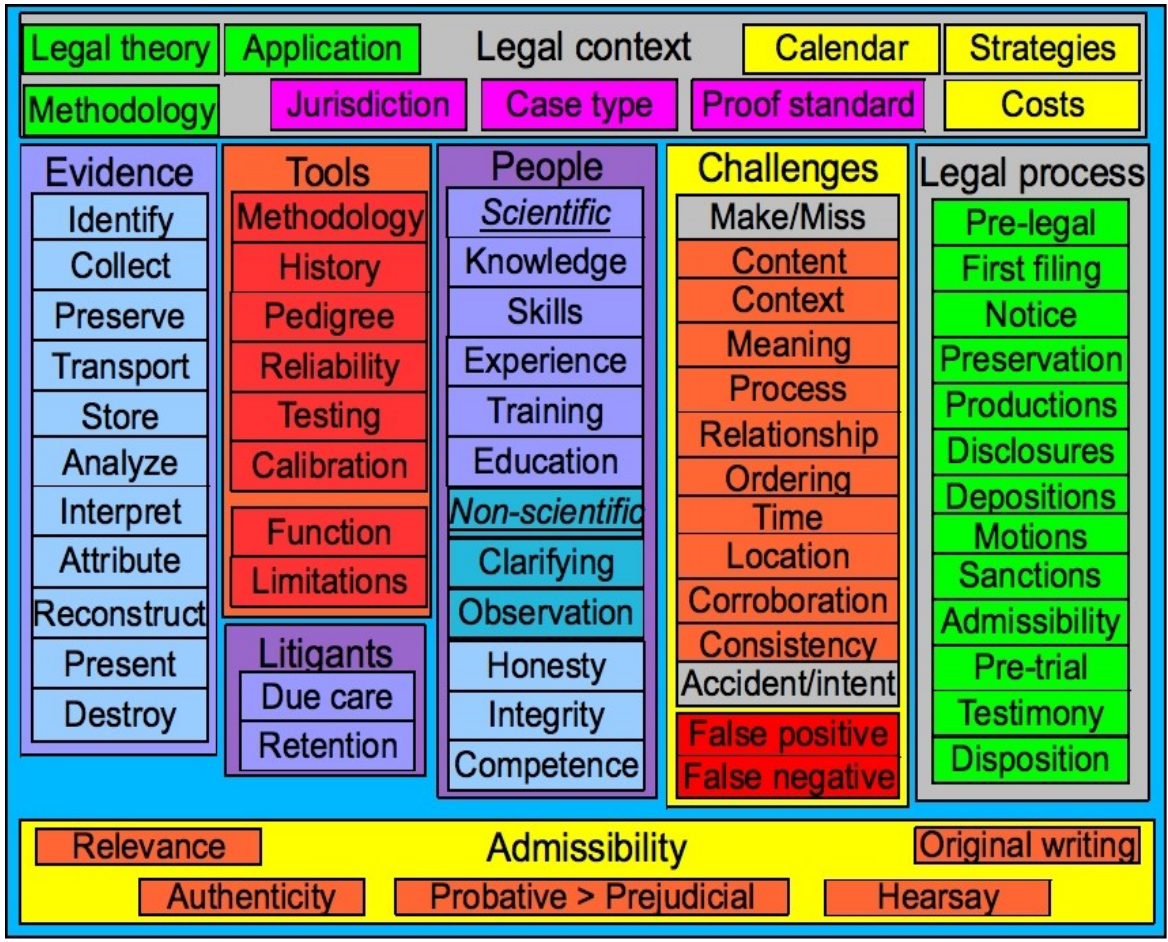
Digital Forensic Evidence Examination
The State of the Science and Where to Go From Here

NEFX – Sep 13, 2010

Dr. Fred Cohen
President - California Sciences Institute
CEO – Fred Cohen & Associates

Outline

- Introduction
- Epistemology?
- Theory?
- Methodology?
- Experimental basis?
- Physics?
- Where do we agree?



Your speaker

- Education:
  - B.S. Electrical Engineering (C-MU '77)
  - M.S. Information Science (Pitt '81)
  - Ph.D. Electrical Engineering (USC '86)
- Experience:
  - >30 years of information protection R&D, design, engineering, testing, implementation, and operation
  - >20 years since first digital forensics case
- CEO - Fred Cohen & Associates
  - Enterprise information protection architecture
  - Digital forensics for high-valued legal cases

