

Information Theory in Visual Analytics

Min Chen

Professor of Scientific Visualization

including recent work in collaboration with
Amos Golan, American University, USA

Oxford e-Research Centre
University of Oxford



UNIVERSITY OF
OXFORD

<http://www.bslhands4u.com/fingerspelling/4545036827>

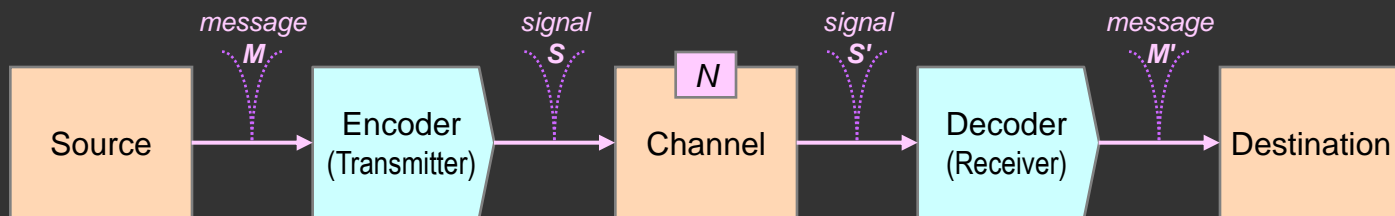
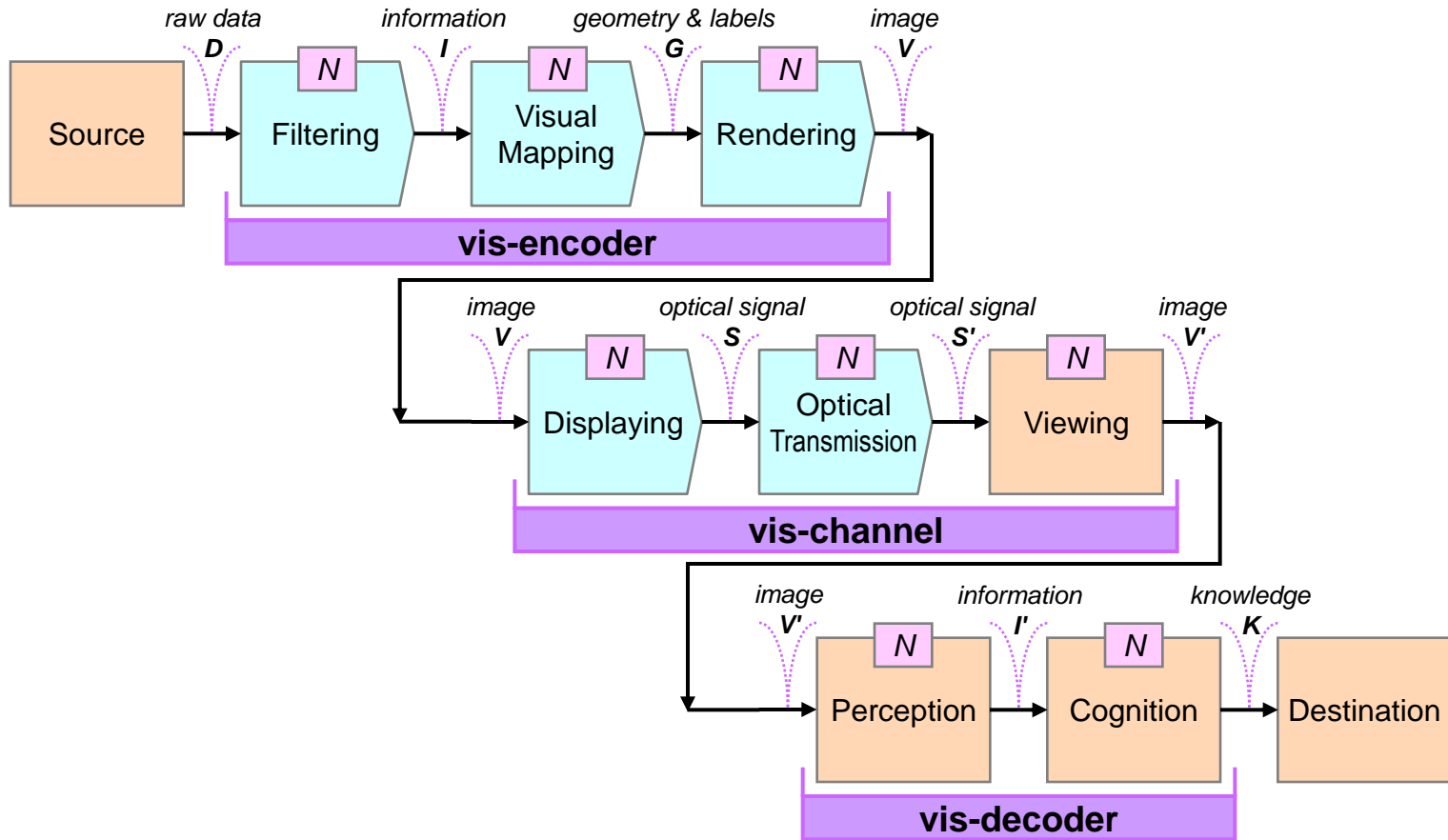
<http://www.infoplease.com/ipa/A0200808.html>



min.chen@oerc.ox.ac.uk

The 41st CREST Open Workshop
UCL, London, 27-28 April 2015

Three Visualization Subsystems



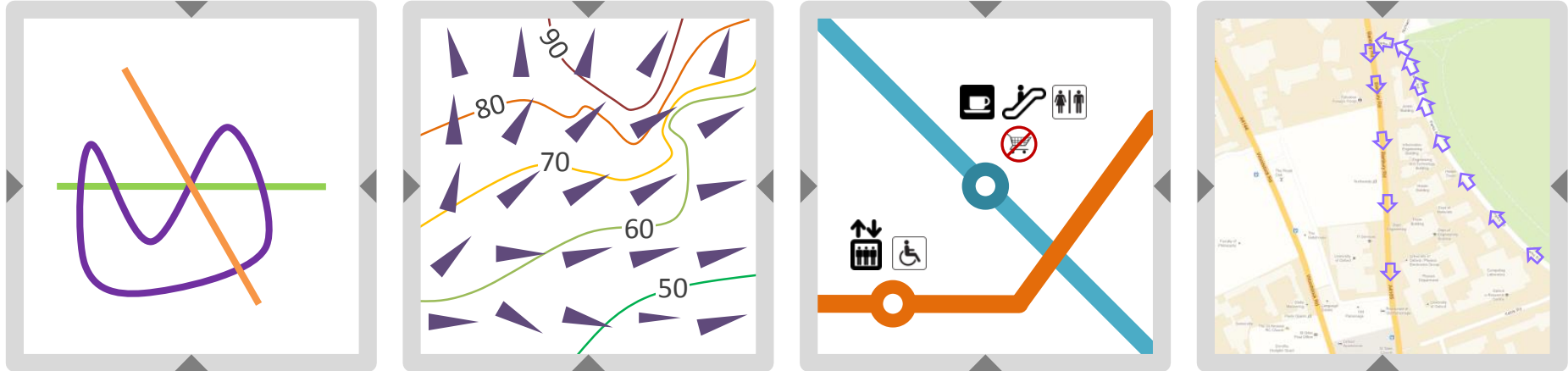
A General Communication System

Existing Uses of Information Theory

- Data processing
- View optimization
- Glyph design
- ...

- Theoretical framework
 - *Measuring visualization capacity, and related quantities*
 - *Explaining phenomena in visualization processes*
 - *Defining laws (mathematically-validated guidelines)*
 - *Defining algorithm- or data-driven metrics*
 - *Confirming the significance of visual analytics*

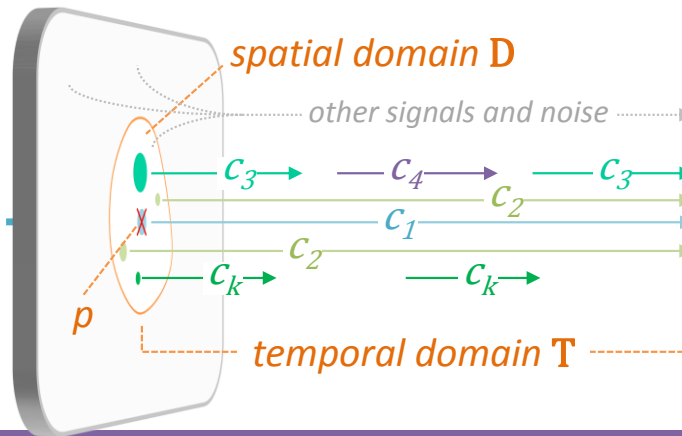
Example: Visual Multiplexing



Location p can be associated with X in the source data or determined by a spatial mapping.

$X = \langle x_1, x_2, \dots, x_k \rangle$ at p

MUX



Perceived information may include estimated values and relationships with data conveyed by other signals.

