A Model-Hybrid 'Platform Technology' for Accelerating and Advancing Life Science Research ~ D.A.S.E.I.N. ~ The Future of Bio-Science Experimentation

Author: Michael Anthony Ricciardi
(Nov., 2013; latest edit: June, 2015)

= =

PROLOGUE:

The past decade or so has seen an explosion of advances and innovations in scientific research technologies – covering nearly every field of scientific endeavor. These advances and innovations are too numerous to list here. This concept proposal utilizes/exploits innovations in three main sub-fields of computer science (data management/analytics, automation, and Artificial Intelligence) and "situates" these in the field of Life Sciences – specifically, in the context of laboratory experimentation. Of the three noted sub-fields, the first (data management/analytics) is embodied in an Information Technology "architecture" that permits distributed network (database) monitoring and analysis, while the second and third noted sub-fields are embodied in a proto-typed (and validated) automated experimentation platform (i.e., a robotic scientist)..

The primary purpose of this model "hybridization" is three-fold: **1**] to greatly increase the rate of scientific/experimental *productivity* (by one or more orders of magnitude, due to the efficient automation and vast scalability of the system), **2**] to improve and accelerate the refinement of experimental *process* (via real-time feedback and reprogramming and enabling precise reproducibility of experimental parameters), and, **3**] to increase the rate of scientific *discovery* (through the rapid and intelligent collection and analysis of "big data" and the real-time modification of experimental protocols). It is proposed and asserted here that, together, these improvements will radically transform ("revolutionize") and advance laboratory-based Life Science research.

In its original manifestation, this concept design was as a proposal submitted to an open innovation challenge in 2013. It was subsequently rejected. This author believes this was due to proprietary/commercial development factors (due to the fact that the two key technologies are either open source and/or already prototyped and publicly disclosed). Additionally, considerations of its structural scale and relative complexity – partly reflecting its essentially "interdisciplinary" nature – likely contributed to its rejection.

None-the-less, the essential/core idea of this proposal is valid, viable, and inestimably transformative, given sufficient vision, will, and resources.

The noted prior challenge submission, having been done under a confidentiality agreement, does not constitute a public disclosure. <u>This present publication does</u>

constitute a public disclosure. Whether or not a provisional or formal patent of this concept design is possible, insofar as it utilizes two cases of prior art (public domain and open source), remains to be seen. At this time of publication, this author has not decided if pursuing any patent claim (e.g., within the one year window permitted after public disclosure or provisional patent application) is warranted or desired. What *is* warranted here is a deeper look at the concept's potential. What is desired (more than any proprietary control) is its adoption and implementation for the advancement of Life Science research and the betterment of the human condition.

This author request only NAME CREDIT in any use or reference (in ANY media) to this concept and its design.

This concept paper does not follow standard research proposal or peer-review journal submission formats and has been organizationally modified from its original submitted form. The primary content is, however, substantively the same (with a few new addenda).

This Proposal contains substantial END NOTES regarding various technologies potentially amenable to integration with the (herein) proposed platform technology.

It is hoped here that readers will "grok" the enormous and transformative potential of this concept, and perhaps advocate its refinement and construction, if not one day utilize it in their own future experimental efforts.

CONCEPT: A MODEL HYBRID (AI + IT) PLATFORM TECHNOLOGY:

The Two Model ("hybridized") Technologies:

1] The Model (prototype) **Artificial Intelligence** System: '*Adam*' – A Robotic (Automated) Scientist [see **pages 14-18** for details, notes, schematic] capable of formulating and testing its own hypotheses and recording all experimental meta data; originator/inventor: Ross King, Department of Computer Science, University of Aberystwyth (2007); most recent version: March, 2009.

2] The Model **Information Technology** Platform: *Hadoop* (a distributed file system and data management platform) [see **pages 18-23** for details, notes]; an open source 'Big Data' management platform.

This model platform is here dubbed 'DASEIN' which stands for:

Distributed **A**utomated **S**cientist **E**xperimentation and **I**ntelligence **N**etwork

Conceptually, we have: HADOOP + ADAM = A distributed robotic scientist network

NOTE: The name DASEIN derives from the German (phenomenology) term *dasein* ("being") which indicates the state of *being in this world* and is sometimes translated as "being there" [see: Hegel, Heidegger, Hüsserl, Merleau-Ponty, etc.)

Description of the Hybrid System/Network:

In the DASEIN, the main/master robot (MADAM) and automated scientist (the approximate equivalent to a Hadoop main "stack"/server, HBase, that runs the program monitoring all databases in the system) monitors all other 'ADAMS' (robots) in a distributed network, compiling/aggregating, assimilating and analyzing their data outputs while each ADAM forms hypotheses, perform experiments (e.g., bioassays), and assesses/refines said hypotheses; each ADAM also collects *metadata* about every experiment it performs and MADAM collects this data/metadata, processes it (including "meta-analyses"), and (if warranted according to programmed criteria) reprograms all other ADAMs to optimize their performance (i.e., make improvements, parameter changes, even suggest a new hypothesis based upon the metadata analysis, etc.*) thus facilitating rapid and *complete evolutionary feedback* in the system…giving us:

Data + Feedback = Improved Performance = LEARNING

…and *learning* at a rate far-outpacing anything that could be accomplished via even the most talented scientific research team (This is not to neglect sheer creative insight that comes from years of knowledge and experience. Indeed, some of said knowledge/experience can be embodied in novel algorithms and then programmed into DASEIN).

DASEIN is a "mashup"/hybrid system that represents Artificial Intelligence and Distributed Network Management design taken to the next logical level.

DASEIN is both a natural (obvious) extrapolation and a radical expansion (and combining) of existing technology.

* Each Adam in the network collects/records its experimental metadata and is capable of "self-reprogramming".

NOTE: the central stack/server of a Hadoop system – given the appropriate software -- can perform/send basic commands to each component server/database in the distributed file system. This function can be expanded to include data analytics (indeed, it is already being done with some networks). In the DASEIN, this role is accomplished/facilitated by MADAM [*madam i'm adam*], the central (main/master Adam) AI component.

MODEL / SCHEMATIC (DASEIN): See **page 14** for a schematic-diagram of an idealized DASEIN).

Quote: "[In this way, citation prediction represents one step on the path]…to creating algorithmic or robot "scientists" that are more creative, risky, persistent, and wide-reading than ourselves." (James A Evans, Computer Science, Perspective Article ~ 'Future Science', *Science*, 4 October, 2013, pgs. 44-45).

How DASEIN Will Impact Life Science Research:

DASEIN will radically increase scientific *process*, *productivity* and *discovery*, realistically, by one or more orders of magnitude.

GENERAL APPLICATION & RATIONALE:

As a general example, suppose we wish to test every element (molecule) of every cell signaling pathway in every model cell used in any given lab, or, to test a therapeutic compound on every metabolic network [e.g., RECON2; see **page.23**] or cell-signaling pathway in every human cell type (re: cell-signaling blockades for cancer treatment). If one were to consider doing this in a conventional manner (even with rapid assays), it would require an enormous amount of time, effort and money, as well as enormous experimental precision (and probably would not be undertaken, certainly not in one project). But, as even *one* Adam in a lab could greatly facilitate such an ambitious endeavor; a network of said Adams centrally controlled and monitored (DASEIN), could realistically accomplish this task in the span of a few weeks.

If updates or refinements to experimental software* are needed, scientists would be able to automatically distribute software improvements (or individualized improvements for a given ADAM) via uploading these to MADAM (who would simultaneously begin monitoring the installation and operational phases of the software installment).

What DASEIN Would Enable:

- Acceleration and increased precision of experimental activity and results,
- Augmentation of scientific ability (enhances hypothesis forming and experimental design),
- Increased productivity (multiplies returns)
- Decrease medium to long-term costs (via automation and efficiency)
- Increased potential for scientific discovery [see "orphan enzymes", **pages. 14-18**].
- Collection of all experimental metadata (thus prevent repetition of flawed hypotheses)

*State of the art bioinformatics software – for data compiling/processing and analysis -- can be readily integrated with the MADAM core software (i.e., if written in the data logic language used by Adam; see **pages 14-18**).

NOTE: Example: *Infosphere BigInsights* (Oracle) – A data/database analytics platform that operates ("lives") on top of Hadoop [see **pages 18-22** for other data

analysis augmentations] is an example of a software augmentation of Hadoop (HDFS) that could also be integrated with DASEIN (e.g., as part of the MADAM component, working with the metadata fed to it from all ADAM components of the DASEIN).

In Nature, in living systems, innovation is driven by (genetic) recombination. This proposal for DASEIN, insofar as it combines and (recombines) existing technologies (and their data/metadata) into a novel platform (a distributed AI experimentation network), *schematically imitates Nature*, while it also accelerates this innovation process.

DASEIN is a true 'omni-inter-accommodative' platform technology (a term coined by R. Buckminster Fuller) that would literally transform Life Science research.

EXAMPLES of Transformative IMPACTS of DASEIN on LIFE SCIENCE Research – Specific & General:

Apart from Adam's proven utility for genetics research in identifying "orphan enzymes" (and their respective genes), a task even more readily accomplished through DASEIN's increased quantitative analysis potential, the proposed platform technology can serve a similar purpose in other Life Science disciplines, such as immunology.