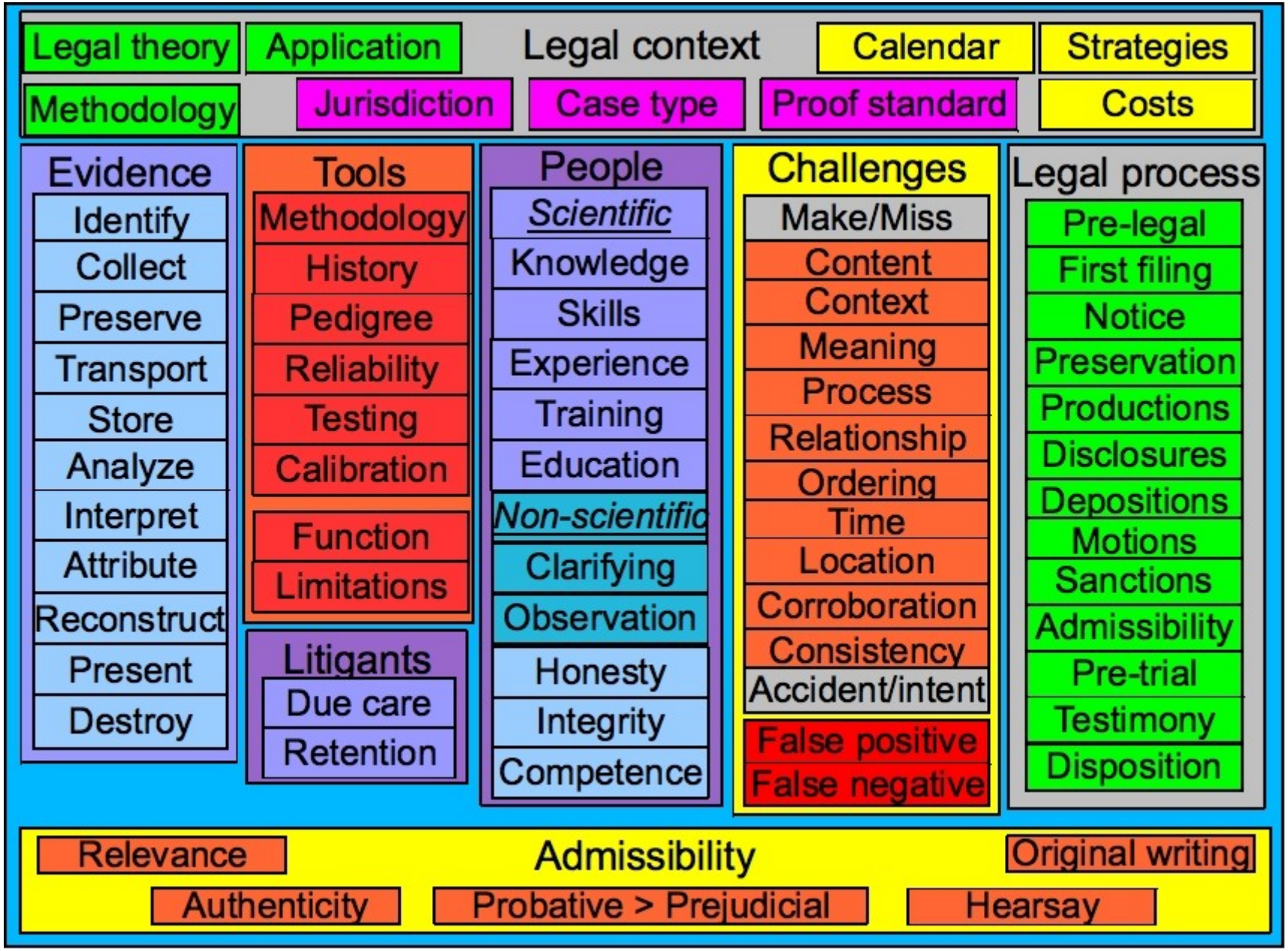
CalSci

- President – California Sciences Institute
  - Started doctoral classes in 2010-07
- M.S. And Ph.D. Program in National Security
  - Technical aspects of these fields
- M.S. In Advanced Investigation
- Ph.D. In Digital Forensics
  - The first Ph.D. program in Digital Forensics in the United States
- calsci.org

