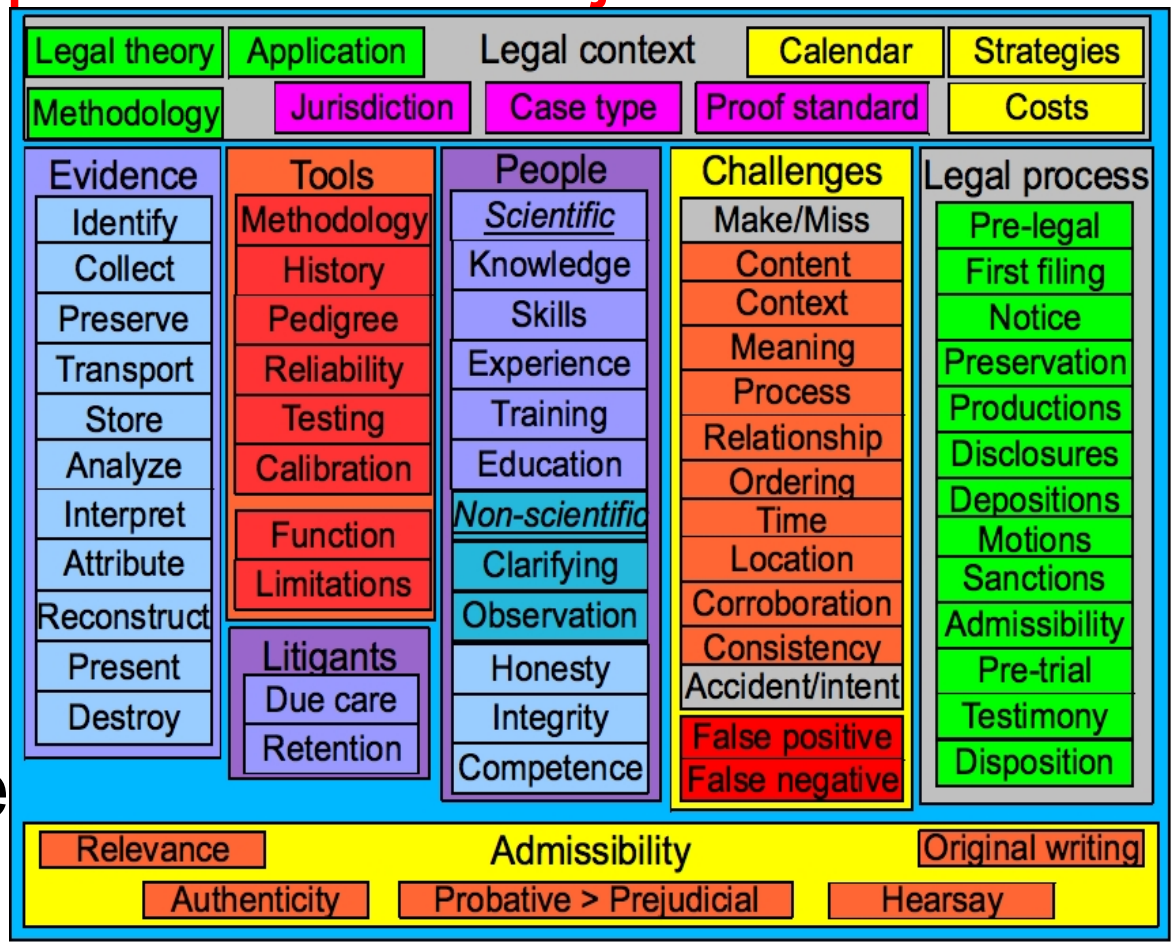
The Science of DFE Examination

IFIP Digital Forensics Conference – Jan 4, 2010

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

Outline

- Background of the speaker and subject
- Definitions
- Epistemology
- Theory
- Methodology
- Experimental basis
- State-of-the-science
- Your turn!



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
  - Starting doctoral classes in 2009-10
- 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

What does he know about the subject?

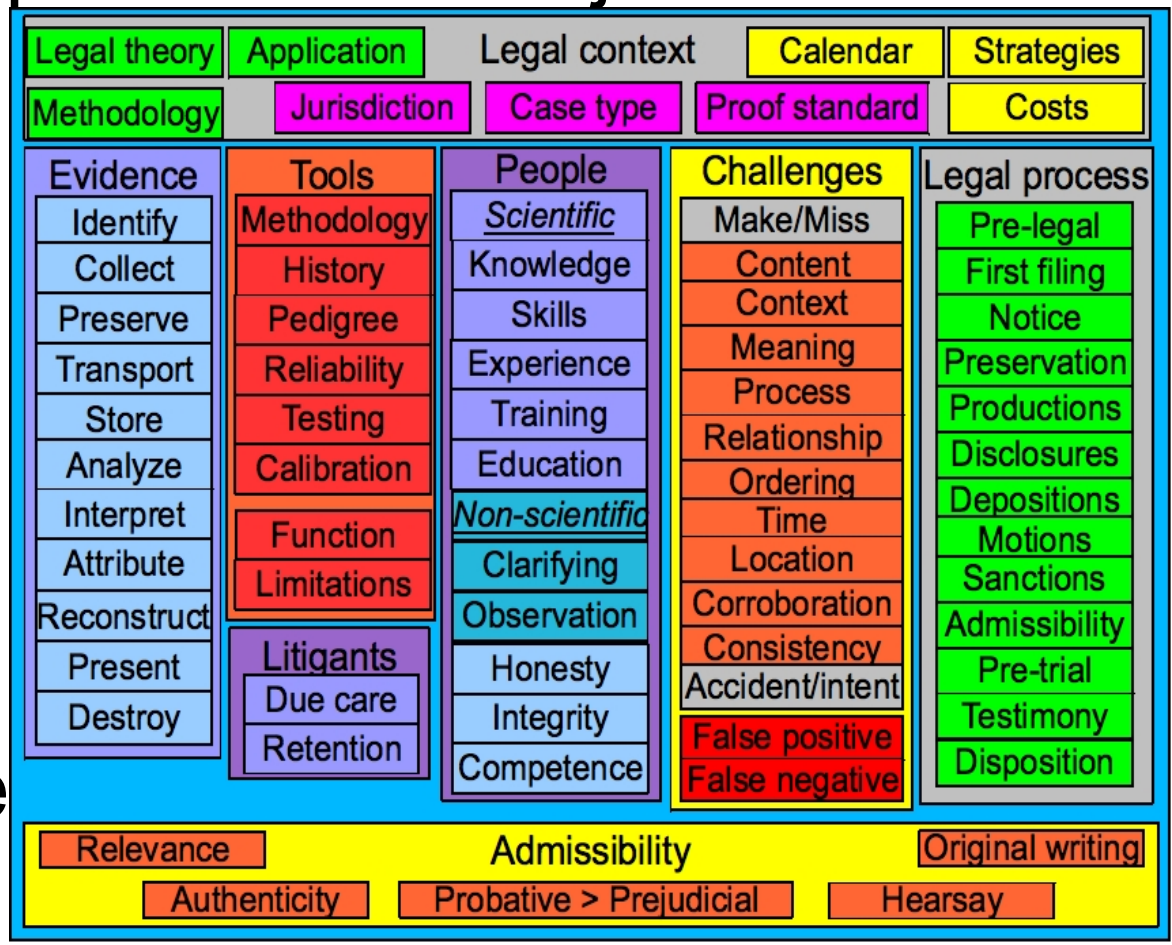
- Knowledge, skill, experience, training, or education
  - Federal Rules of Evidence 701-706
- Knowledge, Skills, and Experience:
  - Created commercial email servers and processing, industry analyst analyzing these mechanisms, operates email processing systems for high volumes of email, teaches, lectures, etc. in the area, POST certified trainer in these areas, admitted to testify as an expert in Federal, State, and Local Criminal and Civil digital forensics matters including related to emails, published refereed and other articles, etc.
- Education:
  - B.S., M.S., and Ph.D. in relevant field