EXAMPLES:

1] Immunological Experimental Procedure: Generating Broadly Neutralizing Antibodies (Variation, Specificity, Avidity, etc.)

RE: Regarding the discovery and clinical testing of an initial group of broadly Neutralizing antibodies (bNAbs) for an HIV-1 Vaccine:

Reference: (Klein F. *et al*) **Antibodies in HIV-1 Vaccine Development and Therapy** (Science, 13 September 2013; pg. 1199)

Relevant Excerpts/Quotes:

"*In vitro*, the most broad and potent antibodies in the initial groups were **b12**, 2G12, 2F5, and 4E10 [which achieved *in vitro* neutralization]…and passive transfer of **b12** [etc.] protected against SHIV [simian immunodeficiency viruses {SIVs} that express the HIV-1 envelope glycoprotein] infection in macaques."

"…**b12** is a phage-derived antibody generated by random pairing of heavy and light chains that may have never existed in nature."

Although attempts by the cited researchers to generate this particular antibody (and the others noted) via vaccination failed, their increased potency (i.e., broadly neutralizing activity) had been successfully demonstrated in both macaques and mice models, and

so they remain, at the very least, models for future development (patterning) and refinement of bNAbs. But note that the "random pairing of heavy and light chains" (located on the antibody molecule, e.g., IgG) refers to a (nucleotide/amino acid variation-generating) process that could be automated, and thus the correct (broader affinity*) antibody sequence could be achieved in a vastly faster time frame (note also: the sentence: "may have never existed in nature" would seem the ideal and appropriate domain for a robot's generative output).

> NOTE: For the generation of genomic variation, an *automated genome engineering platform* [e.g., see: MAGE, **page 29**] could be partly or entirely integrated into ADAM's hardware/software systematic design and repertoire. DASEIN, being an automated platform technology, with expandable capacity, is ready-made to incorporate other automated platforms.

Imagine if a research team had a network of Adams manipulating phage antibody genes (each gene sequence *variant* of the antibody encoding region would be the equivalent of a new hypothesis) and then performing *in vitro* testing (the new experiment) of each variant against HIV-1 infected (blood) cell lines (monkey, mouse, human), recording outcomes, collecting data/metadata, refining each hypothesis (and the related experimental protocol)…all towards the goal of finding the most broadly neutralizing (highest, variable affinity) antibody sequence…and all done efficiently, rapidly, and with the collection of precise and maximal data (thus also allowing exact replication of the experiment). [see 'Reproducibility', **page 8**]

As noted, all Adams in the network would feed their metadata to MADAM, which would analyze the data {and experimental outcomes} to determine the most promising direction (neutralizing sequence, and then send commands to all Adams to proceed {refine}from that sequence (onward) for future recombinant *in vitro* experiments…All at a rate of scientific productivity unachievable by humans. This is just one example of how DASEIN could be put to use and how it could accelerate discovery and development of one class of therapeutic molecules (bNAbs) targeting a specific disease.

> * Affinity maturation is the process by which an antibody's (e.g., IgG) antigen-binding amino acid sequences (found on its H/L chain, CDR I, II, III regions) come to more closely match that of the antigen, resulting in greater 'affinity" (also *avidity*) and thus greater neutralizing potential. This maturation process is quite time-consuming (and in-exact, often resulting unwanted/uncontrolled variation), but could be greatly accelerated if DASEIN were integrated with, for example, an automated genome engineering platform (e.g., a modified MAGE, or "clonetegration" platform, see **pgs. 29-34**; see also: 'Affinity Maturation**', pgs. 34-35**).

2] Rapid Assaying for Preempting Climate Change-Induced Disease Epidemics in Vulnerable Species & Ecosystems:

RE: (Ecology) Inhibition Assays to identify microbes and/or molecules/compounds that inhibit/cure pathogenic infections in target animal populations.

– Experiments to stop or mitigate the on-going, nearly world-wide decimation of *amphibia* (e.g., frogs, salamanders and caecilians) due to a *chytrid* fungal infection (*Batrachochytrium dendrobatidis*) currently focus on identifying species of probiotics (e.g., *Janthinobacterim lividum*, and *pseudomonas*; see R. Harris, M. Bletz, A. Loudon, 2013) that inhibit or kill the lethal fungus. Testing these probiotics involves conducting inhibition assays (i.e., assays to determine a "zone of exclusion" around a given probiotic species);

This procedure is ideally suited to Adam's programming and automated functionality – including various alterations in the assay protocols (note: Adam is "programmed to carry out over 6.6 million types of bioassays").

This line of experimentation was accomplished over a lengthy time period at great effort and cost and all without the aid of an automated scientist (or network of same). Climatologists and zoonotic scientists predict that climate change impacts (e.g., and the invasion of habitats by climate-induced displacement of "alien" species), combined with the global trade in biological specimens (as well as habitat destruction), will produce many more forms of lethal diseases (due to bacterial and fungal invasions, or ecosystem "regime changes") that hold the potential to disrupt ecosystems and drive many animals to extinction.

A distributed network of robotic scientists, conducting *millions* of assays on a myriad of bacteria species (and/or gene variants; SNPs or SNVs of same) -- many potentially identified through DASEIN's stored data libraries* (of phyla, species, gene sequences, proteomes, etc.) -- could greatly aid scientists in *preempting* the pandemic spread of these predicted diseases, or, at the very least, in radically mitigating their impacts, through having already done the time-consuming (often trial and error) work of identifying candidate probiotics (and genetic variants of the same) and testing for their anti-pathogenic gene products (e.g., the discovery of violacein in *J. lividum*).

> *Although not specifically described in this proposal, data libraries should be considered a given when discussing automation and AI technology in the context of laboratory experimentation. The latent (data storage) capacity of a DASEIN should be "built in" but could also be augmented readily at any time with additional data drives.

3] Unlocking the Mystery of Cellular Organization and Life's Origins

For two final "Big Idea" examples, the emerging science of Synthetic Biology (and its over-arching quest to synthesize living systems -- cells), would be given a tremendous boost through proper implementation of DASEIN. Both 'bottom-up' (*de novo*) and 'top-down' (modularity) approaches to achieving this goal would be greatly facilitated (especially the former approach, through testing various molecular "scaffolds").

As molecular biologist Robert L. Dorit noted in a recent *American Scientist* essay (Sept.-Oct. 2013, pg.342-345):

> "Technological breakthroughs including the automation of laboratory tasks (and dramatic increases in computational power)…have expanded the scope of synthetic biology."

How much more so (this expansion) with a network of automated scientists on your side?

The famed 1958 Miller-Urey experiments (and later attempts to replicate these) demonstrated a key amino acid (i.e., adenine) could "spontaneously self-organize" under the right chemical-molecular conditions and a jolt of energy. But these experiments did not produce self-replicating molecules. This molecular achievement would not come until much later (2009) when researchers Lincoln and Scott were able to produce self-replicating ("cross-catalyzing") RNA enzymes that did so without stopping. The time between the earliest attempts to reproduce living matter and more recent ones was half a century (!)…and still, key steps leading to biogenesis and the creation of *autopoietic* systems remain missing and mysterious.

Imagine a DASEIN performing simultaneous biogenesis experiments (e.g., different mixtures of different fundamental molecules, environmental conditions [pH, temperature, pressure, etc.] and energetic inputs) – able to imitate, monitor and modify multiplicitous chemical-catalytic scenarios over greatly contracted time frames (or varying time frames)! How soon would the mysteries of biogenesis fall away?

And, looking exceedingly far forward…with data input from planetary science research, a DASEIN could synthetically reproduce conditions (or plausible conditions) on newly found exoplanets based upon spectroscopic analysis of their molecular makeups…and hence even one day helping to plan interstellar (colonizing) missions! With this application, the field of exobiology would be given a tremendous technological boost.

General Examples of the Transformative Impact of DASEIN

1] Overcoming a Big Problem in Bio-Medical Research: Reproducibility

A widely acknowledged problem/issue in biomedical research is that of *reproducibility*; many published studies cannot be replicated (due to several possible reasons). Further, many researchers are simply not motivated to replicate existing work (despite the acknowledged importance of doing so). A Reproducibility Initiative was recently announced and funded (by a private foundation) to commence an effort to reproduce "50 landmark cancer studies published between 2010 and 2012." (see: http://scim.ag/Reprod); Science Exchange, an "on-line marketplace", will "farm out

the experiments to different companies" [source: news blurb: *Science*, 25 October, 2012; pg. 406-7].

One main reason for the reproducibility issue is that many scientific studies, despite formalized experimental protocols, still allow/introduce variations in their procedures which are not (or not adequately) documented. Slight modifications (to equipment, assays, data generation/analysis, etc.) made in the course of a study can have significant impact on future replication attempts.

With an automated scientist like Adam, metadata about every aspect of an experiment is collected "as a natural consequence of it programming". All this data is preserved, and can be accessed (by another robotic scientist or human researchers) as needed for any reproducibility study. This prevents unintentional and undocumented variations from entering the experimental procedure. On the other hand, any *known* variations or modifications (e.g., from an earlier study/experiment by humans) can be programmed into MADAM, distributed to all network Adams, and faithfully replicated (again, with all metadata preserved) *multiple times*. [see also: End Notes, 'Meta-Knowledge', **pgs. 24-28**]

Further, the so-called "file drawer" issue of unpublished (or undisclosed) null results/negative findings would no longer be an issue, as all data (results) are preserved and may be searched at a future time – thus preventing researchers from repeating flawed experiments or failed hypotheses. Experimental results that do not meet specified criteria (data parameters) would be "tagged" by MADAM (via its analytics program) for later analysis or refinement.

