

Quantum entropy

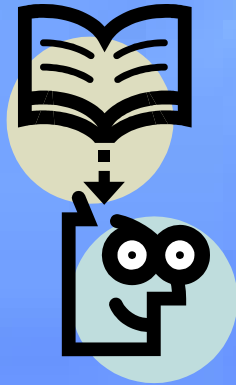
Michael A. Nielsen

University of Queensland

Goals:

1. To define entropy, both classical and quantum.
2. To explain data compression, and its connection with entropy.
3. To explain some of the basic properties of entropy, both classical and quantum.

What is an information source?

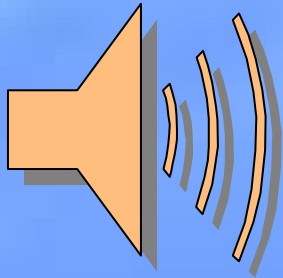


011000101110011100101011100011101001011101000

We need a **simple** model of an information source.

The model might not be realistic, but it should give rise to a **theory of information** that can be applied to realistic situations.

Discrete iid sources



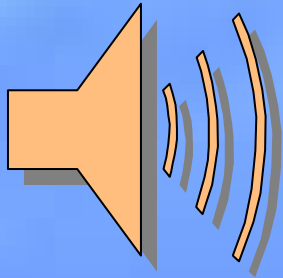
01100010111001110010101110001...

Definition: Each output from a **discrete information source** comes from a finite **set**.

We will mostly be concerned with the case where the alphabet consists of **0** and **1**.

More generally, there is no loss of generality in supposing that the alphabet is $0, \dots, n-1$.

Discrete iid sources



01100010111001110010101110001...

We will model sources using a **probability distribution** for the output of the source.

Definition: Each output from an **iid (independent and identically distributed) source** is independent of the other outputs, and each output has the same distribution.

Example: A sequence of coin tosses of a biased coin with probability p of heads, and $1-p$ of tails.

More generally, the distribution on alphabet symbols is denoted p_0, p_1, \dots, p_n .

What other sources are discrete iid?

Most interesting sources are not.

“What a piece of work is a man! how noble in reason! how infinite in faculties! in form and moving how express and admirable! in action how like an angel! in apprehension how like a god! the beauty of the world, the paragon of animals! And yet to me what is this quintessence of dust?”

However, lots of sources can be **approximated** as iid - even with English text this is not a bad approximation.

Many sources can be described as **stationary, ergodic** sequences of random variables, and similar results apply.

Research problem: Find a good quantum analogue of “stationary, ergodic sources” for, and extend quantum information theory to those sources.
(Quantum Shannon-Macmillan-Breiman theorem?)

How can we quantify the rate at which information is being produced by a source?

Two broad approaches

Axiomatic approach: Write down desirable axioms which a measure of information "should" obey, and find such a measure.

Operational approach: Based on the "fundamental program" of information science.

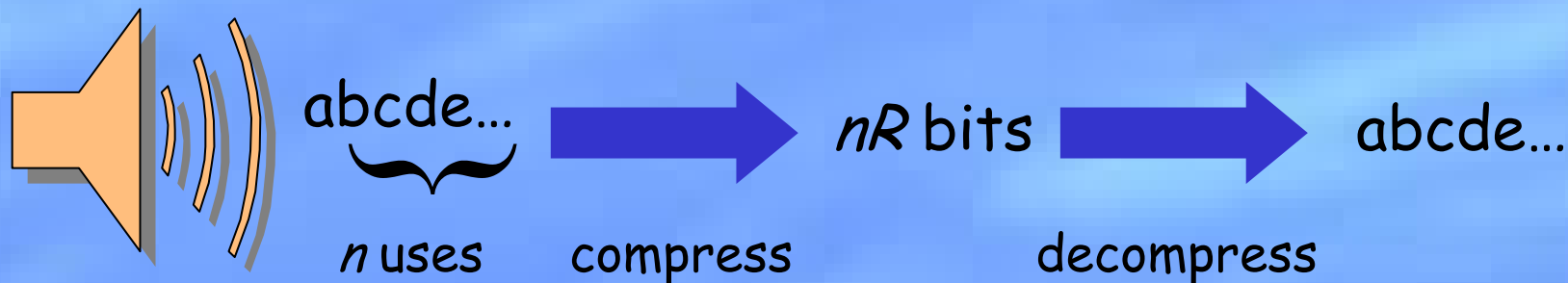
How many bits are needed to store the output of the source, so the output can be reliably recovered?

Historical origin of data compression



"He can compress the most words into the smallest ideas of any man I ever met."

Data compression



What is the minimal value of R that allows reliable decompression?

We will **define** the minimal value to be the information content of the source.

Shannon's noiseless channel coding theorem: The minimal achievable value of R is given by the **Shannon entropy** of the source distribution,