DEMUX

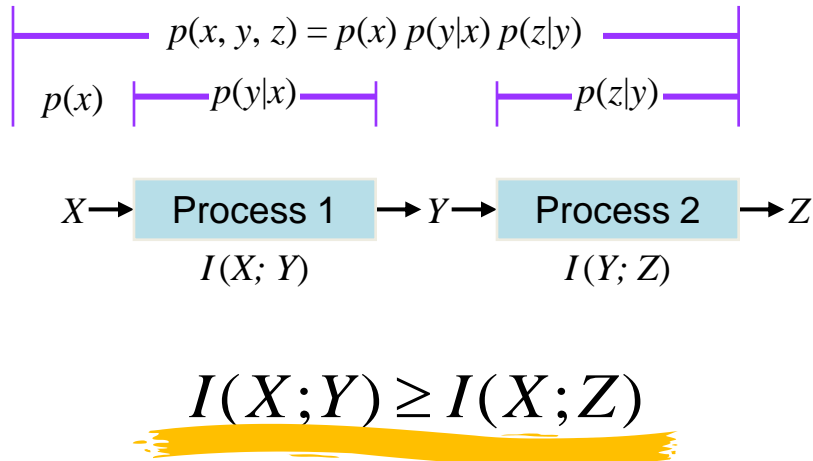
information about X at p

vis-encoder

vis-link (consisting of many vis-channels)

vis-decoder

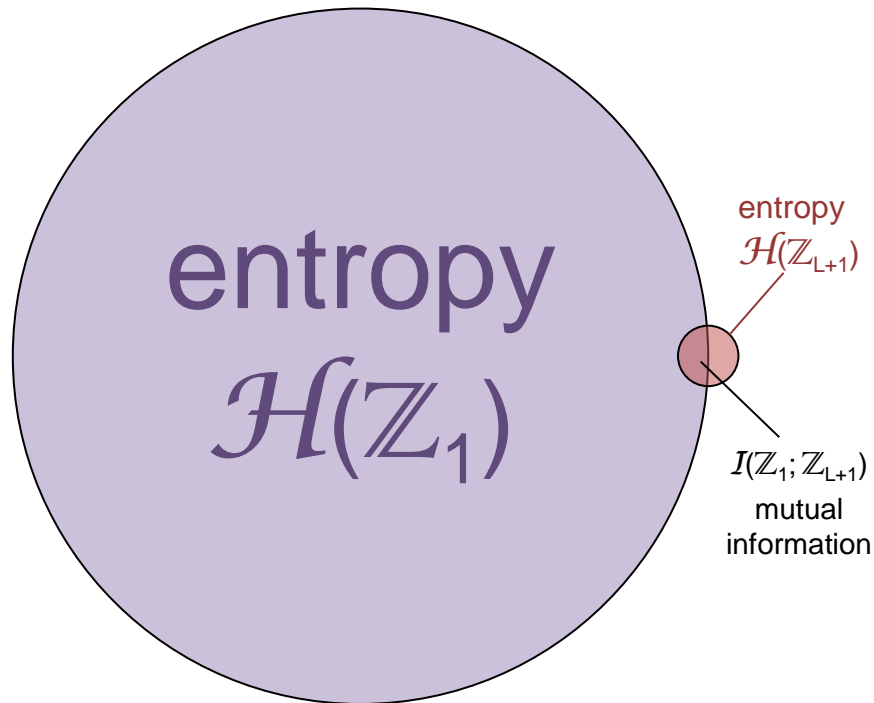
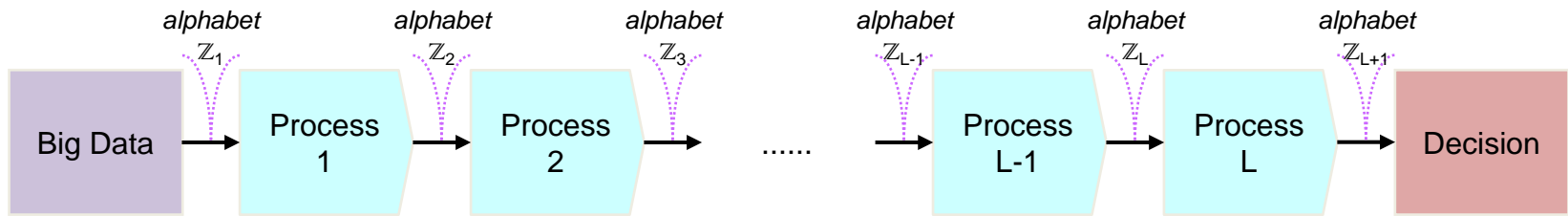
Data Processing Inequality



- “No clever manipulation of data can improve the inferences that can be made from the data”
[Cover and Thomas, 2006]

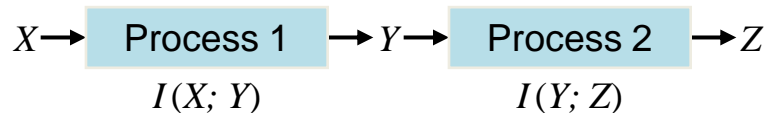


Data Processing Inequality: Big Data Input?

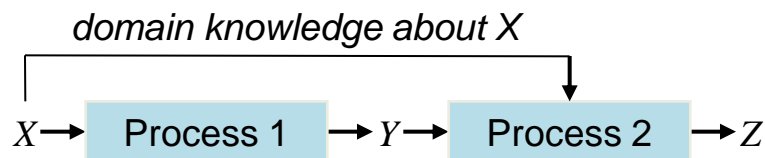
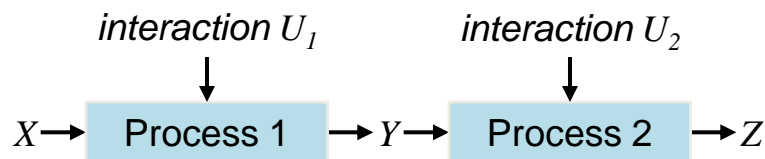


$$p(x, y, z) = p(x) p(y|x) p(z|y)$$

Diagram illustrating the joint probability distribution $p(x, y, z)$ as a product of marginal and conditional probabilities: $p(x)$, $p(y|x)$, and $p(z|y)$.



$$I(X; Y) \geq I(X; Z)$$

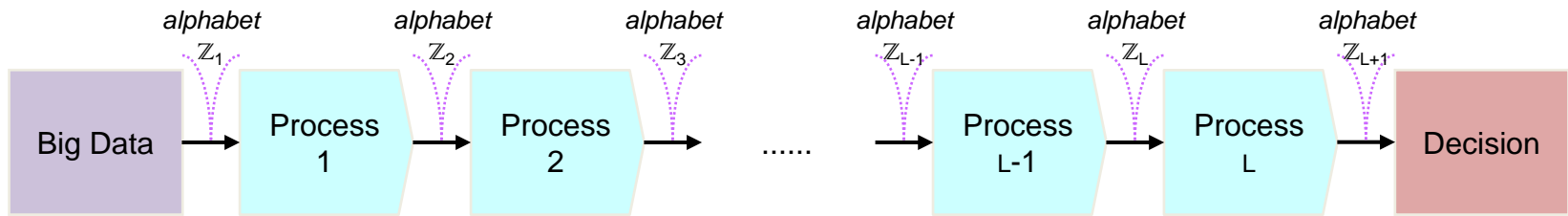


$$I(X; Y) \not\geq I(X; Z)$$

DPI is not Ubiquitous

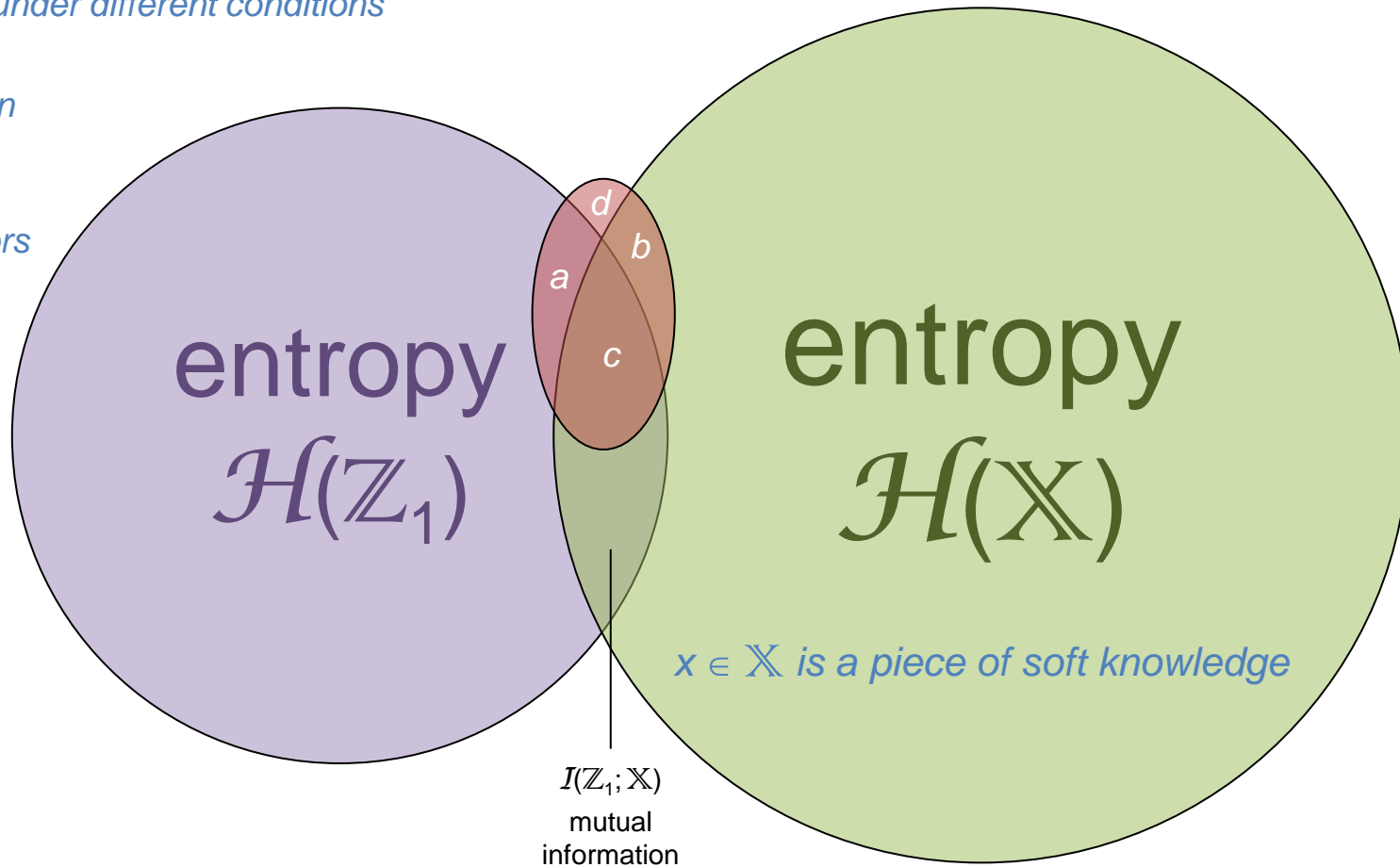
- Markov chain conditions
 - *Closed coupling: $(X, Y), (Y, Z)$*
 - *X and Z are conditionally independent*
- What if one of the conditions is broken?
- In visual analytics, both conditions are usually broken.