Instead of "farming out" reproducibility studies to multiple companies (anyone of which could unintentionally miss, or incorporate, said variation), utilization of a DASEIN type system (which would be scalable to the required size or number of replication studies sought) would vastly improve overall efficiency, cost-competitiveness, obviate the issue of researcher low motivation/enthusiasm (for replicating studies), faithfully and precisely reproduce all documented aspects of an (existing) experimental protocol, and preserve all metadata for future reference and further replication.

DASEIN's impact on the validation of existing studies (through enhanced reproducibility) would be nothing short of unprecedented and "revolutionary".

2] The Power of (Robotic) Team Work - Robots in Knowledge Production:

Reference: (S. Wuchty, B.F. Jones, B. Uzzi) 'The Increasing Dominance of Teams in Production of Knowledge' (*Science*, 18 May, 2007; pgs. 1036 – 1039).

Summary (accompanying the published paper):

> "We have used 19.9 million papers over five decades and 2.1 million patents to determine that teams increasingly dominate solo authors in the production of knowledge. Research is increasingly done in teams across nearly all fields. Teams typically produce more frequently cited research than individuals do, and this advantage has been increasing over time. Teams now also produce the exceptionally high-impact research, even when that distinction was once the domain of solo authors. These results are detailed for science and engineering, social science, arts and humanities, and patents, suggesting that *the process of knowledge creation has fundamentally changed*. [emphasis added]

Imagine if a given research team had a DASEIN to aid its research, or, more radically, if a given research lab, institute or incubator engineered and implemented their own (exclusive) automated scientist network as an *independent research team* dedicated to a given research goal!

If human teams (aided by computers, but which are not AI/automated scientists) can out-produce solo researchers (to the extent described above), how much more would the "process of knowledge production" be changed – indeed, *revolutionized* – with the constructing and implementing of a DASEIN-like system?

> NOTE: While it makes logistic sense to locate the DASEIN "in house", theoretically, the DASEIN could be globally distributed/controlled (via the Internet), as with a common Hadoop arrangement of files distributed across multiple databases located on multiple remote servers. Thus, different entities could host (and share the cost of) an individual Adam component of the larger DASEIN (see: Cost/Funding, next section).

Risks and Challenges in Implementing an Automated Experimentation Network

Big Data - This increased productivity does create a Big (or Mega) Data issue. But we must recall that DASEIN is modeled after the Hadoop architecture and Hadoop is "the world's *de facto* Big Data platform"; DASEIN is not just modeled after Hadoop, but would actually include its own Hadoop "stack" and system software (recall that each ADAM in our DASEIN is the structural equivalent to each database {or remote server} in a Hadoop system).

Any Big Data issues that arise separate from the immediate management and processing of DASIEN generated data (such as off-site/off-network storage) can be handled with a combination of new data compression techniques (which can be performed by the robots) and cloud storage (provided and shared by the sponsoring research institutions, companies, academies). Any additional "number crunching" can be facilitated through use of distributed computing and/or super computers, or possibly novel storage mechanisms.* [see also: End Notes, **pages 21-22**

> *Since each ADAM is designed to handle biological materials, in can certainly handle DNA and probably even DNA *data storage* functions.

This ability (note: storage capacity of DNA = a million times more data storage capacity than on a typical microchip; see reference) will help alleviate some of the large data storage requirements (for a given experiment), but perhaps only transiently until the core data can be compressed and extracted).

Reference: EMBL-European Bioinformatics Institute (EMBL-EBI), *Nature*, January 23rd, 2013 (Goldman *et al*)
http://www.sciencenews.org/view/generic/id/347702/description/DNA_stores_poems_a_photo_and_a_speech

New / More Powerful Software – Clearly, any implementation of a DASEIN-type system for conducting increasingly complex experimental protocols, and/or introducing new assay technology, etc., will require the engineering/design of more powerful and robust (flexible) software – especially the main AI programs (note that this need has already been anticipated by King *et al*; see **pg. 14**). Hardware components can be added as needed (also with some degree of modularity) as each component of the network has built in latent capacity (see the diagram of Adam, **pg. 18**).

Cost / Funding – Perhaps the biggest challenge of implementing a DASEIN platform technology is its cost, which, in the short-term, could be large. However, as there is already a tested prototype to model ('Adam'; King *et al*), costs associated with prototyping/testing a new system would be significantly decreased (although a period of prototype testing *for this proposed purpose* would still be necessary, for both second generation Adams and building of the complete DASEIN). Costs beyond this would be primarily for scaling up components to the network level, and, refinement of the AI operating system. Funding sources could include a government-academia-corporate venture partnership, or private foundation collaboration, or some combination of the above. Ultimately, building and implementing a DASEIN may be one of those 'Big Science' projects developed primarily through government funding (like the Human Genome Project) or corporate R & D capital (like IBM's *Big Blue*).

Scientific Culture – Scientific culture may prove to be the greatest challenge to solve prior to the adoption of DASEIN. While scientists are certainly accepting of computers (most research could not be conducted without them), the idea of a robot scientist (apart from some satirical asides) – and especially a distributed network of robot scientists – might be viewed as something too "disruptive" -- boding ill for the human scientific endeavor (e.g., something approaching Vinge's 'Singularity'). But apart from the real economics issue of replacing human work/jobs (which this author is sensitive to; see "put scientists out of work", next section), robots currently aid several other scientific disciplines, primarily the Earth and Marine sciences, but also, increasingly, the medical and bio-med engineering fields. An application (and expansion) of this robotic presence to the Life Sciences would not seem so surprising then. That said, the full impact of DASEIN on scientific culture in the medium to long-term is unpredictable.

FINAL RATIONALES & SUMMATION

The Fear of Replacing Scientists:

The implementation of DASEIN will not "put scientists out of work", *en mass*, rather it will free up scientists from the necessary tedium of experimental design, procedurals and execution (and much of the analysis that follows), so as to pursue more basic research, theoretical investigations, and translational applications (which will, consequently, also be sped up) …And, it will mean robots joining human scientists in their research endeavors (robots as research team members; see: "Teams", **pg. 9**) which is the implicit promise and power of Artificial Intelligence.

Not Just a 'Revolutionary Technology' ~ An 'Exponential Technology':

An 'exponential technology' is a technology that demonstrates continued accelerating growth of capabilities (speed, efficiency, cost-effectiveness, and/or power), driven both by advances in the individual technologies themselves, as well as through their interplay and synergies (e.g., with other platform technologies). [reference: the Singularity University webpage <http://singularityu.org/graduate-studies-program>]

The proposed DASEIN concept, insofar as it offers the potential to transform Life Sciences research methodologies and experimentation, and, demonstrates the (inherent) potential to accelerate "growth of capabilities", is both *revolutionary* (in its functional expansion) and *exponential* in its synergistic possibilities (e.g., through integration of other platform technologies and/or software, new assay techniques, data collection/metadata analysis, and feedback/learning, etc.).

ADDENDUM of SPECIAL NOTE ~ A Final Integration?

Within just the past six to seven years, a new platform technology has arrived on the scene that is already transforming the design and fabrication fields. It is called 3D printing (AKA rapid prototyping or "additive" printing) and it is finding wide application in fields as diverse a nanotechnology and tissue regeneration. Indeed, the 3D printing tech has already been used to print customized DNA-based nanostructures (see: parabon.com), living human stem cells (Faulkner-Jones *et al*, 2013), and organ tissues. The tech's completely digital programming (typically utilizing CAD-based 3D models) and' layer by layer' operation allows a highly efficient (low waste), bottom-up manufacturing process that can be scaled up or down as needed. One need only do a simple Web search to learn about the myriad uses for this rapidly expanding (and declining cost-per-unit) technology.

As noted previously, DASEIN, being an automated platform, is ready-made to integrate other automated technologies.

Imagine if each ADAM in our DASEIN was connected (proximally) to its own 3D printing apparatus, whereupon, under the meta-analytic "eye" of MADAM (or each individual ADAM's AI analytics program), it could rapidly manufacture specific cellular tissues or complete "organoids" – each with its own functional extracellular matrix – and proceed to test various experimental chemicals/compounds (e.g., new pharmaco-therapeutics; small molecules, "macrocycles" – or even gene therapy-manipulated products [proteins, receptors, RNA species]) -- on these living tissue structures…and then efficiently perform various assays (e.g., endocytotic, immunoprecipitative, etc.), collect and analyze all experimental and meta data…and then refine each and every hypothesis and experimental protocol with great (AI-enhanced) precision and operational plasticity!

