Datenbanken: „The Bricks of Cyberspace“

Andreas Reuter
EML Research gGmbH

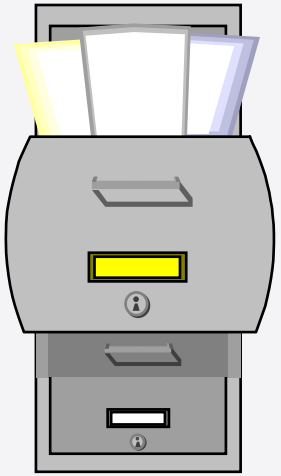
Dagstuhl, 30. Juni 2005

Platten sind billiger als Papier

- Aktenschrank:

Schrank (4 Auszüge)	250€
Papier (24,000 Blatt)	250€
Platz (0,6 m^2)	180€
Summe	680€

2,8 cent/Blatt

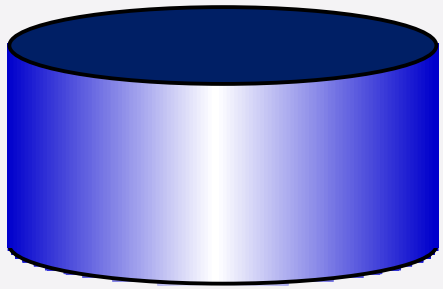


- Platte:

Kapazität: 120 GB 800€

Text (ASCII): 60 Mio Blatt
0,13 cent/Blatt (21 x billiger)

Bild (1MB): 120.000 Bilder
0,67 cent/Bild (5 x billiger)



=> Alles sollte auf Platte gespeichert werden.

Die Rolle von Datenbanken

- Alle Information wird in Datenbanken gespeichert.
- Man kann alles speichern, was man ...
  - liest: 10MB/Tag, 400 GB/Person
  - hört: 400MB/Tag, 16 TB/Person
  - sieht: 1MB/s, 40GB/Tag, 1.6 PB/Person
- Speicherorganisation und Suchen ist nicht trivial.
- Trotzdem: Das ist die Aufgabe von Datenbanken.
- Datenbanken sind gut für stark strukturierte Daten.
- Die nächsten Stationen sind Text, Raum/Zeit, Bilder und Töne.
- Das ist sehr Prozessor-intensiv.

Wie spannend ist ein Ziegel?

- Butler Lampson stellte vor einigen Jahren fest: „The greatest failure of the systems research community over the past ten years was that we did not invent the web”.
- Das Web hat mittlerweile eine Neuausrichtung der IT verursacht (und eine globale Wirtschaftskrise ausgelöst). Hat die Datenbank-Community etwas vergleichbares erfunden?
- Paul Larson sagt: “Database systems are as interesting as the household plumbing to most people. But if they don't work right, that's when you notice them.”
- Sind Datenbanken also weniger die „bricks of cyberspace“, sondern eher die Kanalisation?

Zuerst die Erfolge

- Typerweiterbarkeit
- Neue Methoden der Inhaltsadressierung
- Automatische Optimierung (inkl. Parallelisierung) komplexer Funktionen
- Integration in objektorientierte Programmierumgebungen durch Nutzung von CLR-Konzepten
- Vermeidung des „impedance mismatch“
- Übernahme von Betriebssystemfunktionen durch „stored procedures“, recoverable queues usw.
- Erweiterung von Programmierumgebungen durch mengenorientierte Konzepte, Recovery-Funktionen usw.

Aber das alles begeistert höchstens die Techniker.

Einige Arten privater Datensammlungen

Art	Plattform/Tool	Eigenschaften des Speichersystems	Datenmodell	Bezug zu anderen Arten
Email	Email-System	Dateisystem, Datenbank	Hierachie; schwach strukt. Texte	Viele: Struktur, Wert, Konzept
Adressen	Email-System, Directory	Prop. oder offenes Dateisystem	Quasi-relational; viele “Standards”	Viele: Struktur, Wert
Termine	Kalender-System	Prop. Dateisystem	Hierarchie; unstrukt. Texte	Viele: Struktur, Wert, Konzept
Projektplanung	Planungs-Tool	Prop. Dateisystem oder Datenbank	Abhängigkeitsgraph schwach strukt. Texte	Viele: Struktur, Wert
Budgetplanung	Spreadsheet	Prop. Dateisystem oder Datenbank	Matrix; arithm. Ausdruck; unstrukt. Texte	Mehrere: Wert
Persönliche Bestandsverwaltung	4GL-Tool	Offene Datenbank	Relational	Mehrere: Wert
Persönliche Kontoverw.	Web-Frontend zur Bank-Anwendung	Prop. Datenbank	Formular-orientiert	Viele: Wert
Sonstiges (Rechn., Quittungen)	Pappschachtel, Ordner	./.	./.	Viele

Und dann gibt es noch ...