Overview and thesis

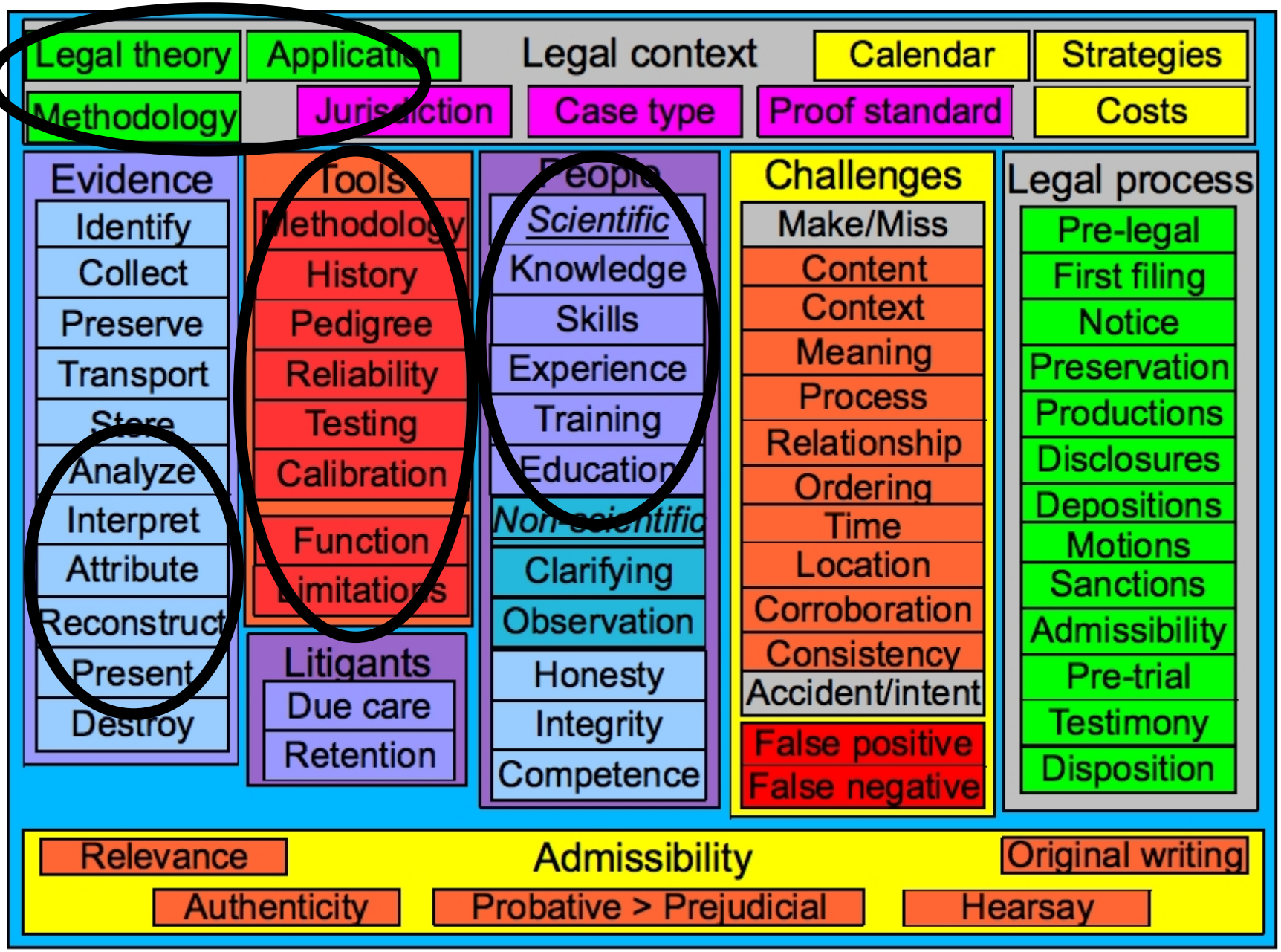
- Digital Forensic Evidence (DFE) Examination
  - Is not operating as “normal science”
    - There is only very limited community consensus
    - Terms are not well defined and consistently used
    - We don't have a well understood epistemology
    - We don't have widely used theory / methodology
    - We don't have a strong experimental basis
    - We don't have an agreed-upon physics
  - This we must change
    - By creating a community consensus
    - By defining and using terms consistently
    - By agreeing on an epistemology, theory, methodology, experimental basis, and physics

Outline

- Background of the speaker and subject
- Definitions
- Epistemology
- Theory
- Methodology
- Experimental basis
- State-of-the-science
- Your turn!