Soft Knowledge in Decision Space



All possible decisions under different conditions

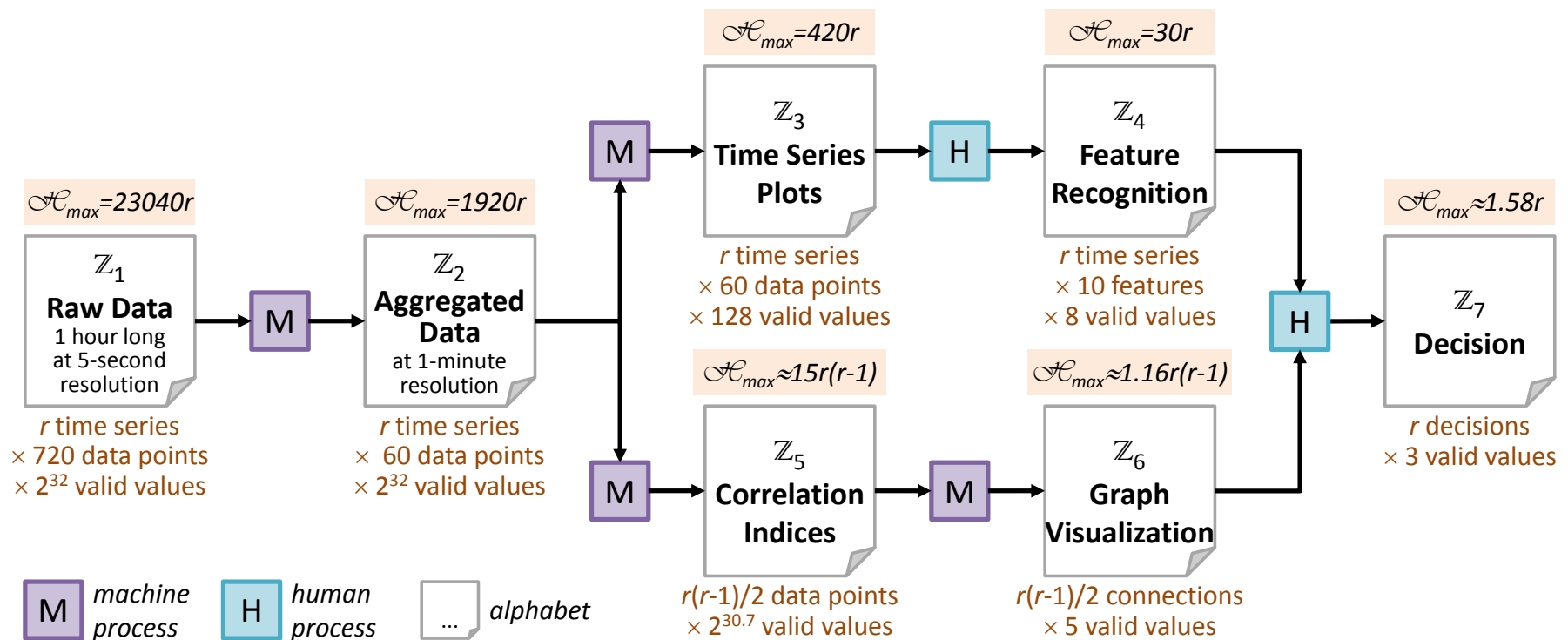
- a) *totally data-driven*
- b) *totally instinct-driven*
- c) *data-informed*
- d) *due to unknown or uncontrollable factors*



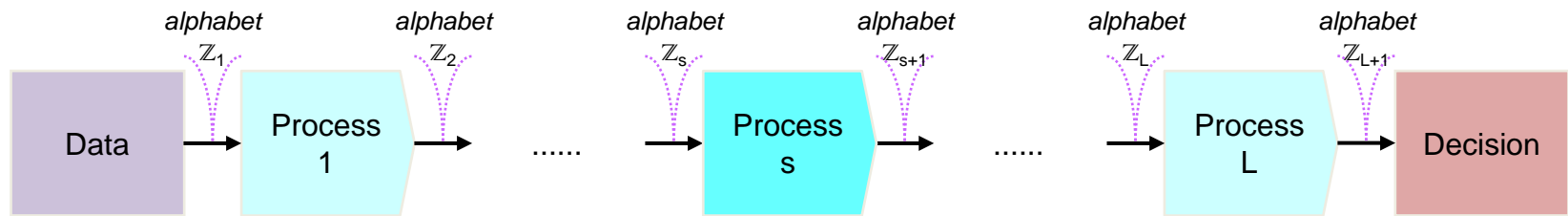
An Example Data Analysis and Visualization Process

- r time series
- 720 data point each series
- 2^{32} valid value each point

- r decisions
- 3 valid values each (e.g., buy, sell, hold)



A Sequential Workflow and Two Basic Metrics



- The s^{th} Function (Process):

$$F_s : \mathbb{Z}_s \longrightarrow \mathbb{Z}_{s+1}$$

- Alphabetic Compression Ratio (ACR):

$$\Psi_{ACR}(F_s) = \frac{\mathcal{H}(Z_{s+1})}{\mathcal{H}(Z_s)}$$

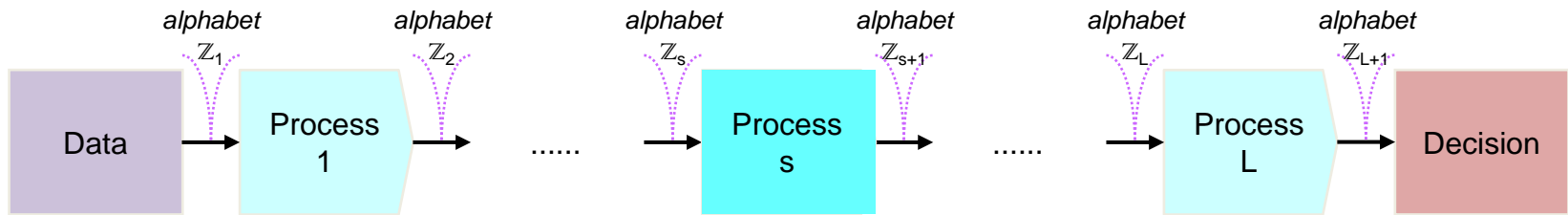
- A Reverse “Guessing” Process:

$$G_s : \mathbb{Z}_{s+1} \longrightarrow \mathbb{Z}'_s$$

- Potential Distortion Ratio (PDR):

$$\Psi_{PDR}(F_s) = \frac{\mathcal{D}_{KL}(Z'_s || Z_s)}{\mathcal{H}(Z_s)}$$

Cost-Benefit Ratio



- Effectual Compression Ratio (ECR):

$$\Psi_{ECR}(F_s) = \frac{\mathcal{H}(Z_{s+1}) + \mathcal{D}_{KL}(Z'_s || Z_s)}{\mathcal{H}(Z_s)}$$

- Incremental Cost-Benefit Ratio (ICBR):