The challenge

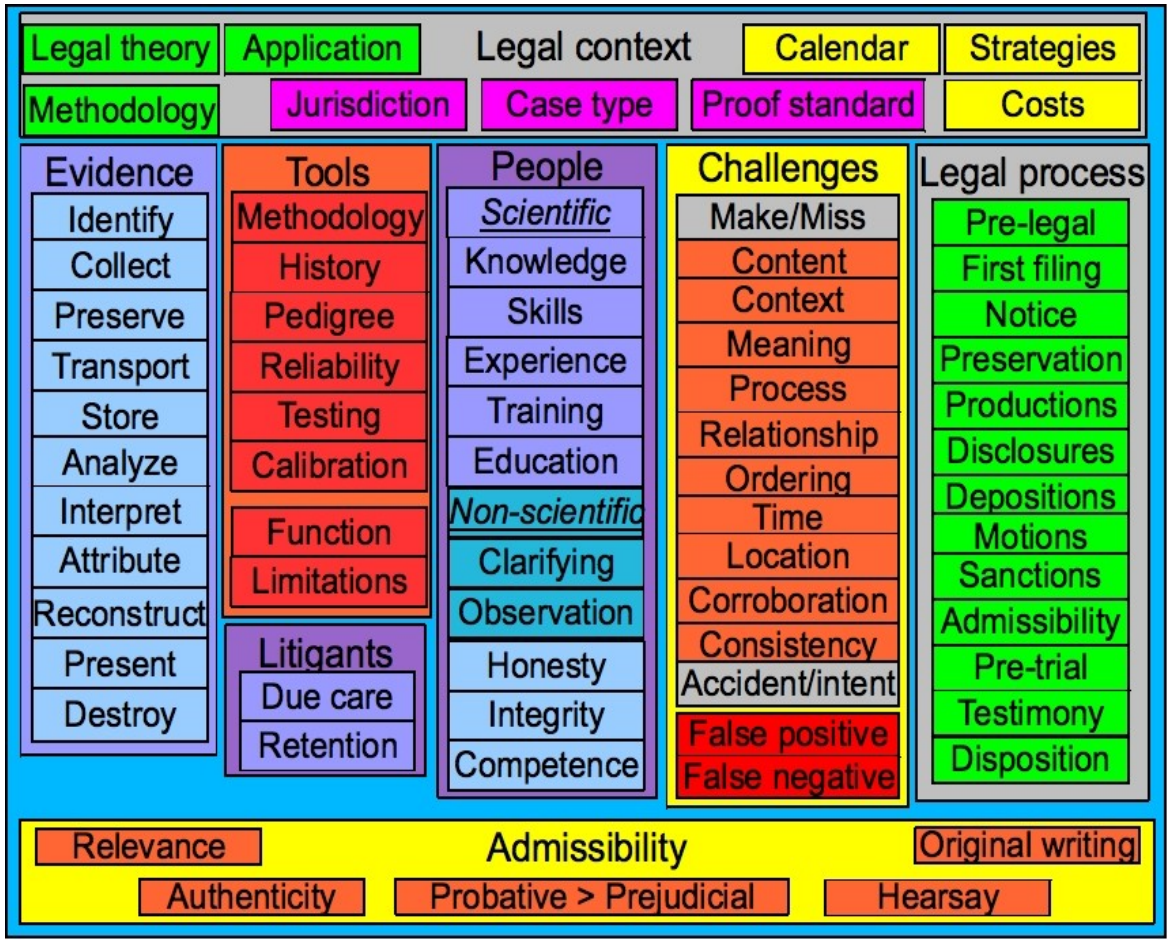
- Digital Forensic Evidence (DFE) Examination
  - Building a consensus in the scientific community
    - A common body of knowledge
    - Well defined and consistently used terms
    - A well understood epistemology, theory, and methodology
    - A strong experimental basis
    - An agreed upon physics
  - This review
    - Identifies select elements that I think should meet that set of requirements
    - Asks for your views – consensus or not?

Observation: the picture today



Outline

- Introduction
- Epistemology?
- Theory?
- Methodology?
- Experimental basis?
- Physics?
- Where do we agree?



Epistemology

- The branch of philosophy that studies the nature of knowledge, its presuppositions and foundations, and its extent and validity.
- In the case of the science of digital forensic evidence examination:
  - Digital evidence is entirely sequences of bits.
  - Physics different than matter and energy.
  - Finite (fairly small) granularity in space and time.
  - Observation without alteration.
  - Duplication without removal.