The big picture



The issues at hand

- Legal theory
- Methodology
- Application

- Knowledge
- Skills
- Experience
- Training
- Education

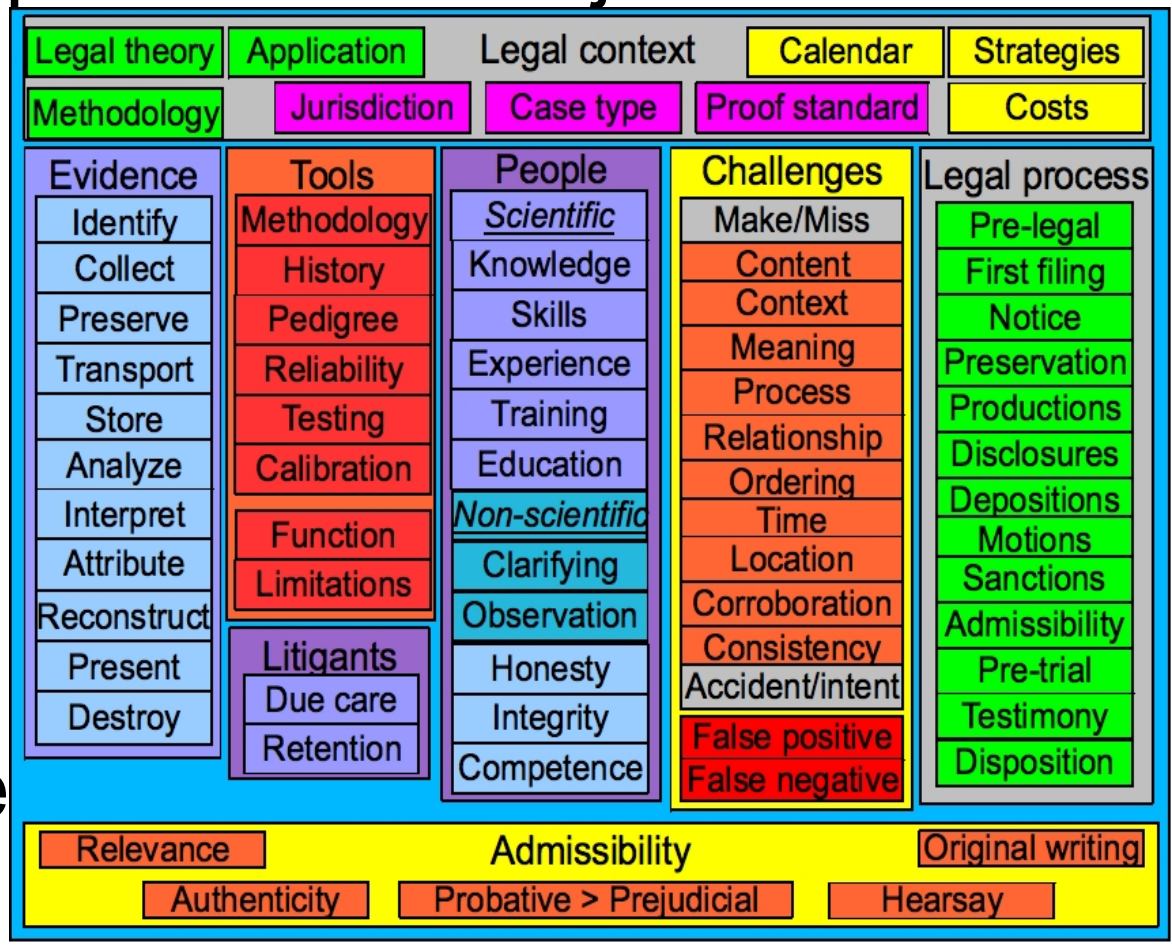
Examination:
- Analyze
- Interpret
- Attribute
- Reconstruct
- (Present)

- Methodology
- History
- Pedigree
- Reliability
- Testing
- Calibration

- Functions
- Limitations

Outline

- Background of the speaker and subject
- Definitions
- Epistemology
- Theory
- Methodology
- Experimental basis
- State-of-the-science
- Your turn!



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, this implies:
  - Digital evidence is entirely sequences of bits.
  - Physics different than matter and energy.
  - Finite granularity.
  - Observation without alteration.
  - Duplication without removal.
  - Finite, but short, times.

Epistemology 2

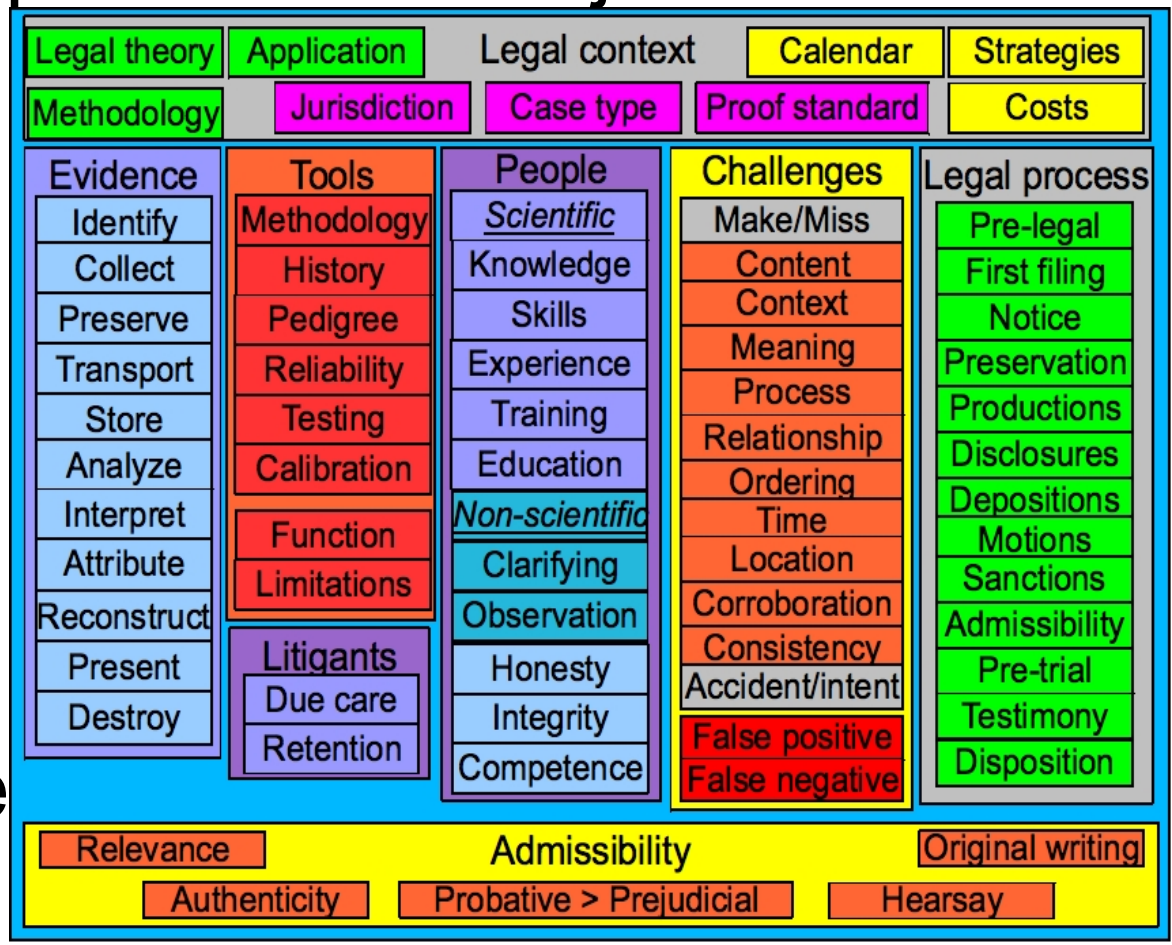
- All DFE is trace, but not transfer.
- DFE is normally latent in nature.
  - Thus it can only be observed through the use of tools.
  - This implies many issues with respect to those tools.
- DFE is produced by FSMs.
  - FSMs have specific properties that define a portion of the physics of DFE.
  - Finite granularity implies limits on accuracy and precision based on representation.
  - FSMs are syntactic in nature so semantics is driven entirely by context.

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.
- Benny Hill: “When you assume, you make an ASS of U and ME.”
  - Be careful what you say, you may be making assumptions that are provably wrong.

Outline

- Background of the speaker and subject
- Definitions
- Epistemology
- Theory
- Methodology
- Experimental basis
- State-of-the-science
- Your turn!



