tech visionaries came together in Manhattan, New York following the ID2020 Summit at the UN to talk about the future of decentralized trust on the internet with the goal of writing 3-5 white papers and specs. This is one of them.

Workshop Sponsors: Blockstack, Blockstream, Evernym, IPFS, Microsoft, Netki, Tierion, ID2020

Workshop Producer: Christopher Allen

Workshop Facilitators: Christopher Allen with graphic facilitation by Sue Shea, additional paper editorial & layout by Shannon Appelcline, and additional support by Kiara Robles.

What's Next?

The design workshop and this paper are just starting points for Rebooting the Web of Trust. If you have any comments, thoughts, or expansions on this paper, please post them to our GitHub issues page:

https://github.com/WebOfTrustInfo/ID2020DesignWorkshop/issues

The next Rebooting the Web of Trust design workshop is scheduled for October 19th-21st in San Francisco, California. If you'd like to be involved or would like to help sponsor these events, email:

ChristopherA@LifeWithAlacrity.com