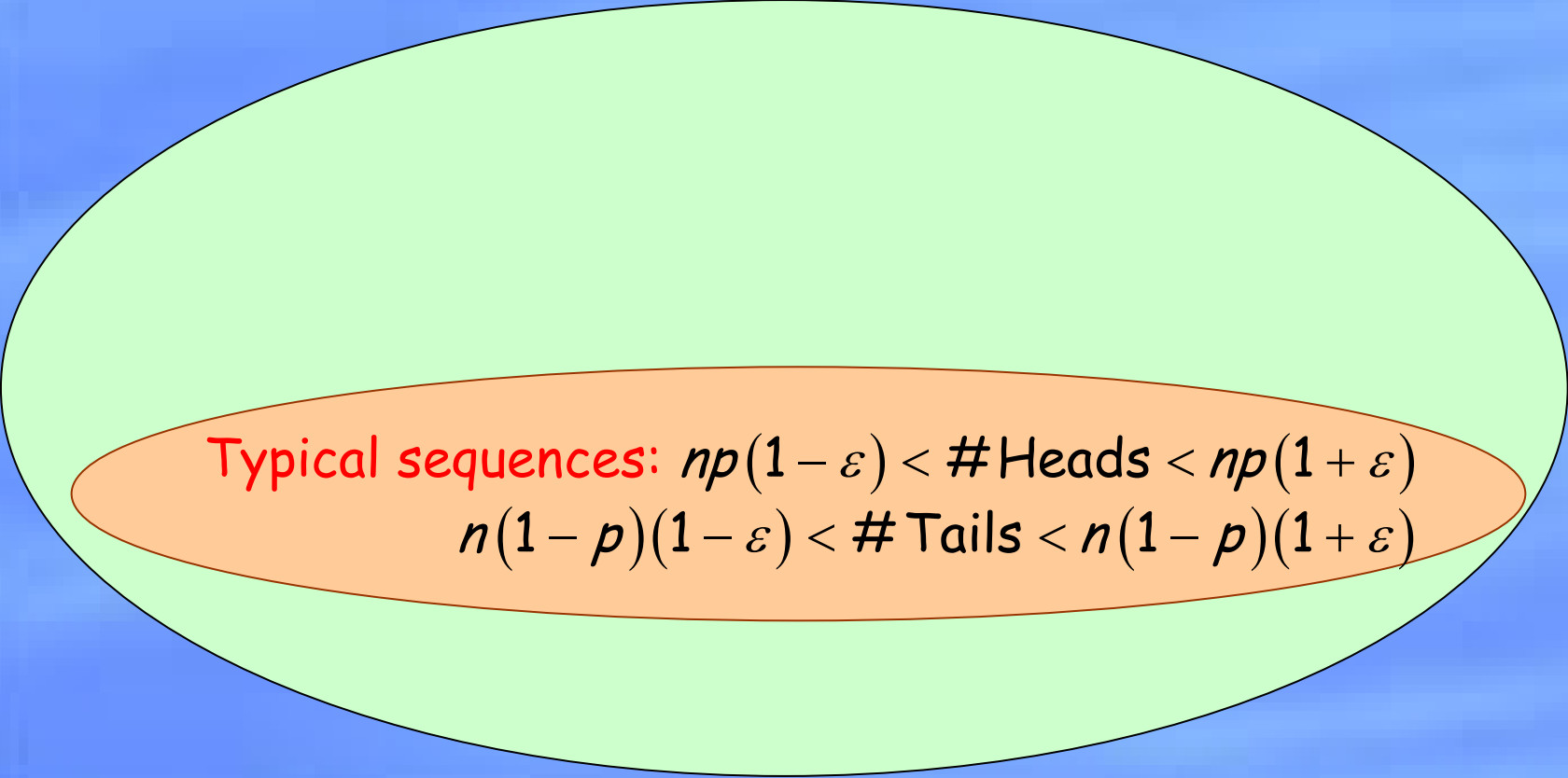
$$H(X) \equiv H(p_x) \equiv -\sum_x p_x \log(p_x),$$

where logarithms are taken to base two.

Data compression

Suppose we flip coins, getting heads with probability p , and tails with probability $1-p$.

For large values of n , it is very likely that we will get roughly np heads, and $n(1-p)$ tails.



Typical sequences: $np(1 - \varepsilon) < \# \text{Heads} < np(1 + \varepsilon)$
 $n(1 - p)(1 - \varepsilon) < \# \text{Tails} < n(1 - p)(1 + \varepsilon)$

Data compression

$$p^{np(1+\varepsilon)} (1-p)^{n(1-p)(1+\varepsilon)} < \Pr(x) < p^{np(1-\varepsilon)} (1-p)^{n(1-p)(1-\varepsilon)}$$

$$\Pr(x) \approx 2^{np \log(p) + n(1-p) \log(1-p)} \approx 2^{-nH(p, 1-p)}$$

$$\# \text{ Typical sequences} \approx 2^{nH(p, 1-p)}$$

Sequence is typical
with probability $\rightarrow 1$.

Atypical sequences

Typical sequences: $np(1-\varepsilon) < \# \text{ Heads} < np(1+\varepsilon)$
 $n(1-p)(1-\varepsilon) < \# \text{ Tails} < n(1-p)(1+\varepsilon)$

Data compression: the algorithm

The two critical facts

Sequence is typical with probability $\rightarrow 1$

Typical sequences $\approx 2^{nH(p,1-p)}$

1. x_1
2. x_2
3. x_3
4. x_4
- ...

In principle it is possible to construct a **lookup table** containing an **indexed list** of all $2^{nH(p,1-p)}$ typical sequences.

Let y be the source output

If y is atypical then

send the bit 0 and then the bit string y

$n+1$ bits

$nH(p,1-p)+1$ bits

else

send 1 and the index of y in the lookup table

On average, only $H(p,1-p)$ bits were required to store the compressed string, per use of the source.

Variants on the data compression algorithm

Our algorithm is for **large n** , gives **variable-length** output that achieves the Shannon entropy on **average**. The algorithm **never** makes an error in recovery.

Algorithms for **small n** can be designed that do almost as well.