Theory

- Scientific theories are not casual theories.
  - They are constructs that are testable by nature.
  - Refutation can destroy any theory, but confirmation cannot prove it.
  - Scientific theories change slowly, and normally, once accepted, only change because of dramatic changes in underlying understanding of physics, and 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

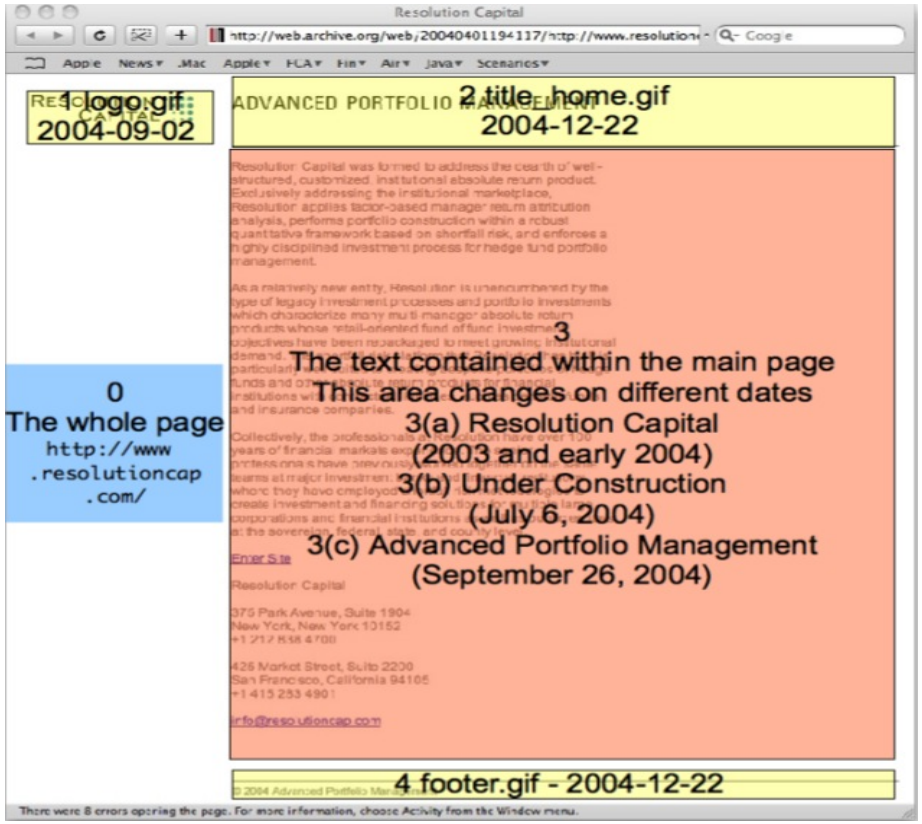
- Theories in DFE examination.
  - Form a physics of information.
  - Many of them are based on mathematical results that have long been widely accepted.
  - Some of them are still conjectures, that may be proven or maybe disproven with time.
- Most of these theories stem from computer engineering, computer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

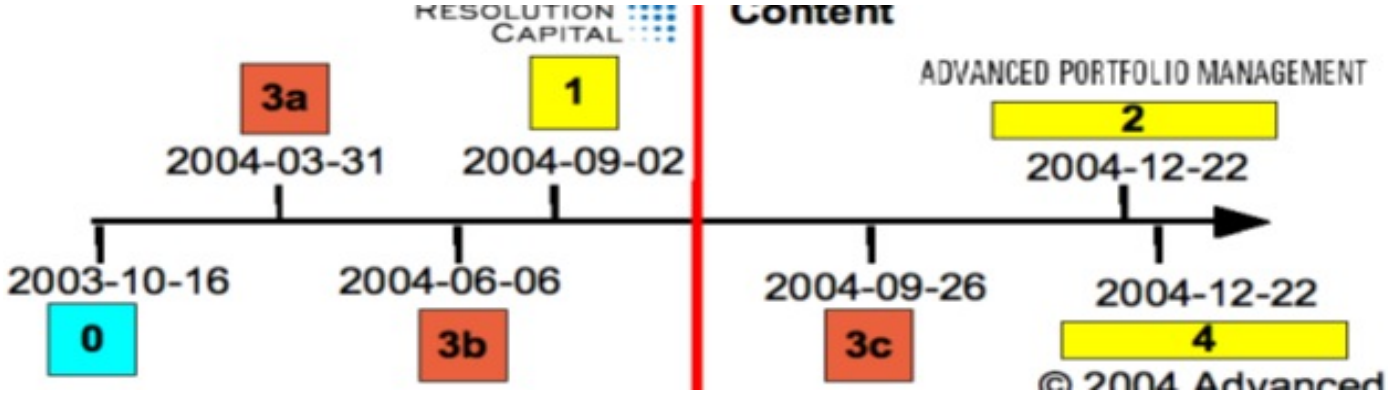
Information physics examples

- Digital space converges with time
- Time is a partial ordering
- Forward time yields perfect prediction
- Backward time yields non-unique prior states
- Thorough examination, in a sense of looking at all possibilities, is almost never feasible
- Time and space are discontinuous
- Input and output are not repeatable across the interface
- Precision is always necessarily finite

Example of time-related issue

- Wayback machine
  - archive.org
  - Historical snapshots of Web sites
  - But depictions contain images from different time frames
  - Can be deceptive