- SMS-Nachrichten
- Bücher
- Bilder
- Anrufe
- Medizinische Befunde
- Tickets
- Urkunden
- ....

Zitat

Data models are tools. They do not contain in themselves the "true" structure of information. What really goes on when we present a data model, e.g., hierarchies, to a user? Does he say "Aha! Of course my information is hierarchically structured; I see how the model fits my data"? Of course not. He has to learn how to use it. We generally presume that this learning is required only because of the complexity of the tool. Difficulties are initially perceived as a failure to fully understand the theory; there is an expectation that perseverance will lead to a marvelous insight into how the theory fits the problem. In fact, much of his "learning" is really a struggle to contrive some way of fitting his problem to the tool: changing the way he thinks about his information, experimenting with different ways of representing it, and perhaps even abandoning some parts of his intended application because the tool won't handle it. Much of this "learning" process is really a conditioning of his perceptions, so that he learns to accept as fact those assumptions needed to make the theory work, and to ignore or reject as trivial those cases where the theory fails.

Wissenschaftliche Datenbanken

- Die Rohdaten werden einmal geschrieben und nicht mehr geändert (z.T. auch aus rechtlichen Gründen).
- Die Rohdaten kommen (je nach Sensor) als Datenströme mit teilweise hohen Bandbreiten (> 100 MB/s). Sie müssen auf jeden Fall gespeichert werden, da viele Messungen nicht wiederholt werden können.
- Für die Mehrheit der Anwendungen sind die Rohdaten uninteressant. Sie brauchen stattdessen Aggregate, abgeleitete Werte oder – im Falle von Textfeldern – eine Art “Abstract”.
- In vielen Fällen hat das Schema tausende von Attibuten, aber jeder Satz hat typischerweise ur einige zehn Attributwerte.
- Das Schema des strukturierten Teils der Datenbank ist oft nicht stabil. Mit dem Fortschritt der Disziplin werden neue Phänomene entdeckt, alte werden neu bewertet, es werden neue Messungen gemachte, für andere werden neue Einheiten eingeführt, und gelegentlich gibt es sogar Änderungen auf der Konzeptebene. Alles das muss dynamisch akkomodiert werden.

Daten für die Simulation metabolischer Netzwerke

- Allgemeine Daten über biochemisches Reaktionen aus diversen Quellen (Datenbanken, Dateien, verschiedene Formate und Referenzen, umstrittene Daten usw.)
- Daten über kinetische Parameter, die typischerweise in Papierform und nicht in Datenbanken vorliegen.
- Die Nutzung dieser Daten für Simulationen erfordert ein semantisch reiches Modell des Gegenstandsbereiches, d.h. der Biochemie, der Molekularbiologie, der Genforschung.
- Ein “vollständiges” semantisches Modell erleichtert auch die Informationsextraktion aus der Literatur und die (semi-automatische) Validierung der Simulationsergebnisse.