Fixed-length compression

Let y be the source output

If y is atypical then

send $(nH(p, 1-p) + 1)$ 0's

else

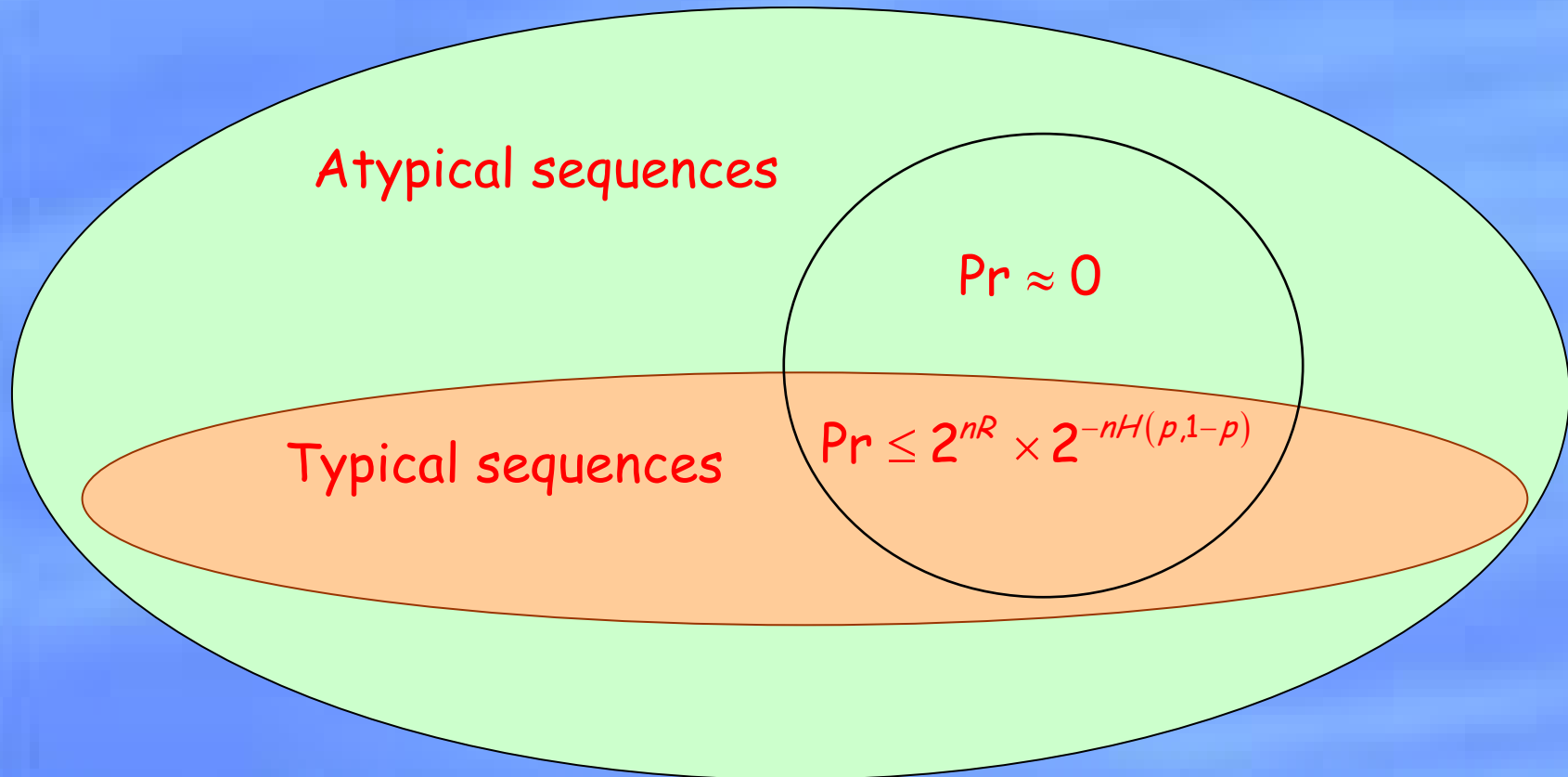
send 1 and the index of y in the lookup table

Errors must always occur in a fixed-length scheme, but it does work with probability approaching one.

Why it's impossible to compress below the Shannon rate

Suppose $R < H(p, 1-p)$ \longrightarrow $\Pr \leq 2^{n(R-H(p,1-p))} \rightarrow 0$

At most 2^{nR} sequences can be correctly compressed and then decompressed by a fixed-length scheme of rate R .



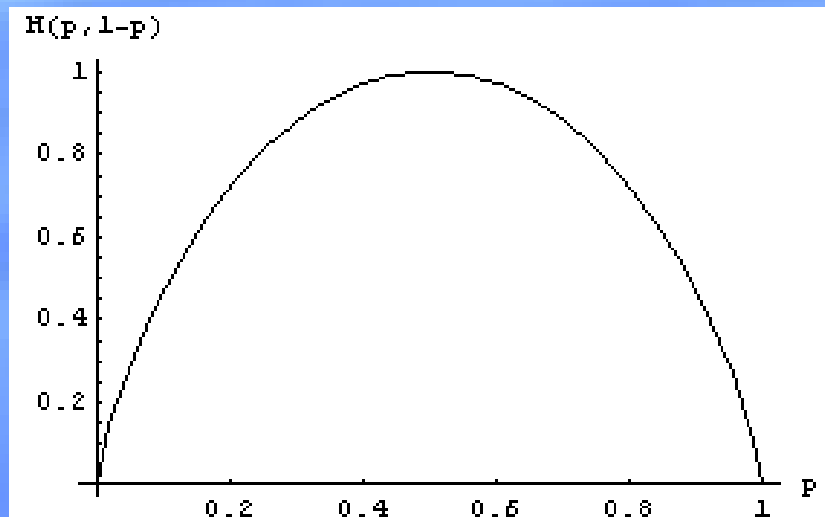
Basic properties of the entropy

$$H(X) \equiv H(p_x) \equiv -\sum_x p_x \log(p_x)$$

$$0 \log 0 \equiv 0$$

The entropy is non-negative and ranges between 0 and $\log(d)$.

$H(p) \equiv H(p, 1-p)$ is known as the **binary entropy**.



Why's this notion called entropy, anyway?

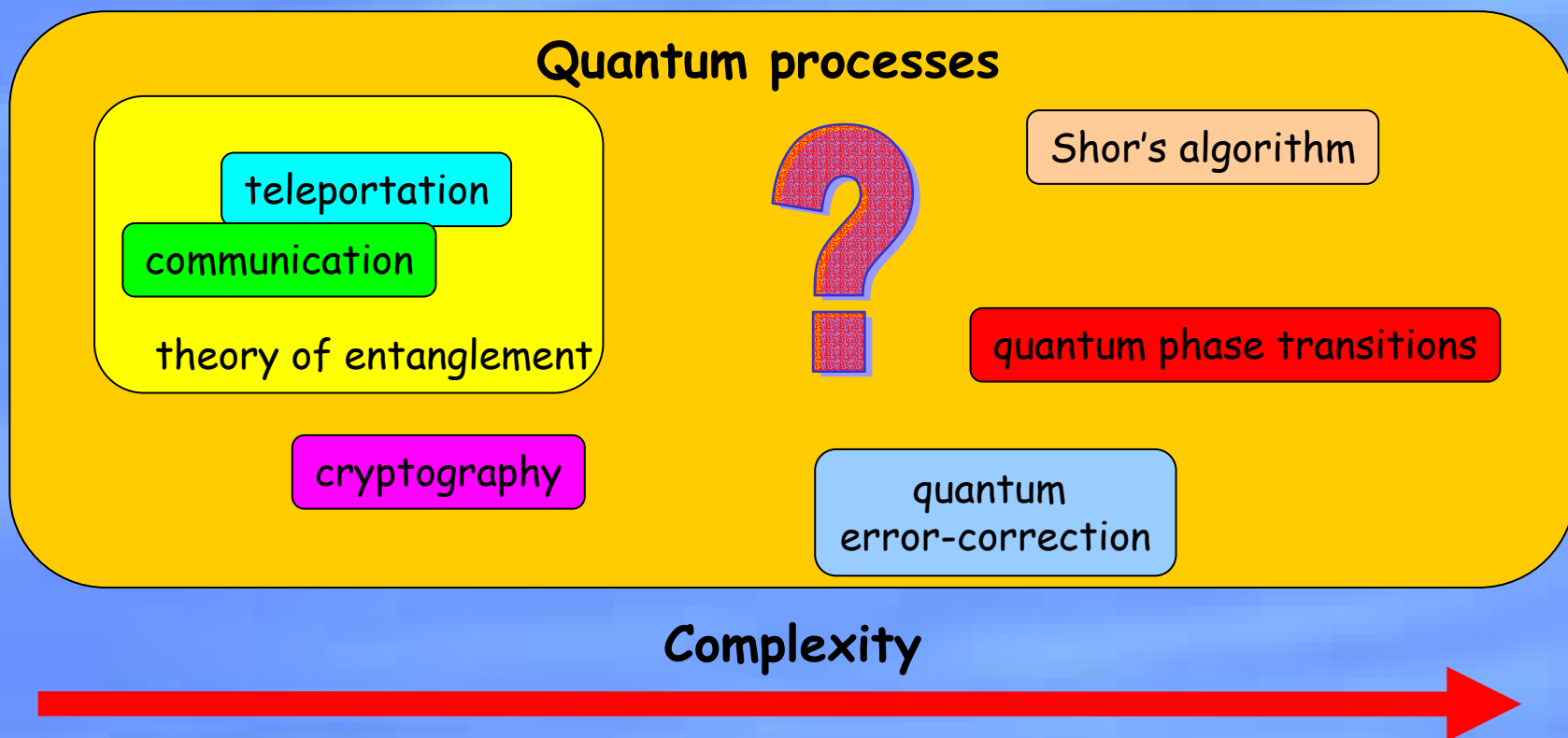
From the *American Heritage Book of English Usage* (1996):