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

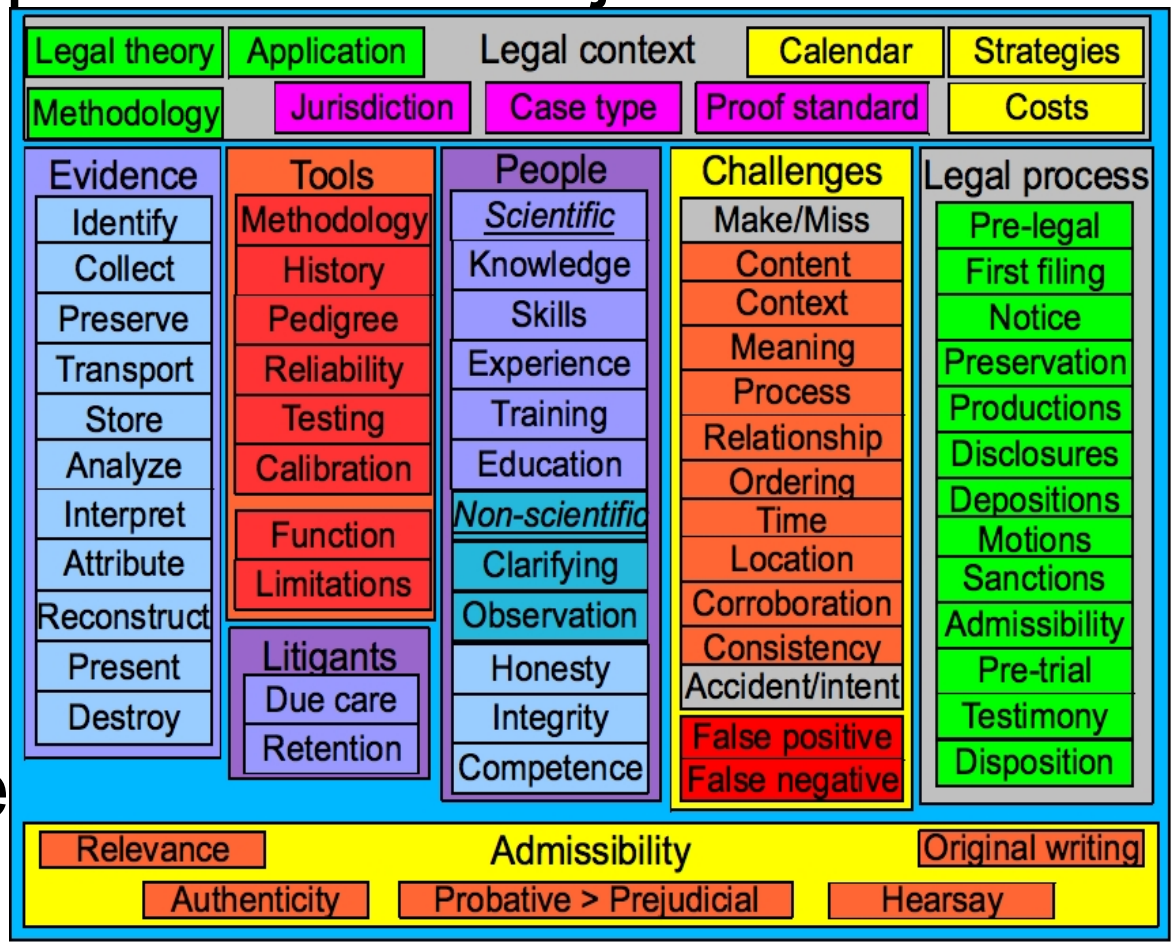
- There is a wide variety of digital forensic evidence, typically in the form of sequences of bits
  - These are called “traces” (T)
- The DFE examiner identifies consistencies and inconsistencies
  - Between and within traces (TxT)
  - Between traces and events (TxE)
- To do this, the examiner uses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

Outline

- Background of the speaker and subject
- Definitions
- Epistemology
- Theory
- Methodology
- Experimental basis
- State-of-the-science
- Your turn!



Scientific methodology in DFE

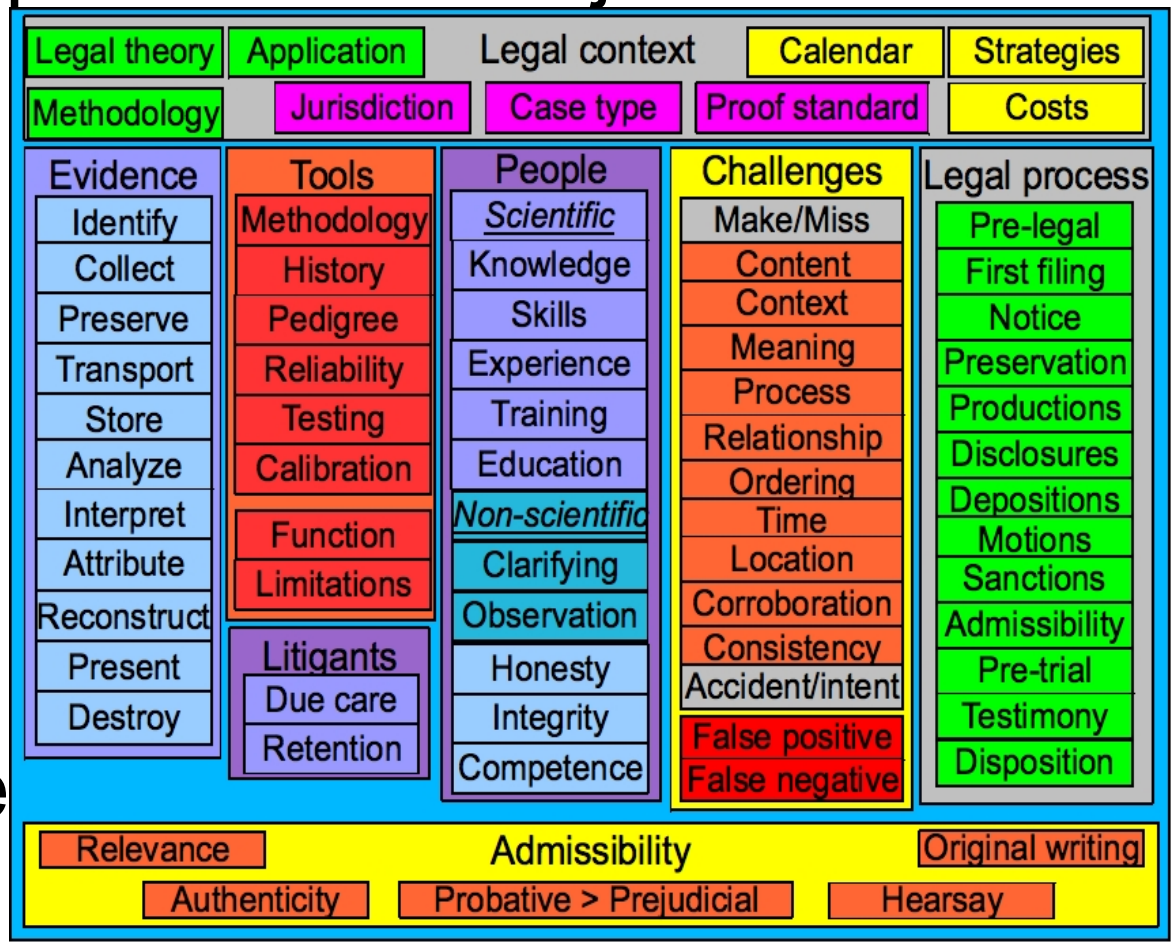
- The fundamental theorem of DFE examination:
  - **What is not consistent is not true**
- DFE examination consists of testing hypotheses to try to refute them.
  - No matter how many tests are performed, except for special cases, you can't prove that anything is true.
  - The best you can do, is show that the tests you undertook 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...
  - A single refutation disproves a hypothesis definitively.
  - The least you can say based on refutation is that the hypothesis is not true.
- Thus the methodology consists of:
  - Devising testable hypotheses
  - Testing those hypotheses against the evidence
    - A scientific test should seek to refute the hypotheses and not to confirm them
  - Inductive and deductive logic are valuable tools for testing hypotheses
  - As is experimental technique

Outline

- Background of the speaker and subject
- Definitions
- Epistemology
- Theory
- Methodology
- Experimental basis
- State-of-the-science
- Your turn!



The experimental basis is limited

- As an area of science, DFE has a relatively small number of peer reviewed and repeated scientific experiments.
  - Most of these have very limited applicability.
  - Most of these are not focused on fundamental understanding.
  - Most of the experiments don't meet the standards of scientific rigor typically apply in other fields.
  - Most of the experiments are oriented toward confirmation rather than refutation, which makes them scientifically dubious in the extreme.