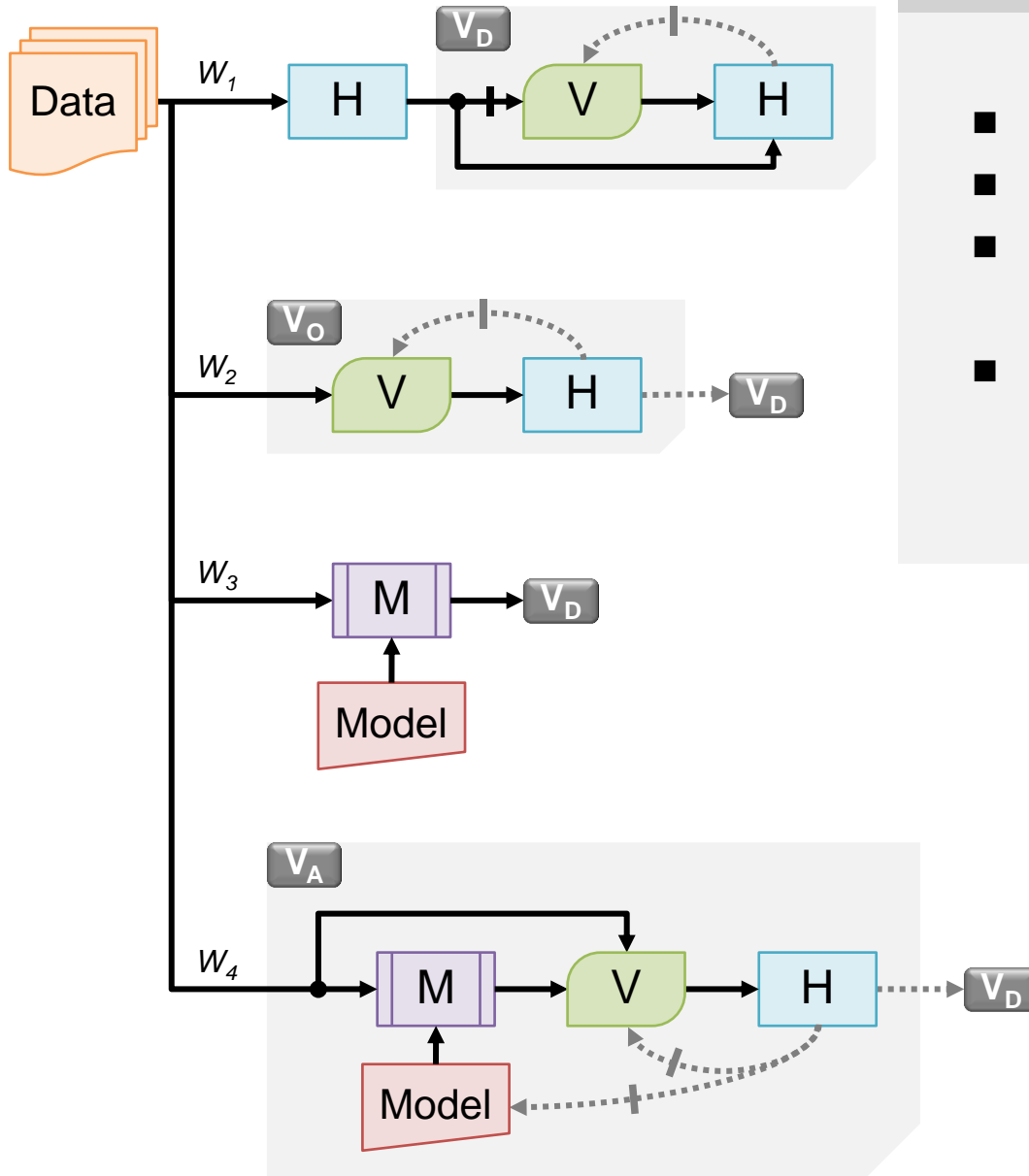
$$\Upsilon(F_s) = \frac{\mathcal{B}(F_s)}{\mathcal{C}(F_s)} = \frac{\mathcal{H}(Z_s) - \mathcal{H}(Z_{s+1}) - \mathcal{D}_{KL}(Z'_s || Z_s)}{\mathcal{C}(F_s)}$$

- *Cost can be measured in energy, time, money, etc.*

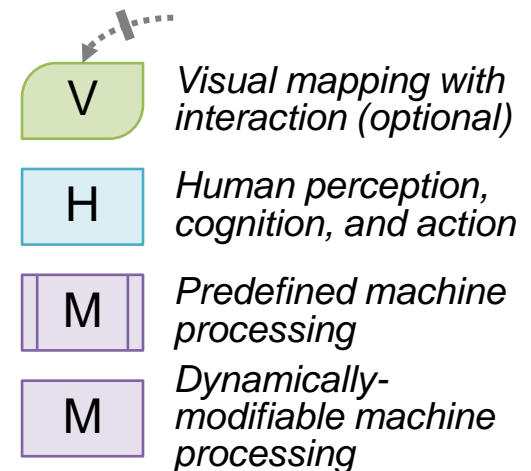
Four Levels of Visualization

1. Disseminative Level (This is A!)
 - *A presentational aid for disseminating information or insight to others.*
 - *The creator does not expect to gain much new knowledge.*
2. Observational / Operational Level (What, when, where?)
 - *An operational aid that enables intuitive and/or speedily observation of captured data. Often part of routine operations.*
 - *Confirmatory observation, anomaly detection., etc.*
3. Analytical Level (Does A relate to B? Why)
 - *An investigative aid for examining and understanding complex relationships (e.g., correlation, causality, contradiction).*
 - *Evaluating hypotheses, models, methods, algorithms and systems.*
4. Model-developmental Level (How does A lead to B?)
 - *A developmental aid for improving existing models, methods, algorithms and systems, as well as the creation of new ones.*

Levels 1, 2, 3

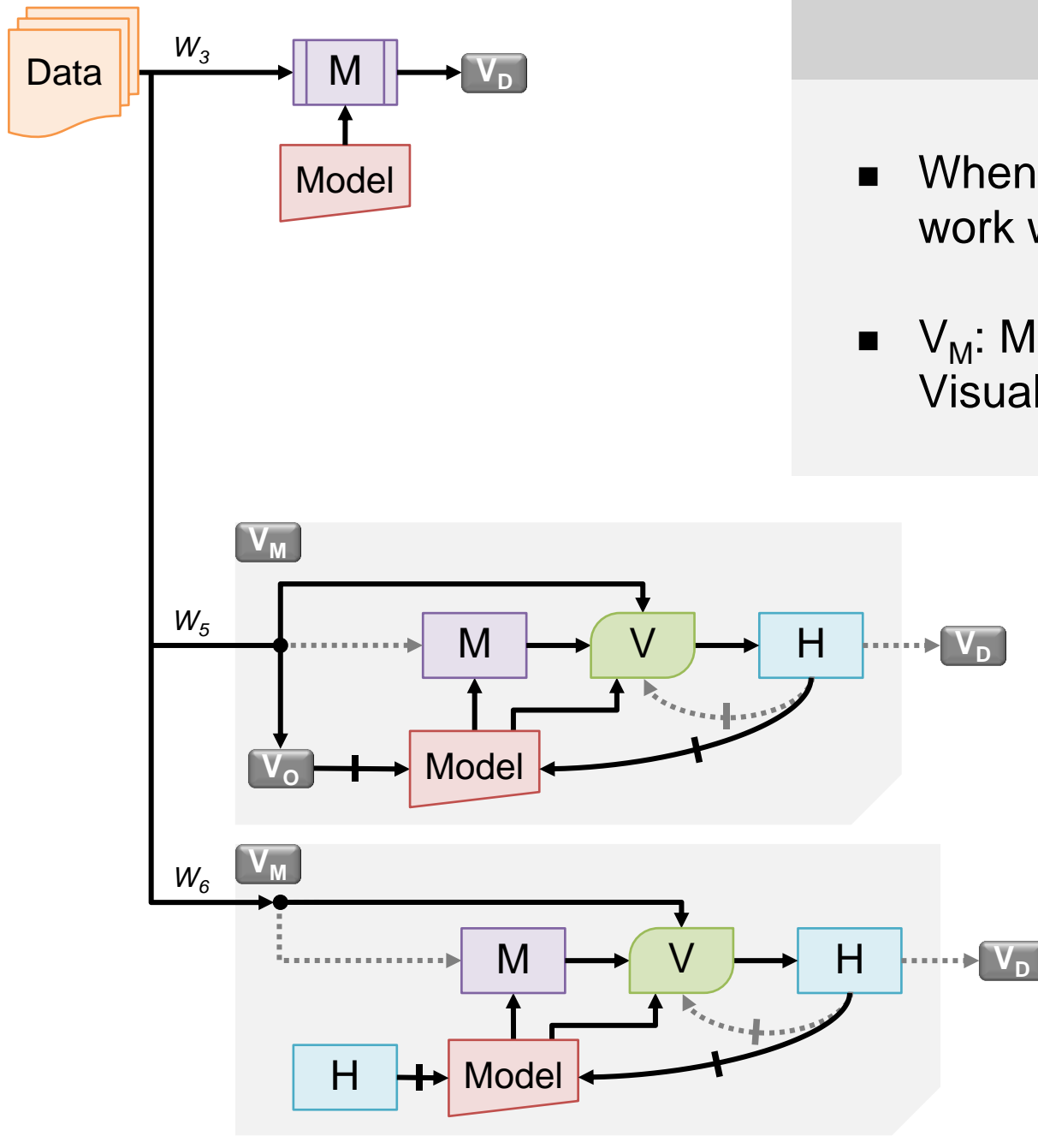


- V_D: Disseminative Visualization
- V_O: Observational Visualization
- V_A: Analytical Visualization
- When will workflow W₃ work and when will not?