"When the American scientist Claude Shannon found that the mathematical formula of Boltzmann defined a useful quantity in information theory, he hesitated to name this newly discovered quantity *entropy* because of its philosophical baggage. The mathematician John Von [sic] Neumann encouraged Shannon to go ahead with the name *entropy*, however, since 'no one knows what entropy is, so in a debate you will always have the advantage.' "

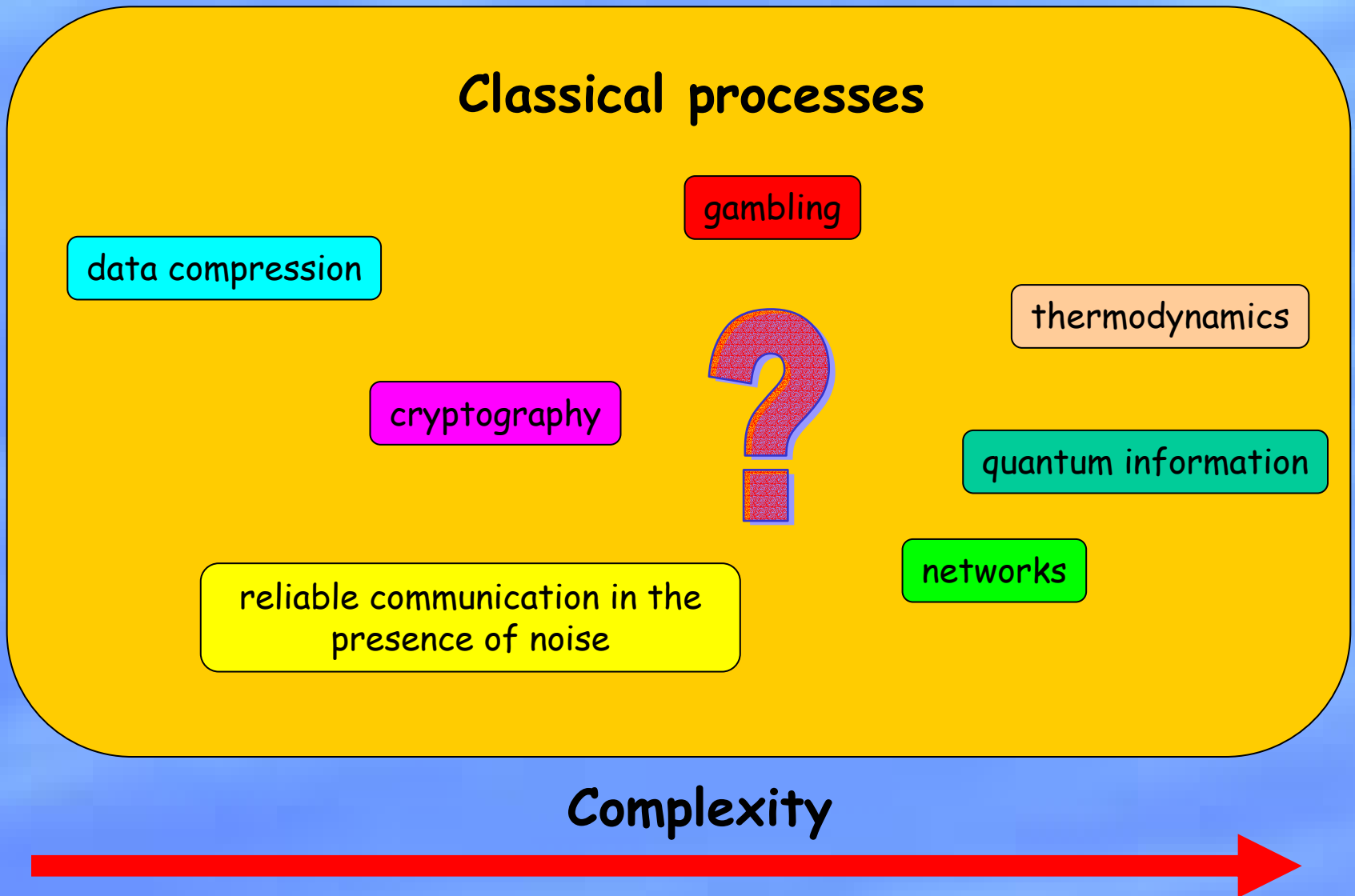
What else can be done with the Shannon entropy?

1. Identify a **physical resource** - energy, time, bits, space, entanglement.
2. Identify an **information processing task** - data compression, information transmission, teleportation.
3. Identify a **criterion for success**.