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 in order 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

Presentation

- Presentation is intimately tied to, but not directly part of, examination
  - Because DFE is latent, presentation is always necessarily an issue
    - For the examiner in examining results of experiments
    - For the jury in understanding the presentation
    - For the judge in evaluating admissibility
    - For the opposition in evaluating expert reports
- Today, there is no standard for even presenting the most common representations of DFE
  - Even something as simple as presenting a text file is fraught with potential errors.

Alternative presentations

- Plaintiff's sworn statements are inconsistent with the evidence.
- If Plaintiff's sworn statements are to be believed, the evidence is not.
- If the evidence is to be believed, Plaintiff's sworn statements are not.
  - The first of these statements encompasses the second two
  - The second seems to say that the evidence is lying
  - The third seems to say that the Plaintiff is lying

Technical presentation errors

- Which of these lines is not like the other?
  - Test 1
  - Test TWO
  - Test three
  - Test 4
- The real answer is...
  - I can't tell by this presentation
  - And I know the answer.

- Which of these lines is not like the other?
  - Test I
  - Test TW0
  - Test three
  - Test 4
- The real answer is...
  - I can't tell by this presentation
  - And I know the answer.

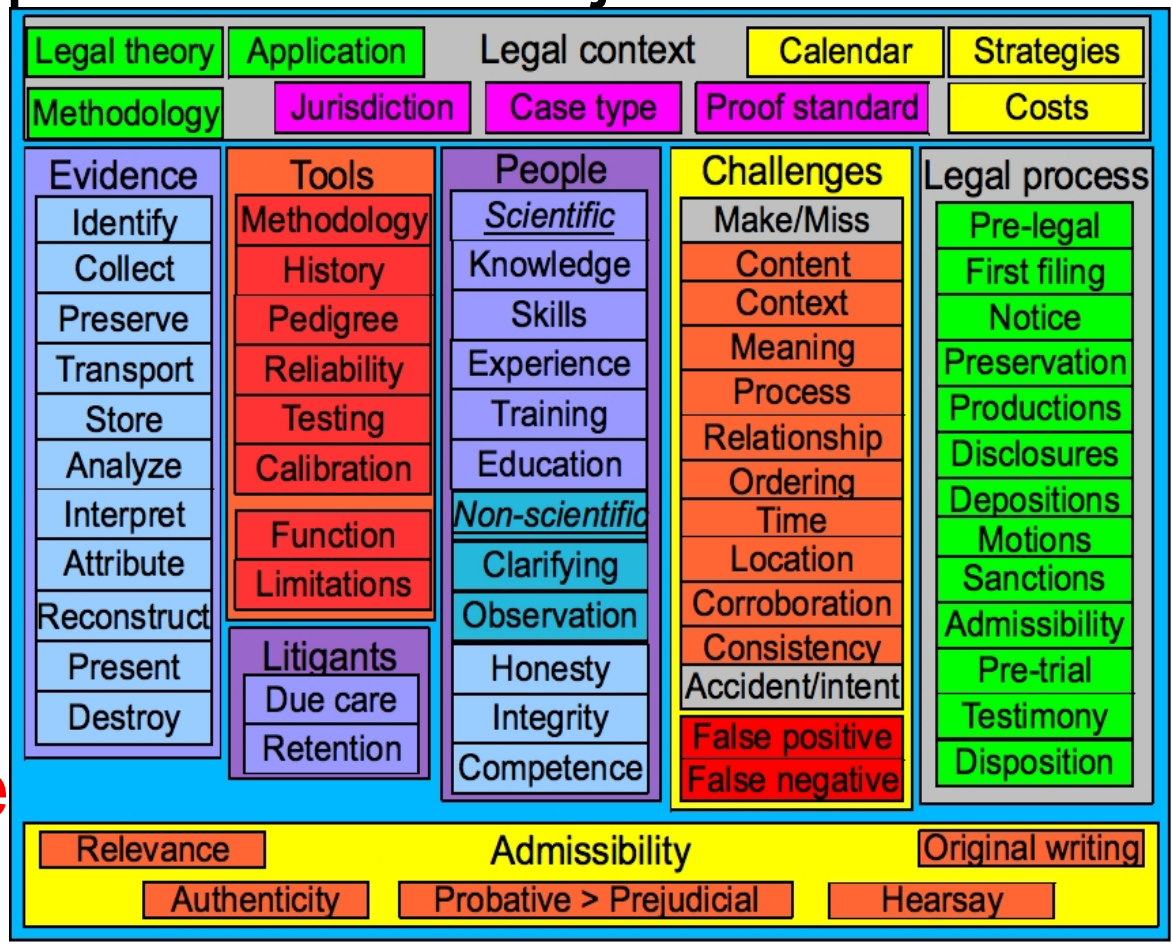
What's the diff?

- Two files that look identical when printed or viewed on the screen

Figure 4 - the output of a "diff" command using a forensic font

Outline

- Background of the speaker and subject
- Definitions
- Epistemology
- Theory
- Methodology
- Experimental basis
- State-of-the-science
- Your turn!



What is the current state?

- Analysis Interpretation Attribution Reconstruction
  - Analysis consists of a set of techniques that provide definitive answers to specific questions.
  - Existing analytical techniques are limited in quantity and in the nature of definitive results that they produce.
  - Analytical techniques are computational in nature, have defined complexity, defined input and output limitations, and can, theoretically, be verified as to their accuracy.
  - By my way of thinking, anything that does not meet these criteria, is not analysis.
    - It is something else

What is the current state?

- Analysis Interpretation Attribution Reconstruction
  - We start with a bag of bits
  - Redundancy in the bag of bits confirms or refutes hypotheses about what the bag of bits is
  - Features and characteristics are detected
  - Symbol sets are identified
  - Trace typing is undertaken
  - Content is parsed, normalized, and structure identified
  - Indicators are analyzed
  - Any of these may return you to a bag of bits

What is the current state?

- Analysis Interpretation Attribution Reconstruction
  - Characteristics and features are analyzed for consistency
  - Ordering assumptions and detection of out of order entries is undertaken
  - Sourcing and travel patterns are identified
  - Consistency is checked across related records
  - Anchor events are used for external bounding
  - Time differentials and jitter are considered
  - All of these are compared to hypotheses to identify consistency and inconsistency

What is the current state?

- Analysis Interpretation Attribution Reconstruction
  - Sieves are built and things are counted
  - Derived traces are formed for analytical convenience
  - Counts are made of various features and characteristics of interest to the issue
  - Mechanisms are combined and resulting errors identified and mitigated
  - Results are verified by independent means wherever feasible

What is the current state?

- Analysis Interpretation Attribution Reconstruction
  - Intentionally hidden items of interest are identified and sought
  - Placement in hard-to-find locations are checked
  - Steganographic other transforms are sought
  - Recursive embedded languages are considered
  - Indicators are identified and sought
- Cognitive methods, both automated and human, are combined to identify consistencies and inconsistencies
  - Available methods and computational complexity limit analytical processes

What is the current state?

