CERN とあいまいな情報処理

副理事長　渡辺貞一

１．はじめに

ファンデルワールス（van der Waals）という名前は、理科の授業に出てくるので、多くの人にとって、一度は耳にしたことのある名前ではないかと思う。

この人名（オランダの物理学者、19 世紀末）のついたファンデルワールスの力は、分子間に働くごく近距離の弱い力（距離のマイナス 7 乗）で、電気双極子の相互作用や凝集力、高分子間力の解明に役立っている。

自然界には多くの力があるが、人名のついた“力”は珍しく、ファンデルワールスは、印象に残る名前の一つである。

T 社の研究所にいたとき、同姓のオランダの物理学者の訪問を受けたことがある。ナイメーヘン大学の教授で、且つ CERN（欧州素粒子・原子核研究機構）の研究員でもある素粒子物理の専門家であった。

国の KEK（高エネルギー加速器研究機構）からの紹介のようで、経緯は分からないが、パターン認識や画像処理に興味があり、研究状況を聞きたいとして訪ねてこられた。

丁度、がん細胞の画像認識の研究を行っていたときで、その話を中心にデジタル画像処理の話をした。

これを契機に、ファンデルワールス先生と少し長い交流が始まった。

先生は、年に 1~2 回、筑波にある KEK 研究所を訪問されており、技術交流を兼ねたプライベートな会を持った。こちらから、パターン認識や画像処理、AI の話をし、先生から CERN の話や素粒子物理の現状を紹介してもらった。大学では、少し物理を勉強したので、場やゲージ粒子の話、また CERN で発見された Z や W ボゾンの話、対称性やその自発的な破れ、さらには最新の標準模型（Standard Model）の話など、量子物理の現状を知る少し懐かし個人授業の時間でもあった。

オランダの人達は、語学が堪能で、ほとんどの人が数カ国語を話す。特に英語は、英米人よりも標準的で分かりやすく、こちらは少し英語が上手になったような気がして、会話を楽しんだ。

これが縁で、CERN に招待され、欧州出張時に、思いきって足を伸ばし訪問することにした。

２．CERN

CERN は、スイスのジュネーブ郊外にあり、一部フランスの国境を跨ぐ、巨大な欧州の

素粒子物理学研究所である。地下 100m ほどのところに、27km の円周のトンネルが掘られていて、粒子を円周状に高速に加速し衝突させる衝突型加速器が設置されている。

訪問したときは、LEP (Large Electron Positron collider) と呼ばれる電子・陽電子加速器で、質量を持つゲージボゾンの Z ボゾンと W ボゾンを発見して、標準模型を大きく裏付ける検証をしたとして活気があった。

地下に降りて見学をしたが、国境を跨ぐので、パスポートを持参した。

この加速器を用いた実験は、年に 2 回行われていて、1 回の実験で、80 テラバイトのデータがテープに記録され、半年かけて解析されていた。

現在は、この加速器は、陽子・陽子の衝突型に改良され、大型ハドロン衝突型加速器 LHC (Large Hadron Collider) となっているので、全てにおいて、高エネルギー化され、衝突によって発生する粒子も多く、巨大で複雑なシステムになっている。

粒子を検出する円筒型の検出器 (ATLAS) は、高さ 22m、長さ 44m で、重量は 7 千トンと云われている。

この中心には、発生した粒子を選別するのに必要な磁場を発生させる超伝導ソレノイドがある。大型のシリンダーの内側に取り付けられた円筒コイルで、2 万ガウスの磁場を発生させることができる。

超伝導は、液体ヘリウムで、アルミ配管されている。また、この検出器の前後には、陽子ビームを絞り込む (20μmφ) ための超伝導 4 極電磁石が設置されている。これらは、いずれも日本製 (T 社) と云われている。

この装置により、数年前に、標準模型で予測された高エネルギーの粒子、ヒッグス粒子が発見された。

現在、標準模型は、自然界を表すほぼ正確なモデルとして受け入れられている。

その方程式 (L) は、基本素粒子、電磁気力、弱い核力、強い核力、ヒッグス粒子の項目からなり、石に刻まれて、CERN の一角に建てられている。

CERN 訪問時のもう一つの関心事は、情報処理であった。

今世界で広く使われているインターネットの HTML (Hyper Text Markup Language) や WWW (World Wide Web) は、ここが発祥の地である。早くから膨大なデータの処理と研究を円滑に進めるために、新しい仕組みが考案され、実用化されてきた。

HTML は、数千人の研究者が、ネットワークを通じて、必要な文献やデータを容易に検索できるように作られた規約である。

文書に目印をつけ、文書を相互にリンクさせるハイパーテキストシステムは、直観的でわかりやすいコンピュータ検索システムである。

この仕組みは、イギリス人の情報処理技術者 Berners Lee が考案し、中心となって開発したと云われている。

仕事や組織が変化、拡大していく場合には、自由に追加、拡張できる柔軟な仕組みが必要で、蜘蛛の巣を張るようなハイパーシステムは非常に有効であったと思う。

このシステムは、現在、WWW として、広く世界に普及している。

CERN の研究所は、世界各国から研究者や技術者、その家族が集まり、一つの街を構成していた。通りには、アインシュタインやボーア、シュレディンガー、パウリ、湯川など有名な物理学者の名前がついていて、物理学が身近な学問の街であった。周囲は、ゆるやかな丘陵地で、一面にブドウ畑が広がっていた。

３．あいまいな情報の処理

量子物理学は、よく知られているように、根底に、揺らぎ、状態の重ね合わせや量子もつれを伴う確率の世界である。しかしその不思議な曖昧さを含めて、理論と実験が支え合い厳密な体系を構築している。