Epistemology 2

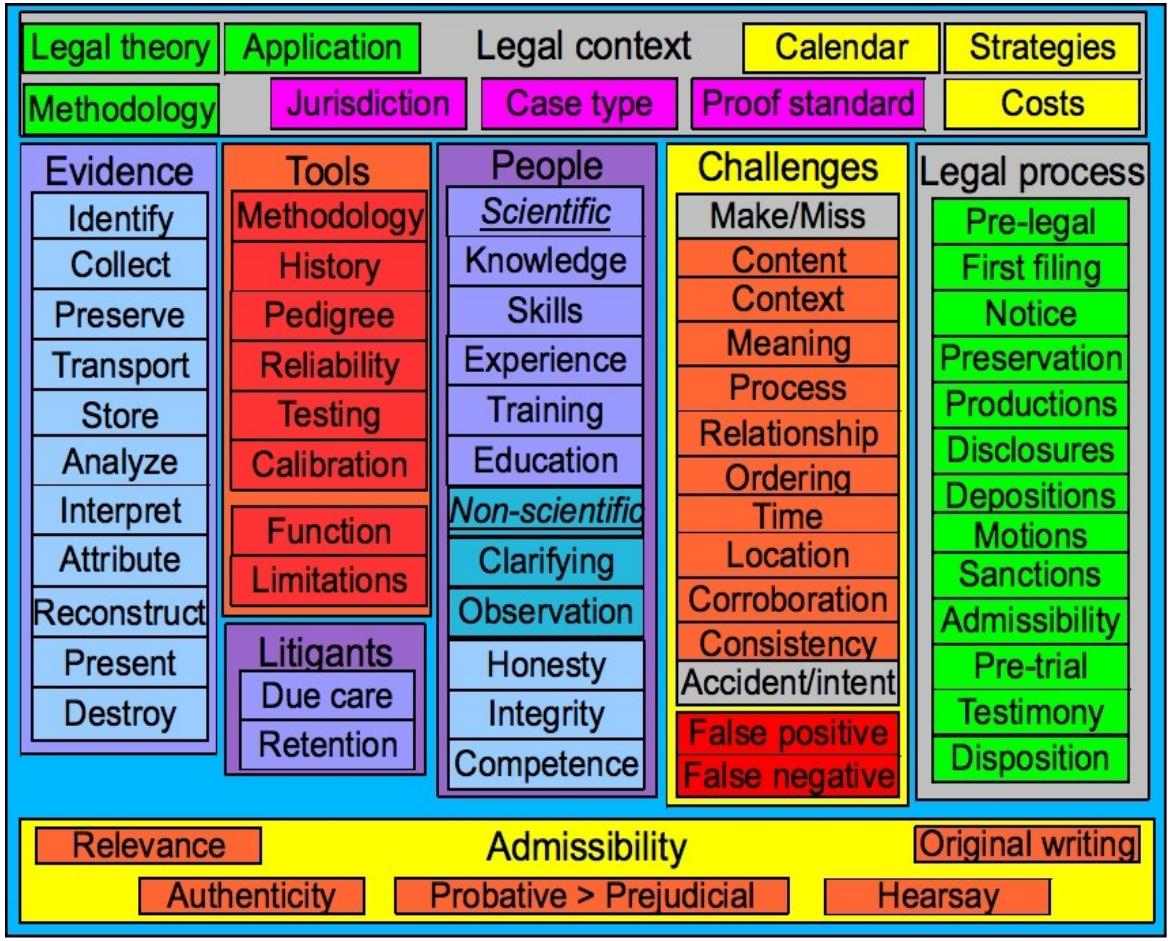
- DFE is trace, but not transfer.
  - Traces produced by the execution of FSMs.
- FSMs have properties that define a physics.
  - Finite granularity implies limits on accuracy and precision based on representation.
  - FSMs are syntactic in nature so semantics is driven entirely by context.
- DFE is normally latent in nature.
  - Can only be observed through use of tools.

Epistemology 3

- There are fundamental limits on what can be done.
  - Computational complexity is like the speed of light in DFE examination.
  - DFE can never directly speak to the physical world except in limiting what FSMs can do.
  - At the edge between digital and physical systems there are assumptions.

Outline

- Introduction
- Epistemology?
- Theory?
- Methodology?
- Experimental basis?
- Physics?
- Where do we agree?



Theory

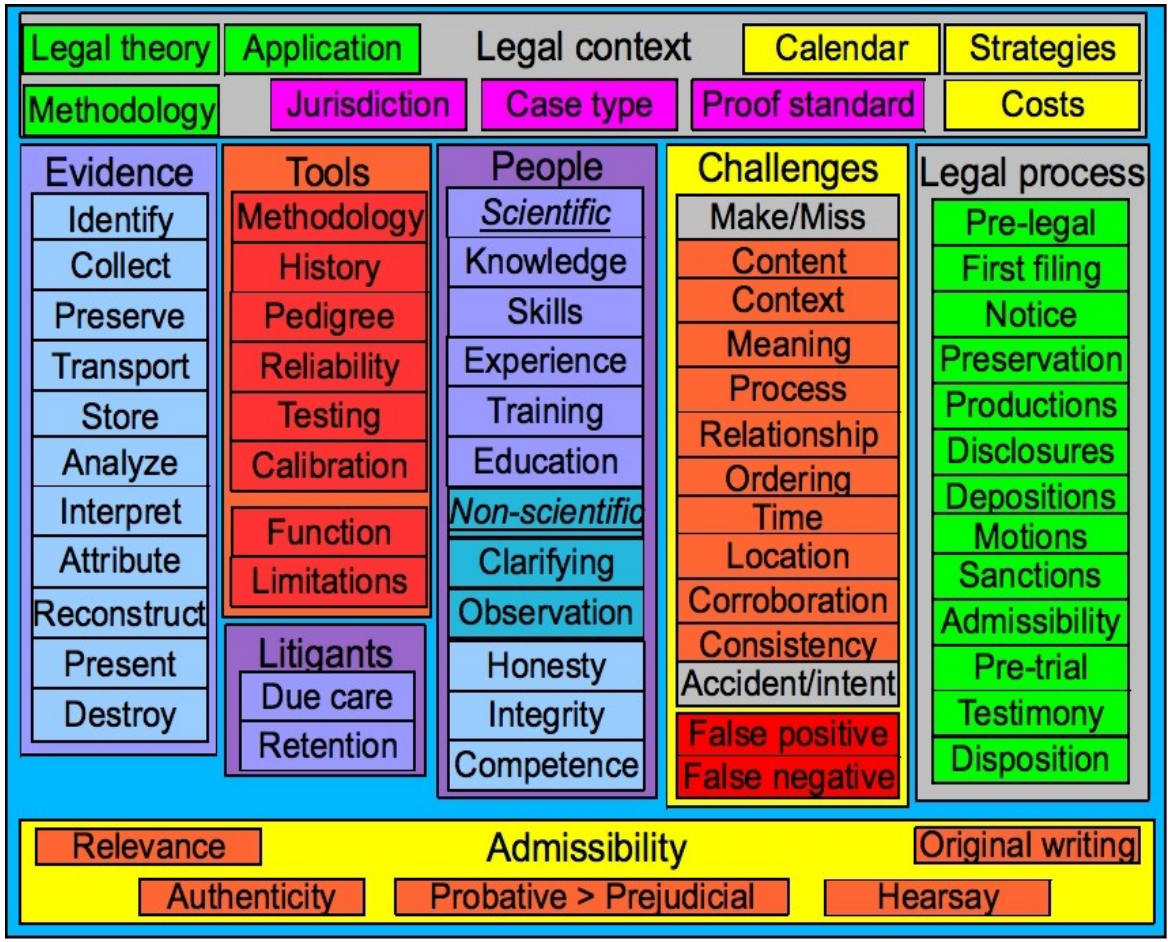
- Scientific theories are not casual theories.
  - They are constructs that are testable by nature.
  - Refutation can destroy a theory, but confirmation cannot prove it except in finite cases.
  - Scientific theories change slowly, and normally, once accepted, only change because of dramatic changes in underlying understanding of physics
  - Those changes are normally only related to special or rarely seen cases.
  - Theories are different than hypotheses, which come up all the time, on a case-by-case basis.