Level 4

- When workflow W_3 does not work well, then ...
- V_M : Model-developmental Visualization



Example: Level 1

- Entropy of Data Alphabet

$$H(Z) = - \sum_{t=0}^{64} \sum_{i=0}^{255} \frac{1}{256} \log_2 \frac{1}{256} = 512$$

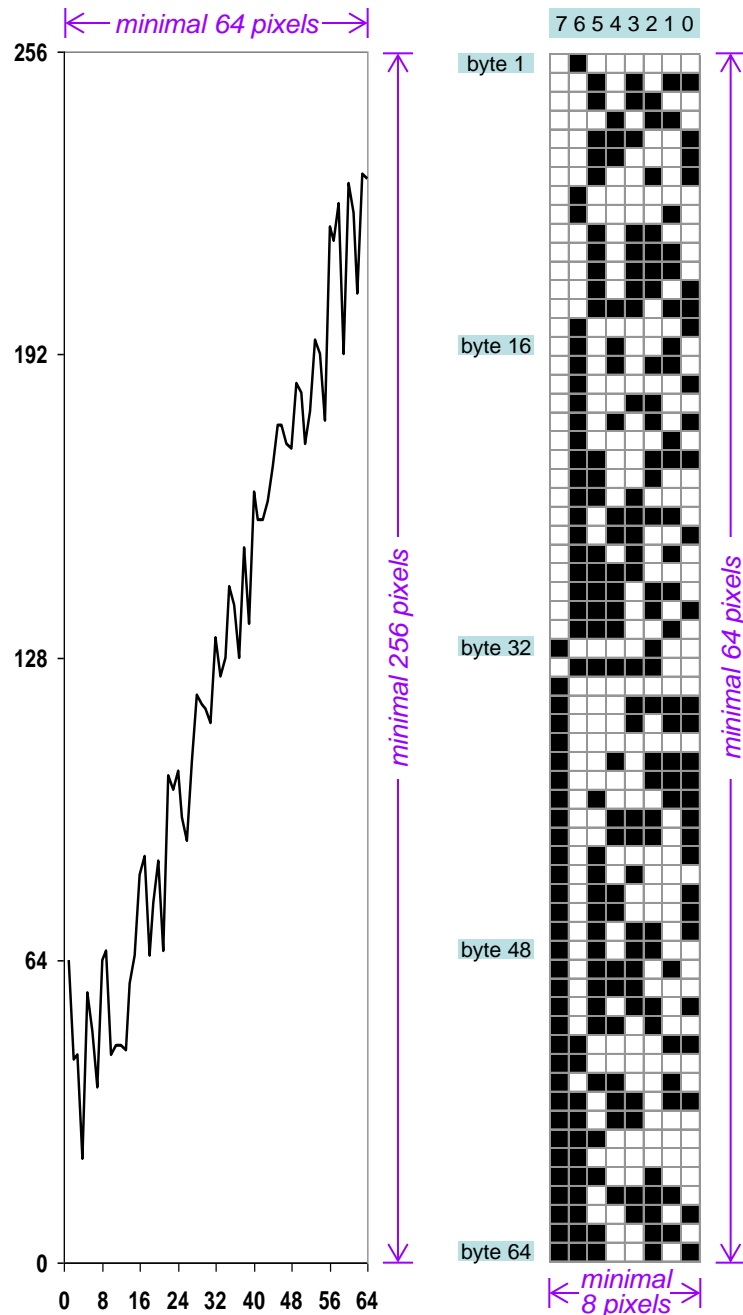
- Binary Pixel Plot

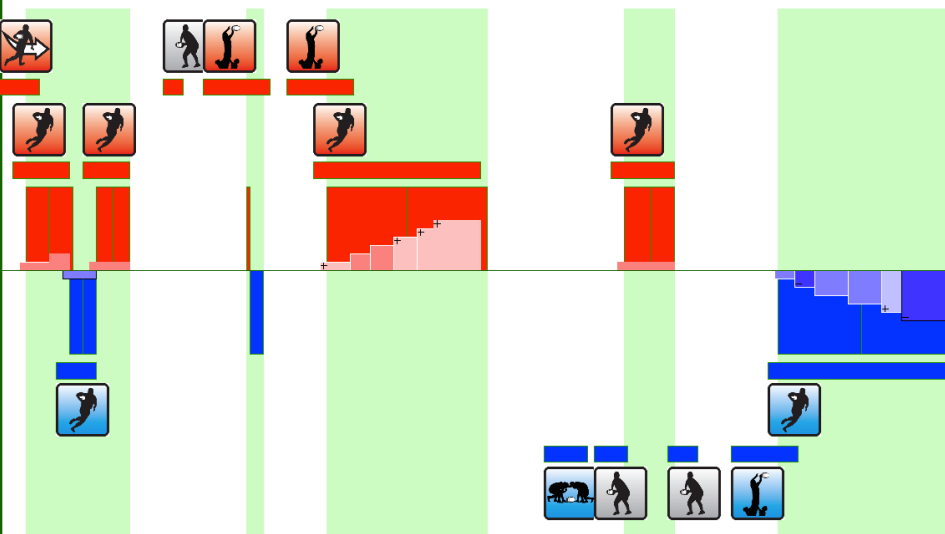
- *4x4 pixels per bit $\rightarrow 2^{13}$ bits*

- Time Series Plot

- *Minimal 256x64 pixels (2^{14} bits)*

- The more compact, the better?





Example: Level 2

- Real-time or offline annotation results in a huge spreadsheet of events

Wales' Rugby World Cup team using Swansea Uni app

Welsh coaches are using technology developed for them by experts at Swansea University to improve their match analysis at the Rugby World Cup.

Coach Warren Gatland and his backroom team are bombarded with statistics about scrums, line-outs and tackles during each game.

Now they are using an 'app' to simplify the information they need and understand the team's performance.

It also allows the coaches to review video of key moments during play.

The Welsh team employs three analysts whose job it is to collate data about all aspects of each game from the set pieces and restarts to tackles made or missed.

Dr Philip Legg of Swansea University's college of engineering and department of computer science said the problem was coaches were suffering from an "information overload".

The university was approached by the Welsh Rugby Union (WRU) and has developed an app it calls the MatchPad which runs on an Apple iPad.

It produces a visual timeline during the game so analysts and coaches can review video and additional detail on the events they are most interested in simply by pressing an icon.

The portability of the device means they can access the information in the analysis box, in the changing rooms, or even at pitch-side.

Dr Legg said: "During each game the team analysts are busy recording each of the events that happen."

"They look at each scrum, line out, restart, possession won and lost and tackles."

"They collect so much information - that's the basic problem and the app just tries to simplify."

Top Stories

- Italy's Berlusconi faces key vote
- Lloyds hit by cost of PPI claims
- Boxing heavyweight Frazier dies
- Sarkozy called Israeli PM 'liar'
- Doctor guilty over Jackson death

Features & Analysis

- Staging da Vinci**
How the National Gallery's blockbuster came together
- Patriotic purchase**
Why aren't people told to buy British?
- Break down**
Who is really profiting from the high cost of petrol and oil?
- Brain teaser**
Debunking the myths about our minds

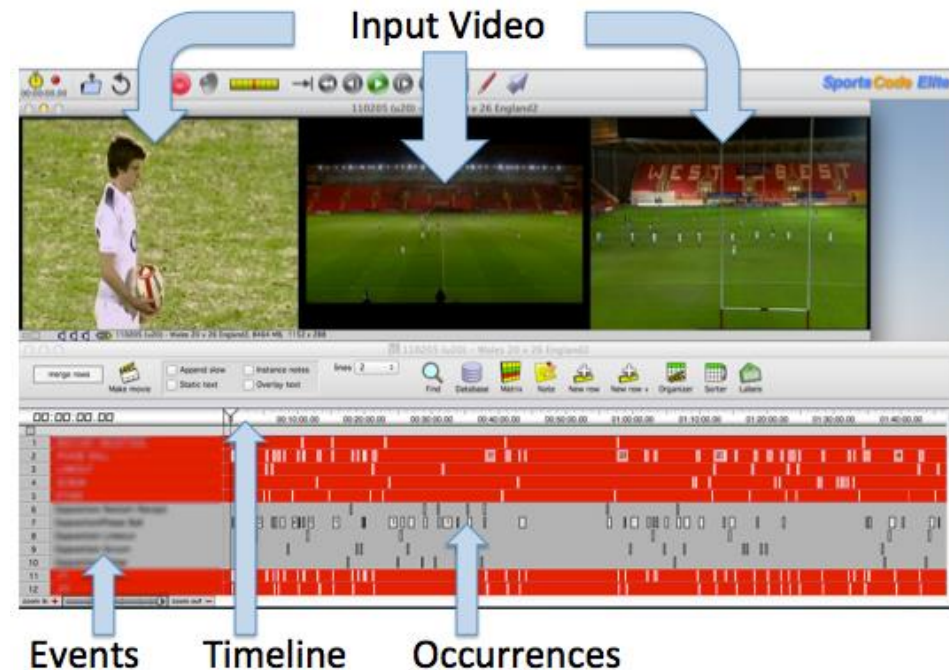
Related Stories

- Wales 'strongest' Euro cup hope
- Fiji 0-66 Wales
- Wales 61-7 Namibia

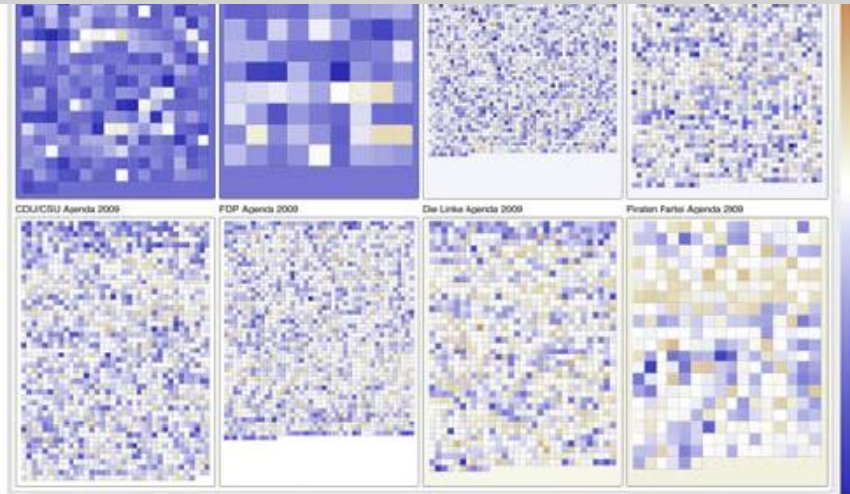
Most Popular

Shared	Read	Video/Audio	Rank
Sarkozy called Israeli PM 'liar'			1
Boxing heavyweight Frazier dies			2
Giant asteroid to pass near Earth			3
Prostitutes found in Mexico jail			4
Why aren't people being told to buy British?			5