And, given the integration of one or multiple bioreactors to provide for the incubation of test tissues/meta-cellular structures, and the "modding" of the DASEIN's micro-titer apparatus to handle larger sample sizes, scientists could thereby engage in highly advanced living cell/tissue experimentation; the DASEIN would precisely monitor all vital cell/tissue/organoid parameters to maintain optimal growth and metabolic conditions for a prescribed duration – and naturally allow exact reproducibility of these parameters and conditions. Such a capability would radically transform pharmacological/bio-medical experimental practice – not least of which is potentially eliminating the need for testing drugs (or gene products) on (live) animal models prior to human test subject validation.

This "final" technology integration* would thereby decidedly move laboratory experimentation into a new *ethical* era of Life Science research (where animals are used, if at all, only in highly specific and necessary experiments).

I will conclude this concept proposal here, noting only that the model AI used here (Adam) is but one prototyped model of an automated scientist, and no doubt, other models will come along in the future. That said, Adam has proven its scientific/experimental prowess already (see End Notes, **pages 15-16**), and further, is designed for expanded capacity and extended integrations of laboratory technologies. It is the ideal automation technology to base an intelligent network upon, and thus to commence our journey into the Brave New World of Automated Intelligent Experimentation Networking.

DASEIN is the Future of Life Science Research…*Be There*.

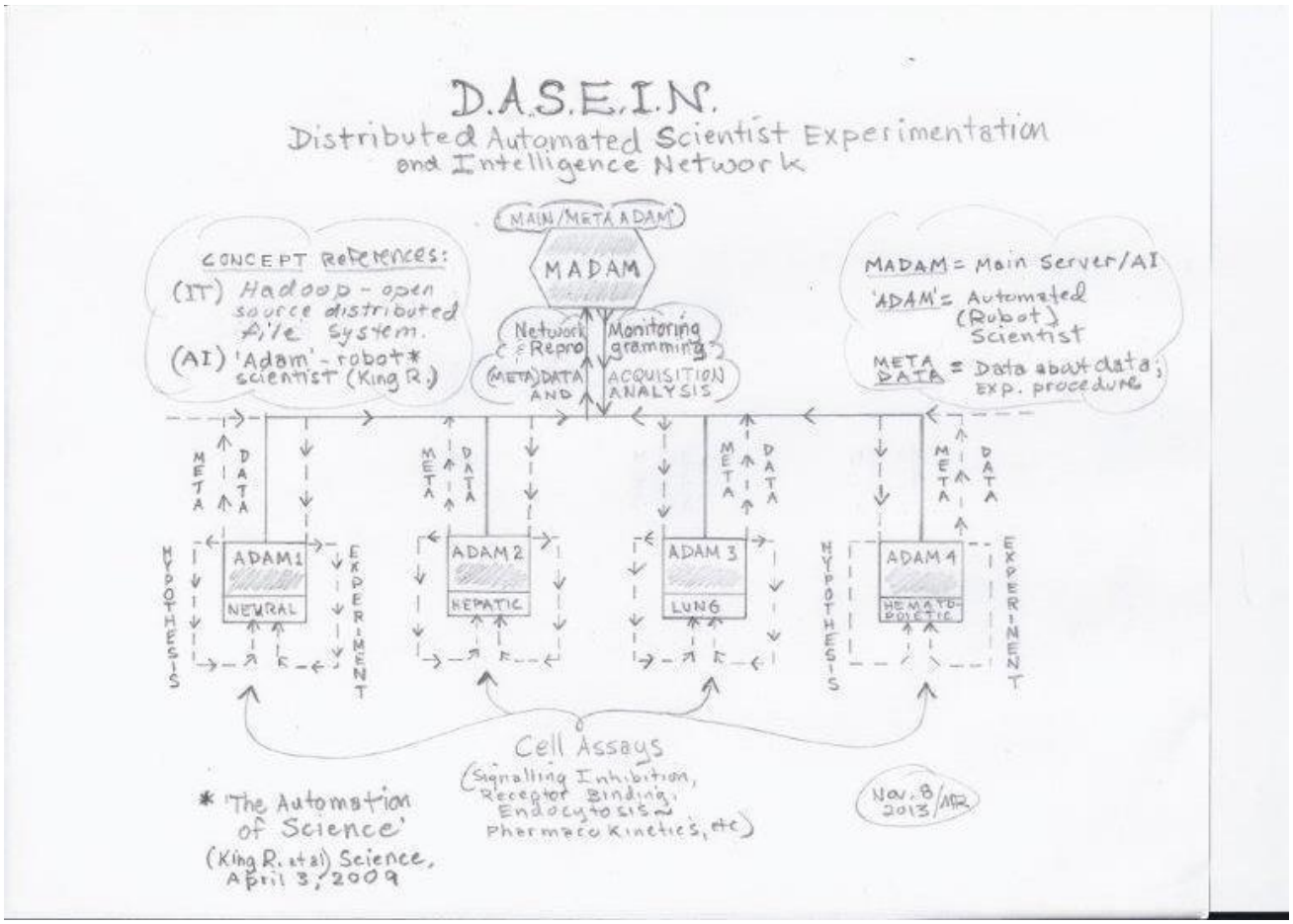
> *Given the "synthetic" or synthesis potentials of 3D Printing technology (including more recent advances such as CLIP), other automated synthesis platforms may be integrated with it to expand the synthesis capability and the over-all functionality of DASEIN. An example of this is the 'building-block based synthesis platform' (Li *et al*) [citation: **'Synthesis of many different types of small molecules using one automated process'** (Junqi Li *et al*); *Science*, 13 March, 2015; pg. 1221-1226.

= = = = = = = = = = == =

END NOTES, REFERENCES, CITATIONS & PAPER EXCERPTS

= = = = = = = = = = == =

CONCEPT DESIGN - SCHEMATIC DIAGRAM for DASEIN:



MAIN TECHNOLOGY REFERENCES (MODELS) for DASEIN:

Model Innovation/Technology: 'Adam' - A Robotic Scientist

Description: Adam is a physically implemented laboratory automation system that exploits techniques from the field of Artificial Intelligence.

Key Attribute/Features of Adam:
Existing high-throughput methods are inadequate for areas such as systems biology because even though large numbers of experiments can be executed, each individual experiment cannot be designed to test a hypothesis about a model. Robot scientists have the potential to overcome this limitation.

The complexity of biological systems necessitates the recording of experimental metadata in as much detail as possible. With robot scientists, comprehensive metadata are produced as a natural consequence of the way they work.

Use of a robot scientist enables all aspects of scientific investigation to be formalized in logic

> NOTE: for Adam, the engineers {King et al} developed LABORS, a customized version of EXPO, expressed in the description logic language OWL-DI; applications of LABORS produce experimental descriptions in the logic programming language Datalog.

Additional – Quotes / Notes:

> [Selected quotes from the report: 'The Automation of Science' (King, *et al*)]
>
> "The complexity of biological systems necessitates the recording of experimental metadata in as much detail as possibleYWith robot scientists, comprehensive metadata are produced as a natural consequence of the way they work. [because experiments are conceived and executed by computer] It is possible to completely capture and digitally curate all aspects of the scientific process and [they] can even *contribute to scientific knowledge* (such as identifying genes that encode "orphan enzymes"*; in the case reported, in the yeast *Saccharomyces cerevisiae*, via autonomously designing and testing hypotheses (i.e., candidate genes)"
>
> > Note: Adam's programming permits the execution of up to 6.6 million biomass measurements.
> >
> > > *****Orphan enzymes: enzymes that catalyze biological reactions but for which no encoding genes are known
>
> "Adams's hardware is designed to automate the high-throughput execution of individually designed microbial batchwork growth experiments in microtiter plates. **Adam measures growth curves (phenotypes] of selected microbial strains (genotypes) growing in defined media (environments)."**
>
> "Adam is capable of designing and initiating over a thousand new strain and defined-growth-medium measurements each day (from **a selection of thousands of yeast strains**) with each experiment lasting up to 5 days."

"Adam has autonomously generated functional genomics hypotheses about the yeast *S. cerevisiae* and experimentally tested these hypotheses by using laboratory automation."

"Adam formulated and tested 20 hypotheses concerning genes encoding 13 orphan enzymes; 12 of these hypotheses, with no previous evidence, were confirmed with $P<0.05$ for the null hypothesis."

> Note: subsequent purification of the gene's protein products, and used in manual *in vitro* enzyme assays, confirmed Adam's conclusions.

"The advances that distinguish Adam from other complex laboratory systems are the individual design of the experiments to test hypotheses and the utilization of complex internal cycles."

> **Note: Adam collects "metadata" about the structure of the experiments and its own execution of them. Such <u>metadata is highly valuable for assessing performance and potentially for identifying key components of the design/tests that can be modified to enhance the overall performance, and thus potentially expedite discovery and new knowledge.</u>**
>
> **IMPORTANT NOTE**: **Adam's success at identifying the genes encoding orphan enzymes could be applied to the identification of *histone-modifying enzymes* (e.g., methyltransferase, demethyltransferase, etc.), as the emerging field of epigenetics has become the focus of much research and investment to identify new classes of therapeutic compounds that target the components of the epigenome**.