DFE theories

- Theories in DFE examination.
  - Form a physics of information.
  - Many of them are based on mathematical results that have long been widely accepted.
  - Some of them are still conjectures, that may be proved or disproved with time.
- Most such theories stem from computer engineering, computer science, finite mathematics, and related fields.
- Many such theories lead substantially limit what can be truly stated about DFE.

Outline

- Introduction
- Epistemology?
- Theory?
- Methodology?
- Experimental basis?
- Physics?
- Where do we agree?



A standard model

- We generally interpret theory in terms of a model -- I will call it “the standard model”
  - But it's hardly standard at this point in time
- The standard model assumes laws, a judicial system with various standards
  - These are called “the legal environment” (L,R,V)
- Claims made by parties, documents, statements, and a wide variety of other non-digital information, and hypotheses are made by examiners
  - These are called “events” (E)

The standard model 2

- There is a wide variety of digital forensic evidence, typically in the form of sequences of bits
  - These are called “traces” (T)
- The DFE examiner identifies consistencies and inconsistencies
  - Between and within traces (TxT)
  - Between traces and events (TxE)
- To do this, the examiner uses forensic methods
  - These are called “procedures” (P)

The standard model 3

- Examiners work within constraints
  - There are limits on available resources (R)
  - There is an ever changing schedule (S)
- There are various implications of this model
  - The sizes of the model components
  - Available computing power and its implication on thoroughness
  - Limitations due to resources and schedule
  - Limits of currently available procedures
  - Legal limitations on what can be used, how, when, and probative versus prejudicial value