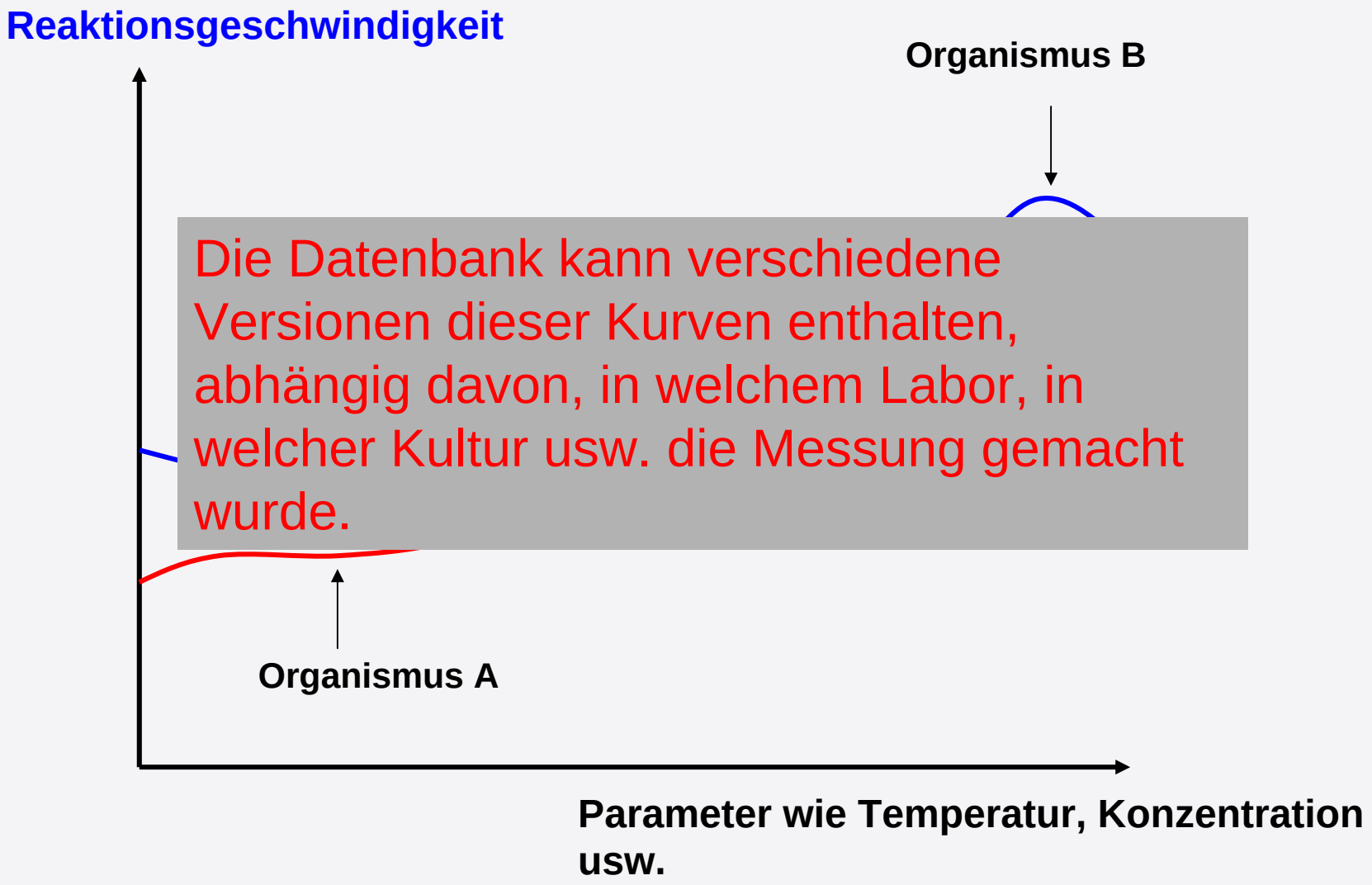
PBE: Problems by Example

- In Labors und Firmen werden etliche hundert “Datenbanken” über Gene, Proteine, biochemische Reaktionen usw. betrieben.
- Die meisten davon sind “flat files”.
- Es gibt kein gemeinsames Schema.
- Die Datensammlungen beruhen auf unterschiedlichen Anwendungskonzepten (die sich zudem ändern).
- Oft steckt die interessante Information in Textfeldern, die durch Abkürzungen und Labor-Jargon weiter verschlüsselt werde.
- Gleichwohl möchten viele Wissenschaftler einen transparenten Zugriff zu allen diesen Daten.

Zitat

Our notions of reality are overwhelmingly dominated by the accidental configurations of our physical senses. We are very parochial in our sense of scale. Bacteria and viruses and sub-atomic particles are not very real to most of us, nor are galaxies. We don't really know how to comprehend them. Our concept of motion is bounded by the physiology of our eyes: the continental plates don't move, but motion pictures (sequences of still pictures!) do. Most of us think of continents and islands as permanent and discrete entities -- rather than as accidents of the current water level in the oceans. Are islands and mountains such different things? Have you ever had the opportunity to observe a reservoir get filled, or emptied?

Reaktionskinetische Daten



Zitat

There's a catch right there: the implicit assumption that there is just one "technology" by which all people perceive information, and hence which is most natural and easy for everybody to use. There probably isn't. Human brains undoubtedly function in a variety of ways. We know that some people do their thinking primarily in terms of visual images; others hear ideas being discussed in their heads; still others may have a different mode of intuiting concepts, neither visual nor aural. Analogously, some people may structure information in their heads in tabular form, others work best with analytic subdivisions leading to hierarchies, and others naturally follow paths in a network of relationships.
This may well be the root of the debates over which data model is best, most natural, easiest to learn and use, most machine independent, etc. The camps are probably divided up according to the way their brains function -- each camp advocating the model that best approximates their own brain technology.