How much of 1 do I need to achieve 2, while satisfying 3?



What else can be done with the Shannon entropy?



What is a quantum information source?

Example: "Semiclassical coin toss"

$|0\rangle$ with probability $\frac{1}{2}$

$|1\rangle$ with probability $\frac{1}{2}$

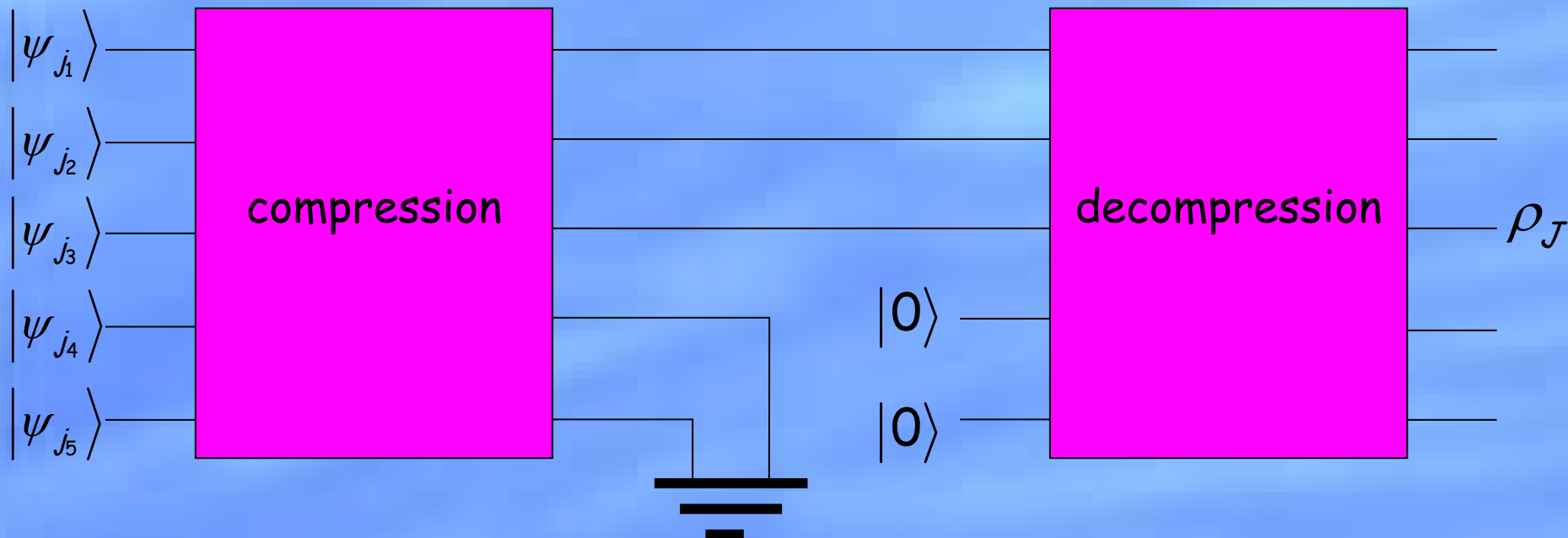
Example: "Quantum coin toss"

$|0\rangle$ with probability $\frac{1}{2}$

$\frac{|0\rangle + |1\rangle}{\sqrt{2}}$ with probability $\frac{1}{2}$

General definition: A quantum information source produces states $|\psi_j\rangle$ with probabilities p_j .

Quantum data compression



$$\mathcal{J} \equiv (j_1, \dots, j_n)$$

$$p_{\mathcal{J}} \equiv p_{j_1} \times \dots \times p_{j_n}$$

$$|\psi_{\mathcal{J}}\rangle \equiv |\psi_{j_1}\rangle \dots |\psi_{j_n}\rangle$$

$$\bar{F} \equiv \sum_{\mathcal{J}} p_{\mathcal{J}} F(|\psi_{\mathcal{J}}\rangle, \rho_{\mathcal{J}})$$

$$(\text{Recall that } F \equiv \sqrt{\langle \psi_{\mathcal{J}} | \rho_{\mathcal{J}} | \psi_{\mathcal{J}} \rangle}.)$$

$$\bar{F} \rightarrow 1$$

What's the best possible rate for quantum data compression?

"Semiclassical coin toss" $|0\rangle$ w. p. $\frac{1}{2}$, $|1\rangle$ w. p. $\frac{1}{2}$

Answer: $H\left(\frac{1}{2}\right) = 1.$

"Quantum coin toss" $|0\rangle$ w. p. $\frac{1}{2}$, $\frac{|0\rangle + |1\rangle}{\sqrt{2}}$ w. p. $\frac{1}{2}$

~~Answer: $H\left(\frac{1}{2}\right) = 1?$~~

Answer: $H\left(\frac{1 + 1/\sqrt{2}}{2}\right) \approx 0.6.$

In general, we can do **better** than Shannon's rate $H(p_j)$.

Quantum entropy

$$\rho \equiv \sum_j p_j |\psi_j\rangle\langle\psi_j|$$

Suppose ρ has diagonal representation

$$\rho \equiv \sum_k \lambda_k |e_k\rangle\langle e_k| \quad (\equiv -\text{tr}(\rho \log \rho)).$$

Define the **von Neumann entropy**, $S(\rho) \equiv H(\lambda_k) = -\sum_k \lambda_k \log \lambda_k$.

Shumacher's noiseless channel coding theorem: The minimal achievable value of the rate R is $S(\rho)$.

Basic properties of the von Neumann entropy

$S(\rho) \equiv H(\lambda_k)$, where λ_k are the eigenvalues of ρ .

$$0 \leq S(\rho) \leq \log(d)$$



ρ_A



ρ_B



ρ_{AB}



Exercise: Show that

$$S(\rho_A \otimes \rho_B) = S(\rho_A) + S(\rho_B).$$

Subadditivity:

$$S(\rho_{AB}) \leq S(\rho_A) + S(\rho_B).$$

The typical subspace

Example: $\rho = p|0\rangle\langle 0| + (1-p)|1\rangle\langle 1|$, $S(\rho) = H(p)$.



A diagram consisting of two nested ellipses. The outer ellipse is light green and contains the text 'Atypical sequences'. The inner ellipse is light orange and contains the text 'Typical sequences: $x_1, \dots, x_{2^{nS(\rho)}}$ '. The ellipses are centered horizontally and vertically on the slide.