Scientific methodology in DFE

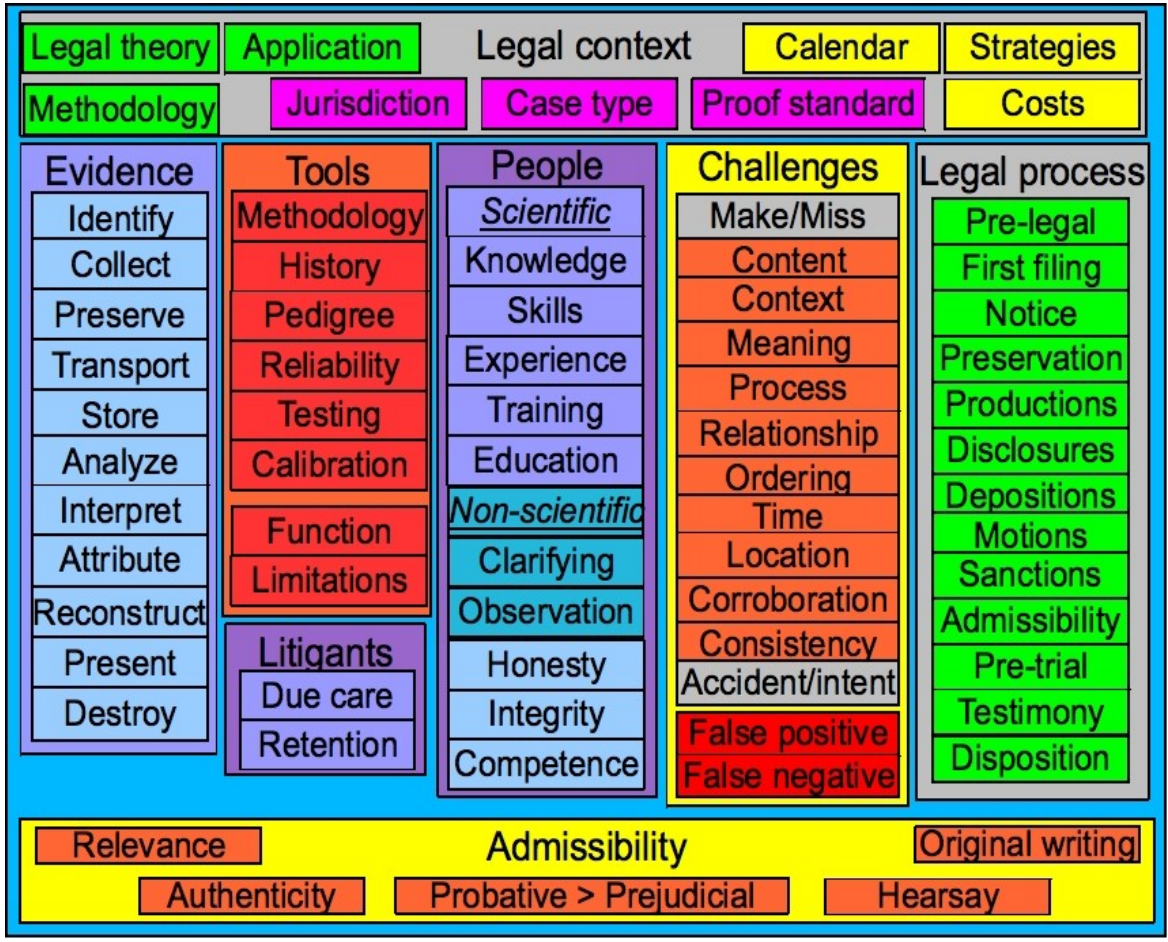
- The fundamental theorem of DFE examination:
  - **What is inconsistent is not true**
- DFE examination consists of testing hypotheses to try to refute them.
  - No matter how many tests are performed, except for special cases, ***you can't prove that anything is true***.
  - The ***best*** you can do, is show that your ***tests failed to refute*** the ***hypotheses*** at issue.
  - The ***most*** you can say (in proof) is that the ***results*** of the tests you did were ***consistent with*** some set of ***hypotheses***.

Refutation is king

- On the other hand...
  - One refutation disproves a hypothesis.
  - The ***least*** you can say based on refutation is that the ***hypothesis is not true***.
- Thus the methodology consists of:
  - Devise testable hypotheses (A ***consistent*** with B)
  - Test those hypotheses against the evidence
    - A scientific test should seek to refute a hypothesis and not to confirm it (seek ***inconsistency***)
  - Inductive and deductive logic are valuable tools for testing hypotheses
  - As is experimental technique

Outline

- Introduction
- Epistemology?
- Theory?
- Methodology?
- Experimental basis?
- Physics?
- Where do we agree?



The experimental basis is limited

- As an area of science, DFE has a relatively small number of peer reviewed and repeated scientific experiments.
  - The total corpus is <500 serious papers.
  - Most of these have very limited applicability.
  - Most not focused on fundamental understanding.
  - Most experiments don't meet the standards of scientific rigor typical of other fields.
  - Most experiments are oriented toward confirmation rather than refutation, which makes them scientifically dubious at best.

Experiments and tools

- DFE is latent, therefore
  - Experiments require tools
- Experiments are limited by the tools, therefore
  - We need to understand the limits of the tools to understand the limits of the experiments.
- We need a methodology to evaluate tools
  - Without a methodology, regardless of what the tools tell us, we don't know how to interpret it.
- What's involved in this methodology?

Tools must be...

- We must understand the nature of errors made by tools.
  - To do this, we need an error model.
- We must understand how to calibrate tools, how to test tools, and create a systematic approach to doing so.
  - The calibration process typically involves validation with known samples.
  - The testing process typically involves verification of the software, which normally involves mathematical proofs combined with tests that exploit the error models.