Sensor- und Realzeit-Datenbanken

- Provide flexible query functionality, aggregation and extraploation of the respective processes that can’t be achieved on the physical objects proper.
- Accommodate significant fluctuations in the data rate, including sharp bursts.
- Incoming data has to be related to the existing data in complex ways in order to compute the type of derived information.
- For example, subscribers need to be notified of relevant changes in the incoming data in real time, so in case of many subscribers there is a correspondingly huge stream of outgoing data, i.e. notification messages.
- Both space and time need to be first-class citizens of the data model rather than just another set of attributes. References to value histories must be supported at the same level as references to the current value.

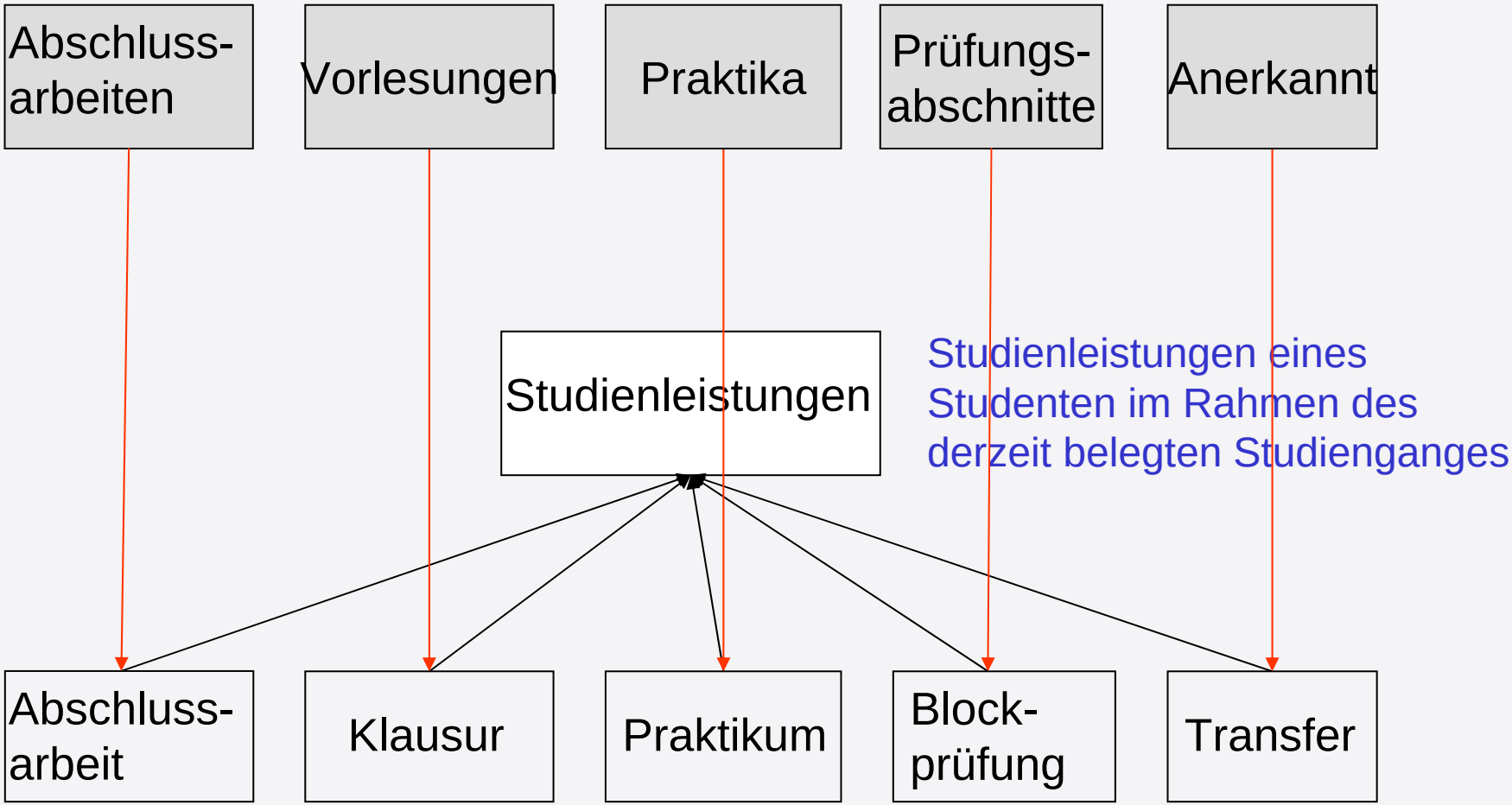
Bestandsaufnahme

Üblicherweise werden alle diese Anwendungen separat behandelt, mit unterschiedlichen Konzepten, Tools, Methoden usw. Datenbanken gelten als Mittel zur Verwaltung strukturierter Daten (was auch noch weitgehend stimmt), wogegen für Textsammlungen “information retrieval systems” zuständig waren/sind. Mit der Idee der “semi-strukturierten Datenbanken” wurde versucht, diesen Spalt zu überbrücken, denn, wie Jennifer Widom beobachtet: “For most people, the initial schema looks like ‘/stuff’ “. Die Konvergenz ist gleichwohl noch nicht erreicht worden. In ähnlicher Weise haben die Verfechter der temporalen Datenbanken, der Realzeit-Datenbanken, der aktiven Datenbanken versucht, ihre speziellen Proble zu lösen und sich nicht um eine Integration der Konzepte gekümmert. Das war wahrscheinlich eine gute Idee, denn die Lösung des Integrationsproblems ist eine gewaltige Aufgabe.

Zitat

Theories tend to distinguish phenomena. A theory tends to be analytical, carefully identifying all the distinct elements and functions involved. Unifying explanations are abstracted, relationships and interactions are described, but the distinctness of the elements tends to be preserved.
Good tools, on the other hand, intermingle various phenomena. They get a job done (even better, they can do a variety of jobs). Their operation tends to intermix fragments of various theoretical phenomena; they embody a multitude of elementary functions simultaneously. That's what it usually takes to get a real job done. The end result is useful, and necessary, and profitable.

Ein „triviales“ Beispiel



Zitat

Many of the concerns about the semantics of data seem relevant to any record keeping facility, whether computerized or not. I wonder why the problems appear to be aggravated in the environment of a computerized data base. Is it sheer magnitude? Perhaps there is just a larger mass of people than before who need to achieve a common understanding of what the data means. Or is it the lost human element? Maybe all those conversations with secretaries and clerks, about where things are and what they mean, are more essential to the system than we've realized. Or is there some other explanation?

Zitat