Reference/Citation: 'The Automation of Science' (King, R. *et al*), *Science* Magazine {News Report}, April 3, 2009, pgs: 85-89

<u>Robot Scientist (Adam) Discovers Enzyme Gene</u>

<u>Interview with the originator (Ross King):</u>

[interviewer] Adam's accomplishments are impressive (e.g., conducting millions of biomass measurements), most notable of which was the identifying of the genes for several Aorphaned@ enzymes (in the yeast *saccharomyces cerevisiae*.). Adam's work was later confirmed by manual experiments. Additionally, Adam records what is known as 'meta-data' (data about the actual structure and performance of the experiment) which offers scientists additional valuable information for future experiments and their analyses.

Questions: 1] What future uses for (this version of) ADAM do you have planned (beyond the yeast genome/tracking orphan enzymes)?

[King] **We are using it to help build systems biology models of yeast. We plan to use it to work with the worm *C. elegans*.**

2] Has there been more follow up analysis of ADAM's work since the paper (e.g., have its hypotheses regarding non-orphaned enzymes received any further validation)?

> **Not to my knowledge**.

3] For the lay (but scientifically interested) reader, could you clarify exactly what ADAM discovered, or was able to do that other/human scientists were not?

> **Adam discovered which genes encode certain enzymes in yeast. Of course, it would have been possible for a human scientist to have done the same thing. The probable reason that this scientific knowledge had not previously been discovered is that it is generally the case that there is a one to one relationship between genes and enzymes (one gene encodes one enzyme), while in many of the cases that Adam investigated the relationship is many to many (one gene encoding more than one enzyme, and one enzyme encoded by more than one gene).**

4] Regarding ADAM's collection/recording of meta-data (i.e., information on the structure of the experiments)...has this proven useful yet to other human scientists (in the same field or in other scientific fields)?

> **We have found it useful to investigate questions about systems biology, and what is more we have also reused the knowledge from these systems biology investigations to investigate other problems. We are also collaborating with experts in data mining to reuse the data/meta-data.**

5] How do you see ADAM being improved (future versions)?

> **I would like to make future systems much more general purpose so that by simply *changing the software* most molecular biology experiments could be executed. The AI component also needs to be made much more intelligent**.

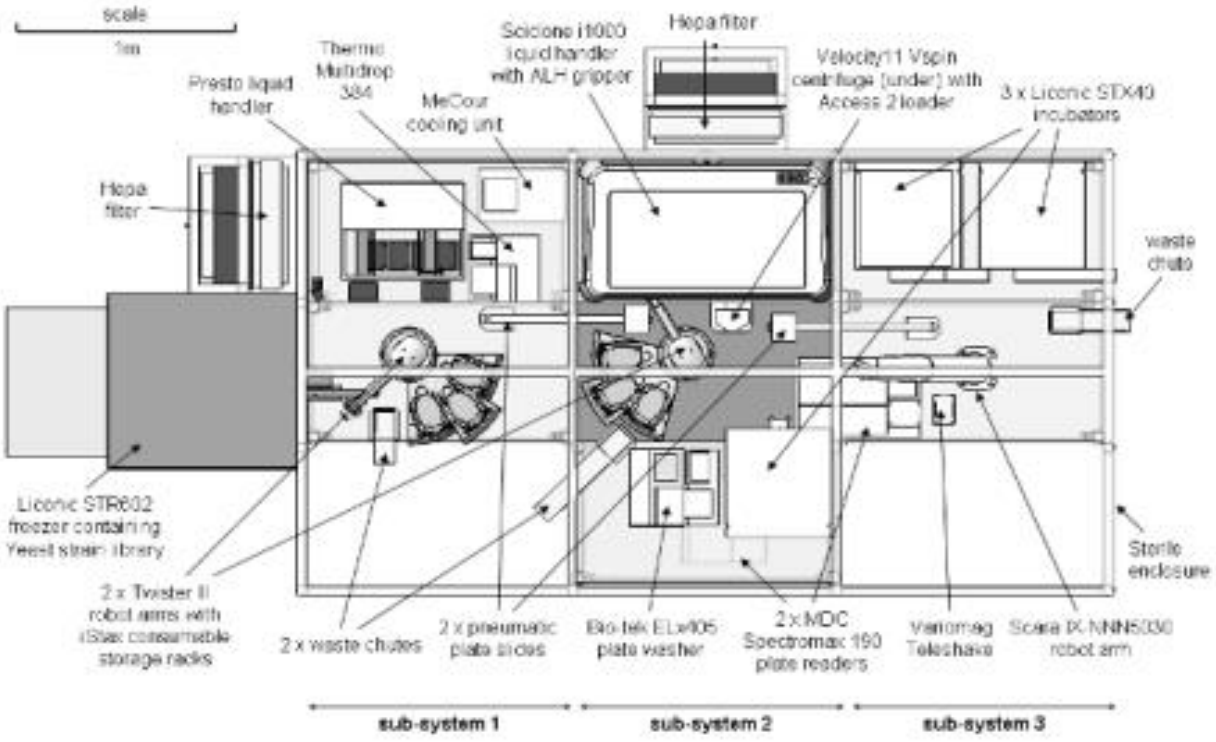
Any additional thoughts or comments are welcome!

> **I think the advantage of automation in producing comprehensive and explicit descriptions in the *logic of experiments* needs to be emphasized. This makes the knowledge generated much easier for humans and other machines to use.**

Ross King
Department of Computer Science University of Aberystwyth
http://www.aber.ac.uk/compsci/Research/bio/robotsci/press/photos/

Notes [excerpted text, diagram] from the *Science* Magazine report: 'The Automation of Science' (3 April, 2009)

[*diagram*]: **Adam, top view, labelled**; source: http://www.aber.ac.uk/en/cs/research/cb/projects/robotscientist/pictures/3d-july07/items/image-65327-en.html]



= =

HADOOP – A DISTRIBUTED FILE SYSTEM

Reference (excerpts from): *Too Big to Ignore - The Business Case for Big Data* by Phil Simon (Wiley & Sons, 2013)

RE: a (model) distributed file system (centrally commanded) for monitoring and managing large data content from multiple sources (i.e., databases, not necessarily interconnected). Note: the DASEIN "architecture" mimics this system.

HADOOP (*Apache*) – A distributed file system (HDFS) that can access and compile data drawn from multiple databases/servers; HDFS can also run analysis software on multiple datasets across multiple DBs/servers.

The Details:

Hadoop is a distributed file system (Hadoop Distributed File System, HDFS)

HDFS stores vast amounts of data used by other parts of the Hadoop 'stack'.

HDFS works closely with MapReduce (MR) a distributed programming framework designed to run on commodity hardware.

Google's MapReduce (+ Google file system):

- Breaks Big data problems into much more manageable subproblems
- Distributes those subproblems to myriad "processing nodes" *
- Reaggregates them (subproblems) into more digestible datasets

Normally, MR does not provide access to real-time data, but this is changing (re: new Hadoop components like HBase – an open-source implementation of Google's NoSQL architecture).

HBase is becoming a key part of the Hadoop stack.

The scale of HBase is enormous - billions of rows and millions of columns; ensure that both read and write performance remains constant.

The HBase NoSQL database is built on top of HDFS (see also: NewSQL which = old SQL + NoSQL)

*"Sub-problems" and "processing nodes" can be viewed as the conceptual equivalents of experimental variants and additional ADAMs in the DASEIN network, respectively.

Additional Hadoop details (drawn from the cited book):

- large collection of open-source projects that distribute and process data
- Hadoop stack and its different components allow organizations to store and make sense of vast amounts of semi-structured and unstructured data
- "world's de facto Big Data platform" (used by FB, Linkedin, AA, IBM, twitter)

Features:

- **can handle many different types and source of data** (structured, unstructured, log files, pictures, audio files, communication records, and email; see also: *Splunk* <*www.**splunk**.com/*>)

- **scales "easily" and across multiple servers (i.e., it is schema-less)**
- high fault tolerance
- **extremely flexible**
- open-source project (has own "ecosystem", **enables** community **improvement**)

Problems/Issues/Challenges (Hadoop):

- no single or official stack or standard configuration
- more than a dozen dynamic components or sub-projects (which are complex to deploy and manage)
- installation, configuration, and production deployment at scale is often challenging.

> **Note: This scaling challenge (apart from cost) would not be the case with DASEIN, as envisioned; scalability of DASEIN is based upon the scalability of the ADAM/robotic scientist design and individual components can be locally (proximally) connected (i.e., not connected via the Internet or cloud server), making scalability a function of the space in which DASEIN resides.**

Quote:

> "…the ability to do real-time queries on a massive data storage and processing framework is finally becoming a reality…we now can worry less about how to handle the data and more the actual insights* that we can derive from it." [Scott Kahler, BD architect for Acknowledge; quote source: cited book, above]

> **Note**: **"insights" – various data exploration methods (e.g., to discover novel relationships in data sets) can be employed here (e.g., MIC/MINE statistical methods; [see Reshef *et al*, 2012*]) by the network "master" brain (MADAM), via remote/local access to databases containing these analytic programs, each rendered in the native programming language (LABORS; see previous End Note; 'Adam').**

> **Note2**: **Other examples (DB architectures/management):**

> *Rainstor* – A cloud DB management service which permits data querying and analytics (of structured and unstructured data) directly by SQL, BI tool, or MapReduce on Hadoop; stores data in partitions (large blocks for easy management via HDFS) also has the advantage of offering a 95% reduction in one's "storage footprint" due to its data compression and de-duplication technology (thus, saving much money on Big Data storage costs).