Tool interpretation

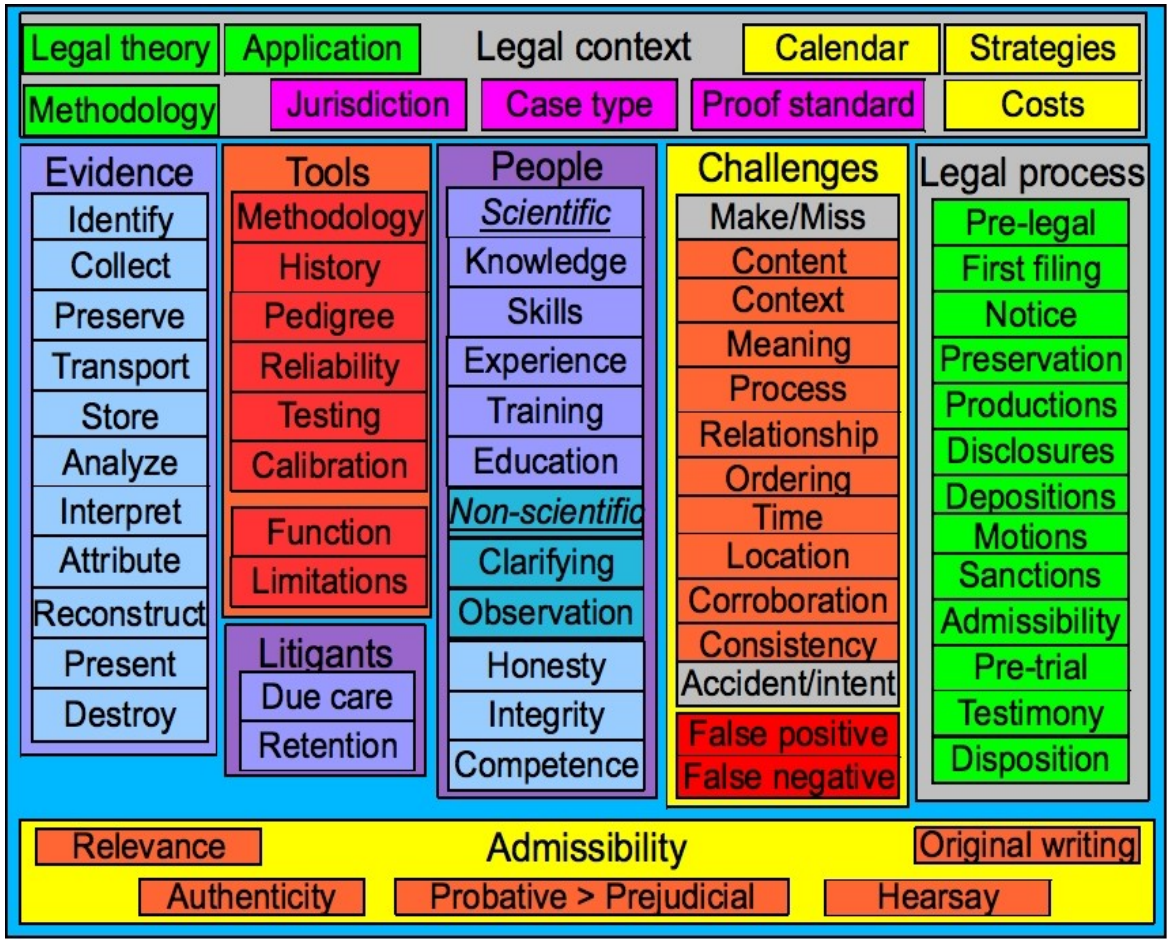
- Regardless of how "good" the tool is:
  - It must be properly used
  - The results must be meaningfully interpreted
  - The limits of the tools must be understood
- This implies expertise by the examiner:
  - Knowledge
  - Skills
  - Experience
  - Training
  - Education

Note the need for a theory of measurement and its application in the context of tool usage...

- What does the ruler measure?
- Do I need the same ruler to test it?
- Can I use the same ruler to test it?
- Can I use a tool that doesn't reveal the mechanisms producing its outputs?

Outline

- Introduction
- Epistemology?
- Theory?
- Methodology?
- Experimental basis?
- Physics?
- Where do we agree?



Information physics examples

- Digital space converges with time
  - FSM: (I,O,S,m:{IxS}→{O,S'}) IF |I|>(|O|+|S|) THEN ∃(i,i')∈I:∃(o)∈O,∃(s)∈S, i⟶(o,s) and i'⟶(o,s)
  - Also note that $h(O) \leq h(I+S)$ (Shannon's h)
  - *Energy and matter space diverges with time (2nd law of thermodynamics)*
  - *Digital space converges with time*
- You can't normally identify I^n from traces T
  - T: $|T|<|I^n|$, $\exists(i,i')\in I^n:\exists(t)\in T$, i⟶(t) and i'⟶(t)
  - In digital space, history is not uniquely determined by the present

Information physics examples

- Time is a partial ordering
  - FSM outputs are strictly sequential as sets but...
    - Traces as recorded are subject to Δt
    - When multiple FSMs are present, A≈B may apply
    - Trace time stamps subject to delays, etc.
- Time directional asymmetry
  - Given {IxS}, {O,S'} are unique and known
  - But... given {O,S'}, {IxS} are (usually) non-unique
- Given T, you cannot uniquely derive the FSM
  - $\forall n$, $M^n:(I^n,O^n,S^n,m^n:\{I^n x S^n\}\rightarrow\{O^n,S^{n\prime}\})$, $\exists$ infinite k: $M^k \approx^o M^n$ (by construction, add redundant states).