- Analysis Interpretation Attribution Reconstruction
  - Traces and analytical results must be interpreted
  - Interpretation involves selecting between alternative explanations (hypotheses)
  - Structured and unstructured traces involve very different interpretation methods
  - Over-interpretation is quite common (a bridge too far)
  - Special care must be taken in making statistical claims
    - Such claims often ignore data not present
    - Such claims often conceal assumptions related to random stochastic processes and other similar things
    - Such claims often assert precision exceeding accuracy

What is the current state?

- Analysis Interpretation Attribution Reconstruction
  - Tools commonly interpret traces
  - These interpretations make assumptions, often without basis or with a wrong basis
  - These interpretations may present false results based on these assumptions
  - These interpretations include presentations and other depictions that may mislead the examiner
  - No methodology currently exists for evaluating interpretation by tools

What is the current state?

- Analysis Interpretation Attribution Reconstruction
  - Interpretation includes the identification and explanation of missing traces
  - Redundancy is used to mitigate interpretation errors
  - Because of its interpretive nature, interpretation is highly subject to individual variations
  - Proper interpretation is properly limited, and properly couched in terms of its accuracy and applicability
  - There is a tendency to grasp onto casual theories and embrace them as if they were something more
  - Don't believe what you think is right unless you really know

What is the current state?

- Analysis Interpretation Attribution Reconstruction
  - Assume each system and situation is in fact different from others
  - Experience is a great teacher, but can also be misleading; be careful not to over-interpret
  - When interpretation is unclear, reconstruction is a viable and worthwhile mediator
    - Reviewed in some depth later
  - All interpretation should be made within the context of information physics
  - As a working assumption, assume that every interpretation will be tested against information physics and refuted if inconsistent

What is the current state?

- Analysis Interpretation Attribution Reconstruction
  - Events are also subject to interpretation
  - Words are interpreted by examiners
    - Cognitive errors in interpretation are common
    - Examiners assume the context of their knowledge
    - Words mean different things to different people
    - Careful interpretation may be viewed as “picky”
  - Events in light of information physics
    - Claims are often inconsistent with (refuted by) physics
    - Causality is particularly tricky
    - Time requirements for actions deals with complexity
    - Reviewing information physics for events may help

What is the current state?

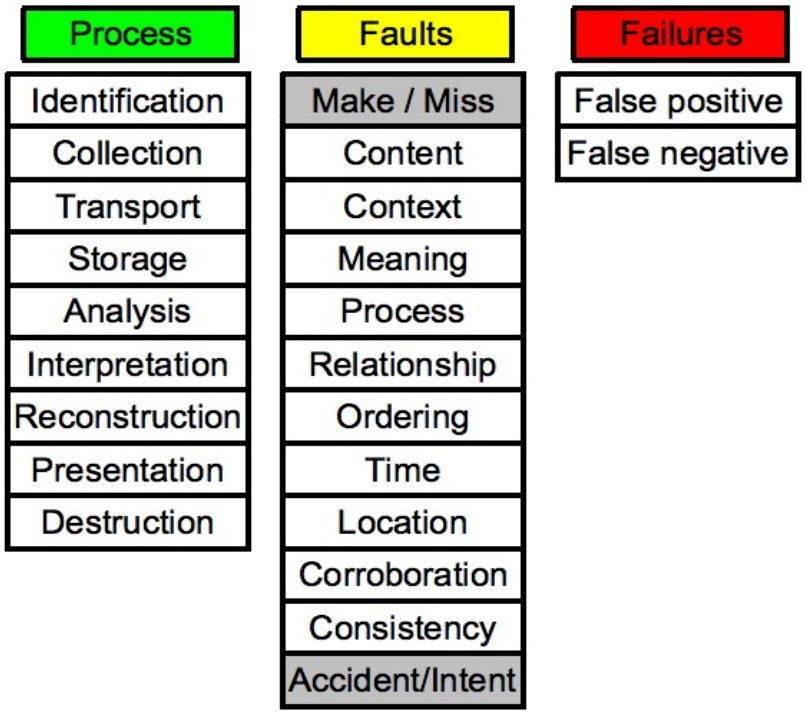
- Analysis Interpretation Attribution Reconstruction
  - Resources limit interpretation
    - Schedule may prohibit contemplation of the issues
    - Computational resources and costs may be prohibitive
    - Available traces may change with time and legal actions
  - Statements are often unnecessarily interpretive
    - Careful wording is highly desirable in written and verbal communications
    - Few “standard wordings” exist for DFE examination
    - Examiners often make statements that end up being wrong but could be stated differently
    - “I found X Bs in file Q” is true even if there are more than X of them present – because you didn't find them

What is the current state?

- Analysis Interpretation Attribution Reconstruction
  - Similarity is almost always interpretive
    - We lack adequate similarity metrics or criteria
    - We lack experimental basis for similarity claims
    - We lack techniques for detecting similarity
    - Automated results often differ dramatically from human observations (e.g., in pictures or sounds)
  - Assumptions in interpretations are critical
    - They are often not detailed in reports or statements
    - Thus the statements/reports lack the basis for the interpretations
    - It may be hard to identify all of your assumptions
    - But confirming/refuting assumptions is the process

What is the current state?

- Analysis Interpretation
  Attribution Reconstruction
  - Visualization is interpretive
    - Format assumptions are often wrong
    - Appearances may be deceptive:
      - O or 0?
      - I or 1?
    - Recent results (Forensic Fonts) are only a start
  - Interpretation is subject to errors and challenges



What is the current state?

- Analysis Interpretation Attribution Reconstruction
  - Correlation is not causality
  - Before is not because
- Fundamental assumption of causation (digital)
  - Traces come about by the execution of finite state automata that follow the physics of the digital world.
- Physical version:
  - Effects come about by the execution of mechanisms that follow the physics of the physical world.

What is the current state?

- Analysis Interpretation Attribution Reconstruction
  - Some basics for the digital world
    - A caused B implies A before B
    - How much before is determined by the digital version of the speed of light
      - Computational complexity
      - Performance levels of devices
      - Mechanisms available
    - Statistics do not apply in most cases
    - The past is what it is; there is no “might have been”
  - Establishing a causal chain is non-trivial
    - It often involves redundant records and almost never eliminates all other possibilities

What is the current state?

- Analysis Interpretation Attribution Reconstruction
  - FSMs are highly predictable
    - FSMs converge with time (real world diverges)
    - Simulation can produce identical outputs
  - BUT BEWARE – reverse time is not unique
    - In the physical world, reverse time may be unique
    - In the digital world, convergence implies many paths lead to the same traces
  - Beware the sensors
    - Physical to digital is highly nonlinear
    - Small differences expanded near the nonlinearity
    - Large differences reduced far from nonlinearity

What is the current state?

- Analysis Interpretation Attribution Reconstruction
  - Attributing actions to actors (human)
    - Authentication methods are of limited value
    - Most biometrics are not good for identification – only for selecting a known out of a group of a few thousand knowns when deception is not in use
    - Something the user has can be taken or exploited
    - Something the user knows can be known by others
    - Something the user can do may be done by others
  - DFE can almost never put a person at a keyboard at a time
  - But other events may be able to help do so

What is the current state?