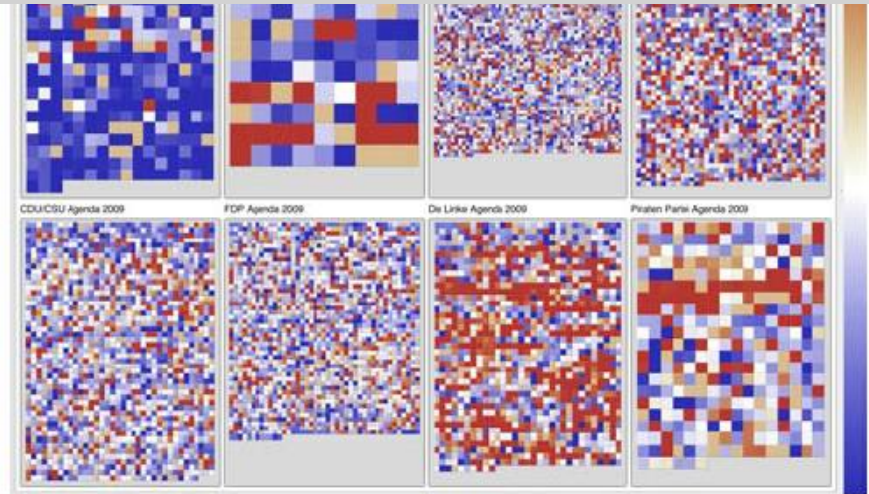
Legg *et al.*, "MatchPad: interactive glyph-Based visualization for real-time sports performance analysis," *Computer Graphics Forum*, 2012



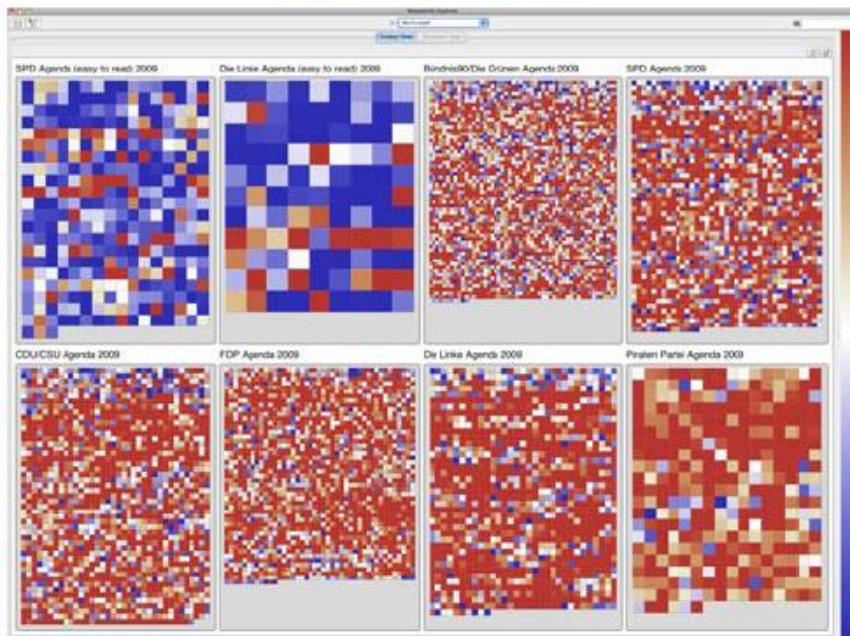
Example: Level 3



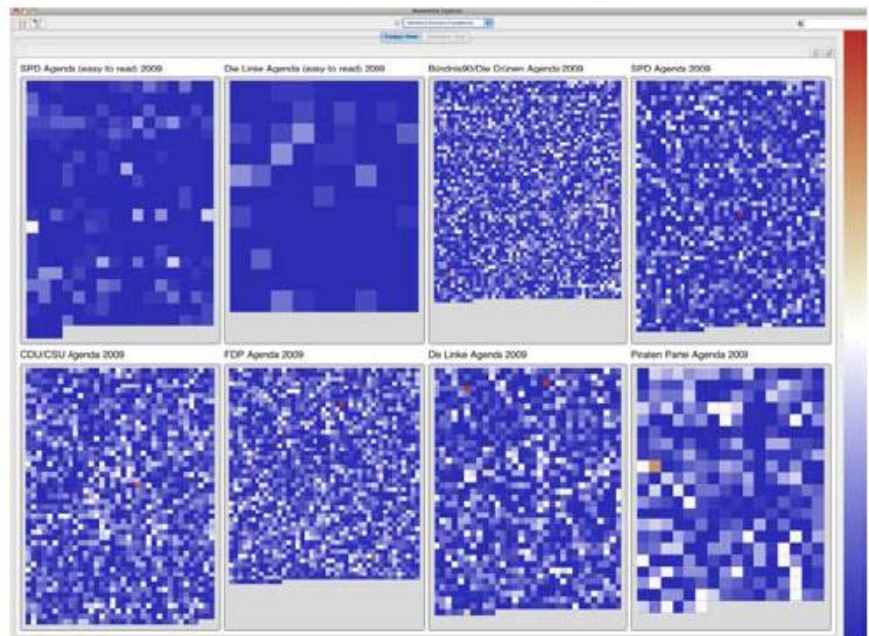
(a) Average Readability Score



(b) Feature: Vocabulary Difficulty

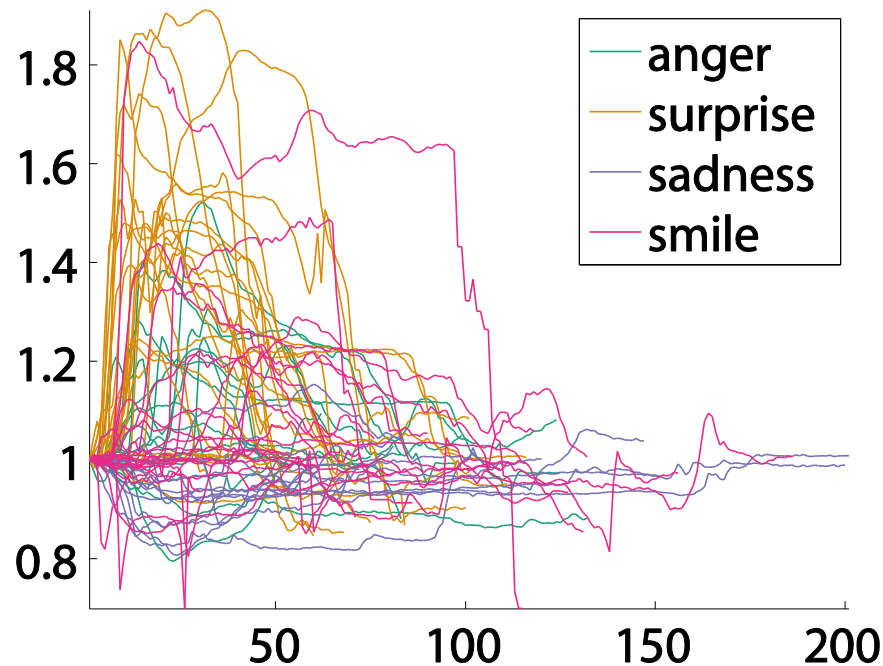
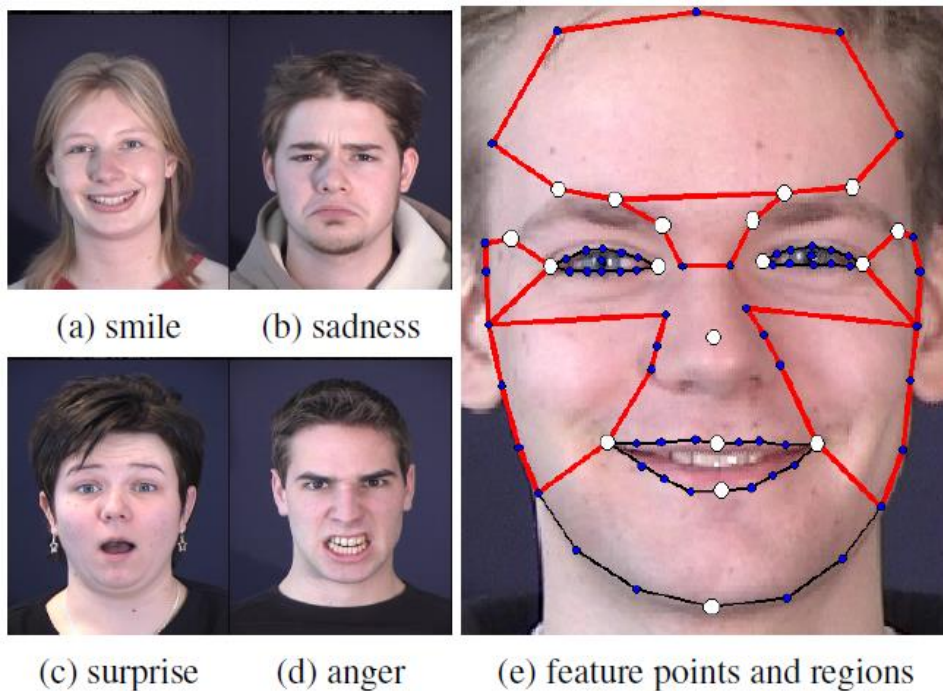