Information physics examples

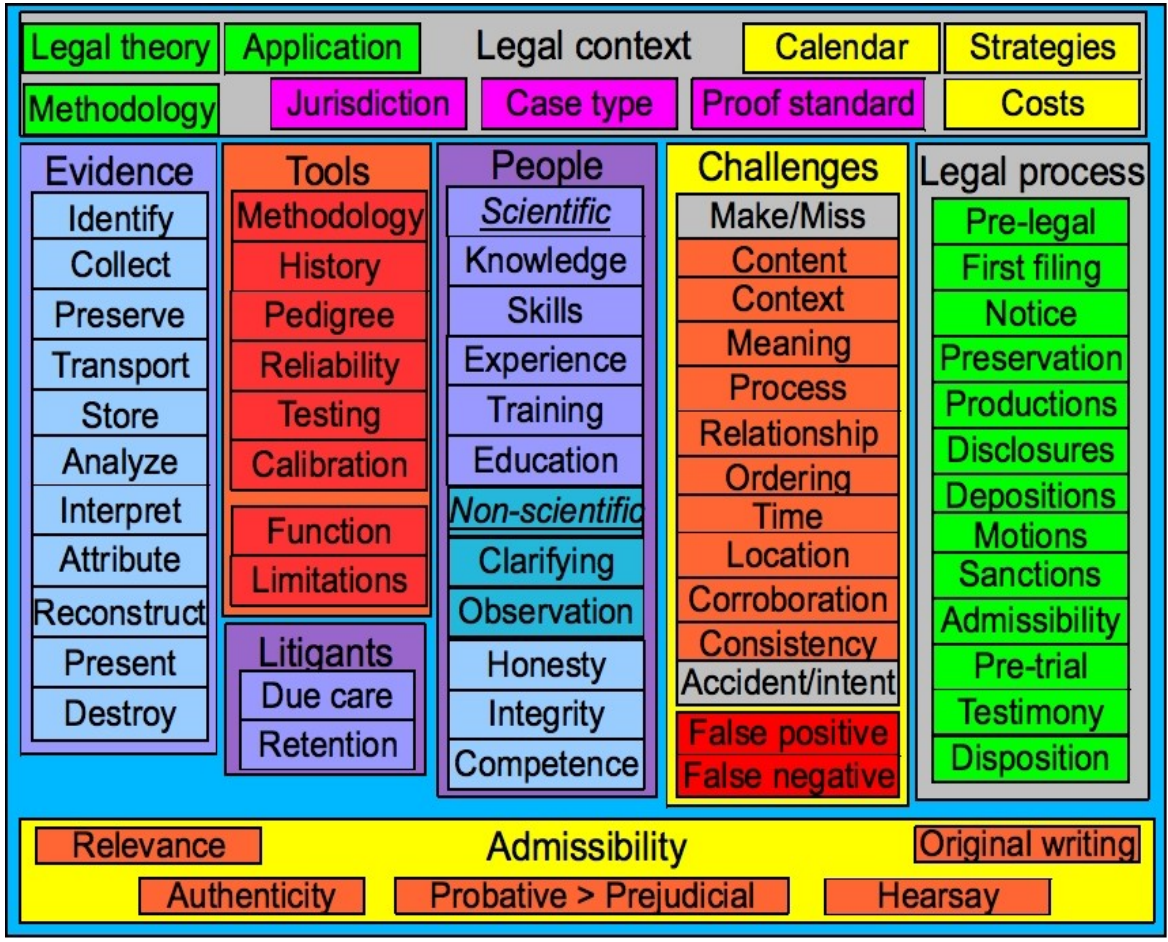
- Thorough examination, in a sense of looking at all possibilities, is almost never feasible
  - 100 instruct. (i) @ 10^9 i/s → $10^{10^{18}}$ 1s i-sequences
- Time and space are discontinuous
  - Finite granularity space (bits) and time (clocks)
- I/O not repeatable across the D-A-D interface
  - Experiment: Print a JPG, scan it back, compare
  - Experiment: Scan the same page multiple times
- Precision is always necessarily finite
  - (/ 1 3) is precise, but try for Pi

Information physics examples

- Hash functions and their limitations (lossy)
  - $|I|=2^n$, $|O|=2^m$, $n>m \rightarrow \exists i,i' \in I, o \in O \;\; i \rightarrow o, \;\; i' \rightarrow o$
  - and $\exists o \in O, \;\; \exists \ddot{I} \subset I, \;\; |\ddot{I}| \geqslant 2^{n-m}$
- Regular physics still applies in most cases…
  - A signal traveling from San Diego to San Francisco cannot get there in less than so many milliseconds
- Computational complexity ≈ c in computers
  - After it got there, it has to perform a computation, and that also takes time!

Outline

- Introduction
- Epistemology?
- Theory?
- Methodology?
- Experimental basis?
- Physics?
- Where do we agree?



Thank You

http://calsci.org/ - calsci at calsci.org
http://all.net/ - fc at all.net