Atypical sequences

Typical sequences: $x_1, \dots, x_{2^{nS(\rho)}}$

Typical subspace: spanned by $|x_1\rangle, \dots, |x_{2^{nS(\rho)}}\rangle$, $P = \sum_j |x_j\rangle\langle x_j|$.

Outline of Schumacher's data compression



Measure $|\psi_j\rangle$ to determine whether we're in the typical subspace or not.
 $P, Q = I - P$

P

Q

Unitarily transform
 $|x_j\rangle \rightarrow |j\rangle|0\rangle\dots|0\rangle$

Send $|0\rangle^{\otimes nS(\rho)}$.



Send $|j\rangle$.

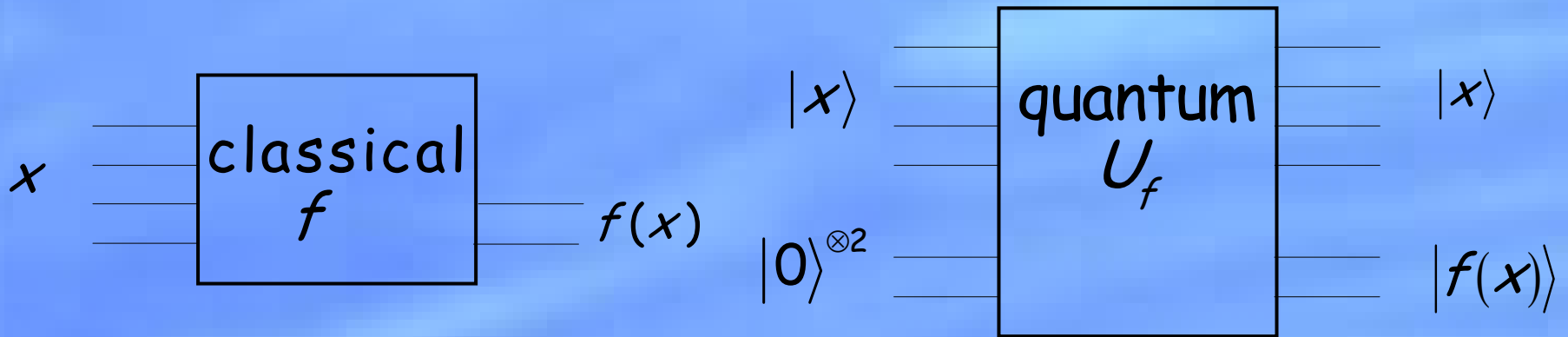
Append $|0\rangle$'s: $|j\rangle|0\rangle\dots|0\rangle$.

Inverse transform $|j\rangle|0\rangle\dots|0\rangle \rightarrow |x_j\rangle$.

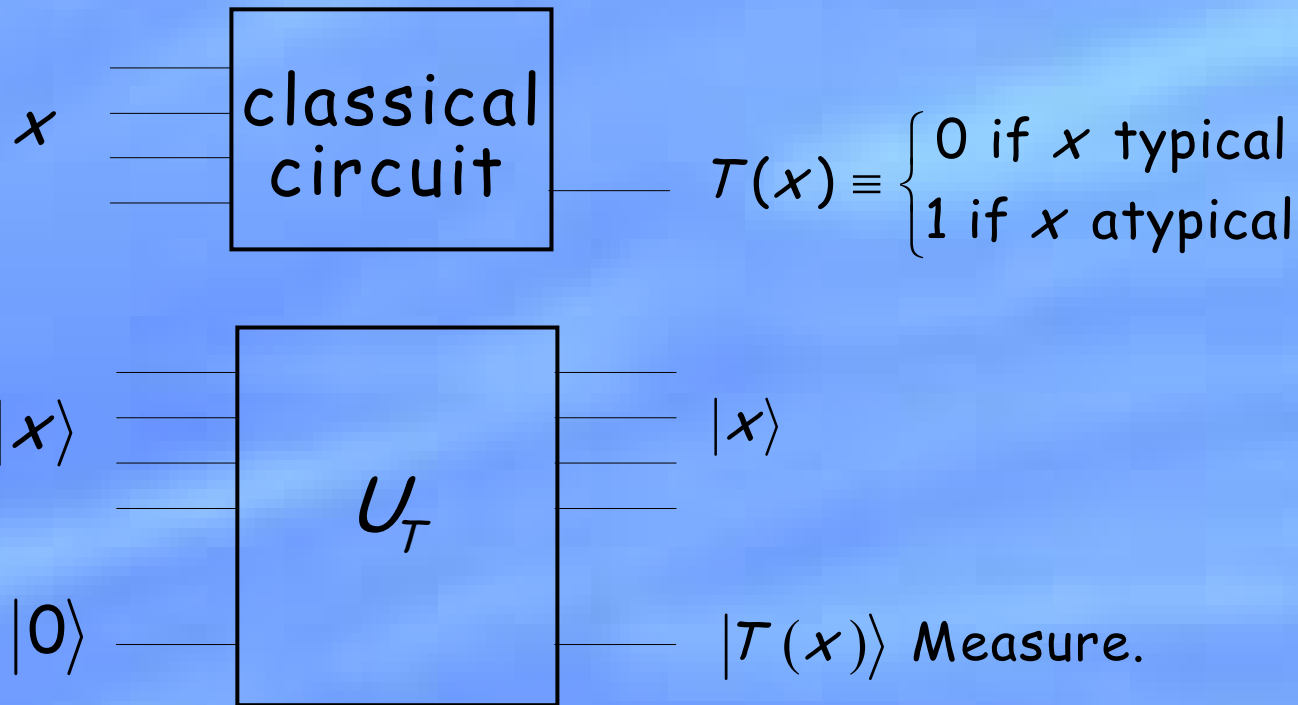
Claim $\bar{F} \rightarrow 1$.



Recall classical to quantum circuits



How to measure P , Q



Exercise: Verify that the effect on the first register is

$$|\psi\rangle \rightarrow \frac{P|\psi\rangle}{\sqrt{\langle\psi|P|\psi\rangle}} \text{ with probability } \langle\psi|P|\psi\rangle$$

$$|\psi\rangle \rightarrow \frac{Q|\psi\rangle}{\sqrt{\langle\psi|Q|\psi\rangle}} \text{ with probability } \langle\psi|Q|\psi\rangle$$

Outline of Schumacher's data compression



Measure $|\psi_j\rangle$ to determine whether we're in the typical subspace or not.
 $P, Q = I - P$

P

Q

Unitarily transform
 $|x_j\rangle \rightarrow |j\rangle|0\rangle\dots|0\rangle$

Send $|0\rangle^{\otimes nS(\rho)}$.