(c) Feature: Word Length



(d) Feature: Sentence Structure Complexity

Example: Level 4

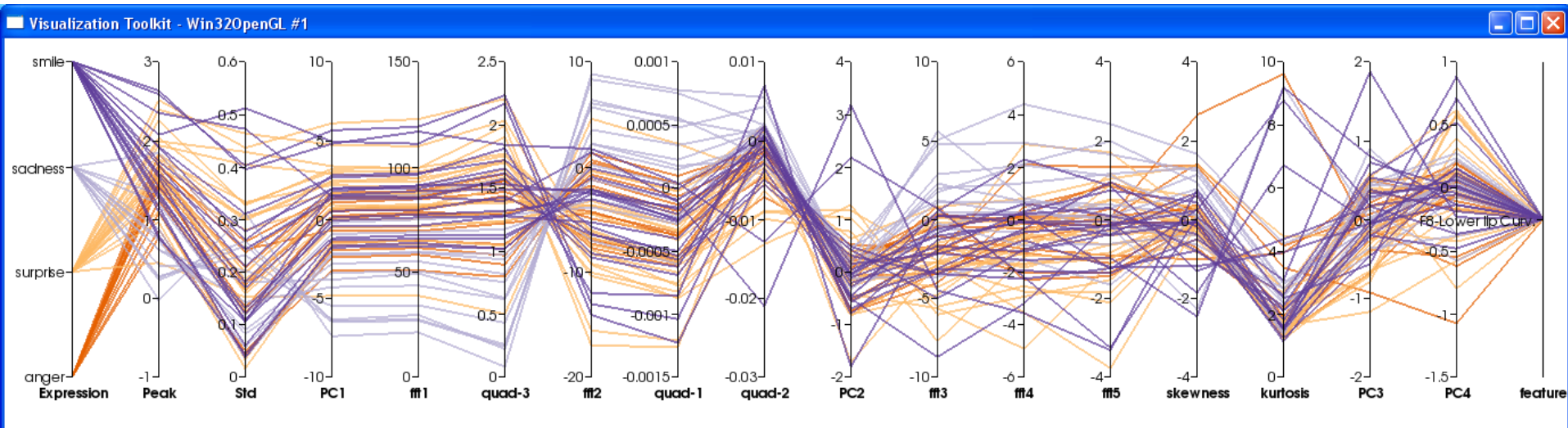
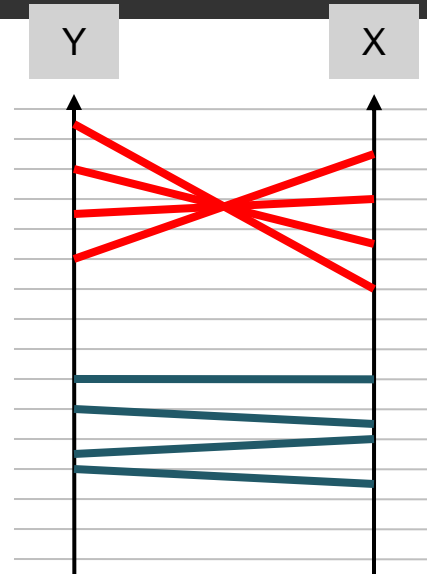
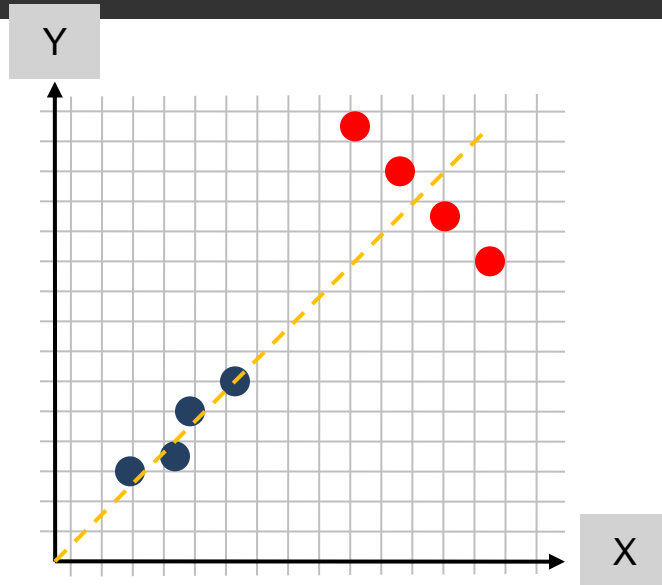


- Expression Recognition
 - *Humans are very good at*
 - *Machine vision is far behind*
 - *Limited understanding*
- Data
 - *Video*
 - *Feature changes*
 - *Time series*
- Challenges
 - *A lot of features*
 - *A lot of ways of measuring features*
 - *Non-uniform temporal behavior*

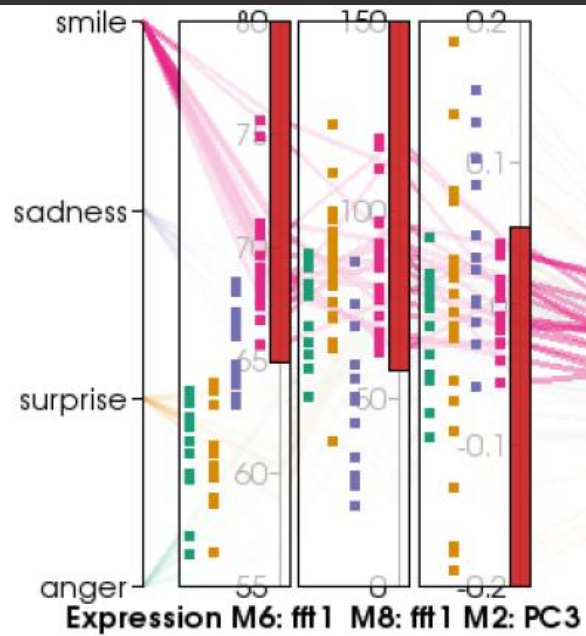
Tam *et al.*, "Visualization of time-series data in parameter space for understanding facial dynamics,"
Computer Graphics Forum, 2011

Parallel Coordinates

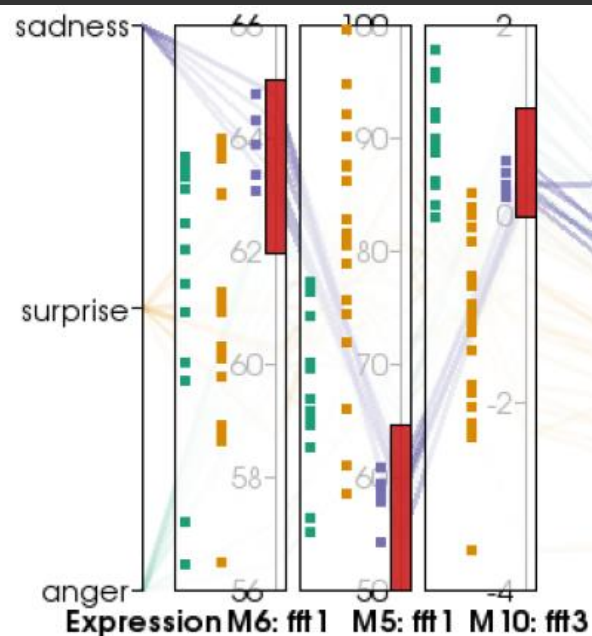
■ Multi-dimensional data visualization



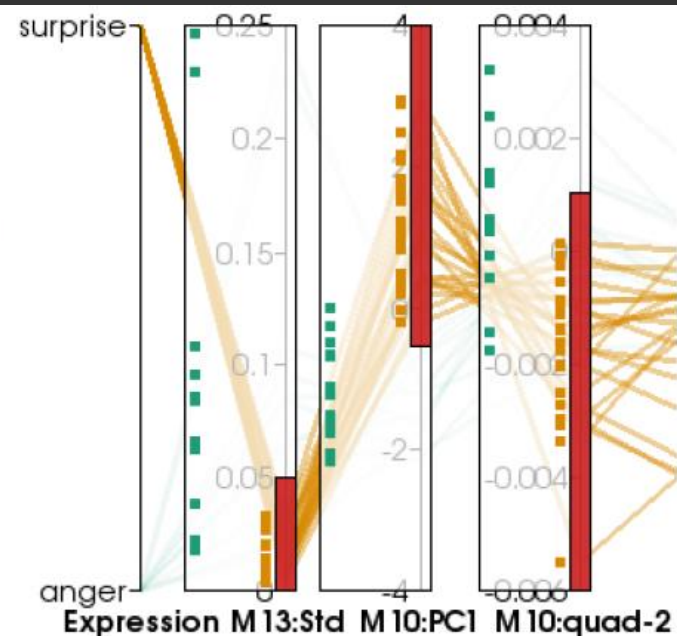
Interactive Visualization: Formulating Decisions