Infosphere BigInsights (Oracle) – A data/database analytics platform that operates ("lives") on top of Hadoop.

*** 'Detecting Novel Associations in Large Data Sets'** (Reshef D.N. *et al*); *Science*, 16 December, 2011; pgs. 1518 – 1524.

RELATED QUESTIONS: Can a Hadoop platform/HDFS architecture, and its search and real-time analytics APIs, be implemented to effectively manage the growing volume of genomics data (due largely to next gen sequencing technologies which greatly reduce sequencing time and greatly *increase* total sequencing data)…data that is and will be increasingly critical to genome analysis efforts and the discovery of genomics laws and relationships?

SOLUTION:

IDEA: A 'Meta-Hadoop' architecture for managing multiple HDFS 'stacks' (which themselves manage multiple DBs) and allowing for massive (inter-database) data accessing, compiling and real-time analytics.

Note: **'Meta Hadoop' architecture is "mapped" onto the DASEIN design.**

RELATED PROBLEM/CHALLENGE: Cost of designing, building and maintaining such a *meta-hadoop* architecture would be large, but if the ("multi-cloud") architecture were globally distributed (e.g., amongst research institutes), its cost could be globally shared, as well.

RELATED QUESTION: **What to do about the growing volume of genomics data**; **(storage space, costs, etc.)?**

REFERENCE: (perspective article**) 'On the Future of Genomic Data'** (Scott D. Kahn), *Science*, 11 Feb., 2011; pg. 728.

Quotes (from the article):

"The Study of genomics increasingly is becoming a field that is dominated by the growth in the size of data and the response by the broader scientific community to effectively use and manage the resulting derived information."

Factoid (from article graph, fig. 1, not shown): A doubling of sequencing output every 9 months has outpaced and overtaken performance improvements within the disk storage and high-performance computation fields.

Note: DASEIN component 'MADAM' AI software integrated with data compression software, etc.

"Such a growth in raw output [from next-gen sequencing] has outstripped the Moore's Law advances in information technology and storage capacity…" (note: "…in which a standard analysis requires 1 to 2 days on a compute cluster and several weeks on a typical workstation.")

[this growth] "…is driving a discussion and the value and definition of "raw data" in genomics, the mechanisms for sharing data, the provenance of the tools that effectively define the derived information, and the nature of community data repositories in the years ahead."

OTHER SOLUTIONS:

> **Centralization** of data on the Cloud ("a positive start") – this implies some type of central management/infrastructure to support it [see also: Related Problem/Challenge, above]

Note: **Thus, this centralization approach would seem to map onto DASEIN Readily (at least for this purpose) and would serve as a potential 'dual use' for the DASEIN system**.

An Alternative:

> **Service-Oriented Architecture** (SOA) – "moving the computation to the data instead of the data to the computation" – SOA encapsulates computation into "transportable compute objects" (run on computers that store the targeted data; these 'objects' function like applications, only they are automatically uninstalled after performing the analytic operation).

An Additional Solution – Data Compression:

> (re: minimizing data storage) **Reference Genomes** – only genomic (base sequencing) data to be stored are the *differences from the reference genome* – (i.e., storing only the base mutations which represent just 0.1% of the genome data).

> (Re: redefining 'raw data') **'Compressed' Raw Data Output** – real-time processing of raw (e.g., sequencing images) data as bases and qualities only.

"Although there are computational challenges with such real-time analysis, this processing affords a two-orders-of-magnitude reduction in data needing to be stored, archived, and processed further."

Note: **"third gen" sequencing methods also output in bases and qualities**.

Note2: "Once the physical challenges in storage and access of genomics data are solved, the issues involving the quality and provenance of the derived information will persist." (Scott D. Kahn). These issues would seem to be obviated with implementation of the proposed DASEIN as all "derived information" comes from the AI programming and is stored, "tagged" and fully searchable

= == = = = = = = = = = = = = = =

EXAMPLE OF A DATA SET / DATABASE Amenable to Quantitative and Qualitative Analysis, Exploration, and Experimentation by DASEIN:

PROBLEM/CHALLENGE: Within any complex network of molecules…how do we identify those molecular 'species' that are 'primary' (or most 'critical') in a given metabolic pathway or causal chain of molecular interactions? Identifying said 'critical players' would aid therapeutic drug targeting.

RATIONALE/NEED: "**Pathway analysis"** [see: Seeker Challenge Details]

Example Database/Data set:

RECON 2 - A biomolecular "Google-like" metabolic map that features 7,400 metabolic reactions (accounting for 1800 genes). [Thiele *et al*, 2013; expanded version of Recon 1, Palsson *et al*, 2007, UC San Diego; contact: chockmuth@ucsd.edu]

Reference URL: http://www.eurekalert.org/pub_releases/2013-03/uoc--icb030113.php

QUOTE:

> **"**One of the most promising applications for the network reconstruction is the **ability to identify specific gene expressions and their metabolic pathways for targeted drug delivery. Large gene expression databases are available for human cells that have been treated with molecules extracted from existing drugs as well as drugs that are in development.** Recon 2 allows researchers to use this **existing gene expression data and knowledge of the entire metabolic network to figure how certain drugs would affect specific metabolic pathways found to create the conditions for cancerous cell growth**, for example. They could **then conduct virtual experiments* to see whether the drug can fix the metabolic imbalance causing the disease."**

Note: this type of modeling could be readily incorporated into DASEIN with each 'virtual lab rat'/model separately and

rapidly tested (simulated) for its response to a given (for example) therapeutic molecule (macrocycle, small molecule, mAb, kinase inhibitor, etc.)

QUESTION: (Within any complex network of molecules or cell 'actors')…How do we identify/discover those molecular 'species' that are 'primary' (or most 'critical') in a given metabolic pathway or causal chain of molecular interactions?)

ANSWER: Use of DASEIN to confirm known and discover novel relationships in the various **metabolic pathway networks**.

= =

FROM METADATA to META KNOWLEDGE (with notes relevant to DASEIN)

RATIONALES: **Clinical trial design and management, refined hypotheses, improved experimental procedures, better quality outcomes, higher productivity**

'META KNOWLEDGE'

General Reference: (Special Feature: '**Dealing with Data – Challenges and Opportunities'**, Science, 11 Feb., 2011; article: '**Metaknowledge**' {Evans J.A., Foster J.G., pg. 721)

Article Abstract (underlining/emphasis added by Solver):

The growth of electronic publication and informatics archives makes it possible to harvest vast quantities of knowledge about knowledge, or "**metaknowledge**". We review the expanding scope of metaknowledge (MK) research, which uncovers regularities in scientific claims and infers the beliefs, preferences, research tools, and strategies behind those regularities. MK research also investigates the effect of knowledge context on content. Teams and collaboration networks, institutional prestige, and new technologies all shape the substance and direction of research. We argue that **as MK grows in breadth and quality, it will enable researchers to reshape science – to identify areas in need of reexamination, reweight former certainties, and *point out new paths* that cut across revealed assumptions, heuristics, and disciplinary boundaries**.

Relevant Excerpts / Notes from 'Metaknowledge', *Science* Perspective article by Evans and Foster (underlining/highlighting/annotations by the Solver):

"…consider the perspective that could be gained by a computer trained to extract and systematically analyze information across millions of scientific articles."

Note: **DASEIN extracts metadata from potentially millions of assay experiments**.

"Metaknowledge (MK) results from the critical scrutiny of what is known, how, and by whom. It can now be obtained on large scales, enabled by a concurrent informatics revolution."

Note: **And proposed here: an *automated scientist revolution***

"Using informatics archives* spanning the scientific process, from data and preprints to publications and citations, researchers can now track knowledge claims across topics, tools, outcomes, and institutions."

"Such investigations yield MK about the *explicit* content of science, but also expose *implicit* content, i.e. beliefs, preferences and research strategies that shape the directions, pace, and substance of scientific discovery."

"MK research further explores the interaction of knowledge *content* with knowledge *context*, from features of the scientific system such as multi-institutional collaboration [**Jones, Wuchty, Uzzi, 2008**] to global trends and forces such as the growth of the Internet." [**Evans J.A., 2008**]

"The quantitative study of MK builds on a large and growing corpus of qualitative investigations into the conduct of science…"

"Such investigations reveal the existence of many intriguing processes in the production of scientific knowledge."

Further Pitfalls and Perils: Implicit Preferences, Heuristics, and Assumptions

THEOREM: **Unstated, influencing factors (biases/beliefs, knowledge contexts, social processes, etc.) on the production of scientific knowledge can be revealed through Metaknowledge investigations.**

Additional relevant excerpts from the 'Metaknowledge' article:

(A range of factors) "…from unstated preferences, tastes, beliefs to the social processes of communication and citation."

Example: "The file-drawer problem [i.e., "hiding" of negative findings] is driven by the well-attested ***preference for publishing positive results*** and statistical findings that exceed arbitrary, field-specific thresholds."

'Such preferences may lead to a massive duplication of scientific effort through retesting doomed hypotheses."