Send $|j\rangle$.

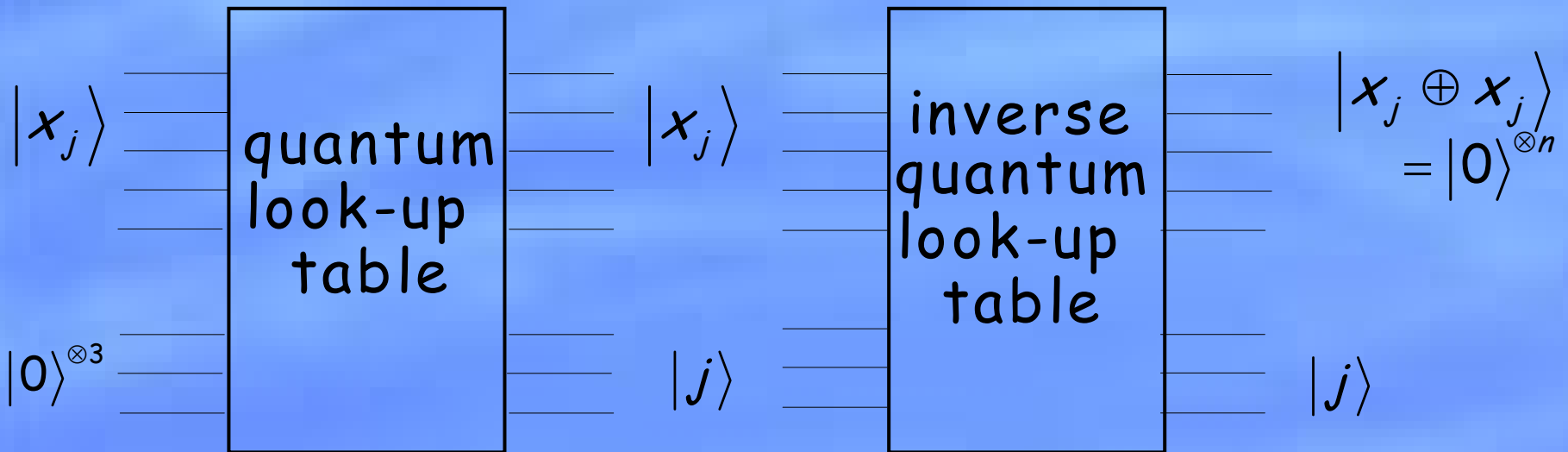
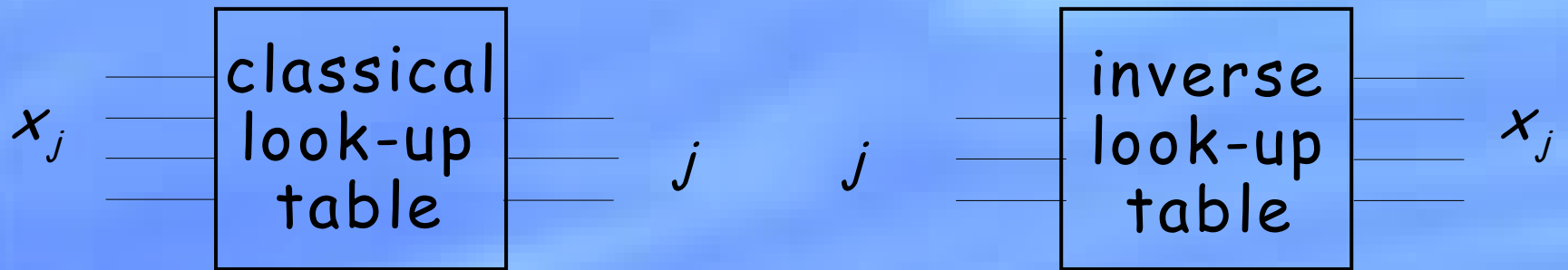
Append $|0\rangle$'s: $|j\rangle|0\rangle\dots|0\rangle$.

Inverse transform $|j\rangle|0\rangle\dots|0\rangle \rightarrow |x_j\rangle$.

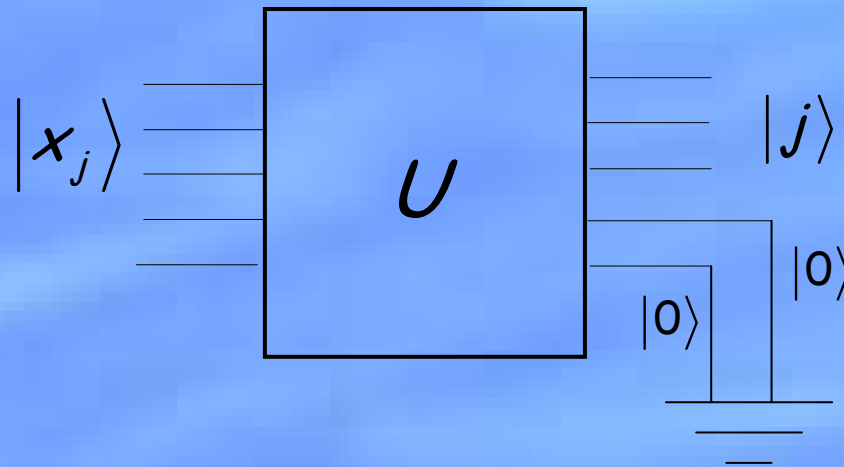
Claim $\bar{F} \rightarrow 1$.



How to unitarily transform $|x_j\rangle \rightarrow |j\rangle|0\rangle\dots|0\rangle$



How to unitarily transform $|x_j\rangle \rightarrow |j\rangle|0\rangle\dots|0\rangle$



Outline of Schumacher's data compression



Measure $|\psi_j\rangle$ to determine whether we're in the typical subspace or not.
 $P, Q = I - P$

P

Q

Unitarily transform
 $|x_j\rangle \rightarrow |j\rangle|0\rangle\dots|0\rangle$

Send $|0\rangle^{\otimes nS(\rho)}$.