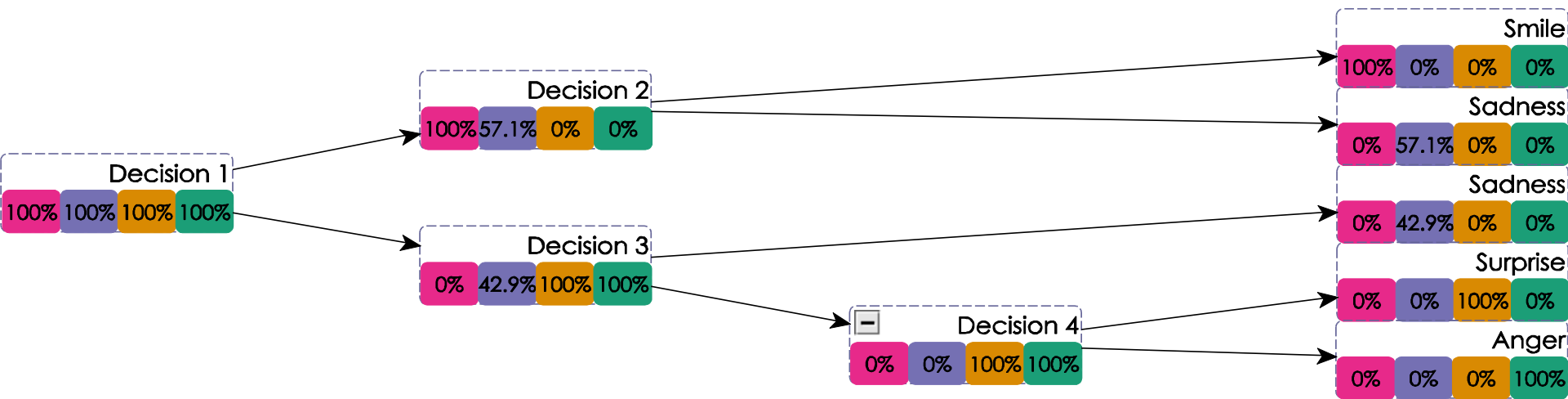
(a) Separate smile from all expressions



(b) Identifying sadness at Decision 3



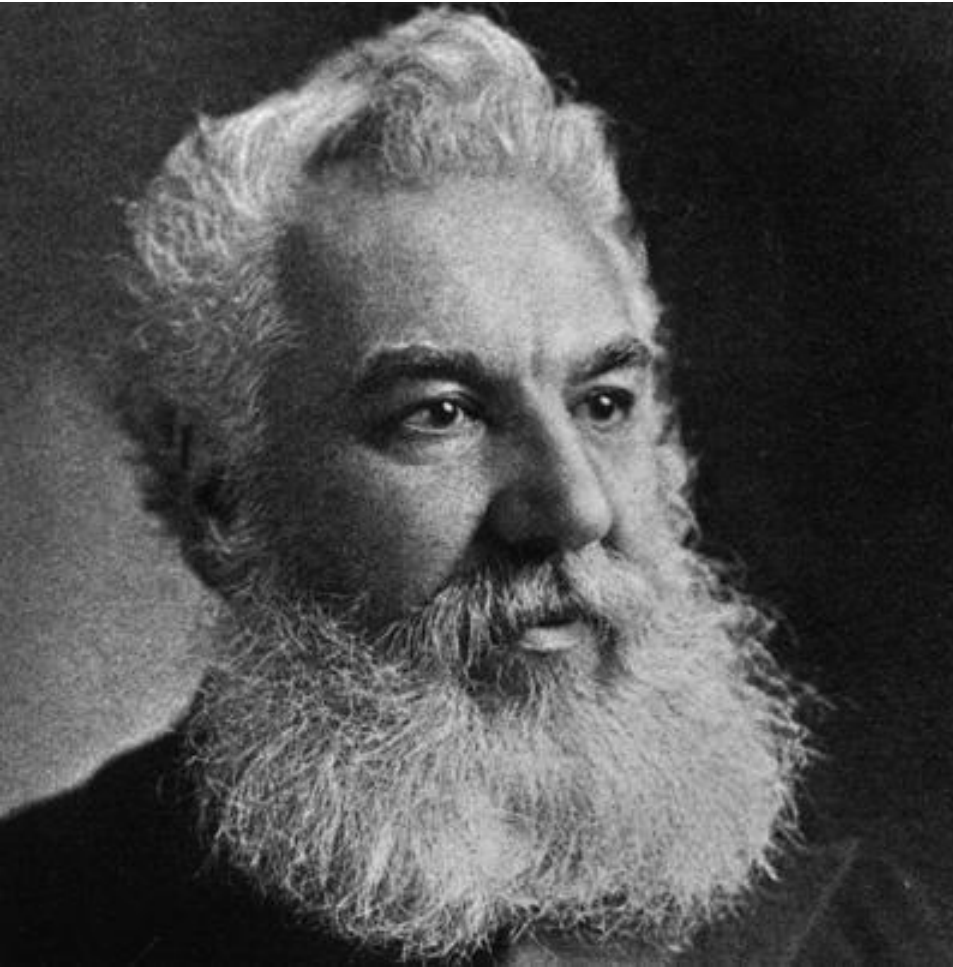
(c) Identifying surprise at Decision 4



Telephone



- In the 1870s, Bell travelled around to give demos 'in concert halls, where full orchestras and choruses played "America" and "Auld Lnag Syne" into his gadgetry.'
- Around 1880, Queen Victoria installed a pair of telephones at Winsor and Buckingham Palace



Alexander Graham Bell (1847-1922)

- Had Alexander Bell invented visualization, he would probably have said:

“Mr. Information,
come here.
I want to see you.”

Acknowledgement

University of Oxford

- Hui Fang
- Saiful Khan
- Simon Walton
- TBA: EPSRC/Airbus Studentship
- *Colleagues in OeRC, OCCAM, Oii, CompSc, EngSc, ...*

Swansea

- Rita Borgo
- Phil W. Grant
- Iwan Griffiths
- Mark W. Jones
- Bob Laramie
- Adrian Morris
- Tavi Murray
- Irene Reppa
- Kilian Scharrer
- Ian Thornton
- *ROs and PhDs (below)*

Stuttgart

- Tom Ertl
- Daniel Weiskopf
- Ralf Botchen ...

Rutgers

- Deborah Silver
- Carlos Correa

Purdue (VACCINE)

- David Ebert

Heidelberger

- Heike Jänicke

Utah

- Chris Johnson, Kate Coles, Julie Lein, Miriah Meyer
- Chuck Hansen

Cardiff

- Andrew Aubrey
- Dave Marshall
- Paul Rosin
- Gary Tam

RIVIC

- Nigel John
- Ralph Martin
- Reyer Zwiggelaar

Past PhDs and ROs:

- C.-Y. Wang (PhD, 1989-1992)
- Mark W. Jones (PhD, 1991-1994)
- Abdula Haji Tablib (PhD, 1990-1994)
- Mike Bews (PhD, 1992-1996)
- Malcolm Price (MPhil, 1997-1998)
- Adrain Leu (PhD, 1996-1999)
- Simon Michael (PhD, 1996-1999)
- Steve Treavett (PhD, 1997-2000)
- Mark Kiddell (RA, 1999-2001)
- Ben Smith (TCA, 1999-2001)
- S.-S. Hong (PhD, 1998-2002)
- Abdul Haji-Ismail (PhD, 1998-2002)
- H.-L. Zhou (MPhil, 2000-2002)
- Andrew S. Winter (PhD, 1999-2002)
- David Rogeman (PhD, 1999-2003)
- Paul Adams (TCA, 2002-2004)

- Tim Lewis (RA, 2004-2005)
- Gareth Daniel (PhD, 2001-2004)
- David P. Clark (PhD, 2001-2005)
- Dave Bown (RA, 2005)
- Ann Smith (PhD, RA, 2001-2006)
- Siti Z. Zainal Abidin (PhD, 2003-2007)
- Alfie Abdul Rahman (PhD, RA, 2004-7)
- Joanna Gooch (PhD, 2004-2007)
- Shoukat Islam (PhD, RA, 2004-2009)
- David Chisnall (PhD, RA, 2005-2008)
- Phil Roberts (RA, 2005-2008)
- Rudy R. Hashim (PhD, 2005-2008)
- Dan Hubball (MPhil, 2007-2008)
- Owen Gilson (PhD, 2006-2009)
- Lindsey Clarke (PhD, 2007-2010)
- Heike Jänicke (RO, 2009-2010)
- Farhan Mohamed (PhD, 2008-)
- Ed Grundy (PhD, 2009-)

- Rita Borgo (2009-2011)
- Hui Fang (2009-2011)
- Yoann Drocourt (PhD, 2010-2011)
- Karl Proctor (PhD, 2009-2011)
- Andrew Ryan (PhD, 2010-2011)
- Phil Legg (RO, 2010-2011)
- David Chung (PhD, RA, 2010-2011)
- Matthew Parry (MPhil, RA, 2010-2011)
- Richard M. Jiang (RO, 2010-2011)
- Brian Buffry (RO, 2011, 2013)
- Kai Berger (RO, 2012-2013)
- Karl Proctor (RO, 2012-2013)
- Jeyan Thiyagalingam (RO, 2013)
- Eamonn Maguire (PhD, RO, 2011-15)
- Alfie Abdul-Rahman (RO, 2012-2015)