(magnifying this effect) "…a trend towards agreement with earlier results, which leads scientists to censor or reinterpret their data to be consistent with the past." (note: this is a form of "**knowledge distortion**")

Example: "Early results that fossilize into facts through a cascade of positive citation, forming "**microparadigms**" (in choosing what parts of past knowledge to certify through positive citation, scientists are likely to accept authors with a history of success more readily" [Rshetsky *et al*, 2006; Shwed, Bearman, 2010] (also noted: a focus on established "**hubs of knowledge**" [Cokol *et al*, 2005]).

(Re: Mitigating the trend towards "assent and convergence") The **"Proteus" phenomenon**: the "rapid alternation in conflicting findings." (noted: this applies to research "staking novel claims" high-profile research leads to "more eye balls" and "reliable negation of these findings attracts considerable interest." (noted: "The incentive to publish in prestigious journals may itself be a distorting preference, potentially leading to a higher incidence of overstated results.") [several studies: Ioannidis, Trikalinos, 2005]

"More powerful methods of NLP and statistical analysis will be essential for revealing subtle content." e.g., "cognitive short cuts, such as the '**availability heuristic'**, in which data and hypotheses are weighted on the basis of how easily they come to mind." [Taversky, Kahneman, 1974]

"Identifying **the distribution of these heuristics across scientific investigations will allow consideration of their consequences, expose possible biases, and recalibrate scientific certainty in particular propositions.**" [Rzhetsky et al, 2006]

(Re: **"ghost theories")** "Subtle but systematic regularities across articles within a scientific domain may signal the presence of [GTs] – unstated theories, assumptions, or disciplinary paradigms that shape the type of reasoning and evidence deemed acceptable." (e.g., human psychological properties are viewed as "universal"; 67% of studies in the *Journal of Personality and Social Psychology* used results from "typical experimental subjects" {i.e., American undergraduates} which were often extended towards the entire human race); A recent **meta-analysis** [Henrich *et al*, 2010] "demonstrated that this assumption is false in several domains (e.g., perception) and recommends expensive changes in sampling to correct for the resulting bias."

> **Note: It is conceivable that most, if not all, of these "unstated preferences and biases" would be eliminated in DASEIN (designed to be "purely" objective).**

Knowledge Context ~ Putting Knowledge In Its Proper Place

Postulate/Theorem: "Reliability of results increases if it is produced in several disparate labs rather than a few linked by shared methods or mentorship"

"The changing organization of research also shapes research content, with teams increasingly producing the most cited research." [Wuchty, Jones, Uzzi, 2007] (note: studies of the **structure of collaborative networks*** reveal intriguing disciplinary differences; also: "research dynasties", how larger institutions influence scientific knowledge [Kevles, 1978; Lenoir, 1997; Coffey, 2008]).

> **Note**: **One would need to design DASEIN – a type of team/ "collaborative" network – with this issue in mind. But it would seem to be a lessor concern with a robotic scientific network**

"Extraction of MK on the distribution of these influences would enable estimation of their aggregate capacity to channel the next generation of research." (re: the importance of understanding the relationship between social and scientific structures, e.g., "chemists and biologists whose collaborations bridge scientific subgroups tend to investigate reactions that themselves bridge distinct clusters of molecules"; Need: "…explore the degree to which chemical structure shapes social structure, and vice versa.").

(Re: **"knowledge distortion"** [Allesina, 2009; Thurner, Hanel, 2010]) "Universities, institutes and companies vary in prestige, access to resources, and cultures of scientific practice. Institutional reputations likely color the acceptance of research findings." (note: "research on multi-institutional collaboration demonstrates that Institutes tend to collaborate with others of similar prestige, potentially exacerbating this effect." (i.e., **knowledge distortion**; also noted: the "influence of shared resources -- databases, accelerators, telescopes -- on the **organization of related research and the pace of advance**."). [Jones B.F. *et al*, 2008]

> **Note: Seems relevant to DASEIN; "organization: and "pace of advance"**

[On 'Breakthrough Investment' - The Role of the Entity ~ The 'Borg' Factor]

QUESTIONS (quoted from the 'Metaknowledge' article): "Can focused investment unleash sudden breakthroughs, or is it the slow development of community-shared culture and a toolkit more important to nurture a flow of discoveries?"

> **Note**: **"slow development of a community-shared culture" would seem to be a challenge to the nature of a DASEIN-type system. Is there any scientific or humanistic ethic at stake here?**

"There is evidence from MK research that embedding research in the private or public sector modulates its path. Company projects tend to eschew dogma in an impatient hunt for commercial breakthroughs, **leading to rapid but unsystematic accumulation of knowledge**, whereas public research focuses on the careful accumulation of consistent results."[Evans J.A., 2010]

Note: this critique of "unsystematic accumulation of knowledge" would not seem to be an issue with an automated scientist (or network of same) which is designed to be systematic in everything it does.

Why MK? Why Now?

"The ecology* of modern scientific knowledge constitutes a complex system; apparently complicated, involving strong interaction between components and predisposed to unexpected collective outcomes." [Foote R., 2007] (noted: "the growing number of 'global scientists' increasingly connected via multiple channels, international conferences, online publications, email, and science blogs has increased this complexity. Rising complexity in turn makes the changing focus of research and the resolution of consensus less predictable."

"The informatics turn in the sciences offers a unique opportunity to mine knowledge for metaknowledge in order to identify and measure the complex processes of knowledge production and consumption." (also noted: "As such, MK research provides a high-throughput complement to existing work in social and historical studies of science by tracing the distribution and relative influence of distinct social, behavioral and cognitive processes on science.")

(Caveat)

"MK investigations will miss subtle regularities accessible to deep, interpretive analysis, an should draw on such work for direction."

However, the authors argue: "…that some regularities will only be identifiable in the aggregate, especially those involving interrelations between competing processes. Once identified, those could become fruitful subjects for interpretive investigation."

(Re: successfully executing a MK program) Need: "…**further improvements in machine reading and inference technologies. Systematic analysis of some elements of scientific production will remain out of reach."**

"As MK grows in sophistication and reliability, it will provide new opportunities to recursively shape science – to use measured biases, revealed assumptions, and previously unconsidered research paths to revise our confidence in bodies of knowledge and particular claims, and to suggest novel hypotheses."

"The computational production and consumption of MK will allow researchers and policy-makers to leverage more scientific knowledge – explicit, implicit, contextual – in their efforts to advance science. This will become essential in an era when so many investigations are linked in so many ways."

Final Notes: DASEIN would go far in minimizing this "systematic analysis" that "remains out of reach"…and will aid in revealing "previously

unconsidered research" and expand the ability of scientific research "to suggest novel hypotheses" [see: Adam]

= =

POSSIBLE TECHNOLOGY INTEGRATIONS WITH THE DASEIN PLATFORM:

Technology: Multiplex Automated Genome Engineering (MAGE)

Description: Multiplex automated genome engineering (MAGE) is a technology/technique_for "large-scale programming and evolution of cells. MAGE simultaneously targets many locations on the chromosome for modification in a single cell or across a population of cells, thus producing combinatorial genomic diversity. Because the process is cyclical and scalable, we constructed prototype devices that automate the MAGE technology to facilitate rapid and continuous generation of a diverse set of genetic changes (mismatches, insertions, deletions)."

Key Features (re: Cell/Genome Variants and Drug Identification):

As many as 15 billion genetic variants (4.33 x 10^8 bp variations per cycle for 35 MAGE cycles) were generated using the MAGE system.

MAGE expedites the design and evolution of organisms with new and improved properties.

MAGE accelerates the rate of [mutation] accumulation in any individual cell, thus increasing the likelihood of finding sets of mutations that may interact synergistically to produce a surprisingly beneficial phenotype.

Rationale/Relevant Notes:

[excerpts from the paper, citation below]

"…genomic diversity is difficult to generate in the laboratory and new phenotypes do not easily arise on practical timescales. Although in vitro and directed evolution methods have created genetic variants with usefully altered phenotypes, these methods are limited to laborious and serial manipulation of single genes and are not used for parallel and continuous directed evolution of gene networks or genomes."

"We **isolated variants with more than fivefold increase in lycopene production within 3 days,** a significant improvement over existing metabolic engineering techniques**. Our multiplex approach embraces engineering in the context of evolution by expediting the design and evolution of organisms with new and improved properties."**

"With the advent of next-generation fluorescent DNA sequencing, our ability to sequence genomes has greatly outpaced our ability to modify genomes. **Existing cloning-based technologies are confined to serial and inefficient introduction of single DNA constructs into cells, requiring laborious and outdated genetic engineering techniques. Whereas in vivo methods such as recombination-based genetic engineering (recombineering) have enabled efficient modification of single genetic targets using single-stranded DNA (ssDNA)"**

"…no such attempts have been made to modify genomes on a large and parallel scale. **MAGE provides a highly efficient, inexpensive and automated solution to simultaneously modify many genomic locations (for example, genes, regulatory regions) across different length scales, from the nucleotide to the genome level."**

"MAGE is also an accelerated evolution platform that permits the repeated introduction and maintenance of many neutral (or deleterious) mutations in the cell population. Although these mutations would normally disappear in the population via genetic drift or natural selection, MAGE accelerates the rate of their accumulation in any individual cell, thus increasing the likelihood of finding sets of mutations that may interact synergistically to produce a surprisingly beneficial phenotype."