Send $|j\rangle$.

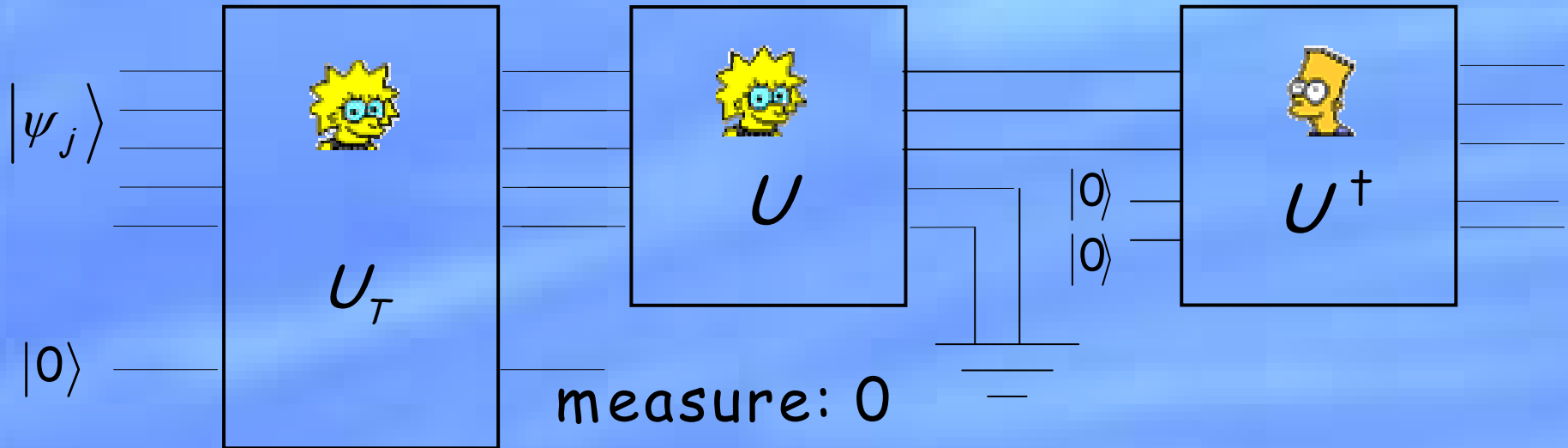
Append $|0\rangle$'s: $|j\rangle|0\rangle\dots|0\rangle$.

Inverse transform $|j\rangle|0\rangle\dots|0\rangle \rightarrow |x_j\rangle$.

Claim $\bar{F} \rightarrow 1$.



Schumacher compression



$$\frac{P|\psi_J\rangle}{\sqrt{\langle\psi_J|P|\psi_J\rangle}} \quad \text{with probability } \langle\psi_J|P|\psi_J\rangle$$

Fidelity for $|\psi_J\rangle$

$$F_J = F(|\psi_J\rangle, \rho_J) = \sqrt{\langle \psi_J | \rho_J | \psi_J \rangle}$$

Ensemble for ρ_J : $\frac{P|\psi_J\rangle}{\sqrt{\langle \psi_J | P | \psi_J \rangle}}$ with probability $\langle \psi_J | P | \psi_J \rangle$

$|\text{junk}\rangle$ with probability $\langle \psi_J | Q | \psi_J \rangle$

$$\begin{aligned} \rho_J &= \langle \psi_J | P | \psi_J \rangle \frac{P|\psi_J\rangle\langle \psi_J | P}{\langle \psi_J | P | \psi_J \rangle} + \langle \psi_J | Q | \psi_J \rangle |\text{junk}\rangle\langle \text{junk}| \\ &= P|\psi_J\rangle\langle \psi_J | P + \langle \psi_J | Q | \psi_J \rangle |\text{junk}\rangle\langle \text{junk}| \end{aligned}$$

$$\begin{aligned} F_J &\geq \sqrt{\langle \psi_J | P | \psi_J \rangle \langle \psi_J | P | \psi_J \rangle} \\ &= \langle \psi_J | P | \psi_J \rangle \end{aligned}$$

Reliability : $\bar{F} \rightarrow 1$

$$\begin{aligned}\bar{F} &= \sum_J p_J F(|\psi_J\rangle, \rho_J) \geq \sum_J p_J \langle \psi_J | P | \psi_J \rangle \\ &= \sum_J p_J \text{tr}(P |\psi_J\rangle \langle \psi_J|)\end{aligned}$$

$$\text{But } \sum_J p_J |\psi_J\rangle \langle \psi_J| = \overbrace{\rho \otimes \dots \otimes \rho}^{n \text{ times}}.$$

$$\begin{aligned}\bar{F} &\geq \text{tr}(P \rho^{\otimes n}) = \sum_{x \text{ typical}, y} p_y \text{tr}(|x\rangle \langle x| y\rangle \langle y|) \\ &= \sum_{x \text{ typical}, y} p_y \delta_{xy} = \sum_{x \text{ typical}} p_x \rightarrow 1\end{aligned}$$

$$\rho = p_0 |0\rangle \langle 0| + p_1 |1\rangle \langle 1|; \quad p_0 \equiv p; \quad p_1 \equiv 1 - p.$$

$$\rho^{\otimes n} = \sum_y p_y |y\rangle \langle y| \quad P = \sum_{x \text{ typical}} |x\rangle \langle x|$$

Proof that it's impossible to compress to a rate below $S(\rho)$

The idea of the proof is similar to Shannon's proof.

Two known proofs:

One is a complicated kludge, done from first principles.

The other proof is an elegant "easy" proof that relies on other deep theorems.

Research problem: Find an easy first-principles proof that $S(\rho)$ is the best achievable rate.

Worked exercise: Prove that the von Neumann entropy satisfies the inequality

$$S\left(\sum_j p_j |\psi_j\rangle\langle\psi_j|\right) \leq H(p_j).$$

(**Hint:** You may find it useful to use the Schumacher and Shannon noiseless channel coding theorems.)

Exercise: Prove that $S\left(\sum_j p_j \rho_j\right) \leq H(p_j) + \sum_j p_j S(\rho_j)$.

Research problem(?): Find a low-rate quantum data compression scheme that, with probability **exactly one**, produces decompressed states with fidelity to the source output approaching one.

Research problem - mixed-state quantum data compression: Suppose a source outputs ρ_j with probability p_j . What is the best achievable rate of compression for such a source?