I have tried to describe information as it "really is" (at least, as it appears to me), and have kept tripping over fuzzy and overlapping concepts. This is precisely why system designers and engineers and mechanics often lose patience with academic approaches. They recognize, often implicitly, that the complexity and amorphousness of reality is unmanageable. There is an important difference between truth and utility. We want things which are useful -- at least in this business; otherwise we'd be philosophers and artists.

Wohin führt das?

Die Frage müsste natürlich genauer heißen: Wohin könnte/sollte das führen, aber das kann jeder für sich umformulieren.

- Datenbanken müssen sehr viel mehr Unterstützung bei der Integration heterogener Begriffs- und Modellwelten leisten.
- Datenbanken müssen Schema-Evolution als reguläre Operation unterstützen.
- Alles dies etwas wesentlich anderes als Schema-Integration bzw. regelbasierte Sichtenverwaltung; es erfordert die Definition (tiefer) Semantik als Teil des Schemas (statt der rein syntaktischen Schemata).
- Wenn Datenbanken das nicht tun, werden es irgendwelche Anwendungsschichten auf irgendwelche verschiedenen Arten tun (und damit das Problem langfristig vergrößern); Datenbanken dagegen werden auf Dauer nur komfortable Behälter für Tupel und Objekte sein.

In diesem Sinne:

Manche behaupten, Komplexität sei wie Energie: Man könne sie umwandeln, transportieren, aber nicht erzeugen und schon gar nicht vernichten (reduzieren).

Schaumer mal.

Nichts Neues unter der Sonne

"We do not, it seems, have a very clear and commonly agreed upon set of notions about data -- either what they are, how they should be fed and cared for, or their relation to the design of programming languages and operating systems. This paper sketches a theory of data which may serve to clarify these questions. It is based on a number of old ideas and may, as a result, seem obvious. Be that as it may, some of these old ideas are not common currency in our field, either separately or in combination; it is hoped that rehashing them in a somewhat new form may prove to be at least suggestive."

G. H. Mealy: Another look at data, FJCC 1967, pp. 525-534.

Das letzte Zitat

"I do not know where we are going, but I do know this -- that wherever it is, we shall lose our way." (Sagatsa)
"If you're confused, it just proves you've been paying attention." (G. Kent)

This book projects a philosophy that life and reality are at bottom amorphous, disordered, contradictory, inconsistent, non-rational, and non-objective. Science and much of western philosophy have in the past presented us with the illusion that things are otherwise. Rational views of the universe are idealized models which only approximate reality. The approximations are useful. The models are successful often enough in predicting the behavior of things that they provide a useful foundation for science and technology. But they are ultimately only approximations of reality, and non-unique at that.

So we shrug it off, shake it away as nonsense, philosophy, fantasy. What good is it? Maybe if we shut our eyes the notion will go away.