Reference/Citation:

Programming cells by multiplex genome engineering and accelerated evolution (http://arep.med.harvard.edu/pdf/Wang.Isaacs.Nature09.pdf) (Harris H. Wang, Farren J. Isaacs1, Peter A. Carr, Zachary Z. Sun, George Xu, Craig R. Forest & George M. Church) (2009).

SEE ALSO: newer/augmented versions of MAGE: COsMAGE, CAGE

- -

Technology: RNA-induced Pluripotent Stem Cells (RiPS)

Description: A recent improvement ("major advance") over established iPS technology; the RiPS technique utilizes short segments of messenger RNA (mRNA) to target the reprogramming genes of the target cell(s)

Key Features (re: Cell models / Drug Discovery):

The (RiPS) technique cuts the reprogramming time in half (e.g., compared to viral iPS cells; see iPS Cell chart for different delivery types, Endnotes, pg. 17), and increases the yield of RiPS cells by almost 100 times...RiPS cells are a much closer match to their source cells than regular iPS cells.

> **Note: one may thus presume here that RiPS cells have the same/greater potential (e.g., for patient-specific disease modelling) as standard non-RNA iPS cells.**

Rationale/Further Notes:

[full paper summary; highlighting added]

> Clinical application of induced pluripotent stem cells (iPSCs) is limited by the low efficiency of iPSC derivation and the fact that most protocols modify the genome to effect cellular reprogramming. Moreover, safe and effective means of directing the fate of patient-specific iPSCs toward clinically useful cell types are lacking. Here we describe a simple, non-integrating strategy for reprogramming cell fate based on administration of synthetic mRNA modified to overcome innate antiviral responses. We show that this approach can reprogram multiple human cell types to pluripotency with efficiencies that greatly surpass established protocols. We further show that the same technology can be used to efficiently direct the differentiation of RNA-induced pluripotent stem cells (RiPSCs) into terminally differentiated myogenic cells. **This technology represents a safe, efficient strategy for somatic cell reprogramming and directing cell fate that has broad applicability for basic research, disease modeling**, and regenerative medicine.
>
> Further Notes [from on-line news article, plus Solver notes]:
>
> > The advantage to this advanced technique is that the RNA used to induce pluripotency is quickly broken down and eliminated within the new cells**.**
> >
> > **Note: This feature of RiPS cells (i.e., utilization of RNA) may decrease or eliminate many (future) failed drug trials which may result from the use of iPS cells that contain traces of the parent cell DNA and thus may generate (in vitro or in animal models) misleading results or effects -- so-called "insertional" defects or artefacts.**
>
> The researchers plan on exploring ways to use RiPS cells to replace proteins in patients with various genetic disorders.
>
> > **Note2: If the RiPS technique can be used to generate functional proteins as replacements for anomalous, unstructured, and/or**

misfolded proteins (due to the stated genetic disorder, i.e., mutation), then these are potentially the bases of new pharmaceutical agents.

Reference/Citation:

Highly Efficient Reprogramming to Pluripotency and Directed Differentiation of Human Cells with Synthetic Modified mRNA. [http://www.cell.com/cell-stem-cell/abstract/S1934-5909%2810%2900434-0] *Cell Stem Cell* (Rossi, D. et al); Volume 7, Issue 5, 618-630, 30 September 2010.

- - -- -

(RADseq) – (Cresko W., Johnson E., 2008/9; Univ. of Oregon) [see **pg. 57** for fuller description of RADseq, excerpts]

[reference: **'Using DNA to Reveal a Mosquito's History'** / Science, 25 February, 2011, pg. **1006]**

RADseq is a 'tagging system' developed to catalog SNPs in stickleback fish; **a short cut SNP-discovery method (restriction site-associated DNA sequencing**) which exploits low-cost **next gen sequencing tech to quickly generate thousands of SNPs that distinguish populations and individuals.**

(The technique was tested on) animal samples from multiple populations, and **uses restriction enzymes to chop up the genomes into short fragments**. Each fragment is joined to a unique barcode: a synthetic 5-base strand of DNA; the sequence reveals which animal the non-bar-code DNA came from. **All fragments are pooled together for mass processing (next gen seq.)**. Because barcodes allow resulting sequences to be associated with individual animals…*bio-informatics software* can **quickly identify genetic differences among individuals and populations.**]

RADseq also used to identify mosquito SNPs (Cresko, Johnson, 2010, PNAS ("cheaper, faster and delivers thousands of markers"). There are numerous SNP projects underway for many animals; example: diamondback moth (Blaxter et al, Univ. of Edinburgh), a crop pest; **RADseq was used to identify a gene that makes the moth resistant to a certain insecticide** (a "transformative technology").

= =

One-Step Cloning and Chromosomal Integration of DNA – "clonetegration"

What: pOSIP - simultaneous cloning and integration; technique treat chromosomes as large cloning vectors (technique combines the integrase-expressing and integration plasmids into a single vector)

François St-Pierre †‡http://pubs.acs.org/doi/abs/10.1021/sb400021j - notes-1, Lun Cui §http://pubs.acs.org/doi/abs/10.1021/sb400021j - notes-1, David G. Priest §, Drew Endy †, Ian B. Dodd §, and Keith E. Shearwin *§

†Department of Bioengineering and ‡Department of Pediatrics, Stanford University, California 94305, United States

§ Department of Biochemistry, School of Molecular and Biomedical Sciences, The University of Adelaide, SA 5005, Australia

ACS Synth. Biol., Article ASAP
DOI: 10.1021/sb400021j
Publication Date (Web): May 6, 2013

= =

ADDITIONAL AUTOMATABLE TECHNIQUES for GENE SEQUENCING AMENABLE TO INTEGRATION with DASEIN:

MIP – Molecular Inversion Probe – "a strategy with novel algorithms for MIP design; an **optimized automatable work flow** with robust performance and minimal DNA input; **extensive multiplexing of samples** while sequencing and reagent costs of less than $1 per gene per sample."

Re: contribution of rare variants and de novo mutations to the genetic basis of complex phenotypes. Quote: "Because of extreme genetic heterogeneity, the sample sizes required to implicate any single gene in a complex phenotype are extremely large."

> Reference/citation:
> 'Multiplex Targeted Sequencing Identifies Recurrently Mutated Genes in Autism Spectrum Disorder' (O'Roak *et al*) Science, 21 Dec., 2012; pg. 1619 [MIP schematic, pg. 1620, fig. 1A]
>
> **Note: DASEIN integrated with MAGE + MIP could generate and manage the large sample sizes (and their assays). DASEIN (with its AI software) would seem amenable to handling such large sample sizes, efficiently and with great precision, and by intelligently assigning and allocating data set processing (compression, etc.) analysis and storage to designated/dedicated servers.**

Additional:

Principle: (in a given cell population) Genomic heterogeneity drives evolution, and cancer.

Need: A method of single cell (whole genome) sequencing

Technique: **MAL BAC** (Zong *et al*, *Science*, 21 Dec. 2012; pg. 1622) – Multiple Annealing and Looping-Based Amplification Cycles

> **Note**: **such "cycles" could be programmed into any given ADAM (assuming that the necessary hardware was installed); see: Adam "internal cycles"**

= =

AFFINITY MATURATION (of ANTIBODIES):

BMC Bioinformatics. 2013 May 30;14(1):168. [Epub ahead of print]

MAPs: a database of modular antibody parts for predicting tertiary structures and designing affinity matured antibodies.

Pantazes RJ, Maranas CD.

Abstract

BACKGROUND: The de novo design of a novel protein with a particular function remains a formidable challenge with only isolated and hard-to-repeat successes to date. Due to their many structurally conserved features, antibodies are a family of proteins amenable to predictable rational design. Design algorithms must consider the structural diversity of possible naturally occurring antibodies. The human immune system samples this design space (2 1012) by randomly combining variable, diversity, and joining genes in a process known as **V-(D)-J recombination**.

DESCRIPTION: By analyzing structural features found in affinity matured antibodies, a **database of Modular Antibody Parts (MAPs) analogous to the variable, diversity, and joining genes has been constructed for the prediction of antibody tertiary structures.** The database contains 929 parts constructed from an analysis of 1168 human, humanized, chimeric, and mouse antibody structures and **encompasses all currently observed structural diversity of antibodies**.

CONCLUSIONS: The generation of 260 antibody structures shows that **the MAPs database can be used to reliably predict antibody tertiary structures** with an average all-atom RMSD of 1.9 A. Using the **broadly neutralizing anti-influenza antibody CH65** and **anti-HIV antibody 4E10** as examples, **promising starting antibodies for affinity maturation are identified and amino acid changes are traced as antibody affinity maturation occurs.**

PMID: 23718826 [PubMed - as supplied by publisher]

Note: **This is an important and highly useful resource for antibody research and <u>one that readily integrate with DASEIN's AI component and the Hadoop architecture.</u>**

Note2: One could also integrate data mining methodologies and algorithms into the MADAM/AI central brain (e.g., maximal information co-efficient, MIC, see: 'Detecting Novel Associations in Large Data Sets' [see: earlier citation: Reshef *et al*]) to explore any (DASEIN derived or accessed) database to reveal <u>new</u> functional (linear and non-linear) relationships and thereby form new hypotheses for testing, and confirm known relationships (validate experimental results).

~~~~~~
~~~~~~