- Analysis Interpretation Attribution Reconstruction
  - Attack attribution is fairly limited today
    - Level 1 – system it came directly from
    - Level 2 – system originating it
    - Level 3 – individual originating it
    - Level 4 – organization behind the individual
  - Behavioral attribution
    - The words and word sequences they use
    - The commands they run and usage patterns
    - Keyboard patterns, etc.
    - All fail under deception
    - Limited to known individuals of small known groups

What is the current state?

- Analysis Interpretation Attribution Reconstruction
  - Device authentication and attribution
    - Various indicators are often present if sought
  - Operating environment identification
    - As for devices
  - Predicted behaviors of programs
    - Many known programs have highly predictable behaviors
  - BUT!!!
    - Most of these are readily subverted by deceptions
    - Redundant traces make deceptions harder to do without detection

What is the current state?

- Analysis Interpretation Attribution Reconstruction
  - Attribution of damages to parties
    - Physical damage, Conversion
    - Deprivation (the key DFE issue)
    - Lost value, Lost rights
  - If attribution can be done of cause to effect
    - Calculation methods are available to identify the extent to which deprivation is done
    - Cost per usage of electricity
    - Cost in reduced lifetime of equipment
    - Cost in demonstrable lost business
    - Cost in reduced life of equipment

What is the current state?

- Analysis Interpretation Attribution Reconstruction
  - Driving time backwards
    - Information physics makes this problematic
    - Lack of adequate traces over time leads to very large envelopes of possible histories
    - No way to tell what was “original”
    - Theft may not be identifiable
    - Travel time and jitter produce ordering uncertainties
    - Unique reverse time through homing sequences is impossible without all relevant prior traces
    - Error accumulation means large reverse expansion
    - The list goes on and on...
  - Experimental demonstration of operation

What is the current state?

- Analysis Interpretation Attribution Reconstruction
  - Experimental demonstration of operation
    - Reconstruction is used to test hypotheses about the particular case
    - It can confirm, refute, or be unrevealing
- Methodology (Constructed and Original Traces):
  - Similarity measures used to define (in advance) "match" of C-trace to O-trace and implications
  - Create reconstruction(s) based on hypotheses
  - Constructed traces (C-trace) compared to original traces (O-trace) to confirm/refute hypotheses

What is the current state?

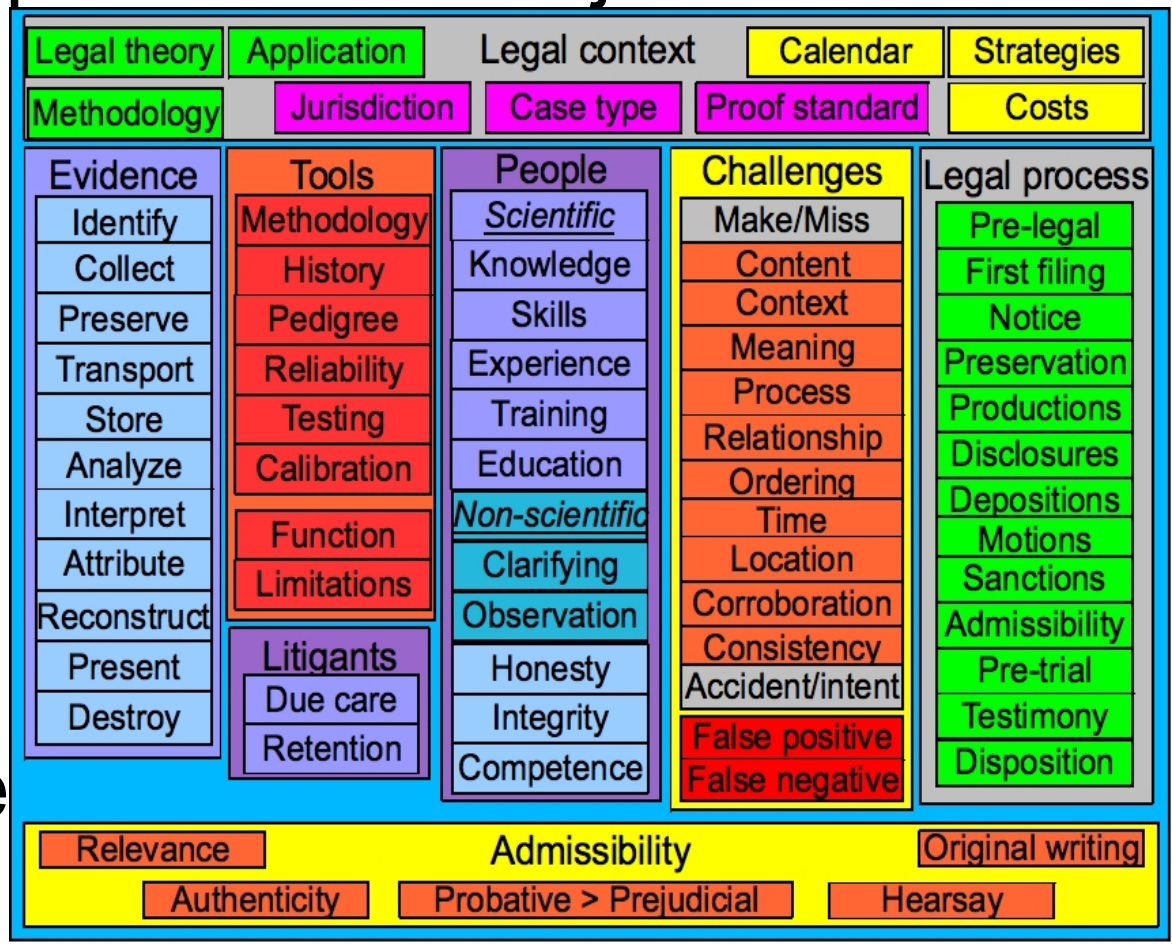
- What we can reasonably say (Who goes further?)
  - I did X and observed Y
  - I [did not find | found] X in Y
  - X is [in]consistent with the claim that Y [because...]
- Tools and processes (Who knows how to do this?)
  - Methods and materials used (complete & thorough)
  - Procedures applied
  - Results of those procedures
  - Conclusions
  - Sources and magnitudes of uncertainty in procedures and conclusions (e.g., confidence intervals)

Review of thesis

- Digital Forensic Evidence (DFE) Examination
  - **Is not operating as "normal science"**
    - What can we build community consensus for?
    - What well-defined and consistently used terms?
    - What well-understood epistemology?
    - What theory / methodology should we choose?
    - What strong experimental basis should we build?
    - What agreed-upon physics should we use?
- How do we move into normal science?
  - Build a community consensus!
  - Is the path I have outlined the right start?
  - What can we embrace / should we change?

Outline

- Background of the speaker and subject
- Definitions
- Epistemology
- Theory
- Methodology
- Experimental basis
- State-of-the-science
- Your turn!



Thank You

http://calsci.org/ - calsci at calsci.org

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