CERN で、ヒッグス粒子が発見されたときは、標準模型から計算される理論値と実験値（2 光子のエネルギーの和）が比較されて、その偶然の揺らぎは、174 万回に 1 回以下に設定され（6σ 相当)、独立な検証を経てようやく新粒子と認定されたと云われている。

また、身近なコンピュータやデジタル情報システムは、アルゴリズムが決まるとプログラムに従って、同じ出力が得られるシステムである。

いずれも正確で信頼性の高い情報処理システムで、再現性は極めて高い。

これに比べると、人間の情報処理は、非常に曖昧である。情報そのものの曖昧さに加えて、その表現も処理も曖昧である。

パターン認識の研究を始めて、まず手書き数字の読み取りを始めたが、文字パターンは曖昧模糊としていて、とりつくしまもなかった。自分自身がどのように読んでいるのか分からない。

手書き数字を多数集めて、0 から 9 まで分類してみると、典型的な文字パターンからくずれて判別が難しいものまで多様であった。変形の多い “2” や “8” は、3 にも 4 にも 5 にも類似して、区別のつかないグレーゾーンのパターンが多数存在した。これは、どの文字にも存在した。

これらは解像度を上げても、観測方法を変えても存在する本質的なグレーゾーンで、人間の情報処理は、この曖昧さを含むものであった。

文字パターンには、もう一つ難しい問題があった。文字は人間が決めた約束のパターンである。そのため、同じ算用数字でも、国や民族、歴史の影響を受け、国ごと地域ごとに違いがある。

標準の字形はあるが、手書き文字では変形が大きく、はっきりした認識のアルゴリズム

を求めることは難しいと思った。

そして認識の方針は、もろもろの人間の判断を反映したデータ依存の方式とし、判断が難しいグレーゾーンは、積極的に分からないとすることにした。

郵便番号読み取り装置は、全国から集めた手書き数字 30 万字をベースに、部分的な特徴を、より大きな特徴列にまとめ上げ、変形を吸収するオートマトン型の決定グラフで認識する方式とした。

手書き文字の変形は、数億、数十億と存在する。読めない文字は、決定グラフ（認識辞書）を修正することで改善を図った。手間のかかる方式だったが、大量のデータと格闘する中で、読み取り率は向上した。

後年、認識方式は、もう少し汎用な複合類似度法（部分空間法）に改善し、数字のみならず英字や記号、漢字まで認識できるようになった。

この方式は、多数の文字データがつくる部分空間と入力の文字パターンの距離を、類似度とするもので、認識対象のパターンデータを多数集めれば、パターンの統計的なばらつきや曖昧さを内包した認識辞書を自動的につくることができる。

手書き文字は、国や地域によって違いがある。ヨーロッパの数字、例えば、“1”、“7”、“9”は、カギや独特の丸みがある文字で、またアルファベットも地域固有のクセがある文字パターンが多い。これは、各国の歴史や教育が反映している。

近年、上記の郵便番号読み取り装置は、世界各国に輸出されているが、国ごとの違いは、文化を反映したデータを多数集めることにより解決されている。

この装置は、その後、より複雑で多様な郵便書状の住所の読み取りもできるようになり、郵便システムの自動化は大きく進んだ。

住所の記載方法は、規約がないので、慣習を反映した統計的、AI 的な（住所の知識など）工夫が必要であった。

我が国では、宛名は、手書きまたは活字の漢字や、かな、英数字の混じり文で、縦書きと横書きが混在する。封書も、色付きや窓付き、ラベルなど多様なので、欧米に比べると、技術開発は厳しい環境であったが、これを乗り越えると、極めて強い技術と製品が出来上がる。この装置は、幸い、世界で活躍している。

パターンは、自然界の至る所に存在し、人間はそれを情報の基本単位として認識しているように思う。画像や音、匂いや触覚、味覚の中に、パターンを感じる。また、行動や思考の中にもパターンがあり、それを感知できる。

パターンは、型や部分的なまとまりであるが、その中で規則性や再現性の高いものが、法則となり、さらに方程式になるように思う。

パターンは、情報を担う媒体であるが、はっきりとした意味を担うもの、意味のないもの、曖昧なもの、また多様な意味を内包するものがある。

パターンを中心とした意味の情報処理は、これからの重要課題であると思う。

パターン認識では、脳の構造を模擬して認識を行おうとする研究がある。古くからあるニューラルネットワーク（Neural Network 神経回路網）と呼ばれる分野で、最近では、ディープラーニング（Deep Learning）技術が、画像認識などで良い成果を出し、注目を集めている。

入力と出力との間に多層の中間層（数十～百数十層）を設け、オートエンコーダという技術により情報圧縮と特徴抽出を行い、ある種の認識構造を構築できるようになったと見なされている。

しかし解析が難しい複雑系なので、分からないことが多く、研究途上の分野であると思う。

パターン認識は、人間（生物）が持つ本質的な機能である。人間はそれを基本として、情報のやりとりをしている。その情報は、曖昧で不正確であるが、繰り返しと学習の中で、意味を確かめつつ、揺らぎのある曖昧な生活を楽しんでいるのかも知れない。

５．まとめ

ファンデルバール先生、CERN の話、パターン認識と曖昧な情報処理の話、人工知能（AI）に繋がる話を、思い出すままに書いてみた。

人間にとって、曖昧な情報処理は大切である。人間を含む情報処理システムは、曖昧さを容認する体系でありたいと思っている。